**Here are some** **questions to consider answering in your report and presentation:**

* 1. What methods did you consider (you don’t have to actually try all of these methods; just ones that you think would work for this problem)?
  2. What method did you choose in the end, and why?

We initially considered the Tweedie GLM Regression to model claimcst0 using a single model. We consider the log-link with a var.power of 1.3 (a range of (1,2) corresponds to the compound Poisson model, otherwise known as a Tweedie distribution).

However, after further research, we decided an alternative approach to modeling claimcst0. We thought of the “Loss” as a product of the Count of claims and the Severity of the claim. So we introduced a two part model, one that would predict the Count of claims, and another that would predict the Severity of the claim.

The first model is a Poisson GLM with a log link on all available data for the training data, and the second model is an Inverse Gaussian GLM with a log link on the subset of the training data where there was at least one claim. Our justification for the use of the models is:

* Count Model: The Count of the number of claims (numclaims) should have a Poisson distribution (since it is count data). We can use the entire training data for this model since we want to consider all cases, regardless of the occurrence of the claim. For modeling Count, we used log(exposure) as an offset because log(numclaims/exposure) would theoretically give us the estimated count in a fixed period of time (in this case one year), but we are only interested in the numclaims, hence in the model is treated as a response while offset(log(exposure)) is a predictor. We used exposure as an log(offset) instead of as a weight, because when predicting for new values of x, we would not be able to use exposure as a predictor if it is treated as a weight in the model.
* Severity Model: For modeling severity, we only want to consider the data for which there was a claim, so the appropriate description of the model would be the mean function of the severity of a claim given that there was a claim. We considered different distributions for this model, but we found that the inverse Gaussian model was the best. Our runner-up was the Gamma regression, but we noticed that the predictions for the Gamma predictions were too dispersed, and gave us some extreme predictions (the right-side tail of the prediction distributions were heavy-tailed). After testing the Inverse Gaussian method (suggested in the book Generalized Linear Models for Insurance Data), we saw that the predictions of claimcst0 for the method had more conservative predictions, so we decided to proceed with Inverse Gaussian GLM.
  1. How did you test the assumptions of this method?

Autocorrelation: two individuals may have had been in the same accident, in which case those two would be correlated, but from the information provided we would not know.

Distribution: response should have the distribution specified in the GLM model.

Variance as a function of mean

Link function = log

* 1. How did you evaluate your model (e.g. fit statistics, over-fitting, etc.)?

We evaluated our model fit using cross validation, computing the gini coefficient on the predictions of the training data from the k-fold cross validation models, and after discussion we agreed that the gini coefficient of this would be similar to the gini coefficient of the new predictions for the V or H data. We also used a bootstrapping method to compute an “unbiased” set of coefficients after deciding on a model after model selection.

* 1. How did you do you variable selection?

We performed variable selection first by using the step function (both direction) starting with the full model (considering all variables except veh\_body). After, we evaluated the p-values and use the Anova function from the car package to consider removal of further variables (to prevent overfitting), and evaluating the model on the cross-validation Gini coefficient. We went through a lot of trial and error.

The reason we did not consider veh\_body as a predictor in either models is because the grouping in veh\_body was too sparse. Some categories contained as little as 9 observations, and we decided that model coefficients computed would not be very accurate, perhaps overfitting on those small sample, which would not be appropriate for this analysis.

* 1. Any concerns about the resulting model?
  2. What questions to you have about the data?

We had some concern about the data, for example the 0 values in the predictor veh\_value. We were not sure whether these 0’s truly meant that the value of a vehicle was 0, or if the 0 meant something else (like NULL or MISSING). We ended up using the 0’s in the model, but if we truly understood what those meant we may have changed how we handled those observations.

* 1. What variables help explain pure premium (explain to a non-statistician; please include this in your presentation for your business partner)?

I would say vehicle value is one of the most crucial variable in estimating the pure premium, because insuring something of a higher value means that the loss from accidents would be larger, so conversely the premium should be higher. Vehicle age and vehicle body (model) would also be important variables, because intuitively we can say that older vehicles tend to have more frequent problems with vehicle performance that may lead to losses, and probably vehicle performance also varies by body type or model. In terms of the drivers, the age of the driver and possibly gender may be important as well. We can with some confidence say that younger drivers and senior drivers tend to have a higher accident rate than those that fall closer in the middle of the age range. We might also see a difference in accident rate by gender. We could also argue that accident rates differ by area as different neighborhoods could have different levels of safeness of driving, but this might be correlated with the type of drivers that live in the area, in which case gender and age could explain the differences in accident rates by neighborhoods.

* 1. What other variables not in the data set do you think might be useful?

If we had age as a numeric value rather than categories might have a better predictive power (gives us more information).

Another variable not included in this data set that could have a high potential for estimating pure premium is the driver history or record (driving record or any sort of convictions) which could help estimate how much of a liability the driver could have.

Credit score

Income?

https://www.casact.org/pubs/forum/07wforum/07w263.pdf