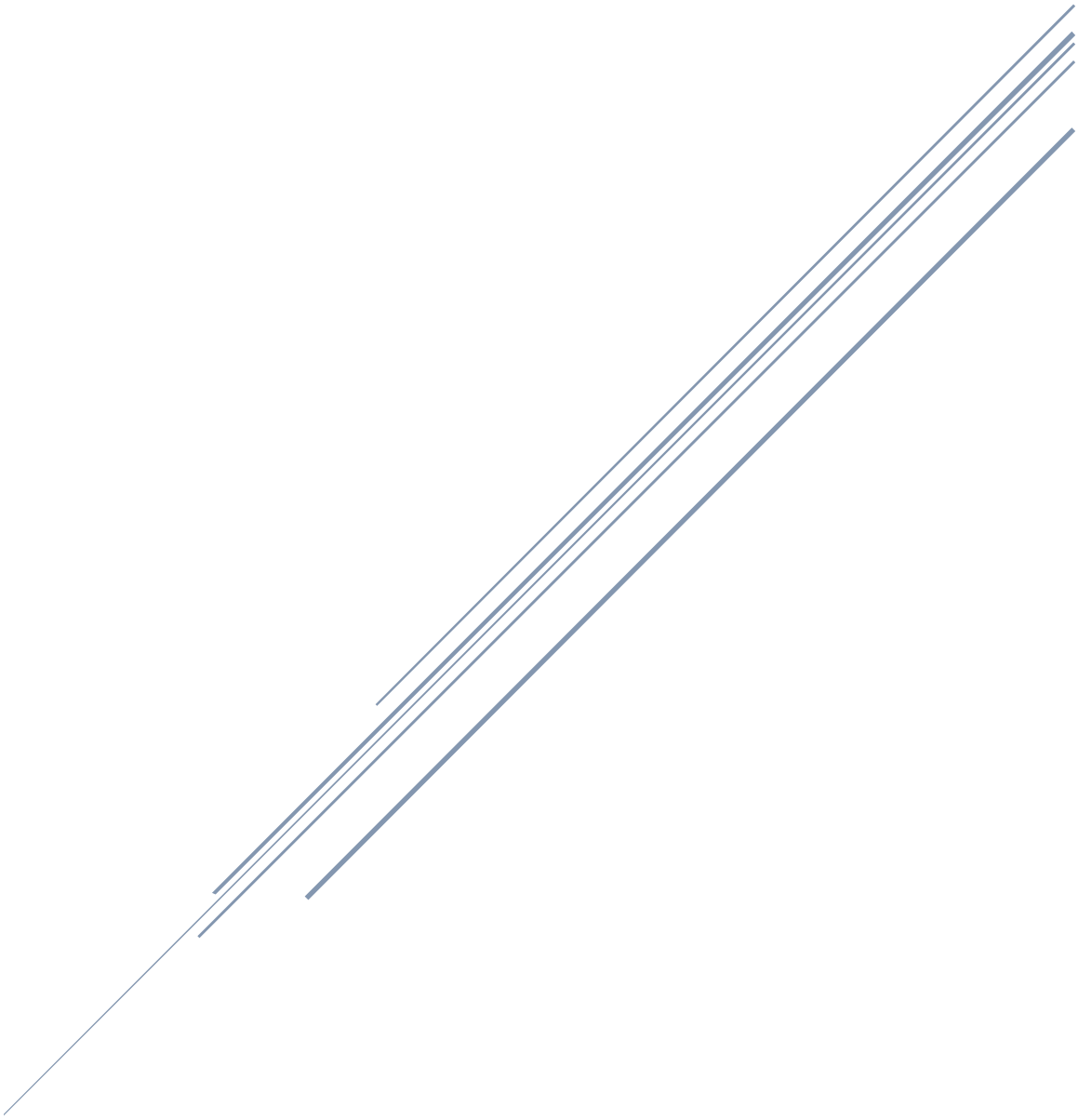
LOSS COST MODELLING

ACTSC 489 Final Project



Vani Singh
Aurelya Wibowo

Table of Contents

Executive Summary

This process discusses the methodology and results of building a loss cost model that will then be used for pricing auto insurance policies. To achieve this, firstly, this report will build and compare a Tweedie model, which models loss cost directly, and a Frequency X Severity model, which models collision frequency and severity independently and then multiplies them together to achieve loss cost. Both models are built using the training data. After comparing both of these models' performance on a holdout dataset, this report finds that the Tweedie model has a better performance and is therefore chosen to model loss cost using non-geographical predictors from the dataset.

Afterwards, the report attempts to improve the model by introducing geographic predictors and territoria/residual modelling. The geographic predictors were derived from the 2011 Canadian census. However, the findings of this analysis revealed that the inclusion of these predictors resulted in overfitting of the model to the training dataset. Therefore, it was concluded that the Tweedie model that did not include any territorial or geographic predictors yielded the most optimal performance and was subsequently recommended for utilization in modeling collision loss cost.

Finally, the loss cost model was transformed into an auto insurance pricing model by incorporating the deductibles. Additionally, the report deliberated on the methods to extrapolate this data for upcoming years, specifically for the accident year 2022.

Introduction and Purpose

Loss cost modelling in insurance refers to the process of estimating the expected costs of insurance claims of losses that an insurance company is likely to incur based on historical data analysis. These models are usually used by insurance companies to determine appropriate premiums for insurance policies. These models may incorporate many factors such as policyholder characteristics, geographic location, and other relevant variables. Loss cost modelling is an important tool for insurers to manage risk and ensure that premiums are adequately priced to cover potential losses.

The purpose of this report is to build the best loss cost model for Rightprice that could be used to segment and price its auto insurance policies. The historical data used to build the model is *collDatasetTrain* which contains policy information from accident year 2019 to 2021. This report will compare two loss cost model building approaches, Frequency-Severity and Tweedie. Then, the best out of the two models will be further refined using geographic predictors and territorial/residual modelling, when appropriate.

Data Treatment

Data Assumption

The data are assumed to be independent and fully credible.

Missing & Improper Values Treatment

Response Variable

The response variable (**`Loss_cost`**, further discussed later) has no missing and improper values (negative) found in response variable. There are many 0 values as seen in the histogram in Appendix A.1. Its summary could also be found in Appendix A.2.

Predicting Variable

There are some missing values found in the data (See Appendix A.3). Metadata is used to confirm if there are any meaning to these missing values. For **`Num_demerit_points`**, it means 0 and for **`Num_insufficient_funds_last_3_yrs`**, **`Num_insufficient_funds_last_5_yrs`** and **`Num_yrs_since_last_at_fault_claim`**, it means that the policyholder/driver never had an insufficient fund or never had an at-fault claim. Thus, those NA values are replaced according to this.

For others, if the variable is continuous, the missing values are replaced with the variable's mean/median and if the variable is categorical or ordinal, the missing values are replaced with the group with the highest amount of exposure.

The same treatment is used to treat improper values.

New Features Created

1. **`Age`**: Age of the principal driver at the beginning of the policy term. Calculated as the difference between **`Term_effective_date`** and **`Insured_birth_date`** in years.
2. **`Vehicle_age`**: Age of the car model at the beginning of the policy term. Calculated as the difference between **`Term_effective_date`** and **`Vehicle_model_year`**.
3. **`New_policy`**: **`Yes`** if it is the policyholder's first term joining Rightprice (i.e. **`Term_effective_date == Inception_date`**); **`No`** otherwise.
4. **`Has_partner`**: **`Yes`** if the principal driver has a partner; **`No`** otherwise.

Training and Holdout Data

The data are randomly divided into two groups — **`Training`** (70%) & **`Holdout`** (30%). **`Training`** is used for EDA and modelling, and **`Holdout`** is used for model testing.

Modelling Approaches

Tweedie Model

The Tweedie approach directly builds a model from the Loss Cost. It assumes the collision frequency and collision severity for any given policy are positively correlated. This assumption is most likely to be violated in real life and fortunately, Tweedie model is still robust despite this violation in assumption.

Frequency-Severity Model

The Frequency-Severity Model is built on the assumption that collision frequency and collision severity for any given policy is independent of one another. Therefore, since loss cost is equivalent to frequency multiplied by severity, it can be predicted by firstly predicting the collision frequency for any given policy and then multiplying it by the severity prediction for that policy. Frequency is modelled using quasi-Poisson distribution and severity is modelled using Gamma distribution.
The frequency model used to generate frequency predictions is below:

```
Collision_claim_count ~ Accident_year + Age_imputed + Age_imputed_squared +
Vehicle_age_cap_30 + Num_minor_convictions_cap_5 + New_policy + Has_partner
```

Variable Description:

Variate Name	Description
`Accident_year`	A factor variate for accident year of each policy.
`Age_imputed`	A continuous variate for the age of the policyholder.
`Vehicle_age_cap_30`	A continuous variate for the age of the insured vehicle..
`Num_minor_convictions_cap_5+`	A continuous variate with the number of minor convictions the policyholder has.
`New_policy`	A two-level categorical variate indicating whether the term effective date of a policy is the same as the inception date.
`Has_partner`	A two level categorial variate that indicates whether the policyholder has a partner or not.

Error Assumptions

Since both Tweedie and Severity are modelled using generalized linear models, normality, homoscedastic and randomness assumptions are made for the error functions of both models. These assumptions will be tested later in this report.

Model Building

EDA Approach

For each variate added to the model, the following was verified:

- There is a significant pattern in the one-way analysis and the variable captures the signals from the residuals.
- There is an increase in the mean cross validated Gini coefficient on training dataset.
- All 4 tests are passed – reasonability, significance, parsimony (AIC, BIC & F-Test) and time consistency.

Tweedie Model

Preparing the Response Variate

The response variate for the Tweedie approach is $Loss\ Cost = \frac{Total\ Losses}{Total\ Number\ of\ Exposure}$. In the data provided, it is made by dividing **`Collision_incurred_amount`** with **`Collision_earned_count`** directly.

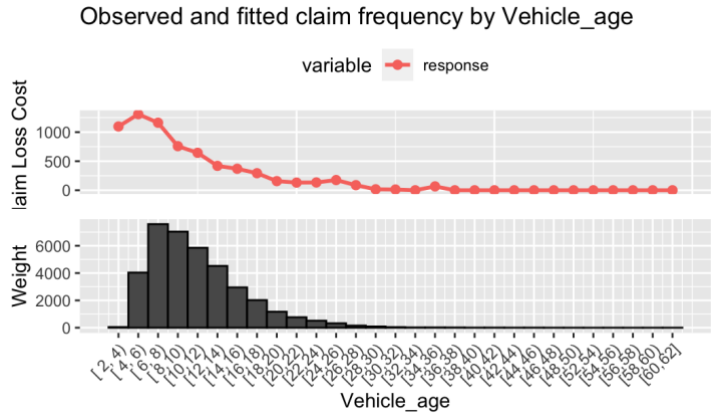
Base Model

`model_base = Loss_cost ~ Accident_year (p = 1.5)`

For the base model, Tweedie's p is set to p = 1.5 first since it is in the middle of p = 1 (Poisson) and p = 2 (Gamma). This means the effect of frequency and severity are looked equally. The p would be adjusted later.

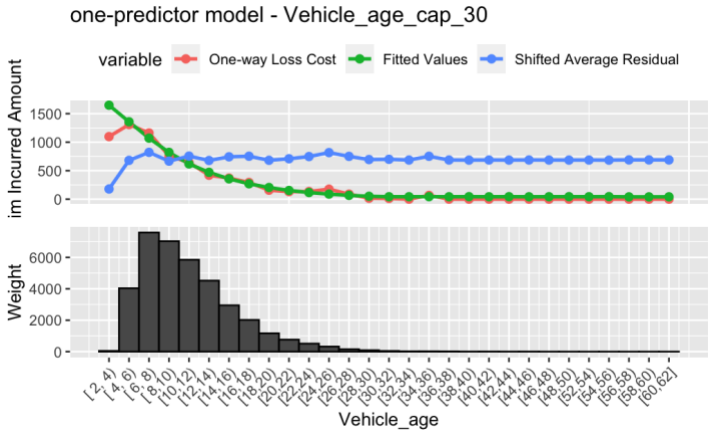
Full EDA for the 1st Variable Added

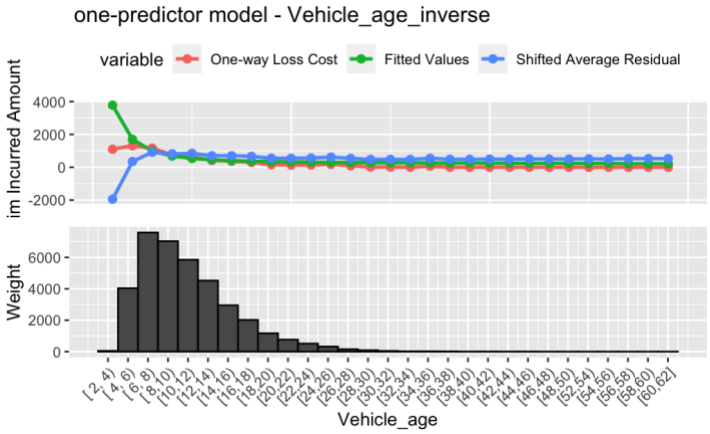
As discussed above, **`Vehicle_age`** is a variable made from **`Vehicle_model_year`** and **`Term_effective_date`**, in which both are concrete measurements and have no missing value. Thus, this makes **`Vehicle_age`** is very credible. **`Vehicle_age`** also shows a significant pattern decreasing pattern in the one-way analysis as shown below. This means that there is a significant relationship between **`Vehicle_age`** and the loss cost. Due to these reasons, **`Vehicle_age`** is looked further.



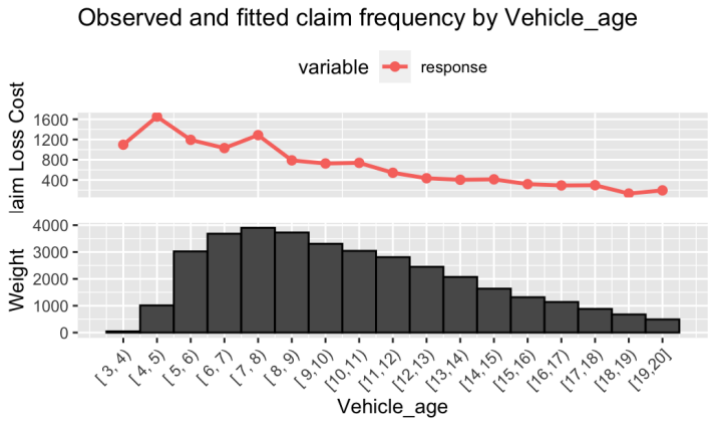
An approach to model this decreasing pattern is just to have a linear model with **`Vehicle_age`**. However, notice that the number of exposures after year 30 is very low at 139.725, making the data not credible. Due to this reason, **`Vehicle_age`** will be capped at 30.

Taking a close look, the pattern is decreasing at a decreasing rate. Thus, another approach that would be considered is 1/x. By doing this, the loss cost will converge to 0 as **`Vehicle_age`** increases.





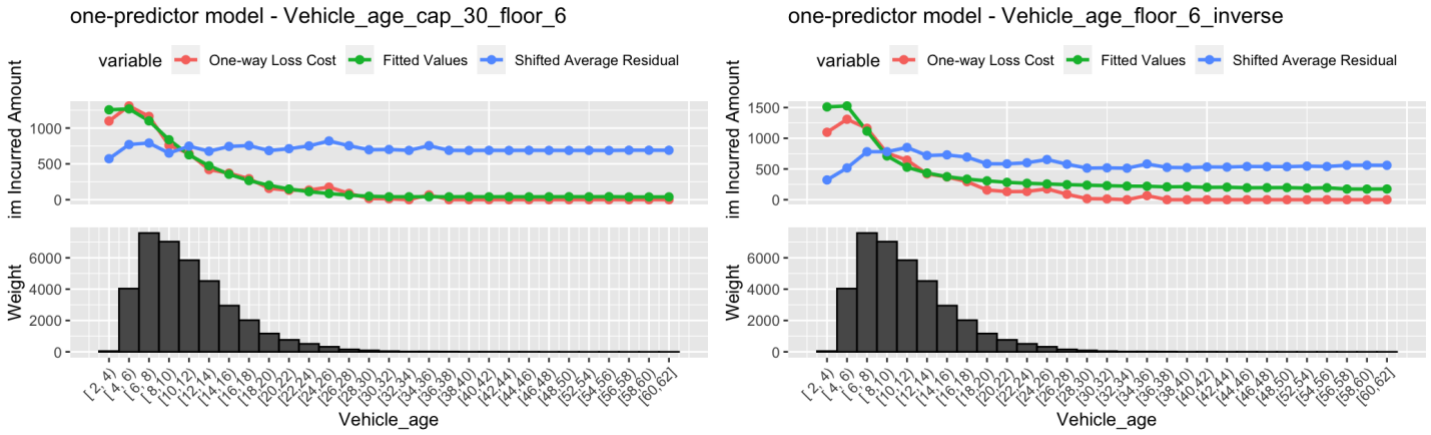
Another thing to notice is at the beginning of the graph, both models over-predict. Zooming into that part as shown below, there is some variability in the loss cost (i.e. the pattern is not smooth). Thus, both models are floored at **`Vehicle_age`** = 6 to reduce this variability.



Hence, we obtain the 2 models below:

`LC_1q5 = Loss_cost ~ Accident_year + Vehicle_age_cap_30_floor_6`
`LC_1q7 = Loss_cost ~ Accident_year + Vehicle_age_floor_6_inverse`

Comparing the residuals graphs of the two models as shown below, model **`LC_1q5`** has a better fit.



The training CV Gini for both models are also similar at 0.288 and 0.292 respectively. Furthermore, it is easier to interpret the former compared to the latter. Thus, model **`LC_1q5`** is chosen over model **`LC_1q7`**.

All 4 Tests for `Vehicle_age`

Reasonability: Newer vehicle models (i.e. small **`Vehicle_age`**) tends to have a more advanced-technology, which tends to be more costly to replace compared to an older vehicle models.

Significance: From the model summary (see Appendix B.1), **`Vehicle_age_cap_30_floor_6`** is significant.

Parsimony:

AIC

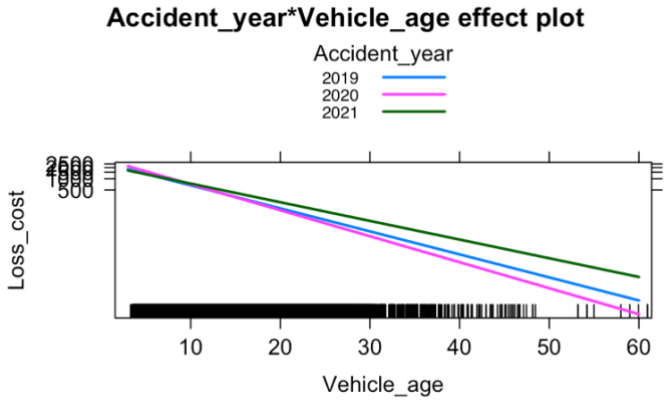
model_base	LC_1q5
84,591	84,388

The AIC is lower compared to the base model. This means that **`Vehicle_age`** provides enough improvement in fit to justify for additional parameter. Also note that when **`Vehicle_age`** is added to **`model_base`**, it has the highest decrease in AIC compared to other variables (See Appendix B.2), which further solidifies the reasoning to include **`Vehicle_age`** in our model.

F-Test

The F-Test shows p-value = 2.2e-16 < 0.05, which means that the longer model is preferred. (i.e. model **`LC_1q5`** is preferred over **`model_base`**.

Time Consistency: Although are some intersections in the time-consistency graph, the gradients of the three lines are actually very similar. Thus, **`Vehicle_age`** passes the time-consistency test. Meaning, the trend is consistent in all three years (2019 to 2021).



In conclusion, Model **`LC_1q5`** passes all 4 tests and captures all the residual signals as well. Thus, **`Vehicle_age`** would be added to the base model.

Other Variables Added

The model after adding all other variables is:

`LC_1q5_2o_3g_4i_5m_6j` = **`Loss_cost`** ~ **`Accident_year`** + **`Vehicle_age_cap_30_floor_6`** + **`Years_driving`** + **`Years_driving_squared`** + **`Num_minor_convictions_cap_5`** + **`Num_yrs_since_last_at_fault_claim_num_cap_13`** + **`New_policy`** + **`Has_partner`**

All other variables are added using the same steps as **`Vehicle_age`**. They need to capture all the residual signals well, have a lower training CV Gini and pass all the 4 tests. The summary of the rationale of adding each variable in the model is discussed below. The details of these addition could be found in Appendix B.3. Summary of progression could also be found in Appendix B.4.

Variate Name	Rationale
`Accident_year`	To incorporate the changes that happen between each year.
`Vehicle_age_cap_30_floor_6`	Clear decreasing pattern, most significant improvement in both AIC and CV Gini.
`Years_driving` + **`Years_driving_squared`**	Clear U-shaped pattern in the one-way analysis.
`Num_minor_convictions_cap_5`	Clear increasing pattern.
`Num_yrs_since_last_at_fault_claim_num_cap_13`	Clear increasing pattern.
`New_policy`	Clear increasing pattern from 'Yes' to 'No'.
`Has_partner`	Clear increasing pattern from 'No' to 'Yes'.

As more variables are added to the model, the impact becomes less noticeable (e.g. the difference in the AIC converges) and some variables become not significant anymore. Hence, the addition of more variables into the model stops here.

Interactions

The interactions between the variable are shown in Appendix B.5. The p-value that are a little significant are Years_driving*Has_partner and New_policy*Has_partner at 0.0336 and 0.0437 respectively. Since it still > 0.01, we are still going to continue with all the variables above.

Trying Different Tweedie's p Values

P	1.30	1.35	1.40	1.45	1.50	1.55	1.60	1.65
Training CV Gini	0.36061	0.36067	0.36070	0.36074	0.36078	0.36084	0.36087	0.36085

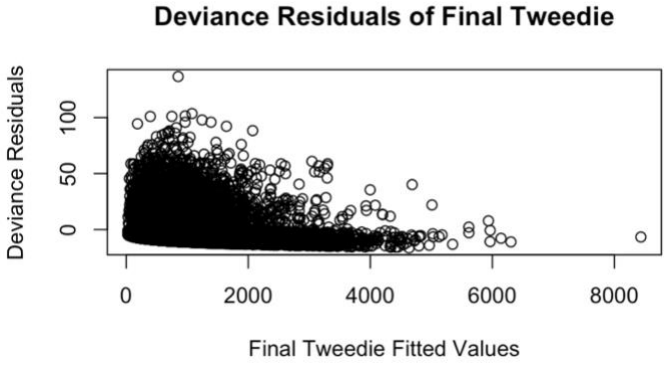
From the table above, there is not significant difference in the Training CV Gini for different p values. Thus, p = 1.5 will still be used.

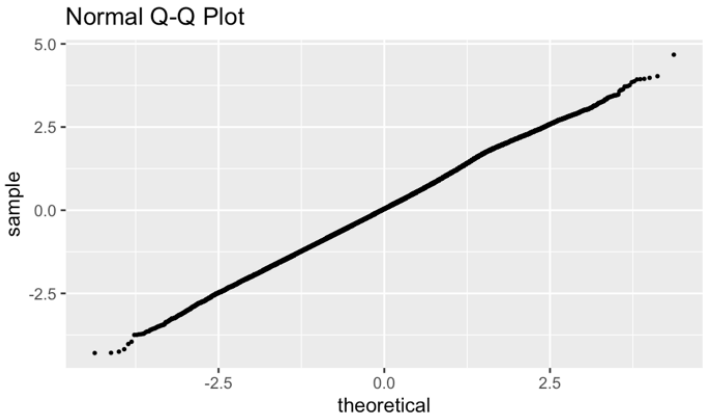
Testing Multicollinearity (VIF)

	GVIF	Df	GVIF^(1/(2*Df))
Accident_year	1.028504	2	1.007051
Vehicle_age_cap_30_floor_6	1.026565	1	1.013195
Years_driving	13.602292	1	3.688128
Years_driving_squared	13.181825	1	3.630678
Num_minor_convictions_cap_5	1.042637	1	1.021096
Num_yrs_since_last_at_fault_claim_num_cap_13	1.024213	1	1.012034
New_business_imputed	1.074514	1	1.036588
Has_partner	1.103096	1	1.050284

All the VIF's are under 10, except for **`Years_driving`** & **`Years_driving_squared`** since they are made from one another. Thus, there is no multicollinearity in the model.

Error Assumptions for the Tweedie Model





Checking homoscedasticity, there is no cone and fanning pattern in the Deviance Residual Plot, indicating constant variance. For randomness the Deviance Residual plots are above random, thus validating the assumption. Checking normality, the QQ-plot shows a straight line. Thus, residuals follow a Normal distribution. Therefore, in conclusion, all error assumptions – Homoscedasticity, randomness & normality are satisfied.

Holdout Testing

To test for overfitting, the final model, **`LC_1q5_2o_3g_4i_5m_6j`**, would be compared to the shorter models (shorter models used to build **`LC_1q5_2o_3g_4i_5m_6j`**) using the holdout data.

The tests are done in Appendix B.6. The Gini coefficients for both training and holdout datasets increases as we add more variables into our model. Compared to other models, it also performs better in both Simple Quantile Plots and Double Lift Charts. These results indicate that there is no overfitting for **`LC_1q5_2o_3g_4i_5m_6j`**.

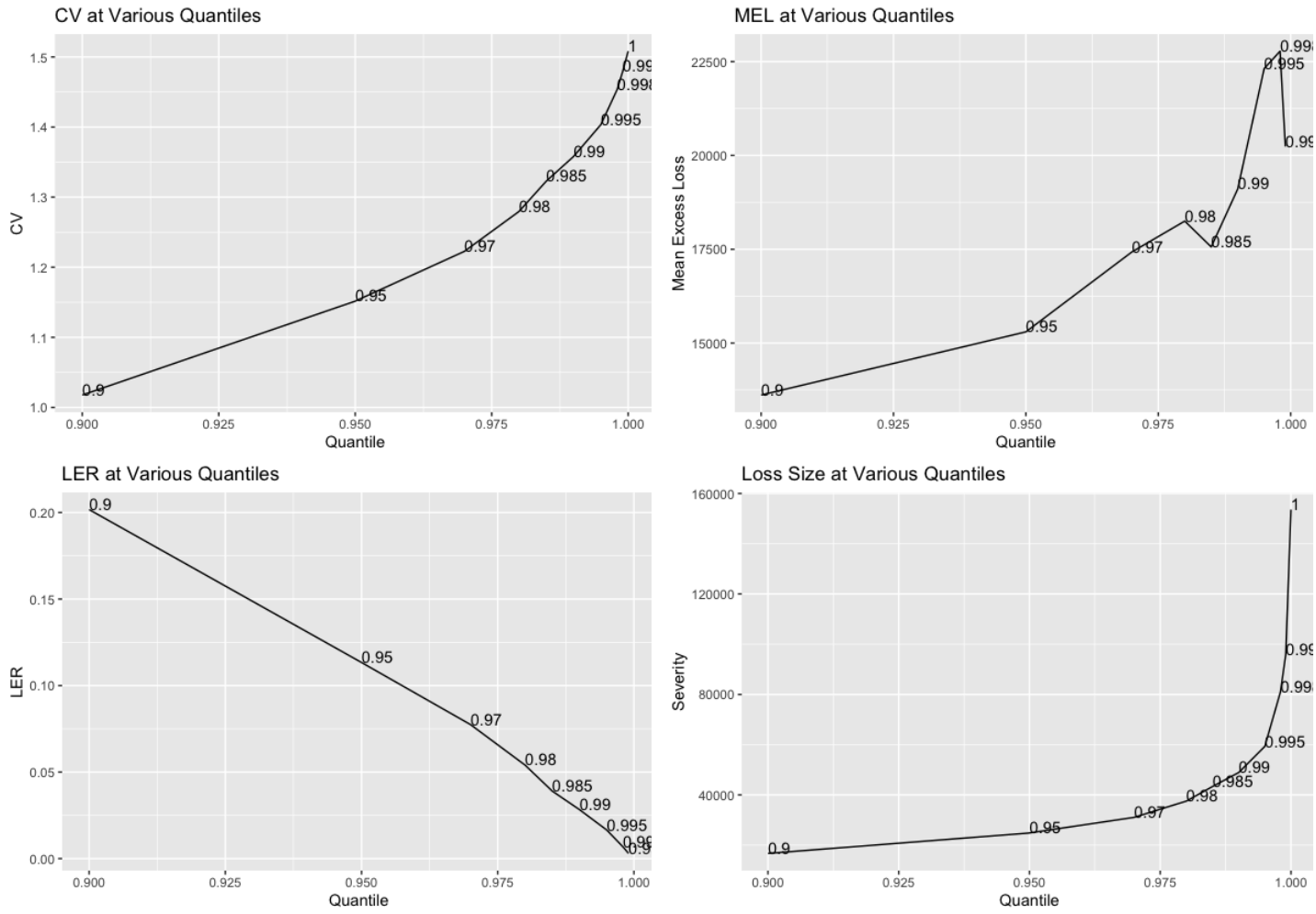
Thus, from the holdout testing, the chosen Tweedie model is **`LC_1q5_2o_3g_4i_5m_6j`**.

Frequency X Severity Model

Preparing the Severity Variate

Using fact that $Severity = \frac{Total\ Losses}{Number\ of\ Claims}$ the severity variate is created by dividing the **`Collision_incurred_amount`** variate by the **`Collision_claim_count`** variate. To avoid dividing errors and to use the gamma error distribution in the model building process, the training dataset must first be filtered to only contain entries that have at minimum 1 collision count. This decreases the training dataset to 5,666 variables.

The next step in preparing the **`severity`** variate is to analyze where the variate should be capped. This was done by plotting the coefficient of variance (CV), mean excess loss (MEL), loss expected ratio (LER) and Loss Size for various quantiles of the response variate.



For the CV and Loss size charts, a hinge at the 0.995 quantile can be observed. This suggests that the severity variate should be capped at this quantile. Since the LER is small at the 0.995 quantile and the MEL is large at the 0.995 quantile, it further confirms that the variate should be capped here. The capped value is 59319.

Summary of EDA for Severity

Variate Name	Description	Included in Model?	Notes
`New_policy` (Appendix C.3)	A 2-level categorical variate indicating whether the term effective date of the policy is the same as the inception date.	N	The coefficient for the variate was no longer significant once other variates were added.
`Vehicle_age` (Appendix C.4)	A continuous variate for the age of the vehicle being insured. Calculated by finding the time between model year and the policy's term effective date.	Y	Capped at 23 years.
`Agecat` (Appendix C.5)	An ordinal variate grouping policyholders based on their age.	Y	group 1: less than 25, group 2: 25-35, group 3: 36-50, group 4: 51-65, group 5: over 65. Missing age was replaced with the mean.
`Liab_driving_record_Ind` (Appendix C.6)	A 2-level categorical variate indicating whether the policyholder's driving record is rated a '6' or not.	Y	NAs were assumed to not be ranked '6'.

Has_partner (Appendix C.7)	A 2-level categorical variate indicating whether the policyholder is married or not.	Y	Common law and same sex marital status's were assumed to be married and all others were grouped into single. Not as significant as other predictors.
Num_at_fault_claims_past_1_yr_Ind (Appendix C.8)	A 2-level categorical variate indicating whether the policyholder has an at fault claim in the past year prior to the policy's effective date.	Y	
Num_minor_convictions (Appendix C.10)	A continuous variate for the number of minor convictions that policyholder has.	N	No clear pattern on how it affects the response variate, not reasonable.
Num_yrs_since_fault (Appendix C.9)	A categorical variate grouping policyholders by whether their last at fault claim was less than 5 years ago or more.	N	Not significant to the model.
Vehicle_retail_price (Appendix C.12)	A continuous variate for the retail price of the insured vehicle for the policy	Y	Missing values were replaced with the mean retail price.
Vehicle_horsepower (Appendix C.13)	A continuous variate for the horsepower of the insured vehicle for the policy	N	Too heavily correlated with retail price. The Gini for the model containing retail price was higher.
Vehicle_wheelbase (Appendix C.14)	A continuous variate for the wheelbase of the insured vehicle for the policy	N	Too heavily correlated with retail price. The Gini for the model containing retail price was higher.

Results of Cross Validated Gini Coefficient Tests for the Severity Model

	Model	Training CV Gini
7	**severity ~ AY_factor + Vehicle_age + Agecat + Liab_driving_record_Ind + Marital_status_Ind + Num_at_fault_claims_past_1_yr_Ind + Vehicle_retail_price**	**0.30461**
8	severity ~ AY_factor + Vehicle_age + Agecat + Liab_driving_record_Ind + Marital_status_Ind + Num_at_fault_claims_past_1_yr_Ind + Vehicle_horsepower	0.30276
9	severity ~ AY_factor + Vehicle_age + Agecat + Liab_driving_record_Ind + Marital_status_Ind + Num_at_fault_claims_past_1_yr_Ind + Vehicle_wheelbase	0.29778

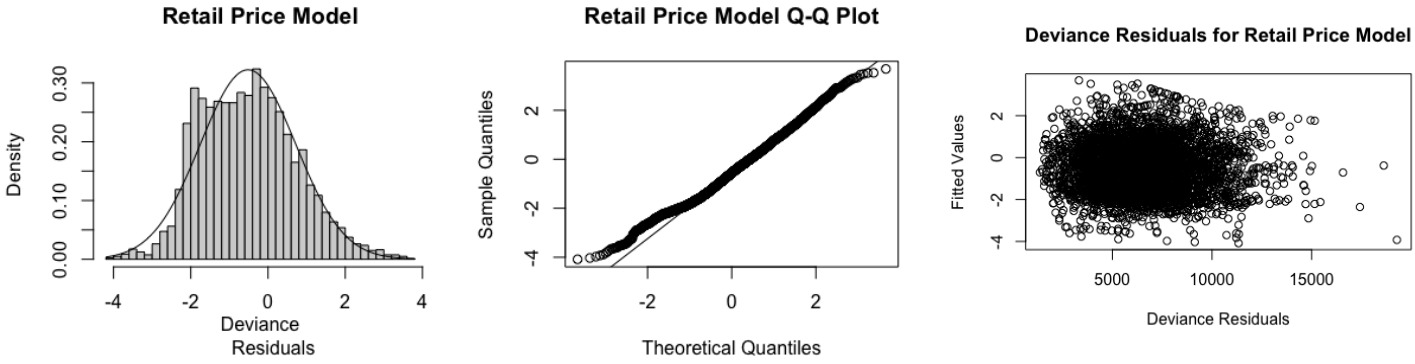
As shown in Appendix C.15, as variates chosen through EDA are added to the model, the mean cross-validated Gini coefficient on the training data increases as a result. Models 7, 8 and 9 have the same number of variates, except one has **Vehicle_retail_price**, one has **Vehicle_horsepower**, and one has **Vehicle_wheelbase**. All these variates cannot be in the model since they are heavily correlated, as shown in the correlation matrix in Appendix C.11. The table above shows that the model containing **Vehicle_retail_price** (8) has the highest mean cross validated Gini on the training data and thus, this is the model that will be tested for overfitting on the holdout data.

Results of Holdout Testing for the Severity Model

According to the model output for the model containing **Vehicle_retail_price** (7) in Appendix C.12.2, the least significant variate in the model is **Liab_driving_record_Ind**. To test for overfitting, a reduced model with that variate removed will be testing against the full model. Firstly, referring to the output in Appendix C.17.1, the full model has lower values for AIC and BIC compared to the full model on the holdout data. Secondly, the simple quantile plots shown in Appendix C.17.2 suggest that the fit for the full model is better, there are less reversibles in the full model, and

the lift for both models are similar. Therefore, the simple-quantile plots suggest that the full model is better. The double lift chart in Appendix C.17.3 shows that the full model fits better on the left and right tail of the data and there is a lot of overlap between the 2 models' performance in the middle of the data. Therefore, the double lift chart also suggests that the full model performs better. Lastly, the Gini coefficient on the holdout data for the full model is 0.33151 whereas for the reduced model it is 0.32636 (Appendix C.17.4). Therefore, the full model (i.e. Model 7 in the table above) is chosen as the best Frequency X Severity model.

Error Assumptions for Severity Model



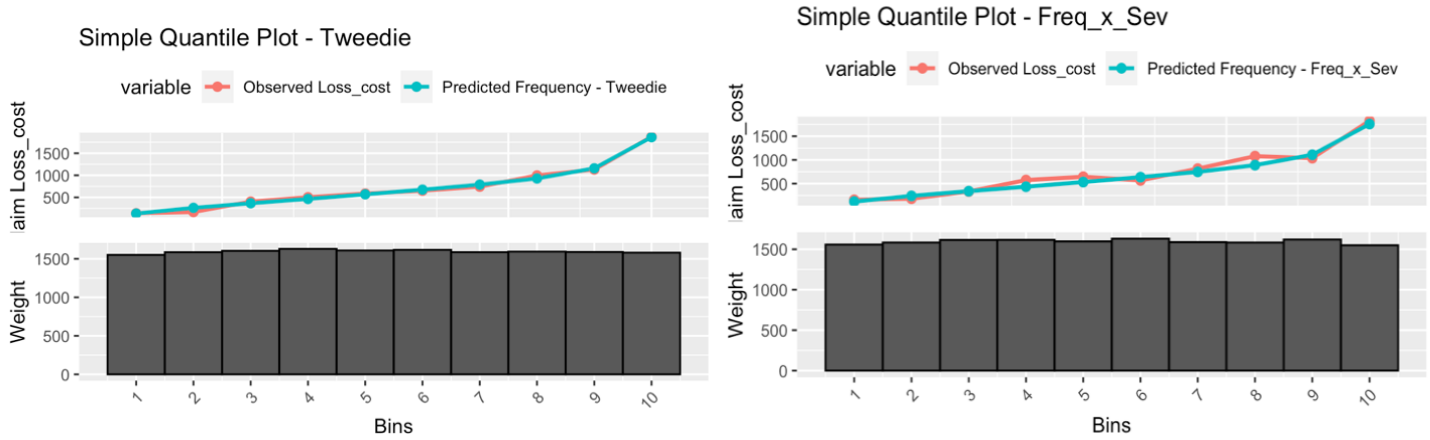
The plots above suggest that the gamma error function is the correct function for the error distribution of this model. The Q-Q plot and plot of the deviance residual density suggest that the assumption of normality is valid and the fitted-values and deviance residual plot suggest that the assumption of homoscedasticity is valid.

Combining the Severity Model Predictions with the Frequency Model Predictions
The Gini coefficient on the holdout data for the frequency model was 0.23783 and the holdout Gini coefficient for the severity model is 0.33151. When the predictions of these two models are multiplied together to create predictions for collision loss cost, the resulting holdout Gini coefficient is 0.36379 (Appendix C.18).

Comparing the Tweedie and Frequency X Severity Approach

The aim of this step is to compare the final Tweedie and Frequency X Severity models mentioned above and eventually pick the best model out of the two.

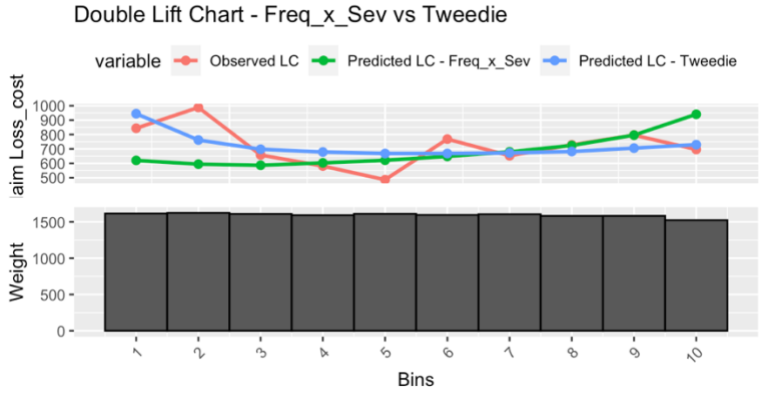
Simple Quantile Plots



Tweedie	Frequency X Severity
Fit: Small deviations in bin 2 & bin 8.	**Fit:** 2 large deviations in bin 3 to 5 & bin 7 to 9 and small deviations in bin 2 & bin 8.
Lift: Similar fit to observed.	**Lift:** Similar lift to observed.
Monotonicity: No reversal.	**Monotonicity:** Reversals in bin 6 & bin 9.

Based on the Simple Quantile Plots, the Tweedie performs better compared to Frequency X Severity.

Double Lift Chart



- Tweedie performs better at the end points.
- Frequency X Severity performs better in the middle bins but not much difference than Tweedie.
- Tweedie is closer to observed in 6 bins whilst Frequency X Severity in only 4 bins.

Thus, based on the Double Lift Charts, Tweedie performs better compared to Frequency X Severity.

Gini Coefficient

Tweedie	Frequency x Severity
0.37665	0.36379

Since Tweedie's holdout Gini is higher compared to Frequency X Severity's, the Tweedie model is preferred.

Chosen Model, Benefits and Limitations

Since all the three tests are in favor of the Tweedie model, the Tweedie model, **`LC_1q5_2o_3g_4i_5m_6j`**, is chosen. a limitation of the Tweedie model is that it provides less insight into loss cost compared to the Frequency X Severity model as it does not model frequency and severity separately. Moreover, the collision loss cost values within the dataset are predominantly skewed towards 0, and thus, when loss cost is over 0, it is considered a rare occurrence. Consequently, the resulting model exhibits a diminished level of stability.

Despite these limitations, Tweedie is built upon only a single model, unlike Frequency X Severity which are built upon two models. This makes Tweedie model's standard errors not as amplified, resulting to a more accurate parameter. Additionally, this also means that building a Tweedie model is more efficient than Frequency X Severity.

Geographic Predictors

The aim of this step is to analyze whether any geographic variates should be added to further refine our current best model, **`LC_1q5_2o_3g_4i_5m_6j`**. The data used for this was the 2011 census data from Statistics Canada, which are grouped by FSA.

Using forward selection, the 7 socio-economic variates in the dataset were added to the model one by one. It was observed however, that none of these variates were significant once added to the model (See Appendix D.1). A reason to this is due to the very outdated census data, which may not be relevant to collision loss cost in years 2019, 2020 and 2022 – the timeframe of the data in which the Tweedie model is built upon. This reasoning is further solidified with the fact that Ontario has one of the fastest growing populations in Canada (including in immigration), which could alter the demographic of the regions.

Additionally, the census data set comprises of numerous variables that are temporal in nature, including the proportion of the population in a specific age group. As the population ages, these proportion will naturally shift over time, making the value inside each threshold to be unknown, if those age group are the specific age groups that are needed.

Therefore, our best model is still **`LC_1q5_2o_3g_4i_5m_6j`**.

Territorial/ Residual Modelling

The multiplicative residual of each training entry could be calculated using $Residual = \frac{Observed\ Loss\ Cost}{Predicted\ Loss\ Cost}$. The multiplicative residual of a territory (i.e. FSA) is the weighted average of all the multiplicative residuals of the entries that are in that specific territory, where the weight is the amount of exposure (i.e. `Collision_earned_count`). From *Figure 1* below, it could be observed which territories our current best model overpredict and underpredict. Thus, utilizing these residuals, the aim of this step is to analyze whether territorial/residual modelling could further enhance our best model.

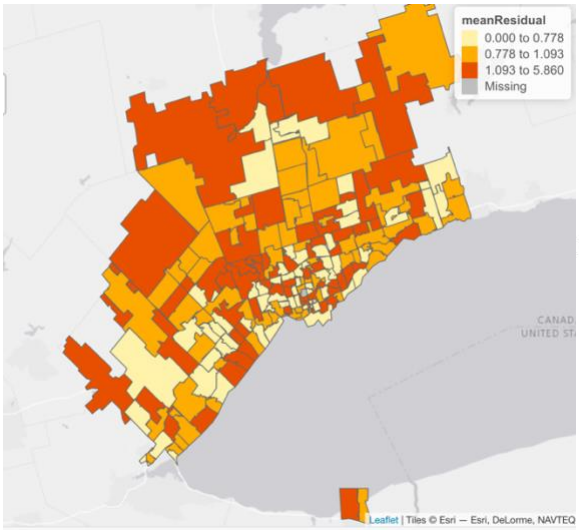


Figure 1: Mean Residual Smoothing

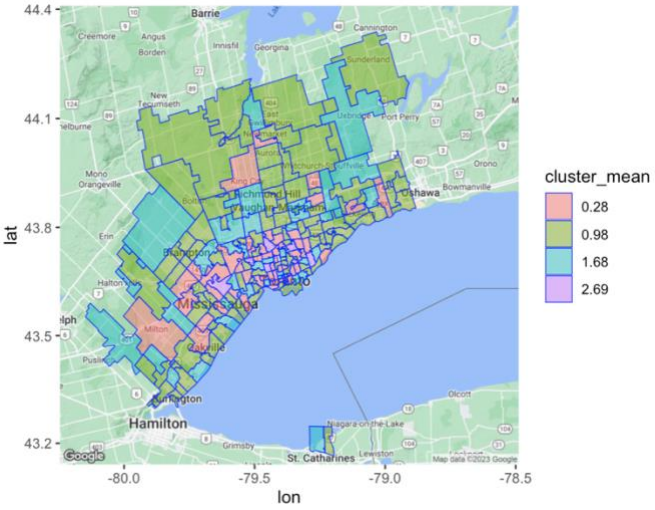


Figure 2: k-Means (k = 4) Cluster Map – No Smoothing

Performing k-means (k = 4), *Figure 2* above is obtained. However, notice that from the table below, there are many FSAs with only a small number of exposures (the 3rd quantile is only 172.35). Due to this reason, it is not preferable to use the k-means directly and add the variables into the current best model. This leads to the use of *Credibility Smoothing*.

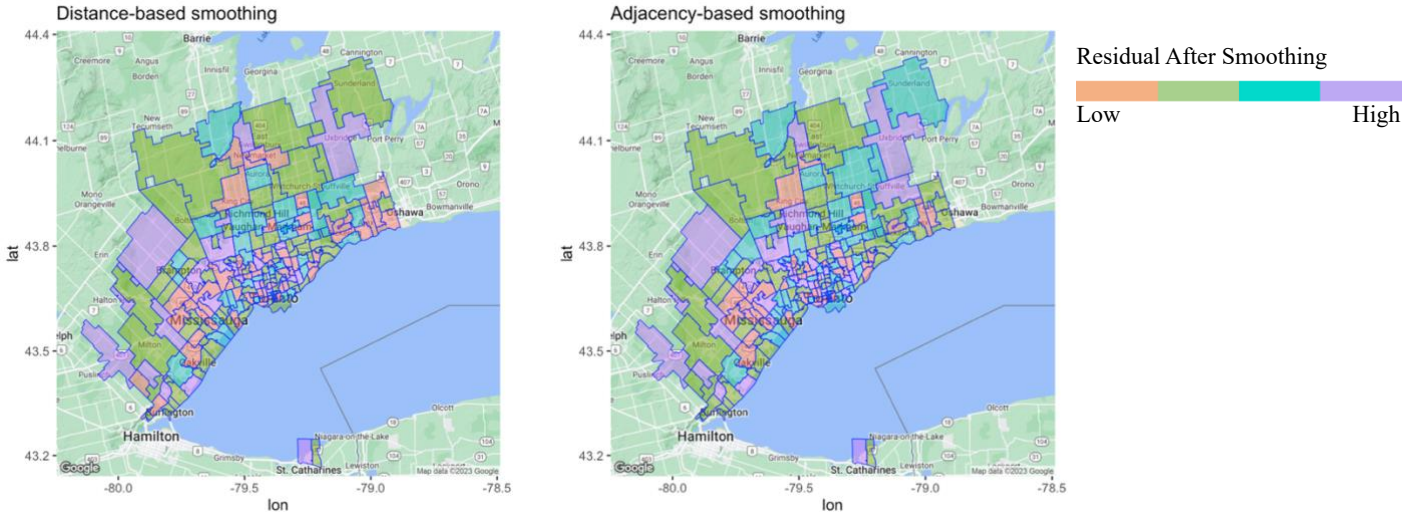
Summary of FSA Exposures

Minimum	1st Quantile	Median	Mean	3rd Quantile	Maximum
1.21	33.12	72.42	201.04	172.35	4,750.48

Credibility Smoothing

Cross validation and the function *territorialSmoothing* are used to tune the values of J and k to find the most ideal combinations that increases the training CV Gini of the model. The combination of values of J = {50, 150, 200} and k = {2, 3, 4} are used to tune. Note that J = 0 is not used, since it implies no smoothing is used, which is not preferred. J = 50 and k = 4 results to the highest training CV Gini. The full results are shown in Appendix E.1.

The residual maps after the smoothing with J = 50 and k = 4 are shown below.



Holdout Testing using J = 50 and K = 4 Found from Tuning

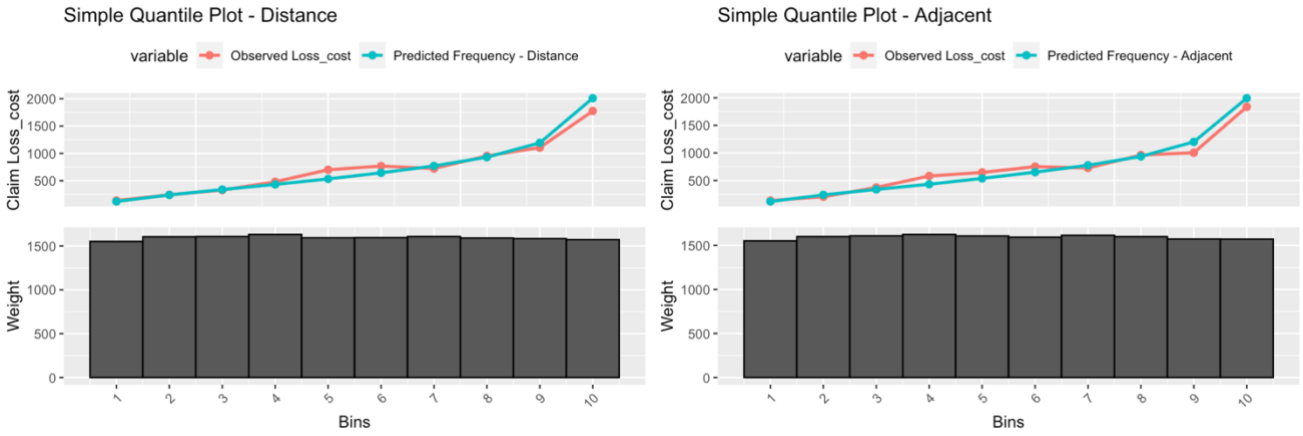
Gini

J = 50 , K = 4

	Tweedie	Distance Based Residual	Adjacency Based Residual
Holdout	0.37665	0.35641	0.35122
Training	0.36691	0.39426	0.39538

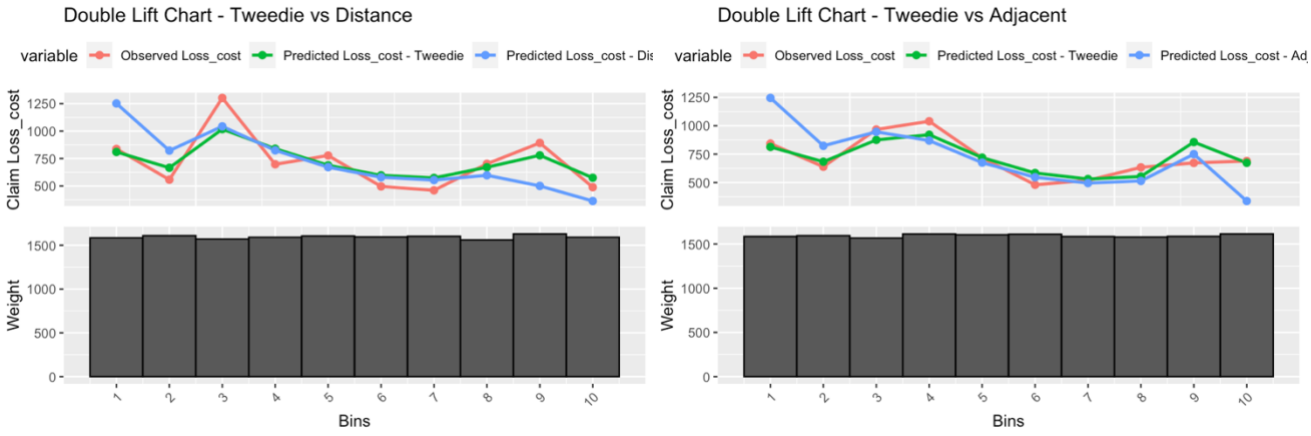
Compared to Tweedie, the training Gini improves a lot for both smoothing methods, but the holdout Gini worsen.

Simple Quantile Plot



Compared to Tweedie's simple quantile plot on page 11, all fit, lift and monotonicity is better in Tweedie compared to both of the smoothing methods.

Double Lift Plot



For both smoothing methods, Tweedie and the smoothing method performs similarly in the middle bins, but Tweedie performs better at the edge.

Conclusion

Based on Gini, Single Lift Plot and Double Lift Plot, Tweedie performs better compared to both of the smoothing methods. A reason to this could be overfitting.

Trying J = 250 and K = 2

The overfit above could be caused by the low J = 50 and a higher k = 4 in addition to having many FSAs with small number of exposures. This is because there is less smoothing when J is lower, leading to residuals getting closer to their true residual value rather than being smoothed out. The higher the k (i.e. more clusters) also means that each cluster is more curated, which leads the same result. Hence, J = 250 and k = 2 is tried out.

Looking at the table below the training Gini coefficients for both smoothing methods are not as high as when J = 50 and k = 4. The holdout Gini coefficients also do not decrease as much. However, it should also be observed that in each case, residual modelling increases the training Gini and lowers the holdout Gini. A reason for this may be because the original

Tweedie model already contained many variates and thus, adding more variates would result in the model being overfit to the training dataset, lowering its performance.

J = 250, K = 2

	Original Tweedie	Distance Based Residual	Adjacency Based Residual
Holdout	0.37665	0.36272	0.36558
Training	0.36691	0.38060	0.37820

Concluding Remarks

Since the original Tweedie model without any geographical or territorial predictors has the best performance on the holdout dataset, that is the final model chosen to predict loss cost. Fitting the model on the entire dataset, as shown in Appendix F.1 the predictor variates are still significant and still hold the same reasonable interpretation as they did in the training model. Therefore, the Tweedie model **`LC_1q5_2o_3g_4i_5m_6j`** is chosen as the final model.

Transforming into Pricing Structure

To convert the loss cost model into rating structure, deductibles must be added to the model. This was done by adding the **`Collision_deductible`** variate in the dataset as an offset term in the model. Once added, the same checks for predictor significance and reasonability were done so ensure consistency (Appendix F.2). The offset for policies with a $1000 deductible is -0.09065, which means that all other things remain consistent, a policy with a $1000 deductible will have a lower rate than a policy with a $500 deductible.

Adjustments to be Valid for 2022

```
  Accident_year `sum(Collision_earned_count)` `mean(Loss_cost)`
1          2019                       19604.            1100.
2          2020                       17573.            1069.
3          2021                       15958.             803.
```

As seen above, the mean loss cost is decreasing from 2019 to 2021. However, the decrease from 2020 to 2021 is much larger than the decrease from 2019 to 2020. By purely following this trend, to trend the model for 2022, the Accident_Year factor for 2022 must be smaller than the Accident_year factor for 2021. This would be consistent with the negative trend of loss cost as displayed by this dataset. Some other factors to consider would include:

- Updating data sources: Insurance companies may need to gather more recent and relevant data sources to reflect current trends in collision frequency, such as changes in driving patterns due to the COVID-19 pandemic, which was not reflected in the simulated dataset.
- Incorporating new factors: Changes in technology, such as the increasing prevalence of driver assistance systems or electric vehicles, may need to be factored into the frequency model's base rate.
- Considering changes in driving behavior: as more people are working from home, roads during rush hour are less busy, which will probably result in a large decrease in loss cost.

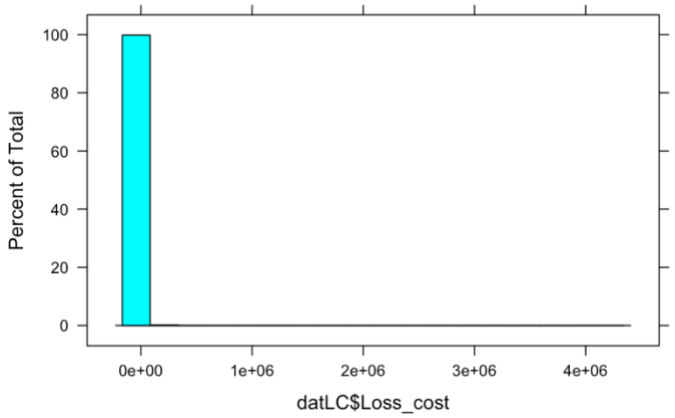
Limitations and Improvements

On top of a Tweedie model's limitations mentioned above, other limitations and improvements include:

- None of the variables in the geographical data was significant, resulting to not being able to further refine the model. Finding another source for geographical variables that are more relevant to the data could be used.
- Territorial/residual modelling could actually significantly improve a model. Making a smaller Tweedie model (i.e. lesser variable) and then doing the territorial/residual modelling could be tried out and used as a comparison.
- The holdout data are used many times – building frequency, severity and Tweedie models, comparing Tweedie and Frequency X Severity, adding geographical model and territorial/residual modelling. This could significantly reduce the power of the holdout data in measuring out-of-sample model performance.

Appendices

Appendix A.1: Histogram of Loss_cost



Appendix A.2: Summary of Loss_cost

Minimum	1st Quantile	Median	Mean	3rd Quantile	Maximum
0	0	0	1,000	0	4,178,000

Appendix A.3: Total NA of Each Variables

```
## # A tibble: 42 x 2
##    `Variable Name`      `NA Count`
##    <chr>                     <int>
##  1 id                            0
##  2 Accident_year                 0
##  3 Policy_number                 0
##  4 Term_effective_date           0
##  5 Vehicle_number                0
##  6 Inception_date               53
##  7 Policy_term                   0
##  8 Insured_postal_code           0
##  9 Risk_postal_code          60718
## 10 Collision_deductible          0
## # ... with 32 more rows
```

Appendix B.1: Model `LC_1q5` Summary

```
Call:
glm(formula = Loss_cost ~ Accident_year + Vehicle_age_cap_30_floor_6,
    family = tweedie(var.power = 1.5, link.power = 0), data = datLC_train,
    weights = Collision_earned_count)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-12.096   -8.276   -6.304   -4.206  151.102

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                7.949737   0.105692  75.216   <2e-16 ***
Accident_year2020          0.076473   0.083347   0.918    0.359
Accident_year2021          0.109834   0.085260   1.288    0.198
Vehicle_age_cap_30_floor_6 -0.143458   0.009199 -15.595   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Tweedie family taken to be 1167.422)

    Null deviance: 6474677  on 79788  degrees of freedom
Residual deviance: 6176959  on 79785  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 7
```

Appendix B.2: AIC of Different Variables when Added to `model_base`

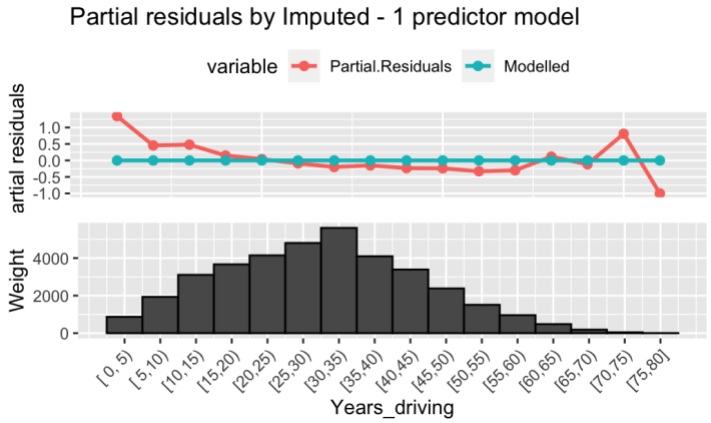
Note that AIC here is the AIC of the best model found with the modified version of the variables.
E.g. Vehicle_age here is Vehicle_age_cap_30_floor_6

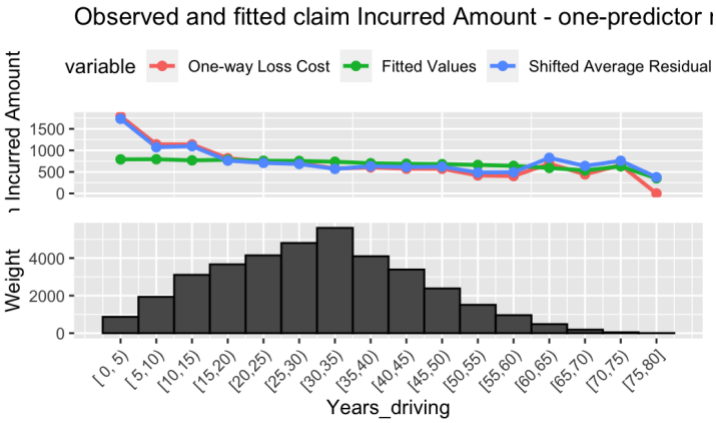
Variable (Accident_year + _)	AIC
	84,591
Vehicle_age	84,387
Vehicle_retail_price	84,576
Vehicle_horsepower	84,566
Num_minor_convictions	84,555
Has_partner	84,568
New_policy	84,563
Years_driving	84,538
Age_imputed	84,547

Appendix B.3: Adding Other Variables to the Model

Adding 2nd Variable to the Model – `Years_driving` + `Years_driving_squared`

The residual plots below show a U-shaped pattern. This means that there is a quadratic relationship between **`Years_driving`** and loss cost.

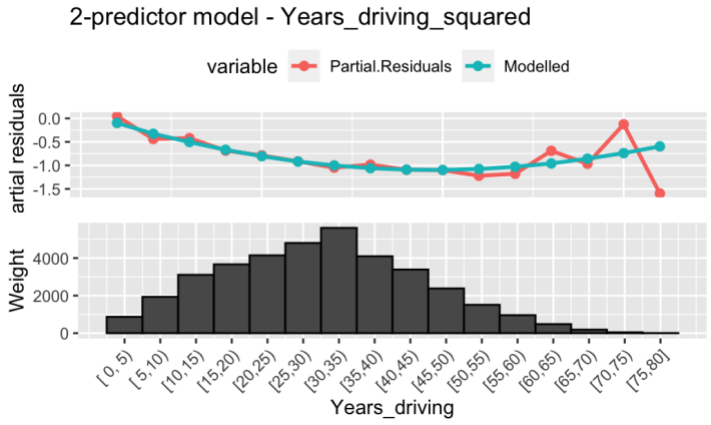


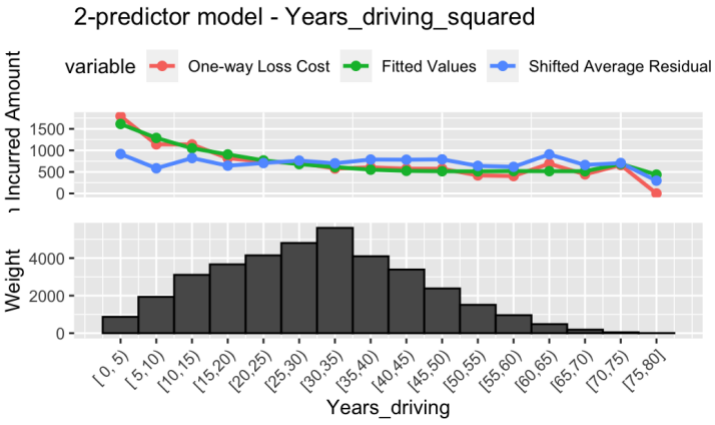


Model:

`LC_1q5_2o` = **`Loss_cost`** ~ **`Accident_year`** + **`Vehicle_age_cap_30_floor_6`** + **`Years_driving`** + **`Years_driving_squared`**

From the residual plots below, the model captures the residual signals well.





Training CV Gini

LC_1q5	LC_1q5_2o
0.28842	0.32866

Since the training CV Gini increases, the longer model is preferred.

Reasonability

When a driver just started driving, he has little driving experience/skills, which could lead to higher loss cost. After driving for some time, the driver gains some skills and the expected loss cost decreases. However, as the driver gets older, the driver's physical ability deteriorates, leading to a deteriorated driving skill.

Parsimony

AIC

LC_1q5	LC_1q5_2o
84,387	84,346

The AIC decreases. Thus, the longer model is preferred.

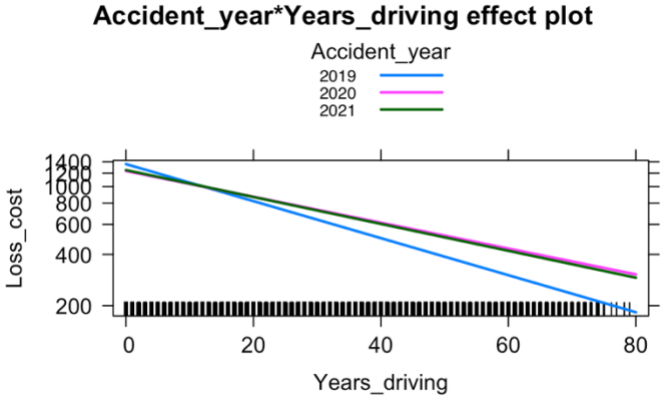
F-Test

p-value = 4.226e-15 < 0.05
This means that the longer model is preferred.

Significance
Both variables are very significant.

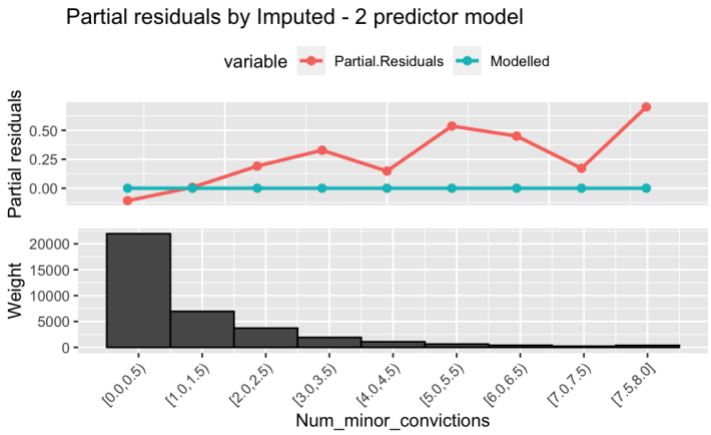
```
Call:
glm(formula = Loss_cost ~ Accident_year + Vehicle_age_cap_30_floor_6 +
    Years_driving + Years_driving_squared, family = tweedie(var.power = 1.5,
    link.power = 0), data = datLC_train, weights = Collision_earned_count)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-14.253   -8.196   -6.237   -4.163  121.997

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                8.6924536  0.1475727  58.903  < 2e-16 ***
Accident_year2020          0.1161320  0.0806112   1.441   0.1497
Accident_year2021          0.1669045  0.0826134   2.020   0.0434 *
Vehicle_age_cap_30_floor_6 -0.1388095  0.0089743 -15.467  < 2e-16 ***
Years_driving              -0.0483444  0.0083993  -5.756 8.66e-09 ***
Years_driving_squared       0.0005317  0.0001359   3.912 9.15e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Tweedie family taken to be 1079.276)

    Null deviance: 6474677  on 79788  degrees of freedom
Residual deviance: 6105486  on 79783  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 7
```

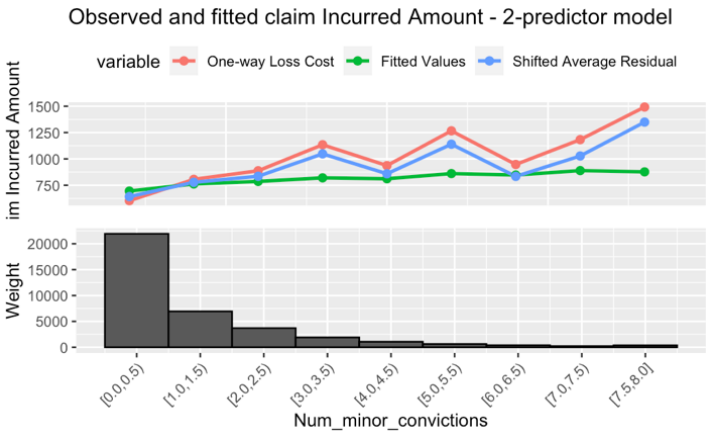
Time-Consistency
From the graph below, although the three lines intersect, the gradients don’t differ by much. Thus, it still passes the time-consistency test.



Adding 3rd Variable to the Model – `Num_minor_convictions_cap_5`

The residual plots below show an obvious increasing pattern. This means that there is a significant relationship between **`Num_minor_convictions`** and loss cost.

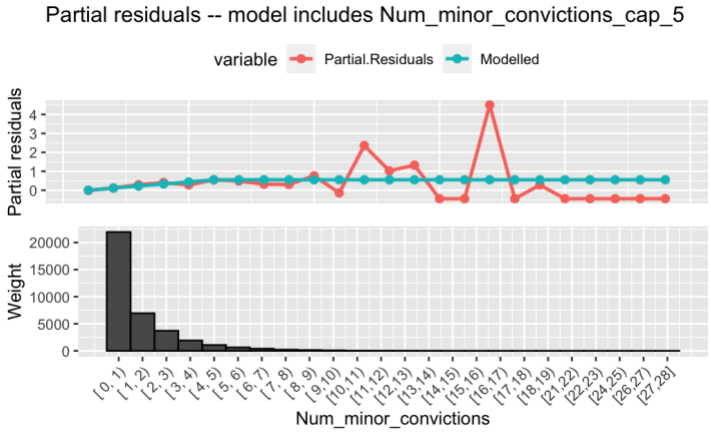


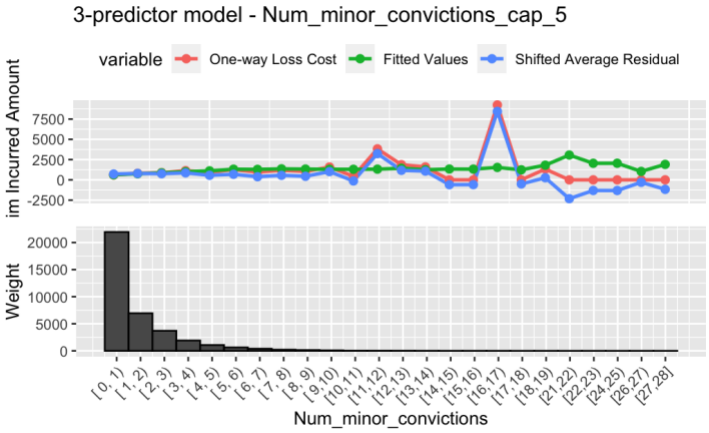


Since there is very little exposure after 5, a cap of 5 is set.

Model:
`LC_1q5_2o_3g = Loss_cost ~ Accident_year + Vehicle_age_cap_30_floor_6 + Years_driving + Years_driving_squared + Num_minor_convictions_cap_5`

From the residual plots below, the model captures the residual signals well.





Training CV Gini

LC_1q5_2o	LC_1q5_2o_3g
0.32866	0.34322

Since the training CV Gini increases, the longer model is preferred.

Reasonability

The higher the number of minor convictions means that the driver is more often to not obey the traffic rules, which includes driving above the speed limit. Thus, that driver would be more likely to get into an accident or a more severe accident.

Parsimony

AIC

LC_1q5_2o	LC_1q5_2o_3g
84,346	84,328

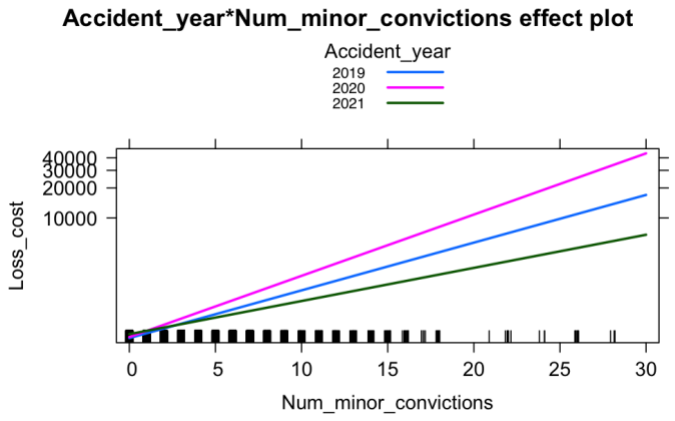
The AIC decreases. Thus, the longer model is preferred.

F-Test
p-value = 1.578e-06 < 0.05
This means that the longer model is preferred.

Significance
The variable is very significant.

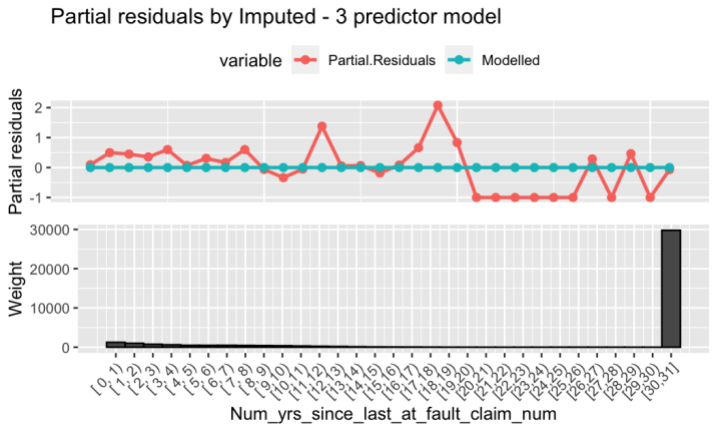
```
Call:
glm(formula = Loss_cost ~ Accident_year + Vehicle_age_cap_30_floor_6 +
    Years_driving + Years_driving_squared + Num_minor_convictions_cap_5,
    family = tweedie(var.power = 1.5, link.power = 0), data = datLC_train,
    weights = Collision_earned_count)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-14.872   -8.171   -6.211   -4.157  122.224

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                8.5385586  0.1488900  57.348  < 2e-16 ***
Accident_year2020          0.1128404  0.0793998   1.421   0.1553
Accident_year2021          0.1739404  0.0812694   2.140   0.0323 *
Vehicle_age_cap_30_floor_6 -0.1360236  0.0088472 -15.375  < 2e-16 ***
Years_driving              -0.0496487  0.0082759  -5.999 1.99e-09 ***
Years_driving_squared       0.0005811  0.0001340   4.337 1.45e-05 ***
Num_minor_convictions_cap_5 0.1102887  0.0227391   4.850 1.24e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Tweedie family taken to be 1042.483)

    Null deviance: 6474677  on 79788  degrees of freedom
Residual deviance: 6081452  on 79782  degrees of freedom
AIC: NA
```

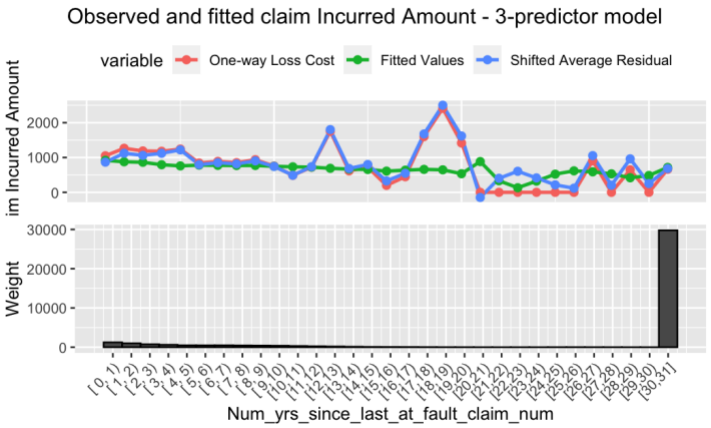
Time-Consistency
From the graph below, the three lines do not intersect. Thus, it passes the time-consistency test.



Adding 4th Variable to the Model – `Num_yrs_since_last_at_fault_claim_num_cap_13`

The residual plots below show an obvious decreasing pattern. This means that there is a significant relationship between **`Num_yrs_since_last_at_fault_claim`** and loss cost.

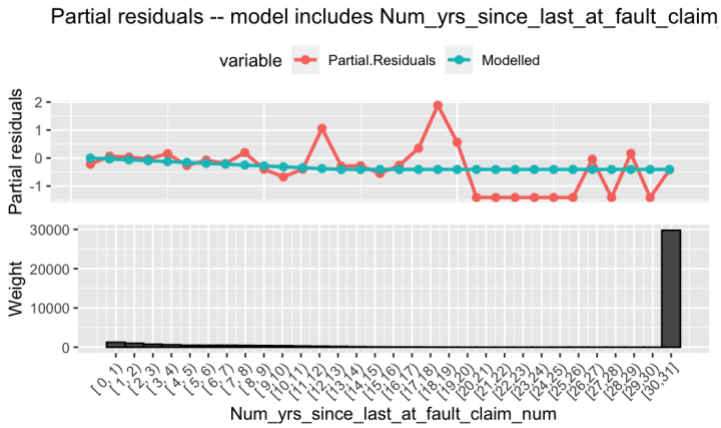


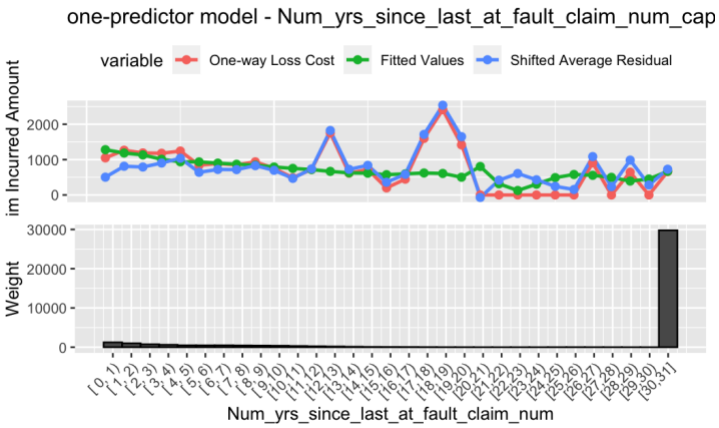


Since there is very little exposure after 13 and many variability, a cap of 13 is set.

Model:
`LC_1q5_2o_3g_4i = Loss_cost ~ Accident_year + Vehicle_age_cap_30_floor_6 + Years_driving + Years_driving_squared + Num_minor_convictions_cap_5 + Num_yrs_since_last_at_fault_claim_num_cap_13`

From the residual plots below, the model captures the residual signals well.





Training CV Gini

LC_1q5_2o_3g	LC_1q5_2o_3g_4i
0.34322	0.34779

Since the training CV Gini increases, the longer model is preferred.

Reasonability

The longer it's been since the driver has a claim that is his fault, it means that the driver tends to drive more carefully, resulting to a lower lost cost.

Parsimony

AIC

LC_1q5_2o_3g	LC_1q5_2o_3g_4i
84,328	84,315

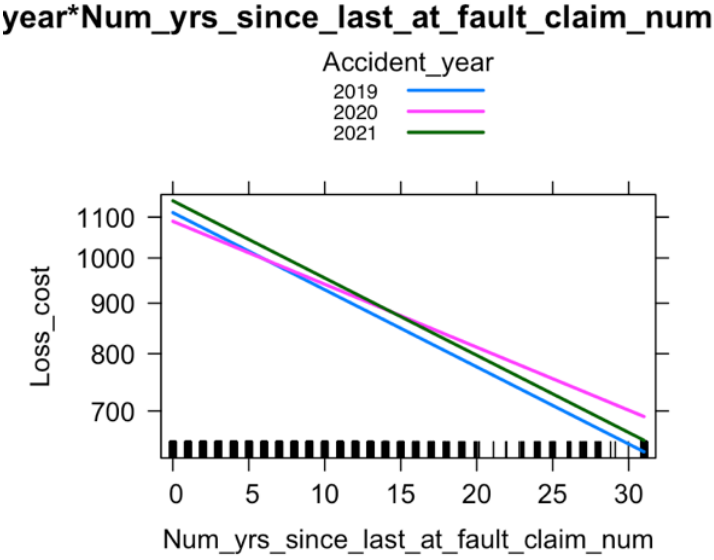
The AIC decreases. Thus, the longer model is preferred.

F-Test
p-value = 0.0001298 < 0.05
This means that the longer model is preferred.

Significance
The variable is significant.

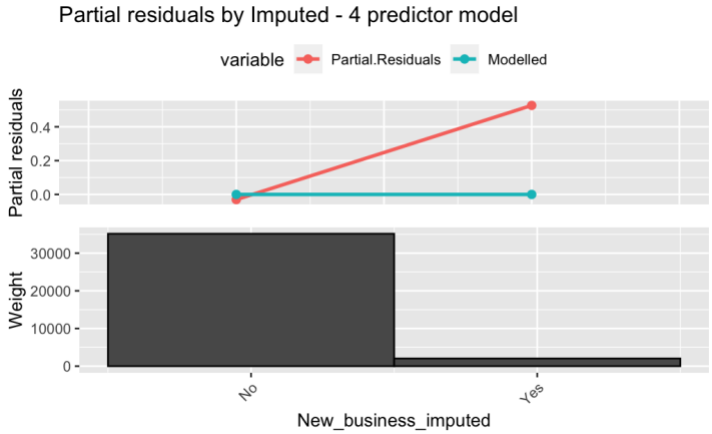
```
Call:
glm(formula = Loss_cost ~ Accident_year + Vehicle_age_cap_30_floor_6 +
    Years_driving + Years_driving_squared + Num_minor_convictions_cap_5 +
    Num_yrs_since_last_at_fault_claim_num_cap_13, family = tweedie(var.power = 1.5,
    link.power = 0), data = datLC_train, weights = Collision_earned_count)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-15.979   -8.159   -6.197   -4.143  125.515

Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                              8.8481833  0.1714224  51.616  < 2e-16 ***
Accident_year2020                        0.1084175  0.0806722   1.344 0.178976
Accident_year2021                        0.1647264  0.0826130   1.994 0.046161 *
Vehicle_age_cap_30_floor_6              -0.1343653  0.0090026 -14.925  < 2e-16 ***
Years_driving                           -0.0471264  0.0084369  -5.586 2.33e-08 ***
Years_driving_squared                    0.0005418  0.0001365   3.970 7.21e-05 ***
Num_minor_convictions_cap_5              0.1009722  0.0232445   4.344 1.40e-05 ***
Num_yrs_since_last_at_fault_claim_num_cap_13 -0.0313673  0.0081473  -3.850 0.000118 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Tweedie family taken to be 1072.977)

    Null deviance: 6474677  on 79788  degrees of freedom
Residual deviance: 6065737  on 79781  degrees of freedom
AIC: NA
```

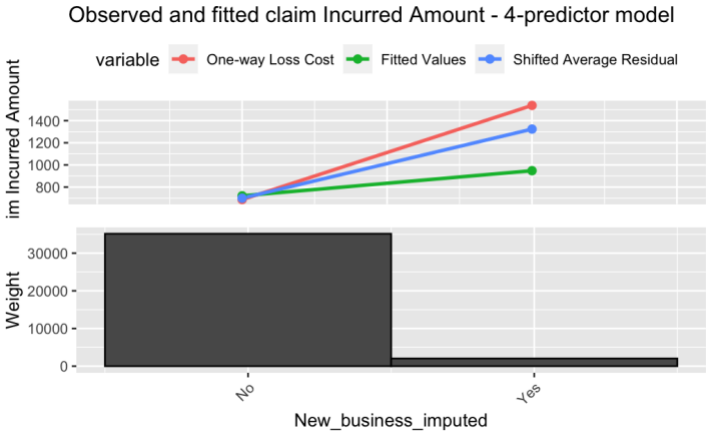
Time-Consistency
From the graph below, although the three lines intersect, the gradients don’t differ by much. Thus, it still passes the time-consistency test.



Adding 5th Variable to the Model – `New_policy`

The residual plots below show an obvious increasing pattern. This means that there is a relationship between **`New_policy_imputed`** and loss cost.





Model:
`LC_1q5_2o_3g_4i_5m`
= `Loss_cost` ~ `Accident_year` + `Vehicle_age_cap_30_floor_6` + `Years_driving` + `Years_driving_squared` + `Num_minor_convictions_cap_5` + `Num_yrs_since_last_at_fault_claim_num_cap_13` + `New_policy`

From the residual plots below, the model captures the residual signals well.

Training CV Gini

`LC_1q5_2o_3g_4i`	`LC_1q5_2o_3g_4i_5m`
0.34779	0.35657

Since the training CV Gini increases, the longer model is preferred.

Reasonability
New business is divided into two – new drivers who just join and experienced drivers who moved from another insurance company. New drivers tend to be riskier than experienced drivers since they have little driving experience and there might be a reason why a driver switches insurance company, which includes a cheaper premium from Rightprice, since other companies see the risk that the driver has that Rightprice doesn't.

Parsimony
AIC

LC_1q5_2o_3g_4i	LC_1q5_2o_3g_4i_5m
84,315	84,301

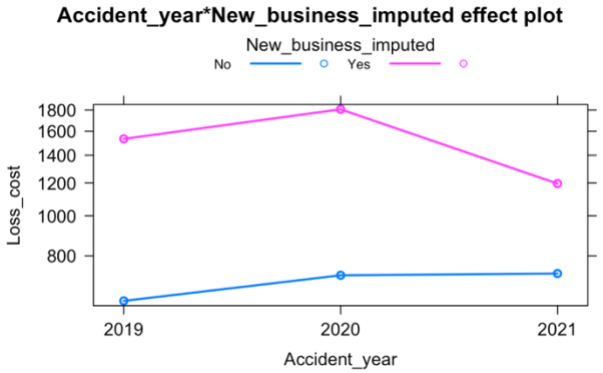
The AIC decreases. Thus, the longer model is preferred.

F-Test
p-value = 1.747e-0.5 < 0.05
This means that the longer model is preferred.

Significance
The variable is significant.

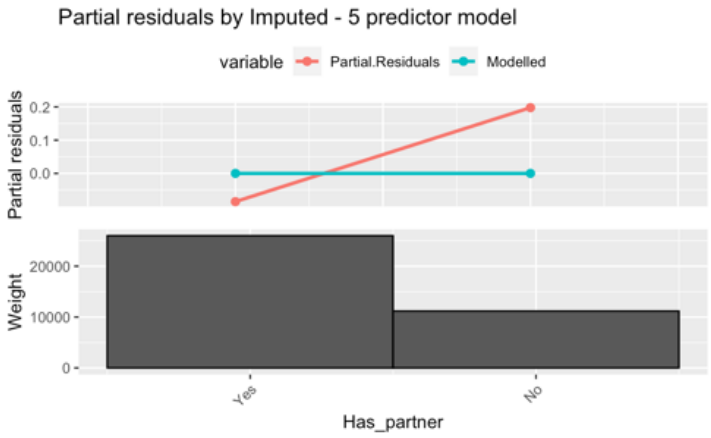
```
Call:
glm(formula = Loss_cost ~ Accident_year + Vehicle_age_cap_30_floor_6 +
    Years_driving + Years_driving_squared + Num_minor_convictions_cap_5 +
    Num_yrs_since_last_at_fault_claim_num_cap_13 + New_business_imputed,
    family = tweedie(var.power = 1.5, link.power = 0), data = datLC_train,
    weights = Collision_earned_count)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-16.220   -8.134   -6.189   -4.143  129.146

Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                              8.6597959  0.1728037  50.113  < 2e-16 ***
Accident_year2020                        0.1482944  0.0798087   1.858 0.063155 .
Accident_year2021                        0.2122581  0.0817303   2.597 0.009404 **
Vehicle_age_cap_30_floor_6              -0.1314562  0.0088288 -14.889  < 2e-16 ***
Years_driving                           -0.0418989  0.0083720  -5.005 5.61e-07 ***
Years_driving_squared                    0.0004859  0.0001347   3.606 0.000311 ***
Num_minor_convictions_cap_5              0.1065552  0.0228223   4.669 3.03e-06 ***
Num_yrs_since_last_at_fault_claim_num_cap_13 -0.0320420  0.0079940  -4.008 6.12e-05 ***
New_business_imputedYes                  0.5429058  0.1260904   4.306 1.67e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Tweedie family taken to be 1031.806)

    Null deviance: 6474677  on 79788  degrees of freedom
Residual deviance: 6046701  on 79780  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 7
```

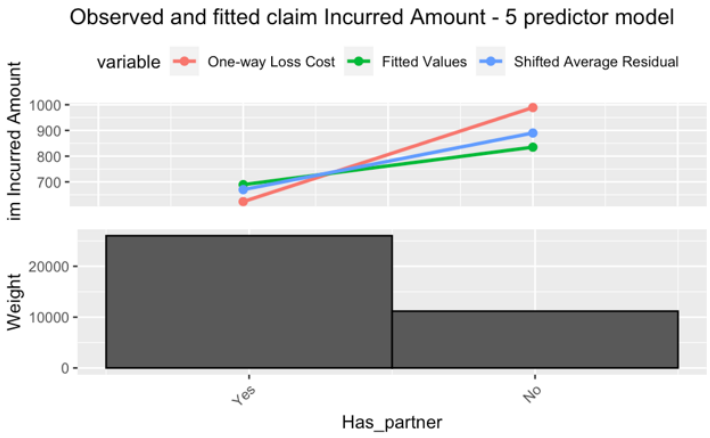
Time-Consistency
‘Yes’ always stays on top of ‘No’ in all three years, which means that the time-consistency test is passed.



Adding 6th Variable to the Model – `Has_partner`

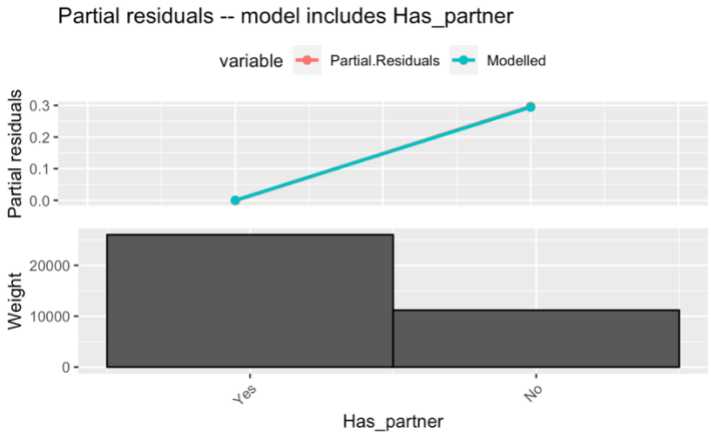
The residual plots below show an obvious increasing pattern. This means that there is a relationship between **`Has_partner`** and loss cost.





Model:
`LC_1q5_2o_4i_5m_6j`
= `Loss_cost` ~ `Accident_year` + `Vehicle_age_cap_30_floor_6` + `Years_driving` + `Years_driving_squared` + `Num_minor_convictions_cap_5` + `Num_yrs_since_last_at_fault_claim_num_cap_13` + `New_policy` + `Has_partner`

From the residual plots below, the model captures the residual signals well.



Training CV Gini

`LC_1q5_2o_3g_4i_5m`	`LC_1q5_2o_3g_4i_5m_6j`
0.35657	0.36078

Since the training CV Gini increases, the longer model is preferred.

Reasonability

When people have a partner, they tend to go out less and participate in fewer dangerous activities, like speeding and drunk driving. Thus, they are less risky compared to people who live alone (i.e. who have no partner).

Parsimony

AIC

`LC_1q5_2o_3g_4i_5m`	`LC_1q5_2o_3g_4i_5m_6j`
84,301	84,289

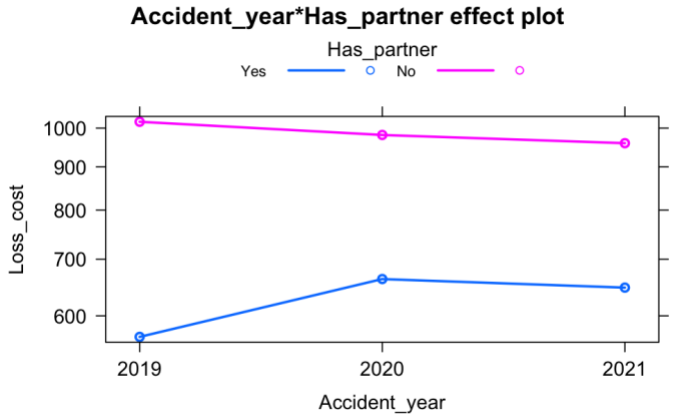
The AIC decreases. Thus, the longer model is preferred.

F-Test
p-value = 5.448e-0.5 < 0.05
This means that the longer model is preferred.

Significance
The variable is significant.

```
Call:
glm(formula = Loss_cost ~ Accident_year + Vehicle_age_cap_30_floor_6 +
    +Years_driving + Years_driving_squared + Num_minor_convictions_cap_5 +
    Num_yrs_since_last_at_fault_claim_num_cap_13 + New_business_imputed +
    Has_partner, family = tweedie(var.power = 1.5, link.power = 0),
    data = datLC_train, weights = Collision_earned_count)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-16.435   -8.113   -6.170   -4.132  136.618

Coefficients:
                                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                                 8.4293058  0.1820629  46.299  < 2e-16 ***
Accident_year2020                           0.1504872  0.0797666   1.887 0.059218 .
Accident_year2021                           0.2107864  0.0817054   2.580 0.009887 **
Vehicle_age_cap_30_floor_6                 -0.1333890  0.0088320 -15.103  < 2e-16 ***
Years_driving                              -0.0355645  0.0085412  -4.164 3.13e-05 ***
Years_driving_squared                       0.0004220  0.0001359   3.105 0.001907 **
Num_minor_convictions_cap_5                 0.1062970  0.0228360   4.655 3.25e-06 ***
Num_yrs_since_last_at_fault_claim_num_cap_13 -0.0286920  0.0080279  -3.574 0.000352 ***
New_business_imputedYes                     0.5059012  0.1265399   3.998 6.39e-05 ***
Has_partnerNo                               0.2944596  0.0724774   4.063 4.85e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Tweedie family taken to be 1028.507)

    Null deviance: 6474677  on 79788  degrees of freedom
Residual deviance: 6029950  on 79779  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 7
```

Time-Consistency
‘No’ always stays on top of ‘Yes’ in all three years, which means that the time-consistency test is passed.



Appendix B.4: Tweedie Summary of Progression

#	Variable	Training CV Gini	AIC
1	Accident_year	-0.01800	84,591
2	+ Vehicle_age_cap_30_floor_6	0.28842	84,387
3	+ Years_driving + Years_driving_squared	0.32866	84,346
4	+ Num_minor_convictions_cap_5	0.34322	84,328
5	+ Num_yrs_since_last_at_fault_claim_num_cap_13	0.34779	84,315
6	+ New_policy	0.35657	84,301
7	+ Has_partner	0.36078	84,289

Appendix B.5: Interaction p-values

	Vehicle_age	Years_driving	Num_minor	Num_yrs_since	New_policy	Has_partner
Vehicle_age		0.5851	0.5010	0.8940	0.1768	0.6961
Years_driving			0.7390	0.5551	0.6161	0.0336
Num_minor				0.4392	0.6060	0.9760
Num_yrs_since					0.7632	0.2101
New_policy						0.0437
Has_partner						

Appendix B.6: Holdout Testing for Tweedie Model

Models

Model Name	Variables
LC_1q5_2o	Accident_year + Vehicle_age + Years_driving + Years_driving_squared
LC_1q5_2o_3g	Accident_year + Vehicle_age + Years_driving + Years_driving_squared + Num_minor_convictions_cap_5
LC_1q5_2o_3g_4i	Accident_year + Vehicle_age + Years_driving + Years_driving_squared + Num_minor_convictions_cap_5 + Num_years_since_last_at_fault_claim_num_cap_13
LC_1q5_2o_3g_4j	Accident_year + Vehicle_age + Years_driving + Years_driving_squared + Num_minor_convictions_cap_5 + Has_partner
LC_1q5_2o_3g_4m	Accident_year + Vehicle_age + Years_driving + Years_driving_squared + Num_minor_convictions_cap_5 + New_policy
LC_1q5_2o_3g_4i_5j	Accident_year + Vehicle_age + Years_driving + Years_driving_squared + Num_minor_convictions_cap_5 + Num_years_since_last_at_fault_claim_num_cap_13 + Has_partner
LC_1q5_2o_3g_4i_5m	Accident_year + Vehicle_age + Years_driving + Years_driving_squared + Num_minor_convictions_cap_5 + Num_years_since_last_at_fault_claim_num_cap_13 + New_policy
LC_1q5_2o_3g_4i_5m_6j	Accident_year + Vehicle_age + Years_driving + Years_driving_squared + Num_minor_convictions_cap_5 + Num_years_since_last_at_fault_claim_num_cap_13 + New_policy + Has_partner

Gini

Holdout Gini Result

LC_1q5_2o	LC_1q5_2o_3g	LC_1q5_2o_3g_4i	LC_1q5_2o_3g_4j	LC_1q5_2o_3g_4m
0.3251347	0.3364621	0.3546620	0.3569521	0.3453938
LC_1q5_2o_3g_4i_5j	LC_1q5_2o_3g_4i_5m	LC_1q5_2o_3g_4i_5m_6j		
0.3706503	0.3633376	0.3766492		

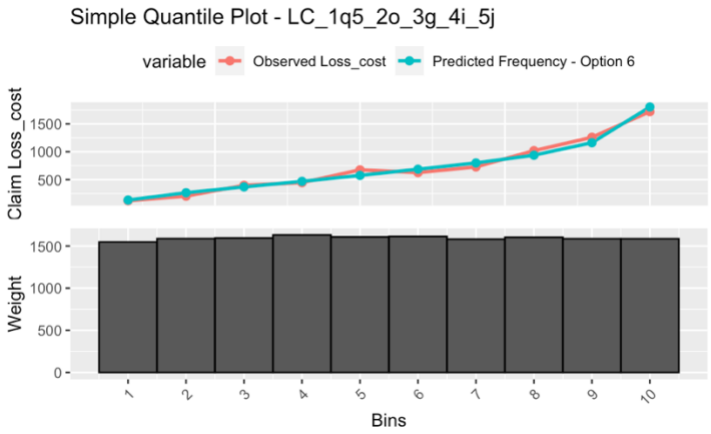
Training Gini Result

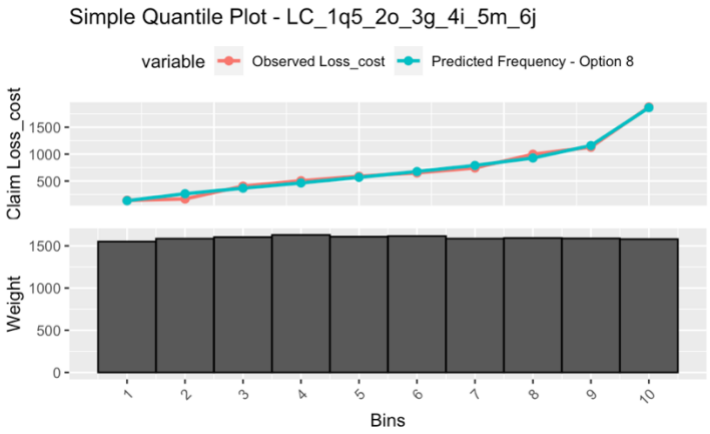
LC_1q5_2o	LC_1q5_2o_3g	LC_1q5_2o_3g_4i	LC_1q5_2o_3g_4j	LC_1q5_2o_3g_4m
0.3325644	0.3476488	0.3529855	0.3570960	0.3564412
LC_1q5_2o_3g_4i_5j	LC_1q5_2o_3g_4i_5m	LC_1q5_2o_3g_4i_5m_6j		
0.3609183	0.3620305	0.3669061		

The Gini coefficients for both training and holdout datasets increases as we add more variables into our model. This indicates that there is no overfitting for **LC_1q5_2o_3g_4i_5m_6j**.

The two best holdout Ginis belong to models **LC_1q5_2o_3g_4i_5j** & **LC_1q5_2o_3g_4i_5m_6j**. Thus, we want to further evaluate these two models.

Simple Quantile Plots

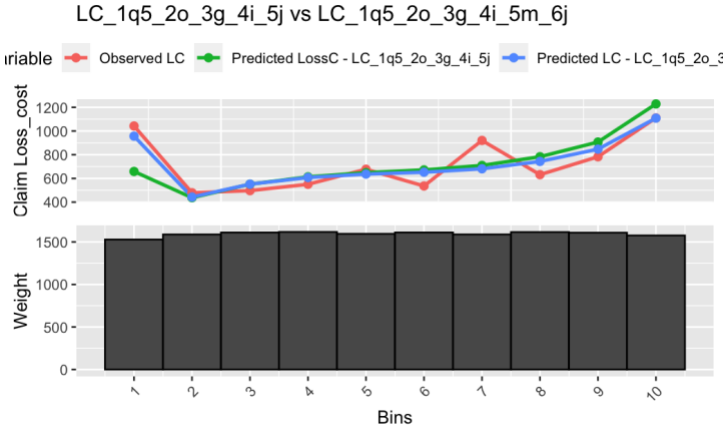




LC_1q5_2o_3g_4i_5j	LC_1q5_2o_3g_4i_5m_6j
Fit: Deviations in bin 5, bin 6 to 7 and bin 8 to 9.	**Fit:** Deviations in bin 2.
Lift: Less similar fit to observed.	**Lift:** Similar lift to observed.
Monotonicity: Reversal in bin 6.	**Monotonicity:** No reversal.

Based on the Simple Quantile Plots, model **LC_1q5_2o_3g_4i_5m_6j** performs better compared to model **LC_1q5_2o_3g_4i_5j**.

Double Lift Charts



- **LC_1q5_2o_3g_4i_5m_6j** performs better at the end points.
- Both models perform similarly in the middle bins.
- **LC_1q5_2o_3g_4i_5m_6j** is closer to observed in 8 bins whilst **LC_1q5_2o_3g_4i_5j** in only 2 bins.

Thus, based on the Double Lift Charts, model **LC_1q5_2o_3g_4i_5m_6j** performs better compared to model **LC_1q5_2o_3g_4i_5j**.

Appendix C.1: Severity Variate Creation and Capping

```
##       quantile num_claims_above
## 90%      16708              400
## 95%      24852              200
## 97%      31062              120
## 98%      37538               80
## 98.5%    43338               60
## 99%      48898               40
## 99.5%    59319               20
## 99.8%    81109                8
## 99.9%    95932                4
## 100%    153603                0
```

Appendix C.2: Creating the Base Model

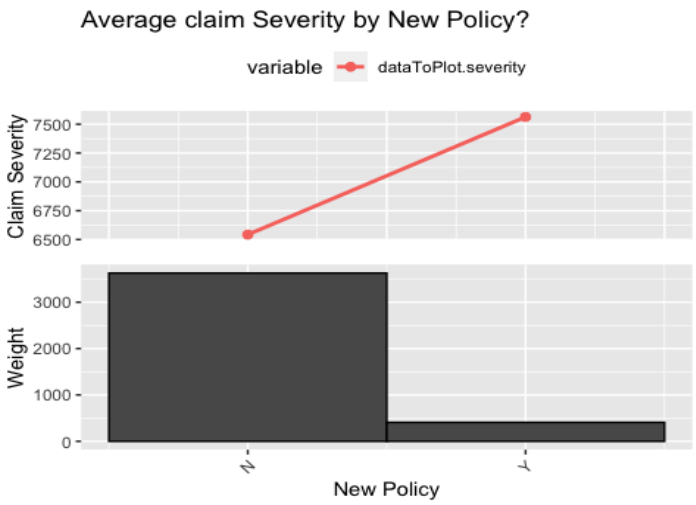
```
##
## Call:
## glm(formula = severity ~ AY_factor, family = Gamma(link = "log"),
##     data = coll_dataset_claims, weights = Collision_claim_count,
##     subset = (partition == "Training"), na.action = "na.pass",
##     x = TRUE)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.9781  -1.5575  -0.6334   0.2499   3.7760
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    8.64136    0.03404 253.859  < 2e-16 ***
## AY_factor2020  0.27128    0.05272   5.145 2.80e-07 ***
## AY_factor2021  0.25424    0.05414   4.696 2.74e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 1.97446)
##
##     Null deviance: 7725.1  on 3999  degrees of freedom
## Residual deviance: 7657.8  on 3997  degrees of freedom
## AIC: 78551
##
## Number of Fisher Scoring iterations: 7
```

Appendix C.2.1: CV Testing for Base Model

```
## [1] 0.08766285
```

Appendix C.3: Adding new_policy

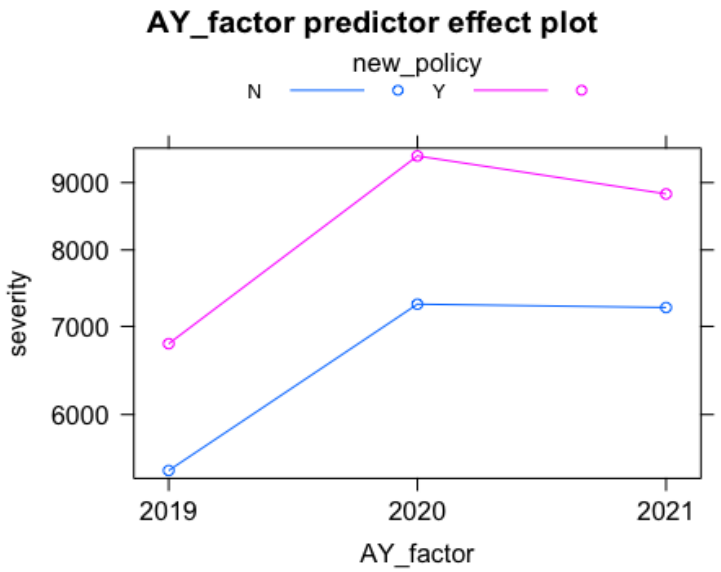
Appendix C.3.1: One-way analysis for new_policy



Appendix C.3.2: Significance Test for new_policy

```
##
## Call:
## glm(formula = severity ~ AY_factor + new_policy, family = Gamma(link = "log"),
##     data = coll_dataset_claims, weights = Collision_claim_count,
##     subset = (partition == "Training"), na.action = "na.pass",
##     x = TRUE)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.9736  -1.5574  -0.6299   0.2592   3.8788
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    8.60077    0.03574 240.675  < 2e-16 ***
## AY_factor2020  0.29391    0.05258   5.590 2.43e-08 ***
## AY_factor2021  0.28464    0.05427   5.245 1.64e-07 ***
## new_policyY    0.22722    0.07373   3.082  0.00207 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 1.929562)
##
##     Null deviance: 7725.1  on 3999  degrees of freedom
## Residual deviance: 7638.4  on 3996  degrees of freedom
## AIC: 78540
##
## Number of Fisher Scoring iterations: 7
```

Appendix C.3.3: Time Consistency for new_policy



Appendix C.3.4: F-Test for new_policy

```
## Analysis of Deviance Table
##
## Model 1: severity ~ AY_factor + new_policy
## Model 2: severity ~ AY_factor
##   Resid. Df Resid. Dev Df Deviance      F   Pr(>F)
## 1      3996     7638.4
## 2      3997     7657.8 -1  -19.398 10.053 0.001533 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Appendix C.3.5: Parsimony for new_policy

```
## [1] "AIC"

## [1] 78551.26 78576.44

## [1] "BIC"

## [1] 78540.37 78571.84
```

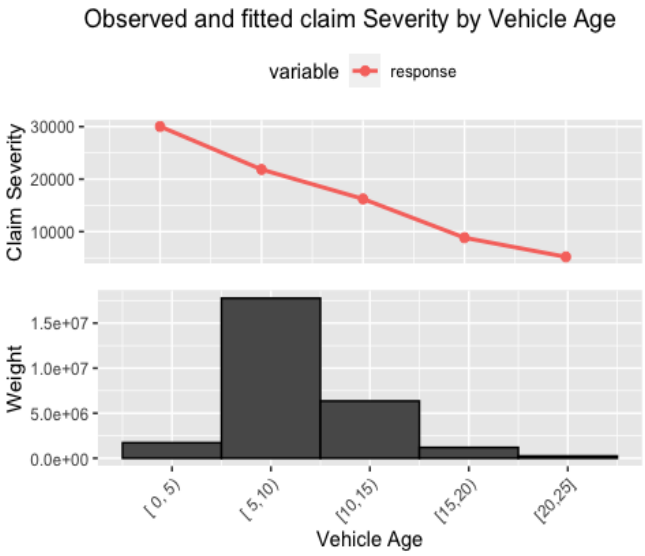
Appendix C.3.6: CV for new_policy

```
## [1] 0.1041012
```

Appendix C.4: Adding vehicle_age

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.289   6.807   8.775   9.796  11.862  34.220

##      80%      85%      90%      95%      96%      97%      98%      99%
## 12.81750 13.82092 15.33671 17.54258 18.28615 19.42699 21.37280 23.19114
##     100%
## 34.22010
```

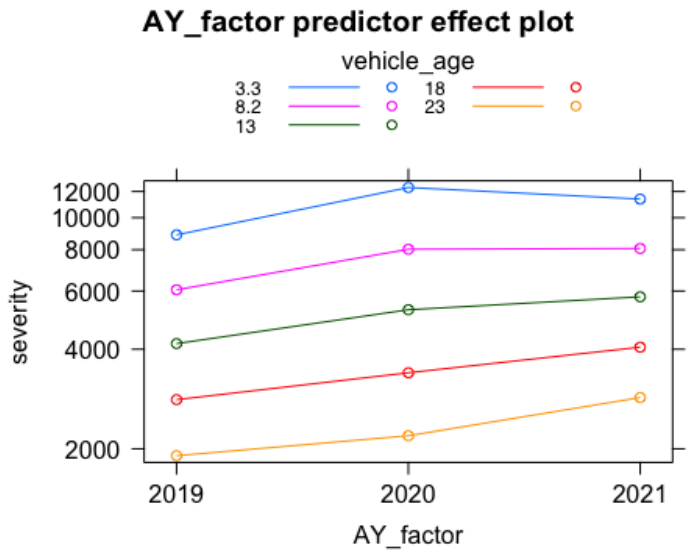
Appendix C.4.1: One-way Analysis for vehicle_age



Appendix C.4.2: Significance for vehicle_age

```
##
## Call:
## glm(formula = severity ~ AY_factor + vehicle_age, family = Gamma(link = "log"),
##     data = coll_dataset_claims, weights = Collision_claim_count,
##     subset = (partition == "Training"), na.action = "na.pass",
##     x = TRUE)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.0571  -1.5045  -0.5791   0.2941   4.0337
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    9.351604   0.059398  157.44  < 2e-16 ***
## AY_factor2020  0.268653   0.049563    5.42 6.30e-08 ***
## AY_factor2021  0.303934   0.050991    5.96 2.73e-09 ***
## vehicle_age   -0.078418   0.005242  -14.96  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 1.744564)
##
##     Null deviance: 7725.1  on 3999  degrees of freedom
## Residual deviance: 7323.5  on 3996  degrees of freedom
## AIC: 78327
##
## Number of Fisher Scoring iterations: 7
```

Appendix C.4.3: Time Consistency for vehicle_age



Appendix C.4.4: Parsimony for vehicle_age

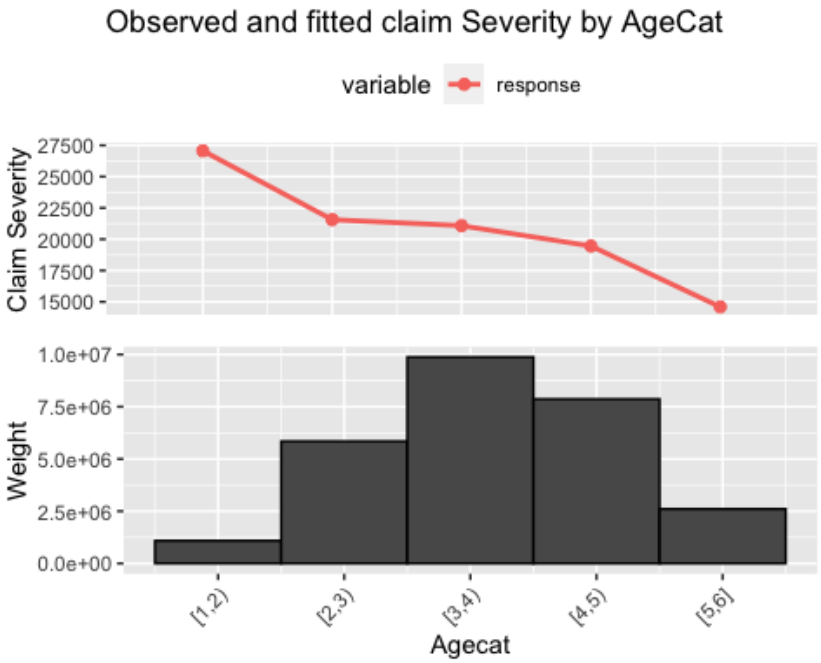
```
## [1] "AIC"

## [1] 78540.37 78571.84

## [1] "BIC"

## [1] 78326.97 78358.44
```

Appendix C.4.5: CV for vehicle_age

```
## [1] 0.2609556
```

Appendix C.5: Adding agecat

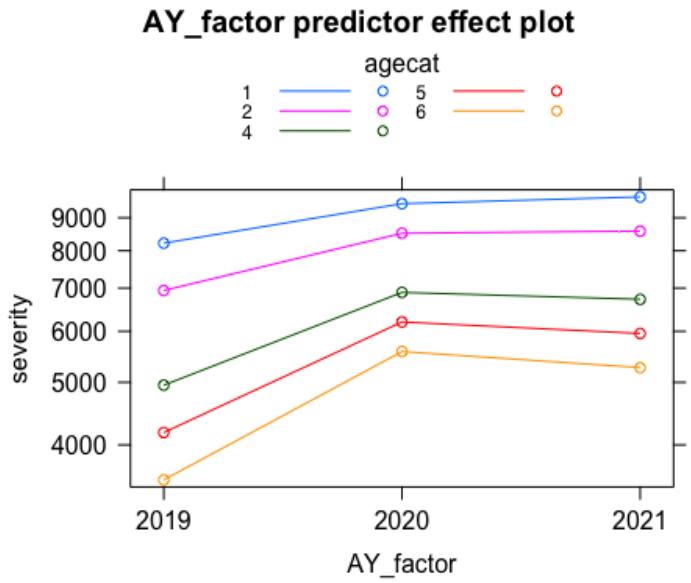
Appendix C.5.1: One-way Analysis for agecat



Appendix C.5.2: Significance Test for agecat

```
##
## Call:
## glm(formula = severity ~ AY_factor + vehicle_age + agecat, family = Gamma(link = "log"),
##     data = coll_dataset_claims, weights = Collision_claim_count,
##     subset = (partition == "Training"), na.action = "na.pass",
##     x = TRUE)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.0627  -1.4998  -0.5700   0.2898   3.9675
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    9.687387   0.083982 115.351  < 2e-16 ***
## AY_factor2020  0.283343   0.049072   5.774 8.33e-09 ***
## AY_factor2021  0.318336   0.050492   6.305 3.20e-10 ***
## vehicle_age   -0.075225   0.005223 -14.402  < 2e-16 ***
## agecat        -0.114468   0.019883  -5.757 9.20e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 1.709384)
##
##     Null deviance: 7725.1  on 3999  degrees of freedom
## Residual deviance: 7265.2  on 3995  degrees of freedom
## AIC: 78289
##
## Number of Fisher Scoring iterations: 7
```

Appendix C.5.3: Time Consistency for agecat



Appendix C.5.4: F-Test for agecat

```
## Analysis of Deviance Table
##
## Model 1: severity ~ AY_factor + vehicle_age
## Model 2: severity ~ AY_factor + vehicle_age + agecat
##   Resid. Df Resid. Dev Df Deviance      F    Pr(>F)
## 1      3996     7323.5
## 2      3995     7265.2  1   58.282 34.095 5.665e-09 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Appendix C.5.5: Parsimony for agecat

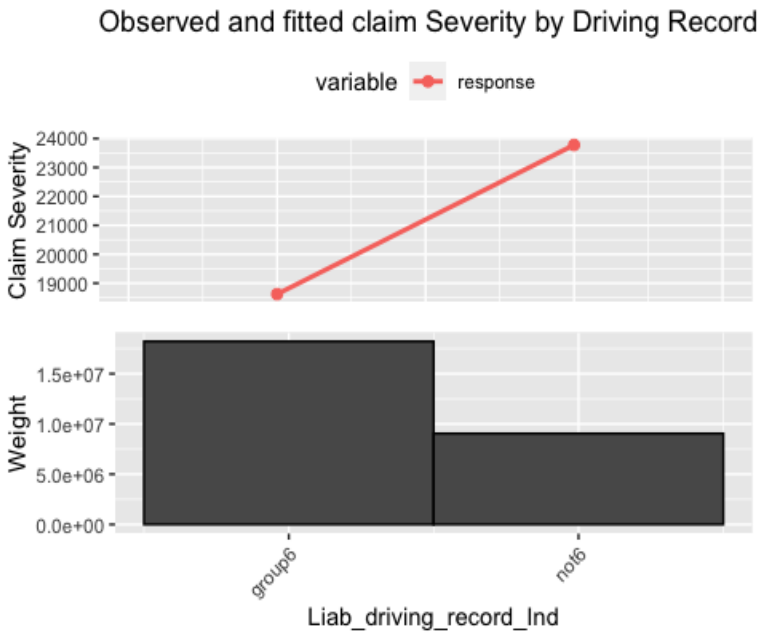
```
## [1] "AIC"

## [1] 78326.97 78358.44

## [1] "BIC"

## [1] 78288.64 78326.40
```

Appendix C.5.6: CV for agecat

```
## [1] 0.281299
```

Appendix C.6: Adding Liab_driving_record_Ind

Appendix C.6.1: One-way Analysis for Liab_driving_record_Ind

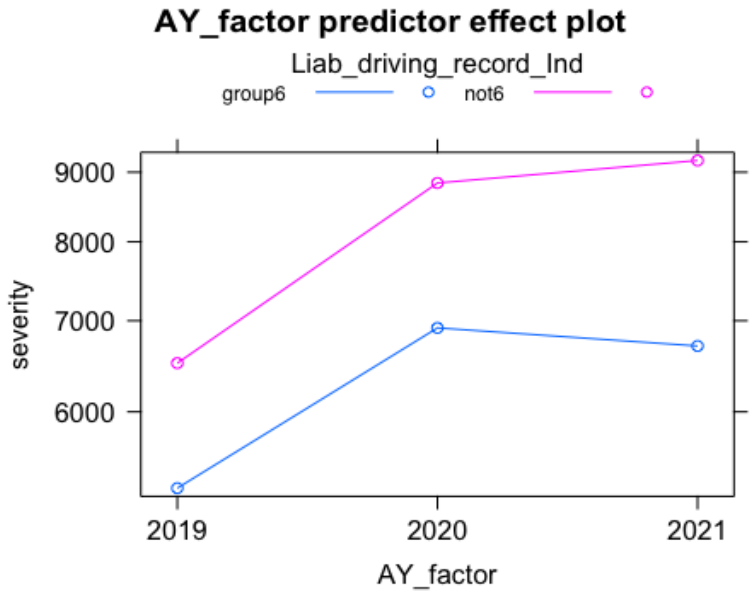


Appendix C.6.2: Significance Test for Liab_driving_record_Ind

```
##
## Call:
## glm(formula = severity ~ AY_factor + vehicle_age + agecat + Liab_driving_record_Ind,
##     family = Gamma(link = "log"), data = coll_dataset_claims,
##     weights = Collision_claim_count, subset = (partition == "Training"),
##     na.action = "na.pass", x = TRUE)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.0504  -1.4928  -0.5640   0.2883   4.1026
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  9.582173   0.089564 106.987  < 2e-16 ***
## AY_factor2020                0.289801   0.049041   5.909 3.72e-09 ***
## AY_factor2021                0.325381   0.050509   6.442 1.32e-10 ***
## vehicle_age                 -0.074279   0.005216 -14.240  < 2e-16 ***
## agecat                      -0.099819   0.020432  -4.886 1.07e-06 ***
## Liab_driving_record_Indnot6  0.147667   0.047357   3.118  0.00183 **
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 1.704545)
##
##     Null deviance: 7725.1  on 3999  degrees of freedom
## Residual deviance: 7248.3  on 3994  degrees of freedom
## AIC: 78279
##
## Number of Fisher Scoring iterations: 7
```

Appendix C.6.3: Time Consistency for Liab_driving_record_Ind



Appendix C.6.4: F-Test for Liab_driving_record_Ind

```
## Analysis of Deviance Table
##
## Model 1: severity ~ AY_factor + vehicle_age + agecat
## Model 2: severity ~ AY_factor + vehicle_age + agecat + Liab_driving_record_Ind
##   Resid. Df Resid. Dev Df Deviance      F   Pr(>F)
## 1      3995     7265.2
## 2      3994     7248.3  1   16.889 9.9084 0.001657 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Appendix C.6.5: Parsimony for Liab_driving_record_Ind

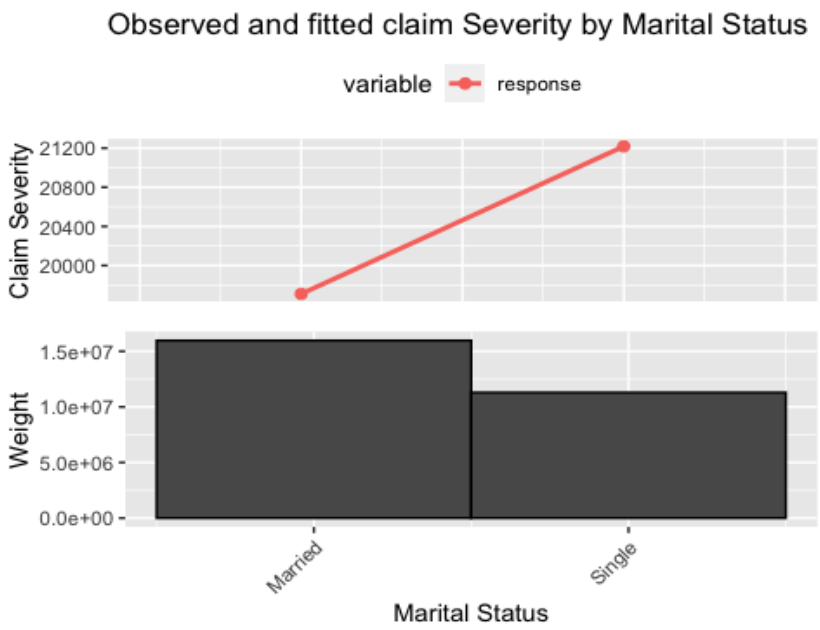
```
## [1] "AIC"

## [1] 78288.64 78326.40

## [1] "BIC"

## [1] 78278.90 78322.96
```

Appendix C.6.6: CV for Liab_driving_record_Ind

```
## [1] 0.2876676
```

Appendix C.7: Adding Marital_status_Ind

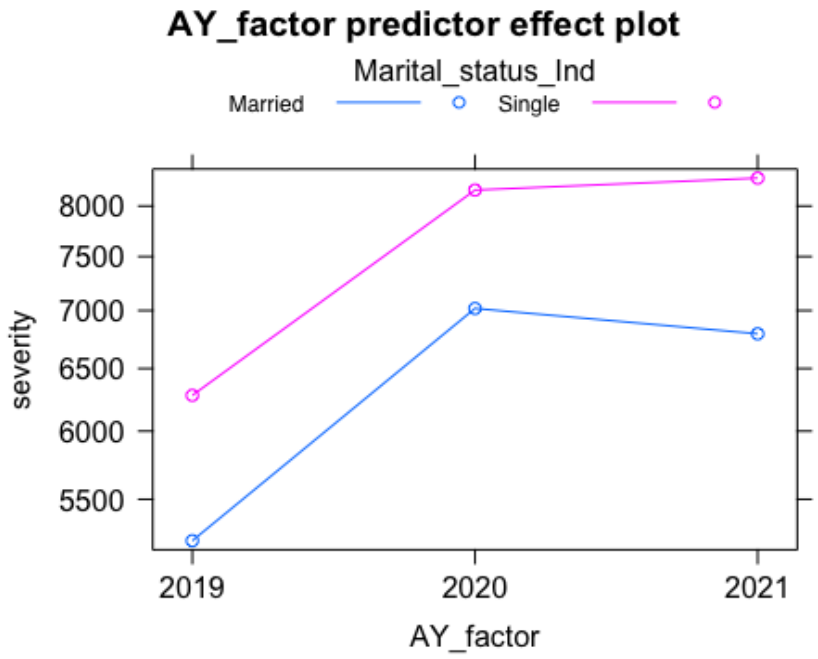
Appendix C.7.1: One-way Analysis for Marital_status_Ind



Appendix C.7.2: Significance Test for Analysis for Marital_status_Ind

```
##
## Call:
## glm(formula = severity ~ AY_factor + vehicle_age + agecat + Liab_driving_record_Ind +
##     Marital_status_Ind, family = Gamma(link = "log"), data = coll_dataset_claims,
##     weights = Collision_claim_count, subset = (partition == "Training"),
##     na.action = "na.pass", x = TRUE)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.0424  -1.4859  -0.5766   0.2936   3.9451
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  9.50939    0.09462 100.497  < 2e-16 ***
## AY_factor2020                0.29502    0.04912   6.006 2.07e-09 ***
## AY_factor2021                0.32969    0.05062   6.513 8.26e-11 ***
## vehicle_age                 -0.07515    0.00523 -14.369  < 2e-16 ***
## agecat                      -0.08668    0.02109  -4.111 4.03e-05 ***
## Liab_driving_record_Indnot6  0.12931    0.04791   2.699  0.00699 **
## Marital_status_IndSingle     0.10251    0.04486   2.285  0.02235 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 1.706467)
##
##     Null deviance: 7725.1  on 3999  degrees of freedom
## Residual deviance: 7239.6  on 3993  degrees of freedom
## AIC: 78275
##
## Number of Fisher Scoring iterations: 7
```

Appendix C.7.3: Time Consistency for Marital_status_Ind



Appendix C.7.4: F-Test for Marital_status_Ind

```
## Analysis of Deviance Table
##
## Model 1: severity ~ AY_factor + vehicle_age + agecat + Liab_driving_record_Ind
## Model 2: severity ~ AY_factor + vehicle_age + agecat + Liab_driving_record_Ind +
##     Marital_status_Ind
##   Resid. Df Resid. Dev Df Deviance      F  Pr(>F)
## 1      3994     7248.3
## 2      3993     7239.6  1   8.7387 5.1209 0.02369 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Appendix C.7.5: Parsimony for Marital_status_Ind

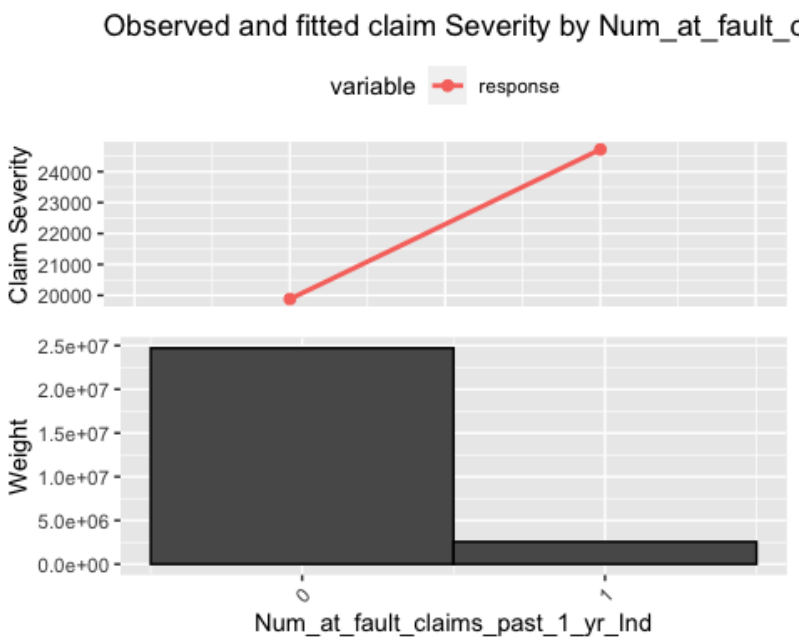
```
## [1] "AIC"

## [1] 78278.90 78322.96

## [1] "BIC"

## [1] 78274.81 78325.17
```

Appendix C.7.6: CV for Marital_status_Ind

```
## [1] 0.290259
```

Appendix C.8: Adding Num_at_fault_claims_past_1_yr_Ind

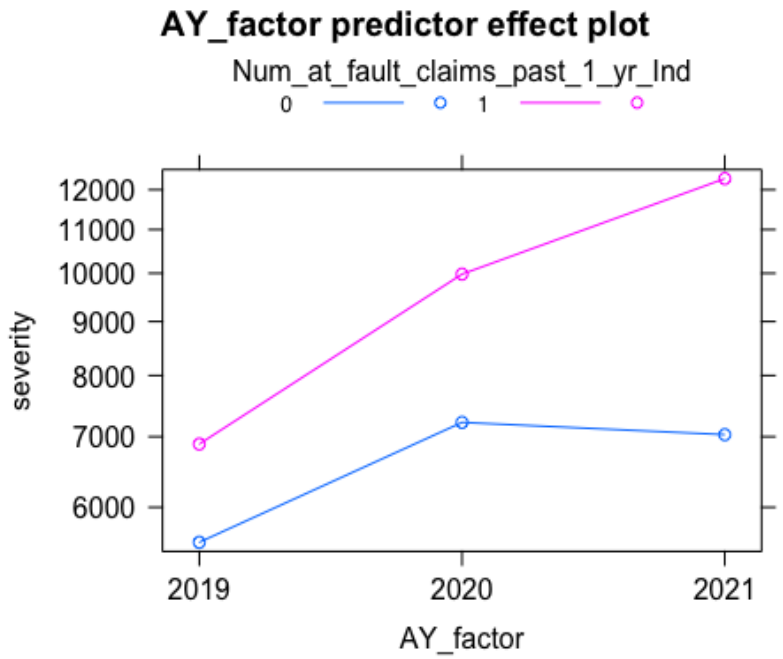
Appendix C.8.1: One-way Analysis for Num_at_fault_claims_past_1_yr_Ind



Appendix C.8.2: Significance Test for Num_at_fault_claims_past_1_yr_Ind

```
##
## Call:
## glm(formula = severity ~ AY_factor + vehicle_age + agecat + +Marital_status_Ind +
##     Liab_driving_record_Ind + Num_at_fault_claims_past_1_yr_Ind,
##     family = Gamma(link = "log"), data = coll_dataset_claims,
##     weights = Collision_claim_count, subset = (partition == "Training"),
##     na.action = "na.pass", x = TRUE)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.0468  -1.4910  -0.5777   0.2958   3.9562
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      9.513184   0.094668 100.490  < 2e-16 ***
## AY_factor2020                    0.292676   0.049146   5.955 2.82e-09 ***
## AY_factor2021                    0.328568   0.050649   6.487 9.82e-11 ***
## vehicle_age                     -0.075033   0.005234 -14.335  < 2e-16 ***
## agecat                          -0.088042   0.021111  -4.170 3.10e-05 ***
## Marital_status_IndSingle         0.097314   0.044879   2.168   0.0302 *
## Liab_driving_record_Indnot6      0.085846   0.051073   1.681   0.0929 .
## Num_at_fault_claims_past_1_yr_Ind1  0.196954   0.086944   2.265   0.0235 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 1.70804)
##
##     Null deviance: 7725.1  on 3999  degrees of freedom
## Residual deviance: 7230.5  on 3992  degrees of freedom
## AIC: 78271
##
## Number of Fisher Scoring iterations: 7
```

Appendix C.8.3: Time Consistency for Num_at_fault_claims_past_1_yr_Ind



Appendix C.8.4: F-Test for Num_at_fault_claims_past_1_yr_Ind

```
#F-Test
anova(glm6,glm7,test="F") #passed

## Analysis of Deviance Table
##
## Model 1: severity ~ AY_factor + vehicle_age + agecat + Liab_driving_record_Ind +
##     Marital_status_Ind
## Model 2: severity ~ AY_factor + vehicle_age + agecat + +Marital_status_Ind +
##     Liab_driving_record_Ind + Num_at_fault_claims_past_1_yr_Ind
##   Resid. Df Resid. Dev Df Deviance      F  Pr(>F)
## 1      3993     7239.6
## 2      3992     7230.5  1   9.0266 5.2848 0.02156 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Appendix C.8.5: Parismony for Num_at_fault_claims_past_1_yr_Ind

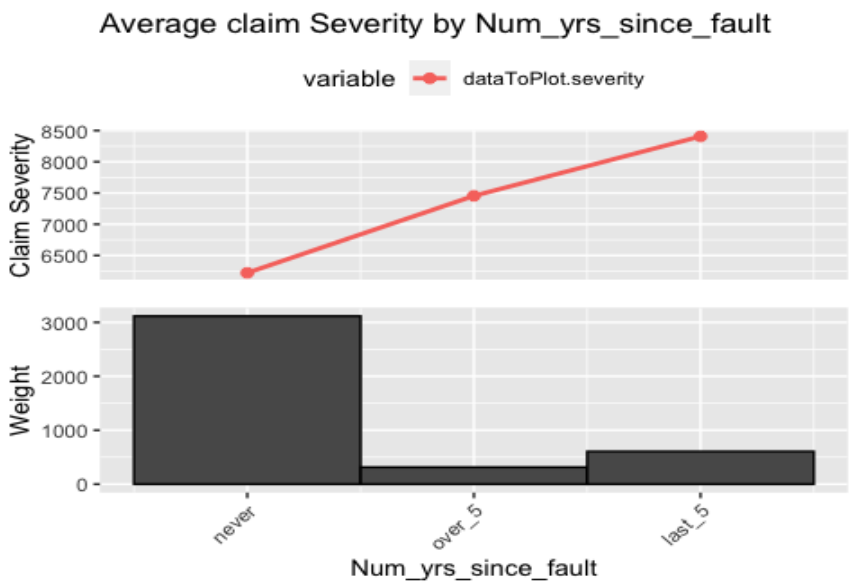
```
## [1] "AIC"

## [1] 78274.81 78325.17

## [1] "BIC"

## [1] 78270.52 78327.17
```

Appendix C.8.6: CV for Num_at_fault_claims_past_1_yr_Ind

```
## [1] 0.2926023
```

Appendix C.9: Not Adding Num_yrs_since_fault

Appendix C.9.1: One-way Analysis for Num_yrs_since_fault

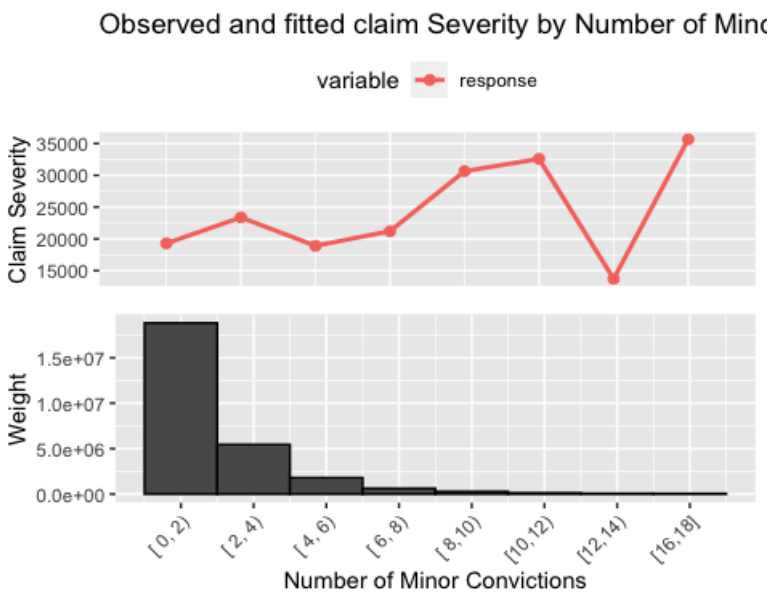


Appendix C.9.2: Significance Test for Num_yrs_since_fault

```
##
## Call:
## glm(formula = severity ~ AY_factor + vehicle_age + agecat + Liab_driving_record_Ind +
##     Marital_status_Ind + Num_at_fault_claims_past_1_yr_Ind +
##     Num_yrs_since_fault, family = Gamma(link = "log"), data = coll_dataset_claims,
##     weights = Collision_claim_count, subset = (partition == "Training"),
##     na.action = "na.pass", x = TRUE)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.0472  -1.4941  -0.5644   0.2927   4.0143
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        9.511700   0.094932 100.195  < 2e-16 ***
## AY_factor2020                      0.291490   0.049192   5.926 3.38e-09 ***
## AY_factor2021                      0.317484   0.050738   6.257 4.33e-10 ***
## vehicle_age                       -0.074700   0.005239 -14.259  < 2e-16 ***
## agecat                            -0.094267   0.021253  -4.435 9.44e-06 ***
## Liab_driving_record_Indnot6        0.054386   0.056756   0.958   0.3380
## Marital_status_IndSingle           0.091705   0.044923   2.041   0.0413 *
## Num_at_fault_claims_past_1_yr_Ind1 0.088474   0.106753   0.829   0.4073
## Num_yrs_since_faultover_5          0.189449   0.077855   2.433   0.0150 *
## Num_yrs_since_faultlast_5          0.160191   0.083986   1.907   0.0565 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 1.709511)
##
##     Null deviance: 7725.1  on 3999  degrees of freedom
## Residual deviance: 7214.4  on 3990  degrees of freedom
## AIC: 78263
##
## Number of Fisher Scoring iterations: 7
```

Appendix C.10: Not Adding Num_minor Convictions

Appendix C.10.1: One-way Analysis for Num_minor_Convictions



Appendix C.11: Correlation Check

```
##                      Vehicle_horsepower Vehicle_retail_price Vehicle_wheelbase
## Vehicle_horsepower            1.0000000            0.8476642         0.7119041
## Vehicle_retail_price          0.8476642            1.0000000         0.5646098
## Vehicle_wheelbase             0.7119041            0.5646098         1.0000000
```

Appendix C.12: Vehicle_retail_price Model

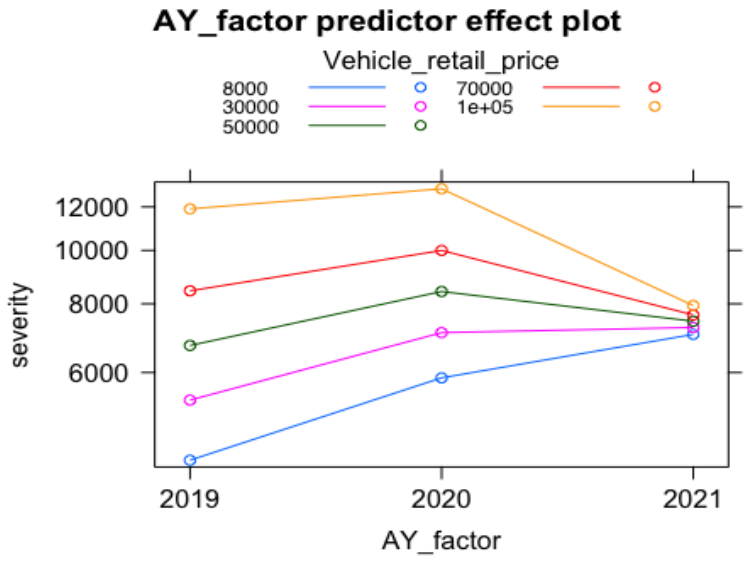
```
##       80%       85%       90%       95%       96%       97%       98%     98.5%
##  44093.60  46828.00  52018.30  60811.00  63494.00  68798.10  75478.00  77790.81
##       99%     99.5%      100%
##  84311.05  96469.37 176550.00
```

Appendix C.12.1: One-way Analysis for Vehicle_retail_price

Appendix C.12.2: Significance Test for Vehicle_retail_price

```
##
## Call:
## glm(formula = severity ~ AY_factor + vehicle_age + agecat + Liab_driving_record_Ind +
##     Marital_status_Ind + Num_at_fault_claims_past_1_yr_Ind +
##     Vehicle_retail_price, family = Gamma(link = "log"), data = coll_dataset_claims,
##     weights = Collision_claim_count, subset = (partition == "Training"),
##     na.action = "na.pass", x = TRUE)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.0830  -1.4851  -0.5613   0.3026   3.6855
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       9.224e+00  1.110e-01  83.131  < 2e-16 ***
## AY_factor2020                     2.925e-01  4.855e-02   6.023 1.86e-09 ***
## AY_factor2021                     3.273e-01  5.010e-02   6.532 7.31e-11 ***
## vehicle_age                      -7.219e-02  5.209e-03 -13.859  < 2e-16 ***
## agecat                           -8.551e-02  2.086e-02  -4.100 4.21e-05 ***
## Liab_driving_record_Indnot6       1.065e-01  5.057e-02   2.106   0.0352 *
## Marital_status_IndSingle          1.092e-01  4.452e-02   2.453   0.0142 *
## Num_at_fault_claims_past_1_yr_Ind1  1.854e-01  8.591e-02   2.159   0.0309 *
## Vehicle_retail_price              6.916e-06  1.473e-06   4.694 2.77e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 1.666811)
##
##     Null deviance: 7725.1  on 3999  degrees of freedom
## Residual deviance: 7194.3  on 3991  degrees of freedom
## AIC: 78247
##
## Number of Fisher Scoring iterations: 7
```

Appendix C.12.3: Time Consistency for Vehicle_retail_price



Appendix C.12.4: F-Test for Vehicle_retail_price

```
## Analysis of Deviance Table
##
## Model 1: severity ~ AY_factor + vehicle_age + agecat + +Marital_status_Ind +
```

```
##     Liab_driving_record_Ind + Num_at_fault_claims_past_1_yr_Ind
## Model 2: severity ~ AY_factor + vehicle_age + agecat + Liab_driving_record_Ind +
##     Marital_status_Ind + Num_at_fault_claims_past_1_yr_Ind +
##     Vehicle_retail_price
##   Resid. Df Resid. Dev Df Deviance      F    Pr(>F)
## 1      3992     7230.5
## 2      3991     7194.3  1   36.221 21.731 3.24e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Appendix C.12.5: Parsimony for Vehicle_retail_price

```
## [1] 78270.52 78327.17

## [1] 78247.22 78310.16
```

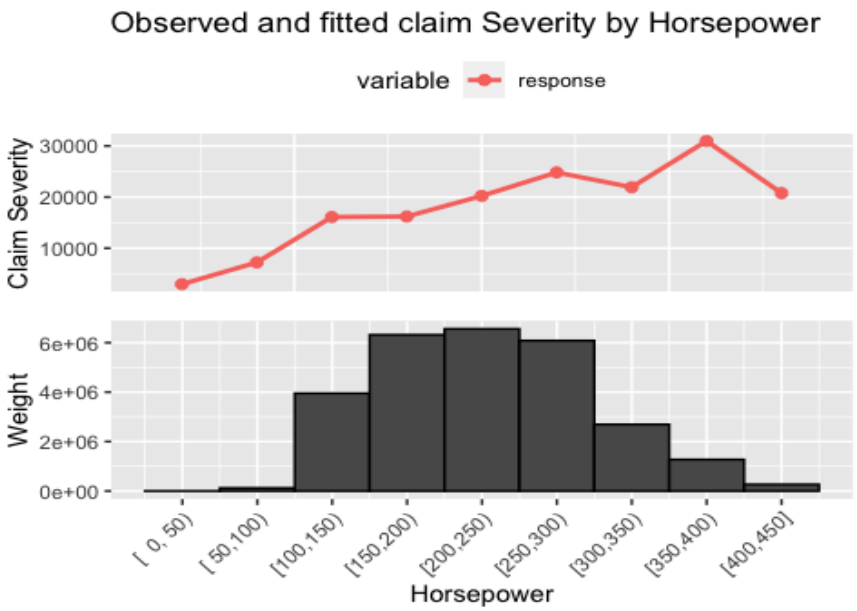
Appendix C.12.6: CV for Vehicle_retail_price

```
## [1] 0.3042531
```

Appendix C.13: Vehicle_horsepower Model

```
##   80%   85%   90%   95%   96%   97%   98% 98.5%   99% 99.5%  100%
##   274   290   306   340   345   350   360   365   390   403   507
```

Appendix C.13.1: One-way Analysis for Vehicle_horsepower

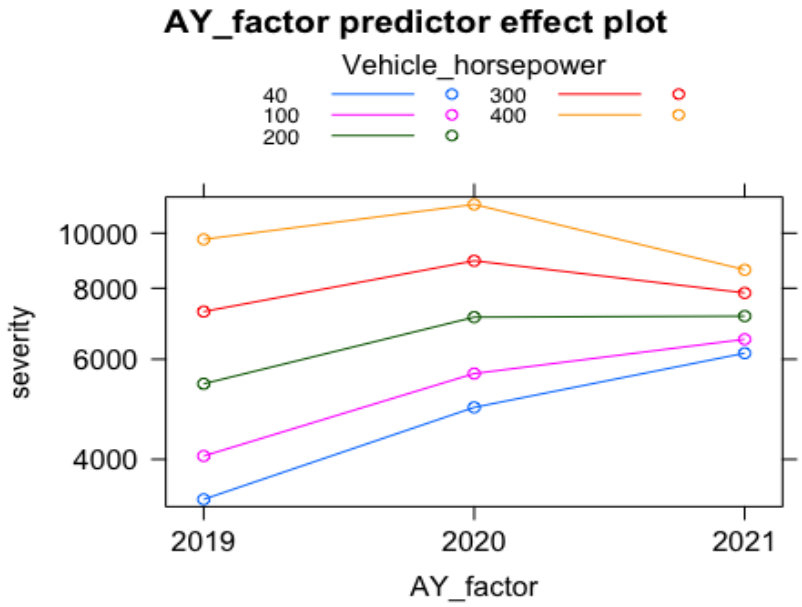


Appendix C.13.2: Significance Test for Vehicle_horsepower

```
##
## Call:
## glm(formula = severity ~ AY_factor + vehicle_age + agecat + Liab_driving_record_Ind +
##     Marital_status_Ind + Num_at_fault_claims_past_1_yr_Ind +
##     Vehicle_horsepower, family = Gamma(link = "log"), data = coll_dataset_claims,
##     weights = Collision_claim_count, subset = (partition == "Training"),
##     na.action = "na.pass", x = TRUE)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.0766  -1.4868  -0.5720   0.2893   3.7607
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      9.1831188  0.1278322  71.837  < 2e-16 ***
```

```
## AY_factor2020                          0.2883806  0.0487462   5.916 3.58e-09 ***
## AY_factor2021                          0.3196113  0.0504293   6.338 2.59e-10 ***
## vehicle_age                           -0.0689630  0.0054381 -12.682  < 2e-16 ***
## agecat                                -0.0826326  0.0209540  -3.944 8.17e-05 ***
## Liab_driving_record_Indnot6            0.0993641  0.0507075   1.960 0.050118 .
## Marital_status_IndSingle               0.1046879  0.0446033   2.347 0.018969 *
## Num_at_fault_claims_past_1_yr_Ind1     0.1887991  0.0861897   2.191 0.028545 *
## Vehicle_horsepower                     0.0011638  0.0003153   3.691 0.000227 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 1.678193)
##
##     Null deviance: 7725.1  on 3999  degrees of freedom
## Residual deviance: 7208.7  on 3991  degrees of freedom
## AIC: 78257
##
## Number of Fisher Scoring iterations: 7
```

Appendix C.13.3: Time Consistency for Vehicle_horsepower



Appendix C.13.4: F-Test for Vehicle_horsepower

```
## Analysis of Deviance Table
##
## Model 1: severity ~ AY_factor + vehicle_age + agecat + +Marital_status_Ind +
##     Liab_driving_record_Ind + Num_at_fault_claims_past_1_yr_Ind
## Model 2: severity ~ AY_factor + vehicle_age + agecat + Liab_driving_record_Ind +
##     Marital_status_Ind + Num_at_fault_claims_past_1_yr_Ind +
##     Vehicle_horsepower
##   Resid. Df Resid. Dev Df Deviance      F    Pr(>F)
## 1      3992     7230.5
## 2      3991     7208.7  1   21.818 13.001 0.0003151 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Appendix C.13.5: Parsimony for Vehicle_horsepower

```
## [1] "AIC"

## [1] 78270.52 78327.17
```

```
## [1] "BIC"

## [1] 78257.29 78320.23
```

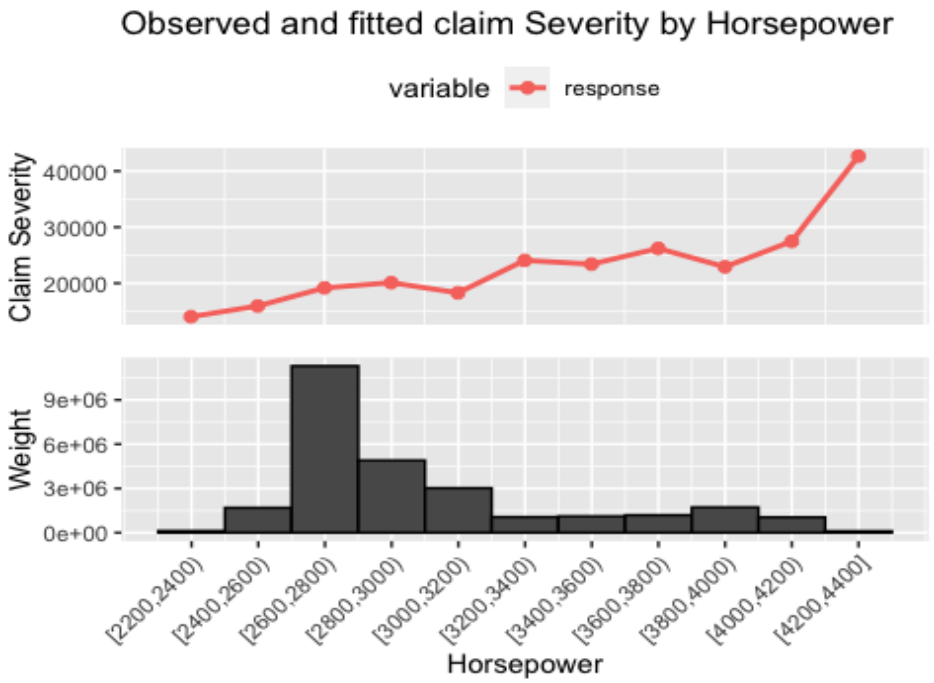
Appendix C.13.6: CV for Vehicle_horsepower

```
## [1] 0.302763
```

Appendix C.14: Vehicle_wheelbase Model

```
##     80%     85%     90%     95%     96%     97%     98%   98.5%     99%   99.5%
## 3198.00 3531.00 3664.00 3891.15 3941.00 3998.06 4064.00 4077.00 4165.00 4173.00
##    100%
## 4369.00

##    0.5%      1%      2%      3%      4%      5%      7%     10%
## 2373.00 2428.96 2465.00 2480.00 2511.00 2519.75 2570.00 2600.00
```

Appendix C.14.1: One-way Analysis for Vehicle_wheelbase

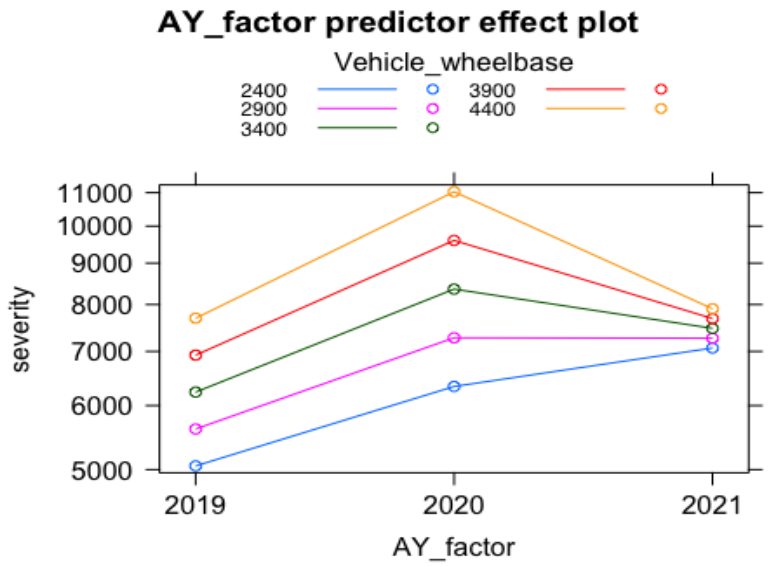


Appendix C.14.2: Significance Test for Vehicle_wheelbase

```
##
## Call:
## glm(formula = severity ~ AY_factor + vehicle_age + agecat + Liab_driving_record_Ind +
##     Marital_status_Ind + Num_at_fault_claims_past_1_yr_Ind +
##     Vehicle_wheelbase, family = Gamma(link = "log"), data = coll_dataset_claims,
##     weights = Collision_claim_count, subset = (partition == "Training"),
##     na.action = "na.pass", x = TRUE)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.0495  -1.4921  -0.5722   0.2976   4.0924
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     9.100e+00  1.777e-01  51.223  < 2e-16 ***
## AY_factor2020                   2.853e-01  4.889e-02   5.835 5.79e-09 ***
## AY_factor2021                   3.237e-01  5.044e-02   6.418 1.54e-10 ***
## vehicle_age                    -7.384e-02  5.220e-03 -14.147  < 2e-16 ***
## agecat                         -8.403e-02  2.103e-02  -3.996 6.56e-05 ***
```

```
## Liab_driving_record_Indnot6         9.312e-02  5.082e-02   1.832  0.06697 .
## Marital_status_IndSingle            9.602e-02  4.464e-02   2.151  0.03154 *
## Num_at_fault_claims_past_1_yr_Ind1  1.964e-01  8.647e-02   2.271  0.02319 *
## Vehicle_wheelbase                   1.320e-04  4.874e-05   2.709  0.00678 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 1.689569)
##
##     Null deviance: 7725.1  on 3999  degrees of freedom
## Residual deviance: 7218.2  on 3991  degrees of freedom
## AIC: 78264
##
## Number of Fisher Scoring iterations: 7
```

Appendix C.14.3: Time Consistency for Vehicle_wheelbase



Appendix C.14.4: F-Test for Vehicle_wheelbase

```
## Analysis of Deviance Table
##
## Model 1: severity ~ AY_factor + vehicle_age + agecat + +Marital_status_Ind +
##     Liab_driving_record_Ind + Num_at_fault_claims_past_1_yr_Ind
## Model 2: severity ~ AY_factor + vehicle_age + agecat + Liab_driving_record_Ind +
##     Marital_status_Ind + Num_at_fault_claims_past_1_yr_Ind +
##     Vehicle_wheelbase
##   Resid. Df Resid. Dev Df Deviance      F  Pr(>F)
## 1      3992     7230.5
## 2      3991     7218.2  1   12.354 7.3117 0.00688 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Appendix C.14.5: Parsimony for Vehicle_wheelbase

```
## [1] "AIC"

## [1] 78270.52 78327.17

## [1] "BIC"

## [1] 78263.90 78326.85
```

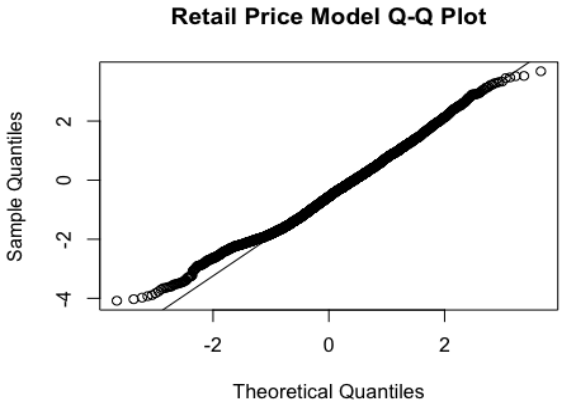
Appendix C.14.6: CV for Vehicle_wheelbase

```
## [1] 0.2977308
```

Appendix C.15: Results of CV on Training Dataset

```
##       models mean_ginis
##  [1,]      1 0.08766285
##  [2,]      2 0.10410122
##  [3,]      3 0.26095556
##  [4,]      4 0.28129904
##  [5,]      5 0.28766763
##  [6,]      6 0.29025896
##  [7,]      7 0.29260229
##  [8,]      9 0.30425314
##  [9,]     10 0.30276298
## [10,]     11 0.29773083
```

Appendix C.16: Model Assumptions for Vehicle_retail_price Model (9)



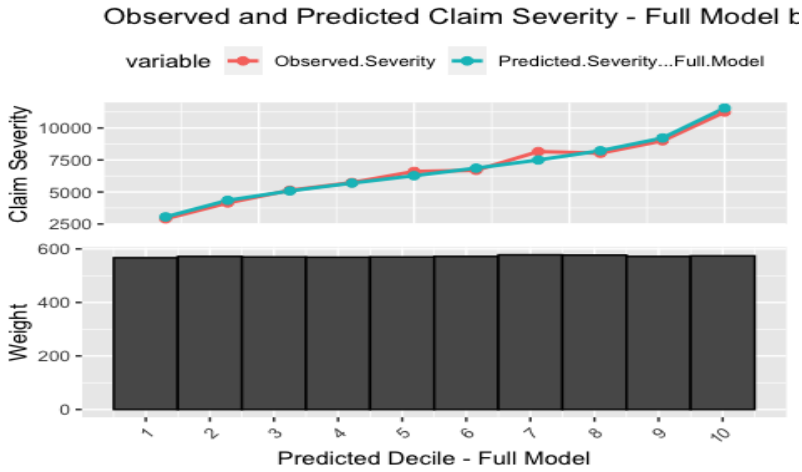
Appendix C.17: Holdout Testing

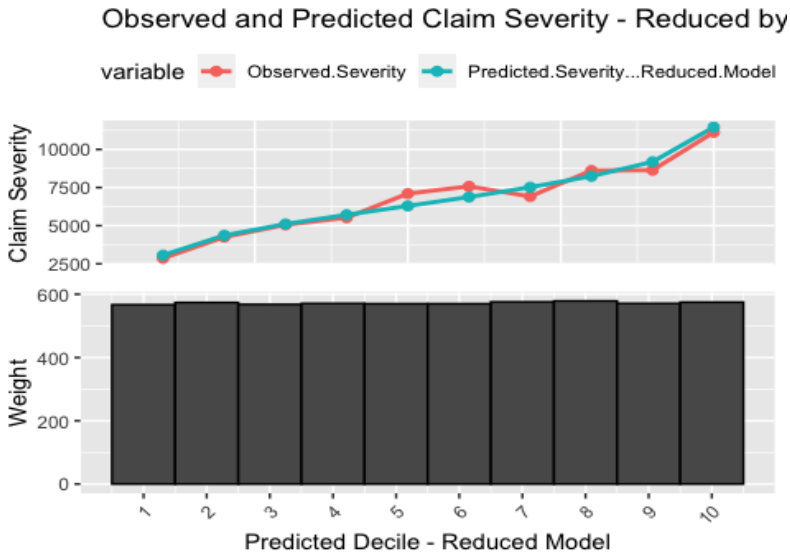
Appendix C.17.1: Parsimony Tests for Full Model vs Reduced Model

```
## [1] "AIC"

## [1] 109981.4 110047.8

## [1] "BIC"

## [1] 109991.2 110051.0
```

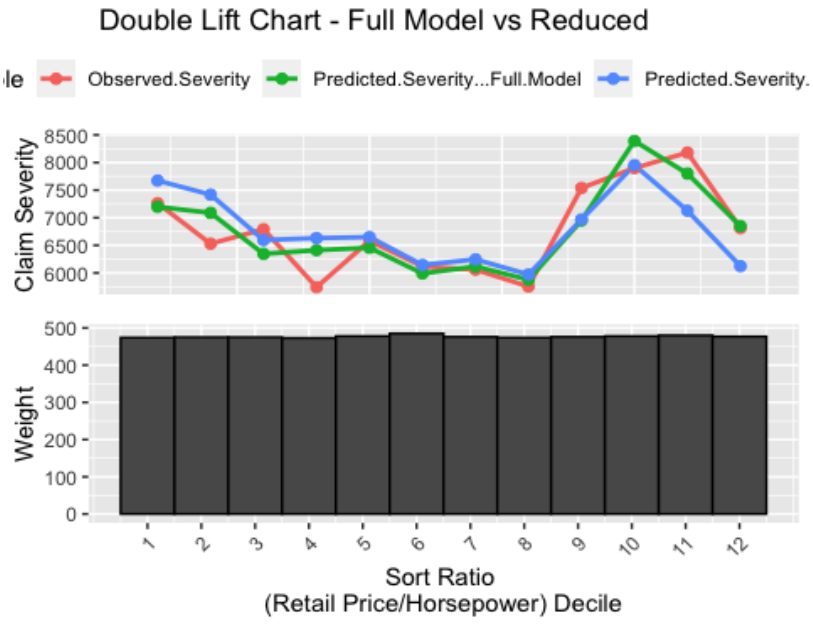
Appendix C.17.2: Simple Quantile Plots for Full Model and Reduced Model

```
## [1] 1.002557
```





Appendix C.17.3: Double Lift Chart



Appendix C.17.4: Holdout Gini Coefficients

```
## [1] 0.3315122
```

```
## [1] 0.3263574
```

Appendix C.18: Combining Sev Predictions with Freq Predictions

```
## [1] 0.3637885
```

```
## [1] 0.2378316
```

Appendix D.1: Adding Geographic Predictors to Model

Adding age_15: Proportion of the population aged 15 and over

```
Call:
glm(formula = Loss_cost ~ Accident_year + Vehicle_age_cap_30_floor_6 +
    Years_driving + Years_driving_squared + age_15, family = tweedie(var.power = 1.5,
    link.power = 0), data = dat_cencus %>% filter(partition ==
    "Training"), weights = Collision_earned_count)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-14.281   -8.196   -6.231   -4.168  121.080

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                9.6526653  0.7793731  12.385  < 2e-16 ***
Accident_year2020          0.1147367  0.0803228   1.428   0.1532
Accident_year2021          0.1633370  0.0823387   1.984   0.0473 *
Vehicle_age_cap_30_floor_6 -0.1388184  0.0089435 -15.522  < 2e-16 ***
Years_driving              -0.0482699  0.0083704  -5.767 8.11e-09 ***
Years_driving_squared       0.0005311  0.0001354   3.921 8.81e-05 ***
age_15                     -1.1734655  0.9360334  -1.254   0.2100
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Tweedie family taken to be 1071.709)

    Null deviance: 6474677  on 79788  degrees of freedom
Residual deviance: 6103793  on 79782  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 7
```

Adding age15_19: Proportion of the population aged 15 to 19 years

As a continuous variable

```
Call:
glm(formula = Loss_cost ~ Accident_year + Vehicle_age_cap_30_floor_6 +
    Years_driving + Years_driving_squared + age15_19, family = tweedie(var.power = 1.5,
    link.power = 0), data = dat_cencus %>% filter(partition ==
    "Training"), weights = Collision_earned_count)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-14.247   -8.195   -6.238   -4.163  122.182

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                8.7500807  0.2737582  31.963  < 2e-16 ***
Accident_year2020          0.1164585  0.0805651   1.446   0.1483
Accident_year2021          0.1675724  0.0825614   2.030   0.0424 *
Vehicle_age_cap_30_floor_6 -0.1388008  0.0089685 -15.477  < 2e-16 ***
Years_driving              -0.0483428  0.0083944  -5.759 8.49e-09 ***
Years_driving_squared       0.0005315  0.0001358   3.913 9.11e-05 ***
age15_19                   -0.7937296  3.1592731  -0.251   0.8016
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Tweedie family taken to be 1077.915)

    Null deviance: 6474677  on 79788  degrees of freedom
Residual deviance: 6105420  on 79782  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 7
```

As a categorical variable grouped as high or low, in which the middle point is the median

```
Call:
glm(formula = Loss_cost ~ Accident_year + Vehicle_age_cap_30_floor_6 +
    Years_driving + Years_driving_squared + age15_19_high, family = tweedie(var.power = 1.5,
    link.power = 0), data = dat_cencus %>% filter(partition ==
    "Training"), weights = Collision_earned_count)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-14.269   -8.196   -6.237   -4.164  121.844

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                8.6867420  0.1526441  56.908  < 2e-16 ***
Accident_year2020          0.1159361  0.0806349   1.438   0.1505
Accident_year2021          0.1666352  0.0826377   2.016   0.0438 *
Vehicle_age_cap_30_floor_6 -0.1388044  0.0089769 -15.462  < 2e-16 ***
Years_driving              -0.0483578  0.0084015  -5.756 8.65e-09 ***
Years_driving_squared       0.0005319  0.0001359   3.913 9.12e-05 ***
age15_19_highYes            0.0103045  0.0677992   0.152   0.8792
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Tweedie family taken to be 1079.911)

    Null deviance: 6474677  on 79788  degrees of freedom
Residual deviance: 6105461  on 79782  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 7
```

Adding age18_24: Proportion of the population aged 18 to 24 years

```
Call:
glm(formula = Loss_cost ~ Accident_year + Vehicle_age_cap_30_floor_6 +
    Years_driving + Years_driving_squared + age18_24, family = tweedie(var.power = 1.5,
    link.power = 0), data = dat_cencus %>% filter(partition ==
    "Training"), weights = Collision_earned_count)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-14.272   -8.197   -6.232   -4.167  121.119

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                9.046897   0.272802  33.163  < 2e-16 ***
Accident_year2020          0.116792   0.080052   1.459   0.1446
Accident_year2021          0.166579   0.082051   2.030   0.0423 *
Vehicle_age_cap_30_floor_6 -0.138805   0.008912 -15.576  < 2e-16 ***
Years_driving              -0.048267   0.008342  -5.786 7.24e-09 ***
Years_driving_squared       0.000531   0.000135   3.934 8.37e-05 ***
age18_24                   -1.545642   1.000382  -1.545   0.1223
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Tweedie family taken to be 1064.094)

    Null deviance: 6474677  on 79788  degrees of freedom
Residual deviance: 6103009  on 79782  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 7
```

Adding `age40_44`: Proportion of the population aged 40 to 44 years

```
Call:
glm(formula = Loss_cost ~ Accident_year + Vehicle_age_cap_30_floor_6 +
    Years_driving + Years_driving_squared + age40_44, family = tweedie(var.power = 1.5,
    link.power = 0), data = dat_cencus %>% filter(partition ==
    "Training"), weights = Collision_earned_count)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-14.346   -8.194   -6.232   -4.166  120.354

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                8.3974213  0.2985694  28.126  < 2e-16 ***
Accident_year2020          0.1148197  0.0801558   1.432   0.1520
Accident_year2021          0.1656245  0.0821458   2.016   0.0438 *
Vehicle_age_cap_30_floor_6 -0.1387794  0.0089226 -15.554  < 2e-16 ***
Years_driving             -0.0482035  0.0083530  -5.771 7.92e-09 ***
Years_driving_squared      0.0005306  0.0001351   3.926 8.64e-05 ***
age40_44                   3.7767696  3.3376639   1.132   0.2578
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Tweedie family taken to be 1066.969)

    Null deviance: 6474677  on 79788  degrees of freedom
Residual deviance: 6104118  on 79782  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 7
```

Adding `age50_54`: Proportion of the population aged 50 to 54 years

```
Call:
glm(formula = Loss_cost ~ Accident_year + Vehicle_age_cap_30_floor_6 +
    Years_driving + Years_driving_squared + age50_54, family = tweedie(var.power = 1.5,
    link.power = 0), data = dat_cencus %>% filter(partition ==
    "Training"), weights = Collision_earned_count)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-14.195   -8.198   -6.235   -4.164  121.882

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                8.9026861  0.2603326  34.197  < 2e-16 ***
Accident_year2020          0.1165392  0.0803662   1.450   0.1470
Accident_year2021          0.1663873  0.0823727   2.020   0.0434 *
Vehicle_age_cap_30_floor_6 -0.1388172  0.0089468 -15.516  < 2e-16 ***
Years_driving             -0.0483004  0.0083744  -5.768 8.07e-09 ***
Years_driving_squared      0.0005310  0.0001355   3.919 8.90e-05 ***
age50_54                  -2.5532541  2.6014175  -0.981   0.3264
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Tweedie family taken to be 1072.684)

    Null deviance: 6474677  on 79788  degrees of freedom
Residual deviance: 6104463  on 79782  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 7
```

Adding per_2: Proportion of private households with 2 persons

```
Call:
glm(formula = Loss_cost ~ Accident_year + Vehicle_age_cap_30_floor_6 +
    Years_driving + Years_driving_squared + per_2, family = tweedie(var.power = 1.5,
    link.power = 0), data = dat_census %>% filter(partition ==
    "Training"), weights = Collision_earned_count)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-14.202   -8.195   -6.236   -4.166  121.836

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 8.8450603  0.2319994  38.125  < 2e-16 ***
Accident_year2020           0.1159838  0.0803106   1.444   0.1487
Accident_year2021           0.1655566  0.0823155   2.011   0.0443 *
Vehicle_age_cap_30_floor_6 -0.1388662  0.0089416 -15.530  < 2e-16 ***
Years_driving              -0.0483804  0.0083669  -5.782 7.39e-09 ***
Years_driving_squared       0.0005327  0.0001354   3.935 8.33e-05 ***
per_2                      -0.5046186  0.5980220  -0.844   0.3988
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Tweedie family taken to be 1071.3)

    Null deviance: 6474677  on 79788  degrees of freedom
Residual deviance: 6104716  on 79782  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 7
```

Adding per_5: Proportion of private households with 5 persons

```
Call:
glm(formula = Loss_cost ~ Accident_year + Vehicle_age_cap_30_floor_6 +
    Years_driving + Years_driving_squared + per_5, family = tweedie(var.power = 1.5,
    link.power = 0), data = dat_census %>% filter(partition ==
    "Training"), weights = Collision_earned_count)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-14.243   -8.196   -6.235   -4.165  122.065

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 8.6584726  0.1881514  46.019  < 2e-16 ***
Accident_year2020           0.1161556  0.0806206   1.441    0.150
Accident_year2021           0.1664241  0.0826266   2.014    0.044 *
Vehicle_age_cap_30_floor_6 -0.1388210  0.0089756 -15.466  < 2e-16 ***
Years_driving              -0.0483510  0.0084000  -5.756 8.64e-09 ***
Years_driving_squared       0.0005320  0.0001359   3.914 9.07e-05 ***
per_5                       0.3820649  1.3034616   0.293    0.769
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Tweedie family taken to be 1079.563)

    Null deviance: 6474677  on 79788  degrees of freedom
Residual deviance: 6105391  on 79782  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 7
```

Appendix E.1 Training CV Gini of Different Combinations of J and k for Credibility Smoothing

Note: numClusters = k

J <dbl>	numClusters <int>	cvGini <dbl>
250	2	0.3621887
150	2	0.3674586
250	4	0.3678689
150	3	0.3704747
250	3	0.3707244
50	2	0.3720954
150	4	0.3731371
50	3	0.3757557
50	4	0.3777735

Appendix F.1: Final Model with All Data

```
Call:
glm(formula = Loss_cost ~ Accident_year + Vehicle_age_cap_30_floor_6 +
    Years_driving + Years_driving_squared + Num_minor_convictions_cap_5 +
    Num_yrs_since_last_at_fault_claim_num_cap_13 + New_business_imputed +
    Has_partner, family = tweedie(var.power = 1.5, link.power = 0),
    data = datLC, weights = Collision_earned_count)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-16.745   -8.090   -6.152   -4.127  136.626

Coefficients:
                                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                                8.4258773  0.1483408  56.801  < 2e-16 ***
Accident_year2020                          0.1428226  0.0655278   2.180 0.029291 *
Accident_year2021                          0.2427263  0.0668480   3.631 0.000282 ***
Vehicle_age_cap_30_floor_6                -0.1300840  0.0072097 -18.043  < 2e-16 ***
Years_driving                             -0.0349625  0.0069781  -5.010 5.44e-07 ***
Years_driving_squared                      0.0004020  0.0001111   3.618 0.000297 ***
Num_minor_convictions_cap_5                0.1052627  0.0186512   5.644 1.67e-08 ***
Num_yrs_since_last_at_fault_claim_num_cap_13 -0.0338773  0.0065242  -5.193 2.08e-07 ***
New_business_imputedYes                    0.5301164  0.1028209   5.156 2.53e-07 ***
Has_partnerNo                              0.3345716  0.0590643   5.665 1.48e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Tweedie family taken to be 981.8045)

    Null deviance: 9222052  on 113784  degrees of freedom
Residual deviance: 8572389  on 113775  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 7
```

Appendix F.2: Final Model with All Data and Deductible

```
Call:
glm(formula = Loss_cost ~ Accident_year + Vehicle_age_cap_30_floor_6 +
    Years_driving + Years_driving_squared + Num_minor_convictions_cap_5 +
    Num_yrs_since_last_at_fault_claim_num_cap_13 + New_business_imputed +
    Has_partner + Collision_deductible, family = tweedie(var.power = 1.5,
    link.power = 0), data = datLC, weights = Collision_earned_count)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-16.933   -8.090   -6.147   -4.128  135.715

Coefficients:
                                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                                 8.4577286  0.1511315  55.963  < 2e-16 ***
Accident_year2020                           0.1435219  0.0656972   2.185 0.028920 *
Accident_year2021                           0.2446290  0.0670400   3.649 0.000263 ***
Vehicle_age_cap_30_floor_6                 -0.1305695  0.0072396 -18.035  < 2e-16 ***
Years_driving                              -0.0358361  0.0070461  -5.086 3.66e-07 ***
Years_driving_squared                       0.0004114  0.0001118   3.678 0.000235 ***
Num_minor_convictions_cap_5                 0.1067602  0.0187498   5.694 1.24e-08 ***
Num_yrs_since_last_at_fault_claim_num_cap_13 -0.0341931  0.0065417  -5.227 1.73e-07 ***
New_business_imputedYes                     0.5421117  0.1034457   5.241 1.60e-07 ***
Has_partnerNo                               0.3373463  0.0592352   5.695 1.24e-08 ***
Collision_deductible1000                   -0.0906490  0.0793763  -1.142 0.253450
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Tweedie family taken to be 986.4508)

    Null deviance: 9222052  on 113784  degrees of freedom
Residual deviance: 8571114  on 113774  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 7
```