Casualty Actuarial Society E-Forum, Spring 2013

The CAS *E-Forum*, Spring 2013

The Spring 2013 Edition of the CAS *E-Forum* is a cooperative effort between the CAS *E-Forum* Committee and various other CAS committees, task forces, or working parties.

The CAS Committee on Ratemaking presents for discussion 10 papers prepared in response to the 2013 call for ratemaking papers. Some of the Ratemaking Call Papers will be discussed by the authors at the 2013 CAS Ratemaking and Product Management Seminar on March 11-13, 2013, in Huntington Beach, CA.

Committee on Ratemaking

CAS *E-Forum,* Spring 2013

Table of Contents

2013 Ratemaking Call Papers

For information on submitting a paper to the *E-Forum*, visit
[http://www.casact.org/pubs/forum/.](http://www.casact.org/pubs/forum/)

Beyond the Cost Model: Understanding Price Elasticity and Its Applications

Serhat Guven, FCAS, MAAA, and Michael McPhail, FCAS, MAAA

Abstract

Once cost models have been constructed, insurers spend a significant amount of time translating those expected cost models into a rating algorithm. Today, competitive analytics are widely used to support this effort. However, companies often fail to fully integrate competitive analytics into the pricing process. The intent of this paper is to provide the basic tools needed for insurers to make more effective pricing decisions using customer price elasticity of demand. To achieve this, we will explore demand modeling techniques, as well as practical applications of demand modeling in pricing.

Keywords. Price elasticity; demand; generalized linear modeling; price simulation; price optimization

1. INTRODUCTION

We learned in our introductory economics courses that price elasticity of demand (PED) is loosely defined as the change in demand for a given change in price. It measures a consumer's sensitivity to a change in price for a given good or service, ranging from high sensitivity (elastic) to low sensitivity (inelastic). In equation form,

$$E = -\frac{(\mu_{P1}-\mu_{P0})/\mu_{P0}}{(P1-P0)/P0} \tag{1.1}$$

Where E = PED

$P0$ = Initial Price

$P1$ = New Price

μ_{P0} = Demand at initial price

μ_{P1} = Demand at new price

One of the benefits of PED is that it enables a firm to enhance its pricing strategies through a better understanding of the price sensitivity of its target market. Let's look at a basic example. Suppose an accountant has been operating a private firm for the past five years and has built a base of loyal customers. Although sales are stable, growth is stagnant and he is becoming concerned with the long-term viability of the firm. He decides to implement a 20% discount on first-time tax returns to bring in new business. As an accountant, he understands the need to maintain profitability and simultaneously raises the price of a tax return from \$250 to \$275. Knowing that his current customers are extremely loyal, he believes the majority of them will accept the slight increase rather than face the risks associated with switching accountants (i.e. lower quality). In other words, he is altering his pricing strategy based on the price elasticity of his target market. For new business, which tends to be more price sensitive, he is lowering the price of first-time tax returns to $220 ($275 x 80%). He offsets the discount by raising the price for his existing business, which tends to be less price sensitive. Through this pricing strategy, he expects to increase his customer base and revenue in the long term. In reality, there are a number of factors that must be considered when altering pricing strategy, but hopefully this example provides some insight into how PED can be utilized.

Another benefit of PED is the accurate measurement of revenue changes in response to pricing changes. In the above example, the accountant increased the price of tax returns for current customers by 10% (275/250 - 1). Although a majority of customers will accept the increase, some customers are expected to switch accountants. Ignoring growth, consider the following:

(1) Current Price	(2) Proposed Price	(3) # of Current Customers	(4) # of Customers Expected to Leave Due to Rate Change	(5) = (1) * (3) Expected Revenue Under Current Rates	(6) = (2) * ((3) – (4)) Expected Revenue Under Proposed Rates	(7) = (6)/(5) - 1 Expected % Change in Revenue
$250	$275	100	5	$25,000	$26,125	4.5%

Table 1.1 Price change vs. revenue change

Although the accountant is still profitable, the expected percentage change in revenue is less than half of the proposed rate change. But what if the number of customers expected to leave were ten rather than five? In that case, the expected percentage change in revenue would actually be negative, resulting in an overall loss to the business. Understanding the effects of demand on revenue is crucial to making pricing decisions that improve the overall position of the firm.

As an actuary in the P&C insurance market, PED opens up a whole new realm of possibilities. Although competitive analytics are widely used in the industry today, many companies fail to fully integrate them into the pricing process. For a specific firm, what does it mean to be competitive? Is it purely a function of price or should more abstract concepts such as brand and loyalty also be considered? How can a firm systematically incorporate that information into its pricing strategy? The rest of this paper will attempt to answer these questions, with a focus on basic techniques that can be used to model demand. In Section 2, we discuss demand modeling including model form and model structure. In Sections 3 & 4, we explore practical applications of PED in the context of price simulation and price optimization. Finally, in Section 5, we provide our concluding remarks.

Before moving into demand modeling, the authors would like to provide a clarification on the intent of this paper. Due to minimal actuarial literature regarding PED, this paper strives to offer an introductory review of the topic as opposed to an actual practitioner's guide. As such, concepts deserving of numerous pages receive little treatment for the sake of brevity. Our hope is that this paper provides a springboard for further work on PED.

2. DEMAND MODELING

In statistical models, there are two basic types of data variables. The fitted response variable is the output of the model whereas predictor variables are inputs to the model. Given a set of predictors, the model will provide a fitted response. The actual response in the data has two elements. The signal represents the predictive behavior and the noise is the random behavior. The goal of the model is to separate the signal from the noise. Thus, we are trying to identify a function that, when applied to the predictors, produces a fitted response which represents the signal in the actual response.

When modeling the price elasticity of demand, the first task is to clearly define the actual response. In addition, it is useful to classify the different types of predictors that are to be studied. There are significant challenges in clarifying the data for elasticity modeling. For example, many companies wrestle with the definition of a quote – questions such as how long a quote can stay open before it is considered a rejection, how do you deal with multiple changes to endorsements, among many other such issues. These issues are outside of the scope of this paper.

Price elasticity modeling builds a model in which the actual response is the individual customer's acceptance or rejection of a quote or renewal offering. The predictors in these models can be classified as price and non price related. The fitted elasticity is derived from the coefficients that are associated with the price related predictors in the model. Note it can be useful to further categorize price factors as external (reflecting pricing activities of competitors) and internal (reflecting pricing activities of the company).

In the following sections, we will be discussing the structural form of a model where, given the actual response and predictors, we will be able to calculate fitted responses and derive fitted elasticity. In discussing the model form, we will focus on the following three areas:

- Error distribution of the actual response
- Functional form that relates the actual response to the predictors
- Structural design of the predictors

Furthermore, we will discuss key validation techniques to assess the quality of the projections. The result of the form and modeling exercise is to improve understanding of a consumer's behavior toward price.

2.1. Error Distribution

Unlike many other responses that practitioners model, it is a relatively straight forward exercise to select the error distribution function of a price elasticity model. Since the actual response is of the form yes/no, the binomial distribution is the most appropriate choice. Note that it is possible to have intermediate phases that are associated with different degrees of acceptance or rejection, but that is outside the scope of this article.

Before continuing, it is useful to understand what it means to have a binomial distribution. Error distributions are defined in empirical terms as well as functional forms. Empirically, the actual response can only be 0 or 1, meaning the selection of a binomial structure is clearly supported. The functional form of an error distribution relates the variance of the actual response to the fitted response. Another interesting interpretation is to understand how the variance function plays a part in model form. The relationship of the variance to the mean in a binomial distribution assumes the following form:

Variance = Mean x (1 – Mean)

Because a binary response is modeled, the fitted response will lie between 0 and 1 (more on this later). Given that constraint, the following picture shows the relationship between the mean and the variance:

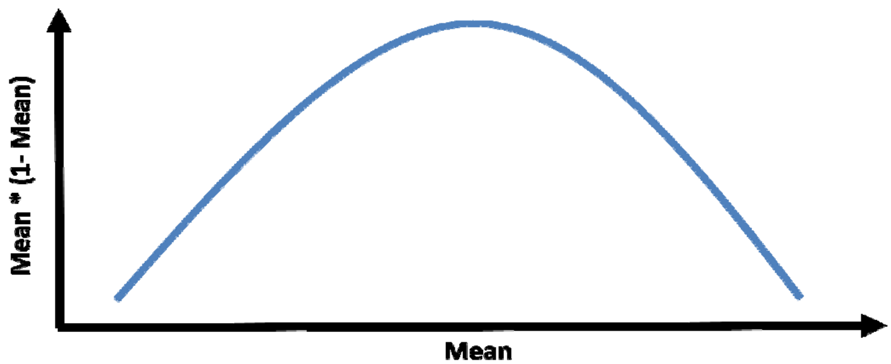


Figure 2.1 Variance mean relationship of binomial distribution

The variance of a result represents the degree of uncertainty associated with the predictions of that result. In a cost model, it is common to assume that the larger the fitted value, the greater the uncertainty (variability) of the estimate. In elasticity models, both large and small estimates have lower variability whereas mid-range estimates have higher variability. This is the inherent nature of binomial models. If all the records are always rejecting offers (actual response is low - zero), then the modeler can feel fairly confident that future offers will be rejected. Similarly, if all records are always accepting offers, (actual response is high - one) then the same high level of confidence can be said about forecasted results. Now if the actual response is mid range, i.e. 50%, then there would be much greater uncertainty associated with the fitted value. Once the distribution function has been selected, the next key structure is the link function.

2.2. Link Function

The link function is the functional form that relates the response to the predictors. There are two key requirements when selecting a link function for elasticity modeling. The first is that the fitted response of the probability of acceptance or rejection of a price offer should lie between 0 and 1. The requirement of the link function is to transform the resulting coefficients from the structural design into a result that is consistent with the laws of probability. Note that in common statistical terminology, the combination of coefficients from the structural design of the model is called a linear predictor. It is not necessary for the linear predictor to be purely linear, as will be discussed in following sections.

The second requirement, though less stringent, is that the fitted responses approach pure acceptance or rejection but never actually reach it. This asymptotic behavior is best visualized as the following S-shape curve:

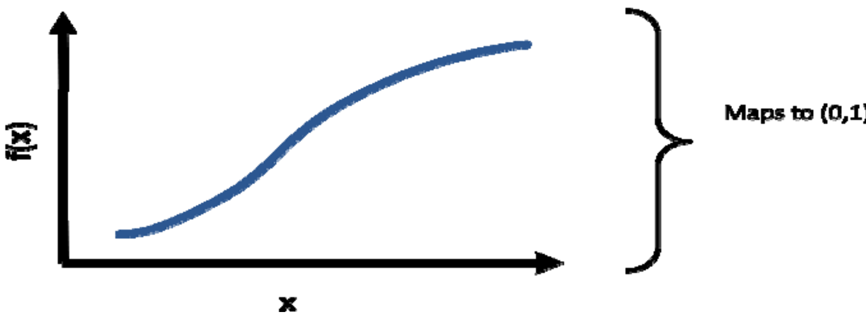


Figure 2.2 Shape of modeled response for elasticity modeling

Allowing the fitted response to reach pure acceptance or rejection implies too much certainty in the forecast.

Unlike error distributions, there are many choices of link functions that satisfy these requirements. Two functions commonly used in practice are:

1. The logit or logistic function

$$\mu = f(x) = \frac{1}{1 + 1/\exp(x)} \tag{2.1}$$

   Where x is the linear predictor

2. The probit or normal function

$$\mu = f(x) = \Phi(x) \tag{2.2}$$

   Where Φ is the cumulative normal distribution

When selecting between different functions, it can be useful to perform validation tests using the elasticity concept. Recall that because we cannot observe individual elasticity, we have to derive it from the fitted responses from the model. There are many ways to define elasticity, and we will focus on the following two:

1. Classical Elasticity, which is the percentage change in demand due to the percentage change in price. As defined in the introduction, for a given customer,

$$E = -\frac{(\mu_{P1} - \mu_{P0})/\mu_{P0}}{(P1 - P0)/P0} \quad (2.3)$$

   Assume P1 > P0 so that the numerator is negative and the denominator is positive. From this formula, the initial expected demand is 'fixed' so elasticity has a linear relationship with demand. Note that μ_{Pi} is the expected probability of success associated with the price P_i.

2. Linear Predictor Elasticity, which we define as the change in the linear predictor due to the percentage change in price. Specifically, for a given customer:

$$E = -\frac{\left(\beta_0 + \propto \times \frac{P1}{P0}\right) - \left(\beta_0 + \propto \times \frac{P0}{P0}\right)}{(P1 - P0)/P0} \quad (2.4)$$

   Assuming a simple linear predictor with no nonlinear components and one price factor:

$$E = -\propto \quad (2.5)$$

   Where α is the coefficient associated with the price factor

   Note that in this simple case, E does not vary with demand. Thus, as the expected demand increases, we expect E to be constant. This relationship holds true regardless of the number of non-price factors in the model structure.

Comparing these two definitions with respect to demand yields the following two lines on this theoretical graph:

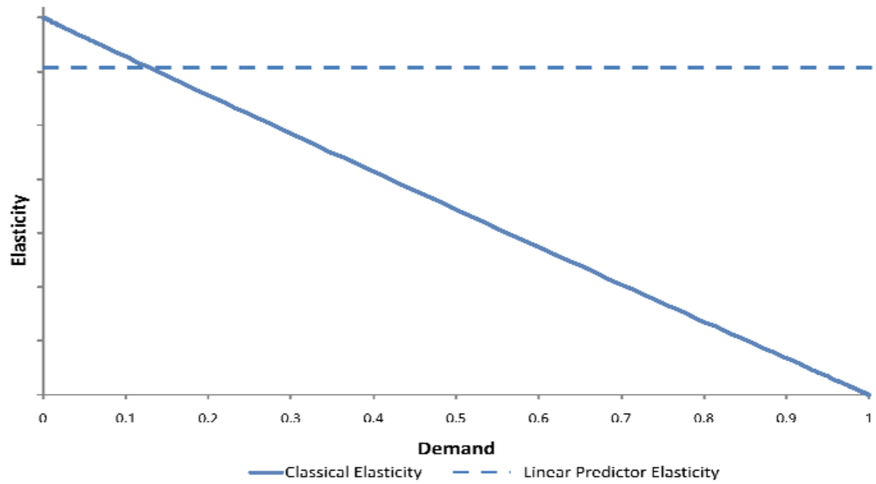


Figure 2.3 Classical and linear predictor elasticity vs. demand

As shown in Figure 2.3, the linear predictor elasticity (dashed) does not vary by demand whereas classical elasticity (solid) has a linear relationship with demand. While there is limited interpretative appeal of linear predictor elasticity, we developed this particular definition because it has useful properties when assessing link functions.

Consider the following validation exercise. A model was built using quote data from months one through six. During months seven and eight, the number of quotes and converted policies were captured. Also assume that a rate change took effect between months seven and eight. The model was used to score the expected close rates of the month seven and month eight quotes. Using these scores, the two months of data were ranked from highest to lowest and placed into 17 bins each having approximately the same number of quotes. For each bin, the average linear predictor elasticity was calculated as follows:

Month 7 Linear Predictor = $\ln(CR_7/(1-CR_7))$

Month 8 Linear Predictor = $\ln(CR_8/(1-CR_8))$

Where CR_i is the Close Rate in month i

Then, the change in the observed linear predictor divided by the change in average price between the two months was calculated. This yields the following graph where the model used a logit link function:

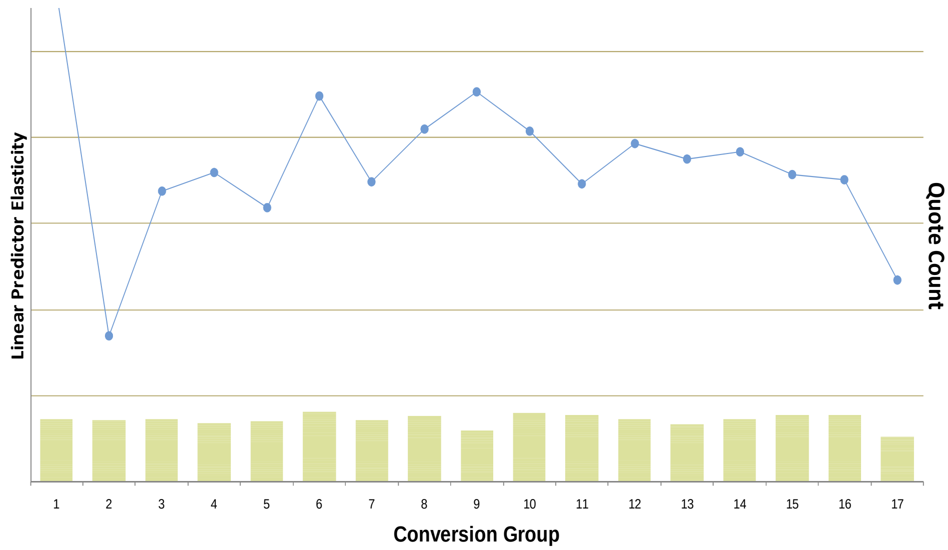


Figure 2.4 Linear predictor elasticity vs. demand for logit link function

In this graph, the x-axis is the expected conversion for the quotes in both months. This score was used to create approximately equal weighted buckets as identified by the bars from the right hand y-axis. On the left hand y-axis, the average linear predictor elasticity was calculated for each group as the change in the aggregate observed linear predictor divided by the change in average price between months seven and eight.

Building a similar graph with the same model structure using a probit link function produces the following:

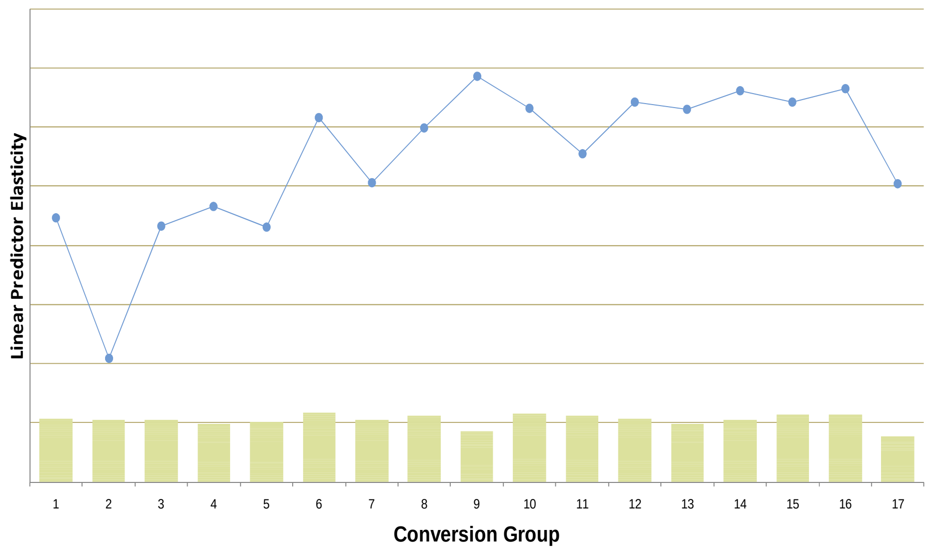


Figure 2.5 Linear predictor elasticity vs. demand for probit link function

For this particular data set, the actual linear predictor elasticity line in the logistic model is flatter than the actual linear predictor elasticity line in the probit model. From the prior theoretical discussion, the linear predictor elasticity would be expected to be flat when plotted against the expected demand. In this case, the logistic link function outperforms the probit link function as it results in a linear predictor elasticity that is more constant relative to demand.

As with any modeling exercise, the practitioner's choice should be validated to ensure the best possible model. Now that the distribution and link function have been selected, the next step is to identify the structural design of the predictors.

2.3. Model Structure

Building the model structure is a balance between over-fitting and under-fitting. When the model structure is too complex, there is a greater likelihood of over-fitting the data. This generally results in the loss of predictive power. However, when the structure is too simple, there is a likelihood of under-fitting the data, also resulting in the loss of explanatory power. When building the model structure, there are a variety of tests the analyst performs to ensure that each element in the model reflects this balance. These include:

- Statistical Tests - these include but are not limited to basic Type III tests, standard errors, etc.
- Consistency Tests – typically these test consistency across different time frames in the data. Alternatively, consistency can be tested across different random data splits.
- Judgment – the best models reflect the balance of technical information and knowledge of the underlying process.

The structural design of the model is often called the linear predictor. It reflects the combinations of parameters or coefficients that are applied to a particular observation. Note that the modeler specifies the structural design, the link function, and error distribution. The model then calculates a set of parameters that are based on these assumptions.

The linear predictor contains a variety of different objects to reflect the different predictors. In addition to categorizing predictors as price and non-price factors, they can also be classified as categorical and continuous elements. When a predictor is considered categorical, a parameter is calculated for each level in the predictor. When a predictor is considered continuous, a parameter is based on the form of the function introduced. The following example illustrates this concept. Without loss of generality, assume the following simple model structure:

$$\text{Base} + \text{RatingArea} + \text{NamedInsuredAge} + f_1(\text{RateChange}) + f_2(\text{CompetitivePosition})$$

In this case, the CompetitivePosition is defined as the MarketPrice/CompanyPrice, where the MarketPrice is a straight average of five competitors. This particular metric was useful in this model. There are many other competitive metrics that can be used in demand modeling (rank, cheapest ratios, etc.) and there are many challenges in obtaining this external data. These concepts are outside of the scope of this paper.

Looking back at the model form in the example, RatingArea and NamedInsuredAge are treated as categorical elements whereas RateChange and CompetitivePosition are continuous elements which are introduced as first degree polynomials. Consider the following output from a model:

Level	Predictor	Component
Base		-0.37876
NamedInsuredAge	25	0.39105
RatingArea	1A	0.07411
RateChange	1.09	-0.9235 * 1.09 + 1.0389
CompetitivePosition	1.07	-0.9414 * 1.07 + 1.1767
Linear Predictor		0.28809
Fitted Value		0.57153

Table 2.1 Model structure example; no interactions

The component represents the parameter applied to the predictor. For categorical factors, the parameter and component are the same. For continuous factors, the parameter is a coefficient applied to the predictor to yield the component. So the linear predictor for this particular observation is the sum of the components. These components are either specific parameters for categorical concepts OR functions of continuous concepts. The fitted value is then the inverse link function applied to the linear predictor. In the example, the model is using a logit link function as defined earlier.

The prior example reflects a model structure in which the predictors are all treated as main effects. This is a strictly linear concept in which the relationship between predictors does not vary based on other predictors. Adding main effects to the model should be done with great care to ensure the main effect is both statistically valid and the resulting pattern in the parameters is explainable. These results should be validated within the framework of the three classes of tests described previously.

To remove this constraint in the linear predictor, it is common to introduce interaction terms. An interaction term is a non-linear construct that allows the results of one predictor to become dependent on the value of another predictor. Building on the prior example, assume we have the following model structure:

$$\text{Base} + \text{RatingArea} + \text{NamedInsuredAge} + f_1(\text{RateChange}) + f_2(\text{CompetitivePosition}) + f_2(\text{CompetitivePosition.RatingArea})$$

For this model structure the curve parameters for the competitive position will vary based on rating area. Specifically:

Level	Predictor	Component
Base		-0.37781
NamedInsuredAge	25	0.38846
RatingArea	1A	0.09457
RateChange	1.09	-0.9216 * 1.09 + 1.0368
CompetitivePosition	1.07	-0.9234 * 1.07 + 1.1542
CompPosit.RatingArea	1A - 1.07	-0.4548 * 1.07 + 0.5685
Linear Predictor		0.38550
Fitted Value		0.59520

Table 2.2 Model structure example; interactions between competitive position and rating area

Similar to the discussion related to the main effect, adding interaction terms to the model should be done within the context of model testing. Identifying interactions are a great way to complicate the model structure to improve its explanatory power. This process is generally a non-trivial task. The various strategies for identifying interactions are outside the scope of this paper.

Recall the main purpose in building a demand model is to project the expected price sensitivity of the in-force customer or new business quote – in other words, the elasticity. Using the definitions from the prior section, we see that the elasticity is a slope function associated with the price factors in the model. Oftentimes, when building a model structure, we are looking for interactions (i.e. non-linear relationships) between the price and non-price factors. Within a traditional linear predictor, the analyst could build a model structure where all price factors interact with all non-price factors. This model would have a strong chance of over-fitting the data, and there is a possibility that a derived negative elasticity could occur from such a model.

Arbitrarily adding interactions is generally not a good idea. This could lead to over-fitting which would weaken the predictive power of the model. In addition, adding any type of complexity to a model should be validated based on statistical, consistency and judgment tests. As in other models, if a structure does not produce coherent results, then that structure should not be included in the model form. However, there are technical solutions available to build these complex interactions forms without resulting in the negative elasticity issue.

Localization

The first approach is to build localized models. Essentially, the analyst would use a decision tree tool to split the data into various subsets. The local main effects models will be built for each subset. By splitting the data into more homogenous subgroups, the modeler can build simpler models on smaller sets of data which would have the same effect of building complex models on larger sets of data. The localization procedure is easy to explain and implement; however, depending on the split, the model could produce discontinuities in some of the continuous factors. For example, if the split was based on rate increases versus rate decreases, then the resulting elasticity projections around the no rate change would not likely be continuous.

Customer Scoring

The second approach is to use a form of customer scoring to build the model structure. In this approach, a main effect model is built to include price and non-price factors. Interactions among the non-price factors are then incorporated as necessary. The parameters from the non-price factors are then combined into a customer score. This score is then interacted with the price factors to yield the necessary complexity. Clarifying this with an example, assume we have the following model form:

$$\text{Base} + \text{RatingArea} + \text{NamedInsuredAge} + f_1(\text{RateChange}) + f_2(\text{CompetitivePosition})$$

Using the parameters from RatingArea and NamedInsuredAge, the following model form is built:

$$\text{Base + Score} + f_1(\text{RateChange}) + f_2(\text{CompetitivePosition})$$

Where Score = RatingArea + NamedInsuredAge

Note at this point both of these models would produce the exact same fitted values. The final step is to introduce the interaction:

$$\text{Base + Score} + f_1(\text{RateChange}) + f_2(\text{CompetitivePosition}) + \text{Score}.f_1(\text{RateChange}) + \text{Score}.f_2(\text{CompetitivePosition})$$

The following picture illustrates this effect:

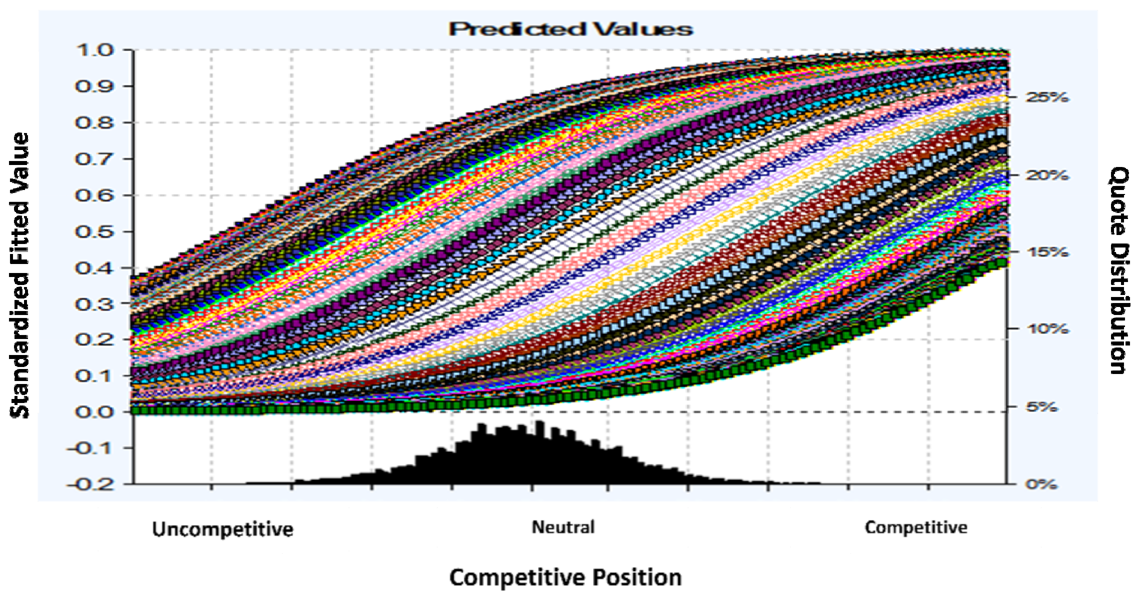


Figure 2.6 Interaction between customer score and competitive position

The x-axis represents the competitive position defined as MarketPrice/CompanyPrice. The MarketPrice is defined as a complex weighted average of a collection of competitors' prices. The left hand side of the x-axis represents less competitive rates and the right hand side represents more competitive rates. The left hand y-axis is the fitted value from the model and the right hand y-axis is the quote distribution. The score of the customer is derived from the non-price factors as described earlier. Sometimes this is labeled as the loyalty of a customer because it reflects the component of the expected demand that is not based on price. Each line represents a customer score where the higher the score the greater the 'loyalty' of the customer. When this customer score is interacted with the competitive price factor, we see that the steepest lines (i.e. most price sensitive) are evident in the middle of the tapestry. This makes sense because the most loyal customers (the flatter lines on the top end of the tapestry) are less price sensitive as they identify with either the brand or other services provided by the company. On the opposite end, the least loyal customers (the flatter lines on the bottom end of the tapestry) are less price sensitive as they do not identify with the company brand or do not seem to be interested in the additional services provided by the company. The middle of the spectrum represents customers who are more influenced by the pricing and less so by branding.

The customer scoring approach allows the modeler to introduce a significant level of complexity into the model form. There are interpretative challenges of what makes a more loyal customer for the non-price factors. In addition, the two staged approach implies an amount of quantitative certainty associated with parameters from the original main effects model.

Non-linear Modeling

The third approach defines the scoring parameters and the main effects parameters within the same fitting algorithm. The following nonlinear model form is constructed:

$$\text{Base} + \text{RatingArea} + \text{NamedInsuredAge} + f_1(\text{RateChange}) + f_2(\text{CompetitivePosition}) + \text{Score}.f_1(\text{RateChange}) + \text{Score}.f_2(\text{CompetitivePosition})$$

Where $\text{Score} = \exp(\text{RatingArea} + \text{NamedInsuredAge})$

Essentially, there are two sets of non-price coefficients that are being calculated within the model. The first represent the parameters to be applied to the main effects. The second set correspond to the embedded score function. This solution is different from the customer scoring approach because it is based on deriving the score parameters at the same time as the main effect parameters. Recall the customer scoring solution is a recursive process where the customer score is parameterized in the main effect. This score is then introduced in a subsequent model both as a main effect and an interaction term. The advantage of the non-linear model is to remove the element of recursion. Note that the more stable the main effects parameters, the less the difference between the prior two model forms.

As with the customer scoring approach, there are interpretive challenges to the non-linear structure. In addition, the specific structure can produce convergence issues in the fitting algorithm; however, this can be partly alleviated by simplifying the structure. Finally, this approach (along with the customer scoring strategy) ignores the potential for real negative elasticity. This possibility can exist when a company takes massive rate decreases resulting in cancellations. When customers see a large decrease, they tend to believe that they were previously and unfairly priced too high. This jeopardizes the customer/insurer relationship and causes many customers to shop and switch insurers.

2.4. Validation

The resulting models from these approaches need to be validated using out of time samples to test their performance. In earlier sections, validation techniques which were associated with specifying the correct link function were discussed. In this section, we will focus on two additional validation strategies associated with a validation data set.

The validation data set can either be a random hold out sample from the original data OR an out of time sample OR both. The value of an out of time sample is that it reflects an accurate measure of predictive accuracy. The models built on historical data are applied to subsequent data. What better way to assess the model performance than to use 'future' data. The disadvantage of an out of time sample is that there could be a significant change in the way business operates that is not reflective of the historical data (e.g. sudden increased presence of internet aggregators). So, the lack of predictive power is not a fault of the model, but rather due to the fact that the times are changing. A similar converse statement can be made about a random sample – it is a good measure of explanatory power and will not be skewed by distributional drifts in the trend; however, the measure provides a weak insight into how well the model will perform in the future. Ultimately, the modeler decides on variations of these two themes.

Once the validation data set has been defined, the model structure and resulting parameters that were built from the training and testing data can be used to score the validation records. These scores can then be compared to the actual values to assess model performance. Consider the following example:

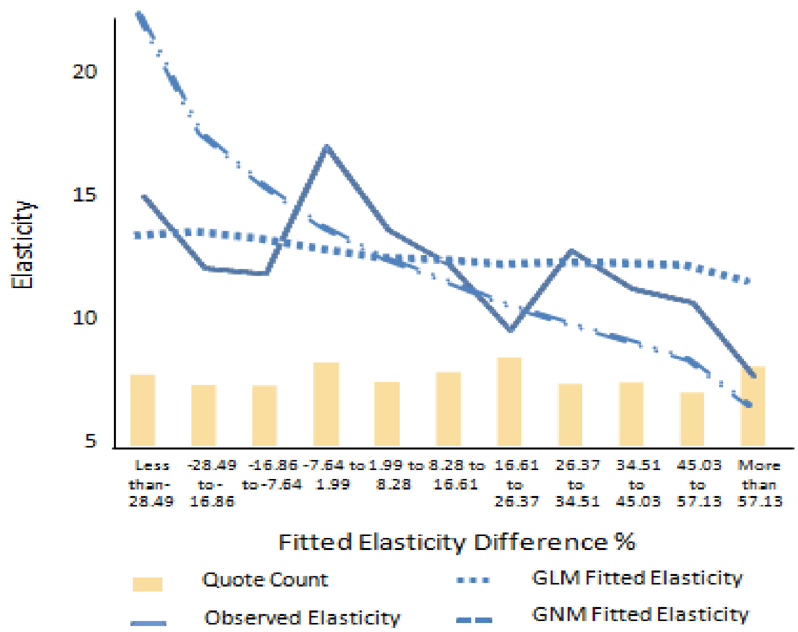


Figure 2.7 Validation of elasticity models against observed elasticity

In this example, two models were compared to actual results based on an out of time validation data set in much the same way that was described in the previous section. As before, the out of time data is the number of quotes from months seven and eight. The first model (dotted) was built using a standard logistic regression within a GLM framework. The model form was localized and several interactions were introduced into the localized components. The second model (dashed) was built using the non-linear method described in the prior section. Applying the structure and parameters to the out of time validation data set, a modeled elasticity score was generated from both models. The quote data was then ranked into approximately 11 equal weighted buckets based on the ratio of the score from both models. Specifically:

(GLMFittedElasticty/GNMFittedElasticity – 1) * 100

This ratio represents the x-axis in Figure 2.7. For each of these buckets, the aggregate modeled elasticities were calculated as described earlier (the change in close rate divided by the change in price). The dashed line and the dotted line intersect when both models produce the same score. To the left of the intersection, the dashed line is higher than the dotted. This represents the set of observations in which the GNM produces a higher expected elasticity. To the right of the intersection, the dashed line is less than the dotted which represents when the GNM approach produces a lower expected elasticity.

Finally, the observed elasticities (solid line) for each of these buckets were plotted. Comparing these different lines allow the modeler to assess which structural form is doing a better job of predicting the elasticity. In this particular case, it appears as if the non-linear model structure is over-fitting the historical data. Thus, it has weaker predictive power. This result is more apparent when the GNM predicts a higher elasticity than the GLM.

In the prior example, the model structure and resulting parameters were validated on an out of time sample. From this approach, we can make general statements about the appropriateness of the model form; however, we are unable to make specific statements about the strengths and weaknesses of particular aspects of the model structure. For example, how predictive is the set of parameters coming from a particular interaction? An approach to deal with this problem is to use the validation data to reparameterize the model structure that was built from the historical data. The modeler can then compare the parameters calculated from the training and testing data to the parameters calculated from the validation data. The inconsistencies between the sets of the parameters indicate where there are weaknesses in the predictive power of certain aspects of the model structure.

Modeling the demand and deriving elasticity provides a powerful understanding of the customer's reaction to the presented price. This understanding is crucial in setting the right rate for the risk. These models require different technical expertise than those of standard costs models, as was described by their structure and form. Ultimately, when building a model it is imperative to realize that, regardless of strategy, the best models tend to be a mixture of technical expertise combined with business acumen.

3. PRICE SIMULATION

In the following two sections, we will look at practical applications of PED, beginning with price simulation. With the help of our demand models, price simulation allows us to "predict" how our book of business will change over time with respect to key metrics such as profitability and demand. To gain an understanding of how price simulation can be utilized in the pricing process, the remainder of this section will focus on a specific example.

3.1. Scenario Testing

During a rate review, it is not uncommon for an insurance company to have a few different scenarios that it would like to test. Although the company may have a high-level idea in mind (such as a 5% rate decrease), there are a number of ways that 5% decrease can be allocated. Price simulation allows us to determine which scenario will help us best achieve our goals. For example, suppose an insurer writes personal auto insurance. Based on initial countrywide planning efforts, the insurer has decided to pursue a 5% rate decrease in state X. In addition, it would like to simulate two price scenarios to help determine which one should be implemented. The scenarios are as follows:

- Scenario 1 – 5% base rate decrease
- Scenario 2 – 15% decrease for operators aged 25-30 off-balanced to an overall 5% decrease

Once the scenarios have been determined, the assumptions that feed the simulator must be developed. For this example, assume the following:

- Datasets
  - Quotes
  - Renewals
- Conversion Model
  - Price Component - Premium
- Retention model
  - Price Component – Market Competitiveness (fixed)
- Time Horizon – Four periods, each lasting six months
- Quote Growth Rate – 5% each period
- Quotes do not enter simulation until new rates go into effect at the beginning of period 1
- Quote distribution constant over time (i.e. if 21 year olds represent 15% of the quotes in period 0, they will represent 15% in subsequent periods as well)
- Aging Assumptions
  - Operators age by 1 every other period
  - Vehicles age by 1 every other period

3.2. Running the Simulation

Now that the assumptions have been identified, the insurer can run a simulation. For educational purposes, we will examine "quotes" and "renewals" separately to gain an understanding of the mechanics underlying price simulation. Let us begin with "quotes":

Quotes								
	(1) Period	**(2) Policies Offered**	**(3) Policies Written**	**(4) = (3) / (2) Conversion Rate**	**(5) Policies Retained**	**(6) = (5) / (3) Retention Rate**	**(7) Profit Margin**	**(8) Elasticity**
Scenario 1	0	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	1	20,000	5,493	27.5%	4,669	85.0%	1.9%	1.8
	2	21,000	5,767	27.5%	4,902	85.0%	1.9%	1.8
	3	22,050	6,058	27.5%	5,150	85.0%	1.9%	1.8
	4	23,153	6,360	27.5%	5,406	85.0%	1.9%	1.8
Scenario 2	0	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	1	20,000	5,646	28.2%	4,743	84.0%	1.8%	2.4
	2	21,000	5,928	28.2%	4,980	84.0%	1.8%	2.4
	3	22,050	6,228	28.2%	5,231	84.0%	1.8%	2.4
	4	23,153	6,538	28.2%	5,492	84.0%	1.8%	2.4

Table 3.1 Quote simulation

In period 0, there are no **Policies Offered** since we assumed quotes do not enter the simulation until period 1. In period 1, the initial quote dataset of 20,000 enters the simulation. In period 2, 21,000 new quotes enter the simulation based on the assumed 5% quote growth rate. Periods 3 & 4 follow similarly. **Polices Written** shows the number of quotes converted based on the assumed conversion model. Once policies written are determined, an aggregate **Conversion Rate** can be calculated, as shown in Column (4). **Policies Retained** gives the number of converted quotes that persist to the next period based on the assumed retention model. Once policies retained are determined, an aggregate **Retention Rate** can be calculated, as shown in Column (6). Column (7) shows the **Profit Margin** changes over time in each scenario. The conversion rate, retention rate and profit margin are the same for each period because the quote distribution remains constant over time per our assumptions. Lastly, Column (8) contains the aggregate elasticity for the quote dataset. Since elasticity is a function of demand and demand is higher in Scenario 2, it follows that the Scenario 2 elasticity is higher than the Scenario 1 elasticity. Based on Table

3.1, which scenario is better? From a growth perspective, Scenario 2 is superior. However, from a profit margin perspective, the insurer would recommend Scenario 1. Before making a final decision, we must consider "existing business":

Renewals					
	(1) Period	(2) Policies Offered	(3) Policies Retained	(4) = (3) / (2) Retention Rate	(5) Profit Margin
Scenario 1	0	50,000	44,000	88.0%	2.5%
	1	44,000	41,287	93.8%	2.4%
	2	45,956[1]	44,162	96.1%	2.3%
	3	49,064	47,147	96.1%	2.2%
	4	52,296	49,315	94.3%	2.2%
Scenario 2	0	50,000	44,000	88.0%	2.5%
	1	44,000	41,287	93.8%	2.4%
	2	46,030	44,155	95.9%	2.5%
	3	49,135	47,121	95.9%	2.6%
	4	52,352	49,263	94.1%	2.7%

Table 3.2 Renewal simulation

In period 0, policies offered equal the initial renewal dataset of 50,000 policies. In period 1, policies offered equal the policies retained from period 0. This number is the same in both scenarios since we assumed the price component in the retention model was fixed (i.e. the market price and competitor price were perfectly correlated). In period 2, converted quotes from period 1 have transitioned to renewals. Thus, policies offered equal the sum of period 1 policies retained from the quotes in Table 3.1 and the renewals table in Table 3.2. Periods 3 and 4 follow similar logic. **Policies Retained** shows the number of renewal policies that are retained to the next period based on the assumed retention model. Once policies retained are determined, an aggregate **Retention Rate** can be calculated, as shown in Column (4). Lastly, Column (5) shows how the **Profit Margin** changes over time in each scenario. Based on Table 3.2, which scenario is better? From a retention perspective, Scenario 1 is best. However, from a profit margin perspective, the insurer would choose Scenario 2. Since Tables 3.1 and 3.2 failed to provide a consistent solution, we must examine the total results to determine which scenario to recommend.

[1] 45,956 = 41,287 + 4,669

Total (Renewals + Quotes)							
	(1) Period	**(2)** Policies Offered	**(3)** Policies Written	**(4)** Policies Retained	**(5)** Earned Premium	**(6)** Profit Margin	**(7)** Absolute Profit
Scenario 1	0	50,000	50,000	44,000	$35,250,000	2.5%	$881,250
	1	64,000	49,493	45,956	$34,486,258	2.3%	$810,152
	2	66,956	51,723	49,064	$36,412,258	2.3%	$822,930
	3	71,114	55,122	52,296	$38,800,399	2.2%	$842,146
	4	75,449	58,657	54,722	$40,949,798	2.2%	$888,759
Scenario 2	0	50,000	50,000	44,000	$35,250,000	2.5%	$881,250
	1	64,000	49,646	46,030	$34,729,064	2.3%	$812,026
	2	67,030	51,958	49,135	$36,692,114	2.4%	$891,271
	3	71,185	55,363	52,352	$39,087,466	2.5%	$985,029
	4	75,505	58,890	54,755	$41,236,423	2.6%	$1,076,159

Table 3.3 Total business simulation

After viewing the combined results in Table 3.3, the answer is clear. Over the stated time horizon of two years, Scenario 2 outperforms Scenario 1 from growth, retention, and profit perspectives. Whereas traditional analysis provides little distinction between the scenarios, price simulation allowed the insurer to "predict" how its book will change over time in each scenario, resulting in the better choice.

Please note that the above example is intended to illustrate the concept of price simulation. With a slight difference in rate allocation, we see that Scenario 2 yields more policies (~30) and higher profit (~$400,000) than Scenario 1 by the end of period 4. Not surprisingly, further rate allocation with a larger book of business could lead to a more significant difference in pricing scenarios.

4. PRICE OPTIMIZATION

Suppose ABC Insurance wants to increase its new business conversion rate. One option is to decrease the overall premium level. This will certainly increase the company's conversion rate, but it's probably not the optimal solution because it affects each insured equally. Another approach is for the company to be more surgical in its approach by targeting specific market segments with higher growth potential. Although this option is superior to the first, current techniques rely on sub-optimal human judgment when deciding which segments to target. Price optimization is a mathematical procedure that allows insurers to determine the appropriate rates to charge to maximize some metric over some specified time horizon. As stated above, ABC Insurance wanted to increase its overall conversion rate on new business. Although the approaches described will do that, the result will be inferior to a price optimization routine that analyzes thousands of different pricing options. Examples of metrics a company might want to maximize include:

- Profitability
    - Percentage basis (margin)
    - Absolute basis (total dollars)
- New Business Conversion
- Renewal Business Retention
- Competitive Position

It is important to note that these metrics are not mutually exclusive. Typically, a company sets a constraint on one metric while maximizing another. For example, a company may want to maximize its new business conversion rate while maintaining its current profitability. Similarly, a company may want to maximize its profitability while maintaining its current conversion rate.

4.1. Price Optimization Methods

In this section, we will discuss two price optimization methods. The first method optimizes prices at the rating structure level by determining optimal rating factors for the structures the company has chosen to revise. This method optimizes on the current rating structure directly, and is therefore easy to implement. However, it fails to identify gaps in the current rating structure where growth potential is not being realized. The second method optimizes prices irrespective of the current rating structure by determining an optimized premium for each insured in a firm's book of business. This has a number of advantages:

- Optimizes premium at the individual insured level
- Provides opportunity for improvements in a firm's rating structure by optimizing rates outside of the current pricing system
- Produces an efficient frontier of optimized premiums representing the highest possible demand at varying profit levels

Despite these advantages, this method has a few drawbacks. First, it requires more time since a rating structure must be reverse engineered from the optimized premiums. Second, some of the benefit gained from the optimization routine is lost during the reverse engineering process due to modeling limitations. In addition to the disadvantages listed above, both methods suffer from regulatory constraints, which will be discussed in a later section. The remainder of this section will focus on the second method.

4.2. The Benefit Function

Assume a firm wants to maximize its new business conversion rate while maintaining its current profitability. The firm has decided to use price optimization to achieve this goal. In order to optimize prices at the individual insured level, we begin with the benefit function. In basic terms, the benefit function calculates the expected profit for each insured over a specific time horizon. In equation form, the one-year benefit function might look like this:

$$BF_i = CD_i * (Q_i - L_i - E_i) \tag{4.1}$$

Where BF = Benefit Function

CD = Cumulative Demand

Q = Proposed Premium

L = Pure Premium

E = Expenses

i = i^{th} insured

Let us examine each piece of the equation starting with cumulative demand. In order to calculate **cumulative demand**, we need a conversion model and a retention model. Each of these models can be created using the techniques described in Section 2 of this paper. Once the models are built, we simply accumulate total demand for each individual over the stated time horizon. For a one-year time horizon, the cumulative demand for new business is the probability that the firm acquires the insured in year one. If the time horizon were three years, then the cumulative demand for new business would be the probability that the firm acquires the insured in year one, multiplied by the probability that the firm retains the insured for two more years. The **proposed premium** is derived by solving for the optimized premium. The **pure premium** is an estimate of the expected loss costs for each insured. Lastly, we must account for **expenses**. Once we have defined each piece of the benefit function, we can solve for the proposed premium (i.e. optimized premium) by maximizing the benefit function. This results in an optimized premium for each insured that maximizes his or her expected profit over the stated time horizon irrespective of the conversion rate. Although this certainly qualifies as an optimized premium, it is not the optimized premium sought by the firm in our example. Remember, our firm wanted to increase its new business conversion rate **while maintaining its current profitability.** So, where does the firm go from here?

4.3. The Efficient Frontier

The efficient frontier is a curve that plots the expected loss ratio against demand. It is "efficient" because each point on the curve represents a set of optimized premiums. The maximized benefit function exists on the efficient frontier at the point where profitability is the greatest. As we saw earlier, it has a set of "optimized" premiums associated with it; namely, the ones that maximize profitability over the stated time horizon. As you move up the curve, demand increases and profitability decreases. Points inside the curve, such as the firm's current position, are considered sub-optimal because the company could improve demand while maintaining profitability, improve profitability while maintaining demand, or some combination thereof. If a firm wants to maximize its new business conversion rate while keeping its current profitability constant, the firm moves its position up the curve until it reaches the optimized location. To move up the curve, the optimization routine analyzes the PED curve for each individual insured to determine the proper price to charge them in order to achieve the target expected loss ratio.

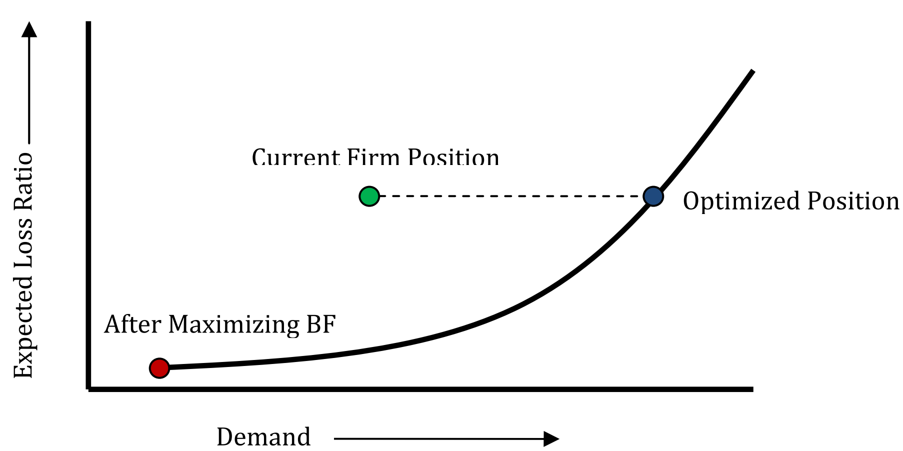


Figure 4.1 The Efficient Frontier

Once a set of optimized premiums has been determined that meets the firm's goals, the firm can finally reverse engineer a rating structure by modeling the optimized premiums. In general, this just involves fitting the current rating structure onto the optimized premiums. However, thanks to the flexible nature of optimizing at the individual insured level, other predictors may be used to account for gaps in the current rating structure found during the reverse engineering process. Although there are a number of ways to complete the reverse engineering process, we have chosen to omit that discussion due to the limited scope of this paper.

4.4. Implementing Optimized Rates in a Regulatory Environment

One substantial challenge associated with price optimization is implementing optimized rates within the constraints of a regulatory environment. The important thing to remember is that price optimization is a tool designed to help firms make better pricing decisions. As with any rate filing, a company starts by developing indications based on the loss propensities, expense loads and rate of return objectives for various segments of its book of business. At this point, the company can choose to stay at the current rates, use the indicated rates, or deviate from the indicated rates based on business judgment. In this regard, price optimization can be seen as a deviation from the indicated rates. Of course, this doesn't mean that a firm can file anything that comes out of the optimization routine. Since price optimization relies on PED in addition to loss propensity, it could conflict with traditional pricing methods. For example, suppose a mature segment of a firm's book of business has a traditional actuarial rate indication of -10%. In addition to this, assume the segment has a low PED relative to the rest of the book. Now, assume the firm runs an optimization routine designed to increase overall demand. Due to the low PED of the mature segment, the optimized rate will most likely be higher than the current rate since that particular segment is less price sensitive. However, this conflicts with the indicated decrease of -10%. In this case, the optimized rates could be subject to regulatory challenge and may need to be revised to coincide with the loss indications. As a result, some of the benefit from the optimization routine would be lost. This example is not meant to deter the use of price optimization; it's meant as a reminder that the final rates decided upon still require actuarial justification and approval by state agencies.

5. CONCLUDING REMARKS

In this paper, we detailed basic techniques for developing demand models with emphasis on model form and model structure. Next, we examined the use of PED in the context of two practical applications: price simulation and price optimization. Lastly, we described how to implement optimized rates in a regulatory environment.

Despite the sophistication present in the insurance industry today, the integration of competitive data within the pricing process continues to invite further exploration. This paper provides the reader with the basic tools needed to begin this process. With these tools, insurers possess the necessary knowledge and skills to implement more targeted pricing decisions that better achieve their goals.

Acknowledgments

We would like to thank the following individuals for their comments and suggestions: William Frierson, Rob Kane, David Maurer, & Ryan Voge. We would also like to give special thanks to Rob Kane for his numerous contributions to the price simulation section and to Michael Korner for his invaluable and ongoing research in demand modeling techniques.

6. REFERENCES

1. Anderson, Duncan, et al. "A Practitioner's Guide to Generalized Linear Models." *CAS Discussion Paper Program* (2004): 1-115.

2. Brockman, Michael J. "Statistical Motor Rating: Making Effective Use of Your Data." *Journal of the Institute of Actuaries* (April 1992): 457-543.

3. Hastie, T., R. Tibshirani, and J.H. Friedman. *The Elements of Statistical Learning*. Springer, 2001.

4. Holler, Keith D., David Sommer, and Geoff Trahair. "Something Old, Something New in Classification Ratemaking with a Novel Use of GLMs for Credit Insurance." *CAS Forum* (Winter 1999): 31-84.

5. McCullagh, P., and J.A. Nelder. *Generalized Linear Models*. 2nd ed. Chapman and Hall, 1989.

6. Mildenhall, Steve. "Systematic Relationship Between Minimum Bias and Generalized Linear Models." *CAS Proceeding LXXXVI* (1999): 393-487.

7. Murphy, Karl P., Michael J. Brockman, and Peter K. Lee. "Using Generalized Linear Models to Build Dynamic Pricing Systems." *CAS Forum* (Winter 2000): 107-139.

Author Biographies

- **Serhat Guven** is a Senior Consultant with Towers Watson. He is a Fellow of the CAS and a member of the American Academy of Actuaries. He can be reached at serhat.guven@towerswatson.com.
- **Michael McPhail** is a Pricing Director with USAA. He is a Fellow of the CAS and a member of the American Academy of Actuaries. He can be reached at michael.mcphail@usaa.com.

Estimation of Claim Count Data using Negative Binomial, Generalized Poisson, Zero-Inflated Negative Binomial and Zero-Inflated Generalized Poisson Regression Models

Noriszura Ismail, Ph.D. and Hossein Zamani, Ph.D.

Abstract: This study relates negative binomial and generalized Poisson regression models through the mean-variance relationship, and suggests the application of these models for overdispersed or underdispersed count data. In addition, this study relates zero-inflated negative binomial and zero-inflated generalized Poisson regression models through the mean-variance relationship, and suggests the application of these zero-inflated models for zero-inflated and overdispersed count data. The negative binomial and generalized Poisson regression models were fitted to the Malaysian OD claim count data, whereas the zero-inflated negative binomial and zero-inflated generalized Poisson regression models were fitted to the German healthcare count data.

Keywords: negative binomial regression, generalized Poisson regression, zero-inflated negative binomial regression, zero-inflated generalized Poisson regression.

---

1. INTRODUCTION

Poisson regression has been widely used for fitting count data. As examples, in insurance area, Aitkin et al. (1990) and Renshaw (1994) fitted Poisson regression to two different sets of U.K. motor claim data, whereas in healthcare area, Riphahn et al. (2003) fitted the model to German Socioeconomic Panel (GSOEP) data. However, count data often display overdispersion and inappropriate imposition of Poisson regression may underestimate the standard errors and overstate the significance of regression parameters. Quasi Poisson regression has been suggested to accommodate overdispersion in count data and the advantage of using this model is that the model can be fitted without knowing the exact probability function of the response variable, as long as the mean is specified to be equivalent to Poisson mean and the variance can be written as a linear proportion of Poisson mean. To account for overdispersion, quasi Poisson regression produces regression estimates equivalent to Poisson regression, but standard errors larger than Poisson regression (Ismail and Jemain 2007). In insurance applications, McCullagh and Nelder (1989) fitted quasi Poisson regression to damage incidents of cargo-carrying vessels and Brockman and Wright (1992) fitted the model to U.K. own damage motor claim data.

Besides quasi Poisson regression, several mixed Poisson regressions have been considered as alternatives for handling overdispersed count data. The details of mixed Poisson regressions such as Poisson-inverse Gaussian (PIG), Poisson-lognormal (PLN) and Poisson-gamma (or negative binomial) can be found in Denuit et al. (2007), who applied these models for accommodating overdispersed claim data. For the case of both overdispersed and underdispersed count data, generalized Poisson and COM-Poisson regression models have been suggested and fitted by several researchers (Consul 1989, Consul and Famoye 1992, Famoye et al. 2004, Wang and Famoye 1997, Conway and Maxwell 1962, Shmueli et al. 2005, Lord et al. 2008, Lord et al. 2010).

Several parameterizations have been performed for negative binomial (NB) regression, and the two well known models, NB-1 and NB-2, have been applied (Cameron and Trivedi 1986, Lawless 1987, McCullagh and Nelder 1989, Cameron and Trivedi 1998, Winkelmann 2008, Hilbe 2007). However, both NB-1 and NB-2 regressions are not nested, and appropriate statistical tests to choose a better model cannot be carried out. Recently, the functional form of NB regression has been extended and introduced as NB-P regression, where both NB-1 and NB-2 regressions are special cases of NB-P when $P=1$ and $P=2$ respectively (Greene 2008). The advantage of using NB-P regression is that it parametrically nests both NB-1 and NB-2 regressions, and hence, allowing statistical tests of the two functional forms against a more general alternative. In particular, likelihood ratio test can be implemented for choosing between NB-1 against NB-P regressions, or NB-2 against NB-P regressions. Applications of several parameterizations for NB, PIG and PLN regressions can be found in Boucher et al. (2007), who fitted NB (NB-1, NB-2 and NB-K+1), PIG (PIG-1, PIG-2 and PIG-K+1) and PLN (PLN-1, PLN-2 and PLN-K+1) regressions to Spanish motor claim data.

Several parameterizations have also been performed for generalized Poisson (GP) regression where the two well known models, GP-1 and GP-2, have been applied for dealing with overdispersed as well as underdispersed count data (Consul and Jain 1973, Consul 1989, Consul and Famoye 1992, Famoye et al. 2004, Wang and Famoye 1997, Ismail and Jemain 2007). Similar to NB-1 and NB-2 regressions, both GP-1 and GP-2 regressions are not nested and appropriate statistical tests to choose a better model cannot be performed. Recently, the functional form of GP regression has been extended and introduced as GP-P regression, where both GP-1 and GP-2 regressions are special cases of GP-P when $P=1$ and $P=2$ respectively (Zamani and Ismail 2012). The advantage of using GP-P is that it parametrically nests both GP-1 and GP-2, and therefore, allowing statistical tests of the two functional forms against a more general alternative. In particular, likelihood ratio test can be implemented for choosing between GP-1 against GP-P regressions, or GP-2 against GP-P regressions.

In terms of properties, there is no big difference between NB and GP distributions when the means and variances are fixed, but, GP distribution has heavier tail whereas NB distribution has more mass at zero (Joe and Zhu 2005). In addition, the score tests for overdispersion in Poisson vs. NB (NB-1 and NB-2) regressions and Poisson vs. GP (GP-1 and GP-2) regressions are equal (Yang et al. 2007, Yang et al. 2009a).

In several cases, count data often have excessive number of zero outcomes than are expected in Poisson regression. As an example, the proportion of zero claims in motor insurance data may increase due to the conditions of deductible and no claim discount that discourage insured drivers to report small claims (Yip and Yau 2005). In healthcare area as another example, services of psychiatric outpatient may report a large proportion of zero utilization of such services for many patients (Neelon et al. 2010). Zero-inflated datasets can also be found in other areas such as environmental sciences (Agarwal et al. 2002), medicine (Bohning et al. 1996) and manufacturing (Lambert 1992). Zero-inflation phenomenon is a very specific type of overdispersion, and zero-inflated Poisson (ZIP) regression has been suggested to handle purely zero-inflated data. ZIP regression mixes a distribution degenerate at zero with a Poisson distribution, by allowing the incorporation of explanatory variables in both the zero process and the Poisson distribution.

As an alternative to ZIP regression, one may consider zero-inflated negative binomial (ZINB) regression if the count data continue to suggest additional overdispersion. ZINB regression is obtained by mixing a distribution degenerate at zero with a NB distribution, by allowing the incorporation of explanatory variables in both the zero process and the NB distribution. Applications of ZINB-1 and ZINB-2 regressions can be found in Ridout et al. (2001).

Besides ZINB, zero-inflated generalized Poisson (ZIGP) regression has been proposed as an alternative to handle zero-inflation and additional overdispersion in count data. ZIGP regression, which mixes a distribution degenerate at zero with a GP distribution and allows the incorporation of explanatory variables in both the zero process and the GP distribution, have been applied by Famoye and Singh (2006) for domestic violence data and by Yang et al. (2009b) for apple shoot propagation data. However, both ZIGP-1 and ZIGP-2 regressions are not nested and appropriate statistical tests to choose a better model cannot be carried out. Recently, the functional form of ZIGP regression has been extended and introduced as ZIGP-P regression (Zamani and Ismail 2013a), where ZIGP-1 and ZIGP-2 regressions are special cases of ZIGP-P regression when $P = 1$ and $P = 2$ respectively. The advantage of using ZIGP-P regression is that it parametrically nests both ZIGP-1 and ZIGP-2 regressions, and hence, allowing statistical tests of the two functional forms against a more general alternative. In particular, likelihood ratio test can be implemented for choosing between ZIGP-1 against ZIGP-P regressions, or ZIGP-2 against ZIGP-P regressions.

Even though ZIGP regression is a good competitor of ZINB regression, in several cases, ZINB regression may not provide converged values in the iterative technique of the fitting procedure, and thus, ZIGP regression may be considered as an alternative (Famoye and Singh 2006). In addition, when both zero-inflation and overdispersion exist in count data, ZIGP regression behaves similarly to ZINB regression. As examples, Yang et al. (2009b) proved that the score statistics for testing overdispersion in ZIP against ZIGP (ZIGP-1 and ZIGP-2) regressions and ZIP against ZINB (ZINB-1 and ZINB-2) regressions are equal, whereas Joe and Zhu (2005) proved that ZIGP distribution provides better fit than ZINB distribution when there is a large zero fraction and heavy tail, implying that the ZIGP regression can be used as an alternative for modeling zero-inflated and overdispersed count data.

The objectives of this study are:

- to relate NB and GP regressions through the mean-variance relationship,
- to suggest applications of these models for overdispersed or underdispersed claim count data,
- to relate ZINB and ZIGP regression models through the mean-variance relationship, and finally,
- to suggest applications of these zero-inflated models for zero-inflated and overdispersed claim count data.

2. NB REGRESSION MODELS

Let $(Y_1, Y_2, ..., Y_n)^T$ be the vector of count random variables and n be the sample size. The probability mass function (p.m.f.) for Poisson regression is,

$$\Pr(Y_i = y_i) = \frac{\exp(-\mu_i)\mu_i^{y_i}}{y_i!}, \quad y_i = 0,1,... \tag{1}$$

with mean and variance $E(Y_i) = Var(Y_i) = \mu_i$. The mean or the fitted value can be assumed to follow a log link, $E(Y_i) = \mu_i = \exp(\mathbf{x_i}'\boldsymbol{\beta})$, where $\mathbf{x_i}$ denotes the vector of explanatory variables and $\boldsymbol{\beta}$ the vector of regression parameters. The maximum likelihood estimates can be obtained by maximizing the log likelihood.

The latent heterogeneity can be incorporated by rewriting the conditional mean of Poisson regression as (Cameron and Trivedi 1986, Cameron and Trivedi 1998, Winkelmann 2003, Greene 2008),

$$E(Y_i \mid \varepsilon_i) = \exp(\mathbf{x_i}'\boldsymbol{\beta} + \varepsilon_i) = k_i \mu_i,$$

where $k_i = \exp(\varepsilon_i)$ is assumed to follow gamma distribution with mean 1 and variance $v^{-1} = a$, with probability density function (p.d.f.),

$$f(k_i) = \frac{v^v k_i^{v-1} \exp(-v k_i)}{\Gamma(v)}, \quad k_i \geq 0, v > 0,$$

so that the conditional Poisson regression is,

$$\Pr(Y_i = y_i \mid k_i) = \frac{\exp(-k_i\mu_i)(k_i\mu_i)^{y_i}}{y_i!}, \quad y_i = 0,1,...$$

The marginal distribution is NB regression with p.m.f.,

$$\begin{aligned} \Pr(Y_i = y_i) &= \int_0^\infty \frac{e^{-k_i\mu_i}(k_i\mu_i)^{y_i}}{y_i!} \frac{v^v e^{-vk_i} k_i^{v-1}}{\Gamma(v)}\, dk_i \\ &= \frac{\Gamma(y_i + v)}{y_i!\Gamma(v)} \left(\frac{v}{v+\mu_i}\right)^v \left(\frac{\mu_i}{v+\mu_i}\right)^{y_i}, \quad y_i = 0,1,2,..., \end{aligned} \qquad (2)$$

where the mean is $E(Y_i) = \mu_i$, the variance is $Var(Y_i) = \mu_i(1 + v^{-1}\mu_i) = \mu_i(1 + a\mu_i)$, and $v^{-1} = a$ denotes the dispersion parameter. The NB regression in (2) is also referred as NB-2 regression. Another parameterization for NB regression is by letting $v = a^{-1}\mu_i$ in (2) to produce NB-1 regression, with mean $E(Y_i) = \mu_i$ and variance $Var(Y_i) = \mu_i(1 + a)$. Another parameterization is the NB-P regression, which is produced by letting $v = a^{-1}\mu_i^{2-P}$ in (2), so that the mean is $E(Y_i) = \mu_i$ and the variance is $Var(Y_i) = \mu_i(1 + a\mu_i^{P-1})$, where a denotes the dispersion parameter and P the functional parameter (Greene 2008).

NB regressions reduce to Poisson regression in the limit as $a \to 0$, and display overdispersion when $a > 0$. In addition, NB-P regression reduces to NB-1 when $P = 1$ and reduces to NB-2 when $P = 2$. Therefore, NB-P regression parametrically nests both NB-1 and NB-2, and allows statistical tests of the two functional forms against a more general alternative. The mean of NB regressions can also be assumed to follow the log link, $E(Y_i)=\mu_i=\exp(\mathbf{x_i}'\boldsymbol{\beta})$, and the maximum likelihood estimates can be obtained by maximizing the log likelihood.

NB-P regression can be fitted to count data using *R program*. For faster convergence, estimated parameters from fitting Poisson regression can be used as initial values. Both NB-1 and NB-2 regressions can also be fitted using the same fitting procedure, by letting the functional parameter, *P*, to be fixed at *P*=1 and *P*=2 respectively for NB-1 and NB-2.

We can also use *MASS package* in *R* to fit Poisson and NB-2 regression models. If we use *glm* function and set *family=poisson*, we can get Poisson regression. If we use *glm.nb* function, we can get NB-2 regression.

3. GP REGRESSION MODELS

The p.m.f of GP distribution is (Consul and Famoye 1992),

$$\Pr(Y_i=y_i)=\frac{\theta(\theta+vy_i)^{y_i-1}\exp(-\theta-vy_i)}{y_i!},\quad y_i=0,1,2,\ldots, \tag{3}$$

where $\theta>0$ and $\max(-1,-\frac{\theta}{4})<v<1$. The mean and variance are $E(Y_i)=\mu=(1-v)^{-1}\theta$ and $Var(Y_i)=(1-v)^{-3}\theta=(1-v)^{-2}\mu$, where $(1-v)^{-2}$ denotes the dispersion factor and v the dispersion parameter.

There are two well known parameterizations for GP regression, referred as GP-1 and GP-2. By letting $\theta_i=(1-v)\mu_i$ in (3), GP-1 regression is produced with mean $E(Y_i)=\mu_i$ and variance $Var(Y_i)=(1-v)^{-2}\mu_i$. GP-1 regression reduces to Poisson regression when $v=0$, allows overdispersion when $v>0$, and allows underdispersion when $v<0$. A new form of GP-1 regression, which has the same properties but different form of p.m.f., was recently proposed in Zamani and Ismail (2012) by rewriting $v=a(1+a)^{-1}$ and $\theta_i=(1+a)^{-1}\mu_i$ in (3). The mean and variance of new GP-1 regression are $E(Y_i)=\mu_i$ and $Var(Y_i)=(1+a)^2\mu_i$, where a denotes the dispersion parameter. The other parameterization is GP-2 regression, involving the parameterization of $v=(1+\mu_i a)^{-1}a\mu_i$ and $\theta_i=(1+\mu_i a)^{-1}\mu_i$ in (3), with mean $E(Y_i)=\mu_i$ and variance $Var(Y_i)=(1+a\mu_i)^2\mu_i$. Another parameterization is GP-P regression, which was recently proposed in Zamani and Ismail (2012), obtained using $\theta_i=(1+a\mu_i^{P-1})^{-1}\mu_i$ and $v=(1+a\mu_i^{P-1})^{-1}a\mu_i^{P-1}$ in (3).

The mean and variance of GP-P regression are $E(Y_i) = \mu_i$ and $Var(Y_i) = (1 + a\mu_i^{P-1})^2 \mu_i$, where a denotes the dispersion parameter and P the functional parameter.

GP regressions reduce to Poisson regression when $a = 0$, display overdispersion when $a > 0$, and display underdispersion when $a < 0$. In addition, GP-P reduces to new GP-1 when $P = 1$, and reduces to GP-2 when $P = 2$. Therefore, GP-P regression parametrically nests both new GP-1 and GP-2 regressions, and allows statistical tests of the two functional forms against a more general alternative. The mean of GP regressions can be assumed to follow the log link, $E(Y_i) = \mu_i = \exp(\mathbf{x_i}'\boldsymbol{\beta})$, and the maximum likelihood estimates can be obtained by maximizing the log likelihood.

GP-P regression can be fitted using R *program*. For faster convergence, estimated parameters from fitting Poisson regression can be used as initial values. The Poisson, new GP-1 and GP-2 regressions can also be fitted using the same fitting procedure, by letting the dispersion parameter, *a*, to be fixed at *a*=0 for Poisson and by letting the functional parameter, *P*, to be fixed at *P*=1 and *P*=2 respectively for new GP-1 and GP-2. An example of *SAS code* for fitting GP-1 and GP-2 regressions can be found in Yang et al. (2007) and Yang et al. (2009a) respectively.

We can also use *ZIGP package* in *R* to fit GP-2 regression models. If we use *est.zigp* function and set *fm.Z=NULL* (which is the weight for zero-inflation), we can get GP-2 regression. Table 1 summarizes the p.m.f., mean and variance of Poisson, NB and GP regression models.

Table 1: Poisson, NB and GP regression models

Regression model	P.m.f.	Mean and variance
Poisson	$\frac{\mu_i^{y_i}}{y_i!}\exp(-\mu_i)$	$E(Y_i)=\mu_i$ $Var(Y_i)=\mu_i$
NB-1	$\frac{\Gamma(y_i+\mu_i a^{-1})}{y_i!\Gamma(\mu_i a^{-1})}\left(\frac{\mu_i a^{-1}}{\mu_i a^{-1}+\mu_i}\right)^{\mu_i a^{-1}}\left(\frac{\mu_i}{\mu_i a^{-1}+\mu_i}\right)^{y_i}$	$E(Y_i)=\mu_i$ $Var(Y_i)=\mu_i(1+a)$
NB-2	$\frac{\Gamma(y_i+a^{-1})}{y_i!\Gamma(a^{-1})}\left(\frac{a^{-1}}{a^{-1}+\mu_i}\right)^{a^{-1}}\left(\frac{\mu_i}{a^{-1}+\mu_i}\right)^{y_i}$	$E(Y_i)=\mu_i$ $Var(Y_i)=\mu_i(1+a\mu_i)$
NB-P	$\frac{\Gamma(y_i+a^{-1}\mu_i^{2-P})}{y_i!\Gamma(a^{-1}\mu_i^{2-P})}\left(\frac{a^{-1}\mu_i^{2-P}}{a^{-1}\mu_i^{2-P}+\mu_i}\right)^{a^{-1}\mu_i^{2-P}}\left(\frac{\mu_i}{a^{-1}\mu_i^{2-P}+\mu_i}\right)^{y_i}$	$E(Y_i)=\mu_i$ $Var(Y_i)=\mu_i(1+a\mu_i^{P-1})$
GP-1	$\frac{((1-v)\mu_i)((1-v)\mu_i+vy_i)^{y_i-1}}{y_i!}\exp(-((1-v)\mu_i+vy_i))$	$E(Y_i)=\mu_i$ $Var(Y_i)=\mu_i(1-v)^{-2}$
New GP-1	$\frac{\mu_i(\mu_i+ay_i)^{y_i-1}}{(1+a)^{y_i}y_i!}\exp\left(-\frac{\mu_i+ay_i}{1+a}\right)$	$E(Y_i)=\mu_i$ $Var(Y_i)=\mu_i(1+a)^2$
GP-2	$\frac{\mu_i(\mu_i+a\mu_i y_i)^{y_i-1}}{(1+a\mu_i)^{y_i}y_i!}\exp\left(-\frac{\mu_i+a\mu_i y_i}{1+a\mu_i}\right)$	$E(Y_i)=\mu_i$ $Var(Y_i)=\mu_i(1+a\mu_i)^2$
New GP-P	$\frac{\mu_i(\mu_i+a\mu_i^{P-1}y_i)^{y_i-1}}{(1+a\mu_i^{P-1})^{y_i}y_i!}\exp\left(-\frac{\mu_i+a\mu_i^{P-1}y_i}{1+a\mu_i^{P-1}}\right)$	$E(Y_i)=\mu_i$ $Var(Y_i)=\mu_i(1+a\mu_i^{P-1})^2$

Several general comparisons can be made regarding NB and GP regression models:

- NB regressions reduce to Poisson regression in the limit as $a \to 0$, whereas GP regressions reduce to Poisson regression when $a = 0$.
- The mean-variance relationships of NB-1 and GP-1 are linear, NB-2 is quadratic, GP-2 is cubic, NB-P is to the *p*th power, and GP-P is to the (2*p*-1)th power.
- Likelihood ratio test (LRT) can be performed for testing overdispersion in Poisson vs. NB (NB-1 and NB-2) regressions, $H_0 : a = 0$ vs. $H_1 : a > 0$. Since the null hypothesis is on the boundary of parameter space, the asymptotic distribution for LRT statistic is a mixture of half of probability mass at zero and half of chi-square with one degree of freedom (Lawless 1987).
- Likelihood ratio test (LRT) can also be performed for testing dispersion (over or underdispersion) in Poisson vs. GP (GP-1 and GP-2) regressions, $H_0 : a = 0$ vs. $H_1 : a \neq 0$, where the LRT statistic is asymptotically distributed as a chi-square with one degree of freedom (Consul and Famoye 1992, Wang and Famoye 1997).
- In terms of properties, there is no big difference between NB and GP distributions when the means and variances are fixed. However, GP distribution has heavier tail, whereas NB distribution has more mass at zero (Joe and Zhu 2005).

4. ZIP REGRESSION MODEL

Zero-inflated Poisson (ZIP) regression has been used by researchers for handling purely zero-inflated count data. ZIP regression can be obtained by mixing a distribution degenerate at zero with the Poisson distribution, by allowing the incorporation of explanatory variables in both the zero process and the Poisson distribution. The p.m.f. of ZIP regression is,

$$\Pr(Y_i = y_i) = \begin{cases} \omega_i + (1-\omega_i)\exp(-\mu_i), & y_i = 0 \\ (1-\omega_i)\dfrac{\mu_i^{y_i}}{y_i!}\exp(-\mu_i), & y_i > 0 \end{cases} \tag{4}$$

where $0 \le \omega_i < 1$ and $\mu_i > 0$, with mean $E(Y_i) = (1-\omega_i)\mu_i$ and variance $Var(Y_i) = (1-\omega_i)\mu_i(1+\omega_i\mu_i)$. ZIP regression reduces to Poisson regression when $\omega_i = 0$, and exhibits overdispersion when $\omega_i > 0$. The covariates can be incorporated by using a log link for μ_i and a logit link for ω_i,

$$\log(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta} \text{ and } \log\left(\frac{\omega_i}{1-\omega_i}\right) = \mathbf{z}_i^T \boldsymbol{\gamma}, \tag{5}$$

where $\mathbf{x}_i$ and $\mathbf{z}_i$ are the vectors of explanatory variables, and $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are the vectors of regression parameters. Maximum likelihood estimates can be obtained by maximizing the log likelihood.

5. ZINB REGRESSION MODELS

Zero-inflated negative binomial (ZINB) regressions have been used by researchers for handling both zero-inflation and overdispersion in count data. ZINB-1 regression can be obtained by mixing a distribution degenerate at zero with the NB-1 distribution, by allowing the incorporation of explanatory variables in both the zero process and the NB-1 distribution. The mean and variance of ZINB-1 regression are $E(Y_i) = (1-\omega_i)\mu_i$ and $Var(Y_i) = (1-\omega_i)\mu_i(1+a+\omega_i\mu_i)$.

ZINB-2 regression can be obtained by mixing a distribution degenerate at zero with the NB-2 distribution, by allowing the incorporation of explanatory variables in both the zero process and the NB-2 distribution. The mean and variance of ZINB-2 regression are $E(Y_i) = (1-\omega_i)\mu_i$ and $Var(Y_i) = (1-\omega_i)\mu_i(1+a\mu_i+\omega_i\mu_i)$.

The ZINB-P regression can be obtained by mixing a distribution degenerate at zero with the NB-P distribution, by allowing the incorporation of explanatory variables in both the zero process and the NB-P distribution. The mean and variance of ZINB-P regression are $E(Y_i) = (1-\omega_i)\mu_i$ and $Var(Y_i) = (1-\omega_i)\mu_i(1+a\mu_i^{P-1}+\omega_i\mu_i)$.

ZINB regressions reduce to NB regressions when $\omega_i = 0$, and reduce to ZIP regression in the limit as $a \to 0$. The variance of ZINB regressions exhibits overdispersion when $a > 0$ and $\omega_i > 0$, allowing the models to be used for handling both zero-inflated and overdispersed count data. The link functions for ZINB regressions can also be written as (5). Maximum likelihood estimates can be obtained by maximizing the log likelihood.

ZINB-P regression can be fitted using *R program*. For faster convergence, estimated parameters from fitting ZIP regression can be used as initial values. Both ZINB-1 and the ZINB-2 regressions can also be fitted using the same fitting procedure, by letting the functional parameter, *P*, to be fixed respectively at *P*=1 and *P*=2 for ZINB-1 and ZINB-2.

We can also use *PSCL package* in *R* to fit ZIP and ZINB-2 regression models. If we use *zeroinfl* function and set *dist="poisson"*, we can get ZIP regression. If we set *dist="negbin"*, we can get ZINB-2 regression.

6. ZIGP REGRESSION MODELS

As an alternative for handling zero-inflation and overdispersion in count data, zero-inflated generalized Poisson (ZIGP) regressions can be fitted. ZIGP-1 regression is obtained by mixing a distribution degenerate at zero with the GP-1 distribution, by allowing the incorporation of explanatory variables in both the zero process and the GP-1 distribution. The mean and variance of ZIGP-1 regression are $E(Y_i) = (1-\omega_i)\mu_i$ and $Var(Y_i) = (1-\omega_i)\mu_i((1-v)^{-2} + \omega_i\mu_i)$. A new form of ZIGP-1 regression has been proposed by Zamani and Ismail (2013a), by mixing a distribution degenerate at zero with the new GP-1 distribution, and allowing the incorporation of explanatory variables in both the zero process and the new GP-1 distribution. The mean and variance of new ZIGP-1 are $E(Y_i) = (1-\omega_i)\mu_i$ and $Var(Y_i) = (1-\omega_i)\mu_i((1+a)^2 + \omega_i\mu_i)$.

ZIGP-2 regression is obtained by mixing a distribution degenerate at zero with the GP-2 distribution, by allowing the incorporation of explanatory variables in both the zero process and the GP-2 distribution. The mean and variance of ZIGP-2 regression are $E(Y_i) = (1-\omega_i)\mu_i$ and $Var(Y_i) = (1-\omega_i)\mu_i((1+a\mu_i)^2 + \omega_i\mu_i)$.

A functional form of ZIGP, which is referred as ZIGP-P, was recently proposed by Zamani and Ismail (2013a). ZIGP-P regression is obtained by mixing a distribution degenerate at zero with the GP-P distribution, by allowing the incorporation of explanatory variables in both the zero process and the GP-P distribution. The mean and variance of ZIGP-P regression are $E(Y_i) = (1-\omega_i)\mu_i$ and $Var(Y_i) = (1-\omega_i)\mu_i((1+a\mu_i^{P-1})^2 + \omega_i\mu_i)$. When $\omega_i = 0$, ZIGP regressions reduce to GP regressions, and when $a = 0$, ZIGP regressions reduce to ZIP regression. In addition, when $P = 1$ and $P = 2$, ZIGP-P regression reduces to ZIGP-1 and ZIGP-2 regressions respectively. The variance of ZIGP regressions exhibit overdispersion when $a > 0$ and $\omega_i > 0$, indicating that they can be used for handling both zero-inflated and overdispersed count data. The link functions for ZIGP regressions can also be written as (5), and the maximum likelihood estimates can be obtained by maximizing the log likelihood.

ZIGP-P regression can be fitted using *R program*. For faster convergence, estimated parameters from fitting ZIP regression can be used as initial values. The ZIP, ZIGP-1 and ZIGP-2 regressions can also be fitted using the same fitting procedure, by letting the dispersion parameter, *a*, to be fixed at *a*=0 for ZIP and by letting the functional parameter, *P*, to be fixed respectively at *P*=1 and *P*=2 for ZIGP-1 and ZIGP-2. An example of *SAS code* for fitting ZIGP-2 regression can be found in Yang et al. (2009b).

We can also use *ZIGP package* in *R* to fit ZIP and ZIGP-2 regression models. If we use *est.zigp* function and set *fm.W=NULL* (which is the weight for dispersion), we can get ZIP regression. We can get ZINB-2 regression if both *fm.Z* (weight for zero-inflation) and *fm.W* (weight for dispersion) are not set to *NULL*. Table 2 summarizes the p.m.f., mean and variance of ZIP, ZINB and ZIGP regression models.

Several general comparisons can be made regarding ZINB and ZIGP regression models:

- ZINB regressions reduce to ZIP regression in the limit as $a \to 0$, whereas ZIGP regressions reduce to ZIP regression when $a = 0$.
- LRT can be performed for testing overdispersion in ZIP vs. ZINB regressions, where $H_0 : a = 0$ vs. $H_1 : a > 0$. Since the null hypothesis is on the boundary of parameter space, the asymptotic distribution for LRT statistic is a mixture of half of probability mass at zero and half of chi-square with one degree of freedom (Stram and Lee 1994, 1995).
- The signed square root of likelihood ratio test (SSN-LRT) can be performed for testing overdispersion in ZIP vs. ZIGP regressions, $H_0 : a = 0$ vs. $H_1 : a > 0$, where the statistic is asymptotically distributed as a standard normal (Yang et al. 2009b, Famoye and Singh 2006).
- Since GP distribution has heavier tail and NB distribution has more mass at zero, both ZINB and ZIGP regressions can be used for fitting zero-inflated and overdispersed count data (Joe and Zhu 2005).

Table 2: ZIP, ZINB and ZIGP regression models

Model	P.m.f.	Mean and variance
ZIP	$\begin{cases}\omega_i+(1-\omega_i)\exp(-\mu_i), & y_i=0\\ (1-\omega_i)\dfrac{\mu_i^{y_i}}{y_i!}\exp(-\mu_i), & y_i>0\end{cases}$	$E(Y_i)=(1-\omega_i)\mu_i$ $Var(Y_i)=E(Y_i)(1+\omega_i\mu_i)$
ZINB-1	$\begin{cases}\omega_i+(1-\omega_i)\left(\dfrac{\mu_i a^{-1}}{\mu_i a^{-1}+\mu_i}\right)^{\mu_i a^{-1}}, & y_i=0\\ (1-\omega_i)\dfrac{\Gamma(y_i+\mu_i a^{-1})}{y_i!\Gamma(\mu_i a^{-1})}\left(\dfrac{\mu_i a^{-1}}{\mu_i a^{-1}+\mu_i}\right)^{\mu_i a^{-1}}\left(\dfrac{\mu_i}{\mu_i a^{-1}+\mu_i}\right)^{y_i}, & y_i>0\end{cases}$	$E(Y_i)=(1-\omega_i)\mu_i$ $Var(Y_i)=E(Y_i)(1+a+\omega_i\mu_i)$
ZINB-2	$\begin{cases}\omega_i+(1-\omega_i)\left(\dfrac{a^{-1}}{a^{-1}+\mu_i}\right)^{a^{-1}}, & y_i=0\\ (1-\omega_i)\dfrac{\Gamma(y_i+a^{-1})}{y_i!\Gamma(a^{-1})}\left(\dfrac{a^{-1}}{a^{-1}+\mu_i}\right)^{a^{-1}}\left(\dfrac{\mu_i}{a^{-1}+\mu_i}\right)^{y_i}, & y_i>0\end{cases}$	$E(Y_i)=(1-\omega_i)\mu_i$ $Var(Y_i)=E(Y_i)(1+a\mu_i+\omega_i\mu_i)$
ZINB-P	$\begin{cases}\omega_i+(1-\omega_i)\left(\dfrac{a^{-1}\mu_i^{2-P}}{a^{-1}\mu_i^{2-P}+\mu_i}\right)^{a^{-1}\mu_i^{2-P}}, & y_i=0\\ (1-\omega_i)\dfrac{\Gamma(y_i+a^{-1}\mu_i^{2-P})}{y_i!\Gamma(a^{-1}\mu_i^{2-P})}\left(\dfrac{a^{-1}\mu_i^{2-P}}{a^{-1}\mu_i^{2-P}+\mu_i}\right)^{a^{-1}\mu_i^{2-P}}\left(\dfrac{\mu_i}{a^{-1}\mu_i^{2-P}+\mu_i}\right)^{y_i}, & y_i>0\end{cases}$	$E(Y_i)=(1-\omega_i)\mu_i$ $Var(Y_i)=E(Y_i)(1+a\mu_i^{P-1}+\omega_i\mu_i)$

Model	P.m.f.	Mean and variance
ZIGP-1	$\begin{cases} \omega_i + (1-\omega_i)\exp(-(1-v)\mu_i), & y_i = 0 \\ (1-\omega_i)\dfrac{((1-v)\mu_i)((1-v)\mu_i + vy_i)^{y_i-1}}{y_i!}\exp(-((1-v)\mu_i + vy_i)), & y_i > 0 \end{cases}$	$E(Y_i) = (1-\omega_i)\mu_i$ $Var(Y_i) = E(Y_i)((1-v)^{-2} + \omega_i\mu_i)$
New ZIGP-1	$\begin{cases} \omega_i + (1-\omega_i)\exp\left(-\dfrac{\mu_i}{1+a}\right), & y_i = 0 \\ (1-\omega_i)\dfrac{\mu_i(\mu_i + ay_i)^{y_i-1}}{(1+a)^{y_i} y_i!}\exp\left(-\dfrac{\mu_i + ay_i}{1+a}\right), & y_i > 0 \end{cases}$	$E(Y_i) = (1-\omega_i)\mu_i$ $Var(Y_i) = E(Y_i)((1+a)^2 + \omega_i\mu_i)$
ZIGP-2	$\begin{cases} \omega_i + (1-\omega_i)\exp\left(-\dfrac{\mu_i}{1+a\mu_i}\right), & y_i = 0 \\ (1-\omega_i)\dfrac{\mu_i(\mu_i + a\mu_i y_i)^{y_i-1}}{(1+a\mu_i)^{y_i} y_i!}\exp\left(-\dfrac{\mu_i + a\mu_i y_i}{1+a\mu_i}\right), & y_i > 0 \end{cases}$	$E(Y_i) = (1-\omega_i)\mu_i$ $Var(Y_i) = E(Y_i)((1+a\mu_i)^2 + \omega_i\mu_i)$
New ZIGP-P	$\begin{cases} \omega_i + (1-\omega_i)\exp\left(-\dfrac{\mu_i}{1+a\mu_i^{P-1}}\right), & y_i = 0 \\ (1-\omega_i)\dfrac{\mu_i(\mu_i + a\mu_i^{P-1} y_i)^{y_i-1}}{(1+a\mu_i^{P-1})^{y_i} y_i!}\exp\left(-\dfrac{\mu_i + a\mu_i^{P-1} y_i}{1+a\mu_i^{P-1}}\right), & y_i > 0 \end{cases}$	$E(Y_i) = (1-\omega_i)\mu_i$ $Var(Y_i) = E(Y_i)((1+a\mu_i^{P-1})^2 + \omega_i\mu_i)$

7. TESTS FOR NB AND GP REGRESSION MODELS

7.1 Likelihood Ratio Test (LRT)

Since NB regressions reduce to Poisson regression in the limit as $a \to 0$, the test of overdispersion in Poisson vs. NB-1 regressions and Poisson vs. NB-2 regressions, $H_0 : a = 0$ vs. $H_1 : a > 0$, can be performed using LRT, $T = 2(\ln L_1 - \ln L_0)$, where $\ln L_1$ and $\ln L_0$ are the models' log likelihood under their respective hypothesis. Since the null hypothesis is on the boundary of parameter space, the LRT is asymptotically distributed as half of probability mass at zero and half of chi-square with one degree of freedom (Lawless 1987). In other words, to test the null hypothesis at significance level α, the critical value of chi-square distribution with significance level 2α is used, or reject H_0 if $T > \chi^2_{1-2\alpha,1}$. As an example, for 0.05 significance level, the critical value is $\chi^2_{0.90,1} = 2.7055$ instead of $\chi^2_{0.95,1} = 3.8415$.

A two-sided test can be performed for dispersion (over or underdispersion) in Poisson vs. new GP-1 regressions and Poisson vs. GP-2 regressions, $H_0 : a = 0$ vs. $H_1 : a \neq 0$, using LRT which is asymptotically distributed as a chi-square with one degree of freedom. If we are interested in a one-sided test for overdispersion in Poisson vs. new GP-1 regressions and Poisson vs. GP-2 regressions, $H_0 : a = 0$ vs. $H_1 : a > 0$, we can use signed square root of likelihood ratio statistic (SSR-LRT), $\text{sgn}(a)\sqrt{T} = \text{sgn}(a)\sqrt{2(\ln L_1 - \ln L_0)}$, where $\text{sgn}(.)$ is a sign function which indicates value of 1 when $a > 0$ and value of -1 when $a < 0$. Under H_0, SSR-LRT is asymptotically distributed as a standard Normal distribution.

A two-sided test for testing NB-1 vs. NB-P, NB-2 vs. NB-P, new GP-1 vs. GP-P and GP-2 vs. GP-P regressions can be performed using LRT. The hypothesis are $H_0 : P = 1$ vs. $H_1 : P \neq 1$ or $H_0 : P = 2$ vs. $H_1 : P \neq 2$, and the LRT is asymptotically distributed as a chi-square with one degree of freedom.

7.2 Wald Test

The test of overdispersion in Poisson vs. NB-1 regressions and Poisson vs. NB-2 regressions can also be performed using Wald statistic which is, $\dfrac{\hat{a}^2}{Var(\hat{a})}$, where $\hat{a}$ is the estimate of dispersion parameter and $Var(\hat{a})$ is its variance. Since the null hypothesis is on the boundary of parameter space, the Wald statistic is asymptotically distributed as half of probability mass at zero and half of chi-square with one degree of freedom. For testing dispersion (over or underdispersion) in Poisson vs. GP-1 regressions and Poisson vs. GP-2 regressions, the Wald statistic can also be applied, and the statistic is asymptotically distributed as a chi-square with one degree of freedom.

Using similar approach, the adequacy of NB-1 vs. NB-P, NB-2 vs. NB-P, new GP-1 vs. GP-P and GP-2 vs. GP-P regression can be performed using Wald statistic which is, $\frac{\hat{P}^2}{Var(\hat{P})}$, where $\hat{P}$ is the estimate of functional parameter and $Var(\hat{P})$ is its variance. The Wald statistic is asymptotically distributed as a chi-square with one degree of freedom.

7.3 Vuong Test

For non-nested models, a comparison between models with p.m.f. $p_1(.)$ and $p_2(.)$ can be performed using Vuong test (Vuong 1989), $V = \frac{\bar{m}\sqrt{n}}{sd(m)}$, where $\bar{m}$ is the mean of m_i, $sd(m)$ is the standard deviation of m_i, n the sample size and $m_i = \ln\left(\frac{p_{1i}(y_i)}{p_{2i}(y_i)}\right)$. The Vuong test statistic follows a standard normal. As an example, for 0.05 significance level, the first model is "closer" to the actual model if V is larger than 1.96. In the other hand, the second model is "closer" to the actual model if V is smaller than -1.96. Otherwise, neither model is "closer" to the actual model and there is no difference between using the first or the second model.

For models with unequal number of parameters, the equation for m_i in Vuong test is slightly modified to account for the difference in the number of parameters, $m_i = \ln\left(\frac{p_{1i}(y_i)}{p_{2i}(y_i)}\right) - \frac{k_1 - k_2}{2}\ln(n)$, where k_1 and k_2 are the number of parameters in model 1 and model 2 respectively.

7.4 AIC and BIC

When several models are available, one can compare the models' performance based on several likelihood measures which have been proposed in statistical literatures. Two of the most regularly used measures are Akaike Information Criteria (AIC) and Bayesian Schwartz Information Criteria (BIC). The AIC penalizes a model with larger number of parameters, and is defined as $\text{AIC} = -2\ln L + 2p$, where $\ln L$ denotes the fitted log likelihood and p the number of parameters. The BIC penalizes a model with larger number of parameters and larger sample size, and is defined as $\text{BIC} = -2\ln L + p\ln(n)$, where $\ln L$ denotes the fitted log likelihood, p the number of parameters and n the sample size.

8 TESTS FOR ZINB AND ZIGP REGRESSION MODELS

8.1 Likelihood Ratio Test

Since ZINB regressions reduce to ZIP regression in the limit as $a \to 0$, the test of overdispersion in ZIP vs. ZINB-1 regressions and Poisson vs. ZINB-2 regressions, $H_0 : a = 0$ vs. $H_1 : a > 0$, can be performed using LRT. Since the null hypothesis is on the boundary of parameter space, the LRT statistic is asymptotically distributed as half of probability mass at zero and half of chi-square with one degree of freedom (Stram and Lee 1994, 1995).

A two-sided test can be performed to test dispersion (over or underdispersion) in ZIP vs. new ZIGP-1 regressions and ZIP vs. ZIGP-2 regressions using LRT, $H_0 : a = 0$ vs. $H_1 : a \neq 0$. Since ZIP regression model is nested within new ZIGP-1 and ZIGP-2 regressions, the boundary problem does not exist and the statistic is asymptotically distributed as a chi-square with one degree of freedom. If we are interested in a one-sided test for overdispersion in ZIP vs. new ZIGP-1 regressions and ZIP vs. ZIGP-2 regressions, $H_0 : a = 0$ vs. $H_1 : a > 0$, we can use SSR-LRT which is asymptotically distributed as a standard Normal distribution.

A two-sided test for adequacy of ZINB-1 vs. ZINB-P, ZINB-2 vs. ZINB-P, new ZIGP-1 vs. ZIGP-P and ZIGP-2 vs. ZIGP-P regressions can be performed using LRT, where the hypothesis are $H_0 : P = 1$ vs. $H_1 : P \neq 1$ or $H_0 : P = 2$ vs. $H_1 : P \neq 2$. Since both ZINB-1 and ZINB-2 regressions are nested within ZINB-P regression, and both ZIGP-1 and ZIGP-2 regressions are nested within ZIGP-P regression, the LRT statistic is asymptotically distributed as a chi-square with one degree of freedom.

8.2 Wald Test

The test of overdispersion in ZIP vs. ZINB-1 regressions and ZIP vs. ZINB-2 regressions can also be performed using Wald statistic. Since the null hypothesis is on the boundary of parameter space, the Wald statistic is asymptotically distributed as half of probability mass at zero and half of chi-square with one degree of freedom. For dispersion (over or underdispersion) in ZIP vs. ZIGP-1 regressions and ZIP vs. ZIGP-2 regressions, the Wald statistic can also be applied, where the statistic is asymptotically distributed as a chi-square with one degree of freedom.

Using similar approach, the adequacy of ZINB-1 vs. ZINB-P, ZINB-2 vs. ZINB-P, new ZIGP-1 vs. ZIGP-P and ZIGP-2 vs. ZIGP-P regressions can be performed using Wald statistic, where the statistic is asymptotically distributed as a chi-square with one degree of freedom.

9. EXAMPLES

9.1 Malaysian Own Damage Claim Counts

The dataset for private car Own Damage (OD) claim counts analyzed in Zamani and Ismail (2012) is reconsidered here for fitting NB and GP regression models. The data was based on 1.01 million private car policies for a three-year period of 2001-2003, the exposures were expressed in car-year units, and the incurred claims consisted of claims already paid as well as outstanding. Table 3 shows the rating factors and rating classes for the exposures and incurred claims. The estimates of Poisson regression were used as initial values for fitting NB and GP regressions.

Table 3: Rating factors and rating classes (Malaysian OD data)

Rating factors	Rating classes
Vehicle age	0-1 year
	2-3 years
	4-5 years
	6-7 years
	8+ years
Vehicle c.c.	0-1000
	1001-1300
	1301-1500
	1501-1800
	1801+
Vehicle make	Local type 1
	Local type 2
	Foreign type 1
	Foreign type 2
	Foreign type 3
Location	North
	East
	Central
	South
	East Malaysia

Table 4 shows the parameter estimates and their *t*-ratios for the fitted models. The best Poisson model was chosen using backward stepwise based on both AIC (chose model 2 if $AIC_{\text{model 2}} < AIC_{\text{model 1}}$) and *p*-values (drop a covariate if it is not significant). The same Poisson covariates were then utilized for fitting NB and GP regressions to ensure that the LRT can be performed. The results in Table 4 indicate that the regression parameters for all models have similar estimates. As expected, NB and GP models provide similar inferences for the regression parameters, i.e. their t-ratios, in absolute value, are smaller than Poisson model. In particular, the estimates and *t*-ratios of NB-1 are closer to GP-1, and similar results are also observed between NB-2 and GP-2, and between NB-P and GP-P. These results are expected since the mean-variance relationships for NB-1 and GP-1 are linear, NB-2 is quadratic, GP-2 is cubic, and both NB-P and GP-P are in the *p*th power and the (2*p*-1)th power respectively.

For testing overdispersion in Poisson versus NB-1 regressions, $H_0 : a = 0$ vs. $H_1 : a > 0$, the likelihood ratio and the Wald *t* respectively are 2[-2182.75-(-3613.71)]=2861.92 and 14.44, indicating that the null hypothesis is rejected and the NB-1 is more adequate. The likelihood ratio and Wald *t* for testing overdispersion in Poisson versus NB-2, Poisson versus GP-1 and Poisson versus GP-2 regressions are (2840.50, 2896.96, 2691.08) and (9.93, 19.22, 12.27) respectively, also indicating that the data is overdispersed and NB-2, GP-1 and GP-2 regressions are better than Poisson regression.

Table 4: NB and GP regression models (Malaysian OD data)

Parameters	Poisson		NB-1		NB-2		NB-P		GP-1		GP-2		GP-P	
	Est.	*t*-ratio	Est.	*t*-ratio	Est.	*t*-ratio	Est.	*t*-ratio	Est.	*t*-ratio	Est.	*t*-ratio	Est.	*t*-ratio
Intercept	-3.04	-195.43	-3.05	-70.55	-3.20	-45.78	-3.09	-54.19	-3.06	-67.78	-3.17	-48.22	-3.09	-54.03
2-3 year	0.51	41.09	0.53	15.07	0.57	8.75	0.54	11.36	0.53	14.62	0.58	9.11	0.55	11.47
4-5 year	0.52	39.96	0.51	14.03	0.52	8.02	0.52	10.70	0.51	13.29	0.51	8.10	0.52	10.56
6-7 year	0.43	33.63	0.45	12.55	0.40	6.14	0.44	9.19	0.46	12.20	0.37	5.92	0.45	9.34
8+ year	0.24	19.11	0.24	6.84	0.27	4.19	0.25	5.40	0.24	6.55	0.29	4.48	0.26	5.45
1001-1300 cc	-0.31	-24.64	-0.28	-8.30	-0.12	-2.03	-0.24	-5.20	-0.27	-7.69	-0.08	-1.36	-0.23	-5.05
1301-1500 cc	-0.16	-14.83	-0.16	-5.17	0.10	1.69	-0.12	-2.86	-0.15	-4.81	0.19	3.03	-0.12	-3.04
1501-1800 cc	0.14	12.93	0.13	4.36	0.25	4.61	0.15	3.90	0.13	4.08	0.27	5.18	0.14	3.69
1801+ cc	0.11	10.83	0.11	3.77	0.33	5.92	0.16	4.05	0.11	3.64	0.34	6.05	0.15	3.95
Local type 2	-0.46	-32.41	-0.45	-11.55	-0.27	-4.04	-0.40	-7.94	-0.45	-11.07	-0.35	-5.47	-0.41	-8.23
Foreign type 1	-0.20	-19.17	-0.18	-6.12	-0.27	-5.38	-0.20	-5.32	-0.17	-5.54	-0.34	-6.79	-0.20	-5.16
Foreign type 2	0.18	11.50	0.21	4.80	0.34	6.34	0.26	5.19	0.22	4.96	0.31	5.93	0.25	5.10
Foreign type 3	-	-	-	-	-	-	-	-	-	-	-	-	-	-
East	0.35	19.91	0.39	8.29	0.31	4.71	0.36	6.28	0.41	8.37	0.28	4.83	0.37	6.56
Central	0.32	29.43	0.31	10.21	0.31	5.28	0.31	7.50	0.31	9.63	0.28	4.76	0.31	7.39
South	0.26	20.40	0.26	7.20	0.36	5.87	0.30	6.36	0.26	6.92	0.39	6.77	0.30	6.42
East Malaysia	0.13	8.87	0.12	3.04	0.11	1.80	0.12	2.26	0.12	2.80	0.10	1.77	0.11	2.15
a	-	-	7.07	14.44	0.14	9.93	1.57	6.60	2.01	19.22	0.02	12.27	0.64	7.95
p	-	-	1.00	-	2.00	-	1.39	39.00	1.00	-	2.00	-	1.27	46.23
Log likelihood		-3613.71		-2182.75		-2193.46		-2114.38		-2165.23		-2268.17		-2111.91
AIC		7259.42		4399.50		4420.92		4264.77		4364.45		4570.33		4259.83
BIC		7328.30		4472.67		4494.10		4342.25		4437.63		4643.51		4337.31

For testing the adequacy of NB-1 against NB-P and GP-1 against GP-P regressions, $H_0 : P = 1$ vs. $H_1 : P \neq 1$, the likelihood ratio and Wald *t* respectively are (136.74, 106.64) and (39.00, 46.23), implying that the null hypothesis is rejected and both NB-P and GP-P are more adequate. The likelihood ratios for testing adequacy of NB-2 versus NB-P and GP-2 versus GP-P regressions, $H_0 : P = 2$ vs. $H_1 : P \neq 2$, are (158.16, 312.52), also indicating that both NB-P and GP-P are better models.

Based on AIC and BIC, GP-P model has the lowest value for both criteria, indicating that the GP-P is the best model. However, based on Vuong test between GP-P as the first model and NB-P as the second model, the statistic is 1.0647 (less than 1.96), indicating that neither model is preferred over the other.

9.2 German Healthcare Data

The German Socioeconomic Panel (GSOEP) data (Riphahn et al. 2003) which was analyzed in Zamani and Ismail (2013a) is reconsidered here for fitting ZINB and ZIGP regression models. For an illustration purpose, only the first 438 individual data were fitted, where the response variable is the number of doctor visit in the last three month and the covariates were gender, age, health satisfaction, marital status, working status and education years. Table 5 shows the mean and standard deviation of the selected variables.

Table 5: Descriptive summary (German healthcare data)

Variable	Measurement	Mean	Standard deviation
DOCVIS	Number of doctor visit in last three months	2.93	33.09
GENDER	Female=1; Male=0	0.51	0.25
AGE	Age in years	43.30	116.23
HSAT	Health satisfaction	6.84	4.89
MARRIED	Married=1; else=0	0.54	0.25
WORKING	Employed=1; else=0	0.79	0.17
EDUC	Years of schooling	12.45	9.77

As mentioned previously, the covariates of μ_i and ω_i for ZIP, ZINB and ZIGP regressions can be included via the log and logit link functions, as shown in (5), where the vectors $\mathbf{x}_i$ and $\mathbf{z}_i$ may or may not share the same components. For the case where the covariates are not utilized in ω_i, the log and logit functions can be rewritten as (Yip and Yau, 2005; Ozmen and Famoye, 2007),

$$\log(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta} \text{ and } \log\left(\frac{\omega}{1-\omega}\right) = \tau . \qquad (6)$$

The data were fitted to ZIP, ZINB and ZIGP regressions using link functions (6), and the estimates of ZIP regression were used as initial values for fitting ZINB and ZIGP regressions.

Table 6 shows the parameter estimates and their t-ratios for the fitted models. The best ZIP model was chosen based on *p*-values (drop a covariate if it is not significant). The same ZIP covariates were then utilized for fitting ZINB and ZIGP models to ensure that the LRT can be performed. Since ZINB-P model did not provide converged solutions, the results are not displayed in the table. The results in Table 6 indicate that the regression parameters for all models have similar estimates. As expected, both ZINB and ZIGP regressions provide similar inferences for the regression parameters, i.e. their t-ratios, in absolute value, are smaller than ZIP.

For testing overdispersion in ZIP versus ZINB-1 and ZIP versus ZINB-2 regressions, $H_0 : a = 0$ vs. $H_1 : a > 0$, the likelihood ratio and Wald *t* respectively are (393.40, 525.30) and (6.16, 3.99), indicating that the null hypothesis is rejected and both ZINB-1 and ZINB-2 regressions are more adequate. The Vuong test for choosing between non-nested models of ZINB-2 as the first model and ZINB-1 as the second model is 5.5287 (more than 1.96), indicating that the ZINB-2 is a better model.

The likelihood ratio and Wald *t* for testing overdispersion in ZIP versus ZIGP-1 and ZIP versus ZIGP-2 regressions are (524.00, 521.44) and (7.68, 6.31) respectively, also indicating that the data is overdispersed, and both ZIGP-1 and ZIGP-2 are better models than ZIP. For testing the adequacy of ZIGP-1 against ZIGP-P regressions, $H_0 : P = 1$ vs. $H_1 : P \neq 1$, and ZIGP-2 versus ZIGP-P regressions, $H_0 : P = 2$ vs. $H_1 : P \neq 2$, the likelihood ratio and Wald *t* respectively are (7.34, 9.90) and 9.22, implying that the null hypothesis is rejected and ZIGP-P is more adequate.

Table 6: ZIP, ZINB and ZIGP regression models (German healthcare data)

Parameter	ZIP		ZINB-1		ZINB-2		ZIGP-1		ZIGP-2		ZIGP-P	
	est.	*t*-ratio	est.	*t*-ratio	est.	*t*-ratio	est.	*t*-ratio	est.	*t*-ratio	est.	*t*-ratio
Intercept	2.49	26.26	2.33	13.46	2.29	8.00	2.33	10.10	2.34	8.36	2.42	9.13
GENDER	0.30	4.85	0.21	1.90	0.58	3.83	0.52	3.73	0.52	3.52	0.60	3.90
HSAT	-0.22	-18.67	-0.16	-7.23	-0.25	-7.95	-0.22	-8.19	-0.24	-7.44	-0.25	-8.10
MARRIED	0.26	4.25	0.14	1.42	0.20	1.37	0.13	1.06	0.18	1.23	0.17	1.17
WORKING	0.15	2.35	0.13	1.13	0.05	0.27	-0.06	-0.44	0.14	0.81	0.00	-0.02
τ	-0.33	-3.16	-0.22	-2.28	-1.69	-2.65	-1.38	-4.92	-0.99	-4.25	-1.29	-4.49
a	-	-	2.53	6.16	1.39	3.99	1.52	7.68	0.33	6.31	0.77	3.80
P	-	-	1.00	-	2.00	-	1.00	-	2.00	-	1.46	9.22
Log likelihood	-1136.48		-939.78		-873.83		-874.48		-875.76		-870.81	
AIC	2284.96		1893.56		1761.66		1762.97		1765.52		1757.62	
BIC	2309.45		1922.13		1790.24		1791.54		1794.09		1790.28	

Based on AIC and BIC, ZIGP-P model has the lowest AIC but ZINB-2 model has the lowest BIC. For chosing between ZINB-2 as the first model and ZIGP-P as the second model which involves non-nested models with different number of parameters, the Vuong test statistic is 1.7303 (less than 1.96), indicating that neither model is preferred over the other.

Table 7 shows the parameters, log likelihood, AIC and BIC for the best ZIGP-P and ZINB-2 models, chosen based on *p*-values (drop a covariate if it is not significant). Based on AIC and BIC, ZIGP-P model has the lowest value for both criteria. However, the Vuong test statistic for comparing between ZINB-2 as the first model and ZIGP-P as the second model, which are non-nested models, is 1.6803 (less than 1.96), indicating that neither model is preferred over the other.

Table 7: ZINB-2 and ZIGP-P models with significant covariates (German healthcare data)

Parameter	ZINB-2			ZIGP-P		
	est.	*t*-ratio	*p*-value	est.	*t*-ratio	*p*-value
Intercept	2.46	9.87	0.00	2.53	11.01	0.00
GENDER	0.55	3.68	0.00	0.59	3.91	0.00
HSAT	-0.26	-8.16	0.00	-0.26	-8.27	0.00
τ	-1.75	-2.67	0.00	-1.28	-4.72	0.00
a	1.43	4.11	0.00	0.78	3.94	0.00
P	2.00	-	-	1.46	9.31	0.00
Log likelihood	-874.77			-871.50		
AIC	1759.54			1755.01		
BIC	1779.95			1779.50		

10. CONCLUSIONS

This study has related NB and GP regressions through the mean-variance relationship and has shown applications of these models for overdispersed count data. In addition, this study has related ZINB and ZIGP regressions through the mean-variance relationship and has shown applications of these zero-inflated models for zero-inflated and overdispersed count data.

The Malaysian data for private car Own Damage (OD) claim counts analyzed in Zamani and Ismail (2012) has been reconsidered for fitting Poisson, NB and GP regression models. The results indicate that the regression parameters of all models have similar estimates and the t-ratios, in absolute value, for NB and GP models are smaller than Poisson model. The likelihood ratio and Wald t for testing overdispersion indicate that the data is overdispersed, and NB-1, NB-2, GP-1 and GP-2 models are better than Poisson model. The likelihood ratio and Wald t also imply that both NB-P and GP-P are the best two models. Based on AIC and BIC, GP-P model has the lowest value for both criteria. However, based on Vuong test between GP-P and NB-P models, neither model is preferred over the other.

The German Socioeconomic Panel (GSOEP) data (Riphahn et al. 2003) which was analyzed in Zamani and Ismail (2013a) has been reconsidered for fitting ZIP, ZINB and ZIGP regression models. Unfortunately, ZINB-P model did not provide converge solutions and the results were not displayed. For other models, the results indicate that the regression parameters of all models have similar estimates and the t-ratios, in absolute value, for ZINB and ZIGP models are smaller than ZIP model. The likelihood ratio and Wald t for testing overdispersion indicate that the data is overdispersed, and ZINB-1, ZINB-2, ZIGP-1 and ZIGP-2 models are better than ZIP model. For ZINB models, the Vuong test indicates that ZINB-2 regression is better than ZINB-1 regression, whereas for ZIGP models, the likelihood ratio and Wald t indicate that ZIGP-P regression is better than both ZIGP-1 and ZIGP-2 regressions. Based on AIC and BIC, ZIGP-P model has the lowest AIC but ZINB-2 model has the lowest BIC. However, based on Vuong test between ZINB-2 and ZIGP-P regressions, neither model is preferred over the other.

Several general comparisons can be made regarding NB and GP regression models. Firstly, NB regression reduces to Poisson regression in the limit as $a \to 0$, whereas GP regression reduces to Poisson regression when $a = 0$. Secondly, the mean-variance relationships of NB-1 and GP-1 are linear, NB-2 is quadratic, GP-2 is cubic, NB-P is to the pth power, and GP-P is to the $(2p-1)$th power. Thirdly, LRT can be performed for testing overdispersion in Poisson vs. NB regressions, where the statistic is asymptotically distributed as a mixture of half of probability mass at zero and half of chi-square with one degree of freedom (Lawless 1987). LRT can also be performed for testing dispersion (over or underdispersion) in Poisson vs. GP regressions, where the statistic is asymptotically distributed as a chi-square with one degree of freedom (Consul and Famoye 1992, Wang and Famoye 1997). And finally, in terms of properties, there is no big difference between NB and GP distributions when the means and variances are fixed. However, GP distribution has heavier tail, whereas NB distribution has more mass at zero (Joe and Zhu 2005).

Several general comparisons can also be made regarding ZINB and the ZIGP models. Firstly, ZINB regression reduces to ZIP regression in the limit as $a \to 0$, whereas ZIGP regression reduces to ZIP regression when $a = 0$. Secondly, LRT can be performed for testing overdispersion in ZIP vs. ZINB regressions, where the statistic is asymptotically distributed as a mixture of half of probability mass at zero and half of chi-square with one degree of freedom (Stram and Lee 1994, 1995). SSN-LRT can be performed for testing overdispersion in ZIP vs. ZIGP regressions, where the statistic is asymptotically distributed as a standard normal (Yang et al. 2009b, Famoye and Singh 2006). And finally, in terms of properties, GP distribution has heavier tail whereas NB distribution has more mass at zero, indicating that both ZINB and ZIGP regressions can be used for fitting zero-inflated and overdispersed count data (Joe and Zhu 2005).

In this study, we have fitted a variety of models to two different datasets, involving several forms of NB, GP, ZINB and ZIGP regressions. The selection of the best model depends on many considerations. First, we can check for overdispersion, and if the data is slightly overdispersed, we can fit the data to quasi-Poisson regression. Secondly, if our data is largely overdispersed and it is not caused by excessive zeros but due to variation in the data, we can fit NB and GP regressions. Thirdly, if our data is both overdispersed and zero-inflated, we can fit zero-inflated (ZINB, ZIGP) and hurdle (HNB, HGP) regressions. Nevertheless, the hurdle regression models were not discussed in this study. Finally, the choice between zero-inflated (ZINB, ZIGP) and hurdle (HNB, HGP) models should be based upon a priori knowledge of the cause of excessive zeros in the data. Zero-inflated models are interpreted as a mix of structural and sampling zeros from two processes; the process that generates excess zeros from a binary distribution which are the structural zeros, and the process that generates both non-negative and zero counts from Poisson or NB distributions which are the sampling zeros. In the other hand, hurdle models assume that all zeros are sampling zeros. Therefore, as a crude guideline, if occurrences of count event do not depend on any condition and may occur at any time, the hurdle models should be fitted. However, if occurrences of count events depend on specific conditions and/or time, such as the case of deductible or no claim discount in insurance data, the zero-inflated models are more appropriate.

Acknowledgement

The authors gratefully acknowledge the financial support received in the form of research grants (UKM-ST-06-FRGS0108-2009 and GUP-2012-024) from the Ministry of Higher Education (MOHE), Malaysia. The authors are also pleasured to thank Insurance Services Malaysia Berhad (ISM) for supplying the data.

REFERENCES

[1.] Agarwal, D.K., Gelfand, A., Citron-Pousty, S. 2002. Zero-inflated model with application to spatial count data. *Environmental and Ecological Statistics*. 9: 341-355.

[2.] Aitkin, M., Anderson, D., Francis, B., Hinde, J. 1990. *Statistical modelling in GLIM*. New York: Oxford University Press.

[3.] Bohning, D., Dietz, E., Schlattman, P., Mendonca, L., Kirchner, U. 1996. The zero-inflated Poisson model and the decayed, missing and filled teeth index in dental epidemiology. *Journal of the Royal Statistical Society, Series A*. 162: 195-209.

[4.] Boucher, J.P., Denuit, M., Guillen, M. 2007. Risk classification for claim count: a comparative analysis of various zero-inflated mixed Poisson and hurdle models. *North American Actuarial Journal*. 11(4): 110-131.

[5.] Brockmann, M.J., Wright, T.S. 1992. Statistical motor rating: making effective use of your data*, Journal of the Institute of Actuaries*. 119(3): 457-543.

[6.] Cameron, A.C., Trivedi, P.K. 1986. Econometric models based on count data: comparisons and applications of some estimators and tests. *Journal of Applied Econometrics*. 1: 29-53.

[7.] Cameron, A.C., Trivedi, P.K. 1998. *Regression Analysis of Count Data*. New York: Cambridge University Press.

[8.] Consul, P.C. 1989. *Generalized Poisson Distribution: Properties and Application*. New York: Marcel Dekker.

[9.] Consul, P.C., Famoye, F. 1992. Generalized Poisson regression model. *Communications in Statistics (Theory & Method)*. 2(1): 89-109.

[10.] Consul, P.C., Jain, G.C. 1973. A generalization of the Poisson distribution. *Technometrics*. 15: 791-799.

[11.] Conway, R. W., Maxwell, W. L. 1962. A queuing model with state dependent service rates. *Journal of Industrial Engineering*. 12: 132–136

[12.] Denuit, M., Marechal, X., Pitrebois, S., Walhin, J.F. 2007. *Actuarial Modeling of Claim Counts: Risk Classification, Credibility and Bonus-Malus Systems*. John Wiley and Sons: England.

[13.] Famoye, F., Wulu, J.T., Singh, K.P. 2004. On the generalized Poisson regression model with an application to accident data. *Journal of Data Science*. 2: 287-295.

[14.] Famoye, F., Singh, K.P. 2006. Zero-inflated generalized Poisson regression model with an application to domestic violence data. *Journal of Data Science*. 4: 117-130.

[15.] Greene, W. 2008. Functional forms for the negative binomial model for count data. *Economics Letters*. 99: 585-590.

[16.] Hilbe, J. 2007. *Negative Binomial Regression*. Cambridge, UK: Cambridge University Press.

[17.] Ismail, N., Jemain, A.A. 2007. Handling overdispersion with negative binomial and generalized Poisson regression models. *Casualty Actuarial Society Forum*. Winter: 103-158.

[18.] Joe, H., Zhu, R. 2005. Generalized Poisson distribution: the property of mixture of Poisson and comparison with negative binomial distribution. *Biometrical Journal*. 47: 219-229.

[19.] Lambert, D. 1992. Zero-inflated Poisson regression, with an application to random defects in manufacturing. *Technometrics*. 34: 1-14.

[20.] Lawless, J.F. 1987. Negative binomial and mixed Poisson regression. *Canadian Journal of Statistics*. 15(3): 209-225.

[21.] Lord, D., Guikema, S.D., Geedipally, S.R. 2008. Application of the Conway–Maxwell–Poisson generalized linear model for analyzing motor vehicle crashes. *Accident Analysis & Prevention*. 40(3): 1123–1134.

[22.] Lord, D., Geedipally, S.R., Guikema, S.D. 2010. Extension of the application of Conway-Maxwell-Poisson models: analyzing traffic crash data exhibiting under-dispersion. *Risk Analysis*. 30(8): 1268-1276.

[23.] McCullagh, P., Nelder, J.A. 1989. *Generalized Linear Models (2nd Edition)*. Chapman and Hall: London.

[24.] Neelon, B.H., O'Malley, A.J., Normand, S.T. 2010. A Bayesian model for repeated measures zero-inflated count data with application to outpatient psychiatric service use. *Statistical Modelling*. 10(4): 421–439.

[25.] Ozmen, I., Famoye, F. 2007. Count regression models with an application to zoological data containing structural zeros. *Journal of Data Science*. 5: 491-502.

[26.] Renshaw, A.E., 1994. Modelling the claims process in the presence of covariates. *ASTIN Bulletin*. 24(2): 265-285.

[27.] Ridout, M.S., Hinde, J.P., Demetrio, C.G.B. 2001. A score test for testing a zero-inflated Poisson regression model against zero-inflated negative binomial alternatives. *Biometrics*. 57: 219-223.

[28.] Riphahn, R., Wambach, A., Million, A. 2003. Incentive effects in the demand for health care: a bivariate panel count data estimation. *Journal of Applied Econometrics*. 18(4): 387-405.

[29.] Shmueli, G., Minka T., Kadane, J.B., Borle, S., Boatwright, P.B. 2005. A useful distribution for fitting discrete data: revival of the Conway–Maxwell–Poisson distribution. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 54: 127–142.

[30.] Stram, D.O., Lee, J.W. 1994. Variance components testing in the longitudinal mixed effects model. *Biometrics*. 50: 1171-1177.

[31.] Stram, D.O., Lee, J.W. 1995. Correction to "Variance components testing in the longitudinal mixed effects model". *Biometrics*. 51: 1196.
[32.] Vuong, Quang H. 1989. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*. 57(2): 307–333.
[33.] Wang, W., Famoye, F. 1997. Modeling household fertility decisions with generalized Poisson regression. *Journal of Population Economics*. 10: 273-283.
[34.] Winkelmann, R. 2008. *Econometric Analysis of Count Data*. Heidelberg: Springer Verlag.
[35.] Yang, Z., Hardin, J.W., Addy, C.L. Vuong, Q.H. 2007. Testing approaches for overdispersion in Poisson regression versus the generalized Poisson model. *Biometrical Journal*. 49: 565-584.
[36.] Yang, Z., Hardin, J.W., Addy, C.L. 2009a. A score test for overdispersion in Poisson regression based on the generalized Poisson-2 model. *Journal of Statistical Planning and Inference*. 139: 1514-1521.
[37.] Yang, Z., Hardin, J.W., Addy, C.L. 2009b. Testing overdispersion in the zero-inflated Poisson model. *Journal of Statistical Planning and Inference*. 139: 3340-3353.
[38.] Yip, K.C.H., Yau, K.K.W. 2005. On modeling claim frequency data in general insurance with extra zeros. *Insurance: Mathematics and Economics*. 36: 153-163.
[39.] Zamani, H., Ismail, N. 2012. Functional form for the generalized Poisson regression model. *Communications in Statistics (Theory and Methods)*. 41(20): 3666–3675.
[40.] Zamani, H., Ismail, N. 2013a. Functional Form for the zero-inflated generalized Poisson regression model. *Communications in Statistics (Theory and Methods)*. (in press)
[41.] DOI: 10.1080/03610926.2012.665553.
[42.] Zamani, H., Ismail, N. 2013b. Score tests for overdispersion in Poisson regression model against generalized Poisson alternatives and zero-inflated Poisson regression model against zero-inflated generalized Poisson alternatives. (submitted).

Biographies of Authors

Noriszura Ismail is an Associate Professor in Universiti Kebangsaan Malaysia, teaching Actuarial Science courses since July 1993, and is currently pursuing research activities in Actuarial Science and Statistics areas with her PhD students. She received her PhD (Statistics) in 2006 from Universiti Kebangsaan Malaysia, and MSc (Actuarial Science) and BSc (Actuarial Science) in 1993 and 1991 from University of Iowa.

Hossein Zamani is a lecturer in Hormozgan University, Iran, teaching Statistics courses and is currently pursuing research activities in Statistics. He received his PhD (Statistics) in 2011 from Universiti Kebangsaan Malaysia, MSc (Statistics) in 2004 from Shiraz University, Iran and BSc (Statistics) in 2002 from Razi University, Iran.

Extending the Asset Share Model: Recognizing the Value of Options in P&C Insurance Rates

Greg McNulty

Abstract: In this paper we will present a refinement of the well-known asset share model for ratemaking. The new method for calculating premiums and premium relativities accounts for risk classification transition probabilities. The relationship between risk class transition and options on insurance coverage is discussed. Some simple examples will be worked which will demonstrate how risk class transition can cause problems for the traditional asset share model which are remedied by our extended asset share model. We will also show how the new method can be used to determine the price for insurance policies with the popular "accident forgiveness" feature.

Keywords. Ratemaking, risk classification, asset share model, option pricing

1. INTRODUCTION

Many P&C actuaries are familiar with the asset share model through the work of Sholom Feldblum, [1]. The method demonstrates how lower prices can be justified for more loyal customers by essentially crediting some of the expected profits on renewal terms to the first term. However, there is a disconnect between the theory and practice. The asset share model assumes a constant (or perhaps gradually inflating) premium over the life of the policy, whereas in practice the rate may change at policy renewal, sometimes drastically, as more information about the insured is revealed. In this paper we will present a method for calculating premiums and premium relativities that accounts for risk classification transition probabilities.

We begin in Section 2 with a discussion of the merits and shortcomings of the established asset share model. Section 3 will present a very simple example for ratemaking with two risk classes. Policies will be limited to one renewal, but the twist is that policies can change from one risk class to another. We will show that in carefully accounting for risk class transition, we arrive at different indicated rates than both the traditional method and the standard asset share model.

In Section 4 we will discuss options, and demonstrate that the difference between the extended asset share method developed here and the standard asset share model is that the assumption of constant premium inflation causes the standard model to include the cost of options on future coverage in the indicated prices. The extended model allows us to correctly price the "floating rate" policy, and separately determine the cost of the options.

In Section 5 we will present the full mathematical framework for the extended asset share model and in Section 6 we will apply the extended asset share model in two worked examples, including pricing insurance policies with the "accident forgiveness" feature popular in personal lines insurance. We will conclude in Section 7 with some additional considerations

2. THE STANDARD ASSET SHARE MODEL

Hopefully the reader is familiar with Feldblum's paper [1] on the asset share model which presents a method for pricing insurance contracts by looking at lifetime expected premium, loss and expense. Traditional methods view only a single contract term. The mechanics of the model are straightforward: for the first and all subsequent renewal terms estimate the losses and expenses expected to be incurred. An estimate of the probability of the policy renewing is also required. The required premium can then be calculated using the traditional premium formula:

$$P = \frac{F + L}{1 - U - V}$$

where P, F, and L represent the expected lifetime present value of premium, fixed expense and loss, respectively, while V is the variable cost ratio and U is the desired underwriting profit provision as a percent of premium. Feldblum uses examples to demonstrate how differing probabilities of renewal combined with decreasing loss costs as policy tenure increases can cause optimal premium relativities to be very different from loss relativities; these were traditionally assumed to be the same except for fixed expense loading.

2.1 Why the Asset Share Model is an Improvement

The asset share model is a major improvement over the traditional method of ratemaking. It has been well documented, in the paper by Wu and Lin [4], that average loss costs improve as policyholder tenure increases. The paper shows that this phenomenon is present in most lines of business, not just private passenger auto which was the subject of the Feldblum paper.

So why not just charge the traditional premium for each term of a policy? Premiums will be higher for the first term and decrease from there, with the company securing an expected profit for every term of a customer's time with the company. Consumer behavior gives insurers an opportunity for a better strategy: charge a lower premium to new customers when they are more sensitive to price, and charge higher than actuarial premiums for renewal terms when customers are likely to renew and less sensitive to price. Note that this is not necessarily irrational consumer behavior; it is simply a matter of preferring cash now (in the form of a lower new business policy premium) to a stream of contingent future cashflows (renewal discounts which are realized only if the policy is renewed). The asset share model determines the required premium under the better strategy.

2.2 Shortcomings of the Asset Share Model

Another well documented phenomenon in insurance is that past loss experience is indicative of future loss experience. Ratemaking schemes in personal lines, commercial lines and reinsurance businesses all charge higher premiums to policyholders who experience losses, or greater than expected losses. In the asset share model it is assumed that premium gradually inflates over time so that if a customer is charged a premium P at inception, they will be expected to pay $(1+r)^n * P$ for the nth policy term, where r is the rate of inflation. In actuality, a cohort of policies that were identical at inception will, due to experience rating, be paying a wide range of premiums for their future renewal policies.

To see why this is problematic, think about how we would use the asset share model to determine optimum relativities for experience rating factors in personal lines. We would set up tables of expected loss, expenses, renewal probabilities and premium for two groups, "Low" and "High" expected future loss. There would be two independent variables, the premium for the low group and the premium for the high group. We could determine the correct premium for each and the differential between the two would be our new experience rating factor. Suppose the new factor amounts to a 20% increase in premium for "high" risk customers. The asset share model we set up is likely no longer valid because renewal rates should decrease with increasing premiums. Also, the expected inflation trend in the premium for the low group should increase because the small fraction of them that receive the "high risk" premium on renewal after having adverse experience will pay 20% higher premiums. It may be possible to iterate the process and come to an answer, but it seems that this may not be the best solution.

The assumption of constantly inflating premiums can lead us astray in other scenarios as well. Suppose we use the asset share model to determine optimal premium relativities for new drivers versus those with 5 years of experience, and we determine that premiums for new drivers need to be 30% higher over the life of the policy. What happens when the driver attains 5 years of experience and wishes to renew the policy? In most cases the company would need to charge the customer the lower "experienced driver" rate. The original asset share model is now invalidated because it was counting on profitable policy years six and beyond priced at the higher rate to subsidize expected underwriting losses on the first few years of the policy. Even if the new rate is consistent with an asset share model somehow recalculated for this risk, we can never go back in time and charge more for policy terms that are already over for which our rate was inadequate. This error would lead to the company systematically undercharging customers over their lifetimes and poor profitability.

2.3 Remedy for the Above Issues

The problems noted above are just symptoms of the fact that a group of policies that were identical in terms of risk classification as new customers will have a range of different characteristics at renewal and be charged different premiums. The solution is simply to track the movement of policies into different risk classes as they renew. Note that "risk classifications" need not be limited to different expected losses, but also renewal rates, loss cost trends or any other parameter that will have an effect on premium.

A helpful analogy is the binomial model of stock movements from basic finance. A stock starts at time $t = 0$ at price S_0, and at time 1 it increases to price $S_0\ e^{r+s}$ with probability p or decreases to price $S_0\ e^{r-s}$ with probability $1 - p$.

For insurance contract pricing purposes we will need several additional pieces of information at each "node": premium, loss, expense and probability of moving into each of the next nodes. Since the policy may not renew, the probabilities need not add to 1. We will refer to this method as either the extended asset share model, or the option model.

3. SIMPLE EXAMPLE

Let's illustrate a simple example to show how the extended model works. Importantly, we'll also show that the extended model indicates a different premium amount and relativities even in the simplest of cases.

The example is set up as follows:

- Insurer sells annual policies, and the rate plan has two risk classes: high and low.
- Policies renew 80% of the time after 1 year, and none renew after 2 years.
- Premium P collected and expense $20\% \ * \ P$ paid at time 0 and 1
- Fixed expense 10 paid at time 0
- Losses of $L_{low} \ = \ 50$ or $L_{high} \ = \ 70$ paid at times 1 and 2, depending on the policy's risk classification
- There is a 10% chance after 1 year that a low risk policy will be reclassified as high risk and charged the corresponding premium; all high risk policies remain high risk; renewal probability is independent of reclassification

- Cash flows discounted at 5% interest, and the company desires a 5% UW profit on premium

The goal is to determine what premium to charge for a low risk policy and a high risk policy, under the assumption that the rate can vary only by a policy's risk classification at the beginning of each term and is independent of the original risk classification.

3.1 Traditional Ratemaking

We will begin with traditional ratemaking, which uses the premium loading formula:

$$P = \frac{F + L}{1 - U - V}$$

Taking all inputs literally we would get:

$$P_{low} = (10 + 50)/(1 - 20\% - 5\%) = 80$$

$$P_{high} = (10 + 70)/(1 - 20\% - 5\%) = 106.67$$

The obvious problem with this is that if this premium is also applied at renewal, the customer must pay again for the fixed expenses, even though they are incurred only at the initial policy purchase. The most elementary way to correct for this would be to amortize fixed expenses over the policy's expected lifetime. In this example we have an expected lifetime of 1.8, so we would replace the fixed expenses of 10 in the formulas above with 10/1.8 = 5.56 and determine premiums of:

$$P_{low} = (5.56 + 50)/(1 - 20\% - 5\%) = 74.08$$

$$P_{high} = (5.56 + 70)/(1 - 20\% - 5\%) = 100.75$$

Note that the loss relativity is $70/50 = 1.400$ and the implied premium relativity is $100.75/74.08 = 1.3600$.

3.2 Standard Asset Share Model

Next we will use the standard asset share model. We will again use the premium loading formula, but the inputs will be expected present value of lifetime premium, fixed expense and loss. We also need to have an estimate of premium and loss trend, for which we will need some additional assumptions.

The classification ratemaking example in Feldblum's paper uses the same premium and loss trend for both classes, so we will do the same here. Suppose the company does not currently charge different rates for the two classes, but the book of business is made up mostly of low risk policies. The observed premium trend will then be 0%, since at renewal the average insured's premium stays the same. The loss trend however could be approximated at 4%, since the typical insured is low risk and their loss cost increases by 40% in 10% of cases. Below, we calculate the premium for each class of policyholders by balancing the expected present value of premium to the expected present value of loss and expense.

$$EPV(P_{low}) = P_{low} * (1 + \frac{0.8}{1.05}) =$$
$$(10 + \frac{50}{1.05} + \frac{0.8 * 50 * 1.04}{1.05^2})/(1 - 20\% - 5\%)$$
$$\Rightarrow P_{low} = 72.16$$
$$EPV(P_{high}) = P_{high} * (1 + \frac{0.8}{1.05}) =$$
$$(10 + \frac{70}{1.05} + \frac{0.8 * 70 * 1.04}{1.05^2})/(1 - 20\% - 5\%)$$
$$\Rightarrow P_{high} = 97.99$$

The premium relativity implied here is $97.99/72.16 = 1.3580$. Notice that the premiums here are lower than for the traditional ratemaking case. That's because losses are discounted to time 0 at 5% in the asset share model. Also notice that the premium relativity is lower, so that premiums for the low and high risk classes are proportionally closer to each other. That's because when losses are discounted, a larger portion of the premium goes toward the time 0 fixed expense which is equal for the two risk classes.

3.3 Extended Asset Share/Option Model

Now we will calculate premiums using the extended asset share/option model. Like the traditional asset share model, we will use the expected present value of premium, loss and expense, accounting for probability of renewal. The difference in this solution is that we will explicitly recognize the effect of low risk policies moving to the high risk class at time 1.

We first calculate the premium for high risk policies:

$$EPV(P_{high}) = P_{high} * (1 + \frac{0.8}{1.05}) =$$
$$(10 + \frac{70}{1.05} + \frac{0.8 * 70}{1.05^2})/(1 - 20\% - 5\%)$$
$$\Rightarrow P_{high} = 96.46$$

Notice that we actually arrived at a lower indicated premium than in the standard asset share model. The difference is that before we had determined a 4% average loss trend and applied that to all policies, whereas here we recognize that high risk policies have zero loss trend.

For low risk class policies, the calculation differs from the standard model in the renewal term expected losses and premium. Since there is a 10% chance that the risks are reclassified as high risk, we have:

$$EPV(L_2 \mid risk\ class\ =\ low\ at\ t = 0)\ =\ (0.8\ *\ 0.9\ *\ 50\ +\ 0.8\ *\ 0.1\ *\ 70)/1.05^2$$
$$=\ 37.73$$

Note that, not coincidentally, this is exactly the same result as for the standard asset share model because this process is what we observed in order to determine the 4% overall loss trend.

For premium, we also recognize that the 10% of low risk policies that are reclassified will also be charged the high risk class premium upon renewal. Following the same logic as above we have:

$$EPV(P_2 \mid risk\ class\ =\ low\ at\ t = 0)\ =\ (0.8\ *\ 0.9\ *\ P_{low}\ +\ 0.8\ *\ 0.1\ *\ 96.46)/1.05$$

where we have used the premium calculated above for high risk policies. Putting it all together:

$$P_{low}\ +\ (0.8\ *\ 0.9\ *\ P_{low}\ +\ 0.8\ *\ 0.1\ *\ 96.46)/1.05\ =$$
$$(10\ +\ \frac{50}{1.05}\ +\ 37.73)/(1 - 20\%\ -\ 5\%)$$
$$\Rightarrow\ P_{low}\ =\ 71.06$$

This implies a premium relativity of 1.3574, slightly less than in the standard asset share case. Also, the premiums for both classes are lower. They are lower for high risk policies because we are not incorrectly applying a loss trend factor, and they are lower for the low risk class policies because the higher expected losses in term two are paid for by increased premiums on those policies that transition to the high risk class group. In the standard asset share case, the increased second term losses were partially prepaid by higher premiums in the first term, even on policies that remain low risk.

3.4 Reclassification and Trend

The simplicity of this example allows us to closely examine what went wrong with the standard asset share model. The problem was our calculation and application of trend. For loss trend we saw that for 10% of policies loss costs increased 40% at renewal, so assuming that most of our book of business was low risk we would have observed a 4% annual loss trend. This trend is a function of risk class loss costs and transition probability. For a good paper describing this phenomenon, see [7]. In reality, there will also be a general inflation factor which impacts losses. The advantage that the extended model gives us is that we can explicitly break the average trend down into its two components: risk classification transition and inflation.

4. OPTIONS

4.1 Implicit Options in the Standard Asset Share Model

The standard asset share model assumes that premium will stay the same, or perhaps gradually inflate over the life of the policy, while the extended model recognizes that the rate can change at renewal. Even if all loss costs, expenses, renewal rates and trends were identical these two policies, one where the rate is the same or gradually inflates at renewal and one that has no premium restriction, would have different values.

Let's refer back to the example. What would be the indicated premium for a policy where the renewal rate was the same as the new business rate, even if the risk is reclassified at renewal? The rate would be the same for the high risk class policy as we calculated above for the extended asset share model. The expected lifetime premium, where the nominal premium is the same in both terms, is set equal to the loaded expected present value of fixed expenses and loss costs. For the high risk policies we have:

$$EPV(P_{high}) = P_{high} * (1 + \frac{0.8}{1.05}) =$$

$$(10 + \frac{70}{1.05} + 0.8 * \frac{70}{1.05^2})/(1 - 20\% - 5\%)$$

$$\Rightarrow P_{high} = 96.46$$

This is the same as in the option model because there is no risk reclassification for high risk policies in our example.

For the low risk class, we recognize the transition probability and corresponding high risk loss costs in the second term, but we do not recognize any increase in premium for the second term because there is none, as we are assuming. Borrowing from above we have:

$$EPV(P_{low}) = P_{low} * (1 + \frac{0.8}{1.05}) =$$

$$(10 + \frac{50}{1.05} + 37.73)/(1 - 20\% - 5\%)$$

$$\Rightarrow P_{low} = 72.16$$

Note that this is the same premium we obtained in the standard asset share model. That is because here the expected loss in the second term given that there is a second term is 50 * 1.04. This is the same term appearing in the calculation for the standard asset share model since we estimated a 4% loss trend. Also, when we calculated the standard asset share model we had assumed a 0% premium trend, which is exactly what the premium trend would be for an insurance policy with a guaranteed renewal rate as we have here.

So we obtain the same premium for a guaranteed rate contract as we get for a traditional contract priced using the standard asset share model. What should the difference in price be between an insurance contract with a guaranteed renewal price and one without? The difference will be the value of the implicit option that comes with a guaranteed renewal rate; the option, but not obligation, for the insured to renew the policy for a predetermined price.

Let's determine the value of this option in our example from the perspective of the customer. If they start out classified as low risk, the price for insurance using the extended asset share model is 71.06. The price has an 8% chance of increasing to 96.46 at renewal (10% chance of our company's premium increasing and an 80% chance that the policy renews), otherwise it remains at 71.06. This is exactly like a simple example from option pricing using the binomial stock price model.

The possible values of the insurance contract at time 1 are 96.46 with probability 8%, 71.06 with probability 72%, and 0 with probability 20%. The strike price of the call option is the low risk price which is 71.06. Since the option is purchased at time zero and the payoff is at time one we must discount, so the value of the option is:

$$C = 0.08 * (96.46 - 71.06) / 1.05 = 1.935$$

The reader may note that things don't seem to add up. We said that the guaranteed rate contract is a standard contract plus an option, so the value of the guaranteed rate contract should be the value of the standard contract plus the value of the option, but we have:

$$71.06 + 1.935 = 73.00 \neq 72.16$$

The disconnect is that since the rate for the guaranteed rate contract is fixed for the life of the policy, the price of the option is amortized and paid for in equal amounts in the first and second terms. The expected NPV of a dollar of premium charged to a low risk customer with a guaranteed price contract is:

$$1 + \frac{0.8}{1.05} = 1.762$$

so the price per term for the option is:

$$1.935/1.762 = 1.10$$

and we have:

$$71.06 + 1.10 = 72.16$$

So the price of the guaranteed rate contract is the sum of the price of the standard contract and the amortized cost of the option.

4.2 The Renewal Rate Change Option

It's also possible to view a standard insurance contract as a guaranteed rate contract where the policyholder has sold the implicit call option back to the insurer. How would that transaction work exactly? Suppose the policy has been reclassified as high risk before renewing. The policyholder can renew at the guaranteed price. The insurer will then choose to exercise the call option, since the price of insurance exceeds the option's strike price of the guaranteed rate. The option seller, the customer, must pay the insurer the difference between the strike price (low risk policy premium) and the asset price (high risk policy premium), the asset being an insurance policy. The net payment upon renewal by the customer is the high risk class price, as is the case in a standard insurance contract. This is a convenient way of viewing insurance options as insurance company assets; renewal rate change options.

4.3 Business Relevance

The above is not just abstract theory. We have shown that in a relatively simple example, we are able to charge actuarially justified lower prices for insurance coverage. We are also able to price the value of options on insurance coverage. In any insurance contract where the price at renewal is not completely predetermined, the insurer holds a renewal rate change option of some value, but there are many examples of true insurance options that are being sold to insurance customers in the marketplace today.

In personal lines, some companies offer special features that act to reduce or eliminate rate increases upon renewal. One such feature is “accident forgiveness”. In this case, the company will decline to reclassify the policy as higher risk even though there has been a claim, meaning that the premium would be increased under the standard rating plan. The company in this case is selling an option to the customer, or equivalently, failing to buy back the renewal rate change option on a fixed rate contract. The option is a bit more complex however because other types of rate changes would still be allowed.

Another interesting product in personal lines is perpetual deposit fire insurance. For this type of policy, the insured places a deposit with the insurer and in return receives “free” fire insurance coverage. The deposit is returned when the policy terminates, no matter what the loss experience has been. This product is a swap between interest payments and insurance payments. One could imagine some bells and whistles added, much like has been done in life insurance, that incorporate complex option-like features such as minimum investment returns.

In commercial insurance and reinsurance, there are many types of policy features which involve the implicit selling of options. Some common features are policy extensions, see [5], and multi-year coverage. Multi-year coverage is the most obvious example of options in insurance because it is precisely the guaranteeing of future insurance coverage availability at a predetermined price. Different contracts may have features which oblige the insured to purchase coverage in the future years, which would make these contracts more like forward contracts.

An interesting example in reinsurance is a contingently automatically renewing policy. How it works is that if the financial result (e.g. loss ratio or combined ratio) remains within a certain range, then the reinsurer is obligated to renew the treaty at a predetermined price. If the cedent has the right but not obligation to buy at the strike price, then this reinsurance contract contains options on future reinsurance coverage with a trigger. It may also contain compound options if the renewed treaty again includes options.

Another area where insurance options are present is in the regulation of insurance rates. There are many laws which limit cancellation of policies, limit the magnitude of rate changes, and prevent or slow rate changes. Some of the most hotly debated regulations are those which either allow or disallow certain risk classifications. These are all examples of options which are required to be included in the regulated contracts, or limitations on the exercise of renewal rate change options.

5. GENERAL FRAMEWORK FOR THE EXTENDED ASSET SHARE/OPTION MODEL

We will now present the full mathematical framework for applying the extended asset share model in the general case. The general case is an extension of the simple example in a number of ways. There can be an unlimited number of risk classes and policy terms. Risk classes can transition to and from any other class. The expected loss for any risk class can be anything in any policy term, independent of previous or future terms and of other risk classes. The same will be true for renewal rates and transition probabilities.

The mechanics are essentially just linear algebra, with each dimension representing a risk class. Premium, losses and fixed expenses will be represented by vectors, or series of vectors. We track the movements of risks through the risk classes with a series of matrices. These matrices will represent the transition probabilities between each risk class and probability of termination at each renewal. We will continue to use the simple premium loading formula, with expected present lifetime value for all inputs.

Given the loss cost and fixed expense vectors, the risk class transition matrices, the discount factors, premium trend and the variable expense and profit loads, the method will solve for the current premium level and rate relativities which result in the expected present value of lifetime premium equaling the loaded expected present value of loss costs and expenses simultaneously for all risk classes.

Suppose there are n distinct risk classifications. Let R be an n-dimensional vector such that $R[i]$

is the premium relativity for the i^th^ risk class. R is fixed and does not vary over time. We can require that $R[1] = 1.00$. Let $P(t)$ be a premium index over time (a real function, not a vector), so that $P(t)$ is the expected average rate level independent of risk class transitions. Basically, if a company assumes it will take a 5% rate increase each year, then we would have

$$P(n) = P(0) * 1.05^n$$

We do not need to assume or specify a $P(0)$, as the method will solve for the correct current average rate. Let L_t be a series of n-dimensional vectors so that $L_t[i]$ is the expected loss costs for risk class i at time t. Let F be an n-dimensional vector of fixed expenses so that $F[i]$ is the fixed expense, assumed to all be incurred at time 0, for risk class i. Let U and V be the usual profit and variable expense provisions in the premium. These are real numbers, independent of time and risk class. Let $v(t)$ be a series of discount factors, such that the present value of $1 at time t is $v(t)$ at time 0. Obviously we must have $v(0) = 1$. Finally, let A_t be a series of n-dimensional square matrices such that the probability that a risk in class i at time 0 is in class j at time t is $A_t[i, j]$. The matrices are cumulative transition probability matrices; it is equivalent to define a series of incremental transition probability matrices.

Many readers will notice that under the above construction the policies' risk classification forms a Markov Chain. It may be possible to put more of this exposition into the language already developed in the extensive literature on Markov models, but the goal of this paper is to build on the asset share model as developed in the actuarial literature.

5.1 Premium Formula

We use the standard premium loading formula and as usual equate premiums with loaded loss plus fixed expense, but we use the expected present lifetime value for all quantities as in the standard asset share model:

$$\sum_{n=0}^{\infty} P(n) * v_n * A_n(R) = \frac{1}{1-U-V}\left[F + \sum_{n=0}^{\infty} v_n * A_n(L_n)\right] \quad (1)$$

Let's first break down the left hand side to see why it equals the expected present lifetime value of premium. More precisely, it is an n-dimensional vector, such that the i^th^ entry is the expected lifetime present value of premium for a risk in class i at time 0. $P(n)$ is the average "rate" at time n, and v_n is the discount factor to time zero. The i^th^ entry of $A_n(R)$, which is a matrix applied to a vector, is the inner product of the ith row of the matrix A_n with the vector R. Since A_n is the cumulative transition probability matrix, the rows represent the distribution among the risk classes at time n of risks starting in the class corresponding to that row (as we have defined it). That means the i[th] entry of $A_n(R)$ is the average premium relativity factor at time n for risks in class i at time 0, weighted by the probability that the risk is in each class at time n. So when we sum over all time periods n (the initial new business policy purchase and all subsequent renewals), the i[th] entry on the left hand side is the expected present lifetime value of premium for risks in class i at time 0, taking into account discounting, probability of renewal, premium trend and the difference in premium due to future expected risk classification changes.

What about the right hand side? Likewise, it is a vector. The i[th] entry equals the expected present value of lifetime losses and fixed expense, loaded for profit and variable expense. The profit and expense portions are clear, so we will discuss the sum. Again, v_n is the discount factor to time 0. The i[th] entry of $A_n(L_n)$ is the average loss costs at time n for a risk that was in class i at time 0. That's because the i[th] entry is the inner product of the i[th] row of A_n, which is the risk class distribution at time n of risks starting in class i at time 0, with the vector L_n, which is the vector of loss costs by risk class at time n. Therefore the right hand side is a vector representing the loaded expected present value of lifetime expense and loss costs for each risk class.

Despite the summation, the left hand side is just a matrix applied to the premium relativity vector. To get premium alone on the left hand side we apply the inverse of the matrix (assuming it has one):

$$P(0) * R = \frac{1}{1-U-V}\left(\sum_{n=0}^{\infty} \frac{P(n)}{P(0)} * v_n * A_n\right)^{-1} \left(F + \sum_{n=0}^{\infty} v_n * A_n(L_n)\right) \qquad (2)$$

5.2 Return to Traditional Ratemaking

We can recreate the traditional ratemaking formula from the extended asset share model premium formula above by making some simplifying assumptions for the parameters. Suppose that losses gradually inflate over time with premium trending in step. Also suppose losses trend with general inflation, which is expected to match the risk free investment return. In the notation above we would have:

$$\frac{P(n)}{P(0)} = \frac{L_n[i]}{L_0[i]} = {v_n}^{-1}, \forall\, i$$

The premium formula then simplifies to:

$$P_0 * R = \frac{1}{1-U-V}\left(\sum_{n=0}^{\infty} A_n\right)^{-1} \left(F + \sum_{n=0}^{\infty} A_n(L_0)\right)$$

Distributing the matrix on the right side to the fixed expense and loss terms separately, we can cancel the matrix and its inverse:

$$P_0 * R = \frac{1}{1-U-V}\left[\left(\sum_{n=0}^{\infty} A_n\right)^{-1}(F) + L_0\right]$$

Now premium and loss are unaffected by the matrix sum. Supposing again that we ignored risk class transitions, then each A_n would be a diagonal matrix, with entries equal to the renewal probability for each risk class. We would then have that the sum of A_n over all n would be a diagonal matrix with entries equal to the expected lifetime at time 0 of the policy given its risk class. The i[th] entry of the $(\sum_{n=0}^{\infty} A_n)^{-1}(F)$ vector would then be the fixed expense for the i[th] risk class divided by the expected policy lifetime for risk class i. This is the same expense amortization that we have seen previously. So what we have shown is that traditional ratemaking with fixed expense amortization is the equivalent of the extended asset share model with simplified premium and loss trend assumptions.

6. IN-DEPTH EXAMPLES

We will work through two more complex examples to demonstrate the potential of the extended asset share model. First we will broaden the assumptions in the two risk class case we analyzed at the beginning of the paper, and explore how changes in the parameters affect the premium relativities. Then we enhance the example by adding a "medium risk" class with slightly higher expected losses but the same premium as the low risk class. This will demonstrate how to use the extended asset share/option model to determine the proper premium to charge for "accident forgiveness" or other anti-experience rating policy features. We will again show that the indicated premium will be the sum of the standard premium plus the amortized cost of the options.

6.1 Two Risk Class Case

We can use the extended asset share model premium formula to calculate premiums and relativities for a more complex two risk class example. The low risk class will be given all the characteristics of more desirable underwriting risks such as decreasing loss costs and longer expected lifetime, while the high risk group will be given less desirable characteristics. After calculating the premium, we will compare the derivative of the premium with respect to the various parameters. The assumptions are as follows:

- Insurer sells annual policies, and the rate plan has two risk classes: high and low

- Low risk policies renew with 90% probability at the end of each term, indefinitely
- High risk policies renew with 70% probability at the end of each term, indefinitely
- Premium P collected and expense $20\% * P$ paid at times 0, 1, 2, etc.
- Fixed expense 10 paid at time 0
- Losses of $L_{low}(0) = 50$ or $L_{high}(0) = 70$ paid at time 1 for the first term
- Losses inflate at 1% per year for low risk policies and 3% per year for high risk policies
- At the end of each and every year there is a 10% chance that a low risk policy will be reclassified as high risk and charged the corresponding premium; all high risk policies remain high risk
- Low risk class policies renew with only a 70% probability at the time of reclassification to high risk
- Premiums are expected to inflate at 4% per year.
- Cash flows discounted at 5% interest, and the company desires a 5% UW profit on premium

The only area of ambiguity is that we have said low risk policies renew 90% of the time, transition to high risk 10% of the time, and renew 70% of the time after transition (due to their aversion to rate increases). We will assume the 90% renewal rate is the total renewal rate so that at renewal 10% do not renew, 7% (=10% * 70%) transition to high risk and renew, and the remaining 83% remain low risk and also renew.

Notice that we have assumed premium trend will outpace loss trend on policies that start today. Accounting for this phenomenon of improving loss costs over policy tenure was the original motivation behind the asset share model.

The starting point is the extended asset share premium formula (2) where we had isolated the premium on the left hand side. All necessary parameters are defined above, and the solution is an exercise in linear algebra. Detail on the calculations can be found in Appendices 1 and 3 and were completed using Sage, the Python-based open source mathematics software.

As a point of reference, let's determine what premiums we would charge if our ratemaking method was traditional with expense amortization:

$$P_{low} = (50 + \tfrac{10}{10})/(1 - .2 - .05) = 68.00$$
$$P_{high} = (70 + \tfrac{10}{3.33})/(1 - .2 - .05) = 97.33$$

where 10 and 3.33 are the expected lifetimes of low and high risk policies, respectively, given a

90% and 70% probability of renewal.

The extended asset share model determines the optimal premiums as follows:

$$P_{low} = 56.22$$

$$P_{high} = 90.92$$

The method indicates discounts relative to the traditional premium, just as the standard asset share model would. The discount is higher for low risk policies because their loss trend is more greatly outpaced by premium trend.

It is more interesting here to determine the sensitivity of the low risk premium to the new parameter we have introduced with the extended model: risk classification transition probability. Also, because we are recognizing the chance for reclassification, the low risk premium may also depend on the parameters of the high risk class.

In fact, it isn't even obvious a priori whether the derivative of P_{low} is positive or negative with respect to transition probability. Let us present seemingly valid arguments for both positions. We will call the transition probability inclusive of renewal probability, 7% in our example here, q:

- $\partial P_{low}/\partial q < 0$: Since P_{low} is smaller than P_{high} and total lifetime profit is set at 5% of premium, the dollar amount of required profit on low risk policies is lower than for high risk policies. When policies transition to high risk, they are priced to a higher dollar value of profit, so the extra profit made on transitioning low risk policies subsidizes the price for low risk policies that do not transition.
- $\partial P_{low}/\partial q > 0$: Since high risk policies have short lifetimes, the profit must be front loaded in the first few terms. When low risk policies transition, it is into the later terms of the high risk class which have low profitability or may even be unprofitable. The low risk price will need to be raised to subsidize the relative loss on transitioning policies.

There are some obvious flaws in both of these arguments, but either could be valid under certain assumptions for this simple example. For the parameters we have assumed we have:

$\partial P_{low}/\partial q = -0.28$ (q expressed as a percentage)

That means that keeping all other parameters the same, a 1 point increase in transition probability (inclusive of renewal rate) should result in a reduction in premium of 28 cents for all low risk policies. That makes sense in terms of argument number 1: since premium trend outpaces loss trend for high risk policies, they become more profitable as time goes on and that additional profit subsidizes the price of all low risk policies. If we change the parameters of the example so that high risk loss trend is 4% and premium trend is only 2%, then we see that:

$\partial P_{low}/\partial q \ = \ 0.19$ (q expressed as a percentage)

In this case argument number 2 is correct. Since high risk loss trend outpaces premium trend, later terms of high risk policies are less profitable and may even result in loss. Those terms are precisely when low risk classes are expected to transition, so those losses must be subsidized by higher rates for all low risk policies.

To give some context to the values above we will show the derivative of the low risk premium with respect to some of the other variables. The reader should keep in mind that a 1 point change in trend is a much larger variation proportionally than a $1 change in expected loss, or a 1 point change in renewal probability. To put these measures on an even basis of deviations from the mean, we should look at derivatives of the natural log of the quantities. These can be interpreted as the percentage change in low risk class premium for every percent change in the independent variable. To obtain the figures below from the raw derivatives calculated in Appendix 1 one simply needs to scale by the ratio of the quantities using:

$$dln(y)/dln(x) = (dy/y)/(dx/x) = dy/dx * x/y$$

We have:

$\partial ln(P_{low})/\partial ln(q) \ = \ -0.04$

$\partial ln(P_{low})/\partial ln(p) \ = \ -0.94$ (p = low risk renewal and non-transition probability)

$\partial ln(P_{low})/\partial ln(L_{low}) \ = \ 0.99$

$\partial ln(P_{low})/\partial ln(r_0) \ = \ 3.91$ (r_0 = low risk class loss trend)

$\partial ln(P_{low})/\partial ln(r_1) \ = \ 1.73$ (r_1 = high risk class loss trend)

The above shows that the calculated premium is much more highly leveraged on loss trends than the other variables. While the risk class transition probability appears relatively unimportant, the high risk class loss trend has more leverage on the low risk class premium than either the low risk renewal rate or the low risk loss costs. There must be some significant second order effects as well, since if the transition probability were zero, then the high risk loss trend would have no effect on low risk premium.

6.2 Anti-Experience Rating

As mentioned above, a popular feature of personal lines insurance is "accident forgiveness" or anti-experience rating. A private passenger auto rating plan typically segments risks finely by accident and driving violation history. What would typically happen after an accident is that at renewal the policy would be reclassified based on the new information and charged a higher premium. The promise made by companies offering anti-experience rating to select groups of customers is that in the event of an accident, the company will decline to reclassify the risk and the renewal premium will remain the same. This is a very good example of an option, or more precisely a call option held by the customer and purchased from the insurer.

We will calculate the correct premium for policies with anti-experience rating options using the extended asset share model. We will utilize the previous example with some modifications as follows:

- There is a third risk class, medium or "med" as a variable subscript
- $L_{med} = 55$ and loss trend over time is 2%
- Low risk policies transition to medium risk only with the same 10% probability at the end of each and every year
- Medium risk policies are charged the same premium as low risk, i.e. their slightly higher expected loss is "forgiven"
- Medium risk policies renew with 95% probability, including the term in which they transition
- Medium risk policies transition to high risk with probability 25% at each renewal, otherwise they remain medium risk

We will try to answer the following simple questions: What price would we charge to each of the three classes, low, medium and high, if we ignore anti-experience rating? Then, given that medium risk policies are "forgiven", and charged the same price as low risk policies, what should that price be? What is the value of the call options offered to the low risk policies? How do these quantities relate?

It is worth taking a moment to think about how difficult it would be to answer these questions using just the traditional ratemaking formula or the standard asset share model. Traditional ratemaking clearly falls short, since it only considers one policy period of loss costs. If we add fixed expense amortization, then it would actually indicate a lower price for low risk policies with anti-experience price guarantees, since the expected policy lifetime is longer. Although it is possible for a policy feature which increases insurer costs but also increases policy tenure to pay for itself in increased lifetime profit, the traditional formula is not capable of truly identifying these opportunities and this example is not one of them.

The standard asset share model incorporates persistency rates and trends, so it gives a better chance for modeling this example. The option would affect premium trend since premiums increase less because of option exercise, and customer persistency since customers offered below market prices will be more likely to renew. The standard asset share model would be able to pick up these effects; it's just a question of how accurate it would be. Since the two effects are likely to be small and in opposite directions the analyst may only be confident that the premium should be different, but unsure whether it should be higher or lower.

Another problem with the standard asset share model is that it assumes constant trends throughout time. Since the low risk policies are guaranteed the same rate at the first renewal, the premium trend due to risk reclassification is 0%. There will still be the 4% overall annual premium inflation we assumed. Premium trend will increase for subsequent renewals as policies transition to high risk and are charged higher premiums. Eventually, all policies will be classified as high risk and premium trend will again be 0% attributable to reclassification, plus premium inflation. This clearly cannot be modeled properly with a constant trend. Likewise the standard asset share model assumes a constant loss trend, but in this example we have the same phenomenon of changing trends throughout time.

Even if we were to recognize the changing loss and premium trend throughout time in the standard asset share model, how would we calculate those trends? What is the average premium increase between the third and fourth renewal for policies starting out as low risk at time zero? It is a function of the transition matrix and the premiums charged, but we haven't determined the proper premiums yet. The same complications arise with persistency rates. We seem to be doomed before we even begin.

This is a relatively simple problem when using the extended asset share model, solved in just a few lines of code which are detailed in Appendix 2. The correct premiums are:

$$P_{low} \text{ (no option)} = \$54.98$$

$$P_{med} \text{ (no option)} = \$64.25$$

$$P_{low} \text{ (with option)} = P_{med} \text{ (with option)} = \$57.65$$

$$P_{high} = \$90.92$$

Note that for policies starting out as high risk, there is no difference between this example and the previous one where we only had high and low risk policies. All trends and persistency rates are the same, so the extended asset share model has indicated the same premium. The premiums indicated for the low risk class, either with or without the option, are not comparable with the previous example because we have changed the parameters. In particular, the expected lifetime is different and the rate of transition to high risk is different.

The second part of the problem is to determine the value of the options embedded in the low risk class policy with forgiveness, and to relate it to the price we determined for the policy with and without the options. We know the option is a call option held by the insured. What is the strike price? Because the low policyholder is guaranteed a price at renewal of $57.65 we might guess that to be the strike price, but the $57.65 is the price for another term of insurance coverage plus another call option on the price of insurance at the next renewal. The $57.65 really represents $54.98 for insurance coverage, plus $2.67 for call options with strike price $54.98.

We can easily determine the value of the option with exercise at the first renewal. There is a 10% chance that they will be reclassified as medium risk, and in that event a 95% chance of renewing for a combined 9.5% chance of option exercise. The "market price" for medium risk insured is what the company would like to charge, $64.25. After incorporating the 4% premium trend and 5% cash flow discounting, the present value of this option at time zero is:

$$0.095 * (64.25 - 54.98) * e^{0.04-0.05} = \$0.8719$$

That's a lot less than the $2.67 we've determined should be charged to provide low risk policies with the option, so where is the disconnect? We need to consider options provided on future renewals as well. What is the value of the option to a policy that has already moved to the medium classification? The market price is $64.25 and we are offering them an in-the-money option to purchase insurance at renewal for $54.98. We have assumed they have a 95% total persistency rate with a 25% chance of being reclassified as high risk and a 70% rate of renewal if that happens. The probability of option exercise is 95% - 25% * 70% = 77.5%. With discounting and trending, the NPV of this option at the beginning of a term is:

$$0.775 * (64.25 - 54.98) * e^{0.04-0.05} = \$7.1128$$

This is a lot more than the difference in premium we determined for adding the option to a low risk policy. Recall the simple example from Section 3. There was only one option provided for the life of the policy since all policies terminated at the second renewal, but the cost of the option was spread over both the first and second term premiums. We will show the same for this example, that the $2.67 extra charged for the low and medium risk policies with forgiveness is the amortized cost of options offered over the life of the policy.

The lifetime value of the options can be calculated directly using the binomial model from basic finance. The value of the option is the expected present value of the payoff, which is the difference between the strike price and the market price at time of option exercise, times the probability of option exercise. At the n[th] renewal, the option pays off for all low risk class policies transitioning to medium risk, and all low risk policies that transitioned to medium risk in the 1[st] though n-1[st] renewal and remain medium risk. The reader can verify the following expression with increased decimal precision as the lifetime present value of the option payoffs:

$$(64.24538 - 54.98420) \sum_{n=1}^{\infty} \sum_{k=0}^{n-1} 0.805^{n-1-k} * 0.095 * 0.775^{k} * e^{-0.01n}$$

Evaluating either algebraically or numerically, we find that the lifetime present value of options offered to a low risk policy at time zero is approximately $18.4379.

From the Sage output (see Appendix 2), we have that the present value of $1 charged to low and medium risk policies over their lifetime, until termination or transition to high risk, is $6.92. Again increasing decimal precision, we find that the present value of the $2.67 charge is equal to the above:

$$2.66569 * 6.91675 = \$18.4379$$

So what we have shown is that the extended asset share model calculates the premium required for an insurance policy with accident forgiveness as the price of a standard policy plus the amortized cost of the options.

7. ADVANCED CONSIDERATIONS AND DISCUSSION

As we have seen, there are a number of advantages with the extended asset share/option model. We have been able to more precisely price traditional insurance contracts based on expected lifetime profit, and we can also price more complex policy features such as options on future coverage prices. This paper is meant to be an introduction of the method, not the final word. Below are some issues worth further consideration.

7.1 Dynamic Persistency

The biggest shortcoming of the method as presented here is that renewal rates are not independent of premium levels in reality. Obviously, the higher the price, the lower the probability of renewal, especially if the price is high relative to competitors. In the development above our transition matrices A_n are not functions of $R[i]$, which effectively assumes independence so that the premium can be isolated on the left hand side of the equation. This was a problem we identified with the original asset share model. There is nothing stopping us from defining A_n as being a matrix of functions of the premium relativities, $R[i]$, but this would require more complex numerical methods for determining the optimal premium relativities.

7.2 More Sophisticated Premium Formulas

Throughout we assumed that the premium was given by a simple loading formula. What the asset share model was able to do for us was more precisely determine the effect of risk reclassification on the optimal premiums. In practice, it may be desirable to use one of the many other premium formulas, see [6], such as IRR or RORAC with capital allocation. One could extend the current method by determining an allocated capital vector C, one amount for each risk class. The desired rate of return times the allocated capital would then be an expense accounted for on the right hand side of the premium equation.

7.3 Regulatory Constraints

A final issue worth discussing is the constraint on rates that may prevent the use of the extended (or even traditional) asset share models. It is conceivable that two risks could have identical expected loss costs, expenses and renewal rates but be charged different premiums. For example, it could be that one risk is much more likely to transition to a state that is not able to be charged actuarial rates and cannot be canceled. Our method would call for that risk's current rate to be increased to pay for expected future losses. This phenomenon may violate some state insurance laws, and even "Actuarial Standard Of Practice Number 12, Risk Classification" could be interpreted as prohibiting this.

The disconnect is that, as we mentioned in our discussion of insurance options, there is an implicit option being sold with the policy since the insured will have the right but not obligation to buy at a predetermined price which is less than "market price" (the rate the company would like to charge) in case of transition to the subsidized risk class. A potential defense of the method to the above criticism would be that the price difference reflects a true difference in costs, with that difference being the cost of the option. Just as options on stock prices or commodities are real and quantifiable, options on insurance coverage are real and quantifiable as I hope this paper has demonstrated. It would therefore be actuarially appropriate and in line with the ASOP to charge an increased premium for this additional cost.

If the company were prohibited from explicitly charging for legally mandated insurance options bundled with the policy, then a solution would be to selectively offer a policy with a longer term to the risks with lower transition probability to the subsidized class. The rate for the standard policy would be increased to cover the option value, and all policyholders who qualified would opt for the longer term policy whose rate would be relatively lower because the implicit options are cheaper.

8. BIBLIOGRAPHY

[1] Feldblum, Sholom, "Personal Automobile Premiums: An Asset Share Pricing Approach for Property/Casualty Insurance," *Proceedings of the Casualty Actuarial Society*, 1996, Vol. LXXXIII, No. 159, 190-296.

[2] Brealey, Richard A., Stewart C. Myers, and Franklin Allen. *Principles of Corporate Finance*. Boston: McGraw-Hill/Irwin, 2006. Print.

[3] Cunningham, Robin J., Thomas N. Herzog, and Richard L. London. *Models for Quantifying Risk*. Winsted (Connecticut): ACTEX, 2006. Print.

[4] Wu, Cheng-Sheng Peter, Hua Lin. "Large Scale Analysis of Persistency and Renewal Discounts for Property and Casualty Insurance," *Casualty Actuarial Society E-Forum,* Winter 2009, 396-408.

[5] Newton, Benjamin R., "Multi-Year Policy Pricing," *Casualty Actuarial Society E-Forum,* Winter 2011, Vol. 2, 1-15.

[6] Robbin, Ira, "The Underwriting Profit Provision," *CAS Exam Study Note,* Casualty Actuarial Society, Arlington, Virginia, 1992, 1-86.

[7] McCarthy, Timothy L., "Premium Trend Revisited," *Casualty Actuarial Society Forum*, Casualty Actuarial Society, Arlington, Virginia, 2000: Winter, 47.

[8] Actuarial Standards Board, "Actuarial Standard of Practice, No. 12, Risk Classification," *www.actuarialstandardsboard.org*, Dec. 2005.

Author Biography

Greg McNulty is a pricing actuary in the U.S. Treaty division of SCOR Re. Prior to entering reinsurance he worked as an actuarial analyst in personal auto. He received a B.S. in Mathematics from UCLA in 2005 and an M.S. in Mathematics from University of Michigan in 2007. He is a Fellow of the Casualty Actuarial Society.

Appendix 1. Code for Example of Section 6.1

Below is a copy of the Sage code used to calculate the results in Section 6.1:

```
sage: a,b = var('a,b')
sage: p,q,r,s = var('p,q,r,s')
sage: v,d = var('v,d')
sage: A = matrix(2, [p,q,r,s])
sage: D = matrix(2, [a,0,0,b])
sage: x,y,z,w = var('x,y,z,w')
sage: S = matrix(2, [x,y,z,w])
sage: I = matrix(2, [1,0,0,1])
sage: T = v*I + v*A*S*D - S
sage: g = var('g')
sage: g = solve([ e == 0 for e in T.list() ], S.list(), solution_dict = True)
sage: S_solve = S.subs(*g)
sage: reset('e')
sage: M_left = I-v*d*A
sage: M_right = M_left*S_solve
sage: l_1,l_2 = var('l_1,l_2')
sage: L = vector([l_1,l_2])
sage: P_1,P_2,U,V, F_1,F_2 = var('P_1,P_2,U,V, F_1,F_2')
sage: P =vector([P_1,P_2])
sage: F = vector([F_1, F_2])
sage: P = (M_left*S_solve*L + M_left*F)/(1-U-V)
sage: example = [{l_1:50, l_2:70, F_1:10, F_2:10, v:e^-0.05, d:e^0.04, a:e^0.01, b:e^0.03, U:0.05, V:0.20,
sage: p:0.83, q:0.07, r:0, s:0.7}];
sage: P.subs(*example)
(56.2179531176269, 90.9236721460172)
sage: P[0].diff(q).subs(*example)/100
-0.281473365369170
sage: example2 = [{l_1:50, l_2:70, F_1:10, F_2:10, v:e^-0.05, d:e^0.02, a:e^0.01, b:e^0.04, U:0.05, V:0.20,
sage: p:0.83, q:0.07, r:0, s:0.7}];
sage: P[0].diff(q).subs(*example2)/100
0.188685376430719
sage: P[0].diff(p).subs(*example)
-63.9320830023675
sage: P[0].diff(l_1).subs(*example)
1.11622969670737
sage: P[0].diff(a).subs(*example)
217.553423159780
sage: P[0].diff(b).subs(*example)
94.3772606495679
sage: P[0].diff(l_2).subs(*example)
-0.0149466720869281
sage: P[0].diff(d).subs(*example)
-359.635554630114
sage: P[0].diff(s).subs(*example)
-3.26752863057154
sage: P[0].diff(v).subs(*example)
-2.68704335235142
```

Appendix 2. Code for Example of Section 6.2

Below is a copy of the Sage code used to calculate the results in Section 6.2:

```
sage: v,d = var('v,d')
sage: A = matrix(3, [.805,.095,0, 0,.775,.175, 0,0,.7])
sage: D = matrix(3, [e^0.01,0,0,0,e^0.02,0,0,0,e^0.03])
sage: s1,s2,s3,s4,s5,s6,s7,s8,s9 = var('s1,s2,s3,s4,s5,s6,s7,s8,s9')
sage: S = matrix(3, [s1,s2,s3,s4,s5,s6,s7,s8,s9])
sage: I = matrix(3, [1,0,0,0,1,0,0,0,1])
sage: v =e^-0.05
sage: d =e^0.04
sage: T = v*I + v*A*S*D - S
sage: g = var('g')
sage: g = solve([ e == 0 for e in T.list() ], S.list(), solution_dict = True)
sage: S_solve = S.subs(*g)
sage: reset('e')
sage: M_left = (I-v*d*A).inverse()
sage: L = vector([50,55,70])
sage: P_1,P_2,P_3, U, V= var('P_1,P_2,P_3, U, V')
sage: P =vector([P_1,P_1,P_3])
sage: F = vector([10,10,10])
sage: M = matrix(2,3,[1,0,0,0,0,1])
sage: U = 0.05
sage: V= 0.2
sage: G1 = M*(M_left*P - (S_solve*L + F)/(1-U-V))
sage: G1
(6.91675209043679*P_1 + 1.12370402721951*P_3 - 500.921305786324, 3.25769915424678*P_3 -
296.201975547860)
sage: solns = solve( [e == 0 for e in G1.list()], [P_1, P_3], solution_dict = True)
sage: [[s[P_1].n(25), s[P_3].n(25)] for s in solns]
[[57.64989, 90.92367]]
sage: P2 =vector([P_1,P_2,P_3])
sage: G2 = (M_left*P2 - (S_solve*L + F)/(1-U-V))
sage: solns2 = solve( [e == 0 for e in G2.list()], [P_1, P_2, P_3], solution_dict = True)
sage: [[s[P_1].n(25), s[P_2].n(25), s[P_3].n(25)] for s in solns2]
[[54.98420, 64.24538, 90.92367]]
```

Appendix 3. Matrix Math

Here we will bridge the extended asset share model premium formula in the paper with the code in the previous two appendices.

We begin with the premium side matrix sum, which is simpler. In formula (2) of Section 5.1, the extended asset share premium formula where we have isolated the premium on the left hand side, the right hand side contained the term:

$$\left(\sum_{n=0}^{\infty} \frac{P_n}{P_0} * v_n * A_n\right)^{-1}$$

Since the incremental transition probabilities are the same for all time periods, we have $A_n = A^n$. Likewise, $v_n = v^n$ for the present value factors and $d_n = d^n$ for the premium trend. The familiar geometric series formula can then be used:

$$\sum_{n=0}^{\infty} x^n = \frac{1}{1-x}$$

Since the sum is inverted on the right side of the formula, that is where we get the term:

$$M_{left} = 1 - v * d * A$$

The matrix sum for the loss is a little trickier. In the general case, we let L_n be a series of independent vectors. We cannot just sum the matrices because they are being applied to a series of vectors, not just one as in the premium sum:

$$\sum_{n=0}^{\infty} v_n * A_n(L_n)$$

In these examples however, we assume constant loss inflation trends for each risk class, so we have:

$$L_n = D^n(L_0);\ D = \begin{pmatrix} e^{r_1} & 0 \\ 0 & e^{r_2} \end{pmatrix}$$

So the problem is now to simplify the matrix sum:

$$S = \sum_{n=0}^{\infty} v^{n+1} * A^n * D^n$$

Here we use $v_n = v^{n+1}$ since losses are paid at the end of the period, instead of the beginning as premiums are.

We cannot simply apply the geometric series however because the matrices A and D do not necessarily commute, i.e. $A^n * D^n \neq (A * D)^n$. We do however have the following which produces a simple identity for S:

$$v * I + v * ASD = v * I + v * A \left(\sum_{n=0}^{\infty} v^{n+1} * A^n * D^n \right) D$$

$$= v * I + \sum_{n=0}^{\infty} v^{n+2} * A^{n+1} * D^{n+1}$$

$$= v * I + \sum_{n=1}^{\infty} v^{n+1} * A^n * D^n$$

$$= \sum_{n=0}^{\infty} v^{n+1} * A^n * D^n = S$$

We let Sage solve this equation numerically.

PEBELS: Policy Exposure-Based Excess Loss Smoothing

Marquis J. Moehring, ACAS

Abstract: The invention of PEBELS, or policy exposure based excess loss smoothing, was motivated by the need to develop estimates of high layer expected loss cost for extremely small, non-credible segments of a primary property book of business. The existing actuarial literature provides methods for estimating high layer excess loss cost for large property portfolios in aggregate, but does not provide a method to produce similar provisions for smaller subsets of such a book. PEBELS was developed to be just such a method. PEBELS generalizes existing pricing theory from published property per risk reinsurance exposure rating methods and leverages increasingly available exposure data to produce a method which allows the practitioner to develop accurate high layer expected loss cost estimates down to the policy level. Once the theoretical framework required to implement this method in practice has been formulated, this paper continues to explore applications of PEBELS outside of primary insurance pricing where granular segmentation of expected loss by layer would refine results, such as adjusting modeled catastrophe average annual losses (AALs) for bias from implicit linearity assumptions, improving the predictiveness of property predictive models including generalized linear models (GLMs), and refining the published property per risk reinsurance exposure rating method.

Motivation. Develop a practical method to determine property large loss provisions.
Method. Property exposure analysis.
Results. The PEBELS method of developing property large loss provisions, for use in any application requiring segmentation between primary and excess property loss layers.
Conclusions. PEBELS is a practical theory of property large loss exposure.
Availability. N/A

Keywords. Property, Ratemaking, Large Loss, Excess Loss, Catastrophe, Average Annual Loss, AAL, Predictive Modeling, Generalized Linear Model, GLM, Reinsurance, Exposure Rating, Property Per Risk Reinsurance, PPR

1. INTRODUCTION

Anyone who has been responsible for property pricing has faced the issue of how to incorporate experience from very infrequent but very large claims into their analysis. This is particularly an issue for non-weather losses for which modeled average annual losses are not available. While it is relatively easy to determine an aggregate provision for *non-catastrophe large losses* (NCLL) for a large credible property book [9], it is far less obvious how to determine an appropriate provision for a smaller, less credible, subset of such a book, e.g. a specific program in a specific state.

The actuarial literature is eerily silent on how to produce such a provision even though very infrequent but very large NCLLs drive a disproportionate amount of total loss in property lines [4]. Determining appropriate NCLL provisions for smaller subsets of a larger book is an important step when developing rate level indications specifically for these subsets. PEBELS was developed as a method for producing such provisions by leveraging insured value data (which, thanks to the prevalence of catastrophe modeling, continues to become increasingly accurate and available), exposure curve theory (which was developed for pricing property per risk reinsurance treaties), and the aggregate NCLL provision for the more credible aggregate book (which is routinely determined using published methods).

The method given in this paper was designed for use in producing property rate level indications for primary insurers (personal or commercial). It will be demonstrated that PEBELS is actually a generalization of per risk reinsurance exposure rating techniques in the published actuarial literature, which allows further refinement of existing reinsurance pricing techniques and facilitates the development of additional exposure-based ratemaking applications for both primary insurers and reinsurers. Furthermore, despite the method being developed specifically for use with non-catastrophe large loss experience, it will be shown that the method is equally applicable for catastrophe ratemaking applications given appropriate exposure curve assumptions.

1.1 Research Context

PEBELS was born of the practical problem described in the introduction. Although the aggregate NCLL provision for an entire property book was available with full credibility, similar provisions for subsets of the book by program and state had very low credibility. Furthermore, it was clear by inspection of the underlying exposures in some of these smaller, non-credible subsets that their fundamental exposure to NCLL varied dramatically. Since the business imperative was to develop reliable indications at the program and state level, it became necessary to determine reliable NCLL provisions at this level.

The first step undertaken was a literature review, with the intention of adopting a previously published method for determining the NCLL provision. However, none was found. Hurley [4] and Werner [9] provide methods for determining the overall NCLL provision for an entire book, either explicitly or implicitly, but that issue had already been resolved in this application. A two-sided percentile approach to excess loss loading was also identified in Dean [3], but it was also designed to smooth experience for an aggregate book. What was needed was a way to determine appropriate NCLL provisions by program and state even when such segments had extremely low credibility.

A second step was to expand the literature review beyond that available for primary property ratemaking. The literature for developing excess loss provisions in liability lines was much more developed than that for property lines. However, this was also of little help in solving the problem since the ground-up severity distribution of liability risks was generally assumed similar regardless of whether the insured selects a $1M limit or a $10M limit [7], while this is far from the case with property insurance. In property insurance the insured value (loosely analogous to the "limit" of a liability policy) is strongly correlated with the size and value of the exposure. Thus the underlying severity distribution of the $1M policy is likely to be much different than that of the $10M policy. This distributional dependence on insured values will be formulated in terms of exposure curves in the body of Section 2.5.

Continuing to expand the literature review to applications in reinsurance pricing led to the most relevant methods and theory, though these were not directly applicable to the problem at hand. Clark [2] and Ludwig [5] were the most significant as they laid out the state of the art for exposure rating for property per risk reinsurance which became the basis for deriving PEBELS. Bernegger [1] afforded valuable insight into options available to determine appropriate exposure curves and how to vary these curves between heterogeneous risks. Most notably Bernegger provided specific analytical curves that are sufficient for a very simple implementation of PEBELS.

Finally, a survey of current methods was conducted to determine whether any unpublished methods in common practice could be adapted to solve this problem. Commonly used methods fell into two broad categories: (1) empirical methods, and (2) allocation methods.

Empirical methods, which have been extensively used by ISO, directly compute excess loss factors as total loss to loss below a specified threshold. The data are segmented as desired (in the example given above this would be by program and state, while ISO's commercial property filings have historically segmented by construction and protection classes). In addition, the data are further segmented by insured value interval. Segmenting the excess loss factor estimates by insured value improves estimates by partially controlling for bias in the empirical estimate from the distributional dependence on the insured value noted above. The main challenge with this method is process variance. Even with ISO's massive volume of data, it had to make two adjustments to manage the extensive volatility in its commercial property analysis: (1) it used 10 years of experience in computing excess loss factor estimates (twice as much as it used in the experience period for rate level indications), and (2) it used the fitted values from curves fitted to the excess loss factor estimates rather than the estimates themselves as the large loss provisions in their rate level indications.

Large process variance in our data was apparent from the fact that the empirical excess loss factor estimates even at the state level (before subdividing by program or insured value interval) were extremely volatile. In fact, this was the reason a new method for determining excess loss provisions was needed. The fact that ISO with all of its industry data had to double its experience period and use fitted curves to stabilize their results only reinforced the conclusion to search for a more stable approach.

Another obstacle to using this method was the presence of mix shifts. ISO's use of an extended experience period is reasonable only to the extent the implicit assumption that the industry portfolio is relatively stable over the ten year large loss experience period is reasonable. While it is plausible to assume this is true for bureau data which roughly correspond to the industry's total data, it is far less reasonable to assume that any single carrier's mix of business is stable for such a long period of time. Furthermore, it was well known that the book under consideration had experienced significant mix shifts over the last several years.

Although it is tempting to add experience from additional years to increase data volume we must be cognizant that adding older years of historical data will not improve our loss projection when the historical book of business which produced that experience is not representative of the prospective book of business being rated. This phenomenon is demonstrated by Mahler [6] in his study which concluded that using additional years of data when exposure risk characteristics are known to be shifting can actually decrease the predictive accuracy of the resulting projections.

Finally, allocation methods were considered where the goal was to allocate the NCLL provision for a large credible property book down to the level of program and state using a "reasonable" allocation base. This approach was intuitively appealing because it was simple to explain, easy to implement, and it tied directly to the credible overall NCLL provision of the total book of business under consideration. Furthermore, it was believed that an appropriate exposure base would yield stable results and control for mix shifts. Thus it was decided to use the allocation approach.

However, even though we had decided to pursue this approach, it was not clear what a "reasonable" allocation base would be. Given the intention to allocate this provision down to a level where there was little credibility and large heterogeneity in underlying exposure to NCLL, it was critical to develop a reliably predictive allocation base.

Taking the survey one step further, practitioners were interviewed as to the allocation bases in current use and their satisfaction with each. Allocation bases in common usage were: (1) loss (either total or capped below a specified threshold) and (2) earned premium. Practitioners surveyed were generally aware that neither of these bases were particularly satisfactory, the reasons for which will be elaborated on in Section 2.1 below.

1.2 Objective

The goal of this work was to develop a method to determine appropriate non-catastrophe large loss provisions for low credibility segments of business. Since a method to produce such provisions was not found in the existing actuarial literature (see research context above), it became necessary to develop a new technique. The result of this work was the development of PEBELS.

Sections 2.1-2 continue where Section 1.1 ended by demonstrating problems with the allocation bases in common usage. It goes on to describe issues with insured value which was the first unique allocation base explored. Section 2.3 defines and formulates PEBELS. Sections 2.4-5 provide background necessary to understand the theory behind the many generalizations to traditional per risk reinsurance pricing theory that PEBELS relies on to make accurate estimates of excess loss exposure. Sections 2.6-10 describe the dimensions in which PEBELS is generalized relative to traditional per risk reinsurance pricing. Sections 2.11-14 focus on applications of PEBELS in catastrophe modeling, predictive modeling, and per risk reinsurance pricing. Section 2.15 is a discussion of practical issues that must be grappled with to successfully implement a PEBELS analysis.

1.3 Outline

2.1 Allocation Bases in Common Usage

2.2 Insured Value

2.3 The Formulation of PEBELS

2.4 Property Per Risk Reinsurance Exposure Rating

2.5 Property Exposure Curves

2.6 The PEBELS Exposure Base: PEBEL and NLE

2.7 Per Policy Generalization

2.8 Heterogeneity Generalization

2.9 Historical versus Prospective Exposure Distribution Generalization

2.10 Low Credibility Generalization

2.11 The Scope of PEBELS

2.12 PEBELS and Adjusting Modeled Catastrophe AALs

2.13 PEBELS and Predictive Models

2. BACKGROUND AND METHODS

2.1 Allocation Bases in Common Usage

The exhibits below were created based on simulated data from 10,000 hypothetical homeowners insurance exposures for a hypothetical insurer who writes homeowners business in three states (X, Y and Z) and in three programs: one for each of Barns, Houses and Estates. NCLL exposure is assumed identical by state, but is assumed to vary by program.

State X and Y are not catastrophe exposed and are assumed to have identically distributed rates; State Z is catastrophe exposed and is assumed to have 50% higher rates than States X and Y on average to compensate for this exposure. Barns are assumed to be small structures with insured values not exceeding $25,000, Houses are assumed to be single-family dwellings with insured values not exceeding $500,000, and Estates are considered to be large luxury dwellings with insured values not exceeding $10,000,000. The total exposures are equally distributed among States X, Y and Z. Houses comprise 90% of the exposures while Barns and Estates each comprise 5%. Ultimate losses at prospective levels for these exposures were simulated for these risks using loss distributions inferred from exposure curves assumed for each program. More details of the simulated exposures and the calculations underlying the allocations below are contained in the simulation and exhibit work files which are attached to this research paper. In each of the exhibits below, loss amounts in excess of $100,000 are considered "excess" and the overall NCLL provision for the entire book is defined as the total excess loss during the experience period, which is $16,920,439 (out of $23,172,176 in total loss) based on the simulation described above.

The actual simulated excess loss experience for this book by state and program is summarized in Table 1 below.

Table 1: Actual NCLL by State and Program

	Program			
State:	Barn	House	Estate	Total
X	0	1,760,617	2,414,135	4,174,752
Y	0	1,214,428	2,830,850	4,045,278
Z	0	806,427	7,893,983	8,700,409
Total	0	3,781,471	13,138,968	16,920,439

It can be immediately inferred from the data that not every cell can be fully credible. There are large differences between House and Estate NCLL experience by state, which we would not expect a priori since our assumptions imply that we have the same expected NCLL in each state and in each program. This is the motivation for allocating or "smoothing" the total experience as opposed to relying solely on historical results. Our first attempt at the allocation method uses losses capped at $100,000 as a proxy for NCLL exposure. The results are summarized in Table 2 below.

Table 2: NCLL Allocation by State and Program Based on Capped Loss

	Program			
State:	Barn	House	Estate	Total
X	65,617	4,454,532	927,711	5,447,860
Y	157,712	5,743,197	541,304	6,442,214
Z	83,916	3,757,569	1,188,881	5,030,366
Total	307,245	13,955,299	2,657,895	16,920,439

This first attempt suffers from many problems. The first problem is that it allocates NCLL to the Barns Program (because Barns generate losses below $100,000). However, since no Barns have insured values over $25,000, these exposures are not able to produce losses in excess of the $100,000 threshold and thus should have $0 allocated to them. This is the first red flag that capped loss is not an appropriate proxy for NCLL exposure. Another issue is that the allocated provision for Houses varies quite a bit, which is unexpected given that we are assuming that each state has the same expected NCLL exposure on average. Finally, Estates seem to be allocated much less than Houses which seems wrong since Estates generated most of the actual excess losses. In short, this allocation base does not appear to be appropriately correlated with NCLL exposure.

A second closely related attempt is to allocate NCLL based on losses in excess of the $100,000 threshold. The results of this approach are summarized below in Table 3.

Table 3: NCLL Allocation by State and Program Based on Excess Loss

		Program		
State:	Barn	House	Estate	Total
X	0	1,760,617	2,414,135	4,174,752
Y	0	1,214,428	2,830,850	4,045,278
Z	0	806,427	7,893,983	8,700,409
Total	0	3,781,471	13,138,968	16,920,439

The problem with the results of this allocation base is that it is numerically equal to the actual NCLL in Table 1 which we already decided had insufficient credibility. This implies that the use of total losses which is the sum of capped and excess losses will yield a result between that of the actual NCLL and the NCLL allocated using capped losses, both of which we have identified as undesirable options.

Searching for a more stable base leads us to earned premium. Table 4 below summarizes the NCLL allocation based on earned premium.

Table 4: NCLL Allocation by State and Program Based on Earned Premium

		Program		
State:	Barn	House	Estate	Total
X	26,206	2,398,755	2,356,210	4,781,170
Y	26,191	2,403,745	2,443,075	4,873,011
Z	32,884	3,679,817	3,553,556	7,266,257
Total	85,281	8,482,317	8,352,841	16,920,439

As expected these NCLL allocations are much more stable. However, we are back to allocating NCLL to the Barn program. An even more ominous problem is that we are systematically biased towards allocating more NCLL exposure to State Z. Recall that rates in State Z are assumed to be

50% higher than States X and Y to compensate for catastrophe exposure, even though all three states have the same expected NCLL exposure.

While we could achieve a satisfactory result in the example above by adjusting the premiums from State Z, that is only possible due to our simplistic 50% assumption. In practice it is much harder to identify, quantify and adjust for the many considerations that would bias earned premium as a measure of NCLL exposure. Frequency of low severity claims and rate adequacy are prominent examples of such potential biases.

2.2 Insured Value

Having exhausted the traditionally used allocation bases, insured value was considered as a potential allocation base. Historically, insured value data tended to be poorly captured or otherwise unavailable because it is not reported on financial statements. However, given the increasing reliance of insurers on catastrophe modeling, both the availability and quality of insured value data has dramatically increased in recent years. Interestingly, while it was commonly understood that insured value is highly predictive of catastrophic loss potential, the fact that it was also highly predictive of non-catastrophe large loss potential has gone largely unnoticed. This observation is fundamental to the formulation of PEBELS given in Section 2.3. Table 5 below summarizes the NCLL allocation based on insured value.

Table 5: NCLL Allocation by State and Program Based on Insured Value

		Program		
State:	Barn	House	Estate	Total
X	9,807	1,598,954	3,958,740	5,567,502
Y	9,664	1,602,367	4,101,806	5,713,837
Z	10,102	1,645,086	3,983,912	5,639,100
Total	29,573	4,846,408	12,044,458	16,920,439

Aside from the fact that we are again allocating NCLL to Barns, this base seems to be a marked improvement in terms of stability and the proportion of NCLL allocated to Estates versus Houses.

A small modification that would alleviate the non-zero allocation to Barns would be to allocate based on *layer insured value* defined as the amount of insured value in the exposed layer. For example, in the tables above, the layer of interest is from $100,000 to infinity so a barn with insured value of $25,000 would have a layer insured value of $0, while a $300,000 house would have a layer insured value of $300,000 - $100,000 = $200,000. Similarly a $5,000,0000 estate will have a layer insured value of $4,900,000 Table 6 below summarizes the NCLL allocation based on layer insured value.

Table 6: NCLL Allocation by State and Program Based on Layer Insured Value

		Program		
State:	Barn	House	Estate	Total
X	0	1,157,291	4,402,339	5,559,630
Y	0	1,163,078	4,564,989	5,728,067
Z	0	1,201,790	4,430,953	5,632,742
Total	0	3,522,159	13,398,280	16,920,439

At this point it seems that all obvious allocation issues have been resolved. Furthermore, there is no doubt that NCLL exposure is strongly dependent on layer insured value. However, there remains a subtle issue as to whether the allocation between Houses and Estates is appropriate. Using layer insured value as the allocation base implies that the potential for loss in the layer from $100,000 to $200,000 on a $500,000 house contributes the same amount of NCLL exposure as the layer from $9,900,000 to $10,000,000 on a $10,000,000 estate. Is that "reasonable" to assume?

Intuitively, it seems far more likely that a $500,000 single-family dwelling will experience a partial loss of 40% ($200,000 / $500,000) or more than that a $10,000,000 estate will suffer a total loss. This is because:

- Total losses are less common than smaller, partial losses
- Larger losses, as measured by the ratio of insured loss over insured value, become less likely as the insured value of the property increases

The first bullet follows from the observation that people tend to put out the fire, or otherwise stop whatever damage is being incurred. The second bullet is a bit more subtle, but can be conceptualized in terms of how "spread-out" the property exposure is.

Barns are not very "spread-out" at all. The chances of a total loss to a barn given a fire is almost 100% (the barn would likely be engulfed in the time that it would take to detect the fire). At the other extreme, college campuses are highly "spread-out". A fire is unlikely to cause damage to a large proportion of the campus since it will almost certainly be localized to a single building, or in extreme cases it may spread to a few nearby buildings before the fire is contained. Even extreme catastrophes are unlikely to totally destroy all the buildings on a campus. Thus, given a loss to a college campus the loss is very likely to be a small percentage of the total insured value.

2.3 The Formulation of PEBELS

PEBELS, or policy exposure based excess loss smoothing, is based on the notion that exposure to excess loss should be directly computed based on the characteristics of the policy itself. A technique was devised based on a generalization of the traditional per risk reinsurance exposure rating algorithm to directly quantify the exposure to excess loss for a given policy which is defined as the *policy exposure based excess loss* (PEBEL) for that policy. When PEBEL is taken as the base for allocating or "smoothing" aggregate NCLL we call that the PEBELS method of allocating excess loss.

For ease of illustration, the general formulation of PEBEL for the i^{th} policy is copied below from Section 2.6 where it is derived,

$$PEBEL_i = P_i * ELR_i * EF_i, where \tag{2.7}$$

P_i is the premium of the i^{th} policy,
ELR_i is the expected loss ratio, which is assumed to be constant for all policies (for now),
EF_i = the proportion of total loss cost expected in the excess layer.

Note that there is no theoretical reason why $PEBEL_i$ could not be used as a direct estimate of NCLL for a given policy. However, in practice it is preferable to use it as an allocation base because (1) we are assuming that actual NCLL is fully credible at the aggregate level, which makes it desirable for the sum of the NCLL estimates for all segments to equal the aggregate NCLL, and (2) using $PEBEL_i$ directly as our estimate of NCLL implies we have full confidence in the estimate from Equation 2.7, which would require a great deal of precision in specifying inputs as will be demonstrated in the following sections.

The balance of this paper will develop the theory behind Equation 2.7 and the practical considerations that must be addressed to ensure a reasonable NCLL allocation using PEBELS. The emphasis will be on understanding the many complex dynamics of the property excess loss distributions by policy, and how PEBELS can be leveraged to reflect these dynamics leading to refined policy level NCLL estimates.

2.4 Property Per Risk Reinsurance Exposure Rating

Exposure rating for high property loss layers has been common in reinsurance pricing since 1963 when Ruth Salzmann published a study [8] using building losses from fire claims on a book of homeowners policies to quantify the relationship between size of loss and insured value. This study was widely utilized in reinsurance pricing despite its deliberately narrow scope. Among the results of Salzmann's work were the first exposure curves. Ludwig [5] greatly expanded on Salzmann's work in 1991 by considering perils other than just fire, first party coverages other than just building, and commercial property exposures.

Clark [2] states in his 1996 study note on reinsurance pricing that "the exposure rating model is fairly simple, but at first appears strange as nothing similar is found on the primary insurance side", although no implication is made that something similar could not exist on the primary insurance side. He then goes on to produce an example of how a reinsurer might use exposure rating in practice which is reproduced below in Table 7 for background on this pricing technique.

Table 7: Property Per Risk Exposure Rating Example from David Clark's 1997 Study Note

Treaty Retention: $100,000
Treaty Limit: $400,000

(1) Insured Value Range ($000s)	(2) Midpoint ($000s)	(3) Retention as a % of Insured value	(4) Retention + Limit as a % of Insured value	(5) Exposure Factor	(6) Subject Premium	(7) Expected Loss Ratio	(8) Expected Primary Losses	(9) Expected Reinsurer Losses
20-100	60	167%	833%	0%	682,000	65%	443,300	0
100-250	175	57%	286%	26%	161,000	65%	104,650	27,209
250-1,000	625	16%	80%	41%	285,000	65%	185,250	75,953
1,000-2,000	1,500	7%	33%	33%	1,156,000	65%	751,400	247,962
Grand Total					2,284,000	65%	1,484,600	351,124

(1) = ranges in which property exposures were banded based on their insured value
(2) = average of range given in (1)
(3) = 100 / (2)
(4) = (100 + 400) / (2)
(5) = [Exposure Curve Evaluated at (4)] - [Exposure Curve Evaluated at (3)]
(6) = sum of primary premiums of property exposures in insured value range
(7) = Assumption
(8) = (6) * (7)
(9) = (5) * (8)

Clark provided an illustrative exposure curve for the purpose of the example. Note that column (5) develops the exposure factor using that exposure curve. The next section is an introduction to practical considerations for selecting and using exposure curves.

2.5 Property Exposure Curves

An exposure curve, G(x), gives the proportion of the total loss cost at or below a specified threshold, x, as a function of the threshold itself. When using exposure curves the threshold is stated as a percent of a total measure of exposure. Most authors have used insured value as the denominator, though Bernegger [1] varied his formulation by specifying maximum possible loss to be the denominator in order to facilitate his analytical distributional analysis by ensuring that the range of possible threshold values is always less than or equal to 100%, which is not necessarily the case when using insured value as the denominator. This paper uses insured value as the denominator, which is common in practice.

Given a retention of interest, (say $100,000 to be consistent with our preceding examples), an exposure curve can help answer the question of what percent of total loss cost will be eliminated by the retention. The threshold can be expressed mathematically as:

$$x = (\textit{Gross Loss Amount in Dollars}) \,/\, (\textit{Insured Value}) \tag{2.1}$$

In this formulation the exposure curve, G(x) is given by:

$$G(x) = \frac{\int_0^x [1 - F(y)]dy}{\int_0^1 [1 - F(y)]dy}, where\ F(x) = the\ cumulative\ distribution\ function\ of\ x \tag{2.2}$$

Because these curves describe the relationship between the proportion of loss under the retention relative to the total, they are equivalent to the loss distribution function conditional on having a loss. In fact Bernegger [1] demonstrates that the cumulative distribution function of the loss distribution, F(x), can be stated mathematically as a function of the exposure curve using the following relationship:

$$F(x) = 1 - \frac{G'(x)}{G'(0)}\ when\ x\ is\ in\ [0,1);\ or\ F(x) = 1\ when\ x = 1 \tag{2.3}$$

This observation is important because it implies that determining the appropriate exposure curves for a given risk is indirectly equivalent to specifying a loss distribution for that risk in a manner analogous to the way that making ILF selections is indirectly analogous to specifying a loss distribution [7].

Consequently, for the same reasons that it may be appropriate to select different ILFs between different classes of liability risks based on their underlying loss distributions, it is also appropriate to vary selected exposure curves by risk class to appropriately reflect the underlying loss distribution. For example, it will almost certainly be appropriate to vary the selected exposure curve by insured values where significant differences arise for the reasons outlined in Section 2.2.

For the sake of the illustrative examples given in the sections below we will use Bernegger's analytical exposure curve formulation [1] derived from the Maxwell-Boltzmann, Bose-Einstein, Fermi-Dirac, and Planck, or MBBEFD distribution from statistical mechanics, which he found to be "very appropriate for the modeling of empirical loss distributions on the interval [0, 1]." Bernegger's parameterization of the MBBEFD for use in modeling exposure curves with the maximum loss limited to a maximum possible loss which we have assumed to be insured value in Equation 2.1:

$$G(x) = \frac{\ln\left[(g-1)b + (1-gb)b^x\right] - \ln(1-b)}{\ln(gb)}, where\ g\ and\ b\ are\ the\ parameters \tag{2.4}$$

In Bernegger's paper this form is only one of four possible piecewise expressions given for G(x). However, the other three specify endpoint conditions or mathematical extremes, so we need only focus on this form to use the curve in practical applications.

Bernegger went on to fit this parameterization of the MBBEFD distribution he formulated to well-known exposure curves that were the result of a study conducted by Swiss Re, a well-known reinsurer. Four such exposure curves were fit to Equation 2.4, each curve loosely corresponding to risks of increasing size or how "spread-out" the exposure is, as discussed in Section 2.2 above. In addition a fifth curve was fit based on data from Lloyd's of London, the well-known insurance exchange. The curve resulting from these data is considered to correspond to larger and more spread out risks than are contemplated in any of the Swiss Re curves.

Having computed the pairs of (b, g) parameters from fits to each of the five exposure curves described above, Bernegger devised a parametric mapping of the pairs that transformed them from two dimensions to a single dimension which allowed linear interpolation between the well-known exposure curves. This interpolation scheme is given below:

$$g_c = e^{(0.78+0.12c)c} \tag{2.5}$$

and

$$b_c = e^{3.1-0.15(1+c)c} \tag{2.6}$$

In his study Bernegger concluded that the five fitted curves were adequately modeled using values of the parameter c = 1.5, 2.0, 3.0, 4.0, or 5.0 with 1.5 corresponding to the Swiss Re curve appropriate for the smallest properties and 5.0 corresponding to the Lloyd's of London curve.

The example given above in Table 7 assumed an illustrative exposure curve provided by Clark to facilitate his example. Unless specified otherwise, the numerical examples provided in the tables below use exposure factors derived from the continuous Bernegger exposure curve given in Equation 2.4 with parameters g = 4.221 and b = 12.648, which were derived using Bernegger's fitted interpolation of the "well known" Swiss Re and Lloyd's of London exposure with c = 1.5, which most closely corresponds to the first Swiss Re curve in Bernegger's formulation. This c value was selected for convenience and ease of illustration.

2.6 The PEBELS Exposure Base: PEBEL and NLE

As stated in Section 2.3, PEBELS maintains that exposure to excess loss should be directly computed based on the characteristics of the policy itself. Up to this point, only indirect measures of excess loss exposure have been tested. However, the non-proportional per risk reinsurance pricing background developed in Sections 2.4 and 2.5 can be used to develop the direct measure of excess loss exposure given in Section 2.3, in Equation 2.7.

To get the allocation base, we simply compute the expected excess loss implied by the exposure rating methodology developed in Section 2.4, for each policy individually. The result of this computation for each policy is defined to be the PEBEL for each policy which is taken as the PEBELS base for allocating aggregate NCLL. Table 8 demonstrates this computation using four representative example risks from our simulation.

Table 8: Property Per Risk Exposure Rating of Selected Individual Policies for Illustration (Under Assumption of Homogeneous Risk Characteristics)

Layer Lower Bound: $100,000
Layer Upper Bound: Unlimited

(1) Policy Insured Value ($000)	(2) Policy Program	(3) Layer Lower Bound as a % of Insured value	(4) Layer Upper Bound as a % of Insured value	(5) Exposure Factor	(6) Policy Premium	(7) Expected Loss Ratio	(8) Expected Primary Losses	(9) PEBEL
20,319	Barn	492%	Infinite	0%	127	65.0%	83	0
313,398	House	32%	Infinite	53%	1,567	65.0%	1,019	535
220,278	House	45%	Infinite	40%	1,652	65.0%	1,074	433
8,883,554	Estate	1%	Infinite	97%	1,156,000	65.0%	751,400	728,260

Notice that this calculation mimics the computation shown in Table 7 in every respect except that the calculation is done per policy; just as in Section 2.4, all policies are assumed to have the same expected loss ratio, and the same exposure curve was used for each exposure. The reasonableness of these assumptions will be explored in the sections below.

The NCLL allocation implied by using this basis is summarized in Table 9 below:

Table 9: NCLL Allocation by State and Program Using PEBELS (Under Assumption of Homogeneous Risk Characteristics)

	Program			
State:	Barn	House	Estate	Total
X	0	1,651,927	3,109,857	4,761,784
Y	0	1,661,574	3,229,621	4,891,195
Z	0	2,572,479	4,694,981	7,267,460
Total	0	5,885,980	11,034,459	16,920,439

Now this allocation is starting to become refined in that it is not only capturing where there is exposure to excess loss, but it is also contemplating the likelihood of excess loss from the exposure by contemplating exposure factors derived from the exposure curve.

Unfortunately, this allocation still has at least one problem. Notice how the allocations to State Z are biased high relative to States X and Y. This is because premiums in State Z are higher than X and Y due to catastrophe exposure which is skewing this allocation high relative to its true non-catastrophe excess loss potential which is the same as the other states. The generalizations given in Section 2.8 will provide the flexibility necessary to correct for this and another more subtle distortion. However, before diving into techniques to refine the allocation we will decompose the PEBEL analytically to derive a simplified formulation which will allow this method to be used when exposure profiles used for catastrophe modeling are available but the data detail required to reliably attach policy premiums is not available.

We begin by abstracting the tabular computation illustrated in Table 8 as:

$$PEBEL_i = P_i * ELR_i * EF_i, where \tag{2.7}$$

P_i is the premium of the i^{th} policy
ELR_i is the expected loss ratio, which is assumed to be constant for all policies (for now)
$EF_i = G(x_{\text{Upper Bound}}) - G(x_{\text{Lower Bound}})$, is the exposure factor of the i^{th} policy

Equation 2.7 can be expanded into:

$$PEBEL_i = (E_i * r) * ELR_i * EF_i, where \tag{2.8}$$

E_i is the insured value of the i^{th} policy
r is the base rate such that $E_i * r = P_i$, which is assumed to be the same for all policies in the absence of policy-specific premium information. (This is the key assumption for deriving the simplified formulation).

Since the intended application is to use $PEBEL_i$ as an allocation base, the implied excess loss potential attributable to the i^{th} policy will be:

$$\frac{PEBEL_i}{PEBEL_{total}} = \frac{E_i * r * ELR_i * EF_i}{\sum_i E_i * r * ELR_i * EF_i} = \frac{E_i * EF_i}{\sum_i E_i * EF_i} \tag{2.9}$$

Thus the simplified formulation is a reduced form allocation base equal to:

$$NLE_i = E_i * EF_i, where \qquad (2.10)$$

NLE_i is defined to be the *Net Layer Exposure* for a given policy.

The allocation based on NLE will be numerically equal to the allocation based on PEBEL whenever the ELR and r are constant for each risk. The ELR assumption is common, but the rate assumption is less so and suffers from the obvious problem that it is not true. However, it is a fair a priori assumption in the absence of premium detail. If we had rate detail such as average premium modification we could incorporate such information into this allocation base improving its accuracy, but the chances that we would have such specific modifier detail without having policy premiums is exceedingly unlikely.

2.7 Per Policy Generalization

The first generalization to the exposure rating algorithm has already been introduced. It is simple and subtle, but should be pointed out since it is one of the cornerstones of the PEBELS approach. The per policy generalization is simply the insistence that excess loss exposure be quantified at the policy level.

Note that computing PEBEL at the policy level is a break from computing it for groups of policies with similar insured values, which is how it is computed in the published reinsurance approach reproduced in Section 2.4. While the use of grouped insured value bands was, no doubt, an artifact of some data or operational limitation, it is sub-optimal in at least three ways.

Most obviously, it is approximate. The exposure factors are computed using the midpoints of the insured value bands which will be incorrect for most if not all risks in the band. As we become increasingly reliant on the distributional information in the exposure curves to set differences in excess loss exposure between risks, this inaccuracy becomes decreasingly acceptable. Luckily it is easily overcome by computing PEBEL per policy.

Second, it would be insufficient for our application. We are trying to allocate excess loss exposure down to the granular level of state and program. Attempting to use PEBEL implied by the broad grouping approach for this purpose would only compound the inaccuracy described directly above.

The third problem with a grouped approach is that it requires an assumption of homogeneity of all risks within the band, which, as Section 2.6 alluded, might not be reasonable. In the example from Section 2.6, one band may include risks in both States X and Z which vary in their exposure to catastrophes and thus overall premium levels even though their exposure to non-catastrophe excess loss potential is identical. In that example, assuming homogeneity is too restrictive.

2.8 Heterogeneity Generalization

Up to this point no effort has been made to tailor the PEBEL assumptions by policy to reflect policy-specific characteristics even though failing to do so has introduced distortions in our NCLL allocation. This section will provide the tools to refine the allocation base for heterogeneity between policies.

The preferred method to correct for differing rate levels not driven by non-catastrophe loss exposure is to vary the ELR by policy. Specifically, we want to use an ELR representative of the true non-catastrophe expected loss (note: by varying the ELR by policy we are segmenting policies by expected total loss as measured by the ELR). If we assume that the 65.0% ELR assumed so far is appropriate for States X and Y, then the appropriate non-catastrophe ELR for State Z is 65% / 1.5 = 43.3% which follows from our assumption that Z has the same ex-catastrophe exposure as X and Y but has rates that are 50% higher due to its catastrophe exposure. In practice you would not know a priori how much rates varied so the ex-catastrophe ELR may be estimated as the total ELR minus the catastrophe LR implied by modeled AALs or whatever measure is deemed appropriate. The NCLL allocation derived by varying the ELR for cat exposure is summarized below in Table 10.

Table 10: NCLL Allocation by State and Program Using PEBELS
(Reflect ELR Heterogeneity for Catastrophe Exposure Only)

	Program			
State:	Barn	House	Estate	Total
X	0	1,927,950	3,629,488	5,557,438
Y	0	1,939,209	3,769,264	5,708,473
Z	0	2,001,546	3,652,982	5,654,528
Total	0	5,868,705	11,051,734	16,920,439

Now this allocation looks very smooth, perhaps too smooth. If State Z really is catastrophe exposed, it probably requires a higher profit margin than States X and Y to support the additional capital required to write in that state. Suppose the additional profit margin required was 5%, then the true ex-catastrophe ELR for state Z is actually (65.0% - 5.0%) / 1.5 = 40.0%.

But we cannot stop there. Once we have opened the Pandora's Box of rate adequacy adjustments we have to consider other possible rate differentials. For example, we may require an additional profit margin, say another 5%, for Estates based on the additional capital required to write such large exposures. We could go even farther down this path specifying ELR differentials between different states, but at a certain point this gets impractical since we would need exactly the type of rate level indications we plan to use this analysis to produce. However, that does not keep us from reflecting significant deviations in ELR that we can reflect with confidence based on broad characteristics such as catastrophe-exposure, program, or pronounced rate inadequacy. In all other situations broad ELR assumptions are used. Table 11 below gives the ELR assumptions for each segment used in the final allocation.

Table 11: Final Non-Catastrophe ELR Assumptions by State and Program

	Program		
State:	Barn	House	Estate
X	65.0%	65.0%	60.0%
Y	65.0%	65.0%	60.0%
Z	40.0%	40.0%	33.3%

The NCLL allocation using these final ELR assumptions is summarized below in Table 12.

Table 12: NCLL Allocation by State and Program Using PEBELS
(Reflect Final Non-Catastrophe ELR Assumptions by State and Program)

	Program			
State:	Barn	House	Estate	Total
X	0	2,086,412	3,625,663	5,712,075
Y	0	2,098,596	3,765,292	5,863,888
Z	0	1,999,437	3,345,039	5,344,475
Total	0	6,184,445	10,735,994	16,920,439

The magnitude of the impact of adjusting for differing expected profit loads is much less than the magnitude of the impact of reflecting differences in cat-exposures which reinforces the decision to only reflect significant differences in broad groups of policies.

The last significant source of heterogeneity we will discuss is that stemming from differences in underlying loss distributions by policy. By using the same exposure curve for all risks we are implicitly assuming that all risks are homogeneous with respect to their underlying loss distributions.

In particular Clark [2] identified this issue in his study note on reinsurance pricing, cautioning that "an implicit assumption in the exposure rating approach outlined above is that the same exposure curve applies regardless of the size of the insured value…This assumption of scale independence may be appropriate for homeowner's business, for which this technique was first developed but may be a serious problem when applied to large commercial risks."

Each of the three programs referenced throughout this paper can be expected to have dramatically differing distributions of insured values. Thus it is not reasonable to expect that they would have the same loss distributions and thus the same exposure curves.

In the simulation examples above the exposure factors have been derived using the continuous Bernegger exposure curve given in Equation 2.4 with parameters g = 4.221 and b = 12.648, which were derived using Bernegger's fitted interpolation of the "well known" Swiss Re and Lloyd's of London exposure with c = 1.5, which was chosen because it most closely corresponds to the first Swiss Re curve in Bernegger's formulation.

Recall from Section 2.5 that according to Bernegger's formulation we can reflect an increase in the size or how "spread out" a risk is by selecting larger values of c. Reflecting this fact in the characterization of which curve is appropriate for a given risk will have the effect of producing thinner tails in the exposure curves of larger properties. This can produce the counterintuitive, yet empirically observable, phenomenon that larger properties can have less excess loss potential than smaller properties in a given loss layer.

For our purposes it is necessary to appropriately vary exposure curves by risk to reflect distributional differences in underlying policy exposure to excess loss. For the purpose of the illustration we will assume that the Bernegger interpolation with c = 1.5 is appropriate for Houses, but not for Barns or Estates. For Barns we will assume that given a loss, a barn will suffer a total loss which is consistent with the very concentrated nature of such a small exposure; For Estates we will assume that the Bernegger interpolation with c = 4.0 (loosely corresponding to the fourth Swiss Re curve according to Bernegger) is appropriate. The results of the NCLL allocation reflecting these refined exposure curve selections are summarized below in Table 13.

Table 13: NCLL Allocation by State and Program Using PEBELS
(Reflect all Heterogeneity Adjustments by State and Program)

		Program		
State:	Barn	House	Estate	Total
X	0	2,420,543	3,284,039	5,704,582
Y	0	2,434,679	3,432,505	5,867,184
Z	0	2,319,639	3,029,034	5,348,674
Total	0	7,174,861	9,745,577	16,920,439

The results of this allocation are interesting in that NCLL exposure has shifted modestly from Estates to Houses. The larger the difference in underlying loss distribution, and thus exposure curve, the more dramatic this NCLL shift will be, and in practice it can be quite dramatic. Note that Section 2.2 discussed the phenomenon whereby larger exposures have a lower likelihood of having larger losses as a percentage of insured value, so this shift is consistent with our intuition of the underlying loss process. By reflecting the differences in loss distributional characteristic via selection of an appropriate exposure curve, we can quantify how much exposure a nominal amount of high layer of exposure incurs for a small property versus a larger property, a quantity we were not previously able to measure.

It is worth noting that determining an appropriate exposure curve requires care and judgment. Furthermore, the universe of available exposure curves is less expansive than one would hope for. That being said, there are reasonable curves available for these applications that at least allow the practitioner to reflect major loss distributional differences for significantly different classes of risks. While a comprehensive discussion of exposure curves is beyond the scope of this paper, it is also worth noting that there exists significant opportunity to expand the library and understanding of exposure curves. Currently available exposure curve options are discussed in Section 2.15.

2.9 Historical versus Prospective Exposure Distribution Generalization

To this point we have made no effort to specify when the exposure distribution to be used in computing the PEBEL or NLE should be evaluated. For common analyses requiring an exposure distribution such as catastrophe modeling or reinsurance quoting, the latest available exposure distribution is desired, which is reasonable given the prospective nature of those tasks. Therefore we will define the latest available exposure distribution to be the *prospective exposure distribution*.

For loss ratio rate making we are uniquely interested in allocating the *historical losses* stated at their *prospective level*. By virtue of producing loss ratio based rate level indications, we are using historical experience we have determined to be relevant as the basis for our loss ratio projection. The exercise now is to allocate that experience to where exposure was earned. This is a fundamentally different task than catastrophe modeling where simulations are used as the basis for the projection instead of historical experience.

The losses used in the examples throughout this paper are stated at their prospective level by virtue of the assumption that they have been developed and trended as specified in Section 2.1. This is a fair assumption because developing and trending property losses are both routine tasks in practice. The challenge this paper was written to address is how to allocate the historical NCLL to the *historical exposures* whose experience forms the basis for our loss ratio based rate level indications. Using the prospective exposure distribution to allocate historical NCLL can lead to significantly skewed and in some cases nonsensical loss ratio based rate level indications.

For example, consider a situation where the company has decided to non-renew all policies in State Z over the last year in order to curtail catastrophe risk. If that were the case then the prospective exposure distribution would reflect the fact that there is no longer any business in force in State Z and thus all historical NCLL experience would be allocated between states X and Y, dramatically overstating their rate level indications. This example was deliberately dramatic (though not unrealistic) in order to demonstrate the phenomenon. In practice even small shifts will distort the NCLL allocation if the prospective exposure distribution is used in place of historical.

Another example is an insurer entering a new state. Suppose the company has decided to expand its writings into State N this last year. The company's new product has proven to be competitive in State N and the insurer now has a comparable number of policies in State N as it does in X, Y and Z. Using the prospective exposure distribution will cause roughly a fourth of the NCLL experience from the entire experience period to be allocated to state N even though it has not yet earned even a full year's exposure. This will cause the rate level indications for states X, Y and Z to be understated.

Thus we need a meaningful definition of *historical exposures* which can be used to develop the NCLL allocations needed for our rate level indications. To accomplish this we define the historical exposures to be the sum of the "earned" exposures over the experience period. Of course since earned exposures will never be available in our data sources, we will have to determine a reliable proxy for it. To do this we can use an accountant's trick. We observe that exposures (and written premiums) of in force policies are stated as of a specified date in the same way that assets on a balance sheet are stated as of a specified date. For accounting purposes the average value of assets during the year is often taken as the average of the asset values at the beginning and end of a year. Extrapolating this rationale, we can estimate earned exposures of a specific segment during a given year as the average of the historical exposures at the beginning and end of the year.

Of course this is not to say that there is no value to computing PEBEL using the prospective exposure distribution. Such an analysis could be quite useful for NCLL exposure monitoring, analogous to the catastrophe exposure modeling frequently performed by insurance companies. It is just not the right exposure distribution for loss ratio ratemaking.

2.10 Low Credibility Generalization

Keep in mind that all of the analysis above assumes that historical NCLL experience is fully credible in total. If this is not the case, then credibility methods must be applied in order to develop a reliable total expected historical NCLL provision that can be allocated.

The first thing to note is that we can and should band historical NCLL experience in layers to get more credibility out of our data. In the examples derived above we have considered only two bands: the primary layer and the excess layer from $100,000 to infinity which was selected to keep the examples simple. However, in practice we will want to band to maximize credibility in lower bands and give ourselves an opportunity to use reasonable complements in less credible bands.

Consider the following simple banding scheme that can be used to illustrate the method:

- $0 to $100,000 – the primary layer
- $100,001 to $500,000 – the first excess layer which we will assume fully credible
- $500,001 to infinity – the high excess layer assumed to be only 30% credible

The PEBEL per policy will now have to be computed by layer. The only difference in the computation will be that the argument of the exposure curve, G(x) must be limited to the upper and lower bounds of the layer when determining the exposure factor EF_i for the given layer (the example reproduced in Table 7 provides a numerical example of this). A separate allocation will then be performed for each layer using the techniques developed in this paper.

The only new complication in this scenario is that the NCLL experience in the $500,001 to infinity layer is only 30% credible. Thus it must be credibility weighted with an appropriate complement before the allocation can be completed.

The task then is to identify a meaningful complement of credibility. One method suggested by Clark [2] for pricing non-credible per risk reinsurance layers was to use the actual experience in a credible lower loss layer to approximate the expected experience in a non-credible higher loss layer by applying the relationship implied by the PEBEL in each layer. We can apply this approach in our problem to derive reasonable complements of credibility from the $100,001 to $500,000 layer as follows:

$$NCLL_{\$0.1M\ to\ \$0.5M}^{Historical} * \frac{PEBEL_{\$0.5M\ to\ infinity}^{Historical}}{PEBEL_{\$0.1M\ to\ \$0.5M}^{Historical}} \tag{2.11}$$

Or alternately using the primary layer as follows:

$$NCLL_{\$0\ to\ \$0.1M}^{Historical} * \frac{PEBEL_{\$0.5M\ to\ infinity}^{Historical}}{PEBEL_{\$0\ to\ \$0.1M}^{Historical}} \tag{2.12}$$

An alternate complement can be inferred from the quoted cost of per risk excess of loss reinsurance rates using the formulation below:

$$(Direct\ Earned\ Premium) * (Reinsurance\ Rate) * (Reinsurer's\ Permissible\ LR) \tag{2.13}$$

The reinsurance rate in this context would be specific to the company's property per risk excess of loss treaty which would probably be quoted by layer. This complement is messy for several reasons including:

- It will likely cover multiple property lines as opposed to just the line under review
- The rate will likely cover catastrophic excess loss in addition to non-catastrophic
- The quoted layers may not align with those being used for PEBELS analysis
- The reinsurer's permissible loss ratio will not be obvious

However, despite all these complications this complement has the very desirable property that it reflects the *actual* cost to the insurer for the reinsured layer of loss because this is what the insurer *actually* pays to insure those layers. This gives it a very concrete interpretation, makes it easy to explain, and easy to defend. In fact, the smaller an insurer's property portfolio is, and thus the lower the credibility implicit in the insurer's actual experience, the more attractive this complement becomes relative to the alternative.

2.11 The Scope of PEBELS

Property actuaries do not currently have a systematic framework to quantify the non-linear relationship between expected losses in primary and excess loss layers. For liability insurance, increased loss factors (ILFs) have long been the framework for quantifying this relationship. In practice ILFs are often developed based on a broad group of risks and then imposed on similar risks in a straight forward manner which makes them easy to use and explain. A key reason for the crispness of this formulation is due to the large degree of independence between the loss process and the limit purchased by the insured which simplifies the formulation greatly [7].

Property is not afforded any such simplifying assumptions. On the contrary, property lines are plagued by a high level of dependence between insured value and the underlying loss process. Furthermore, property lines are also beset by complications such as catastrophic exposure, diverse capital requirements, etc. PEBELS provides a framework from which to reflect and quantify all of these moving parts simultaneously in determining the relationship between the expected primary and excess loss.

Even though PEBELS was created to solve a very specific allocation problem, it evolved into the theory of property excess loss rating described above. The theoretical framework of inputs and levers for allocating property losses between primary and excess layers given in the sections above is general enough that it can be tailored for use in any application that requires segmentation between primary and excess property loss layers given appropriate actuarial input assumptions.

The sections below provide some examples of applications which could benefit from the segmentation between property loss layers that PEBELS affords.

2.12 PEBELS and Adjusting Modeled Catastrophe AALs

Traditionally, catastrophe models have assumed that *average annual losses* (AALs) are linearly proportional to exposure. However, we know both from practice and from the theory presented in this paper that this is not a valid assumption. All of our discussion about the theory of exposure curves and the property loss distributions that underlie them suggest that property loss costs are non-linear. This a priori reasoning was empirically validated for catastrophe loss costs by the hurricane exposure curve study completed by Ludwig using Hurricane Hugo [5].

As such, we would expect, given that total AAL results are calibrated to total industry catastrophes, that the catastrophe loss ratios implied by modeled AALs should be understated for property books with smaller insured values (such as personal lines) and overstated for property books with larger insured values (such as commercial lines). This reasoning is consistent with anecdotal observations.

As discussed in Section 2.11, the methods developed from PEBELS provide the framework from which to quantify an appropriate measure of the amount by which modeled AALs should increase with a linear increase in exposure. Adjusting modeled AALs for this PEBELS effect would eliminate the systematic bias from implicit modeling assumption that AALs increase linearly with exposure.

To implement this type of refinement the practitioner would need to identify appropriate exposure curves to be used for the given catastrophe exposure under consideration. This would require care and judgment, but no more so than any of the myriad of other assumptions that comprise a typical catastrophe model.

Of course there exists an opportunity for catastrophe modelers to directly enhance their models to reflect the non-linearity of expected catastrophe loss cost with insured value in their models. However, until such a refinement is implemented within the models, insurers can directly adjust their own AALs estimates using a post-modeling PEBELS adjustment.

2.13 PEBELS and Predictive Models

It is common practice when creating predictive models such as generalized linear models (GLMs) to use separate models to predict the frequency and severity components of the reviewed line of business's expected loss. Not surprisingly, the severity model tends to be tremendously less stable than the frequency model for lines and coverages with significant exposure to very infrequent but very large losses, such as the property exposures this paper has focused on. This phenomenon is amplified when property coverage is modeled by peril, because total loss for perils such as fire and lightning are disproportionately driven by very infrequent but very large losses relative to other perils.

My hypothesis is that modeling on PEBELS for representative layers in the models would be more predictive than modeling on insured value alone for the reasons explored in the this paper. This is plausible because PEBELS was developed to contemplate actual expected exposure within a layer of insured value and thus contains more information about layer exposure than raw insured value does, assuming that reasonable exposure curves and heterogeneity assumptions are applied.

While PEBELS estimates do contain frequency information from the exposure curve assumptions, the biggest benefit from using PEBELS in predictive modeling would likely be from refining the severity model where large loss observations are so sparse. This is true whether a single uncapped severity model is used, or if uncapped severity is modeled using a combination of 1) a capped severity model (say the cap is $100,000 to be consistent with our preceding examples), 2) a propensity model (the probability that the claim amount will exceed $100,000 given that there is a claim), and 3) an excess model (the severity of the excess portion of the claim given that the claim amount exceeds $100,000). When the combination model is used I would hypothesize that PEBELS would be most predictive in the excess model.

2.14 PEBELS and Revised Property Per Risk Reinsurance Exposure Rating

Besides moving to a per policy computation and reflecting heterogeneity by policy which are discussed in Sections 2.7 and 2.8 above, a major point in the PEBELS formulation discussed in Section 2.9 is the importance of distinguishing between the historical exposure distribution and prospective exposure distribution, and applying the appropriate distribution for the appropriate application.

Clark [2] states that property per risk exposure rates are generally used to determine the exposure relativities between a credible and non-credible layer. This relativity is then multiplied by the ratio to historical losses in the credible layer to develop the indicated exposure base rate as detailed below:

$$NCLL_{Non-Credible\ Higher\ Layer}^{Expected\ Prospective} = NCLL_{Credible\ Lower\ Layer}^{Historical} * \frac{PEBEL_{Non-Credible\ Higher\ Layer}^{Prospective}}{PEBEL_{Credible\ Lower\ Layer}^{Prospective}} \quad (2.14)$$

Based on the reasoning presented in Section 2.9 there are two problems with this approach. First, there is no reason that the ratio of PEBEL between the higher and lower layers based on the prospective exposure distribution is representative to the relationship based on the historical exposure distribution (which is the relationship we need to establish). Second, there is no adjustment to reflect the relative change in exposure between prospective and historical exposure distributions for the layer being rated.

An alternate formulation that would reconcile both issues is:

$$NCLL_{Non-Credible\ Higher\ Layer}^{Expected\ Prospective} = (NCLL_{Credible\ Lower\ Layer}^{Historical}) * \left(\frac{PEBEL_{Non-Credible\ Higher\ Layer}^{Historical}}{PEBEL_{Credible\ Lower\ Layer}^{Historical}}\right) * \left(\frac{PEBEL_{Non-Credible\ Higher\ Layer}^{Prospective}}{PEBEL_{Non-Credible\ Higher\ Layer}^{Historical.Annualized}}\right) \quad (2.15)$$

The first ratio adjusts for the *historical* ratio of the high layer exposure relative to low layer exposure, while the second ratio allows us to quantify the prospective layer exposure relative to the historical, which for all the reasons outlined in Section 2.9 is more accurate than the traditional approach. The historical PEBEL can be *annualized* by simply dividing by the total number of years in the experience period.

Note that for reinsurance pricing we are using historical experience to project a prospective loss cost so there is an additional layer of computation to bring the estimate to the level of the prospective exposure distribution, which is not done in loss ratio ratemaking as described in Section 2.9.

My anticipation is that the most difficult part of implementing this more refined exposure rating algorithm will be obtaining the historical exposure profiles from the reinsured. However, in many cases the reinsurer should have a history of exposure profiles from past quotes if the insured has renewed multiple times. Furthermore, given the prevalence of catastrophe modeling, it is becoming increasingly likely that the primary insurer may have a history of historical exposure profiles readily available.

2.15 A Final Word on Exposure Curves

In practice specifying the exposure curves is a key step in a PEBELS analysis. Furthermore, the practitioner should be aware that the more complex the underlying policy is, the more judgment is required. An example of an area where judgment must be applied is in defining "insured value" for large commercial risks with a stand-alone contents limit. A reasonable approach is to add this value to the stand-alone building limit in determining the total insured value since they are both exposed to a total loss; however, this is not necessarily the optimal choice since the building and contents may not share the same underlying loss distribution, thus the need for actuarial judgment. Another wrinkle is the question of how to incorporate the limit for business interruption coverage, especially when coverage is provided on an "actual loss sustained" basis, in which case the exposure is essentially unlimited.

Despite the additional judgment required for more complex risks, analyses involving these more complex risks tend to benefit most strongly from a PEBELS approach because it is with these risks that data are least credible and heterogeneity among policies is greatest, affording them the most utility from PEBELS's ability to segment excess loss exposure based on policy characteristics.

Another critical area where judgment is required is in specifying the exposure curve (or curves) to be used in the analysis. The more homogeneous the book of business being analyzed, the more likely it is that using a single exposure curve in the analysis is a reasonable approach. However, regardless of how many curves are required, a decision must ultimately be made as to what curve to use for each given policy. Ideally we would select curves that are particularly appropriate for the risks under consideration, but the universe of available curves is limited so we must operate within our available choices.

There are a few publically available curves published in the actuarial literature, but only the fitted Bernegger curves offer more than a few points, which is why I used them for the illustrations in this paper. There are also some well-known curves such as the Swiss Re or Lloyd's of London curves. Additionally, commercially produced curves such as ISO's PSOLD curves may be available for purchase. A motivated insurer may also choose to create proprietary curves.

Lastly, an interesting property of exposure curves that was highlighted in Ludwig's study [5] is that exposure curves vary by peril. While this conclusion may seem somewhat obvious, it is significant in that most of the available exposure curves are given on a combined peril basis. In most cases they are not segmented by many attributes at all, even ones that might obviously seem predictive of the underlying loss distribution such as class, occupancy or protection.

3. RESULTS AND DISCUSSION

The original intention of the project that led to the development of PEBELS was to identify a method from the existing literature via literature review. The expectation was to find a method or theory analogous to that of ILFs used in liability insurance or ELFs used in workers compensation insurance, but no comparable method or theory for property insurance was identified. However, theory on property exposure curves and methods for pricing property per risk treaties in the property reinsurance literature was identified, and came to form the theoretical foundation upon which PEBELS was built.

The key result of this work was the development of PEBELS, which was designed to be an accurate and practical theory of property non-catastrophe excess loss exposure. Although the scope of the original problem was quite narrow, the numerous challenges worked through to develop an allocation basis that passed the necessary reasonability tests led to the development of an increasingly sophisticated model in order to quantify the non-linearity between the expected loss costs in different loss layers for a given property risk. PEBELS provides a framework that is general enough to apply in a variety of situations.

Because PEBELS is a theory of expected property excess loss by layer, it can be used in any application where such tool would be of value. Some applications explored in this paper were

- Catastrophe Modeling – to correct for implicit assumption that AAL is linearly proportional to insured value
- Predictive Modeling – to use PEBELS to improve predictiveness of severity models
- Property Per Risk Reinsurance Pricing – to use findings to refine the existing model

4. CONCLUSIONS

PEBELS is a theory of property excess loss exposure that can be used to quantify the non-linear relationship between expected loss costs in different loss layers. PEBELS is the closest analog that property insurance has to ILFs for liability insurance.

PEBELS requires far more care and judgment to use than ILFs do, mainly due to the presence of many more sources of heterogeneity between policies and the interactions between those sources of heterogeneity present in property insurance relative to liability. The PEBELS formulation in this paper provides the framework of theory and practical levers required to utilize this theory in practical applications.

Acknowledgment

My thanks to David VanKoevering, my manager while in Commercial Property Pricing, for giving me the opportunity to research and develop the method that became the foundation for this paper, and to his manager Eric Drummond-Hay for giving him the opportunity to do so.

And my special thanks to my wife Asanka, my parents Mark and Karin, and my wife's parents Anura and Nalani without whose abundant support this paper, among so many other things, would not be a reality.

Supplementary Material

Work files attached: Data simulation, and exhibits.

5. REFERENCES

[1] Bernegger, S. (1997). Swiss Re exposure curves and the MBBEFD distribution class. *ASTIN Bulletin, Volume 27, No. 1*.

[2] Clark, D.R. (1996). Basics of reinsurance pricing. *Casualty Actuarial Society Exam Study Note*.

[3] Dean*, C.G.,* Hafling, D.N., Wenger, M.S., Wilson, W.F. (1998). Smoothing weather losses: A two-sided percentile model. *Proceedings of the Casualty Actuarial Society LXXXV*.

[4] Hurley, R.L. (1954). A credibility framework for gauging fire classification experience. *Proceedings of the Casualty Actuarial Society XLII.*

[5] Ludwig, S.J. (1991). An exposure rating approach to pricing property excess-of-loss reinsurance. *Proceedings of the Casualty Actuarial Society LXXVIII.*

[6] Mahler, H.C. (1990). An example of credibility and shifting risk parameters. *Proceedings of the Casualty Actuarial Society*

LXXVII.
[7] Miccolis, R.S. (1977). On the theory of increased limits and excess of loss pricing. *Proceedings of the Casualty Actuarial Society LXIV*.
[8] Salzmann, R. (1963). Rating by layer of insurance. *Proceedings of the Casualty Actuarial Society L*.
[9] Werner, G., Modlin C. (2010). Basic Ratemaking. *Casualty Actuarial Society Exam Study Note*.

Abbreviations and notations
NCLL, Non-Catastrophe Large Loss
PEBEL, Policy Exposure Based Excess Loss
PEBELS, Policy Exposure Based Excess Loss Smoothing

Biography of the Author

Marquis J. Moehring is an assistant actuary at Liberty Mutual Personal Insurance in Seattle, WA. He works on the Safeco Strategic Auto Research Team, prior to that he completed pricing rotations in Commercial Property, Commercial General Liability, and Personal Auto. Marquis is an Associate of the Casualty Actuarial Society.

Catastrophe Pricing: Making Sense of the Alternatives

Ira Robbin, Ph.D.

Abstract:
This paper examines different ways of pricing catastrophe (CAT) coverage for reinsurance treaties and large insurance accounts. While all the methods use CAT loss simulation model statistics, they use different statistics and different algorithms to arrive at indicated prices. This paper will provide the reader with the conceptual foundations and practical insights for understanding alternative approaches.

Keywords: Catastrophe Pricing, Risk Measure, Coherence, Capital Allocation

1. INTRODUCTION

This paper will examine different ways of pricing the property catastrophe (CAT) coverage provided by reinsurance treaties and by insurance policies written on large accounts. All the approaches utilize CAT loss simulation models to estimate key statistics. However, they differ in the choice of statistics and their ways of translating statistics into premiums.

For the type of business under discussion, all widely-accepted methods use a Return on Risk-adjusted Capital (RORAC) approach. Indicated prices are defined as those that hit the specified target return on an amount of required capital that has been adjusted for risk. Where the methods differ is in how they arrive at the amount of required capital. They employ different risk measures and they use risk measures in different ways.

One objective of this paper is to bridge the gap of understanding between abstract formulas and black-box simulation software. It will provide definitions and also demonstrate how to apply them. It will show, in particular, how to properly define relevant risk measures on sets of discrete data such as those found in a deck of simulated CAT losses by year. The paper will feature a comparison of pricing indications for a few hypothetical accounts. In the end, the reader should have a better understanding of the formulas and a balanced framework for evaluating the alternatives.

1.1 Existing Literature

In 1990, Kreps [13] published an influential paper on Marginal Capital methods for pricing reinsurance treaties. He used Marginal Variance as a metric for determining Marginal Capital. Mango [15] observed that applying Marginal Variance or Marginal Standard Deviation to CAT pricing led to pricing that was Order Dependent. He proposed an application of game theory to eliminate the order dependency.

In the late 1990's, the Value at Risk (VaR) concept moved from the financial literature into rating agency capital requirement models. It was then but a short step before Incremental VaR (also referred to as Marginal VaR) approaches to required capital made their way into CAT pricing algorithms. However, theoreticians, such as Artzner, Delbaen, Eber, and Heath [4], had proposed a set of fundamental axiomatic properties called *coherence* and observed that VaR is not coherent. Others, including Acerbi and Tashce [1] and Wang [24], also objected to VaR, citing its incoherence along with other objections. Order dependence is also a problem for Marginal VaR. Meyers, Klinker, and Lalonde [19] also highlighted the point that the Marginal VaR method does not calibrate automatically with the portfolio and introduced an overall adjustment factor to achieve such calibration.

Rockafellar and Uryasev [22], Acerbi and Tasche [1] and others have promoted use of Tail Value at Risk (TVaR) as a risk metric, citing its coherence as a key advantage. As Uryasev and Rockafellar [22] explained, TVaR may not be the same as Conditional Tail Expectation (CTE) in discrete cases.[1] This paper will explain and demonstrate the difference.

Kreps [14] developed a general riskiness leverage model and also proposed co-statistics as an alternative to incremental statistics. Co-statistics provide a method for allocating portfolio capital based on account contributions to portfolio results. This approach has great intuitive appeal, but the application of the contribution concept to tail statistics such as VaR and TVaR does not produce perfectly well-behaved risk measures. As will be shown later with a simple example, Co-VaR is unstable in that modest changes in the inputs can lead to wild swings in the resulting statistic[2]. A new result in this paper is that Co-TVaR fails subadditivity, a defining property of coherence. This is perhaps a bit surprising since TVaR is coherent and one might have hoped co-measures would inherit coherence properties.

Other authors such as Bodoff [6] and Wang [24] have questioned the exclusive focus on the tail inherent in VaR and TVaR. Bodoff proposed a Percentile Allocation approach and Wang proposed use of Distortion Measures.

Venter [23] and Goldfarb [12] have published excellent survey articles that provide useful

---

[1] Many authors incorrectly assume TVaR and CTE are the same, but this is only true for continuous distributions. For example, D'Arcy [9] wrote, "…Tail-Value-at-Risk (TVaR), which is also termed Tail Conditional Expectation (TCE), takes the average of all values above a particular percentile….TVaR only provides the average loss if a loss in excess of the TVaR threshold were to occur."

[2] From informal discussions, the author is aware that this behavior is known to several analysts. However, it does not seem to have been previously noted in the literature.

background and discussion. What is clear from a review of existing literature is that the field is not settled. Practitioners use approaches decried by theoreticians and shrug off anomalies. Yet, the experience of 2010-2012 suggests that sub-optimal approaches were used that may have inadvertently promoted "de-worsification".[3] Overall, there is no want of proposals and opinions.

1.2 Organization of the Paper

The discussion will begin in Chapter 2 with a definition of the indicated account premium based on required CAT capital. Chapter 3 will present an overview of risk measure theory and then examine several commonly used risk measures. Chapter 4 will discuss the different algorithms used to set CAT capital for an account.

Then in Chapter 5 several particular algorithms will be reviewed. Finally in Chapter 6, simulation will be used to generate hypothetical years of CAT loss data for a portfolio and two sample accounts. Then risk measures and algorithms will be applied to this set of simulated data to arrive at prices based on different procedures. Chapter 7 will provide several general observations and Chapter 8 will have a summary and conclusion.

2. INDICATED PREMIUM ALGORITHMS

Let X denote the CAT Loss for a particular account and let P(X) denote the indicated premium prior to any loading for expenses. We may write the indicated premium as the sum of the expected CAT Loss, E[X] plus a risk load, RL(X).

$$\boldsymbol{P(X) = \mathrm{E}[X] + \mathrm{RL}(X)} \qquad \text{Equation (2.1)}$$

We define some basic properties that an indicated pricing algorithm formula should obey.

Table 1

<u>Premium Algorithm Basic Properties</u>

1. **Monotonic: If $X_1 \leq X_2$, then $P(X_1) \leq P(X_2)$.**
2. **Pure: If $X \equiv \alpha$ a constant, then RL(X) =0 and P(X)=E[X] = α.**

---

[3] "De-worsification" is a term used (see article by Brodsky [7]) to describe overzealous diversification to the point that business is written with minimal or no profit load in order to diversify a portfolio.

3. **Bounded: If X ≤ K, then P(X) ≤ K.**

4. **Continuous(Stable): P(X) is a continuous function[4] of X.**

We say a premium calculation algorithm is **coherent** if it also satisfies:

Table 2

Premium Algorithm Coherence Properties

1. **Scalable: P(λX) =λ·P(X)**

2. **Translation Invariant: P(X+α) = P(X)+α.**

3. **Subadditive: $P(X_1 + X_2) \le P(X_1) + P(X_2)$**

Other authors first define coherence as a set of properties of risk measures, but from our perspective it seems more natural to define it first with respect to the indicated premiums.

2.1 RORAC Pricing Formula

Write C(X) to stand for the Required CAT Capital for the account. Applying the RORAC approach to pricing, the risk load is given by applying a target return, r_{Target}, against required capital. Thus the indicated premium is given as:

$$P(X) = \mathrm{E}[X] + r_{Target} \cdot C(X) \qquad \text{Equation (2.1.1)}$$

Note Equation 2.1.1 produces a premium equal to the Expected Loss when the required capital is zero.

3. RISK MEASURES

The different premium calculations use different formulas for computing Required CAT Capital for an account. They employ different functions, called risk measures, to quantify risk. The risk measures are also used in different algorithmic procedures.

3.1 Risk Measures

We define a risk measure simply as a mapping from real-valued random variables to the non-negative real numbers. It is desirable that a risk measure, ρ, obey basic properties that correspond to

[4] Continuity is with respect to the L^1 topology.

the basic premium algorithm properties in Table 1:

Table 3

Risk Measure Basic Properties

1. **Monotonic: If $X_1 \leq X_2$, then $E[X_1]+\rho(X_1) \leq E[X_2]+\rho(X_2)$**
2. **Pure: If $X \equiv \alpha$ a constant, then $\rho(X)=0$**
3. **Bounded: If $X \leq K$, then $\rho(X) \leq K$.**
4. **Continuous(Stable): $\rho(X)$ is a continuous function of X.**

Note monotonicity for risk measures is more complicated than the corresponding property for indicated premiums. The reason is that if one random variable always exceeds another, it need not have more risk. However, under Property #1 in Table 3, the sum of its measured risk plus its expected value should exceed the corresponding sum for the smaller random variable. The idea of purity in a risk measure is that, if the outcome is known and fixed, a pure risk metric will say there is no risk.[5] Some popular risk measures such as VaR are not pure. However, unless a risk measure is pure, it may produce a pricing algorithm that does not satisfy Basic Property 2 in Table 1 and thus may generate a strictly positive risk load even if the loss is fixed and constant. Boundedness requires that the measured risk be no greater than the largest possible loss, if there is such a largest loss. Finally, the property of continuity means that small changes in the points or probabilities do not lead to large changes in the measured value of risk. As we will see later, some commonly used premium algorithms are not stable.

Going beyond the basic properties, we now turn to the special properties of coherence. We say a risk measure is **coherent** if it is scalable, translation invariant, and subadditive.

Table 4

Coherence Properties of Risk Measures

1. **Scalable: $\rho(\lambda X) = \lambda \cdot \rho(X)$**
2. **Translation Invariant: $\rho(X+\alpha) = \rho(X)$ when $0 \leq \alpha$.**
3. **Subadditive: $\rho(X_1 + X_2) \leq \rho(X_1) + \rho(X_2)$**

[5] Panjer and Jing [20] present a similar property called "riskless allocation" under which a risk with no uncertainty gets a zero capital allocation.

These definitions are arranged so they are consistent with the comparable coherent premium calculation properties.

Some (see Panjer and Jing [20]) believe scalability is a fundamental property expressing the idea that currency conversion should not impact measured risk. Unfortunately, some commonly-used measures of risk, such as variance, do not obey the scaling property. Acerbi and Tasche [1] refuse to refer to a function as a risk measure if it does not obey coherence properties. While we agree with the points these authors have made, we choose to use a very broad definition of a risk measure. This is for convenience. We need a ready way to refer to the various functions that are used, rightly or wrongly, to measure risk in pricing CAT covers.

3.1.1 Sign Conventions

Our approach to sign conventions is the same as that taken in most papers on property and casualty insurance. However, in reviewing the academic and financial literature, a reader may find it confusing because sign conventions are the reverse of what we have used. For example, many authors, including Artzner, Delbaen, Eber and Heath [4] as well as Acerbi and Tashce [1], state the property of translation invariance as $\rho(X + \alpha) = \rho(X) - \alpha$.

From this perspective, the random variable X represents the net outcome: so a positive result is favorable. Adding a constant reduces the measured amount of risk. For our work, the random variable X represents the CAT loss; so larger positives are less favorable.

Many of the other financial and academic authors also allow the risk measure to be negative, while for our purposes we always want the risk measure to be non-negative. Again most of the writers focusing on property and casualty insurance assume, implicitly or explicitly, that risk measures are non-negative.[6]

3.1.2 Risk Measure Definitions on Distributions with Mass Points

In a CAT pricing context, it is important to understand just how to apply the definition of any particular risk measure to a loss distribution that has mass points. There are three reasons for this. First, it is common practice to use a CAT simulation model to generate thousands of simulated years of results. So any measure needs to be defined on such a set of discrete sample points. Second, the impact of catastrophic events on underlying building values may produce real mass points. For

---

[6] In Wang [24] the translation invariance property is presented with all positive signs and in Venter [23] all the risk measures are non-negative.

example, if a risk has a $100 million beachfront hotel in a hurricane prone locale and no other property nearby, there could be a natural mass point at $100 million. Finally, limits and layering could also give rise to mass points.

3.2 Specific Risk Measures

We will now list several risk measures, discuss them briefly, and then show how they work on a small set of hypothetical sample loss data.

Table 5

Specific Risk Measures

1. **Variance:** $Var(X) = E[(X-\mu)^2]$
2. **Semivariance:** $Var^+(X) = E[(X-\mu)^2 | X \geq \mu] \cdot Prob(X \geq \mu)$
3. **Standard Deviation:** $\sigma = \sqrt{Var(X)}$
4. **Semi Standard Deviation:** $\sigma^+ = \sqrt{Var^+(X)}$
5. **Value at Risk:** $VaR(\theta) = \inf\{x \mid F(x) \geq \theta\}$ **for** $0 < \theta < 1$
6. **Tail Value at Risk:**

$$TVaR(\theta) = \frac{1}{1-\theta}\begin{bmatrix} E[X|X > VaR(\theta)] \cdot (1 - F(VaR(\theta))) \\ + \\ \{1 - \theta - (1 - F(VaR(\theta)))\} \cdot VaR(\theta) \end{bmatrix}$$

7. **Excess Tail Value at Risk:** $XTVaR(\theta) = TVaR(\theta) - \mu$
8. **Distortion Risk Measure:** $E^*[X] = E[X^*]$
   where F*(x) = g(F(X)) for g a distortion function
9. **Excess Distortion Risk Measure:** $E^*[X] - \mu$
   where E*[X] is the mean under a distortion risk measure
   and μ = E[X]

3.2.1 Variance and Standard Deviation

Variance and Standard Deviation are well-known. Both are pure risk measures in the sense we have defined that term. However, Variance is not scalable and is therefore not coherent, but Standard Deviation is.

Semivariance has been advocated by Fu and Khury [11] and SemiStandard Deviation does obey the scaling property (Property 1 in Table 5). A major appeal of both is that they do not count favorable deviations as part of the risk. Exhibit 1, Sheet 1 shows the computation of these measures in an example with twenty sample points.

3.2.2 VaR, TVaR, and XTVaR

VaR is a well-known metric that was developed in financial and investment risk analysis settings. VaR(θ) is the θ^{th} percentile. It is intuitively the "best of the worst", the least unfavorable outcome from the worst (1- θ) % of outcomes. This may not be precisely correct when there are mass points and the cumulative distribution jumps discontinuously. To make the definition work in the general case, the "infimum" formulation is used[7]. VaR is not pure: its value on a constant is equal to that constant. VaR is not subadditive and therefore not coherent.

TVaR is intuitively the average loss in the worst (1- θ) % of outcomes. As the definition in Table 5 indicates, the computation of this average entails evaluating two terms. The first term is the product of the Conditional Tail Expectation (CTE) where CTE $= E[X|X > VaR(\theta)]$ and the probability weight corresponding to the tail of X values strictly larger than VaR(θ). The second term is the product of VaR(θ) times the residual probability needed so that the total of the two probabilities equals (1- θ). The sum of these two terms is then divided by (1 -θ) to obtain TVaR. For example, suppose VaR(99.0%) = 150 and assume the probability of X values strictly larger than 150 is 0.8%. It follows the mass at 150 is equal to or larger than 0.2%. Also assume the CTE is 200. Then TVaR(99.0%) equals 190 since 190 =(200*0.8% + 150*0.2%)/1.0%.

In the general case, the residual probability will be non-zero only when there is a strictly positive mass point at VaR(θ). When there is no mass point, TVaR is equal to CTE. However, in the discrete case, 1-F(VaR(θ)), the probability of exceeding VaR(θ), may be less than 1- θ. If that scenario, CTE is not truly an average of the worst (1-θ) of outcomes and the second term is needed. Our definition follows that of several authors (Rockafellar and Uryasev [22], and Acerbi and Tasche [1]) by including a portion of the mass point at VaR(θ) in the definition of TVaR.

When evaluating VaR and TVaR on a sample of trial data as might be generated by a simulation model, the formulas can be stated in a more straightforward way using the rankings of ordered data.

---

[7] The infimum is this context is the lower bound of the tail.

Table 6

Ranking Definitions of VaR and TVaR

Let $X_1 \geq X_2 \geq \ldots \geq X_n$ be an ordering of n trails of X.

If k=(1-θ)n, then:

1. $\mathbf{VaR(\theta) = X_k}$
2. $\mathbf{TVaR(\theta)} = \frac{1}{k}\left(\sum_{j=1}^{k} X_j\right)$

For instance, if there are 20 trials, the largest trial is VaR(95%) since 1= (1-.95)*20 and the smallest trial is VaR(0%) since 20= (1-0)*20. Also observe that TVaR(0%) is the average of all 20 sample points. Under these definitions VaR(75%) is the 5th largest point and TVaR(75%) is the average of the 5 largest points.[8] Exhibit 1, Sheet 2 shows the computation of these measures using the "ranking" definitions in the simple example of twenty samples points. Note the repetition of some of the sample point values does not cause any difficulty for the ranking definitions.

TVaR is not pure, but it is subadditive. As Acerbi and Tasche [1] and others have noted, the CTE may fail subadditivity in the discrete case[9]. XTVaR captures the average amount by which the worst (1-θ)% of loss outcomes exceeds the mean. Note TVaR is automatically larger than the mean so XTVaR is non-negative. In cases where X is constant, TVaR will be equal to the expectation, while XTVaR will be zero. Thus XTVaR is a pure risk measure, while TVaR is not.

3.2.3 Distortion Risk Measures

Wang [24] has proposed use of distortion risk measures and shown that they are coherent when g is continuous. We have not seen the excess risk measure in the literature, but we list it as an obvious extension that leads to a coherent and pure risk measure.

One particular distortion is the Wang shift[10]:

$$\mathbf{g(u) = \Phi[\Phi^{-1}(u) - \lambda]} \qquad \text{Equation (3.2.3.1)}$$

---

[8] One alternative is to use the average of the 4 largest points out of 20 for TVaR(75%) and more generally use the average of the k-1, instead k largest points, as the definition for TVaR(θ). In our view, the definition with k points is better, for one, because it results in TVaR(0) being equal to the average.

[9] Acerbi and Tasche[1] use the terminology Tail Conditional Expectation (TCE) where we use CTE.

[10] Wang also introduced the Proportional Hazards transform [26].

Here Φ is the standard unit normal and λ is the Wang shift parameter.

Exhibit 1, Sheet 3 has an example with columns showing how the original cumulative distribution is transformed. To pick a particular row, the second ranking point out of twenty sample points has an empirical cumulative distribution (cdf) of 0.950. The standard unit normal inverse of 0.950 is 1.645. The Wang shift in the example is 0.674. Subtracting this from 1.645 yields 0.971 and the standard normal has a cdf of 0.834 at 0.971. So the cdf is transformed from 95.0% to 83.4%. New transformed mass densities are derived by subtracting the transformed cdfs in sequence. Using the resulting transformed densities moves the mean from 10.0 to 16.7; so the resulting excess mean is 6.7.

4. ALGORITHMIC PROCEDURES

There are three different ways of using risk measures to arrive at a capital calculation algorithm. In this chapter, we will review these and define properties of such algorithms.

First is a **Standalone** capital computation. This simply involves looking at the value of a selected risk measure on an account's own CAT loss distribution. Even if this is not the final selected approach, it is usually a good idea to know the standalone value for any account as it provides a baseline for comparison. Some analysts would stop there and say the standalone capital from a coherent pure risk measure is the right answer. A variation of the standalone approach is the market equilibrium approach. This method attempts to find the theoretical price that should be charged for a given risk in a hypothetical fair and efficient market in equilibrium. While this may seem the direct opposite of a standalone approach, it can be rightly classed as a variant of a standalone concept in that the indicated price for an account does not depend on other risks in the portfolio.

Other methods do reflect the portfolio in their computations. The second general algorithm, the **Marginal** approach, determines required capital for an account by computing how it changes required capital for the portfolio. The third general algorithm, **Real Allocation,** entails use of a risk measure as an allocation base to allocate portfolio capital. This methodology may use one risk measure to compute required portfolio capital and then possibly another one as an allocation base.

Let X stand for the CAT Loss for a particular account and let R denote the portfolio CAT Loss excluding that account. Write C(X) to stand for the Account CAT Capital for the account. Equations for the three general procedures are:

Table 7

General CAT Capital Calculation Procedures

1. **Standalone :** $\mathbf{C(X) = \rho(X)}$
2. **Marginal:** $\mathbf{C(X \mid R) = \rho(X + R) - \rho(R)}$
3. **Real Allocation:** $\mathbf{C(X \mid R) = \left(\dfrac{\rho_2(X)}{\sum\limits_{Y \in (X+R)} \rho_2(Y)} \right) \cdot \rho_1(X + R)}$

We say the Marginal and Real Allocation procedures are **Portfolio Dependent** because the required capital amounts depend on the portfolio as well as on the account. We define several properties of Portfolio Dependent Algorithms.

Table 8

Portfolio Dependent Capital Properties

1. **Standalone Capital Cap:** A portfolio dependent algorithm is capped by standalone capital if $\mathbf{C(X \mid R) \leq C(X)}$.
2. **Automatically Calibrated:** A portfolio dependent algorithm is automatically calibrated if $\mathbf{\sum C(X|R) = C(R)}$
3. **Order Dependent:** A portfolio dependent algorithm is order dependent if $\mathbf{C(X_1 \mid R+X_2) \neq C(X_1 \mid R)}$ for some X_1 and X_2..

.

4.1 Calibration and Consolidation Benefits or Penalties

With both Standalone and Marginal approaches, there is a potential mismatch between the sum of individual account required capital amounts and the required capital for the portfolio taken as a whole. In other words, they are not automatically calibrated. If the portfolio requires less capital than the sum of the account capital requirements, as would be true for a subadditive measure, the difference will be called the **consolidation benefit.** However, not all risk measures are subadditive and therefore it is possible to have a **consolidation penalty**. So, even though it is a near universal truth in insurance that risk is reduced by pooling, that may not necessarily be the case with respect to

CAT coverage. Further, when subadditivity fails with a marginal algorithm, it is possible that required standalone capital is less than the indicated marginal capital. Such an algorithm would not satisfy the standalone cap.

Exhibit 2 has an example showing VaR is not subadditive. In the exhibit there are 20 trials with results for an account, "A", and a Reference Portfolio. The next column shows the computed sum, "A+ Ref", for each trial. Then there is a block labeled "Ordered Loss Data" showing the resulting rank-ordered losses for risk A, the Reference Portfolio, and the sum. The ordering is done separately for each. Since the 5th largest loss for account A is 4 and the 5th largest loss for the Reference Portfolio is 34, it follows that $VaR_A(75\%)$ is 4.and $VaR_{Ref}(75\%)$ is 34. Yet the 5th largest loss for the combined portfolio after A is added is 39. So the Marginal VaR for account A is 5 (= 39-34), which exceeds its Standalone VaR.

There are two general options for achieving calibration if desired when it does not occur automatically. A simple and direct idea is to apply a calibration factor to each initial required capital amount.[11] The other approach is to use game theoretic methods or other approaches to "auction off" any portfolio consolidation benefit (or penalty). Another view is that there is no need to reconcile the sum of the individual account required capital amounts and required capital for the portfolio. The key point is that users should be aware of the calibration properties of any proposed method and make adjustments or not according to their own philosophy.

4.2 Incoherence and Order Dependence

Marginal methods are also prone to generating required capital amounts that fail one or more of the coherence properties. They inherit incoherence if they are based on an incoherent measure, and they are also subject to order dependence[12] and other vulnerabilities arising from the nature of the incremental process.

For example, the scaling property does not necessarily hold if capital is based on Marginal VaR. This means that capital for a 50% share of a treaty is not necessarily twice the capital needed for a 25% share. See Exhibit 3 for an example demonstrating this.

TVaR has been shown to be coherent. However, if one is not careful with definitions and uses TCE (Tail Conditional Expectation) instead of TVaR, the resulting capital formula is not necessarily

---

[11] Meyers, Klinker, and LaLonde [19] proposed such a factor to calibrate capital requirements based on the Marginal VaR.

[12] Mango [15] showed the order dependence of marginal variance and marginal standard deviation.

even monotonic. In other words, adding an account may decrease the conditional tail expectation. Exhibit 4 demonstrates this.

4.3 Portfolio Allocation Methodologies

True allocation methodologies are less subject to calibration and order dependence issues than marginal methods. There are several different approaches to real allocation. Some are effectively a reprise of the methods used to set account capital; only in this context they are used to allocate a portfolio total already determined.

Table 9

Allocation Procedures

1. **Stand Alone**: Allocate portfolio capital in proportion to account Stand-alone risk measures

2. **Marginal:** Allocate portfolio capital in proportion to account marginal impact on portfolio risk measure.
   - Game Theory Modified Marginals: Adjust Marginals via Game Theory so allocations are not order dependent.

3. **Co-Measure**: Allocate portfolio capital using a co-measure.

4. **Percentile Allocation**: Allocate surplus based on allocations of each percentile of portfolio loss.

4.4 Standalone and Marginal Allocations

The first method uses an allocation base equal to the standalone value of a risk measure. This measure could be the same or different than the one used to compute required portfolio capital. The second approach uses allocations proportional to the account marginal increments. The order in which an account is added to a portfolio can have a strong influence on its marginal impact. This can lead to pricing anomalies. Mango [15] has developed a refinement of the marginal allocation approach in which game theoretic averaging over all orderings produces a more stable allocation.

4.5 Contribution Statistics

The third allocation concept is based on **Contribution Statistics** (Co-Statistics). The Co-Statistic for an account is conceptually the amount it contributes to the portfolio statistic, where the computation is based on an examination of the scenarios that determine the portfolio statistic.

Given that the portfolio had an adverse result, how much of it was due to a particular account? If portfolio capital is determined by a particular risk measure, the associated account co-measures are automatically additive and consistent with the portfolio total.

4.5.1. Co-VaR Instability

Though intuitively appealing, Co-Statistics may exhibit troublesome behavior. One significant problem is that Co-VaR can be unstable.[13] Specifically, this means Co-VaR can change dramatically due to small changes in the data. Even more troubling, different sets of simulations can yield quite different answers and increasing the number of trials may or may not yield convergence.

How does this arise? Consider the simulated events for the portfolio put in descending order. For each event, the contribution due to a specific treaty is also known from the simulation. However, depending on how the events line up at the portfolio level, the Co-stats for the treaty may vary considerably. The treaty may contribute a sizeable portion of loss to the event corresponding to the portfolio 100-year VaR, yet it may add nothing to many of the events close by in the list. Or the opposite may be true.

Exhibit 5 shows an example in which the portfolio 100-year VaR is $405 and the associated Co-VaR for the account is $20. But for the events on either side, the account loss is $0. Averaging over a band of nearby points (the set {6, 0, 20, 0, 4} in the example in Exhibit 5) will produce a more stable and meaningful answer. This will be called the **Co-Stat Band** approach. How large a neighborhood to include is a matter of judgment.

4.5.2. Co-TVaR Fails to be Subadditive

To calculate Co-TVaR for Treaty A when it is added to a given Reference Portfolio, the first step is to rank-order events by the sum of combined Reference Portfolio plus Treaty A losses. This is shown in Exhibit 6, Sheet 1 where the ordered total is displayed in the far right column. Then the contribution of Treaty A to the sum is posted in the column labeled Co-A. To obtain Co-TVaR of the 75th percentile, we average the Co-A amounts for the first 5 out of 20 of the events ordered on the sum[14] and arrive at Co-TVaR = (1+7+0+4+3)/5= 3.

It is known that TVaR is coherent and that has helped spark interest in using Co-TVaR as an

---

[13] The behavior of co-VaR is important because VaR is often used to define required capital at the portfolio level. Perhaps the most natural way to allocate VaR-based portfolio capital would be to use co-VaR.

[14] In cases where several events have the same total loss equal to the VaR of the target percentile, results from all orderings of those events should be averaged to get the resulting co-stat.

allocation metric. However, as we demonstrate by example in Exhibit 6, Co-TVaR is not coherent because it is not necessarily subadditive. This is a new result that has not been previously presented in the literature. Continuing with the example, we already know Co-TVaR of Treaty A is 3.0 and Sheet 2 and Sheet 3 of Exhibit 6 show the Co-TVaR of Treaty B is 3.0 and the Co-TVaR for the combination of the two, Treaty A+B, is 11.0. To be subadditive, the Co-TVaR for Treaty A+B would needs to be less than 6.0.

4.6 Percentile Allocation

Bodoff [6] has proposed the concept of percentile allocation. This starts with the observation that capital is needed to cover not just the large event losses, but also all losses up to and including those due to large events. For the first dollar of capital that is needed to cover losses, an allocation is done proportional to how often an account will use that first dollar relative to the other accounts. This entails looking at all events, seeing which ones tap the first dollar of capital, and then looking at account losses for each of those events. Event probabilities are then used to allocate the first dollar of capital. After the first dollar is allocated, the same procedure is followed to allocate the second dollar and so forth. This method is fairly stable and is not order dependent.

5. PRACTICAL APPLICATION ISSUES

Before going further, we need to address several practical issues that directly impact which methods can realistically be used to price CAT business.

5.1 Reference Portfolios

The first is that all the portfolio approaches are fundamentally impractical in their pure form. This is because reality differs from the one-by-one pricing paradigm implicit in portfolio approaches. Under this paradigm, the model is run, a quote is generated, and the account is bound in an instant during which no other changes take place to the portfolio. In the real world, time is not frozen. The portfolio changes as other accounts expire or are bound[15] in the time lag between quoting and binding a particular account. So the indicated price at the time of binding could be different than the indicated price at the time of quoting.

In principle, one should rerun and derive updated indications for all existing unbound quotes

[15] Or possibly those that are authorized but not yet bound. Since the authorized share of treaty may not be the final share, a more careful treatment would use an estimate of the expected final share.

each time the portfolio changes. But this is unworkable. It would slow down the process of providing quotes to underwriters and put a huge strain on models and analysts during peak renewal season. Further, even if the modeling could be done, the market will not accept quotes that change after they are made. Thus, in practice, it is likely very few companies attempt to use the actual up-to-date portfolio as the base for their pricing calculations. None that we are aware of revise existing open quotes. Rather most of them use a defined Reference Portfolio. This is fixed for a week or a month or a quarter. The possible conceptual advantage in using a portfolio-based pricing algorithm may be partially undercut if the Reference Portfolio differs materially from the real one.

In some shops, a renewal account is run against a Reference Portfolio that includes the expiring policy. While that is technically wrong and can lead to inaccurate results, the alternative entails setting up a custom Reference Portfolio for each account. That can become somewhat cumbersome.

5.2 Impracticality of Game Theoretic Averaging

In theory, averaging over all possible orders of accounts is the ideal solution to order dependence anomalies. However, in practice that is generally unworkable. While Mango found a closed form solution for a particular risk measure so that one did not need to actually run impact statistics for each ordering, such closed form solutions do not exist for many risk measures. The number of different orderings explodes in factorial fashion with the number of accounts so that performing all the brute-force computations is usually not feasible. This is one reason the order dependence issue is often ignored and Marginal statistics are computed against the Reference Portfolio with each account added on a last-in basis.

6. DEMONSTRATION

We will now demonstrate various pricing algorithms on hypothetical CAT loss data. We start with 50 events as shown in Exhibit 7. Each has a 2% annual probability of occurring. We also show the loss amounts for a Reference Portfolio and for two risks, denoted Treaty A and Treaty B. Both these risks have the same theoretical mean loss; however Treaty A losses are much more volatile. Total Annual Treaty A losses are theoretically independent of the Annual Reference Portfolio losses as Treaty A was constructed so that it has zero loss for events that generate losses for the Reference Portfolio. Under this construction, events with IDs from 1 to 25 have zero loss for Treaty A and non-zero loss for the Reference Portfolio. The situation is reversed for events with IDs from 26 to 50. Treaty B was constructed so that it has significant correlation with the Reference Portfolio because it has losses on many of the events that give rise to Reference Portfolio losses. It was also constructed to be theoretically independent of Treaty A. To summarize Treaty A is independent and volatile, while Treaty B is relatively well-behaved but correlated.

We then simulated 1,000 trial years and created a matrix with 1,000 rows and 50 columns to record which events if any occurred in each year. An excerpt of that matrix is shown in Exhibit 8. This sample size of 1,000 is too small to drive sampling error down to the fine decimals. This is true, even though with 2% annual probabilities each event should turn up in roughly 50 of the 1,000 trial years. We computed linear and rank correlations on the simulated data for each of our two risks. These are consistent with our construction. In the 1,000 trial years, Treaty A has a 1% linear correlation with the Reference Portfolio, while Treaty B has a 60% linear correlation. The corresponding sample rank correlations over these trials are -2% and 93% for treaties A and B respectively.

Indicated capital amounts and indicated premiums were then computed using Standalone and Incremental algorithms with the VaR, TVaR, and XTVaR risk measures. Required capital amounts and premium indications were also computed using Real Allocation based on the standalone amounts and the co-statistics of these risk measures. For Co-VaR, we used a banded approach for stability. Results are detailed in Exhibit 9 and a pricing comparison for the two risks is shown on Sheet 3 of Exhibit 9.

As might be expected, the incremental portfolio and co-statistic methods produce very modest risk loads for Treaty A. Indications using Co-XTVaR in particular have very little risk load. The standalone pricing for Treaty A is much higher. It is up to the reader to decide whether it makes sense to charge negligible risk load on this volatile account because it diversifies the portfolio or to go with substantial risk load because it is risky on a standalone basis. The situation is reversed for Treaty B. It has significant risk load under all the portfolio based methods, but standalone pricing is lower than it was for Treaty A.

7. GENERAL OBSERVATIONS

This chapter contains several key general observations that impact catastrophe pricing.

7.1 Sampling Error

In the background one should always be aware that simulation models are being used to generate the results and that simulation statistics are prone to sampling error. When adding a small account to a large initial portfolio, such sampling error may mask or distort the impact of adding the new account. The resulting relative error in pricing may be quite significant. As well, the pricing of relatively small changes in treaty layers may be prone to inaccuracy due to sampling error.

The event probabilities in simulation models are often quite small. Many events have return periods on the order of 10,000 years. To get even a roughly accurate estimate that has a representative number of such events, one needs to run the model with the number of trials set as a multiple of the return period of rare events in the models.

Beyond that, there are other pricing pitfalls stemming from sampling error. One of the most common is pricing a layer by taking differences of other layers that were priced using results from several different simulations. Statistical fluctuations could lead to an inconsistent result. All layers of interest need to be run in one set of simulation trials to achieve more accurate differentials in layer prices. Of course, sampling error can be reduced by running more simulation trials.[16] The analyst needs to balance the extra time and cost against the need for accuracy.

7.2 Tails

A key area of distinction concerns how much influence the tail has on the answer. Under some methods, capital allocation is determined solely by the tail. "Tail-Only" advocates implicitly or explicitly believe the tail should be regarded as the one true indicator of risk. Others argue that any event that could potentially consume capital needs to be considered; not just the tail events that would consume all the capital.

Beyond the philosophical divide, one should be aware that tail dependence tends to increase the volatility in the results. This is partly due to statistical sampling error. Further, as models change over time, the tails often move quite a bit more than the expected CAT losses.

For strict empiricists, an additional concern is that there is effectively no empirical way to validate the distribution of events in the extreme tail. What is the average size of portfolio losses that happen less frequently on average than once in every 250 years? It would take more than twenty samples of 1,000 simulated years to obtain a tentative estimate of this average. The simulation

[16] The author is advocating the analyst estimate beforehand how many simulated years are needed to achieve a desired level of accuracy in the estimated value of the risk measure. This is not the same as the flawed procedure of running sets of trials and stopping when results seem to stabilize. When the trial sets are too small to reliably capture the big but rare events, results from two or three sets of trials may be fairly stable. However, such "stable" results could underestimate the true tail of losses.

modeling software can do this without any problem, but our available empirical history is insufficient to validate the estimate.

This line of thinking also highlights the high-end cut-off problem. The size and frequency of mega-events included in a model may have an inordinate impact on any particular tail metric. Whether there are enough extreme events in a model is subject to some debate.[17] However, once very rare events are put into a model, it is not clear on what philosophical grounds we can exclude geologically significant asteroid strikes or comet impacts or events relating to ice ages. Many practicing analysts feel that discussion of such calamities is a vast digression and that such events should not be in the model as they could unduly impact tail metrics.

7.3 Diversification Benefit

If required capital for the portfolio is less than the sum of the standalone requirements, then there is effectively a diversification benefit to the portfolio as a whole. Many pricing methods would translate this into a reduction in the overall level of pricing of the whole portfolio as compared to the total pricing that would otherwise result if each account were priced on a standalone basis. For example, if the portfolio required 10% less capital than the sum of the standalone capital requirements, indicated standalone account pricing for each account could be given a 10% haircut to arrive at the final indication reflecting the portfolio diversification benefit.

Another sense of diversification benefit arises at the account level even if there is no overall benefit to the portfolio. Some methods effectively surcharge risks in peak zones, and provide discounts for those written in off-peak zones. It had been an accepted truism for many that CAT pricing indications should reward such geographic diversification. But the experience of recent years (2010-12) has led some to question the uncritical acceptance of diversification and warn of the dangers of "de-worsification".[18]

Our comparisons indicate that some algorithms lead to bargain-basement prices for exposures in zones where the company has little existing exposure. This highlights the issue: how much of a discount in the indicated price is justifiable in order to achieve the benefits of diversification?

---

[17] No vendor we are aware has said it included an event as large as the Japanese earthquake of March 2011 in its event set for events associated with faults in that region of Japan.

[18] Brodsky [7] quoted the noted CAT modeling expert Karen Clark as saying , "… some reinsurers may want diversification a little too much …" and "How much do you want to write of this underpriced risk just to get diversification?".

7.4 Treatment of Premiums, Commissions, and Loss-sensitive Features

As argued in Robbin and DeCouto [21], capital should be based on the distribution of Bounded Underwriting Loss, B, where

$$B = \text{Max}(0, \text{Loss} + \text{Commission} - \text{Premium}) \qquad \text{Equation (7.6.1)}$$

Notice there is a floor in the definition which prevents any negative result. In effect this prevents reduction in capital for scenarios where a profit is made.

In practice, capital is often based on just the loss and no adjustment is made for the premium or expense. The advantage is that this is straightforward to implement. An excess metric, such as XTVaR, captures the amount by which loss is above the mean and that partly incorporates the more complete approach. However, the reader is cautioned that the more sophisticated methodology is needed to model reinstatements and loss-sensitive features.

8. SUMMARY AND CONCLUSION

We started with the theory and basic equations of CAT pricing within the RORAC framework and saw that the crux of the issue comes down to how to set capital for an account. We looked at basic properties of risk measures, examined several specific ones, and demonstrated that some fail key theoretical properties. We have also systematically reviewed the Standalone, Marginal, and Real Allocation algorithms for computing CAT capital. We have shown the theoretical formulas, demonstrated how they work with accessible discrete examples, and illustrated why some need to be modified to be practical. We emphasized the "ranking" definitions as a clear way to define and compute the VaR and TVaR metrics. We also explained why TVaR on a discrete data set of CAT loss data is not the same as the Conditional Tail Expectation, and why a banded version of Co-VaR is necessary. We presented the new result that Co-TVaR is not subadditive.

Our work led to a concrete comparison of how various alternatives performed on two hypothetical accounts. A key goal of our presentation was to provide an understandable illustration of the process for generating the answers. We started with an event loss table showing account and portfolio losses for each event in a modest set of hypothetical CAT events. We then used this to generate simulated random results by year. With this generated data, we computed capital under several alternatives. Our examples showed that incremental or co-statistic tail based pricing could lead to pricing barely above expectation for a non-correlated, but risky, account.

If nothing else, we would urge any reader considering different CAT pricing methodologies to examine their properties and see how they work on simple examples. We hope this paper helps to increase understanding of the alternatives. We also hope it motivates readers to investigate, test, and achieve insights beyond those we have offered.

Exhibit 1
Sheet 1

Risk Measure Definitions -Discrete Example
Variance, Standard Dev, SemiVariance and SemiStnd Dev

Statistic	Value	Statistic	Value
Trials	20	Variance	88.4
Average	10.0	Standard Dev	9.4
		Semivariance	64.2
		SemiStnd Dev	8.0

Ordered Loss Data			
Rank	Loss	Variance Contribution	Semivariance Contribution
1	40.0	900	900
2	26.0	256	256
3	18.0	64	64
4	14.0	16	16
5	14.0	16	16
6	14.0	16	16
7	14.0	16	16
8	10.0	0	0
9	8.0	4	0
10	8.0	4	0
11	6.0	16	0
12	6.0	16	0
13	6.0	16	0
14	4.0	36	0
15	4.0	36	0
16	2.0	64	0
17	2.0	64	0
18	2.0	64	0
19	2.0	64	0
20	0.0	100	0

Exhibit 1
Sheet 2

Risk Measure Definitions -Discrete Example
VaR, TVaR and XTVaR

Statistic	Value	Statistic	Value
Trials	20	Rank for VaR	5.0
Average	10.0	VaR	14.0
Percentage	75.00%	TVaR	22.4
		XTVaR	12.4

Ordered Loss Data			
Rank	Loss	VaR Percentage	Conditional Tail Avg
1	40.0	95%	40.0
2	26.0	90%	33.0
3	18.0	85%	28.0
4	14.0	80%	24.5
5	14.0	75%	22.4
6	14.0	70%	21.0
7	14.0	65%	20.0
8	10.0	60%	18.8
9	8.0	55%	17.6
10	8.0	50%	16.6
11	6.0	45%	15.6
12	6.0	40%	14.8
13	6.0	35%	14.2
14	4.0	30%	13.4
15	4.0	25%	12.8
16	2.0	20%	12.1
17	2.0	15%	11.5
18	2.0	10%	11.0
19	2.0	5%	10.5
20	0.0	0%	10.0

Exhibit 1
Sheet 3

Risk Measure Definitions -Discrete Example
Transformed Mean and XSTransformed Mean

Statistic	Value	Statistic	Value
Trials	20	Wang Shift Parameter	0.674
Average	10.0	Transformed Mean	16.7
Percentage	75.00%	XS Transformed Mean	6.7

Ordered Loss Data				
Rank	Loss	Empirical CDF	Transformed CDF	Transformed Density
1	40.0	100.0%	100.0%	16.6%
2	26.0	95.0%	83.4%	10.6%
3	18.0	90.0%	72.8%	8.7%
4	14.0	85.0%	64.1%	7.5%
5	14.0	80.0%	56.6%	6.6%
6	14.0	75.0%	50.0%	6.0%
7	14.0	70.0%	44.0%	5.4%
8	10.0	65.0%	38.6%	4.9%
9	8.0	60.0%	33.7%	4.5%
10	8.0	55.0%	29.2%	4.2%
11	6.0	50.0%	25.0%	3.8%
12	6.0	45.0%	21.2%	3.5%
13	6.0	40.0%	17.7%	3.2%
14	4.0	35.0%	14.5%	2.9%
15	4.0	30.0%	11.5%	2.7%
16	2.0	25.0%	8.9%	2.4%
17	2.0	20.0%	6.5%	2.1%
18	2.0	15.0%	4.4%	1.8%
19	2.0	10.0%	2.5%	1.5%
20	0.0	5.0%	1.0%	1.0%

Exhibit 2

Failure of VaR Subadditivity

Statistic	Value		Mean	VaR
Trials	20	Risk A Standalone	2.50	4.00
Percentage	75.00%	Reference Portfolio	25.00	34.00
Rank	5	Sum	27.50	38.00
		Combined Portfolio	27.50	39.00
		Consolidation Benefit	0.00	-1.00
		Marginal VaR for A		5.00

Loss Data by Trial				**Ordered Loss Data**			
Trial	A	Ref	A+Ref	Rank	A	Ref	A+Ref
1	0.00	12.00	12.00	1	8.00	37.00	41.00
2	0.00	37.00	37.00	2	8.00	36.00	40.00
3	4.00	36.00	40.00	3	7.00	35.00	40.00
4	0.00	35.00	35.00	4	6.00	34.00	40.00
5	6.00	34.00	40.00	5	4.00	34.00	39.00
6	2.00	17.00	19.00	6	4.00	32.00	37.00
7	1.00	16.00	17.00	7	4.00	31.00	35.00
8	8.00	32.00	40.00	8	3.00	30.00	34.00
9	0.00	27.00	27.00	9	2.00	27.00	30.00
10	0.00	14.00	14.00	10	2.00	27.00	27.00
11	3.00	27.00	30.00	11	1.00	26.00	26.00
12	4.00	15.00	19.00	12	1.00	23.00	24.00
13	0.00	20.00	20.00	13	0.00	20.00	20.00
14	4.00	30.00	34.00	14	0.00	18.00	20.00
15	8.00	31.00	39.00	15	0.00	17.00	19.00
16	2.00	18.00	20.00	16	0.00	16.00	19.00
17	1.00	23.00	24.00	17	0.00	16.00	17.00
18	0.00	26.00	26.00	18	0.00	15.00	16.00
19	7.00	34.00	41.00	19	0.00	14.00	14.00
20	0.00	16.00	16.00	20	0.00	12.00	12.00

Exhibit 3
Sheet 1

Failure of Marginal VaR Scalability - Risk A + Portfolio

Statistic	Value		Mean	VaR
Trials	20	Risk A Standalone	2.50	4.00
Percentage	75.0%	Reference Portfolio	25.00	34.00
Rank	5	Sum	27.50	38.00
		Combined Portfolio	27.50	37.00
		Consolidation Benefit	0.00	1.00
		Marginal VaR for A		3.00

Loss Data by Trial				**Ordered Loss Data**			
Trial	A	Ref	A+Ref	Rank	A	Ref	A+Ref
1	4.00	12.00	16.00	1	8.00	37.00	41.00
2	0.00	37.00	37.00	2	8.00	36.00	40.00
3	0.00	36.00	36.00	3	7.00	35.00	40.00
4	0.00	35.00	35.00	4	6.00	34.00	39.00
5	6.00	34.00	40.00	5	4.00	34.00	37.00
6	2.00	17.00	19.00	6	4.00	32.00	36.00
7	1.00	16.00	17.00	7	4.00	31.00	35.00
8	8.00	32.00	40.00	8	3.00	30.00	34.00
9	0.00	27.00	27.00	9	2.00	27.00	30.00
10	0.00	14.00	14.00	10	2.00	27.00	27.00
11	3.00	27.00	30.00	11	1.00	26.00	26.00
12	4.00	15.00	19.00	12	1.00	23.00	24.00
13	0.00	20.00	20.00	13	0.00	20.00	20.00
14	4.00	30.00	34.00	14	0.00	18.00	20.00
15	8.00	31.00	39.00	15	0.00	17.00	19.00
16	2.00	18.00	20.00	16	0.00	16.00	19.00
17	1.00	23.00	24.00	17	0.00	16.00	17.00
18	0.00	26.00	26.00	18	0.00	15.00	16.00
19	7.00	34.00	41.00	19	0.00	14.00	16.00
20	0.00	16.00	16.00	20	0.00	12.00	14.00

Exhibit 3
Sheet 2

Failure of Marginal VaR Scalability - Risk 2*A + Portfolio

Statistic	Value		Mean	VaR
Trials	20	Risk 2A Standalone	5.00	8.00
Percentage	75.0%	Reference Portfolio	25.00	34.00
Rank	5	Sum	30.00	42.00
		Combined Portfolio	30.00	38.00
		Consolidation Benefit	0.00	4.00
		Marginal VaR for 2A		4.00

Loss Data by Trial				**Ordered Loss Data**			
Trial	2A	Ref	2A+Ref	Rank	2A	Ref	2A+Ref
1	8.00	12.00	20.00	1	16.00	37.00	48.00
2	0.00	37.00	37.00	2	16.00	36.00	48.00
3	0.00	36.00	36.00	3	14.00	35.00	47.00
4	0.00	35.00	35.00	4	12.00	34.00	46.00
5	12.00	34.00	46.00	5	8.00	34.00	38.00
6	4.00	17.00	21.00	6	8.00	32.00	37.00
7	2.00	16.00	18.00	7	8.00	31.00	36.00
8	16.00	32.00	48.00	8	6.00	30.00	35.00
9	0.00	27.00	27.00	9	4.00	27.00	33.00
10	0.00	14.00	14.00	10	4.00	27.00	27.00
11	6.00	27.00	33.00	11	2.00	26.00	26.00
12	8.00	15.00	23.00	12	2.00	23.00	25.00
13	0.00	20.00	20.00	13	0.00	20.00	23.00
14	8.00	30.00	38.00	14	0.00	18.00	22.00
15	16.00	31.00	47.00	15	0.00	17.00	21.00
16	4.00	18.00	22.00	16	0.00	16.00	20.00
17	2.00	23.00	25.00	17	0.00	16.00	20.00
18	0.00	26.00	26.00	18	0.00	15.00	18.00
19	14.00	34.00	48.00	19	0.00	14.00	16.00
20	0.00	16.00	16.00	20	0.00	12.00	14.00

Exhibit 4

Conditional Tail Expectation is Not Monotonic

Statistic	Value		Mean	VaR	TVaR	CTE
Trials	20	Risk A Standalone	2.50	4.00	6.60	7.25
Percentage	75.0%	Reference Portfolio	25.00	34.00	35.20	36.00
Rank	5	Sum	27.50	38.00	41.80	43.25
		Combined Portfolio	27.50	34.00	35.40	35.75
		Consolidation Benefit	0.00	4.00	6.40	7.50
		Marginal Measure		0.00	0.20	-0.25

Loss Data by Trial				**Ordered Loss Data**			
Trial	A	Ref	A+Ref	Rank	A	Ref	A+Ref
1	8.00	12.00	20.00	1	8.00	37.00	37.00
2	0.00	37.00	37.00	2	8.00	36.00	36.00
3	0.00	36.00	36.00	3	7.00	35.00	35.00
4	0.00	35.00	35.00	4	6.00	34.00	35.00
5	1.00	34.00	35.00	5	4.00	34.00	34.00
6	2.00	17.00	19.00	6	4.00	32.00	34.00
7	7.00	16.00	23.00	7	4.00	31.00	34.00
8	0.00	32.00	32.00	8	3.00	30.00	33.00
9	4.00	27.00	31.00	9	2.00	27.00	32.00
10	4.00	14.00	18.00	10	2.00	27.00	31.00
11	6.00	27.00	33.00	11	1.00	26.00	26.00
12	8.00	15.00	23.00	12	1.00	23.00	24.00
13	0.00	20.00	20.00	13	0.00	20.00	23.00
14	4.00	30.00	34.00	14	0.00	18.00	23.00
15	3.00	31.00	34.00	15	0.00	17.00	20.00
16	2.00	18.00	20.00	16	0.00	16.00	20.00
17	1.00	23.00	24.00	17	0.00	16.00	20.00
18	0.00	26.00	26.00	18	0.00	15.00	19.00
19	0.00	34.00	34.00	19	0.00	14.00	18.00
20	0.00	16.00	16.00	20	0.00	12.00	16.00

Exhibit 5

Instability of Co-VaR

Rank	VaR Percentage	Portfolio Loss	Risk A Loss
1			
98	99.02%	$422	$6
99	99.01%	$408	$0
100	99.00%	$405	$20
101	98.99%	$395	$0
102	98.98%	$390	$4
10,000			

Exhibit 6
Sheet 1

Co-TVaR Subadditivity Failure
Risk A Co-TVaR Calculation

Stat	Value
Trials	20
Pct	75.0%
Rank	5

Results	A	Ref	A+Ref
Mean	2.50	25.00	27.50
VaR	4.00	33.00	36.00
TVaR	6.60	35.80	38.00
Co-TVaR	3.00	35.00	38.00

Loss Data by Trial				Separately Ordered			Co-Stats			
Trial	A	Ref	A+Ref	Rank	A	Ref	Trial	Co- A	Co-Ref	A+Ref
1	2.00	8.00	10.00	1	8.00	39.00	7	1.00	39.00	40.00
2	0.00	38.00	38.00	2	8.00	38.00	3	7.00	32.00	39.00
3	7.00	32.00	39.00	3	7.00	35.00	2	0.00	38.00	38.00
4	0.00	35.00	35.00	4	6.00	34.00	14	4.00	33.00	37.00
5	2.00	14.00	16.00	5	4.00	33.00	6	3.00	33.00	36.00
6	3.00	33.00	36.00	6	3.00	33.00	4	0.00	35.00	35.00
7	1.00	39.00	40.00	7	3.00	32.00	19	0.00	34.00	34.00
8	2.00	16.00	18.00	8	2.00	30.00	15	3.00	30.00	33.00
9	8.00	25.00	33.00	9	2.00	27.00	11	6.00	27.00	33.00
10	0.00	11.00	11.00	10	2.00	26.00	9	8.00	25.00	33.00
11	6.00	27.00	33.00	11	2.00	25.00	12	8.00	22.00	30.00
12	8.00	22.00	30.00	12	1.00	23.00	18	0.00	26.00	26.00
13	0.00	20.00	20.00	13	1.00	22.00	17	1.00	23.00	24.00
14	4.00	33.00	37.00	14	1.00	20.00	16	2.00	18.00	20.00
15	3.00	30.00	33.00	15	0.00	18.00	13	0.00	20.00	20.00
16	2.00	18.00	20.00	16	0.00	16.00	8	2.00	16.00	18.00
17	1.00	23.00	24.00	17	0.00	16.00	20	1.00	16.00	17.00
18	0.00	26.00	26.00	18	0.00	14.00	5	2.00	14.00	16.00
19	0.00	34.00	34.00	19	0.00	11.00	10	0.00	11.00	11.00
20	1.00	16.00	17.00	20	0.00	8.00	1	2.00	8.00	10.00

Exhibit 6
Sheet 2

Co-TVaR Subadditivity Failure

Risk B Co-TVaR Calculation

Stat	Value
Trials	20
Pct	75.0%
Rank	5

Results	B	Ref	B+Ref
Mean	2.50	25.00	27.50
VaR	4.00	33.00	36.00
TVaR	6.40	35.80	38.40
Co-TVaR	3.00	35.40	38.40

Loss Data by Trial				Separately Ordered			Co-Stats			
Trial	B	Ref	B+Ref	Rank	B	Ref	Trial	Co- B	Co-Ref	B+Ref
1	0.00	8.00	8.00	1	9.00	39.00	7	5.00	39.00	44.00
2	0.00	38.00	38.00	2	7.00	38.00	2	0.00	38.00	38.00
3	4.00	32.00	36.00	3	7.00	35.00	14	4.00	33.00	37.00
4	2.00	35.00	37.00	4	5.00	34.00	4	2.00	35.00	37.00
5	2.00	14.00	16.00	5	4.00	33.00	3	4.00	32.00	36.00
6	1.00	33.00	34.00	6	4.00	33.00	19	1.00	34.00	35.00
7	5.00	39.00	44.00	7	3.00	32.00	11	7.00	27.00	34.00
8	1.00	16.00	17.00	8	2.00	30.00	9	9.00	25.00	34.00
9	9.00	25.00	34.00	9	2.00	27.00	6	1.00	33.00	34.00
10	0.00	11.00	11.00	10	2.00	26.00	15	0.00	30.00	30.00
11	7.00	27.00	34.00	11	1.00	25.00	12	7.00	22.00	29.00
12	7.00	22.00	29.00	12	1.00	23.00	18	2.00	26.00	28.00
13	1.00	20.00	21.00	13	1.00	22.00	17	1.00	23.00	24.00
14	4.00	33.00	37.00	14	1.00	20.00	16	3.00	18.00	21.00
15	0.00	30.00	30.00	15	1.00	18.00	13	1.00	20.00	21.00
16	3.00	18.00	21.00	16	0.00	16.00	8	1.00	16.00	17.00
17	1.00	23.00	24.00	17	0.00	16.00	20	0.00	16.00	16.00
18	2.00	26.00	28.00	18	0.00	14.00	5	2.00	14.00	16.00
19	1.00	34.00	35.00	19	0.00	11.00	10	0.00	11.00	11.00
20	0.00	16.00	16.00	20	0.00	8.00	1	0.00	8.00	8.00

Exhibit 6
Sheet 3

Co-TVaR Subadditivity Failure

Combined Risk A + Risk B Co-TVaR Calculation

Stat	Value
Trials	20
Pct	75.0%
Rank	5

Results	A+B	Ref	A+B +Ref
Mean	5.00	25.00	30.00
VaR	8.00	33.00	40.00
TVaR	12.80	35.80	42.20
Co-TVaR	11.00	31.20	42.20

Loss Data by Trial				**Separately Ordered**			**Co-Stats**			
Trial	A+B	Ref	A+B +Ref	Rank	A+B	Ref	Trial	Co-A+B	Co-Ref	A+B +Ref
1	2.00	8.00	10.00	1	17.00	39.00	7	6.00	39.00	45.00
2	0.00	38.00	38.00	2	15.00	38.00	3	11.00	32.00	43.00
3	11.00	32.00	43.00	3	13.00	35.00	9	17.00	25.00	42.00
4	2.00	35.00	37.00	4	11.00	34.00	14	8.00	33.00	41.00
5	4.00	14.00	18.00	5	8.00	33.00	11	13.00	27.00	40.00
6	4.00	33.00	37.00	6	6.00	33.00	2	0.00	38.00	38.00
7	6.00	39.00	45.00	7	5.00	32.00	12	15.00	22.00	37.00
8	3.00	16.00	19.00	8	4.00	30.00	6	4.00	33.00	37.00
9	17.00	25.00	42.00	9	4.00	27.00	4	2.00	35.00	37.00
10	0.00	11.00	11.00	10	3.00	26.00	19	1.00	34.00	35.00
11	13.00	27.00	40.00	11	3.00	25.00	15	3.00	30.00	33.00
12	15.00	22.00	37.00	12	2.00	23.00	18	2.00	26.00	28.00
13	1.00	20.00	21.00	13	2.00	22.00	17	2.00	23.00	25.00
14	8.00	33.00	41.00	14	2.00	20.00	16	5.00	18.00	23.00
15	3.00	30.00	33.00	15	2.00	18.00	13	1.00	20.00	21.00
16	5.00	18.00	23.00	16	1.00	16.00	8	3.00	16.00	19.00
17	2.00	23.00	25.00	17	1.00	16.00	5	4.00	14.00	18.00
18	2.00	26.00	28.00	18	1.00	14.00	20	1.00	16.00	17.00
19	1.00	34.00	35.00	19	0.00	11.00	10	0.00	11.00	11.00
20	1.00	16.00	17.00	20	0.00	8.00	1	2.00	8.00	10.00

Exhibit 6
Sheet 4

Co-TVaR Subadditivity Failure Summary

Statistic	Value
Trials	20
Pct	75%
Rank	5

Results	A	B	A+B	Ref	A+B +Ref
Mean	2.50	2.50	5.00	25.00	30.00
VaR	4.00	4.00	8.00	33.00	40.00
TVaR	6.60	6.40	12.80	35.80	42.20
Co-TVaR	3.00	3.00	11.00	31.20	42.20

Loss Data by Trial

Trial	A	B	A+ B	Ref	A+ B +Ref
1	2.00	0.00	2.00	8.00	10.00
2	0.00	0.00	0.00	38.00	38.00
3	7.00	4.00	11.00	32.00	43.00
4	0.00	2.00	2.00	35.00	37.00
5	2.00	2.00	4.00	14.00	18.00
6	3.00	1.00	4.00	33.00	37.00
7	1.00	5.00	6.00	39.00	45.00
8	2.00	1.00	3.00	16.00	19.00
9	8.00	9.00	17.00	25.00	42.00
10	0.00	0.00	0.00	11.00	11.00
11	6.00	7.00	13.00	27.00	40.00
12	8.00	7.00	15.00	22.00	37.00
13	0.00	1.00	1.00	20.00	21.00
14	4.00	4.00	8.00	33.00	41.00
15	3.00	0.00	3.00	30.00	33.00
16	2.00	3.00	5.00	18.00	23.00
17	1.00	1.00	2.00	23.00	25.00
18	0.00	2.00	2.00	26.00	28.00
19	0.00	1.00	1.00	34.00	35.00
20	1.00	0.00	1.00	16.00	17.00

Exhibit 7
Sheet 1

Event Loss table

Event ID	Annual Probability	Treaty A Loss	Treaty B Loss	Reference Portfolio Loss
1	2%	0	500	125,000
2	2%	0	1,000	100,000
3	2%	0	1,000	90,000
4	2%	0	2,000	80,000
5	2%	0	2,500	75,000
6	2%	0	3,000	70,000
7	2%	0	3,000	60,000
8	2%	0	3,000	50,000
9	2%	0	3,000	50,000
10	2%	0	2,500	40,000
11	2%	0	2,000	38,000
12	2%	0	1,000	36,000
13	2%	0	1,000	34,000
14	2%	0	1,000	32,000
15	2%	0	1,000	30,000
16	2%	0	2,000	25,000
17	2%	0	2,500	20,000
18	2%	0	3,000	15,000
19	2%	0	3,000	10,000
20	2%	0	3,000	5,000
21	2%	0	3,000	5,000
22	2%	0	2,500	4,000
23	2%	0	2,000	3,000
24	2%	0	1,000	2,000
25	2%	0	500	1,000

Exhibit 7
Sheet 2

Event Loss table

Event ID	Annual Probability	Treaty A Loss	Treaty B Loss	Reference Portfolio Loss
26	2%	12,500	0	0
27	2%	10,000	0	0
28	2%	7,500	0	0
29	2%	5,000	0	0
30	2%	4,000	0	0
31	2%	3,000	0	0
32	2%	2,000	0	0
33	2%	1,000	0	0
34	2%	900	0	0
35	2%	800	0	0
36	2%	700	0	0
37	2%	600	0	0
38	2%	500	0	0
39	2%	400	0	0
40	2%	300	0	0
41	2%	200	0	0
42	2%	100	0	0
43	2%	90	0	0
44	2%	80	0	0
45	2%	70	0	0
46	2%	60	0	0
47	2%	50	0	0
48	2%	50	0	0
49	2%	50	0	0
50	2%	50	0	0

Exhibit 8

Simulation of 1,000 Trial Years for 50 events

Event ID	1	2	3	4	5
Treaty A Loss	0	0	0	0	0
Treaty B Loss	500	1,000	1,000	2,000	2,500
Treaty C Loss	500	0	1,000	0	2,500
Reference Portfolio Loss	125,000	100,000	90,000	80,000	75,000
Ref + A + B	125,500	101,000	91,000	82,000	77,500
Simulated Annual Prob	1.80%	1.60%	1.90%	2.60%	2.40%
Prob Occ	2.00%	2.00%	2.00%	2.00%	2.00%

Event Indicator Matrix

Trial Year	**Number of events**	**Event ID 1**	**Event ID 2**	**Event ID 3**	**Event ID 4**	**Event ID 5**
1	0					
2	1					
3	0					
4	2					
5	2					
6	2					
7	1					
8	1		1			
9	1					
10	1			1		
11	1			1		
12	1					
13	1					
14	1					
15	0					
16	2					
17	0					
18	0					
19	0					
20	0					

Exhibit 9
Sheet 1

Capital and Pricing Indications
Treaty A

Percentage	95.0%
Target Return	15.0%

	E[L]	Linear Correlation	Rank Correlation
Reference Portfolio	19,651		
Treaty A	1,060	1%	-2%

Capital Calculation Method	Treaty A Standalone Capital	Reference Portfolio Capital	Reference + Treaty A Capital	Treaty A Required Capital	Treaty A Indicated Premium	Treaty A Indicated Risk Load	Risk Load % of Premium
Standalone VaR	7,590	N/A	N/A	7,590	2,198	1,139	52%
Standalone TVaR	11,255	N/A	N/A	11,255	2,748	1,688	61%
Standalone XTVaR	10,195	N/A	N/A	10,195	2,589	1,529	59%
VaR Increment	N/A	100,000	102,000	2,000	1,360	300	22%
TVaR Increment	N/A	130,800	132,206	1,406	1,271	211	17%
XTVaR Increment	N/A	111,149	111,495	346	1,112	52	5%
Allocation via Standalone VaR	7,590	100,000	102,000	7,196	2,139	1,079	50%
Allocation via Standalone TVaR	11,255	100,000	102,000	10,318	2,608	1,548	59%
Allocation via Standalone XTVaR	10,195	100,000	102,000	9,437	2,475	1,416	57%
Band Co-VaR Allocation	N/A	N/A	102,000	3,237	1,545	486	31%
Co-TVaR Allocation	N/A	N/A	102,000	1,239	1,246	186	15%
Co-XTVaR Allocation	N/A	N/A	102,000	1,519	1,288	228	18%

Exhibit 9
Sheet 2

Capital and Pricing Indications
Treaty B

Percentage	95.0%
Target Return	15.0%

	E[L]	Linear Correlation	Rank Correlation
Reference Portfolio	19,651		
Treaty B	1,054	60%	93%

Capital Calculation Method	Treaty B Standalone Capital	Reference Portfolio Capital	Reference + Treaty B Capital	Treaty B Required Capital	Treaty B Indicated Premium	Treaty B Indicated Risk Load	Risk Load % of Premium
Standalone VaR	4,000	N/A	N/A	4,000	1,654	600	36%
Standalone TVaR	5,480	N/A	N/A	5,480	1,876	822	44%
Standalone XTVaR	4,426	N/A	N/A	4,426	1,718	664	39%
VaR Increment	N/A	100,000	104,000	4,000	1,654	600	36%
TVaR Increment	N/A	130,800	134,010	3,210	1,536	482	31%
XTVaR Increment	N/A	111,149	113,305	2,156	1,377	323	23%
Allocation via Standalone VaR	4,000	100,000	104,000	4,000	1,654	600	36%
Allocation via Standalone TVaR	5,480	100,000	104,000	5,403	1,864	810	43%
Allocation via Standalone XTVaR	4,426	100,000	104,000	4,408	1,715	661	39%
Band Co-VaR Allocation	N/A	N/A	104,000	2,588	1,442	388	27%
Co-TVaR Allocation	N/A	N/A	104,000	2,491	1,428	374	26%
Co-XTVaR Allocation	N/A	N/A	104,000	2,571	1,440	386	27%

Exhibit 9
Sheet 3

Capital and Pricing Indications Summary
Treaty A vs Treaty B Comparison

Percentage	95.0%
Target Return	15.0%

	Simulated Value E[L]	Linear Correlation	Rank Correlation
Reference Portfolio	19,651		
Treaty A	1,060	1%	-2%
Treaty B	1,054	60%	93%

Capital Calculation Method	Treaty A Required Capital	Treaty A Indicated Premium	Risk Load % of Premium	Treaty B Required Capital	Treaty B Indicated Premium	Risk Load % of Premium
Standalone VaR	7,590	2,198	52%	4,000	1,654	36%
Standalone TVaR	11,255	2,748	61%	5,480	1,876	44%
Standalone XTVaR	10,195	2,589	59%	4,426	1,718	39%
VaR Increment	2,000	1,360	22%	4,000	1,654	36%
TVaR Increment	1,406	1,271	17%	3,210	1,536	31%
XTVaR Increment	346	1,112	5%	2,156	1,377	23%
Standalone VaR Allocation	7,196	2,139	50%	4,000	1,654	36%
Standalone TVaR Allocation	10,318	2,608	59%	5,403	1,864	43%
Standalone XTVaR Allocation	9,437	2,475	57%	4,408	1,715	39%
Band Co-VaR Allocation	3,237	1,545	31%	2,588	1,442	27%
Co-TVaR Allocation	1,239	1,246	15%	2,491	1,428	26%
Co-XTVar Allocation	1,519	1,288	18%	2,571	1,440	27%

BIBLIOGRAPHY

[1] C. Acerbi and D. Tasche, "Expected Shortfall: a natural coherent alternative to Value at Risk", 2001.

[2] D. Appell, "Quantifying Risk Load for Property Catastrophe Exposure", Presentation *Casualty Actuarial Society Ratemaking Seminar,* 2010.

[3] P. Artzner, "Application of Coherent Risk Measures to Capital Requirements in Insurance", *North American Actuarial Journal*, Volume 3 Number 2, p 11-25.

[4] P. Artzner, F. Delbaen, J. Eber, and D. Heath, "Coherent Measures of Risk," *Math. Finance 9* (1999), no. 3, p 203-229.

[5] M. Barges, H. Cossette, and E. Marceau, "TVaR-based capital allocation with copulas", *Insurance: Mathematic and Economics*, Vol 45 Issue 3 December 2009, p 348-361.

[6] N. Bodoff, "Capital Allocation by Percentile Layer," *Variance*, 2009, Volume 3/ Issue 1, p 13-30.

[7] M. Brodsky, "Diversification or Deworsification?", *Risk and Insurance*, June 1, 2011.

[8] R. Butsic, "Allocating the Cost of Capital", *Casualty Actuarial Society Forum,* 2002 Fall.

[9] S. D'Arcy, "Capital Allocation in the Property-Liability Insurance Industry", *Variance*, 2011, Volume 5/Issue 2, p 141-157.

[10] M. Denault, "Coherent Allocation of Risk Capital", *The Journal of Risk*, Volume 4, Number 1, Fall 2001.

[11] L. Fu and C.K. Khury, "Optimal Layers for Catastrophe Reinsurance", *Variance,* 2010, Volume 4/Issue 2, p 193-208.

[12] R. Goldfarb, "Risk Adjusted Performance Measures for P&C Insurers", *Casualty Actuarial Society Study Note,* 2010.

[13] R. Kreps, "Reinsurer Risk Loads from Marginal Surplus Requirements", *Proceedings of the Casualty Actuarial Society* Vol LXXVII, 1990, p. 196-203.

[14] R. Kreps, "Riskiness Leverage Models", *Proceedings of the Casualty Actuarial Society*, Vol XCII, 2005, p. 31-60.

[15] D. Mango, "An Application of Game Theory: Property Catastrophe Risk Load", *Proceedings of the Casualty Actuarial Society* Vol LXXXV, 1998, 157-186.

[16] D. Mango, "Insurance Capital as a Shared Asset", *Casualty Actuarial Society Forum*, 2006 Fall, p. 573-576.

[17] D. Mango and J. Sandor, "Dependence Models and the Portfolio Effect", *Casualty Actuarial Society Forum*, 2002 Winter, p 57-72.

[18] G. Meyers, "Setting Capital Requirements with Coherent Measures of Risk –Part 1 and Part 2". *Actuarial Review*, August and November, 2002.

[19] G. Meyers, F. Klinker and G. Lalonde, "The Aggregation and Correlation of Reinsurance Exposure," *Casualty Actuarial Society Forum*, 2003.

[20] H. Panjer and J. Jing, "Solvency and Capital Allocation", *International Symposium on Insurance Regulatory Environment and Actuarial Practice in China*, October 2001.

[21] I. Robbin and J. DeCouto, "Coherent Capital for Treaty ROE Calculations", *Casualty Actuarial Society Forum*, Spring 2005.

[22] R. Rockafellar and S. Uryasev, "Conditional value-at-risk for general loss distributions", *Journal of Banking and Finance* 26 (2002), p 1443-1471.

[23] G.Venter, "A Survey of Capital Allocation Methods with Commentary," ASTIN Colloquia, 2003.

[24] S. Wang, "A Risk Measure That Goes Beyond Coherence", 2002.

[25] S. Wang, V. Young, and H. H. Panjer, "Axiomatic Characterization of Insurance Prices", *Insurance Mathematics and Economics,* 21 (1997), pp 173-182.

[26] S. Wang, "Implementation of Proportional Hazards Transforms in Ratemaking", *Proceedings of the Casualty Actuarial Society,* Vol LXXXV, 1998, p. 940-979.

[27] J.Wirch and M. Hardy, "Distortion Risk Measures: Coherence and Stochastic Dominance", *International Congress on Insurance: Mathematics and Economics*, July 15-17, 2002.

[28] N. Zhao, "Modeling Dependence in Catastrophe Risk Portfolio and Optimizing Capital Loading", *ETH Thesis*, 2010.

Abbreviations and notations

CAT, Catastrophe
cdf, Cumulative Distribution Function
Co-Statistic or Co-Stat, Contribution Statistic
CTE, Conditional Tail Expectation
PML, Probable Maximum Loss
RORAC, Return on Risk-Adjusted Capital
TVaR, Tail Value at Risk
VaR, Value at Risk
XTVaR, Excess Tail Value at Risk

Biography of the Author

Ira Robbin is currently at AIG after having previously held positions at P&C Actuarial Analysts, Endurance, Partner Re, CIGNA PC, and INA working in several corporate, pricing, research, and consulting roles. He has written papers and made presentations on many topics including risk load, capital requirements, ROE, Coherent Capital, price monitors, and One-year Reserve Risk. He has a PhD in Math from Rutgers University and a Bachelor degree in Math from Michigan State University.

Disclaimers

Opinions expressed in this paper are solely those of the author. They may or may not be consistent with the views of the authors' current or prior clients or employers. No warranty is given that any formula or assertion is accurate. No liability is assumed whatsoever for any losses, direct or indirect, that may result from use of the methods described in this paper or reliance on any of the views expressed therein.

Acknowledgments

The author gratefully acknowledges the contributions of Joshua London, David Clark, and Neil Bodoff. Their insightful commentary, illustrative spreadsheets, and constructive reviews were critical in helping the author draft and improve this paper. The author also wishes to thank the review committee members for their thoughtful comments and corrections.

Loss Cost Components and Industrial Structure

Frank Schmid

Abstract
Motivation. Concomitant with the 2007-2009 recession, the U.S. economy experienced profound changes in industrial structure that led to widely varying growth rates of employment by industry. Whereas some of these changes may be temporary, others are likely to be permanent or keep progressing. Among the sectors exhibiting the most significant shifts in employment levels are manufacturing, health care, and construction. Quantifying the impact of the changes to the industrial structure on loss cost components is important for understanding trends in NCCI aggregate ratemaking.
Method. A statistical model is employed to measure the impact of the rates of employment growth by industry and the change in the rate of private nonfarm employment growth on the growth rates of frequency and the indemnity and medical severities. Frequency is defined as the ratio of the lost-time claim count (developed to ultimate) to on-leveled and wage-adjusted premium. Severity is defined as the ratio of (on-leveled, developed-to-ultimate, and wage-adjusted) losses to the number of lost-time claims (developed to ultimate). In a sensitivity analysis, ridge regression is applied to control for possible multicollinearity.
Results. The industries whose changes in the rates of employment growth are most pertinent to variations in the rate of frequency growth are health care and construction (but not manufacturing). The industrial structure is of little import to the growth rate of indemnity severity, but of consequence for the growth rate of medical severity. Further, the evidence regarding the effect on the growth rate of frequency of changes in the rate of employment growth agrees with the findings on the impact of job flows previously documented in Schmid [6].
Availability. The model was implemented in R (http://cran.r-project.org/) using the sampling platform JAGS (Just Another Gibbs Sampler, http://www-ice.iarc.fr/~martyn/software/jags/). JAGS was linked to R by means of the R package rjags (http://cran.r-project.org/web/packages/rjags/index.html).

Keywords. Aggregate Ratemaking, Industrial Structure, Loss Cost Components, Ridge Regression

1. INTRODUCTION

Concomitant with the 2007-2009 economic recession, the U.S. economy experienced profound changes in industrial structure that led to widely varying growth rates of employment by industry. Whereas some of these changes may be temporary, others are likely to be permanent or keep progressing. Among the sectors exhibiting the most significant shifts in employment levels are manufacturing, health care, and construction. Quantifying the impact of the changes to the industrial structure on loss cost components is important for understanding trends in NCCI aggregate ratemaking.

The impact of changes in the industrial structure on loss cost components is quantified at the state level using a Bayesian statistical model. The model relates the logarithmic rates of frequency growth on the logarithmic rates of employment growth by industry and on the first difference in the logarithmic rate of total private nonfarm employment growth. This analysis is then repeated for the rates of indemnity and medical severity growth. In a sensitivity analysis, ridge regression is applied to these three loss cost components to control for possible multicollinearity among the covariates.

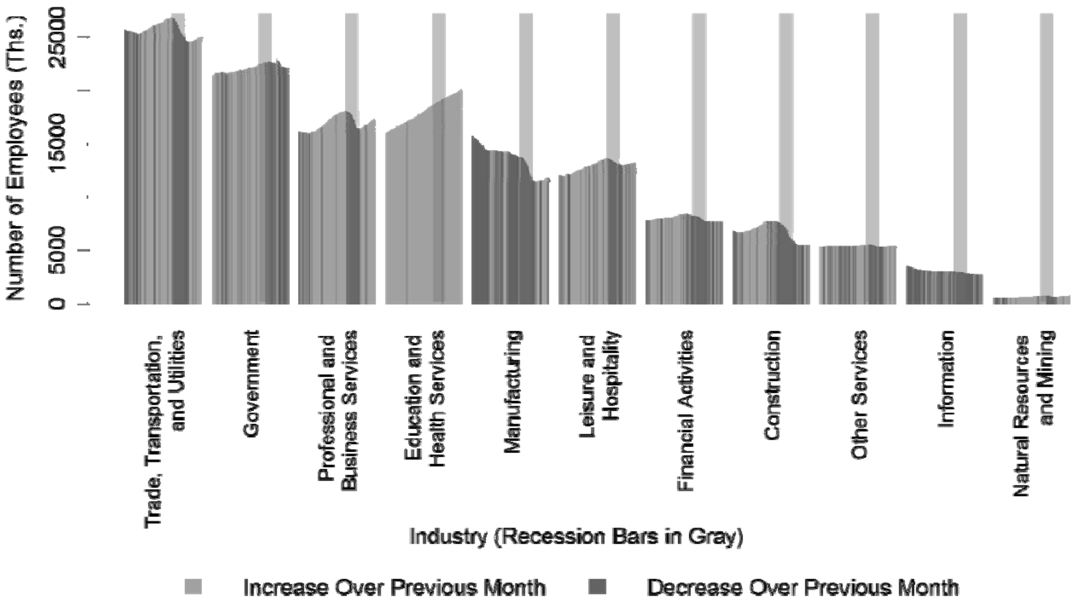
In keeping with NCCI trend selection methodology in aggregate ratemaking, frequency is defined as the ratio of the lost-time claim count (developed to ultimate) to on-leveled and wage-adjusted premium. Severity is defined as the ratio of (on-leveled, developed-to-ultimate, and wage-adjusted) losses to the number of lost-time claims (developed to ultimate). With frequency and severity defined in this way, the loss ratio equals the product of frequency and the pertinent severity. Because frequency and the severities share the claim count as an influence, albeit in opposite ways (as claim count is the numerator of frequency and the denominator of the severities), some of the economic forces bearing on frequency have the opposite effect on the severities. Although these opposing effects do not necessarily offset each other entirely, economic conditions oftentimes affect the loss ratio much less than they affect frequency and the pertinent severity.

The aggregate ratemaking data employed in the analysis are on a paid basis; state funds are included where applicable. There are 37 states in the data set, 35 of which are on policy years, the remaining two are on accident years. The model makes use of employment forecasts (by industry and private nonfarm) from a professional forecasting firm in generating forecasts for the loss cost components. These forecasts for the loss cost components extend through Policy (or, where applicable, Accident) Year 2015.

The model shows that construction, among all industries, exerts the most influence on loss cost components during the studied time period. This influence is most manifest in the time window of the recession (which lasted from December 2007 to June 2009) and its immediate aftermath (Policy Year 2009 and Accident Year 2010). The four states that experienced the steepest decline in construction employment in the United States between 2007 and 2009 (based on annual numbers) are Nevada (56 percent), Arizona (51 percent), Florida (44 percent), and Idaho (40 percent), all of which are included in the data set. By comparison, the overall percentage decline in construction employment in the United States during the same time period amounted to only 28 percent.

Chart 1 presents industry histograms that display monthly numbers of nonfarm employees broken down by Bureau of Labor Statistics (BLS) sectors and supersectors. The histograms start in November 2001, which marks the trough in employment following the 2001 recession. The light gray bars indicate increases in employment relative to the respective previous month; similarly, the dark gray bars indicate decreases in employment relative to the previous month.

Chart 1: Change in Industrial Structure Following the 2001 Post-Recession Low in Employment

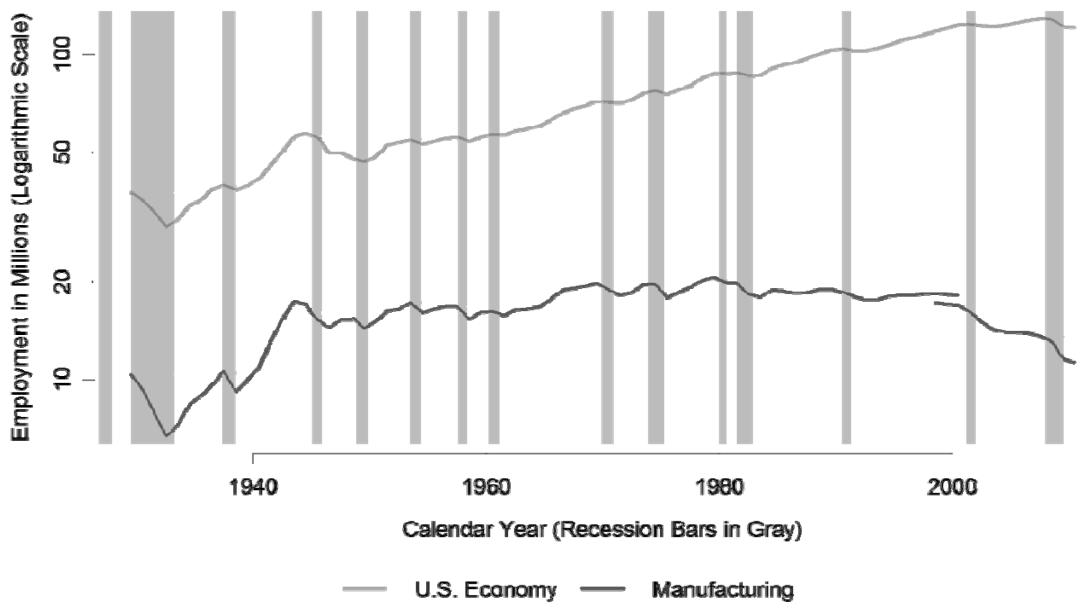


Note: Nonfarm Employment, seasonally adjusted, monthly observations, November 2001 through September 2011. Source: FRED, https://research.stlouisfed.org/fred2; U.S. Bureau of Labor Statistics (BLS), http://www.bls.gov.

Most noticeable are the comparatively steep percentage declines in employment in the manufacturing and construction sectors during and immediately following the latest recession (which is indicated by the gray rectangles). At the same time, the employment increases in the supersector Education and Health Services showed little signs of slowing during the latest recession. Note that the public school system is included in the government sector; more than 80 percent of the employees in the supersector Education and Health Services fall into the category Health Care and Social Assistance (or, health care, for short), the remainder comprising Educational Services.

Although manufacturing employment started recovering in 2010, this recovery may fail to restore the pre-recession level of employment. Such has been the experience since the double-dip recession of the early 1980s. As Chart 2 shows, over the past 30 years, manufacturing has never reached the pre-recession level during the subsequent economic recovery.

Chart 2: Manufacturing Employment



Note: Employment is from National Income and Product Account (NIPA) Tables 6.5 (A [1929-1948], B [1948-1987], C [1987-2000], and D [1998-2009]) and measured in numbers of full-time equivalent employees. The break in the manufacturing series indicates the switch from SIC (Standard Industrial Classification) to NAICS (North America Industry Classification System). Frequency of observation: annual; latest available data point: 2010. Source: Bureau of Economic Analysis (BEA), http://www.bea.gov.

1.1 Research Context

Understanding loss cost trends and forecasting these trends is challenging. First, the data series are comparatively short and, second, these series exhibit a high degree of volatility.

There are two types of models that have been developed for the purpose of eliciting loss cost trends. First, there are models that estimate the trend rate of growth of the loss cost component under the assumption of stationarity on time intervals between structural breaks. Among these models are the Exponential Trend model traditionally considered in NCCI aggregate ratemaking. Another class of models employs covariates in forecasting trends in lost cost components. The shortness of the time series and their high volatilities pose a particular problem for such structural models. In small samples, only the most extreme contributions of covariates are able to breach the threshold of statistical significance; see Gelman and Weakliem [2]. This is because the shorter the sample, the higher the standard error of the estimated regression coefficient, all else being equal. Specifically, the risk associated with regression on small samples is that the correlation between the dependent variable and the covariate is a random outcome, which does not repeat itself in systematic ways in the future. As a result, the covariate may adversely affect the quality of the forecast. All else being equal, the more covariates are tested during the model building process, the higher is the probability that a covariate will be included based on random correlation.

Among models that employ covariates for the purpose of forecasting loss cost trends are the model discussed here and the model developed by Brooks [1]. The covariate in the Brooks model that has the most explanatory power is the ratio of cumulative injury claims to total indemnity claims. This raises another problem of structural models, which is the availability of forecasts for the covariates.

1.2 Objective

The objective is to quantify the effect of changes in the industrial structure on the lost cost components in NCCI aggregate ratemaking. This quantification is primarily for the purpose of improving the understanding of loss cost trends during the aftermath of the 2007–2009 economic recession. It is not the objective of the model to deliver trend estimates that feed directly into the trend selection in aggregate ratemaking; this is because this model is geared toward an economic situation that is likely to be transitory. The degree of change in the industrial structure may not perpetuate itself at the current pace, which would deprive the model of its value added. Likewise, had this model been developed prior to the latest recession, it may have been dismissed due to a lack of explanatory power.

1.3 Outline

What follows in Section 2 is a description of the model. Section 3 then discusses the data and presents the estimated effects. Section 4 concludes.

2. THE MODEL

In three independent regressions, the logarithmic growth rates of frequency and the severities are related to the logarithmic rates of employment growth by industry and the first difference in the logarithmic rate of growth of private nonfarm employment. For every state, the conditional mean of the logarithmic rate of growth of the respective loss cost component is assumed to be normally distributed. All data are state-level observations.

The states are estimated simultaneously in a pooled time-series cross-section framework. All states share the same regression coefficients (of the employment growth by industry and the change in the rate of private nonfarm employment growth); this is to reduce the risk of mistaking noise for a signal in a situation where the time series are short and volatile, as discussed. The variance of the normal distribution of the conditional mean is allowed to vary by state, thus resulting in a discrete scale mixture of normal distributions as the likelihood of the Bayesian model.

In addition to the growth rates of employment by industry, the first difference in the rate of private nonfarm employment growth is the only other covariate. The inclusion of this covariate in the regression model is motivated by the study by Schmid [6] on job flows. In this study, it was shown that the rate of frequency growth is affected by changes in the rates of job creation and job destruction. Because forecasts for job creation and destruction are not available, only the net job flows effect (that is, net job creation as it manifests itself in employment growth) is included in the model.

In a sensitivity analysis, a ridge regression version of the model was estimated. Ridge regression was developed by Hoerl and Kennard [3][4] for the purpose of reducing the mean square error of estimated regression coefficients in the presence of nonorthogonality. Specifically, ridge regression is an approach to mitigate the adverse effects of multicollinearity, which arises from correlations among the set of covariates. Such multicollinearity may cause the estimated regression coefficients to be excessively large in magnitude or have the wrong sign.

The use of ridge regression is motivated by the possibility that the industry growth rates are highly correlated with each other, as they are subject to common shocks (associated with economic recessions and expansions). On the other hand, if the differences between the ridge estimates and the conventional estimates are small, this indicates that these correlations are mild and the estimated regression coefficients can be relied upon.

All covariates were centered, which implies that the intercept of the regression equation delivers the growth rate of the respective loss cost component in the steady state.

For the purpose of the ridge regression, all covariates were standardized, which means that they were not only centered but also normalized by their sample standard deviations. Following the estimation, the regression coefficients were transformed back to non-standardized (yet centered) coefficients for the purpose of making them comparable to the conventional estimates.

Ridge regression is straightforward to implement in Bayesian regression models, as documented in The BUGS Project [7]. After standardizing the covariates, the prior distributions for the covariates are assigned a common precision (which is the reciprocal value of the variance). In the model, a common precision is shared by the parameters measuring the impact of the employment growth by industry (but not by the parameter that quantifies the effect of the change in the rate of private nonfarm employment growth).

In order to align the time window of the covariates with policy years, the rate of employment growth by industry was measured Q2 (second quarter of the calendar year) over Q2. For instance, the rate of frequency growth in construction employment for Policy Year 2008 is calculated by comparing the employment level in Q2/2009 to the one in Q2/2008. For accident year data, the corresponding growth rate was measured Q4 over Q4. The same principle applies to the difference in the rate of growth of private nonfarm employment.

For the purpose of statistical modeling, the industries displayed in Chart 1 were aggregated. The resulting broader industry categorization reduces the number of covariates and thus the risk of fitting to noise. Further, the government sector was dropped because this sector is not pertinent to the analysis.

Chart 3 displays the industry aggregation. The services sectors were grouped into a single industry, except for Trade, Transportation, and Utilities and Education and Health Services. Trade, Transportation, and Utilities is a large sector and, unlike some other services sectors, its level of employment is highly susceptible to variations in economic activity. For this reason, the Trade, Transportation, and Utilities sector is treated as a stand-alone industry. Education and Health Services shows a trajectory of employment growth different from any other services industry (as shown in Chart 1) and hence deserves special attention. The three goods-producing sectors Construction, Manufacturing, and Natural Resources and Mining were also left disaggregated.

The model was implemented in R (http://cran.r-project.org/) using the sampling platform JAGS (Just Another Gibbs Sampler, http://www-ice.iarc.fr/~martyn/software/jags/). JAGS was linked to R by means of the R package rjags (http://cran.r-project.org/web/packages/rjags/index.html).

Chart 3: Industry Aggregation

	BLS Sectors and Supersectors	Industry Breakdown
Service-providing	Trade, Transportation, and Utilities	**Trade, Transportation, and Utilities**
	Education and Health Services	**Education and Health Services**
	Financial Activities Information Leisure and Hospitality Professional and Business Services Other Services	**Services NEC (Not Elsewhere Classified)**
Goods-producing	Construction	**Construction**
	Manufacturing	**Manufacturing**
	Natural Resources and Mining	**Natural Resources and Mining**

3. FINDINGS

The data (expressed as growth rates) are from the 2010 NCCI ratemaking season and run from 1991 through 2008 for policy years, and from 1992 through 2009 for accident years. Employment data are on a quarterly basis and provided by a professional data provider and forecaster. Note that forecasts are subject to revision; the forecasts used in this study are as of July 6, 2011. From these quarterly observations, 12-month growth rates are calculated, as discussed. Historical employment observations run through Q1/2011; employment forecasts start in Q2/2011.

The measured quantitative effects for the influence of the industrial structure on loss cost components are presented in Table 1 (frequency growth), Table 2 (indemnity severity growth), and Table 3 (medical severity growth). The displayed regression coefficients have a straightforward interpretation. For instance, in Table 1, the regression coefficients measure the change of frequency growth in percentage points in response to a one-percentage point change in employment growth in the industry in question. As an example, let us start out in a situation where frequency declines at an annual rate of four percent and construction employment grows at an annual rate of one percent. Then, the construction sector shrinks, its growth rate dropping by seven percentage points to a negative six percent. If nothing else changed, frequency growth would increase by approximately one percentage point (a negative 0.07 multiplied by a negative 0.154) to a negative three percent from the original negative four percent.

At the same time, the decrease in the growth rate of construction employment, if not offset by changes in employment growth in other industries, changes the rate of private nonfarm employment growth in the state. This implication points to the second effect, which is related to job flows. For the frequency growth regression, the posterior distribution of the coefficient that reflects the influence of changes to the rate of employment growth is displayed in Chart 4.

Table 1: Effect on Frequency Growth in Percentage Points in Response to a One-Percentage Point Change in Industry Employment Growth Change

	Discrete Scale Mixture of Normal Distributions		
	Standard Approach	Ridge Regression	Probability that Variable Has Explanatory Power (Percent)
	Estimated Coefficient	Estimated Coefficient	
Trade, Transportation, and Utilities	-0.132	-0.129	13.17
Education and Health Services	-0.430	-0.422	87.47
Services NEC	0.149	0.135	23.68
Construction	-0.154	-0.148	87.61
Manufacturing	0.125	0.125	23.69
Natural Resources and Mining	0.002	0.001	8.53

Note that the public school system is part of Government, not Education and Health Services. The bulk of jobs in Education and Health Services fall into the category Health Care and Social Assistance.

Chart 4: Effect of Change in Net Job Creation on Frequency Growth

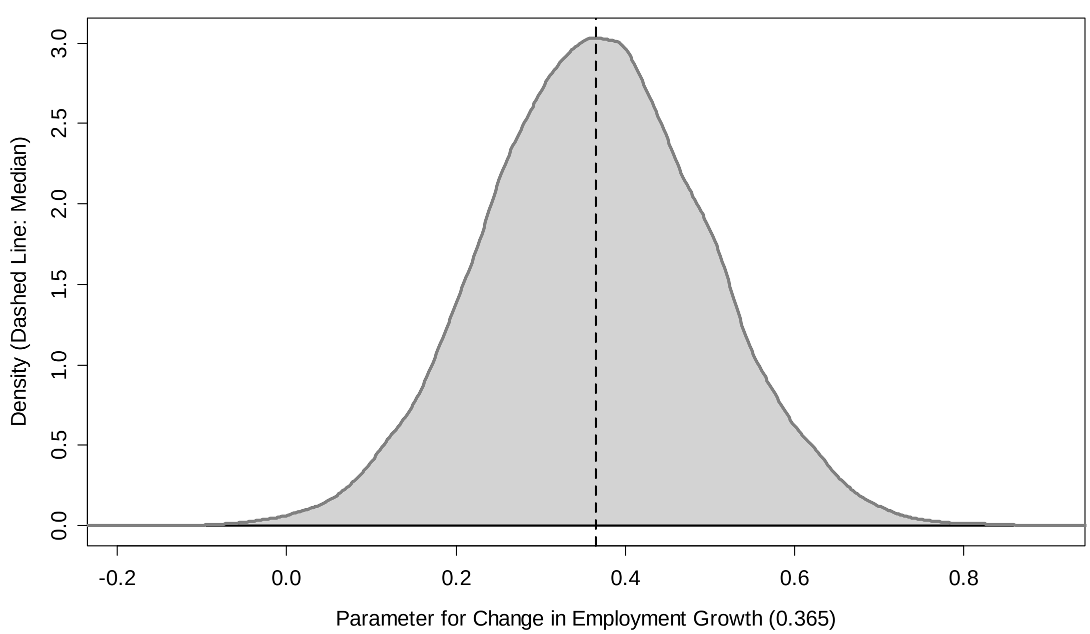


Chart 4 shows that a drop in the rate of private nonfarm employment growth, as typically associated with a slowdown in economic growth, causes the rate of frequency growth to decline. Specifically, economic recessions tend to accelerate the decline in frequency, whereas recoveries from recessions tend to slow down this decline or reverse it temporarily.

Table 2: Effect on Indemnity Severity Growth in Percentage Points in Response to a One-Percentage Point Change in Industry Employment Growth Change

	Discrete Scale Mixture of Normal Distributions		
	Standard Approach	Ridge Regression	Probability That Variable Has Explanatory Power (Percent)
	Estimated Coefficient	Estimated Coefficient	
Trade, Transportation, and Utilities	0.120	0.104	4.73
Education and Health Services	0.035	0.027	1.91
Services NEC	-0.248	-0.229	4.48
Construction	0.058	0.054	2.43
Manufacturing	-0.022	-0.023	2.21
Natural Resources and Mining	-0.015	-0.013	2.13

Note that the public school system is part of Government, not Education and Health Services. The bulk of jobs in Education and Health Services fall into the category Health Care and Social Assistance.

Table 3: Effect on Medical Severity Growth in Percentage Points in Response to a One-Percentage Point Change in Industry Employment Growth Change

	Discrete Scale Mixture of Normal Distributions		
	Standard Approach	Ridge Regression	Probability That Variable Has Explanatory Power (Percent)
	Estimated Coefficient	Estimated Coefficient	
Trade, Transportation, and Utilities	0.425	0.383	96.30
Education and Health Services	0.148	0.143	25.89
Services NEC	-0.508	-0.455	93.22
Construction	0.151	0.142	96.27
Manufacturing	-0.198	-0.191	91.07
Natural Resources and Mining	-0.072	-0.070	98.20

Note that the public school system is part of Government, not Education and Health Services. The bulk of jobs in Education and Health Services fall into the category Health Care and Social Assistance.

To continue with the numerical example, if the mentioned hypothetical contraction of the construction sector reduced the rate of private nonfarm employment by three tenths of a percentage point, then this effect of a decline in the rate of job creation would accelerate the frequency decline by approximately 0.1 percentage points (a negative 0.003 multiplied by 0.365), thus bringing the total effect on frequency growth to about 0.9 percentage points. In summary, the rate of frequency decline would slow to 3.1 percent from the original four percent.

Tables 1 through 3 display the ridge regression results alongside the conventional regression outcomes. Because ridge regression shrinks the regression coefficients toward zero, these coefficients are smaller in absolute value than those generated by the conventional regression approach. The comparatively modest differences in magnitudes between the coefficients of the two approaches validate the conventional regression estimates.

Evidence of explanatory power of the covariates was established by means of the Generalized Linear Spike-and-Slab Stochastic Search Variable Selection approach developed by Pang and Gill [5]. By means of variable selection, it is possible to provide a probability that the null hypothesis is true (and, hence, a probability that the null hypothesis is false). For instance, in Table 1, the probability that variation in construction employment growth has power in explaining variation in frequency growth equals 87.61 percent. The only other industry with a high probability of having explanatory power is health care; the probability that variations in health care employment are relevant for variations in frequency growth equals 87.47 percent. Interestingly, the probability that variations in the growth rate of manufacturing employment contribute to variations in the growth rate of frequency runs at only 23.69 percent.

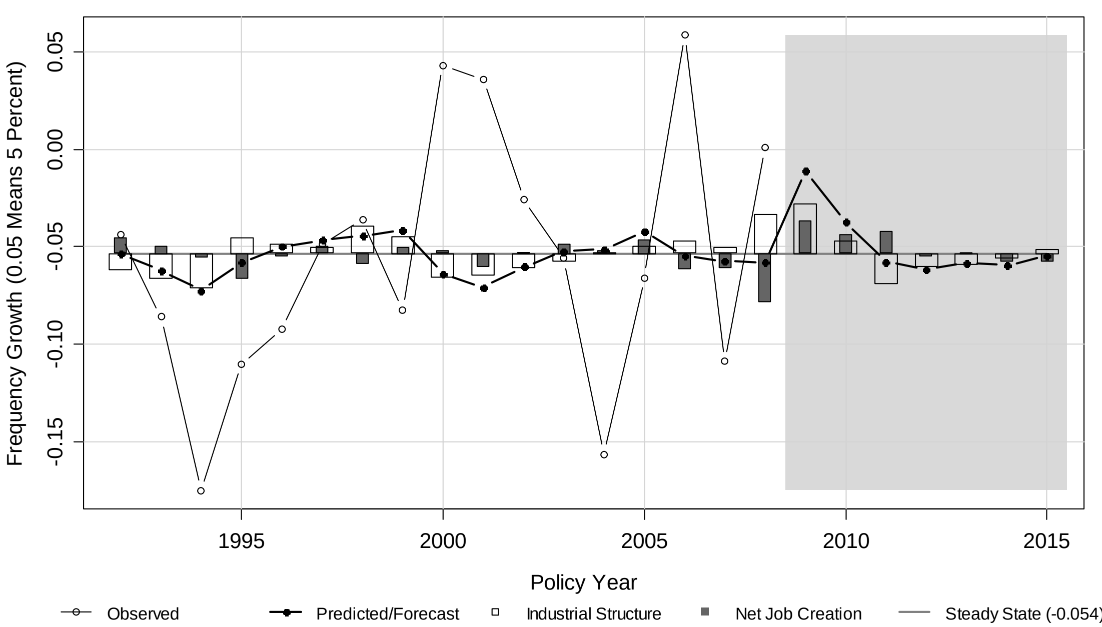
The reason that a decline in the growth rate of construction employment has a positive impact on frequency growth (that is, decelerates the decline in frequency) is that construction is a low-frequency industry in the context of NCCI aggregate ratemaking. This is because, for aggregate ratemaking purposes, frequency is defined as claim count normalized by premium, as opposed to claim count normalized by wages or the number of full-time equivalent employees.

The interpretation of the regression coefficients in Table 2 (indemnity severity) and Table 3 (medical severity) is analogous to the reading of the regression coefficients in Table 1 (frequency) delivered above. It is noteworthy that all the regression coefficients switch signs when going from frequency (Table 1) to the severities (Tables 2 and 3). This confirms that the effects on frequency growth of changes in the rates of employment growth by industry are at least partially offset by effects on the severities. This offset originates primarily from the medical severity where all but one regression coefficient exhibit high probabilities of having explanatory power. This is in stark contrast to the regression coefficients for indemnity severity, none of which musters more than a five-percent probability of being statistically relevant.

As discussed, the change in the rate of private nonfarm employment growth is the only covariate aside from the rates of employment growth by industry. The probability that changes in private nonfarm employment growth have statistical power to explain frequency growth equals 90.73 percent. This statistical evidence agrees with the role of jobs flows for the business cycle behavior of frequency discussed in Schmid [6]. The probabilities that changes in the rate of private nonfarm employment growth are relevant to the rates of indemnity growth equals 100 percent; the corresponding number for medical severity reads 32.57 percent.

Chart 5 summarizes for the first analyzed state the regression results for frequency growth. In this state, which is on policy years, the (percentage) decline in construction employment approximately equals the reading for the United States overall. The observed rates of frequency growth are plotted alongside the values predicted by the model. The gray box represents the future from the perspective of the 2010 ratemaking season. Within the gray box, the predicted values for the rate of frequency growth are forecasts. All displayed growth rates are logarithmic.

Chart 5: State No. 1: Job Losses in Construction Were Close to the National Average

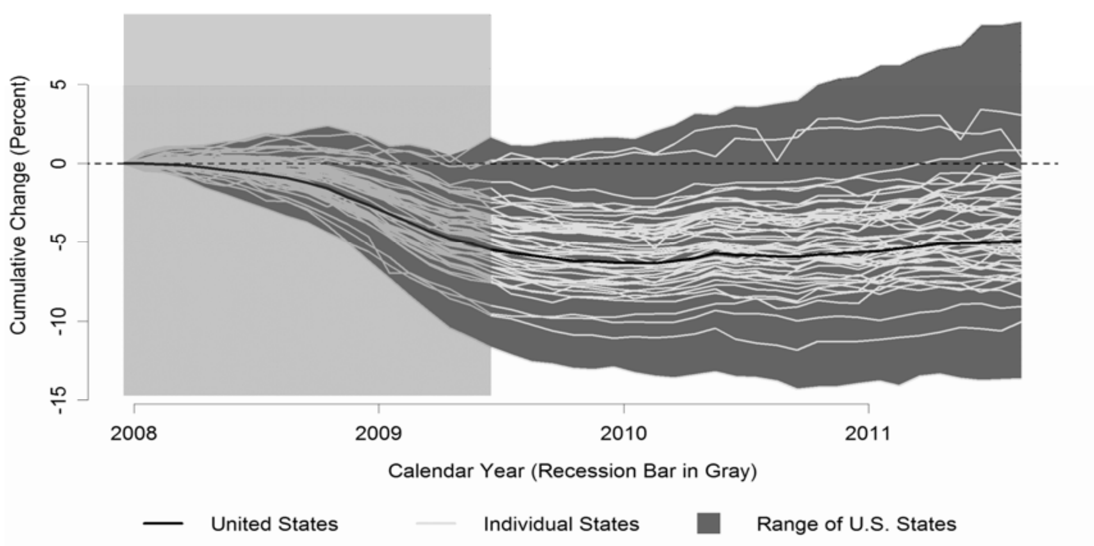


The steady state rate of frequency growth of the state displayed in Chart 5 runs at a negative 5.4 percent and is represented by a horizontal line. By comparison, the equally weighed average of the steady state rates of frequency growth among the 37 states in the data set equals a negative 4.4 percent; the corresponding premium-weighted average comes to a negative 4.3 percent. (Premiums are as of Policy Year 2008 or Accident Year 2009, as applicable.)

In Chart 5, the transparent box represents the influence of changes in the industrial structure, as measured by variations in the rates of employment growth by industry; the solid gray box stands for the effect of variations in the change of the rate of private sector employment growth. Both boxes measure the influence relative to the steady state.

There is little action in the estimated effects prior to Policy Year 2008. Although the recession started in December 2007, most states did not experience meaningful declines in employment before mid-2008. The delayed response of employment to the recession is evident from Chart 6, which displays the cumulative employment growth since the onset of the recession for every U.S. state, the District of Columbia, and the United States. In this chart, the recession is represented by a gray box.

Chart 6: Cumulative Change in Nonfarm Employment Since Onset of Recession



Note: Nonfarm Employment, seasonally adjusted. The set of individual states and the range of U.S. states include the District of Columbia. Frequency of observation: monthly; latest available data point for US: August 2011. Tick marks indicate beginning of year. Sources: U.S. Bureau of Labor Statistics (BLS), http://www.bls.gov; NBER, http://www.nber.org/cycles.html.

As shown in Chart 5, in Policy Year 2009, the change in industrial structure contributed to an increase in frequency, which was more than offset by the influence originating from the slowdown in net job creation. As hiring moderates, the share of short-tenured workers on the payroll declines, which contributes to an acceleration of the frequency decline. This reasoning is supported by Chart 7, which states that about a quarter of workers are responsible for about one-third of all workplace injuries—these are the workers that have been with the current employer for one year or less.

Chart 7: Workplace Injury Proportions by Job Tenure

Note: Workplace injuries represent nonfatal injuries and illnesses involving days away from work. Short job tenure means 11 months or less (workplace injuries) or 12 months or less (employment). Intermediate length of service means one to five years and 13 months to four years, respectively. Long length of service means more than five years or five years or more, respectively. Percentages for workplace injuries do not account for a small "residual category." Job tenure information for employment is available bi-annually (for January only). Percentages may not add to 100 due to rounding. Source: U.S. Bureau of Labor Statistics (BLS), http://www.bls.gov.

Chart 5 shows two effects on the rate of frequency growth that holds for nearly all of the 37 analyzed states. First, in Policy Year 2008 (or Accident Year 2009, where applicable), the industrial structure effect and the net job creation effect opposed each other; while the industrial structure effect contributed to an increase in the rate of frequency growth, the net job creation effect subtracted from it. Second, in Policy Year 2009 (or Accident Year 2010, where applicable), the two effects pointed in the same direction by contributing to an increase in frequency growth. The reason for the net job creation effect changing direction is that employment growth turned positive in many states in late 2009 and early 2010, as depicted in Chart 6.

For further five states, the regression results for frequency growth are summarized in Charts 8 through 12. Similar to the state exhibited in Chart 5, the state displayed in Chart 8 experienced a (percentage) decline in construction employment that was about equal to the decline in the United States overall. Things are different in Charts 9 through 12, which represent states with sharply contracting construction sectors; note that states that experienced steep declines in construction employment were also subject to above-average (percentage) losses in private nonfarm employment. As is evident from Charts 9 through 12, both the industrial structure effect and the net job creation effect are more pronounced (when measured in changes of percentage points) than in the states shown in Charts 5 and 8. Taking the two effects together, frequency growth in Policy Year 2009 (Accident Year 2010) is forecast to be several percentage points higher (but not necessarily positive) for states with worse-than-average slumps in the housing market.

Chart 8: State No. 2: Job Losses in Construction Were Close to the National Average

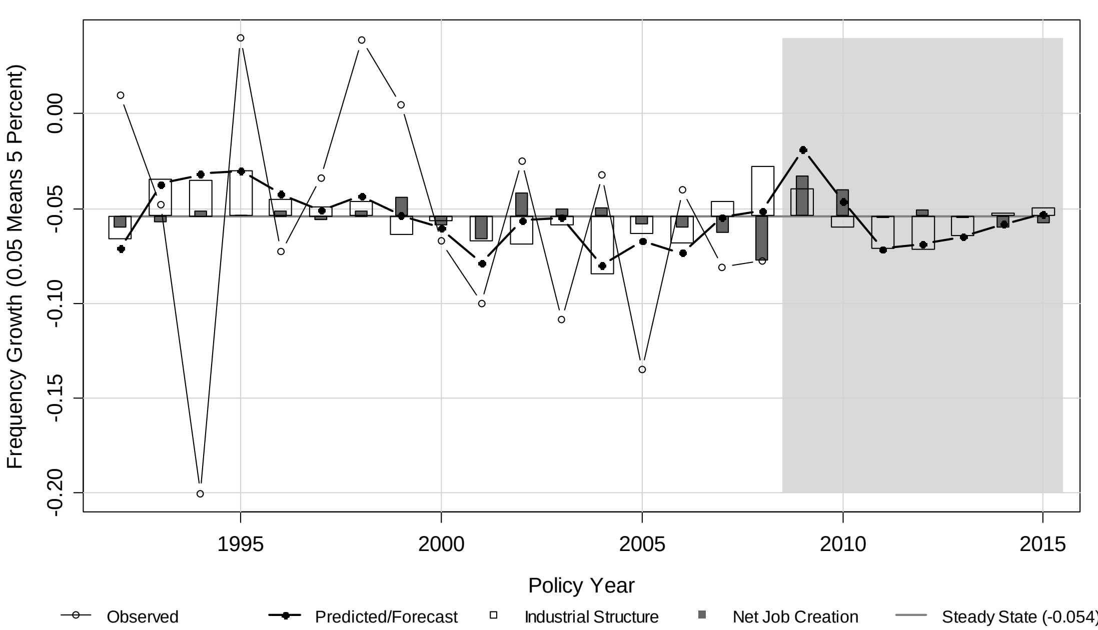


Chart 9: State No. 3: Job Losses in Construction Exceed the National Average

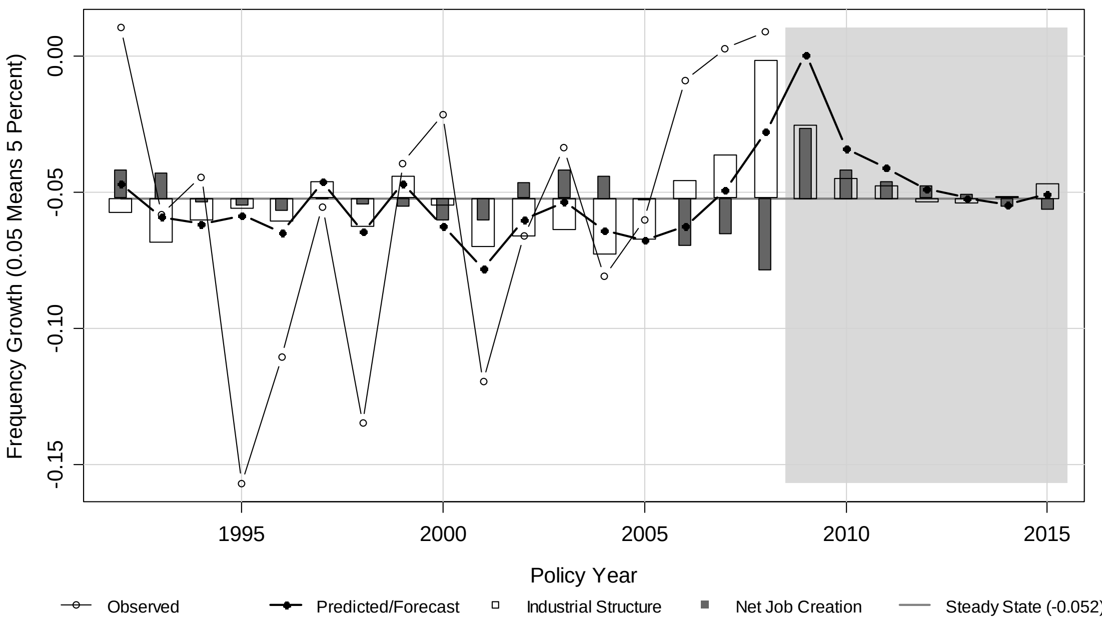


Chart 10: State No. 4: Job Losses in Construction Exceed the National Average

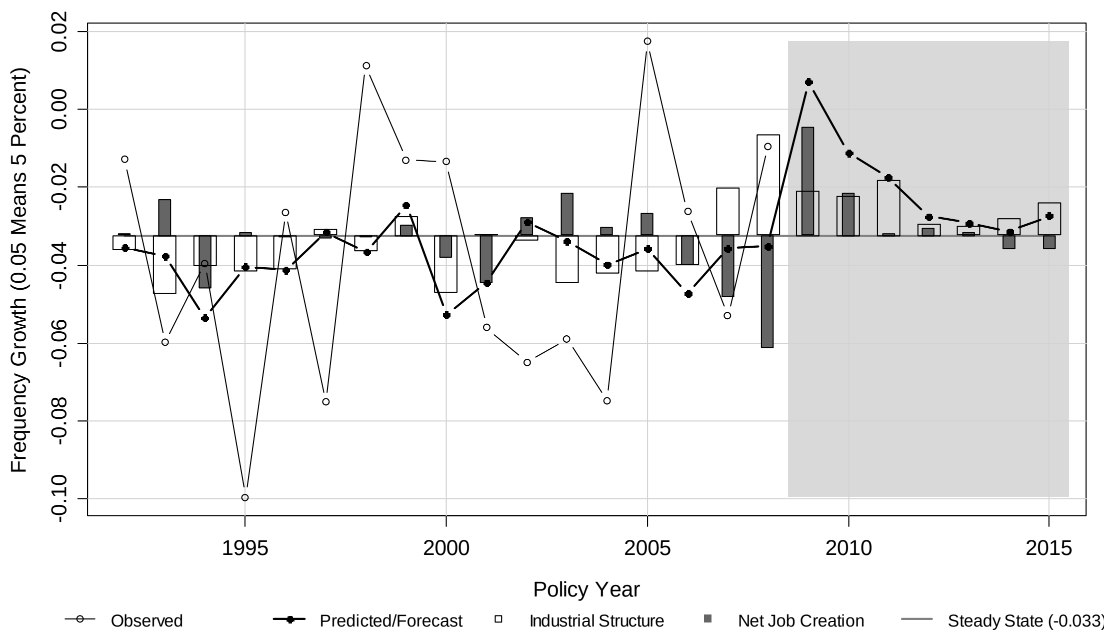


Chart 11: State No. 5: Job Losses in Construction Exceed the National Average

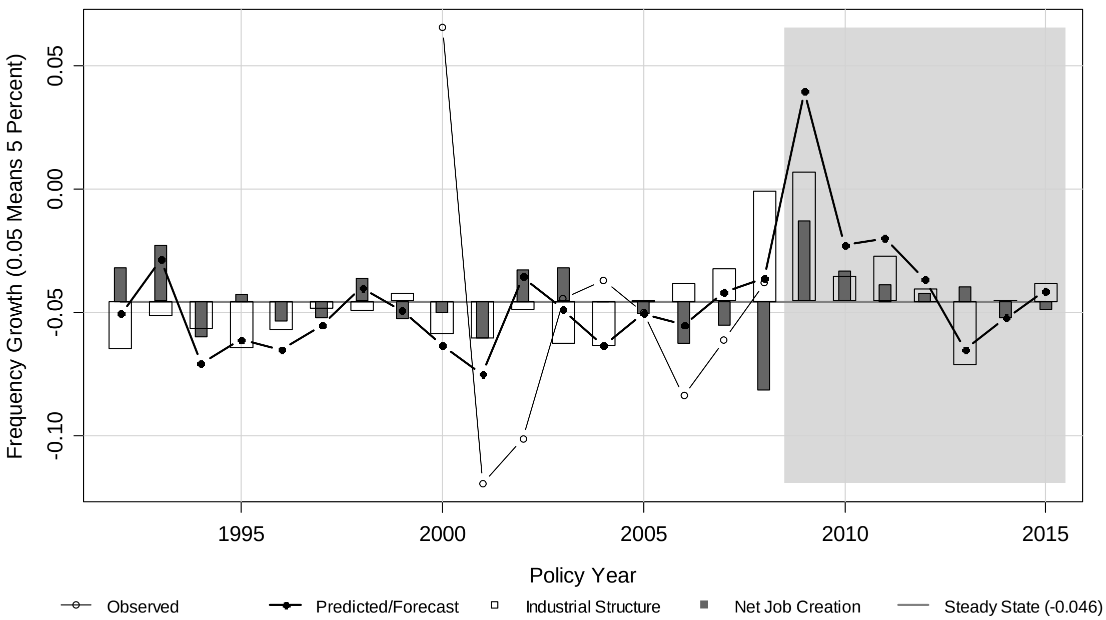


Chart 12: State No. 6: Job Losses in Construction Exceed the National Average

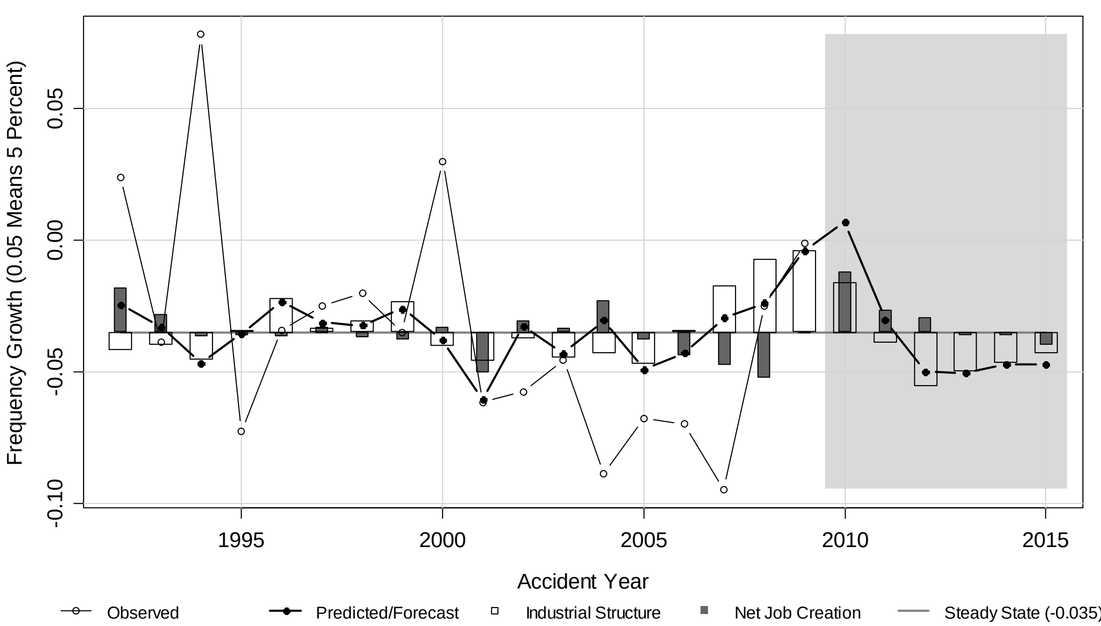


Chart 5 and Charts 8 through 12 show that the economic effects that emanate from the changes in the industrial structure effect and the net job creation are expected to taper off quickly after Policy Year 2009 (Accident Year 2010). Based on the forecasts for employment growth used in the model, the effects will have largely dissipated by Policy (Accident) Year 2015.

4. CONCLUSIONS

The change in industrial structure concomitant with the 2007-2009 recession and its aftermath affects ratemaking loss cost components in important ways. This effect is largely due to the pre-recession expansion and subsequent contraction of the construction sector. Thus, the effect is most pronounced in states that experienced a steep downturn in the housing market.

There are offsetting effects on the part of the severities. Medical severity appears to be more responsive to changes in industrial structure than indemnity severity, although this finding should not be considered conclusive.

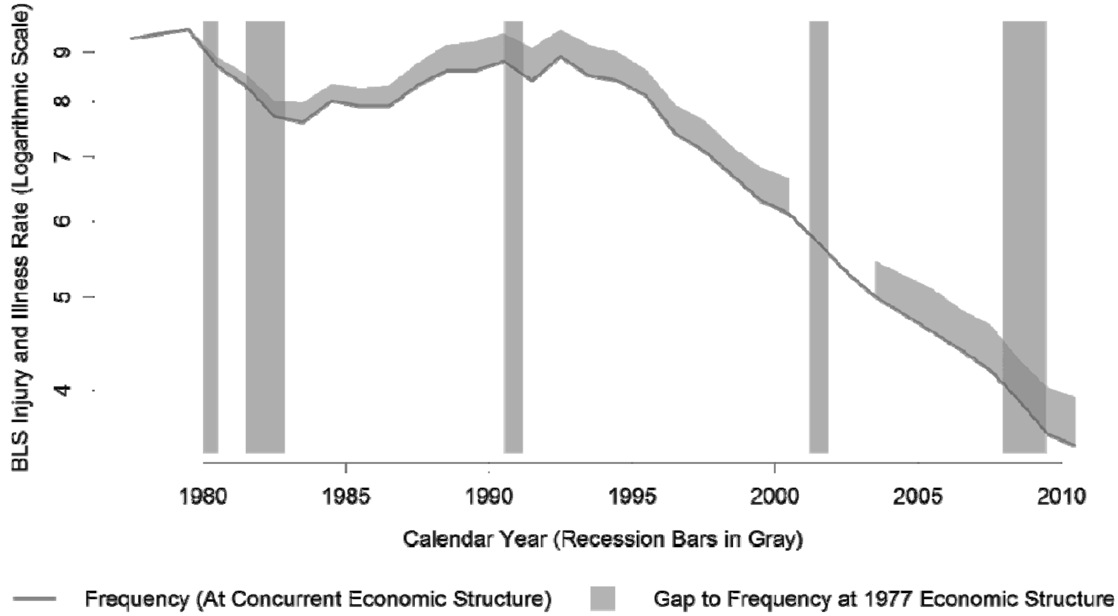
The model has not been validated using a holdout period. This is because, within the history of available aggregate ratemaking data, there have been no changes in the industrial structure of the economy comparable in magnitude.

The effect is likely to be transitory. In the model, the rates of frequency and severity growth revert to their steady states as the economy returns to its trend rate of growth. It is important to note that this return to the steady state is a model assumption, not a model outcome.

In fact, an important question raised by this analysis is how the recent changes in industrial structure may have altered the steady state of frequency growth. Although this question is beyond the scope of the analysis presented here, there is evidence that the role of the industrial structure in determining the rate of frequency decline in the steady state is rather small.

Chart 13 shows the BLS injury and illness rate for all private industry on an annual basis since 1977, the first year for which this data series is available. The edge of the gray shading reports the injury and illness rate that we would have observed had the industrial structure not changed in 1977. The chart shows that the bulk of the frequency decline happened within the industries, and only 7.5 percent of the frequency decline that was observed between 1977 and 2010 is due to the change in industrial structure. This finding motivated the assumption of an unaltered steady state in the presented statistical model.

Chart 13: BLS Injury and Illness Rate and Change in Industrial Structure



Note: Injury and Illnesses Cases per 100 Full-Time-Equivalent Workers, Total Recordable Cases, All Private Industry. Frequency of observation: annual; latest available data point: 2010. No data points are available for 2001 and 2002 due to changes in industry classification. Tick marks indicate beginning of year. The data points are mid-year. Sources: Bureau of Economic Analysis (BEA), http://www.bea.gov; U.S. Bureau of Labor Statistics (BLS), http://www.bls.gov/iif.

Acknowledgment

Thanks to Sikander Nazerali for research assistance.

5. REFERENCES

[1] Brooks, Ward, "A Study of Changes in Frequency and Severity in Response to Changes in Statutory Workers Compensation Benefit Levels," Casualty Actuarial Society *Proceedings*, **1999**: Vol. LXXXVI, No. 164 & 165, www.casact.org/pubs/proceed/proceed99/99046.pdf.

[2] Gelman, Andrew, and David Weakliem, "Of Beauty, Sex, and Power: Statistical Challenges in Estimating Small Effects," http://www.stat.columbia.edu/~gelman/research/unpublished/power.pdf, September 8, 2007.

[3] Hoerl, Arthur E., and Robert W. Kennard, "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics,* **1970**, Vol. 12, No. 1, 55-67.

[4] Hoerl, Arthur E., and Robert W. Kennard, "Ridge Regression: Applications to Nonorthogonal Problems," *Technometrics,* **1970**, Vol. 12, No. 1, 69-82.

[5] Pang, Xun, and Jeff Gill, "Spike and Slab Prior Distributions for Simultaneous Bayesian Hypothesis Testing, Model Selection, and Prediction, of Nonlinear Outcomes," Working Paper, Washington University in St. Louis, http://polmeth.wustl.edu/mediaDetail.php?docId=914, July 13, 2009.
[6] Schmid, Frank A., "Workplace Injuries and Job Flows," www.ncci.com/Documents/WorkplaceInjuries-0709.pdf, July 31, 2009.
[7] The BUGS Project, *Examples Vol. 1*, http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/Vol1.pdf, [no date].

Abbreviations and notations

BEA, Bureau of Economic Analysis
BLS, Bureau of Labor Statistics
BTS, Bayesian Trend Selection
BUGS, Bayesian Inference Using Gibbs Sampling
FRED, Federal Reserve Economic Data
JAGS, Just Another Gibbs Sampler
MCMC, Markov-Chain Monte Carlo Simulation
NAICS, North American Industrial Classification System
NBER, National Bureau of Economic Research
NCCI, National Council on Compensation Insurance
NIPA, National Income and Product Accounts
SIC, Standard Industrial Classification

Biography of the Author

Frank Schmid, Dr. habil., was, at the time of writing, a Director and Senior Economist at the National Council on Compensation Insurance, Inc.

Bayesian Trend Selection

Frank Schmid, Chris Laws, and Matthew Montero

Abstract

Motivation. Selecting loss ratio trends is an integral part of NCCI aggregate ratemaking. The trend selection draws on an exponential trend (ET) regression model that is applied, alternatively, to the latest 5, 8, and 15 observations (dubbed 5-point, 8-point, and 15-point ET). Then, using actuarial judgment (which may account for a variety of influences), the three estimates are aggregated into a single forecast. This process of decision making under uncertainty can be formalized using Bayesian model selection.

Method. A Bayesian trend selection (BTS) model is introduced that averages across the three ET models. Using a double-exponential likelihood, this model minimizes the sum of absolute forecast errors for a set of (overlapping) holdout periods. The model selection is accomplished by means of a categorical distribution with a Dirichlet prior. The model is estimated by way of Markov chain Monte Carlo simulation (MCMC).

Results. The BTS is validated on data from past ratemaking seasons. Further, the robustness of the model is examined for past ratemaking data and a long series of injury (and illness) incidence rates for the manufacturing industry. In both cases, the performance of the BTS is compared to the 5-point, 8-point, and 15-point ET, using the random walk as a benchmark. Finally, for the purpose of illustration, the BTS is implemented for an unidentified state.

Availability. The model was implemented in R (cran.r-project.org/), using the sampling platform JAGS (Just Another Gibbs Sampler, www-ice.iarc.fr/~martyn/software/jags/). JAGS was linked to R via the R package rjags (cran.r-project.org/web/packages/rjags/index.html).

Keywords. Model Selection, Model Averaging, Model Robustness, Aggregate Ratemaking, Trend Estimation, Workers Compensation

1. INTRODUCTION

Selecting loss ratio trends is an integral part of NCCI aggregate ratemaking. The trend selection draws on an exponential trend (ET) regression model that is typically applied to the latest 5, 8, and 15 observations; these three alternative regressions are dubbed the 5-point, 8-point, and 15-point ET. Then, using actuarial judgment (which may account for a variety of influences), the three estimates are aggregated into a single forecast. This process of decision making under uncertainty can be formalized using Bayesian model selection.

In what follows, a Bayesian trend selection (BTS) model is introduced that averages across the three ET models. Using a double-exponential likelihood, this model minimizes the sum of absolute forecast errors for a set of (overlapping) holdout periods. The model selection is accomplished by means of a categorical distribution with a Dirichlet prior. The model is estimated by way of Markov chain Monte Carlo simulation (MCMC).

The BTS is validated on data from past ratemaking seasons. Further, the robustness of the model is examined for past ratemaking data and a long series of injury (and illness) incidence rates for the manufacturing industry. In both cases, the performance of the BTS is compared to the 5-

point, 8-point, and 15-point ET, using the random walk as a benchmark. Finally, for the purpose of illustration, the BTS is implemented for an unidentified state.

The BTS, as any model, is limited in the scope of information that it is capable of processing—at the same time, this statistical model is not subject to potential biases of human decision making. On one hand, human judgment is capable of considering influences beyond the scope of statistical models, such as changing economic conditions. On the other hand, research pioneered by Amos Tversky and Daniel Kahneman shows that human decision making is subject to systematic errors. Although decision making under uncertainty is liable to random errors (if only due to measurement errors in the data), systematic errors lead to predictable biases. Most interestingly, Tversky and Kahneman [10] argue that the prevalence of biases is "not restricted to the layman. Experienced researchers are also prone to the same biases—when they think intuitively."

Among the biases identified by Tversky and Kahneman [10] is one related to the availability of events, that is, the ease with which occurrences can be retrieved from the memory bank. The more available an event is, the more this event weighs on the decision, all else being equal. Among the factors shaping the availability of an event is its salience. Events that occur more recently or have greater degrees of immediacy (for instance, seeing a house burn, as opposed to reading about the fire in the newspaper) loom larger on the human mind. The salience bias poses a risk of overreacting to recent events and extreme outcomes. For an extensive discussion of the availability bias, see Taylor [9].

1.1 Research Context

Past work on trend modeling in the context of ratemaking in workers compensation includes Brooks [2], Evans and Schmid [4], and Schmid [7]. All of these approaches face a major challenge in the shortness and comparatively high volatility of the available time series.

Brooks [2] studied the determinants of indemnity frequency for California using ordinary least squares regression. Indemnity frequency is defined as the ratio of indemnity-related claim count to payroll at the 1987 wage level. The explanatory variables comprise measures of indemnity benefit and medical benefit levels, macroeconomic variables, and the ratio of cumulative injury claims to total indemnity claims; the latter variable accounts for more than half the explanatory power of the presented models.

In Brooks [2], all variables are transformed into first differences (on the raw scale; on the scale of the natural logarithm, such first differences would represent continuously compounded rates of growth). The author runs 84 indemnity frequency regressions (the maximum possible number, given the set of available explanatory variables), orders them by the adjusted R-squared, and then publishes the seven highest ranking models.

Brooks [2] aims at a high R-squared (adjusted for the degrees of freedom), by having "accounted for as much variation as possible." Brooks [2] reads a high explanatory power of the model as an indication that the estimates "are not distorted due to a misspecified model with a large portion of unaccounted-for variance." Yet, there is little relation between a small forecast error and a high R-squared. For instance, in a game of (two six-faced) dice, seven is the best forecast for any toss, because this is the forecast that minimizes the expected forecast error; at the same time, regressing 49 sevens on the 49 possible outcomes of this game of dice generates an R-squared of zero. See also Armstrong [1], who argues against the use of the R-squared (and, more generally, measures of in-sample fit) in selecting the most accurate time-series model.

The approach of considering a multitude of models, as pursued by Brooks [2], creates a condition of multiple comparisons. Considering a set of statistical hypotheses simultaneously increases the probability of false positives in the context of traditional significance tests, and thus invalidates conventional measures of statistical significance. For instance, when applying a type 1 error probability of 10 percent in an *F*-test (that is, in an analysis of variance, which tests the statistical significance of the model), 10 percent of a random set of models are statistically significant by chance, on average. Specifically, among the 84 models that Brooks [2] ran, in expected value terms, eight will turn out statistically significant by chance.

Evans and Schmid [4] discuss the use of the Kalman filter for frequency and severity forecasting in NCCI aggregate ratemaking. Although, in general, the Kalman filter is a powerful tool for breaking down time-series data into signal and noise, the time series employed in NCCI ratemaking are too short to estimate reliably the variances of innovation and white noise. Specifically, in short time series, the Kalman filter may overestimate the innovation variance, thus favoring the random walk (over alternative data-generating processes).

Schmid [7] tries to overcome the problem associated with the Kalman filter in short time series by using a random-walk smoother. The innovation variance of this smoother is calibrated such that the root mean squared forecast error for the time horizon of three years is minimized, on average. Although this model performs as expected, it is difficult to communicate to the decision maker.

The advantage of the approach by Brooks [2] is its use of covariates, which, in a single-equation model, may improve the forecasts considerably. Forecasts for macroeconomic variables that may serve as covariates may be readily available from professional forecasting firms. Conversely, the model developed by Brooks [2] requires forecasts for the ratio of cumulative injury claims to total indemnity claims, which is the covariate that wields the most explanatory power (Brooks [2]). This ratio may be just as difficult to forecast as the variable of interest. From this perspective, the use of covariates can pose an obstacle when using single-equation regression models for the purpose of forecasting.

1.2 Objective

The objective is to formalize the actuarial decision making process of trend selection, rather than developing a new trend model. Statistically, this decision making process under uncertainty is a problem of model selection. The models that compete in this selection process are the 5-point, 8-point, and 15-point ET. The resulting forecast is a weighted average of the three ET estimates, where the weights are the posterior probabilities generated by the model selection process.

1.3 Outline

What follows in Section 2 is a description of the forecasting task and the data. Section 3 offers the theoretical framework of Bayesian model selection. Section 4 presents the validation of the model on past ratemaking data. Section 5 examines the robustness of the model; in addition to ratemaking data, the model is applied to a long time series of manufacturing injury (and illness) incidence rates. Section 6 applies the model to an unidentified state. Section 7 concludes.

2. FORECASTING TASK AND DATA

The objective is to forecast, in the context of aggregate ratemaking, compound annual growth rates for the indemnity and medical loss ratios or, alternatively, compound annual growth rates for frequency and the indemnity and medical severities. Severity is defined as the ratio of (on-leveled, developed-to-ultimate, and wage-adjusted) losses to the number of lost-time claims (developed to ultimate). Frequency is defined as the ratio of the lost-time claim count (developed to ultimate) to on-leveled and wage-adjusted premium. The loss ratio is the product of frequency and the respective severity.

The forecasts apply to the time period between the midpoint of the respective policy (accident) year included in the experience period and the midpoint of the proposed rate effective period—this time interval is called the trend period. For the latest policy (accident) year that is included in the experience period, the length of the trend period usually amounts to slightly more than three years. The rate effective period is the time period during which the rates are expected to be effective, based on the rate filing. In NCCI rate filings, the experience period comprises at least two policy (accident) years.

All data are state-level observations and are on an annual basis. From the perspective of the model, there is no difference between processing policy year and accident year data. (In NCCI ratemaking, nearly all states operate on policy year information.)

The ET and, hence the BTS, do not make use of covariates or an autoregressive process. As a result, the forecast compound annual growth rate (CAGR) can be scaled up to the trend period simply by means of compounding.

In NCCI ratemaking, where the experience period comprises more than one policy (or accident) year, the ET models make use of policy (or accident) year observations through the end of the experience period. Then, different scale factors are applied for the individual policy (or accident) years in the experience period when compounding the CAGR.

3. STATISTICAL MODELING

The BTS is a tool for decision making under uncertainty. Statistically, this concept of decision making is implemented by means of model selection in a Bayesian framework. The objective of the approach is to arrive at a forecast by means of weighted averaging across a set of candidate models. This set of models arises from the ET applied to a set of time intervals, all of which end with the most recent observation but differ by the degree to which they extend into the past. The weights of the average originate in the posterior probabilities by which each of the approaches is deemed to be the “true” model.

3.1 Exponential Trend (ET)

The ET model regresses the (natural) logarithm of (for instance) frequency on a linear trend. The forecast CAGR is then backed out of the regression coefficient that captures this linear time trend.

The semi-logarithmic regression equation reads

$$\ln\left(Y_t\right) = \alpha + \beta \cdot t + \varepsilon_t \ , \ t = 1,\ldots,T \ , \tag{1}$$

where $\ln\left(Y_t\right)$ is the dependent variable (e.g., the natural logarithm of frequency), α is the intercept, β is the parameter that captures the time trend, ε_t is an error term, and T equals the number of observations of the time series.

The forecast CAGR equals

$$e^{\hat{\beta}} - 1 , \tag{2}$$

where $\hat{\beta}$ is the ordinary least squares estimate of the parameter that reflects the influence of time.

Note that the least squares approach makes no assumption about the distribution of the conditional mean of the dependent variable. In particular, the conditional mean of y need not be normally distributed for the ET to deliver meaningful estimates.

Due to the loss ratio being equal to the product of frequency and the respective severity, the following relation holds for the CAGR forecasts:

$$e^{\hat{\beta}_{\text{loss ratio}}} - 1 = \left(e^{\hat{\beta}_{\text{frequency}}}\right) \cdot \left(e^{\hat{\beta}_{\text{severity}}}\right) - 1 \ . \tag{3}$$

In NCCI ratemaking, the ET is typically applied to the latest 5, 8, and 15 data points. The forecast growth rate is judgmentally determined and may be interpreted as a weighted average of the resultant three CAGR. By applying judgment, it is possible to factor in influences that lie outside the model, such as additional trend estimates or changes in economic conditions.

3.2 Bayesian Trend Selection (BTS)

In what follows, the nature of the BTS approach is illustrated for the frequency trend selection; the loss ratio trend selections are implemented accordingly and independently. The CAGR for the severity rates of growth are then backed out of the CAGR of frequency and the respective loss ratio. Loss ratio modeling is given preference over the modeling of severity, since loss ratio rates of growth are generally smoother than severity (and frequency) rates of growth. This is because the numerator of frequency equals the denominator of severity and, as a result, some of the variation of the frequency and the severity series cancels out at the level of the loss ratio.

In a first step, for a given state, three values for the CAGR of frequency are estimated using the 5-point, 8-point, and 15-point ET. A holdout period of three years applies to each of the three ET specifications, which means that the time windows of the ET models shift into the past by three years. This estimation process is performed for the S most recent ratemaking data sets (inclusive of the ongoing ratemaking season). Currently, S equals three but may be allowed to increase with future ratemaking seasons. By employing such a holdout period, the CAGR forecasts can be compared to the realized values three years hence. The three-year holdout period is motivated by the length of the trend period in aggregate ratemaking, which typically amounts to slightly more than three years, as mentioned. (Note that the S holdout periods are overlapping, since they are three-year time windows that move in increments of one year.) The realized CAGR three years hence that is compared to the ET forecast defined in Equation (2) is calculated as $\sqrt[3]{Y_T / Y_{T-3}} - 1$, where Y is the observed value of frequency.

In the BTS framework, the forecasts and the realized values three years hence are related to each other by means of a double exponential distribution. In this model, there are S observations, since every frequency data set generates one data point. The three ET forecasts compete for being considered as the location parameter of the double-exponential distribution. This selection process is accomplished by means of a categorical distribution with a Dirichlet prior.

In relating forecasts to realized values, the double exponential distribution minimizes the sum of absolute forecast errors, as opposed to the sum of squared forecast errors. In the context of the BTS, minimizing the sum of squared forecast errors would be equivalent to minimizing the root mean squared forecast error, which Armstrong [1] has shown to be "unreliable, especially if the data might contain mistakes or outliers." Instead, by minimizing the sum of absolute forecast errors, the BTS is robust to outliers.

The BTS model reads

$$y_s \sim \text{Dexp}\left(\mu_s, \tau\right),\ s = 1,\ldots,S\ , \tag{4}$$

$$\mu_s = \vec{y}_s^{\,f}\left[\lambda\right],\ s = 1,\ldots,S\ , \tag{5}$$

$$\lambda \sim \text{Cat}\left(\vec{p}\right), \tag{6}$$

$$\vec{p} \sim \text{Dirch}\left(\vec{\alpha}\right), \tag{7}$$

$$\vec{\alpha} = \left(1,1,1\right), \tag{8}$$

$$\tau \sim \text{Ga}\left(0.001, 0.001\right), \tag{9}$$

$$\sigma = \sqrt{2} / \tau\ , \tag{10}$$

where the suffix $s=1,\ldots,S$ indicates the respective ratemaking data set. The most recent data set pertains to the ongoing ratemaking season, whereas the remaining $S-1$ data sets originate from ratemaking seasons of the immediate past.

Equation (4) presents the double exponential distribution, which defines the likelihood of the Bayesian model; as mentioned, a double exponential likelihood minimizes the sum of absolute errors, which are the S forecast errors from as many holdout periods.

The model selection mechanism comprises Equations (5) through (8). In Equation (5), the conditional mean of the double exponential distribution is chosen from the vector of the three competing CAGR forecasts $\vec{y}^f$ that arise from as many ET models. The parameter λ, which takes on integer values between (and inclusive of) unity and three, selects the element of the vector $\vec{y}^f$ that is to serve as the location parameter of the double exponential distribution. This parameter λ is generated by a categorical distribution (Equation [6]) with a Dirichlet prior (Equation [7]). The parameters of the Dirichlet distribution (Equation [8]) are identical, thus affording each of the three ET models equal prior probability of being the "true" model.

Equation (9) depicts the gamma prior for the scale parameter of the double exponential distribution. Equation (10) denotes the standard deviation of the double exponential errors, as an informational item.

The model is estimated by means of MCMC. Every iteration of the Markov chain generates a draw from the categorical distribution. The relative frequency with which a model is selected equals the posterior probability of this model representing the "true" data-generating process. The CAGR for frequency (and, similarly, for the loss ratios) is then computed as a weighted average of the CAGR values of the three ET models estimated without a holdout period, using the latest available observations. The weights of this average are the relative frequencies of the values assumed by the parameter λ. The CAGR values for the severities are backed out of the CAGR values for frequency and the loss ratios according to Equation (3).

Finally, in a sensitivity analysis, the BTS is re-estimated S times, each time leaving out one of the S data sets. This leave-out-one cross-validation (CV) generates a range of forecasts that do not necessarily include the original forecast (that was obtained when including all S data sets). Further, this range of forecasts can be comparatively narrow, because only the posterior probabilities (of the ET models) are cross-validated, not the ET estimates themselves. A CV range that lies below the BTS forecast indicates that the risk is skewed toward the downside, thus implying that the decision maker should be cautious about selecting a value greater than the BTS forecast. Similarly, a CV range that lies entirely above the BTS forecast indicates a risk of the forecast being too low.

The JAGS code of the BTS outlined in Equation (4) through Equation (10) is documented in Appendix 8.1.

4. MODEL VALIDATION

The BTS is validated by means of comparing its forecast performance to the 5-point, 8-point, and 15-point ET. The forecasts of all four models are benchmarked to the random walk. If a time series follows a random walk, then the best forecast for any future time period is the latest observed value.

Implementing the BTS requires S data sets of a contiguous time period—the ongoing and past $S-1$ ratemaking seasons. As discussed, S is set to three but may be increased as the number of available data sets grows with future ratemaking seasons.

Validating the BTS necessitates a total of $S+3$ data sets. For instance, applying the BTS during the 2012 ratemaking season (where the latest available policy year is 2010) necessitates data from the 2011 and 2010 ratemaking seasons as well. Model validation requires, in addition, data sets from the ratemaking seasons 2007 through 2009.

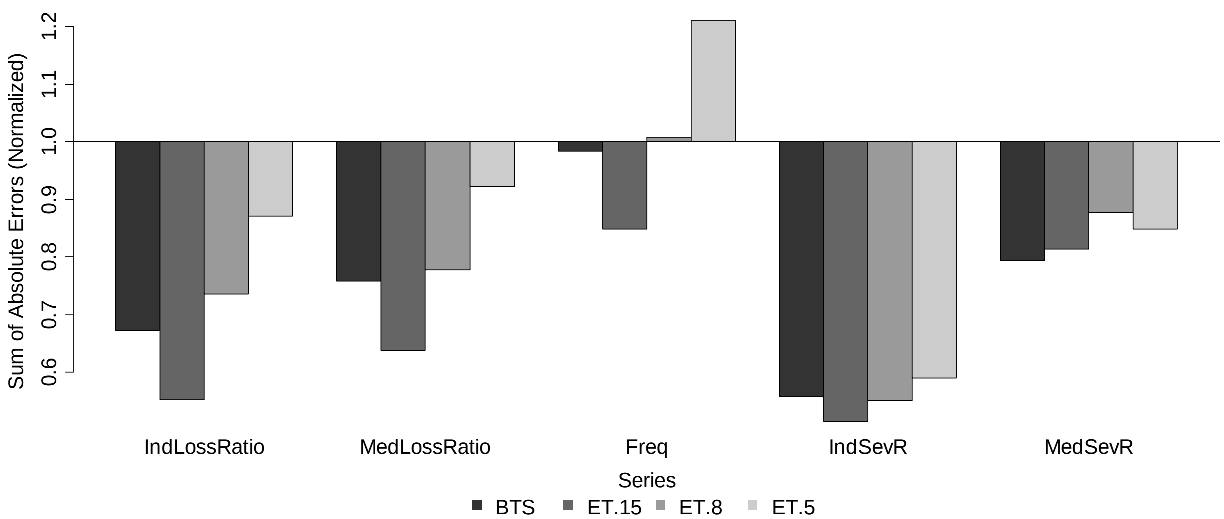
In a first step, the BTS is applied retrospectively to the 2008 ratemaking season, thus making use of the 2008, 2007, and 2006 ratemaking data sets. Then, in a second step, the three-year CAGR forecasts associated with the three ET models and the BTS are compared to the CAGR that were observed for the time period 2009 through 2011. For this purpose, the sums of absolute forecast errors are computed across the set of analyzed states and normalized by the forecast error associated with the random walk. (For every state, there is one three-year forecast for each model and data series; for instance, there is one BTS forecast for the three-year CAGR of frequency.)

The model validation makes use of 29 states. States with an insufficient number of observations or insufficient number of past data sets could not be considered. Such data limitations may arise in states that switched from accident year to policy year information in the recent past or in states that have only recently been added to the data collection process.

Chart 1 depicts the absolute forecast errors, summed up across the 29 states and normalized by the forecast error of the random walk. Values greater than unity imply that the respective model underperforms the random walk, thus adding no information that is not yet comprised in the latest observed data point. As judged by the loss ratio forecasts, the BTS is the second-best model, behind the 15-point ET. As an informational item, the BTS comes in first for medical severity, second for frequency, and third for indemnity severity, each time outperforming the random walk.

The performance of the ET models relative to the random walk is most interesting. On one hand, it appears that the frequency rate of growth is highly nonstationary, given the performance of the random walk compared to the BTS and the ET models. On the other hand, the only model that beats the random walk is the 15-point ET, which is the model that is most appropriate when the frequency growth rate is stationary.

Chart 1: Sum of Absolute Forecast Errors, Validated on the 2011 Ratemaking Season



5. ROBUSTNESS

The concept of model robustness has been pioneered by Lars Hansen and Thomas Sargent. According to this principle, a decision rule should be designed such that it "works well enough (is robust) despite possible misspecification of its model" (Ellison and Sargent [3]). In particular, the performance of the model should be adequate even in a worst-case scenario. For an extensive treatment of the concept of robustness, see Hansen and Sargent [5].

In the context of the BTS, robustness has two dimensions. First, the performance of the BTS in forecasting the CAGR of the loss ratios should be adequate for any state. Specifically, for the loss ratios, the largest absolute forecast error of the BTS should be smaller than the largest absolute forecast error within the set of four alternative models, that is, the three ET models and the random walk. Chart 2 shows that the BTS satisfies this condition although, for the indemnity loss ratio, the advantage of the BTS over the worst-performing ET (5-point) and the random walk is slender. Regarding the ratio that defines the normalized largest absolute forecast error in Chart 2, numerator (model) and denominator (random walk) may not refer to the same state.

Second, the performance of the BTS should be adequate in an environment where the nature of the time series changes. As shown by Schmid [8], the data-generating process of the injury (and illness) incidence rate in the manufacturing industry has changed over the course of the 20th century from one of high variance and small serial correlation in the error term to one with low variance and high persistence in the error. The structural break in the degree of serial correlation occurred in the early 1960s and coincided with the end of a two-decade decline in the variance. In part, the change in the nature of the manufacturing incidence rate may be related to changes in the data collection process, although it is worthy of note that OSHA (Occupational Safety and Health Administration) did not become effective until 1971.

Chart 2: Maximum Absolute Forecast Error, Validated on the 2011 Ratemaking Season

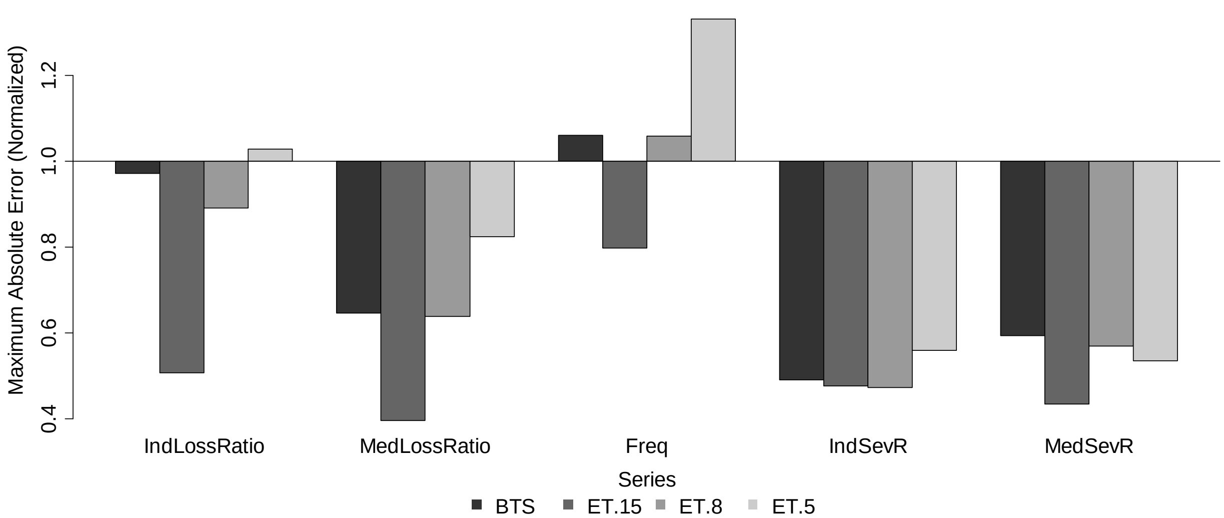


Chart 3 depicts the incidence rate of workplace injuries (and illnesses) for the manufacturing industry; the series runs from 1926 through 2010. The incidence rate is on a logarithmic scale, which implies that a straight line signifies a constant rate of growth. The gray vertical bars indicate economic recessions, as defined by the NBER (National Bureau of Economic Research). With OSHA becoming operational in 1971, the incidence rate started including work-related illnesses. For methodological details on the incidence rate and for data sources, see Appendix 8.2.

Chart 3: Manufacturing Injury (and Illness) Incidence Rate, 1926–2010, per 100 FTE Employees

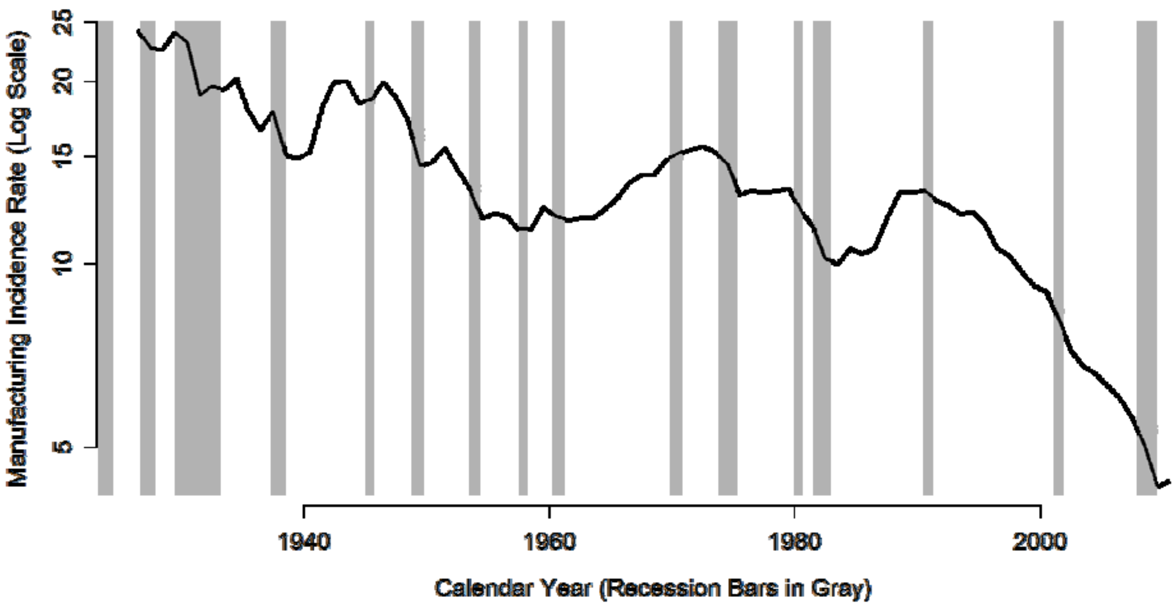


Chart 4 presents the logarithmic rates of growth (that is, first differences in natural logarithms) of the manufacturing injuries (and illnesses) incidence rate; again, the gray bars indicate recessions. The decline in variance and the increase in persistence of variations around zero (as a proxy for the negative, possibly time-varying mean) are apparent. Starting in the early 1960s, frequency increases are more likely to be followed by further increases than they were in the first half of the 20th century; likewise, frequency decreases show more persistence as well. Clearly, the nature of the time series changed over time.

In the context of the manufacturing incidence rate, a defining characteristic of a robust decision rule is its adequate forecast performance in both the early high-variance low-persistence regime and the later low-variance high-persistence regime. In what follows, the three ET models and the BTS are applied to the full series (1926–2010); the early regime, inclusive of the transition period (1926–1964); and the second regime.

Similar to studying model robustness for ratemaking data, the 5-point, 8-point, and 15-point ET models are applied with a holdout period of three years, using time windows that roll forward at increments of one year. The three forecast errors that are associated with the three neighboring time windows are the basis for the model selection of the BTS. The BTS forecast is then calculated as a weighted average of the 5-point, 8-point, and 15-point ET forecasts, where the ET models have been run with data leading up to the start of the forecasting horizon (and, hence, without the three-year holdout period). The resulting BTS forecast error is compared to the forecast errors of the 5-point, 8-point, and 15-point ET that enter the weighted average of the BTS forecast.

Chart 4: Manufacturing Injury (and Illness) Incidence Rate, Log Growth Rate, 1927–2010

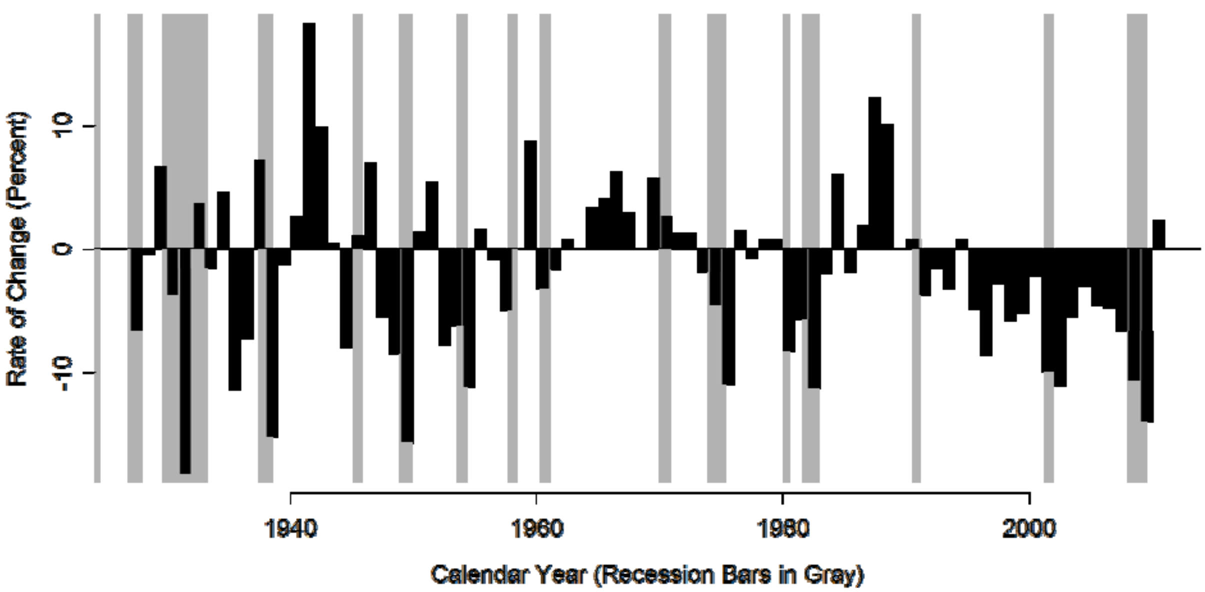


Chart 5 depicts for the full manufacturing incidence rate series the sum of absolute errors of the three ET models and the BTS, normalized by the sum of absolute forecast errors of the random walk. The BTS and the 5-point ET perform about equally well, outperforming the random walk and the 15-point and 8-point ET models, although the performance differences to the 8-point ET is slim.

Chart 5: Manufacturing Incidence Rate, Sum of Absolute Errors, Normalized, 1926–2010

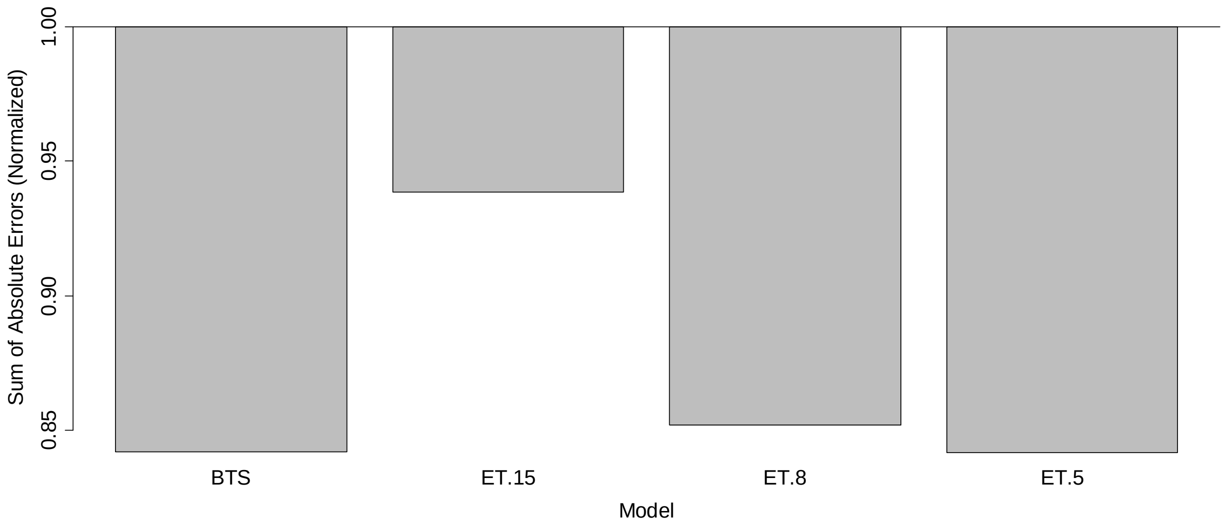


Chart 6 presents the analysis of the forecast error for the first, high-variance low-persistence regime and the subsequent transition period (1926–1964). Here again, the BTS and the 5-point ET perform about equally well, beating the random walk and the 15-point and 8-point ET models.

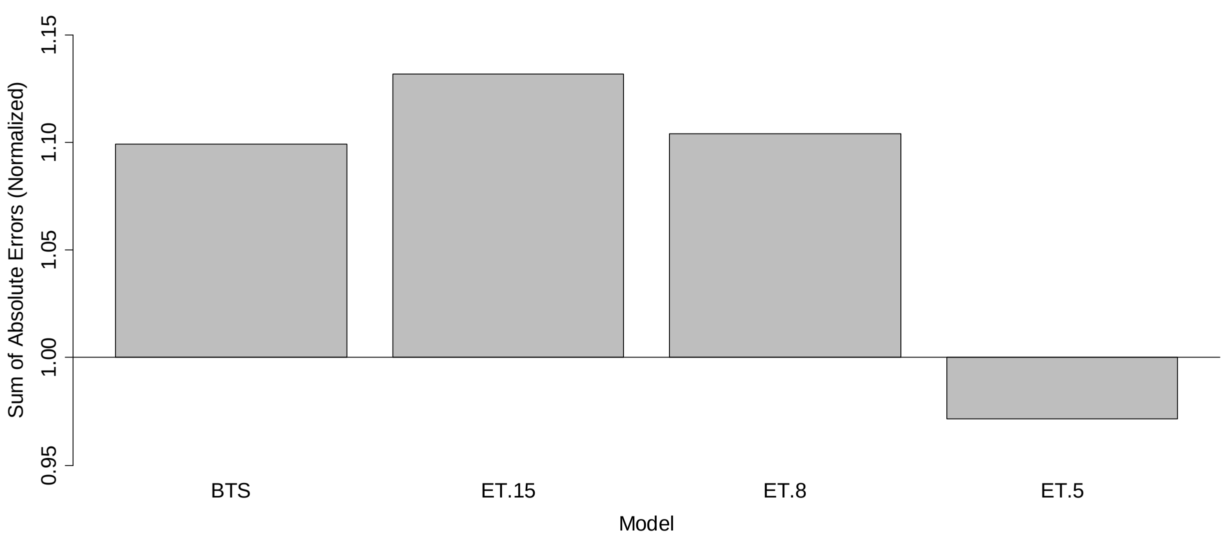
Chart 6: Manufacturing Incidence Rate, Sum of Absolute Errors, Normalized, 1926–1964

Finally, Chart 7 offers the analysis for the second regime (1965–2010). In this low-variance high-persistence setting, only the 5-point ET outperforms the random walk. The BTS comes in second, although the performance differences to the 8-point and 15-point ET models are slender.

Whereas, for the ratemaking data, the 15-point ET emerged as the highest performing model, for the manufacturing incidence rate, it is the 5-point ET. At the same time, the BTS outperforms the 15-point ET for the manufacturing series and beats the 5-point ET on the ratemaking data. The BTS performs well in both environments. Although the BTS never emerges as the winner, it proves to be the most robust model.

The BTS delivers a robust decision rule due to its ability to draw on any of the ET models; the BTS makes its model selection based on recent forecast performance. As the nature of a time series changes, the BTS adapts and leans toward the most suitable model. At the same time, the BTS is conservative in its model selection in that it never unequivocally adopts the highest performing model (which explains why it underperforms the 15-point ET on the loss ratios and the 5-point ET on the manufacturing incidence rate series).

Chart 7: Manufacturing Incidence Rate, Sum of Absolute Errors, Normalized, 1965–2010



6. APPLICATION

The BTS is applied to an unidentified state. The data originates from the 2011 NCCI ratemaking season.

There are two types of charts. The first type of chart displays the observed growth rates, the 5-point, 8-point, and 15-point ET estimates, and the BTS forecast. To avoid clutter, the CAGR values associated with the three ET models are indicated by boxes (instead of lines), where the horizontal distances between the boxes indicate the time intervals associated with 4, 7, and 14 rates of growth. Further, there are the CV forecast ranges, which may be close to vanishing in some of the charts.

The second type of chart displays the posterior probabilities related to the three ET models. As mentioned, these posterior probabilities are the relative frequencies with which any of the three ET models was selected during the BTS estimation process. This second type of chart is available only for frequency and the loss ratios—this is because the CAGR forecasts for the severities are backed out of the CAGR forecasts for frequency and the respective loss ratio.

Chart 8 depicts the analysis of the indemnity loss ratio, which displays a clear upward drift. The 5-point ET estimate exceeds the 8-point ET estimate, which, in turn, exceeds the 15-point ET estimate. As a consequence of this apparent nonstationarity (that is, time-variation in the mean rate of growth), the BTS affords high probabilities to ET models that do not reach deep into the past (see Chart 9).

Chart 10 presents the results for the medical loss ratio. The growth rate of this time series is comparatively stable, as indicated by the close proximity of the 5-point, 8-point and 15-point ET estimates to each other. As a result, the BTS leans toward stationarity (that is, the presumption of a time-invariant mean rate of growth), thus preferring estimates based on long time series to estimates that rely on only the most recent data points (see Chart 11). Clearly, if a time series is stationary, the CAGR should be calculated from a high number of data points, since this increases the accuracy of the estimate.

Chart 12 displays the frequency analysis. The rate of frequency growth appears to drift up, as indicated by the observation that the CAGR forecasts of the 8-point and 5-point ET models exceed the estimate of the 15-point ET. The BTS forecast is close to the 5-point and 8-point ET estimates, as implied by the high posterior probabilities afforded to the two models (see Chart 13).

Chart 8: Indemnity Loss Ratio, Growth Rates, 1996–2009

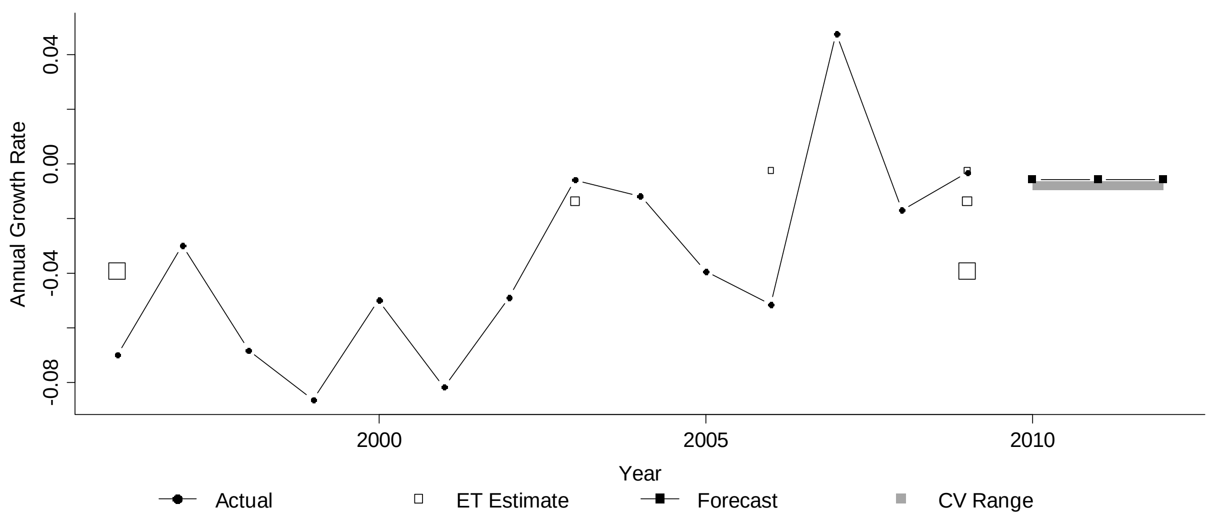


Chart 9: Indemnity Loss Ratio, Posterior ET Probabilities

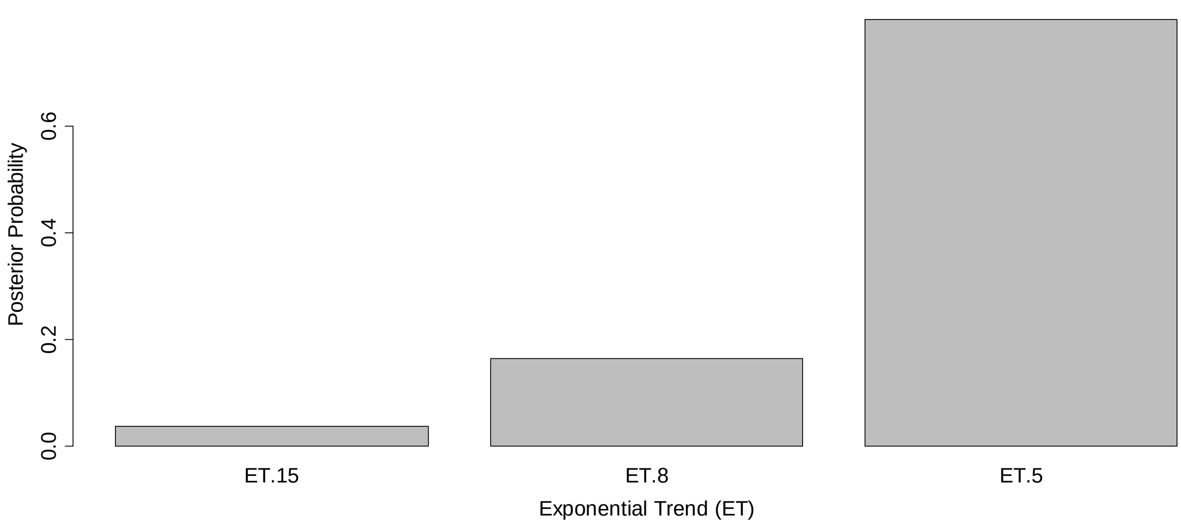


Chart 10: Medical Loss Ratio, Growth Rates, 1996–2009

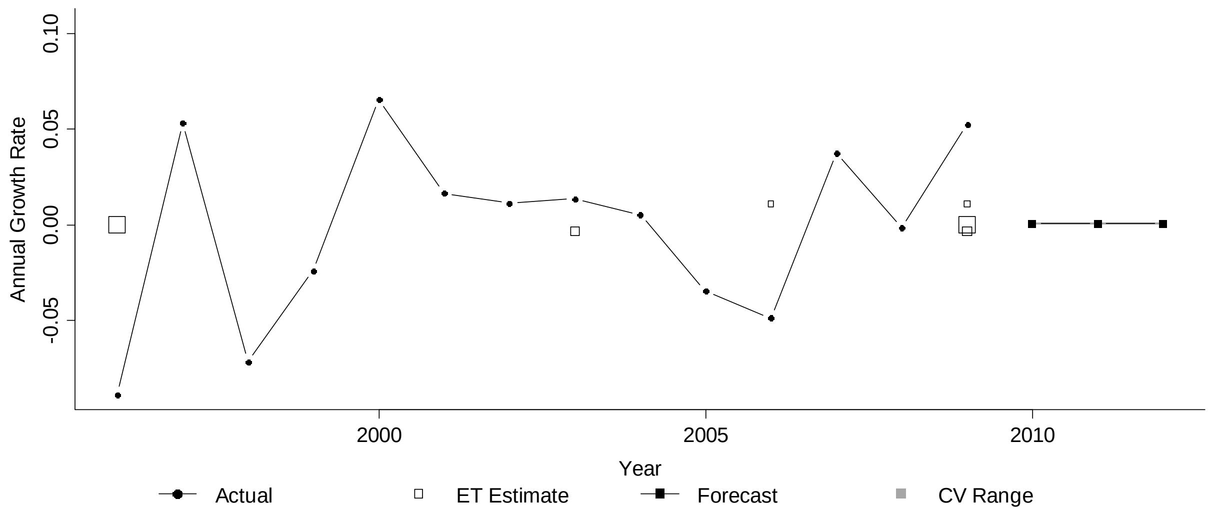


Chart 11: Medical Loss Ratio, Posterior ET Probabilities

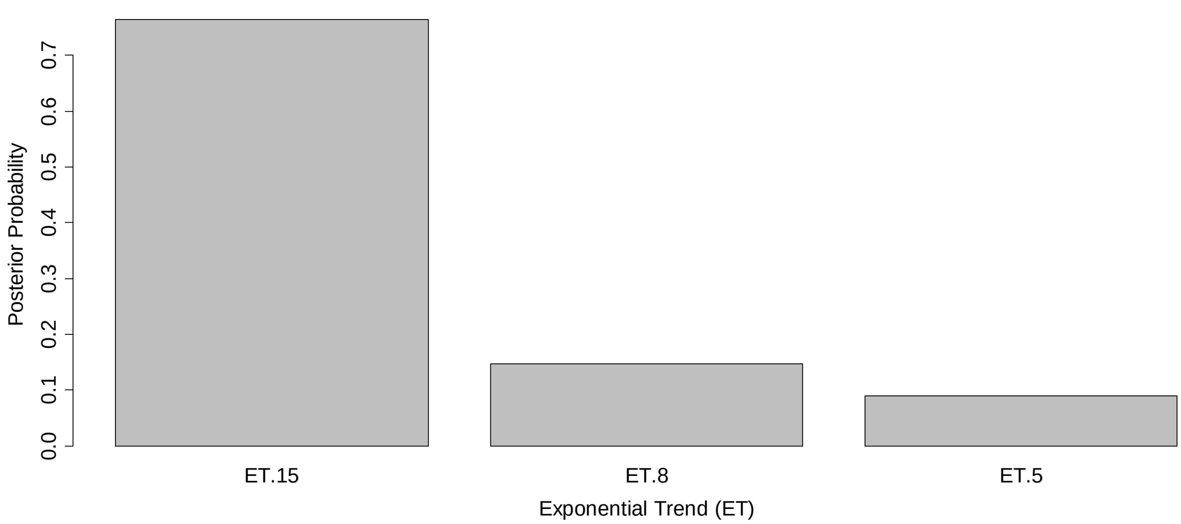


Chart 12: Frequency, Growth Rates, 1996–2009

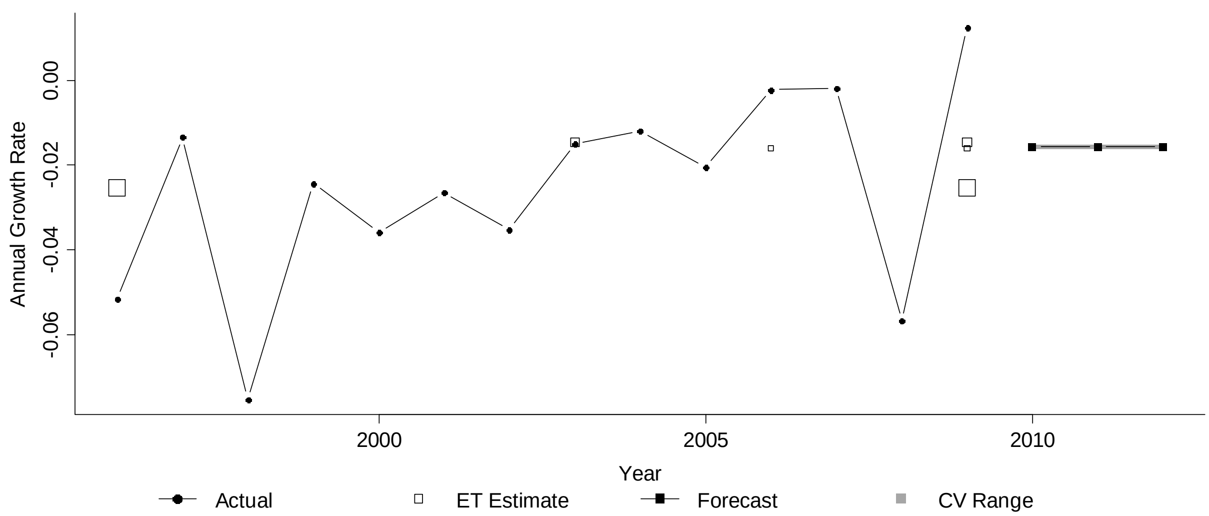
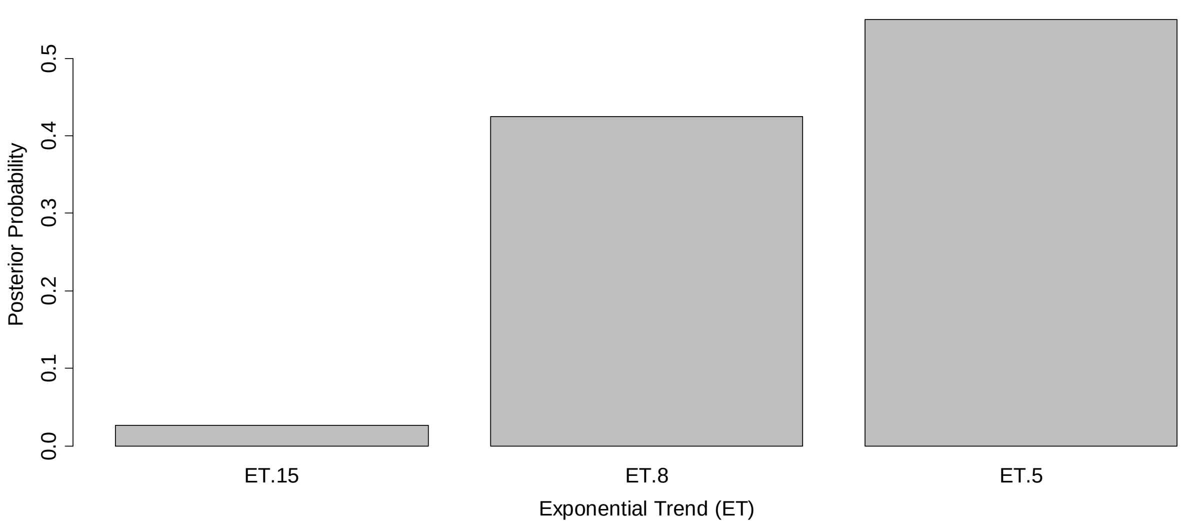


Chart 13: Frequency, Posterior ET Probabilities



Charts 14 and 15 display the findings for the severities. As mentioned, the ET estimates and the BTS forecast for the severities are backed out of the CAGR values of frequency and the pertinent loss ratio. Chart 14 reveals the upward drift in indemnity severity that contributes to the drift in the severity loss ratio. Chart 15, on the other hand, indicates that the growth rate of the medical loss ratio is highly stationary

Chart 14: Indemnity Severity, Growth Rates, 1996–2009

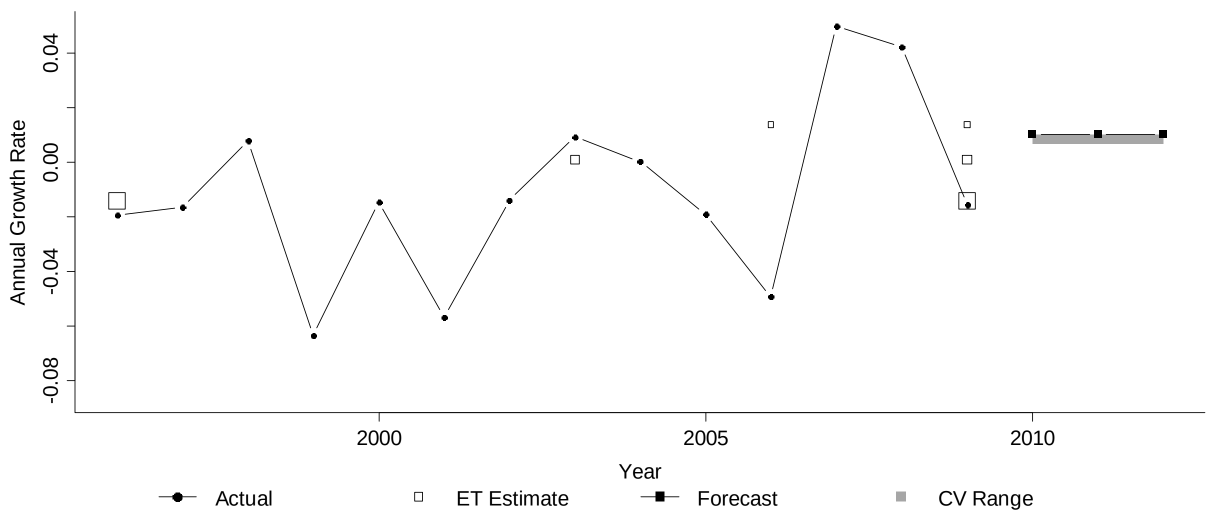
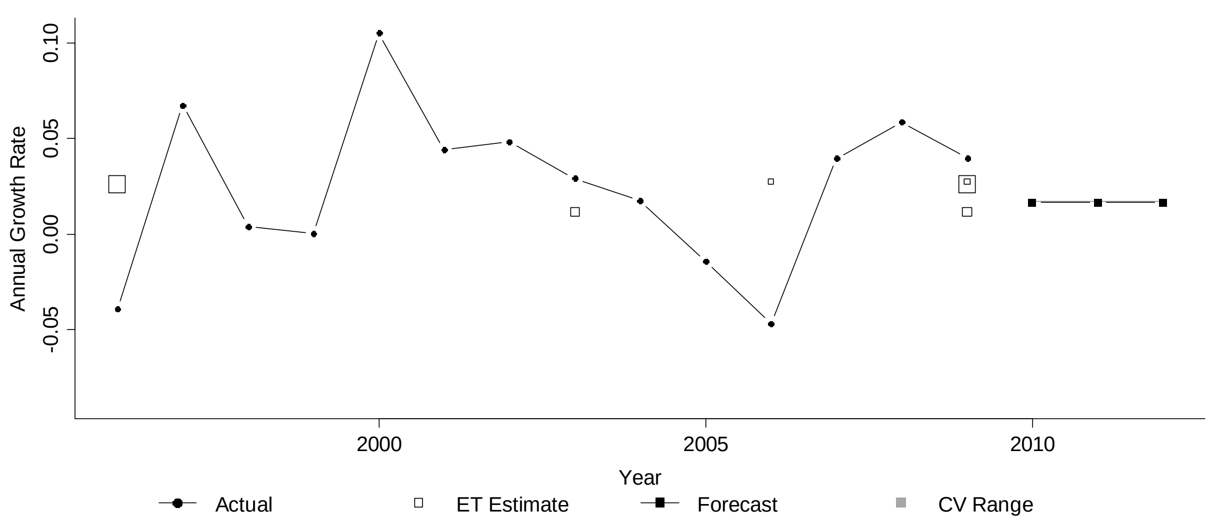


Chart 15: Medical Severity, Growth Rates, 1996–2009



7. CONCLUSION

The BTS is an attempt to formalize the trend selection process in NCCI aggregate ratemaking. The information processed by the BTS is confined to past forecast errors, which represents both a strength and a weakness of this model. On one hand, the BTS is not prone to biases in human decision making. On the other hand, the model is not capable of processing information that is not incorporated in the data, such as changes in the economic or legislative environments that occurred between the end of the experience period and the time of decision making.

Given the relative shortness of the loss ratio data series available in NCCI aggregate ratemaking, and considering the high degree of volatility of these series, the use of covariates bears the risk of fitting to noise; such overfitting can cause considerable forecast errors. Further, economic conditions tend to mean-revert; for instance, economic recessions usually last less than a year. The comparatively long trend period of slightly more than three years typically suffices for average economic conditions to reestablish themselves. Thus, mean reversion limits the loss of information associated with the absence of macroeconomic covariates as long as the decision maker relies on long-term averages.

A major strength of the BTS is its robustness. Although the BTS underperforms the 15-point ET on recent ratemaking data sets, the BTS performs well in an environment where the nature of the data-generating process is changing. Because there is always a degree of uncertainty surrounding the process that generates the growth rates of the loss ratios (or alternatively, frequency and the severities), the BTS emerges as a robust decision rule.

The strength of the BTS becomes apparent when, for instance, in a given state, the growth rate of frequency repeatedly falls short of its long-term average. Such was the case during the boom in the housing market in the early 2000s, as documented in Schmid [6]. On one hand, the 15-point ET generates the smallest forecast errors in general; on the other hand, observing the frequency rate of growth falling short of trend for many years running poses a challenge to the decision maker. Although these frequency growth rates ultimately mean-reverted, the time it takes for the trend to reestablish itself in the observed rates of growth may be longer than a decision maker is prepared to tolerate.

8. APPENDIX

8.1 JAGS Code (BTS)

```
## Bayesian Trend Selection (BTS)
model
{
  for(series.i in 1:n.series){
    for(year.i in 1:n.model.years){
      actual[series.i,year.i] ~ ddexp(mu[series.i,year.i],tau[series.i])
      mu[series.i,year.i] <- exp.trend[series.i,year.i,model.id[series.i]]
    }

    tau[series.i] ~ dgamma(0.001,0.001)
    sigma[series.i] <- sqrt(2)/tau[series.i]

    model.id[series.i] ~ dcat(model.id.p[series.i,1:n.exp.trend])
    model.id.p[series.i,1:n.exp.trend] ~ ddirch(lambda[series.i,])
  }
}
```

8.2 BLS Manufacturing Injury and Illness Incidence Rate, 1926–2010

The long series of manufacturing injury and illness rates (1926–2010) is inspired by research at the Federal Reserve Bank of Dallas. In its Annual Report, authored by Michael Cox and Richard Alm (dallasfed.org/assets/documents/fed/annual/2000/ar00.pdf), the Bank published a series of injury rates per 1,000 full-time workers in manufacturing for the period 1926 through 1999 (page 8).

The Dallas Fed series of manufacturing injury and illness incidence rates runs from 1926 through 1999 and draws on four sources:

- *Historical Statistics of the United States: Colonial Times to 1970*, Census Bureau, 1975, www.census.gov/prod/www/abs/statab.html; incidence rates are available from 1926 through 1970
- *Statistical Abstract of the United States*, various years, www.census.gov/prod/www/abs/statab.html
- Webster [11]; incidence rates are available from 1977 through 1997
- BLS

From 1926 through 1970, the rate is the average number of disabling workplace injuries per million man hours worked; effective 1958, the industry definition was revised to conform to the 1957 edition of the Standard Industrial Classification (SIC) Manual. (The SIC was developed in the 1930s; www.census.gov/epcd/www/sichist.htm.)

From 1972 on, the number of injuries and illnesses per 100 full-time workers is computed as (N/EH) × 200,000, where N is the number of injuries and illnesses and EH is the total hours worked by all employees during the calendar year; 200,000 is the base for 100 equivalent full-time workers (working 40 hours per week, 50 weeks per year).

The *Statistical Abstract of the United States* (1976) reports the old incidence series through 1970 and the new series for 1972. The 1971 incidence rate in our data set, which equals 15.4, was obtained by means of linear interpolation, since no incidence rate is reported by the *Statistical Abstract of the United States*, issues 1972 through 1974, 1976 through 1977. The 1970 number equals 15.2, and the 1972 number equals 15.6. Although, due to methodological differences, the new series is not comparable to the old series, the structural break between the two series is insignificant. It is worthy of note that the record keeping and data reporting process for the incidence rate changed significantly with OSHA (Occupational Safety and Health Administration, www.osha.gov/) becoming operational in 1971.

Starting in 2003, the manufacturing industry is no longer defined by the Standard Industrial Classification Manual, www.osha.gov/pls/imis/sic_manual.html, but instead is based on the newly created NAICS, www.census.gov/eos/www/naics/. Prior to adopting NAICS, the definition of manufacturing had been updated periodically to conform to SIC Manual revisions.

Acknowledgment

Thanks to Eric Anderson and Sikander Nazerali for their research assistance.

9. REFERENCES

[1] Armstrong, J. Scott, "Evaluating Forecasting Methods," in: J. Scott Armstrong, ed., *Principles of Forecasting: A Handbook for Researchers and Practitioners*. Norwell (MA): Kluwer Academic Publishers, **2001**, 365–382.

[2] Brooks, Ward, "A Study of Changes in Frequency and Severity in Response to Changes in Statutory Workers Compensation Benefit Levels," Casualty Actuarial Society *Proceedings*, **1999**: Vol. LXXXVI, No. 164 and 165, www.casact.org/pubs/proceed/proceed99/99046.pdf.

[3] Ellison, Martin, and Thomas J. Sargent, "A Defence of the FOMC," **2010**, July 10, files.nyu.edu/ts43/public/research/fomc_tom_5.pdf.

[4] Evans, Jonathan P., and Frank Schmid, "Forecasting Workers Compensation Severities and Frequency Using the Kalman Filter," Casualty Actuarial Society *Forum*, **2007**: Winter, 43–66, www.casact.org/pubs/forum/07wforum/07w49.pdf.

[5] Hansen, Lars Peter, and Thomas J. Sargent, *Robustness*. Princeton (NJ): Princeton University Press, **2008**.

[6] Schmid, Frank, "Loss Cost Components and Industrial Structure," **2011**, National Council on Compensation Insurance, Inc., www.ncci.com/documents/Loss_Cost_Components_and_Industrial_Structure-2011.pdf.

[7] Schmid, Frank, "Statistical Trend Estimation with Application to Workers Compensation Ratemaking," Casualty Actuarial Society *Forum*, **2009**: Winter, 330–40, www.casact.org/pubs/forum/09wforum/schmid.pdf.

[8] Schmid, Frank, "Workplace Injuries and Job Flows," **2009**, National Council on Compensation Insurance, Inc., www.ncci.com/Documents/WorkplaceInjuries-0709.pdf.

[9] Taylor, Shelley E., "The Availability Bias in Social Perception and Interaction," in: Daniel Kahneman, Paul Slovic, and Amos Tversky, eds., *Judgment under Uncertainty: Heuristics and Biases*. Cambridge (UK): Cambridge University Press, **1982**, 190–200.

[10] Tversky, Amos, and Daniel Kahneman, "Judgment under Uncertainty: Heuristics and Biases," *Science* 185, **1974**, 1124–1131, reprinted in: Daniel Kahneman, Paul Slovic, and Amos Tversky, eds., *Judgment under Uncertainty: Heuristics and Biases*. Cambridge (UK): Cambridge University Press, **1982**, 3–20.

[11] Webster, Timothy, "Work-related Injuries, Illnesses and Fatalities in Manufacturing and Construction," *Compensation and Working Conditions*, BLS, **1999**: Fall, 34–37, www.bls.gov/opub/cwc/archive/fall1999brief3.pdf.

Abbreviations and Notations

BLS	Bureau of Labor Statistics
BTS	Bayesian Trend Selection
CAGR	Compound Annual Growth Rate
CV	Cross-Validation
ET	Exponential Trend
FTE	Full-Time Equivalent
MCMC	Markov Chain Monte Carlo Simulation
NAICS	North American Industry Classification System
NBER	National Bureau of Economic Research
NCCI	National Council on Compensation Insurance, Inc.
OSHA	Occupational Safety and Health Administration
RMSE	Root Mean Squared Error
SIC	Standard Industrial Classification

Biographies of the Authors

Frank Schmid, Dr. habil., was, at the time of writing, a Director and Senior Economist at the National Council on Compensation Insurance, Inc.
C**hris Laws** is a Research Consultant at the National Council on Compensation Insurance, Inc.
Matthew Montero is a Senior Actuarial Analyst at the National Council on Compensation Insurance, Inc.

Indemnity Benefit Duration and Obesity

Frank Schmid, Chris Laws, and Matthew Montero

Abstract
Motivation. There is mounting evidence of obesity contributing to the cost of workers compensation. Longitudinal studies by Duke University (Østbye, Dement, and Krause [5]) of its own employees—and by Johns Hopkins University (Pollack et al. [7]) of employees of a multi-site U.S. aluminum manufacturing company—point to substantially higher odds of injury for workers in the highest obesity category. Further, a 2011 Gallup survey (Witters and Agrawal [9]) of 109,875 full-time employees found that obese employees account for a disproportionately high number of missed workdays, thus causing a significant loss in economic output. Finally, an NCCI study (Laws and Schmid [4]) of workers compensation claims established that where claimants are assigned a comorbidity code indicating obesity, the medical costs of the claim are a multiple of what is observed otherwise. In the following study, using a methodology similar to the one employed by Laws and Schmid [4], but accounting for a possible immortal time bias, it is shown that the indemnity benefit duration of claimants with an obesity-related comorbidity indicator is a multiple of what is observed for comparable non-obese claimants.
Method. The study makes use of a matched-pairs research design, where obese claims are matched with comparable non-obese claims in the data set. Exact matching applies to all claim characteristics, except age at injury, where proximity matching is employed. The set of matched pairs is then analyzed using a semiparametric Bayesian Weibull proportional hazard model, the nonparametric component of which accounts for the possible nonlinear influence of age. Aside from age, an indicator variable signifying obesity is the only covariate in the model—this is because net of these two covariates (and duration, which serves as the dependent variable), the claims within each set of matched pairs are identical. The model is estimated by means of MCMC (Markov chain Monte Carlo simulation).
Results. The study shows that, based on Temporary Total and Permanent Total indemnity benefit payments, the duration of obese claimants is more than five times the duration of non-obese claimants, after controlling for primary ICD-9 code, injury year, U.S. state, industry, gender, and age. When Permanent Partial benefits are counted toward indemnity benefit duration as well, this multiple climbs to more than six.
Availability. The model was implemented in R (cran.r-project.org/) using the sampling platform JAGS (Just Another Gibbs Sampler, mcmc-jags.sourceforge.net/). JAGS was linked to R by means of the R package rjags (cran.r-project.org/web/packages/rjags/index.html).

Keywords. Obesity, Duration, Proportional Hazard Model, Semiparametric Model, Workers Compensation

1. INTRODUCTION

There is mounting evidence of obesity contributing to the cost of workers compensation. Longitudinal studies by Duke University (Østbye, Dement, and Krause [5]) of its own employees—and by Johns Hopkins University (Pollack et al. [7]) of employees of a multi-site U.S. aluminum manufacturing company—point to substantially higher odds of injury for workers in the highest obesity category. Further, a 2011 Gallup survey (Witters and Agrawal [9]) found that obese employees account for a disproportionately high number of missed workdays, thus causing a significant loss in economic output. Finally, an NCCI study (Laws and Schmid [4]) of workers compensation claims found that claimants with a comorbidity code indicating obesity experience medical costs that are a multiple of what is observed for comparable non-obese claimants.

In contrast to the mentioned studies, there is research that finds no evidence for a link between work-related injuries and obesity. Shaw et al. [8] "assess the effect of body mass index (BMI) on pain and function outcomes in the acute and sub-acute stages of work-related low back pain (LBP)," finding that "BMI is not a useful prognostic factor during the acute and sub-acute stages of work-related LBP." Similarly, Pollack and Cheskin [6], who review the early literature on the relation between workplace injuries and obesity, report only weak evidence for such a link.

In the following study, using a methodology similar to the one employed by Laws and Schmid [4], but accounting for a possible immortal time bias, it is shown that the indemnity benefit duration of claimants with an obesity-related comorbidity indicator is a multiple of what is observed for comparable non-obese claimants. The study makes use of a matched-pairs research design, where obese claims are matched with comparable non-obese claims in the data set. Exact matching applies to all claim characteristics, except age at injury, where proximity matching is employed. The set of matched pairs is then analyzed using a semiparametric Bayesian Weibull proportional hazard model, the nonparametric component of which accounts for the possible nonlinear influence of age. Aside from age, an indicator variable signifying obesity is the only covariate in the model—this is because net of these two covariates (and duration, which serves as the dependent variable), the claims within each set of matched pairs are identical. The model is estimated by means of Markov chain Monte Carlo simulation (MCMC). The study shows that, based on Temporary Total and Permanent Total indemnity benefit payments, the duration of obese claimants is more than five times the duration of non-obese claimants, after controlling for primary ICD-9 code, injury year, U.S. state, industry, gender, and age. When Permanent Partial benefits are counted toward duration as well, this multiple climbs to more than six.

1.1 Research Context

The Duke University study by Østbye, Dement, and Krause [5] is a comprehensive statistical analysis of the effect of obesity on the cost of workers compensation. This study makes use of a longitudinal data set, which was obtained by monitoring a cohort of 11,728 employees of Duke University and the Duke University Health System from January 1, 1997, through December 31, 2004. The cohort was defined by all employees who had at least one health risk assessment (HRA) during this time period; taking an HRA is voluntary and available to employees eligible for healthcare benefits. (Note that the number of members in the study may have shrunk over time due to employee termination or disability.) The members of this cohort were assigned to body mass index (BMI) categories based on the first HRA they participated in during the time of the study.

At the end of the eight-year time window, the number of workers compensation claims, the number of lost workdays, and the indemnity and medical costs related to workers compensation were tallied for each employee; then, this information was matched to the BMI category (and other characteristics) of the claimant. There are six BMI categories, ranging from underweight to recommended weight, overweight, and three classes of obesity. The highest level of obesity is class III, which comprises the morbidly obese, identified by a BMI of 40 or higher. The Duke University study finds that for the morbidly obese employees, the medical costs are 6.8 times the costs for employees of recommended weight. At the same time, an employee in this morbidly obese group is twice as likely to have a claim, while the number of lost workdays is almost 13 times higher. For obese classes II (BMI of at least 35 but less than 40) and I (BMI of at least 30 but less than 35), the medical costs per employee are (respectively) 3.1 and 2.6 times the medical costs for employees of recommended weight; the respective multiples for the number of claims read 1.9 and 1.5, while the respective multiples for the lost workdays are 8.3 and 5.3. (The numbers cited above rest on the bivariate analysis presented in Østbye, Dement, and Krause [5], Table 3.)

Another way of presenting the findings of the Duke University study is on an approximate per-claim basis. Transforming the medical costs per 100 full-time equivalent (FTE) employees into costs per claim based on the reported means shows that this amount is 3.4 (obesity class III), 1.7 (obesity class II), and 1.7 (obesity class I) times the magnitude recorded for employees of normal weight. The corresponding numbers for indemnity read 5.5 (obesity class III), 3.4 (obesity class II), and 2.9 (obesity class I). Finally, the number of lost workdays per claim for obesity classes III, II, and I are 6.4, 4.5, and 3.5 (respectively) times the amount for claimants of recommended weight. (Here, too, the numbers are based on the bivariate analysis presented in Table 3 of Østbye, Dement, and Krause [5].)

The Johns Hopkins University study (Pollack et al. [7]) offers no direct evidence of an effect of obesity on lost workdays. The authors analyze the distribution and odds of occupational injury among 7,690 workers at eight plants of a U.S. aluminum manufacturing company from January 1, 2002, through December 31, 2004. The study finds that the odds of injury in the highest obesity group, as compared to the ideal body weight, runs at 2.21; injuries to the leg or knee were particularly prevalent among this very obese group.

Finally, the Gallup survey (Witters and Agrawal [9]) focused on "unhealthy days" per month, which are then converted into missing workdays at a rate of 0.31 missed workdays per unhealthy day. The survey was conducted from January 2, 2011, through October 2, 2011, covering 109,875 full-time employees. Employees of normal weight and no chronic conditions establish the baseline rate; the chronic conditions included in the survey comprise a past heart attack, high blood pressure, high cholesterol, cancer, diabetes, asthma, depression, and certain types of recurring pain. The baseline rate of unhealthy days per month runs at 0.34. At 0.36, this number is slightly higher for the group of overweight or obese and no chronic conditions. Similarly, the number of unhealthy days per month for those with one or two chronic conditions is higher for the overweight and obese (at 1.08) than those of normal weight (1.07). Finally, for employees with three or more chronic conditions, the numbers of unhealthy days per month read 3.51 for the overweight and obese and 3.48 for those of normal weight. Although correlation does not establish causation, it is of interest that among the full-time working population, of the employees who have one or two chronic conditions, there are twice as many obese or overweight employees than there are employees of normal weight. When it comes to the group of employees with three or more chronic conditions, the number of obese or overweight employees exceeds those of normal weight by a factor of more than three.

1.2 Objective

The effect of obesity on indemnity benefit duration is studied using nonparametric and semiparametric statistical approaches. Indemnity benefit duration is measured in two alternative ways. Both statistical analyses account for interval-censoring and right-censoring of duration. The nonparametric analysis rests on Kaplan-Meier plots; the semiparametric model employs a partially linear regression in a Weibull proportional hazard framework.

1.3 Outline

Section 2 offers a detailed description of how the data set was compiled, how obesity was identified, and how indemnity benefit duration was calculated. This is followed by a discussion of challenges that arise from matching by age, and by descriptive information of the matched pairs that enter the analysis. Section 3 then offers a nonparametric analysis based on Kaplan-Meier plots and a semiparametric analysis using a partially linear Weibull proportional hazard model. Section 4 concludes. An appendix offers technical details on the immortal time bias correction, the choice of the primary ICD-9 code, the conversion of benefit time into calendar time for the waiting period, and the calculation of duration where there are multiple matches; further, the appendix displays the JAGS computer code for the Weibull proportional hazard model.

2. THE DATA

We use a large data set of workers compensation claims provided by a set of insurance companies. The data comprise records from 40 states (AK, AL, AR, AZ, CO, CT, DC, FL, GA, HI, IA, ID, IL, IN, KS, KY, LA, MD, ME, MI, MO, MS, MT, NC, NE, NH, NM, NV, OK, OR, RI, SC, SD, TN, TX, UT, VA, VT, WI, and WV) and 11 injury years (1998–2008). The transaction dates run through December 31, 2009, thus allowing for the accumulation of at least one year of indemnity and medical transactions for any given claim. Further, the data is evaluated as of June 30, 2010, thereby accommodating a reporting lag of six months. Claimants who are reported as being 15 years of age or younger are excluded.

A claim is categorized as "obese" if, for any diagnostic field (other than the primary one) on any medical transaction that occurs within 12 months after injury, the three leading digits of the ICD-9 code equal 278. The ICD-9 code 278, which is "non-reimbursable," denotes overweight, obesity, and other hyperalimentation. Note that the claim characteristic of interest rests on medical transactions. For instance, a claim whose indemnity benefits ends after one month without having acquired the ICD-9 code 278 may still qualify as obese if, within the following 11 months, there is a medical transaction with a 278 comorbidity indicator.

Allowing claims 12 months to acquire the ICD-9 code 278 creates "immortal time," that is, time during which the claim could not have ended—otherwise, the claim would not have been able to acquire the ICD-9 code 278. The problem that arises from there being immortal time may best be illustrated for the case where there is no difference between indemnity benefit and medical benefit durations. Assume that obesity has no impact on the duration of a claim and that any given claim on any given day has a certain probability of ending and, independent of that, a certain probability of assuming the ICD-9 code 278. The longer the duration of a claim is, the more time this claim has to acquire the code; as a result, the median duration of claims with an ICD-9 code 278 tends to exceed the median duration of claims that did not acquire this code. This difference in median durations is known as "immortal time bias."

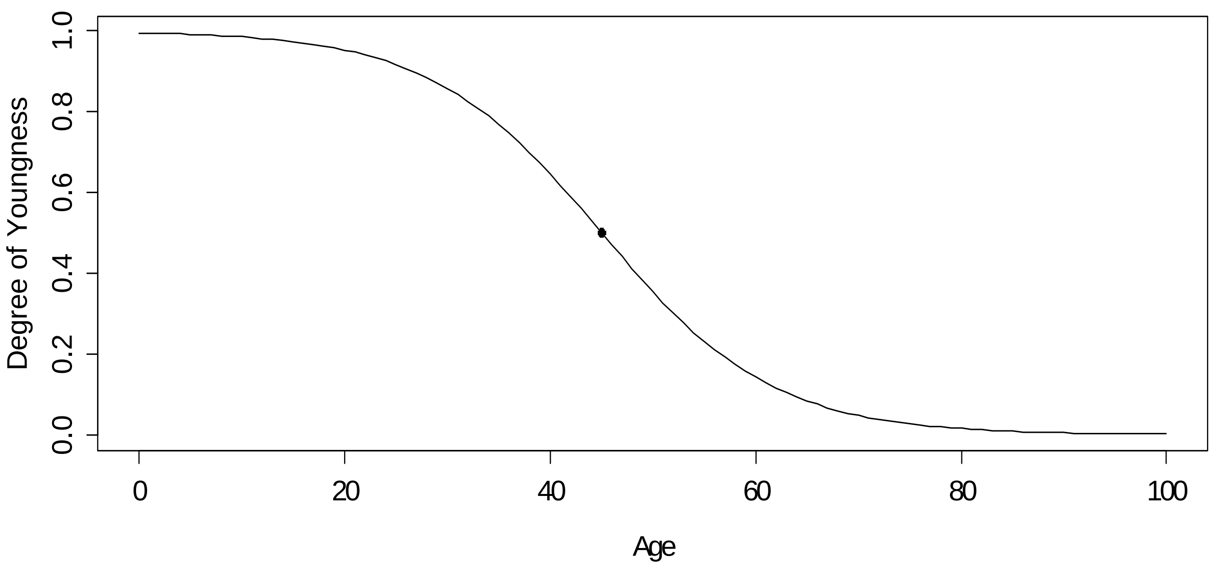
In order to correct for the immortal time bias, it is stipulated that a non-obese claim be a potential match for an obese claim only if this non-obese claim had at least one medical transaction past the time span it took the obese claim to acquire the ICD-9 code 278. As shown in Appendix 5.1, this immortal time bias correction (ITBC) almost perfectly removes the bias, given the nature of the data set. This does not imply that the applied ITBC approach will perform well in general.

The data set offers no information on the body mass index (BMI), which is a standard measure of obesity. As a result, we are not able to differentiate among degrees of obesity. It can be assumed that the comorbidity indicator identifies the claimant as severely obese, thus putting this claimant into one of the higher obesity classes. This conjecture is supported by the fact that only 0.15 percent of the claims in the data set acquire the ICD-9 code 278 (within 12 months of the date of injury). By comparison, in the Duke University study (Østbye, Dement, and Krause [5]), the proportion of morbidly obese claimants amounts to 4.9 percent.

The small percentage of obese claims in the total number of claims, along with the very high total number of claims (4,821,562) is most suited for a matched-pairs research design. In such an analytical setting, each obese claim in the data set is matched with a non-obese claim of the same primary ICD-9 code (at the three-digit level), injury year, U.S. state, industry, gender, and age at injury. Except for age at injury, all matching criteria are categorical in nature, which allows for exact matching (thus obviating the need for propensity matching). At the same time, extending the concept of exact matching to age at injury (which is measured on a scale of integers) leaves many obese claims without matches. Thus, for age at injury, we use proximity matching.

Proximity matching rests on the concept of the nearest neighbor. When exact matching for age is not feasible (for lack of exact matches when using integers for years of age), researchers often resort to matching by age bracket. In matching by age bracket, an obese claim is paired with a non-obese claim that belongs to the same (for instance) five-year age bracket (subject to being identical by all exact-matching characteristics). A disadvantage of matching by age bracket is that many claims are not matched with the closest neighbor. For instance, a 20-year-old obese claimant may be matched (within the 20–24 age bracket) with a 24-year-old non-obese claimant but is prevented from being matched with an otherwise identical 19-year-old claimant. Proximity matching avoids this problem by looking for the nearest neighbor. At the same time, it may not be appropriate to have the concept of the nearest neighbor rest on the simple age difference. Because aging is a nonlinear process, it may be preferable to match an obese 25-year-old claimant with a 20-year-old instead of a 30-year-old. Similarly, matching an obese 55-year-old with a non-obese 60-year-old may be more appropriate than matching this person with a non-obese 50-year-old. For this reason, we use a sigmoid function to create a fuzzy set for youngness; the sigmoid function reads $1/(1+\exp(-\sigma \cdot (h-45)))$, where h is the age at injury and σ was chosen to equal 0.12. (For the concept of fuzzy sets, see, for instance, Kasabov [2].) Chart 1 depicts the fuzzy set; the degree of youngness is set to 50 percent at age 45 and then, in an "S-shaped" manner, it gets near zero and 100 percent as the years of age approach 100 and zero, respectively. The nearest neighbor of an obese claim to an otherwise identical non-obese claim is defined by the smallest difference (in absolute value terms) in the degree of youngness.

Chart 1: Degree of Youngness, Defined by Sigmoid Function



In establishing the primary ICD-9 code, there is an order of priority based on (1) providers, (2) the time in the duration of a claim at which the medical service was provided, and then (3) the paid amount. The logic of establishing the primary ICD-9 code is detailed in Appendix 5.2.

Matching by industry relies on an NCCI industry classification; the five industries comprise Manufacturing, Contracting, Office and Clerical, Goods and Services, and Miscellaneous.

To summarize, we define as neighbors to a given obese claim the set of non-obese claims that match exactly based on primary ICD-9 code, injury year, U.S. state, industry, and gender; only non-obese claims are considered that have at least one medical transaction following the acquisition of the ICD-9 code 278 by the pertinent obese claim. In this matching process, a given non-obese claim may be used as a neighbor to more than one obese claim. Among the thus identified set of neighbors, the nearest neighbor is chosen based on the degree of youngness.

This nearest neighbor may not be unique because of ties in the youngness distance, in which case there are multiple matches. In the event of multiple non-obese matches for a given obese claim, a synthetic claim is generated using the median duration from the set of multiple matches. The algorithm for calculating the median indemnity benefit duration for a set of multiple matches is provided in Appendix 5.3.

2.1 Indemnity Benefit Duration

We calculate indemnity benefit duration (duration, for short) based on recorded indemnity transactions. The transactions are from the payment categories Temporary Total and Permanent Total; claims with transactions in the category Fatal are excluded. In a sensitivity analysis, we count transactions in the Permanent Partial category toward duration as well.

The duration of indemnity benefits may be censored. Censoring occurs when the duration is only partially known. In the event of right-censoring, the duration is only known to be greater than a certain value. In the case of interval-censoring, the duration falls into a given interval, as is the case with a claim that does not breach the waiting period. Note that for the purpose of this study, we count the waiting period toward the indemnity benefit duration even where the retroactive period has not been reached.

We employ two concepts of duration. Duration concept I aggregates compensated time intervals based on the "from-date" and "through-date" information associated with benefit payments. If more than 20 percent of the transactions associated with a given claim miss the from-/through-date information, then the claim is categorized as right-censored at the waiting period.

Duration concept II measures the time interval between injury date and most recent through-date. Transaction dates substitute for through-dates where the latter are missing. On one hand, duration concept II appears to be less sensitive to missing observations; on the other hand, if a claimant returned to work intermittently, duration concept II fails to capture this information.

For claims without indemnity transactions, it is assumed that the claim did not breach the waiting period. Such claims are treated as interval-censored on the waiting period. For claims with indemnity transactions, benefit duration is established by means of the concept of dormancy; only transactions prior to the date of dormancy are considered. The date of dormancy is the first transaction date that is not followed by another transaction within 180 days; reopenings past the date of dormancy are not accounted for. The latest date for which such a transaction-free time interval can be established is 180 days prior to December 31, 2009 (which is the latest transaction date considered in this study). Claims for which no 180-day transaction-free interval can be established are categorized as right-censored.

Both duration concepts measure calendar time (as opposed to number of missed workdays). Where there are transactions of indemnity payments available, measuring calendar time is straightforward; this is because the time intervals covered by the from-/through-date sets of recorded transactions tend to refer to calendar time rather than the number of compensated workdays; this calendar year information can then be linked to the date of injury. Measuring duration when there are no indemnity transactions is more demanding. Appendix 5.4 offers an algorithm for measuring the elapsed calendar time that leads up to the end of the waiting period.

2.2 Matching-Toward-the-Center Problem

Matching by the degree of youngness tends to pair old claimants with younger claimants and young claimants with older claimants; this is because the numbers of claimants on the edges of the age distribution tend to be comparably small. For instance, within the age bracket 60–64, the number of observations tends to decline with age. Thus, a claimant at the center of this age bracket has a comparatively high chance of being "matched down the age distribution," as opposed to being "matched up"; this is simply because the number of potential matches within this bracket is greater at lower ages than at higher ages. Conversely, young claimants tend to be "matched up the age distribution." Note that this matching toward the center (of the age distribution) is not unique to proximity matching, but it is also characteristic of the traditional approach of matching by the age bracket. Although proximity matching by age, as applied in this analysis, mitigates the matching-toward-the-center-problem, it turns out not to eliminate the problem entirely.

In Laws and Schmid [4], the dependent variable is the (natural) logarithm of the ratio of obese to non-obese claim costs; age at injury serves as a covariate. In such a regression design, matching toward the center of the age distribution affects the interpretation of the regression results for age (and age alone). As pointed out by the authors, if the medical costs of workers compensation claims increase with age, then the influence of age on the cost of obesity is underestimated for young claimants and overestimated for old claimants. As a consequence, the estimated effect of age may be distorted on the edges of the age distribution, taking on an "S-shaped" form.

In this study, and unlike in Laws and Schmid [4], for each pair of obese/non-obese claims, both observations enter the statistical analysis. As a result, the estimated effect of age is not adversely affected by the matching-toward-the-center problem as long as the influence of age is properly accounted for in the regression equation.

2.3 Descriptive Characteristics

We start out with a data set of 4,821,562 claims, of which 7,145 (or 0.1482 percent) are categorized as obese. We are able to match 6,528 obese claims (or 91 percent) to non-obese claims, among which we identified 487,603 potential matches before correcting for the immortal time bias and applying proximity matching by the degree of youngness. Correcting for the immortal time bias reduces the number of usable obese claims to 6,435; the number of potential non-obese matches drops to 356,423. After proximity matching by the degree of youngness, the number of potential matches reduces to 14,929, which translates into an approximate 2.3 non-obese matches per obese claim.

Table 1 details the number of matched pairs by injury year. The non-obese claim of a matched pair may be synthetic where there are multiple matches, as discussed above.

Table 1: Numbers of Matched Pairs by Injury Year

Injury Year	Number of Matched Pairs
1998	333
1999	319
2000	391
2001	472
2002	550
2003	548
2004	591
2005	678
2006	759
2007	841
2008	953
Total	6,435

Chart 2 offers the relative claim frequency by U.S. state, pooled over the studied 11 injury years. Florida and Texas are the most highly represented states, whereas Rhode Island and West Virginia are the least highly represented. The representation of a state in the data set depends primarily on the size of its labor force, but it also depends on the combined market share of the insurance companies that contribute to the analyzed data set. Another contributing factor is the share represented by the self-insured.

Chart 2: Relative Frequency of Matched Pairs by State, 1998–2008

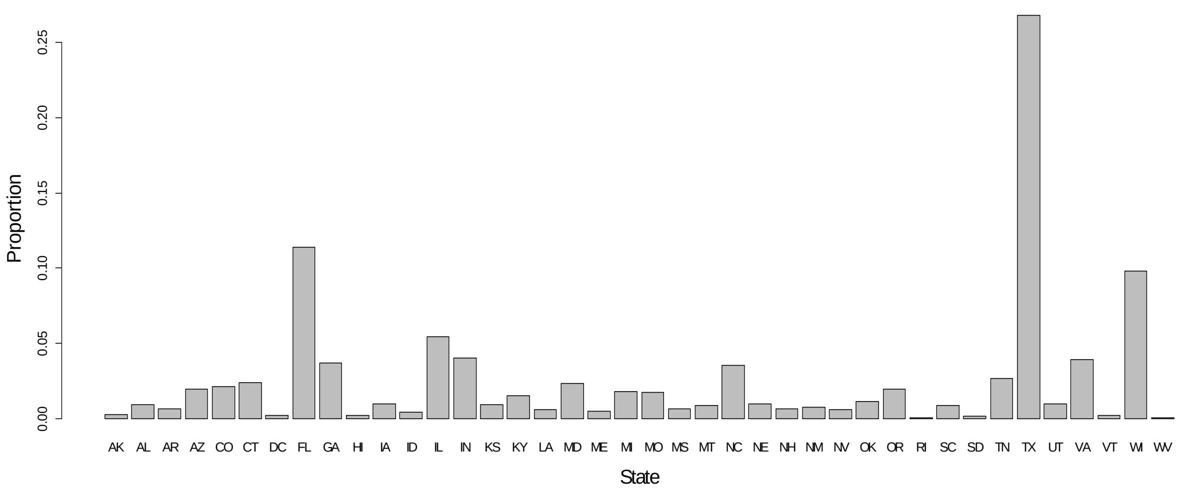


Chart 3 presents a distribution of age at injury of the set of studied (obese and non-obese) claims; again, the observations are pooled over the 11 injury years. The median age at injury is in the neighborhood of 40.

Chart 3: Relative Frequency of Claims by Age at Injury, 1998–2008

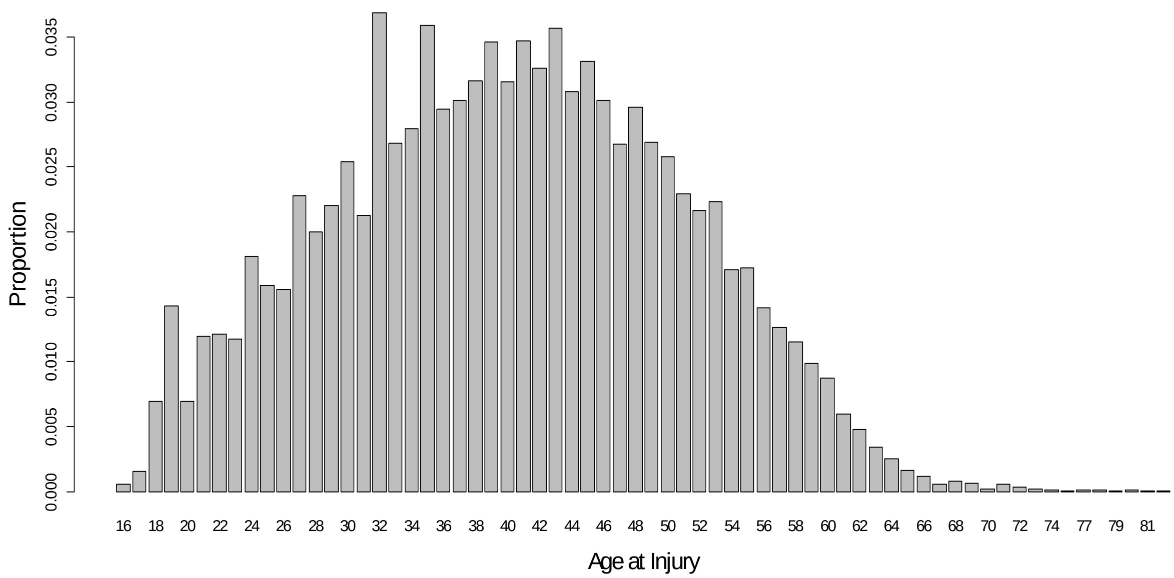
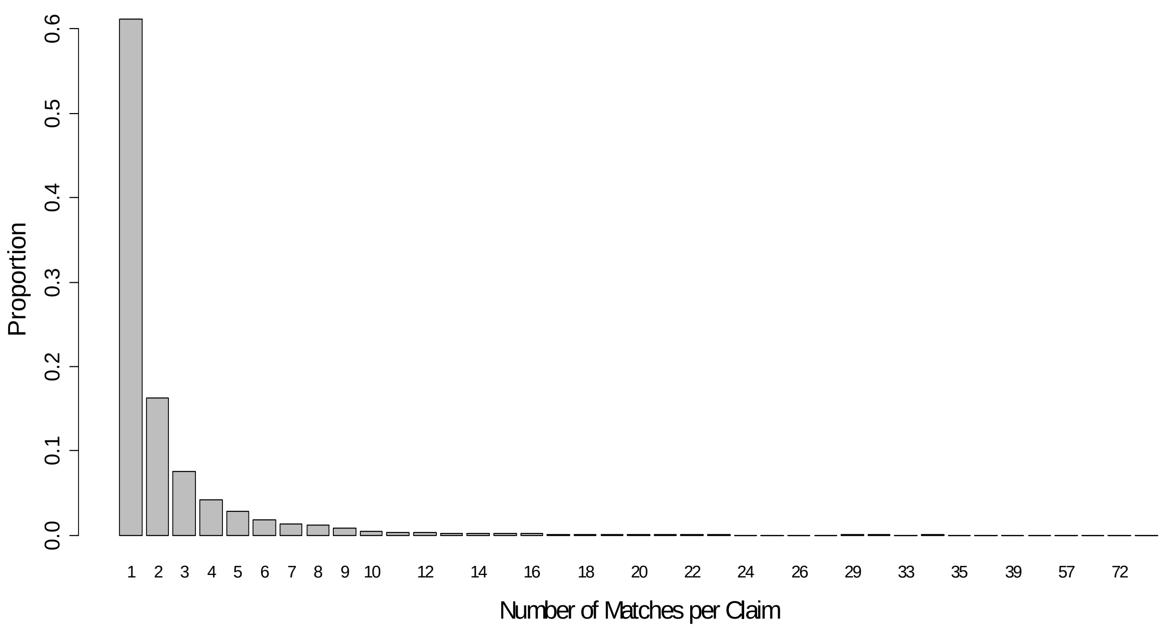


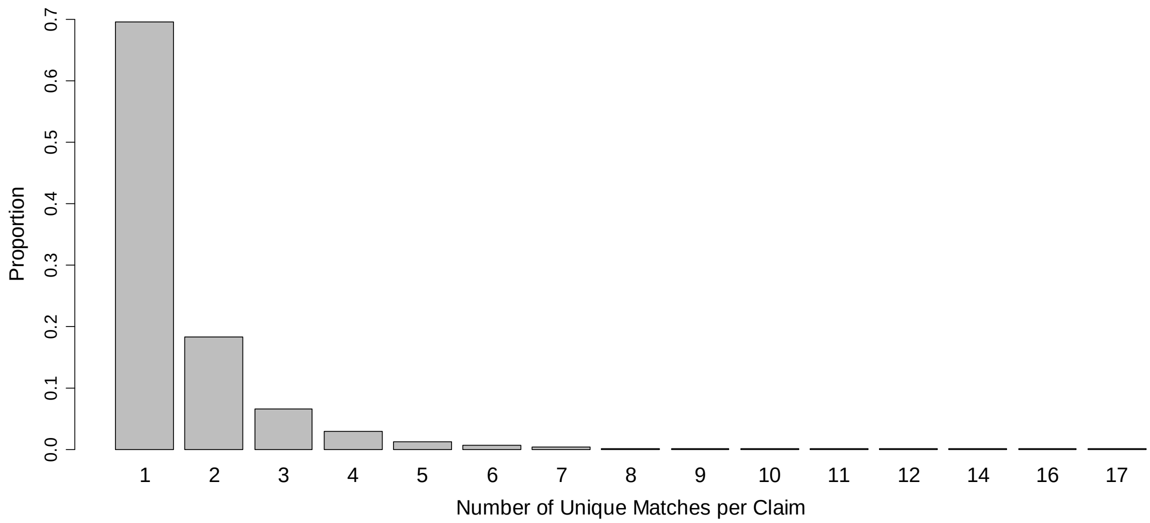
Chart 4 displays the relative frequency of the number of matches per obese claim. As mentioned, on average, there are 2.3 matches per obese claim, but, as the chart illustrates, this distribution is highly skewed. The majority of obese claims (around 60 percent) have only one match; obese claims with more than 10 matches are rare. As discussed, where matches are not unique, the median duration is calculated using the algorithm detailed in Appendix 5.3.

Chart 4: Relative Frequency of Numbers of Matches per Obese Claim



Finally, Chart 5 depicts a frequency distribution for the number of unique matches per obese claim. The multiple matches shown in Chart 4 are not necessarily unique in the sense that they may have different durations (but are otherwise identical). In Chart 5, a match is unique if there is no other (non-obese) claim that is identical in all respects (disregarding the type of censoring). For instance, two claims with durations of two weeks, one of them being right-censored, do not count as only one observation in the chart. Chart 5 indicates that in the vast majority of cases, an obese claim has a unique non-obese match.

Chart 5: Relative Frequency of Numbers of Unique Matches per Obese Claim



3. THE STATISTICAL ANALYSIS

Conceptually, the statistical problem of quantifying the effect of obesity on the duration of indemnity benefits is one of survival, that is, continuation of benefits. The hazard in this survival framework manifests itself in the termination of benefits. In a first step, the data is analyzed by means of Kaplan-Meier plots. Following this nonparametric analysis, a semiparametric Weibull proportional hazard model is estimated. In both analyses, the obese claims enter the statistical model along with their non-obese match. Also, both statistical frameworks account for interval-censoring and right-censoring. Finally, in both analytical settings, benefit duration is measured in calendar time.

3.1 Kaplan-Meier Plots

The Kaplan-Meier estimator is a simple way of computing a survival curve, which is a graph that depicts the survival experience (i.e., continuation) of indemnity benefit payments by claim. The percentage of surviving claims is displayed on the vertical axis. The horizontal axis depicts calendar time since injury.

Chart 6 presents the Kaplan-Meier plots for duration concept I. The vertical axis of the plot indicates the proportion of claims that survive any given day displayed on the timeline (which is on the log10 scale). The reading on the vertical axis of the starting point of the survival curve indicates the proportion of claims that do not breach the waiting period. As shown, only about 25 percent of obese claims do not breach the waiting period, whereas more than 50 percent of the non-obese claims do not breach the waiting period.

The waiting period varies by state, from a minimum of three days to a maximum of seven days. After translating lost workdays into calendar time (see Appendix 5.4), the upper bound of this interval grows to 11 days. This is because a worker whose regular workweek runs from Monday through Friday, and whose first missed workday is a Friday, will have missed eleven calendar days by the time he will have missed seven workdays. In the Kaplan-Meier plot, during the waiting period, the hazard of the affected, interval-censored claims is constant.

The whiskers at the bottom of the Kaplan-Meier plots indicate the location of non-censored claims ("Observed") and censored claims ("Censored"); censored claims may be either interval-censored (in which case the whiskers indicates the end of the waiting period) or right-censored (at the waiting period or later).

Chart 6: Kaplan-Meier Plot for Duration Concept I

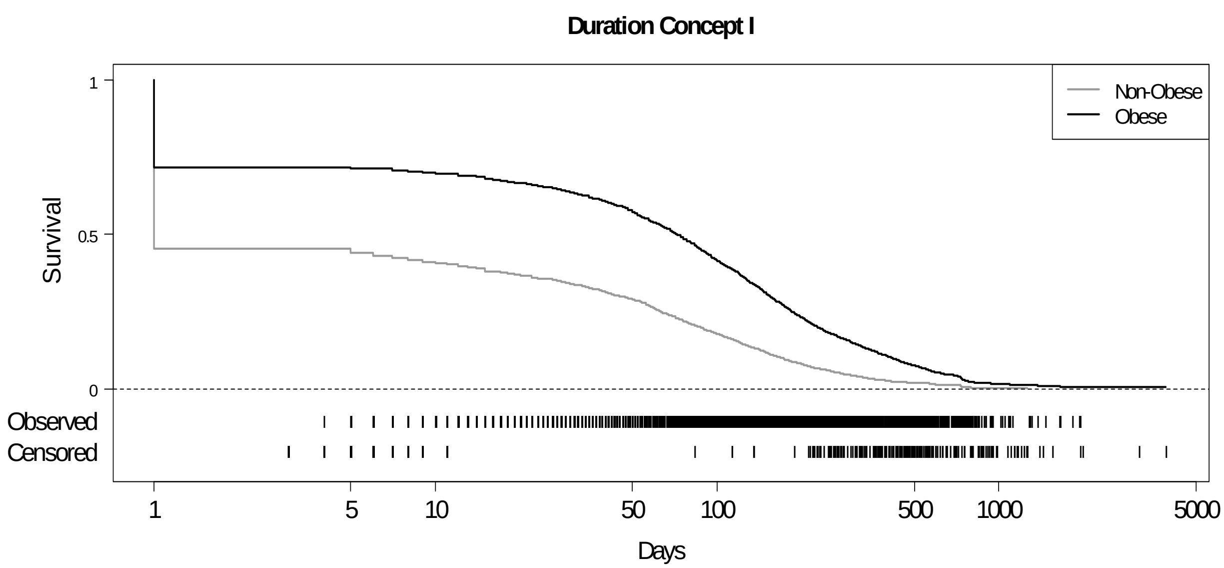
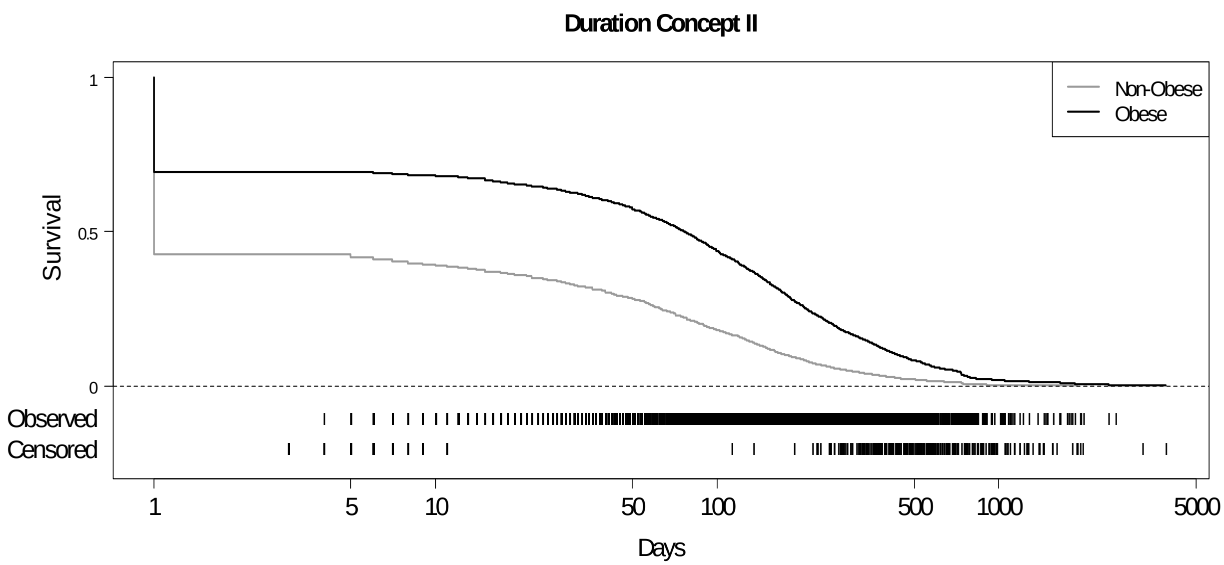


Chart 6 shows that for any duration displayed on the horizontal axis, the proportion of claimants still collecting indemnity benefits is considerably higher for the obese than for the non-obese. To the degree that obesity correlates with age, the vertical distance between the survival curves may be distorted slightly by the matching-toward-the-center effect.

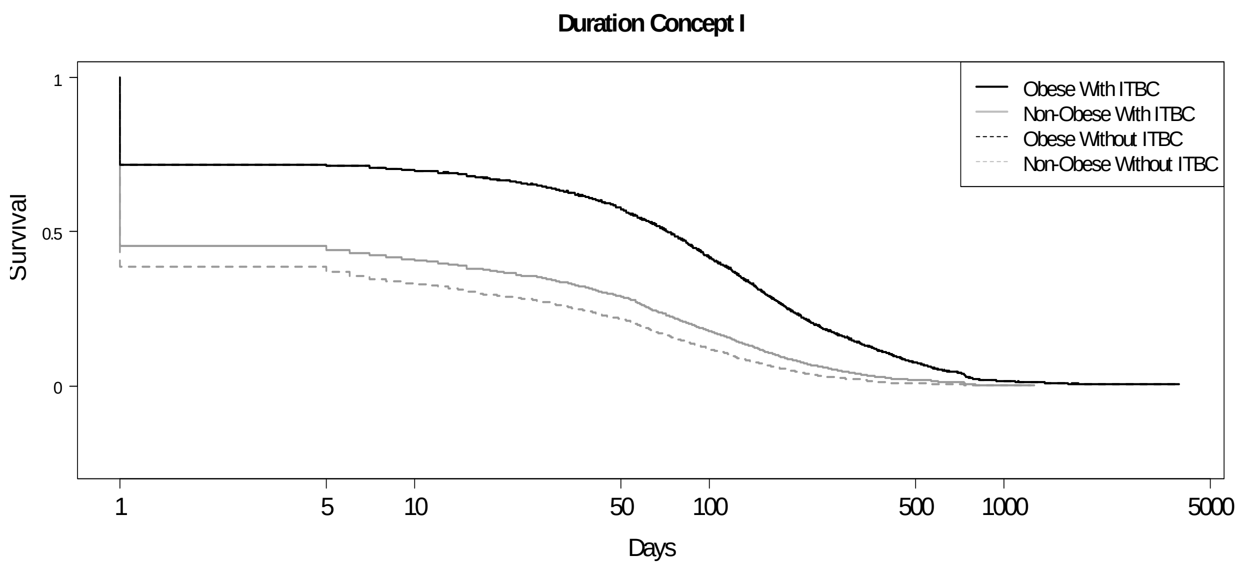
Chart 7 displays the survival curves for duration concept II. These curves look very similar to those obtained using duration concept I.

Chart 7: Kaplan-Meier Plot for Duration Concept II



In an effort to ascertain the impact that the immortal time bias would have on the results if left uncorrected, we applied the Kaplan-Meier estimator to the data without the immortal time bias correction (ITBC). Chart 8 displays (for duration concept I) the survival curves with and without ITBC. As mentioned, the ITBC correction reduces the data set to 6,435 obese claims and 14,929 non-obese claims. Without the ITBC, the number of obese claims is slightly larger at 6,528; by comparison, there are 19,397 matching non-obese claims. Not surprisingly, the Kaplan-Meier estimators for the two sets of obese claims are nearly identical to the point of being indistinguishable in the chart. For the non-obese matches, however, the survival curves are distinctly different and, considering the log10 scale, this difference is nonnegligible.

Chart 8: Kaplan-Meier Plot for Duration Concept I, Immortal Time Bias Correction



3.2 Weibull Proportional Hazard Model

The purpose of the proportional hazard model is to quantify the effect of obesity on the mean (and median) duration while controlling for the influence of age. Controlling for age is necessitated by the matching-toward-the-center effect.

A proportional hazard model implies that the hazard in one group (e.g., the obese) is, on the timeline (calendar time since injury), a constant proportion of the hazard in the other group (e.g., the non-obese); this proportion is known as the hazard ratio. The baseline hazard as a function of time is assumed to follow a Weibull distribution.

As specified, the Weibull proportional hazard model accommodates interval-censoring and right-censoring. Further, the model is semiparametric in that it consists of a parametric component (akin to a standard, linear regression model) and a nonparametric component. The linear component includes only one explanatory variable, which is the indicator variable for obesity. The purpose of the nonparametric component is to capture the potentially nonlinear influence of age.

The model is estimated by means of MCMC, using the sampling platform JAGS in R. The JAGS code is displayed in Appendix 5.5.

Following Dellaportas and Smith [3], the hazard function can be written as

$$\lambda(t;\boldsymbol{X}) = \lambda_0(t) \cdot \exp\left(\boldsymbol{X\beta} + f(z)\right) , \tag{1}$$

where $\lambda_0(t)$ is an unknown function of time and $\boldsymbol{X\beta} + f(z)$ is a semiparametric regression model. The matrix $\boldsymbol{X}$ comprises the covariates at time t that are included in the parametric component of the partially linear model, and z represents the single covariate that enters the nonparametric component. The nonparametric component is represented by a linear spline with knots at ages 25, 35, 45, 55, and 65. For the purpose of parameter identification, the spline is centered on age 45.

For $\lambda_0(t) = \rho t^{\rho-1}$, we obtain the Weibull hazard function

$$\lambda(t;\boldsymbol{X}) = \rho t^{\rho-1} \cdot \exp\left(\boldsymbol{X\beta} + f(z)\right) , \tag{2}$$

where $\rho > 0$ is the shape parameter of the Weibull distribution.

With the cumulative hazard given by $\Lambda(t) = t^{\rho}$, we can write the Weibull density function as

$$f(t) = \rho t^{\rho-1} \cdot \exp\left(\boldsymbol{X\beta} + f(z) - t^{\rho} \exp\left(\boldsymbol{X\beta} + f(z)\right)\right) . \tag{3}$$

Repeatedly substituting the scale parameter λ for the generalized semiparametric regression model $\exp\left(\boldsymbol{X\beta} + f(z)\right)$ delivers the Weibull density function and its familiar form:

$$\begin{aligned} f(t) \quad = \quad & \rho t^{\rho-1} \cdot \exp\left(\boldsymbol{X\beta} + f(z) - t^{\rho}\lambda\right) \\ & \rho t^{\rho-1} \cdot \exp\left(\boldsymbol{X\beta} + f(z)\right) \cdot \exp\left(-t^{\rho}\lambda\right) \\ & \rho t^{\rho-1} \cdot \lambda \cdot \exp\left(-t^{\rho}\lambda\right) . \end{aligned} \tag{4}$$

The mean of the Weibull distribution is known to read

$$\left(\frac{1}{\lambda}\right)^{\frac{1}{\rho}} \cdot \Gamma\left(1 + \frac{1}{\rho}\right) , \tag{5}$$

where Γ is the gamma function. Similarly, the median of the Weibull distribution equals

$$\left(\frac{1}{\lambda}\right)^{\frac{1}{\rho}} \cdot \left(\ln(2)\right)^{\frac{1}{\rho}} . \tag{6}$$

Most interestingly, the ratio of the mean to the median is solely a function of the shape parameter ρ and, thus, independent of the hazard λ and the associated generalized semiparametric regression model $\exp\left(\boldsymbol{X\beta} + f(z)\right)$. By implication, any multiplicative impact of a covariate on the mean equals the multiplicative impact on the median.

The effect of obesity is captured by an indicator variable and manifests itself in a change in intercept; this indicator variable is the only covariate in the parametric part of the partially linear regression model.

Let $\hat{\beta}_1$ be the estimated intercept associated with obese claims, and let $\hat{\beta}_0$ be the corresponding value for non-obese claims. In this case then, the ratio of mean (and median) durations of obese claims to non-obese claims (or, equivalently, the multiplicative impact of obesity on mean and median durations) equals

$$\frac{\left(\dfrac{1}{\exp\left(\hat{\beta}_1 + \hat{f}(z)\right)}\right)^{\frac{1}{\hat{\rho}}}}{\left(\dfrac{1}{\exp\left(\hat{\beta}_0 + \hat{f}(z)\right)}\right)^{\frac{1}{\hat{\rho}}}} = \exp\left(\hat{\beta}_0 - \hat{\beta}_1\right)^{\frac{1}{\hat{\rho}}} , \tag{7}$$

where $\exp\left(-\left(\hat{\beta}_0 - \hat{\beta}_1\right)\right)$ is the associated hazard ratio.

Chart 9 depicts the multiplicative impact of obesity. As shown in the chart, the mean and median durations of obese claimants are more than five times the average durations of non-obese claimants. When Permanent Partial benefit payments are counted toward indemnity benefit duration as well, this multiple climbs to more than six.

Chart 10 displays the hazard ratio. When excluding Permanent Partial benefit payments, the probability of the indemnity benefits of an obese claim ending on any given day is a little more than half the probability of this event occurring for a non-obese claim. Including Permanent Partial benefit payments in the analysis has little impact on the estimated hazard ratio.

Finally, Chart 11 depicts the effect of age on indemnity duration for the set of matched pairs analyzed in this study. The effect of age is net of the effect of obesity. As mentioned, the spline is centered on age 45 for the purpose of parameter identification; as a result, the credible intervals are relative to a claimant with an age at injury of 45. (For the difference between credible intervals and confidence intervals, see Carlin and Louis [1].) The chart shows that a claimant whose age at injury falls into the age bracket 45 to 65 tends to have indemnity benefit durations that are about twice the indemnity durations of claimants who sustain a workplace injury at age 25. As a caveat, the research design employed in this study aims at discerning the effect of obesity on indemnity benefit duration, not the effect of age. For the purpose of arriving at definitive statements on the effect of age, it is preferable to use a more comprehensive data set, as opposed to the employed matched pairs.

Chart 9: Multiplicative Impact of Obesity on Mean Duration; Exclusive and Inclusive of Permanent Partial Indemnity Benefits; Duration Concept I and II; 80 Percent Credible Intervals

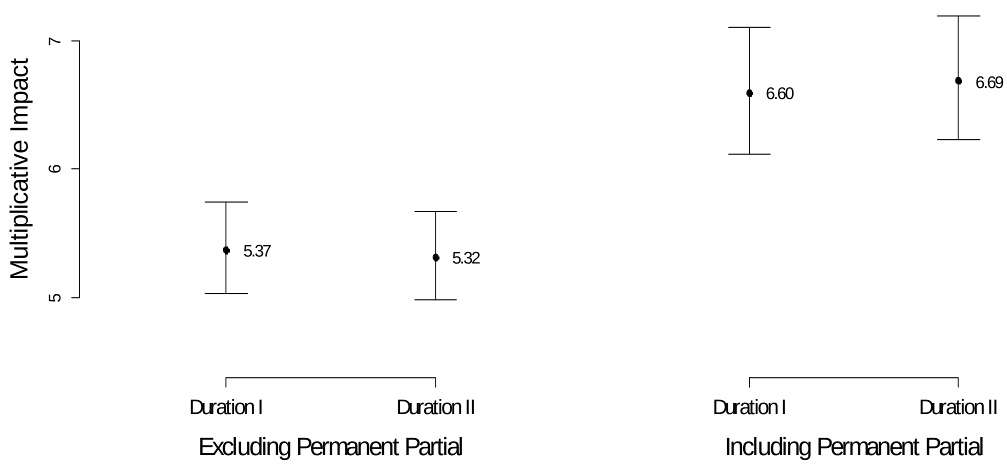


Chart 10: Hazard Ratio; Exclusive and Inclusive of Permanent Partial Indemnity Benefits; Duration Concept I and II; 80 Percent Credible Intervals

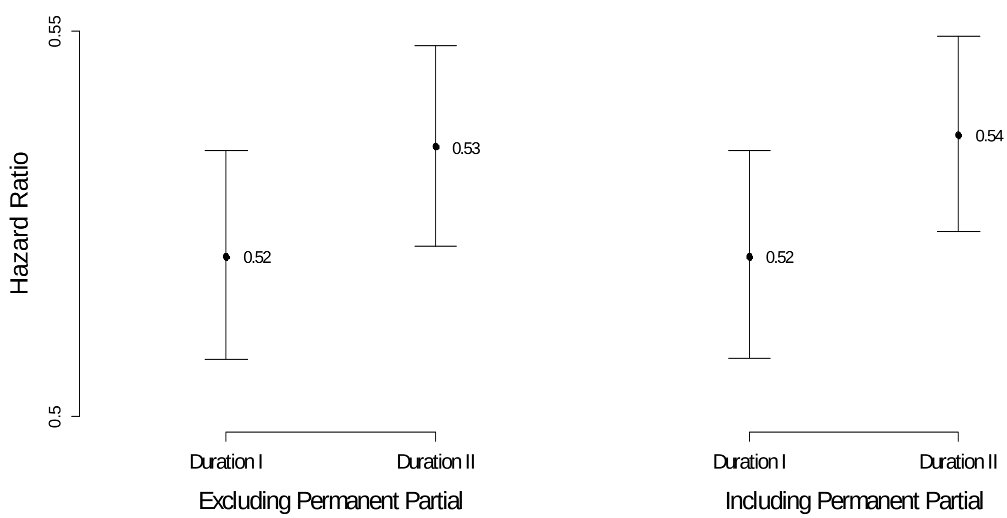
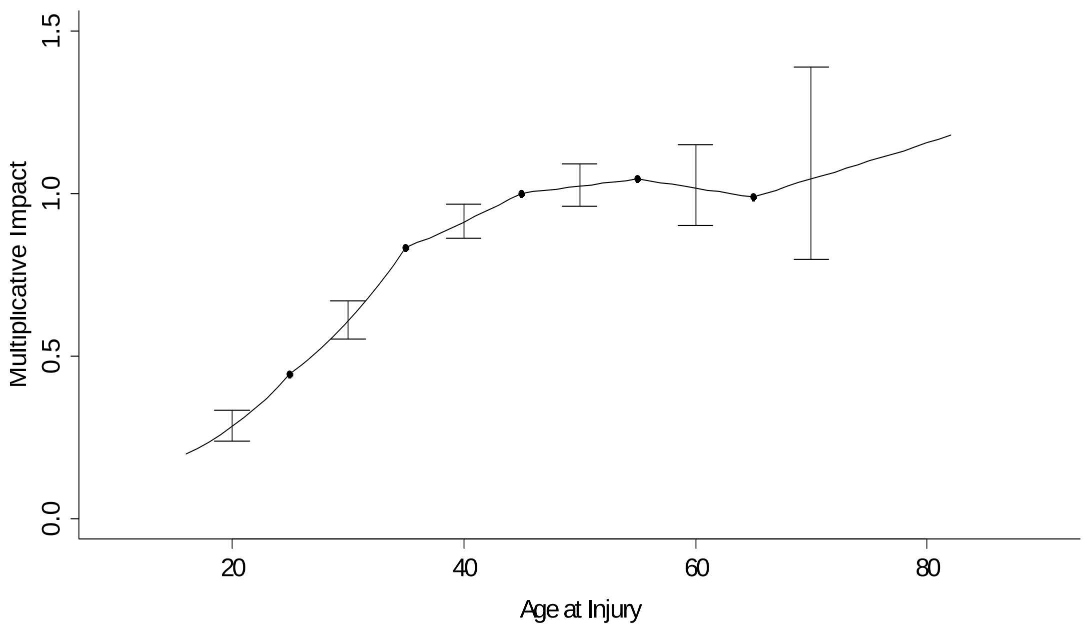


Chart 11: Effect of Age on Mean Duration, Duration Concept II, 80 Percent Credible Intervals



4. CONCLUSION

The study finds that obesity contributes in significant ways to the length of time during which claimants receive indemnity benefits. Indemnity duration was measured based on Temporary Total and Permanent Total indemnity benefit payments; in a sensitivity analysis, Permanent Partial benefits were counted toward indemnity benefit duration as well. Two concepts of aggregating observed indemnity benefit transactions into duration were employed, with little difference in the results. Further, the statistical analysis accounted for the presence of interval-censoring and right-censoring, both in the nonparametric framework of Kaplan-Meier plots and in the Bayesian semiparametric proportional hazard model.

The statistical analysis showed that claimants with a comorbidity indicator pointing to obesity have an indemnity benefit duration that is more than five times the value of claimants who do not have this comorbidity indicator but are otherwise comparable. Inclusive of Permanent Partial indemnity payments, this multiple climbs to more than six.

Clearly, the limiting factor in this study is the lack of information on the body mass index of the claimant. On one hand, it can be argued that the analysis overestimates the effect of obesity if the assignment of the comorbidity indicator, the ICD-9 code 278, is related to the arrival of obesity-related medical complications, as opposed to the condition of obesity. Medical complications provide an incentive to the physician to assign this "non-reimbursable" ICD-9 code. From this perspective, a selection bias may be introduced, as a result of which claimants who acquire this comorbidity indicator may disproportionately belong to the highest obesity category—the morbidly obese; this conjecture is supported by the fact that only 0.15 percent of the claims in the data set acquire the ICD-9 code 278 (within 12 months of the date of injury). On the other hand, a case can be made that the effect of obesity is underestimated. This is because many of the claimants who are categorized as non-obese in this study may, in fact, be overweight or obese, thereby diminishing the measured contribution of obesity to duration.

Despite the limitations of the data set employed in this study, the results obtained for the effect of obesity on indemnity duration are close to what has been established by Duke University for the morbidly obese. Based on the reported means in the Duke University study, for the morbidly obese, the number of lost workdays per claim amounts to 6.4 times the value observed for claimants of recommended weight. By comparison, for duration concept I [II], the multiples established here equal 5.4 [5.3] (Temporary Total and Permanent Total) and 6.6 [6.7] (Temporary Total, Permanent Total, and Permanent Partial).

5. APPENDIX

5.1 Simulation for Validating the Immortal Time Bias Correction (ITBC)

A Monte-Carlo simulation was carried out to investigate the validity of the immortal time bias correction described earlier. The simulation maintains the primary features of the underlying data-generation process under the assumptions that, first, the data-generating process for benefit duration is the same for all claims and, second, the probability of acquiring a 278 ICD-9 code is governed solely by the duration of a claim. The simulation assumes that (1) claims only generate ICD-9 codes while they are open, (2) claims are only considered open over their indemnity benefit duration, (3) there is a fixed probability, π, that on any day that a claim is open, it will produce a 278 ICD-9 code, and (4) the acquiring of a 278 ICD-9 on day x is independent of the acquiring of a 278 on day y, provided that $x \neq y$. In this simulation, the parameters were calibrated to reflect important characteristics of the data and the estimated statistical model.

The algorithm generates pairs of obese and non-obese claims as follows:

(1) Draw the indemnity benefit duration and the time to the first occurrence of a ICD-9 code 278 for a claim classified as obese

   (a) Randomly generate an indemnity benefit duration from a Weibull distribution and round up to the nearest integer

   (b) Randomly draw the time to the first ICD-9 code 278 from a geometric distribution with parameter π (Note: This step b is independent of step a)

   (c) If the time to the first ICD-9 code 278 is greater than the minimum of the drawn indemnity benefit duration or one year, then reject the draw and return to step 1; otherwise, accept the draw and proceed to step 2

(2) Draw a suitable non-obese claim to pair with the obese claim generated in step 1

   (a) Randomly generate an indemnity benefit duration from a Weibull distribution and round up to the nearest integer

   (b) Randomly draw the time to the first ICD-9 code 278 from a geometric distribution with parameter π (Note: This step b is independent of step a)

   (c) If the time to the first ICD-9 code 278 is less than the minimum of the drawn indemnity benefit duration or one year, then reject the draw and return to step 2; otherwise, accept the draw and return the obese/non-obese pair

Under the aforementioned simulation, the average duration for claims flagged as obese is 98.5 days; the average duration for claims flagged as non-obese is 98.2 days.

Carrying out the simulation without the immortal time bias correction is equivalent to all generated claims passing step 2c. Without the immortal time bias correction, the average duration for claims flagged as obese is 98.5 days and the average duration for claims flagged as non-obese is 9.5 days.

The simulation result demonstrates that the proposed correction removes the bulk of the immortal time bias.

5.2 Logic of Establishing Primary ICD-9 Code

The primary ICD-9 code of a given claim is the first ICD-9 code associated with reported medical transactions (using a valid service date) based on the following order of priority:

(1) Paid amount must be greater than zero

(2) Provider Group must be one of the following four:

01 (Physician), 04 (Hospital, Ambulatory Surgical Center, X-Ray, Lab), 05 (Clinic), 07 (Non-Medical)

When no such transaction exists, transactions from all other providers are considered:

02 (Chiropractor), 03 (Therapist), 06 (Pharmacies and Durable Medical Equipment Center), 08 (Other)

(3) Transactions are selected from the first remaining

(4) service date, subject to data cleansing considerations

(5) From the ICD-9 codes resulting from steps 1 through 3, the ICD-9 codes associated with the highest paid amount over the (heretofore recorded) life of the claim are selected

The paid amount of a transaction is assigned in full to any ICD-9 code associated with the transaction

(5) If there are ties after applying the mentioned criteria, an ICD-9 code is randomly selected from the set of tied codes

5.3 Determining the Median Duration among Multiple Non-Obese Matches

Assume that an obese claim has multiple matching non-obese claims. In this case, the matching non-obese claim assumes the median duration of the set of multiple matches. This median duration is computed by first sorting the matching claims by duration. Then, in the event that the number of matching claims is odd, all claims are identified that have the same duration as the central value. If more than one such claim exists, a choice is made among these claims based on the following hierarchy, listed in descending order: right-censored, non-censored, interval-censored. In the event that the number of matching claims is even, the two central values are identified. Then, all claims with a duration that equals the longest duration of these two values are identified. If more than one such claim exists, a choice is made among these claims based on the following hierarchy, listed in descending order: right-censored, non-censored, interval-censored.

5.4 Converting the Waiting Period into Calendar Time

Waiting periods are stated in numbers of workdays. These workdays were converted into calendar time by assuming employees work five consecutive days a week. The conversion algorithm reads as follows:

(1) In the event that the day of injury is between (and inclusive of) Monday and Friday, the days off work are Saturday and Sunday; else, it is Tuesday and Wednesday

(2) Determine the shortest vector of consecutive days of the week, starting with the day of injury, such that the length of the vector, excluding the days off work, equals the waiting period

5.5 JAGS Code (Weibull Proportional Hazard Model)

```
## Defining is.censored (from JAGS Manual)
## X ~ dinterval(failure.time, waiting.period.or.last.observed)
## Left Censored X = 0; Right Censored X = 1; Not Censored X = NA
model
{
for(i.obs in 1:n.obs){
   failure.time[i.obs] ~ dweib(r,mu[i.obs])
   is.censored[i.obs] ~ dinterval(failure.time[i.obs], waiting.period.or.last.observed[i.obs])

    mu[i.obs] <- exp(alpha[is.obese[i.obs]+1] + f[i.obs])

    f[i.obs] <- age[i.obs] * beta[1] +
        (age[i.obs] >= knot[1])*(age[i.obs] - knot[1])*b[1] +
            (age[i.obs] >= knot[2])*(age[i.obs] - knot[2])*b[2] +
                (age[i.obs] >= knot[3])*(age[i.obs] - knot[3])*b[3] +
                    (age[i.obs] >= knot[4])*(age[i.obs] - knot[4])*b[4] +
                        (age[i.obs] >= knot[5])*(age[i.obs] - knot[5])*b[5]

    median[i.obs] <- (log(2) * 1/mu[i.obs])^(1/r)
}

##Spline
##Random regression coefficients corresponding to the truncated polynomial functions
for(i.knot in 1:n.knots){
    b[i.knot] ~ dnorm(0, tau.b)
}

##Fixed regression coefficient corresponding to the ‘plus’ functions
beta ~ dnorm(0,1.0E-2)

##Priors
r ~ dexp(1)
alpha[1] ~ dnorm(0,0.01) ##Not Obese
alpha[2] ~ dnorm(0,0.01) ##Obese

tau.b ~ dgamma(1.0E-3,1.0E-3)
sigma.b <- 1/sqrt(tau.b)

obese.control <- alpha[2] - alpha[1]
}
```

Acknowledgment

Thanks to Chun Shyong and Nathan Lord for research assistance.

6. REFERENCES

[1] Carlin, Bradley P., and Thomas A. Louis, *Bayes and Empirical Bayes Methods for Data Analysis.* 2nd ed., Boca Raton: Chapman & Hall/CRC, 2000.

[2] Kasabov, Nicola K., *Foundations of Neural Networks, Fuzzy Systems, and Knowledge Engineering*, Cambridge, MA: MIT Press, 1996.

[3] Dellaportas, P., and Adrian F.M. Smith, "Bayesian Inference for Generalized Linear and Proportional Hazards Models via Gibbs Sampling," *Applied Statistics* **42**(3), 1993, pp. 443–459.

[4] Laws, Chris, and Frank Schmid, "Reserving in the Age of Obesity," November 1, 2010, www.ncci.com/documents/research-age-of-obesity.pdf.

[5] Østbye, Truls, John M. Dement, and Katrina M. Krause, "Obesity and Workers' Compensation: Results From the Duke Health and Safety Surveillance System," *Archives of Internal Medicine* **167***,* 2007, pp. 766–773.

[6] Pollack, Keshia M., and Lawrence J. Cheskin, "Obesity and Workplace Traumatic Injury: Does the Science Support the Link?" *Injury Prevention* **13**(5), 2007, pp. 297–302.

[7] Pollack, Keshia M., Gary S. Sorock, Martin D. Slade, Linda Cantley, Kanta Sircar, Oyebode Taiwo, and Mark R. Cullen, "Association Between Body Mass Index and Acute Traumatic Workplace Injury in Hourly Manufacturing Employees" (with discussion), *American Journal of Epidemiology* **166**, 2007/2008, pp. 204-211 & **167**, pp. 123-124.

[8] Shaw, William S., Torill H. Tveito, Mary Jane Woiszwillo, and Glenn Pransky, "The Effect of Body Mass Index on Recovery and Return to Work Following Onset of Work-Related Low Back Pain," *Journal of Occupational and Environmental Medicine* **54**(2), 2012, pp. 192–197.

[9] Witters, Dan, and Sangeeta Agrawal, "Unhealthy U.S. Workers' Absenteeism Costs 153 Billion," Gallup, October 17, 2011, www.gallup.com/poll/150026/unhealthy-workers-absenteeism-costs-153-billion.aspx.

Abbreviations and notations

BMI, Body Mass Index
FTE, Full-Time Equivalent
HRA, Health Risk Assessment
ITBC, Immortal Time Bias Correction
ICD-9, International Classification of Diseases, Ninth Revision
LBP, Low back pain
MCMC, Markov chain Monte Carlo simulation
NCCI, National Council on Compensation Insurance

Biographies of the Authors

Frank Schmid, Dr. habil., was, at the time of writing, a Director and Senior Economist at the National Council on Compensation Insurance, Inc.
C**hris Laws** is a Research Consultant at the National Council on Compensation Insurance, Inc.
Matthew Montero is a Senior Actuarial Analyst at the National Council on Compensation Insurance, Inc.

The Impact of Physician Fee Schedule Introductions in Workers Compensation: An Event Study

Frank Schmid and Nathan Lord

Abstract
Motivation. Quantifying the impact of the introduction of physician fee schedules in workers compensation is an integral part of NCCI legislative pricing. Although the vast majority of U.S. states have physician fee schedules in place, a small number of jurisdictions operate without such a legal provision. Several studies have attempted to measure the impact of fee schedule introductions on the price and utilization of medical services provided by physicians. None of these analyses delivered evidence of an aggregate utilization increase in response to fee schedule introductions, although some point to changes in utilization for isolated procedures. Further, this prior research established overwhelming evidence that fee schedules contribute to lower price levels and lower rates of price level increases.
Method. This study analyzes the introduction of workers compensation physician fee schedules in two states: Tennessee (which adopted a fee schedule in July 2005) and Illinois (February 2006). Event study methodology is used to quantify the aggregate effect of these legislative actions. This aggregate effect is measured by a change in the severity index, which comprises both the price and utilization responses. Time series modeling is used to forecast the severity and utilization indexes that would have been observed absent the fee schedule introduction. The differences between observed (net of noise) and forecast severity and utilization indexes in the third month of fee schedule operation then serve as estimates for the impacts of the fee schedule implementation. The price response to this cost containment measure is backed out of the severity and utilization responses. In a sensitivity analysis, the price level response is obtained by comparing the price level of the third post-implementation month to the price level of the final pre-implementation month; the change in the price level thus measured is then combined with the estimated utilization effect to arrive at the severity response.
Results. In both jurisdictions, the fee schedule introductions contribute to a marked decline in the price level of medical services provided by physicians, as well as a permanent weakening in the rate at which this price level subsequently increases. In Tennessee, the price level declines in excess of 7 percent, and the annual rate of inflation lessens by 0.3 percentage points. By comparison, in Illinois, the price level drops by slightly more than 5 percent and the annual rate of inflation decreases by 0.6 percentage points. In Tennessee, there is a negative utilization response, which may be related to a restriction on physician choice that was implemented prior to the fee schedule introduction. For Illinois, the utilization response is essentially nil.
Availability. The statistical model was implemented in R (cran.r-project.org), using the R package *forecast* (cran.r-project.org/web/packages/forecast/index.html).

Keywords. Physician Fee Schedules, Utilization, Workers Compensation

1. INTRODUCTION

Quantifying the impact of the introduction of physician fee schedules in workers compensation is an integral part of NCCI legislative pricing. Although the vast majority of U.S. states have physician fee schedules in place, a small number of jurisdictions operate without such a legal provision. Several studies have attempted to measure the impact of fee schedule introductions on the price and utilization of medical services provided by physicians. None of these analyses delivered evidence of an aggregate utilization increase in response to fee schedule introductions, although some point to changes in utilization for isolated procedures. Further, this prior research established overwhelming evidence that fee schedules contribute to lower price levels and lower rates of price level increases.

This study analyzes the introduction of workers compensation physician fee schedules in two states: Tennessee (which adopted a fee schedule in July 2005) and Illinois (February 2006). Event study methodology is used to quantify the aggregate effect of these legislative actions. This aggregate effect is measured by a change in the severity index, which comprises both the price and utilization responses. Time series modeling is used to forecast the severity and utilization indexes that would have been observed absent the fee schedule introduction. The differences between observed (net of noise) and forecast severity and utilization indexes in the third month of fee schedule operation then serve as estimates for the impacts of the fee schedule implementation. The price response to this cost containment measure is backed out of the severity and utilization responses. In a sensitivity analysis, the price level response is obtained by comparing the price level of the third post-implementation month to the price level of the final pre-implementation month; the change in the price level thus measured is then combined with the estimated utilization effect to arrive at the severity response.

1.1 Research Context

Studies on the impact of physician fee schedules fall into two categories. First, there are analyses that compare, for a given point in time, states with a physician fee schedule to states without such a legal provision. Second, there are single-state studies that focus on the time windows before and after the fee schedule introduction or, alternatively, compare workers compensation medical care to Group Health. Cross-state studies may be most suited for discerning the long-term effects of fee schedule introductions, although isolating the contribution of a physician fee schedule from other cost containment measures can be challenging. Paradoxically, now that the vast majority of states utilize physician fee schedules in workers compensation, gauging the degree to which these legal provisions succeed in containing the medical costs of workers compensation claims becomes an increasingly demanding task—this is because the counterfactual becomes progressively difficult to establish. Then again, single-state studies of the long-term effects of fee schedule introductions have their own challenges—the reason is that, over time, the analyzed jurisdictions may experience a host of changes to the legislative and economic environment that affect the provision of medical services by physicians in workers compensation. Examples of such changes include restrictions on physician choice and programs that provide incentives for returning to work.

Early cross-state studies of the effect of physician fee schedule introductions on the medical costs of workers compensation claims include a 1986 paper by Borba [2], research published in 1989 by the National Council on Compensation Insurance [14], an analysis by Boden and Fleischman [1] of the same year, and a paper by Durbin and Appel [4] from 1991. Further studies were published by Pozzebon [15] in 1994 and by Robertson and Corro [21] in 2007.

In a preliminary analysis, Borba [2] compares the medical costs per claim of fee schedule states to non-fee schedule states. The author draws on workers compensation claims filed in 1980 and 1983, as recorded in the Detailed Claims Information (DCI) database of the National Council on Compensation Insurance (NCCI). The 1980 data set comprises three fee schedule jurisdictions (Florida, Massachusetts, and New York) and 10 non-fee schedule jurisdictions; the 1983 data set covers four fee schedule jurisdictions (Hawaii, Florida, Massachusetts, and Oregon) and 8 non-fee schedule jurisdictions. The author standardizes the medical care expenditures of the various states using information from the Health Care Financing Administration (HCFA). Borba finds that in the 1980 (1983) data set, the costs per claim are 9.0 (15.4) percent lower in fee schedule states than in non-fee schedule states.

The 1989 article by the National Council on Compensation Insurance [14] relates preliminary findings of a study undertaken by the NCCI Research Division. This study covers the time period 1965 through 1984 and an undisclosed number of jurisdictions; there are 512 annual observations. Based on an econometric time series–cross section approach, physician fee schedules are found to reduce the average medical costs per claim by 11.9 percent. Following this preliminary analysis, Durbin and Appel [4] published final results two years later, using a data set that spans the same time period (1965 through 1984) and comprises 33 jurisdictions. In this refined study, the authors arrive at a long-term effect of physician fee schedules on the medical costs per claim of a negative 3.5 percent; in a sensitivity analysis, a variation of the employed econometric model delivers a cost reduction of 5.4 percent.

The two multi-state studies, Boden and Fleischman [1] and Pozzebon [15], find no evidence for cost containment effects of physician fee schedules in workers compensation. The study by Boden and Fleischman provides an extensive analysis of the medical costs of workers compensation claims in 43 jurisdictions over the time period 1965 through 1985. The data were obtained from the National Council on Compensation Insurance, independent rating bureaus, and exclusive state funds. In a statistical analysis, the authors correlate for 41 jurisdictions the average annual rate of medical cost growth with the presence of a fee schedule. The measured correlation, although negative, is weak. Boden and Fleischman state that this lack of a statistical relation does not provide evidence for fee schedules having no effect and call for a more elaborate statistical analysis.

The approach pursued by Pozzebon [15] differs from the six discussed cross-state studies in that it makes use of individual claims data, as opposed to aggregates. The analyzed set of workers compensation claims consists of a random sample of 31,707 claims, which was drawn from a total of 316,280 claims recorded in the NCCI DCI database. Only claims that could be considered closed 18 months after the date of injury were included in the sample. The observations range from injury years 1979 through 1987 and comprise 17 jurisdictions. The author reports eight versions of a time series–cross section model, all of which establish a statistically significant influence of physician fee schedules on the medical costs of workers compensation claims. In seven specifications, this influence is positive; the estimated percentage effect varies between 6.2 percent and 10.5 percent. Then again, in one specification, this influence is negative, measuring minus 6.8 percent. According to the author, the study suggests that physician fee schedules (among other measures of cost control) are not effective in containing the medical costs of workers compensation claims.

Robertson and Corro [21] use the least granular data of all cross-state studies when comparing individual medical transactions in workers compensation to Group Health. The data set of medical transactions related to workers compensation claims was provided by a set of insurance carriers. The Group Health data were obtained from MEDSTAT Group, Inc. (now known as Thomson Medstat) and comprise medical transactions related to PPOs, HMOs, and traditional health care plans. The analysis covers the time period 1997 through 2004 and encompasses 14 jurisdictions. Robertson and Corro show that in the jurisdictions without a fee schedule (Illinois, Indiana, and Tennessee), medical prices in workers compensation have the highest markup over Group Health.

An early single-state study on the effect of fee schedule introductions was sponsored by the Minnesota Department of Labor and Industry [13] and released in 1990. Further single-state studies were published in 1994 by Roberts and Zonia [20] on Michigan, and in 2010 by Radeva et al. [17] [18] on Illinois and Tennessee.

The 1990 study sponsored by the Minnesota Department of Labor and Industry [13] contrasts the costs of medical care associated with the workers compensation claims of a large insurance carrier to the costs of medical care experienced by Blue Cross/Blue Shield Minnesota for similar injuries. Further, this research investigates how medical costs per claim and charges for certain types of injury in the Minnesota workers compensation system compare to other states. The study concludes that there is no evidence that the physician fee schedule in Minnesota had a "significant direct impact on medical expenditures." The authors acknowledge that the "ineffectiveness is due in large part to faulty design of the medical fee schedule. Strategies that may increase effectiveness include limiting reimbursements to hospitals and mandating utilization review."

Roberts and Zonia [20] analyze the introduction of a physician fee schedule in Michigan. This fee schedule became effective in August 1989, along with a utilization review process. The authors compare the medical costs of workers compensation claims of the post-implementation time period October 1, 1989 through March 31, 1990 to the corresponding six-month time window in the prior year. The data set was provided by an insurer that accounted for about 17 percent of premium dollars in the voluntary workers compensation market in this jurisdiction at the time. Only claims for which the medical treatment started in the stated time intervals were included. The sample size in the pre-implementation time window equaled 6,997, compared to 9,620 in the time interval following the implementation. Roberts and Zonia find that the median medical costs of workers compensation claims were 25.7 percent higher in the post-implementation time window; by comparison, from October 1988 to March 1990, the medical care component of the Consumer Price Index (CPI) increased by only 12.4 percent. The authors conclude that there is no evidence that the implementation of a physician fee schedule in Michigan decreased the medical costs of workers compensation claims.

In two studies, Radeva et al. [17] [18] analyze the impact of the introduction of physician fee schedules in Illinois and Tennessee. The Tennessee fee schedule took effect in July 2005; the Illinois fee schedule became operational in February 2006. In both jurisdictions, the physician fee schedule was part of a comprehensive set of policy actions concerning the reimbursement of medical care provided by physician and hospitals in the context of workers compensation. The authors quantify the impact of these legislative reform packages using workers compensation claims with more than seven days of lost time. For Tennessee, the authors are able to show that, following the fee schedule introduction, the medical costs per claim stabilized at the pre-implementation level. By comparison, during the three years prior to this legislative action taking effect, the medical costs per claim had increased at an annual rate of 7 percent, on average. Similarly, for Illinois, the authors find that the rate of growth of medical payments per claim slowed to 5 percent from the 12 percent annual average that was observed during the three years leading up to the fee schedule.

None of the discussed research publications provides persuasive findings of a systematic effect of fee schedule introductions on the utilization of medical care in workers compensation. Of the cross-state studies, only Borba [2] and Robertson and Corro [21] investigate the degree to which the impact of fee schedules on the cost of medical care may be due to differences in utilization. Borba reports that the number of office visits is higher in fee schedule states than in non-fee schedule states but cautions that this finding may be related to differences in compensability associated with waiting periods. Robertson and Corro, who study a dozen common workers compensation injuries for 14 jurisdictions, find no evidence of an impact of fee schedules on utilization.

Among the mentioned single-state studies, Roberts and Zonia [20] and Radeva et al. [18] address possible utilization effects of fee schedule introductions. Roberts and Zonia discover a decrease in the number of procedures to treat patients when comparing the post-implementation time window to the pre-implementation time interval surrounding the Michigan fee schedule implementation. Although this decrease in the supply of medical care is not statistically significant, in the context of a statistically significant decrease in the duration of treatment, the finding points to a reduction in utilization. Then again, the authors document a "dramatic change … in the use of procedures for which there was no specific maximum fee in the fee schedule." Finally, Radeva et al. state that, following the Tennessee fee schedule introduction, "utilization by major nonhospital services … did not change significantly."

1.2 Objective

The objective of this study is to evaluate the impacts of fee schedule introductions on the medical costs of workers compensation claims. The medical costs of these claims are measured using a concept of contemporaneous severity. This concept of severity encompasses both the price and the quantity of medical services consumed within a given time window, where the consumption is normalized by the number of claims active during this time window. As a consequence of the contemporaneous nature of the severity measure, changes in the consumption of medical services associated with variations in claim duration are not accounted for.

This being an event study on a small set of jurisdictions, the impact of fee schedule introductions is measured on an unconditional basis. Specifically, the impact is not conditioned on the legislative environment of the jurisdiction, its demographic characteristics, or its economic conditions. Most important, the analysis does not attempt to isolate the contribution of the physician fee schedule to the efficacy of the whole body of cost containment measures; some of these measures may have been introduced alongside the fee schedule or may have been in place at the time this legal stipulation became effective. Among such cost containment initiatives may be mandatory utilization reviews, restrictions on physician choice, and hospital fee schedules.

1.3 Outline

What follows in Section 2 is a description of the data on fee schedules and medical transactions that are employed in this study. Section 3 outlines the severity, price, and utilization indexes used to gauge the responses to the fee schedule introductions; further, this section introduces the concept of price departure and offers a description of the time series model that establishes the counterfactual, defined as the situation that would have been observed absent the fee schedule introduction. Next, Section 4 presents the measured effects of the fee schedule introductions for the two analyzed states. Section 5 concludes.

2. THE DATA

The study uses data from two sources. First, physician fee schedules were obtained from Ingenix Inc. (now known as OptumInsight, Inc.) and employed in the analysis as stated by this data provider. Second, the effects of fee schedule introductions were measured on a large set of medical transactions associated with workers compensation claims. These transaction records cover the time period January 1, 2000 through December 31, 2010 and were provided by a set of insurance carriers.

The jurisdictions analyzed in this study are Tennessee and Illinois. The jurisdiction state criterion and provider zip code information are used when linking medical transactions to a given state. Neither of the two states operates a state fund; both Tennessee and Illinois operate second injury funds, which are not associated with significant numbers of transactions and not included in the data set. As measured by written premium, for the year 2006, the market share of the carriers contributing to the medical transaction data set equals 30.7 percent for Tennessee and 32.8 percent for Illinois. In this market share computation, the self-insured are excluded.

The data set excludes transactions associated with medical services provided by hospitals and ambulatory surgical centers, but includes transactions related to services delivered by physicians (as the provider type) at these places of service. The medical transactions data were edited using expert knowledge on billing and reimbursement practices, and the data set was cleansed using statistical tools of outlier detection. For an overview on the employed data cleansing tools, see the appendix (Section 6).

The medical services associated with the obtained transaction records can be categorized into the American Medical Association (AMA) service categories Evaluation and Management Services, Anesthesia, Surgery, Radiology, Pathology and Laboratory, and Medicine. Transactions related to Anesthesia were excluded from the study due to difficulties in quantifying the units of service that are associated with the individual records. Pathology and Laboratory tends to be sparsely populated and, although included in All Categories, is consequently not studied as an independent service category. Physical Medicine, as a subcategory of Medicine, is defined by Current Procedural Terminology (CPT) codes ranging from 97001 through 97799.

For the purpose of this study, medical services are identified by a combination of CPT code and modifier; only modifiers that are recognized by fee schedules are considered. For transactions associated with a thus identified medical service, the maximum allowable reimbursement (MAR) may vary by geozip (geographic areas identified by arrays of zip codes; Illinois only) and place of service. Where such variation exists, the MAR of a given medical service cannot be read directly from the fee schedule but instead needs to be calculated for any given month as a weighted average across geozips and places of service, where the weights are the number of units of service provided.

For a given medical service, the study recognizes a MAR only if the fee schedule specifies a dollar amount (dubbed fixed-value MAR), as opposed to defining the MAR as usual and customary or in percent of billed charges, for instance. When a fee schedule change occurs mid-month, for the purpose of calculating the average MAR of a given medical service for that month, the pertinent fee schedules are prorated based on the numbers of units of service provided under each regime.

Because the MAR that applies to a given medical transaction may differ with the place of service and (in Illinois) across geozips, the average monthly MAR of a given medical service may vary over time for a given fee schedule. This is because the distribution of transactions by place of service and geozip may change from month to month.

3. RESEARCH FRAMEWORK

The impact of fee schedule introductions is analyzed in a time window surrounding the fee schedule effective date. The counterfactual is established using a time series model that is calibrated to a multi-year pre-implementation time period. This way, possible changes in behavior associated with the anticipation of the fee schedule on the part of claimants, physicians, and insurers have only a dampened impact on the measured effect.

3.1 Price, Utilization, and Severity Indexes

The impact of fee schedule introductions is read from indexes. These indexes are on a monthly basis and calculated for the mentioned AMA categories and, these service categories taken together, for All Categories. Further, these indexes are computed for Physical Medicine, which constitutes a subset of the AMA category Medicine.

Price indexes are calculated for two sets of prices. First, there is a price index calculated from reimbursed amounts, encompassing all medical services. Second, there is a price index based on the fee schedule, comprising only medical services subject to a fixed-value MAR.

Both price indexes are Fisher indexes—prices and quantities are measured on the level of the medical service. Technically, the Fisher index is the geometric mean of the Laspeyres and Paasche indexes. The Laspeyres index compares the set of prices of the current month to the set of prices of the previous month, using as weights the quantities of medical services of the previous month. The Paasche index undertakes this comparison by using as weights the quantities of the current month.

The formula for the Laspeyres index reads

$$P_L = \frac{\sum_i p_i q_{i-1}}{\sum_i p_{i-1} q_{i-1}} \quad . \tag{1}$$

By comparison, the equation for the Paasche index is given by

$$P_P = \frac{\sum_i p_i q_i}{\sum_i p_{i-1} q_i} \ . \qquad (2)$$

In the equations above, swapping prices for quantities and vice versa delivers the Laspeyres and Paasche quantity indexes. Analogous to the Fisher price index, the Fisher quantity index is obtained as the geometric mean of the Laspeyres and Paasche quantity indexes. For details on price and quantity indexes, see International Labour Office [10].

The Fisher index has a host of desirable properties. The most important property of the Fisher index in the context of this study is its ability to break down accurately the price and quantity responses to changes in relative prices. For instance, if the structure of consumed medical services changes in response to some services increasing in price less than others, then the Laspeyres (Paasche) price index overestimates (underestimates) the rate of inflation in the event that medical consumption shifts toward services that have increased in price comparatively less. This bias occurs because the Laspeyres index is based on the quantities of the previous month, which do not yet account for the shift in consumption toward the services that have become comparatively less expensive. Conversely, the Paasche index draws on the new quantities, thereby biasing the reading of the rate of inflation in the opposite direction.

The Fisher quantity index corresponds to the Fisher price index at reimbursed amounts—both concepts encompass medical services regardless of their fee schedule treatment. The product of these two Fisher indexes is related to the transaction volume. Specifically, the percentage change from the previous month of the product of the Fisher price and quantity indexes equals the percentage change in the transaction volume on which these indexes are calculated.

A utilization index is calculated by normalizing the Fisher quantity index by the number of active claims in the applicable month. This utilization index is again calculated for the mentioned service categories and for All Categories. Note that the number of active claims in a given month may vary across service categories. Specifically, a claim is considered active in a given service category (in All Categories) if there is a transaction associated with this claim in this service category (in any service category) that enters the price index for this service category (for All Categories) in that month.

A severity index is calculated as the product of the Fisher price index at reimbursed amounts and the utilization index. The following relation holds among the rates of severity growth (s), the rate of inflation (p), and the rate of utilization growth (u):

$$s = (1+p) \cdot (1+u) - 1 \ . \qquad (3)$$

The utilization index represents a contemporaneous concept of utilization, since it builds only on transactions that have been observed in the month under consideration. To the degree to which the consumption of medical services provided by physicians is front-loaded in the lifetime of a claim, the utilization index (and, hence, the severity index) may increase in response to an influx of claims. This property of the utilization index may explain some of the seasonal variation that the utilization and the severity indexes display over the course of the calendar year. Seasonality in the utilization of medical services is to be expected, given the influence of climactic conditions on economic activity, in particular in the construction and leisure and hospitality industries. Finally, to the degree to which there is front-loading of medical services in the lifetime of a claim, the utilization index may decrease in response to a systematic increase in claim duration.

In Tennessee, in Physical Medicine, the MAR of a given medical service may be regressive with respect to the number of units of service provided to the claimant. Due to data limitations, this regression in MAR was not factored into the computation of the price index at fee schedule.

3.2 Price Departure

Typically, medical services provided by physicians are reimbursed at or below fee schedule, although jurisdictions vary by the degree to which the stipulated price ceilings are enforced. For instance, in Tennessee, the Commissioner may assess civil penalties for fee schedule violations at his discretion. By comparison, in Illinois, reimbursement above fee schedule is permitted when agreed to pursuant to a written contract.

Price departure measures the deviation of reimbursement from fee schedule. Specifically, price departure is defined as the ratio of transaction volume at reimbursed amounts to transaction volume at fee schedule, minus 1. In the denominator, for transactions associated with medical services that are not subject to a fixed-value MAR, reimbursed amounts substitute for the fee schedule. Medical services that are not subject to a fixed-value MAR are subject to price ceilings that are defined as a percentage of charges or are subject to usual and customary reimbursement (which may be defined as a function of charges). The transaction volume that contributes to the price departure calculation is slightly more comprehensive than the transaction volume that finds its way into the computation of the price and quantity indexes. This is because medical services enter the price (or quantity) index of a given month only if there is at least one transaction associated with these services in both the current month and the prior month.

Similar to the price, utilization, and severity indexes, price departure is calculated on a monthly basis for All Categories, the mentioned AMA categories, and Physical Medicine. To provide a numerical example, a price departure of minus 0.05 (or, equivalently, a negative 5 percent) states that the reimbursed amounts are 5 percent below fee schedule, on average.

Changes in the difference between the price indexes at reimbursed amounts and at fee schedule do not map exactly into changes in price departure. For instance, for a given array of prices, the price index does not respond to changes in quantities. By contrast, the price departure may increase for given reimbursement practices and a given fee schedule if the distribution of transactions shifts in favor of medical services that exhibit a comparatively large spread between MAR and reimbursed amount.

For Tennessee, in keeping with the computation of the price index at fee schedule, the regressive nature of the fee schedule in Physical Medicine was not factored into the computation of the price departure.

3.3 Time Series Model

The impact of a fee schedule introduction is measured by the difference between the observed severity index net of noise and the severity index that is estimated to have materialized absent the fee schedule. This aggregate impact is broken down into the price and utilization responses. The utilization response is again measured by the difference between the observed value net of noise and the counterfactual. After establishing the severity and utilization responses in this way, the price response is obtained as a residual. No prorating of the responses is necessary because, in both jurisdictions, the fee schedule introduction took effect on the first day of the month.

Visual inspection of the severity and price index charts indicates that the responses of these indexes to the fee schedule introductions have come to completion by the third post-implementation month; this holds for both analyzed jurisdictions and across all service categories. This short time lag in the response to fee schedule introductions agrees with the response to fee schedule changes documented in Schmid and Lord [23]. According to these authors, the bulk of the price and severity adjustments to fee schedule changes occur within the first two months of the new fee schedules having taken effect.

For the severity and utilization indexes, observed index values net of statistical noise and counterfactual index values are established using a univariate time series model. The statistical model is applied to the respective index through the final pre-implementation month for the purpose of generating forecasts for the post-implementation months. The third month of the forecast time period establishes the counterfactual of interest. Next, the statistical model is applied to the post-implementation observations, starting in the third month of the fee schedule operation. The first fitted value in this post-implementation time series delivers the observed value net of noise that is compared to the counterfactual.

The time series model employed in this analysis is a univariate exponential smoothing state-space model called TBATS. This model was developed by De Livera, Hyndman, and Snyder [5] and is available as part of the R package *forecast*. TBATS stands for Trigonometric, Box-Cox transform, ARMA errors, Trend, and Seasonal components, which summarizes the basic characteristics of this model. The seasonal components of the model are based on a trigonometric specification, which distinguishes this approach from the competing BATS model—the trigonometric specification limits the number of parameters in the seasonal components of the model. The number of seasonal periods was set to 12, thus acknowledging the monthly nature of the time series. Auto-regressive moving average (ARMA) errors were permitted, as well as the Box-Cox transform (which corrects for possible non-normality). For the purpose of forecasting, trend dampening was employed, although this may have little impact, given the short forecasting horizon of only three data points. Automatic model selection was employed using the Akaike information criterion (AIC).

4. IMPACT OF PHYSICIAN FEE SCHEDULE INTRODUCTIONS

This section presents the impacts of physician fee schedule introductions on the price, utilization, and severity indexes of medical care provided by physicians in the context of workers compensation. These impacts are discussed in their overall effect (All Categories) and, for illustration, in their effect on an individual service category or subcategory. For Tennessee, this category is Evaluation and Management Services and, for Illinois, it is Physical Medicine. The monthly price, utilization, and severity indexes run from January 2000 through December 2010. The fee schedule introductions in the two states occur approximately in the middle of the observed 11-year time period. Further, the events of interest occurred well ahead of the economic downturn that started in December 2007 (www.nber.org/cycles.html).

4.1 Tennessee

In 2004, the Tennessee Workers' Compensation Act underwent a sweeping reform. As a result of this legislative action, a physician fee schedule was introduced, effective July 1, 2005. This fee schedule became mandatory on January 1, 2006.

The Tennessee fee schedule was designed as a multiple of the applicable Medicare fee schedule, where the multiplicative factors vary by procedure. For repeat services in occupational and physical therapy, these multiples regress to unity, thereby gradually lowering the price ceiling to 100 percent of the amount reimbursable under Medicare. Reimbursement of procedures not explicitly covered by the fee schedule is limited to the lesser of usual and customary and 100 percent of Medicare. Prior to the fee schedule introduction, reimbursement was subject to usual and customary practices, as defined by the Medical Care and Cost Containment Committee.

As implemented, the Tennessee fee schedule updated annually based on the Medicare Economic Index, upon review by the Commissioner. Having been tied to Medicare in this way, the MAR for a given medical service could change without further legislative action. The first Medicare-related update of the Tennessee physician fee schedule took effect on January 1, 2006.

As part of the 2004 legislative reform, Tennessee limited provider choice. It is mandatory that the employer offer a panel of three physicians. From this list of providers, the employee has the privilege to choose one. The physician may be changed during ongoing treatment only with consent from both the employee and the employer. The employer has to give proof of the existence of the panel by signing a legal form, a copy of which has to be provided to the employee. This provision became effective for injuries on or after July 1, 2004.

The Tennessee data set contains 66 observations prior to the fee schedule implementation, and an equal number following it. All pre-implementation observations feed into the time series model that establishes the counterfactual for the third post-implementation month. All but the first two post-implementation observations are employed in the time series model that eliminates the statistical noise from the observed severity and utilization indexes in the third month of fee schedule operation.

4.1.1 All Categories

Chart 1, top panel, depicts the Fisher price indexes of medical services, at reimbursed amounts and at fee schedule. As mentioned, the price index at fee schedule is based on the fixed-value MAR. When no fixed-value MAR is available (for instance, because the medical service is subject to usual and customary reimbursement or the price ceiling is defined as a percentage of billed charges), this medical service is not involved in the computation of the price index at fee schedule.

In Chart 1, top panel, the first box (June 2005) associated with the price index at reimbursed amounts indicates the final month prior to the fee schedule implementation; the second box (September 2005) points to the third month of fee schedule operation. For this third post-implementation month, the impact of the fee schedule is measured by the difference between the counterfactual and the fitted value.

The displayed price index at fee schedule starts in the third post-implementation month and, for that month, this index is normalized to the price index at reimbursed amounts. If, subsequently, the price index at reimbursed amounts rises above the price index at fee schedule, this implies that since that first common month, on average, the inflation rate of the former has exceeded the inflation rate of the latter (as measured by the compound annual growth rate). As mentioned, the price index at fee schedule may change from month to month, even as the fee schedule does not. This is because the distribution of transactions by place of service may vary over time.

Chart 1: Tennessee, Price Indexes and Price Departure, All Categories

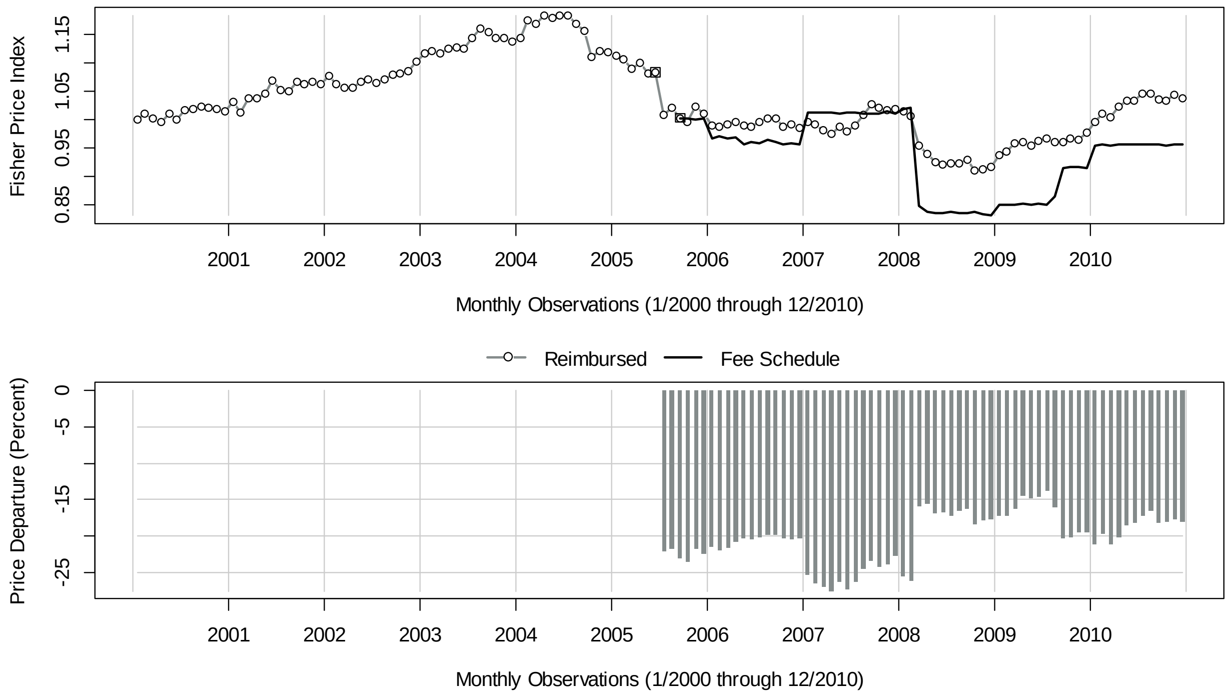


Chart 1, bottom panel, displays the price departure for All Categories, starting in the month the fee schedule took effect. As discussed, the price departure is calculated as the ratio of transaction volume at reimbursed amounts to transaction volume at fee schedule, minus 1. For a given fee schedule, an increase in the price index at reimbursed amounts tends to diminish the magnitude of the price departure, thereby shortening the depicted bars. Yet, as pointed out, changes in the price departure do not perfectly correlate with changes in the vertical distance between the price indexes.

The chart reveals that between the July 2005 fee schedule introduction and the end of 2006, the price departure averaged around minus 21 percent. Then, in January 2007, following a fee schedule increase, the price departure deepened to a value beyond a negative 25 percent. As the reimbursed amounts rose in mid-2007, the price departure retreated slightly. In early 2008, in the course of a major fee schedule decrease, only about half of the stipulated reduction in MAR manifested itself in a decline of the price level at reimbursed amounts. As a result of the limited adjustment, the price departure shrank to about a negative 16 percent.

The fee schedule reduction in early 2008 holds a lesson regarding the interaction between policy decisions and pricing behavior. Following a muted price level response to the 2008 fee schedule reduction, the fee schedule was raised in the second half of 2009. Clearly, during the time period

2008 through 2009, the price level at reimbursed amounts was more stable than the price ceilings imposed by policy actions. Further, the episode shows that the relation between fee schedules and reimbursed amounts is reciprocal.

Chart 2 depicts the severity and the utilization indexes in two panels. The panels, which are drawn on the same scale, display the observed, raw data (light gray) alongside the fitted values (black) generated by the time series model. The time window surrounding the fee schedule introduction is highlighted by a gray box. At the center of this box are the forecasts (dark gray), which are the estimated values in the first three post-implementation months for the hypothetical situation that no fee schedule was introduced. The fee schedule impact is measured by the vertical distance between the final value of these three forecasts and the first fitted value available for the post-implementation time period—both values are from the third month of fee schedule operation. This vertical distance between the observed value net of noise and the counterfactual indicates a 9.3 percent drop in severity (top panel) and a comparatively small 2.0 percent decrease in utilization (bottom panel). As implied by Equation (3), the decline in the price index at reimbursed amounts equals 7.4 percent.

The severity index displayed in Chart 2, top panel, displays a fair amount of statistical noise, which adds to the uncertainty associated with the estimated severity response. Thus, a sensitivity analysis is performed where the severity response is backed out of the observed change in the price level and the estimated utilization effect. The observed change in the price level is obtained by comparing the price level of the third post-implementation month to the price level of the final pre-implementation month; the price level response thus measured equals minus 7.4 percent, which (as a rounded number) equals the price level response established above. By way of Equation (3), based on unrounded numbers, the severity response amounts to a negative 9.2 percent, which is nearly identical to the previously established value.

Chart 2: Tennessee, Severity and Utilization Responses, All Categories

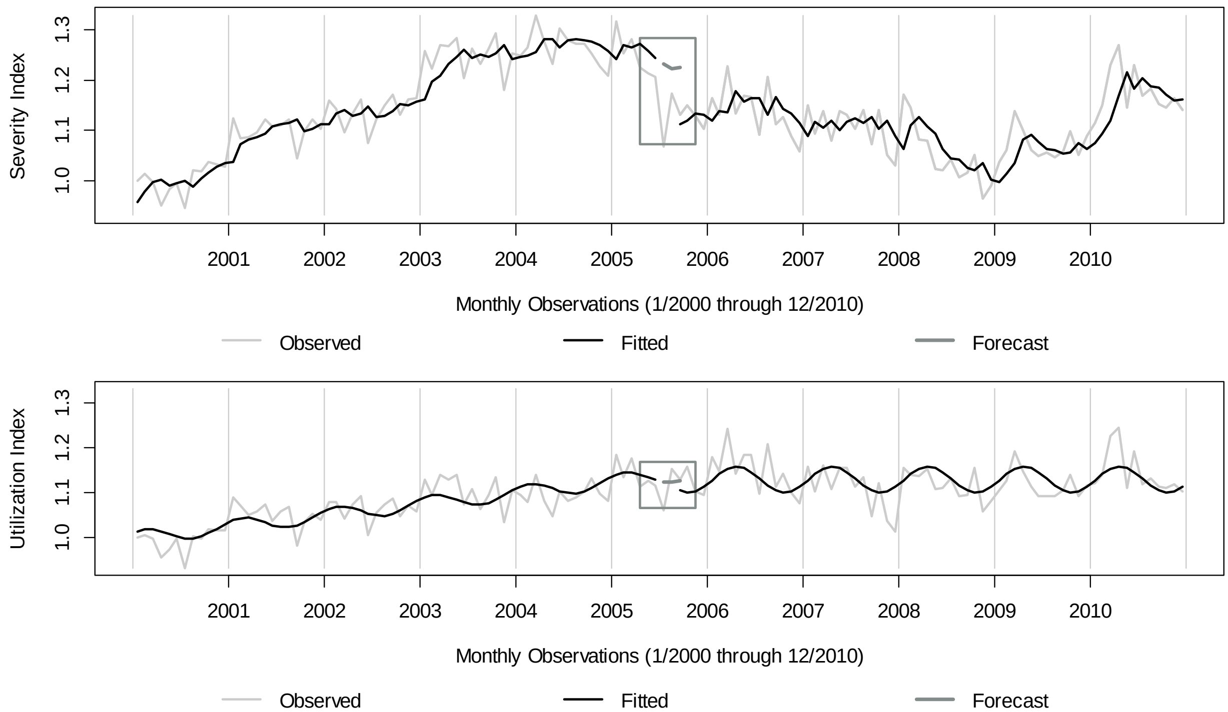


Chart 2, bottom panel, displays the utilization index, which exhibits a pronounced seasonality. Because the severity index is the product of the utilization index and the price index, the seasonality inherent in the utilization index carries over to the severity index (top panel).

Aside from this seasonal variation, the utilization index remains flat following the fee schedule introduction. This absence of a post-implementation utilization trend is in stark contrast to the steady utilization increase during the years leading up to the fee schedule. The estimated drop (compared to the counterfactual) and subsequent ebbing of the utilization index may be more related to the restriction on physician choice introduced in mid-2004 than to the fee schedule. Empirical studies by Durbin and Appel [4] and Neumark, Barth, and Victor [15] provide clear evidence of the cost containment effect of restrictions on provider choice. However, these studies focus on the overall cost impact and do not explicitly address the contribution of utilization changes.

For the purpose of comparing fee schedule prices to pre-implementation reimbursement at large, the price departure was calculated for the final pre-implementation month on an *as and if implemented* basis. This concept of *ex ante* price departure is analogous to the price departure displayed in the bottom panel of Chart 1 in that reimbursed amounts again substitute for the fee schedule where medical services have not been assigned a fixed-value MAR. With the *ex ante* price departure calculated in this way, in the final pre-implementation month, the price level at reimbursed amounts was 15.7 percent below the price level at the impending fee schedule. By comparison, the price level at reimbursed amounts was 23.1 percent below the price level at fee schedule in the third month of fee schedule operation—the *ex post* price departure. The differential between *ex ante* and *ex post* price departures is attributable to the decline in the price level at reimbursed amounts.

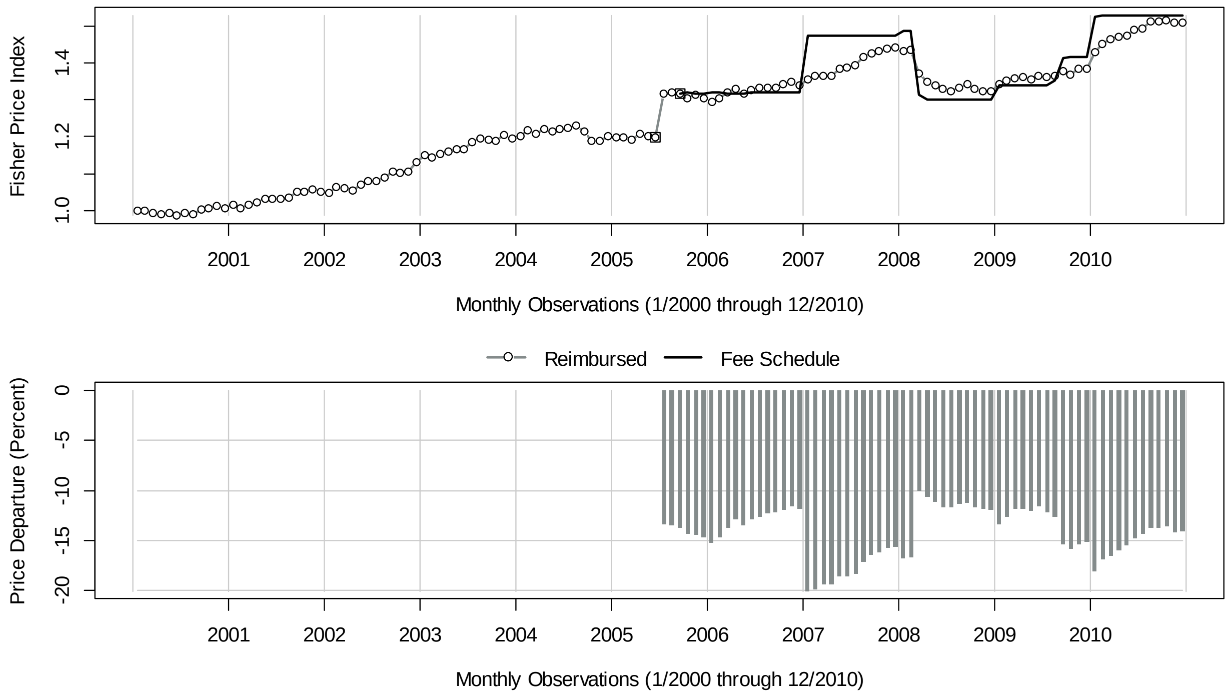
It is worth noting that the share of the transaction volume of the final pre-implementation month that was going to be subject to the fee schedule equaled 95.4 percent. In keeping with existing methodology, in this computation, only medical services that were to receive a fixed-value MAR were regarded as subject to the impending fee schedule.

4.1.2 Evaluation and Management Services

The impact of the fee schedule introduction on an individual service category is exemplified on Evaluation and Management Services. The analysis performed on this service category is analogous to the inquiry undertaken for All Categories. When it comes to normalizing the quantity index by the number of active claims for a given month, only claims for which there were transactions recorded in Evaluation and Management Services (in that month) were considered. In this service category, the share of the transaction volume of the final pre-implementation month that was going to be subject to the fee schedule amounted to 97.1 percent.

Chart 3 exhibits for Evaluation and Management Services the price index at reimbursed amounts and at fee schedule (top panel), along with the price departure (bottom panel). In this service category, contrary to the discussed All Categories (and all other service categories), the price level at reimbursed amounts increased following the fee schedule implementation. This price level increase measures 11.0 percent or, based on the discussed sensitivity analysis, 10.0 percent.

Chart 3: Tennessee, Price Indexes and Price Departure, Evaluation and Management Services



For a given medical service, in a given month, the reimbursed amounts are not necessarily uniform. Instead, these amounts describe a distribution, which frequently is multimodal. The price ceiling of a newly introduced fee schedule may pose a binding constraint for the top percentiles of the distribution or reimbursed amounts. By means of capping the top percentiles at the fixed-value MAR, the fee schedule lowers the average reimbursed amount, all else being equal. From this perspective, the introduction of a fee schedule is expected to depress the average reimbursement for a given medical service or, more broadly, the price index at reimbursed amounts of a given service category. But then, in Evaluation and Management Services, the price index at reimbursed amounts increased. Given that this service category accounts for 25.6 percent of the volume in the final pre-implementation month, this finding can hardly be treated as an anomaly in the data.

A possible explanation of the atypical response of the price level for Evaluation and Management Services may be found in the heuristic of anchoring discussed in psychology. In this cognitive process, an individual adjusts its numerical judgment or attitude toward a value that it considers as a candidate answer. As argued by Kahneman, Ritov, and Schkade [11], "anchoring effects are among the most robust observations in the psychological literature." These authors identify two "necessary and apparently sufficient conditions for the emergence of anchoring effects." One condition is "the presence of some uncertainty about the correct or appropriate response." The other condition is "a procedure that causes the individual to consider a number as a candidate answer." In the context of physician fee schedules, the introduction of a MAR may serve as a candidate answer to the question of appropriate reimbursement.

Clearly, anchoring is not the only process that is capable of generating the observed price level response in Evaluation and Management Services. In industrial economics, it has been argued that price ceilings serve as focal points that solve the coordination problem inherent in tacit collusion. The concept of focal points has been introduced by Schelling [22], who showed how mutually recognized signs are able to facilitate solutions in coordination games. In the context of physician fee schedules, the MAR may serve as a focal point for the coordination of billing behavior among physicians. In a study of the credit card market in the 1980s, Knittel and Stango [12] argue that price ceilings solved the coordination problem that made tacit collusion possible. Then again, there are doubts that price ceilings can serve as focal points. For instance, Engelmann and Normann [6] analyze in an experimental setting whether price ceilings have a collusive effect. Based on results established in laboratory markets, the authors reject the focal-point hypothesis, according to which price ceilings solve the coordination problem associated with tacit collusion. The experiments show that markets with price ceilings have lower prices than markets without such constraints.

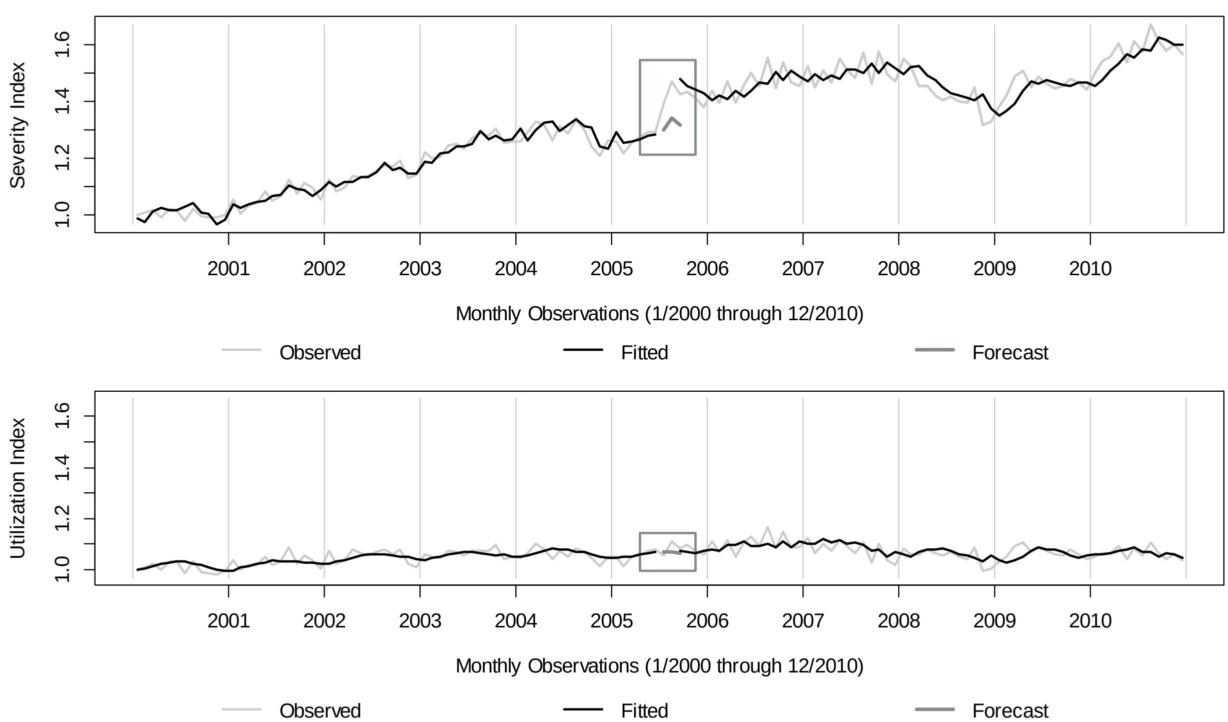
As argued, capping the upper percentiles of the distribution of reimbursed amounts reduces its mean, all else being equal. Conversely, to the extent that the MAR anchors the judgment regarding an appropriate reimbursement, the distribution of reimbursed amounts moves toward the MAR, thereby increasing the mean of this distribution. The net effect of these two opposing forces determines whether the average reimbursement for a given medical service or the overarching price index of the applicable service category increases in response to a fee schedule introduction.

When compared to All Categories, the *ex ante* price departure in Evaluation and Management Services was extensive, measuring a negative 21.1 percent. This large differential between the center of the pre-implementation distribution of reimbursed amounts and the newly introduced MAR may have afforded the anchoring effect an elevated degree of leverage. The price level increase that followed the fee schedule introduction reduced this differential to an *ex post* price departure of minus 13.7 percent (as measured in the third post-implementation month, according to its definition).

Chart 3 shows that in Evaluation and Management Services, the adjustment of the reimbursed amounts to the fee schedule implementation was swift. In spite of sluggish responses to subsequent fee schedule changes, by the end of the year 2010, the rate of inflation at reimbursed amounts that had materialized since the third post-implementation month was in close proximity to the rate of inflation implied by the stipulated price ceilings.

Chart 4 displays the fee schedule impacts on the severity and utilization indexes. The increase in severity reads 12.1 percent or, using the sensitivity analysis, 11.0 percent. The utilization response is positive but contained, measuring 1.0 percent. Remarkably, for Evaluation and Management Services, the utilization response agrees with classical economic theory, which posits that supply increases with price. Similarly, for Surgery (not shown) and Radiology (not shown), the utilization effect in response to a price level decrease is negative, albeit slender. Assuming that physicians have some discretion over the amount of medical care they provide to a claimant, it can be expected that the supply of these services will increase with the reimbursed amounts (and, correspondingly, decrease in response to a payment reduction). Then again, for the service category Medicine (not shown), the estimated utilization effect in response to a price level increase is positive and is thus at odds with classical economic theory. It is worth noting that at the level of the service category, there is a fair degree of statistical noise in the data, which poses a challenge to measuring an effect that, if it exists, is likely to be small in magnitude. As a result of the elevated level of randomness, the estimated utilization effects for the individual services categories have to be interpreted with caution.

Chart 4: Tennessee, Severity and Utilization Responses, Evaluation and Management Services



A negative utilization effect in response to a price increase agrees with the income targeting hypothesis put forward by Camerer at al. [3] in a study of New York City taxi drivers. Based on this proposition, the supply of services decreases with price since a given income target can be reached with less input. But then, as argued by Schmid and Lord [23] in a study of workers compensation fee schedule changes, there is no indication of income targeting on the part of physicians. More important, Radeva et al. [19], in the context of the Tennessee fee schedule introduction, provide persuasive evidence that, in the category Evaluation and Management Services, physicians respond to changes in reimbursed amounts according to classical economic theory. The authors show that physicians substituted comparatively complex established patient office visits (CPT code 92214) for less complex ones (CPT code 92213). At the same time, the reimbursed amount for the more complex office visit increased by 29 percent, on average, compared to a mere 18 percent for the less complex of these services; the price increases were measured over the time period 2004 through 2006.

Aside from the two competing economic hypotheses, there is also the formative force of best medical practice. Although fee schedule introductions may alter economic incentives in the short term, it is difficult to argue that medical treatments do not revert to best practice in the longer term, especially as treatment patterns and medical technology evolve over time.

4.2 Illinois

The state of Illinois introduced a physician fee schedule in workers compensation effective February 1, 2006. Prior to this legislative provision taking effect, reimbursement was subject to being reasonable and necessary. When the employer did not agree to the expenses, the employee could file a petition asking the Workers' Compensation Commission to decide the disputed issue. For details, see Illinois Workers' Compensation Commission [8].

The Illinois fee schedule set the MAR of medical services to 90 percent of the 80th percentile of charges observed from August 1, 2002 through August 1, 2004. These charges were adjusted for the CPI for the period August 1, 2004 through September 30, 2005.

The MAR varies by geozip, which is defined by the zip codes that share the same first three digits, and by the site where the treatment occurred. Employers or insurance carriers may contract with providers at prices higher than fee schedule. Procedures for which the physician fee schedule does not stipulate a fixed-value MAR, the default reimbursement is set to 76 percent of actual charge.

Prior to the fee schedule introduction, amendments to the Illinois Workers' Compensation Act ended balance billing. As stated in the Fiscal Year 2005 Annual Report of the Illinois Workers' Compensation Commission [9], effective July 20, 2005, "[a] provider cannot hold an employee liable for costs related to non-disputed services for a compensable injury and shall not bill or attempt to recover from the employee the difference between the provider's charge and the amount paid by the employer or insurer on a compensable injury."

In Illinois, an employee may choose any doctor or hospital. The employee is entitled to two choices of medical provider and their respective chains of referrals. These provisions were in place over the entire time window covered by this study; for details, see Illinois Workers' Compensation Commission [7] [8]. More recently, upon approval by the Department of Insurance, the employer may sponsor a Preferred Provider Program (PPP). Where implemented, the PPP is the employee's first provider choice, by default. By opting out of the PPP, the employee exercises his right on a second choice; the employee is then limited to one choice of doctor and subsequent referrals.

The Illinois data set contains 73 observations prior to the fee schedule introduction and 59 observations following it. As with Tennessee, all pre-implementation observations are employed in the time series model that establishes the counterfactual for the third post-implementation month, and all but the first two post-implementation observations enter the time series model that removes the statistical noise from the observed severity and utilization indexes.

4.2.1 All Categories

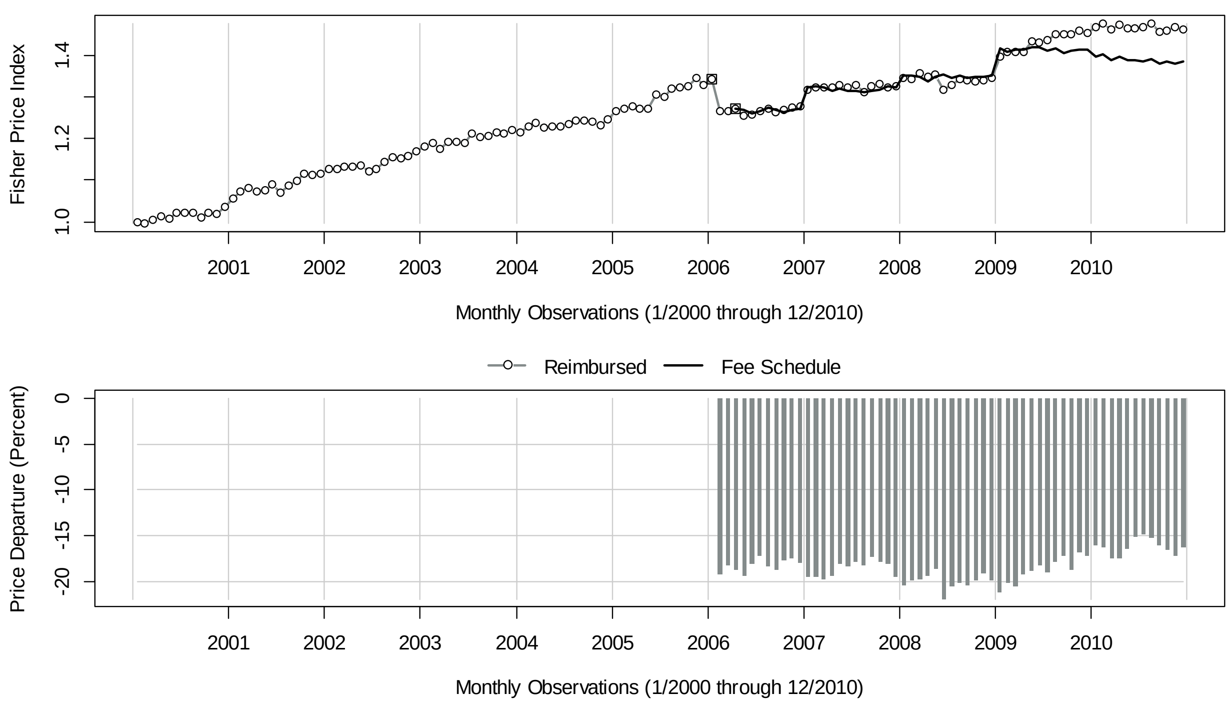
For All Categories, Chart 5, top panel, depicts the Fisher price indexes of medical services, at reimbursed amounts and at fee schedule. Again, the boxes (2006, January and April) indicate the final pre-implementation and the third post-implementation months. And, as before, the price index at fee schedule starts in the third month of the fee schedule operation and is normalized to the price index at reimbursed amounts observed in that month.

As the chart shows, the price index at reimbursed amounts closely tracks the price index at fee schedule during the first four years of the fee schedule operation. Then, several months into the year 2009, the price index at reimbursed amounts starts climbing as the price index at fee schedule begins to soften, thereby opening up a gap between the two indexes. As a result of the emerging difference in the rates of inflation, over the course of the year 2009, the price departure (displayed in the bottom panel) contracted from a value in the neighborhood of about minus 21 percent to approximately minus 17 percent.

Comparing Chart 5 to the corresponding chart for Tennessee (Chart 1) reveals that for a given fee schedule, the price index at fee schedule is less stable in Illinois than in Tennessee. This is because, in Illinois, the fee schedule varies by geozip. As the geographic distribution of transactions varies from month to month, the average MAR for a given medical service may change as a result.

Chart 6 depicts the severity and the utilization indexes in two panels. Again, the two panels are drawn on the same scale and display the observed data (light gray) alongside the fitted (black) and forecast (dark gray) values generated by the time series model. For the third post-implementation month, the vertical distance between the observed value net of noise and the forecast indicates a 4.9 percent drop in severity (top panel). The change in utilization (bottom panel) is essentially nil (at 0.3 percent). By implication, the decline in the price index at reimbursed amounts equals 5.2 percent or, based on the sensitivity analysis, 5.3 percent.

Chart 5: Illinois, Price Indexes and Price Departure, All Categories

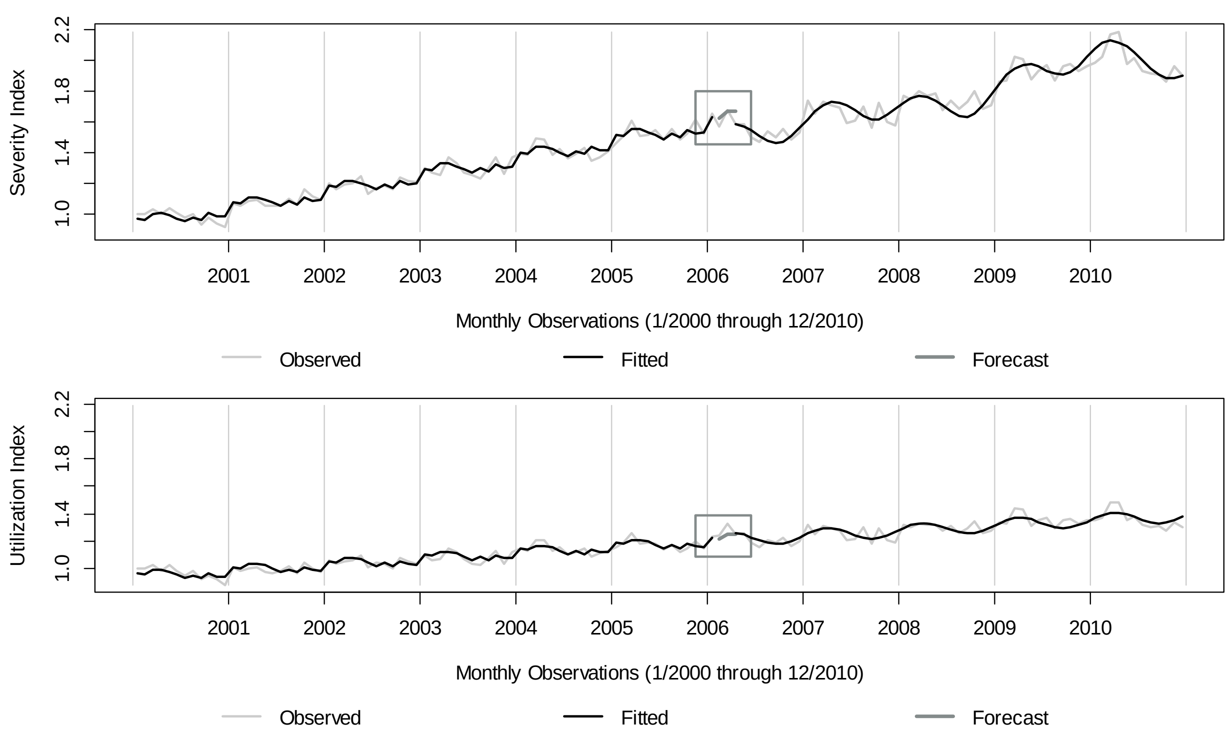


In the final pre-implementation month, the relative deviation of the reimbursed amounts from the MAR of the impending fee schedule—the *ex ante* price departure—ran at minus 14.7 percent. By comparison, the *ex post* price departure came to minus 18.8 percent (as shown in Chart 5, bottom panel). This differential of the *ex post* price departure to the *ex ante* price departure is due to the mentioned 5 percent drop in the price level at reimbursed amounts. Regarding the coverage of the fee schedule, 98.75 percent of the transaction volume in the pre-implementation month was going to be subject to fixed-value MAR.

The utilization index displayed in Chart 6 (bottom panel) exhibits a marked seasonality, similar to what was observed for Tennessee. It is evident from the chart that the seasonality in the severity index is more pronounced in the post-implementation time window. This increase in the amplitude of the seasonal variation is a statistical artifact that arises from the multiplication with increasing price index values. For a higher level of the severity index, a given percentage of seasonal variation translates into a greater magnitude of absolute change. A closer look at the utilization index (bottom panel) reveals no apparent increase in seasonality.

Chart 6, bottom panel, demonstrates that, in Illinois, there is no leveling out of the consumption of medical services (per active claim) following the fee schedule introduction. This lends support to the hypothesis that the receding utilization growth observed in Tennessee is related to the restriction of provider choice, rather than the physician fee schedule.

Chart 6: Illinois, Severity and Utilization Responses, All Categories



4.2.2 Physical Medicine

The impact of the fee schedule introduction is exemplified for Physical Medicine. As mentioned, Physical Medicine is a subcategory of the AMA category Medicine, defined by the CPT code range 97001 through 97799. In this category, the share of the transaction volume of the final pre-implementation month that was going to be subject to the fee schedule equaled 99.5 percent.

Chart 7 exhibits for Physical Medicine the price index at reimbursed amounts and at fee schedule (top panel), along with the price departure (bottom panel). Following the initial price level decrease associated with the fee schedule introduction, the rate of inflation at reimbursed amounts fell short of the rate of inflation at fee schedule, thereby exposing a gap between the two indexes. Then, after being largely flat for Calendar Years 2007 and 2008, the price index at reimbursed amounts rose noticeably in 2009 before softening in the second half of 2010. This surge in the price index in Physical Medicine contributed to the discussed rise in the price index for All Categories (Chart 5).

Chart 7, bottom panel, depicts the price departure for Physical Medicine. The run-up in the price level at reimbursed amounts from early 2009 to about mid-2010 manifested itself in a lessening of the price departure from a negative 15 percent to a mere negative 7.5 percent, approximately.

Chart 7: Illinois, Price Indexes and Price Departure, Physical Medicine

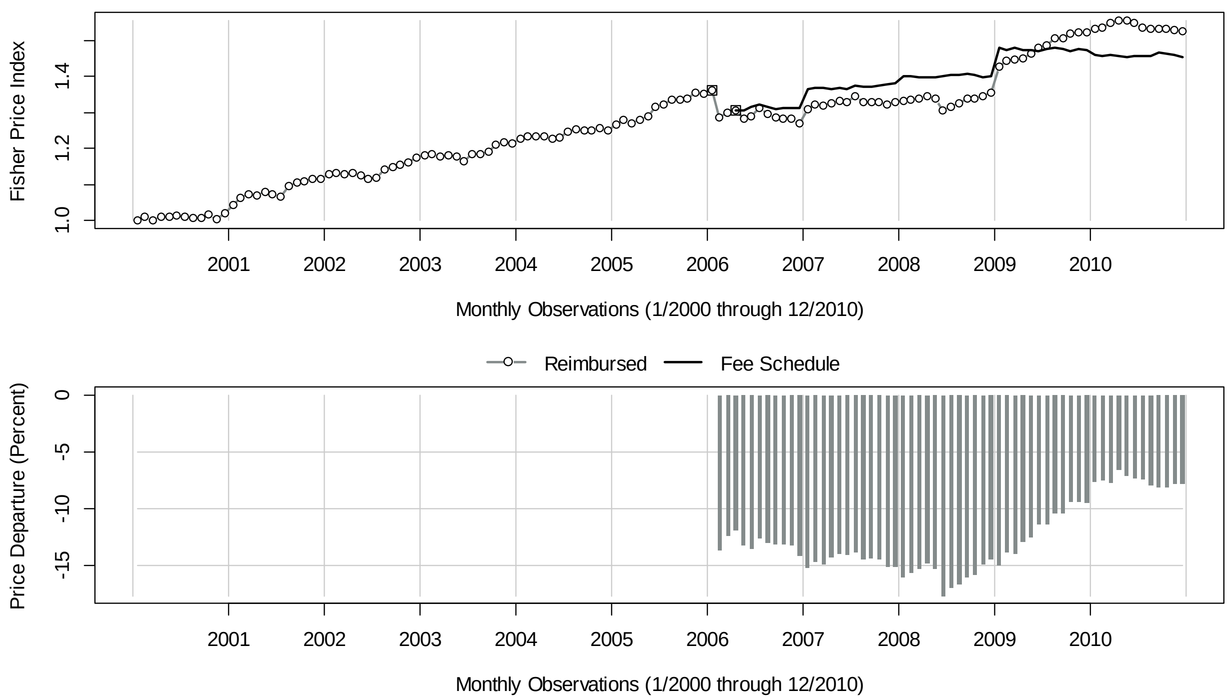
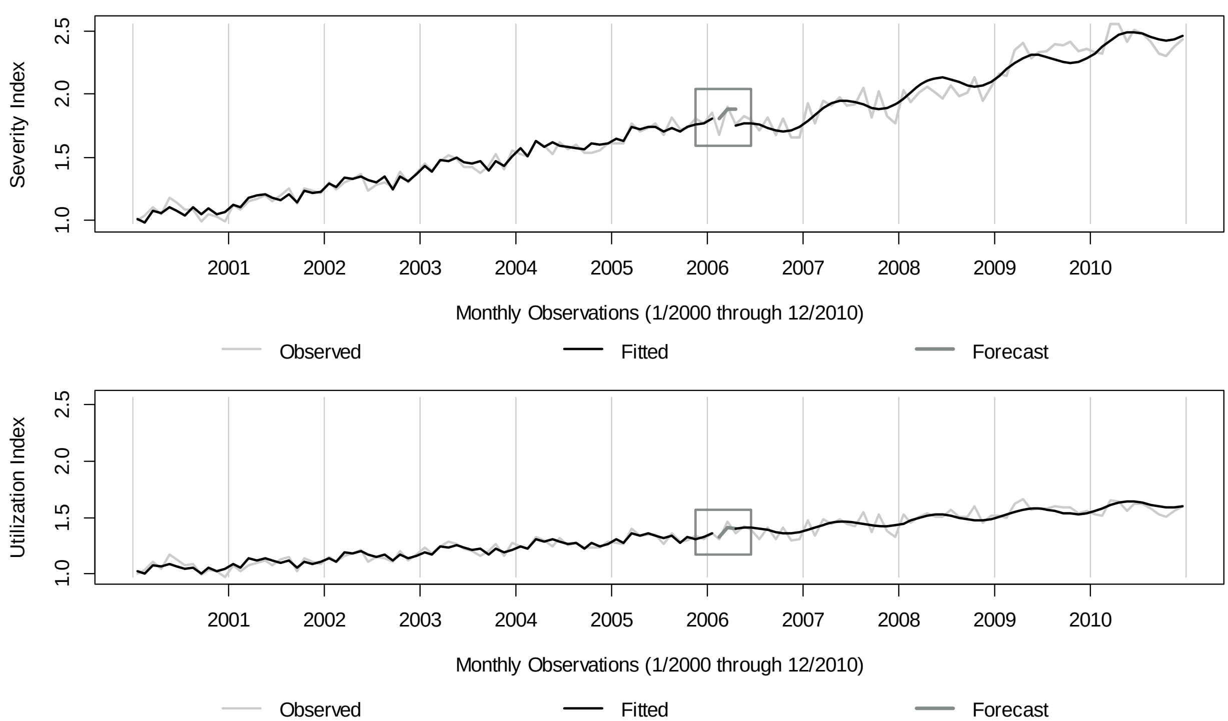


Chart 8 displays the severity (top panel) and utilization indexes. Again, the fee schedule impact is measured by the vertical distance between the final value of the three forecast data points and the first fitted value available for the post-implementation time period—both values are from the third month of fee schedule operation. This vertical distance between the observed value net of noise and the counterfactual indicates a 7.1 percent drop in severity (top panel) and a negligible 0.3 percent increase in utilization (bottom panel). By implication, the decline in the price index at reimbursed amounts equals 7.3 percent. Based on the sensitivity analysis, the price level decrease measures 4.1 percent; the implied severity decline amounts to 3.9 percent.

The insignificant utilization responses in All Categories and Physical Medicine are in accord with similarly immaterial utilization effects (not shown) in Medicine overall, Evaluation and Management Services, and Radiology. Only Surgery displays a material utilization increase, at 8.5 percent; this is in response to a 13.9 percent (sensitivity analysis: 8.6 percent) decline in the price level. It is worth noting that Surgery is a comparatively sparsely populated service category (as measured by number of transactions, not dollars) and thus liable to a fair amount of statistical noise.

Chart 8: Illinois, Severity and Utilization Responses, Physical Medicine



4.3 Summary of Effects

Table 1 summarizes the findings of the event study for All Categories. The table exhibits the responses to the fee schedule introductions of severity, price level, and utilization. Further, the table informs on the *ex post* price departure (defined as relative deviation of reimbursed amounts from fee schedule in the third month of fee schedule operation) as well as the *ex ante* price departure (which, by definition, is computed for the final pre-implementation month).

Although Table 1 does not capture any potential impact of the fee schedule introduction that may have materialized following the third post-implementation, such residual effect may be gauged by studying Charts 2 (Tennessee) and 6 (Illinois).

Both fee schedule introductions elicit negative price level responses—these responses amount to minus 7.4 (7.4) percent for Tennessee and minus 5.2 (5.3) percent for Illinois, where the values in parentheses were obtained in a sensitivity analysis. In Tennessee, the negative utilization effect may be associated with restrictions on provider choice that were implemented a year before the fee schedule introduction. In Illinois, the utilization response is immaterial. As a consequence, the severity responses that can be attributed with confidence to the fee schedule introductions are those associated with the price level.

Table 1: Impact of Fee Schedule Introduction, All Categories

Jurisdiction	Effects (Percent Change)			Price Departure (Percent)		Price Departure (Percentage Points)
	Severity	Price Level	Utilization	*Ex Ante*	*Ex Post*	Difference
Tennessee	–9.3 (–9.2)	–7.4 (–7.4)	–2.0	–15.7	–23.1	–7.4
Illinois	–4.9 (–5.0)	–5.2 (–5.3)	0.3	–14.7	–18.7	–4.0

Note: Values in parentheses pertain to the sensitivity analysis. Calculations are based on unrounded numbers; consequently, the displayed values may differ from those obtained when performing the calculations on the displayed, rounded values. The negative utilization effect in Tennessee may be related to restrictions on physician choice introduced in mid-2004.

The negative price level responses indicate that the hypothesized anchoring effect does not dominate the opposing effect that originates in the price ceilings becoming binding constraints. Then again, as discussed for Evaluation and Management Services in the context of the Tennessee fee schedule introduction, a pronounced *ex ante* price departure may prompt an increase in the price index at reimbursed amounts.

The *ex ante* price departures of the two jurisdictions are remarkably similar, measuring about minus 15 percent. Yet, there is a considerable difference in the *ex post* price departures. Whereas for Tennessee, the difference between the *ex ante* to *ex post* price departures measures 7.4 percentage points, for Illinois, it is 4.0 percentage points.

The absence of a significant utilization response allows for a straightforward evaluation of fee schedule introductions, because it enables the decision-maker to focus on the price level impact. For an *ex ante* price departure in the neighborhood of minus 15 percent, the effect of the fee schedule introduction can be summarized by a price decline of between 5 and 7.5 percent. As an alternative, the effect of the fee schedule introduction can be summarized in a drop in the price departure (from the *ex ante* value to the *ex post* value) by between 4 and 7.5 percentage points.

The discussed reduction in the price level informs on the initial, short-term effect of the fee schedule introduction. The long-term, permanent effect is more difficult to establish, if only because there is more than one way of measuring it. The long-term effect of the fee schedule introduction may be related to the permanence of the initial effect on the price level. Alternatively, the long-term effect may be tied to a lasting reduction in the rate of inflation (compared to the rate that would be observed otherwise). According to this latter definition, the long-term effect includes and extends beyond the permanence of the initial price level decrease. It is important to note that a permanent non-increase in the rate of inflation (compared to what would be observed otherwise) is a sufficient condition for the initial effect on the price level to be permanent.

It could be argued that the permanence of the initial price level effect can be read from the behavior of the post-implementation price departure. Yet, even if the price departure remained at the *ex post* level (apart from short-term fluctuations), this would not constitute evidence of the initial effect being permanent. This is because of the possibility of regulatory capture, as put forward by Stigler [24]. According to this theory of public choice, the regulated (here, the physicians) find ways of employing the regulators to their own economic benefit. Specifically, as a result of regulatory capture, the fee schedule might become so accommodating over time that, despite a stable or even deepening price departure, the price index at reimbursed amounts grows as fast as it would absent the fee schedule. Additionally, the price departure may remain stable as the fee schedule allows for a rate of inflation that, over time, erases the initial effect and restores the price level to what would be observed absent the cost containment measure. Due to the possibility of a fee schedule being diluted as a result of regulatory capture, the magnitude of the price departure cannot serve as a measure of the permanence of the initial price level decrease.

Research findings that point to the permanence of the effect of fee schedules on the price level of medical services provided by physicians have been published by Robertson and Corro [21] and, most recently, by Yang and Fomenko [26]. (The research report by Yang and Fomenko follows earlier editions, which offer similar findings.) As discussed, Robertson and Corro compare workers compensation to Group Health and show that in jurisdictions without a fee schedule, medical prices in workers compensation tend to have a higher markup over Group Health. Yang and Fomenko, in comparing 25 jurisdictions for the first half of 2011, demonstrate that the six non-fee schedule jurisdictions are among the seven states with the highest price level for nonhospital, nonfacility medical services delivered in the context of workers compensation. Illinois is the only fee schedule jurisdiction that ranks with the set of non-fee schedule states.

As mentioned, if the post-implementation rate of inflation is no greater than what would be observed absent the fee schedule, then this can be read as an indication of the permanence of the initial price level decrease. Marked differences in the rates of price inflation for medical services between fee schedule states and non-fee schedule states have been observed by Yang and Fomenko [26]. (Here again, the findings reported by Yang and Fomenko have been published in earlier editions of that research report.) Yang and Fomenko present an annual price index for workers compensation-related nonhospital, nonfacility services for the time period 2002 through (the first half of) 2011. The authors show that, among the 25 studied jurisdictions, the six non-fee schedule states are among the jurisdictions with a higher than average rate of inflation. Similar evidence has been established by Schmid and Lord [23], who study medical services provided by physicians in the context of workers compensation for 32 states in the time period 2000 through 2010. These authors show that states without fee schedules (or with charge-based fee schedules, where reimbursement is determined as a percentage of the charged amount) are liable to experience above-average rates of inflation.

As pertains to the two jurisdictions studied here, the rates of inflation during the years following the fee schedule introduction are considerably below their pre-implementation values. This also holds for the rates of inflation of the medical care components of the applicable regional CPI (regional M-CPI, for short). For Tennessee, the pertinent Bureau of Labor Statistics region is the South; for Illinois, it is the Midwest.

Table 2 exhibits the rates of inflation for the Fisher index at reimbursed amounts and for the applicable regional M-CPI. The pre-implementation values of the price indexes run from January 2000 through the pertinent final pre-implementation month, which is June 2005 for Tennessee and January 2006 for Illinois. By comparison, the post-implementation price index series starts in the third post-implementation month and ends in December 2010. The price indexes are not seasonally adjusted. Technically, the rates of inflation are calculated as compound annual growth rates. In a first step, the difference in the natural logarithms of the final and first monthly values of the applicable time window was divided by the fractional number of years that separates these two observations. In the next step, this ratio was exponentiated and diminished by one.

The impact of the fee schedule introduction on the rate of inflation is determined by means of a difference-in-differences approach. First, for each inflation series, the pre-implementation values are subtracted from the corresponding post-implementation numbers. Second, for every jurisdiction, the resulting difference for the Fisher price index is diminished by the difference obtained for the regional M-CPI.

Table 2: Annual Rates of Inflation: Post-Implementation Compared to Pre-Implementation

Jurisdiction	Fisher Price Index at Reimbursed Amounts (Percent Increase)		Regional M-CPI (Percent Increase)		Difference in Differences (Percentage Points)
	Before	After	Before	After	
Tennessee	1.5	0.6	4.0	3.5	–0.3
Illinois	5.0	3.0	4.8	3.4	–0.6

Note: The values for the difference in differences were computed from unrounded rates of inflation and, as a result, may disagree with the corresponding values obtained from the displayed rounded rates of inflation.

As shown in Table 2, for both jurisdictions, the fee schedule introduction has contributed to a lasting reduction in the rate of inflation for medical services delivered by physicians in the context of workers compensation. This reduction goes beyond what has been observed for medical care in general. For Tennessee, the annual rate of increase in the price index of medical services provided by physicians is 0.3 percentage points lower than what would be observed without the fee schedule. For Illinois, the reduction in the annual rate of inflation amounts to 0.6 percentage points. Although Illinois sustains the smaller decline in the price level, it does experience the larger decrease in the rate of inflation.

5. CONCLUSION

Event study methodology was employed to quantify the impact of physician fee schedule introductions on the price and quantity levels of medical services provided by physicians to workers compensation claimants.

In accordance with previous studies of fee schedule introductions, no material and systematic increase in utilization related to this cost containment measure has been discovered. The decrease in utilization that was estimated for Tennessee may be related to restrictions on provider choice that were introduced a year before the fee schedule was implemented. By implication, the impact of a fee schedule introduction is largely confined to the price level response.

In both jurisdictions, the short-term impact on the price level is substantial; more important, there is a weakening of the rate at which this price level subsequently increases. In Tennessee, the price level declines by 7.4 percent and the annual rate of inflation lessens by 0.3 percentage points. By comparison, in Illinois, the price level dips by 5.2 percent and the annual rate of inflation slows by 0.6 percentage points. Alternative to the change in the price level, the short-term effect of the fee schedule introduction can be gauged by a change in price departure (from the *ex ante* value to the *ex post* value). For Tennessee, the price departure expands by 7.4 percentage points and, for Illinois, it grows by 4.0 percentage points.

For Evaluation and Management Services in Tennessee, the price level increases in the wake of the fee schedule introduction. Possibly, this price level response may be due to an anchoring effect. Yet, no direct evidence in favor of this hypothesis was produced and, hence, competing hypotheses cannot be dismissed without further research.

The analysis was unconditional in that there was no attempt to isolate the fee schedule impact from other cost containment measures that may have been implemented alongside. Further, no economic or demographic differences across the two studied jurisdictions were accounted for. This operational constraint originated in the small number of events available for the analysis. But then, such dissimilarities in the legislative and economic environment between the two jurisdictions may account for the difference in the price level and inflation responses in the presence of nearly identical *ex ante* price departures.

6. APPENDIX

6.1 Recoding Units of Service

For some medical services, the supplied quantity of service is measured in number of minutes. Yet, the units of service, as laid down in the fee schedule, refer to the number time intervals, where a time interval is defined as a multiple of minutes. Medical services where the units of service are defined in such a way are common in Anesthesia (not covered in this study) and Physical Medicine.

In the employed data set, there were transaction records that appeared to report the number of minutes instead of the units of service. Before subjecting these transactions to the data cleansing tools discussed below, an algorithm was applied for the purpose of recoding the reported units of service, if necessary. First, if the reported units equaled 9 or less, the reported value was left in place. Second, if the number of units reported was a multiple of 15, it was assumed that this reported value refers to the number of minutes. Third, if the number of units reported was not a multiple of 15 but was a multiple of 10, it was again assumed that the reported value reflects the number of minutes. Finally, if the reported value was greater than 9 but was not a multiple of 15 or 10, the units of service were treated as unknown—this way, the transaction is excluded from the computation of the price fences detailed below, yet is submitted to the associated outlier detection and management.

6.2 Outlier Detection

Outlier detection is undertaken independently for each jurisdiction and calendar year. At the center of the outlier detection approach is the schematic plot developed by Tukey [25]. The algorithm associated with this plot applies to reimbursement per unit of service on the natural log scale. Percentiles are indicated by the letter p; upper case format indicates a percentile on the raw scale, whereas the lower case points to the logarithmic scale. Outlier detection is performed on the level of the medical service and on the level of the service category. Price fences are defined for the purpose of restating prices and quantities.

For a given medical service, if $P_{75} \neq P_{25}$, then the price fences are set to $p_{75}+0.6$ and $p_{25}-0.6$. Exceptions are transactions for which the paid-to-submit ratio is greater than 0.5; in these cases, the lower fence is set to $p_{25}-0.7$. Conversely, if $P_{75}=P_{25}$, then the price fences are set to $p_{85}+0.2$ and $p_{15}-0.2$. Transactions that fall inside the price fences of a given medical service make it into the calculation of the price fences at the level of the service category. The upper fence at the level of the service category is defined as the maximum of $p_{90}+0.5$ (where the percentile is based on the service category) and the maximum of the upper fences across all medical services (in that service category) that register at least 20 transactions in the applicable calendar year. Similarly, the lower fence at the level of the service category is defined as the minimum of p_{10}–0.5 and the minimum of the lower fences across all medical services that account for at least 20 transactions.

Transactions that are located inside the fences, both at the level of the medical service and the level of the service category, remain unedited. These transactions enter the computation of the average price per unit of service and the median units of service per transaction for the applicable medical service in a given month.

Medical services that register less than 12 records (in a given year and jurisdiction) are excluded from the analysis.

6.3 Outlier Management

Outlier management is undertaken on a calendar year basis, in keeping with the price fences, which were computed from data for the calendar year.

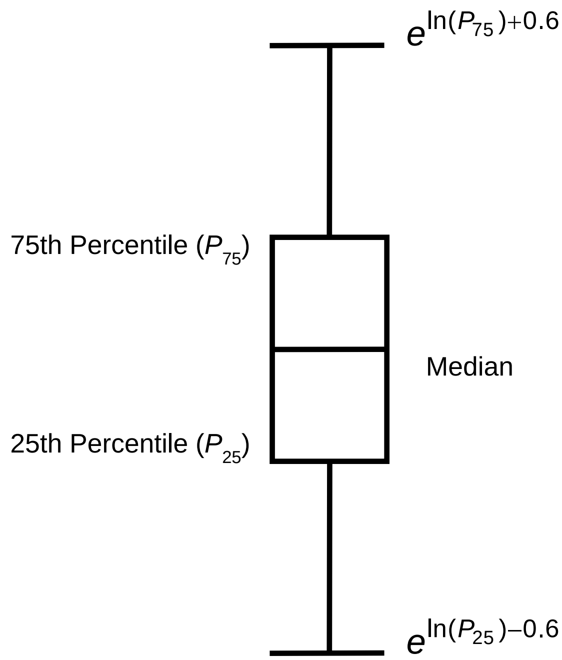
Transactions that come with prices per unit of service above the upper fences of the applicable medical service or the overarching service category are reset to the mean price of this medical service; the units of service associated with these transactions remain unaltered.

Transactions with prices per unit of service below the lower fences of the respective medical service or the applicable service category have their units of service restated based on the following process—that process also applies to transactions with unknown units. First, the number of units is set to the median number of units per transaction for this medical service; this median number of units is typically unity. Then, the price is recalculated based on the units of service thus restated. If this recalculated price falls below the lower fences of the medical service or the service category, then the units of service associated with this transactions is set to unity. Then again, the price is recalculated. If this price still falls below any of the two applicable lower price fences, then the record is discarded as a nuisance transaction. Conversely, if, during this iterative process, the recalculated price exceeds any of the two upper fences, then this price is reset to the mean of the applicable medical service and no further restatement is done.

6.4 Tukey's Schematic Plot (Boxplot)

Chart A.1 depicts Tukey's schematic plot, also known as box-and-whiskers plot or boxplot. The objective of the boxplot is to report major location parameters (such as the median and the 25th and 75th percentiles) and to identify outliers. In this graph, the hinges that define the upper and lower limits of the box identify the inter-quartile range (IQR), which comprises 50 percent of the data. For the purpose of this study, the fences are defined on the logarithmic scale. The upper fence signifies the sum of the 75th percentile and 0.6 on the natural log scale (which corresponds to multiplying the 75th percentile on the raw scale by 1.8, approximately). The lower fence equals the difference between the 25th percentile and 0.6 on the log scale (which amounts to dividing the 25th percentile on the raw scale by 1.8, approximately). Values beyond the fences are considered outliers.

Chart 9: Tukey's Schematic Plot (Box Plot)



Acknowledgment

Thanks to James Bonk, Raji Chadarevian, Susan Kirk, Natasha Moore, and John Potter for comments, and to Adam Jensen, Jerry Hsieh, Chandra Lakkaraju, Linda Li, Gary Nelson, and Jimmy Pillow for research assistance.

7. REFERENCES

[1] Boden, Leslie, and Charles Fleischman, *Medical Costs in Worker's Compensation: Trends and Interstate Comparisons*, Cambridge (MA): Workers Compensation Research Institute, **1989**: December.
[2] Borba, Philip S., "Can Medical Fee Schedules Control Medical Care Expenditures," *NCCI Digest*, **1986**: 1(2), 1–13.
[3] Camerer, Colin, Linda Babcock, George Loewenstein, and Richard Thaler, "Labor Supply of New York City Cabdrivers: One Day at a Time," *Quarterly Journal of Economics*, **1997**: 112(2), 407–441.
[4] Durbin, David, and David Appel, "The Impact of Fee Schedules and Employer Choice of Physician," *NCCI Digest*, **1991**: 6(3), 39–59.
[5] De Livera, Alysha M., Rob J. Hyndman, and Ralph D. Snyder, "Forecasting Time Series with Complex Seasonal Patterns Using Exponential Smoothing," *Journal of the American Statistical Association*, **2011**: 106(496), 1513–1527.
[6] Engelmann, Dirk, and Hans-Theo Normann, "Price Ceilings as Focal Points? An Experimental Test," in: Jeroen Hinloopen and Hans-Theo Normann, eds., *Experiments and Competition Policy*. Cambridge (UK): Cambridge University Press, **2009**, 61–80.
[7] Illinois Workers' Compensation Commission, *Handbook on Workers' Compensation and Occupational Diseases: For Injuries and* Illnesses *from 2/1/06–6/27/11*, **2011**: April 19, www.iwcc.il.gov/handbook020106.pdf.
[8] Illinois Workers' Compensation Commission, *Handbook on Workers' Compensation and Occupational Diseases: For Injuries and Illnesses Before 2/1/06*, [**no date**], www.iwcc.il.gov/handbook.pdf.
[9] Illinois Workers' Compensation Commission, *FY 2005 Annual Report*, **2006**: July, www.iwcc.il.gov/annualreport05.pdf.
[10] International Labour Office, *Consumer Price Index Manual: Theory and Practice*, Geneva (Switzerland): International Labour Office, **2004**.
[11] Kahneman, Daniel, Ilana Ritov, and David Schkade, "Economic Preferences or Attitude Expressions? An Analysis of Dollar Responses to Public Issues," *Journal of Risk and Uncertainty*, **1999**: 19(1–3), 203–235.
[12] Knittel, Christopher R., and Victor Stango, "Price Ceilings as Focal Points for Tacit Collusion: Evidence from Credit Cards," *American Economic Review*, **2003**: 93(5), 1703–1729.
[13] Minnesota Department of Labor and Industry, *Report to the Legislature on Health Care Costs and Cost Containment in Minnesota Workers' Compensation,* St. Paul (MN): Minnesota Department of Labor and Industry, **1990**: March.
[14] National Council on Compensation Insurance, "Cost Containment," *NCCI Digest*, **1989**: 4(4), 25–49.
[15] Neumark, David, Peter S. Barth, and Richard A. Victor, "The Impact of Provider Choice on Workers' Compensation Costs and Outcomes," *Industrial and Labor Relations Review*, **2007**: 61(1), 121–142.
[16] Pozzebon, Silvana, "Medical Cost Containment under Workers' Compensation," *Industrial and Labor Relations Review*, **1994**: 48(1), 153–167.
[17] Radeva, Evelina, Nicole M. Coomer, Stacey Landes, Carol A. Telles, Rui Yang, and Ramona P. Tanabe, *Monitoring the Impact of Illinois Regulatory Changes: CompScope™ Medical Benchmarks, 10th Edition*, Cambridge (MA): Workers Compensation Research Institute, **2010**: July.
[18] Radeva, Evelina, Nicole M. Coomer, Bogdan Savych, Carol A. Telles, Rui Yang, and Ramona P. Tanabe, *Monitoring the Impact of Regulatory Changes in Tennessee: CompScope™ Medical Benchmarks, 10th Edition*, Cambridge (MA): Workers Compensation Research Institute, **2010**: July.
[19] Radeva, Evelina, Carol A. Telles, Rui Yang, and Ramona P. Tanabe, *Monitoring the Impact of the Tennessee Fee Schedules: CompScope™ Medical Benchmarks, 9th Edition*, Cambridge (MA): Workers Compensation Research Institute, **2009**: June.
[20] Roberts, Karen, and Susan Zonia, "Workers' Compensation Cost Containment and Health Care Provider Income Maintenance Strategies," *Journal of Risk and Insurance*, **1994**: 61(1), 117–131.
[21] Robertson, John, and Dan Corro, "Making Workers Compensation Medical Fee Schedules More Effective," *NCCI Research Brief*, **2007**: December, www.ncci.com/documents/WC_Medical_Fee_Schedule.pdf.
[22] Schelling, Thomas C., *The Strategy of Conflict*, Cambridge (MA): Harvard University Press, **1960**.
[23] Schmid, Frank, and Nathan Lord, "The Impact of Physician Fee Schedule Changes in Workers Compensation: Evidence from 31 States," **2012**, www.ncci.com, *forthcoming*.

[24] Stigler, George, "The Theory of Economic Regulation," *Bell Journal of Economics and Management Science*, 1971: **3**, 3–18.
[25] Tukey, John W., *Exploratory Data Analysis*, Reading (MA): Addison-Wesley, **1977**.
[26] Yang, Rui, and Olesya Fomenko, *WCRI Medical Price Index for Workers' Compensation, Fourth Edition (MPI-WC)*, Cambridge (MA): Workers Compensation Research Institute, **2012**: March.

Abbreviations and Notations

AIC, Akaike information criterion
AMA, American Medical Association
ARMA, Auto-regressive moving average
BATS, Box-Cox transform, ARMA errors, trend, and seasonal components
CPI, Consumer Price Index
CPT, Current Procedural Terminology
DCI, Detailed Claims Information
HCFA, Health Care Financing Administration
HMO, Health Maintenance Organization
IQR, Inter-Quartile Range
MAR, Maximum Allowable Reimbursement
M-CPI, Medical care component of the CPI
NCCI, National Council on Compensation Insurance
PPO, Preferred Provider Organization
PPP, Preferred Provider Program
TBATS, Trigonometric, Box-Cox transform, ARMA errors, trend, and seasonal components

Biographies of the Authors

Frank Schmid, Dr. habil., was, at the time of writing, a Director and Senior Economist at the National Council on Compensation Insurance, Inc.
Nathan Lord is a Senior Actuarial Analyst at the National Council on Compensation Insurance, Inc.

Applications of Convex Optimization in Premium Rating

Dimitri Semenovich

Abstract: In this paper we discuss the application of modern mathematical optimization techniques to some of the common problems in insurance premium rating. The computationally tractable setting of convex optimization [6] is particularly attractive as it encompasses parameter estimation in generalized linear models and offers means to address practical challenges such as variable selection, coefficient smoothing, spatial and hierarchical priors, constraints on relativities and the time evolution of model parameters. Recent advances in modelling systems for convex optimization make these methods not only eminently practical but also in many respects more flexible than what is presently offered by statistical software.

Keywords: Convex optimization, generalized linear models, credibility, graduation, spatial smoothing, dynamic regression, revenue management.

1 Introduction

In this paper we formulate several common models arising in premium rating as convex optimization problems and describe the use of constraints and regularization to address practical issues such as variable selection, coefficient smoothing, hierarchical credibility, parameter evolution and spatial priors in a unified framework. The resulting optimization problems can be solved using efficient algorithms [65] developed for convex programming[1].

Many classical actuarial techniques such as Whittaker graduation and various credibility models can be interpreted as performing regularized or constrained fitting [10, 38, 29]:

$$\underset{\mathbf{w}}{\text{minimize}} \quad \mathcal{L}(\mathbf{y};\mathbf{w}) + R(\mathbf{w}) \qquad\qquad \begin{array}{ll} \underset{\mathbf{w}}{\text{minimize}} & \mathcal{L}(\mathbf{y};\mathbf{w}) \\ \text{subject to} & R(\mathbf{w}) \leq \epsilon, \end{array} \tag{1}$$

where $\mathcal{L}(\mathbf{y};\mathbf{w})$ is the term penalizing model error relative to the data $\mathbf{y}$ and the regularization term $R(\mathbf{w})$ measures the lack of smoothness or some other desired property of the model $\mathbf{w}$. While not a part of the classical theory, regularization has important implications for the practical use of generalized linear models (GLMs), by now a nearly universal tool in premium rating. Namely, rather than carrying out manual feature design and selection it may often be far more effective to control the degrees of freedom of the model by imposing penalties or constraints on the coefficients. Recent advances in solvers and open source modelling software for convex optimization [21] have made it exceedingly easy to develop such custom models.

The idea of regularization itself has been developed independently many times in many different fields, e.g. the work of A. Tikhonov on operator equations in the 1960s [58]. It is also the principal reason for the remarkable performance of the "support vector machines" family of algorithms which implicitly generate feature spaces of high dimension through the kernel functions [59].

[1] Video lectures for the Stanford course EE364A Convex Optimization, made publicly available via the Stanford Engineering Everywhere initiative, are highly recommended as background for this paper.

In recent years there has been a resurgence of interest in a particular form of regularization known as "ℓ_1-norm" regularization. Its early applications appeared in geophysics in 1970s [55] where its sparsity inducing properties had been employed for signal recovery. Around the same time in the actuarial literature Schuette proposed an ℓ_1-norm formulation of the graduation problem [45] which allowed it to be solved by linear programming; this idea was further developed in econometrics as quantile smoothing splines [35]. It was not until much later that ℓ_1-norm regularization had become a widely adopted technique in signal processing with applications including computing transform coefficients ("basis pursuit") [12] and signal recovery from incomplete measurements ("compressive sensing") [14]. In statistics, the idea of ℓ_1-norm regularization was popularized by the well-known "lasso" procedure [56] for linear regression and its many extensions [57, 68, 67, 37]. While for a long time ℓ_1-norm regularization has been viewed as little more than a useful heuristic in optimization, recent theoretical results (e.g. [9]) have provided surprising guarantees on its performance in certain restricted settings.

What the above models have in common is that they all, together with many others[2] [6], can be formulated as convex optimization problems. Recognizing convexity and its implied properties offers a unifying perspective on a collection of seemingly unrelated ideas from many different fields and dramatically reduces the need to develop special purpose algorithms as many instances can be handled by standard solvers. Moreover, convex problems constitute, perhaps, the widest known class of optimization problems for which exist efficient algorithms guaranteed to find a global solution, making convexity especially desirable when reliable numerical solutions are a requirement.

This paper consists of two main parts. In the first half we outline the basics of mathematical optimization and convex calculus and briefly describe the connection between convex optimization and statistical estimation. In the second part we focus on the application of convex optimization to some of the problems in technical premium rating. We discuss variable selection, curve fitting, spacial clustering and smoothing, additive models, hierarchical credibility, time evolution of model parameters and stochastic optimization.

1.1 Related work

A number of unifying approaches along similar lines have been previously proposed in the actuarial literature but using as the foundation algorithmic developments from statistics, rather than mathematical optimization. These include Bayesian multilevel extensions of generalized linear models [19, 18] and the so called "mixed effect" models [1]. Unlike convex optimization, however, these have arguably less broad scope of applicability (e.g. ability to handle constraints is lacking) and do not take into account computational complexity of the proposed methods.

It should be noted at the same time that stateful sampling algorithms such as variants of geometric Markov chain Monte Carlo [25], used for inference in Bayesian models, bear close resemblance to iterative optimization methods and their hybridization is an active area of research, e.g. [63]. It seems not unreasonable to hope that one day an effective synthesis will be attained.

The most complete treatment of actuarial models from the perspective of convex optimization was developed in a visionary paper of Brockett [7], albeit in the *dual* form[3] and some 15 years before advances in algorithms and computing power have made such schemes truly practical for all insurance applications.

[2] In particular it is worth noting that one of many fields where convex analysis has proven to be the key tool is mathematical economics; see [61] for an accessible introduction from this point of view.

[3] The equivalence between "information theoretic" maximum entropy principle and maximum likelihood estimation for exponential families is discussed in standard references, e.g. [8].

2 Optimization and statistical inference

2.1 Mathematical optimization

A constrained optimization problem has the the following form:

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \quad & f_0(\mathbf{x}) \\ \text{subject to} \quad & f_i(\mathbf{x}) \leq b_i, \ i = 1, \ldots, m. \end{aligned} \tag{2}$$

The vector $\mathbf{x} \in \mathbb{R}^n$ is the optimization variable, the function $f_0 : \mathbb{R}^n \to \mathbb{R}$ is the objective function, the functions $f_i : \mathbb{R}^n \to \mathbb{R}$, $i = 1, \ldots, m$ are the inequality constraint functions[4], and the constants $b_1, \ldots, b_m$ are the limits, or bounds, for the constraints. A vector $\mathbf{x}^*$ is called a *solution* or a *global minimum* of the optimization problem (2) if no other vector satisfying the constraints achieves a smaller objective value, that is for any $\mathbf{z} \in \mathbb{R}^n$ with $f_1(\mathbf{z}) \leq b_1, \ldots, f_m(\mathbf{z}) \leq b_m$ we have $f_0(\mathbf{z}) \geq f_0(\mathbf{x}^*)$.

2.1.1 Convex optimization

While the general problem (2) is computationally completely intractable, we can mitigate this by restricting the class of functions $f_0, \ldots, f_m$. For example, if we take the objective and the constraints to be linear, the optimization problem (2) is called a linear program and can be written as:

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \quad & \mathbf{c}^T\mathbf{x} \\ \text{subject to} \quad & \mathbf{a}_i^T\mathbf{x} \leq b_i, \ i = 1, \ldots, m. \end{aligned} \tag{3}$$

Despite the seeming simplicity surprisingly many problems in business operations and engineering (e.g. optimal network flow) can be expressed in this form [3]. There are also applications in statistics and econometrics (least absolute deviations, quantile regression [34]). The subject was developed in 1930s and 40s by L. Kantorovich and G. Dantzig. The latter introduced the simplex algorithm for solving linear programs that for many applications remains unsurpassed to this day and can routinely solve problems with millions of variables and constraints.

A less restrictive class of tractable optimization problems is the one in which the objective and constraint functions are convex, namely:

$$f_i(\alpha\mathbf{x} + \beta\mathbf{y}) \leq \alpha f_i(\mathbf{x}) + \beta f_i(\mathbf{y}) \tag{4}$$

for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and all $\alpha, \beta \in \mathbb{R}$ with $\alpha + \beta = 1, \ \alpha \geq 0, \ \beta \geq 0$. Convex functions have the essential property that every local minimum is also a global minimum. A function $g(\mathbf{x})$ is called a concave function if $-g(\mathbf{x})$ is convex. It is easy to check that linear functions are convex (they are also concave) and therefore linear programs are a special case of convex optimization problems.

There are efficient polynomial time algorithms (e.g. so called interior point methods [6]) for global minimization of convex functions subject to convex inequality constraints. Indeed convex problems are effectively the widest class of optimization problems for which such algorithms exist at this time. And while convexity is not an essential property of a successful optimization model, it is worthwhile to be aware of the trade-off between efforts to make a model more realistic and ensuing difficulties with numerical methods. To quote Y. Nesterov [42], one of the key figures in the development of convex programming:

Every year I meet Ph.D students of different specializations who ask me for advice

[4] Equality constraints are omitted for brevity but are implied, i.e. an equality constraint $f_i(\mathbf{x}) = 0$ can be represented as $f_i(\mathbf{x}) \geq 0$ and $-f_i(\mathbf{x}) \geq 0$.

on reasonable numerical schemes for their optimization models. And very often they seem to have come too late. In my experience, if an optimization model is created without taking into account the abilities of numerical schemes, the chances that it will be possible to find an acceptable numerical solution are close to zero.

2.1.2 Convex calculus

It is generally labor intensive to check convexity of a function directly from the definition. In most cases it is much easier to see wheteher a given function is built up of known convex functions using transformations that preserve convexity, just as classical calculus permits effective computation without explicitly working with infinite series.

Familiarity with the basics of convex analysis and a few heuristics will permit effective creation of custom models for many applications. The tedious (and ultimately mechanical) task of converting the resulting formulation to one of standard forms understood by solvers can be handled by the modelling system [21].

We describe some convex functions of one and many variables together with operations that maintain convexity in Appendix A. Much more detailed treatments can be found in [44, 6, 2, 42].

2.2 Convex optimization and statistical estimation

Statistical inference can often reduced to solving certain optimization problems. Below we discuss two such principles.

2.2.1 Maximum likelihood and loss minimization:

A familiy of probability density functions on $\mathbb{R}^n$ denoted $p(\mathbf{y};\mathbf{w})$ with parameter vector $\mathbf{w} \in \mathbb{R}^m$ is called a likelihood function when taken as a function of $\mathbf{w}$ only for a fixed $\mathbf{y}$. It is, however, often more convenient to deal with the logarithm of the likelihood function or the log-likelihood, $\log p(\mathbf{y};\mathbf{w})$. The negative log-likelihood is sometimes also called the "loss function":

$$\mathcal{L}(\mathbf{y};\mathbf{w}) = -\log p(\mathbf{y};\mathbf{w}). \tag{5}$$

It is worth noting, however, that not all loss functions are directly motivated by a priori distributional assumptions e.g. quantile loss (27).

A remarkably effective method for estimating the parameter $\mathbf{w}$ given an observation $\mathbf{y}$ consists of maximizing the log-likelihood (equivalently, minimising the loss function) with respect to $\mathbf{w}$:

$$\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmin}}\, \mathcal{L}(\mathbf{y};\mathbf{w}). \tag{6}$$

In many practical applications we have prior information that can be represented in the form of constraints on the admissible values of $\mathbf{w}$. These constraints can be defined explicitly by specifying a set $C \subseteq \mathbb{R}^m$ such that $\mathbf{w} \in C$ or incorporated into the likelihood function by setting $p(\mathbf{y};\mathbf{w}) = 0$ and correspondingly $\mathcal{L}(\mathbf{y};\mathbf{w}) = \infty$ for all $\mathbf{w} \notin C$. When C is given, the constrained maximum likelihood estimation problem can be written as follows:

$$\begin{aligned} \underset{\mathbf{w}}{\text{minimize}} \quad & \mathcal{L}(\mathbf{y};\mathbf{w}) \\ \text{subject to} \quad & \mathbf{w} \in C. \end{aligned} \tag{7}$$

While computationally intractable in general, maximum likelihood estimation is reduced to a convex optimization problem if the loss function $\mathcal{L}(\mathbf{y};\mathbf{w})$ is convex in $\mathbf{w}$ and C is a convex set.

2.2.2 Bayesian estimation

Maximum likelihood procedure has an analogue in the Bayesian setting known as maximum a posteriori estimation. Here the parameter vector $\mathbf{w}$ and the observation $\mathbf{y}$ are both considered to be random variables with a joint probability density $p(\mathbf{y}, \mathbf{w})$. The density of $\mathbf{w}$ is then given by

$$p(\mathbf{w}) = \int p(\mathbf{y}, \mathbf{w})\, dy_1 \ldots dy_n. \tag{8}$$

This is referred to as the prior distribution of $\mathbf{w}$ and represents the information about $\mathbf{w}$ before $\mathbf{y}$ is observed. We can similarly define $p(\mathbf{y})$, the prior distribution of $\mathbf{y}$. The conditional probability density of $\mathbf{y}$ given $\mathbf{w}$ is as follows:

$$p(\mathbf{y}|\mathbf{w}) = \frac{p(\mathbf{y}, \mathbf{w})}{p(\mathbf{w})}. \tag{9}$$

Being a function of $\mathbf{w}$, it is equivalent to the likelihood function in employed in maximum likelihood estimation. The conditional probability density of $\mathbf{w}$ given $\mathbf{y}$ can then be written as:

$$p(\mathbf{w}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{w})p(\mathbf{w})}{p(\mathbf{y})} \tag{10}$$

If we substitute the observed value of $\mathbf{y}$ into $p(\mathbf{w}|\mathbf{y})$ we obtain the posterior density of $\mathbf{w}$, representing the updated information about $\mathbf{w}$. The maximum a posteriori estimate of $\mathbf{w}$ is the one that maximizes the posterior probability ($p(\mathbf{y})$ does not depend on $\mathbf{w}$ and can be omitted):

$$\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmax}}\, p(\mathbf{w}|\mathbf{y}) = \underset{\mathbf{w}}{\operatorname{argmax}}\, p(\mathbf{y}|\mathbf{w})p(\mathbf{w}). \tag{11}$$

After taking the logarithm the expression for $\mathbf{w}^*$ can be written as:

$$\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmax}}\ \log\big(p(\mathbf{y}|\mathbf{w})p(\mathbf{w})\big) = \underset{\mathbf{w}}{\operatorname{argmin}}\ -\log\big(p(\mathbf{y}|\mathbf{w})\big) - \log\big(p(\mathbf{w})\big). \tag{12}$$

This is equivalent to minimising a data dependent loss function $-\log(p(\mathbf{y}|\mathbf{w}))$ with the additional regularization term $-\log(p(\mathbf{w}))$.

It is also a basic consequence of Lagrange duality [6, Chapter 5] that under some mild regularity conditions (12) has the same solution $\mathbf{w}^*$ as the following constrained optimization problem for some instance dependent value of ϵ:

$$\begin{array}{ll} \underset{\mathbf{w}}{\text{minimize}} & -\log\big(p(\mathbf{y}|\mathbf{w})\big) \\ \text{subject to} & -\log\big(p(\mathbf{w})\big) \leq \epsilon. \end{array} \tag{13}$$

For example if the prior density of $\mathbf{w}$ has support over a a set C and is uniform then finding the maximum a posteriori estimate is the same as loss minimization subject to the constraint $\mathbf{w} \in C$.

For any estimation problem with a convex loss function we can add a convex regularization term (corresponding to a prior density on $\mathbf{w}$ that is log-concave) and the resulting optimization problem will be convex.

2.3 Convex loss functions

A large number of statistical problems can be reduced to minimizing convex loss functions, with conditional exponential families being perhaps the key example. Several approaches not based directly on the maximum likelihood principle are also mentioned.

2.3.1 Conditional exponential families

It is a standard result that the log-likelihood of distributions in the exponential families is concave in the natural parameters [8, 62]. Below we discuss convexity properties of conditional exponential families, closely related to generalized linear models.

Consider an exponential family distribution on $\mathcal{Y} \times \mathcal{X}$:

$$\begin{aligned} p(y, x|\mathbf{w}) &= h_0(y, x) \exp\Big(\sum_{k=1}^{m} w_k \phi_k(y, x) - A(\mathbf{w})\Big) \\ &= h_0(y, x) \frac{\exp\Big(\sum_{k=1}^{m} w_k \phi_k(y, x)\Big)}{\exp\big(A(\mathbf{w})\big)}. \end{aligned} \tag{14}$$

In this context the non-negative function h_0 is the base or *carrier* measure, $\mathbf{w} \in \mathbb{R}^k$ are the model parameters, $\phi(y, x) = [\phi_1(y, x), \ldots, \phi_k(y, x)]^T$ is the vector of ***sufficient statistics*** and $A(\mathbf{w})$ is the logarithm of the normalizing constant or the ***log-partition*** function, namely:

$$A(\mathbf{w}) = \log\Big(\int_{(y,x)\in\mathcal{Y}\times\mathcal{X}} \exp\Big(\sum_{k=1}^{m} w_k \phi_k(y, x)\Big) h_0(y, x)\, dx\, dy\Big), \tag{15}$$

with summation replacing the integral for discrete distributions. For reasons such as data and computational limitations, we may instead wish to directly estimate the conditional probability:

$$p(y|x, \mathbf{w}) = h_0(y, x) \exp\Big(\sum_{k=1}^{m} w_k \phi_k(y, x) - A(\mathbf{w}|x)\Big), \tag{16}$$

with conditional log-partition function given by:

$$A(\mathbf{w}|x) = \log\Big(\int_{y\in\mathcal{Y}} \exp\Big(\sum_{k=1}^{m} w_k \phi_k(y, x)\Big) h_0(y, x)\, dy\Big). \tag{17}$$

Note that the sufficient statistics that do not depend on y can effectively be omitted as the choices of associated parameters do not influence the conditional densities. Given a collection of independent samples $(y_i, x_i) \in \mathcal{Y} \times \mathcal{X}$ for $i = 1, \ldots, n$, the joint conditional probability can be written as:

$$\prod_{i=1}^{n} p(y_i|x_i, \mathbf{w}) = \prod_{i=1}^{n} \Big(h_0(y_i, x_i) \exp\Big(\sum_{k=1}^{m} w_k \phi_k(y_i, x_i) - A(\mathbf{w}|x_i)\Big)\Big), \tag{18}$$

giving rise to the following maximum log-likelihood estimation problem to find parameters $\mathbf{w}$:

$$\underset{\mathbf{w}}{\text{minimize}} \quad \sum_{i=1}^{n} \Big(A(\mathbf{w}|x_i) - \sum_{k=1}^{m} w_k \phi_k(y_i, x_i)\Big). \tag{19}$$

The above objective is convex being a sum of linear terms and the convex log-partition functions. The latter are convex in $\mathbf{w}$ by an extension of the soft max rule (see A.3).

In this formulation $\mathcal{Y}$ is not restricted to be equal to $\mathbb{R}$ or to a small set of discrete outcomes but can also represent more complex structured objects, e.g. all possible parts of speech assignments for a particular sentence. This type of models is commonly[5] referred to as a *conditional random field* [49] and is likely to prove quite fruitful in insurance applications.

To recover the (now classical) generalized linear models of Wedderburn and Nelder, consider

[5] At least in computer science literature on "machine learning" (a variant of computational statistics with a strong focus on out of sample performance) and natural language processing.

the case when $\mathcal{X} = \mathbb{R}^k$, the carrier measure h_0 does not depend on x and with a particularly simple choice of sufficient statistics:

$$\phi_k(y, \mathbf{x}) = yx_k. \tag{20}$$

We can then rewrite (16) as a single parameter exponential family with respect to $\mathbf{w}^T\mathbf{x}$ in the following way:

$$\begin{aligned} p(y|\mathbf{x}, \mathbf{w}) &= p(y|\mathbf{w}^T\mathbf{x}) \\ &= h_0(y) \exp\Big(y\mathbf{w}^T\mathbf{x} - B(\mathbf{w}^T\mathbf{x})\Big), \end{aligned} \tag{21}$$

with the log partition function:

$$B(\theta) = \log\Big(\int_{y\in\mathcal{Y}} \exp(\theta y) h_0(y)\, dy\Big) \tag{22}$$

and giving rise to the following maximum likelihood parameter estimation problem:

$$\underset{\mathbf{w}}{\text{minimize}} \quad \mathcal{L}(y; X\mathbf{w}) = \sum_{i=1}^{n} \big(B(\mathbf{w}^T\mathbf{x}_i) - y_i\mathbf{w}^T\mathbf{x}_i\big). \tag{23}$$

The usual relation between the natural parameter and the expected value of y obtains:

$$\mathbb{E}(y|\mathbf{x}, \mathbf{w}) = \nabla_\theta B(\mathbf{w}^T\mathbf{x}), \tag{24}$$

with $\nabla_\theta B^{-1}$ being the canonical link function.

There are two remaining incompatibilities between conditional random fields and generalized linear models, however. Firstly GLMs are based on the so called *exponential dispersion* families [31]:

$$p(y|\mathbf{x}, \mathbf{w}, \lambda) = h_0(y, \lambda) \exp\Big(\lambda\big(y\mathbf{w}^T\mathbf{x} - B(\mathbf{w}^T\mathbf{x})\big)\Big), \tag{25}$$

rather than exponential families considered up to this point.

For the fixed dispersion parameter $\sigma^2 = \lambda^{-1}$ this class of models coincides with single parameter exponential families. This is often the case, e.g. when σ^2 represents a known number of observations for the binomial distribution. If σ^2 is not known and is to be estimated, the resulting log-likelihood is not in general jointly concave in w and σ^2, unlike that for conditional random fields. One way to overcome this limitation is to consider the overlapping class of two parameter exponential families, which includes many standard distributions and also provides means to deal with overdispersion [13].

Another difficulty is with regard to the link function - convexity of the negative log-likelihood does not necessarily hold for choices other than the canonical link. Even in this case, however, for all of the models described in this paper local solutions can be obtained using sequential quadratic approximation (variants of which are known as iteratively reweighted least squares and Fisher scoring). It is worth remembering that none of the methods implemented in existing statistical software provide guarantees of global optimality in this situation either.

Convex formulations for parameter estimation in some common GLMs are shown in Appendix B.

2.3.2 Huber loss

Huber loss function [26] is often used to make least squares estimates more robust to outliers, see also [6, Chapter 6] for an illuminating informal discussion. It agrees with the squared ℓ_2-loss (96) for $|u| < M$ and for $|u| \geq M$ the Huber loss function reverts to linear growth which gives lowest attainable sensitivity to outliers while still maintaining convexity.

$$\mathcal{L}(\mathbf{y}; X\mathbf{w}) = \sum_{i=1}^{n} \phi(y_i - \mathbf{w}^T \mathbf{x}_i), \quad \phi(u) = \begin{cases} u^2, & |u| < M \\ M(2|u| - M), & |u| \geq M. \end{cases} \tag{26}$$

2.3.3 Quantile loss

One interpretation of the least squares procedure (96) is that it estimates the conditional mean of y_i given the data vector $\mathbf{x}_i$. Regression with the asymmetric quantile [34] loss function ρ_τ on the other hand results in estimates approximating the conditional τ-th quantile of the response variables y_i:

$$\mathcal{L}(\mathbf{y}; X\mathbf{w}) = \sum_{i=1}^{n} \rho_\tau(y_i - \mathbf{w}^T \mathbf{x}_i), \quad \rho_\tau(u) = \begin{cases} \tau u, & u > 0 \\ -(1-\tau)u, & u \leq 0. \end{cases} \tag{27}$$

When τ is equal to 0.5 and corresponds to the median, quantile regression is equivalent to the method of least absolute deviations which estimates $\mathbf{w}$ by seeking to minimize $\frac{1}{2}\|\mathbf{y} - X\mathbf{w}\|_1$. We should note that quantile regression appears quite attractive for insurance applications [36] as it can provide a non-parametric estimate of the full conditional distribution of the dependent variable and can deal with such issues as concetration of probability mass at a certain point. In addition to directly modelling claim costs per policy, it may also be applicable in situations when mean of the dependent variable is not easily interpreted, e.g. when fitting a model to a sample of competitor rates.

3 Applications in premium rating

3.1 Variable selection

One compelling application of constrained parameter estimation is variable selection. Consider a regression type problem with an arbitrary convex loss function $\mathcal{L}(\mathbf{y}; X\mathbf{w})$ where y is the vector of response variables, X is the design matrix and w a vector of parameters to be estimated. This can be accomplished by restricting the ℓ_1-norm of the coefficient vector (assuming that the explanatory variables are standardized with mean 0):

$$\begin{array}{ll} \underset{\mathbf{w}}{\text{minimize}} & \mathcal{L}(\mathbf{y}; X\mathbf{w}) \\ \text{subject to} & \|\mathbf{w}\|_1 \leq \epsilon. \end{array} \tag{28}$$

This generally results in a sparse estimate $\mathbf{w}^*$ with the number of non zero entries controlled by the magnitude of ϵ. It is possible to motivate this formulation as a convex approximation (or *relaxation*) of the computationally intractable best subset selection problem [6]. Using an ℓ_1-norm penalty to obtain a sparse solution has been a well known heuristic in optimization and its applications going back at least to the 1970s [55, 45]. It was popularized in the statistics literature as the "lasso" by Tibshirani [56]. Under some mild conditions the above problem can be equivalently formulated as:

$$\underset{\mathbf{w}}{\text{minimize}} \quad \mathcal{L}(\mathbf{y}; X\mathbf{w}) + \lambda\|\mathbf{w}\|_1 \tag{29}$$

where $\lambda \geq 0$ corresponds to the optimal Lagrange multiplier associated with the inequality constraint in (28). In this form the "lasso" procedure has a Bayesian interpretation as a maximum a posteriori estimate with a Laplacian prior on $\mathbf{w}$ with mean zero and variance $\frac{2}{\lambda^2}$. The value of λ controls the sparsity of $\mathbf{w}*$ and can be chosen via cross-validation.

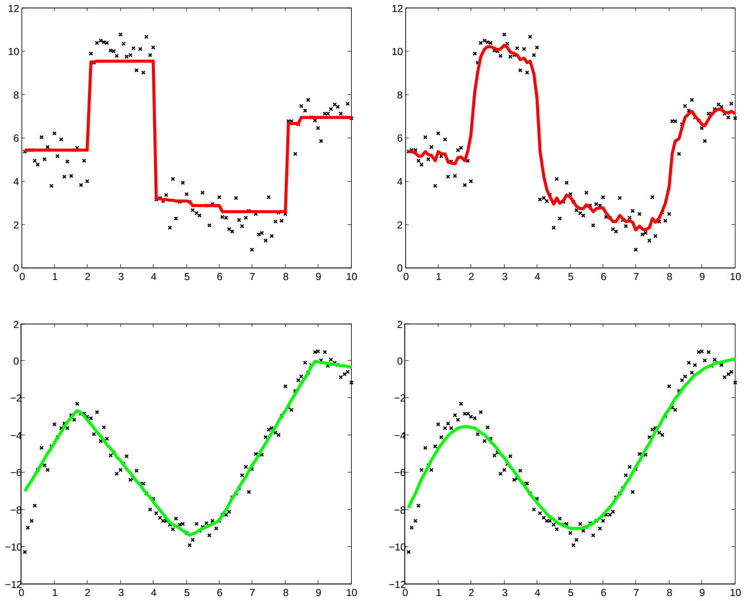


Figure 1: *Left:* Solutions of optimization problem (39) for different choices of D - "fused lasso", $D^{(1,n)}$ (*top*) and "ℓ_1-trend filtering", $D^{(2,n)}$ (*bottom*). *Right:* Solutions of the Whittaker graduation problem (32) with the same choices of D. We set λ to give the same squared error $\|y-w\|_2^2$ as the corresponding solutions on the left.

In the presence of correlated covariates "lasso" will tend to select only one of them. To alleviate this problem we can trade off between penalising ℓ_1 and squared Euclidean norms of $\mathbf{w}$:

$$\underset{\mathbf{w}}{\text{minimize}} \quad \mathcal{L}(\mathbf{y}; X\mathbf{w}) + \lambda\left(\alpha\|\mathbf{w}\|_1 + (1-\alpha)\|\mathbf{w}\|_2^2\right), \tag{30}$$

where $0 \leq \alpha \leq 1$. This method is known as "elastic net" [68] in the statistics literature. The "elastic net" performs variable selection while at the same time pushing together coefficient values of correlated variables. Indeed, for $\alpha = 0$ it is equivalent to ridge regression, a classical procedure for dealing with collinearity:

$$\underset{\mathbf{w}}{\text{minimize}} \quad \mathcal{L}(\mathbf{y}; X\mathbf{w}) + \lambda\|\mathbf{w}\|_2^2, \tag{31}$$

see [40] for connections with credibility.

3.2 Graduation or curve fitting

3.2.1 One dimensional data

Before we discuss further extensions to generalized linear models we motivate our approach by examining the classical setting of non-parametric graduation. The goal of the graduation procedure is to smooth a sequence of observations $\mathbf{y} = [y_1, y_2, \ldots, y_n]^T$ which are usually indexed by time or age. As pointed out in [46], as early as 1899 Bohlmann had proposed [4] to perform graduation by solving the following convex optimization problem:

$$\underset{\mathbf{w}}{\text{minimize}} \ \|\mathbf{y}-\mathbf{w}\|_2^2 + \lambda\|D\mathbf{w}\|_2^2, \tag{32}$$

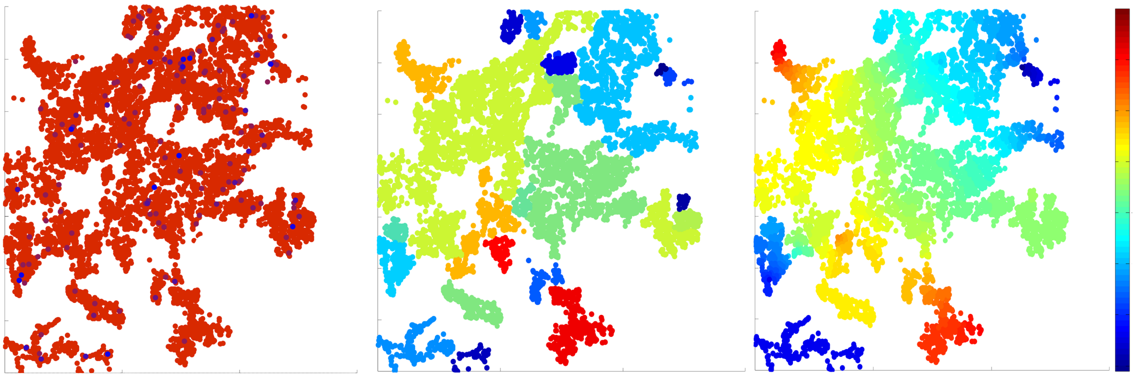


Figure 2: *Left:* Original residuals (motor portfolio). *Center:* Solution to (39) using the incidence matrix of the 10 nearest neighbors graph for regularization. *Right:* Solution using the graph Laplacian.

with $\lambda \geq 0$ and $D = D^{(1,n)}$ being the $(n-1) \times n$ first order finite differences matrix:

$$D^{(1,n)} = \begin{bmatrix} -1 & 1 & & & \\ & -1 & 1 & & \\ & & \cdots & & \\ & & & -1 & 1 \end{bmatrix}. \tag{33}$$

Heuristically, the "fidelity" term $\|\mathbf{y} - \mathbf{w}\|_2^2$ encourages the solution $\mathbf{w}$ to be close to the original data $\mathbf{y}$ and the smoothness term $\|D^{(1,n)}\mathbf{w}\|_2^2$ penalizes non-zero entries of $D^{(1,n)}\mathbf{w}$, first order finite differences (or the discretized first derivative) of $\mathbf{w}$. The value of the parameter λ determines the relative importance of the smoothness term. By Lagrange duality we can obtain the same solution as for (32) with any value of $\lambda \geq 0$ by solving either of the following constrained optimization problems for some values of ϵ_1 and ϵ_2:

$$\begin{array}{llcll} \underset{\mathbf{w}}{\text{minimize}} & \|\mathbf{y} - \mathbf{w}\|_2^2 & \quad & \underset{\mathbf{w}}{\text{minimize}} & \|D\mathbf{w}\|_2^2 \\ \text{subject to} & \|D\mathbf{w}\|_2^2 \leq \epsilon_1 & & \text{subject to} & \|\mathbf{y} - \mathbf{w}\|_2^2 \leq \epsilon_2, \end{array} \tag{34}$$

i.e. the objective and the constraint can be freely interchanged.

Bohlmann's procedure (32) can be extended to penalize k-th order finite differences. In this case the k-th order finite differences matrix $D^{(k,n)} \in \mathbb{R}^{(n-k)\times n}$ can be defined recursively:

$$D^{(k,n)} = D^{(1,n-k)}D^{(k-1,n)}, \; k = 2, 3, \ldots \tag{35}$$

The second order finite differences matrix would then, for example, is as follows:

$$D^{(2,n)} = \begin{bmatrix} -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & \cdots & & & \\ & & & -1 & 2 & -1 \end{bmatrix}. \tag{36}$$

Whittaker [64] had described the underlying probabilistic model and an approximate solution method for the case of third order differences and weighted fidelity term. A Bayesian interpretation of Whittaker graduation is also given by Taylor [51]. Among many extensions to Whittaker graduation most relevant to the present discussion are the work of Schuette [45] and Chan et al. [11, 10]. In his remarkable paper, Schuette proposed the formulation using ℓ_1-norm penalties (weights omitted to simplify presentation):

$$\underset{\mathbf{w}}{\text{minimize}} \ \|\mathbf{y} - \mathbf{w}\|_1 + \lambda \|D^{(k,n)}\mathbf{w}\|_1, \tag{37}$$

After applying a standard transformation this optimization problem can be reformulated as a linear program. In the discussion following [45], S. Klugman had pointed out that the method attempts to make most of the entries of $D^{(k,n)}\mathbf{w}$ (or k-th order differences of $\mathbf{w}$) zero but several of them could be large. This "sparsity inducing" property of ℓ_1-norm penalty motivates its use in the "lasso" variable selection procedure 28.

Chan et al. [10] show that for $p, q \in \{1, 2, \infty\}$ the mixed ℓ_p and ℓ_q norm graduation problem:

$$\underset{\mathbf{w}}{\text{minimize}} \ \|\mathbf{y} - \mathbf{w}\|_p + \lambda \|D^{(k,n)}\mathbf{w}\|_q, \tag{38}$$

can be formulated as a linear program whenever $p, q \in \{1, \infty\}$ and as a quadratic program whenever either p or q is 2.

Smoothing techniques equivalent to Whittaker graduation are known under different names in many fields e.g. "Hodrick-Prescott filter" in economics [24]. More recently, a variant of (38):

$$\underset{\mathbf{w}}{\text{minimize}} \ \|\mathbf{y} - \mathbf{w}\|_2^2 + \lambda \|D^{(1,n)}\mathbf{w}\|_1, \tag{39}$$

with $p = 2$ (or equivalently squared ℓ_2-norm) and $q = 1$ has been popularized in applied statistics literature as "fused lasso" [57]. In signal processing the same formulation is called "total variation denoising". This procedure usually gives a piecewise constant solutions $\mathbf{w}^*$ i.e. discrete first derivative $D^{(1,n)}\mathbf{w}^*$ has mostly zero entries (see top left section of Fig. 1). Similarly, using second order differences $D^{(2,n)}$ often results in a piecewise linear $\mathbf{w}^*$ (see bottom left panel of Fig. 1) and has been described as "ℓ_1 trend filtering" [33] and "quantile splines" [35], the latter replacing the quadratic term with quantile loss. These are effective approaches to change point detection and are considerably simpler than many methods proposed to date.

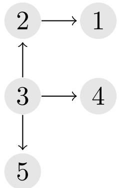
3.2.2 Multidimensional data, spatial smoothing and clustering

We can also apply ℓ_q-norm penalized formulation (38) in situations when the samples $\mathbf{y}$ are over a regular grid or indeed an arbitrary graph (e.g. a k-nearest neighbors graph for objects embedded in a metric space).

To obtain a piecewise constant solution we can use $q = 1$ and the graph incidence matrix for regularization (instead of the first order finite differences matrix $D^{(1,n)}$ in one dimensional case). The incidence matrix A for a graph with n nodes and m edges is a $m \times n$ matrix, with each row representing an edge and composed of a 1 and a -1 in the columns corresponding to the two connected nodes and zeroes elsewhere. See Figure 3 for an example.

To get an equivalent of graduation with second order differences over a regular grid we can consider horizontal and vertical second order differences. As in the one dimensional case, we minimize a weighted sum of the fitting error and,for $q = 1$, a penalty on absolute value of slope changes in the horizontal and vertical directions. The resulting solution tends to be affine over connected regions. The boundaries between regions can be interpreted as curves along which the gradient of the underlying function changes quickly. The approach can not be extended directly to arbitrary graphs. Instead we can use the suitably normalized graph Laplacian (which can be interpreted as a discretization of the Laplace operator) defined as:

$$L = A^T A. \tag{40}$$



$$A = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 0 & 1 \end{bmatrix} \qquad L = A^T A = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 3 & -1 & -1 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 0 & 1 \end{bmatrix}$$

Figure 3: A simple directed graph (*left*), its incidence matrix A (*center*) and Laplacian L (*right*).

Figure 2 shows the results of ℓ_1-norm spatial smoothing applied to geocoded residuals using the incidence matrix of 10 nearest neighbors graph and its Laplacian. The former provides a piecewise constant solution over connected regions, where the regions with the constant fitted value can be interpreted as clusters, and the latter a piecewise affine solution. In the actuarial literature the use of Whittaker graduation to perform spatial smoothing with irregular regions has been proposed by Taylor [53].

It is worth pointing out that the formulation (32) is in fact quite general, given a free choice of matrix D, as it can viewed as a reparametrization of regularized regression. So for the squared ℓ_2 norm penalty the following problems are equivalent (taking $\mathbf{a} = X\mathbf{w}$ and provided X has full row rank):

$$\underset{\mathbf{w}}{\text{minimize}} \quad \|\mathbf{y} - X\mathbf{w}\|_2^2 + \lambda\|\mathbf{w}\|_2^2, \qquad \underset{\mathbf{a}}{\text{minimize}} \quad \|\mathbf{y} - \mathbf{a}\|_2^2 + \lambda\|\mathbf{a}\|_K^2, \tag{41}$$

where $\|\mathbf{a}\|_K^2 = \mathbf{a}^T K\mathbf{a} = \|D\mathbf{a}\|_2^2$ for $K = (XX^T)^{-1} = D^T D$. The second form is known as "Gaussian process regression" [43], or "kriging"[48] in geostatistics literature.

3.3 Additive models

Additive models were first introduced, perhaps, by Ezekiel in 1920s [15] and extended by Hastie some 60 years later [22, 23]. Generalized linear models form predictions based on a linear function of the features:

$$g(\mathbb{E}y) = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_m x_m, \tag{42}$$

where $x_i \in \mathbb{R},\ i = 1, \ldots, m$ are explanatory variables, g is the link function and μ is the expected value of the dependent variable y. Additive models replace the linear combination with a sum of arbitrary functions of explanatory variables:

$$g(\mathbb{E}y) = w_0 + f_1(x_1) + f_2(x_2) + \cdots + f_m(x_m). \tag{43}$$

The richer class of models can lead to overfitting without suitable regularization. The latter is usually achieved by requiring the functions f_i to be (piecewise) smooth.

A similar approach is often followed in the practical applications of GLMs. Continuous variables are discretized into a number of bins. For ordered categorical variables, "natural" levels can be used or if the number of levels is deemed too large, binning is applied to reduce the number of distinct levels. If the total number of bins is not controlled this approach can lead to overfitting and poor predictive performance so it is a standard practice to examine the initial fit and then either manually reduce the number of categories or select a suitable collection of basis functions (e.g. linear or cubic splines) [52].

We observe that it is possible to largely automate this process by casting it into the convex optimization framework and taking advantage of regularization (see Figure 4 for an example). Let $X \in \mathbb{R}^{n\times m}$ be the design matrix for the original problem, then we follow the standard procedure and transform the data by binning each feature into k intervals of equal length (we assume the same number of intervals for every feature for simplicity). This gives a new design matrix $X' \in \mathbb{R}^{n\times km}$

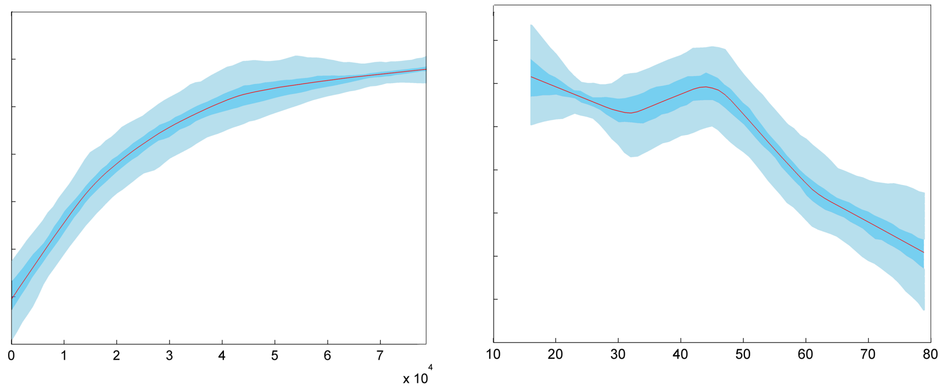


Figure 4: *Left:* The additive effect of sum insured on claims frequency in a motor portfolio. The outer band shows the 90% confidence interval calculated via the bootstrap. *Right:* The additive effect of policyholder age. The anomaly around age 45 could be, for example, due to teenage children driving the family car (also see Figure 5).

and a new parameter vector $\mathbf{w} \in \mathbb{R}^{km}$, with $\mathbf{w}_i \in \mathbb{R}^k$:

$$X' = \begin{bmatrix} X'_1 & ,\dots, & X'_m \end{bmatrix}, \quad \mathbf{w} = \begin{bmatrix} \mathbf{w}_1 \\ \cdots \\ \mathbf{w}_m \end{bmatrix} \tag{44}$$

where X'_p is defined as follows (with $[X]_{i,j}$ denoting x_{ij}):

$$[X'_p]_{i,j} = \begin{cases} 1, & [X]_{i,p} \text{ falls into the } j\text{-th bin} \\ 0, & \text{otherwise} \end{cases} \tag{45}$$

Each observation is now transformed into a sparse vector of dimension km with m non-zero terms. While we lose some information about the features, we can now model non-linear effects in each coordinate. We can write down the parameter estimation problem as:

$$\begin{aligned} \underset{\mathbf{w}}{\text{minimize}} \quad & \mathcal{L}(\mathbf{y}; X'\mathbf{w}) \\ \text{subject to} \quad & \|D\mathbf{w}\|_q \leq \epsilon \end{aligned} \tag{46}$$

or in the equivalent Lagrangian form for some problem specific value of λ:

$$\underset{\mathbf{w}}{\text{minimize}} \quad \mathcal{L}(\mathbf{y}; X'\mathbf{w}) + \lambda \|D\mathbf{w}\|_q \tag{47}$$

where $\mathcal{L}(\mathbf{y}; X'\mathbf{w})$ is an exponential family negative log-likelihood or some other convex loss function and D is a block matrix which evaluates discretized derivatives of the coefficients for each binned explanatory variable:

$$D = \begin{bmatrix} D_1 & & & \\ & D_2 & & \\ & & \cdots & \\ & & & D_m \end{bmatrix}. \tag{48}$$

As shown in section 3.2, we can choose appropriate D_i depending on the structure of the problem and our objectives. Finite difference matrices of up to third order are likely to be sufficient for most applications. The parameter λ can then be chosen by cross-validation. Sometimes it may also

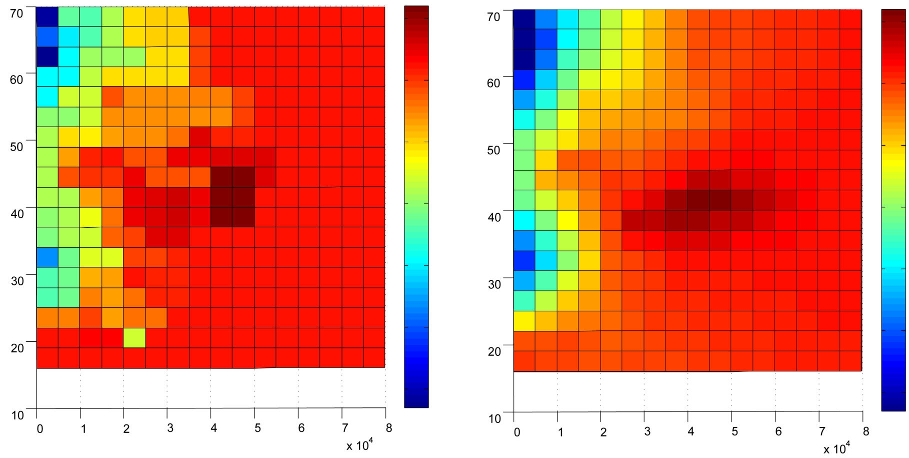


Figure 5: *Left:* Sum insured vs. age interaction for claim frequency in a motor portfolio (same data as Figure 4) using the incidence matrix penalty ("fused lasso"). *Right:* Same interaction but penalising the second order discrete derivative in both coordinates.

be appropriate to introduce independent smoothing parameters (but will make their estimation more difficult):

$$\underset{\mathbf{w}}{\text{minimize}} \;\; \mathcal{L}(\mathbf{y}; X'\mathbf{w}) + \lambda_1 \|D_1 \mathbf{w}_1\|_q + \; \ldots \; + \lambda_m \|D_m \mathbf{w}_m\|_q. \tag{49}$$

Notably we can still perform variable selection analogous to the standard ℓ_1-norm regularisation (28) by adding a "group lasso" penalty [67]:

$$\underset{\mathbf{w}}{\text{minimize}} \;\; \mathcal{L}(\mathbf{y}; X'\mathbf{w}) + \lambda \|D\mathbf{w}\|_q + \mu \sum_{i=1}^{m} \|\mathbf{w}_i\|_2. \tag{50}$$

Here the term $\sum_{i=1}^{m} \|\mathbf{w_i}\|_2$ encourages the individual components of w to go uniformly to zero as the parameter μ is increased.

3.3.1 Variable interactions

In a Gaussian additive model with identity link function the effect of all the explanatory variables is a sum of their individual effects. Individual effects show how the expected response varies as any single explanatory variable is changed with the others held constant at arbitrary values. For example in order to maximize the expected response we only need to separately maximize each of the component functions of the additive model.

In general there are no guarantees that an additive model will provide a satisfactory fit in any given situation. Non-additivity means that, as one explanatory variable is varied, the effect on the expected response depends on the fixed values of the other variables. Below is an example of how we would change equation (43) if the model is non-additive in variables x_1 and x_2:

$$g(\mu) = w_0 + h(x_1, x_2) + f_3(x_3) + \cdots + f_m(x_m). \tag{51}$$

We can model non-additivity by including the corresponding interactions. Using the notation from equation 45 we can define the interaction $X'_{(1,2)}$ as:

$$[X'_{(1,2)}]_{i,*} = [X'_1]_{i,*} \otimes [X'_2]_{i,*}, \tag{52}$$

where e.g. $[X'_1]_{i,*}$ denotes the i-th row of X'_1 and $\otimes$ is the Kronecker product.

In the resulting model, we can penalize first, second or higher order derivatives in x_1 and x_2 coordinates (see Figure 5), cf. smoothing over a regular grid in section (3.2). Another possibility is to work with the graph Laplacian, which in this case has a particularly simple form:

$$L_{(1,2)} = I \otimes D' + D' \otimes I, \tag{53}$$

with D' a $k \times k$ variant of the second differences matrix $D^{(2,k)}$ with zero padding.

3.3.2 Spatial smoothing and clustering

As geographic regions are not usually arranged in a regular grid the approach in the previous section is not directly applicable for spatial smoothing. Instead, we can introduce indicator variables for each geographic grouping present in the data (postcode, census zone or even unique coordinates) and use a suitably constructed graph representing distance and adjacency information for regularization. Depending on whether we use the graph incidence matrix or the Laplacian, for $q = 1$ we can obtain either a piecewise constant or a smoothly varying surface (see Figure 2) jointly with the other model parameters:

$$\underset{\mathbf{w}}{\text{minimize}} \;\; \mathcal{L}(\mathbf{y}; X'\mathbf{w}) + \lambda \left\| D\mathbf{w} \right\|_1, \tag{54}$$

where D is a block diagonal matrix (cf. equation 48) and one of the blocks is either the aforementioned graph incidence matrix or the graph Laplacian. As before, the value of λ can be chosen by cross-validataion.

3.4 Kalman filter and dynamic models

Kalman filter [32] and related ideas have played a central role in the development of state space methods in engineering control through out 1960s (culminating in the linear quadratic Gaussian theory). Remarkably, the first practical application of the Kalman filter was to improve the accuracy of navigation for the Apollo program, quickly followed by adoption for a wide range of aerospace problems. In these applications the goal is typically to track the "state" of a missile or a spacecraft obeying Newtonian dynamics. The state vector would contain the current position together with velocity and acceleration vectors and the goal would be to repeatedly re-estimate the state using measurements coming in from a range of sensors, such as inertial, optical, ground based radar etc.

Kalman filter also has quite a long history in the actuarial literature both as a generalization of the classical linear credibility models [39, 30] and applied to optimal updating of claim reserves [29, 54]. Below we formulate the Kalman filter as an optimization problem and describe some intuitive extensions that make the technique directly applicable to tasks such as ongoing monitoring of conversion rates, claim frequencies or other aspects of portfolio performance.

3.4.1 Kalman filter as an optimization problem

Consider the standard least squares problem (see also Appendix B.1):

$$\underset{\mathbf{w}}{\text{minimize}} \;\; \|\mathbf{y} - X\mathbf{w}\|_2^2, \tag{55}$$

and partition the design matrix X and the response vector $\mathbf{y}$ into m row blocks, corresponding to time periods:

$$X = \begin{bmatrix} X_1 \\ \cdots \\ X_m \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \cdots \\ \mathbf{y}_m \end{bmatrix}. \tag{56}$$

We can then equivalently transform (55) by introducing m copies of the parameter vector and some linear equality constraints:

$$\begin{array}{ll} \underset{\mathbf{w}_1,\ldots,\mathbf{w}_m}{\text{minimize}} & \sum_{t=1}^{m} \|\mathbf{y}_t - X_t\mathbf{w}_t\|_2^2 \\ \text{subject to} & \mathbf{w}_{t+1} - \mathbf{w}_{t+1} = \mathbf{0}, \;\; t = 1,\ldots,m-1. \end{array} \tag{57}$$

The above model can also be written in the state space form with the identity state transition matrix, no state transition noise and i.i.d. Gaussian observation noise[6], where $\mathbf{w}_t$ is the unobserved state vector and $\mathbf{y}_t$ are the observations associated with time dependent observation matrices X_t:

$$\begin{array}{ll} \mathbf{w}_{t+1} = \mathbf{w}_t, & \mathbf{y}_t = X_t\mathbf{w}_t + \boldsymbol{\epsilon}_t, \\ & \boldsymbol{\epsilon}_t \sim \mathcal{N}(0, I). \end{array} \tag{58}$$

By introducing i.i.d. Gaussian state transition noise:

$$\begin{array}{ll} \mathbf{w}_{t+1} = \mathbf{w}_t + \boldsymbol{\nu}_t, & \mathbf{y}_t = X_t\mathbf{w}_t + \boldsymbol{\epsilon}_t, \\ \boldsymbol{\nu}_t \sim \mathcal{N}(0, I), & \boldsymbol{\epsilon}_t \sim \mathcal{N}(0, I) \end{array} \tag{59}$$

we effectively relax the equality constraints in (57) replacing them with a squared ℓ_2-norm penalty term. It is then possible to perform estimation by solving the following convex optimisation problem:

$$\underset{\mathbf{w}_1,\ldots,\mathbf{w}_m}{\text{minimize}} \;\; \sum_{t=1}^{m} \|\mathbf{y}_t - X_t\mathbf{w}_t\|_2^2 + \sum_{t=1}^{m-1} \|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2, \tag{60}$$

which amounts to substituting every independent variable in the model by its interaction with the time index variable t and regularizing the differences in the corresponding parameters. The estimation problem can be either solved directly or transformed back to the standard least squares form:

$$\underset{\mathbf{w}}{\text{minimize}} \;\; \|\mathbf{y}' - X'\mathbf{w}\|_2^2, \tag{61}$$

where the design matrix X' and the response vector $\mathbf{y}'$ are redefined as follows:

$$X' = \begin{bmatrix} X_1 & & & \\ -I & I & & \\ & & \cdots & & \\ & & & -I & I \\ & & & & X_m \end{bmatrix}, \quad \mathbf{y}' = \begin{bmatrix} \mathbf{y}_1 \\ 0 \\ \cdots \\ 0 \\ \mathbf{y}_m \end{bmatrix}, \quad \mathbf{w} = \begin{bmatrix} \mathbf{w}_1 \\ \cdots \\ \mathbf{w}_m \end{bmatrix}. \tag{62}$$

The formulation in (60) simultaneously performs both "filtering" and "smoothing" conditional on all the observations up to time m. If new information becomes available the augmented optimization problem should be solved again to obtain new estimates of the entire history of state transitions $\mathbf{w}^* = [\mathbf{w}_1^*, \ldots, \mathbf{w}_m^*, \mathbf{w}_{m+1}^*]^T$. Standard Kalman filter given information up to time $m+1$, on the other hand, only updates the estimate of the current state $\mathbf{w}_{m+1}^*$ and requires a backward

[6] We can in fact avoid the Gaussianity assumption by positing a quadratic loss function instead. The estimators obtained are the same in both cases.

"smoothing" pass to update estimates of past states $\mathbf{w}_1, \ldots, \mathbf{w}_m$.

Indeed, the Kalman filter followed by a "smoothing" step can be viewed as a computationally efficient recursive procedure for solving the normal equations of the least squares problem (61) which exploits block tri-diagonal structure of the matrix X'. With advances in numerical linear algebra routines for sparse matrices and increasing computer speeds very large problems of this kind can be solved directly.

So far we focused on the special case with the identity state transition matrix and i.i.d noise, however state estimation in the the general linear Gaussian state space model:

$$\begin{aligned} \mathbf{w}_{t+1} &= F\mathbf{w}_t + \boldsymbol{\nu}_t, \qquad & \mathbf{y}_t &= X_t\mathbf{w}_t + \boldsymbol{\epsilon}_t, \\ \boldsymbol{\nu}_t &\sim \mathcal{N}(0, \Sigma_\nu), \qquad & \boldsymbol{\epsilon}_t &\sim \mathcal{N}(0, \Sigma_\epsilon) \end{aligned} \tag{63}$$

can also be expressed as a convex optimisation problem. Denoting $\|\mathbf{a}\|_P = (\mathbf{a}^T P\mathbf{a})^{\frac{1}{2}}$, P-quadratic norm for a positive definite matrix P, it is:

$$\underset{\mathbf{w}_1,\ldots,\mathbf{w}_m}{\text{minimize}} \quad \sum_{t=1}^{m} \|\mathbf{y}_t - X_t\mathbf{w}_t\|^2_{\Sigma_\epsilon^{-1}} + \sum_{t=1}^{m-1} \|\mathbf{w}_{t+1} - F\mathbf{w}_t\|^2_{\Sigma_\nu^{-1}}. \tag{64}$$

Notably we can recover both Whittaker graduation and Jones-Gerber "evolutionary" credibility [28] as state space models by choosing a one dimensional state vector w_t with the observation matrix X_t a constant vector $\mathbf{1}$ [30, 60]:

$$\underset{w_1,\ldots,w_m}{\text{minimize}} \quad \sum_{t=1}^{m} \|\mathbf{y}_t - \mathbf{1}w_t\|_2^2 + \sum_{t=1}^{m-1} \|w_{t+1} - w_t\|_2^2. \tag{65}$$

When there is only a single observation per time step, this is identical to Whittaker graduation with first order differences as in (32), while the Gerber-Jones model allows multiple observations per time period.

3.4.2 Some extensions to the dynamic models:

We can use any convex loss for the observations, such as quantile or logistic, and still end up with a convex optimization problem:

$$\underset{\mathbf{w}_1,\ldots,\mathbf{w}_m}{\text{minimize}} \quad \sum_{t=1}^{m} \mathcal{L}(\mathbf{y}_t; X_t\mathbf{w}_t) + \sum_{t=1}^{m-1} \|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2. \tag{66}$$

In the context of dynamic generalized linear models this corresponds to posterior mode estimation as proposed by Fahrmeier [17, 16]. Non-Gaussian state noise is another possibility. It may be appropriate to apply ℓ_1-norm penalty to state changes provided most of the time parameters stay constant with occasional large jumps (cf. Figure 1):

$$\underset{\mathbf{w}_1,\ldots,\mathbf{w}_m}{\text{minimize}} \quad \sum_{t=1}^{m} \mathcal{L}(\mathbf{y}_t; X_t\mathbf{w}_t) + \sum_{t=1}^{m-1} \|\mathbf{w}_{t+1} - \mathbf{w}_t\|_1. \tag{67}$$

Another possibility is a combination of norms, e.g. an approach combining squared ℓ_2-norm and ℓ_1-norm penalties will attempt to decompose the state trajectory into a smooth and a piecewise constant component:

$$\underset{\mathbf{w},\mathbf{c}}{\text{minimize}} \quad \sum_{t=1}^{m} \mathcal{L}(\mathbf{y}_t; X_t(\mathbf{w}_t + \mathbf{c}_t)) + \lambda \sum_{t=1}^{m-1} \|\mathbf{w}_{t+1} - \mathbf{w}_t\|_1 + \mu \sum_{t=1}^{m-1} \|\mathbf{c}_{t+1} - \mathbf{c}_t\|_2^2. \tag{68}$$

We can also allow linear trends in the parameters (this formulation can be reduced to the standard state space model by expanding the state vector):

$$\underset{\mathbf{w}_1,\ldots,\mathbf{w}_m}{\text{minimize}} \quad \sum_{t=1}^{m} \mathcal{L}(\mathbf{y}_t; X_t \mathbf{w}_t) + \sum_{t=1}^{m-2} \|\mathbf{w}_{t+2} - 2\mathbf{w}_{t+1} + \mathbf{w}_t\|_1. \tag{69}$$

Finally it is possible to add arbitrary convex inequality and linear equality constraints (see section 3.5.2), so e.g. seasonality adjustments can be handled by introducing new variables $\mathbf{c}_1, \ldots, \mathbf{c}_m$ and equality constraints:

$$\begin{array}{ll} \underset{\mathbf{w},\mathbf{c}}{\text{minimize}} & \sum_{t=1}^{m} \mathcal{L}(\mathbf{y}_t; X_t(\mathbf{w}_t + \mathbf{c}_t)) + \sum_{t=1}^{m-1} \|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2 \\ \text{subject to} & \mathbf{c}_t = \mathbf{c}_{t+k}, \ \ t = 1, \ldots, m-k \\ & \sum_{t=1}^{m} \mathbf{c}_t = 0. \end{array} \tag{70}$$

Formulating state space models as regularized regression can make them considerably more intuitive for those lucky not to have a background in control theory.

3.5 Other applications

3.5.1 Hierarchical credibility

We can describe hierarchical credibility [27, 55] models (closely related to both "random effects" from statistics literature [41, 1] and linear filtering [30, 39]) in the optimization framework provided variances are known. This can be achieved by introducing additional variables to the optimization problem. Consider a simple setup with two risks I_1 and I_2 with observations $y_i, \ i \in I_j$, for $j \in \{1, 2\}$ with unit variance and the following group mean:

$$\mathbb{E}y_i = w_j, \quad \forall i \in I_j, \tag{71}$$

where w_j are themselves random variables with group mean w and known variance:

$$\mathbb{E}w_j = w, \ \ \text{Var}(w_j) = \sigma^2, \quad j = 1, 2. \tag{72}$$

We can then obtain the linear credibility estimator as a solution of the following optimization problem:

$$\underset{w_1,w_2,w}{\text{minimize}} \quad \sum_{j=1}^{2} \sum_{i \in I_j} (y_i - w_j)^2 + \frac{1}{\sigma^2} \sum_{j=1}^{2} (w_j - w)^2. \tag{73}$$

While the notation above is rather cumbersome, complex models are easy to implement in practice. This can be achieved by augmenting the data with indicator variables for the lowest level of the hierarchy and regularizing by the squared Euclidean norm of the product of the (tree) graph incidence matrix representing the hierarchical structure and the parameter vector.

3.5.2 Constraints on relativities

Treating maximum likelihood estimation of GLM parameters as a convex optimization problem allows us to introduce arbitrary convex constraints on rate relativities in addition to various smoothness penalties described earlier. One such constraint in the classical GLM theory is the "offset", used e.g. to allow for exposure in Poisson models. It amounts to setting the coefficient associated with the offset term x_i equal to one:

$$\begin{aligned} \underset{\mathbf{w}}{\text{minimize}} \quad & \mathcal{L}(\mathbf{y}; X\mathbf{w}) \\ \text{subject to} \quad & w_i = 1. \end{aligned} \tag{74}$$

Given the range of practical insurance applications of this simple device [66], the overall approach seems promising.

Consider for example an additive model with binned variables and the associated regularization term (47). Possible constraints would include bounds on the absolute magnitude of the effect $\mathbf{w}_i$ associated with the i-th risk factor or its rate of growth (first differences):

$$\begin{aligned} \underset{\mathbf{w}}{\text{minimize}} \quad & \mathcal{L}(\mathbf{y}; X'\mathbf{w}) + \lambda \|D\mathbf{w}\|_1 \\ \text{subject to} \quad & \mathbf{w}_i \geq -\epsilon_1 \mathbf{1} \\ & \mathbf{w}_i \leq \epsilon_1 \mathbf{1} \end{aligned} \qquad \begin{aligned} \underset{\mathbf{w}}{\text{minimize}} \quad & \mathcal{L}(\mathbf{y}; X'\mathbf{w}) + \lambda \|D\mathbf{w}\|_1 \\ \text{subject to} \quad & D^{(1,k)}\mathbf{w}_i \geq -\epsilon_2 \mathbf{1} \\ & D^{(1,k)}\mathbf{w}_i \leq \epsilon_2 \mathbf{1}. \end{aligned} \tag{75}$$

In the above $\geq$ denotes entrywise vector inequality and the constraints can alternatively be written as:

$$|w_{ij}| \leq \epsilon_1, \;\; i = 1, \ldots, k \qquad |w_{i,j+1} - w_{ij}| \leq \epsilon_2, \;\; i = 1, \ldots, k-1. \tag{76}$$

We can also directly control the shape of the additive effect in a particular variable, requiring it, for example, to be non-decreasing or convex (these two might be appropriate for sum insured and driver age respectively):

$$\begin{aligned} \underset{\mathbf{w}}{\text{minimize}} \quad & \mathcal{L}(\mathbf{y}; X'\mathbf{w}) + \lambda \|D\mathbf{w}\|_1 \\ \text{subject to} \quad & D^{(1,k)}\mathbf{w}_i \geq \mathbf{0} \end{aligned} \qquad \begin{aligned} \underset{\mathbf{w}}{\text{minimize}} \quad & \mathcal{L}(\mathbf{y}; X'\mathbf{w}) + \lambda \|D\mathbf{w}\|_1 \\ \text{subject to} \quad & D^{(2,k)}\mathbf{w}_i \geq \mathbf{0} \end{aligned} \tag{77}$$

Here monotonicity constraints[7] can be written without matrix notation as:

$$w_{i1} \leq w_{i2} \leq \; \ldots \; \leq w_{ik} \tag{78}$$

and similarly for convexity we obtain:

$$w_{ij} - 2w_{i,j+1} + w_{i,j+2} \quad j = 1, \ldots, k-2. \tag{79}$$

Finally consider a problem where there are two known risks ξ_1 and ξ_2 for which we want the percentage difference in premium to be within a certain range ϵ (either due to market or regulatory considerations). Assuming the model uses the log link, this yields:

$$\begin{aligned} \underset{\mathbf{w}}{\text{minimize}} \quad & \mathcal{L}(\mathbf{y}; X\mathbf{w}) \\ \text{subject to} \quad & |\mathbf{w}^T\xi_1 - \mathbf{w}^T\xi_2| \leq \log(1+\epsilon). \end{aligned} \tag{80}$$

3.5.3 Demand based pricing

In recent years optimization has most often been mentioned in insurance context in relation to various aspects of "demand driven" pricing which amounts to varying margins in order to maximize underwriting profit. Below we present a very simple model for pricing renewals which admits

[7] Remarkably, there are entire monographs devoted to this topic under the name of *isotonic regression.*

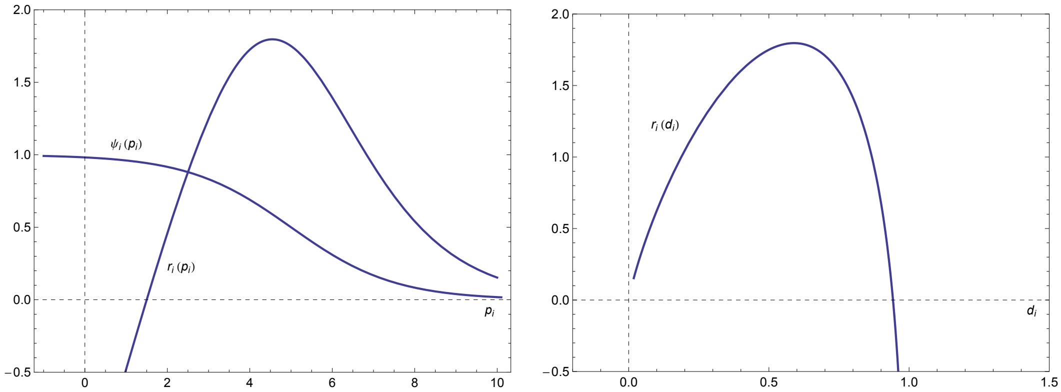


Figure 6: ***Left:*** Logistic demand $\psi_i(p_i)$ and revenue $r_i(p_i) = (p_i - c_i)\, \psi_i(p_i)$ as functions of price p_i. ***Right:*** Revenue $r(d_i) = \left(\psi_i^{-1}(d_i) - c_i\right) d_i$ as a function of demand d_i is concave for $d_i \in [0, 1]$.

a convex reformulation:

$$\begin{aligned} \underset{p_1,\ldots,p_n}{\text{maximize}} \quad & \sum_{i=1}^{n} (p_i - c_i)\, \psi_i(p_i) \\ \text{subject to} \quad & \sum_{i=1}^{n} \psi_i(p_i) \geq C. \end{aligned} \tag{81}$$

The model maximizes the total revenue objective for a cohort of n policies subject to a constraint on minimum retention level, where for i-th policyholder p_i is the proposed premium, $d_i = \psi_i(p_i)$ is the expected demand as a function of premium and c_i is the expected cost of claims. This optimization problem is not convex in general, however, for monotone demand functions from certain parametric families e.g. logistic and probit (see [50] for a detailed discussion) we can obtain a convex equivalent (or rather a maximization problem with a concave objective) parametrized by expected demand d_i (see Figure 6):

$$\begin{aligned} \underset{d_1,\ldots,d_n}{\text{maximize}} \quad & \sum_{i=1}^{n} \left(\psi_i^{-1}(d_i) - c_i\right) d_i \\ \text{subject to} \quad & 0 \leq d_i \leq 1,\ i = 1, \ldots, n \\ & \sum_{i=1}^{n} d_i \geq C. \end{aligned} \tag{82}$$

Here $\psi_i^{-1}(d_i)$ is the inverse demand function which can be uniquely defined for $d_i \in [0, 1]$.

It is worth noting that both claim costs and demand are usually not known precisely or even with reasonable accuracy and the rates obtained from (82) do not in fact maximize the expected revenue, but rather the "certainty equivalent" objective where we have replaced stochastic demand and claim costs with their expected values. A more accurate model would have the following form:

$$\begin{aligned} \underset{p_1,\ldots,p_n}{\text{maximize}} \quad & \mathbb{E}_\omega\left(\sum_{i=1}^{n} (p_i - c_i(\omega_i))\, \psi_i(p_i, \omega_i)\right) \\ \text{subject to} \quad & \mathbb{E}_\omega\left(\sum_{i=1}^{n} \psi_i(p_i, \omega_i)\right) \geq C, \end{aligned} \tag{83}$$

where the expectation is with respect to the joint distribution of individual demand functions and claims costs. Clearly some simplifying assumptions about the distribution will need to be made in order to make its estimation tractable. The interpretation of the volume constraint holding in expectation is also quite difficult. To address the latter point it may be appropriate to consider e.g.

the expected shortfall (reminiscent of conditional risk measures):

$$\mathbb{E}_\omega\Big(\big[C - \sum_{i=1}^{n} \psi_i(p_i, \omega_i)\big]_+\Big) \leq \epsilon. \tag{84}$$

Techniques that attempt to address this type of problems usually go under the name of stochastic optimization [47].

4 Solving convex optimization problems

Until recently, solving convex optimization problems required not inconsiderable subject matter expertise and even increasing availability of high quality open source and commercial solvers (SDPT3, SEDUMI, MOSEK, CVXOPT) did not allay the situation. The reason for this is that these solvers require problems to be converted to one of several restrictive standard forms (e.g. a second order cone program). In some cases such a transformation is not possible and it may be necessary to develop a custom solver. For a potential user with a focus on applications these requirements are challenging to say the least.

This has recently changed with the introduction of CVX [20, 21], a high level specification language for general convex optimization. It provides a convenient interface for specifying convex optimization problems and then automates underlying mathematical steps for analysing and solving them. The underlying solvers can often handle sparse instances with millions of variables. While the idea of a an optimization modelling language is not new (e.g. GAMS), existing commercial offerings can not handle general convex problems.

Below is sample CVX/MATLAB code[8] for the problem (39):

```
n = length(y);  %length of timeseries y
m = n-2;        %length of Dw
lambda = 15;

I2 = speye(m,m);
O2 = zeros(m,1);
D = [I2 O2 O2]+[O2 -2*I2 O2]+[O2 O2 I2];

cvx_begin
   variable w(n)
   minimize(sum_squares(y-w)+lambda*norm(D*w,1))
cvx_end
```

Due to some limitations in the way the current version of CVX deals with problems that are not SDP representable (e.g. loss functions involving logarithms and exponentials) for large problems it may sometimes be necessary to implement an iterative reweighting scheme similar to Fisher scoring, such as in [37], with CVX in the inner loop. For very large problems decomposition methods are available, see [5] for an accessible introduction and examples.

[8] Please contact the author for larger examples described in this paper.

Appendix

A Basics of convex calculus

As mentioned earlier, it is often the easiest to verify convexity by checking whether a given function is composed from known convex primitives by applying convexity preserving transformations. Below we list a subset of these sufficient to verify convexity of problems described in this paper.

A.1 Functions of a single variable:

- *Exponential.* $e^{\alpha x}$ is convex on $\mathbb{R}$ for any $\alpha \in \mathbb{R}$.
- *Powers.* x^{α} is convex on $(0, +\infty)$ for $\alpha \geq 1$ or $\alpha \leq 0$, and concave for $0 \leq \alpha \leq 1$.
- *Logarithm.* $\log x$ is concave on $(0, +\infty)$.

A.2 Vector norms:

All vector norms on $\mathbb{R}^n$ are convex. A function $f : \mathbb{R}^n \to \mathbb{R}$ is called a norm if:

- $f(\mathbf{x}) \geq 0$ for all $\mathbf{x} \in \mathbb{R}^n$,
- $f(\mathbf{x}) = 0$ only if $\mathbf{x} = 0$,
- $f(\alpha\mathbf{x}) = |\alpha| f(\mathbf{x})$, for all $\mathbf{x} \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$,
- f satisfies the triangle inequality: $f(\mathbf{x} + \mathbf{y}) \leq f(\mathbf{x}) + f(\mathbf{y})$, for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.

Such functions are usually denoted $\|\mathbf{x}\|$ rather than $f(\mathbf{x})$ with a subscript to indicate the exact norm being used. They can be interpreted as measuring the length of the elements of $\mathbb{R}^n$. Most common is the Euclidean or the ℓ_2-norm defined as:

$$\|\mathbf{x}\|_2 = \sqrt{\mathbf{x}^T\mathbf{x}} = \left(\sum_i x_i^2\right)^{\frac{1}{2}}. \tag{85}$$

Two other examples of norms on $\mathbb{R}^n$ are the absolute value or ℓ_1-norm:

$$\|\mathbf{x}\|_1 = \sum_i |x_i| \tag{86}$$

and the max or ℓ_∞-norm, given by:

$$\|\mathbf{x}\|_\infty = \max\left(|x_1|, \ldots, |x_n|\right). \tag{87}$$

These can be generalized as the ℓ_p-norm, for $p \geq 1$:

$$\|\mathbf{x}\|_p = \left(\sum_i |x_i|^p\right)^{\frac{1}{p}}, \tag{88}$$

where ℓ_∞-norm obtains as $p \to \infty$. It is easy to check convexity of norms from the definition.

A.3 Other functions on $\mathbb{R}^n$

- *Max function.* The function $\max(x_1, \ldots, x_n)$ is convex on $\mathbb{R}^n$. Note that it is distinct from the ℓ_∞-norm.
- *Soft max.* The so called "soft max" function $\log(e^{x_1} + \cdots + e^{x_n})$ is convex on $\mathbb{R}^n$. It can be viewed as a differentiable approximation of $\max(x_1, \ldots, x_n)$.
- *Geometric mean.* The geometric mean $f(\mathbf{x}) = (\prod_{i=1}^n x_i)^{\frac{1}{n}}$ is concave for $x_i > 0,\ i = 1, \ldots, n$.

A.4 Transformations that preserve convexity

Convexity is maintained for certain compositions of convex functions, for example:

A.4.1 Weighted sums:

A weighted sum of convex functions

$$f = \alpha_1 f_1 + \cdots + \alpha_n f_n \tag{89}$$

is convex if all the weights are non-negative, $\alpha_i \geq 0,\ i = 1, \ldots, n$. Similarly a weighted sum of concave functions is concave for non-negative weights.

A.4.2 Composition with an affine function:

If f is a convex function, $f : \mathbb{R}^m \to \mathbb{R}$, A is an $m \times n$ matrix and $\mathbf{b}$ a vector in $\mathbb{R}^n$ then $g : \mathbb{R}^n \to \mathbb{R}$:

$$g(\mathbf{x}) = f(A\mathbf{x} + \mathbf{b}) \tag{90}$$

is also a convex function. Similarly, g is concave if f is concave.

A.4.3 Composition of multivariable functions:

Suppose that a function $f : \mathbb{R}^n \to \mathbb{R}$ has the following form:

$$f(\mathbf{x}) = h(g(\mathbf{x})) = h(g_1(\mathbf{x}), \ldots, g_m(\mathbf{x})), \tag{91}$$

where $g_i : \mathbb{R}^n \to \mathbb{R},\ i = 1, \ldots, m$. Then the following holds:

- f is convex if functions g_i are convex and h is convex and non-decreasing in each argument,
- f is convex if g_i are concave and h is non-increasing in each argument.

B Some common GLMs as convex problems

B.1 Gaussian

Assume $y_i \in R$ follows a Gaussian distribution with known variance σ^2 and mean μ_i then its probability density function is:

$$p(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(y_i - \mu_i)^2}{2\sigma^2}\right). \tag{92}$$

We can then express μ_i as a linear function (ignoring the intercept term for clarity) of the vector of explanatory variables $\mathbf{x}_i \in \mathbb{R}^m$, parametrized by $\mathbf{w} \in \mathbb{R}^m$:

$$\mu_i = \mathbf{w}^T \mathbf{x}_i. \tag{93}$$

The likelihood function for n independent observations $(y_i, \mathbf{x}_i)$ is given by:

$$p(\mathbf{y}; X\mathbf{w}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(y_i - \mathbf{w}^T\mathbf{x}_i)^2}{2\sigma^2}\right), \tag{94}$$

where X is the design matrix with rows $\mathbf{x}_i$ for $i = 1, \ldots, n$ and the loss or negative log-likelihood function is as follows:

$$\begin{aligned} \mathcal{L}(\mathbf{y}; X\mathbf{w}) &= -\sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(y_i - \mathbf{w}^T\mathbf{x}_i)^2}{2\sigma^2}\right)\right) \\ &= \sum_{i=1}^{n} \frac{1}{2\sigma^2}(y_i - \mathbf{w}^T\mathbf{x}_i)^2 + \frac{1}{2}\log(2\pi\sigma^2). \end{aligned} \tag{95}$$

We can obtain an estimate of $\mathbf{w}$ by solving a convex optimization problem (to check convexity in $\mathbf{w}$, observe that it is a square of ℓ_2-norm composed with an affine function):

$$\underset{\mathbf{w}}{\text{minimize}} \quad \sum_{i=1}^{n}(y_i - \mathbf{w}^T\mathbf{x}_i)^2, \tag{96}$$

which in matrix notation can be restated as:

$$\underset{\mathbf{w}}{\text{minimize}} \quad \|\mathbf{y} - X\mathbf{w}\|_2^2. \tag{97}$$

B.2 Poisson

Take $y_i \in \mathbb{Z}^+$ to be a random variable with Poisson distribution and mean $\mu_i > 0$:

$$P(y_i = k) = \frac{e^{-\mu_i}\mu_i^k}{k!}. \tag{98}$$

The mean μ_i can be modelled via the log link as a linear function of the vector of explanatory variables $\mathbf{x}_i \in \mathbb{R}^m$:

$$\mu_i = \exp(\mathbf{w}^T\mathbf{x}_i). \tag{99}$$

Here $\mathbf{w} \in \mathbb{R}^m$ is the parameter vecor. Given n independent observations $(y_i, \mathbf{x_i}),\ i = 1, \ldots, n$, the likelihood function is as follows:

$$p(\mathbf{y}; X\mathbf{w}) = \prod_{i=1}^{n} \frac{\exp\left(-\exp(\mathbf{w}^T\mathbf{x}_i)\right)\exp(\mathbf{w}^T\mathbf{x}_i)^{y_i}}{y_i!}, \tag{100}$$

with the corresponding loss or negative log-likelihood:

$$\mathcal{L}(\mathbf{y}; X\mathbf{w}) = \sum_{i=1}^{n} \left(\exp(\mathbf{w}^T\mathbf{x}_i) - y_i\mathbf{w}^T\mathbf{x}_i + \log(y_i!)\right). \tag{101}$$

A maximum likelihood estimate of w can then be obtained by solving the convex loss minimization problem (convexity follows directly from the composition rules defined in A.4):

$$\underset{\mathbf{w}}{\text{minimize}} \quad \sum_{i=1}^{n} \left(\exp(\mathbf{w}^T\mathbf{x}_i) - y_i\mathbf{w}^T\mathbf{x}_i\right). \tag{102}$$

B.3 Logistic

Consider a random variable $y_i \in \{0, 1\}$, with the binomial distribution:

$$P(y = 1) = \mu_i, \;\; P(y = 0) = 1 - \mu_i, \tag{103}$$

where $\mu_i \in [0, 1]$ is the expected value of y_i. It depends on a linear function of a vector of explanatory variables $\mathbf{x}_i \in \mathbb{R}^m$ via the logistic link:

$$\mu_i = \frac{\exp(\mathbf{w}^T\mathbf{x})}{1 + \exp(\mathbf{w}^T\mathbf{x})}, \tag{104}$$

where $\mathbf{w} \in \mathbb{R}^m$ is the vector of model parameters. Given n independent observations $(y_i, \mathbf{x}_i), i = 1, \ldots, n$ and defining $I_0 = \{i \mid y_i = 0\}$ and $I_1 = \{i \mid y_i = 1\}$ we obtain the following likelihood function:

$$p(\mathbf{y}; X\mathbf{w}) = \prod_{i \in I_1} \frac{\exp(\mathbf{w}^T\mathbf{x}_i)}{1 + \exp(\mathbf{w}^T\mathbf{x}_i)} \prod_{i \in I_0} \left(1 - \frac{\exp(\mathbf{w}^T\mathbf{x}_i)}{1 + \exp(\mathbf{w}^T\mathbf{x}_i)}\right). \tag{105}$$

The corresponding negative log-likelihood or loss function has the form:

$$\begin{aligned} \mathcal{L}(\mathbf{y}; X\mathbf{w}) &= -\sum_{i \in I_1} \log\left(\frac{\exp(\mathbf{w}^T\mathbf{x}_i)}{1 + \exp(\mathbf{w}^T\mathbf{x}_i)}\right) - \sum_{i \in I_0} \log\left(1 - \frac{\exp(\mathbf{w}^T\mathbf{x}_i)}{1 + \exp(\mathbf{w}^T\mathbf{x}_i)}\right) \\ &= -\sum_{i \in I_1} \mathbf{w}^T\mathbf{x}_i + \sum_{i=1}^{n} \log\left(1 + \exp(w^T\mathbf{x}_i)\right) \end{aligned} \tag{106}$$

It is easy to check that $\mathcal{L}(\mathbf{y}; X\mathbf{w})$ is a convex function in w (composition of soft max and affine). Maximum likelihood estimation for logistic regression is then equivalent to a convex loss minimization problem:

$$\underset{\mathbf{w}}{\text{minimize}} \quad -\sum_{i \in I_1} \mathbf{w}^T\mathbf{x}_i + \sum_{i=1}^{n} \log\left(1 + \exp(\mathbf{w}^T\mathbf{x}_i)\right). \tag{107}$$

References

[1] K. Antonio and J. Beirlant. Actuarial statistics with generalized linear mixed models. *Insurance: Mathematics and Economics*, 40:58–76, 2007.

[2] D. Bertsekas. *Nonlinear programming*. Athena Scientific Belmont, 1999.

[3] D Bertsimas and J Tsitsiklis. *Introduction to Linear Optimization*. Athena Scientific, 1st edition, 1997.

[4] G. Bohlmann. Ein Ausgleichungsproblem. *Nachrichten Gesellschaft Wissenschaften Gottingen. Math.-Phys. Klasse*, pages 260–271, 1899.

[5] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3:1–122, 2011.

[6] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Presss, 2004.

[7] P. Brockett. Information theoretic approach to actuarial science: a unification and extension of relevant theory and applications. *Transactions of the Society of Actuaries*, 43:73–114, 1991.

[8] L. Brown. *Fundamentals of statistical exponential families with applications in statistical decision theory*. Institute of Mathematical Statistics, Hayward, CA, 1986.

[9] E. Candes, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 2006.

[10] F. Y. Chan, L. K. Chan, J. Falkenberg, and M. H. Yu. Applications of linear and quadratic programming to some cases of the Whittaker-Henderson graduation method. *Scandinavian Actuarial Journal*, pages 141–153, 1986.

[11] F. Y. Chan, L. K. Chan, and M. H. Yu. A generalization of Whittaker-Henderson graduation. *Transactions of Society of Actuaries*, 36:183–211, 1984.

[12] S. Chen, D. Donoho, and M. Saunders. Atomic decomposition by basis pursuit. *SIAM Review*, 2001.

[13] D. K. Dey, A. E. Gelfand, and F. Peng. Overdispersed generalized linear models. *Journal of Statistical Planning and Inference*, 64:93–108, 1997.

[14] D. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 2006.

[15] M. Ezekiel. A method of handling curvilinear correlation for any number of variables. *Journal of the American Statistical Association*, 19:431–453, 1924.

[16] L. Fahrmeir. Posterior mode estimation by extended Kalman filtering for multivariate dynamic generalised linear models. *Journal of the American Statistical Association*, 87:501–509, 1992.

[17] L. Fahrmeir and H. Kaufmann. On Kalman filtering, posterior mode estimation and Fisher scoring in dynamic exponential family regression. *Metrika*, 38:37–60, 1991.

[18] E. Frees, P. Shi, and E. Valdez. Actuarial applications of a hierarchical insurance claims model. *ASTIN Bulletin*, 39:165–197, 2009.

[19] A. Gelman and J. Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, 2006.

[20] M. Grant and S. Boyd. Graph implementations for nonsmooth convex programs. In *Recent Advances in Learning and Control*. 2008.

[21] M. Grant and S. Boyd. CVX: Matlab software for disciplined convex programming, version 1.21, September 2010.

[22] T. Hastie and R. Tibshirani. Generalized additive models. *Statistical Science*, 1:297–318, 1986.

[23] T. Hastie and R. Tibshirani. *Generalized aditive models*. Chapman & Hall/CRC, 1990.

[24] R. Hodrick and E. Prescott. Postwar U.S. business cycles: An empirical investigation. *Journal of Money, Credit, and Banking*, 29:1–16, 1997.

[25] M. D. Hoffman and A. Gelman. The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, In press.

[26] P. Huber. *Robust Statistics*. John Wiley & Sons, 1981.

[27] W. Jewell. The use of collateral data in credibility theory: a hierarchical model. *Giornale dell'Instituto Italiano degli Attuari*, 38:1–16, 1975.

[28] D. Jones and H. Gerber. Credibility formulas of the updating type. *Transactions of the Society of Actuaries*, 27:31–46, 1975.

[29] P. De Jong and B. Zehnwirth. Claims reserving state space models and the Kalman filter. *Journal of the Institute of Actuaries*, 110:157–181, 1983.

[30] P. De Jong and B. Zehnwirth. Credibility theory and the Kalman filter. *Insurance: Mathematics and Economics*, 2:281–286, 1983.

[31] B. Jørgensen. Exponential dispersion models (with discussion). *Journal of the Royal Statistics Society Series B*, 49:127–162, 1987.

[32] R. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82:35–45, 1960.

[33] S.J. Kim, K. Koh, S. Boyd, and D. Gorinevsky. ℓ_1-trend filtering. *SIAM Review*, 51:339–360, 2009.

[34] R. Koenker. *Quantile regression*. Cambridge University Press, 2005.

[35] R. Koenker, P. Ng, and S. Portnoy. Quantile smoothing splines. *Biometrika*, 81:673–80, 1994.

[36] A. Kudryavtsev. Using quantile regression for rate–making. *Insurance: Mathematics and Economics*, 45:296–304, 2009.

[37] S. Lee, H. Lee, P. Abeel, and A. Ng. Efficient ℓ_1-regularized logistic regression. In *AAAI*, 2006.

[38] W. B. Lowrie. Multidimensional Whittaker-Henderson graduation with constraints and mixed differences. *Transactions of Society of Actuaries*, 45:215–255, 1993.

[39] R. Mehra. Credibility theory and kalman filtering with extensions. Technical Report RM 75-64, International Institute for Applied Systems Analysis, Schloss Laxenburg, Austria, 1975.

[40] H. Miller and P. Mulquiney. Credibility, penalised regression and boosting; let's call the whole thing off. In *Pricing and Underwriting Seminar*. Casualty Actuarial Society, 2011.

[41] J. Nelder and R. Verrall. Credibility theory and generalized linear models. *ASTIN Bulletin*, pages 71–82, 1997.

[42] Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course (Applied Optimization)*. Springer, 1st edition, 2003.

[43] C.E. Rasmussen and C.K.I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2005.

[44] R.T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

[45] D. Schuette. A linear programming approach to graduation. *Transactions of Society of Actuaries*, 30:407–445, 1978.

[46] H. L. Seal. Graduation by piecewise cubic polynomials: A historical review. *Blätter der DGVFM*, 15:89–114, 1981.

[47] A. Shapiro, D. Dentcheva, and A. Ruszczyski. *Lectures on Stochastic Programming: Modeling and Theory*. SIAM, 2009.

[48] M. L. Stein. *Interpolation of spatial data: Some theory for Kriging*. Springer-Verlag, New York, 1999.

[49] C. A. Sutton and A. McCallum. An introduction to conditional random fields. *Foundations and Trends in Machine Learning*, 4:267–373, 2012.

[50] K. Talluri and G. van Ryzin. *The Theory and Practice of Revenue Management*. Springer, 2004.

[51] G. Taylor. A Bayesian interpretation of Whittaker-Henderson graduation. *Insurance: Mathematics and Economics*, 11:7–16, 1992.

[52] G. Taylor. Technical aspects of domestic lines pricing. Technical report, University of Melbourne, Centre for Actuarial Studies, 1997.

[53] G. Taylor. Geographic premium rating by Whittaker spacial smoothing. *ASTIN Bulletin*, 31:147–160, 2001.

[54] G. Taylor. Second-order Bayesian revision of a generalised linear model. *Scandinavian Actuarial Journal*, pages 202–242, 2008.

[55] H. Taylor, S. Banks, and J. McCoy. Deconvolution with the ℓ_1 norm. *Geophysics*, 1979.

[56] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, 1996.

[57] R. Tibshirani, M. Saunders, S. Rosset, J Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistics Society Series B*, 67:91–108, 2005.

[58] A. Tychonoff and V. Arsenin. *Solution of Ill-posed Problems*. Winston & Sons, 1977.

[59] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, 1995.

[60] R. Verall. A state space formulation of Whittaker graduation with existensions. *Insurance: Mathematics and Economics*, 13:1–14, 1993.

[61] R. V. Vohra. *Advanced Mathematical Economics*. Routledge, London and New York, 2005.

[62] M. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1:1–305, 2008.

[63] M. Welling and Y. W. Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *ICML*, 2011.

[64] E. Whittaker. On a new method of graduation. *Proceedings of the Edinburgh Mathematical Society*, 41:63–75, 1923.

[65] M. Wright. The interior-point revolution in optimization: history, recent developments, and lasting consequences. *Bulletin of the American Mathematical Society*, 42:39–56, 2005.

[66] J. Yan, J. Guszcza, M. Flynn, and C.-S. P. Wu. Applications of the offset in property-casualty predictive modeling. In *Forum*. Casualty Actuarial Society, 2009.

[67] M. Yuan and L. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society*, 2006.

[68] H. Zou and T. Hastie. Regularization and variable selection via the elastic net,. *Journal of the Royal Statistical Society*, 2005.