Case 3
MT7028 - Non-life Insurance Pricing

Majken Gunnarsson - majken.gunnarsson@gmail.com
Oscar Potts - oscar.n.potts@gmail.com

December 11, 2024

1 Part 1

Based on prior analysis we have ended up on the model specified in table 1. The baseline category is (3,1,3,3,2) which corresponds to a 40 plus year old driving a rural 5-9 year old smaller vehicle. The baseline premium is 20.84 SEK and all of these factors have been found to be significant.

For the frequency model, we observe an overdispersion estimate of 1.64 and for severity, we use a lognormal model.

2 Part 2

When we use the duration as an offset, we fix a $\beta_W = 1$ toward the log of the duration. Instead of assuming that $\beta_W = 1$ we can relax the condition and estimate it as a covariate. In doing so, we obtain the observed estimate of $\beta_W = 0.34$ with a confidence interval 95% of $[0.251, 0.429]$.

This makes the assumption not supported by our data. It is interesting that this is not the case once we aggregate the data. In addition, we observe estimated disperions of 1.019, which is more in line with Poisson-distributed data.

For future work, instead of using the observed W, use the expected value of it $E[W|X]$ or, as in this case, see it as a covariate.

3 Part 3

To create a simple model, we use backward selection on the 50% split training data using 10-fold cross-validation to calculate the generalization error as the cross-validated deviance.

We will conduct backward selection per seperately for the frequency and severity model to remove one category at the time until we cannot reduce the generalization error. In current model both severity and frequency use the same formula variables, by conducting this backward selection we may trim the models to improve generalization.

The resulting factors in the training data for the simple and advances model can be seen in table 2. For the frequency model, we ended up only removing vehicle age category 4 without further reduction in the generalization error by removing more variables. For the severity model, we removed driver age categories and vehicle age categories 1 and 2. Generally, the estimates are very similar, with some exceptions for driver age, gender categories, and vehicle age. Interestingly, the factors seem more drastic in the simple model, even though it should generalize better.

Studying the fit of the model, we see in Table 3 the resulting unit deviation that the simpler model performs better on the test data set, but worse on the training data. Specifically, we see an improvement in the severity model compared to the frequency. This may not be a surprise; we have not modeled severity granularly enough, and since it contains quite a few datapoins, we should expect it to overfit more. It thereby seems that the simple models remove some of the overfitting we have done previously to a model which extrapolates better.

4 Part 4

To evelaute the toal fit of our model, there is no specific distribution or distance measure we can use. We therefore try to make the unit deviation defined as $D(y,\mu) = \sum_{i=1}^{n} 2(yln(\frac{y}{\mu} - y + \mu)$, or we can use the Gaussian unit deviation defined as $D(y,\mu) = \sum_{i=1}^{n}(\mu - y)^2$. In doing so, we obtain the results shown in table 4. As can be seen, our initial model is better compared to the simple model. This is a bit surprising, but may not be a surprise. The initial model was already quite optimized with relatively few variables selected.

The simple model had to remove already quite significant factors which may distort more than what it optimizes. We should mention that the difference are marginal, meaning that a simpler model may be preferred as it explains similar amount of variance on smaller amount of variable.

For future work, one should expand the modeling towards continuous variables. For example, splines mixed with some of these categorical values likely improve the model and also make pricing seem more continuous for customers. In pricing, there are always more things that could be done and this is no different. In task 4 we hope to continue with further approaches.

Factor	Class	Class Description	Relative Factor
Zone	1	Urban areas Swedens 3 largest cities	4.67
	2	Suburbs plus middle-sized cities	3.5.74
	3	The rest of sweden	1
MC class	2	EV ratio 20-24	3.2
	1	Other EV-ratios	1
Vehicle Age	1	0-1 years	8.83
	2	1-5 years	1.85
	3	5-9 years	1
	4	9-13 years	0.54
	5	13- years	0.23
Driver Age & Gender	1	Male 0-25 years old	2.79
	2	Male 25-30 years old	3.23
	3	Other Driver Combinations.	1
Driver Age	1	Driver 0-40 years old	3.18
	2	Driver 40- years old	1

Table 1: Resulting Rating factors for the suggested new Tariff

Factor	Class	Initial Model	Simple Model
Zone	1	5.79	5.653
	2	3.49	3.855
	3	1	1
MC class	2	2.78	3.660
	1	1	1
Vehicle Age	1	4.93	8.379
	2	1.48	1.564
	3	1	1
	4	0.49	0.370
	5	0.19	0.173
Driver Age & Gender	1	3.129	4.479
	2	2.68	5.419
	3	1	1
Driver Age	1	3.41	1.838
	2	1	1

Table 2: Resulting Rating factors for the suggested new Tariff

Factor	Frequency		Severity	
	Train	Test	Train	Test
Base Model	0.0917	0.0925	1569085720	1448394752
Simple Model	0.0918	0.0925	1575950733	1455552014

Table 3: Average deviance of our to original model compared to the simple model.

Factor	Unit Deviance		Guassian Deviance	
	Train	Test	Train	Test
Base Model	2546	2652	23026857	22036454
Simple Model	2549	2658	23022361	22041698

Table 4: Average deviance of our to original model compared to the simple model.