A Practitioner's Guide to Generalized Linear Models

A CAS Study Note

Duncan Anderson, FIA
Sholom Feldblum, FCAS
Claudine Modlin, FCAS
Doris Schirmacher, FCAS
Ernesto Schirmacher, ASA
Neeza Thandi, FCAS

Third Edition

February 2007

The *Practitioner's Guide to Generalized Linear Models* is written for the practicing actuary who would like to understand generalized linear models (GLMs) and use them to analyze insurance data. The guide is divided into three sections.

Section 1 provides a foundation for the statistical theory and gives illustrative examples and intuitive explanations which clarify the theory. The intuitive explanations build upon more commonly understood actuarial methods such as linear models and the minimum bias procedures.

Section 2 provides practical insights and realistic model output for each stage of a GLM analysis - including data preparation and preliminary analyses, model selection and iteration, model refinement and model interpretation. This section is designed to boost the actuary's confidence in interpreting GLMs and applying them to solve business problems.

Section 3 discusses other topics of interest relating to GLMs such as retention modeling and scoring algorithms.

More technical material in the paper is set out in appendices.

Acknowledgements

The authors would like to thank James Tanser, FIA, for some helpful comments and contributions to some elements of this paper, Shaun Wang, FCAS, for reviewing the paper prior to inclusion on the CAS exam syllabus, and Volker Wilmsen for some helpful comments on the Second Edition of this paper.

Contents

1 GLMs - theory and intuition

1.1 Section 1 discusses how GLMs are formularized and solved. The following topics are covered in detail:

- background of GLMs - building upon traditional actuarial methods such as minimum bias procedures and linear models
- introduction to the statistical framework of GLMs
- formularization of GLMs - including the linear predictor, the link function, the offset term, the error term, the scale parameter and the prior weights
- typical model forms
- solving GLMs - maximum likelihood estimation and numerical techniques
- aliasing
- model diagnostics - standard errors and deviance tests.

Background

1.2 Traditional ratemaking methods in the United States are not statistically sophisticated. Claims experience for many lines of business is often analyzed using simple one-way and two-way analyses. Iterative methods known as minimum bias procedures, developed by actuaries in the 1960s, provide a significant improvement, but are still only part way toward a full statistical framework.

1.3 The classical linear model and many of the most common minimum bias procedures are, in fact, special cases of generalized linear models (GLMs). The statistical framework of GLMs allows explicit assumptions to be made about the nature of the insurance data and its relationship with predictive variables. The method of solving GLMs is more technically efficient than iteratively standardized methods, which is not only elegant in theory but valuable in practice. In addition, GLMs provide statistical diagnostics which aid in selecting only significant variables and in validating model assumptions.

1.4 Today GLMs are widely recognized as the industry standard method for pricing private passenger auto and other personal lines and small commercial lines insurance in the European Union and many other markets. Most British, Irish and French auto insurers use GLMs to analyze their portfolios and to the authors' knowledge GLMs are commonly used in Italy, the Netherlands, Scandinavia, Spain, Portugal, Belgium, Switzerland, South Africa, Israel and Australia. The method is gaining popularity in Canada, Japan, Korea, Brazil, Singapore, Malaysia and eastern European countries.

1.5 The primary applications of GLMs in insurance analysis are ratemaking and underwriting. Circumstances that limit the ability to change rates at will (eg regulation) have increased the use of GLMs for target marketing analysis.

The failings of one-way analysis

1.6 In the past, actuaries have relied heavily on one-way analyses for pricing and monitoring performance.

1.7 A one-way analysis summarizes insurance statistics, such as frequency or loss ratio, for each value of each explanatory variable, but without taking account of the effect of other variables. Explanatory variables can be discrete or continuous. Discrete variables are generally referred to as "factors", with values that each factor can take being referred to as "levels", and continuous variables are generally referred to as "variates". The use of variates is generally less common in insurance modeling.

1.8 One-way analyses can be distorted by correlations between rating factors. For example, young drivers may in general drive older cars. A one-way analysis of age of car may show high claims experience for older cars, however this may result mainly from the fact that such older cars are in general driven more by high risk younger drivers. Relativities based on one-way analyses of age of vehicle and age of driver would double-count the effect of age of driver. Traditional actuarial techniques for addressing this problem usually attempt to standardize the data in such a way as to remove the distorting effect of uneven business mix, for example by focusing on loss ratios on a one-way basis, or by standardizing for the effect of one or more factors. These methods are, however, only approximations.

1.9 One-way analyses also do not consider interdependencies between factors in the way they affect claims experience. These interdependencies, or interactions, exist when the effect of one factor varies depending on the levels of another factor. For example, the pure premium differential between men and women may differ by levels of age.

1.10 Multivariate methods, such as generalized linear models, adjust for correlations and allow investigation into interaction effects.

The failings of minimum bias procedures

1.11 In the 1960s, actuaries developed a ratemaking technique known as minimum bias procedures.[1] These procedures impose a set of equations relating the observed data, the rating variables, and a set of parameters to be determined. An iterative procedure solves the system of equations by attempting to converge to the optimal solution. The reader seeking more information may reference "The Minimum Bias Procedure: A Practitioner's Guide" by Sholom Feldblum and Dr J. Eric Brosius.[2]

---

[1] Bailey, Robert A. and LeRoy J. Simon, "Two Studies in Automobile Insurance Ratemaking," Proceedings of the Casualty Actuarial Society, XLVII, 1960.

[2] Feldblum, Sholom and Brosius, J Eric, "The Minimum Bias Procedures: A Practitioner's Guide", Casualty Actuarial Society Forum, 2002 Vol: Fall Page(s): 591-684

1.12 Once an optimal solution is calculated, however, the minimum bias procedures give no systematic way of testing whether a particular variable influences the result with statistical significance. There is also no credible range provided for the parameter estimates. The minimum bias procedures lack a statistical framework which would allow actuaries to assess better the quality of their modeling work.

The connection of minimum bias to GLM

1.13 Stephen Mildenhall has written a comprehensive paper showing that many minimum bias procedures do correspond to generalized linear models.[3] The following table summarizes the correspondence for many of the more common minimum bias procedures. The GLM terminology *link function* and *error function* is explained in depth later in this section. In brief, these functions are key components for specifying a generalized linear model.

Minimum Bias Procedures	Generalized Linear Models	
	Link function	**Error function**
Multiplicative balance principle	Logarithmic	Poisson
Additive balance principle	Identity	Normal
Multiplicative least squares	Logarithmic	Normal
Multiplicative maximum likelihood with exponential density function	Logarithmic	Gamma
Multiplicative maximum likelihood with Normal density function	Logarithmic	Normal
Additive maximum likelihood with Normal density function	Identity	Normal

1.14 Not all minimum bias procedures have a generalized linear model analog and vice versa. For example, the χ^2 additive and multiplicative minimum bias models have no corresponding generalized linear model analog.

Linear models

1.15 A GLM is a generalized form of a linear model. To understand the structure of generalized linear models it is helpful, therefore, to review classic linear models.

1.16 The purpose of both linear models (LMs) and generalized linear models is to express the relationship between an observed response variable, Y, and a number of covariates (also called predictor variables), X. Both models view the observations, Y_i, as being realizations of the random variable Y.

[3] Mildenhall, Stephen, "A Systematic Relationship between Minimum Bias and Generalized Linear Models", Proceedings of the Casualty Actuarial Society, LXXXVI, 1999.

1.17 Linear models conceptualize Y as the sum of its mean, μ, and a random variable, ε:

$$Y = \mu + \varepsilon$$

1.18 They assume that

a. the expected value of Y, μ, can be written as a linear combination of the covariates, X, and

b. the error term, ε, is Normally distributed with mean zero and variance σ^2.

1.19 For example, suppose a simple private passenger auto classification system has two categorical rating variables: territory (urban or rural) and gender (male or female). Suppose the observed average claim severities are:

	Urban	Rural
Male	800	500
Female	400	200

1.20 The response variable, Y, is the average claim severity. The two factors, territory and gender, each have two levels resulting in the four covariates: male (X_1), female (X_2), urban (X_3), and rural (X_4). These indicator variables take the value 1 or 0. For example, the urban covariate, (X_3), is equal to 1 if the territory is urban, and 0 otherwise.

1.21 The linear model seeks to express the observed item Y (in this case average claim severity) as a linear combination of a specified selection of the four variables, plus a Normal random variable ε with mean zero and variance σ^2, often written $\varepsilon \sim N(0,\sigma^2)$. One such model might be

$$Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$$

1.22 However this model has as many parameters as it does combinations of rating factor levels being considered, and there is a linear dependency between the four covariates X_1, X_2, X_3, X_4. This means that the model in the above form is not uniquely defined - if any arbitrary value k is added to both β_1 and β_2, and the same value k is subtracted from β_3 and β_4, the resulting model is equivalent.

1.23 To make the model uniquely defined in the parameters β_i consider instead the model

$$Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

1.24 This model is equivalent to assuming that there is an average response for men (β_1) and an average response for women (β_2), with the effect of being an urban policyholder (as opposed to being a rural one) having an additional additive effect (β_3) which is the same regardless of gender.

1.25 Alternatively this could be thought of as a model which assumes an average response for the "base case" of women in rural areas (β_2) with additional additive effects for being male (β_2-β_1) and for being in an urban area (β_3).

1.26 Thus the four observations can be expressed as the system of equations:

$$\begin{aligned} Y_1 &= 800 = \beta_1 + 0 + \beta_3 + \varepsilon_1 \\ Y_2 &= 500 = \beta_1 + 0 + 0 \ + \varepsilon_2 \\ Y_3 &= 400 = 0 + \beta_2 + \beta_3 + \varepsilon_3 \\ Y_4 &= 200 = 0 + \beta_2 + 0 + \varepsilon_4 \end{aligned}$$

1.27 The parameters $\beta_1, \beta_2, \beta_3$ which best explain the observed data are then selected. For the classical linear model this is done by minimizing the sum of squared errors (SSE):

$$\begin{aligned} SSE &= \varepsilon_1^2 + \varepsilon_2^2 + \varepsilon_3^2 + \varepsilon_4^2 \\ &= (800 - \beta_1 - \beta_3)^2 + (500 - \beta_1)^2 + (400 - \beta_2 - \beta_3)^2 + (200 - \beta_2)^2 \end{aligned}$$

1.28 This expression can be minimized by taking derivatives with respect to β_1, β_2 and β_3 and setting each of them to zero. The resulting system of three equations in three unknowns is:

$$\begin{aligned} \frac{\partial SSE}{\partial \beta_1} = 0 \quad &\Rightarrow \quad \beta_1 + \beta_3 + \beta_1 = 800 + 500 = 1300 \\ \frac{\partial SSE}{\partial \beta_2} = 0 \quad &\Rightarrow \quad \beta_2 + \beta_3 + \beta_2 = 400 + 200 = 600 \\ \frac{\partial SSE}{\partial \beta_3} = 0 \quad &\Rightarrow \quad \beta_1 + \beta_3 + \beta_2 + \beta_3 = 800 + 400 = 1200 \end{aligned}$$

which can be solved to derive:

$$\beta_1 = 525$$
$$\beta_2 = 175$$
$$\beta_3 = 250$$

Vector and Matrix Notation

1.29 Formulating the system of equations above quickly becomes complex as both the number of observations and the number of covariates increases; consequently, vector notation is used to express these equations in compact form.

1.30 Let $\underline{Y}$ be a column vector with components corresponding to the observed values for the response variable:

$$\underline{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \end{bmatrix} = \begin{bmatrix} 800 \\ 500 \\ 400 \\ 200 \end{bmatrix}$$

1.31 Let $\underline{X_1}$, $\underline{X_2}$, and $\underline{X_3}$ denote the column vectors with components equal to the observed values for the respective indicator variables (eg the i^{th} element of $\underline{X_1}$ is 1 when the i^{th} observation is male, and 0 if female):

$$\underline{X}_1 = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix} \qquad \underline{X}_2 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \end{bmatrix} \qquad \underline{X}_3 = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \end{bmatrix}$$

1.32 Let $\underline{\beta}$ denote a column vector of parameters, and for a given set of parameters let $\underline{\varepsilon}$ be the vector of residuals:

$$\underline{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} \qquad \underline{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \end{bmatrix}$$

1.33 Then the system of equations takes the form:

$$\underline{Y} = \beta_1 \underline{X}_1 + \beta_2 \underline{X}_2 + \beta_3 \underline{X}_3 + \underline{\varepsilon}$$

1.34 To simplify this further the vectors $\underline{X_1}$, $\underline{X_2}$, and $\underline{X_3}$ can be aggregated into a single matrix $\boldsymbol{X}$. This matrix is called the design matrix and in the example above would be defined as:

$$X = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

1.35 Appendix A shows an example of the form of the design matrix $\boldsymbol{X}$ when explanatory variables include continuous variables, or "variates".

1.36 The system of equations takes the form

$$\underline{Y} = \boldsymbol{X}.\underline{\beta} + \underline{\varepsilon}$$

1.37 In the case of the linear model, the goal is to find values of the components of $\underline{\beta}$ which minimize the sum of squares of the components of $\underline{\varepsilon}$. If there are n observations and p parameters in the model, $\underline{\varepsilon}$ will have n components and $\underline{\beta}$ will have p components ($p<n$).

1.38 The basic ingredients for a linear model thus consist of two elements:

a. a set of assumptions about the relationship between $\underline{Y}$ and the predictor variables, and

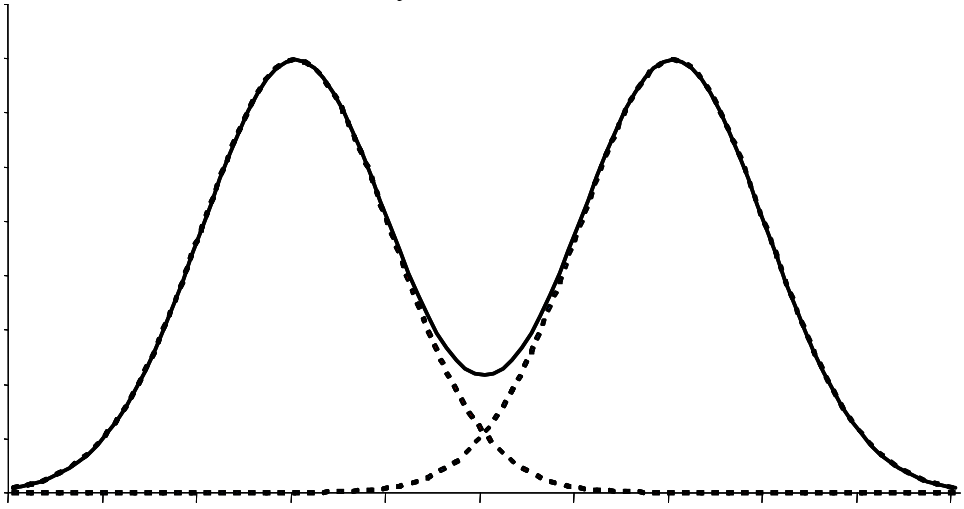
b. an objective function which is to be optimized in order to solve the problem. Standard statistical theory defines the objective function to be the likelihood function. In the case of the classical linear model with an assumed Normal error it can be shown that the parameters which minimize sum of squared error also maximize likelihood.

Classical linear model assumptions

1.39 Linear models assume all observations are independent and each comes from a Normal distribution.

1.40 This assumption does not relate to the aggregate of the observed item, but to each observation individually. An example may help illustrate this distinction.

Distribution of individual observations



1.41 An examination of average claim amounts by gender may identify that average claim amounts for men are Normally distributed, as are average claim amounts for women, and that the mean of the distribution for men is twice the mean of the distribution for women. The total distribution of average claim amounts across all men and women is not Normally distributed. The only distribution of interest is the distribution of the two separate classes. (In this case there are only two classes being considered, but in a more complicated model there would be one such class for each combination of the rating factors being considered.)

1.42 Linear models assume that the mean is a linear combination of the covariates, and that each component of the random variable is assumed to have a common variance.

1.43 The linear model can be written as follows:

$$\underline{Y} = E[\underline{Y}] + \underline{\varepsilon}, \qquad E[\underline{Y}] = \mathbf{X}.\underline{\beta}$$

1.44 McCullagh and Nelder outline the explicit assumptions as follows:[4]

(LM1) *Random component:* Each component of $\underline{Y}$ is independent and is Normally distributed. The mean, μ_i, of each component is allowed to differ, but they all have common variance σ^2

(LM2) *Systematic component:* The p covariates are combined to give the "linear predictor" $\underline{\eta}$:

$$\underline{\eta} = \mathbf{X}.\underline{\beta}$$

(LM3) *Link function:* The relationship between the random and systematic components is specified via a link function. In the linear model the link function is equal to the identity function so that:

$$E[\underline{Y}] \equiv \underline{\mu} = \underline{\eta}$$

1.45 The identity link function assumption in **(LM3)** may appear to be superfluous at this point, but it will become more meaningful when discussing the generalization to GLMs.

Limitations of Linear Models

1.46 Linear models pose quite tractable problems that can be easily solved with well-known linear algebra approaches. However it is easy to see that the required assumptions are not easy to guarantee in applications:

- It is difficult to assert Normality and constant variance for response variables. Classical linear regression attempts to transform data so that these conditions hold. For example, *Y* may not satisfy the hypotheses but *ln(Y)* may. However there is no reason why such a transformation should exist.
- The values for the response variable may be restricted to be positive. The assumption of Normality violates this restriction.
- If the response variable is strictly non-negative then intuitively the variance of *Y* tends to zero as the mean of *Y* tends to zero. That is, the variance is a function of the mean.

[4] McCullagh, P. and J. A. Nelder, *Generalized Linear Models*, 2nd Ed., Chapman & Hall/CRC, 1989.

- The additivity of effects encapsulated in the second **(LM2)** and third **(LM3)** assumptions is not realistic for a variety of applications. For example, suppose the response variable is equal to the area of the wings of a butterfly and the predictor variables are the width and length of the wings. Clearly, these two predictor variables do not enter additively; rather, they enter multiplicatively. More relevantly, many insurance risks tend to vary multiplicatively with rating factors (this is discussed in more detail in Section 2).

Generalized linear model assumptions

1.47 GLMs consist of a wide range of models that include linear models as a special case. The LM restriction assumptions of Normality, constant variance and additivity of effects are removed. Instead, the response variable is assumed to be a member of the exponential family of distributions[5]. Also, the variance is permitted to vary with the mean of the distribution. Finally, the effect of the covariates on the response variable is assumed to be additive on a transformed scale. Thus the analog to the linear model assumptions **(LM1), (LM2),** and **(LM3)** are as follows.

(GLM1) *Random component:* Each component of $\underline{Y}$ is independent and is from one of the exponential family of distributions.

(GLM2) *Systematic component:* The p covariates are combined to give the linear predictor $\underline{\eta}$:

$$\underline{\eta} = \mathbf{X}.\underline{\beta}$$

(GLM3) *Link function:* The relationship between the random and systematic components is specified via a link function, g, that is differentiable and monotonic such that:

$$E[\underline{Y}] \equiv \underline{\mu} = g^{-1}(\underline{\eta})$$

1.48 Most statistical texts denote the first expression in **(GLM3)** with $g(x)$ written on the left side of the equation; therefore, the systematic element is generally expressed on the right side as the inverse function, g^{-1}.

[5] The exponential family is a broader class of distributions sharing the same density form and including Normal, Poisson, gamma, inverse Gaussian, binomial, exponential and other distributions.

Exponential Family of Distributions

1.49 Formally, the exponential family of distributions is a 2-parameter family defined as:

$$f_i(y_i;\theta_i,\phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i,\phi)\right\}$$

where $a_i(\phi)$, $b(\theta_i)$, and $c(y_i,\phi)$ are functions specified in advance; θ_i is a parameter related to the mean; and ϕ is a scale parameter related to the variance. This formal definition is further explored in Appendix B. For practical purposes it is useful to know that a member of the exponential family has the following two properties:

a. the distribution is completely specified in terms of its mean and variance,

b. the variance of Y_i is a function of its mean.

1.50 This second property is emphasized by expressing the variance as:

$$Var(Y_i) = \frac{\phi\, V(\mu_i)}{\omega_i}$$

where V(x), called the variance function, is a specified function; the parameter ϕ scales the variance; and ω_i is a constant that assigns a weight, or credibility, to observation i.

1.51 A number of familiar distributions belong to the exponential family: the Normal, Poisson, binomial, gamma, and inverse Gaussian.[6] The corresponding value of the variance function is summarized in the table below:

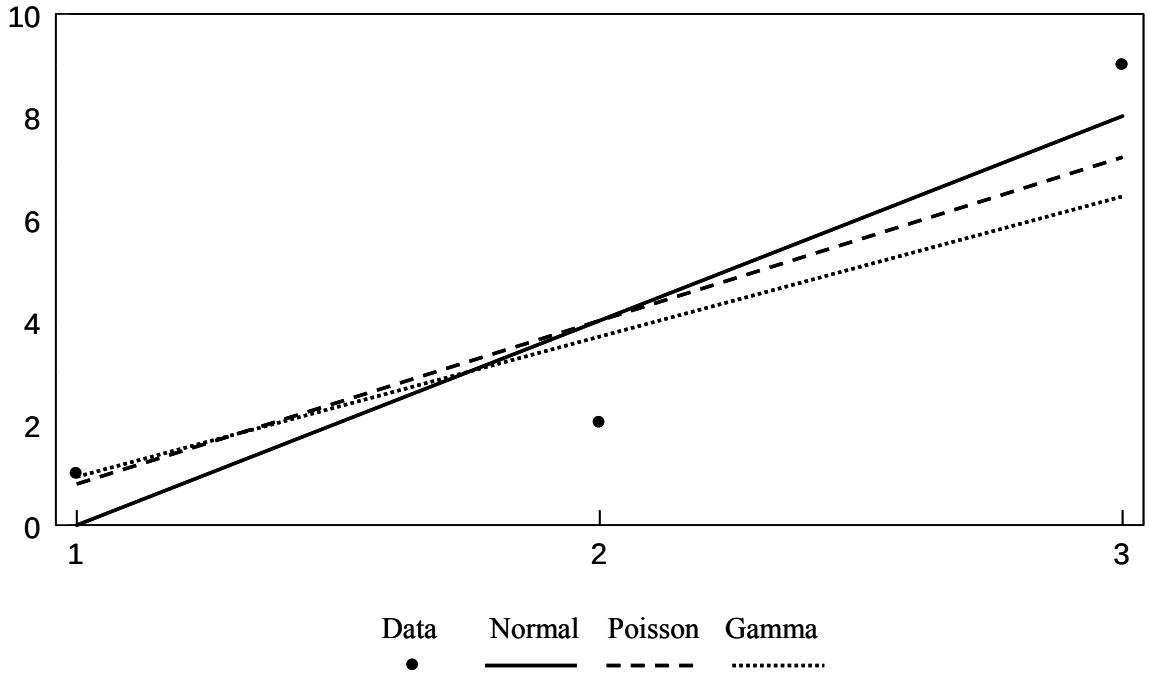
	$V(x)$	
Normal	1	
Poisson	x	
Gamma	x^2	
Binomial	$x(1-x)$	*(where the number of trials = 1)*
Inverse Gaussian	x^3	

1.52 A special member of the exponential family is the Tweedie distribution. The Tweedie distribution has a point mass at zero and a variance function proportional to μ^p (where $p<0$ or $1<p<2$ or $p>2$). This distribution is typically used to model pure premium data directly and is discussed further in Appendix C.

[6] A notable exception to this list is the lognormal distribution, which does not belong to the exponential family.

1.53 The choice of the variance function affects the results of the GLM. For example, the graph below considers the result of fitting three different (and very simple) GLMs to three data points. In each case the model form selected is a two-parameter model (the intercept and slope of a line), and the three points represent the individual observations (with the observed value Y_i shown on the y-axis for different values of a single continuous explanatory variable shown on the x-axis).

Effect of varying the error term (simple example)



1.54 The three GLMs considered have a Normal, Poisson and gamma variance function respectively. It can be seen that the GLM with a Normal variance function (which assumes that each observation has the same fixed variance) has produced fitted values which are attracted to the original data points with equal weight. By contrast the GLM with a Poisson error assumes that the variance increases with the expected value of each observation. Observations with smaller expected values have a smaller assumed variance, which results in greater credibility when estimating the parameters. The model thus has produced fitted values which are more influenced by the observation on the left (with smaller expected value) than the observation on the right (which has a higher expected value and hence a higher assumed variance).

1.55 It can be seen that the GLM with assumed gamma variance function is even more strongly influenced by the point on the left than the point on the right since that model assumes the variance increases with the square of the expected value.

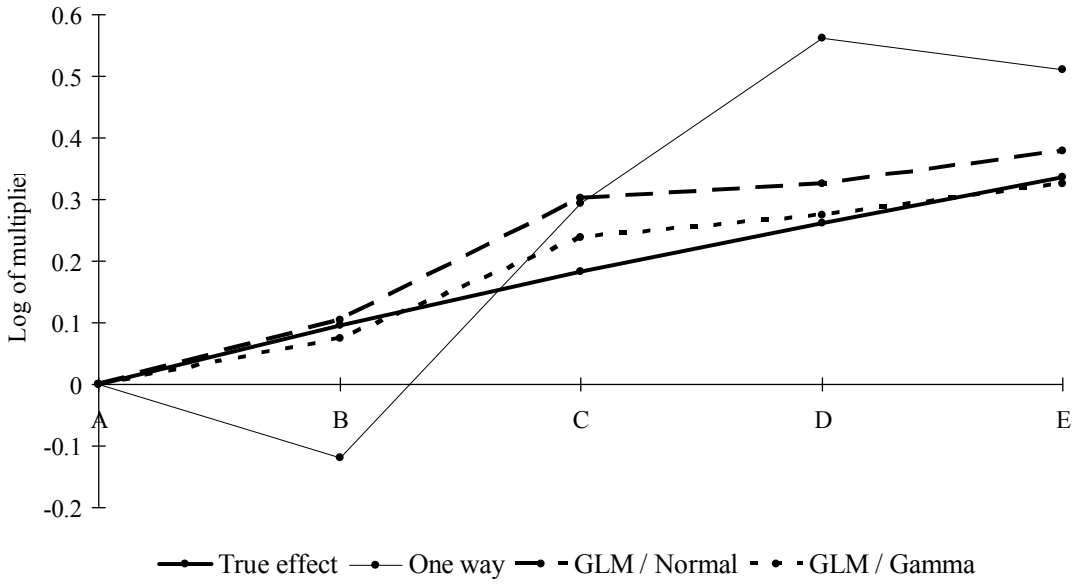
1.56 A further, rather more realistic, example illustrates how selecting an appropriate variance function can improve the accuracy of a model. This example considers an artificially generated dataset which represents an insurance portfolio. This dataset contains several rating factors (some of which are correlated), and in each case the true effect of the rating factor is assumed to be known. Claims experience (in this case average claim size experience) is then randomly generated for each policy using a gamma distribution, with the mean in each case being that implied by the assumed effect of the rating factors. The claims experience is then analyzed using three models to see how closely the results of each model relate to the (in this case known) true factor effect.

1.57 The three methods considered are

- a one-way analysis
- a GLM with assumed Normal variance function
- a GLM with assumed gamma variance function.

1.58 The results for one of the several rating factors considered are shown on the graph below. It can be seen that owing to the correlations between the rating factors in the data, the one-way analysis is badly distorted. The GLM with an assumed Normal distribution is closer to the correct relativities, but it can be seen that it is the GLM with an assumed gamma variance function which yields results that are the closest to the true effect.

Effect of varying the error term (insurance rating factor example)



1.59 In addition to the variance function *V(x)*, two other parameters define the variance of each observation, the scale parameter ϕ and the prior weights ω_i

$$Var[Y_i] = \frac{\phi V(\mu_i)}{\omega_i}$$

Prior weights

1.60 The prior weights allow information about the known credibility of each observation to be incorporated in the model. For example, if modeling claims frequency, one observation might relate to one month's exposure, and another to one year's exposure. There is more information and less variability in the observation relating to the longer exposure period, and this can be incorporated in the model by defining ω_i to be the exposure of each observation. In this way observations with higher exposure are deemed to have lower variance, and the model will consequently be more influenced by these observations.

1.61 An example demonstrates the appropriateness of this more clearly. Consider a set of observations for personal auto claims under some classification system. Let cell i denote some generic cell defined by this classification system. To analyze frequency let:

m_{ik} be the number of claims arising from the k^{th} unit of exposure in cell i

ω_i be the number of exposures in cell i

Y_i be the observed claim frequency in cell i:

$$Y_i = \frac{1}{\omega_i} \sum_{k=1}^{\omega_i} m_{ik}$$

1.62 If the random process generating m_{ik} is Poisson with frequency f_i for all exposures k then

$$E[m_{ik}] = f_i = Var[m_{ik}]$$

1.63 Assuming the exposures are independent then

$$\mu_i = E[Y_i] = \frac{1}{\omega_i}\sum_{k=1}^{\omega_i} E[m_{ik}] = \frac{1}{\omega_i}\cdot \omega_i f_i = f_i$$

$$Var[Y_i] = \frac{1}{\omega_i^2}\sum_{k=1}^{\omega_i} Var[m_{ik}] = \frac{1}{\omega_i^2}\cdot \omega_i f_i = \frac{1}{\omega_i} f_i = \mu_i \cdot \frac{1}{\omega_i}$$

1.64 So in this case $V(\mu_i) = \mu_i$, $\phi = 1$, and the prior weights are the exposures in cell i.

1.65 An alternative example would be to consider claims severity. Let

z_{ik} be the claim size of the kth claim in cell i

ω_i be the number of claims in cell i

Y_i be the observed mean claim size in cell i:

$$Y_i = \frac{1}{\omega_i}\sum_{k=1}^{\omega_i} z_{ik}$$

1.66 This time assume that the random process generating each individual claim is gamma distributed. Denoting

$$E[z_{ik}] = m_i$$

and

$$Var[z_{ik}] = \sigma^2 m_i^2$$

and assuming each claim is independent then

$$\mu_i = E[Y_i] = \frac{1}{\omega_i}\sum_{k=1}^{\omega_i} E[z_{ik}] = \frac{1}{\omega_i}\cdot \omega_i m_i = m_i$$

$$Var[Y_i] = \frac{1}{\omega_i^2}\sum_{k=1}^{\omega_i} Var[z_{ik}] = \frac{1}{\omega_i^2}\cdot \omega_i \sigma^2 m_i^2 = \frac{1}{\omega_i}\sigma^2 m_i^2 = \mu_i^2 \cdot \frac{\sigma^2}{\omega_i}$$

1.67 So for severity with a gamma distribution the variance of Y_i follows the general form for all exponential distributions with $V(\mu_i) = \mu_i^2$, $\phi = \sigma^2$, and prior weight equal to the number of claims in cell i.

1.68 Prior weights can also be used to attach a lower credibility to a part of the data which is known to be less reliable.

The scale parameter

1.69 In some cases (eg the Poisson distribution) the scale parameter ϕ is identically equal to 1 and falls out of the GLM analysis entirely. However in general and for the other familiar exponential distributions ϕ is not known in advance, and in these cases it must be estimated from the data.

1.70 Estimation of the scale parameter is not actually necessary in order to solve for the GLM parameters $\underline{\beta}$, however in order to determine certain statistics (such as standard errors, discussed below) it is necessary to estimate ϕ.

1.71 ϕ can be treated as another parameter and estimated by maximum likelihood. The drawback of this approach is that it is not possible to derive an explicit formula for ϕ, and the maximum likelihood estimation process can take considerably longer.

1.72 An alternative is to use an estimate of ϕ, such as

a. the moment estimator (Pearson χ^2 statistic) defined as

$$\hat{\phi} = \frac{1}{n-p} \sum_i \frac{\omega_i (Y_i - \mu_i)^2}{V(\mu_i)}$$

b. the total deviance estimator

$$\hat{\phi} = \frac{D}{n-p}$$

where D, the total deviance, is defined later in this paper.

Link Functions

1.73 In practice when using classical linear regression practitioners sometimes attempt to transform data to satisfy the requirements of Normality and constant variance of the response variable and additivity of effects. Generalized linear models, on the other hand, merely require that there be a link function that guarantees the last condition of additivity. Whereas **(LM3)** requires that Y be additive in the covariates, the generalization **(GLM3)** instead requires that some transformation of Y, written as $g(Y)$, be additive in the covariates.

1.74 It is more helpful to consider μ_i as a function of the linear predictor, so typically it is the inverse of *g(x)* which is considered:

$$\mu_i = g^{-1}(\eta_i)$$

1.75 In theory a different link function could be used for each observation i, but in practice this is rarely done.

1.76 The link function must satisfy the condition that it be differentiable and monotonic (either strictly increasing or strictly decreasing). Some typical choices for a link function include)

	$g(x)$	$g^{-1}(x)$
Identity	x	x
Log	$\ln(x)$	e^x
Logit	$\ln(x/(1-x))$	$e^x/(1+e^x)$
Reciprocal	$1/x$	$1/x$

1.77 Each error structure has associated with it a "canonical" link function which simplifies the mathematics of solving GLMs analytically. These are discussed in Appendix D. When solving GLMs using modern computer software, however, the use of canonical link functions is not important and any pairing of link function and variance function which is deemed appropriate may be selected.

1.78 The log-link function has the appealing property that the effect of the covariates are multiplicative. Indeed, writing $g(x) = ln(x)$ so that $g^{-1}(x) = e^x$ results in

$$\mu_i = g^{-1}(\beta_1 x_{i1} + ... + \beta_p x_{ip}) = \exp(\beta_1 x_{i1}) \cdot \exp(\beta_2 x_{i2}) ... \exp(\beta_p x_{ip})$$

1.79 In other words, when a log link function is used, rather than estimating additive effects, the GLM estimates logs of multiplicative effects.

1.80 As mentioned previously, alternative choices of link functions and error structures can yield GLMs which are equivalent to a number of the minimum bias models as well as a simple linear model (see section "The Connection of Minimum Bias to GLM").

The offset term

1.81 There are occasions when the effect of an explanatory variable is known, and rather than estimating parameters $\underline{\beta}$ in respect of this variable it is appropriate to include information about this variable in the model as a known effect. This can be achieved by introducing an "offset term" $\underline{\xi}$ into the definition of the linear predictor $\underline{\eta}$:

$$\underline{\eta} = \boldsymbol{X}.\underline{\beta} + \underline{\xi}$$

which gives

$$E[\underline{Y}] = \underline{\mu} = g^{-1}(\underline{\eta}) = g^{-1}(\boldsymbol{X}.\underline{\beta} + \underline{\xi})$$

1.82 A common example of the use of an offset term is when fitting a multiplicative GLM to the observed number, or count, of claims (as opposed to claim frequency). Each observation may relate to a different period of policy exposure. An observation relating to one month's exposure will obviously have a lower expected number of claims (all other factors being equal) than an observation relating to a year's exposure. To make appropriate allowance for this, the assumption that the expected count of claims increases in proportion to the exposure of an observation (all other factors being equal) can be introduced in a multiplicative GLM by setting the offset term $\underline{\xi}$ to be equal to the log of the exposure of each observation, giving:

$$E[Y_i] = g^{-1}\left(\sum_j X_{ij}\beta_j + \xi_i\right) = \exp\left(\sum_j X_{ij}\beta_j + \log(e_i)\right) = \exp\left(\sum_j X_{ij}\beta_j\right).e_i$$

where e_i = the exposure for observation i.

1.83 In the particular case of a Poisson multiplicative GLM it can be shown that modeling claim counts with an offset term equal to the log of the exposure (and prior weights set to 1) produces identical results to modeling claim frequencies with no offset term but with prior weights set to be equal to the exposure of each observation.

Structure of a generalized linear model

1.84 In summary, the assumed structure of a GLM can be specified as:

$$\mu_i = E[Y_i] = g^{-1}(\sum_j X_{ij}\beta_j + \xi_i)$$

$$Var[Y_i] = \frac{\phi V(\mu_i)}{\omega_i}$$

where

Y_i is the vector of responses

$g(x)$ is the link function: a specified (invertible) function which relates the expected response to the linear combination of observed factors

X_{ij} is a matrix (the "design matrix") produced from the factors

β_j is a vector of model parameters, which is to be estimated

ξ_i is a vector of known effects or "offsets"

ϕ is a parameter to scale the function $V(x)$

$V(x)$ is the variance function

ω_i is the prior weight that assigns a credibility or weight to each observation

1.85 The vector of responses Y_i, the design matrix X_{ij}, the prior weights ω_i, and the offset term ξ_i are based on data in a manner determined by the practitioner. The assumptions which then further define the form of the model are the link function $g(x)$, the variance function $V(x)$, and whether ϕ is known or to be estimated.

Typical GLM model forms

1.86 The typical model form for modeling insurance claim counts or frequencies is a multiplicative Poisson. As well as being a commonly assumed distribution for claim numbers, the Poisson distribution also has a particular feature which makes it intuitively appropriate in that it is invariant to measures of time. In other words, measuring frequencies per month and measuring frequencies per year will yield the same results using a Poisson multiplicative GLM. This is not true of some other distributions such as gamma.

1.87 In the case of claim frequencies the prior weights are typically set to be the exposure of each record. In the case of claim counts the offset term is set to be the log of the exposure.

1.88 A common model form for modeling insurance severities is a multiplicative gamma. As well as often being appropriate because of its general form, the gamma distribution also has an intuitively attractive property for modeling claim amounts since it is invariant to measures of currency. In other words measuring severities in dollars and measuring severities in cents will yield the same results using a gamma multiplicative GLM. This is not true of some other distributions such as Poisson.

1.89 The typical model form for modeling retention and new business conversion is a logit link function and binomial error term (together referred to as a logistic model). The logit link function maps outcomes from the range of (0,1) to $(-\infty, +\infty)$ and is consequently invariant to measuring successes or failures. If the y-variate being modeled is generally close to zero, and if the results of a model are going to be used qualitatively rather than quantitatively, it may also be possible to use a multiplicative Poisson model form as an approximation given that the model output from a multiplicative GLM can be rather easier to explain to a non-technical audience.

1.90 The below table summarizes some typical model forms.

$\underline{Y}$	Claim frequencies	Claim numbers or counts	Average claim amounts	Probability (eg of renewing)
Link function $g(x)$	ln(x)	ln(x)	ln(x)	ln(x/(1-x))
Error	Poisson	Poisson	Gamma	Binomial
Scale parameter ϕ	1	1	Estimated	1
Variance function $V(x)$	x	x	x^2	x(1-x)*
Prior weights $\underline{\omega}$	Exposure	1	# of claims	1
Offset $\underline{\xi}$	0	ln(exposure)	0	0

* where the number of trials=1, or x(t-x)/t where the number of trials = t

GLM maximum likelihood estimators

1.91 Having defined a model form in terms of $\boldsymbol{X}$, $g(x)$, $\underline{\xi}$, $V(x)$, ϕ, and $\underline{\omega}$, and given a set of observations $\underline{Y}$, the components of $\underline{\beta}$ are derived by maximizing the likelihood function (or equivalently, the logarithm of the likelihood function). In essence, this method seeks to find the parameters which, when applied to the assumed model form, produce the observed data with the highest probability.

1.92 The likelihood is defined to be the product of probabilities of observing each value of the y-variate. For continuous distributions such as the Normal and gamma distributions the probability density function is used in place of the probability. It is usual to consider the log of the likelihood since being a summation across observations rather than a product, this yields more manageable calculations (and any maximum of the likelihood is also a maximum of the log-likelihood). Maximum likelihood estimation in practice, therefore, seeks to find the values of the parameters that maximize this log-likelihood.

1.93 In simple examples the procedure for maximizing likelihood involves finding the solution to a system of equations with linear algebra. In practice, the large number of observations typically being considered means that this is rarely done. Instead numerical techniques (and in particular multi-dimensional Newton-Raphson algorithms) are used. Appendix E shows the system of equations for maximizing the likelihood function in the general case of an exponential distribution.

1.94 An explicitly solved illustrative example and a discussion of numerical techniques used with large datasets are set out below.

Solving simple examples

1.95 To understand the mechanics involved in solving a GLM, a concrete example is presented. Consider the same four observations discussed in a previous section for average claim severity:

	Urban	Rural
Male	800	500
Female	400	200

1.96 The general procedure for solving a GLM involves the following steps:

a. Specify the design matrix $\boldsymbol{X}$ and the vector of parameters $\underline{\beta}$

b. Choose the error structure and link function

c. Identify the log-likelihood function

d. Take the logarithm to convert the product of many terms into a sum

e. Maximize the logarithm of the likelihood function by taking partial derivatives with respect to each parameter, setting them to zero and solving the resulting system of equations

f. Compute the predicted values.

1.97 Recall that the vector of observations, the design matrix, and the vector of parameters are as follows:

$$\underline{Y} = \begin{bmatrix} \text{Male Urban} \\ \text{Male Rural} \\ \text{Female Urban} \\ \text{Female Rural} \end{bmatrix} = \begin{bmatrix} 800 \\ 500 \\ 400 \\ 200 \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 0 \end{bmatrix}, \quad \text{and} \quad \underline{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}$$

where the first column of ***X*** indicates if an observation is male or not, the second column indicates whether the observation is female, and the last column specifies if the observation is in an urban territory or not.

1.98 The following three alternative model structures are illustrated:

- Normal error structure with an identity link function
- Poisson error structure with a log link function
- Gamma error structure with an inverse link function.

1.99 These three model forms may not necessarily be appropriate models to use in practice - instead they illustrate the theory involved.

1.100 In each case the elements of $\underline{\omega}$ (the prior weights) will be assumed to be 1, and the offset term $\underline{\xi}$ assumed to be zero, and therefore these terms will, in this example, be ignored.

Normal error structure with an identity link function

1.101 The classical linear model case assumes a Normal error structure and an identity link function. The predicted values in the example take the form:

$$E[\underline{Y}] = g^{-1}(X.\underline{\beta}) = \begin{bmatrix} g^{-1}(\beta_1 + \beta_3) \\ g^{-1}(\beta_1) \\ g^{-1}(\beta_2 + \beta_3) \\ g^{-1}(\beta_2) \end{bmatrix} = \begin{bmatrix} \beta_1 + \beta_3 \\ \beta_1 \\ \beta_2 + \beta_3 \\ \beta_2 \end{bmatrix}$$

1.102 The Normal distribution with mean μ and variance σ^2 has the following density function:

$$f(y;\mu,\sigma^2) = \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2)\right\}$$

1.103 Its likelihood function is:

$$L(y;\mu,\sigma^2) = \prod_{i=1}^{n} \exp\{-\frac{(y_i-\mu_i)^2}{2\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2)\}$$

1.104 Maximizing the likelihood function is equivalent to maximizing the log-likelihood function:

$$l(y;\mu,\sigma^2) = \sum_{i=1}^{n} -\frac{(y_i-\mu_i)^2}{2\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2)$$

1.105 With the identity link function, $\mu_i = \Sigma_j X_{ij}\beta_j$ and the log-likelihood function becomes

$$l(y;\mu,\sigma^2) = \sum_{i=1}^{n} -\frac{(y_i - \sum_{j=1}^{p} X_{ij}.\beta_j)^2}{2\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2)$$

1.106 In this example, up to a constant term of $2.\ln(2\pi\sigma^2)$, the log-likelihood is

$$l^*(y;\mu,\sigma^2) = -\frac{(800-(\beta_1+\beta_3))^2}{2\sigma^2} - \frac{(500-\beta_1)^2}{2\sigma^2} - \frac{(400-(\beta_2+\beta_3))^2}{2\sigma^2} - \frac{(200-\beta_2)^2}{2\sigma^2}$$

1.107 To maximize l^* take derivatives with respect to β_1, β_2 and β_3 and set each of them to zero. The resulting system of three equations in three unknowns is:

$$\frac{\partial l^*}{\partial \beta_1} = 0 \quad \Rightarrow \quad \beta_1 + \beta_3 + \beta_1 = 800 + 500 = 1300$$

$$\frac{\partial l^*}{\partial \beta_2} = 0 \quad \Rightarrow \quad \beta_2 + \beta_3 + \beta_2 = 400 + 200 = 600$$

$$\frac{\partial l^*}{\partial \beta_3} = 0 \quad \Rightarrow \quad \beta_1 + \beta_3 + \beta_2 + \beta_3 = 800 + 400 = 1200$$

1.108 It can be seen that these equations are identical to those derived when minimizing the sum of squared error for a simple linear model. Again, these can be solved to derive:

$$\beta_1 = 525$$
$$\beta_2 = 175$$
$$\beta_3 = 250$$

which produces the following predicted values:

	Urban	Rural
Male	775	525
Female	425	175

The Poisson error structure with a logarithm link function

1.109 For the Poisson model with a logarithm link function, the predicted values are given by

$$E[\underline{Y}] = g^{-1}(X\beta) = \begin{bmatrix} g^{-1}(\beta_1 + \beta_3) \\ g^{-1}(\beta_1) \\ g^{-1}(\beta_2 + \beta_3) \\ g^{-1}(\beta_2) \end{bmatrix} = \begin{bmatrix} e^{\beta_1+\beta_3} \\ e^{\beta_1} \\ e^{\beta_2+\beta_3} \\ e^{\beta_2} \end{bmatrix}$$

1.110 A Poisson distribution has the following density function

$$f(y;\mu) = e^{-\mu}\mu^{y} / y!$$

1.111 Its log-likelihood function is therefore

$$l(y;\mu) = \sum_{i=1}^{n} \ln f(y_i;\mu_i) = \sum_{i=1}^{n} -\mu_i + y_i \ln \mu_i - \ln(y_i!)$$

1.112 With the logarithm link function, $\mu_i = exp(\Sigma_j X_{ij}\beta_j)$, and the log-likelihood function reduces to

$$l(y;e^{X\beta}) = \sum_{i=1}^{n} -\exp(\sum_{j=1}^{p} X_{ij}.\beta_j) + y_i \sum_{j=1}^{p} X_{ij}.\beta_j - \ln(y_i!)$$

1.113 In this example, the equation is

$$l(y;\mu) = -e^{(\beta_1+\beta_3)} + 800.(\beta_1 + \beta_3) - \ln 800! \quad - e^{\beta_1} + 500.\beta_1 - \ln 500!$$
$$- e^{(\beta_2+\beta_3)} + 400.(\beta_2 + \beta_3) - \ln 400! \quad - e^{\beta_2} + 200.\beta_2 - \ln 200!$$

1.114 Ignoring the constant of ln800! + ln500! + ln400! + ln200!, the following function is to be maximized:

$$l^*(y;\mu) = -e^{(\beta_1+\beta_3)} + 800(\beta_1+\beta_3) - e^{\beta_1} + 500.\beta_1 - e^{(\beta_2+\beta_3)} + 400(\beta_2+\beta_3) - e^{\beta_2} + 200.\beta_2$$

1.115 To maximize l^* the derivatives with respect to β_1, β_2 and β_3 are set to zero and the following three equations are derived:

$$\frac{\partial l^*}{\partial \beta_1} = 0 \quad => \quad e^{\beta_1}(e^{\beta_3}+1) = 1300$$

$$\frac{\partial l^*}{\partial \beta_2} = 0 \quad => \quad e^{\beta_2}(e^{\beta_3}+1) = 600$$

$$\frac{\partial l^*}{\partial \beta_3} = 0 \quad => \quad e^{\beta_3}(e^{\beta_1}+e^{\beta_2}) = 1200$$

1.116 These can be solved to derive the following parameter estimates:

$$\beta_1 = 6.1716$$
$$\beta_2 = 5.3984$$
$$\beta_3 = 0.5390$$

which produces the following predicted values:

	Urban	Rural
Male	821.1	479.0
Female	378.9	221.1

The gamma error structure with an inverse link function

1.117 This example is set out in Appendix F.

Solving for large datasets using numerical techniques

1.118 The general case for solving for maximum likelihood in the case of a GLM with an assumed exponential distribution is set out in Appendix E. In insurance modeling there are typically many thousands if not millions of observations being modeled, and it is not practical to find values of $\underline{\beta}$ which maximize likelihood using the explicit techniques illustrated above and in Appendices E and F. Instead iterative numerical techniques are used.

1.119 As was the case in the simple examples above, the numerical techniques seek to optimize likelihood by seeking the values of $\underline{\beta}$ which set the first differential of the log-likelihood to zero, as there are a number of standard methods which can be applied to this problem. In practice, this is done using an iterative process, for example Newton-Raphson iteration which uses the formula:

$$\underline{\beta}_{n+1} = \underline{\beta}_n - \boldsymbol{H}^{-1}.\underline{s}$$

where $\underline{\beta}_n$ is the n^{th} iterative estimate of the vector of the parameter estimates $\underline{\beta}$ (with p elements), $\underline{s}$ is the vector of the first derivatives of the log-likelihood and **H** is the (p by p) matrix containing the second derivatives of the log-likelihood. This is simply the generalized form of the one-dimensional Newton-Raphson equation,

$$x_{n+1} = x_n - f'(x_n) / f''(x_n)$$

which seeks to find a solution to $f'(x)=0$.

1.120 The iterative process can be started using either values of zero for elements of $\underline{\beta}_0$ or alternatively the estimates implied by a one-way analysis of the data or of another previously fitted GLM.

1.121 Several generic commercial packages are available to fit generalized linear models in this way (such as SAS®, S+, R, etc), and packages specifically built for the insurance industry, which fit models GLMs more quickly and with helpful interpretation of output, are also available.

Base levels and the intercept term

1.122 The simple examples discussed above considered a three parameter model, where β_1 corresponded to men, β_2 to women and β_3 to the effect of being in an urban area. In the case of an additive model (with identity link function) this could be thought of as either

- assuming that there is an average response for men, β_1, and an average response for women, β_2, with the effect of being an urban policyholder (as opposed to being a rural one) having an additional additive effect β_3 which is the same regardless of gender

or

- assuming there is an average response for the "base case" of women in rural areas, β_2, with an additional additive effects for being male, β_1-β_2, and for being in an urban area, β_3.

1.123 In the case of a multiplicative model this three parameter form could be thought of as

- assuming that there is an average response for men, $exp(\beta_1)$, and an average response for women, $exp(\beta_2)$, with the effect of being an urban policyholder (as opposed to being a rural one) having a multiplicative effect $exp(\beta_3)$, which is the same regardless of gender

or

- assuming there is an average response for the "base case" of women in rural areas $exp(\beta_2)$ with an additional multiplicative effects for being male, $exp(\beta_1-\beta_2)$, and for being in an urban area $exp(\beta_3)$.

1.124 In the example considered, some measure of the overall average response was incorporated in both the values of β_1 and β_2. The decision to incorporate this in the parameters relating to gender rather than area was arbitrary.

1.125 In practice when considering many factors each with many levels it is more helpful to parameterize the GLM by considering, in addition to observed factors, an "intercept term", which is a parameter that applies to all observations.

1.126 In the above example, this would have been achieved by defining the design matrix X as

$$X = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 1 & 0 \end{bmatrix}$$

that is, by redefining β_1 as the intercept term, and only having one parameter relating to the gender of the policyholder. It would not be appropriate to have an intercept term *and* a parameter for every value of gender since then the GLM would not be uniquely defined - any arbitrary constant k could be added to the intercept term and subtracted from each of the parameters relating to gender and the predicted values would remain the same.

1.127 In practice when considering categorical factors and an intercept term, one level of each factor should have no parameter associated with it, in order that the model remains uniquely defined.

1.128 For example consider a simple rating structure with three factors - age of driver (a factor with 9 levels), territory (a factor with 8 levels) and vehicle class (a factor with 5 levels). An appropriate parameterization might be represented as follows:

Age of driver

Factor level	Parameter
17-21	β_2
22-24	β_3
25-29	β_4
30-34	β_5
35-39	β_6
40-49	
50-59	β_7
60-69	β_8
70+	β_9

Territory

Factor level	Parameter
A	β_{10}
B	β_{11}
C	
D	β_{12}
E	β_{13}
F	β_{14}
G	β_{15}
H	β_{16}

Vehicle class

Factor level	Parameter
A	
B	β_{17}
C	β_{18}
D	β_{19}
E	β_{20}

Intercept term	β_1

that is, an intercept term is defined for every policy, and each factor has a parameter associated with each level except one. If a multiplicative GLM were fitted to claims frequency (by selecting a log link function) the exponentials of the parameter estimates $\underline{\beta}$ could be set out in tabular form also:

Age of driver

Factor level	Multiplier
17-21	1.6477
22-24	1.5228
25-29	1.5408
30-34	1.2465
35-39	1.2273
40-49	1.0000
50-59	0.8244
60-69	0.9871
70+	0.9466

Territory

Factor level	Multiplier
A	0.9407
B	0.9567
C	1.0000
D	0.9505
E	1.0975
F	1.1295
G	1.1451
H	1.4529

Vehicle class

Factor level	Multiplier
A	1.0000
B	0.9595
C	1.0325
D	0.9764
E	1.1002

Intercept term	0.1412

1.129 In this example the claims frequency predicted by the model can be calculated for a given policy by taking the intercept term 0.1412 and multiplying it by the relevant factor relativities. For the factor levels for which no parameter was estimated (the "base levels"), no multiplier is relevant, and this is shown in the above table by displaying multipliers of 1. The intercept term relates to a policy with all factors at the base level (ie in this example the model predicts a claim frequency of 0.1412 for a 40-49 year old in territory C and a vehicle in class A). This intercept term is not an average rate since its value is entirely dependent upon the arbitrary choice of which level of each factor is selected to be the base level.

1.130 If a model were structured with an intercept term but without each factor having a base level, then the GLM solving routine would remove as many parameters as necessary to make the model uniquely defined. This process is known as *aliasing*.

Aliasing

1.131 Aliasing occurs when there is a linear dependency among the observed covariates $X_1,\ldots,X_p$. That is, one covariate may be identical to some combination of other covariates. For example, it may be observed that

$$\underline{X}_3 = 4 + \underline{X}_1 + 5\underline{X}_2$$

1.132 Equivalently, aliasing can be defined as a linear dependency among the columns of the design matrix $\boldsymbol{X}$.

1.133 There are two types of aliasing: intrinsic aliasing and extrinsic aliasing.

Intrinsic aliasing

1.134 Intrinsic aliasing occurs because of dependencies inherent in the definition of the covariates. These intrinsic dependencies arise most commonly whenever categorical factors are included in the model.

1.135 For example, suppose a private passenger automobile classification system includes the factor *vehicle age* which has the four levels: 0-3 years (X_1), 4-7 years (X_2), 8-9 years (X_3), and 10+ years (X_4). Clearly if any of X_1, X_2, X_3, is equal to 1 then X_4 is equal to 0; and if all of X_1, X_2, X_3, are equal to 0 then X_4 must be equal to 1. Thus $X_4 = 1 - X_1 - X_2 - X_3$.

1.136 The linear predictor

$$\eta = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$$

(ignoring any other factors) can be uniquely expressed in terms of the first three levels:

$$\begin{aligned}\eta &= \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 (1 - X_1 - X_2 - X_3) \\ &= (\beta_1 - \beta_4) X_1 + (\beta_2 - \beta_4) X_2 + (\beta_3 - \beta_4) X_3 + \beta_4\end{aligned}$$

1.137 Upon renaming the coefficients this becomes:

$$\eta = \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_0$$

1.138 The result is a linear predictor with an intercept term (if one did not already exist) and three covariates.

1.139 GLM software will remove parameters which are aliased. Which parameter is selected for exclusion depends on the software. The choice of which parameter to alias does not affect the fitted values. For example in some cases the last level declared (ie the last alphabetically) is aliased. In other software the level with the maximum exposure is selected as the base level for each factor first, and then other levels are aliased dependent upon the order of declaration. (This latter approach is helpful since it minimizes the standard errors associated with other parameter estimates - this subject is discussed later in this paper.)

Extrinsic Aliasing

1.140 This type of aliasing again arises from a dependency among the covariates, but when the dependency results from the nature of the data rather than inherent properties of the covariates themselves. This data characteristic arises if one level of a particular factor is perfectly correlated with a level of another factor.

1.141 For example, suppose a dataset is enriched with external data and two new factors are added to the dataset: the factors *number of doors* and *color of vehicle.* Suppose further that in a small number of cases the external data could not be linked with the existing data with the result that some records have an unknown color and an unknown number of doors.

Exposures		**# Doors**				
		2	3	4	5	Unknown
Color	Red	13,234	12,343	15,432	13,432	0
	Green	4,543	4,543	13,243	2,345	0
	Blue	6,544	5,443	15,654	4,565	0
	Black	4,643	1,235	14,565	4,545	0
	Unknown	0	0	0	0	3,242

Selected Base: # Doors = 4; Color = Red

Additional Aliasing: Color = Unknown

1.142 In this case because of the way the new factors were derived, the level *unknown* for the factor *color* happens to be perfectly correlated with the level *unknown* for the factor # *doors*. The covariate associated with *unknown color* is equal to 1 in every case for which the covariate for *unknown # doors* is equal to 1, and vice versa.

1.143 Elimination of the base levels through intrinsic aliasing reduces the linear predictor from 10 covariates to 8, plus the introduction of an intercept term. In addition, in this example, one further covariate needs to be removed as a result of extrinsic aliasing. This could either be the *unknown color* covariate or the *unknown # doors* covariate. Assuming in this case the GLM routine aliases on the basis of order of declaration, and assuming that the # *doors* factor is declared before *color*, the GLM routine would alias *unknown color* reducing the linear predictor to just 7 covariates.

"Near Aliasing"

1.144 When modeling in practice a common problem occurs when two or more factors contain levels that are almost, but not quite, perfectly correlated. For example, if the color of vehicle was known for a small number of policies for which the # doors was unknown, the two-way of exposure might appear as follows:

Exposures		**# Doors**				
		2	3	4	5	Unknown
Color	Red	13,234	12,343	15,432	13,432	0
	Green	4,543	4,543	13,243	2,345	0
	Blue	6,544	5,443	15,654	4,565	0
	Black	4,643	1,235	14,565	4,545	5
	Unknown	0	0	0	0	3,242

Selected Base: # Doors = 4; Color = Red

1.145 In this case the *unknown* level of color factor is not perfectly correlated to the *unknown* level of the # doors factor, and so extrinsic aliasing will not occur.

1.146 When levels of two factors are "nearly aliased" in this way, convergence problems can occur. For example, if there were no claims for the 5 exposures indicated in *black color* level and *unknown # doors* level, and if a log link model were fitted to claims frequency, the model would attempt to estimate a very large and negative parameter for *unknown # doors* (for example, -20) and a very large parameter for *unknown color* (for example 20.2). The sum (0.2 in this example) would be an appropriate reflection of the claims frequency for the 3,242 exposures having unknown # doors and unknown color, while the value of the *unknown # doors* parameter would be driven by the experience of the 5 rogue exposures having color black with unknown # doors. This can either give rise to convergence problems, or to results which can appear very confusing.

1.147 In order to understand the problem in such circumstances it is helpful to examine two-way tables of exposure and claim counts for the factors which contain very large parameter estimates. From these it should be possible to identify those factor combinations which cause the near-aliasing. The issue can then be resolved either by deleting or excluding those rogue records, or by reclassifying the rogue records into another, more appropriate, factor level.

Model diagnostics

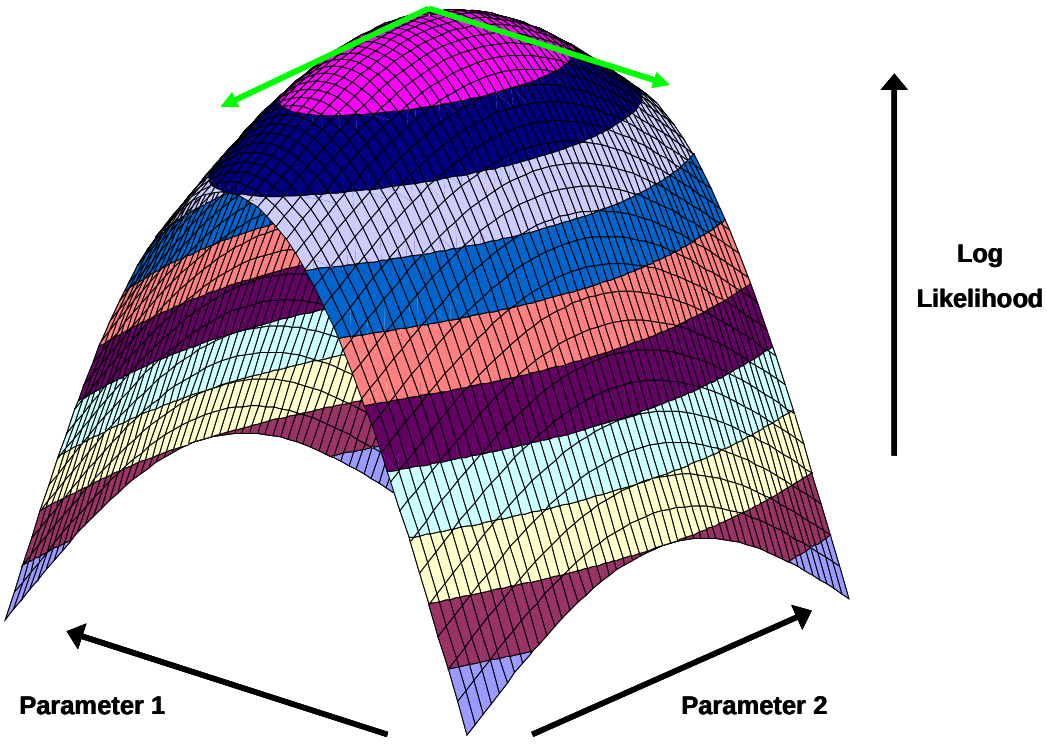
1.148 As well as deriving parameter estimates which maximize likelihood, a GLM can produce important additional information indicating the certainty of those parameter estimates (which themselves are estimates of some true underlying value).

Standard errors

1.149 Statistical theory can be used to give an estimate of the uncertainty. In particular, the multivariate version of the Cramer-Rao lower bound (which states that the variance of a parameter estimate is greater than or equal to minus one over the second derivative of the log likelihood) can define "standard errors" for each parameter estimate. Such standard errors are defined as being the diagonal element of the covariance matrix $-\boldsymbol{H}^{-1}$ where $\boldsymbol{H}$ (the Hessian) is the second derivative matrix of the log likelihood.

1.150 Intuitively the standard errors can be thought of as being indicators of the speed with which log-likelihood falls from the maximum given a change in a parameter. For example consider the below diagram.

Intuitive illustration of standard errors



1.151 This diagram illustrates a simple case with two parameters (β_1 and β_2) and shows how log likelihood varies, for the dataset in question, for different values of the two parameters. It can be seen that movements in parameter 1 from the optimal position reduce log likelihood more quickly than similar movements in parameter 2, that is to say the log likelihood curve becomes steeper in the parameter 1 direction than in the parameter 2 direction. This can be thought of as the second partial differential of log likelihood with respect to parameter 1 being large and negative, with the result that the standard error for parameter 1 (being minus one over the second partial differential) is small. Conversely the second partial differential of log likelihood with respect to parameter 2 is less large and negative, with the standard error for parameter 2 being larger (indicating greater uncertainty).

1.152 Generally it is assumed that the parameter estimates are asymptotically Normally distributed; consequently it is in theory possible to undertake a simple statistical test on individual parameter estimates, comparing each estimate with zero (ie testing whether the effect of each level of the factor is significantly different from the base level of that factor). This is usually performed using a χ^2 test, with the square of the parameter estimate divided by its variance being compared to a χ^2 distribution. This test in fact compares the parameter with the base level of the factor. This is not necessarily a fully useful test in isolation as the choice of base level is arbitrary. It is theoretically possible to change repeatedly the base level and so construct a triangle of χ^2 tests comparing every pair of parameter estimates. If none of these differences is significant then this is good evidence that the factor is not significant.

1.153 In practice graphical interpretation of the parameter estimates and standard errors are often more helpful, and these are discussed in Section 2.

Deviance tests

1.154 In addition to the parameter estimate standard errors, measures of deviance can be used to assess the theoretical significance of a particular factor. In broad terms, a deviance is a measure of how much the fitted values differ from the observations.

1.155 Consider a deviance function $d(Y_i,\mu_i)$ defined by

$$d(Y_i;\mu_i) = 2\omega_i \int_{\mu_i}^{Y_i} \frac{(Y_i - \zeta)}{V(\zeta)} d\zeta$$

Under the condition that V(x) is strictly positive, $d(Y_i,\mu_i)$ is also strictly positive and satisfies the conditions for being a distance function. Indeed it should be interpreted as such.

1.156 Consider an observation Y_i and a GLM that makes a prediction μ_i for that observation. $d(Y_i,\mu_i)$ is a measure of the difference between the fitted and actual observations which gives more weight to the difference between Y_i and μ_i when the variance function $V(x)$ is small. That is, if Y_i is known to come from a distribution with small variance then any discrepancy between Y_i and μ_i is given more emphasis.

1.157 $d(Y_i,\mu)$ can be thought of as a generalized form of the squared error.

1.158 Summing the deviance function across all observations gives an overall measure of deviance referred to as the total deviance D:

$$D = \sum_{i=1}^{n} 2\omega_i \int_{\mu_i}^{Y_i} \frac{(Y_i - \zeta)}{V(\zeta)} d\zeta$$

1.159 Dividing this by the scale parameter ϕ gives the scaled deviance D^*, which can be thought of as a generalized form of the sum of squared errors, adjusting for the shape of the distribution.

$$D^* = \sum_{i=1}^{n} 2\frac{\omega_i}{\phi} \int_{\mu_i}^{Y_i} \frac{(Y_i - \zeta)}{V(\zeta)} d\zeta$$

1.160 For the class of exponential distributions the scaled deviance can be shown to be equal to twice the difference between the maximum achievable likelihood (ie the likelihood where the fitted value is equal to the observation for every record) and the likelihood of the model.

1.161 A range of statistical tests can be undertaken using deviance measures. One of the most useful tests considers the ratio of the likelihood of two "nested" models, that is to say where one model contains explanatory variables which are a subset of the explanatory variables in a second model. Such tests are often referred to as "type III" tests (as opposed to "type I" tests which consider the significance of factors as they are added sequentially to a model with only an intercept term, referred to as a null model).

1.162 The change in scaled deviance between two nested models (which reflects the ratio of the likelihoods) can be considered to be a sample from a χ^2 distribution with degrees of freedom equal to the difference in degrees of freedom between the two models (where the degrees of freedom for a model is defined as the number of observations less the number of parameters), ie

$$D_1^* - D_2^* \sim \chi^2_{df_1 - df_2}$$

1.163 This allows tests to be undertaken to assess the significance of the parameters that differ between the two models (with the null hypothesis that the extra parameters are not important). Expressed crudely this measures whether the inclusion of an explanatory factor in a model improves the model enough (ie decreases the deviance enough) given the extra parameters which it adds to the model. Adding any factor will improve the fit on the data in question - what matters is whether the improvement is significant given the extra parameterization.

1.164 The χ^2 tests depend on the scaled deviance. For some distributions (such as the Poisson and the binomial) the scale parameter is assumed to be known, and it is possible to calculate the statistic accurately. For other distributions the scale parameter is not known and has to be estimated, typically as the ratio of the deviance to the degrees of freedom. This can decrease the reliability of this test if the estimate of the scale parameter used is not accurate.

1.165 It is possible to show that, after adjusting for the degrees of freedom and the true scale parameter, the estimate of the scale parameter is also distributed with a χ^2 distribution. The F-distribution is the ratio of χ^2 distributions. The ratio of the change in deviance and the adjusted estimate of the scale is therefore distributed with an F-distribution.

$$\frac{(D_1 - D_2)}{(df_1 - df_2)D_2 / df_2} \sim F_{df_1 - df_2, df_2}$$

1.166 This means that the F-test is suitable for use when the scale parameter is not known (for example when using the gamma distribution). There is no advantage to using this test where the scale is known.

2 GLMs in practice

2.1 Section 1 discussed how GLMs are formularized and solved. This section considers practical issues and presents a plan for undertaking a GLM analysis in four general stages:

- pre-modeling analysis - considering data preparation as well as a range of helpful portfolio investigations
- model iteration - typical model forms and the diagnostics used in both factor selection and model validation
- model refinement - investigating interaction variables, the use of smoothing, and the incorporation of artificial constraints
- interpretation of the results - how model results can be compared to existing rating structures both on a factor-by-factor basis and overall.

Data required

2.2 GLM claim analyses require a certain volume of experience. Depending on the underlying claim frequencies and the number of factors being analyzed, credible results on personal lines portfolios can generally be achieved with around 100,000 exposures (which could for example be 50,000 in each of two years, etc). Meaningful results can sometimes be achieved with smaller volumes of data (particularly on claim types with adequate claims volume), but it is best to have many 100,000s of exposures. As models fitted to only a single year of data could be distorted by events that occurred during that year, the data should ideally be based on two or three years of experience.

2.3 In addition to combining different years of experience, combining states (or provinces) can also improve stability, assuming inputs are consistent across borders.[7] In the case where one geographic area has sufficient exposure it may be more appropriate to fit a model just to that area's experience. If a countrywide model has been run, the goodness of fit of that model on state data may be investigated, or the state and countrywide model results may be compared side-by-side. Examining the interaction of state with each predictive factor may also identify where state patterns differ from countrywide; interaction variables are discussed later in this paper.

[7] In this sense, inputs refer to explanatory criteria, not necessarily existing rating relativities. Data coding should be reviewed to ensure some level of consistency and care should be taken with recycled coding from one state to another (eg territory 1 in Virginia should not be added to territory 1 in Florida).

2.4 Different types of claim can be affected by rating factors in different ways and so often it is appropriate to analyze different types of claim with separate models. Analyzing different claim elements separately will often identify clearer underlying trends than considering models based on a mixture of claims (for example, liability claims combined with theft claims). Even if a single model is required ultimately, it is generally beneficial to model by individual claim type and later to produce a single model which fits the aggregate of the underlying models by claim type.

2.5 The overall structure of a dataset for GLM claims analysis consists of linked policy and claims information at the individual risk level. Typical data requirements and a more detailed discussion of issues such as dealing with IBNR are set out in Appendix G. In summary, however, the following fields would typically be included in a GLM claims dataset.

- Raw explanatory variables - whether discrete or continuous, internal or external to the company.
- Dummy variables to standardize for time-related effects, geographic effects and certain historical underwriting effects.
- Earned exposure fields - preferably by claim type if certain claim types are only present for some policies. These fields should contain the amount of exposure attributable to the record (eg measured in years).
- Number of incurred claims fields. There should be one field for each claim type, giving the number of claims associated with the exposure period in question.
- Incurred loss amounts fields. There should be one field for each claim type, giving the incurred loss amount of claims associated with the exposure period in question, based on the most recent possible case reserve estimates.
- Premium fields. These give the premium earned during the period associated with the record. If it is possible to split this premium between the claim types then this can be used to enhance the analysis. This information is not directly required for modeling claims frequency and severity, however it can be helpful for a range of post-modeling analyses such as measuring the impact of moving to a new rating structure.

2.6 When analyzing policyholder retention or new business conversion, a different form of data is required. For example to fit GLMs to policyholder renewal experience, a dataset would contain one record for each invitation to renew and would contain the following fields:

- explanatory variables including, for example,
  - rating factors
  - other factors such as distribution channel, method of payment and number of terms with company
  - change in premium on latest renewal[8]
  - change in premium on previous renewal
  - measure of competitiveness on renewal premium
  - details of any mid-term adjustments occurring in the preceding policy period
- number of invitations to renew (typically 1 for each record - this would be the measure of exposure)
- whether or not the policy renewed.

2.7 If several risks are written on a single policy, renewal may be defined at the policy level or at the individual risk level (for example, a personal automobile carrier may write all vehicles in a household on a single policy). An understanding of how the model will be used will aid data preparation. For example, models that will be part of a detailed model office scenario will benefit from data defined at the individual risk level. Models used to gain an overall understanding of which criteria affect policyholder retention (perhaps for marketing purposes) would not require such detail.

[8] Separation of premium change into rate change and risk criteria change would be beneficial.

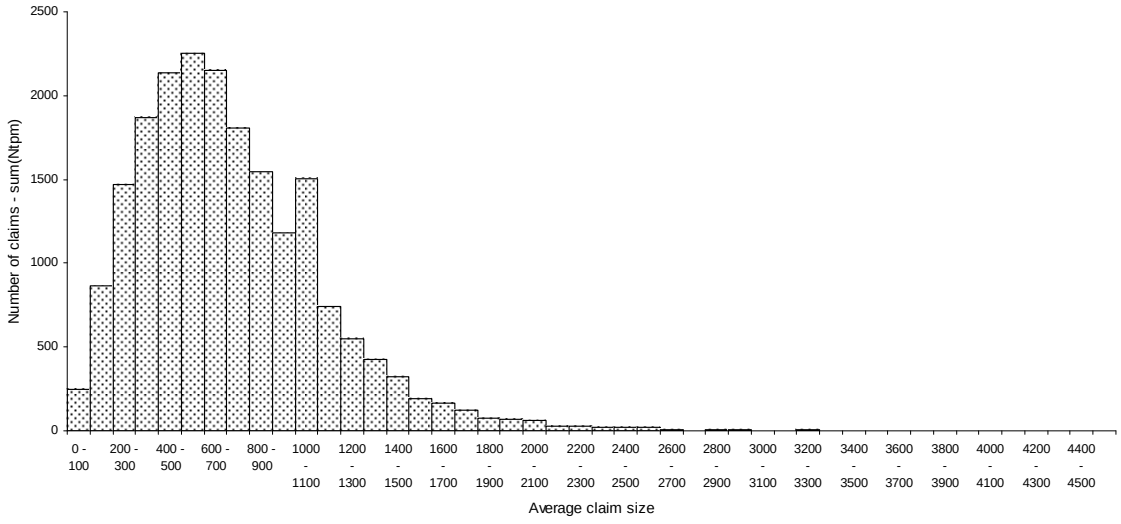
Preliminary analyses

2.8 Before modeling, it is generally helpful to undertake certain preliminary analyses. These analyses include data checks such as identification of records with negative or zero exposures, negative claim counts or losses, and blanks in any of the statistical fields. In addition, certain logical tests may be run against the data - for example, identifying records with incurred losses but with no corresponding claim count.

Analysis of distributions

2.9 One helpful preliminary analysis is to consider the distribution of key data items for the purpose of identifying any unusual features or data problems that should be investigated prior to modeling. Mainly this concerns the distribution of claim amounts (ie number of claim counts by average claim size), which are examined in order to identify features such as unusually large claims and distortions resulting from average reserves placed on newly reported claims. A typical claim distribution graph is shown below.

Distribution of claim amounts



2.10 This distribution, along with a distribution of loss amount by average claim size, will aid in understanding the tail of the distribution for a particular claim type. (When modeling severity, it is often appropriate to apply a large loss threshold to certain claim types, and this helps assess possible thresholds. A tabular representation of the distribution would also help quantify the percent of the claims distribution which would be affected by different large loss thresholds.)

2.11 Distribution analyses can also highlight specific anomalies that might require addressing prior to modeling. For example, if many new claims have a standard average reserve allocated to them, it might be appropriate to adjust the amount of such an average reserve if it was felt that the average level was systematically below or above the ultimate average claims cost.

One and two-way analyses

2.12 Although GLMs are a multivariate method, there is generally benefit in reviewing some one-way and two-way analyses of the raw data prior to modeling.

2.13 Firstly, the one-way distribution of exposure and claims across levels of each raw variable will indicate whether a variable contains enough information to be included in any models (for example, if 99.5% of a variable's exposures are in one level, it may not be suitable for modeling).

2.14 Secondly, assuming there is some viable distribution by levels of the factor, consideration needs to be given to any individual levels containing very low exposure and claim count. If these levels are not ultimately combined with other levels, the GLM maximum likelihood algorithm may not converge (if a factor level has zero claims and a multiplicative model is being fitted, the theoretically correct multiplier for that level will be close to zero, and the parameter estimate corresponding to the log of that multiplier may be so large and negative that the numerical algorithm seeking the maximum likelihood will not converge).

2.15 In addition to investigating exposure and claim distribution, a query of one-way statistics (eg frequency, severity, loss ratio, pure premium) will give a preliminary indication of the effect of each factor.

Factor categorizations

2.16 Before modeling, it is necessary to consider how explanatory variables should be categorized, and whether any variables should be modeled in a continuous fashion as variates (or polynomials in variates). Although variates do not require any artificially imposed categorization, the main disadvantage is that the use of polynomials may smooth over interesting effects in the underlying experience. Often it is better to begin modeling all variables as narrowly defined categorical factors (ensuring sufficient data in each category) and if the categorical factor presents GLM parameter estimates which appear appropriate for modeling with a polynomial, then the polynomial in the variate may be used in place of the categorical factor.

2.17 When using categorical factors consideration needs to be given to the way in which the factors are categorized. If an example portfolio contained a sufficient amount of claims for each for each age of driver (say from age 16 to 99), the categorization of age of driver may consist of each individual age. This is rarely the case in practice, however, and often it is necessary that levels of certain rating factors are combined.

2.18 In deriving an appropriate categorization, the existing rating structure may provide initial guidance (particularly if the GLMs are to be applied in ratemaking), with factor levels with insufficient exposure then being grouped together and levels with sufficient exposure being considered separately. In general such a manual approach tends to be the most appropriate. One particular automated approach within the GLM framework is considered in Appendix H. This approach, however, would not necessarily produce any more appropriate results than the manual approach.

Correlation analyses

2.19 Once categorical factors have been defined, it can also be helpful to consider the degree to which the exposures of explanatory factors are correlated. One commonly used correlation statistic for categorical factors is Cramer's V statistic.[9] Further information about this statistic is set out in Appendix I.

2.20 Although not used directly in the GLM process, an understanding of the correlations within a portfolio is helpful when interpreting the results of a GLM. In particular it can explain why the multivariate results for a particular factor differ from the univariate results, and can indicate which factors may be affected by the removal or inclusion of any other factor in the GLM.

Data extracts

2.21 In practice it is not necessary to fit every model to the entire dataset. For example, modeling severity for a particular claim type only requires records that contain a claim of that type. Running models against data subsets, or extracts, can improve model run speed.

[9] Other correlation statistics for categorical factors include Pearson chi-square, Likelihood ratio chi-square, Phi coefficient and Contingency coefficient. A thorough discussion of these statistics is beyond the scope of this paper.

2.22 The error term assumed for a model can also influence these data extracts. In the case of claim counts, a particular property of Poisson multiplicative model is that the observed data Y_i can be grouped by unique combination of rating factors being modeled (summing exposure and claim counts for each unique combination) and the GLM parameter estimates and the parameter estimate standard errors will remain unchanged. This is helpful in practice since it can decrease model run times. This is not the case for some other distributions.

2.23 A gamma multiplicative model does not produce identical results if the observations are grouped by unique combinations of factors. Such a grouping would not change parameter estimates, but it would affect the standard errors. Depending on the line of business, however, it may be appropriate to group the small number of multiple claims which occur on the same policy in the same exposure period.

Model iteration and the role of diagnostics

2.24 Given data relating to the actual observations and the assumptions about the model form, a GLM will yield parameter estimates which best fit the data given that model form. The GLM will not automatically provide information indicating the appropriateness of the model fitted - for this it is necessary to examine a range of diagnostics. This section reviews model forms typically used in practice and discusses the range of diagnostics which aid in both the selection of explanatory factors and the validation of statistical assumptions.

Factor selection

2.25 One of the key issues to consider is which explanatory factors should be included in the model. The GLM will benefit from including factors which systematically affect experience, but excluding factors which have no systematic effect. To distinguish whether a factor effect is systematic or random (and therefore unlikely to be repeated in the future) there are a number of criteria which can be considered, including

- parameter estimate standard errors
- deviance tests (type III tests)
- consistency with time
- common sense.

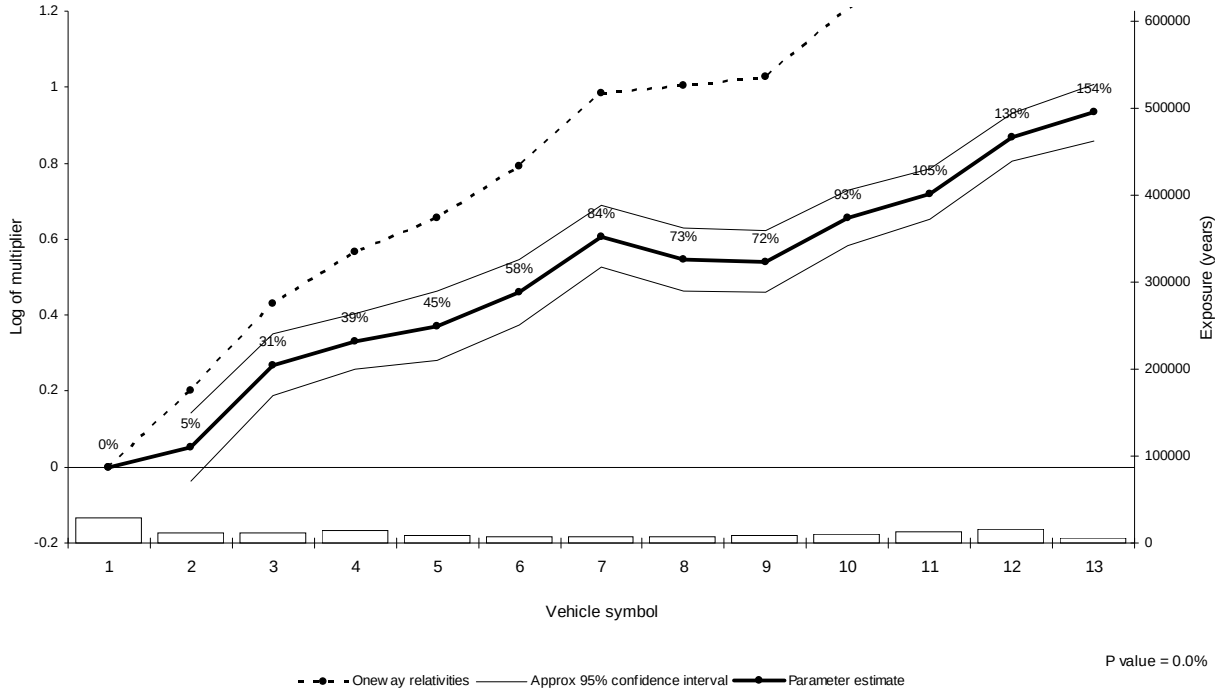
Standard errors

2.26 As discussed in Section 1, as well as deriving parameter estimates which maximize likelihood, a GLM can produce important additional information indicating the certainty of those parameter estimates.

2.27 One such helpful diagnostic is the standard errors of the parameter estimates, defined as being the square root of the diagonal element of $-\boldsymbol{H}^{-1}$ where $\boldsymbol{H}$ (the Hessian) is the second derivative matrix of the log likelihood.

2.28 Although theoretically tests could be performed on individual parameter estimates using standard errors, in practice it is often more helpful to consider for each factor in the GLM the fitted parameter estimates alongside the associated standard errors (for one base level) in a graphical form thus:

GLM output (example of significant factor)



2.29 One such graph would be shown for each factor in the model. In this case the factor in question is Vehicle Symbol with levels running from 1 to 13.

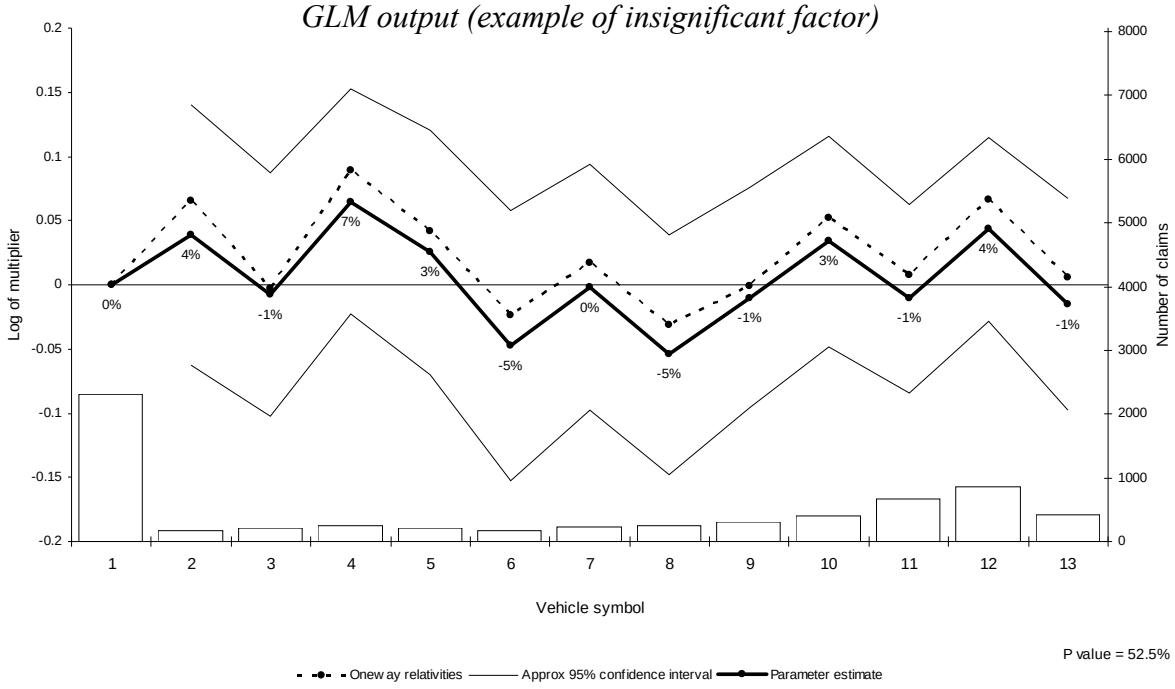
2.30 The thick solid line shows the fitted parameter estimates. In this case the model is a multiplicative model with a log link function and so the parameter estimates represent logs of multipliers. For clarity the implied loadings are shown as labels by each point on the thick solid line. For example - the parameter estimate for Vehicle Symbol 3 has value 0.27. This means that the model estimates that, all other factors being constant, exposures with Vehicle Symbol 3 will have a relativity of $e^{0.27} = 1.31$ times that expected for exposures at the base level (in this example Symbol 1). This multiplier is shown on the graph as a "loading" of 31%.

2.31 The thin solid lines on each graph indicate two standard errors either side of the parameter estimate. Very approximately this means that (assuming the fitted model is appropriate and correct) the data suggests that the true relativity for each level of rating factor will lie between the two thin solid lines with roughly 95% certainty. The two bands will be wide apart, indicating great uncertainty in the parameter estimate where there is low exposure volume, where other correlated factors also explain the risk, or where the underlying experience is very variable.

2.32 The dotted lines shows the relativities implied by a simple one-way analysis. These relativities make no allowance for the fact that the difference in experience may be explained in part by other correlated factors. These one-way estimates are of interest since they will differ from the multivariate estimates for a given factor when there are significant correlations between that factor and one or more other significant factors. The distribution of exposure for all business considered is also shown as a bar chart at the bottom of each graph. This serves to illustrate which level of each factor may be financially significant.

2.33 Even though the standard errors on the graph only indicate the estimated certainty of the parameter estimates relative to the base level, such graphs generally give a good intuitive feel for the significance of a factor. For example in the above case it is clear that the factor is significant since the parameter estimates for Vehicle Symbols 3 to 13 are considerably larger than twice the corresponding standard errors. By contrast the graph below (an example of the same factor in a different model for a different claim type) illustrates an example where a factor is not significant - in this case there are no parameter estimates more than two standard errors from zero.

GLM output (example of insignificant factor)



2.34 Sometimes some levels of a categorical factor may be clearly significant, while other levels may be less so. Although the factor as a whole may be statistically significant, this may indicate that it is appropriate to re-categorize the factor, grouping together the less significant levels with other levels.

Deviance tests

2.35 As discussed in Section 1, comparing measures of deviance of two nested models allows "type III" tests (χ^2 or F-tests depending on whether or not the scale parameter ϕ is known) to be performed to determine the theoretical significance of individual factors.

2.36 In the Vehicle Symbol examples above (which were based on frequency models of two different claim types, each with a Poisson error structure), the resulting probability values (or P values) from the χ^2 tests are shown as footnotes to the graphs. Each χ^2 test compares a model with Vehicle Symbol to one without. In the first case the χ^2 test shows a probability level close to 0 (displayed to one decimal place as 0.0%). This means that the probability of this factor having such an effect on the deviance by chance is almost zero, ie this factor (according to the χ^2 test) is highly significant. Conversely in the second example the probability value is 52.5%, indicating that the factor is considerably less significant and should be excluded from the model. Typically factors with χ^2 or F-test probability levels of 5% or less are considered significant.

2.37 These kinds of type III likelihood ratio tests can provide additional information to the graphical interpretation of parameter estimates and standard errors. For example if other correlated factors in a model could largely compensate for the exclusion of a factor, this would be indicated in the type III test. Also the type III test is not influenced by the choice of the base level in the way that parameter estimate standard errors are.

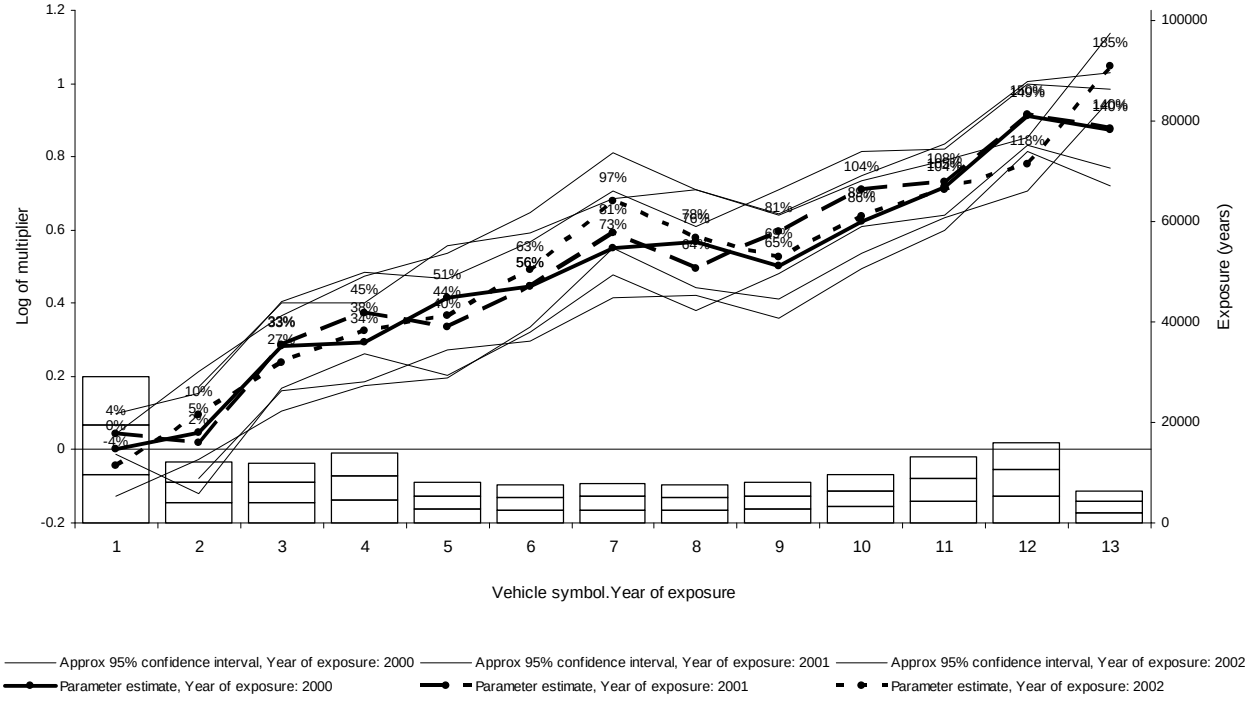
2.38 On the other hand, type III tests can be impractical on occasions - for example if a 20 level factor contained only one level that had any discriminatory effect on experience, a type III test might indicate that the factor was statistically significant, whereas a graphical representation of the model results would show at a glance that the factor contained too many levels and needed to be re-categorized with fewer parameters.

Interaction with time

2.39 In addition to classical statistical tests it can often be helpful to consider rather more pragmatic tests such as whether the observed effect of a rating factor is consistent over time. For example if more than one year's experience is being considered it is possible to consider the effect of a particular factor in each calendar year of exposure (or alternatively policy year). In theory this could be done by fitting separate models to each year and then comparing the results, however this can be hard to interpret since a movement in one factor in one year may to a large extent be compensated for by a movement in another correlated factor. A potentially clearer test, therefore, is to fit a series of models each one of which considers the interaction of a single factor with time. (Interactions are discussed in more detail later in this paper.)

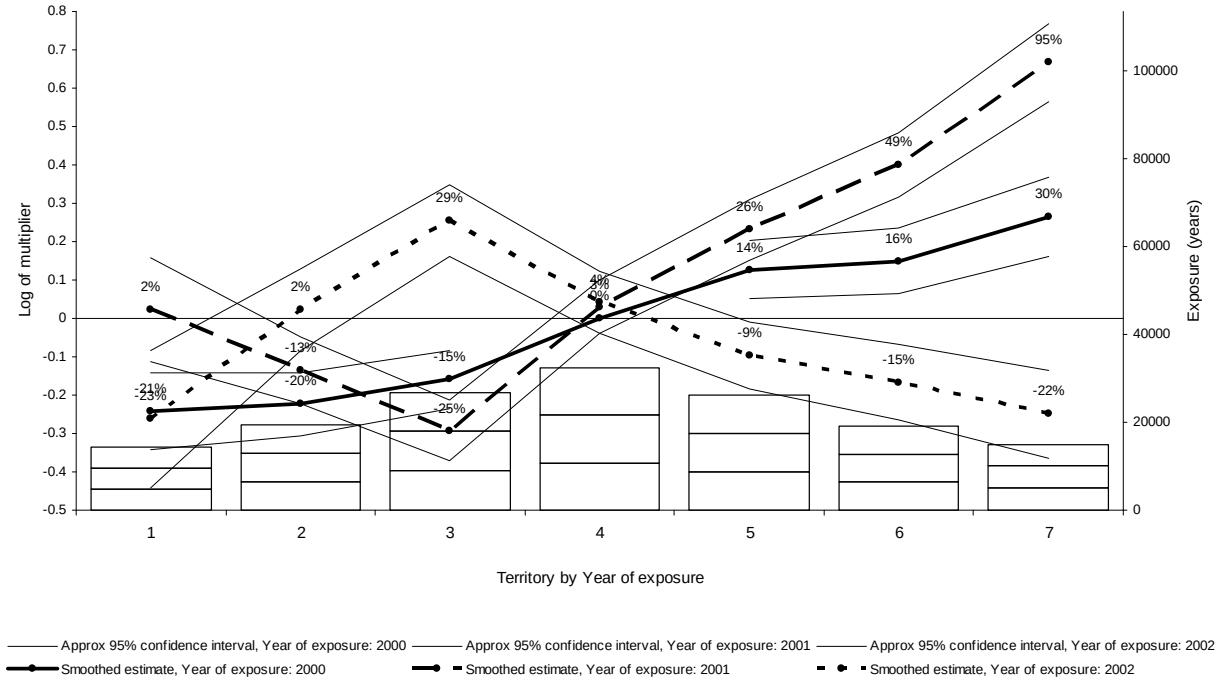
2.40 The below diagram shows one example factor interacted with calendar year of exposure. It is clear from this result showing lines which are largely parallel that the factor effect is mainly consistent from year to year, suggesting that the factor is likely to be a good predictor of future experience.

GLM output - example showing factor consistent over time



2.41 Conversely the graph below shows an example of a factor (in this case territory classification) which, although significant according to classical type III tests, shows a pattern for some levels which differs from year to year. In such a case it would be appropriate to investigate whether there was a possible explanation for such variations. If the variation can be attributed to some known change (for examples some event in one of the territories during one period) then that can be allowed when interpreting the results. If no explanation can be found for variations over time, this may indicate that the factor will be an unreliable predictor of future experience.

GLM output - example showing factor inconsistent over time



Intuition

2.42 In addition to statistical and other pragmatic tests, common sense can also play an important role in factor selection. Issues which should be considered when assessing the significance of a factor include

- whether the observed effect of a factor is similar across models which consider related types of claim (eg auto property damage liability and collision)
- whether the observed effect makes logical sense (given the other factors in the model)

- whether the observed effects of a categorical factor which represents a continuous variable (such as the age of a vehicle) show a natural trend - the model has no way of knowing that factor levels have a natural order, therefore if a trend is observed this may suggest that the factor has a more significant effect than the pure statistical tests alone would suggest.

Model iteration / stepwise macros

2.43 It is not generally possible to determine from a single GLM which set of factors are significant since the inclusion or exclusion of one factor will change the observed effects and therefore possibly the significance of other correlated factors in the model. To determine the theoretically optimal set of factors, therefore, it is generally necessary to consider an iterated series of models.

2.44 Often the model iteration starts with a GLM that includes all the main explanatory variables. Insignificant factors can then be excluded, one at a time, refitting the model at each stage.

2.45 When a factor is identified as being insignificant it is helpful to compare the GLM parameter estimates for that factor with the equivalent one-way relativities. When the GLM parameter estimates are different from the one-way relativities this indicates that the factor in question is correlated with other factors in the model and that the removal of that factor from the model is likely to affect the parameter estimates for other factors and quite possibly also their significance. Conversely if the one-way relativities are very similar to the GLM relativities for the factor to be excluded, it is likely that there will be no such consequences and that therefore to save time a second insignificant factor could be removed at that iteration also.

2.46 If a very large number of factors are to be considered it can be impractical to start the factor iteration process with all possible factors in the model. In such cases it is possible to select a model with certain factors which are known to be important, and then to test all other excluded factors by fitting a series of models which, one at a time, tests the consequences of including each of the excluded factors. The most significant of the excluded factors can then be included in the model, and then the other excluded factors can be retested for significance.

2.47 Where possible it is generally best to iterate models manually by analyzing the various diagnostics described above for each factor. In practice if many factors are being analyzed this can be impractical. In such circumstances automatic "stepwise" model iterating algorithms can be programmed to iterate models on the basis of type III tests alone. Such algorithms start with a specified model, and then:

a. the significance of each factor in the model is tested with a type III test, and the least significant factor is removed from the model if the significance is below a certain specified threshold

b. the significance of each factor not in the model (but in a specified list of potential factors) is tested by (one at a time) creating a new model containing the factors in the previous step plus the potential new factor. The most significant factor not currently in the model (according to a type III test) is then included if the significance is above the specified threshold

c. steps **a.** and **b.** are repeated until all factors in the model are deemed significant, and all factors not in the model are deemed insignificant.

2.48 Such algorithms allow no human judgment to be exercised and can take a significant time to complete. They are also heavily dependent on the type III test which has some practical shortcomings as described previously. Nevertheless they can derive a theoretically optimal model which at the very least could form the starting point for a more considered manual iteration.

Model validation

2.49 As well as considering the significance of the modeled rating factors, there are a number of more general diagnostic tests which allow the appropriateness of other model assumptions to be assessed. Diagnostics which aid in this investigation include:

- residuals which test the appropriateness of the error term
- leverage which identifies observations which have undue influence on a model
- the Box-Cox transformation which examines the appropriateness of the link function

Residuals

2.50 Various measures of residual can be derived to show, for each observation, how the fitted value differs from the actual observation.

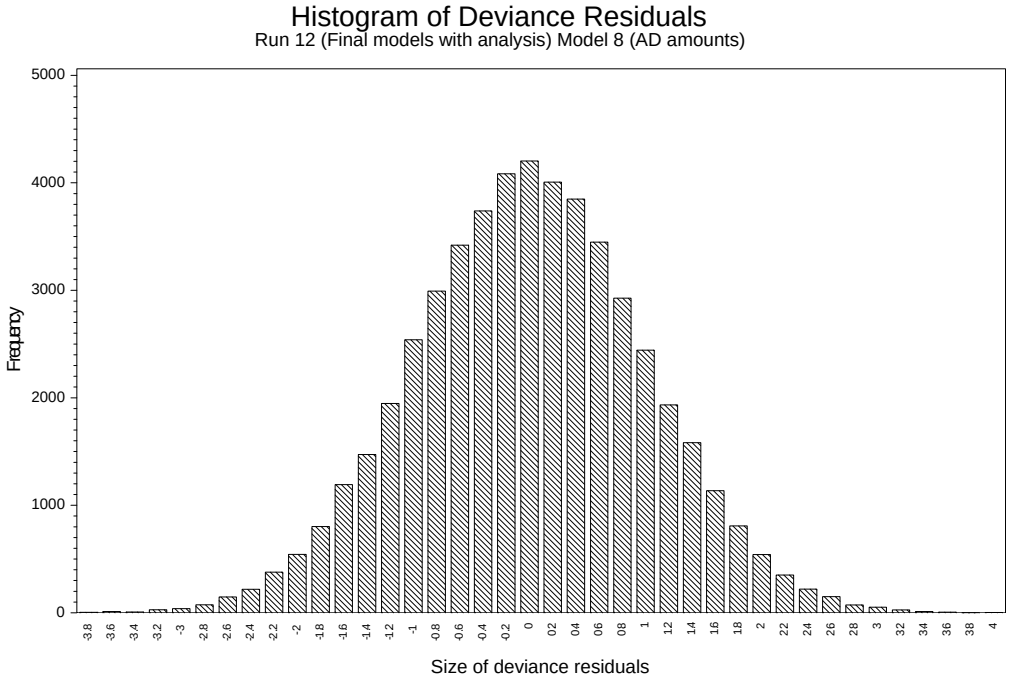
2.51 One measure of residual is the deviance residual

$$r_i^D = sign(Y_i - \mu_i)\sqrt{2\,\omega_i \int_{\mu_i}^{Y_i} \frac{(Y_i - \zeta)}{V(\zeta)} d\zeta}$$

which is the square root of the observation's contribution to the total deviance (ie a measure of the distance between the observation and the fitted value), multiplied by 1 or -1 depending on whether the observation is more than or less than the fitted value.

2.52 The deviance residuals have various helpful properties. In general they will be more closely Normally distributed than the raw residuals (defined simply as the difference between the actual observation and the expected value predicted by the GLM), as the deviance calculation corrects for the skewness of the distributions. For continuous distributions it is possible to test the distribution of the deviance residuals to check that they are Normally distributed. Any large deviation from this distribution is a good indication that the distributional assumptions are being violated.

2.53 The below diagram shows a distribution of deviance residuals from an example model. In this case the residuals appear to be reasonably consistent with a Normal distribution.



2.54 For discrete distributions the deviance residuals based on individual observations tend not to appear Normally distributed. This is because the calculation of the contribution to the deviance can adjust for the shape but not the discreteness of the observations. For example, in the case of fitting a model to claim numbers, a GLM might predict a fitted value for a record of say 0.1 representing an expected claims frequency of 10%. In reality (ignoring multiple claims) either a claim occurs for that record or it does not, with the result that the residual for that record will either correspond to an "actual minus expected" value of (0-0.1) = -0.1, or (with lower probability), the residual will correspond to an "actual minus expected" value of (1-0.1) = 0.9.

2.55 Some practitioners group together the individual residuals into large groups of similar risks. This aggregation can disguise the discreteness allowing some distributional tests to be performed. For example, it is commonly thought that a Poisson with a suitably large mean can be thought of as being almost Normally distributed. At this point the deviance residual calculated on the aggregate data should be smooth enough to test meaningfully.

2.56 The deviance residuals are often standardized before being analyzed. The purpose of this standardization is to transform the residuals so that they have variance 1 if the model assumptions hold. This is achieved by adjusting by the square root of the scale parameter and also by the square root of one minus the "leverage" h_i:

$$r_i^{DS} = \frac{sign(Y_i - \mu_i)}{\sqrt{\phi(1-h_i)}} \sqrt{2\,\omega_i \int_{\mu_i}^{Y_i} \frac{(Y_i - \zeta)}{V(\zeta)} d\zeta}$$

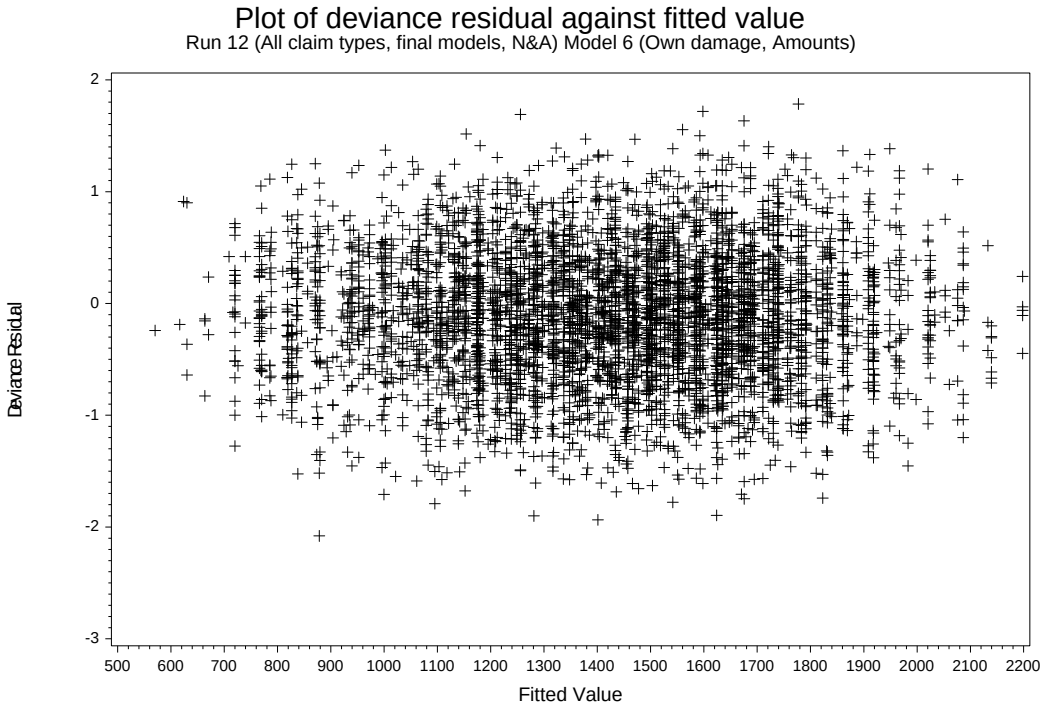
2.57 The leverage h_i is a measure of how much influence an observation has over its own fitted value. Its formal definition is complex but essentially it is a measure of how much a change in an observation affects the fitted value for that observation. Leverage always lies strictly between 0 and 1. A leverage close to 1 means that if the observation was changed by a small amount the fitted value would move by almost the same amount. Where the leverage is close to 1 it is likely that the residual for that observation will be unusually small because of the high influence the observation has on its fitted value. Dividing by the square root of one minus the leverage corrects for this by increasing the residual by an appropriate amount.

2.58 Another commonly used measure of the residual is the Standardized Pearson residual. This is the raw residual adjusted for the expected variance and leverage (as described above):

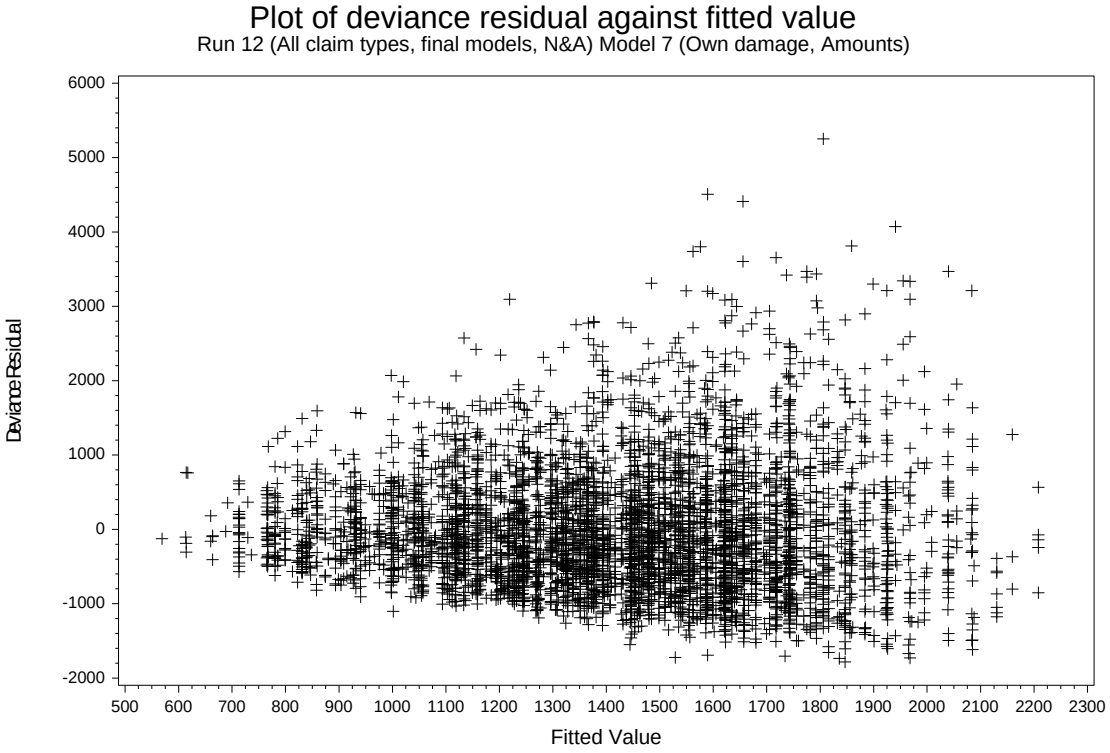
$$r_i^{PS} = \frac{(Y_i - \mu_i)}{\sqrt{\frac{\phi}{\omega_i} V(\mu_i)(1-h_i)}}$$

2.59 This adjustment makes observations with different means comparable, but does not adjust for the shape of the distribution.

2.60 Observing scatter plots of residuals against fitted values can give an indication of the appropriateness of the error function which has been assumed. For example, if the model form is appropriate then the standardized deviance residuals should be distributed Normal (0,1) regardless of the fitted value. The example scatter plot below shows the result of fitting a GLM with a gamma variance function to data which has been randomly generated on a hypothetical insurance dataset from a gamma distribution (with a mean based on assumed factor effects). It can be seen that moving from the left to the right of the graph the general mean and variability of the deviance residuals is reasonably constant, suggesting (as is known to be the case in this artificial example) that the assumed variance function is appropriate.

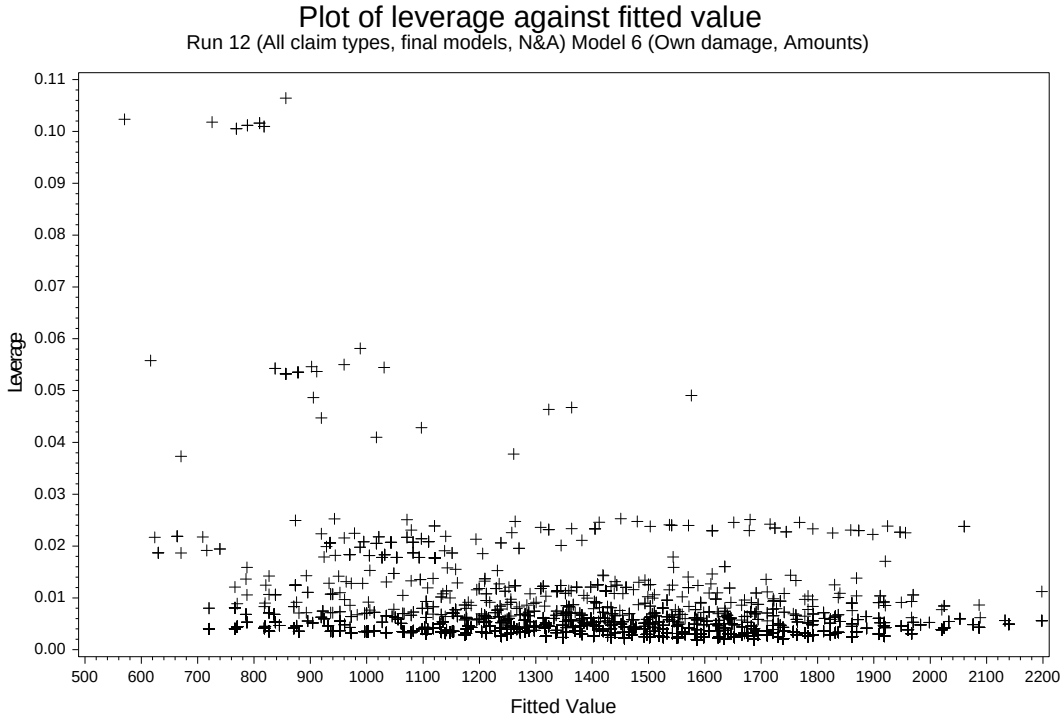


2.61 Conversely the graph below shows the scaled deviance residuals obtained from fitting a GLM with an assumed Normal error to the same gamma data. In this case the variability increases with fitted value, indicating that an inappropriate error function has been selected and that the variance of the observations increases with the fitted values to a greater extent than has been assumed. This could occur, for example, when a Normal model is fitted to Poisson data, when a Poisson model is fitted to gamma data, or (as is the case here) where a Normal model is fitted to gamma data.



Leverage

2.62 As well as being needed to calculate standardized residuals, the leverage statistic is also a helpful diagnostic in its own right, since it can identify particular observations which might have an undue influence on the model. For example the graph below shows a scatter plot of leverage against fitted value. In this case seven particular observations have clearly higher leverage than other observations (around 0.1) and it is possible that they are having an undue influence on the model. An inspection of these observations may indicate whether or not it is appropriate to retain them in the model.



Box Cox transformation and the case for a multiplicative model

2.63 The Box Cox transformation can be used to assess the appropriateness of the assumed link function. The transformation defines the following link function in terms of a scalar parameter λ:

$$g(x) = \begin{cases} \dfrac{(x^{\lambda} - 1)}{\lambda}, & \lambda \neq 0 \\ \ln(x), & \lambda = 0 \end{cases}$$

2.64 If λ=1, g(x) = x-1. This is equivalent to an identity link function (ie an additive model) with a base level shift.

2.65 As $\lambda \to 0$, $g(x) \to \ln(x)$ [10]

$$\lim_{\lambda \to 0} \frac{x^{\lambda} - 1}{\lambda} = \lim_{\lambda \to 0} \left[\frac{\frac{\mathrm{d}}{\mathrm{d}\lambda}(\exp(\lambda \ln(x)) - 1)}{\frac{\mathrm{d}}{\mathrm{d}\lambda}\lambda} \right] = \lim_{\lambda \to 0} \frac{\ln(x).x^{\lambda}}{1} = \ln(x)$$

This is equivalent to a multiplicative model.

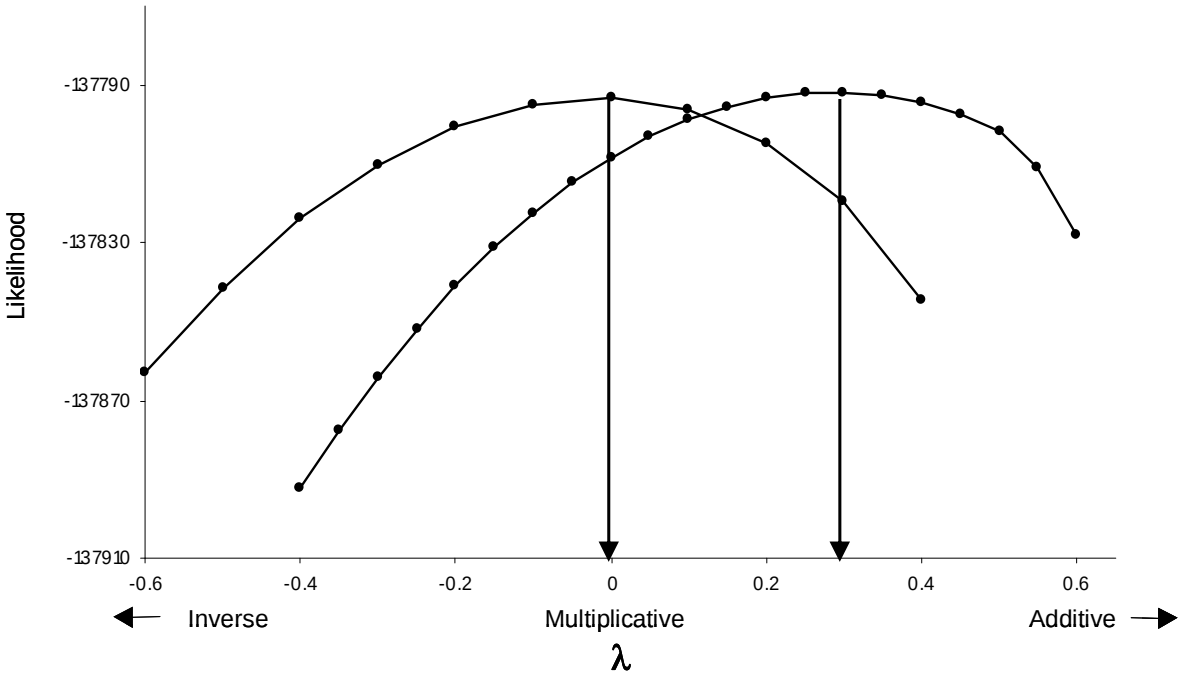
2.66 If λ=-1, $g(x) = 1-x^{-1}$. This is equivalent to an inverse link function with a base level shift.

2.67 By fitting a series of GLMs to the data with many different values of λ (including real values between -1 & 0 and 0 & 1), and with all other model features identical in every other respect, it is possible to assess which value of λ is most appropriate for the dataset in question by seeing which value of λ yields the highest likelihood. Optimal values of λ around 0 would suggest that a multiplicative structure with a log link function would be the most appropriate for the data in question, whereas optimal values of λ around 1 would suggest an additive structure would be best, with values around -1 indicating that an inverse link function would be most appropriate.

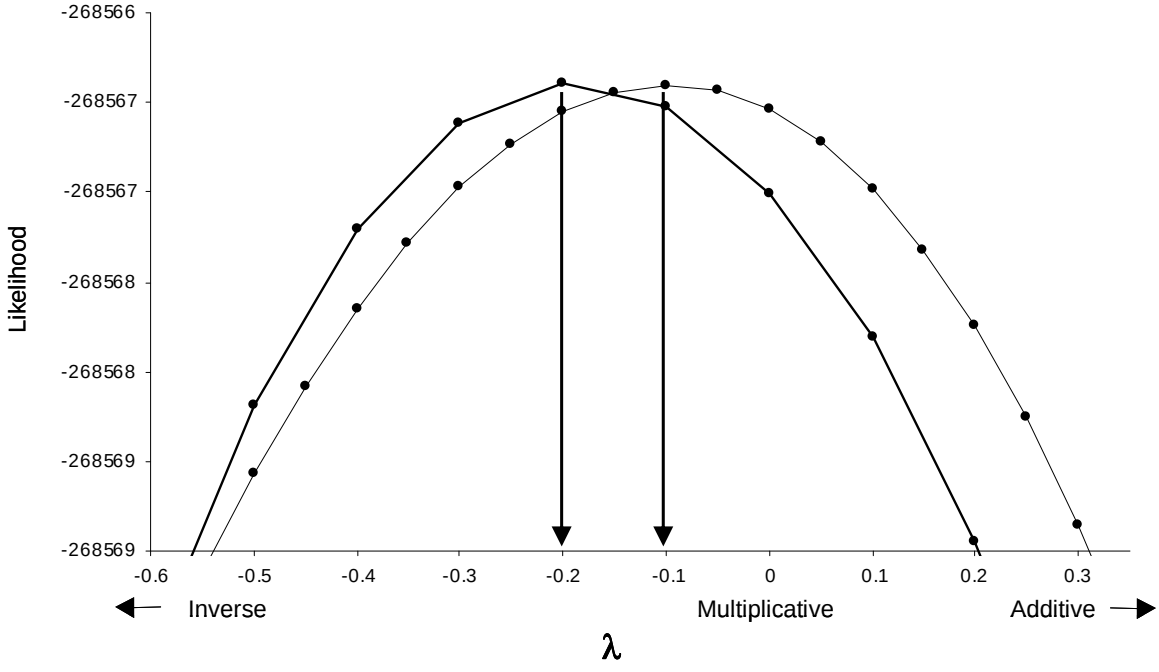
2.68 Examples based on two real datasets are shown below.

[10] Via L'Hôpital's Rule.

Box Cox transformation results on frequency



Box Cox transformation results on severity

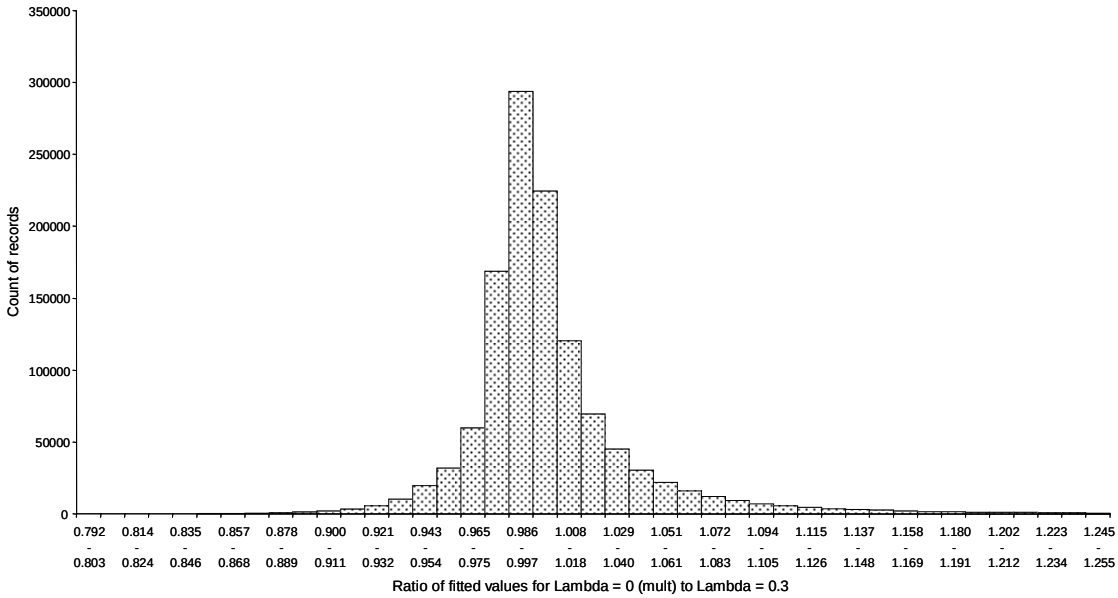


2.69 The first graph shows various values of λ tested on two different datasets containing private passenger automobile property damage liability frequency experience. The optimal λ in one case is very close to zero (suggesting a multiplicative model) but in the other is around 0.3, suggesting that the frequency in that case is largely influenced by explanatory variables in a multiplicative fashion, but to an extent also in an additive fashion.

2.70 The second graph shows the results for claim amounts models for the same data. Here the optimal values of λ are near zero (multiplicative), but this time slightly toward the direction of being partly inverse.

2.71 In order to understand how significant the value of λ is upon the fitted values produced by the model it is helpful to consider the histogram graph below which shows, for one of the two frequency datasets considered above, the distribution of the ratio of fitted values produced by a GLM with λ=0 to an otherwise identical GLM with λ=0.3. It can be seen that there is in fact little difference between the fitted values produced by these two models, with the great majority of fitted values being within 2 or 3% of each other.

Distribution of ratio of fitted values between model with λ = 0 and model with λ = 0.3



2.72 In practice there are many significant advantages with using a multiplicative structure, not least because it is easy to understand. In the above examples it seems that there is no strong evidence to use a structure other than a multiplicative structure.

2.73 While this should be tested in each case, it is often the case that multiplicative structures and log link functions are the most appropriate practical model for modeling insurance risk, and this may explain the high prevalence of multiplicative rating structures, especially in Europe where GLMs have been in use for many years.

Model refinement

Interactions

2.74 Thus far, the discussion has focused on the independent effect of factors in the model. Generalized linear models can also consider the interaction between two or more factors. Interactions occur when the effect of one factor varies according to the level of another factor.

2.75 Interactions relate to the effect which factors have upon the risk, and are not related to the correlation in exposure between two factors. This is illustrated with two examples which consider two rating factors in a multiplicative rating structure.

Example 1 - correlation but no interaction

Earned exposure

	Town	Countryside	Total
Male	200	100	300
Female	100	200	300
Total	300	300	600

Number of claims

	Town	Countryside	Total
Male	80	20	100
Female	20	20	40
Total	100	40	140

Claims frequency

	Town	Countryside
Male	40%	20%
Female	20%	10%

2.76 In this example the exposure is not distributed evenly amongst the different rating cells - a higher proportion of town dwellers are male than is the case in the countryside. The *effect* of the two factors upon the risk, however, does not in this example depend on each other - men are twice the risk of women (regardless of location) and town dwellers are twice the risk of countryside dwellers (regardless of gender). In this example there is therefore a correlation between the two rating factors, but no interaction.

Example 2 - interaction but no correlation

Earned exposure

	Town	Countryside	Total
Male	300	150	450
Female	200	100	300
Total	500	250	750

Number of claims

	Town	Countryside	Total
Male	180	30	210
Female	40	10	50
Total	220	40	260

Claims frequency

	Town	Countryside
Male	60%	20%
Female	20%	10%

2.77 In this example the exposure is distributed evenly amongst the different rating cells - the same proportion of town dwellers are males as are countryside dwellers. The effect of the two factors upon the risk, however, in this example depend on each other - it is not possible to represent accurately the effect of being male (compared with being female) in terms of a single multiplier, nor can the effect of location be represented by a single multiplier. To reflect the situation accurately it is necessary in this case to consider multipliers dependent on the combined levels of gender and location.

2.78 An interaction term can be included within a GLM simply by defining an explanatory variable in terms of two or more explanatory variables. In the above example, rather than declaring location and gender as two explanatory variables (each with a base level and one parameter), a combined "gender-location" variable could be declared with four levels (a base level and three parameters).

2.79 Interaction terms should only be included where there is statistical justification for the inclusion of the additional parameters. In the above example the interaction term only involved the addition of one further parameter to the model, but if an interaction is introduced between two ten level factors (each with a base level and nine parameters), a further 81 parameters would be introduced into the model.

"Complete" and "marginal" interactions

2.80 Interactions can be expressed in different ways. For example consider the case of two factors each with four levels. One way of expressing an interaction is to consider a single factor representing every combination of the two factors (or "complete" interaction). A set of multipliers (in the case of a multiplicative model) could therefore be expressed as follows:

		Factor 1: A	B	C	D
Factor 2:	W	0.72	0.80	0.88	0.96
	X	0.90	1.00	1.10	1.20
	Y	0.97	1.20	1.45	1.66
	Z	1.26	1.40	1.85	2.10

2.81 In this case the base level has been selected to be the level corresponding to level B of factor 1 and level X of factor 2, and the interaction term has 15 parameters.

2.82 An alternative representation of this interaction is to consider the single factor effects of factor 1 and factor 2 *and* the additional effect of an interaction term over and above the single factor effects (or "marginal" interaction). A set of multipliers in this form can be set as follows:

		Factor 1:	A	B	C	D
			0.90	-	1.10	1.20
Factor 2:	W	0.80	1	-	1	1
	X	-	-	-	-	-
	Y	1.20	0.9	-	1.1	1.15
	Z	1.40	1	-	1.2	1.25

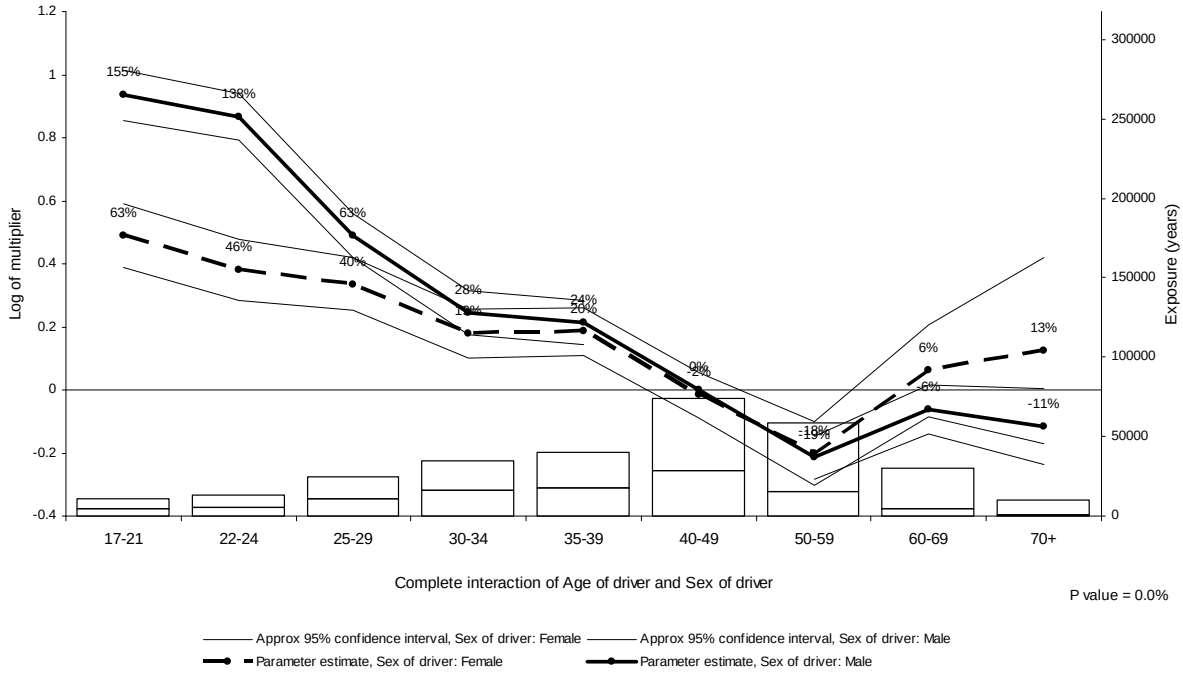
2.83 In this case fewer parameters are present in the additional interaction term because the presence of the single factor effects makes some of the interaction terms redundant. When fitted in a GLM (assuming that the single factor effects were declared first) the redundant terms in the additional interaction term would be aliased. Overall the three terms combined still have 15 parameters, and result in identical predicted values (for example in the case of factor 1 level D and factor 2 level Z, 1.2 * 1.4 * 1.25 = 2.1).

2.84 In practice sometimes it can be helpful to consider the "complete" interaction (ie just a single factor representation of all combinations of the two factors) and sometimes it can be helpful to consider the additional or "marginal" interaction term over and above the single factor effects. While the fitted values from both approaches are identical, what differs is the statistical diagnostics available in the form of parameter estimate standard errors.

Example

2.85 For example, the graph below shows the result of a "complete" interaction between the age of driver and the gender of driver for the claims frequency of a certain type of auto claim, with age relativities for men and women superimposed (with solid and dotted lines respectively) on the same x-axis.

Complete interaction

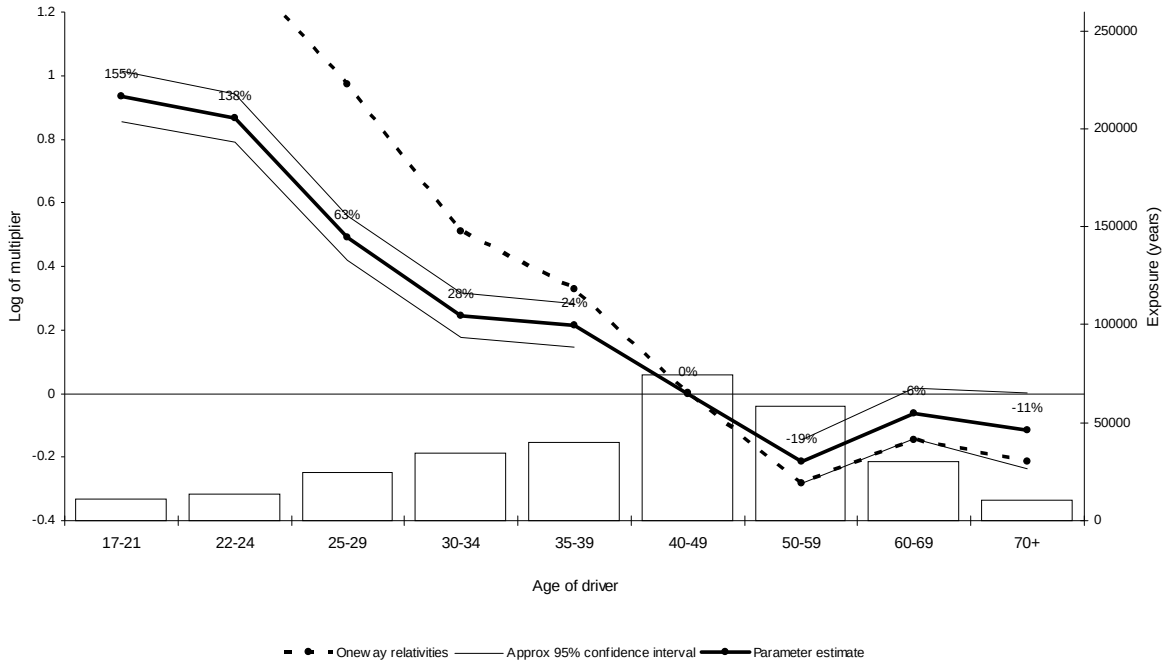


2.86 If there were no significant interaction between these two factors the solid and dotted lines (showing parameters from a log link GLM) would be parallel. In this example they are clearly not parallel, showing that while younger drivers have higher frequencies, and while in general male drivers have higher frequencies, in this example (as in many real cases) young men experience a higher frequency than would be predicted by the average independent effects of the two factors.

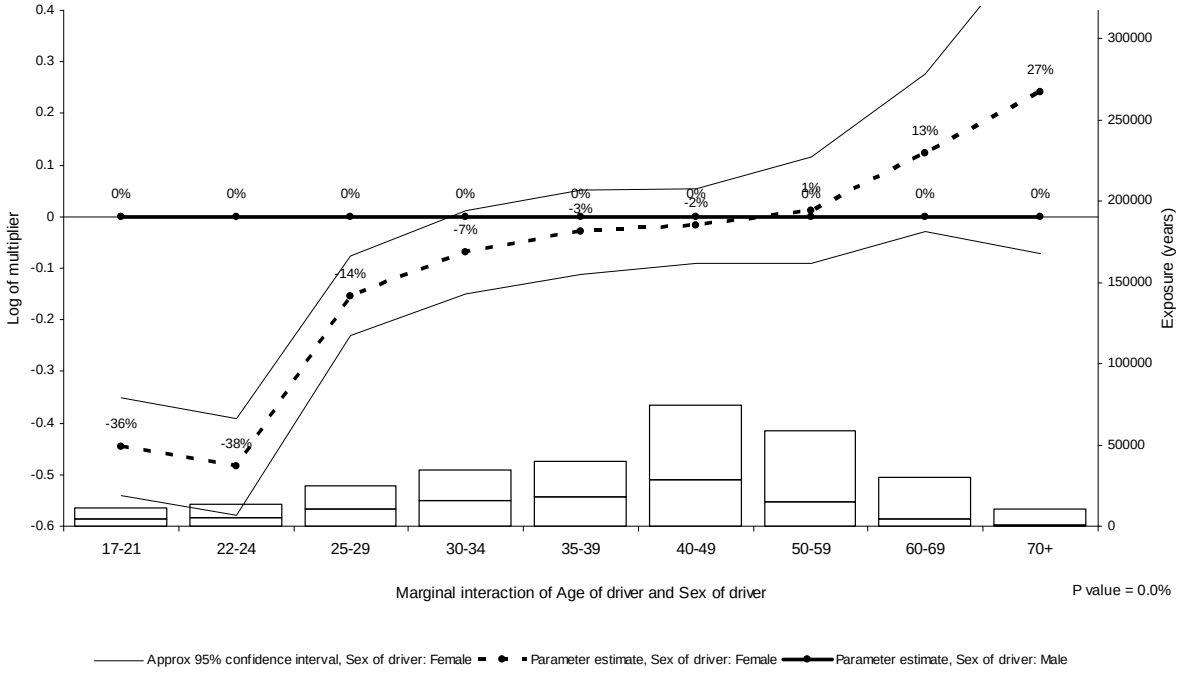
2.87 The narrow standard error bands around the parameter estimate lines suggest the likely statistical significance of the result, however they do not provide any sound theoretical basis for assessing the significance of the factor. A more theoretically appropriate test can be applied if a marginal interaction is considered.

2.88 The graphs below show the results of fitting age and then a marginal interaction of age and sex to the same data.

Main effect



Partial marginal interaction



2.89 The first graph shows the single factor effect for age, and the second shows the marginal interaction term over and above this single factor effect. (In this case the single factor gender of the driver was not included since it proved not to be significant.)

2.90 Since the male level of the gender factor is ordered after the female level, the male levels of the marginal factor have been aliased, with the result that the first graph represents the age effects for males, and the marginal graph shows the additional adjustment which is appropriate for females of different ages.

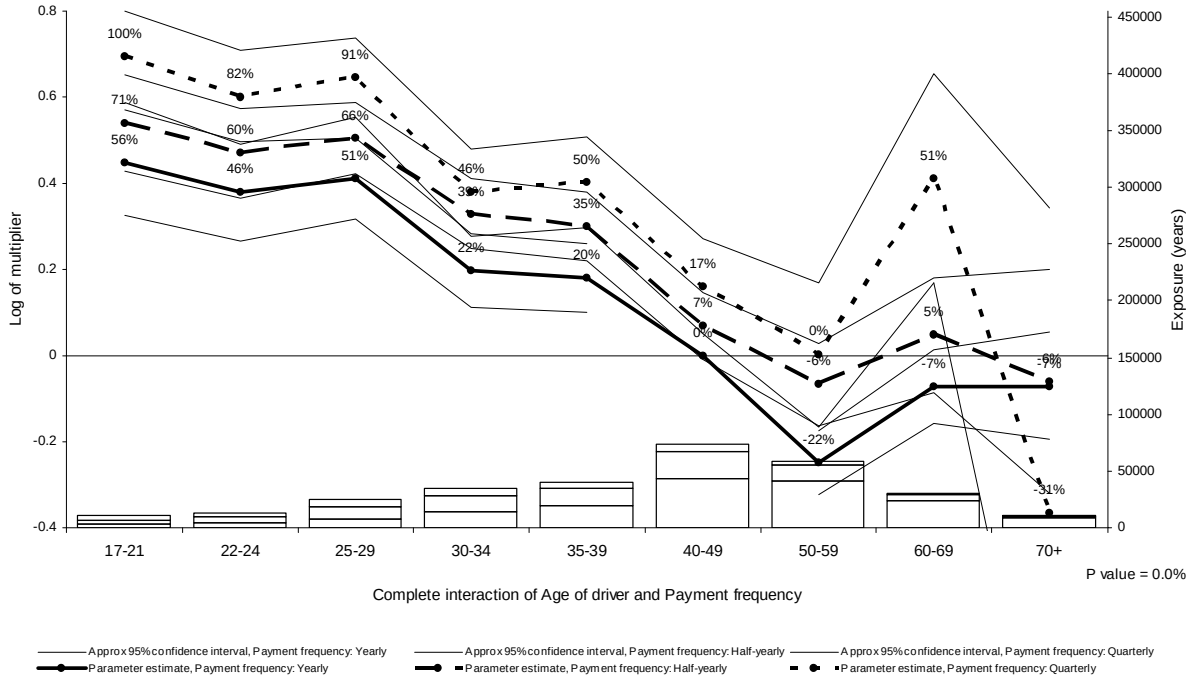
2.91 The implied fitted values from the marginal interaction are the same as the complete interaction - for example:

- Complete interaction effect for age 22-24 female = +46% or multiplier of 1.46
- Marginal approach:
  - Single factor age effect for 22-24 = +138% or multiplier of 2.38
  - Marginal effect of women relative to men at age 22-24 = -38% or multiplier of 0.62
  - Combined effect for age 22-24 female = 2.38 x 0.62 = 1.48 (differences due to rounding)

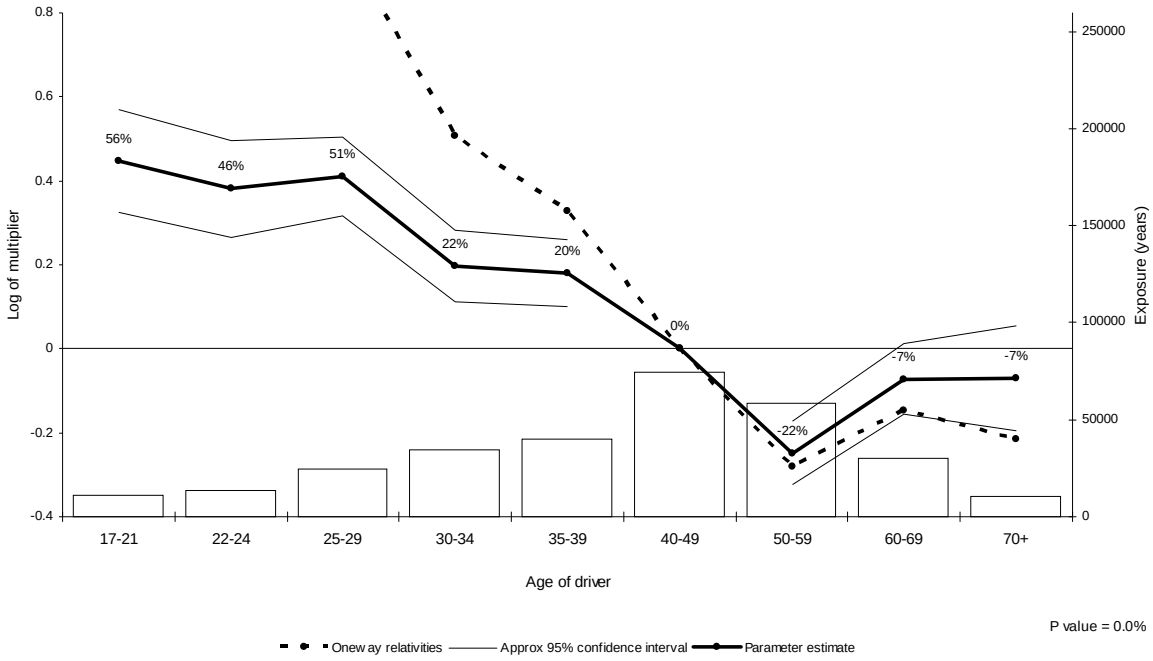
2.92 The marginal approach does however provide more meaningful diagnostics in the form of parameter estimate standard errors and type III tests. The standard errors on the graph of the marginal term indicate that the marginal term is indeed significant, and the Type III P-value of 0.00% for this factor confirms that this is the case.

2.93 An example of an interaction term which is not significant is shown below. The first graph is the complete interaction (where the parameter estimate lines can be seen to be largely parallel). The second and third graphs show the main effects of age of driver and payment frequency, respectively. The fourth graph shows the marginal interaction (where the marginal interaction term can be seen to be insignificant, both visually and because of the type III p-value of 61.9%).

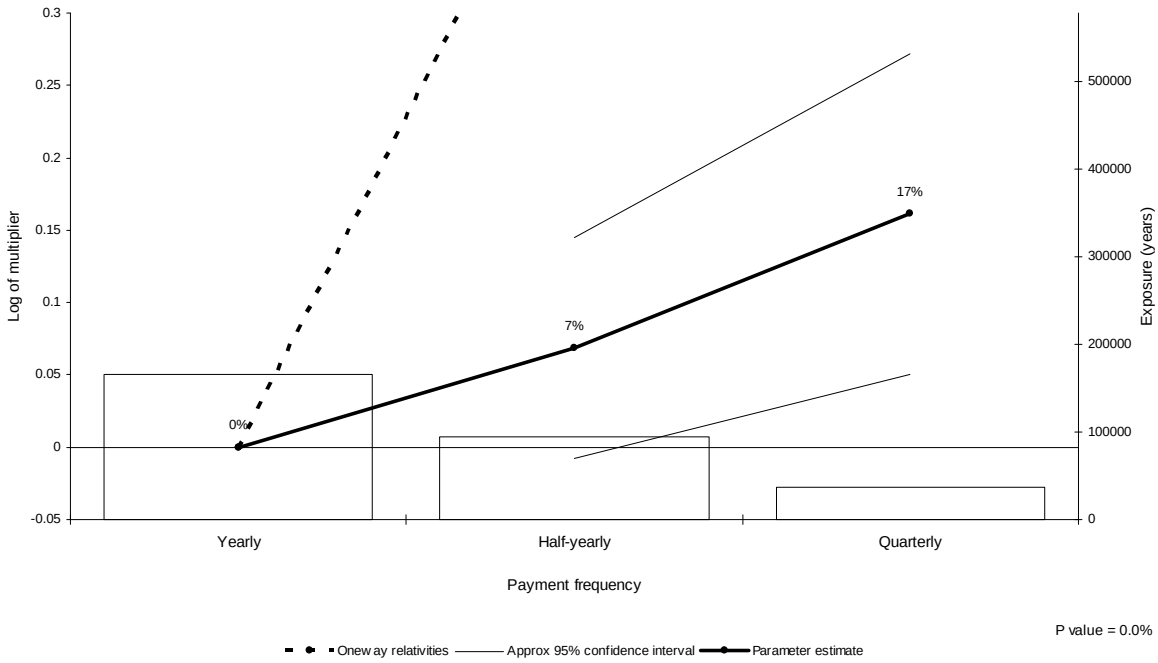
Complete interaction



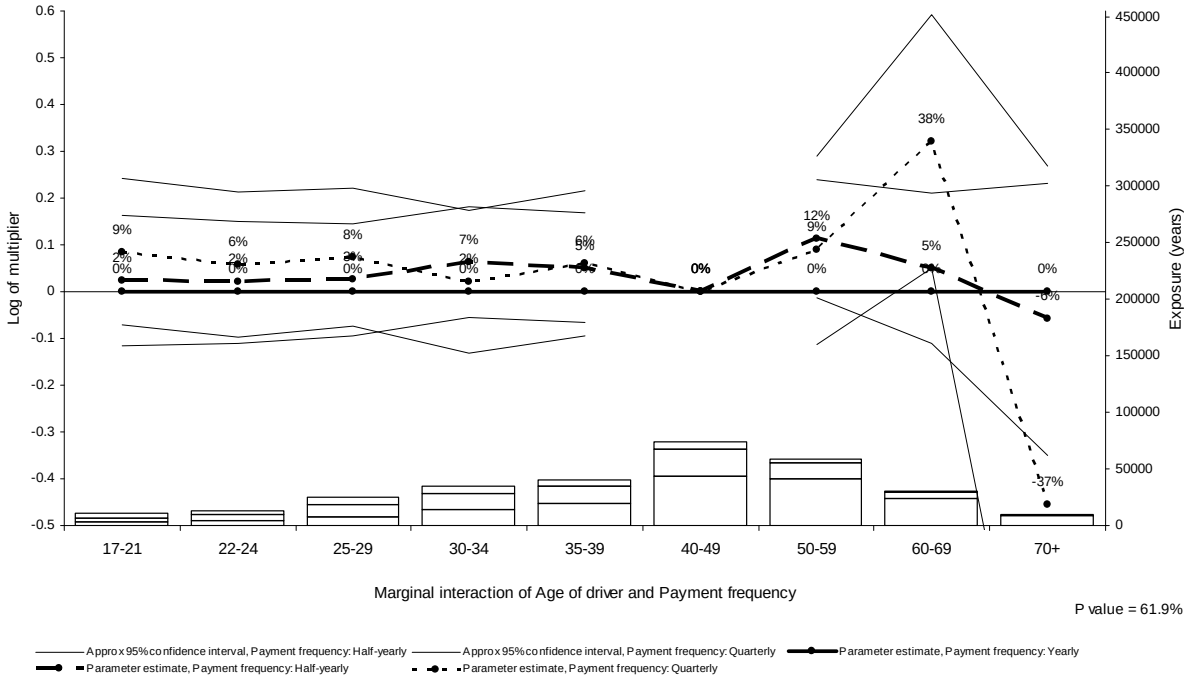
Main effect



Main effect



Marginal interaction



Interpreting marginal interactions

2.94 Although the marginal form of an interaction provides a more sound theoretical basis for assessing the significance of a factor, in practice marginal interactions can be hard to interpret. For example consider the example of two factors each with four levels. It might be the case that the true underlying frequency (all other factors being at a certain level) was as follows:

	Factor 1:	**A**	**B**	**C**	**D**
Factor 2:	**W**	7.2%	8.0%	8.8%	9.6%
	X	9.0%	10.0%	11.0%	12.0%
	Y	9.7%	12.0%	14.5%	16.6%
	Z	12.6%	14.0%	18.5%	21.0%

2.95 However in reality the exposure available for this analysis might be low for some combinations of these two factors, for example:

Exposure	Factor 1:	A	B	C	D
Factor 2:	W	1000	1000	1000	1000
	X	1000	1000	1000	1
	Y	1000	1000	1000	1000
	Z	1000	1000	1000	1000

2.96 If in general the claims experience was in line with the underlying frequencies but the one policy with factor 1 level D and factor 2 level X had one claim (resulting in a very high claims frequency of 100%), a marginal interaction would yield results which could be hard to interpret. Specifically if a marginal interaction were fitted, the GLM would seek the following parameters:

	Factor 1:	A	B	C	D
		β_1	-	β_2	β_3
Factor 2:	W	β_4	β_7	-	β_8	β_9
	X	-	-	-	-	-
	Y	β_5	β_{10}	-	β_{11}	β_{12}
	Z	β_6	β_{13}	-	β_{14}	β_{15}

2.97 Since level X is the base level of factor 2, there is no single term in the marginal interaction which can represent the very high observed frequency for factor 1 level D / factor 2 level X. Instead the model will yield parameter β_3 with a very high value, and parameters β_9, β_{12} and β_{15} with low values. Although theoretically correct, the parameter estimates and standard errors for parameters β_9, β_{12} and β_{15} would be hard to interpret.

Searching for interactions

2.98 In general the significance of an interaction can be assessed by considering

- the standard errors of the parameter estimates of the marginal term
- the type III P-value of the marginal term
- general intuition given the overall "complete" interaction effect
- the consistency of an interaction over time.

2.99 In theory all possible combinations of pairs or triplets of factors could be tested as interactions one at a time in a model. In practice, the design of the current rating plan, the results of two-way analyses and wider experience will influence the choice of what is tested, as will the ease of interpretation and the ultimate application of the model.

2.100 In some cases rather than considering every combination of two factors with many levels it can be appropriate to consider only the strongest effects. For example, a marginal interaction of driver age, car symbol and the interaction of driver age - car symbol (denoted driver age*car symbol) may highlight an interesting effect in one "corner" of the interaction (eg young drivers driving high car symbols). In practice, the interaction may be re-parameterized as a combination of detailed single factors for age of driver and car symbol, and an additional less detailed factor based on the combination of age of driver and car symbol which has the same level for many combinations, and a few levels representing certain combinations of young drivers driving high car symbols.

2.101 The inclusion of several meaningful interactions which share factors (eg age*sex, age*multi-car and territory*multi-car) could provide a theoretically correct model but may be very difficult to interpret. The practitioner may consider creating separate models for single and multi-car, and continue to investigate other interactions.

Smoothing

2.102 Once models have been iterated to include only significant effects and interactions have been investigated, smoothing of the parameter estimates may be considered in order to improve the predictive power of the model. Much like the offset and prior weight terms in the formularization of GLMs, smoothing is used to incorporate some element of the practitioner's knowledge into the model. In this sense, the practitioner may impart knowledge that some factors have a natural order (eg that age of car seven should fall between age of car six and age of car eight). Outliers may also be tempered. This tempering is not based on commercial selections at this point (ie tolerance for rate change) but rather an attempt to adjust an anomaly once a proper investigation has been done to ensure that the outlier is truly an anomaly and not something systematic in the experience.

2.103 The selection of smoothed parameter estimates can be done in an unscientific fashion (for example - a visual modification to a curve) or in a more scientific fashion (for example - fitting polynomials to the observed parameter estimates, or electing to refit a model using polynomial terms as variates within the GLM). If smoothing is rather severe, the practitioner may consider restricting the values of the smoothed factor and re-running a model to allow other factors to compensate. (The concept of restrictions is discussed later in this paper.) In general, however, this technique may only remove the random element from one factor and move it to another factor, and often it can be preferable not to refit using restrictions in this way.

Risk Premium

2.104 Fitting GLMs separately to frequency and severity experience can provide a better understanding of the way in which factors affect the cost of claims. This more easily allows the identification and removal (via smoothing) of certain random effects from one element of the experience. Ultimately, however, these underlying models generally need to be combined to give an indication of loss cost, or "risk premium", relativities.[11]

2.105 In the case of multiplicative models for a single claim type, the calculation is straightforward - the frequency multipliers for each factor can simply be multiplied by the severity multipliers for the same factors (which is analogous to adding the parameter estimates when using a log link function). Alternatively, models may be fitted directly to pure premium data using the Tweedie distribution (discussed in Appendix C). The advantages and disadvantages of this alternative approach are discussed in Appendix J.

2.106 Certain market conditions may warrant the development of a single theoretical risk premium model, even if different types of claim have been modeled separately. An example is the aggregation of homeowners models by peril into a single rating algorithm at point of sale. The derivation of a single model in this situation is not as straightforward since there is no direct way of combining the model results for the underlying claim types into a single overall expected cost of claims model. In this situation, however, it is possible to approximate the overall effect of rating factors on the total cost of claims by using a further GLM to calculate a weighted average of the GLMs for each of the underlying frequency and severity models for each of the claim types. Specifically this can be done by

- selecting a dataset which most accurately reflects the likely future mix of business
- calculating an expected claim frequency and severity by claim type for each record in the data
- combining these fitted values, for each record, to derive the expected cost of claims (according to the individual GLMs) for each record
- fitting a further generalized linear model to this total expected cost of claims, with this final GLM containing the union of all factors (and interactions) in all of the underlying models.

[11] The term "risk premium" is used rather than pure premium in order to differentiate between a model fitted directly on pure premium data and a model derived by combining underlying frequency and severity models.

2.107 An illustrative example is shown below. The top table represents the intercepts and multipliers from underlying frequency and severity models for claim types 1 and 2. The bottom table shows the calculation of the total risk premium, based on the underlying models, for the first four records in the data. The additional GLM is fitted to this last column in this dataset in order have a single theoretical risk premium model.

		Claim type 1		**Claim type 2**	
		Frequency	**Severity**	**Frequency**	**Severity**
Intercept		0.32	1,000	0.12	4,860
Sex:	**Male**	1.00	1.00	1.00	1.00
	Female	0.75	1.20	0.67	0.90
Area:	**Town**	1.00	1.00	1.00	1.00
	Country	1.25	0.72	0.75	0.83

Policy	**Sex**	**Area**	**Fitted freq 1**	**Fitted sev 1**	**Fitted RP 1**	**Fitted freq 2**	**Fitted sev 2**	**Fitted RP 2**	**Total RP**
82155654	M	T	32.00%	1,000.00	320.00	12.00%	4,860.00	583.20	**903.20**
82168746	F	T	24.00%	1,200.00	288.00	8.04%	4,374.00	351.67	**639.67**
82179481	M	C	40.00%	720.00	288.00	9.00%	4,033.80	363.04	**651.04**
82186845	F	C	30.00%	864.00	259.20	6.03%	3,630.42	218.91	**478.11**
……	…….	…….	…….	…….	…….	…….	…….	…….	…….

2.108 In addition to combining frequency and severity across multiple claim types, the technique of fitting an overall GLM to fitted values of other GLMs can be used to incorporate non-proportional expense elements into the modeled relativities. For example, a constant dollar amount could be added to each observation's expected risk premium and then a GLM re-fitted to this new field. The resulting "flattened" risk premium relativities will prevent high risk factor levels from being excessively loaded for expenses.

2.109 Alternatively, the amount added to each observation's expected risk premium could be designed to vary according to the results of a separate retention study. This would allow risks with a high propensity to lapse to receive a higher proportion of fixed expense than those risks with a low propensity to lapse. As above, a further GLM is fitted to the sum of the expected risk premium and a (lapse-dependent) expense load.

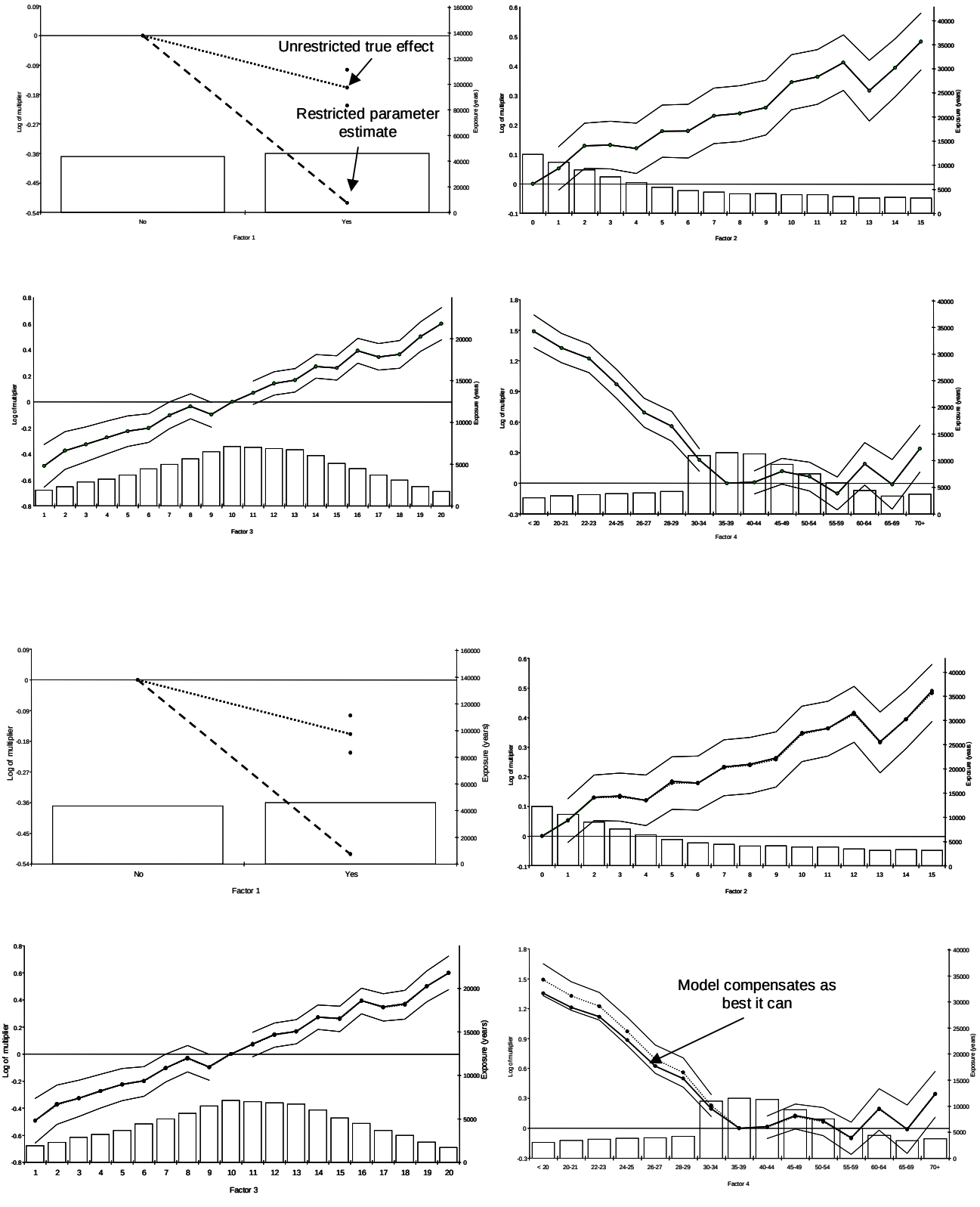
Restrictions

2.110 The theoretical risk premium results from a GLM claims analysis will differ from the rates implemented in practice since consideration needs to be given to price demand elasticity and the competitive situation. There are, however, some situations where legal or commercial considerations may also impose rigid restrictions on the way particular factors are used in practice. When the use of certain factors is restricted, if desired the model may be able to compensate to an extent for this artificial restriction by adjusting the fitted relativities for correlated factors. This is achieved using the offset term in the GLM.

2.111 Specifically, the required parameter estimates (logs of multipliers in the case of a multiplicative model) are calculated for each record and added to the offset term ξ_i. The factor in question is then not included as an explanatory factor in the GLM. (This can intuitively be thought of as fixing some selected elements of β_i to be specified values.)

2.112 The graphs below illustrate the use of a restriction. In the upper series of graphs, the dotted lines display the theoretically correct parameter estimates indicated by a GLM containing these four rating factors. The dashed line in Factor 1 shows the intended restriction. In the lower series of graphs, the solid lines show the output of the GLM after the restriction for Factor 1 has been incorporated and Factors 2, 3, 4 have been allowed to compensate. It can be seen that the parameter estimates in Factors 2 and 3 have hardly changed, suggesting little correlation between these factors and Factor 1. On the other hand the solid line in Factor 4 has moved away from the theoretically correct dotted line, suggesting a correlation between the restricted levels in Factor 1 and those levels in Factor 4 which moved to compensate for the restriction.

Example of restricting a factor



2.113 Although restrictions could be applied either to frequency or amounts models (or in part to both), generally it is more appropriate to impose the restriction on the model at the risk premium stage since this allows a more complete and balanced compensation by the other factors. This can be achieved by calculating the expected cost of claims for each record, according to "unrestricted" GLMs, and then imposing the restriction in the final GLM which is then fitted to the total expected cost of claims. (For restricted risk premium models this approach is necessary even in the case of a single claim type.)

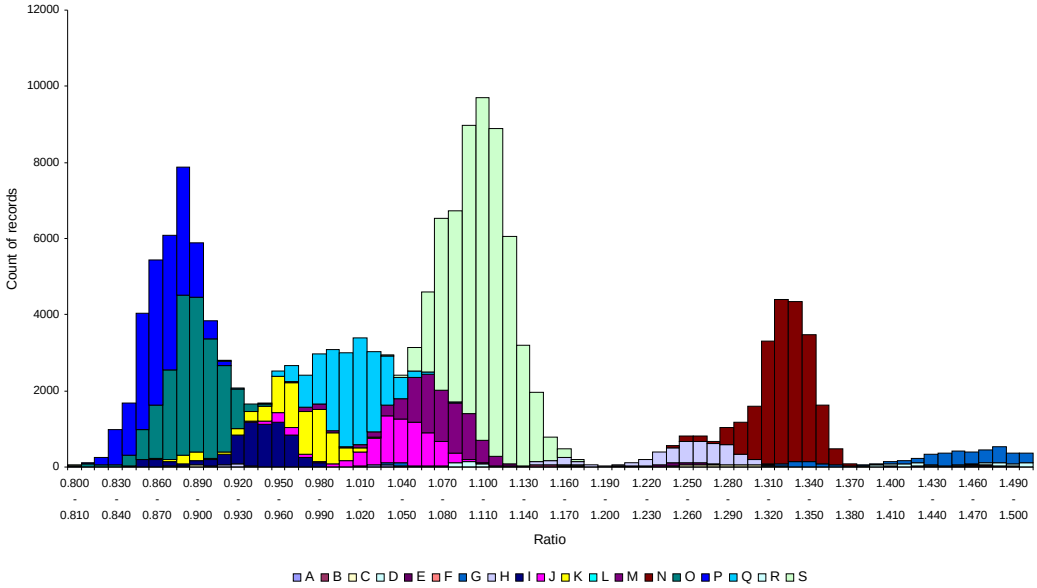
2.114 In the US, many personal lines rating plans contain discounts that were initially implemented for marketing appeal or perhaps mandated by regulation. Today's models may indicate that these discounts are not supported by the claims experience - or in many cases may even indicate a surcharge. If a company chooses to continue offering such discounts, it is important that these restrictions are incorporated into the modeling process since such restrictions can affect the relativities which become appropriate for other correlated factors. Counterintuitive model results may occur on behavioral factors such as factors which policyholders self-select, for example limits and deductibles. These factors may require restriction if they are to be used directly in ratemaking.

2.115 Model restrictions are also used in US ratemaking to mitigate the number of factors which will change in a given rate review. Companies may restrict certain existing rating factors and allow the GLM to measure only the effect of new rating factors. Restrictions may also come into play when applying the results of a countrywide model to a particular state.

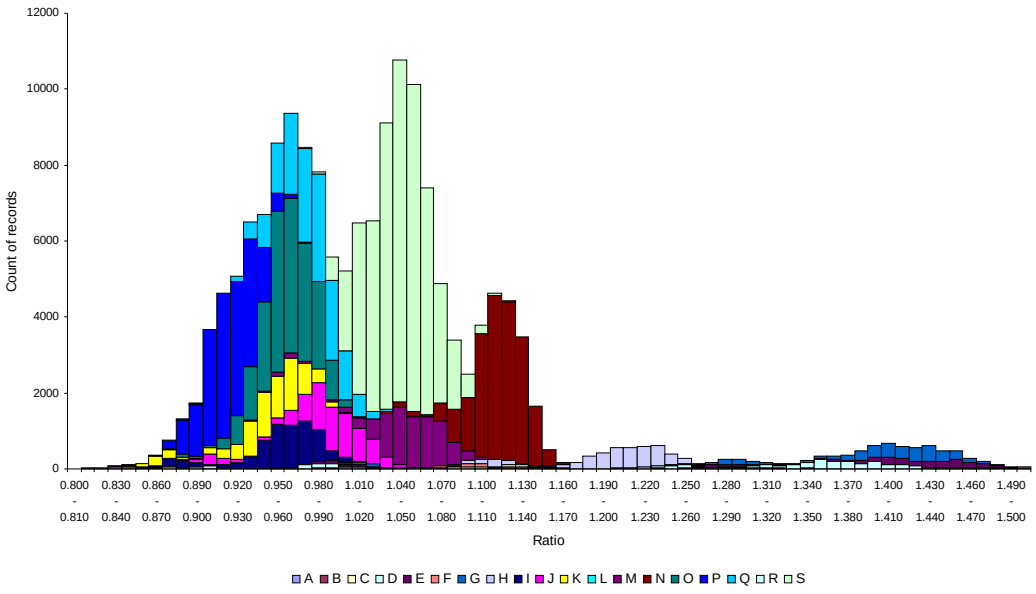
2.116 Prior to incorporating restrictions, it is still important to assess the true effect of all factors upon the risk by initially including them in the analysis as if they were ordinary factors. In addition, a comparison of the fitted values of the theoretical model and the restricted models will demonstrate the degree to which other factors have compensated for the restriction. The examples below show two such comparisons. Each graph shows the number of policies (on the y-axis) that have different ratios of restricted to unrestricted fitted values (on the x-axis). The graph is subdivided by levels of the restricted factor (shown in different shading). If the GLM can compensate well for a factor restriction (because there are many other factors in the model correlated with the restricted factor) then this distribution will be narrow. Conversely if the GLM cannot compensate well for the restriction, this distribution will be wider.

2.117 In this particular example the factors in the upper graph have not compensated well for the restriction. The wide distribution of the restricted to unrestricted ratio implies that the restriction is moving the model away from the theoretical result. The lower graph, on the other hand, shows a model which contains factors that are more correlated with the restricted factor, and which have compensated better for the restriction.

Distribution of ratio of fitted values between restricted and unrestricted models (showing little compensation from other factors)



Distribution of ratio of fitted values between restricted and unrestricted models (showing some compensation from other factors)



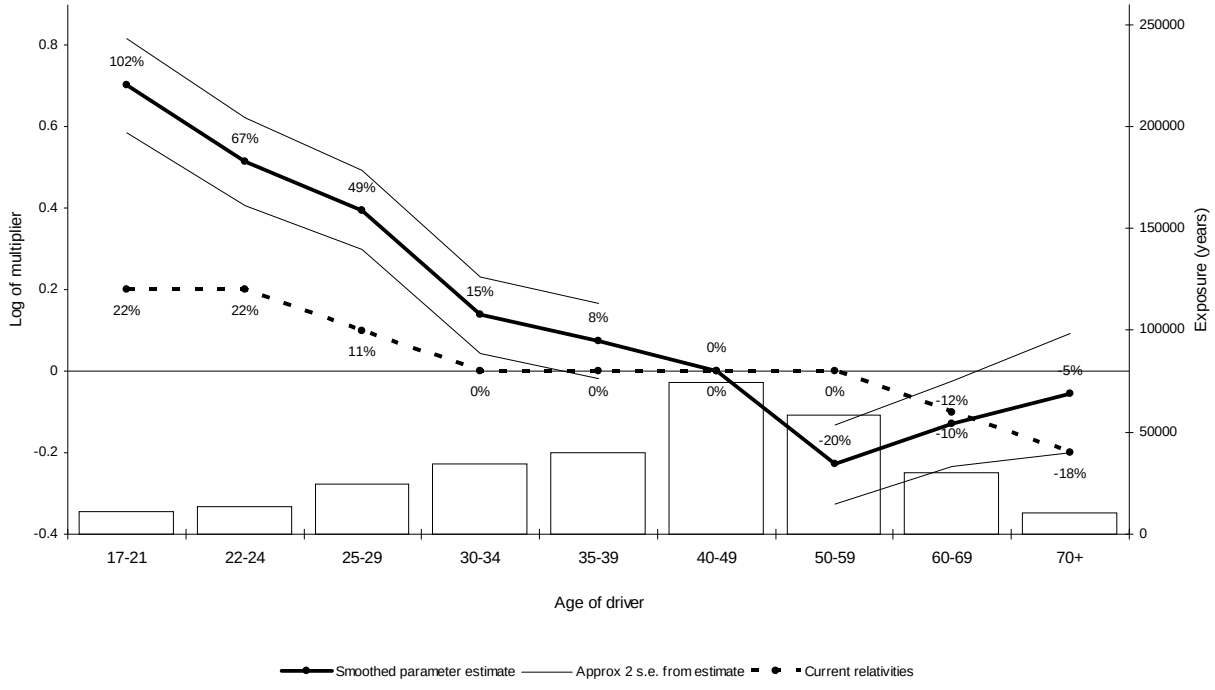
Interpreting the results

2.118 To understand how the results of a GLM claims model differ from the existing rating relativities it is helpful to consider the results both on a factor-by-factor basis and also by measuring the overall effect of all factor differences combined.

Comparing GLM indicated relativities to current relativities

2.119 The final risk premium models can be plotted on graphs similar to those shown in previous sections. Another line can be added to display the relativities implicit in the current rating structure. This allows easy comparison of the relativities indicated by the model and those which are currently used. An example graph is shown below. In this example it can be seen that the current relativities for young drivers (shown as a dotted line) are too low.

Final risk premium model compared to current relativities

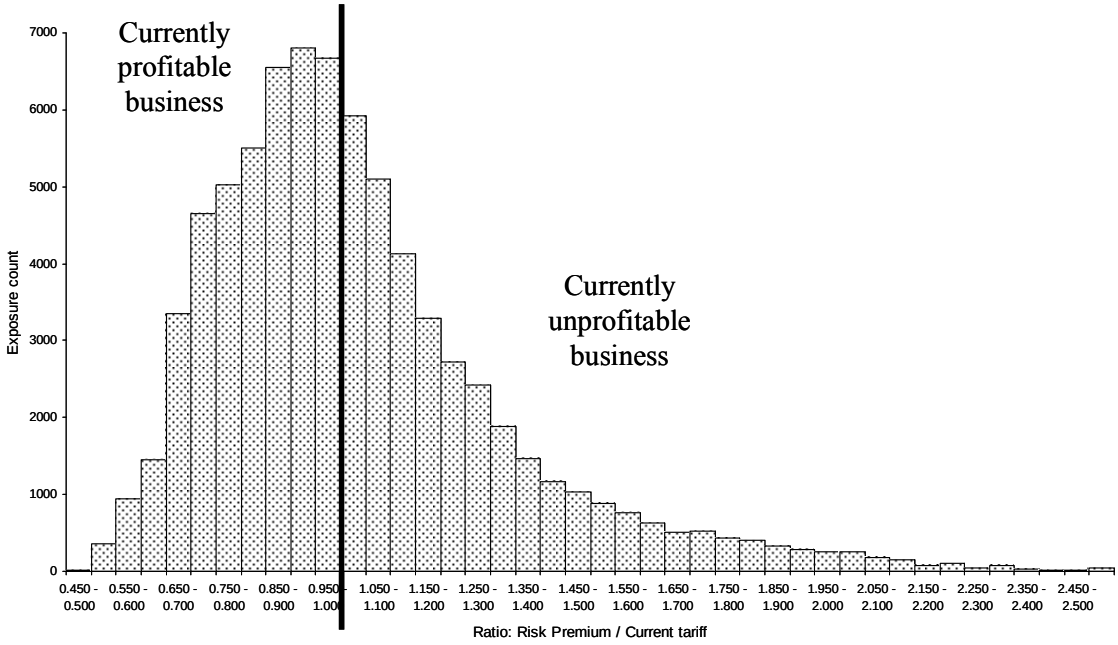


2.120 If the existing rating plan is purely multiplicative, superimposing current relativities on the graph above is very straightforward. Superimposing relativities from a mixed multiplicative/additive rating plan is slightly less straightforward. Some additive components may be re-expressed as an interaction variable (eg {A x B x (C+D)} may be re-expressed to consider the interaction of C and D[12]). Existing rating plans with more complex additive components may be approximated by fitting a multiplicative model to a data field containing existing premium. The appropriateness of this multiplicative proxy to the mixed rating plan can be evaluated by examining the distribution of the ratio of the premium produced by the multiplicative proxy and the actual premium. Proxy models which estimate the rating plan within a narrow distribution (eg +/-5%) may well be appropriate to use.

Impact graphs

2.121 The results of a GLM analysis are interdependent and must be considered together. For example, while a GLM analysis might suggest that young driver relativities are too low, it may also suggest that relativities for inexperienced drivers (eg less than two years licensed) are too high. Although the existing rating structure may be theoretically wrong, it might be the case that to a large extent these errors compensate each other. To understand the true "bottom line" difference between the existing rating structure and the theoretical claims cost, "impact" graphs such as the one below can be considered.

Impact on portfolio of moving to theoretically correct relativities

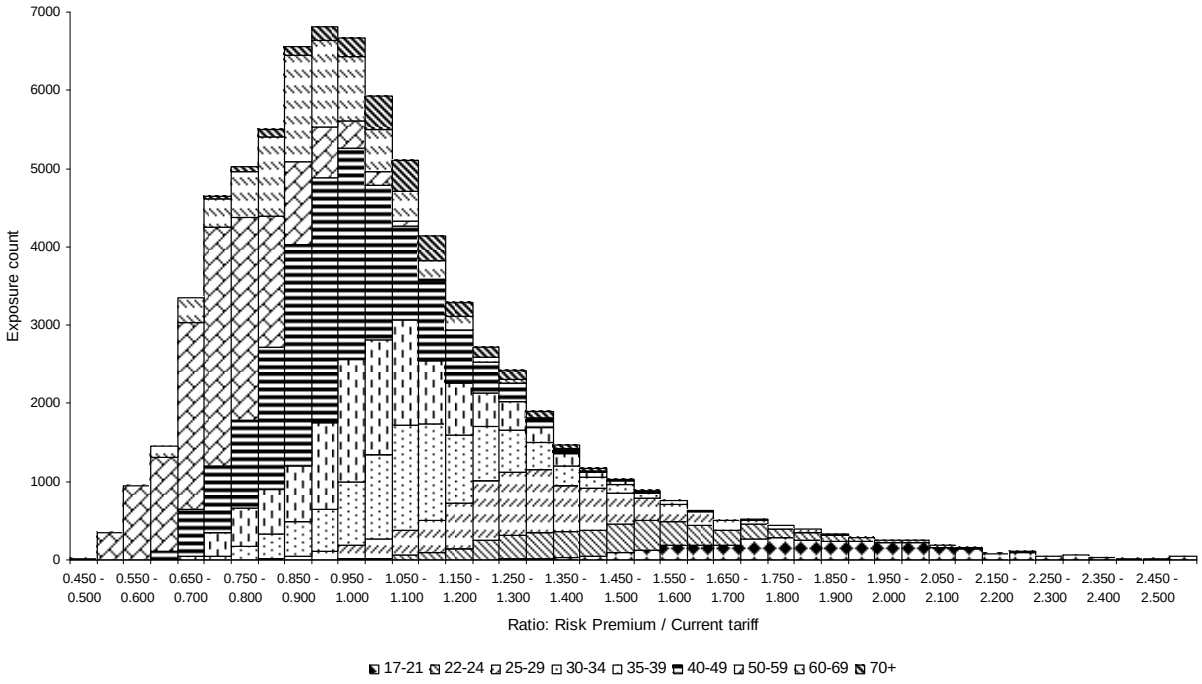


[12] Where A, B, C and D represent factors each of which possibly have a different number of levels.

2.122 This graph above shows the number of exposures in the existing portfolio that would experience different changes in premium if the rating structure were to move from its existing form to the theoretically correct form immediately. It is, of course, exceptionally unlikely that such dramatic change would be implemented in practice. The purpose of this analysis is to understand the magnitude of the existing cross-subsidies by considering the effect of all rating factors at the same time.

2.123 This graph can also be divided by levels of a particular rating factor. (Indeed one such graph can be produced for each rating factor.) This identifies which sectors of the business are currently profitable, and which are currently unprofitable, taking into account the correct theoretical model and considering the effect of all factors at the same time. In the example below, the impact graph is segmented by age of driver (notice the shape does not change, only how the histogram is patterned).

Impact on portfolio of moving to theoretically correct relativities (segmented by age of driver)

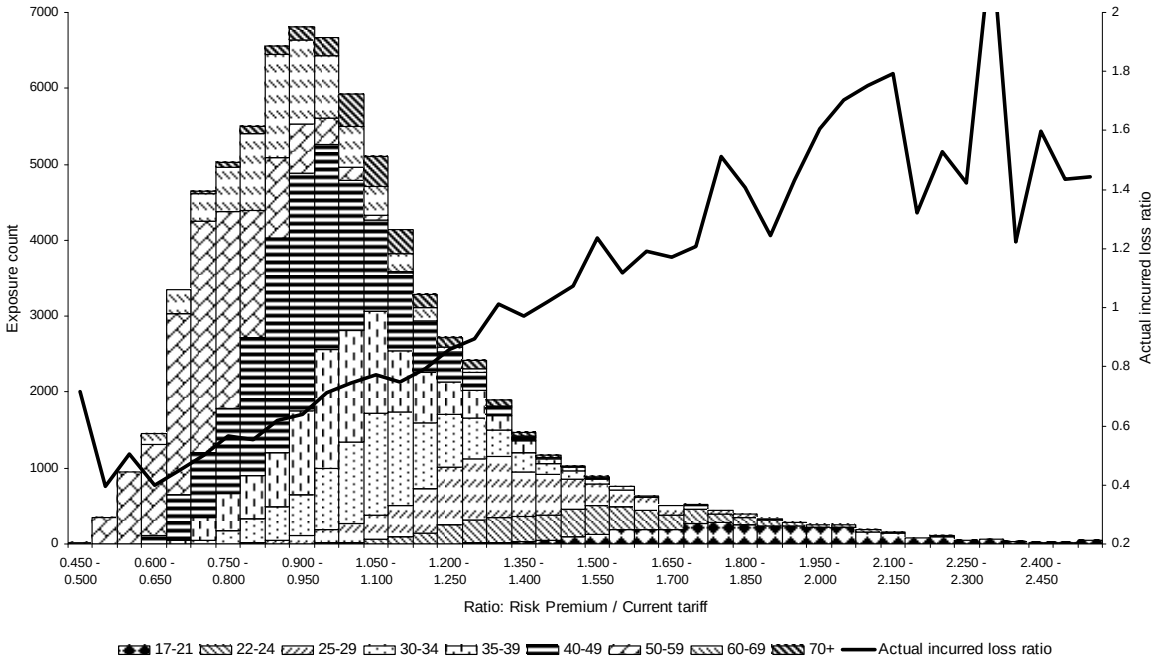


2.124 The histogram shows the impact of all rating factor changes (not just the age of driver factor) by age of driver levels. It can be seen in this example that a large number of exposures which would experience large increases in premium if the rating structure were moved immediately to the theoretically correct structure are young drivers. It had already been seen from the GLM risk premium graphs that young driver relativities were too low. This graph suggests there are no effects from other correlated factors which noticeably mitigate this effect; otherwise, young drivers would not be so strongly on the "unprofitable" side of the impact graph..

2.125 An example may make interpretation of the graph above clearer. Assume the multiplicative claims model uses age, gender, marital status, territory and credit as rating factors. Consider the following young driver profile with indicated rate change for each criterion in parenthesis: age 17-21 (+60%), male (+15%), single (-5%), urban territory (+15%), high credit score (-20%). All factors considered, the total indicated rate change for this risk profile is +61% and so this policy would contribute a count of one to the bar at 1.60-1.65. There are roughly 600 total exposures in this band; roughly one-third of which correspond to drivers age 17-21.

2.126 The graph below adds a second (right hand) y-axis. This y-axis contains the actual loss ratio present in the historical data. This shows very clearly how the GLM has differentiated between segments of differing profitability - each band on the x-axis represents a band of differing expected profitability, and the solid line shows the actual profitability experienced for that band.

Impact on portfolio of moving to theoretically correct relativities (segmented by age of driver, with actual loss ratio also shown)



3 Other applications of GLMs

3.1 This section briefly discusses

- the role of GLMs in the use of credit in personal lines ratemaking
- the use of scoring algorithms in more general terms to consider underwriting and marketing scorecards not necessarily related to credit
- the use of GLMs in retention/conversion analysis.

The role of GLMs in the use of credit-based insurance scores

3.2 Credit-based insurance scores attempt to measure the predictive power of components of consumer credit report data on the cost of insurance claims. The personal lines insurance industry in the US has been using credit-based insurance scoring for over a decade. A 2001 Conning & Company survey reported that 92% of the respondents of a survey of the 100 largest personal automobile insurance writers in the US use some form of credit scoring.[13]

3.3 The early published actuarial studies on the use of credit information in insurance demonstrated clear differences in univariate loss ratio by different bands of a credit-based credit score. Further studies examined this relationship by components of the Insurance Bureau score and also considered how loss ratio by credit component varied across certain traditional rating variables (ie a two-way approach).[14]

3.4 These studies drew early criticism regarding possible double-counting of effects already present in risk classification schemes.[15] Generalized linear models and other multivariate methods have played a critical role in addressing that criticism. A study conducted by EPIC Actuaries LLC on behalf of the property-casualty insurance industry's four national trade associations, offered four major findings about credit-based insurance scores:

---

[13] "Insurance Scoring in Private Passenger Automobile Insurance – Breaking the Silence", *Conning Report,* Conning, (2001).

[14] The reader seeking more information may reference the summaries of the Tillinghast study and the James Monoghan paper in "Does Credit Score Really Explain Insurance Losses? Multivariate Analysis from a Data Mining Point of View" by Cheng-Sheng Peter Wu and James C Guszcza, Casualty Actuarial Society Forum 2003 Vol: Winter Page(s): 120-125.

[15] The use of credit information in insurance underwriting and ratemaking has also drawn serious criticism regarding issues such as social equity, intuitive correlation with loss, disparate impact by race and level of wealth, etc. These issues are beyond the scope of this paper.

a. using generalized linear models to adjust for correlations between factors, insurance scores were predictive of propensity for private passenger automobile insurance loss (particularly frequency);

b. insurance scores are correlated with other risk characteristics, but after fully accounting for those correlations, the scores significantly increase the accuracy of risk assessment;

c. insurance scores are among the three most important risk factors for each of the six automobile claim types studied;

d. an analysis of property damage liability frequencies by insurance score group for each of the fifty states suggest consistent results across states.[16]

3.5 Model vendors and insurance companies have developed credit-based insurance scoring algorithms which vary in complexity, application and proprietary nature. The 2001 Conning & Company survey concluded that smaller insurers were using credit scoring predominantly in their underwriting processes, whereas larger insurers appeared to be focusing on underwriting, pricing and sophisticated market segmentation.

Insurance scores beyond credit

3.6 Other scoring techniques can be used as a way to share vital information between the actuarial departments and the rest of the insurance organization. For example, scores can be used to predict the profitability of an insurance policy given a certain rating structure (regardless of whether or not credit is considered). This information can be used in underwriting, cession decisions, marketing, and agent compensation schemes.

3.7 The most direct way to manage the profitability of a personal lines product is through effective ratemaking. Often, however, regulatory, practical or commercial conditions restrict the degree to which premiums can be set to reflect the risk. In these circumstances a score based on expected loss ratio can be used by insurers to gauge which customers are likely to be more profitable. As various functional areas are familiar with the application of scoring algorithms, this provides a common language for communicating a desired strategy throughout the insurance organization.

[16] "The Relationship of Credit-based Insurance Scores to Private Passenger Automobile Insurance Loss Propensity, an actuarial study by Epic Actuaries LLC"; principal authors Michael J. Miller and Richard A. Smith

3.8 For example, a scoring algorithm could help target marketing campaigns to those customers who are likely to be more profitable. Scores can also be used as part of an incentive scheme for agents, where commission or bonus is linked to the average customer score. Such applications can be particularly useful in highly regulated markets, as the score can include policyholder characteristics that are not permitted in the actual premium.

Producing the score

3.9 One method of deriving a scoring algorithm takes advantage of the "linear" part of generalized linear models (GLMs). The output of a GLM is a series of additive parameters which is then transformed via the link function to give the expected value for an observation. When calculating a score the link function can be omitted, leaving a simple additive structure which orders the risk. A straightforward calculation can then transform the additive structure into a scoring algorithm which produces scores between a desired range, for example 0 to 100.

3.10 To derive a profitability score, the starting point would be a standard analysis of claims experience using GLMs as discussed in detail in Section 2. This would involve fitting a series of GLMs to historic claims data, considering frequency and severity separately for each claim type. These models would include all standard rating factors, as well as any additional information that will be available at the time the score is to be calculated. Such additional information could include geodemographic data.

3.11 The expected cost of claims can then be calculated for each record in the data based upon the GLM claims models. For each policy this can then be divided by the premium which will be charged to yield an expected loss ratio, which can then itself be modeled and re-scaled to derive the profitability score.

3.12 The model of expected loss ratio should only include those factors that will be considered at the time the score is to be applied. For direct mailing campaigns this will usually mean that traditional insurance rating factors used in the premium will have to be excluded at this point (since they are not known at the time of the mailing campaign).

Example results

3.13 The graph below shows how a score can be used to segment very effectively between profitable and unprofitable business. The bars on the graph show the number of policies that have been allocated different scores between 0 and 100. The solid line shows the actual loss ratio experienced for business with differing scores. It can be seen that the business towards the left of the graph, with low profitability scores, is experiencing loss ratios of 100% and above, while the business to the right of the graph, with high scores, is returning loss ratios of 50% and below.

Distribution of score

3.14 Scores are simple to produce, easy to explain and are increasingly used by insurers. Actuaries can play a vital role in the development of scoring models with the aid of generalized linear models.

Retention modeling using GLMs

3.15 Traditional ratemaking techniques focus primarily on loss analysis in a static environment. Rate changes developed by these techniques, especially when they are large, can actually contribute to a shortfall in projected premium volume and profitability if insufficient consideration is given to the effect of the rate change and other policy characteristics on customer retention and/or new business conversion. Modeling retention (or its complement, lapse rate) and new business conversion with GLMs can improve ratemaking decisions and profitability forecasts, as well as improve marketing decisions.

3.16 The data for a retention model must include information on individual policies that have been given a renewal offer, and whether or not each policy renewed.[17] Similarly, data for a conversion model must contain individual past quotes and whether the quote converted to new business. (While most insurers have access to appropriate retention data, many distributing via exclusive agents or independent brokers will not have access to appropriate individual conversion data.) The explanatory variables to include in the data can be divided into three categories: customer information, price change data, and information on the competitive position.

3.17 The first category should encompass more than just the standard rating variables (eg age, territory, claim experience). Other "softer" variables such as number of years with the company, other products held, payment plan and endorsement activity can determine much about a customer's behavior. Distribution channel, too, can have a clear effect on the retention rate - and may interact significantly with other factors (eg the effect of age may be different with internet distribution than with agency distribution).

3.18 Prior rate change, whether measured in percent change or dollar change, is often one of the most significant factors in a retention model. Though it is intuitive that retention is a function of rate change, the slope of the elasticity curve at different rate changes may not be as obvious. In addition, measuring retention using a generalized linear model will adjust for exposure correlations between price elasticity and other explanatory variables (eg a GLM will not show that a particular rating factor level has a low retention rate merely because historically there was aggressive rate activity with that level).

3.19 The third type of variable, information on the competitive position, is often the hardest to gather in practice. An example of a competitive index may be the ratio of the company's renewal quote to the third cheapest quote from a specified selection of major competitors at the time of the quote.

3.20 Tracking the myriad of competitor rate changes in a multitude of states can be overwhelming - even with the availability of third party competitive rating software and advances in quote collection procedures. Fortunately even the most rudimentary competitive index variables can prove to be predictive in a retention model (and more so in a conversion model).

[17] Alternatively, retention data may be organized by risk if more than one risk is written on a single policy.

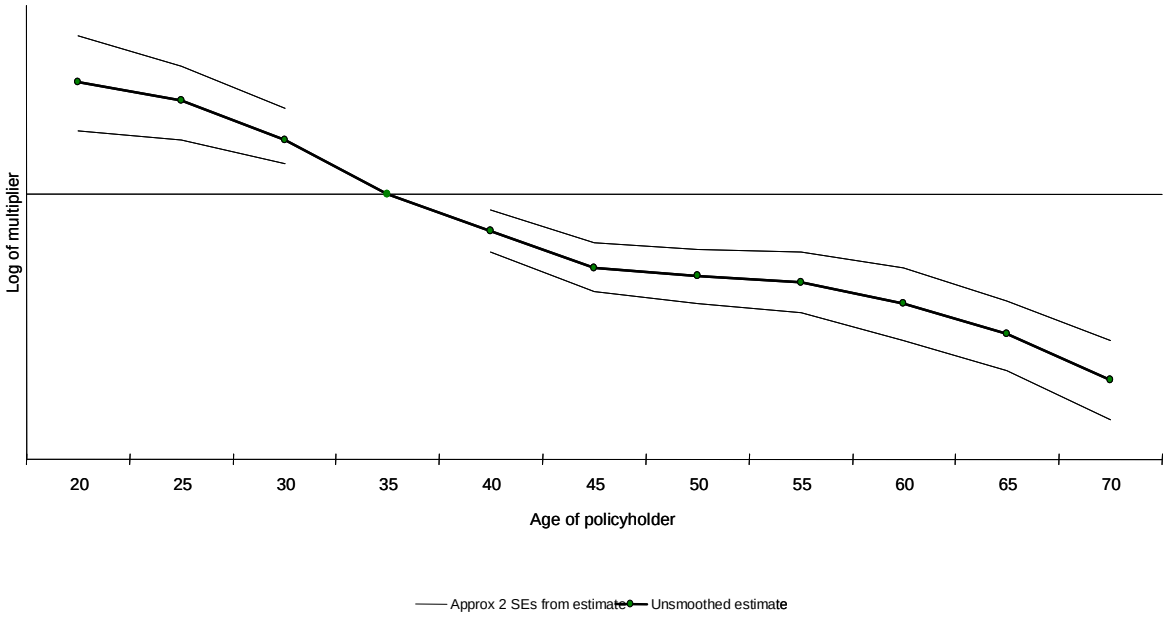
Model form

3.21 As mentioned previously in Section 1, the typical model form for modeling retention (or lapse) and new business conversion is a logit link function and binomial error term (together referred to as a logistic model). The logit link function maps outcomes from the range of (0,1) to $(-\infty, +\infty)$ and is consequently invariant to measuring successes or failures. If the y-variate being modeled is generally close to zero, and if the results of a model are going to be used qualitatively rather than quantitatively, it may also be possible to use a multiplicative Poisson model form as an approximation given that the model output from a multiplicative GLM can be rather easier to explain to a non-technical audience.

Example results

3.22 The graph below shows sample GLM output for a lapse model. The main line on the graph demonstrates (on a log scale) the measured multiplicative effect of age of policyholder upon lapse rate. The effect is measured relative to an arbitrarily selected base level, and the results take into account the effect of all other factors analyzed by the GLM.

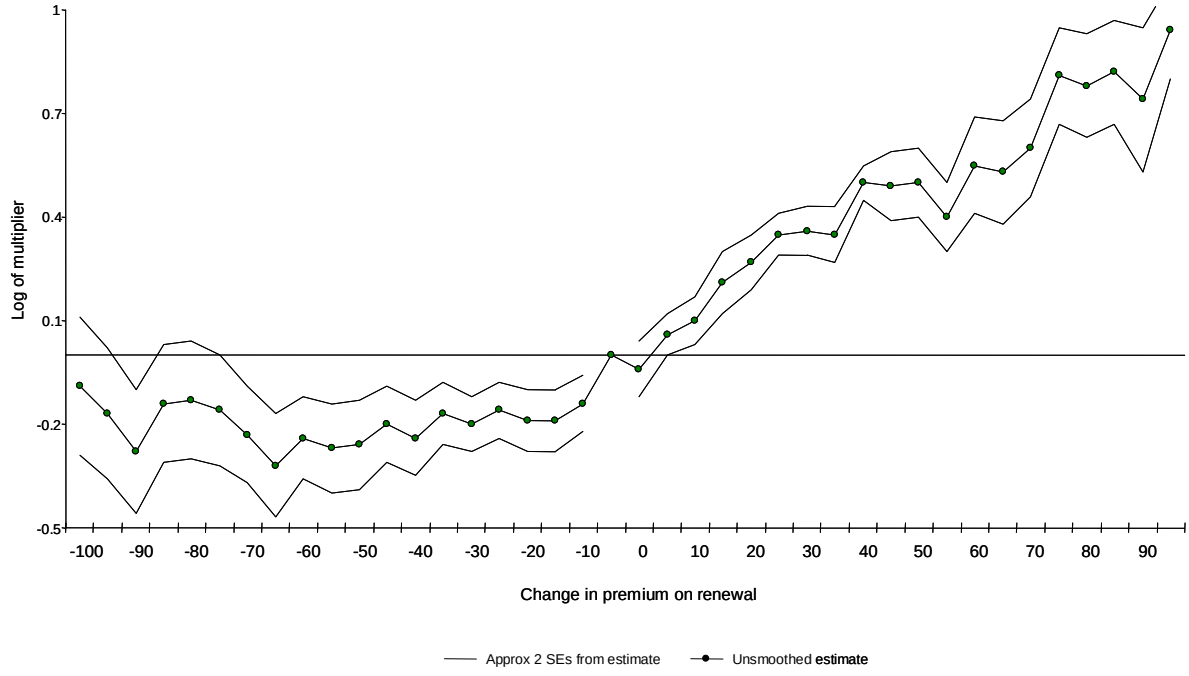
Poisson multiplicative lapse model output



3.23 In this example, which is fairly typical, it can be seen that young policyholders lapse considerably more than older policyholders, perhaps as a result of having more time and enthusiasm in searching for a better quotation, and perhaps also as a result of being generally less wealthy and therefore more interested in finding a competitive price.

3.24 This next graph shows the effect of premium change on lapse rate. This GLM output is from a UK Institute of Actuaries General Insurance Research Organisation (GIRO) study[18] based on around 250,000 policies across several major UK insurers in 1996. The premium change is measured in ranges of monetary units (British pounds in this case), but the model could easily be based on the percentage change in premium. As would be expected, increases in premium increase lapses. The model, however, quantifies this accurately and enables investigations into potentially optimal rate increases to be undertaken. It can be seen in this case (as is often the case) that decreases in premium beyond a small threshold do not decrease lapses.

Poisson multiplicative lapse model output



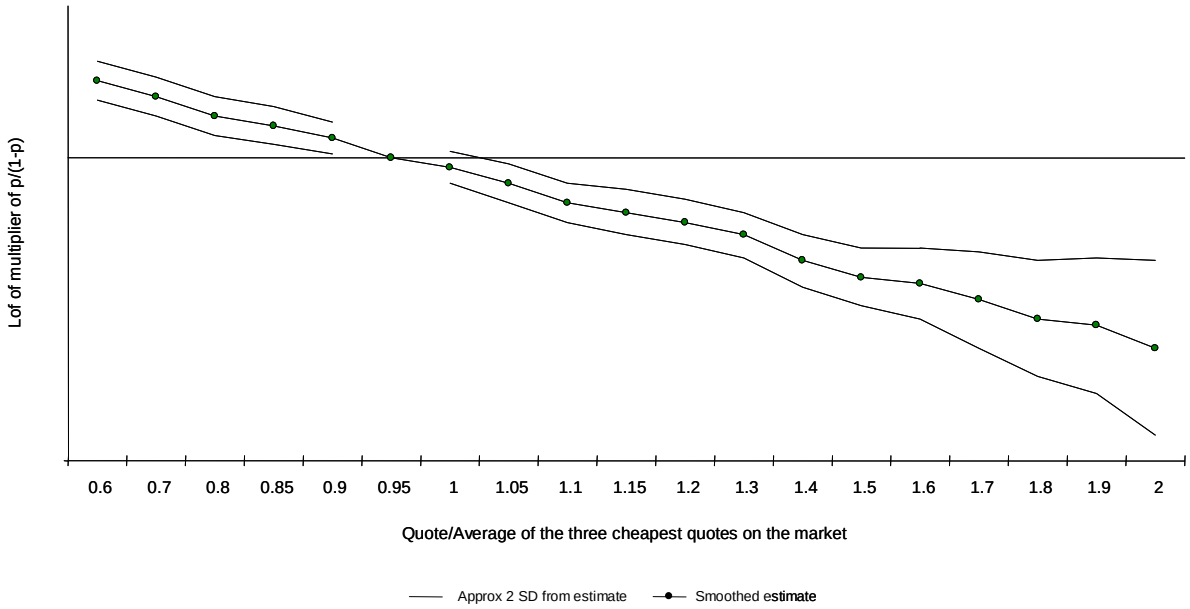
[18] Bland, R.H. et al, Institute of Actuaries GIRO Customer Selection and Retention Working Party, 1997 - ISBN 0 901066 45 1

3.25 Measures of premium change should ideally consider whether customers have an inherent expectation of premium change. For example, customers with recent claims will anticipate a premium increase and may be prepared to accept their renewal offer rather than face the underwriting guides of a new company. Conversely, customers who are rolling off an accident surcharge, hitting a milestone age or a change in marital status may expect a decrease. A possible proxy for customer expectation is to adjust the premium change variable to be the ratio of proposed premium (based on new risk criteria and new rates) to adjusted proposed premium (new risk criteria based on last year's rates).

3.26 In addition to including premium change, absolute premium can also be considered as a factor in a model. This approach, though not theoretically incorrect, may make the model difficult to interpret since many other factors in the model will be a component of premium and therefore highly correlated with premium size. Adding absolute premium to the model may significantly alter the observed relativities for other factors which may make the results hard to interpret. One alternative to including absolute premium in such a case is to fit separate models for different bands of average premium.

3.27 The next graph below shows an example of the effect of competitiveness in a new business conversion model. The measure of competitiveness used in this case is the ratio of the proposed premium to the average of the three cheapest alterative premiums from a selection of alternative insurers. It can be seen that the less competitive the premium, the lower the conversion rate.

Logistic new business conversion model output



3.28 A further analysis which can be undertaken is to superimpose the results of two models on one graph: one model that includes the competitiveness measure and one model that does not. The disparity between these two models will show how much of a factor's effects are simply price-related.

Applications

3.29 In a fully deregulated market such as the UK, insurance companies can set premium rates according to what the market will bear. In most US states and Canadian provinces, insurance companies need to demonstrate that rates are within a reasonable range of loss and expense cost estimates. Companies can, however, measure the sensitivity of various point selections within those ranges (whether the point estimates pertain to overall rate level or classification ratemaking). Future pricing reviews may not only present management with support on actuarial considerations such as trend and loss development, but also a forecast of how various rate change proposals are expected to affect retention, conversion, premium, overall loss ratio (incorporating both overall rate change and portfolio shift of classification changes) and profitability in the short term and/or long term.

3.30 Retention analyses can also lead to operational actions which are unrelated to price. For example, in a highly rate-regulated state, consideration could be given to which segments of the population (given a restricted set of rates) are both profitable and most likely to renew in the future. Such a measure could help form new underwriting guides or targeted marketing and cross-sell campaigns.

3.31 Insurance expense analysis is another field of study that is often over-shadowed by loss analysis. If acquisition expenses are higher than renewal expenses then an understanding of likely retention (and therefore expected life of a policy) can be used to amortize the higher acquisition cost over the expected life of the policy.

Conclusion

3.32 A GLM statistically measures the effect that variables have on an observed item. In insurance, GLMs are most often used to determine the effect rating variables have on claims experience and the effect that rating variables and other factors have on the probability of a policy renewing or a new business quotation being accepted.

3.33 GLMs estimate the true effect of each variable upon the experience, making appropriate allowance for the effect of all other factors being considered. Ignoring correlation can produce significant inaccuracies in rates.

3.34 GLMs incorporate assumptions about the nature of the random process underlying claims experience. Having the flexibility to specify a link function and probability distribution that matches the observed behavior increases the accuracy of the analysis.

3.35 A further advantage of using GLMs is that as well as estimating the effect that a given factor has on the experience, a GLM provides information about the certainty of model results.

3.36 GLMs are robust, transparent and easy to understand. With advances in computer power, GLMs are widely recognized as the industry standard in European personal lines, and fast gaining acceptance from industry professionals in the US and Canada.

3.37 GLMs in insurance are not limited to pricing. Alternative applications of GLM claims analyses include underwriting, selective marketing and agency marketing.

3.38 GLMs are grounded in statistical theory and offer a practical method for insurance companies to attain satisfactory profitability and a competitive advantage.

Bibliography

Bailey, Robert A.; and LeRoy J. Simon, "Two Studies in Automobile Insurance Ratemaking," Proceedings of the Casualty Actuarial Society, XLVII, 1960.

Bland, R.H. et al, Institute of Actuaries GIRO Customer Selection and Retention Working Party, 1997 - ISBN 0 901066 45 1

Brockman, M.J; Wright, T.S., "Statistical Motor Rating: Making Effective Use of Your Data", Journal of Institute of Actuaries 119, Vol. III, pages: 457-543, 1992.

Conning, "Insurance Scoring in Private Passenger Automobile Insurance – Breaking the Silence", Conning Report (2001).

Feldblum, Sholom; and Brosius, Eric J "The Minimum Bias Procedure--A Practitioner's Guide" Casualty Actuarial Society Forum 2002 Vol: Fall Page(s): 591-684

Hardin, James; and Hilbe, Joseph, "Generalized Linear Models and Extensions", Stata Press, 2001

Jørgenses, B and De Souza, M.C.P, "Fitting Tweedie's Compound Poisson Model to Insurance Claims Data", Scand. Actuarial J. 1994 1:69-93.

McCullagh, P. and J. A. Nelder, "Generalized Linear Models", 2nd Ed., Chapman & Hall/CRC, 1989.

Mildenhall, Stephen, "A systematic relationship between minimum bias and generalized linear models", Proceedings of the Casualty Actuarial Society, LXXXVI, 1999.

Miller, Michael J.; and Smith, Richard A., "The Relationship of Credit-based Insurance Scores to Private Passenger Automobile Insurance Loss Propensity", an Actuarial Study by Epic Actuaries LLC, 2003

Wu, Cheng-Sheng Peter; and Guszcza, James C., "Does Credit Score Really Explain Insurance Losses? Multivariate Analysis from a Data Mining Point of View", Casualty Actuarial Society Forum 2003 Vol: Winter, Page(s): 120-125.

A The design matrix when variates are used

Consider the example of a model which is based on two continuous rating variables: age of driver and age of car.

Let $\underline{Y}$ be a column vector with components corresponding to the n observed values for the response variable, for example severity:

$$\underline{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \dots \\ Y_n \end{bmatrix} = \begin{bmatrix} 800 \\ 400 \\ \dots \\ 200 \end{bmatrix}$$

Let $\underline{X}_1$ and $\underline{X}_2$ denote the column vectors with components equal to the observed values for the continuous variables (eg $\underline{X}_1$ shows the actual age of the driver for each observation):

$$\underline{X}_1 = \begin{bmatrix} 18.1 \\ 32.2 \\ \dots \\ 44.4 \end{bmatrix} \qquad \underline{X}_2 = \begin{bmatrix} 12.5 \\ 1.6 \\ \dots \\ 3.8 \end{bmatrix}$$

As before, $\underline{\beta}$ denotes a column vector of parameters, and $\underline{\varepsilon}$ the vector of residuals:

$$\underline{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} \qquad \underline{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_n \end{bmatrix}$$

Then the system of equations takes the form:

$$\underline{Y} = \beta_1 \underline{X}_1 + \beta_2 \underline{X}_2 + \underline{\varepsilon}$$

Or, defining the design matrix **X** as

$$X = \begin{bmatrix} 18.1 & 12.5 \\ 32.2 & 1.6 \\ \dots & \dots \\ 44.4 & 3.8 \end{bmatrix}$$

The system of equations takes the form

$$\underline{Y}=\boldsymbol{X}.\underline{\beta}+\underline{\varepsilon}$$

Polynomials

Rather than assuming that the value of $\boldsymbol{X}.\underline{\beta}$ is linear in the variate, it is also possible to include in the definition of $\boldsymbol{X}.\underline{\beta}$ terms based on polynomials in the variates. For example, a model could be based on a third order polynomial in age of driver and a second order polynomial in age of vehicle. In this case the design matrix would be defined as follows:

$$X=\begin{bmatrix} 1 & 18.1 & 327.61 & 5929.741 & 12.5 & 156.25 \\ 1 & 32.2 & 1036.84 & 33386.25 & 1.6 & 2.56 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & 44.4 & 1971.36 & 87528.38 & 3.8 & 14.44 \end{bmatrix}$$

where

- the first column represents the intercept term (driver age)^0
- the second column represents the values of (driver age)^1
- the third column represents the values of (driver age)^2
- the fourth column represents the values of (driver age)^3
- the fifth column represents the values of (vehicle age)^1
- the sixth column represents the values of (vehicle age)^2

B The exponential family of distributions

Formally the exponential family of distributions is a two-parameter family of functions defined by:

$$f(y;\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}$$

where $a(\phi)$, $b(\theta)$, and $c(y, \phi)$ are specified functions. Conditions imposed on these functions are that

- **a.** $a(\phi)$ is positive and continuous;
- **b.** $b(\theta)$ is twice differentiable with the second derivative a positive function (in particular $b(\theta)$ is a convex function); and
- **c.** $c(y, \phi)$ is independent of the parameter θ.

These three functions are related by the simple fact that f must be a probability density function and so it must integrate to 1 over its domain. Different choices for $a(\phi)$, $b(\theta)$, and $c(y, \phi)$ define a different class of distributions and a different solution to the GLM problem. The parameter θ is termed the canonical parameter and ϕ the scale parameter. The chart below summarizes some familiar distributions that are members of the exponential family:

	$a(\phi)$	$b(\theta)$	$c(y,\phi)$
Normal	ϕ/ω	$\theta^2/2$	$-\frac{1}{2}(\omega y^2/\phi + \ln(2\pi\phi/\omega))$
Poisson	ϕ/ω	e^θ	$-\ln y!$
Gamma	ϕ/ω	$-\ln(-\theta)$	$\omega/\phi \ln(\omega y/\phi) - \ln(y) - \ln(\Gamma(\omega/\phi))$
Binomial (m trials)	ϕ/ω	$m.\ln(1+e^\theta)$	$\ln\binom{m}{y}$
Inverse Gaussian	ϕ/ω	$-\sqrt{-2\theta}$	$-\frac{1}{2}\{\ln(2\pi\phi y^3/\omega) + \omega/(\phi y)\}$

It can be seen that the standard choice for $a(\phi)$ is

$$a(\phi) = \frac{\phi}{\omega}$$

where ω is a *prior weight*, a constant that is specified in advance. For insurance applications common choices for the prior weight are equal to 1 (eg when modeling claim counts), the number of exposures (eg when modeling claim frequency), or the total number of claims (eg when modeling claim severity). It is also clear from the chart that for certain distributions, such as the Poisson and binomial distributions, the scale parameter ϕ is equal to 1 and plays no further role in the modeling problem.

A distribution for each observation Y_i needs to be specified. It is assumed that

$$f(y_i;\theta,\phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi)\right\}$$

Thus each observation has a different canonical parameter θ_i but the scale parameter ϕ is the same across all observations. It is further assumed that the functions $a(\phi)$, $b(\theta)$, and $c(y, \phi)$ are the same for all i. So each observation comes from the same class within the exponential family, but allowing θ to vary corresponds to allowing the mean of each observation to vary.

The parameters θ_i and ϕ encapsulate the mean and variance information about Y_i. It can be shown that for this family of distributions:

$$\mu_i = E(Y_i) = b'(\theta_i)$$
$$Var(Y_i) = b''(\theta_i).a(\phi)$$

where the prime (') denotes differentiation with respect to θ.

The first equation implicitly defines θ_i as a function of μ_i. If an explicit expression for the inverse of $b'(\theta_i)$ is known (as is the case for the familiar distributions) then the first equation can be solved to express the canonical parameter θ_i explicitly as a function of the mean of the distribution μ_i:

$$\theta_i = (b')^{-1}(\mu_i)$$

Thus the canonical parameter is essentially equivalent to the mean.

Section 1 describes how a GLM asserts that μ_i is a function of the linear predictor η_i where the linear predictor is a linear combination of the p covariates $X_{i1}, \ldots X_{ip}$:

$$\mu_i = g^{-1}(\beta_1 x_{i1} + \ldots + \beta_p x_{ip})$$

Thus θ is ultimately a complicated function of the elements of $\underline{\beta}$:

$$\theta_i = (b')^{-1}\left(g^{-1}(\beta_1 x_{i1} + \ldots + \beta_p x_{ip})\right)$$

This derivation makes explicit the manner in which the distribution of Y_i depends on the GLM parameters $\beta_1, \ldots, \beta_p$.

It can be seen from the table above that the expression for $c(y, \phi)$ can be complicated. Fortunately as long as $c(y, \phi)$ does not depend on θ - and hence not on $\underline{\mu}$ and thus not on the GLM modeling parameters $\underline{\beta}$ - then the form of $c(y, \phi)$ is irrelevant to the solution of the maximum likelihood estimator.

Given that θ_i is a function of the mean μ_i the equation

$$Var(Y_i) = b''(\theta_i).a(\phi)$$

can be interpreted as establishing the variance of Y_i as a function of the mean of Y_i times some scaling term $a(\phi)$. Thus the scaling parameter ϕ is a function of the mean and variance of the distribution.

Thus the exponential family has two desirable properties:

- each distribution in the family is completely specified in terms of its mean and variance
- the variance of Y is a function of its mean.

This second property is emphasized by writing

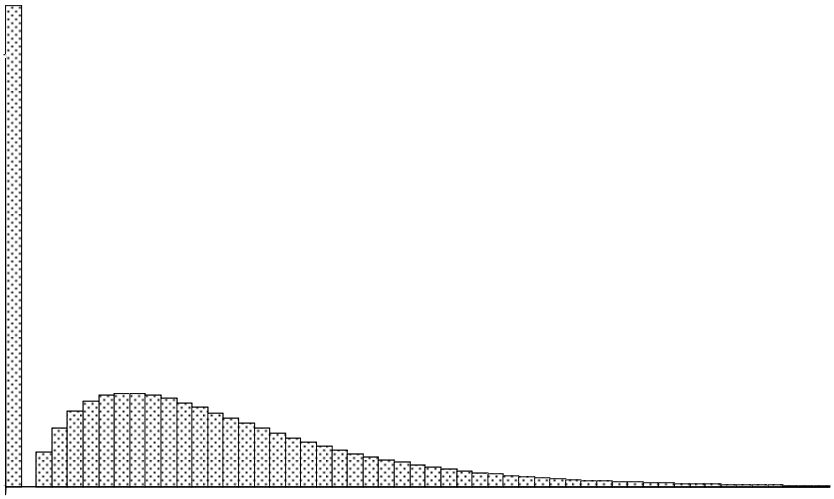
$$Var(Y_i) = \phi \frac{V(\mu_i)}{\omega_i}$$

where the function V is termed the variance function. The chart below summarizes the relationship between the mean and the canonical parameter, expresses f in terms of the standard parameters for the respective distribution, and lists the variance function for the familiar distributions:

	Notation	ϕ	$\mu(\theta)$	$V(\mu)$
Normal	$N(\mu, \sigma^2)$	σ^2	θ	1
Poisson	$P(\mu)$	1	e^{θ}	μ
Gamma	$G(\mu, \nu)$	ν^{-1}	$-1/\theta$	μ^2
Binomial	$B(m, \pi)/m$	$1/m$	$e^{\theta}/(1+e^{\theta})$	$\mu(1-\mu)$
Inverse Gaussian	$IG(\mu, \sigma^2/\omega)$	σ^2	$(-2\theta)^{-1/2}$	μ^3

C The Tweedie distribution

Direct modeling of pure premium or incurred loss data is problematic since a typical pure premium distribution will consist of a large spike at zero (where policies have not had claims) and then a wide range of amounts (where policies have had claims). This is illustrated in the diagram below.



Many of the traditional members of the exponential family of distributions are not appropriate for modeling claims experience from such a distribution since they do not have a point mass at zero combined with an appropriate spread across non-zero amounts.

The Tweedie distribution is a special member of the exponential family which has a variance function proportional to μ^p, with p being an additional parameter.

In the case of $1<p<2$ the Tweedie distribution has a point mass at zero and corresponds to the compound distribution of a Poisson claim number process and a Gamma claim size distribution. The distribution can be Poisson-like (as $p \to 1$) or Gamma-like (as $p \to 2$).

In total the distribution has three parameters - a mean parameter, a dispersion parameter, and the "shape" parameter p, which when $1<p<2$ is often written in terms of α where

$$p = \frac{\alpha - 2}{\alpha - 1}$$

Its density function is rather complex, and in the case of $1<p<2$ is defined as:

$$f_Y(y;\theta,\lambda,\alpha) = \sum_{n=1}^{\infty} \frac{\left\{(\lambda\omega)^{1-\alpha}\kappa_\alpha(-1/y)\right\}^n}{\Gamma(-n\alpha)n!\,y}.\exp\{\lambda\omega[\theta_0 y - \kappa_\alpha(\theta_0)]\} \quad \text{for } y > 0$$

and

$$p(Y=0) = \exp\{-\lambda\omega\kappa_\alpha(\theta_0)\}$$

where

$$\kappa_\alpha(\theta) = \frac{\alpha-1}{\alpha}.\left(\frac{\theta}{\alpha-1}\right)^\alpha$$

$$\theta_0 = \theta.\lambda^{1/(1-\alpha)}$$

ω is the prior weight corresponding to the exposure of the observation in question.

It can be shown that the variance function for the above Tweedie distribution is given by

$$V(\mu) = \frac{1}{\lambda}\mu^p$$

In practice the shape parameter can either be assumed to be a particular value or, more usefully, estimated as part of the maximum likelihood process. Typically values of p just under 1.5 seem to be estimated for auto claims experience.

Further information about the Tweedie distribution can be found in the paper "Fitting Tweedie's Compound Poisson Model to Insurance Claims Data" by Jørgenses, B and De Souza, M.C.P, Scand. Actuarial J. 1994 1:69-93.

D Canonical link functions

Each of the exponential distributions has a natural link function called the canonical link. It has the property that $\underline{\theta} = \underline{\eta}$ where $\underline{\theta}$ is the canonical parameter. This property means that the GLM parameters $\beta_1, \ldots \beta_p$ enter the expression for the distribution function in a simple way. In general

$$f_{Y_i}(y_i) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y,\phi)\right\}$$
$$= \exp\left\{\frac{y_i(b')^{-1}(g^{-1}(\eta_i)) - b((b')^{-1}g^{-1}(\eta_i))}{a(\phi)} + c(y_i,\phi)\right\}$$

but if $\underline{\theta} = \underline{\eta}$ this simplifies to

$$\exp\left\{\frac{y_i\eta_i - b(\eta_i)}{a(\phi)} + c(y,\phi)\right\}$$

and subsequent differentiation with respect to the GLM parameters β_j is thus significantly simplified.

The canonical link functions associated with the familiar distributions are listed below

	Canonical Link
Normal	μ
Poisson	$\ln \mu$
Gamma	$1/\mu$
Binomial	$\ln(\mu/(1-\mu))$
Inverse Gaussian	$1/\mu^2$

Note that the requirement to be a canonical link function:

$$\theta = (b')^{-1}(g^{-1}(\eta)) = \eta$$

implies that the inverse of the link function, g^{-1}, is the inverse of b'.

In practice, with sophisticated software to solve GLM modeling problems there is no imperative to use the canonical link associated with a particular distribution. Instead any arbitrary pairings of the link function and the error structure can be made and such non-canonical pairings can in fact yield more predictive models.

E Solving for maximum likelihood in the general case of an exponential distribution

In the case of the exponential family of distributions the log likelihood takes the form:

$$l = \sum_i \left\{ \frac{y_i \theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right\}$$

Log likelihood is maximized by taking, for each j, the first order partial derivative of *l* with respect to β_i and setting equal to zero:

$$\frac{\partial l}{\partial \beta_j} = 0, \qquad j = 1,..., p$$

If there is an explicit expression for θ_i in terms of $\beta_1,\ldots\beta_p$ one can make this substitution into the log likelihood function and then carry out the differentiation. However, the calculations become complicated quite quickly. It is simpler just to apply the chain rule of calculus three times:

$$0 = \frac{\partial l}{\partial \beta_j} = \sum_i \frac{\partial}{\partial \theta_i} \left(\frac{Y_i \theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right) \cdot \frac{\partial \theta_i}{\partial \mu_i} \cdot \frac{\partial \mu_i}{\partial \eta_i} \cdot \frac{\partial \eta_i}{\partial \beta_i}$$

Recalling the following relationships:

$$\mu_i = b'(\theta_i) \Rightarrow \frac{\partial \mu_i}{\partial \theta_i} = b''(\theta_i) \Rightarrow \frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{b''(\theta_i)}$$

$$\eta_i = g(\mu_i) \Rightarrow \frac{\partial \eta_i}{\partial \mu_i} = g'(\mu_i) \Rightarrow \frac{\partial \mu_i}{\partial \eta_i} = \frac{1}{g'(\mu_i)}$$

$$\eta_i = \beta_1 X_{i1} + ... + \beta_p X_{ip} \Rightarrow \frac{\partial \eta_i}{\partial \beta_j} = X_{ij}$$

It can be deduced that

$$\frac{\partial l}{\partial \beta_j} = \sum_i \frac{(y_i - \mu_i)}{a(\phi)} \cdot \frac{1}{b''(\theta_i)} \cdot \frac{1}{g'(\mu_i)} \cdot x_{ij}\,, \qquad j = 1,..., p$$

$$= \sum_i \frac{\omega_i\,(y_i - \mu_i)\,x_{ij}}{\phi\, V(\mu_i) g'(\mu_i)}\,, \qquad j = 1,..., p$$

Although the theoretical system of equations which must be satisfied in order to maximize the likelihood can be (relatively) easily written, finding the solution to these equations is more complicated.

F Example of solving for maximum likelihood with a gamma error and inverse link function

For the gamma error structure with an inverse link function, the predicted values take the form:

$$E[\underline{Y}] = g^{-1}(X.\underline{\beta}) = \begin{bmatrix} g^{-1}(\beta_1 + \beta_3) \\ g^{-1}(\beta_1) \\ g^{-1}(\beta_2 + \beta_3) \\ g^{-1}(\beta_2) \end{bmatrix} = \begin{bmatrix} (\beta_1 + \beta_3)^{-1} \\ (\beta_1)^{-1} \\ (\beta_2 + \beta_3)^{-1} \\ (\beta_2)^{-1} \end{bmatrix}$$

The gamma error structure has the following density function

$$f(x;\mu,\phi) = \frac{x^{-1}}{\Gamma(1/\phi)}(\frac{x}{\mu\phi})^{1/\phi} e^{(-\frac{x}{\mu\phi})}$$

Its log-likelihood function is

$$l(x;\mu,\phi) = \sum_{i=1}^{n} \ln f(x_i;\mu_i) = \sum_{i=1}^{n} \frac{1}{\phi}(\ln\frac{x_i}{\mu_i} - \frac{x_i}{\mu_i}) - \ln x_i - \frac{\ln\phi}{\phi} - \ln\Gamma(\frac{1}{\phi})$$

With an inverse link function, $\mu_i = 1/(\Sigma_j X_{ij}\beta_j)$ and the log-likelihood function reduces to

$$l(x;1/X\beta,\phi) = \sum_{i=1}^{n} \frac{1}{\phi}(\ln(x_i.\sum_{j=1}^{p} X_{ij}\beta_j) - x_i.\sum_{j=1}^{p} X_{ij}\beta_j) - \ln x_i - \frac{\ln\phi}{\phi} - \ln\Gamma(\frac{1}{\phi})$$

In this example,

$$\begin{aligned} l(x;\mu) = {} & \frac{1}{\phi}(\ln(800.(\beta_1 + \beta_3)) - 800.(\beta_1 + \beta_3)) - \ln 800 - \frac{\ln\phi}{\phi} - \ln\Gamma(\frac{1}{\phi}) \\ & + \frac{1}{\phi}(\ln(500.\beta_1) - 500.\beta_1) - \ln 500 - \frac{\ln\phi}{\phi} - \ln\Gamma(\frac{1}{\phi}) \\ & + \frac{1}{\phi}(\ln(400.(\beta_2 + \beta_3)) - 400.(\beta_2 + \beta_3)) - \ln 400 - \frac{\ln\phi}{\phi} - \ln\Gamma(\frac{1}{\phi}) \\ & + \frac{1}{\phi}(\ln(200.\beta_2) - 200.\beta_2) - \ln 800 - \frac{\ln\phi}{\phi} - \ln\Gamma(\frac{1}{\phi}) \end{aligned}$$

Ignoring some constant terms and multiplying by ϕ, the following function is to be maximized

$$\begin{aligned} l^*(y;\mu) = {} & \ln(800.(\beta_1 + \beta_3)) - 800.(\beta_1 + \beta_3) + \ln(500.\beta_1) - 500.\beta_1) \\ & + \ln(400.(\beta_2 + \beta_3)) - 400.(\beta_2 + \beta_3) + \ln(200.\beta_2) - 200.\beta_2 \end{aligned}$$

Again, to maximize l* take derivatives with respect to β_1, β_2 and β_3. Set the derivatives to zero and the following three equations are derived:

$$\frac{\partial l^*}{\partial \beta_1} = 0 \quad => \quad \frac{1}{\beta_1 + \beta_3} + \frac{1}{\beta_1} = 1300$$

$$\frac{\partial l^*}{\partial \beta_2} = 0 \quad => \quad \frac{1}{\beta_2 + \beta_3} + \frac{1}{\beta_2} = 600$$

$$\frac{\partial l^*}{\partial \beta_3} = 0 \quad => \quad \frac{1}{\beta_1 + \beta_3} + \frac{1}{\beta_2 + \beta_3} = 1200$$

Solving these simultaneous equations gives the following solutions:

$$\beta_1 = 0.00223804$$
$$\beta_2 = 0.00394964$$
$$\beta_3 = -0.00106601$$

which result in the following predicted values:

	Urban	Rural
Male	853.2	446.8
Female	346.8	253.2

G Data required for a GLM claims analysis

The overall structure of a dataset for GLM claims analysis consists of linked policy and claims information at the individual risk level. The definition of individual risk level will vary according to the line of business and the type of model. For instance, in a personal automobile claims model, the definition of risk may be a vehicle. (In a personal automobile retention model, the definition of risk may be a policy containing several vehicles.)

One record should be present for each period of time during which a policy was exposed to the risk of having a claim, and during which all factors remained unchanged. Policy amendments should ideally appear as two records, with the previous exposure curtailed at the point of amendment. Mid-term policy cancellations should also result in the exposure period being curtailed. If this data is not available it is often possible to approximate it from less perfect data - for example the policies in force at one year end could be compared with the policies in force at the previous year end, with matching policies being assumed to be in force for the whole year, and appropriate approximations being made for non-matching policies.

The dataset should contain fields defining the earned exposure and the rating factors applicable at the start of the exposure period. Additionally, premium information (typically earned premium) can be attached to each record. Although premium is not used directly in the development of the claims models, it can provide valuable information for measuring the impact of any new rating or underwriting actions, and for producing summary one-way and two-way analyses including loss ratios.

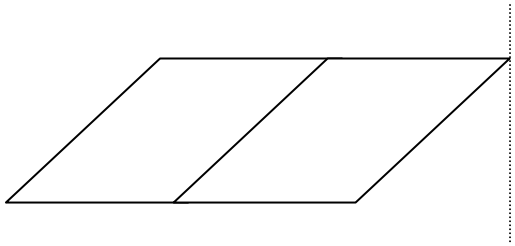
All explanatory variables in the dataset should record the criteria which were applicable at the start of the policy exposure (or, strictly speaking, the point at which the premium was determined for the exposure period in question). In the case of categorical variables such as territory or vehicle class, however, the data recorded should ideally be derived by applying the current method of categorization to the historic situation.

Not all explanatory variables will be used to predict future claims experience. Dummy variables may be used to absorb certain effects that could bias the parameter estimates. For example, if conducting a countrywide study, it may be appropriate to create a dummy variable to standardize for differences in overall loss experience by geography. This dummy variable may be state (province), territory within state (province), or groups of territories within state (province).

Similarly, if combining data from several companies, a company identifier may be an appropriate dummy variable. This dummy variable could absorb differences in underwriting standards and overall quality of business between the companies. Dummy variables could also absorb some historical effect which is not expected to continue in the future. Though dummy variables can be used in such a way, it is still preferable to have an experience period devoid of such disruptive effects.

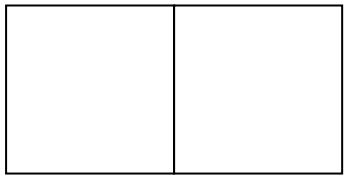
GLM claims datasets are typically either based on a certain policy year period or a certain calendar-accident year period. An example using the traditional parallelogram and rectangle diagrams illustrates the difference between the two.

Dataset A: policy year



PY 01 PY 02

Dataset B: calendar-accident year



CAY 01 CAY 02

Policy year: Annual policies written between 1/1/01 and 12/31/02, earned as of 12/31/03. Claims incurred on these policies before 12/31/03 but losses evaluated as of 6/30/04.

Calendar-accident year: Annual policies earning between 1/1/01 and 12/31/02 in respect of policies written between 1/1/00 and 12/31/02. Claims occurring on policies earning between 1/1/01 and 12/31/02, incurred losses evaluated as of 6/30/03.

There are benefits and disadvantages of each method of organization. The policy year approach has the advantage of relating to a certain period of underwriting and method of selling a product. The earning pattern of any given policy year, however, extends beyond the 12 month period. In order for policies to be fully earned, the cut-off date for exposures needs to extend 12 months (in the case of annual policies) or six months (in the case of semi-annual policies). In addition, the need for some IBNR emergence builds in more delay, resulting in data analyzed being not very recent.

The calendar-accident method of organization requires that each policy be split into its calendar year components (for example, an annual policy written on May 1 will be split into records defined by May 1 through December 31 and January 1 through April 30). Although this adds to system requirements and increases the number of records in the dataset, this allows the creation of an accurate calendar year "dummy" explanatory variable which can be used to absorb trends in claim experience which purely relate to time. If this is not possible, the policy year method of organization can be used, but the effect of any trends can be more difficult to identify.

Claims information

Claim count and loss amount information should be attached to the relevant exposure records, based on the most recent reserve estimates. The choice of definition of incurred claim count, specifically whether this pertains to number of claims or number of claimants, is not particularly important if ultimately the claim frequency and claim severity will be combined to the pure premium level. It is generally easier to model loss information net of deductibles, but should ideally not be truncated according to any large loss threshold at this stage since this allows sensitivity testing of several different large loss thresholds when modeling.

It is appropriate to leave some delay between the end of the experience period and the valuation date to allow for some IBNR claims to emerge and to allow for the case estimates to develop. If there is a regular (annual or quarterly) review of case estimates, or any other known issue surrounding the reserves, the experience period and valuation date should be selected to take advantage of the most accurate information.

The overall base level adjustment for pure IBNR and development of known claims will be made after models are finalized, but it is necessary to consider whether such time-related influences could bias the model rating factor relativities. There is a range of options for investigating the consequences of claims development upon the relativities measured, including:

- ignoring loss development and assuming that parameter estimates are unaffected
- including a dummy variable (eg calendar year or policy year[19]) in the model to absorb time-related influences; once models are finalized, the dummy variable is simply removed and the base levels are adjusted via a separate calculation (this assumes the development of claims is similar for all types of policy)
- before modeling the most recent experience, performing a series of GLM analyses on an older dataset which contains claims statistics as at various periods of development. By comparing GLM relativities based on data as at different development periods it is possible to assess whether claims development differs materially by type of risk - if they do it is possible to use the ratio of two models as at different development periods to derive multivariate development factors which can be applied to analyses based on a more recent dataset.

[19] Dummy variables based on quarters or months may contain an element of seasonality

It is also necessary to consider the treatment of claims closed without payment (also known as CWPs). Before modeling, it is generally most appropriate to remove such claims (setting the claim count field to zero in these cases), perhaps also creating a new claim type consisting of only CWPs (if they are to be modeled for expense allocation purposes). If CWPs are not excluded it can become difficult to model average claim amounts since some common GLM forms (eg those with gamma error functions) cannot be fitted to data containing observations equal to zero.

Generally, one period of policy exposure will have zero or one claim associated with it. Occasionally, there may be two or more accidents occurring in a given period of exposure. There are a number of alternative ways to deal with this situation:

- Multiple claims could be attached to the single exposure record, with the number of such claims and the total amount of such claims being recorded. This is the simplest method. A small amount of information is lost as a result of storing information like this, but such a loss is not generally material.
- Further records could be created in the database in the case of multiple claims. The exposure end date of the original record could be set to be the date of occurrence of the first accident, with the exposure start date of the second record being the day after. Each claim could then be attached to one exposure record (and the "number of claims" fields would always be zero or one). All rating factors recorded in the second record would be identical to the original record.
- Further records could be created in the database as in the second option above, but with the exposure dates in the original record remaining unaltered, and with the exposure start and end dates in the second (and subsequent) copied records being equal to each other, so that the additional records had zero days exposure recorded. (When analyzing claim amounts, the exposure information is not required, and when analyzing claim frequencies the experience could be summarized by unique combination of rating factor levels using an appropriate extract of the data, thus compressing this data to derive the correct exposure).

In practice, the easiest way to program the last two of these three methods produces one extra record for every claim, so policies with one claim would produce two records, and policies with two claims would produce three records. For example, using the second method, the exposure would be split at every claim date, so that there would always be one record with no claims (the last record).

General

In addition to volume requirements, how the model is to be used should also be considered. If the model were to be used to identify inaccuracies in the current rating plan, a line of business which undergoes significant rate intervention at point of sale would not be appropriate (unless being used to guide underwriters on the acceptable range of their intervention). Similarly, if little is collected or stored in the way of explanatory variables, this too would limit the strength of the GLM.

H Automated approach for factor categorization

One automated approach within the GLM framework is to replace a single factor with many levels with a series of factors each containing just two levels which are then tested for significance. For example instead of modeling age of insured with a single factor, a series of binary factors could be created:

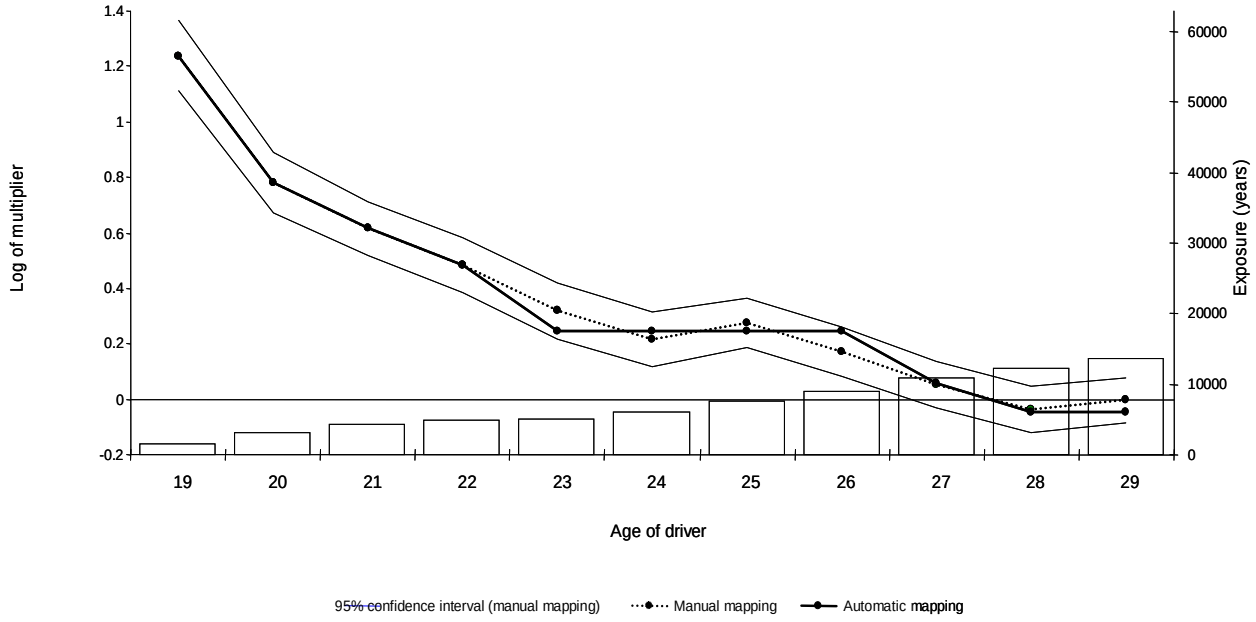
- (binary factor 1) is the age less than 18?
- (binary factor 2) is the age less than 19?
- (binary factor 3) is the age less than 20?
- (binary factor 4) is the age less than 21?
- …
- (binary factor 22) is the age less than 39?
- (age 40 is the base level in this example)
- (binary factor 23) is the age less than 41?
- …
- (binary factor 82) is the age less than 100?

These single parameter binary factors could then be tested for significance using an automatic stepwise algorithm as discussed in Section 2.

If, for example, ages 23, 24, 25 and 26 did not have a statistically different effect on the risk, the factors "is age less than 24", "is age less than 25" and "is age less than 26" would be deemed insignificant and excluded from the model. Those binary factors deemed significant in the model would determine the appropriate age categorization, and implied parameter estimates for each age could then be determined by summing the appropriate binary factors - eg in the above example the implied parameter estimate for "age 20" would be the sum of the parameters for binary factors 4 to 82).

An example result, based on real data, is shown below. The dotted line shows the fitted parameter estimates when age is not grouped (and when a parameter is allocated to each individual age rounded to the nearest integer). The solid line shows the parameter estimates implied by the results of the automatic grouping approach described above. Only results up to age 29 are shown for reasons of confidentiality.

Example of automatic grouping
(part result only - ages over 29, including base, not shown)



In this case it is not at all clear that the automatic approach produces a better categorization than a manual approach - for example it can be seen from the dotted line that age 23 has a parameter estimate between ages 22 and 24, and intuitively it appears wrong to group this level with ages 25 to 26 as the automatic process suggests. It is often the case that a manual approach to categorization can produce more appropriate results than an automated approach.

I Cramer's V

Cramer's V statistic is a measure of correlation between two categorical factors and is defined as

$$\sqrt{\frac{\sum_{i,j}\frac{(n_{ij}-e_{ij})^2}{e_{ij}}}{\min((a-1),(b-1)).n}}$$

where:

a = number of levels of factor one
b = number of levels of factor two
n_{ij} = amount of the exposure measure for the i^{th} level of factor one and j^{th} level of factor two
$n = \Sigma_{ij}(n_{ij})$
$e_{ij} = \Sigma_i(n_{ij}) \,.\, \Sigma_j(n_{ij}) / n$

The statistic takes values between 0 and 1. A value of 0 means that knowledge of one of the two factors gives no knowledge of the value of the other. A value of 1 means that knowledge of one of the factors allows that value of the other factor to be deduced. The two tables below show possible two-way exposure distributions of two categorical factors - each with only two levels, A and B, expressed as either rows or columns. The top table shows a Cramer's V statistic of 0, and the bottom table gives an example of a Cramer's V of 1.

	A	**B**
A	100	100
B	100	100

	A	**B**
A	100	0
B	0	100

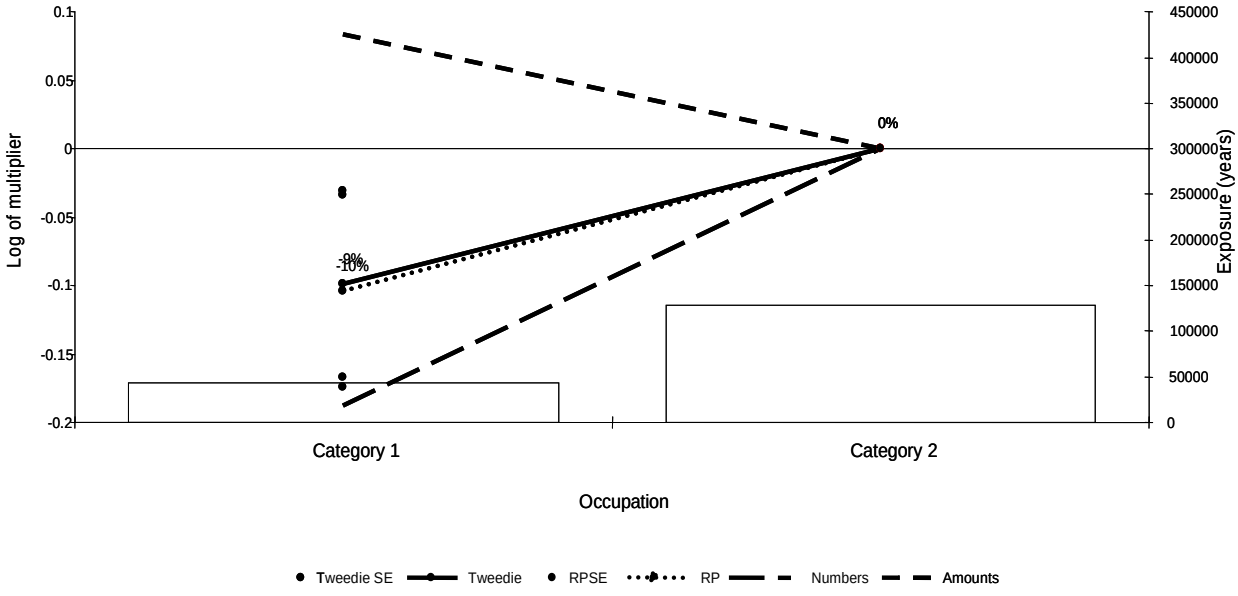
J Benefits of modeling frequency and severity separately rather than using Tweedie GLMs

Tweedie GLMs fitted to pure premium directly can often give very similar results to those derived by the "traditional" approach of combining models fitted to claim frequencies and claim severities separately. In these cases using Tweedie GLMs can reduce the amount of iterative modeling work required to produce satisfactory claims models.

The traditional approach, however, can provide a better understanding of the way in which factors affect the cost of claims, and can more easily allow the identification and removal of certain random effects from one element of the experience, for example via smoothing or by excluding certain factors from one of the frequency or amounts models.

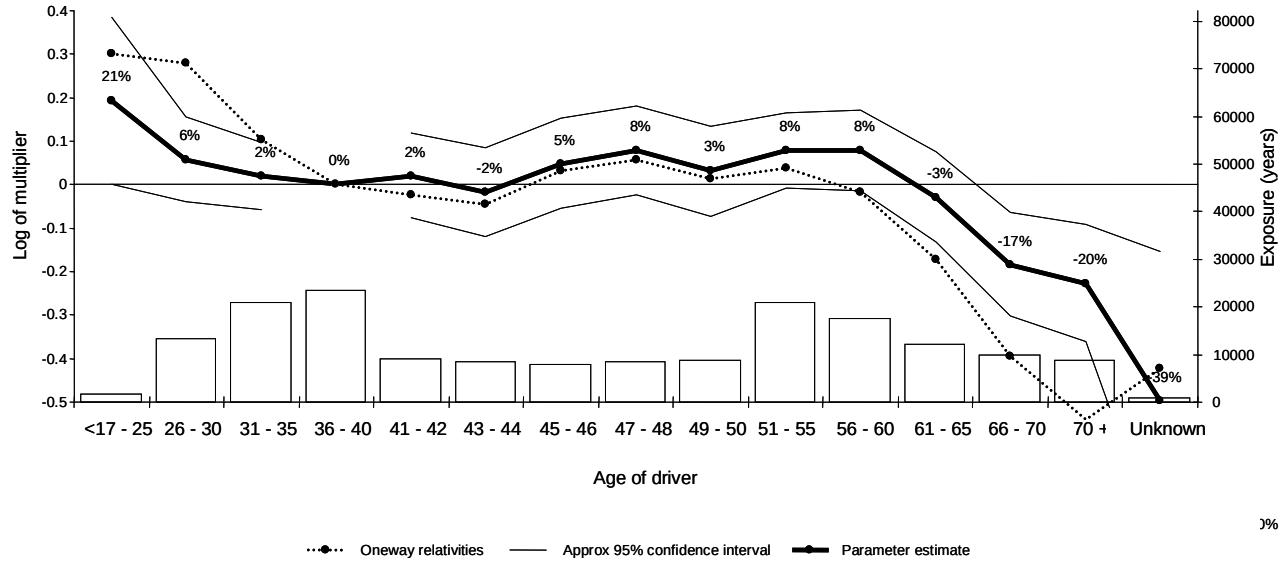
For example, the graph below compares the risk premium results from the Tweedie model to those from the traditional approach for one rating factor. Though the results between the two approaches are nearly identical, the traditional approach does provide additional information about the underlying frequency (numbers) and severity (amounts) effects - in this case the factor affects frequencies and severities in completely opposite ways.

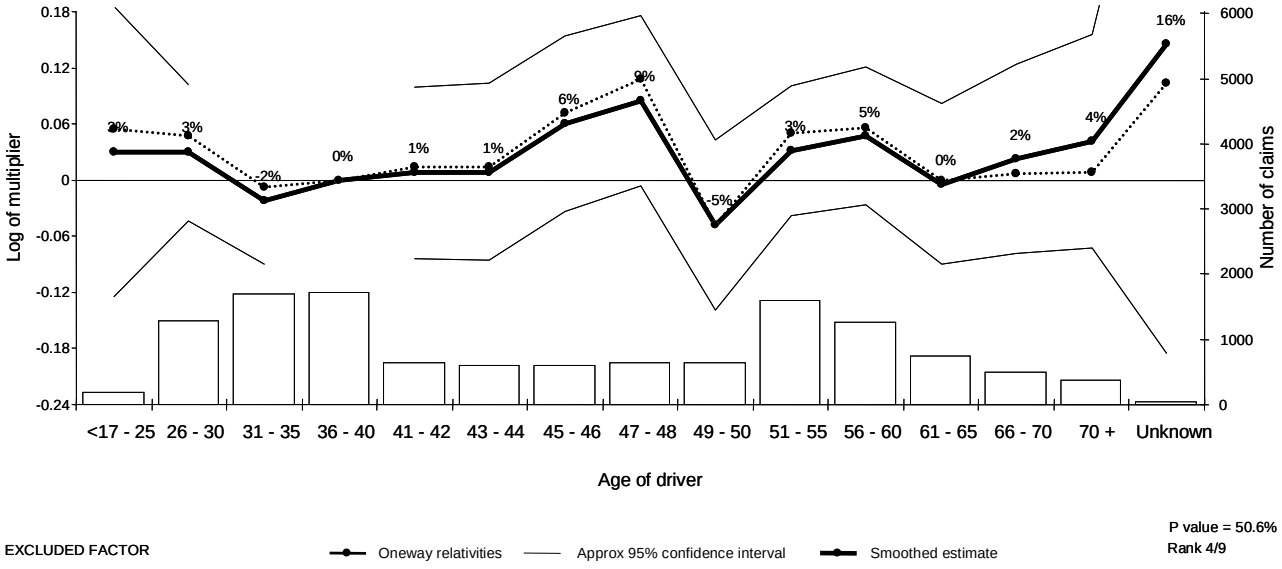
Comparison of Tweedie model with traditional frequency / severity approach



The three graphs below demonstrate a case where the results for a particular factor from a Tweedie GLM differ from those produced by the traditional approach. The first two graphs show the underlying frequency and severity model output from the traditional approach. Because of the wide standard errors, meaningless pattern, and insignificant type III test, the factor has been removed from the severity model. Consequently, the traditional risk premium reflects the underlying frequency experience only. The Tweedie model is more affected by the volatility from the underlying severity experience, and produces results which may be less appropriate.

Comparison of Tweedie model with traditional frequency / severity approach - traditional frequency and severity models





Comparison of Tweedie model with traditional frequency / severity approach