Using Generalized Linear Models to Build Dynamic Pricing Systems

Karl P. Murphy, Michael J. Brockman, and
Peter K. W. Lee

Using Generalized Linear Models to Build Dynamic Pricing Systems for Personal Lines Insurance

by

Karl P Murphy, Michael J Brockman, Peter K W Lee

1. Introduction

This paper explains how a dynamic pricing system can be built for personal lines business, whereby profit loads and risk premiums can be tailored to the individual behavioural characteristics of the customer.

The approach has been developed for a free and competitive rating regulatory environment, although the techniques can be reverse-engineered to provide customer value models in markets where rates are controlled.

We use the statistical technique, generalized linear models (GLMs), for estimating the risk premium and price elasticity components of the model. These techniques are well established in the British and European markets and are recently becoming more widely used in the United States.

The objective is to use as much information as input to these models in order to establish which risk factors are the most predictive. In our experience, every company is different; in their size and organisation, in their underwriting and marketing philosophy, in their claims handling, and in their means of distribution. It is essential that the pricing models built reflect these characteristics.

The paper is intended to be practical, hence we have kept the theory to a minimum, quoting other papers or literature where a more theoretical exposition is required. In particular, much of the contents of this paper follow on from the work of Brockman and Wright which sets out the theory of generalized linear models and its application for personal lines pricing. Since Brockman and Wright, the use of GLMs has become much more common. Whilst GLMs are being widely utilized in the UK and Europe, we do not believe that the results are being fully exploited. In this paper we will be explaining how the results of GLMs can be more effectively employed.

The structure of the paper will be as follows: firstly we are going to discuss the main statistical technique used for predicting behaviour (both claims and demand), namely generalized linear models. We will then look at how GLMs can be used in assessing the risk premium for a policy, followed by a discussion on using GLMs to assess price elasticity curves. Next we will take an overall look at combining the supply and demand models. Finally, we will describe a possible pricing algorithm which brings both the cost and demand sides of the equation together.

2. Generalized Linear Models

2.1. *Traditional rating versus a multiple regression approach*

The insurance industry is quite unique in that, unlike most manufacturing companies, both the cost of writing a policy and the demand for the product are highly dependent on the characteristics of the individual to whom the policy is sold. In addition, the range of factors which affect both claims experience and demand is very large, and this creates problems when deciding on what pricing differentials to apply between different groups.

The traditional way to determine relativities by rating factor is to look at series of one-way tables, either focusing on the relative risk premiums or the relative loss ratios. Holler, Sommer and Trahair (section 1.2) provide a nice example of why one-way reports give very misleading results because of the different mix of business underlying the levels of each factor. This problem should be fairly familiar to most readers, so we do not propose to dwell on it here.

One solution is to use some form of multiple regression approach which removes any distortions caused by different mixes of business. A flexible approach, and which we have found to be very useful in practice, is a regression method known as generalized linear models (GLMs). Many different types of models which suit insurance data fall under this framework, and the more familiar "classical" regression (which assumes that the response variable is a linear combination of the explanatory variables and is normally distributed) is a subset of GLMs.

The additional benefit of using GLMs over one-way tables is that the models are formulated within a statistical framework. This allows standard statistical tests (such as χ^2 tests and F tests) to be used for comparing models, as well as providing residual plots for the purpose of model diagnostic checking.

2.2. *Specification of GLMs*

We outline below the fundamental algebra behind GLMs. The notation that we use is slightly informal, in that we have omitted estimate indicators ^ as well as omitting vector indicators, but this should not detract from the key results.

GLMs take the following form:

$y = h(X\beta) + \text{error}$

where h is a monotonic function, and y is distributed according to a member of the exponential family of distributions. This family includes the normal, Poisson, gamma, inverse Gaussian and binomial distributions. As we will see later, the results can be extended to a more general family.

Notice that this gives a much more flexible model form than the classical case. y is now a function of a linear combination of the explanatory variables rather than just a direct linear combination, and y is not constrained to be normally distributed.

On a point of terminology, X is known as the design matrix. The dimensions of X are n by p, where n is the number of observations available, and p is the number of parameters being estimated. The parameters can either be continuous variables or categorical variables, a distinction which we will detail later. The algebra is the same for both types of variables.

$X\beta$ is know as the linear predictor, and is a straightforward linear combination of the estimated parameters. The linear predictor is usually denoted by η, and is of dimensions n by 1.

$h(X\beta)$ is known as the fitted values, and simply transforms the linear predictor. It is usually denoted by μ.

We discuss the main features of GLMs below.

2.3. Link function

Although the model is usually specified as $y = h(X\beta) + \text{error}$, it is usual to refer to the inverse of h as being the link function, and this is usually denoted by g. If we want to fit a multiplicative model, then $h(x) = \exp(x)$, so $g(x) = \log(x)$, so we have a log link function.

Other common link functions include the identity link function ($h(x) = x$, $g(x) = x$) and the logit link function ($h(x) = 1 / (1 + \exp(-x))$, $g(x) = \log(x / (1 - x))$).

Notice that by fitting a model with a log link function, we ensure that the fitted values are non-negative, and by fitting a model with a logit link function the fitted values fall between 0 and 1. These properties are often very appropriate for insurance data – risk premiums are always non-negative and renewal rates always fall between 0 and 1. Under classical regression, the fitted values can take on any value.

2.4. Variance function

As we will see when it comes to estimating the parameters β, the key feature of the distribution of y is given by the variance function.

The variance function links the variability of y to the fitted values μ. This is the second key difference with classical regression. Under classical regression, the variance of y is constant, no matter what the fitted value. With GLMs, the variance can vary with the fitted values.

Common variance functions are as follows:

Normal:	1 (i.e. constant variance)
Poisson:	μ
Gamma:	μ^2 (i.e. constant coefficient of variance)
Inverse Gaussian:	μ^3
Binomial:	$\mu\,(1 - \mu)$

These variance functions should be quite intuitive to anyone familiar with these distributions.

The property of non-constant variance is often quite appropriate for insurance data. Suppose, for example, a model is being fitted to the average cost of claims. Under a normal error structure, the standard error for an observation with a fitted value of $100 might be (say) $10. Equally, the standard error for an observation with a fitted value of $1,000 would also be $10, as it would be for an observation with fitted value $10,000. It is much more intuitively appealing to have a proportionate standard error, so the standard error for a fitted value of $100 might be 10% (i.e an absolute standard error of $10), for $1,000 would also be 10% (absolute standard error of $100), as it would be for $10,000 (absolute standard error of $1,000). This proportional standard error model has a constant coefficient of variance, and so is a gamma model. In this way, therefore, the variance function can be initially selected based on knowledge of the likely variance relationship of the process being modelled. This can then be tested statistically using standard diagnostic checks. We have found from experience that the correct selection of variance function significantly improves the robustness of the parameter estimates.

2.5. Estimation of the β parameters

β is usually estimated via an iterative process, where the r^{th} estimate of β is given by the following equation:

$$\beta^r = (X^T W^{r-1} X)^{-1} X^T W^{r-1} (z_i^{r-1} + \eta_i^{r-1})$$

where

W^{r-1} (n by n) = diag{ $1 / [g'(\mu_i^{r-1})^2\, V(\mu^{r-1}{}_i) / w_i]$ }

$V(\mu^{r-1}{}_i)$ = variance function

$\eta^{r-1} = X\,\beta^{r-1}$

$\mu^{r-1} = h(\eta^{r-1})$

z_i^{r-1} (n by 1) = $g'(\mu_i^{r-1})\,(y_i - \mu_i^{r-1})$

w_i = weight vector

This equation looks quite complicated, but simplifies quite considerably for many common models.

For example, for a normal model with an identity link function, the W matrix simplifies to a diagonal matrix of the observation weights, and $z_i^{r-1} + \eta_i^{r-1}$ reduces to y, and so we get the familiar general linear model (as opposed to *generalized* linear model) equation:

$\beta = (X^T W X)^{-1} X^T W y$

For a Poisson model with a log link function, the diagonal entries of W reduce to $\mu_i^{r-1} w_i$, and z_i^{r-1} simplifies to $(y_i - \mu_i^{r-1}) / \mu_i^{r-1}$.

For a gamma model with a log link function, the diagonal entries of W reduce to w_i, and z_i^{r-1} simplifies to $(y_i - \mu_i^{r-1}) / \mu_i^{r-1}$ (as for the Poisson model).

The iterative process continues until a statistic known as the *deviance* converges. The deviance is a function of the log likelihood statistics, and is a function of the y values and the fitted values μ. The analogous statistic under classical regression is the sum of squared errors. The concept of "deviance" is described in McCullagh and Nelder (p 33).

2.6. *Extending the range of variance functions*

We listed in section 2.4 above the variance functions for most of the common distributions. The theory works just as well for a slightly broader range of distributions, whereby the variance function takes the form μ^{α}, where α can be any specified value. This family can be useful if any diagnostic checks on the goodness of the model indicate that the fitted distribution may not be appropriate. A value of α between 1 and 2 is often used to model risk premium data; this corresponds to the Tweedie distribution. Mildenhall (sections 8.3 and 8.4) discusses this broader range of distribution further.

2.7. *Categorical and continuous variables*

The use of categorical variables allows separate parameters to be fitted for each level of a rating factor. For example, we can fit a separate parameter for each territory in the rating structure; if we used a continuous variable we would be imposing some sort of relationship (such as a linear or polynomial relationship) across territories, and this may not be appropriate. Indeed, some factors may not have any natural continuous relationship to use.

Factors that have a natural continuous scale can also be treated as categorical variables. For example, a policyholder age scale can be grouped by bands of ages, and each band considered to be a categorical level. If the trends in the categorical level parameters indicate that a continuous scale may be appropriate, then the categorical variable can be converted to a continuous variable. An example of this is given below.

In terms of the design matrix, when fitting a categorical variable we need a column for each level of each rating factor being fitted (with a slight adjustment as described below). The columns then take on the value 1 or 0, depending on whether the particular observation contains that level or not. If we included a column for each level of the rating factor, the design matrix would contain a linear dependency (because for any one categorical factor the sum of all levels would sum to 1), and so one level is omitted from each categorical variable. This level is known as the base level;

we usually choose this level to be the one with the greatest volume of business. In a sense, the base level is subsumed into the first column of the design matrix (the intercept term).

For example, if we are fitting a model with two factors, the first being territory (with levels A, B and C), and the second being policyholder age (grouped into <25, 25-39, 40-59, 60+), then the design matrix would look something like:

Intercept	Territory B	Territory C	Policyholder Age < 25	Policyholder Age 25-39	Policyholder Age 60+
1	0	0	1	0	0
1	1	0	1	0	0
1	1	0	0	1	0
1	0	1	0	1	0
1	0	1	0	0	0
1	0	0	0	0	0
1	1	0	1	0	0
1	1	0	0	0	0
1	1	0	0	1	0
1	1	0	0	0	1
1	0	1	1	0	0
1	0	1	0	1	0
1	0	0	0	0	1
1	1	0	1	0	0
1	1	0	0	1	0
1	1	0	0	0	1

The base level for Territory has been chosen as level A, and the base level for policyholder age has been chosen at level 40-59. The choice of base is somewhat arbitrary, but is often taken to be the level with the most numbers of observations. The parameter estimates are only affected in that they are now relative to a different base; the fitted values are identical.

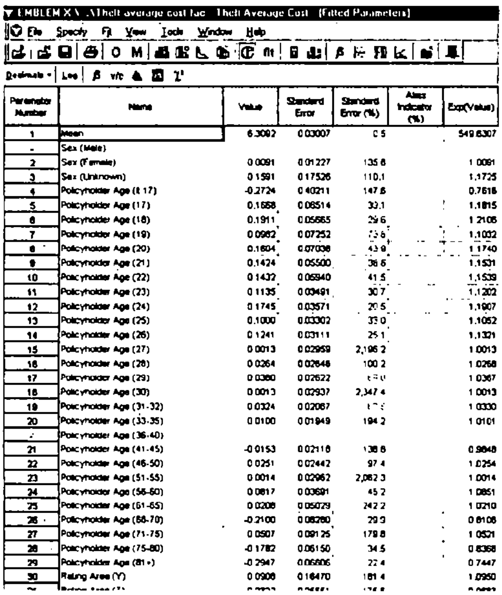
If having estimated the parameters for policyholder age it was felt that the parameters were broadly linear, then the best straight line relationship can be derived by changing the design matrix. If the average policyholder ages within each level are 22, 32, 50 and 69 (say) then the design matrix changes to:

Intercept	Territory B	Territory C	Policyholder Age
1	0	0	22
1	1	0	22
1	1	0	32
1	0	1	32
1	0	1	50
1	0	0	50
1	1	0	22
1	1	0	50
1	1	0	32
1	1	0	69
1	0	1	22
1	0	1	32
1	0	0	69
1	1	0	22
1	1	0	32
1	1	0	69

In other words, we are replacing the categorical (three parameter) policyholder age scale with a (one parameter) linear relationship.

2.8. Practical example

Below is an extract from the output of a generalized linear model fitted to an average cost model.



Parameter Number	Name	Value	Standard Error	Standard Error (%)	Alias Indicator (%)	Exp(Value)
1	Mean	6.3092	0.03007	0.5		549.6307
-	Sex (Male)					
2	Sex (Female)	0.0091	0.01227	135.6		1.0091
3	Sex (Unknown)	0.1591	0.17526	110.1		1.1725
4	Policyholder Age (lt 17)	-0.2724	0.40211	147.6		0.7616
5	Policyholder Age (17)	0.1668	0.06514	33.1		1.1815
6	Policyholder Age (18)	0.1911	0.05665	29.6		1.2106
7	Policyholder Age (19)	0.0962	0.07252	75.6		1.1002
8	Policyholder Age (20)	0.1604	0.07038	43.9		1.1740
9	Policyholder Age (21)	0.1424	0.05500	36.6		1.1531
10	Policyholder Age (22)	0.1432	0.05940	41.5		1.1539
11	Policyholder Age (23)	0.1135	0.03491	30.7		1.1202
12	Policyholder Age (24)	0.1745	0.03571	20.5		1.1907
13	Policyholder Age (25)	0.1000	0.03302	33.0		1.1052
14	Policyholder Age (26)	0.1241	0.03111	25.1		1.1321
15	Policyholder Age (27)	0.0013	0.02959	2,196.2		1.0013
16	Policyholder Age (28)	0.0264	0.02646	100.2		1.0268
17	Policyholder Age (29)	0.0360	0.02622	[illegible]		1.0367
18	Policyholder Age (30)	0.0013	0.02937	2,347.4		1.0013
19	Policyholder Age (31-32)	0.0324	0.02067	[illegible]		1.0330
20	Policyholder Age (33-35)	0.0100	0.01949	194.2		1.0101
-	Policyholder Age (36-40)					
21	Policyholder Age (41-45)	-0.0153	0.02118	138.6		0.9848
22	Policyholder Age (46-50)	0.0251	0.02447	97.4		1.0254
23	Policyholder Age (51-55)	0.0014	0.02962	2,082.3		1.0014
24	Policyholder Age (56-60)	0.0817	0.03691	45.2		1.0851
25	Policyholder Age (61-65)	0.0208	0.05029	242.2		1.0210
26	Policyholder Age (66-70)	-0.2100	0.06280	29.9		0.8106
27	Policyholder Age (71-75)	0.0507	0.09125	179.8		1.0521
28	Policyholder Age (75-80)	-0.1782	0.06150	34.5		0.8368
29	Policyholder Age (81+)	-0.2947	0.06606	22.4		0.7447
30	Rating Area (Y)	0.0908	0.16470	181.4		1.0950

The model fitted has a log link function and a gamma error structure. Because of the log link function, the parameters are on a log scale (the final column calculates the exponential of the parameter values).

The "mean" parameter value (6.3092) represents the intercept parameter, and the exponential of this number (549.6) is the fitted value for a base risk. For this particular model, a base risk is a male aged 36 to 40 (plus appropriate bases for the factors not shown in the extract).

To derive the fitted value for a non base risk, simply add up the parameters for the appropriate levels and then take the exponential this number. For example, the fitted value for a female age 23 is exp(6.3092 + 0.0091 + 0.1135) = 621.3. We can derive the relative claims experience for any one factor by simply looking at the exponential of the parameter. For example, the female relativity is 1.0091 (=exp(0.0091)) compared to males.

The "Standard Error" column gives us an idea of the variability of the parameter estimates.

2.9. Other GLM concepts

We have described above the key concepts underlying generalized linear models. There is naturally considerably more theory underlying the process, as well as other concepts, which we have not discussed.

Concepts which we have not mentioned include residuals graphs, offsetting models to fix parameter values, joint modelling techniques, model diagnostics, dealing with sparse data, and statistical tests. The interested reader is referred to McCullagh and Nelder for a much more thorough explanation.

3. Risk Premium Modelling

Brockman and Wright explore in some detail the use of generalized linear modelling for risk premium analyses. We summarize below some of their main conclusions, and discuss in more detail issues concerning the modelling process and allowing for large claims.

3.1. Splitting the risk premium

Brockman and Wright (sections 2, 3 and 5) recommend that the risk premium is split down into frequency and severity (average cost) by each type of claim covered under the policy. The reasons for doing this include, for example:

- response variables for frequency and severity follow different statistical distributions. Usually a Poisson error structure is used for a frequency model and a gamma error structure for a severity model. A log link function is usually used for both frequency and severity
- a greater insight into the underlying cause of claims experience variability is provided
- certain models are inherently more volatile that others. For example, the average cost of liability claims is likely to be much more volatile than the frequency of own damage claims. By modelling total risk premium rather than splitting it into its constituent parts, we would not be able to identify whether an apparently anomalous trend is the result of a random fluctuation in liability average cost or a genuine trend in the own damage frequency

In a sense, by "best" fitting models with the least parameters to each component process, we are maximizing the information being extracted from (often sparse) datasets.

3.2. Risk premiums and GLMs

Risk premium modelling fits very naturally within the generalized linear model framework, especially when split into its constituent parts (i.e. frequency or average cost by claim type).

The response variable is taken to be the frequency or average cost, with the rating factors available used as the explanatory variables. Usually the rating factors are initially classified as categorical variables; it is a simple task to convert them into continuous variables if it is appropriate, as is described in section 2.7. In addition, we recommend that "time" is included as a rating factor (usually accident year or accident quarter). The parameters indicated by this factor will provide estimates of inflation rates and changes in frequency levels from period to period. Furthermore, we can use the time parameters to test the consistency of parameter trends over time. We discuss this further in section 3.3.

It is possible to build factors into the analysis which are not directly asked of the policyholder, or that are collected but not used in the rating tariff. These might include, for example,

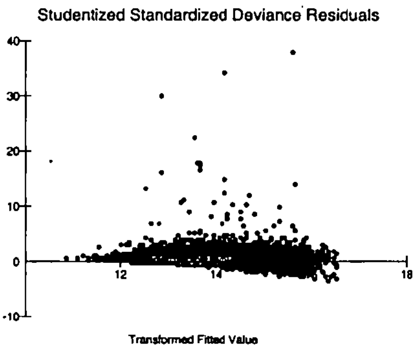
- the length that they have been with the company
- whether or not they have other policies with the company
- the distribution method that attracted the policyholder
- information from external data sources (such as socio-demographic data or credit rating information)

Frequency models are usually fitted assuming a log link function with a Poisson error structure. Claim frequencies fall quite naturally into a Poisson process, and we have found empirically that this is an appropriate model. Using a log link function ensures that the fitted frequencies are positive. Brockman and Wright (section 2.1) go into further detail about why a multiplicative model is appropriate for frequency models.

Average cost models are usually fitted assuming a log link function with a gamma error structure. Once again using a log link function ensures positive fitted values. By having a gamma error structure we have a constant coefficient of variance, and so the standard errors around the fitted values are proportional to the fitted values. For example, if for a fitted value of \$100 the standard error is \$10, then for a fitted value of \$1,000 the standard error would be \$100.

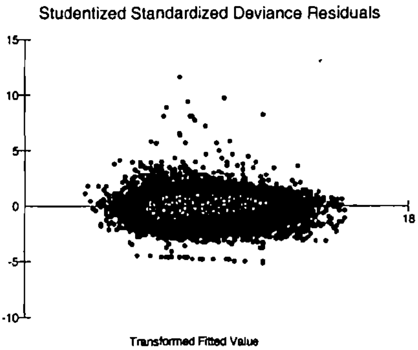
The appropriateness of the error structure can be tested by looking for any evidence of heteroskedasticity within the model residuals. If there is any evidence of the residuals fanning inwards or outwards when plotted against the fitted values, then a different error structure may be more appropriate.

For example, below is a graph showing the deviance residuals for an average cost model (fitted using a Normal error structure). The residuals appear to be fanning outwards, so the variability of observations with higher fitted values is greater than those with lower fitted values. The variance function for a Normal model is 1 (i.e. μ^0), and the shape of the residuals suggests using a variance function with a higher power (such as μ^2).



Deviance residuals are explained in McCullagh and Nelder (p 39). The residuals have been "studentized" and "standardized" (see McCullagh and Nelder section 12.5). The x-axis is a transformation of the fitted values for each observation (in this case, twice the log of the fitted value).

The residuals for a gamma model fitted to the same data are shown below. The obvious fanning outwards has been removed, and this graph would suggest that the gamma model is an appropriate error structure for the data.



3.3. *The modelling process*

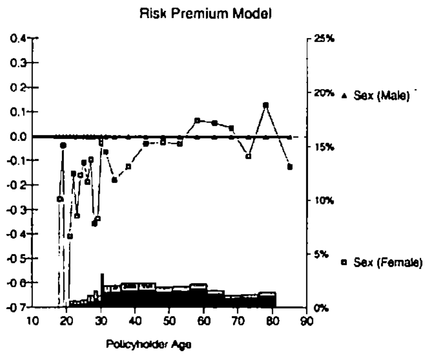
As with any statistical model, we want to find the model which replicates the data quite closely, yet contains as few parameters as possible, i.e. we wish to find the most parsimonious model.

The steps involved in doing this include the following:

- removing rating factors from the model which do not seem to affect variations in claims experience
- grouping levels within rating factors to reduce the number of parameters fitted (e.g. grouping policyholder age into under 25s and over 25s)
- fitting curves to variables which have a natural continuous scale
- making the model more complicated by fitting interaction or combination terms

It is interesting to dwell a bit more on this last point. If within the data there is an interaction effect, then we want to reflect this in the model, even though it involves increasing the number of parameters in the model. An interaction effect occurs when the relative claims experience of a particular factor depends on the level of another factor. For example, a common interaction to find is between policyholder age and gender, whereby the claims experience for young females is considerably better than that for young males, but the experience of older females is either the same or worse than for older males. If an interaction term were not included then there would be a constant difference between males and females across all ages.

The graph below shows the relative claims experience for females compared to males by policyholder age for an overall risk premium model. Although there is an element of randomness, there is a strong upward trend by policyholder age (i.e. the experience of females relative to males worsens with increasing age). A simplification to the interaction would probably be made, to smooth out the random effect.



Interactions are usually tested before simplifying the model; otherwise there is a danger of removing a factor which should be included in an interaction.

There are several ways of assessing whether a simplification to the model (or indeed a complication via fitting interactions) is appropriate. These include:

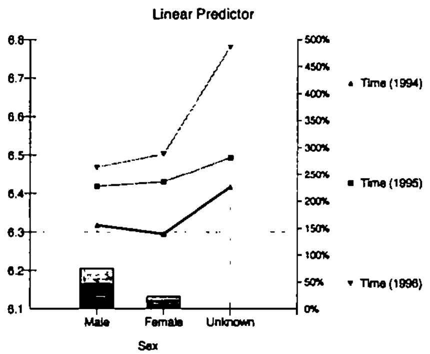
(i) looking at the relative size of the parameters. If the model is only indicating a 0.5% difference in experience between two levels of a rating factor, this would probably be judged to be not significant

(ii) looking at the standard errors of parameters. As a rough rule, if the parameter estimate is within two standard errors of the base rate, then there is no statistical difference between the two

(iii) doing a formal statistical test between two nested models (i.e. where one model is a subset of the other). This is achieved by comparing the change in deviance with the change in degrees of freedom between the two models. If model 2 is a subset of model 1, then model 2 will have a larger deviance, but more degrees of freedom. Our recommended statistical test is to compare the change in deviance with a χ^2 distribution with degrees of freedom being the difference in degrees of freedom between the two models

(iv) fitting an interaction between the factor being considered and the time factor to see the consistency of any trends between different accident years. If the trends are inconsistent over the period being investigated then it is questionable as to whether the trend will exist in the period to which the model is being projected. Equally a consistent trend over time will give more confidence that the same trend will occur in the future.

The graph below shows an interaction fitted between time and gender for an average cost model. Each line represents a separate accident year.



For two out of the three years, the male experience is better than the female experience, but for the other year females have better experience. It is unclear as to what the relative experience is going to be in the following year.

Sometimes a particular trend will be shown to be significant under one test, but not for another. Which test to believe is entirely a matter of judgement. Perhaps the one over-riding criteria is that any trend must make sense and there has to be a rational explanation for it. A good question to ask when making these judgement calls is whether the trend could be explained to an underwriter and whether he/she would believe it.

Another practical point is that eventually the model will be populating a rating tariff, and so the modeller should bear in mind the structure of the tariff when doing the modelling.

3.4. Combining models

Once each frequency and average cost model by claim type has been modelled, it is usual to combine the separate component models to generate one final model.

This can be achieved by calculating the fitted values for each component model for each observation, adding each element together to get a total risk premium, and then fitting a final smoothing model to this total risk premium.

It is usually necessary to make adjustments to the individual components before combining. These adjustments include:

(i) adjusting frequencies and average costs to make allowance for any under- or over-reserving within the data. In addition, we would usually allow for a lag between the period of investigation and the data extraction date so that most claims will at least have been reported
(ii) adjusting frequencies and average costs for projected future trends (i.e. frequency changes or inflation).
(iii) a large claim adjustment (especially if the period of investigation does not contain typical large claims experience). We discuss large claims in more detail in section 3.5.

For the purposes of combining the models, we usually update the claims experience to the latest accident year and project on from there. Effectively at this stage we are not treating time as a rating factor.

The trends from the risk premium model quite naturally are very smooth because each individual component model has been smoothed, and any residual plot should show little variation between the actual and fitted values. Any interactions present in the component models will also be present in the risk premium model. In addition it is possible that additional interactions will be generated from the combining process. This is because the proportions by each claim type across a factor may vary according to the level of another factor.

At this final smoothing stage, it may be necessary to fix the relativities of some factors as certain parts of the rating tariff cannot be changed (perhaps for regulatory or marketing reasons). This process is known as "offsetting". By offsetting we are fixing the parameters for one or more factors, and allowing the other factors to find their correct relativities given this offset. A common example of this in the United Kingdom is offsetting the no claims discount (or bonus malus) scale. Many companies have a fixed scale, and underwriters are reluctant to adjust it. No claims discount (NCD) is very highly correlated with policyholder age, and so by offsetting the NCD scale the policyholder age scale will adjust to compensate. There is a danger that if NCD is not offset, but the fixed scale is used in the rating tariff, then certain segments will receive a double loading or double discount. The effect of the offsetting depends on the degree of correlation with the other factors and the difference between the "theoretical" and offset scale.

3.5. Large claims

The average cost of liability claims is highly variable (especially in the United Kingdom where there is no limit on the size of liability). This makes it difficult to pick out trends by rating factor when modelling the liability average cost.

One way of dealing with large claims is to separate out large and small claims. Fitting a model to the small claim average cost will give more stable results than modelling total average cost.

An additional large claim loading is then required so that the total risk premium is not understated. One method might be to simply increase the small claim average cost by the ratio of total liability average cost to small liability average cost. However, this ignores the possibility of certain segments having a higher propensity to have large claims (for example, young people might have a higher propensity to have large claims than older people). To test this we can formulate a GLM with the following structure. The rating factors are once again the explanatory factors. This time, however, the process is a binomial process, with the number of trials being the number of liability claims (both small and large), and the response variable being the number of liability claims that are large claims. If any difference is shown by any of the rating factors, then the adjustment needs to be made factor-dependent.

Note that it is only necessary to make the adjustment factor-dependent if for certain segments a higher proportion of liability claims become large claims. For example, young people are likely to have a higher liability claim frequency than old people. The factor-dependent adjustment is only required if a higher proportion of those claims are large claims.

3.6. *Building in external data*

It is always possible to supplement the knowledge known about the policyholder by attaching on information from external sources (such as socio-demographic information or credit rating). This is often achieved from the policyholder's zip-code. The external factor is then treated just like any other rating factor.

Many of the external databases available are quite highly correlated with other rating factors already included in the existing rating structure. When evaluating the significance of this information it is essential to do it within a multi-variate framework. This ensures that correlation issues are properly addressed and understood. Simply looking at a one-way table will give very misleading results if the external information is correlated with other factors.

3.7. *Data volumes and sparsity issues*

Our philosophy when fitting GLMs is to include as many factors as is practical into the model and not to group levels within factors to any great extent. The effect of this is to generate quite large data-sets when modelling claim frequencies, with most of the cells having no claims in them. Those that do will usually only have one claim associated with them. We therefore have a sparse data-set.

Sparsity is not necessarily a problem provided that the model is correctly specified and we have found generally that the parameters from a GLM are robust under sparsity conditions. Standard errors produced, however, are less robust, and care needs to be placed when interpreting them. Furthermore, model diagnostic checking is more difficult since the residual plots will not follow the classical shape of being symmetrically spread around zero. Rather most residuals will be small and negative (i.e. those cells which have no claims), and the rest will be large and positive. However, this is not to say that the model is a poor fit – this is the expected shape of the residuals.

Some people believe that the data should be reduced in order to "improve" the look of the residual plots and produce good fitting models, either by limiting the number of factors in the model or by restricting the number of levels within a factor. The extreme example of this is to reduce the data down to one point, the mean and fit a model to this. By definition it will produce a perfect fitting model, but this tells us nothing about the true within-cell variation in claims experience.

By collapsing the data (by reducing the number of factors in the data, or grouping levels within factors to broader bands) we are deceiving ourselves that we have a better statistically fitting model. The within-cell variation arising from individual exposures will be hidden. More importantly, detail about the variations by rating factor is lost.

We believe, therefore, that the analysis should be carried out with as many relevant risk factors in the model as possible. We will typically have around thirty factors in our models for risk premium and renewal/conversion analyses. The model parameters can then be reduced intelligently as part of the analysis rather than by making prejudgements of the data before the analysis is carried out.

Residual plots can still be used for model validation by simply collapsing the residuals after the model has been fitted using the collapsed fitted values and actual values in broader cell definitions. These residuals will be the same as if the data had been collapsed in the first place, and we do not lose the detail of the uncollapsed model.

4. Estimating Price Elasticities

Section 3 above concentrated solely on estimating relative claims experience by rating factor (essentially the supply side in economic terms). This is the traditional area with which actuaries have been involved.

Just as important to the whole business process, however, is the ability to understand how customers respond to price and price change (i.e. the demand side of the equation). Similar techniques to those used for estimating risk premiums can also be used to estimate price elasticity functions for individual customers, and to identify areas of the account with good or bad conversion/renewal experience.

There are two aspects to demand-side modelling, namely new business (or conversion) demand and renewal demand. There are some differences in the types of analyses performed on each, but the fundamental framework is the same for each.

4.1. Demand modelling as a generalized linear model

Renewal rates and conversion rates once again fall naturally into the generalized linear modelling framework. The explanatory factors are the rating factors, plus additional factors which are discussed in more detail below. The response variable is whether or not the policy converts/renews.

The link function is usually taken to be a logit link function, so the model is of the form

$$y = 1 / (1 + \exp(-X\beta)) + \text{error}$$

By having a logit link function we are forcing the fitted values to fall between 0 and 1 (as they should do). Alternative link functions include the probit link function and the complementary log-log function, although in practice the chosen link function makes little difference to the fitted values. The response variable is quite naturally a binomial (0/1) response.

The process of simplifying the model is the same as that for modelling the risk premium element, in that we are trying to find the most parsimonious model for the data. Similar statistical tests and judgmental issues need to be considered.

Because of the 0/1 nature of the response variable, the results from demand models tend to produce less variability than that in risk premium models, and the results derived are often quite smooth. Less data is necessary to generate very good results; indeed demand information is much more dynamic than risk premium information, so it is important that the experience used is not too out of date to be useful.

4.2. *Conversion modelling*

The modelling of conversion (or new business) rates can only be achieved effectively if full information of both quotations that converted to policies and quotations that were declined is known. Currently this can only be achieved effectively by direct response operators; usually companies operating through intermediaries or agents only find out about the quotations that are converted to policies.

Information available includes all questions asked during the quotation (i.e. the traditional rating factors), plus the premium offered. Other information that the insurer might know about the proposer includes:

- information provided by external data providers (such as socio-demographic or credit information)
- the source of the business
- whether or not the proposer has existing policies with the insurer
- information about the market premium available to the proposer (we will discuss this in further detail when discussing renewals)
- methods of payment
- marketing/campaign information

The explanatory factors in the model are the rating factors that are asked of the customer plus any additional information, for example, those listed above. The response is 1 or 0 depending on whether the quotation was accepted or not.

One question about the modelling process is whether or not the premium quoted should be included in the model. The answer to this question depends on the purpose of the analysis. If the purpose of the analysis is to derive price elasticity functions, then the quoted premium should be included in the analysis, and the derived parameters represent the elasticity curve. However, if the purpose of the analysis is to identify the relative conversion rates by segment, then usually only the rating factors would be included, and the parameter estimates derived will indicate the relative conversion rate by segment.

There are some practical considerations to bear in mind when extracting data for a conversion analysis. These are briefly outlined below.

(i) Often one proposal will result in several quotations, for example if the proposer requests alternative quotes under different deductibles. For the accepted quotes we know what deductible is taken, but for the declined quotes we do not. For these quotes a consistent rule is required so as not to double-count quotations (e.g. take the first quotation generated for the declined quotes).

(ii) If quotations are guaranteed for a period of time, then a time delay is required before doing an analysis to allow for development lags.

(iii) Many insurers will refuse to provide a quotation on certain risks that do not fall within their underwriting guidelines. Usually these should be identified and removed from the analysis.

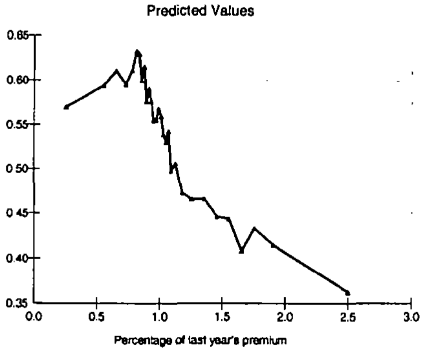
4.3. Renewal modelling

Renewal modelling is in many ways similar to conversion modelling, but usually more information is known about existing policies. In addition, all insurers will know the details of renewals invited and accepted, and so a renewal analysis is not confined to those insurers (mainly directs) who have information about accepted and declined quotes.

Critical to the renewal behaviour of the policyholder is the price movement over the previous year, and so one of the explanatory factors used will be the ratio of this year's premium to last year's. This is a continuous variable, but as discussed earlier this would usually be initially banded into a categorical variable.

It may be necessary to adjust the ratio to allow for the policyholder's expectations of price movements. For example, the policyholder might expect a significant price decrease if they have hit what they perceive to be a critical age. If this decrease is not given, they may shop around elsewhere for alternative quotes. It is impossible to determine exactly what the policyholders' expectations are, but one possible adjustment is to calculate the ratio of this year's premium using this year's rating factors to this year's premium using last year's rating factors (assuming that the policyholders' expectations are reflected in the rating structure). In the UK, the most important adjustment would be for movements up or down the no claims discount (or bonus/malus) scale because policyholders are very aware of like premium movements as the result of having or not having a claim during the policy year.

The graph below shows the shape of a typical elasticity curve. The x-axis is the percentage change in premium on renewal, the y-axis the renewal rate. The graph is plotted for a particular customer segment and this determines the level of the graph. The shape of the graph can also depend on customer segment if the model includes segment related elasticity interaction effects.



Notice that where there has been a large price decrease, the curve tends to flatten off (below 0.8). This is because no matter how cheap this year's premium is, there will always be some policyholders who do not renew (for example if they no longer require auto insurance). Indeed, if they have a very large price decrease they may feel aggrieved by the premium they paid the previous year and seek insurance elsewhere anyhow. Equally, some policyholders are very price inelastic, and for them they will accept virtually any price increase. The steepest part of the graph is around 1 where the premium change has been very small. This makes intuitive sense; the likely impact of a 55% premium increase compared to a 50% premium increase is going to be less than a 10% increase compared to a 5% increase.

Another important input into the model looks at the ratio of the premium quoted to an alternative "market" premium. This ratio is primarily used as an index of competitiveness and is not an absolute measure of renewal rates. There are many reasons why a policy may renew despite this ratio exceeding 1, and equally there may be reasons why a policy may not renew despite this ratio being less than 1. These include:

- depending on how the index is constructed, the policyholder may not have all quotes available, or alternatively may have additional quotes not included when making up the index
- the policyholder may have been with the insurer for many years and is reluctant to switch
- the policyholder may be happy with the brand and good customer service received and so may be willing to pay extra
- the competitive index may not be perfect

The way that the index is constructed depends very much on the distribution channel used and the availability of other insurers' rates. In the United Kingdom, the intermediary market accounts for approximately 60% of the market, with direct response insurers accounting for most of the rest. Quotation systems are available that allow premiums to be calculated across the whole intermediary market for each renewal, and an appropriate competitive level can be chosen. It may not be necessary to choose the cheapest – often the third cheapest, or an average of the top five quotes is more appropriate. In this situation, 40% of the market has been ignored. However, the competitiveness level is only being used as an index, so if the relative competitiveness of the direct market remains constant the model will still be predictive into the future. Even if the direct writers target particular segments of the market (such as older people), the parameter estimates for older people will adjust accordingly.

If other insurers' rates are publicly available, then it is a matter of calculating all alternative rates for each renewal, and choosing an appropriate premium level. Again this appropriate level may not necessarily be the cheapest premium.

In the situation where market premiums are not available, a market index can be constructed from the results of a conversion analysis. Conversion rates are, very often, the best indicator of competitiveness since this is a direct measure of how the company's enquiry profile is responding to their price; an index can be constructed by estimating the expected conversion rate as if the renewal quote was actually a conversion quote.

The importance of market competitiveness depends very much on the distribution method and the easy availability of alternative quotations for the renewals. In the UK, a large proportion of business is written through intermediaries, and so on renewal the intermediary may re-broke if they feel that a better price is available with another insurer. In this case, competitive positioning is very important and is very predictive of renewal rates. For a direct operator, however, the story may be very different. Although the policyholder may be able to get better quotations elsewhere, these quotations are not as readily available; in other words the policyholder is less aware of a better price, and (although still important) competitive positioning is less predictive.

Because of the importance of this factor, and because of the dynamic nature of market premiums, it is necessary to update renewal analyses quite frequently (perhaps monthly or quarterly), and certainly more frequently than a risk premium analysis (risk premiums will not change so rapidly over time).

If a market premium index is not available, then the other parameters will adjust in the areas where the quoted premium is more or less competitive. However, as the market competitive levels change in different segments, so too will the parameter estimates. The model is likely to change quite significantly over time, leading to unstable parameter estimates.

There are many similarities between conversion and renewal analyses, both in terms of the theoretical underpinning and the results that come out of each. Renewal rates tend to be much higher than conversion rates (i.e. renewals are much less price elastic than new business quotations), but we have found in practice that many of the trends seen in a conversion analysis are similar to that in a renewal analysis (albeit at different levels).

5. Developing Optimal Pricing Structures

Section 3 and 4 discussed how to use GLMs to estimate the risk premium and customer demand. In this section we shall start looking at how we can use GLMs to develop optimal pricing structures.

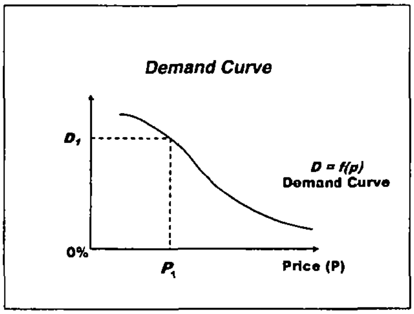
In simplistic terms there are four basic components of estimating the premium to be charged; firstly the risk premium, secondly the direct policy related expenses, thirdly a contribution to fixed overhead expenses and finally a profit margin.

Most premium rating structures do not, however, allow for each of the above elements explicitly. It is normally the case that the profit load and risk premium relativities for different customers are combined in some way. Consequently, in a competitive market, marketing discounts and/or rating action in response to competition become indistinguishable from rating action taken in response to changing claims experience.

However, if it were possible to anticipate an individual customer's response to the new business or renewal terms offered, then a probabilistic approach could be adopted in setting both the contribution to overheads and the profit margin in order to maximise expected profit or some other form of corporate objective. The model needs to be flexible enough to take into account the individual characteristics of each policyholder and be able to respond to the dynamics of the market place. Under certain market conditions or market segments it is quite plausible that the profit loads can be negative.

5.1. Demand curves

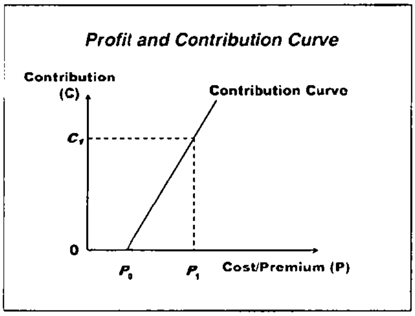
Just considering a one year time horizon initially, the demand for insurance (be it from new business or for renewing policies) for any individual customer is given by the following graph.



Demand curves are generally naturally downward sloping, and the shape and level of the curve depends of the individual characteristics of each customer, on the pricing action taken over the last year (for renewal business), and on the competitive positioning of the market segment involved. The level and shape of the curve will also be significantly different for new business quotations and renewal quotations. As discussed in section 4.1, these functions can be estimated for each customer using GLMs.

5.2. Contribution curves

Ignoring fixed expenses, the contribution curve for a policy looks as follows:



The contribution curve is a 45° line anchored at a point P_0. P_0 represents the expected claims cost arising from the policy (i.e. the risk premium) plus an allowance for direct expenses associated with the policy. P_0 is often referred to as the "breakeven" premium.

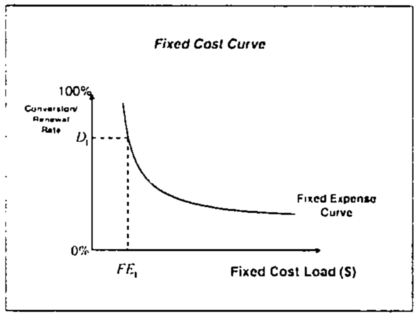
Again, P_0 is very heavily dependent on the risk characteristics associated with each policyholder, because the expected claims costs and associated expenses are heavily dependent on the policyholder. Given that a policy has been written at premium level P_1, the contribution to fixed costs and profit will simply be $P_1 - P_0$.

Again, we have discussed in section 3 how the risk premium and therefore P_0 can be estimated for each customer using GLMs.

5.3. Fixed costs

There is more than one way of building into the rating structure an allowance for fixed expenses. One possible way is to carry out some form of projection of business volume and convert this to a per policy cost. This method is slightly circular in that the volumes written depend on the prices, but equally the loading for expenses in the prices depends on volume. Clearly if more business than expected is written the per policy contribution to fixed costs reduces, the converse being the case if less than expected business volumes are written.

If the probability of conversion/renewal for any individual risk could be predicted, then this could be taken into account in the load for fixed costs prior to issuing the conversion/renewal quote. If a particular policy had a high probability of conversion/renewal compared to average, then the policy contributes a greater expected amount towards fixed costs than if the probability of conversion/renewal was expected to be lower than average. The fixed expense load for each policy can take this into account which can therefore be expected to be inversely related to the probability of conversion/renewal. The graph below gives a typical shape of fixed expense curve.

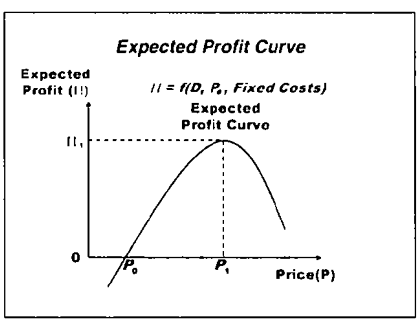


Building in fixed expenses in this manner has the intuitive appeal of attracting policyholders with higher persistency rates.

An alternative way to deal with fixed costs is to ignore them; instead of having an expected profit curve as described below, this changes to an expected contribution to fixed costs and profits curve. It then becomes necessary to check that fixed costs are covered across all policies.

5.4. Expected profit curve

Given the basic components of the demand function and the corresponding costs, it is possible to combine these in order to generate an expected profits curve. The shape of the resulting curve is given in the graph below.



The price that equates to the maximum profit (or minimum loss) can easily be derived, in this case P_1. Other "profit" functions can be estimated. For example, the corporate strategy may not be maximization of profit but rather to achieve satisfactory profit subject to minimum levels of business. These alternative scenarios can easily be calculated by optimising other suitable functions which may place greater value to certain components of the process. Time horizons greater than a year can also be allowed for, and this is discussed further later.

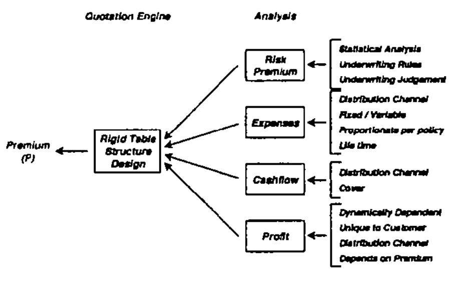
6. Implementing an Optimal Pricing Structure

6.1. The need for flexible pricing engines

We have detailed above the derivation of two of the most important elements of the price, namely the expected claims cost and the demand for the product. We have also briefly looked at fixed expenses. There are of course other considerations, such as a more in-depth treatment of variable expenses, investment income, cash flows and claims development, return on capital etc, but we do not propose to explore these within this paper in any more detail. Instead we will now look at how we can build these different rating components into a tariff system. The statistical analyses undertaken are normally tailored to provide information to help populate the existing tariff structure.

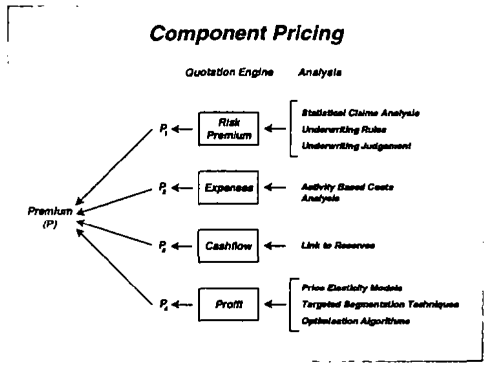
The chart below illustrates how typically the various statistical analyses are currently incorporated into rating tariffs.

Typical Pricing Structures



As you can see, insurers do carry out very sophisticated analyses on each individual element of the premium, but much of this sophistication is compromised in order to fit into a rigid table design from which the premiums are calculated. Much of the analytical benefit is therefore lost at this stage because the lookup tables that are used to derive the actual premium rates are a compromise of the results of the different analyses undertaken. In particular, profit margins cannot be explicitly monitored, leading to a merging of the loads and discounts given for a variety of reasons, for example as a reaction to competitive pressures, for marketing strategy, and for adverse claims experience. This makes it very difficult to manage the profit versus demand equation.

In this paper we have chosen to tackle the problem of implementing pricing structures from a different angle, i.e. how do we best exploit the information obtained from the statistical analyses? We therefore see the need for a much more flexible pricing engine that calculates the premium in several stages. This is shown schematically below.

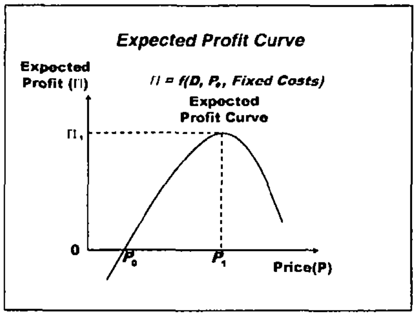


The quotation engine separately calculates each individual element of the premium, and the total premium only calculated at the end. We refer to this approach as component pricing. This approach facilitates much more focused management information. Reports can be designed to specifically manage each premium component. For example, the risk premium component can be directly monitored against the emerging claims cost. Employing such rigorous control on the "cost" components of the premium provides a foundation to exploit the demand for the product given the competitive pressures that exists at the time. At certain parts of the underwriting cycle we may strategically decide to write some segments of the account at a loss. The most important principle here, though, is that informed strategic decisions must be regularly made.

Another way to look at this problem is as a profit optimization process. The optimization will be subject to corporate goal constraints. We shall look at the profit optimization process in more detail in the next section.

6.2. Optimizing Profit

The profit optimization process is the most interesting part of a component pricing system. Whilst the other parts of component pricing can be driven from a series of look-up tables, profit loadings are best implemented by an optimization algorithm. The algorithm will be based on the expected profit curve introduced in section 5, which is reproduced below.



Let us first consider the simple case where the insurer's objective is to maximize the expected profit over a one-year time horizon. The profit optimization process would then look at each customer individually. For any given customer the expected profit curve is estimated by calculating the profit and expected demand (be this conversion or renewal probability) for a series of different premium values. The expected profit at each premium value is then calculated by multiplying the profit with the expected demand. The premium value at which the expected profit is maximized can then be read off. The process can then be repeated for each customer.

The above algorithm simplifies the problem somewhat. It is unlikely that the insurer's objective will be to maximize profits over a one-year time horizon. There may also be some overall corporate goals to consider. For instance, a target of maximizing business volumes whilst achieving satisfactory profitability may be a preferred objective. The algorithm can be easily adjusted to allow for the desired objectives. One way of making the adjustment is to define a utility function that explicitly maps a trade-off between volumes and profitability. The utility function effectively adjusts the expected profit curve so that a new maximum can be derived.

Another factor which the insurer might want to build into the equation is profits made by cross-selling other products. This can be done by either reducing the breakeven premium P_0 by a fixed amount for each policy, or by building a cross-sell model to reflect different cross-sell propensities for different customers.

6.3. Controlling the structure

With such a complex rating structure, it may be viewed as being a difficult task to keep track of and to control the parameters input. In many ways, however, the structure is easier to control than the more traditional lookup table structure because each element of the price can be separately monitored.

The key to successful control involves storing each element of the premium charged in the management information system. This facilitates two things.

Firstly, the actual experience of each component can be monitored against the expected experience. For example, the expected risk premium can be compared against the actual claims experience; the expected conversion and renewal rates can be compared against actual conversion and renewal rates (the latter can be monitored much quicker because of the slow emergence of claims experience).

Secondly, a profit forecast can easily be produced for each month's written business. This will give an early indication to management of reduced levels of profitability; furthermore, the cause of the reduction in profitability can easily be identified. For instance, previously if there was an increase in written policies in a month, the marketing manager would be quite happy because he would be seen to have carried out his job; however, the effect on the bottom-line profitability would be unknown for some time.

By monitoring each element of the price separately, different people or departments can be made accountable for achieving certain goals. The actuaries/statisticians would be held responsible for achieving a 100% loss ratio based on the risk premium component of the total premium; the marketing department would be held responsible for maximizing profits (if they were in control of setting profit levels). Previously if premium rates were too low it would be unclear whether those rates were too low because the actuary underestimated claims experience or whether competitive pressures forced rates down.

7. Adapting the Process in Other Environments

It may not always be possible to implement fully the ideas discussed in section 6 in terms of profit optimization because of regulatory or other restrictions. However, the techniques developed in this paper can be used in a variety of other ways.

One possibility is to reverse engineer the whole process, and use the expected claims costs and demand behaviour to identify the most valuable customers under a set of restricted premium rates. These could be ranked in order of value and scored in terms of customer value. It may be then possible to use marketing means to attract those more desirable customers. For example, it may be possible to market to particular segments in an attempt to improve conversion/renewal rates for those customers, or to be used as the customer value measure in customer relationship management (CRM) programs.

Even if non-pricing means are unavailable, the techniques can be used to effectively cost the consequences of particular pricing action taken, both in terms of expected claims cost and the volumes of business written. The models developed will be based on each individual customer's characteristics and will move dynamically with changing market activity, providing early feedback on pricing decisions taken.

8. Conclusions

We have discussed in the paper the use of generalised linear models (GLM's) and their application in personal lines insurance. We have shown that GLMs have a model and variance structure that closely reflects many of the processes that we often find in insurance. In our experience this leads (if GLM's are used correctly) to reliable and robust parameter estimation. We have also found that these techniques work well on large datasets with many millions of cells and large numbers of risk factors.

Furthermore, we have shown in this paper, how GLMs can be used to estimate both the risk premium and demand functions for individual customers. We then showed how these could be combined and used to set optimal profit loads tailored to the behaviour of each individual customer. Alternatively, the profit load can be estimated for any given premium. The concepts can be extended to cover longer time horizons to create life-time pricing models. Alternatively corporate goals can also be built into the optimization algorithms. In addition, we have demonstrated how, by re-designing the rating tariff, the results of GLMs can be more effectively utilised. The re-designed rating tariff leads to a more rigorous control and management information environment.

Clearly there is more work that can be done. The modelling of the demand function is a new area and the predictiveness of models suggested needs to be more thoroughly road-tested. However, our experience to date is encouraging. Also there are several approaches that could be adopted to defining the optimization algorithm and again these need to be more fully explored. Furthermore, our approach can also be defined in terms of a stochastic environment by incorporating the standard errors of estimates into the algorithms. We hope, however, that this paper will sow the seeds of thought for more ideas and research in these areas.

References

MJ Brockman and TS Wright, *Statistical Motor Rating: Making Effective Use of Your Data*, (JIA 119III)

KD Holler, D Sommer and G Trahair, *Something Old, Something New in Classification Ratemaking With a Novel Use of GLMs for Credit Insurance*, (Casualty Actuarial Forum, Winter 1999)

P McCullagh and JA Nelder, *Generalized Linear Models*, (Chapman and Hall, 1989)

S Millenhall, *A Systematic Relationship Between Minimum Bias and Generalized Linear Models*, 1999 Proceedings of the Casualty Actuarial Society