Estimating Claim Settlement Values Using GLM

Roosevelt C. Mosley Jr., FCAS, MAAA

Estimating Claim Settlement Values Using GLM

by

Roosevelt C. Mosley, Jr., FCAS, MAAA

Abstract: The goal of this paper is to demonstrate how generalized linear modeling (GLM) can be applied in non-traditional ways in property and casualty insurance. Specifically, we will use a property and casualty closed claims database to aid in estimating ultimate claim settlement amounts, evaluating claim trends, and assisting in improving claims handling procedures. This specific example will be used to demonstrate the potential of the application of GLM to different areas of an insurance company.

A GLM will be developed with data from the Insurance Research Council (IRC) closed claims study. The model will be populated with characteristics of closed automobile claims along with final settlement amounts. Using this data, the paper will examine how GLM can be used to identify:

1) *Trends in claims severities over time,*
2) *Differences in severities that exist between current ratemaking characteristics (e.g. state, territory), characteristics of the claims and the injured parties, and other factors (e.g. time from reporting to settlement, attorney involvement, use of arbitration), and*
3) *Interactions between these factors.*

Diagnostics will also be discussed which can be used to test the validity and robustness of the GLM models that are developed, and several applications of the results of this type of analysis will be presented.

Over the last several years, Generalized Linear Modeling (GLM) has seen increased usage among actuaries primarily in traditional ratemaking applications. The benefits of GLM are that it allows for a flexible model structure to be fit to insurance ratemaking data, and it also allows for a multivariate model to be generated that simultaneously incorporates a set of independent variables to determine their impact on a dependent variable. This is an improvement over traditional one-way types of analysis (both loss ratio and pure premium) because it adjusts for the impact of distributional biases that are present in all insurance data sets. The result is a set of indications for whatever you are modeling (class plan relativities, tiering relativities, etc.) that reflect the true impact of each variable being analyzed.

GLM has had immediate appeal in the traditional areas of actuarial practice. Most significantly, insurers have used GLM to refine class plan relativities, establish tiering and underwriting plans, and incorporate commercially available insurance scores into rating and underwriting plans, just to name a few applications. These applications have been addressed quickly as insurers move to this type of analysis for a number of reasons: these areas fall within the actuary's normal area of responsibility, the data for these types of analyses is usually readily available, and this type of analysis can provide the most immediate benefit for an insurer.

However, understanding the general statistical nature of GLM, one realizes that a GLM analysis can be applied to other areas within insurance companies, areas that have not necessarily been within the actuaries' traditional realm of responsibility. Specifically, we have used GLM's for a number of non-traditional applications, including developing custom insurance scores, generating vehicle classification systems, evaluating claims and agency personnel and external service providers, and estimating claim settlement value amounts. These types of analyses can provide benefit to many areas of the company, and can display the actuary's skills to a wider audience.

We will demonstrate the concept of applying GLM to non-traditional areas in this paper using the 1994 Insurance Research Council (IRC) Closed Claim Study database. In this example, we use the characteristics of the closed claims as provided in the IRC database to estimate the ultimate settlement value of a claim; however, we will describe this process in general terms such that it might be applied to a variety of different areas. The goal of this paper is not to provide you with a complete analysis of the IRC database, but to use this database as an example of how this general statistical procedure can be applied to other areas.

The Basics of GLM

GLM is a statistical process by which a model is developed in which a specific dependent, or response variable, is predicted by a number of independent, or explanatory variables. For example, as applied to the insurance ratemaking process, the process of setting class premiums for groups of risks can be thought of graphically as shown in Figure 1.

The goal of the classification ratemaking process is to set premiums by class of risk that reflect the risk of each group. This requires estimating the relative loss potential of each insured characteristic in the classification plan to determine how the factor contributes to the overall risk premium. An insured is then charged a premium based on his or her characteristics, and how these characteristics relate to the risk of loss. The traditional approach to analyzing the variables in the class plan was to analyze each of the variables separately, using a one-way loss ratio or pure premium approach. The inherent assumption in the one-way analysis is that, for **each level** of the factor being analyzed, the distribution of all the other factors in the class plan is constant. This means, for example, if one were analyzing auto symbol, model year, and age using a series of one-way analyses, one would be assuming that the same proportion of teenagers drive 10-year old Ford Escorts and brand new Cadillac Escalades. While this is simply one example, there are a number of other violations of this assumption that can be thought of in an auto or homeowners insurance class plan.

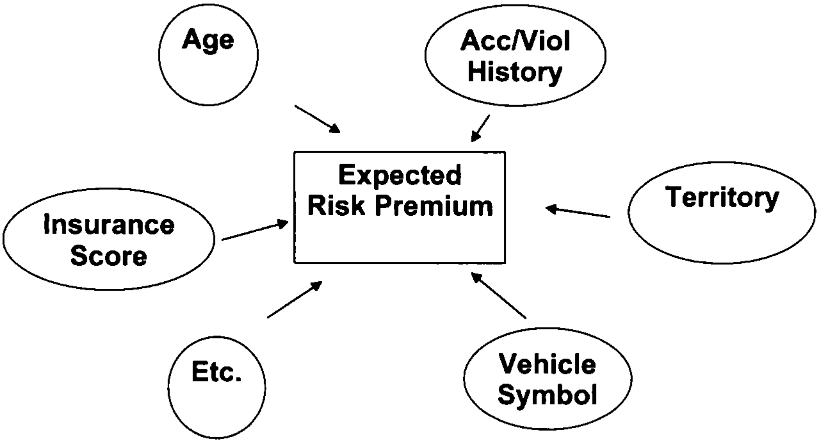
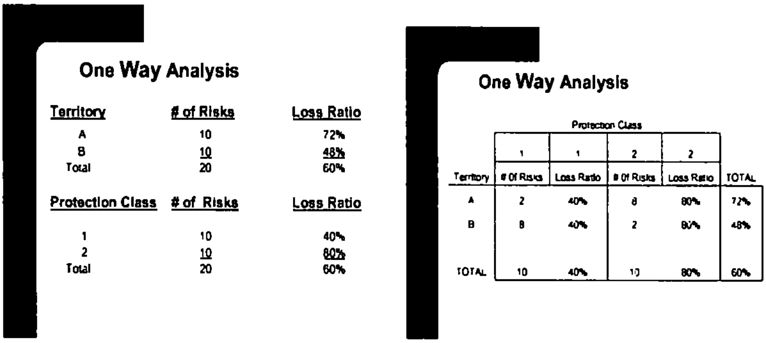


Figure 1: Description of Classification Ratemaking Process

Figure 2 gives an example of how this type of analysis can lead to erroneous results. The first table in Figure 2 gives the results of two separate one-way homeowner's insurance analyses, one for territory and one for protection class. In this particular example, when analyzing the two territories, one assumes that territory A has the same ratio of protection class 1 risks as territory B. The result of the loss ratio analysis shows that the rates for territory A should be increased relative to the territory B rates. Similarly for protection class, the analysis shows that the change in protection class 2 rates should be higher relative to the change in protection class 1 rates. However, when these results are viewed in a two-way table, the true picture becomes clear. The territory loss ratios are identical for both protection classes. The true problem is in the protection class relativities. If one had simply looked at the one-way analysis, the erroneous decision would have been to increase both the territory A rates and the protection class 2 rates, resulting in an over-correction. The reason the one-way loss ratios appear this way is because of the difference in protection class distribution over the two territories. Again, while this is a simple example, one can easily think of the number of different potential scenarios where this can occur in a rating plan.



One Way Analysis

Territory	# of Risks	Loss Ratio
A	10	72%
B	10	48%
Total	20	60%

Protection Class	# of Risks	Loss Ratio
1	10	40%
2	10	80%
Total	20	60%

One Way Analysis

	Protection Class				
	1	1	2	2	
Territory	# Of Risks	Loss Ratio	# Of Risks	Loss Ratio	TOTAL
A	2	40%	8	80%	72%
B	8	40%	2	80%	48%
TOTAL	10	40%	10	80%	60%

Figure 2: Example of one-way loss ratio analysis

GLM corrects for these distributional biases, and also provides a flexible model structure such that it better fits insurance data. One can best think of GLM in terms of one of its simplest forms, classical linear regression. The formula for a simple one-factor linear regression is:

$$y = a + bx + \text{error}$$

This describes the fitting of a line through a series of points, attempting to model a response vanable (y) using an explanatory variable (x). The b represents the relationship of the independent variable x to y. There is also an error term which accounts for the fact that the model will not predict the observations perfectly. Under linear regression, the error is assumed to be normally distributed with a mean of zero and a constant variance. A graphical description of this simple regression model can be seen in Figure 3. In this example, the bodily injury severity is being modeled as a function of the time period.

To extend this to GLM, the more general formula for multiple regression is:

$$y = X\beta + \text{error}$$

In this notation, the Xβ represents a matrix, where X represents a series of independent variables and β represents the relationship of these independent variables to the dependent variable. The error term is more general in that it is not restricted to the assumption of normally distributed error terms (as in simple and multiple linear regression). More general error structures, such as Gamma, Poisson, and Negative Binomial can be used which are more representative of insurance data.

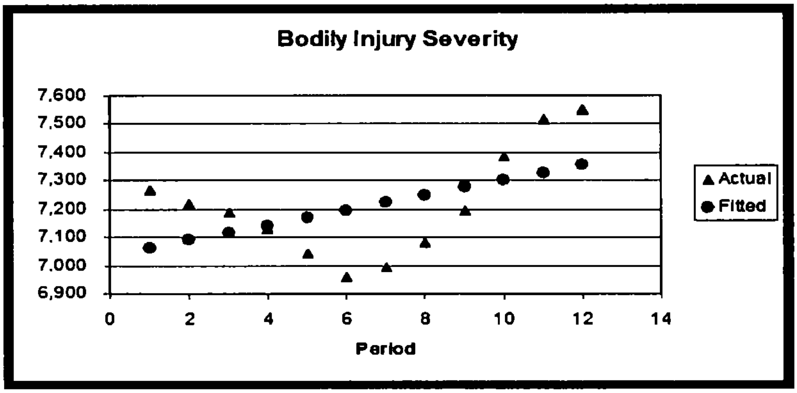


Figure 3: Simple regression example

Non-Traditional Applications

Given the general structure of GLM described above, one can begin to expand the use of GLM beyond the traditional actuarial realm. The general structure of GLM can be described as shown below in Figure 4:

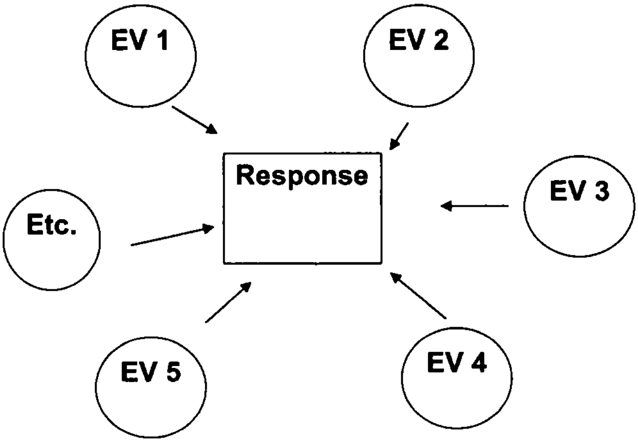


Figure 4: General structure of a GLM model

Because GLM is a general statistical process, it is not limited to estimating class plan relativities. The general structure of the model can be used to describe many different responses by a series of explanatory variables. Depending on what problem GLM is applied to, the explanatory variables and the model error structure will change, but the process of generating and applying the model will remain the same.

CLAIM SETTLEMENT VALUE ESTIMATION

One potential area for the application of GLM in an insurance company is in the estimation of ultimate claim settlement values. The ultimate value of a settled claim can be described as the response variable, and the characteristics of the claim represent the explanatory variables. When a claim is reported to an insurer, the insurer is presented with the facts of the claim. Based on the facts of the claim, an estimate is made of what the final value of that claim will be. This value may be determined based on a claim value estimation software package, guidelines established by the company, the claim persons' expert opinion, or a combination of the three. As the case matures, as payments are made on the case, and as more information regarding the case becomes available, future refinements of that estimate can be made. It is these estimates that are made before the final disposition-of a claim that are reflected in an insurers financial results from year to year.

What this GLM example will do is develop a model to estimate the final amount of the claim settlement, which can then be used as part of the overall information that the claims handler uses to determine the expected final value of a claim. The goal of this analysis is not to replace the claims person, no more than the goal of the analysis of traditional class plan relativities by using GLM is to replace the actuary. The goal of this process would be to provide the claims person with additional information on which to base decisions.

This type of model could be used to help estimate the ultimate settlement value of claims based on the information known. It could also be used to assist claims departments in determining the effectiveness of certain claims handling techniques. It can also provide information on areas of focus such that claim handlers might more efficiently handle claims.

Data

To perform this type of analysis, an insurer would need a database of final closed claim settlement amounts, as well as the characteristics of the claims that have been closed. The characteristics available will likely vary between insurers, but examples of the information that could be used are:

- Insured rating and underwriting characteristics
- Type of injuries involved

- Age of injured parties
- Hospitalization involved
- Location of accident
- Types of treatments
- Treatment providers
- Claim reporting lag
- Claim settlement lag

The list of characteristics to be analyzed could continue, and the goal should be to include all the information that is available that might be useful to the analysis. This could be one potential difficulty for an insurer employing this type of analysis technique. For some insurers, this type of closed claim database might simply not exist, or the information might exist in paper form in the claim files.

For this paper, we have analyzed the IRC 1994 Bodily Injury closed claim database. This database was compiled by the IRC as a sample of claims closed during a specific period during 1992 from a number of insurance companies. The database consists of the ultimate settlement value of these claims, a breakdown of these settlement amounts by type of payment (medical, wage loss, etc.), and a number of characteristics of the claim. The variables analyzed from this database reflect many of the items listed above. A complete list of the factors could be obtained from the IRC.

While not a specific issue with the IRC database, an insurer or claims organization that undertakes this type of analysis will need to be aware of claims that are closed without payment. While these claims do not generate any loss dollars, there are at least two other issues that these claims raise. First, they will generate loss adjustment expense dollars because a claim file will be opened on these claims and a claims person will be assigned to handle the claim. Also, because these claims can generate a series of points with no settlement value or a very small settlement value, this can create some difficulty with the determination of a model error structure. One approach to handling this issue would be to use an analysis similar to a claim frequency analysis, but instead analyze the likelihood of a claim closing without payment. This analysis could then be combined with a settlement value analysis to determine the ultimate expected settlement value.

Additionally, a priori there are some factors that we could analyze that would be significant in our analysis of expected claim value but were not present explicitly in the dataset. For example, in the IRC dataset, we knew the date of the accident and the date of the insurance company's initial contact with the claimant, which allowed us to calculate the contact lag. The a priori expectation was that the longer the period between the accident and the initial contact, the larger the ultimate value of the claim. Another example is a difference between the claimant state and the accident state. We assumed a priori that if a claimant has an accident in a state different than their place of residence, it could potentially increase the ultimate settlement value. In an insurer database, there will be variables like these which the modeler will want to derive from information present in the database.

In addition to data from a closed claim database and data from the rating database, there may be information in other parts of the company or external to the company which might be useful to the GLM process. Potential internal information might include marketing information or underwriting information. External available data might include population and vehicle density, medical inflation rates, wage inflation rates, vehicle repair rates, etc. The ultimate goal of the data process is to be confident that you have compiled as complete and correct a dataset as possible with which to generate the model.

Model Considerations

The overall goal of the modeling process is to generate a model that is complex enough to provide a satisfactory degree of predictive accuracy, yet simple enough that it can be explained and understood by users. This delicate balance can be difficult to maintain, but there are some things that can be done to attempt to make this process easier.

In generating a GLM based on the IRC database, we analyzed 150 potential explanatory factors. Needless to say, when analyzing a dataset of this size, we are fully anticipating that the number of explanatory variables included in the final model will be significantly less that 150. Therefore, we need a process by which to determine which variables provide enough predictive value to the modeling process to remain in the final model. There are a number of different approaches that can be undertaken. Three of these approaches are outlined below:

1. Single Inclusion Process: Beginning with the first potential explanatory variable, we add each variable one by one to the model in order of presence in the dataset, keeping the variables that add predictive power to the model and not using the variables that do not provide predictive power. To determine whether or not predictive power was added to the model, we utilize the chi-square test which is based on the deviance of the model, or the difference between the expected claim settlement value as generated by the model and the actual claim settlement value present in the dataset. The disadvantage of this approach is that the order of addition of explanatory variables to the dataset is generally random, and this could result in a less than optimal set of variables being included in the final model.

2. Stepwise Type I Regression: This process begins with a model including no factors, and then generates a one-factor model for all 150 potential explanatory variables. The factor that produces the lowest deviance and proves to be significant by evaluation of the chi-square test results is added to the model (F1). Next, all 149 potential two factors models are generated which include F1 plus all the other explanatory factors, one at a time. The next factor added to the model is the one that produces the lowest deviance and is also significant based on the chi-square test. The process continues until no other additional factors added to the model produce significant results.

While this process is more time consuming than the first process, it helps assure that the factors that provide the most predictive power will, with high likelihood, make it into the final model. Once you have generated a final model, this process will also require a review of the factors in the final model again for significance. There is the potential that a factor that entered the model early in the modeling process might be proved to be insignificant later by the additional variables. To the extent that the model can be simplified by the removal of these redundant factors, this should be done.

3. Stepwise Type III Regression: This is a variation of the Type I regression that starts with a model which includes all 150 factors, then generates a series of models removing the factors one at a time to determine which factor is the least significant. The factor that is not significant as measured by the chi-square test and has the smallest impact on the deviance will be removed from the model. The process continues until there are no more insignificant factors in the final model.

   This approach is the most time consuming of the three, since it requires models with more explanatory factors to be generated.

If we are working with a dataset with a manageable number of explanatory factors (less than 50), we will generally begin with a model that includes all parameters, and investigate each of the independent variables to determine which factors are significant. For analyses that have a larger number of factors, we usually take an automated approach to determining which factors to further investigate. For the purpose of analyzing the IRC database with a larger number of factors, we employed the Type I regression method.

As a general practice for modeling projects, one should consider developing the model based on a portion of the dataset and testing the model that has been developed on the remaining portion of the dataset. There is the potential in generating models that you can "over-fit" the dataset. The process of splitting your dataset, sometimes referred to as "training and testing," can help avoid interpreting a trend when one really is not there. The optimal split will depend on the size of your dataset, but as a general rule of thumb, using 70% of the data to develop the model and 30% to test it works well. In this particular example, we did not divide the dataset due to the size. There were just under 34,000 records in the dataset, and removal of 30% of these records would have significantly impacted our ability to generate the GLM.

In order to generate the model, one must determine an initial model error structure. For claim settlement values, good a priori distribution assumptions are a gamma or a negative binomial distribution. For purposes of this paper, we have chosen the gamma distribution.

Because we are analyzing liability data, the potential always exists for large claims. Large claims present some difficulty in performing a relativity analysis (such as we are performing here or would be performed in a class plan analysis) because one or two large claims can have a significant impact on an indicated relativity or the indicated impact of a claim characteristic on the final settlement value of a claim. However, large claims cannot be ignored because they are covered as part of the insurance contract. Traditionally, insurers have simply generated relativities based on a limited claim severity analysis, and loaded back a fixed amount to each claim for purposes of covering the large claim amount. However, this ignores the fact that the likelihood of large claims is not constant over all claims that are presented. All liability claims have some potential to become large claims, however, there are certain claims that have a higher than average likelihood of becoming large claims. In this analysis of the IRC data, we have analyzed the likelihood of large claims as a basis for generating a large claim load which varies based on the characteristics of the claim.

To generate the large claim analysis, for claims that pierced a $25,000 threshold, we generated a second model, using a logistic error structure, that attempted to determine the likelihood of a claim to pierce the $25,000 threshold based on its particular characteristics. Each total estimated claim amount would then be a combination of the limited claim settlement value estimate and the adjusted large claim load. A description of the models generated is shown in Figure 5.

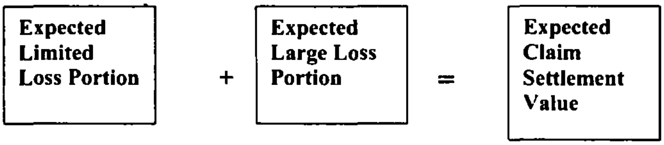

Expected Limited Loss Portion
+
Expected Large Loss Portion
=
Expected Claim Settlement Value

Figure 5: Claim settlement value model structure

LIMITED CLAIM SETTLEMENT VALUE MODEL

Of the 150 variables analyzed, we selected 35 which were determined to be significant for the limited claim settlement value model. Many of the variables and the effects made intuitive sense, however there were some that may have appeared at first glance to be counterintuitive. We provide a few of the results of the model below, as well as some of the simplifications to the factors in the model.

Presence of an Attorney

One of the factors analyzed in the Bodily Injury dataset was whether or not the claimant was represented by an attorney. Insurers have long alleged that the use of an attorney for an auto insurance claim causes the settlement value of that claim to increase. Attorneys have alleged that the settlement value of claims involving attorneys is higher because they are generally involved in the more serious claims. The results for the involvement of an attorney in the claim settlement process are shown in Figures 6 and 7. Figure 6 shows that, all else being equal; the average final claim settlement value for the base claim characteristics for cases involving attorneys (Code 1) was about $9,500, more than double the cost of claims not involving attorneys (Code 2). Figure 7 simply shows the relative cost of these types of claims due to the impact of attorneys, even after removing the impact of the type of injury. This result is helpful in attempting to determine the final value of a claim and it would also be valuable during the claim handling process in determining which claims should be monitored more closely. The bars at the bottom of graph represent the distribution of claims in each category, and relate to the y-axis on the right side of the graph.

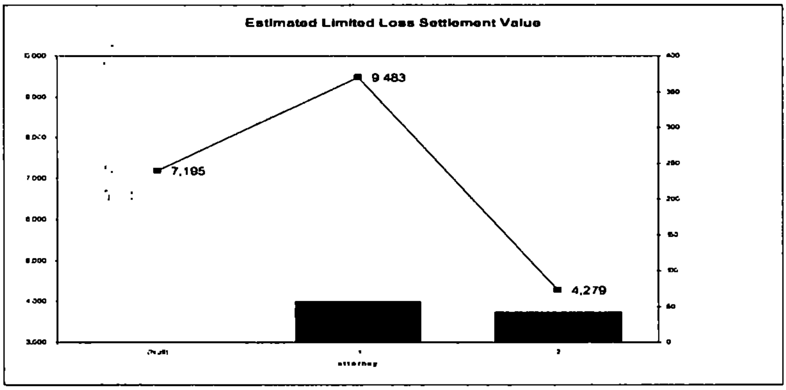


Figure 6: Attorney Involvement

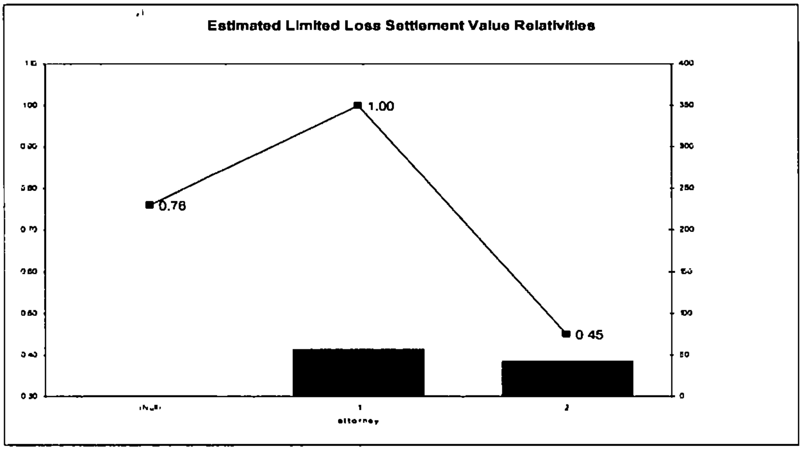


Figure 7: Attorney involvement (Relative to category 1 - yes)

Depending on the type of analysis that you are undertaking, you may have to deal with the issue of unknown explanatory variables. (In the example above the Null category represents an unknown category). Unknown data can come from a couple of different sources. One reason might be that the data collected was just not complete, and therefore there are a number of risks for which you may not have all the desired information. There may also be a systematic reason for unknown variables. For example, in many class plans in the United States, marital status and gender are not used to rate adult risks, so this data is not collected on non-youthful risks. Regardless of the reason for the unknown data, the modeler will need to decide how to handle the unknown values. The best solution would be to try to obtain the missing data fields, however this is not usually feasible. Another option would be to model the unknown variable as a distinct level of a factor, which would make sense if a variable being unknown is a valid occurrence, such as the class plan example given earlier. A third approach would be to group the unknown level with an "average" level, or with the most likely occurrence of the variable. For the purposes of this analysis, since it is likely that there would be information about future claims that is unknown, we chose to model the unknown level as a distinct level.

The graphs shown above represent two different ways of viewing the results of this claim analysis. We will view the results using the relative claim cost method (Figure 7), realizing that we will use the actual claim settlement values when generating the final claim settlement amounts.

Most Significant Injury

Another factor in the analysis dataset is the most significant injury to the claimant. A graph showing the results from this factor is shown in Figure 8. As can be seen from the graph, lower claim amounts were associated with less serious claims, such as minor lacerations (code 3) and various sprains and strains (codes 6-8). The larger claim amounts were associated with more serious claims, such as serious lacerations (code 4), scarring and permanent disfigurement (code 5), Temporomandibular Joint (TMJ) dysfunction (code 16), and loss of senses (code 17).

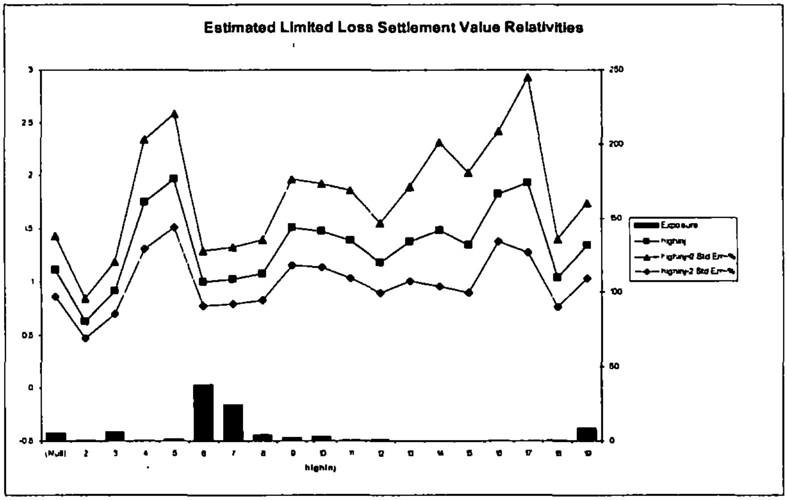


Figure 8: Most significant injury

To assist the modeler in determining the significance of independent variables, the standard error of each parameter estimate is generated. The parameter standard error estimate gives an indication of the reliability of the parameter estimate. For example, the relativity estimate for neck sprains and strains (code 6) is 1.00. Plus and minus two standard errors around this parameter estimate yields 0.77 to 1.29. However, the estimate for the loss of a body part (code 14) is 1.49, two standard errors around this parameter estimate yields a range of 1.13 to 1.93. This is a wider spread, and relecfts the increased uncertainty regarding the serious laceration parameter as compared with the neck sprain/strain parameter. Many times (but not always), increased standard errors for a parameter estimate are caused by a lower number of observations for a particular category. The standard errors will give the modeler information regarding the amount of reliability to place in the estimate.

Year of Accident

The year of the accident occurrence was present in the IRC database, which gives some indication to the length of time the claim had been in the company claim process. Claims were present in this dataset that occurred as far back as 1950. The expectation is that if a claim has been open for a long period of time, it represents a more complex claim, or a claim that may have been contested more fiercely. It is expected that these claims would settle for larger amounts. As can be seen in Figure 9, this trend appears to hold for 1992 back through 1987, but at 1986 the trend appears to break down. This might be a reflection of the trend breaking down, but is more likely a reflection of the data sparseness for years prior to 1988. Due to the lack of data at these points, we decided to combine years 1988 and prior for purposes of this analysis, as shown in Figure 10. Another option would be to potentially extrapolate the trend from 1989 and subsequent onto the 1988 and prior data.

For other variables, levels of the variable that exhibit similar claim settlement values can potentially be combined.

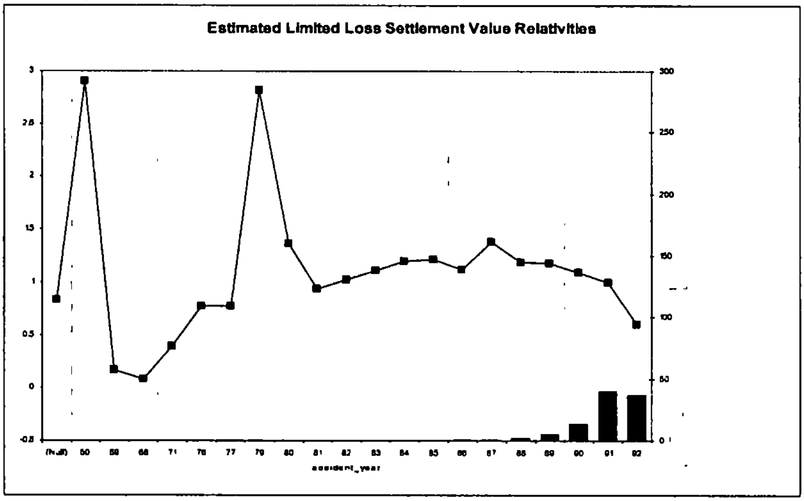


Figure 9: Year of accident

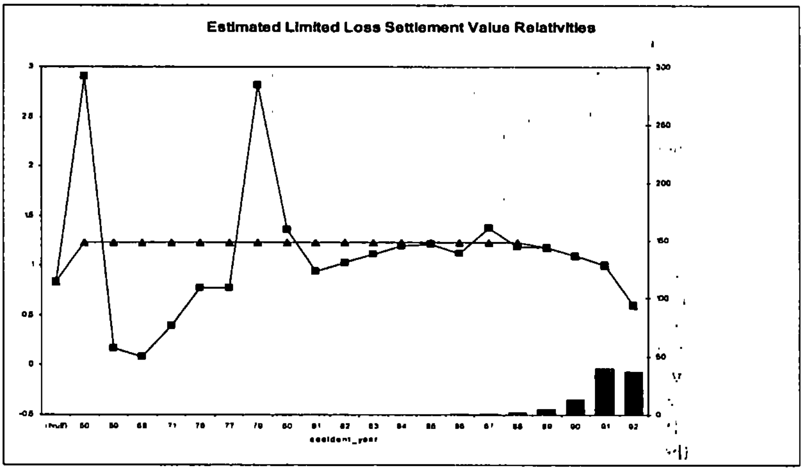


Figure 10: Year of accident grouped

Claimant Age

The age of the claimant was also analyzed as part of the final model. For variables that have a natural scale where successive levels are related, such as age, one can consider fitting a continuous curve representing this factor's impact on the dataset. Figures 11 and 12 represent the initial and final smoothed results of the claimant age factor. In this case, we fit a "mixed" simplification to the claimant age. We fit one curve to ages 0-9, allowed the model to fit separate and distinct factors to ages 10 and 11, and then fit a second curve to ages 12 and over. This demonstrates the flexibility of fitting GLM's. As can be seen, the cost of the ultimate claim tends to increase as the claimant age increases, but then around age 60 begins to decrease again. This may have something to do with the wage earning potential of an injured person. Wages generally tend to increase as a person gets older, and then at older ages the earnings decrease due to retirement.

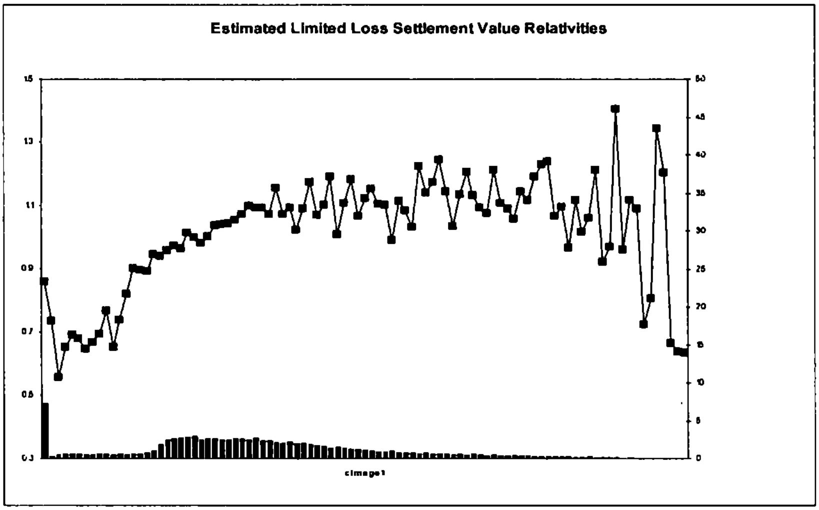


Figure 11: Claimant age

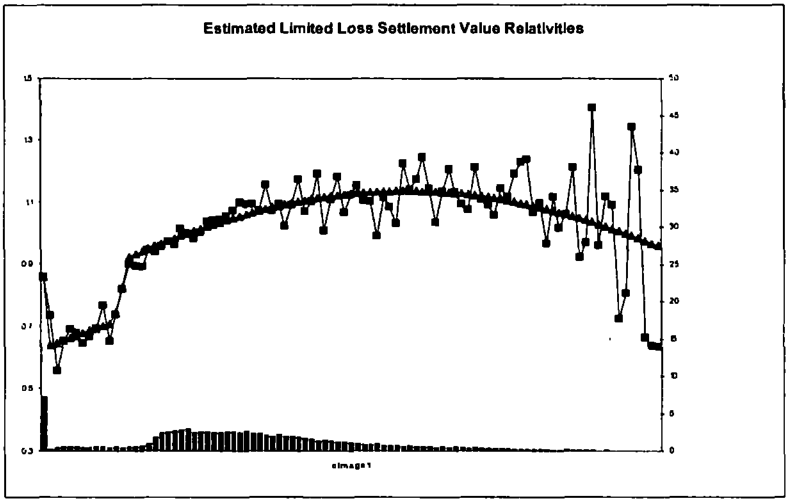


Figure 12: Claimant Age (smoothed)

Injury Type by Attorney Involvement

In addition to the impacts of individual variables on the ultimate settlement value, combinations of factors can have interaction impacts on the final claim settlement amount which can differ from the combined effect of the individual factors. For example, Figures 7 and 8 discussed attorney involvement and injury type, respectively. For a claim that did not involve an attorney, the resulting settlement value was 45% of the value of a claim that did involve an attorney. When considered in combination with injury type, this assumes that all injury types are 45% smaller when an attorney is involved, unless this assumption is specifically relaxed. Figure 13 shows the result of specifically considering this interaction. As can be seen, the presence of an attorney does not have a constant effect when considering different types of injuries. The difference ranges from a 29% increase when dealing with fractures (code 9) to 124% when dealing with other sprains and strains (code 8).

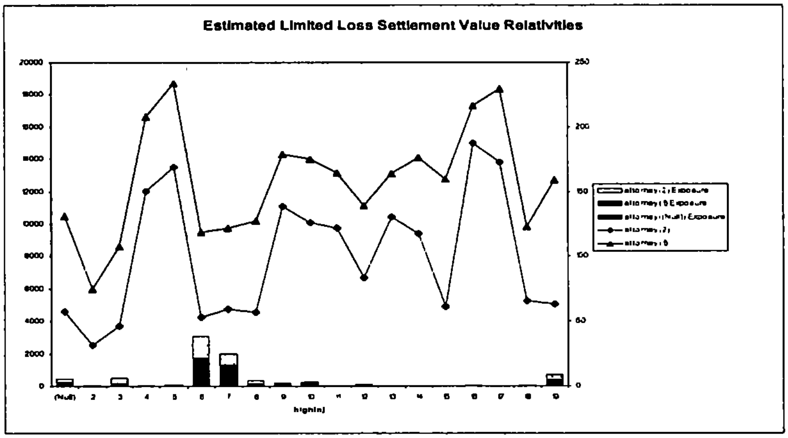


Figure 13: Attorney involvement by injury type

Final Limited Claim Settlement Value Model

An analysis of each of the significant variables was conducted to determine if there were any of the variables that could be simplified, either by grouping of levels of the factor or by fitting of a continuous curve. Also, a series of interactions were tested to determine if they were significant in the final model. After the final limited claim settlement value model is developed, an expected limited claim value is calculated for each record in the dataset. An example of this calculation is shown in Attachment 1. This limited claim settlement value will be combined with the expected excess claim value determined in the next section to reach an overall final expected claim value.

Excess Claim Settlement Value Model

The purpose of developing an excess claim settlement value model is to account for the presence of large claims in the database in a way that recognizes the fact that certain characteristics are more likely to generate large claims than others. We began by generating a model to determine the likelihood of a large claim occurring. This model was developed based on a logistic error structure, with the response variable being whether or not the claim pierced the threshold ($25,000). A logistic model is generally used for the analysis of a yes/no type response variable. We then looked at particular claim characteristics to determine if the presence of certain levels of some characteristics had higher likelihood of large losses than others. The large loss load for each claim was then determined by taking the average excess loss for the base risk and adjusting this excess loss based on the likelihood of a large loss occurring. Below, we show examples of the relative likelihood of large losses for several claim characteristics.

Presence of an Attorney

Again, similar to the limited claim value model, the presence of an attorney significantly increases the likelihood of a large claim. When an attorney is present, the likelihood of a large loss nearly doubles.

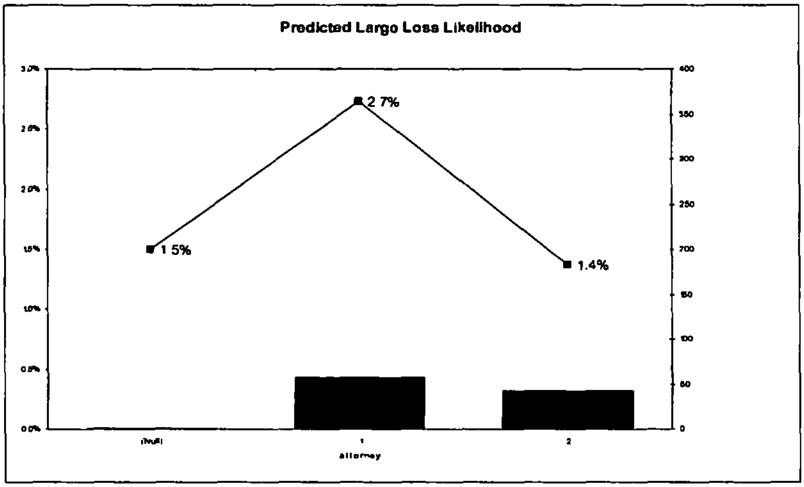


Figure 14: Likelihood of large loss when an attorney is present

Neck Injury

The situation can occur where the results of the large claim frequency analysis might show results that are opposite the results of the limited claim severity. The presence of a neck injury causes a larger limited claim severity. However, the presence of a neck injury is about 15% less likely to produce a large loss (Figure 15).

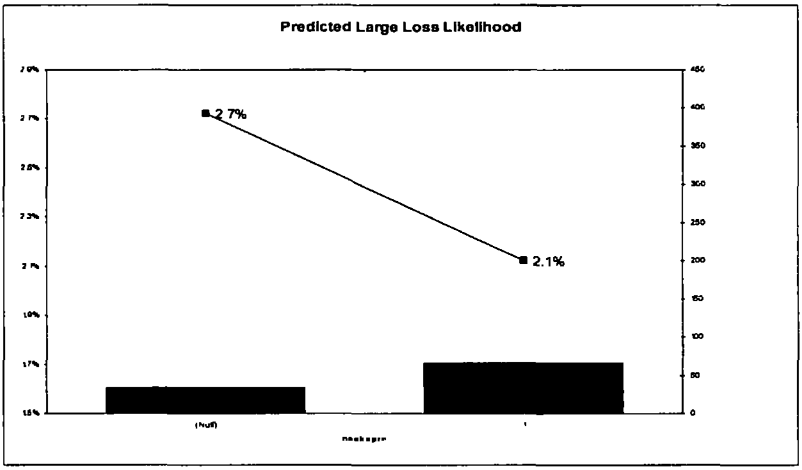


Figure 15: Likelihood of large loss with a neck injury

Accident Year

The accident year was found to be significant in the limited claim settlement analysis, and the older claims had a predicted limited severity of about 25% higher than the base accident year. However, as can be seen in Figure 16, large claims are twice as likely to result from older claims as from less mature claims. This is to be expected, since it is more likely that the more complicated, expensive claims will take longer to settle.

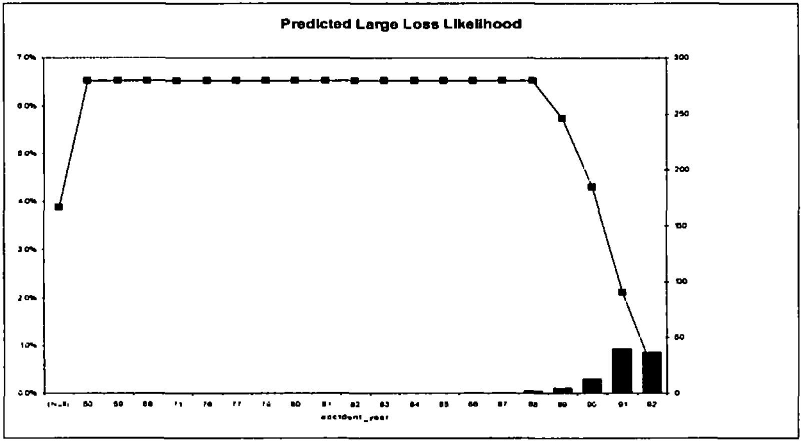


Figure 16: Accident year large claim likelihood

Final Large Claim Load

The final large claim load is calculated by taking the base predicted excess claim amount and adjusting it for the calculated likelihood of a claim turning into an excess claim as determined by the excess claim model. This calculation is different than the limited claim severity example discussed earlier due to the use of the logistic regression model. Taking the product of individual relativities cannot be directly applied here because of the upper limit on the likelihood of 1.0. The formula for the calculation of the likelihood of large loss has natural limits of 0 and 1. For each factor in the excess model, a parameter estimate is developed. The sum of the parameters for risk characteristics of particular claim is then added to the logistic parameter estimate for the base risk, and then the exponent of the negative of this sum is calculated. The final probability is then the inverse of one plus the exponent of the summed parameter. See attachment 2 for the formula and an example of the calculation of the final large claim load. The final expected claim settlement value is simply the sum of the modeled limited claim settlement value and the modeled excess claim settlement value, also shown in Attachment 2.

Evaluating the Overall Model Fit

There are a number of statistical diagnostics that can be applied in order to evaluate the overall fit of the model to the data. These measures include the difference between the observed and fitted values (errors), the standard errors of parameter differences, the chi-square test and the f-test. The last three tests mentioned here are best suited for evaluation of particular factors which may or may not be predictive in the modeling process. There is also the evaluation of the overall model structure which assists in determining if the overall model has been fit with the proper error distribution. In this particular modeling exercise, we modeled the limited claim settlement value data with a Gamma error term. To review the appropriateness of the Gamma model, we looked at residual plots (difference between actual and predicted claim settlement values) to determine whether or not the Gamma assumption makes sense.

Figure 18 shows the resulting residual plot for the limited claim severity model. The residuals have been transformed to adjust for any scale parameter differences in the model so that a better determination can be made regarding the fit of the model. Generally, you would look for a residual plot which is symmetric about 0 on the y-axis and has no obvious asymmetrical tendencies about the x-axis. If you look at the left side of the residual plot in Figure 17, the plot looks reasonable, with a fairly even distribution around 0, and with no obvious distortions, such as a fanning in or fanning out of the plot. However, if you look at the right side of the graph, you will see what appears to be a severe distortion in the residuals. This cut-off along the right side of the graph is due to the fact that we are modeling a capped severity. All of the observed severities have been cut off at $25,000, which causes the residual graph to appear truncated.

Because we are accounting for the excess claim load in a separate model, this residual graph would be acceptable. If there had been other distortions, such as a funnel shaped graph going either way, then these could have been potentially addressed by adjusting the distribution of the error structure. Another potential problem one might see with a residual plot is what appears to be two distinct sets of residuals, aggregating at different places in the residual plot. In this case, there may be a problem with the homogeneity of the underlying data, and segregating the data into more homogenous groups might be the answer. For example, if we attempted to model bodily injury settlement values along with property damage settlement values, we might see a residual plot with two distinct groups of residuals.

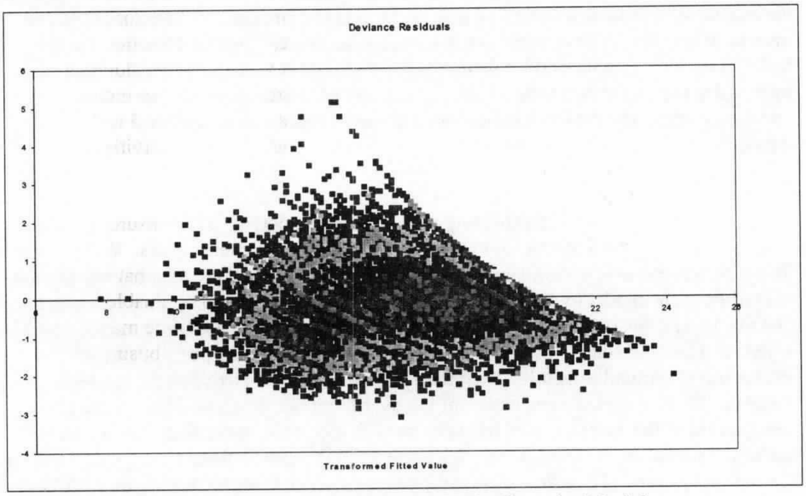


Figure 17: Transformed residual for Limited Claim Severity Model

Applications

There are a number of applications of this type of model to the insurance industry. One potential application would be its use as a tool for claims adjusters in attempting to determine reserve estimates for claims that are made to an insurer or self-insured entity. Once the claim comes in, there are certain characteristics that can be determined. These characteristics could then be used in the claim model to determine an estimated settlement value for the claim. This estimate would not be a replacement for the judgment of the claim adjuster, however, the results of this model would be available as another estimate to assist the claim adjuster in making a final estimate.

Also, there are certain characteristics of the claim that generally lead to larger claim settlement values. As a result of the claim settlement value model, claim persons could be alerted to claims which could potentially become high value claims, and then spend relatively more of their time working on the settlement of these claims. The claims model may simply confirm current common knowledge among claims personnel, such as the presence of an attorney or a fatality would cause the likelihood of a large claim to increase dramatically. It can also provide additional insight into drivers of larger than average claim settlements, especially when considering interactions.

GLM could also allow users to determine trends in claim settlement value estimates. Not only will insurers be able to determine the trend in overall claim settlement values, but it can also be determined if certain factors are increasing in importance over time in estimating the overall claim settlement value. For example, we noticed earlier that the presence of an attorney caused the limited claim settlement value to nearly double. If that relationship between claims with and without attorneys were to be begin to increase from a 2 to 1 ratio with this analysis to 2.25 to 1 with next year's analysis and 2.5 to 1 with the analysis after that, then the company may need to determine why the relativities are trending that way.

Another benefit of this type of claims settlement value model is that the insurer can make use of its own data to determine estimated ultimate claim settlement values. While there may be other models available which have been developed based on data that represents more of the industry, the use of company-specific data can be another valuable estimate that reflects the type of business that the insurance company writes. There may be differences in the claim settlement culture of the company or the type of business the insurer writes which would make a company-specific model valuable.

Conclusion

There are many benefits that the actuary brings to the insurance company. Many of these benefits are thought to be primarily in the area of ratemaking and reserving. However, the ability of the actuary to analyze past statistics and use them to help understand future occurrences has application beyond traditional areas of ratemaking and reserving. Better understanding how to estimate the ultimate claim settlement amount can assist the claims person in better estimating claim reserves. The key here is that the actuary can use his or her unique skills and provide information to claims and other areas. One important tool in providing this assistance is GLM. As actuaries continue to appreciate the potential wide applications of this analysis procedure, innovative solutions can provide value to many areas of the insurance company. Also, this analysis procedure could be applied to many different types of datasets to model different response variables.