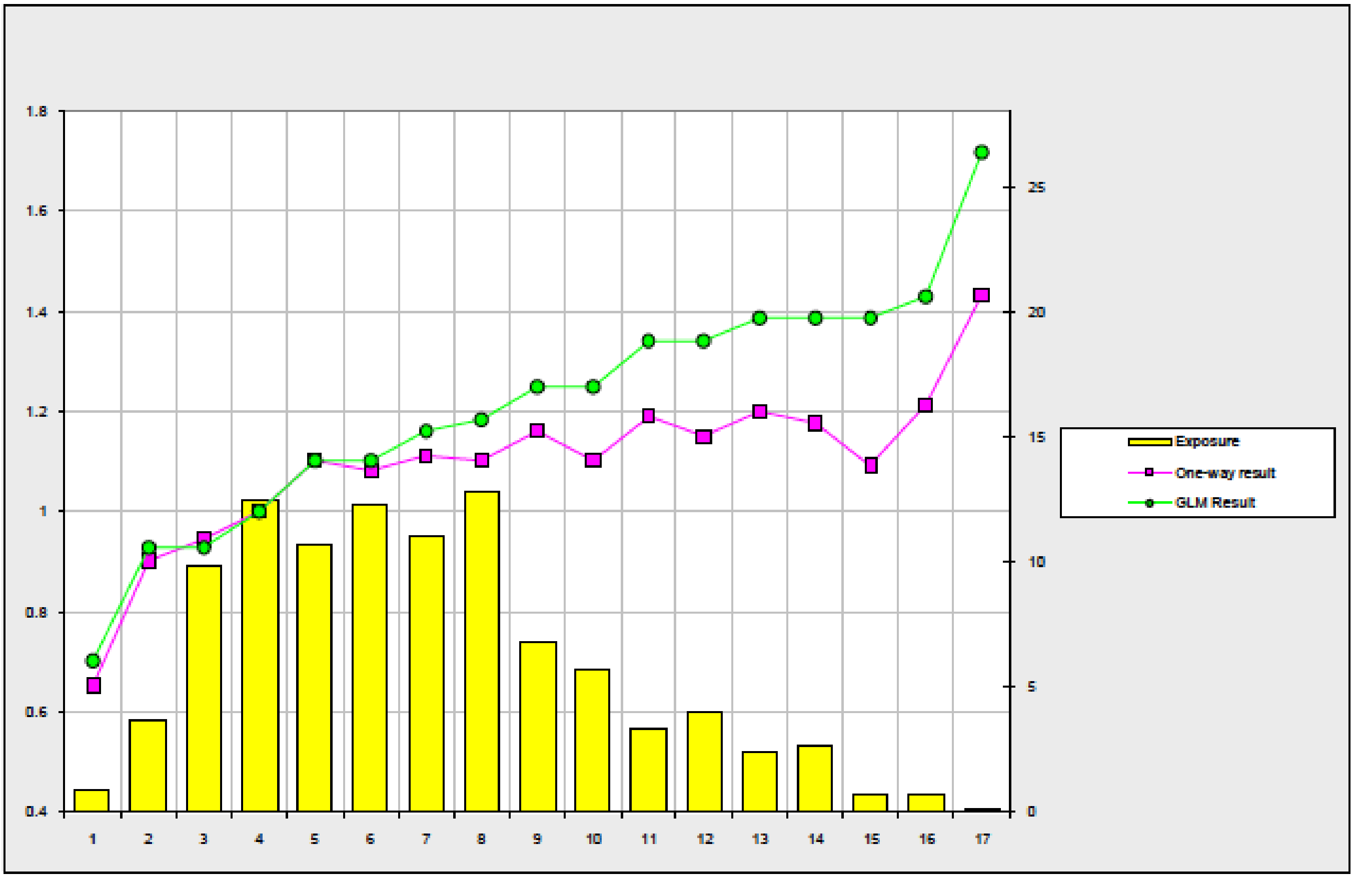
Examples of GLM Output:[1]

Figure 10.1 shows private passenger automobile collision frequency by symbol.[2] [3]
Rectangles represent the volume of exposures. The circles represent a fitted GLM, which includes many more variables than just symbol. So for example, symbol 10 is predicted to have a frequency about 25% higher than symbol 4, with all the other variables being considered.[4]

10.1 Effect of Vehicle Symbol on Automobile Collision Frequency



The squares represent the estimates of a univariate model that only includes symbol. We note that these relativities are significantly different than those from the GLM.

---

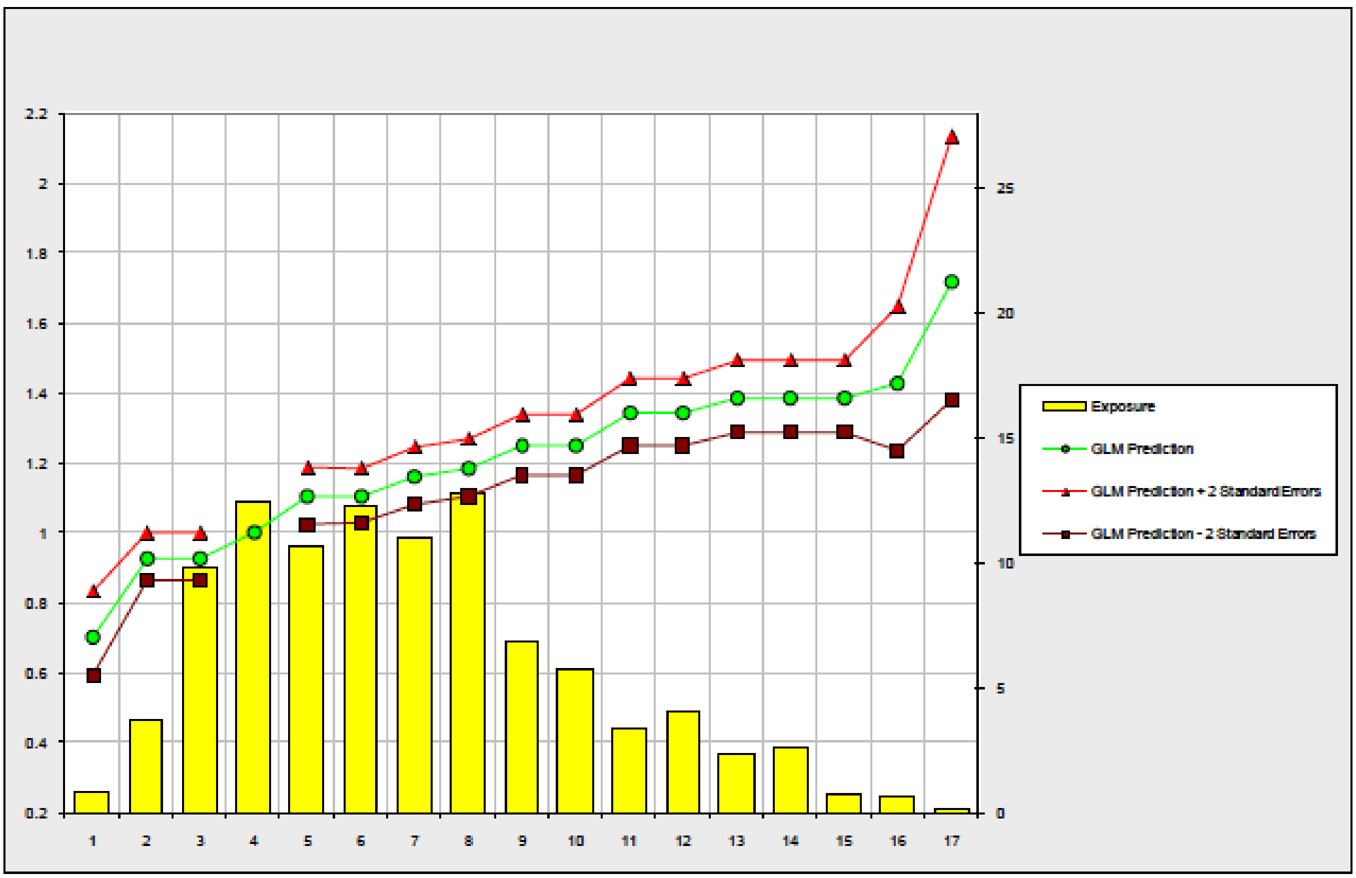
[1] Taken from Chapter 10 and Appendix F of Basic Ratemaking, on the syllabus of Exam 5.
[2] Automobiles have been assigned "symbols" for physical damage coverage for many decades.
All automobiles of a particular make and model have the same symbol.
Each symbol represents a group of vehicles that have been combined based on common characteristics (e.g., weight, number of cylinders, horsepower, cost).
See for example, http://www.iso.com/Products/VINMASTER/Physical-Damage-Rating-Symbols.html
The higher the symbol, the higher the expected pure premium, all else being equal.
[3] Everything is shown relative to symbol 4; symbol 4 has a (multiplicative) relativity of 1. Symbol 4, one of the symbols with a lot of exposures, has been chosen as the base symbol. Choosing as the base level one with lots of exposures makes the denominator of the relativity more stable.
[4] This estimate depends somewhat on which other variables are included in the model.

The difference between the univariate estimates and those from the GLM are probably due to the correlation of symbol with another variable in the model. As discussed previously, univariate analyses can be distorted by such effects.

The GLM results of one variable such as symbol are only meaningful if the results for all other variables are considered at the same time. The indicated relativity of 1.25 for symbol 10 discussed previously will not be valid if variables are removed or added to the model. In other words, the indicated relativities for vehicle symbol are dependent on the other relativities being considered. Also the relativities for one variable usually depends somewhat on the levels of the other variables. The 1.25 relativity for symbol 10 is presumably with all the other variables at their base levels.

Figure 10.2 graphs the fitted relativities and ± 2 standard errors for this GLM.[5] [6]
It is an example of one common diagnostic.

10.2 Standard Errors for Effect of Vehicle Symbol on Automobile Collision Frequency



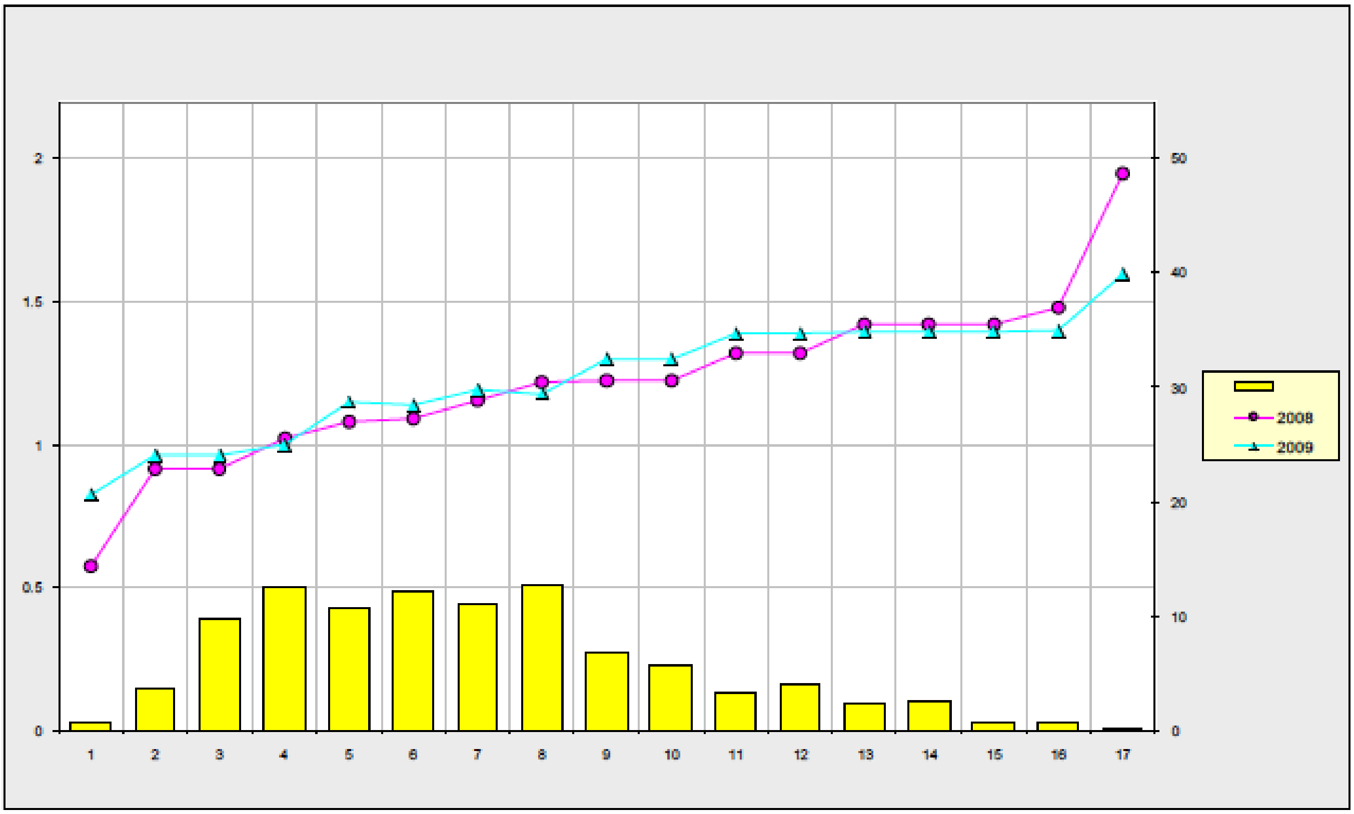
---

[5] This is presumably for all the other variables in the model at some base level.
[6] As discussed previously, we would expect the actual parameter to be within plus or minus two standard errors of the fitted parameter about 95% of the time.

The relativities go up with symbol. The error bars are relatively narrow, although they do get wider for the last few symbols, where there is not much data. Symbol seems to have a systematic effect on claim frequency.

Figure 10.3 shows the model fit to two separate years of data.[7]
We are interested in whether the model results are consistent based on the different years.

10.3 Consistency of Time for Vehicle Symbol



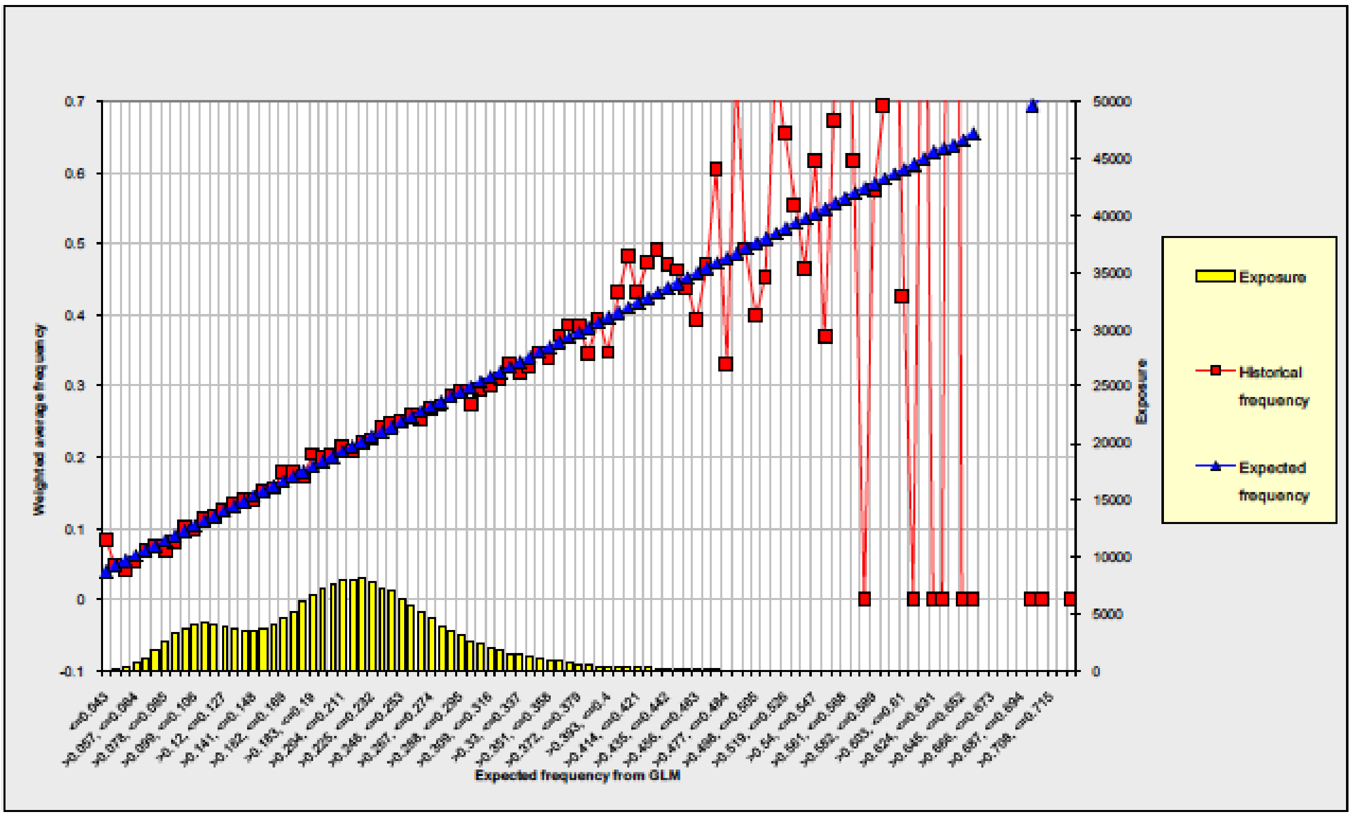
When one splits the original data into separate years like this, each model is based on less data than the original model, so we expect some more random fluctuation. In this case, the results are consistent between the two years, with the exception of symbol 17 where there is very little data. Again the model has been validated.

A key idea is that of a hold-out data set. We intentionally set aside a random portion of the original data, and do not use it to develop and calibrate the model. Then we see how well the model performs at predicting on this hold-out data set.[8]

[7] One could instead split the data into random subsets.
[8] Couret and Venter in their syllabus reading use a hold-out data to test their multidimensional credibility model.

Figure 10.4 compares the fitted model (triangles) to the observed frequency for the hold-out data set (squares). For each exposure in the hold-out data set, the model is used to calculate the expected frequency. Then the exposures in the hold out data set are grouped into intervals by expected frequency.[9] [10] For each of these groups we calculate the observed (historical) frequency.

10.4 Model Validation



In this example, the models performs well. The predicted matches the historical, until we get to the high predicted frequency groups, where there is very little volume, and thus lots of random fluctuation in the historical frequencies.

If in Figure 10.4, we had seen a bad match between the model and historical frequencies, then this might have indicated a model that was either underfit or overfit.
As discussed previously, the actuary wants to avoid both underfitting and overfitting models.

---

[9] Thus, we see the modeled frequencies (triangles) increase smoothly from left to right.
[10] For example, one group contains all exposures with expected frequencies > 14.1% and ≤ 14.8%.

For homeowners insurance it would be common to construct a GLM for each major peril for frequency and severity separately.[11] [12]

The first example models the frequency of claims for water damage on homeowners insurance. The GLM contains many variables, but here we concentrate on the effect of prior claim history.

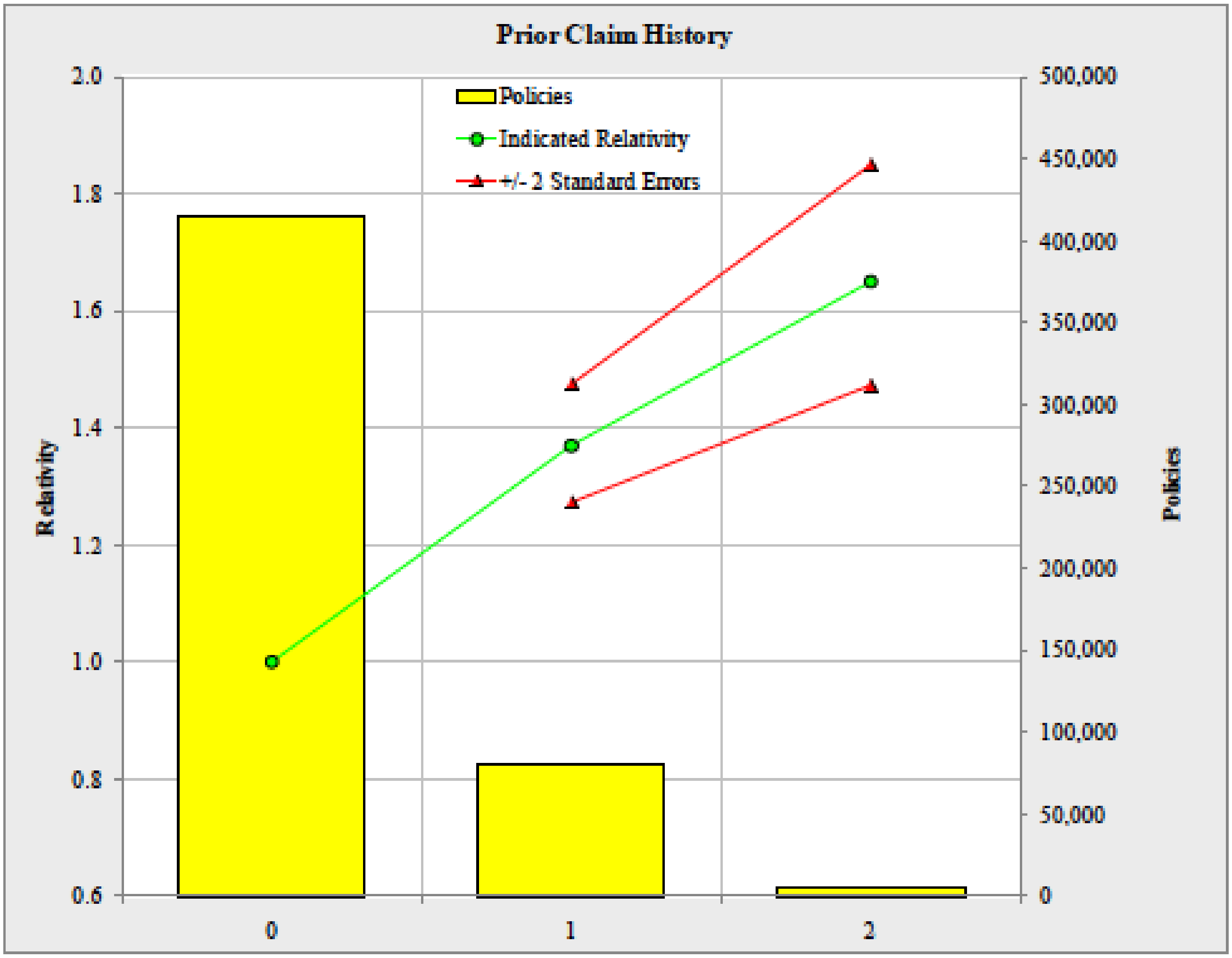
Policies are divided by the number of claims for some unspecified past experience period.[13] Each policy had either 0, 1, or 2 claims.[14] I assume that: each policy covers one home, and that renters and condominium policies are not included.[15]

Figure F.1 shows the fitted model and standard errors.[16] A standard error is the (estimate of) the standard deviation of the underlying errors for the model. If the errors were Normal, which they do not have to be for a GLM, then plus or minus 2 standard errors would cover about 95% probability.

---

[11] Severity has more random fluctuation than frequency, so it is usually harder to model.
[12] Perils would include: Fire, Theft, Wind, Vandalism, Water Damage, Liability, etc.
[13] I assume these are past claims for all perils. It is unclear what period of time is covered. Figure F.2 would lead one to believe that in Figure F.1 we are looking at four years of experience combined, 2011 to 2014.
The labeling of the years is just for illustrative purposes.
[14] While there may have been very few policies with more than 2 claims, they are not shown.
[15] None of these details are essential for interpreting the Figures and validating the model.
[16] This is presumably for all other variables in the model at some base level. Similar to Figure 10.2.

F.1 Main Effect Test for Prior Claim History



First, the fitted model makes sense. Those insureds with more claims in the past are predicted to have a higher expected frequency going forward. Compared to those with no claims, those with 1 prior claim are modeled to have a frequency relativity of about 1.37, in other words 37% more future expected claims from water damage than those with no past claims.

Besides the graph of the model relativities, also shown are ± 2 standard errors.[17] Those with 1 prior claim have a frequency relativity of between about 1.26 and 1.48. This is a relatively tight band, suggesting that it is OK to use prior claim history in the model. Not surprisingly, with a much smaller volume of data, the error bars for those with 2 claims are wider.

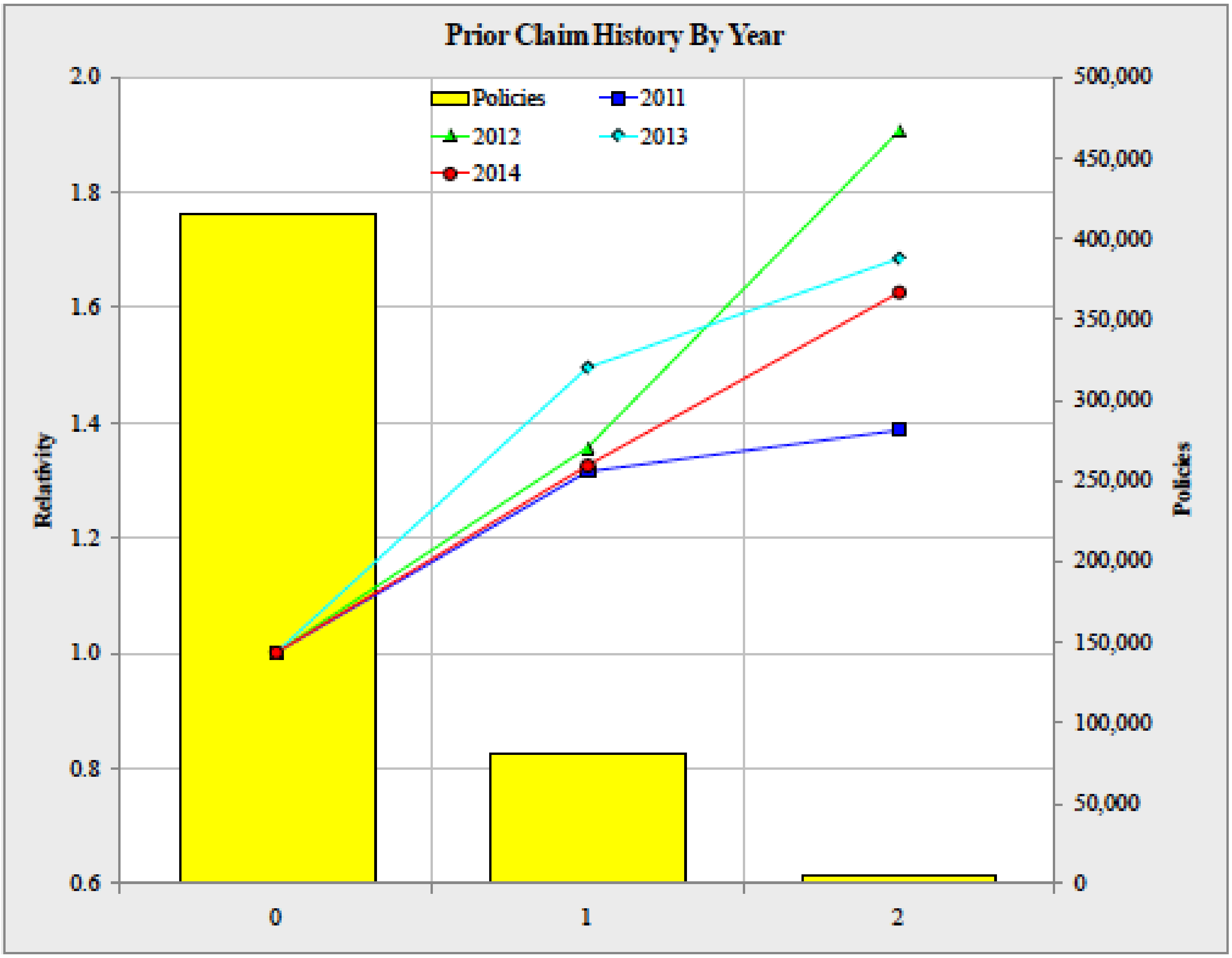
---

[17] As discussed previously, we would expect the actual parameter to be within plus or minus two standard errors of the fitted parameter about 95% of the time.

In general, we want a model that makes sense, and with relatively narrow error bars.

Figure F.2 breaks things down by policy year.[18]

F.2 Consistency Test for Prior Claim History



The models based on a single year each have a larger variance than the model based on all four years combined in Exhibit F.1. However, the lines each slope upwards with similar slopes; the pattern seems consistent over time.

In general, test the consistency of the model by comparing the results on separate subsets of the data base, such as separate years. In general, the actuary should use judgment to check the reasonableness of the results. In this case, it seems reasonable that more past claims would lead to a higher future expected frequency.[19]

---

[18] For the policies from each year, we use the same length of experience period as was used for the previous Exhibit F.1.

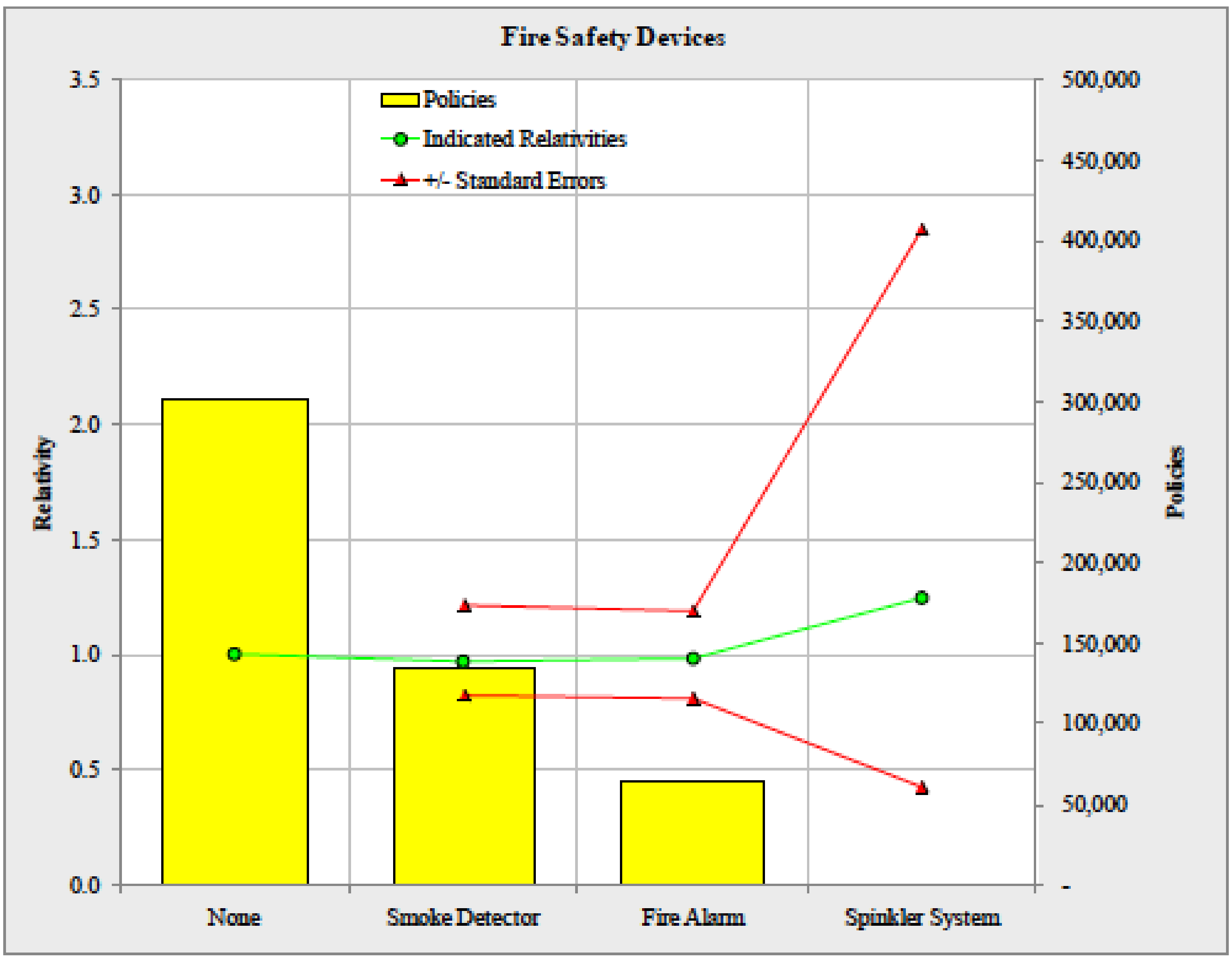
[19] This is the key reason why experience rating is worthwhile.

The second example, is similar to the first, except here we are attempting to predict the frequency of wind losses for homeowners. Again, even though there are many variables in the GLM, we are concentrating on just one, fire safety devices.

First, while we would expect fire safety devices to affect expected fire losses, most actuaries would not expect fire safety devices to significantly affect expected wind losses. In this case, the model does not seem reasonable based on judgement.

Figure F.3 is similar to Figure F.1 from the previous example.
Predicted wind frequency relativities are graphed versus four levels of fire safety device: None, Smoke Detector, Fire Alarm, and Sprinkler System.[20]

F.3 Main Effect Test for Fire Safety Device

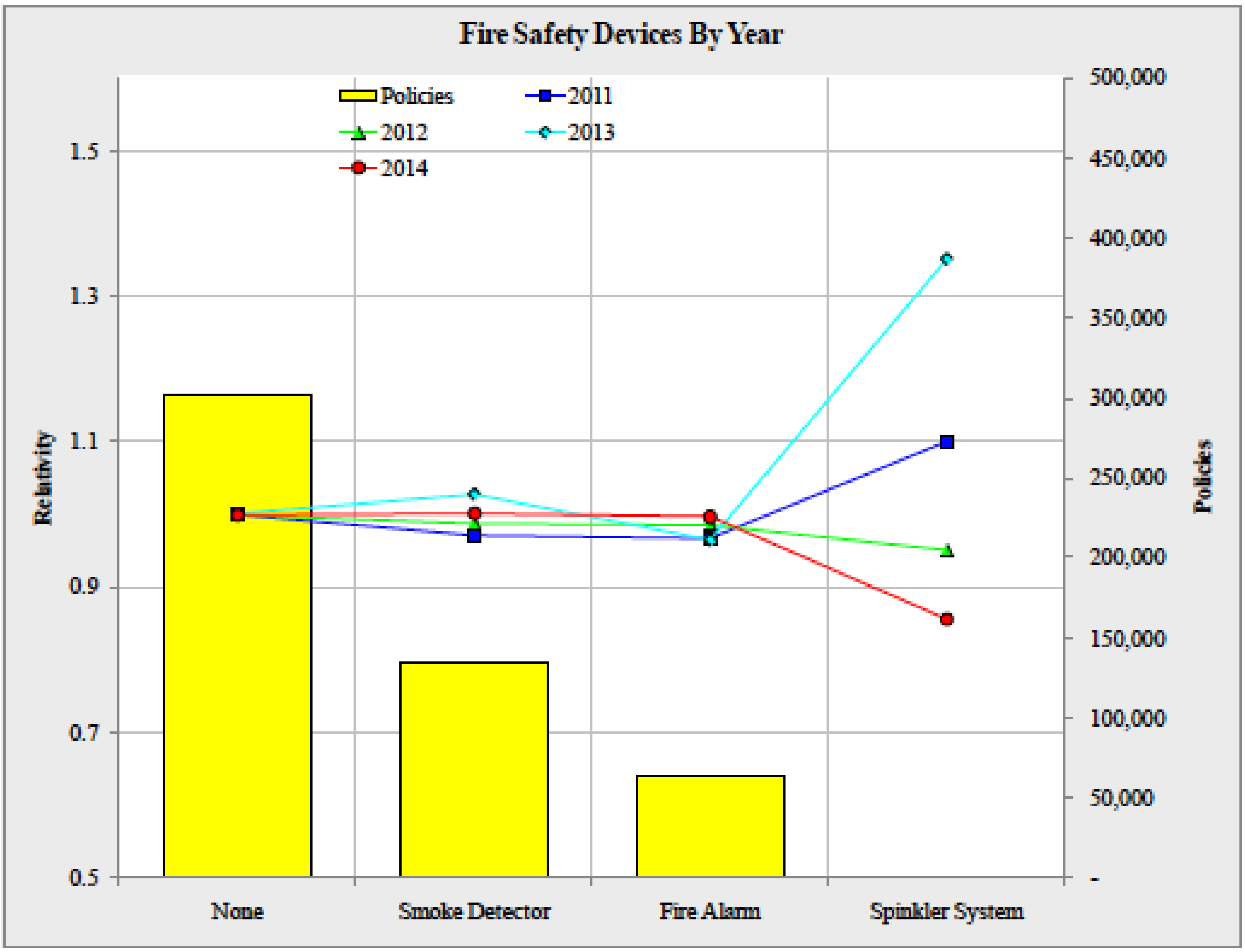


[20] Extremely few homes have a sprinkler system.

The indicated relativities are close to one, except for sprinklers. However, since there is so little data for sprinklers, its standard error is huge. We can conclude that the sprinkler relativity is very likely to be between about 0.45 and 2.9; in other words, this model tells us nothing useful about the relativity for sprinklers. The errors bars on the other relativities are consistent with a relativity of one. We conclude that fire safety devices have no predictive value in this model for frequency of wind losses.

Figure F.4 is similar to Figure F.2 from the previous example.
Again the results are shown for the model run on the data of each of four separate policy years.

F.4 Consistency Test for Fire Safety Device Claim



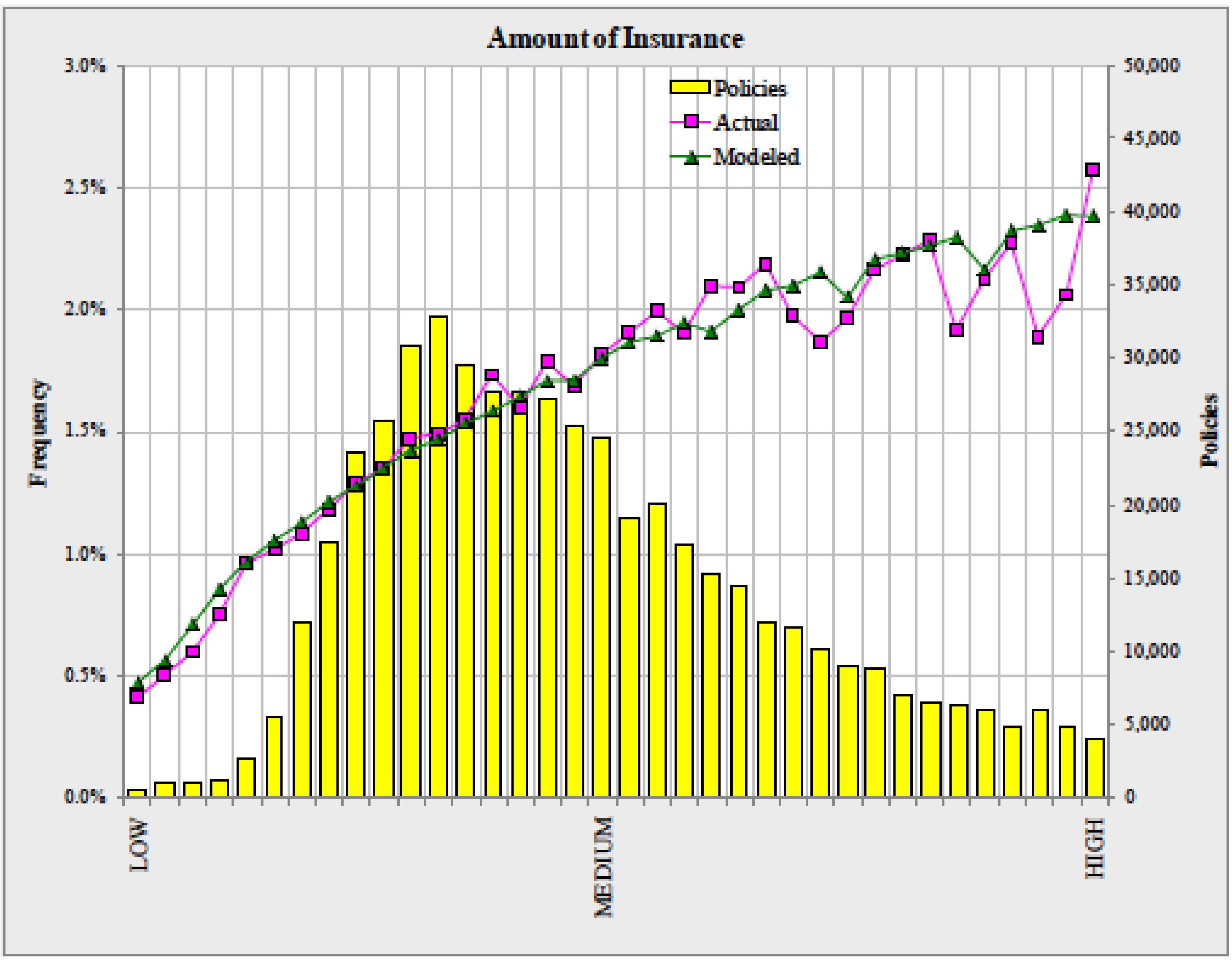
For sprinklers, there is <u>not</u> a consistent pattern across years; the prediction for sprinkles is very volatile due to the small amount of data. From the consistency test we again conclude that the indicated relatively for sprinklers is unreliable, and the model is consistent with a relativity of one for the other fire safety devices.

Next, rather than concentrating on one variable or one peril, we look at output to help us evaluate the performance of the overall model. We are still looking at a model for homeowners insurance.

A key idea is that of a hold-out data set. We intentionally set aside a random portion of the original data, and do not use it to develop and calibrate the model. Then we see how well the model performs at predicting on this hold-out data set. In general, the actuary should test the performance of a GLM on a hold-out data set.

Figure F.5 shows the results of the overall frequency model.

F.5 Actual Results v Modeled Results for AOI

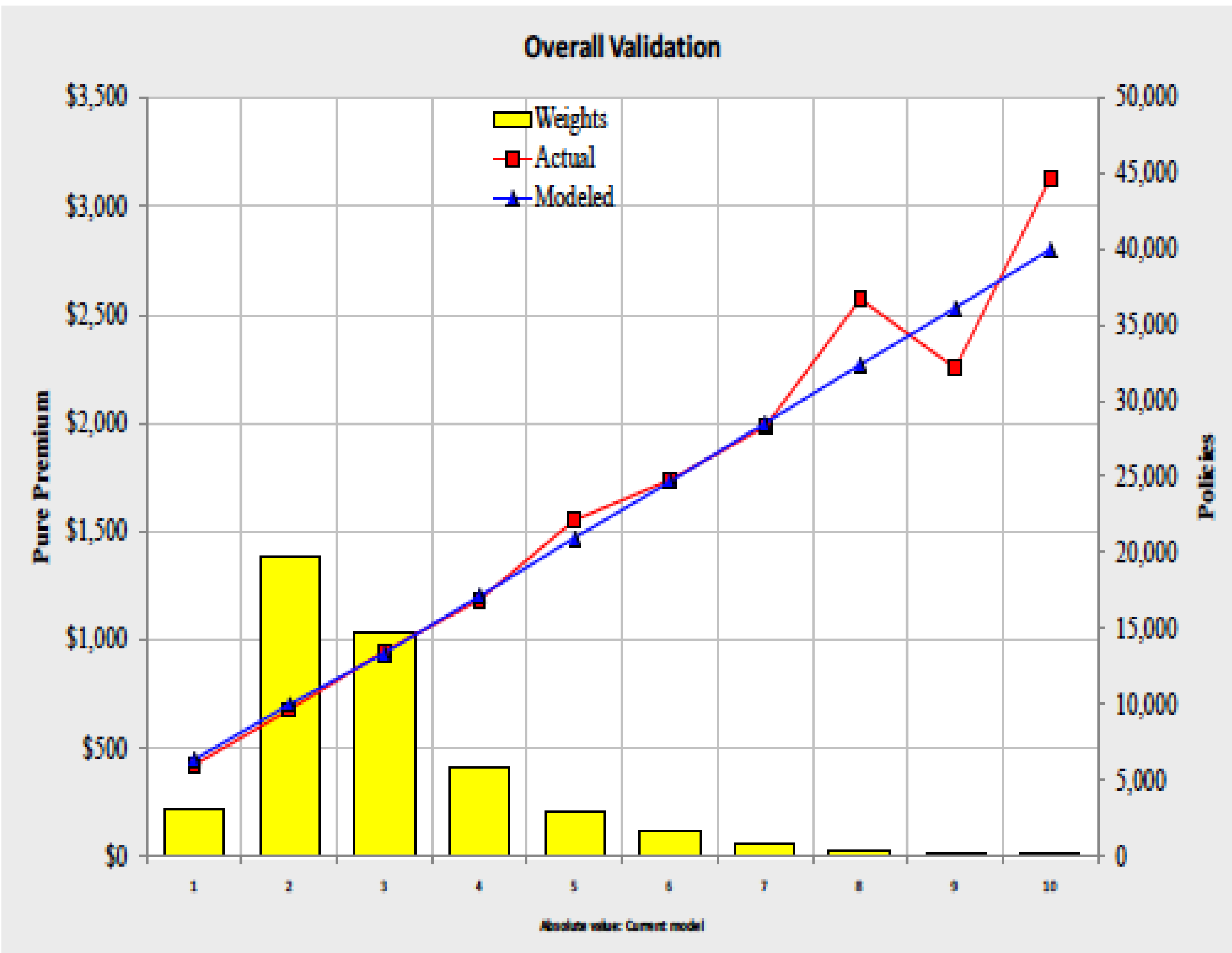


For the hold-out data set, as a function of amount of insurance, the modeled frequency is compared to the actual frequency.[21] We would like a close match between predicted and actual. However, we have a limited amount of policies, particularly for low and high amounts of insurance.

Overall, the match between the model and actual is good. However, the model appears to be underpredicting frequency somewhat for medium sized amounts of insurance and overpredicting frequency somewhat for high amounts of insurance. For extremely low amounts of insurance there is little data and a lot of volatility; however, the graph suggests that the model may be overpredicting for extremely low amounts of insurance.[22]

Figure F.6 displays another way to validate the overall model. This time we compare modeled and actual pure premiums for the hold-out set. We order the hold-out data by modeled pure premium and group it into 10 groups.

F.6 Actual Results v Modeled Results



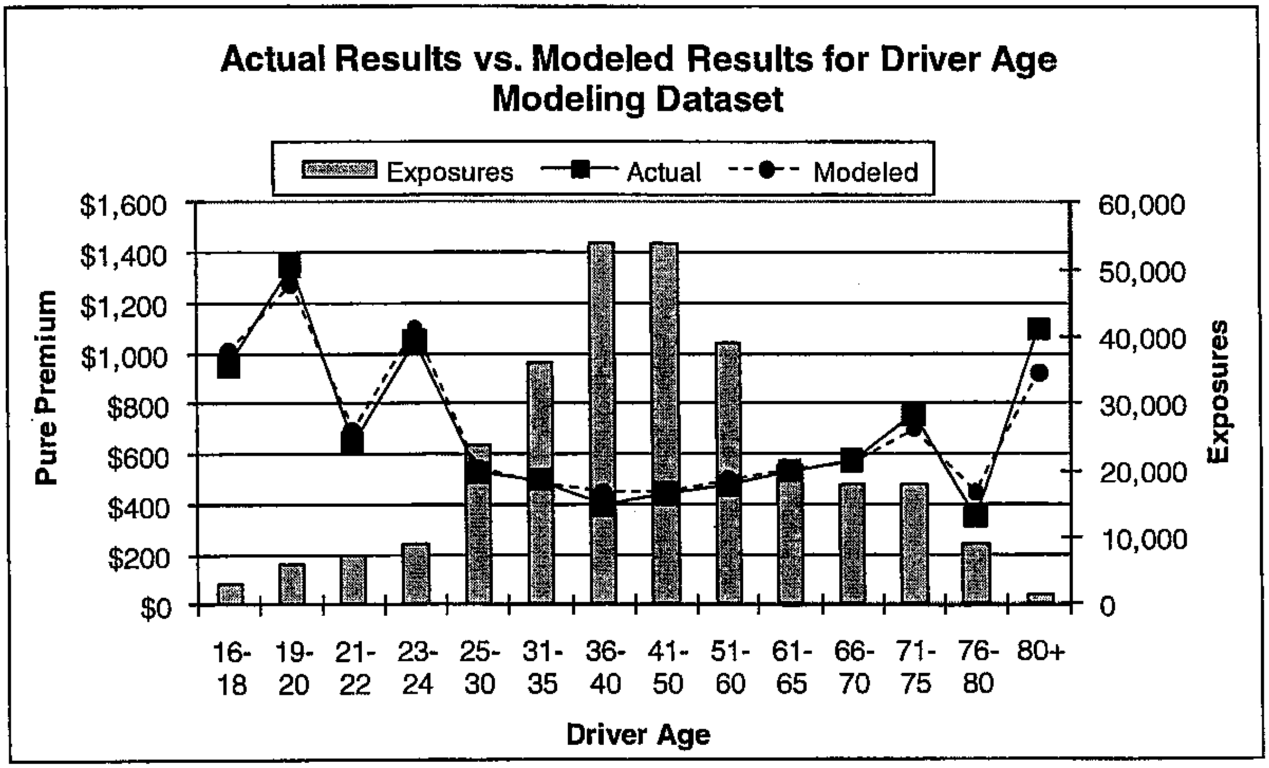
[21] Remember, the GLM was developed and calibrated without this hold-out data.
[22] "Essentially, all models are wrong, but some are useful," George Box.

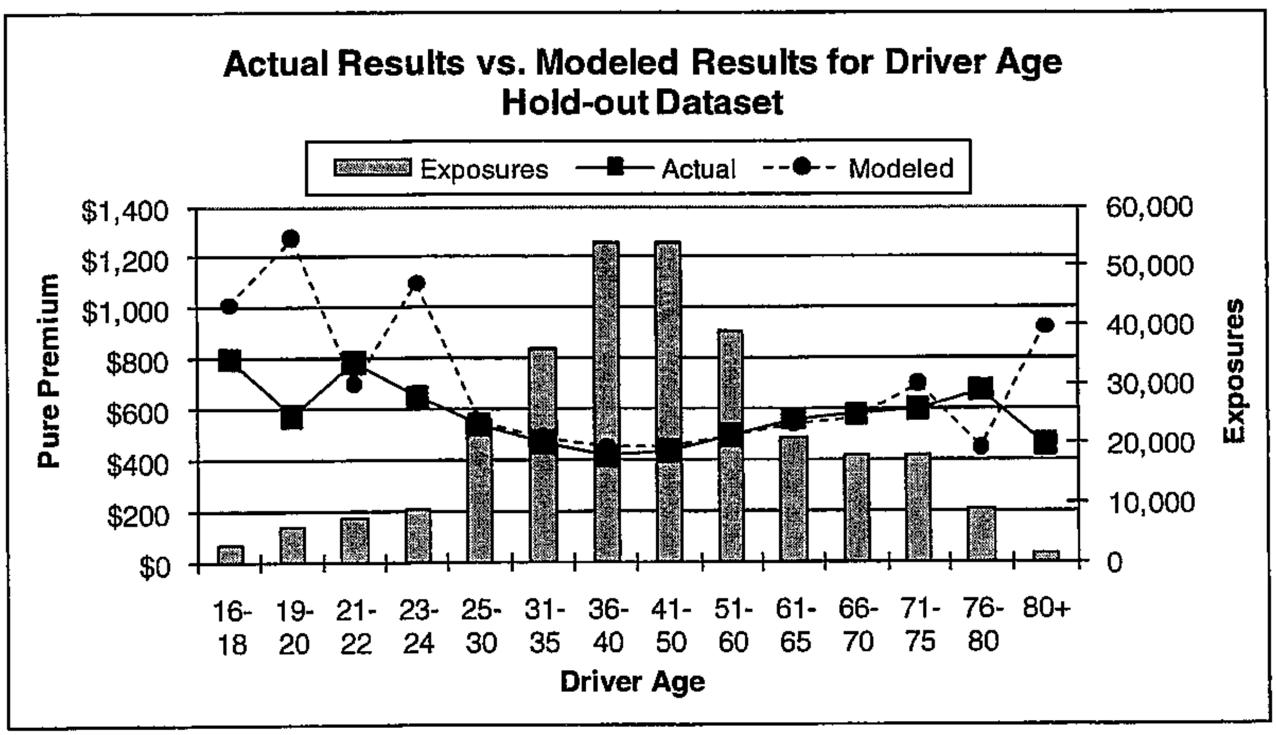
There appears to be a reasonable match between actual and predicted. While they differ for the highest pure premium groups, there is too little data in those groups to draw a definitive conclusion.

In general, in the case of a graph like Figure F.6, the actuary should pay particular attention to the extremes on both ends, since they are usually harder for the model to predict.

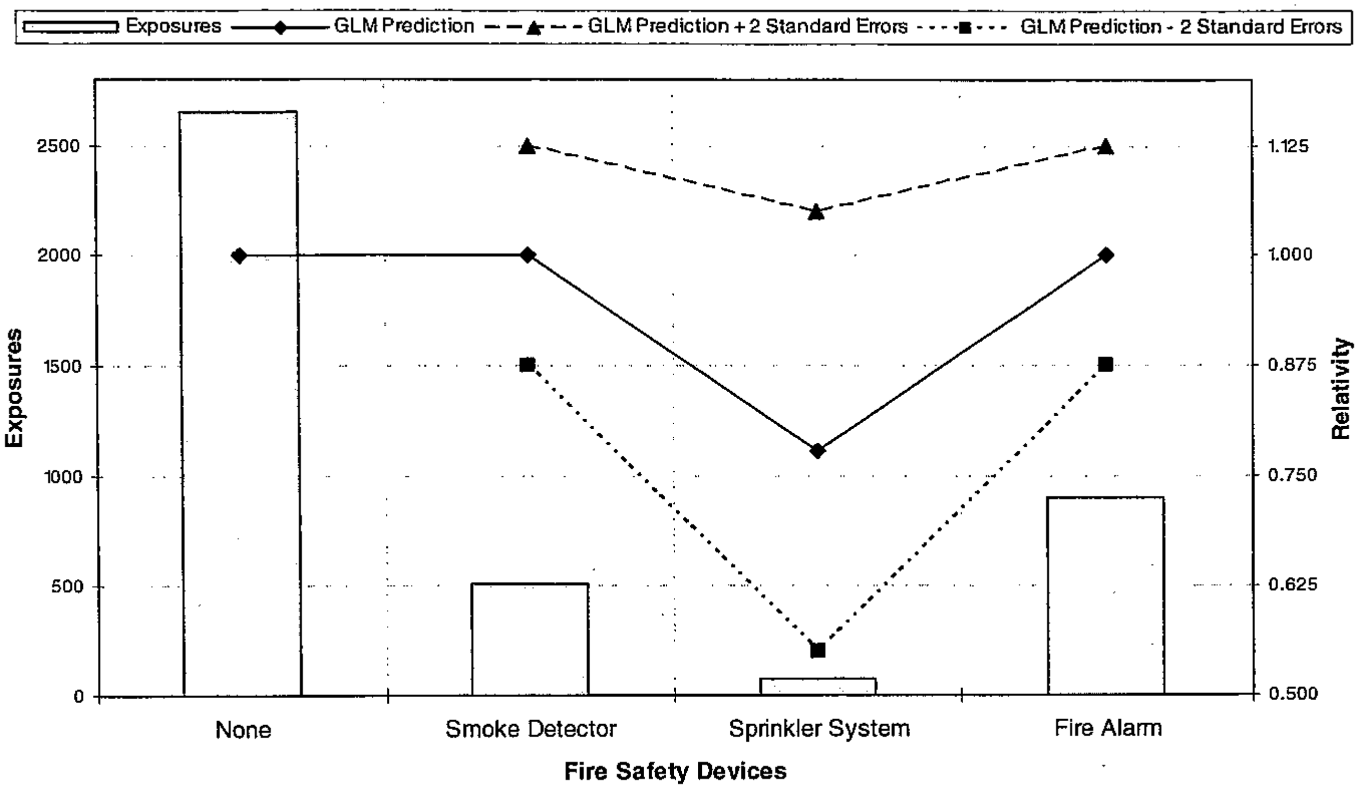
Problems:

5, 5/10, Q.36. (1 point)
Company XYZ applied generalized linear modeling to its personal auto data. Graphs of the actual and modeled pure premiums by the driver groupings were produced by the analysis. The first graph is a plot of the values using the modeling dataset. The second graph is a plot of the values using a hold-out dataset. The modeling dataset and the hold-out dataset have the same number of exposures. Explain whether or not the model appears to be appropriate.



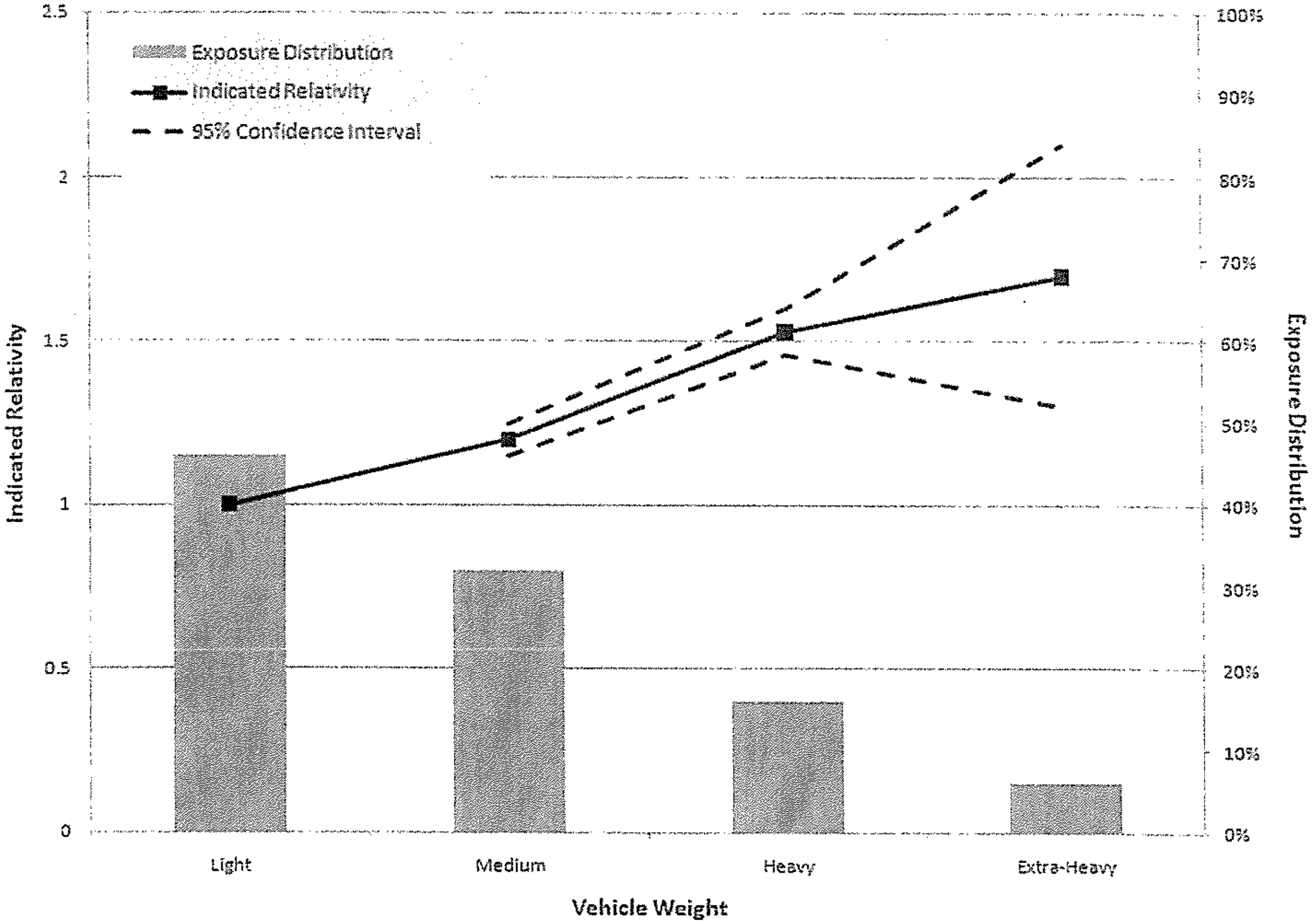


5, 5/11, Q.13. (1 point)
A company applied generalized linear modeling to its homeowners data. A graph of indicated relativities and their standard errors for a fire safety device rating variable is shown below. Evaluate the effectiveness of the variable in the model.



Note that the legend reads: Exposures, GLM Prediction,
GLM Prediction + 2 Standard Deviations, GLM Prediction - 2 Standard Deviations.

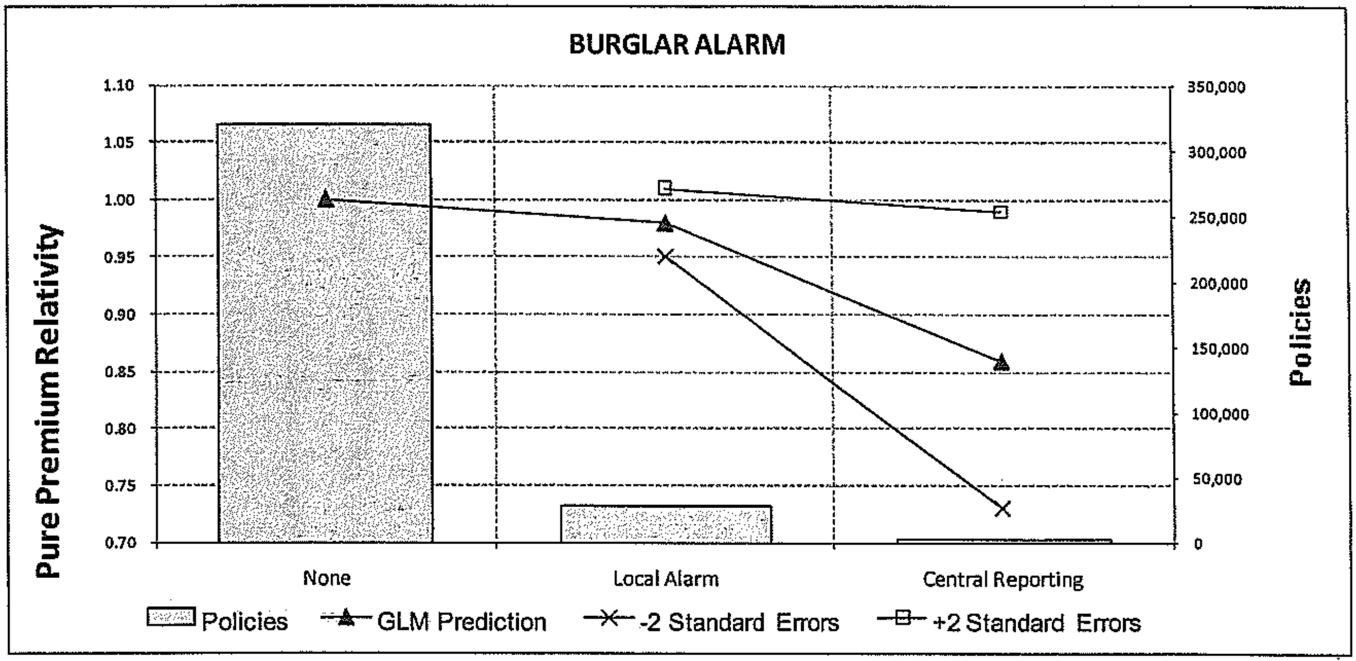
5, 5/12, Q.11. (1.5 points) An insurer uses several rating variables, including vehicle weight, to determine premium charges for commercial automobiles. Your manager has requested a review of the vehicle weight rating relativities. The following diagnostic chart displays the results for vehicle weight from a generalized linear model. (Light, Medium, Heavy, and Extra Heavy.)

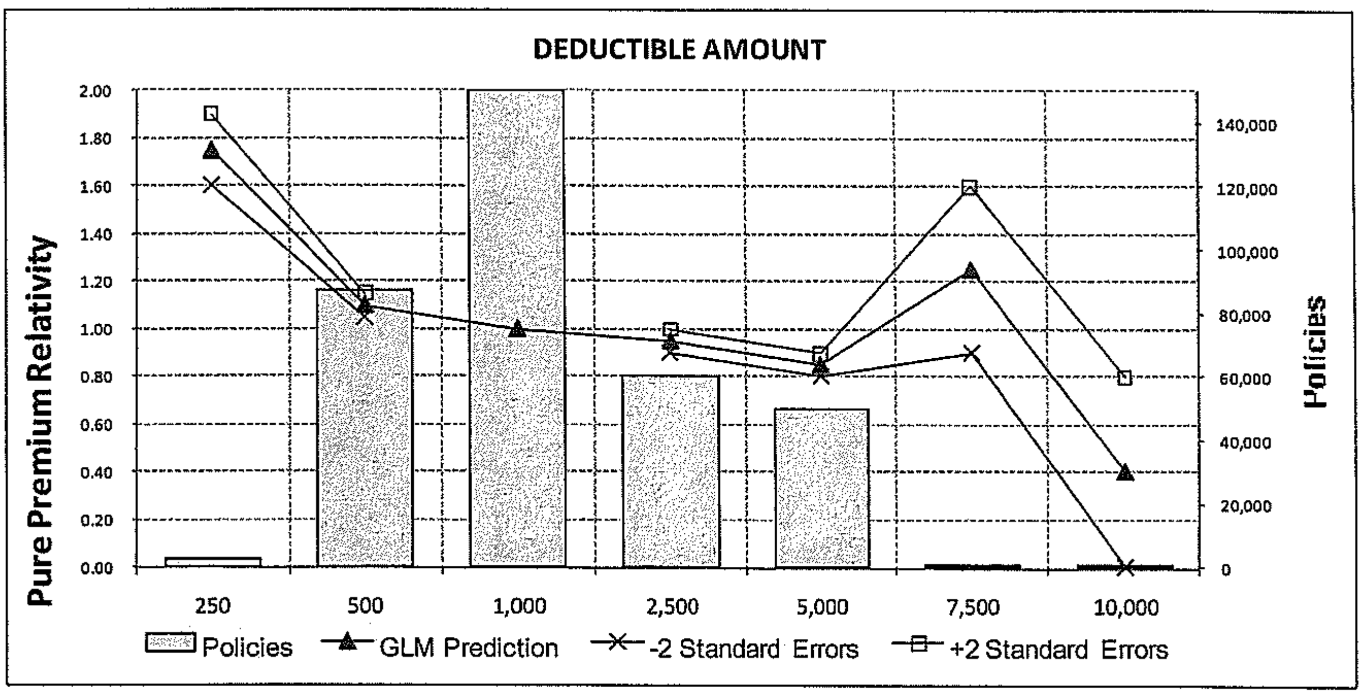


Company management plans to expand its commercial auto marketshare with an emphasis on writing more businesses that operate with extra-heavy weight vehicles. Management wants to charge the same rates for both heavy and extra-heavy weight vehicles.

Based on the model results, provide your recommendation to management and explain the considerations supporting your position. Include a discussion of any potential risks associated with it.

5, 5/13, Q.12. (3 points) An insurer is planning to revise burglar alarm and deductible rating plan factors for its Homeowners program.
Given the following generalized linear model output:





Question is continued on the next page.

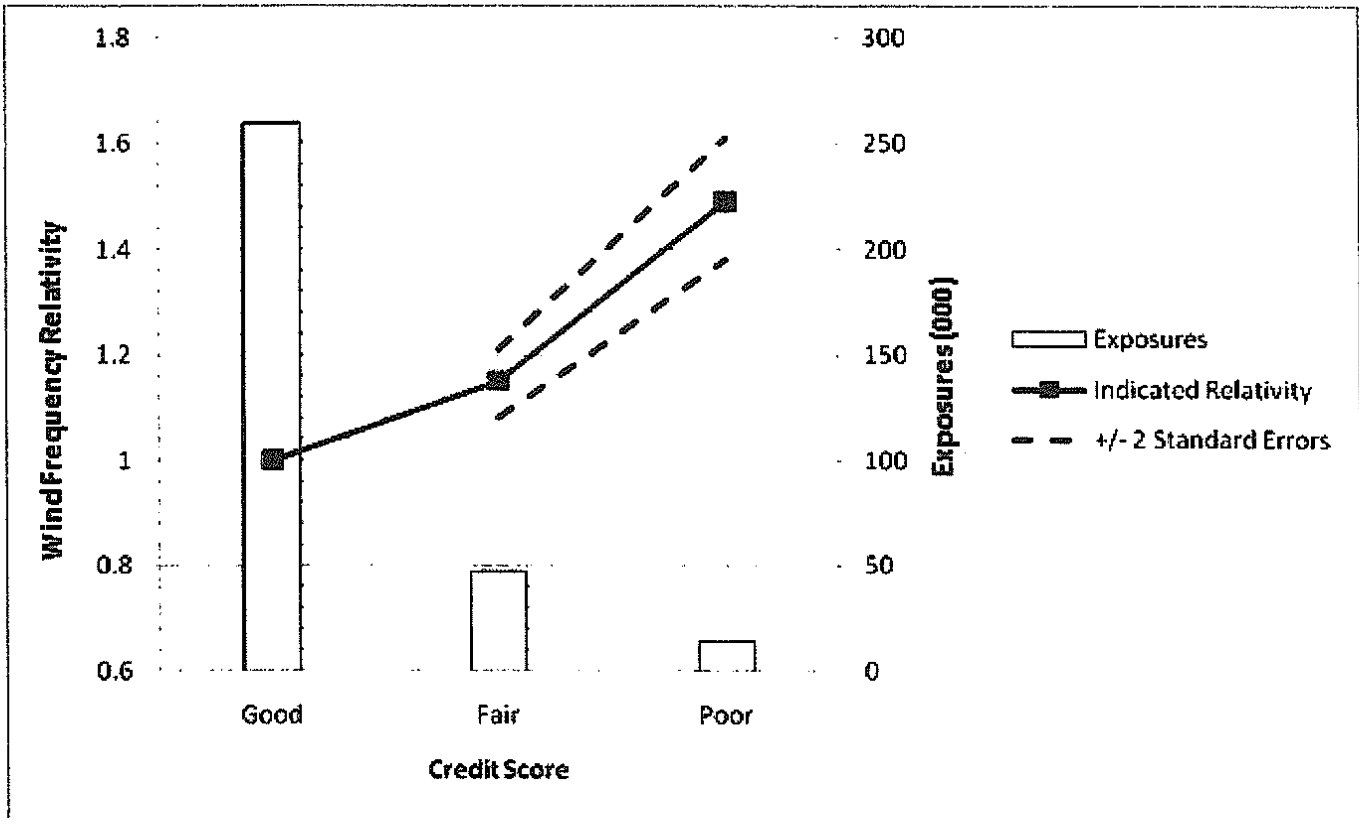
(question continued)

Burglar Alarm	GLM Prediction	-2 Standard Errors	+2 Standard Errors	Policies
None	1.00			320,000
Local Alarm	0.98	0.950	1.010	27,500
Central Reporting	0.86	0.730	0.990	2,500

Deductible	GLM Prediction	-2 Standard Errors	+2 Standard Errors	Policies
$250	1.75	1.60	1.90	2,700
$500	1.10	1.05	1.15	87,000
$1,000	1.00			150,000
$2,500	0.95	0.90	1.00	60,000
$5,000	0.85	0.80	0.90	50,100
$7,500	1.25	0.90	1.60	150
$10,000	0.40	0.00	0.80	50

Propose revised burglar alarm and deductible rating plan factors.
Document the relevant analysis and rationale to support the proposal.

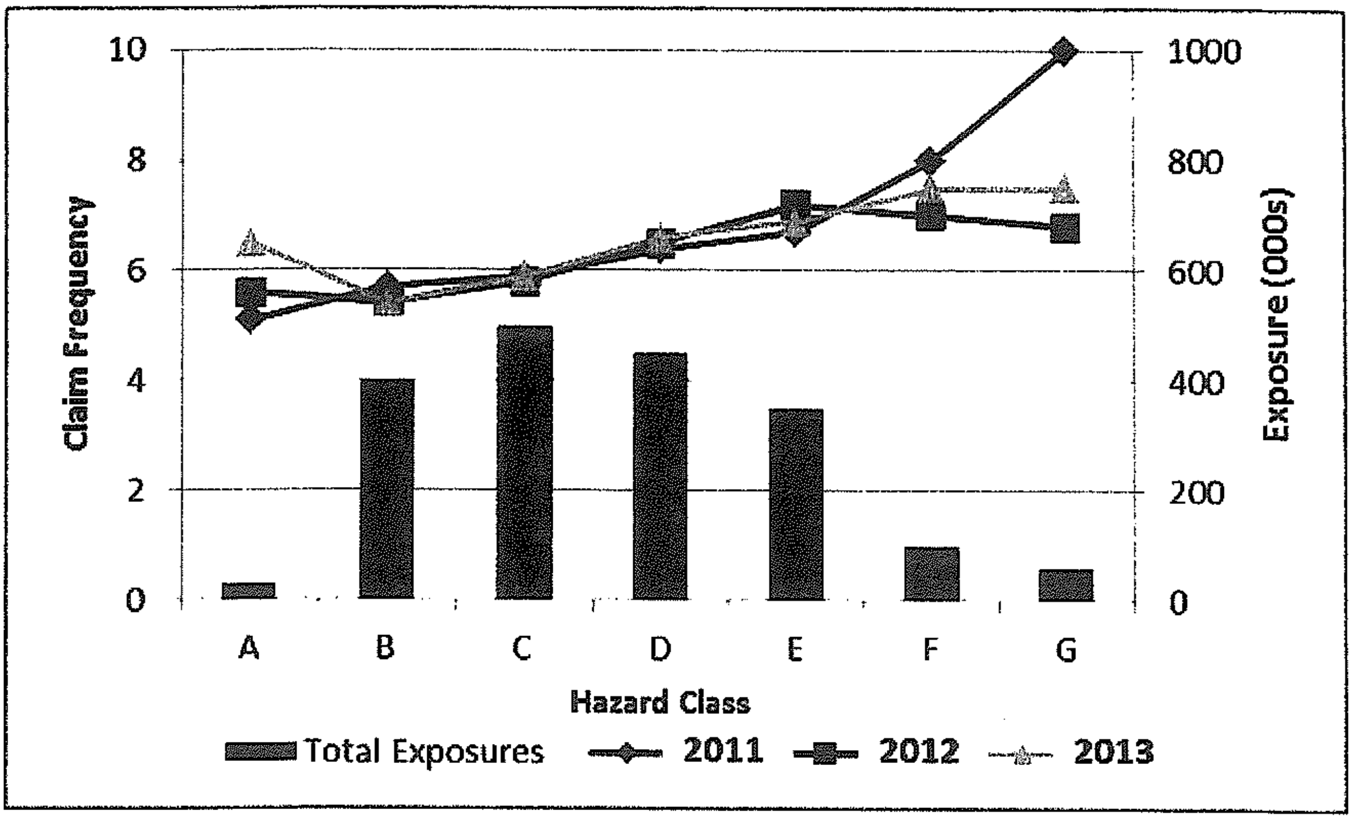
5, 5/14, Q.9. (2 points)
An insurer is considering using credit score to further segment its homeowners book of business. The insurer has developed a generalized linear model to evaluate different variables' contribution to expected frequency of wind claims.
The following diagnostic chart displays the results of a countrywide analysis performed on one year of data from a generalized linear model:



(The solid line is the indicated relativity, while the dashed lines are ± 2 standard errors.)

Using the generalized linear model output, as well as other considerations, justify whether the insurer should add credit score to the homeowners rating plan for the wind peril.

5, 11/14, Q.10a. (0.75 points)
An actuary performed an analysis of a products liability class plan using a Generalized Linear Model (GLM) for the first time on this book of business. The insureds are categorized by hazard classes A through G. The following graph shows claim frequency and exposure data by hazard class.



Fully evaluate the predictive value of hazard class based on the information provided above.

Solutions to Problems:

5, 5/10, Q.36.
The model appears to be appropriate for the ages nearer the center, from about 25 to 70.
The model does not appear to be appropriate for either the younger ages below 25 or the older ages above 70.
The model appears to be over fitted. In the age ranges without many exposures, the model is picking up the random fluctuations in the data, in other words the noise. In these age ranges, the model matches the data to which it was fit but fails to match the holdout dataset to which it was not fit.
If the model had been picking up useful information about a pattern in these age ranges, in other words had been picking up a signal, then it should have matched to some extent the holdout data set as well.
Comment: See Exhibit F.5 in Appendix F of Basic Ratemaking, on Exam 5.

5, 5/11, Q.13. There are relatively wide error bars around the estimates of the model. For example, for the smoke detector class, the relativity could be between about 0.88 and 1.13, or a 12% credit to a 13% debit. The error bars are even wider for the Sprinkler System class, which is not surprising given the smaller number of exposures. A different relativity for Sprinkler System than the others although indicated is not justified by the model. Intuition would lead one to expect all of these devices to lower expected fire losses somewhat. However, these results of the model are not sufficient to justify giving any credit, let alone quantifying the size of such a credit. Based on the data analyzed, this is not an effective variable in the model. One would probably need significantly more exposures than was used here in order to properly analyze this whole situation.
Comment: A much shorter answer should suffice for full credit.
See Exhibit F.3 in Appendix F of Basic Ratemaking on Exam 5, which models frequency of wind losses. We are not told any details about the graph in this exam question. Is it modeling frequency or pure premiums? Is the graph in this exam question modeling a particular peril or all losses?
A good example of where a fancy model can not make up for having too small a volume of data. One would need to observe more homes and/or more years.

5, 5/12, Q.11. The indicated rate for extra-heavy vehicles is 10 to 15 percent higher than that for heavy vehicles. However, the relativity for extra-heavy vehicles is based on few exposures and thus the 95% confidence interval is wide. The 95% confidence interval stretches from about a rate 10% lower to about a rate 40% higher than heavy vehicles. Thus the proposal by management is well within the range indicated by the generalized linear model based on the available data.
On the other hand, based on the pattern for light, medium, and heavy vehicles, the insurance cost increases significantly with weight. Thus it is logical that extra-heavy vehicles would cost more than heavy vehicles.
Thus I would recommend that we charge the indicated relativity.
If we expand the number of extra-heavy vehicles written, we will get more data to better estimate an appropriate relativity in the future.
There is a risk that if we write a lot of new extra-heavy vehicles, they will be on average poorer risks or at least have different risk characteristics.
If management's proposal were adopted, and if the rates for heavy vehicles are lower than costs, we may attract a significant volume of new business, but lose money. If the combined rate for heavy and extra heavy vehicles is higher than costs for heavy vehicles, then we risk adverse selection.
We would be able to write lots of underpriced extra-heavy vehicles and lose a lot of our current heavy vehicles which will be overpriced.
It would be useful to see what competitors are charging for extra-heavy vehicles versus heavy vehicles. If the data used in the GLM is from one state, it would be useful to get information from other states. It would also be useful to investigate the interaction of vehicle weight with the other rating variables in more detail.
Comment: There is no one right answer. Given the limited information available in an exam question, one has to make some assumptions and do the best one can to answer in a sensible manner.

5, 5/13, Q.12. Burglar Alarm:
Based solely on the GLM, there is little evidence to support a discount; there is relatively little data for the non-base classes, particularly for central reporting.
There are wide confidence intervals for both Local Alarm and Central Reporting groups. The Local Alarm standard errors suggest it is not significantly different than the None category; the confidence interval encompass a relativity of one. Central reporting has very few exposures and large standard errors.
I would recommend this variable not be used; in other words, 1.00 factor for all groups.
Alternately, based solely on the GLM, there is little evidence to support a discount.
On the other hand, it is logical that a local burglar alarm will reduce theft losses.
It is logical that a central reporting burglar alarm would be more effective at reducing theft losses than a local alarm.
However, theft losses are only one of many perils covered by Homeowners.
We are given no information on what portion of the expected losses are due to theft; this varies by geographical location.
Based on the logic and the limited statistical support from the GLM, small discounts make sense.
I judgmentally select 0.98 for local alarm and 0.96 for central reporting.
Deductible:
Based solely on the GLM, due to the small amount of data, there is little evidence to support a discount for the $7500 and $10,000 deductibles. Also the discount for a $7500 should be smaller than for a $10,000 deductible; the GLM fitted relativities indicate the opposite.
There is somewhat more data for the $250 deductible, but the error bars are relatively wide.
On the other hand, we know that expected losses paid are more for a lower deductible and are lower for a higher deductible.
For the $7500 and $10,000 deductible, based on the difference between the indicated relativities for 2500 and 5000, I will judgmentally select relativities of 0.78 and 0.74.
(For evenly spaced deductibles, the difference in Loss Elimination Ratios gets smaller as the deductible increases, in the absence of either favorable or adverse selection.)
For the $250 deductible, based on the difference between the $500 and the $1000 relativities, a relatively of something like 1.20 might make sense. The 1.75 prediction from the GLM might be due to adverse selection. So I will judgmentally select a relativity of 1.30.
For the other deductibles, I will use the GLM output.
Thus my selected relativities are: 1.30, 1.10, 1.00, 0.95, 0.85, 0.78, 0.74.
Comment: There are many possible reasonable answers. Additional information besides the GLM output would be very helpful, for example competitor's rates.

5, 5/14, Q.9. Based solely on the given output of the GLM, it makes sense to add credit score to the homeowners rating plan for the wind peril. The ± 2 standard deviation bands around the indicated relativities for fair and poor each do not contain one; the indicated frequency relativities are statistically significantly different than 1. However, I would also want to see an analysis of pure premiums.
Countrywide there are only about 45,000 exposures in the fair category and 15,000 exposures for the poor category. This raises concerns about the credibility of the data from those classes. (The vast majority of the exposures are in the good category. Perhaps some other breakdown of scores into categories would be better.)
I question whether there is casual relationship between credit scores and claim frequency from wind.
This is a countrywide study. Could it be that the average credit scores may vary by state, with those states with higher average wind losses also having lower average credit scores?
(On the other hand, perhaps those with poorer credit scores are less likely to properly maintain the roof of their house, leading to some wind claims that would not have been otherwise made. It is important during hurricanes that the roof remain intact and attached to the home.)
Overall, I would recommend that the variable n<u>ot</u> be added (at this time) based on the lack of causality and the lack of reliable relativities due to the small volume of data for only one year. More data as well as more analysis is needed.
<u>Comment</u>: One should not spend much time commenting on the general issue of using credit scores in rating insurance. I think it should be sufficient to discuss the issue of causality, whether or not there is a logical connection between credit scores and wind losses.
On the general issue of using credit scores in rating insurance:
1. Assuming the insurer writes a reasonable amount of business and credit scores are grouped into intervals that are not tiny, there should be enough data in each rating group to measure costs with sufficient accuracy. The criterion of credibility is fulfilled.
2. Insureds with similar premiums after the use of credit scores have a range of expected costs, just as with any other rating variable. However, the use of rating scores decreases this variation and thus improves the homogeneity.
3. Studies have shown that credit scores are correlated with insurance costs.
Credit scores have been used for several years and the relationship to costs has been reasonably stable over time. Thus the criterion of statistical significance is fulfilled.
4. There are errors in credit reports. Individuals can get copies of their credit reports and try to get the credit bureau to correct any errors. However, the information in the credit report are not subject to manipulation or lying by the insured. The criterion of verifiability is fulfilled.
5. There is considerable expense in obtaining credit reports and turning them into credit scores to use for rating insurance. Either the insurer will incur that cost or pay someone else to do this work. In either case the criterion of low administrative costs is not fulfilled.
6. One can construct credit scores for use in rating insurance using objective definitions, with little ambiguity. Class definitions based on ranges of credit scores can be mutually exclusive and exhaustive. There should not be much administrative error, as the credit scores can be calculated by computer. The criterion of objectivity can be fulfilled.
7. Since when they apply for a home mortgage or a car loan, their credit reports are examined, it is not an issue when these same reports are used for insurance. The criterion of privacy is fulfilled.
8. Both high and low income insureds have good and bad credit reports. The effect of using credit scores should not be correlated with income. The criterion of affordability is fulfilled.

9. The items recorded in a credit report, such as a late payment of a bill, are not responsible for differences in insurance costs. The criterion of causality is not fulfilled.
10. An insured can modify his behavior in order to improve his credit report in the future. The criterion of controllability is fulfilled.

5, 11/14, Q.10a. (a) The indicated frequencies differ significantly by hazard group. (We are not given information in order to determine whether these differences are statistically significant.) Indicated relativities generally increase with Hazard Group, with the exceptions of Hazard Groups A and G which have much less data than the others. The separate indications for the three years are consistent, with the exceptions of Hazard Groups A and G which have much less data than the others. Therefore, I conclude that hazard group is useful for predicting expected frequency.