MATH 3338: Mathematical Modeling (Fall 2019)

Instructor: Dr. Hoa Nguyen
Project Advisor: Courtney Rohde, ACAS

Project Title
Using GLMs to Estimate Auto Insurance Losses by Policy Characteristics

Phase III Information

Task 1: Analyze Dislocation *(1-2 Hours)*

At this point, I want you to start working together to produce one set of final proposed factors and a new base rate. You all have a tentative proposal from the end of Phase II, so it's time to get together, share what different decisions you made in each of your models, and compare how those decisions affected your indicated and proposed rate factors. I don't think you necessarily need to create another model with simplifications that the whole team agrees on, but if that would be helpful you can definitely do that. I'm hoping that perhaps your modeling decisions were similar enough that you can agree on proposed factors that are reasonable in light of each model's indications. If you have trouble with this, let me know. I'd be happy to stop by Trinity again to work with you on it.

In this phase, you will analyze the impact of the changes you are proposing to the rating structure. We are taking a 2.1% increase overall, but this does not mean every policy will see a 2.1% premium increase. Since we are also changing rating factors, some policies will see much larger increases, and some will see decreases. These changes accumulate to 2.1% in total. The percent premium change of an individual policy (calculated as ProposedPrem/CurrentPrem-1) is called **dislocation**. The dislocation analysis can be done within the 'Rating' workbook that was developed in the last phase.

In practice, insurance companies rarely implement the model indications as their rating factors. The indicated factors represent the model's best estimate of the true relativity appropriate for each rating characteristic, so it is definitely desirable to use factors as close to indicated as possible. However, there are always practical business limitations to consider. If the proposed factors are very different from the current factors, certain policies may see dramatic premium increases or decreases. Generally, such premium changes cause the insureds to call and complain, or shop around with competitors for a cheaper policy. This is called **friction**, and we want to limit its effect.

Sometimes, companies will want to keep premiums lower than indicated for certain key segments of their book of business as part of their business strategy. Implicitly, this means they must be charging other segments more than indicated, since they must meet their overall rate level goals. Allowing some segments to "pick up the tab" for other segments' riskiness is called **subsidization**. It can be intentional or unintentional. For our hypothetical company, we will allow some intentional subsidization as part of our strategy. Our key segments are:

- Gender = Male
- Rating Area = 1B
- Rating Area = 2B
- Rating Area = 3B
- Driving Restriction = Named

We want to stay competitive on these segments, which I am defining to mean that the rate increases for these segments must be less than 7% (for the whole segment, not necessarily for every single policy in the segment). To calculate the rate changes by segment, you will need to use the pivot table function in Excel. You can create pivot tables on the "Segments" tab by going to 'Insert > Pivot Table.' For your table/range, select all columns and rows of the data on the "Business Data" tab. Then you should be ready to play around. Pivot tables allow you to slice and dice the data very efficiently. You can also insert column or line charts based on the pivot table, and these should update automatically when your pivot table does. Google pivot tables for more information.

After you have keyed in your proposed factors on the "Rels" tab (use two decimal places) and off-balanced the rates, use a pivot table to check if the overall rate changes for the key segments are less than 7%. If they are not, choose new relativities and run the off-balance procedure again. (You may have to refresh your pivot tables to see the recalculated premiums.) This may be an iterative process, if you have to try several different scenarios before you are happy with your factor selections.

To recap, the goals for your new rates are:

- Overall premium increase of 2.1%
- Proposed factors have moved closer to indicated
- Proposed factors rounded to 2 decimal places
- Proposed factors make sense for the levels (i.e. intuitive, no reversals)
- Rate change on key segments <7%

Once your proposed scenario has accomplished these goals, the last thing to do is look at the distribution of individual policy dislocations. You can create a new column on your business data tab for dislocation (calculated as above), then make another pivot table to see how many policies are getting greater than a 50% premium change (**Tips**: you can use 10% intervals for dislocation, try right-clicking on the pivot table row labels and then "Group…". Also, you can change the metric of the pivot table values – I think it defaults to 'Sum,' but 'Count' may be more useful here. Look for the pivot table 'Field Settings'). Make note of any policies are getting increases greater than 50%, and find out what is causing that. Sometimes outliers are just because of policy characteristics, and other times they are indicative of a mistake in your review. It can be ok to allow them, but you must know how many there are and why it is happening.

Also, I want you to produce a line graph of each variable that shows the Current, Indicated, and Proposed factors all together. These charts will help you make sure your selections are moving in the right direction (toward indicated). Plus, they will be useful when you present your proposed scenario to your classmates (as pricing analysts must do for business leaders).

After all this analysis, the relativities and proposed base rate are ready to be filed with the state department of insurance. Congrats! You've essentially completed a (simple) rate review! Except you don't have to do any of the paperwork, you lucky duck.

Task 2: Evaluate the GLM *(3-4 Hours)*

We will explore three different ways to evaluate the performance of a model. We will use the holdout data set ("ppdata_test.csv") to perform the following tests. Since the model coefficients were estimated without considering the holdout data, we can assume that the model will perform about as well on future observations as it does on the holdout data. This is why the holdout data is valuable for evaluating model performance.

First you will need to apply your final simplifications to the testing data set. Create columns in the "ppdata_test.csv" file that have the same groupings and column names as your final simplified version of the model from Phase II.

Read your holdout data with the appropriate groupings into R. After fitting your model (on the training data, as before), use the following code to calculate predicted values on the holdout data set:

```
ppdata_test$Predicted <- predict(model1,newdata=ppdata_test,type='response')
```

You should only need to update the name of your fitted model, in red. This command creates a new column called "Predicted" in R's working copy of your holdout data. This consists of the model's prediction for each row in the test data set. This way, the predicted pure premiums are calculated using the factors that we were calling the "indicated" factors from Phase II. (The Phase III work doesn't use your proposed factors.)

We will do our model evaluation in Excel, so you will need to export your model predicted values. You can use a command like:

write.csv(ppdata_test,file=" Predictions.csv")

This will write your data (with its new predictions column) to your working directory in R Studio. I think you should be able to download it from the Files section.

I am positive that it's possible to perform all these evaluations within R (more accurately and quickly too!), but unfortunately my attempts at writing code for them were unsuccessful. Perhaps that will be an improvement for future years' projects.

I think it would be interesting and valuable to perform the following tests on these three models and see how they compare: the linear model from Phase I, the unsimplified GLM from Phase I, and your final simplified model from Phase II.

1. Mean Squared Error

This one is easy – it was in the code in Phase I. A lower squared error tells you that the model predictions are closer to the actual observations in absolute value.

It was the line that said "mean(resid(model1)^2)"

2. The Lorenz Curve and Gini Coefficient

[I recommend prioritizing this method of evaluation over #3, simple quantile plots, if short on time, since I think this one may give you better/more useful results.]

The Lorenz curve demonstrates how well your model separates the "good" risks from the "bad" risks. Again, when you have your predictions in Excel, follow these steps:

1) Sort the observations by prediction, smallest to largest.
2) Add more columns calculating:
    a. Cumulative Pure Premium

    b. Cumulative Exposure
    c. Cumulative Pure Premium as % of Total Pure Premium
    d. Cumulative Exposure as % of Total Exposure
3) Using Cumulative Exposure % as your x-values, plot Cumulative Pure Premium %. This should create a concave-up curve – the Lorenz curve. When you have inserted a line graph in Excel, there is a "Select Data" button under Chart Tools > Design. From here, you can specify what you want your x values to be, and what "series" you want to plot (the blue line - the cumulative pure premium %).
4) Using the same x-values, plot Cumulative Exposure %. This is basically drawing on the diagonal line x=y.

I've uploaded an example excel file that demonstrates these calculations & graphs.

The further away the Lorenz curve gets from the diagonal, the better job your model is doing with differentiating good risks from bad risks. For example, since my example Lorenz curve roughly goes through the point (75%,60%), we can infer that the worst 25% of exposure as identified by my example model (on the right side of the x axis) did in fact experience 40% of the actual losses (on the upper part of the y axis). In contrast, the "mean model" line, going through (75%,75%) demonstrates that if our "prediction" for everyone was simply the overall mean, we would expect 25% of the exposure to experience 25% of actual losses (in other words, no differentiation).

The Gini Coefficient is two times the area between the diagonal and the Lorenz curve. In my example excel file, I've calculated the area as the sum of very thin rectangles -one for each observation. The height of the rectangle would be the difference between the two curves for that observation: (Cumulative Exposure % - Cumulative Actual Pure Premium %), and the width of the rectangle would be the % of Total Exposure for that observation. This is not the most accurate way to calculate a Gini Coefficient, but it will work for our purposes.

You can use F2 to follow my calculations within my sample Excel document.

3. Simple Quantile Plots

Quantile plots are used to evaluate how closely the model prediction tracks with the observed values.

Once you have your test data (with predictions) in Excel, follow these steps:

1) Sort the observations by prediction, smallest to largest.

2) Group the observations into ten quantiles, based on quantity of exposure. That is, calculate what is 1/10th of total exposure, and the assign the observations in the 1/10th of exposure with the lowest model prediction to bucket "1", the next tenth to bucket "2", and so on. You can just make another column called "Quantile" for this.
3) For each of the ten quantiles, calculate the average model predicted value and the average observed pure premium.
4) Plot the averages for each quantile on a line graph.
5) We would expect to see both observed and predicted increasing for successive quantiles, and we would hope to see the predicted value tracking well with observed.

Again, there is an example of these calculations in the Excel file I've uploaded.

Note: I had some trouble with this plot when I was making my example. My Observed and Predicted lines should have been basically on top of each other, and they were not. I am unsure if this was a problem with the simple model I used for the example, or a problem with the underlying data (or if I made a mistake! It happens.) Don't sweat it too much if you see this same issue. It's okay for a modeler to say that some of her results are unexplained, and she needs to do more research. ☺️ I may look into this for next year.

I hope you can use these model evaluation methods to make a more confident and informative final presentation. The goal is for you to have some quantitative metrics to point to when you talk about how well your model predicted the pure premiums, and how well your model performed compared to an unsimplified model or a linear model.

Thanks for working on this project with me this semester! You've done great!