APPENDIX F: MULTIVARIATE CLASSIFICATION EXAMPLE

This appendix includes example output from a GLM analysis. It includes several tests used to evaluate the predictive power of a potential rating variable and hold-out sample testing used to evaluate the overall effectiveness of a particular model.

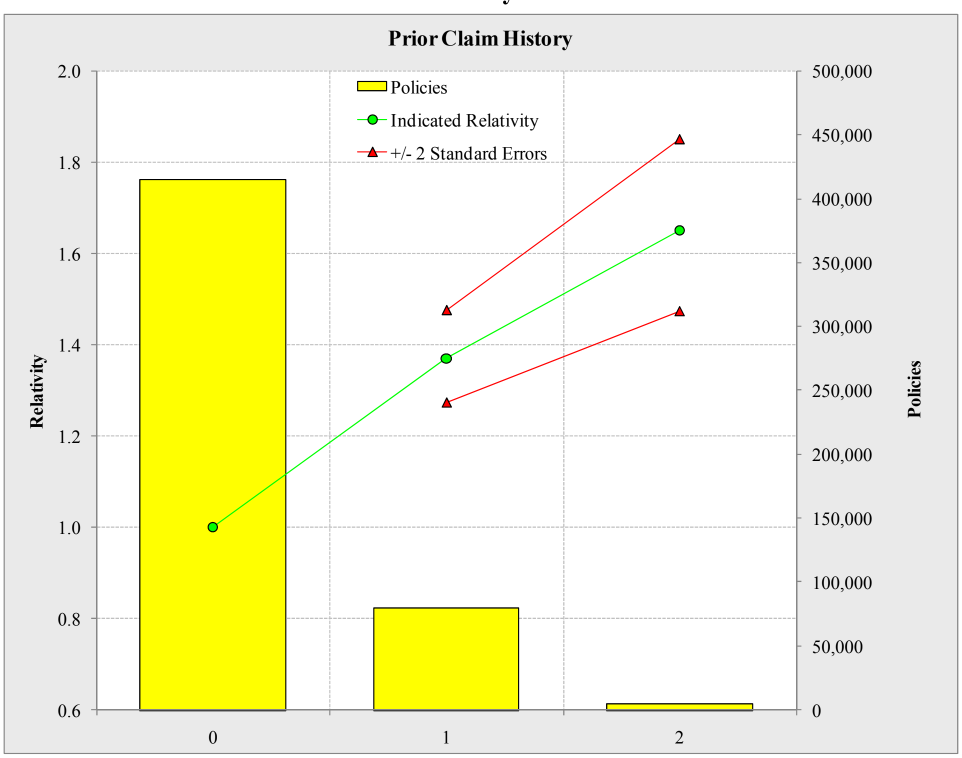
EXAMPLE PREDICTIVE VARIABLE

This section contains sample output from a multiplicative GLM fit to homeowners water damage frequency[58] data. The graphical output isolates the effect of the prior claim history variable as a significant predictor of water damage frequency, though the model contains other explanatory variables that must be considered in conjunction with the prior claims history effect.

Parameters and Standard Errors

The following graph displays the indicated frequency relativities for prior claims history, all other variables considered. The categories on the x-axis represent the levels of the variable (0, 1, or 2 claims). The level for zero prior claims is the base level, and the relativities for the other levels are expressed relative to it. The bars relate to the right y-axis, showing the number of policies in each level. The line with the circle marker shows the indicated relativities, and the lines with the triangle markers represent two standard errors on either side of the indicated relativities.

F.1 Main Effect Test for Prior Claim History



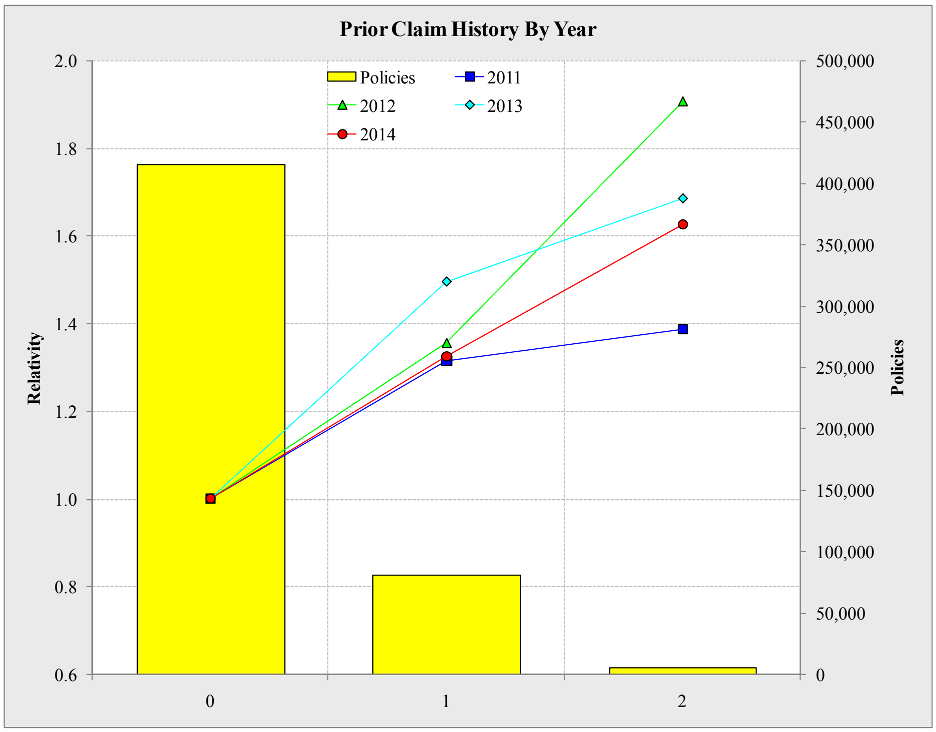
[58] It is common for actuaries to build frequency and severity models for each major peril or cause of loss.

The fact that the indicated relativity line is upward sloping with relatively tight standard errors suggests that the expected frequency is higher for risks with prior claims. More specifically, risks with one or two prior claims have a frequency that is approximately 35% and 65% higher than risks with no prior claims.

Consistency Test

The prior graph shows the indicated relativities for the whole dataset. The following graph shows the pattern of relativities for each of the individual years included in the analysis. (In some cases, the actuary may use random segments of the dataset rather than individual years.) Like the last figure, the categories on the x-axis represent the number of prior claims, and the bars are the number of policies in each level. The lines represent the indicated frequency relativities for prior claims history, separately for each year.

F.2 Consistency Test for Prior Claim History



The fact that each year's indicated line slopes upward with roughly the same shape suggests that the pattern is consistent over time. This provides the actuary with a practical test supporting the stability of this variable's predictive power.

Statistical Test

The actuary can also test the predictive power of a variable using statistical diagnostics such as deviances. One common deviance test is the Chi-Square test. In this test, the actuary fits models with and without the variable being studied and analyzes the trade-off between the increased accuracy of the model with the variable included versus the additional complexity of having additional parameters to estimate. The null hypothesis is that these two models are essentially the same. A Chi-Square percentage is calculated based on the results of the two models. A Chi-Square percentage of less than 5% generally suggests the actuary should reject the null hypothesis that the models are the same and should use the model with the greater number of parameters.

In this example, the Chi-Square percentage is 0.02%. Thus, the actuary rejects the null hypothesis and selects the model with the greater number of parameters. In other words, the actuary selects the model with the prior claims history variable in it.

Judgment

It is important that the actuary evaluate the reasonableness of the model and diagnostic results based on knowledge of the claims experience being modeled. In this case, the statistical results are consistent with the intuitive expectation that frequency is higher with the presence of prior claims.

Decision

All four tests suggest the rating variable is predictive and should be included in the model (and ultimately the rating algorithm).

EXAMPLE UNPREDICTIVE VARIABLE

This section contains sample output from a multiplicative GLM fit to homeowners wind damage frequency data. The output isolates the effect of fire safety devices as an insignificant predictor of wind damage frequency, though the model contains other explanatory variables that must be considered in conjunction with this variable.

Parameters and Standard Errors

The following graph shows the indicated frequency relativities for the fire safety device variable, all other variables considered. The x-axis categories represent the different fire safety devices (the base being the level "none"), and the bars are the number of policies in each level. The lines represent the indicated wind damage frequency relativities and two standard errors on either side of the indicated relativities.

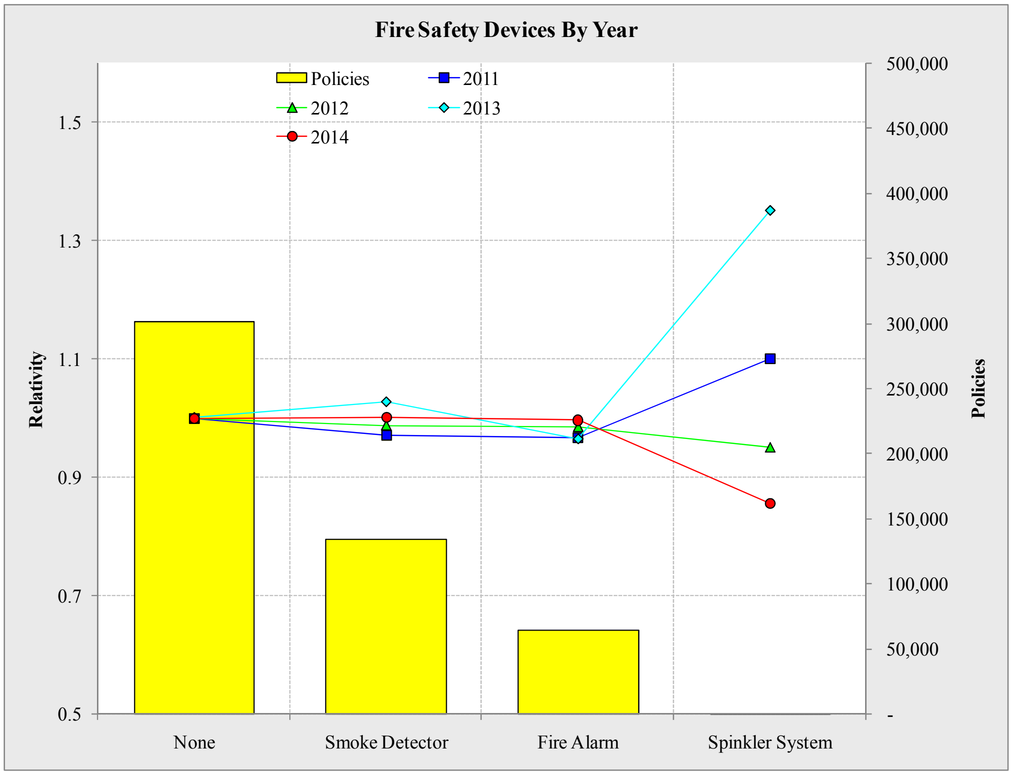
F.3 Main Effect Test for Fire Safety Device

The indicated line is basically flat (i.e., indicated relativities are close to 1.00) for the levels that have a significant number of policies. The one category that has an indication substantially different than 1.0 (sprinkler system) has very wide standard errors around the indicated relativity, which is likely due to the small number of policies in that category. Thus, there appears to be little predictive power in this variable, and it should be removed from the wind damage frequency model.

Consistency Test

The following figure shows the pattern for each of the individual years included in the analysis. Like the last graph, the categories on the x-axis represent different fire safety devices, and the bars are the number of policies in each level. The lines represent the indicated relativities for each year.

F.4 Consistency Test for Fire Safety Device Claim



The patterns are consistent across the years for all categories but the sprinkler system. That category has little data, and the predictions are very volatile. These results confirm the conclusions derived from the parameter results and standard errors.

Statistical Test

The Chi-Square percentage for this variable is 74%. Percentages above 30% indicate that the null hypothesis, which asserts the models are the same, should not be rejected. If the models are "the same," then the actuary should select the simpler model that does not include the additional variable. (Chi-Square percentages between 5% and 30% are often thought to be inconclusive based on this test alone.)

Judgment

The existence of smoke detectors, sprinklers, and fire alarms does not seem to have any statistical effect on the frequency of wind damage losses. This is consistent with intuition.

Decision

All four tests suggest the rating variable is not predictive and should be excluded from the wind damage frequency model.

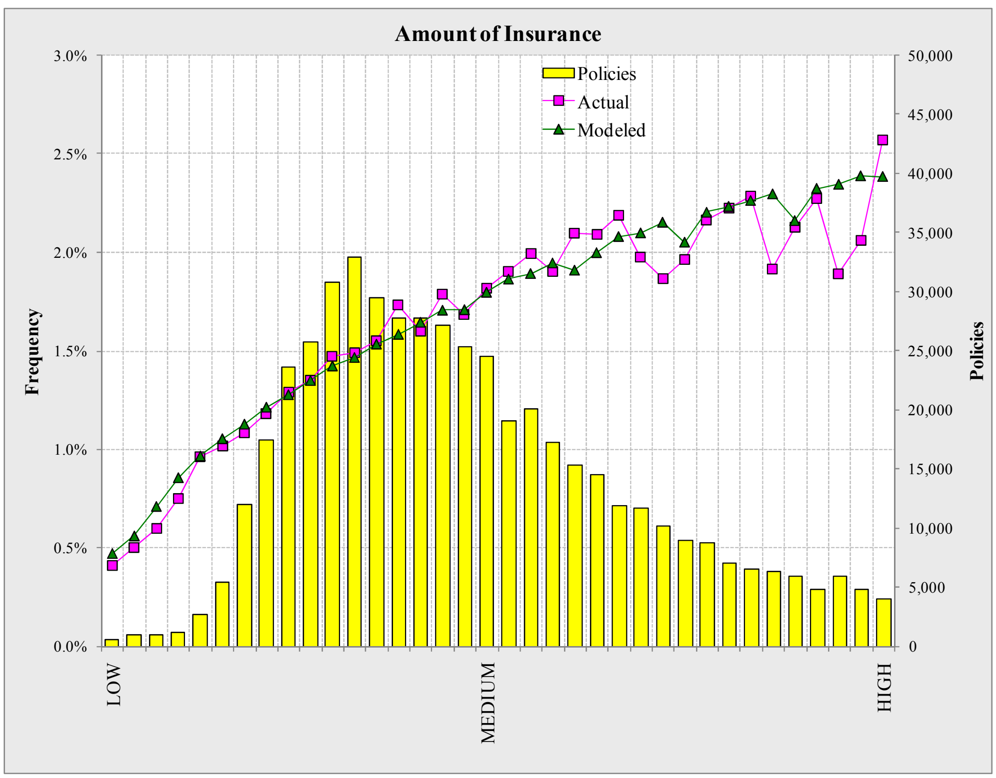
OVERALL MODEL VALIDATION

There are many tests that analyze the overall effectiveness of a given model, the most common of which compares predictions made by the model to actual results on a hold-out dataset (i.e., data not used to develop the model). This test does require that companies set aside a portion of the data for testing; this may not always be possible for smaller companies.

Validation Test Segmented by Variable

The following graph shows the observed and predicted frequencies for various levels of amount of insurance. If the model is predictive, then these frequencies should be close for any level with enough volume to produce stable results. The random nature of the insurance process will create small differences between the lines; however, either large or systematic differences or both should be investigated as possible indicators of an ineffective model. For example, the model may contain too much noise caused by retaining statistically insignificant variables or not have enough explanatory power because statistically significant variables are omitted.

F.5 Actual Results v Modeled Results for AOI

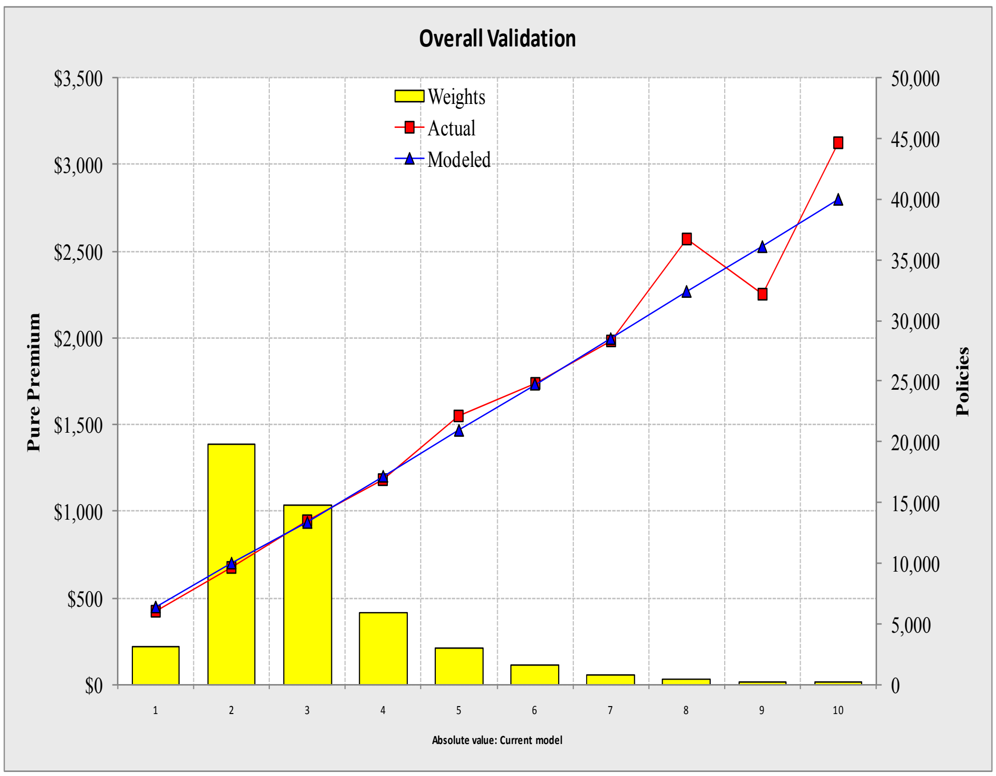


In viewing this graph, it is important to note that amount of insurance is a variable for which there is a natural order to the different levels (i.e., amount of insurance $201,000 is between amounts of insurance $200,000 and $202,000). In general, the results show a close match between expected frequencies from the model and actual claim frequencies. In particular, however, the modeled results for the first four levels appear to be higher than the actual results, suggesting that the model may be over-predicting the frequency for homes with low amounts of insurance. Similar-sized discrepancies can be seen for the medium amounts of insurance (where the actual results appear higher than the modeled results) and the high amounts of insurance (where the actual results appear lower than modeled results but with considerable volatility).

Validation Test Segmented by Fitted Value

In the following figure, the underlying frequency and severity models were used to determine a modeled pure premium for each observation in a hold-out dataset. Then, each observation was ordered according to the modeled pure premium result from the lowest to highest expected value. The observations were then grouped into 10 groups, and the actual and modeled results for each group are compared on the same chart. If the model is predictive, the actual result will be close to the modeled result for each group. Special attention should be paid to the lowest and highest groups where the results are more likely to deviate as models are generally less able to predict observations at the extremes.

F.6 Actual Results v Modeled Results



In this case, the actual results are very close to the modeled results for the first seven groups. There appears to be a lot of difference between actual and modeled results for the last few groups, but the low volume in those groups suggests the results may be distorted by noise and therefore less valid.