GLM Basic Modeling: Avoiding Common Pitfalls

Geoff Werner, FCAS, MAAA, and Serhat Guven, FCAS, MAAA

Abstract Starting in the 1990's many of the larger US personal lines carriers began to implement predictive modeling techniques in the form of generalized linear modeling (GLM). Because of the early success realized by those companies, the vast majority of companies are now rushing to employ these techniques too. In their haste to keep up with competitors, many companies are making mistakes and not getting the full benefit possible.

The following are some of the most common mistakes made by companies beginning to build GLMs:

- Failing to get full buy-in from key stakeholders.
- Relying too heavily on pre-analysis.
- Using loss ratio analysis.
- Modeling raw pure premiums for all coverages directly rather than modeling at the component level.
- Restricting analysis to variables and groupings in the current rating algorithm.
- Misusing offsets.
- Treating the predictive model as a "black box".
- Limiting the use of GLMs to risk models.

This paper will address each of these pitfalls in turn. By being aware of these pitfalls, companies can hopefully minimize the transition period and achieve the full benefits of multivariate pricing as quickly as possible.

BACKGROUND

Predictive modeling has been standard practice for insurance ratemaking in the highly advanced UK marketplace for many years. While a few US companies have been doing predictive modeling for some time, it has not been until the last five years that there has been widespread acceptance of these techniques in the US marketplace.

What Is Predictive Modeling?

Essentially, predictive modeling involves using historical data to construct a statistical model that will be predictive of the future. Each observation in the historical dataset contains information or data elements that are essential in building a predictive model. Figure 1 is a visual representation of predictive modeling.

There are three types of elements in the historical dataset. First, there will be a dependent or response variable which is what the practitioner is trying to predict. For example, when modeling severity, the dependent variable is the loss amount. Second, there will be a weight associated with each observation. When modeling severity, the weight is the number of claims associated with the loss amount of the observation. Finally, there are independent variables or predictors. These are the characteristics of each observation that are being studied to ascertain whether a variable has any predictive power.

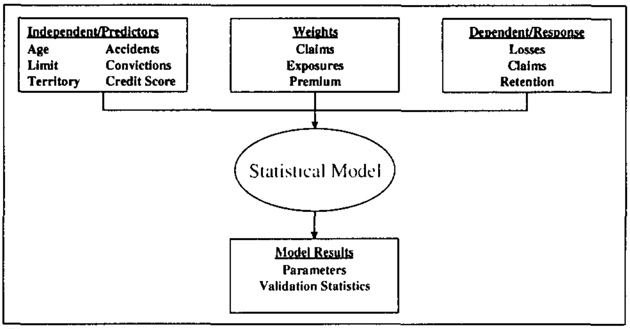

Figure 1. Overview of Predictive Modeling

The practitioner uses the historical data to build a statistical model. The output of the model is a set of parameters and validation statistics. The parameters represent the actual results; for example, when performing class plan analysis, the parameters will be the indicated relativities.[1] The validation statistics provide the practitioner with an understanding of the effectiveness of the model.

The majority of companies using predictive modeling are building such models to identify new rating variables and to better quantify the relationships between these new variables and existing rating variables.

[1] To be precise, practitioners generally use a log link function for class plan analysis. When doing so, the indicated relativities are really calculated as exp(relevant parameters).

Does Predictive Modeling Make a Difference?

Figure 2 shows a comparison of the relative predictions for automobile theft by age derived from a one-way analysis (square markers) and a simple generalized linear model (circle markers). As can be seen, the difference in these lines is quite significant for some age categories; more specifically, the percentage difference between the one-way prediction and the prediction based on the generalized linear model ranges from -38% to +6%. As this is only the difference for one factor for one cause of loss, it is apparent that the differences could be even more significant when the differences in other factors and other causes of loss are compounded.

While the mere fact that the results are different does not prove that the multivariate results are superior to the results based on the univariate analysis, it is commonly accepted that multivariate analysis corrects for methodological flaws inherent in one-way analysis and is more accurate.[2] Thus, the companies who employ multivariate techniques will be able to better predict loss costs and develop more accurate pricing structures. Companies who fail to employ these techniques will not have accurate prices and will be susceptible to adverse selection.

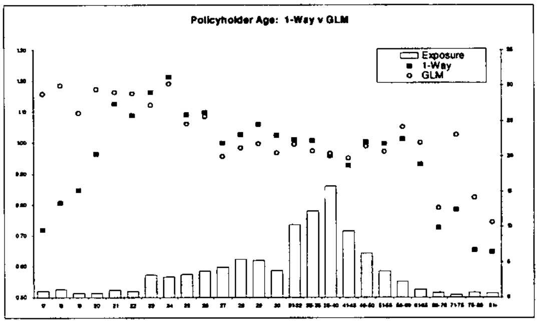

Figure 2. One-way v GLM

What are companies doing?

The number of US companies using multivariate analysis has increased dramatically over the past five years and virtually all of the companies in the top 20 are doing some form of multivariate analysis to gain a competitive advantage with respect to classification and factor analysis. The most commonly used predictive modeling technique is generalized linear modeling (GLM). The popularity of GLMs is likely due to several key advantages of GLMs as compared to traditional and other predictive modeling techniques:

- GLMs can readily adjust for both exposure and response correlations that cause one-way analyses to fail.
- Traditional statistics (e.g., loss ratios) include a systematic and unsystematic component. Like other predictive modeling techniques, GLMs allow the model to separate the components to

[2] In the example shown, the extreme difference between the indications at the youthful ages is caused by distributional biases between age and other variables (e.g., limits and cost of vehicle). A full discussion of the reasons generalized linear model results are more accurate than those produced from traditional analysis is outside the scope of this paper. For more information on this refer to "Something Old, Something New in Classification Ratemaking With Novel Use of GLMs for Credit Insurance" written by Keith Holler, David Sommer, and Geoff Trahir and published in the 1999 CAS Winter Forum.

remove unsystematic variation or the "noise" in the data and identify systematic variation or the "signal" in the data.

- Because GLM is a predictive modeling technique, it allows the user to do more with less data than traditional techniques which require significant amounts of data in each cell for "full credibility". GLMs tend to be more robust than other predictive modeling techniques and are less susceptible to over-fitting (e.g., CART or MARS) that may occur with small data sets.
- GLMs provide the modeler with a battery of diagnostics that allow for decision-making in the context of a solid statistical framework.
- GLMs allow the modeler to assume the process being modeled follows any distribution within the exponential family. The exponential family includes common distributions like Poisson and Gamma that are generally accepted as appropriate for modeling insurance data.
- GLMs are not "black box" models. Unlike some of its predictive modeling counterparts (e.g., neural nets), a GLM is easy to interpret and allows the analyst to clearly understand how each of the predictors are influencing the prediction.

Are Companies Being as Effective as Possible?

Because predictive modeling does make a difference, companies are rushing to employ these techniques. In their haste to keep up with competitors, companies are not always taking the time to perform the analysis appropriately. Consequently, many companies are making mistakes and not getting the full benefit possible. This paper is intended to address some of the most common problems companies encounter when moving from traditional techniques to multivariate analysis, such as:

- Failing to get full buy-in from key stakeholders.
- Relying too heavily on pre-analysis.
- Using loss ratio analysis.
- Modeling raw pure premiums for all coverages directly rather than modeling at the component level.
- Restricting analysis to variables and groupings in the current rating algorithm.
- Misusing offsets.
- Treating the predictive model as a "black box".
- Limiting the use of GLMs to risk models.

This paper will address each of these pitfalls in turn. By being aware of these pitfalls, companies can hopefully minimize the transition period and achieve the full benefits of multivariate pricing as quickly as possible.

KEY STAKEHOLDER BUY-IN

It is important for all key stakeholders to support any major change to pricing techniques. The main reason that companies using multivariate techniques are improving their results is because these techniques more accurately predict the risk. Thus, by definition, the multivariate results are different—and in some cases significantly different—than the univariate results. While all of the key stakeholders do not need to fully comprehend the mathematics associated with GLMs, it is imperative they recognize the benefits of multivariate analysis and that they expect the results to be different than those from prior reviews.

The best way to communicate information of this sort is to provide simple examples based on company data that highlight distortions caused by one way analysis. To do so, the practitioner should build a very simple GLM, look for a variable for which the difference between the one-way and GLM result is materially different, and examine the correlation statistics to find a variable that is likely contributing to the difference. Those two variables can be used as a simple case study of how one-way analysis can lead to inappropriate conclusions. As this represents a major shift in mindset, it is likely that that practitioner will have to go through this process more than once.

Once management is convinced that the multivariate results are superior, the modeler can compare the results of a "quick and dirty" predictive model to the current rates to determine an estimate of the subsidization inherent in the current rates. (If available, indications based on traditional analysis can also be included in the comparison to highlight differences.) Figure 3 is an example of how this type of information can be displayed for non-technical audiences. For each individual observation, the modeler calculates Current Premium/Indicated Premium – 1.00 and categorizes each observation in the range corresponding to the inadequacy or excessiveness associated with that observation. For example, the XXX exposures in the 5%-10% bucket are risks whose current premium is 5% to 10% above the indicated premium. The cross-hatched bars represent risks whose current premium is below the indicated premium (i.e., risks being subsidized) and the solid bars represent risks whose current premium is above the indicated premium (i.e., risks subsidizing). If all risks were being charged the right rate, every observation would be in the bucket containing 0%. If the bars are spread out as in Figure 3, then there is considerable subsidization present in the rating plan. That subsidization represents opportunity for improvement.

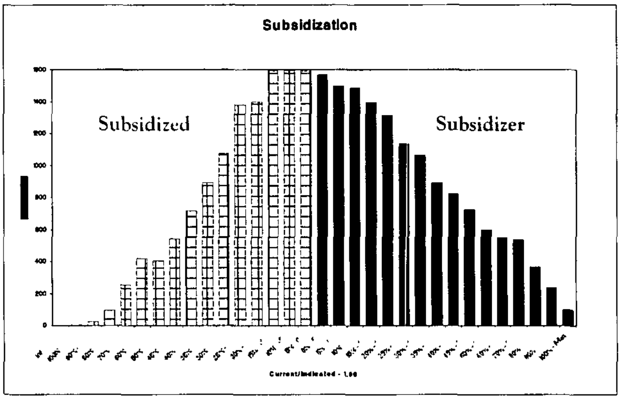

Figure 3. Subsidization Histogram

Unfortunately, in many companies the group responsible for producing the indicated relativities makes the mistake of not doing the appropriate up front communication. Failure to do so invites significant resistance when the indicated multivariate relativities are not in line with the traditional univariate results on which the others within the company have been basing decisions for years. This resistance usually leads to undesirable compromises. For example, senior management may choose to only implement one aspect of the multivariate analysis. In the best case, this weakens the effectiveness of the plan; in the worst case, it can actually result in implementation of plan that is inferior to the current plan.[3] Interestingly, gaining company-wide acceptance is usually most difficult for companies who were the most successful previously. In such companies, senior management is often very reluctant to abandon methodologies that were used to achieve the success.

PRE-ANALYSIS

Performing pre-analysis can be helpful to get a feeling for the data. Pre-analysis oftentimes includes traditional one-way data, volume measures, and correlation statistics.

Traditional one-way data includes items like raw frequencies, severities, pure premiums, and loss ratios. Examining these ratios can help the practitioner in three ways. First, it can help the modeler spot items that may distort the analysis (e.g., extraordinary losses) if left unadjusted. Second, it will highlight what others in the company may be examining, so the practitioner will be better prepared for difficult discussions when the multivariate results are different than the univariate results. Third, it allows the modeler to build a priori expectations that can help when applying judgment during the modeling. The mistake some practitioners make is putting too much emphasis on the univariate results. It is important to recall that the whole point of using multivariate analysis is to correct for the flaws inherent in univariate analysis; therefore, the results will have some differences and it is imperative the modeler does not allow the univariate results to bias judgment and ultimately limit the benefit.

Volume analysis usually includes an examination of the distribution of exposures, claim counts, and premiums. Volumes can help the practitioner decide the appropriate number of years of data necessary and whether there is enough data to do test/training analysis. The mistake some modelers make is to use these volumes to calculate traditional estimates of credibility. Traditional estimates of credibility are not necessary within the context of predictive modeling and are replaced with better diagnostics that indicate the amount of reliance that should be given to individual estimates.

Finally, correlation statistics (e.g., Cramer's V) inform the practitioner which independent predictors have a high degree of exposure correlation. The mistake too many companies make is using this information to eliminate correlated variables from the modeling process before it even starts. The practitioner should not eliminate variables at this stage, but rather note the high correlations and be aware that the inclusion/exclusion or offsetting of one of the factors will have an impact on the other factor. By not ruling out the variable before the analysis starts, the practitioner can test the various variables within the multivariate framework to determine whether only one or a combination of the highly correlated variables should be included.

Basically, if the intent is to do multivariate analysis, then the practitioner should perform multivariate analysis and avoid the temptation to make decisions during the pre-analysis stage. By making decisions before the multivariate analysis begins, the practitioner is only limiting the potential benefit.

[3] The careful use of offsets can help minimize the adverse impact of implementing GLM results on a piecemeal basis. Offsets are discussed in detail in a later section.

PURE PREMIUM V. LOSS RATIO ANALYSIS

When it comes to risk modeling, many companies are doing loss ratio modeling, rather than frequency/severity modeling. There are both practical and theoretical reasons that modeling loss ratios is less preferable.

If loss ratios are being modeled, it is imperative that the loss ratios be calculated on current rate level; failure to do so will make the resulting relativities inappropriate. Because the practitioner is performing multivariate analysis, it is not sufficient to use an on-level approximation method that applies an average current rate level factor to a diverse set of observations. Instead, the loss ratio for each observation must be put on the correct rate level which is done via extension of exposures (i.e., re-rating) and should be "re-underwritten" if underwriting rules were changed. This can be a very difficult--and for many companies --an impossible task. When modeling frequency and severity, the modeler uses exposures rather than premium. Thus, no current rate level adjustment is necessary.

It is widely accepted that there are standard distributions for frequency and severity. Generally speaking, a Poisson error structure is appropriate for frequency modeling and a Gamma error structure is appropriate for severity modeling. Loss ratios, on the other hand, do not follow a typical error structure as the distribution will be highly dependent on the rating structure of the individual company. This unnecessarily adds another level of uncertainty to the modeling process.

Judgment is an important part of the modeling process; therefore, it is helpful if the modeler is able to formulate some a priori expectation. When modeling with a pure premium approach, the practitioner can break the data into frequency and severity components and use their knowledge to formulate reasonable expectations. For example, when modeling auto collision frequency, the modeler may expect the age curve to decrease from youthful to adult and increase again for the most mature drivers.[4] Figure

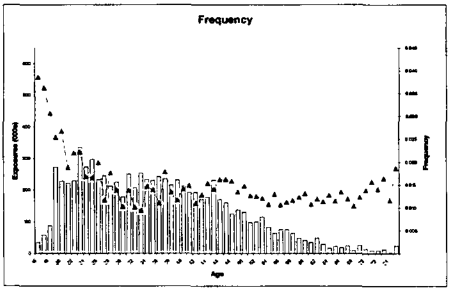

Figure 4. Raw Frequencies By Age

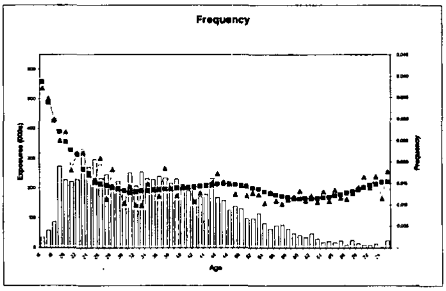

Figure 5. Raw and Modeled Frequencies By Age

[4] Depending on assignment rules, there may be a hump in the middle around the age that teenage drivers are added to the policy.

4 shows an example of raw frequency data by age of driver. To the extent that the pattern is erratic, the modeler will be able to use appropriate techniques (e.g., fit a curve or group levels) and knowledge about insurance to build a model that is captures the signal in the data (see Figure 5). If, on the other hand, the modeler is modeling loss ratios, the only expectation is that the loss ratios will be the same if the rates are perfect. Given that the current rates are probably not perfect, the modeler cannot know whether a resulting erratic pattern in the age results is due to random noise or is a very real pattern due to the underlying rates. Figure 6 shows the raw loss ratios by age for the same dataset. It is clear that it would be difficult to distinguish the noise from the true signal in the data.

Another practical issue is that loss ratio models become obsolete as soon as any change is implemented. Thus, the loss ratio models built during one review cannot be used as a starting point for any subsequent reviews. In contrast, the frequencies and severities of individual observations do not change just because a rate adjustment is made. Thus, the frequency and severity models built with one review should be a very good starting point for the next review.

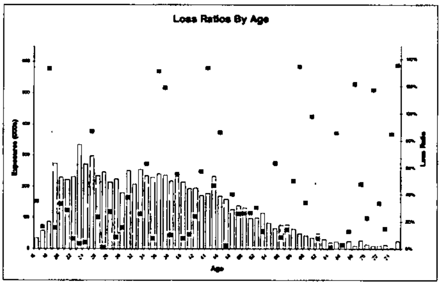

Figure 6. Raw Loss Ratios By Age

If loss ratio modeling has these issues, why do companies do it? There are three basic reasons that practitioners may model loss ratios.

1. Premiums may be readily available and exposures are not. As it is generally easier to obtain exposures than premiums this situation is rare.
2. When modeling pure premiums it is important to include all the variables in the modeling dataset.[5] In some cases, the practitioner may not be able to get all of the variables on the database and uses the premiums to try to account for some of the variation in the missing variables. This, of course, assumes the premiums reflect the true indication, which is often not the case. Additionally, this is only appropriate if the premiums are on level at the granular level.
3. The most common reason seems to be precedent. Historically speaking, companies were performing univariate analysis techniques. With univariate techniques, using loss ratios is more accurate than using pure premiums as loss ratio analysis does a better job of coping with distributional biases in the univariate world. Because of this, companies seem to have gotten into the habit of working with loss ratios. Now that companies are performing multivariate analysis, the reasons that loss ratios outperformed pure premiums in the univariate world are no longer applicable.

So, for these practical and theoretical reasons, loss ratio modeling should only be employed out of necessity. Instead, companies should pursue pure premium modeling at the component level as it will lead to better models and ultimately increased benefits.

[5] At a minimum the practitioner should include all variables that have significant correlation with other independent variables.

COMPONENT LEVEL MODELING

Once a company understands the advantages associated with pure premium modeling relative to loss ratio modeling, the next question is whether the company is going to model frequency and severity separately by cause of loss.

For years, analysts have been performing traditional loss trend analysis. When the data is available, traditional analysis is typically done by coverage or peril and for frequency and severity separately as that is the most effective way to discern underlying trends in the data. Despite that precedent, many companies try to build predictive models on a combined level.

Many companies, especially personal lines homeowners and commercial lines insurers, are tempted to model all coverages or perils combined to save time. Most homeowners insurers can separate the data by peril and should do so. While commercial lines carriers may not have the option of completely separating the data, workers compensation carriers have medical versus indemnity readily available and general liability carriers know property versus liability losses. If a practitioner wants to model on a combined basis, a quick analysis of the residuals highlights whether or not the different perils or coverages can effectively be combined. Figure 7 is an example of a plot of the residual (i.e., actual – predicted) for every observation. The x-axis represents the fitted value and the y-axis represents the magnitude of the residual. In this example, the plot has two separate concentrations. This appearance is typically seen when multiple causes of loss are included and the model is not effectively handling them. If this pattern is seen, the data should be separated and the coverages/perils should be modeled individually. If it is not possible to separate the data, then the practitioner should consider employing dispersion modeling techniques. Dispersion modeling is an advanced topic that is beyond the scope of this paper.

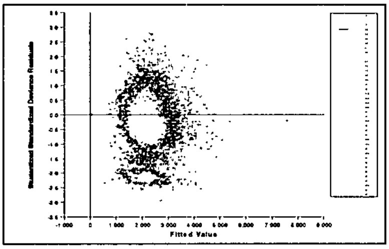

Figure 7. Residual Plot

A more common shortcut companies attempt is to model raw pure premiums directly rather than modeling frequency and severity separately. Interestingly, the anticipated time savings is usually not achieved. When modeling separately, severity modeling is usually very straight-forward and takes little time to find simple trends amongst the noise. More time is usually spent building a good model on the stable frequency data. Practitioners may believe there is time savings based on an assumption that there is only one model being built, but if raw pure premium modeling is done properly, that is not the case. In order to properly reflect the bimodal nature of pure premiums, the analyst should also build a dispersion model that coincides with the Tweedie model.[6] So, whether modeling frequency and severity separately or modeling pure premiums, the analyst is still building multiple models. But with raw pure

[6] A discussion of dispersion modeling is outside the scope of this paper. For more information on this refer to "Fitting Tweedite's Compound Poisson Model to Insurance Claims Data: Dispersion Modelling" written by Gordon Smyth and Brent Jorgensen and published in Astin Bulletin Volume 32, Number 1.

premiums, the practitioner has two challenges. First, the "noise" created by combining frequency and severity makes it more difficult to spot trends in the raw pure premiums. Second, the practitioner has the added complexity of modeling and interpreting the dispersion parameter.

LIMITATIONS ON THE ANALYSIS

When performing a class plan analysis, companies are often tempted to restrict the model to include only those variables that are currently in their rating algorithm. Additionally, for selected variables, companies often group the data to be consistent with the current rating algorithm. For example, if the rating algorithm charges the same rate for ages 30-49, companies will often automatically group ages 30-49 in the model.

The best approach for this type of modeling is to make use of all available data for the initial modeling. This includes importing all the current and potential rating variables, all available underwriting data, and any external data, if available. Some companies will interrogate 200-300 variables. These are the companies that are most likely to find the next really predictive variable.

Figure 8 represents the process when the data is being fully interrogated. The individual frequency and severity models should be built without regard to the current rating algorithm. In other words, the analyst's goal should be to use all available data to build the most predictive frequency and severity models possible. It is not necessary that the frequency and severity models be consistent. In fact, in all likelihood, the model structures for frequency and severity will be different.

After the frequency and severity models are built, the resulting frequency and severity predictions can be combined on an observation by observation basis to form modeled pure premiums. If the underlying component models are built correctly, the unsystematic variation will be removed and the modeled pure premiums will represent the systematic variation in the historical data. The practitioner then builds a pure premium or constraint model using the modeled pure premiums. It is at this stage that the practitioner limits the variables to those that will be used in the proposed rating algorithm, incorporates restrictions on rating variables, and develops underwriting rules that compliment the rates.

By limiting the data to existing variables, companies are limiting their opportunities for major improvement. The goal should not only be to improve the accuracy of existing rating structures, but also to find that new variable that is predictive and can provide a real competitive advantage.

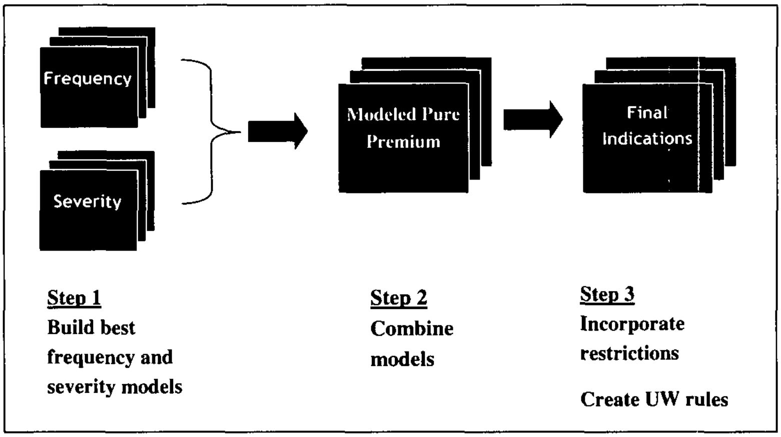

Figure 8. Modeling

INTELLIGENT USE OF OFFSETS

Offsets can be used to specify known relationships between levels of a specific factor or factors within the GLM; consequently, offsets are frequently used when the practitioner wants to fix the relativities of one or more rating factors due to some internal or external constraints. For example, a company may not want to change the multi-policy discount, so the practitioner may use an offset to force the GLM relativities to be consistent with the current discount for that factor. When an offset is used the data is adjusted so as to force the resulting parameters for that variable or variables to the desired values. The parameters for the other variables will change to try to "make up" for the difference and avoid any double-counting. Variables that are highly correlated with the offset variable(s) will change the most.

Companies who do not truly understand the implications of using offsets fail to consider there may be situations where offsets are inappropriate and, therefore, end up with unwanted results. In reality, there are situations when using offsets may be desirable and there are situations when using offsets may be undesirable. It is reasonable to assume that amount of insurance (AOI) and territory are highly correlated (i.e., homes in a particular area tend to have similar AOIs). In light of this, consider the following two examples.

1. Consider the case that a systems constraint forces the analyst to cap the relativity on homes with AOIs over $500K. This represents an undesirable subsidy that the practitioner may want to minimize. The practitioner has two options. On one hand, the practitioner can calculate the indicated GLM relativities without any offsets and cap the indicated relativity for homes over $500K. The impact of this is that the base rate will need to be adjusted to make up for the shortfall. Alternatively, the practitioner can use an offset term. By doing so, the other variables adjust to try to make up for the shortfall. The most significant changes will occur in variables that are highly correlated with the variable being constrained. In this example, the relativities in territories with a relatively heavy concentration of high-valued homes will increase to make up

for the shortfall. Note, as no variable will be perfectly correlated, a minor base rate adjustment will likely be required.

2. Assume the company makes a decision to target high-valued homes. In an effort to increase market share, the practitioner is instructed to implement relativities lower than those indicated for homes $500K and over. In this case, the practitioner does not want to use offsets. As discussed in the preceding paragraph, using offsets increases the relativities in the territories with high-value homes. Since it is immaterial whether the premiums are high due to the AOI curve or the territorial relativity, using an offset simply undoes the desired subsidy. So, if the practitioner is trying to implement a desirable subsidy, then the free-fitted relativities should be left in the model and the desired relativities should be changed outside the modeling process.

The impact of offsets is particularly important to consider in the initial stages of multivariate analysis implementation. At the onset, there are usually significant differences between the current rating plan and that indicated by the multivariate analysis. Due to regulatory constraints and renewal impact considerations, it may be difficult for a company to make all of the changes at once. If that is the case, the company needs to decide whether or not to use offsets for the variables that are not going to be changed. The following example is intended to illustrate the impact of using offsets.

The results of a full GLM analysis will be indicated relativities for all factors. Figure 9 represents the indicated and current relativities for 0, 1, 2, 3, and 4+ years of claims free driving[7] and figure 10 depicts the indicated relativities for drivers age 16-19. Assume that the company wants to implement a change to the youthful relativities with this review, but wants to keep the same relativities for number of years of claims free driving until the next review.

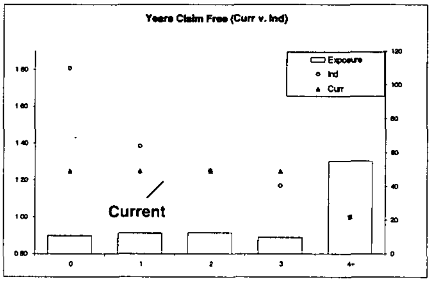

Figure 9. Indicated Years Claim Free Relativities

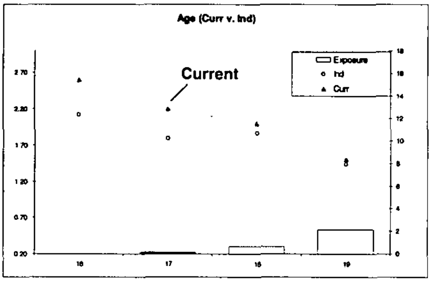

Figure 10. Indicated Age Relativities

The modeler must decide whether to use an offset to minimize the subsidization introduced by maintaining the current relativities for years of claims free driving; if an offset is used, the other factors will adjust to make up for the subsidy. The amount the other factors will adjust is dependent on how correlated the other factors are to the years of claims free driving. Figure 11 is a graphical

[7] In this example, 4+ years of claim free driving is the base (i.e., a factor of 1.00). Thus, there are surcharges (i.e., relativities above 1.00) for 0, 1, 2, and 3+ years of claims free driving.

representation of the exposure distribution by operator age (x-axis) and years claim free (y-axis). Each of the bars stack to 100% with each segment representing a different number of years of claim free driving. If the variables are not correlated, then the segments are consistent for each bar. As can be seen by figure 11, age and years claim free are very highly correlated. More specifically, the younger ages (left side) have a high concentration of drivers with 0 and 1 years of claim free driving and the older ages (right side) have a high concentration of drivers with 4+ years of claim free driving. Thus, the inclusion of an offset for claims free driving will likely have an impact on the age relativities.

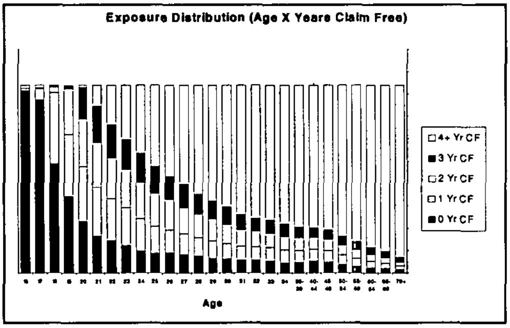

Figure 11. Distribution by Age and Years Claim Free

Figure 12 is the same chart as figure 10 with the addition of the indicated age relativities after the inclusion of the offset for years of claims free driving. In this case, the indicated relativities for ages 16-19 increased significantly to "make up" for the shortfall caused by not implementing the fully indicated surcharge for only 0 and 1 years of claim free driving. This example actually highlights the interesting case where the current is actually between the indicated and indicated with offset. Assuming the company looks at both indications, the company will realize there is an important business decision to be made. First, the company can move toward the indicated with offset as that will minimize the inequity in the rating plan. The con associated with that approach is that if the claims free driving relativities are changed with the next review, the age relativities will need to be lowered to be equitable. This leads to an increase in the age relativities in one year and a decrease in the next, which increases the chances of big premium swings for individual risks. On the other hand, the company can maintain the current age relativities or even move toward the indicated relativities without the offset. This has the benefit of reducing the potential for big premium swings, but does not correct the short run inequities in the rating plan.

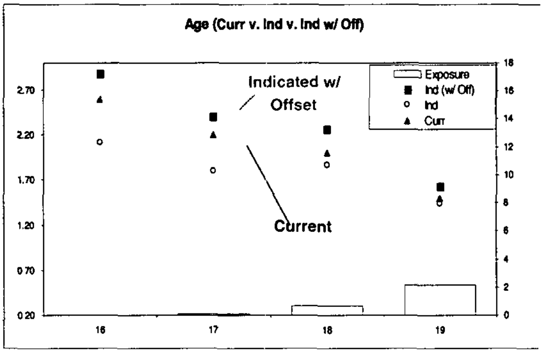

Figure 12. Ages 16-19

The point of these examples is not to encourage companies to avoid offsets. Offsets are a very important aspect of modeling with GLMs. Rather, it should be noted that offsets need to be used with an understanding of the effects so that the practitioner does not get unintended consequences. One way to prevent unintended consequences is to view indicated relativities with and without the offsets.

GLMS ARE NOT A BLACK BOX

One of the advantages of GLMs as compared to other predictive modeling techniques (e.g., neural nets) is that GLMs are not black boxes. Unfortunately, too many practitioners make the mistake of treating GLMs like black boxes. This results in models that are inappropriate and blind implementation of results that may be counterintuitive.

Figure 13 depicts the iterative nature of modeling with GLMs[8]. While certain tasks can be automated, it is important to understand that--when built properly--GLMs require involvement by the analyst in the process. The very process of building GLMs provides the analyst with valuable insights into the data and an ability to provide judgment where necessary.

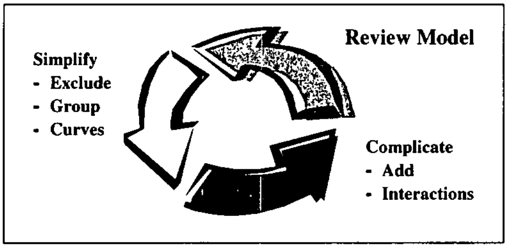

Figure 13. Iterative Modeling Process

One common mistake that companies make is to overcomplicate the models. In other words, the company correctly includes as much data as possible in the model process, but is not appropriately judicious in building the models. A quick review of companies' rating plans highlights that many companies have incredibly complex rating algorithms. It is not uncommon to see variables interacted with many different variables throughout the algorithm. Interestingly, when this is present the relativities included in the various tables tend to all be close to 1.00 suggesting the table was probably an unnecessary complication. This is an area where we should take a lesson from our counterparts in the UK who have been performing this analysis for years. Despite the fact that they have significantly more rating freedom, their rating algorithms are generally much simpler than ours. By examining proper diagnostics (i.e., standard errors, consistency tests, and type III tests) and supplying appropriate judgment throughout the modeling process, the modeler will not introduce unnecessary complications and the resulting rating algorithm will be more predictive and more manageable. Thus, the practitioner will be in a better position to make changes quickly in the future.

In contrast to the companies discussed in the preceding paragraph, there are a significant number of companies who do not introduce enough complexity in the model. A common example of oversimplification is over-smoothing. For example, a company may fit a single curve to smooth out the auto frequency data by age when multiple curves would be a better representation of the true signal. Again, GLM is not a black box. Instead, it is an iterative tool that requires human intervention. While it

[8] The picture shows that the process of building a GLMs involves determining an initial model and then testing to determine what simplifications and complications can be made to improve the performance of the model. Simplifications involve excluding variables that are not predictive, grouping levels within a variable (e.g., ages 75+) that do not add any additional predictive value, and fitting curves to continuous variables. Complications include adding new variables that can help explain variation and incorporating interactions which allow the relationship between levels of one variable to vary by level for another variable (e.g., the relationship between males and females varies by age). The circular nature is intended to show that it is an iterative procedure. As it is multivariate analysis, decisions made at any stage can impact previously made decisions.

may be wise to start slow, the practitioner should ultimately progress to the point where he/she is using the modeling as an exploratory opportunity to identify new material patterns in the data.

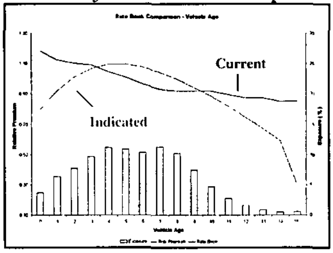

Figure 14. Indicated v Current

The other major mistake that companies make in this regard is blindly implementing results without applying appropriate judgment. A quick review of rating pages of various companies will often uncover patterns in relativities that seemingly make no sense. Figure 14 shows a comparison of indicated and current model year relativities. The current relativities are consistent with expectations as they decrease as the model year gets older. Interestingly, the indications suggest the true risk is better represented by a hump-shaped curve with the highest relativities being in the middle. This, of course, is counterintuitive and the practitioner should dig deeper into the true cause of this pattern. Since these indicated pure premium relativities were derived from the combination of frequency and severity models, the practitioner will gain insights and can work with experts from various areas within the company to determine if there are operational processes that are driving the pattern. If the overall pattern is a frequency phenomenon, then it is prudent to start with the underwriting function. For example, a pattern like this could result from underwriting rules that are especially strict for newer vehicles. If, on the other hand, the overall pattern is really a severity phenomenon, then it is prudent to start with the claims function. For example, a pattern like this could result from claims adjusters who are being relatively generous when settling claims for middle aged vehicles. By understanding the results and having cross-functional discussions about the results, the company can address the issue using the most appropriate lever (pricing, underwriting, or claims).

GLMS ARE A BUSINESS TOOL

As GLMs are generally being championed by actuaries, most of the focus has been on using GLMs to determine relativities for rating structures. Even in this paper, the examples have focused on risk modeling. However, the benefits of GLMs are clearly not restricted to the application of pricing. The following are a few of the other applications for which companies are already using GLMs:

- Practitioners are using GLMs to reduce a variety of risk variables into one score. This has obvious application in regards to creating underwriting tiers, credit scores, fire protection scores, vehicle symbols, etc.
- Many companies have begun to perform elasticity modeling. By building elasticity models for new and renewal business, companies can predict the impact of various actions on market share. A few companies are already linking the profitability and elasticity models to find the optimal pricing decision.
- Claims handlers are starting to see the advantages of GLMs and are using them to help set more accurate reserves and to provide early identification of claims that may be fraudulent or are most likely to end up in a lawsuit.
- Competitive analysis units are using GLMs to reverse-engineer competitors' rates given a large sample of rating quotes.

These are just a few of the other ways that the more advanced companies are using GLMs. Companies should not fall into the trap of thinking GLMs are only for pricing. Instead, companies should realize that GLMs can be used in a variety of circumstances given a historical database with multiple predictors and a response they would like to predict.

CONCLUSION

Predictive modeling is a very powerful tool that has been used effectively by insurance practitioners in other countries for many years. Starting in the 1990's many of the larger US personal lines carriers began to implement predictive modeling techniques in the form of generalized linear modeling (GLM). The early results from the US show that predictive modeling paid large dividends to those companies who embraced it.

Due largely to the success of the first US companies, the US is experiencing a push by many companies who want to implement predictive modeling before they are left behind. With that push companies are already getting benefits. Unfortunately, in a haste to get going, many companies have taken some shortcuts or simply made mistakes that have kept them from realizing the full benefits. By taking a step back and reconsidering some of the prior processes and decisions, companies will be able to maximize the benefits of these new techniques.