STA302H1 Summer Final project report

A practical study on predicting auto insurance monthly premiums using ranges of customer data and the linear regression model

Final projects group 5

November 28, 2024

Contents

1 Introduction

The determination of individual monthly auto insurance premiums is a complex process influenced by various customer data attributes. This report investigates the impact of these variables using an efficient multiple linear regression model to predict auto insurance premiums. The primary objective of this study is to identify the key factors that significantly influence premium costs and to develop a predictive model that can accurately estimate premiums based on customer data.

The dataset used in this research comprises a real-life auto insurance portfolio from Spain, covering the period from November 2015 to December 2018. The data was sourced from the Inter-university Consortium for Political and Social Research (ICPSR) and includes 105,555 records with 30 variables, each representing specific attributes of annually renewable insurance policies. These variables encompass demographic details, vehicle characteristics, and insurance policy specifics, all of which are believed to play a critical role in determining insurance premiums.

However, modeling insurance premiums presents several challenges. The dataset contains missing values, and the relationships between predictors and the response variable are often non-linear, necessitating data cleaning, transformation, and model diagnostics. Additionally, issues such as multicollinearity among predictors and heteroscedasticity in residuals complicate the modeling process. Overcoming these difficulties requires the application of statistical techniques, such as variable transformations, to meet the assumptions of linear regression and improve the model's accuracy.

In this study, a multiple linear regression model was constructed with 19 predictors, carefully selected and transformed to enhance the model's predictive power. Various diagnostic tests were performed to ensure the validity of the model, including residual plots, the Breusch-Pagan test for heteroscedasticity, and the calculation of R-squared values. The model is able to explain the variance in insurance premiums, though some non-linear relationships remained unaccounted for.

This research is particularly relevant to the course STA302 as it applies key concepts such as linear regression, data transformation, and model diagnostics, to a real-world problem. The findings contribute to a deeper understanding of how customer data influences insurance premiums, providing valuable insights for both academic and industry applications.

2 Methodologies

First of all, we will introduce the methodologies for this research. In this section, we will start introducing the data set and our methodologies for data cleaning, we will also explain our ways of data transformations and predictors selections. Finally, we will talk about process of model formulations and statistics procedures in the model.

2.1 Data set description

To adequately address the research topic and question, data was gathered from the reputable Inter-university Consortium for Political and Social Research (ICPSR). The careful selection of a reliable dataset is crucial in mathematical modeling, as it directly influences the accuracy and validity of the research outcomes. The data used is Dataset of an actual motor vehicle insurance portfolio provided by Josep Lledó and Jose M. Pavía from the University of Valencia. This data set records real life auto insurance portfolio from November 2015 to December 2018 in Spain. The dataset can be found at https://www.openicpsr.org/openicpsr/project/193182/version/V1/view;jsessionid=C38C78894508D4679F334EB94D327851

In this data set, it contains 105555 rows and 30 columns. Each row represents a specific annually renewal insurance policy, and each column represents a specific attribute to this specific insurance policy for the insured person. The detailed description of the variables is very long. Thus, it is in the appendix section 5.1 of this report. The first 6 rows of the original data set is also in the appendix section 5.1.

2.2 Cleaning and transforming the data set

The detailed coding of this part is provided in the appendix section 5.2 and 5.3.

First of all, we have to find a way to replace or remove the NA values from the data set. We can inspect the number of NAs categorized by columns. We loading the original data set in R and name the data frame as "data", and run the R code, we see that there is no NA values for any columns except for "Type_fuel" and "Length". "Type_fuel" has 1764 NAs and "Length" has 10329 NAs.

Since "Type_fuel" is a categorical variable, it means we cannot replace NAs with any other values without compromising the integrity of the original data set. Thus, we will use "filter" to remove the rows with NA values in "Type_fuel".

Next, with column "Length", in order to aviod deleting too much rows from the data set, the remaining NAs will be replaced with the mean of "Length".

After cleaning the data set, data transformations and predictors selections are conducted. Since we are performing regression analysis on the "Premium" against other variables in the data set, the variables involving in the analysis should only impact the pricing of the policy. Since the pricing of a auto insurance policy is based on inherent risk and credit of the insured person, the predictors used in the analysis should be limited to those factors that directly influence the evaluation of their risk and credit.

The variable "Premium" is going to be our response variable. For the predictor variables, "ID" will not be considered as it is merely a tracking variable in the data set. "Date_start_contract", "Date_last_renewal", "Date_next_renewal", and "Date_lapse" are covered with similar variables "Lapse" and "Seniority". As for the "Date_driving_licence", "Date_birth", and "Year_matriculation", these time variables will be normalized by taking the number of years between the "Date_next_renewal" and the date recorded in these variables. This ensures that the variables are numeric to and are comparable with the price of the insurance policy. "N_claims_year" and "R_Claims_history" are excluded because variable "N_claims_history" covers the same information. Payment will also not be considered in the regression analysis because the payment method does not interfere with the total premium paid by the insured person. "Max_policies" will be included in the analysis because it covers the data from "Max_products". "Distribution_channel" will also not be included because it does not contribute to the pricing of the insurance.

In summary, we have removed "ID", "Date_start_contract", "Date_last_renewal", "Date_next_renewal", "Date_lapse", "N_claims_year", "R_Claims_history", "Max_products", and "Distribution_channel".

In order to normalize "Date_birth", "Date_driving_licence", and "Year_matriculation". This means we must transform and normalize these information into something we can work with in a regression. In the evaluation of the insurance premium, age, experiences of the insured person, and the age of the insured vehicles are all in the consideration. The first step includes annualizing data that is not recorded on an annual basis. For instance, when payment is semi-annual, "Date_birth," "Date_driving_licence," and "Year_matriculation" are all recorded twice each year. These must be standardized to be comparable with the data when payment is annual. The processing procedure follows:

1. To normalize Date_birth into the brithday_diff, two cases based on payment methods should be considered:

    - When the payment method is annual, the insurance company consider the premium based on the age at Date_last_renewal. Thus, the individual's age at the date of renewal is simply Date_last_renewal minus Date_birth.
    - When the payment is semiannual, the insurance company considers the premium semiannually too. To make the data comparable with the other annualized data, age for semiannual payment should be normalized as an arithmetic average of the two ages recorded in the given year.

    $$\frac{\text{age } 1 + \text{age } 2}{2}$$

2. Date_driving_licence will be normalized to drive_age, following a similar procedure as Date_Birth.

3. For Year_matriculation, we have a little special case because we only have the year of the registration of the vehicles instead of a month and a day. In order to accurately estimate this metric, we will take the year from the Date_last_renewal and use that to minus the Year_matriculation. Although the dataset lacks the specific date of matriculation which causes imprecision when calculating the age of a car, considering vehicles are durable products in general, this estimation should only minimally impact the true metric of measuring risk in general.

Note that the specific realization of the algorithm above is in the appendix. After running the R code, we have normalized the Date_birth into brithday_diff, Date_driving_licence into drive_age, and Year_matriculation into age_car. Now, we will filter the variables we are going to use in a new data set called data3 containing only predictors and the response variable.

Before moving to any formal model construction, we need to consider the fact that our model in one way or another requires transformations in the regresison model. And in log transformation, they only works for positive values. Thus, the first step is to check if the dataset contains any negative values.

By running the R code in the appendix 5.3, we can see that there are 26 data points with negative drive_age after running the R code, this does not mean our calculation of drive_age is wrong, but means after and right on the renewal of the policy, the driver renewed his license. However, this means we need to clean those meaningless data point from the data set by cleaning the whole row. And we do this cleaning by pumping the values we want into a new data frame data4 as shown in the appendix

This concludes our data cleaning and data transformation methodologies. The data4 is now ready for model constructions. You can find the first 6 rows of the data from the head function in the appendix section 5.3.

2.3 Model formulations

The model we are going to be using is a multiple linear regression model with 19 predictors and 1 response variable, which is "Premium". The detailed descriptions of the predictors are in the appendix section.

In order to estimate the response variable with selected predictors to a certain level of accuracy and to understand the extent of linear correlation between predictors and the response variable, the least square

estimator is used to estimate the coefficients of the linear model in the regression. The estimation process will be done automatically by R.

Thus, the equation of our multiple linear model without any transformation is going to look like:

$$\begin{aligned}\text{Premium} &= \beta_0 + \beta_1 \times \text{brithday\_diff} + \beta_2 \times \text{drive\_age} + \beta_3 \times \text{Seniority} + \beta_4 \times \text{Policies\_in\_force} \\ &+ \beta_5 \times \text{Max\_policies} + \beta_6 \times \text{Lapse} + \beta_7 \times \text{Cost\_claims\_year} + \beta_8 \times \text{N\_claims\_history} + \beta_9 \times \text{Type\_risk} \\ &+ \beta_{10} \times \text{Area} + \beta_{11} \times \text{Second\_driver} + \beta_{12} \times \text{age\_car} + \beta_{13} \times \text{Power} + \beta_{14} \times \text{Cylinder\_capacity} \\ &+ \beta_{15} \times \text{Value\_vehicle} + \beta_{16} \times \text{N\_doors} + \beta_{17} \times \text{Type\_fuel is P} + \beta_{18} \times \text{Length} + \beta_{19} \times \text{Weight}\end{aligned}$$

- There are a total of 20 coefficients including the intercept in this basic regression model. Thus, $p = 19$
- There are a total of 103760 number of rows, thus, $n = 103760$.

The significance of the predictors in the linear regression model is judged by its p-value, a very low p value can indicate the predictor is insignificant in the model. However, considering the context of our project and the natural complexity of the real life scenario, it is possible that the p-value of the predictors in our model are small but still contribute to the accuracy of the model. This could be a result of the inherent property of the data set, which requires a higher level of statistical estimation methods such as neural networks and more advanced machine learning algorithms to predict the response to a higher degree of accuracy.

It is expected that a basic linear model is insufficient, thus, $\log$ and square root transformations are applied on the predictors and the response. In case that the four assumptions of linear regression violated:

- From the lecture, we learnt that square root transformation is useful when it comes to counting data and binary data, we did a similar example on the "Cleaning room" data set.

  Thus, square root transformation will be applied to integer data. In addition, one of the another reason we use square root transformation on counting data in our data set is because our counting data involves a lot of zero, which could be troublesome for $\log$ transformation unless we add 1 to each data point. Thus, square root transformation for counting data is the most efficient way in our case. So, we apply the square root transformation on: Seniority, Policies_in_force, Max_policies, Lapse, N_claims_history, age_car, Power, N_doors, and Weight.

- For the rest of the predictors and the response variable, we will apply $\log$ transformations.In addition to apply $\log$, we also add 1 to the variable to avoid $\log(0)$ We apply the $\log$ transformation on: Premium, brithday_diff, Cost_claims_year, Value_vehicle, and Length.

- We perform no transformations on the categorical variables.

Note that we try not to edit existing data set to avoid confusion in the memory of R markdown file, thus, the transformed data set is assigned to a new data frame called data5.

Once those transformations are applied, we will exam the model, and if the normality assumption is still violated, we would perform Box-Cox transformation on the existing model. Unfortunately, we can only visually observe the QQ plot to detect non-normality because we cannot apply Shapiro-Wilk test since it only accepts sample of maximum size of 5000.

During the process of constructing models, we will check potential influential points with cook distance. We will take any points that is above 0.5 a potential influential point. If influential point does exist, we need to view those influential points, and by considering the context of the data set, we might remove these influential points.

In order to detect Multi-collinearity, we will use the VIF inflator test. And we can address the issue using ridge regression.

In order to numerically quantify the degree of Non-constant variance over the process of applying the transformation, we will use the Breusch–Pagan test from the tutorial. If the p-value is increased or above 0.05, we can conclude the transformation has improved the non-constant variance issue.

Finally, we will also apply F statistics, R square and adjusted R square to justify the validity and accuracy of the model. We use F statistics to justify whether we should reject the null hypothesis that there is no statistically significant relationship between the response variable and the predictor variables. Here, we reject the null hypothesis if the p-value is smaller than 0.05, or if the value of the F statistics is bigger than $F_{0.95}(p, n-p-1)$. We use R square and adjusted R square to see what is the percentage of the variance of the response variable is explained by the linear model. For those two metric, the bigger the better.

The specific results from our regression is going to be in the next section.

2.4 Our advantages and limitations

One of the most prominent advantage of our model and method is the fact that we try to remove the smallest among of data during the process. For example, instead of deleting all the rows containing NAs in the data cleaning process as most intuitively, we decided to replace some of the values with their mean. By doing this way, we ensured that the integrity of other data from other columns are unimpaired.

A novelty of our method is using the lubridate package in R to transform date data to something we can work with in the regression. We have successfully implemented the algorithm to calculate the metric at the exact time of paying the premium with two distinct payment method. Thus, we are able to produce a more accurate model.

However, our method contains some limitations. In the section when we pick the transformations for each variable, it contains a certain level of try and error. Initially, we are unsure which transformations we should apply to each kind of data. For example, for "brithday_diff", we are not sure whether we should use square root transformation or $\log$ transformation. We basically tried different transformations on the variables and decided to use the transformation that produces the best result. This "motivation" of our work method is not scientifically accurate.

3 Results

3.1 Initial modeling results

For the first section of the result, we will present the results we got from our initial model called "model" in R by running "lm()". For the initial modeling, we performed linear model directly to the "data4" without any transformation. We obtain the following result:

```
Call:
lm(formula = Premium ~ ., data = data4)

Residuals:
    Min      1Q  Median      3Q     Max
-832.78  -67.10  -17.58   42.02 2510.02

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       1.745e+02  7.149e+00  24.412  < 2e-16 ***
brithday_diff    -2.320e-01  6.028e-02  -3.849 0.000119 ***
drive_age        -6.946e-01  6.253e-02 -11.108  < 2e-16 ***
Seniority        -8.932e-01  6.991e-02 -12.776  < 2e-16 ***
Policies_in_force -1.085e+01  6.765e-01 -16.035  < 2e-16 ***
Max_policies     -6.321e+00  5.715e-01 -11.059  < 2e-16 ***
Lapse             2.649e+01  8.118e-01  32.636  < 2e-16 ***
Cost_claims_year  2.369e-03  2.446e-04   9.684  < 2e-16 ***
N_claims_history  4.856e+00  1.139e-01  42.640  < 2e-16 ***
Type_risk         6.366e+00  1.032e+00   6.170 6.83e-10 ***
Area              1.793e+01  8.204e-01  21.854  < 2e-16 ***
Second_driver     5.281e+01  1.124e+00  47.004  < 2e-16 ***
age_car          -3.493e+00  6.774e-02 -51.571  < 2e-16 ***
Power             4.419e-01  1.848e-02  23.910  < 2e-16 ***
Cylinder_capacity -1.796e-02  1.568e-03 -11.451  < 2e-16 ***
Value_vehicle     4.687e-03  8.284e-05  56.578  < 2e-16 ***
N_doors           1.590e+01  3.436e-01  46.282  < 2e-16 ***
Type_fuelP       -1.759e+00  9.050e-01  -1.943 0.051989 .
Length            1.225e+01  1.476e+00   8.298  < 2e-16 ***
Weight           -1.509e-02  2.043e-03  -7.386 1.53e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 116.8 on 103740 degrees of freedom
Multiple R-squared:  0.3047,    Adjusted R-squared:  0.3046
F-statistic:  2393 on 19 and 103740 DF,  p-value: < 2.2e-16
```

Figure 1: "model" summary

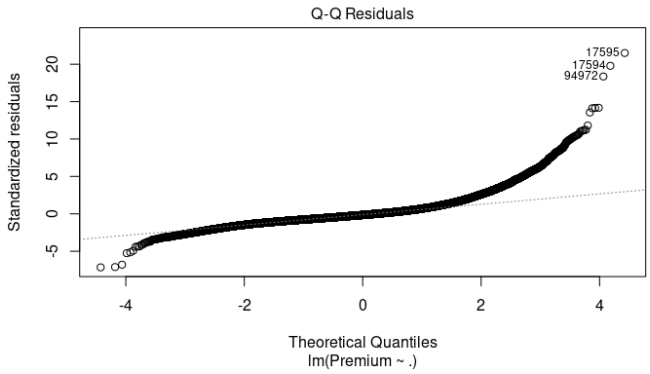

Figure 2: "model" QQ plot

We can see that from the adjusted R square, the result from "model" is not ideal, and there are significant improvement on accuracy we can perform. In addition, from this QQ plot, we can see that we have serious normality assumption violation. In addition, we see that by running the studentized Breusch-Pagan test in R on "model", we have a pretty extreme p-value ($< 2.2e-16$) with the Breusch–Pagan statistics 4580.5, we reject the null hypothesis of homoscedasticity. Thus, we also experience heterscedasticity.

Fortunately, from the F statistics, we can see that since the p-value is low, we fail to reject the null hypothesis and affirm that there is a statistically significant linear relationship between the response and the predictor variables.

Thus, from our introduced methodologies, we will apply the $\log$ and square root transformation accordingly since at least two regression assumptions are violated.

3.2 After apply the Log and Sqrt transformation

We create a new data frame "data5" to host the transformed data, and we run the linear model "lm()" again to construct a new model "model1", and we have the following result:

```
Call:
lm(formula = log_premium ~ ., data = data5)

Residuals:
     Min       1Q   Median       3Q      Max
-2.24133 -0.17909 -0.01083  0.17435  1.90309

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             3.5013587  0.0502832  69.633  < 2e-16 ***
log_brithday_diff       0.0702709  0.0066269  10.604  < 2e-16 ***
log_drive_age          -0.1204028  0.0030639 -39.297  < 2e-16 ***
sqrt_Seniority         -0.0081757  0.0012233  -6.683 2.35e-11 ***
sqrt_Policies_in_force -0.1568064  0.0052779 -29.710  < 2e-16 ***
sqrt_Max_policies      -0.0548472  0.0046789 -11.722  < 2e-16 ***
sqrt_Lapse              0.1010003  0.0024279  41.599  < 2e-16 ***
log_Cost_claims_year    0.0028186  0.0004854   5.806 6.40e-09 ***
sqrt_N_claims_history   0.0548041  0.0012351  44.370  < 2e-16 ***
Type_risk              -0.0517258  0.0030162 -17.149  < 2e-16 ***
Area                    0.0432482  0.0022231  19.454  < 2e-16 ***
Second_driver           0.1578697  0.0030510  51.744  < 2e-16 ***
sqrt_age_car           -0.0663866  0.0011009 -60.300  < 2e-16 ***
sqrt_Power              0.0330358  0.0009955  33.186  < 2e-16 ***
log_Cylinder_capacity  -0.0327451  0.0042862  -7.640 2.20e-14 ***
log_Value_vehicle       0.2117407  0.0047121  44.935  < 2e-16 ***
sqrt_N_doors            0.2634537  0.0028021  94.020  < 2e-16 ***
Type_fuel               0.0076722  0.0024203   3.170  0.00153 **
Length                  0.2194462  0.0223529   9.817  < 2e-16 ***
sqrt_Weight            -0.0063251  0.0003884 -16.287  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3161 on 103740 degrees of freedom
Multiple R-squared:  0.4614,    Adjusted R-squared:  0.4613
F-statistic:  4677 on 19 and 103740 DF,  p-value: < 2.2e-16
```

Figure 3: "model1" summary

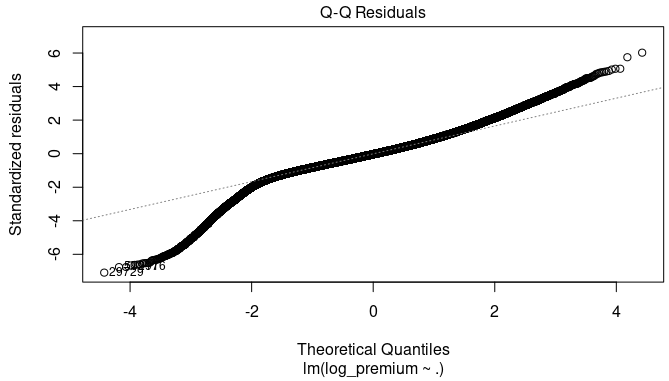

Figure 4: "model1" QQ plot

From the result of the transformed model, we see that we have significant improvements in adjusted R square. It has increased from 0.3046 to 0.4613. However, we are still seeing some normality assumption violation, but it has been improved.

After running the Breusch-Pagan test again on "model1", we have the Breusch–Pagan statistics 4157.1 and a p-value $< 2.2e-16$. Since there is a reduction on the statistics, we see that there has been some improvements in the heteroscedasticity, but we haven't resolve the issue yet.

In addition, since the p-value for F statistics is still lower than the significance level, we still have the same conclusion that that there exists statistically significant linear relationship between the response and the predictor variables.

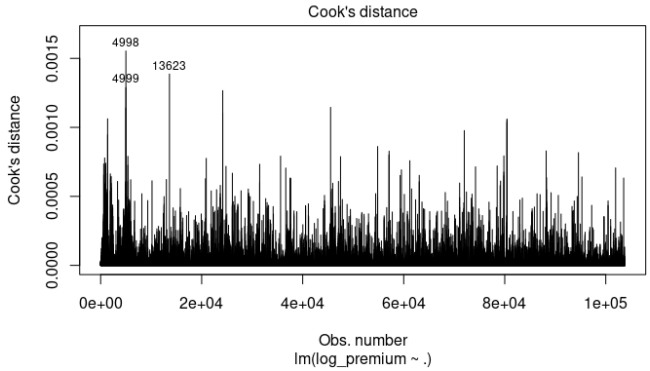

Figure 5: "model1" cook distance plot

By checking the cook distance, we verified that there is no potential influential points in "model1" that requires additional inspection.

Now, according to our methodologies, we need to perform box-cox transformations on "data5". to attempt to ease the non-normality issue.

3.3 The Box-Cox model and model diagnostics

This section introduces the results after applying the Box-Cox transformation.

```
Call:
lm(formula = ((log_premium^lambda - 1)/lambda) ~ ., data = data5)

Residuals:
     Min       1Q   Median       3Q      Max
-0.57827 -0.03915 -0.00056  0.04112  0.35463

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             1.455e+00  1.160e-02 125.494  < 2e-16 ***
log_brithday_diff       1.591e-02  1.528e-03  10.410  < 2e-16 ***
log_drive_age          -2.757e-02  7.066e-04 -39.011  < 2e-16 ***
sqrt_Seniority         -1.826e-03  2.821e-04  -6.473 9.64e-11 ***
sqrt_Policies_in_force -3.836e-02  1.217e-03 -31.511  < 2e-16 ***
sqrt_Max_policies      -1.258e-02  1.079e-03 -11.654  < 2e-16 ***
sqrt_Lapse              2.315e-02  5.599e-04  41.337  < 2e-16 ***
log_Cost_claims_year    5.511e-04  1.119e-04   4.923 8.54e-07 ***
sqrt_N_claims_history   1.279e-02  2.849e-04  44.898  < 2e-16 ***
Type_risk              -1.309e-02  6.956e-04 -18.818  < 2e-16 ***
Area                    9.172e-03  5.127e-04  17.890  < 2e-16 ***
Second_driver           3.544e-02  7.036e-04  50.367  < 2e-16 ***
sqrt_age_car           -1.523e-02  2.539e-04 -59.966  < 2e-16 ***
sqrt_Power              7.624e-03  2.296e-04  33.207  < 2e-16 ***
log_Cylinder_capacity  -3.886e-03  9.885e-04  -3.931 8.46e-05 ***
log_Value_vehicle       4.778e-02  1.087e-03  43.963  < 2e-16 ***
sqrt_N_doors            6.622e-02  6.462e-04 102.474  < 2e-16 ***
Type_fuel               1.959e-03  5.582e-04   3.510 0.000447 ***
Length                  4.860e-02  5.155e-03   9.427  < 2e-16 ***
sqrt_Weight            -1.753e-03  8.956e-05 -19.576  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0729 on 103740 degrees of freedom
Multiple R-squared:  0.4749,    Adjusted R-squared:  0.4748
F-statistic:  4938 on 19 and 103740 DF,  p-value: < 2.2e-16
```

Figure 6: Box-cox model summary

3.3.1 Residuals vs Fitted plot

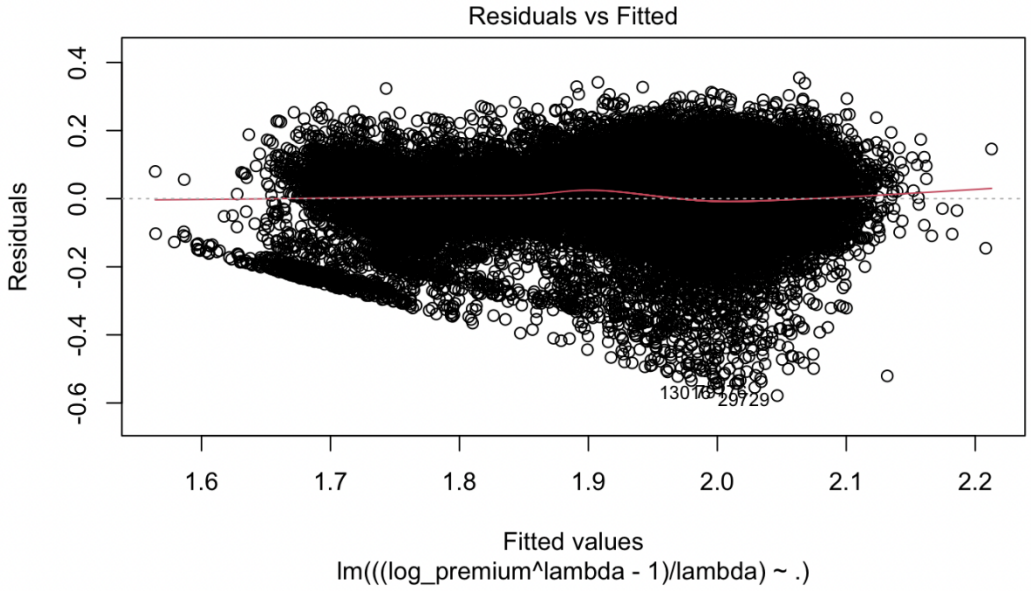

Figure 7: Residuals vs Fitted Plot: Red Line indicates Fitted Line, Grey Dashed Horizontal Line is Zero

This plot shows the standardized residuals against the leverage of each observation.

Leverage measures how far an observation's value on the predictor variables is from the mean of those predictors. Observations with high leverage have a large potential to influence the regression line. These points are further to the right on the x-axis. In this graph, a point with higher leverage is for instance, the one labeled 4998.

The y-axis represents the standardized residuals, which are the residuals divided by their standard deviation. Most of the residuals are concentrated around 0, indicating that the model is capturing the data well. However, there are some points with higher residuals (both positive and negative), indicating some data points where the model is not fitting as well.

Linearity Assumption: The red line in the graph shows the trend of residuals against leverage. Ideally, this line should be flat, indicating that the residuals are randomly scattered and that the relationship is linear. However, the slight downward trend in the red line suggests potential non-linearity, indicating that the linear model might not be fully capturing the relationship between the predictors and the response. This deviation could be a sign that the model might benefit from including polynomial terms or interactions, or it could indicate that a different model might better capture the relationship.

Constant Variance Assumption : If the variance of the residuals were constant, the spread of the points along the y-axis should be roughly the same across all values of leverage. In this graph, there appears to be a greater spread of residuals as leverage increases, particularly in the middle range of leverage values. This suggests that the assumption of homoscedasticity might be violated, with possible heteroscedasticity present in the data. Using a log transformation can address this issue.

This plot does not directly address uncorrelated error and normality of error assumption.

3.3.2 Residuals vs Leverage plot

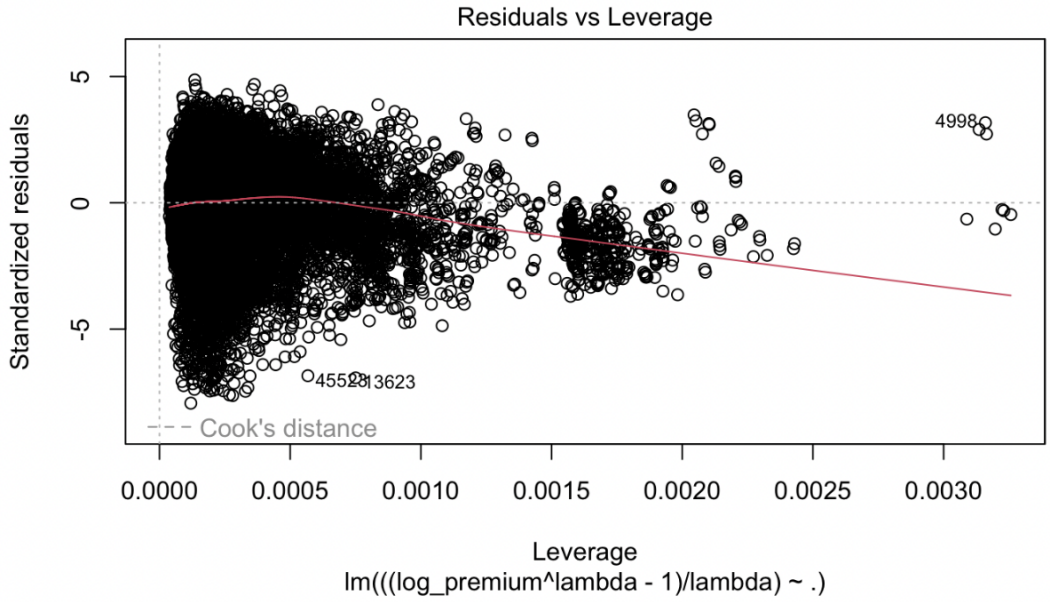

Figure 8: Residuals vs Leverage Plot: Red Line indicates Fitted Line, Grey Dashed Horizontal Line is Zero

Linearity Assumption: In a well-fitted linear model, the residuals should scatter randomly around the horizontal line at zero, with no clear pattern. While most of the residuals are centered around zero, there seems to be a slight curvature in the red line, especially at the extremes of the fitted values. This suggests potential non-linearity in the relationship between the predictors and the response variable, which means that the linearity assumption may not hold perfectly.

Constant Variance Assumption : The spread of the residuals should be roughly the same across all levels of fitted values if the assumption of homoscedasticity is met. However, there appears to be a funnel-like pattern, with a wider spread of residuals as the fitted values increase (particularly on the lower side). This pattern indicates heteroscedasticity, meaning that the variance of the errors is not constant and may increase with the fitted values.

Normality of Error Assumption : Ideally, residuals should be symmetrically distributed around zero. The bulk of the residuals are centered around zero, but there are some outliers and points that deviate significantly from zero. While this does not directly indicate non-normality, it suggests that further checks (such as a Q-Q plot or a normality test) would be prudent, especially given the presence of some large residuals. The QQ plot is subsequently plotted for a better insight.

This plot does not directly address autocorrelation.

3.3.3 Breusch-Pagan test

Constant Variance Assumption: To test heteroscedasticity in the dataset, the Breusch-Pagan test is used. A higher BP value indicates stronger evidence against the null hypothesis of homoscedasticity. The p-value suggests the probability that the observed test statistic could occur under the null hypothesis of homoscedasticity. In all three models (model, model1, and modelboxcox), the very high Breusch-Pagan test statistic and a p-value less than 2.2e-16. This indicates strong evidence against the null hypothesis, suggesting that there is heteroscedasticity present in the models. However, consider the breusch-pagan statistics has improved from 4157.1 to 4130.4, the box-cox transformation also slightly improved the heteroscedasticity problem.

3.3.4 QQ Plot

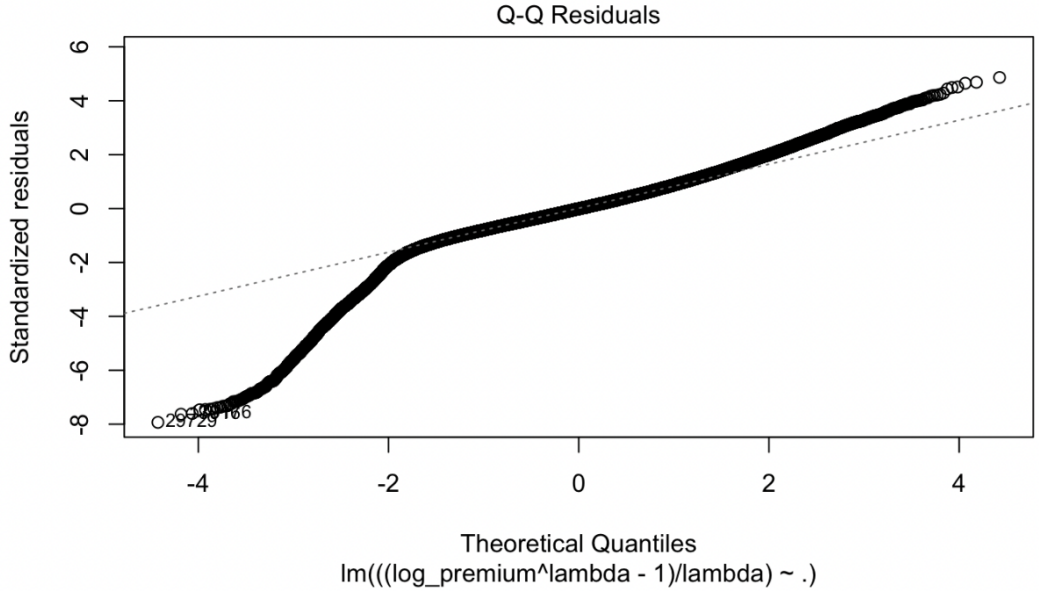

Figure 9: Residuals vs Fitted Plot: Red Line indicates Fitted Line, Grey Dashed Horizontal Line is Zero

Normality of Errors If the residuals are normally distributed, the points should fall approximately along the reference line (the diagonal dashed line).The points in the left tail (representing the lower quantiles) deviate significantly below the line. This suggests that there are more negative residuals than what would be expected under a normal distribution. Similarly, the points in the right tail, representing the higher quantiles, deviate above the line, indicating more positive residuals than expected. The points in the middle of the plot follow the line fairly well, suggesting that the residuals around the mean are approximately normally distributed.

3.3.5 R Squared & Adjusted R Squared

Multiple R-squared is 0.4749, which suggests that approximately 47.5% of the variance in the transformed premium can be explained by the model. The adjusted R-squared is 0.4748, which is very close to the Multiple R-squared, indicating that the model is not overfitted and the predictors are meaningful.

3.3.6 F Statistics

A very high F-statistic (4938) with an extremely low p-value ($<$ 2.2e-16) strongly suggests that at least one predictor in the model is statistically significant.

While a significant F-statistic indicates that the model is statistically significant, it does not guarantee that the relationship between the predictors and the response is linear. Non-linear relationships might still exist, which the linear model might not capture effectively.

3.3.7 Cook distance

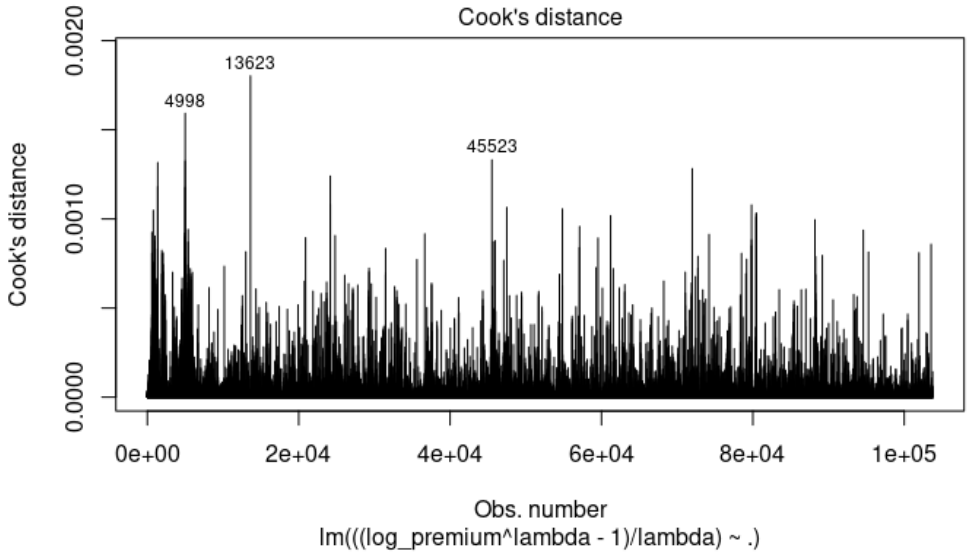

Figure 10: Enter Caption

Cook's distance was calculated to ensure that our model did not contain any high leverage points. Considering there are no points with Cooke's distance above 0.5, we conclude that our model contains no high leverage points.

3.4 The final results: applying ridge regression

So far, data regression yielded exceptionally low p values for all predictive variables indicating a statistically significant relationship between the dependent variable and all predictors; this suggested significant individual impact of every term on the response variable.

By running the VIF inflator test, results yielded that our model variables had VIF greater than 5 signifying multicollinearity issues. To improve our model, ridge regression was performed ensuring the selected model did not suffer from multicollinearity.

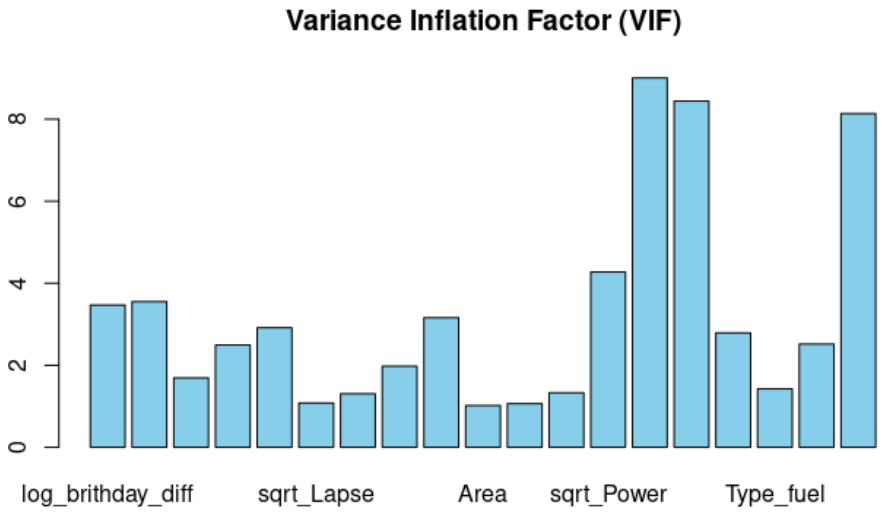

Figure 11: VIF inflator test on the Box-Cox model before ridge regression

After running the ridge regression with "lm.ridge()", we calculate the R square, adjusted R square, and the F statistics.

The coefficient of determination was calculated to be 0.4749. To account for potential overfitting the adjusted R squared was calculated, resulting in an adjusted R Squared value of 0.4748. This suggests that our model explains approximately half of the variability in the outcome. While our model provides some insight into premiums, there is still a portion of variability unaccounted for. In the context of insurance and premium pricing, an R squared of 0.4749 is not exemplary however may be acceptable considering data is complex and regulatory standards have a high impact on insurance premiums.

Performing an F test on the transformed model with 95% significance, 19 predictors and 103740 degrees of freedom yielded a critical F value of 1.5866 and an F statistic of 4937.86. Thus we can reject the null hypothesis and conclude that the final transformed model is significant in predicting premium amounts.

Our final model yielded the following coefficients:

Predictors	Coefficients
Intercept	$-1.258e-02$
sqrt_Max_policies	$-1.258e-02$
Area	$9.172e-03$
log_Value_vehicle	$4.775e-02$
log_birthday_diff	$1.590e-02$
sqrt_Lapse	$2.314e-02$
Second_driver	$3.544e-02$
sqrt_N_Doors	$6.621e-02$
log_drive_age	$-2.756e-02$
log_Cost_claims_year	$5.512e-04$
sqrt_age_car	$-1.522e-02$
Type_fuel	$1.959e-03$
sqrt_Seniority	$-1.827e-03$
sqrt_N_claims_history	$1.279e-02$
sqrt_Power	$7.625e-03$
Length	$4.861e-02$
sqrt_Policies_in_force	$-3.835e-02$
Type_risk	$-1.308e-02$
log_Cylinder_capacity	$-3.878e-03$
sqrt_Weight	$-1.752e-03$

Table 1: Ridge coefficients

It is important to recognize that the coefficient value alone does not necessarily indicate the level of impact. For instance, higher coefficients may be associated with dummy or categorical variables, while lower coefficients, such as those for annual cost of claims, may correspond to high-value continuous variables.

4 Conclusion

4.1 Discussion

Unsurprisingly, factors such as vehicle value, area, lapse, number of claims history, annual cost of claims, and the presence of a second driver on the policy all positively impact the predicted premium.

Higher vehicle value increases claim severity, which subsequently raises premiums. Additionally, claim count and the cumulative annual cost of claims directly affect insurance expenses; our model identifies these variables as highly influential in predicting premiums. Furthermore, areas with greater population density are likely to experience higher claim counts, leading to increased premiums; our model predicts that policyholders from regions with populations exceeding 300,000 face higher premiums. Lastly, the presence of a second driver on the policy introduces greater variability, which in turn results in a higher premium.

Some less intuitive variables that are observed to positively impact driver premiums include the number of doors, fuel type, vehicle length, and the power of the vehicle. Specifically, the number of doors, vehicle length, and vehicle power have a more nuanced effect on premium pricing. These factors are often associated with more expensive or higher-performance vehicle types, which can, in turn, lead to increased premiums. For instance, vehicles with diesel engines typically incur higher repair and maintenance costs; our model reflects this by showing an increase in premiums when the fuel type is diesel.

Unexpectedly, age also had a positive impact on premiums. Traditionally, as a driver's age increases, premiums tend to decrease due to the association with greater risk aversion. This anomaly could indicate a potential error in our model and will be further explored in the limitations section of the report.

Years licensed, car age, seniority, policies in force, max policies, risk type, vehicle weight, and cylinder capacity are all significant factors in premium reduction. Specifically, the number of years a driver has been licensed serves as a direct indicator of driving experience; thus, it is reasonable to anticipate that more experienced drivers present lower risk levels. Additionally, car age is typically inversely correlated with vehicular value; this suggests that as the age of the vehicle increases, premiums tend to decrease due to a reduction in the cost of potential claims.

Furthermore, companies often reward loyalty to promote retention; therefore, it is expected that seniority predicts lower premiums. Another effective tactic used by insurance companies to enhance retention and increase cross-product sales is bundling. Our model reflects this, as the number of policies in force is associated with a decrease in the premium per policy. Following this logic it follows that the maximum number of insurance policies a person can hold also decreases premiums, which is supported by our model findings.

Moreover, our model predicts that higher vehicle weight leads to decreased premiums. This conclusion is intuitive, as larger vehicles generally offer more collision protection, which can contribute to lower premiums. Conversely, and somewhat unexpectedly, our model found that increased cylinder capacity correlates with decreased premiums. Intuitively, increased cylinder capacity usually indicates enhanced performance and speed, which might increase the severity of claims.

4.2 Limitations

Final model interpretation was based on transformed variables making the direct effects of predicting variables harder to interpret.

Another limitation of this model is the questionable interpretation of driver age, theoretically the coefficient for driver age should be negative as driver premiums should decrease with age (as they do in actuality). This could be a limitation of the data set we received or a result of the transformations performed. However, all other variables performed as expected, making the effect of this minute.

Our adjusted R squared is relatively low meaning our model alone can only account for partial variability in the response variable; alone, it will not be a completely accurate prediction. This can be accredited to the nature of our dependent variable. Premiums are calculated with a plethora of factors; much of this data is sensitive and may not be included in a public data set. Moreover, premium pricing regulations vary significantly regionally; different regions offer different rate caps and also have different protective classes

for example, Ontario can use gender as a rating key whereas in British Columbia gender is a protective class.

Furthermore, some of the data suffered from high collinearity. Many variables such as length can be accredited with vehicle value and size. Ridge regression was performed to eliminate this. Further tests should be run to ensure variable selection is done accurately such as partial Ftest and cross validation methods.

Going further, It would be interesting to study the more underlying causes of premium pricing such as claim frequency and claim severity. These topics are less regulated meaning with sufficient data we may be able to provide more accurate predictive models.

5 Appendix

5.1 Detailed descriptions of the variables in the original data set

Using the head function in R, below is the first 6 rows of the original data set:

Description: df [6 × 30]

	ID <int>	Date_start_contract <chr>	Date_last_renewal <chr>	Date_next_renewal <chr>	Date_birth <chr>	Date_driving_licence <chr>	Distribution_channel <chr>	Seniority <int>	Policies_in_force <int>
1	1	05/11/2015	05/11/2015	05/11/2016	15/04/1956	20/03/1976	0	4	1
2	1	05/11/2015	05/11/2016	05/11/2017	15/04/1956	20/03/1976	0	4	1
3	1	05/11/2015	05/11/2017	05/11/2018	15/04/1956	20/03/1976	0	4	2
4	1	05/11/2015	05/11/2018	05/11/2019	15/04/1956	20/03/1976	0	4	2
5	2	26/09/2017	26/09/2017	26/09/2018	15/04/1956	20/03/1976	0	4	2
6	2	26/09/2017	26/09/2018	26/09/2019	15/04/1956	20/03/1976	0	4	2

6 rows | 1-10 of 30 columns

Figure 12: Glimpse of the original data set first 10 columns

Description: df [6 × 30]

Max_policies <int>	Max_products <int>	Lapse <int>	Date_lapse <chr>	Payment <int>	Premium <dbl>	Cost_claims_year <dbl>	N_claims_year <int>	N_claims_history <int>	R_Claims_history <dbl>
2	1	0		0	222.52	0	0	0	0
2	1	0		0	213.78	0	0	0	0
2	1	0		0	214.84	0	0	0	0
2	1	0		0	216.99	0	0	0	0
2	1	0		1	213.70	0	0	0	0
2	1	0		1	215.83	0	0	0	0

6 rows | 11-20 of 30 columns

Figure 13: Glimpse of the data set 11-20 columns

Description: df [6 × 30]

Type_risk <int>	Area <int>	Second_driver <int>	Year_matriculation <int>	Power <int>	Cylinder_capacity <int>	Value_vehicle <dbl>	N_doors <int>	Type_fuel <chr>	Length <chr>
1	0	0	2004	80	599	7068	0	P	NA
1	0	0	2004	80	599	7068	0	P	NA
1	0	0	2004	80	599	7068	0	P	NA
1	0	0	2004	80	599	7068	0	P	NA
1	0	0	2004	80	599	7068	0	P	NA
1	0	0	2004	80	599	7068	0	P	NA

6 rows | 21-30 of 30 columns

Figure 14: Glimpse of the data set 21-30 columns

Below is the detailed description of the columns of the data set:

1. **ID:** A unique identification number to each annual insurance policy hold by each insured person, a single insured person can has multiple policies, thus, their can be multiple rows starting with the same ID.
2. **Date_start_contract:** The start date of the contract with the insurance company for the insured person. It is in DD/MM/YYYY format.
3. **Date_last_renewal:** The last renewal date of the contract. It is in DD/MM/YYYY format.
4. **Date_next_renewal:** The date of the next contract renewal. It is in DD/MM/YYYY format.
5. **Date_birth:** The birthday of the insured person in the contract. it is in DD/MM/YYYY format.
6. **Date_driving_licence:** The date when the insured person receives their driver license. It is in DD/MM/YYYY format.
7. **Distribution_channel:** The way that the insured person got the contract, 1 means a broker, 0 means a agent. This is a categorical variable.
8. **Seniority:** How long the insured person has been in the relationship with the insurance company.

9. **Policies_in_force:** The total number of active policies for the insured person between Date_last_renewal to Date_next_renewal.

10. **Max_policies:** The maximum number of insurance policies a person can hold at any time, which is decided by the insurance company.

11. **Max_products:** The maximum number of insurance policies a person can hold at the same time, which is decided by the insurance company.

12. **Lapse:** The number of insurance policies that can expire if the insured person does not renewal the policy before the Date_next_renewal.

13. **Date_lapse:** The date of the policy terminates, which is the same as the Date_next_renewal. It is in DD/MM/YYYY format.

14. **Payment:** The most recent payment method used by the insured person for the premium, 1 means a half-yearly administrative process and 0 means an annual payment method.

15. **Premium:** The net fees for the specific annual insurance policy. This is going to be the response variable in the study of this project.

16. **Cost_claims_year:** The cost of claims for the insurance company during the annual insurance policy i.e. between Date_last_renewal and Date_next_renewal.

17. **N_claims_year:** The number of claims submitted during the annual insurance policy i.e. between Date_last_renewal and Date_next_renewal.

18. **N_claims_history:** The total number of claims submitted during the entire span of insurance policy.

19. **R_Claims_history:** $\frac{\text{N\_claims\_history}}{\text{Total number of years}}$

20. **Type_risk:** The type of risk for policy, specifically each value represents a different type of vehicle: 1 for motorbikes, 2 for vans, 3 for passenger cars, and 4 for agricultural vehicles.

21. **Area:** The area of the insured person stays, 1 means a urban area with more than 30000 people, and 0 means rural areas.

22. **Second_driver:** The number of drivers included in the insurance policy for the particular vehicle, 0 means only 1 driver, and 1 means more than 1 driver.

23. **Year_matriculation:** The registration year of the vehicles. It is in YYYY format.

24. **Power:** The horse power of the vehicles in the policy.

25. **Cylinder_capacity:** It is a technical term used to describe the volume of all the cylinders of the engine of the vehicle.

26. **Value_vehicle:** The market value of the vehicle on the last day of calendar year 2019.

27. **N_doors:** The number of doors on the vehicle.

28. **Type_fuel:** The type of fuel the vehicle uses, P means petrol and D means diesel.

29. **Length:** The length of the vehicle in meters.

30. **Weight:** The weight of the vehicle in kilograms.

5.2 Complete R code for the data cleaning process

```r
# let us check the number of missing values from the data set
colSums(is.na(data))

# This returns the sum of the data
sum(is.na(data))

# Filter the NAs from Type_fuel
data2 = filter(data, !is.na(Type_fuel))

# Replace the NAs in Length with its mean
data2$Length[is.na(data2$Length)] = mean(data2$Length, na.rm = TRUE)

# We check to make sure there is no NAs in data2
colSums(is.na(data2))
```

5.3 Complete R code for transforming the data set

```r
# Only run this code once. Otherwise clear the output before run
    again.

library(lubridate)

# Function to calculate the age of the insured person
person_age = function(premium_date, birthday) {
  age = as.numeric(difftime(premium_date, birthday, units = "days"))
      / 365
  return(age)
}

# Function to calculate the driving age
time_driving = function(premium_date, licence_date) {
  age1 = as.numeric(difftime(premium_date, licence_date, units = "
     days")) / 365
  return(age1)
}

# Converting to date format for lubridate
data2$Date_birth = dmy(data2$Date_birth)
data2$Date_last_renewal = dmy(data2$Date_last_renewal)
data2$Date_driving_licence = dmy(data2$Date_driving_licence)
data2$Year_matriculation = as.numeric(data2$Year_matriculation)

# Find the age of the insured person using mutate and case_when
data2 = data2 %>%
  mutate(brithday_diff = case_when(
    Payment == 0 ~ person_age(Date_last_renewal, Date_birth),
    Payment == 1 ~ (person_age(Date_last_renewal, Date_birth) +
       person_age((Date_last_renewal %m+% months(6)), Date_birth)) /
        2
  ))

# Find the driving age of the insured person
data2 = data2 %>%
  mutate(drive_age = case_when(
    Payment == 0 ~ time_driving(Date_last_renewal, Date_driving_
       licence),
```

```
    Payment == 1 ~ (time_driving(Date_last_renewal, Date_driving_
      licence) + time_driving((Date_last_renewal %m+% months(6)),
      Date_driving_licence)) / 2
  ))

# Find the vehicle age of the insured person in year
data2$Year_renewal = as.numeric(format(data2$Date_last_renewal, "%Y"
  ))
data2$age_car = data2$Year_renewal - data2$Year_matriculation

# Filer into data3
data3 = data.frame(brithday_diff = data2$brithday_diff, drive_age =
  data2$drive_age, Seniority = data2$Seniority, Policies_in_force =
   data2$Policies_in_force, Max_policies = data2$Max_policies,
  Lapse = data2$Lapse, Premium = data2$Premium, Cost_claims_year =
  data2$Cost_claims_year, N_claims_history = data2$N_claims_history
  , Type_risk = data2$Type_risk, Area = data2$Area, Second_driver =
   data2$Second_driver, age_car = data2$age_car, Power = data2$
  Power, Cylinder_capacity = data2$Cylinder_capacity, Value_vehicle
   = data2$Value_vehicle, N_doors = data2$N_doors, Type_fuel =
  data2$Type_fuel, Length = data2$Length, Weight = data2$Weight)

# Checking for negative values in data3
num_neg_per_col = function(x) {
  sum(x < 0, na.rm = TRUE)
}
sapply(data3, num_neg_per_col)

# Removing the negative values by pumping to a new data frame data4
data4 = data3 %>%
  filter(drive_age > 0)
```

Below is first 6 rows of data4:

Description: df [6 × 20]

	brithday_diff <dbl>	drive_age <dbl>	Seniority <int>	Policies_in_force <int>	Max_policies <int>	Lapse <int>	Premium <dbl>	Cost_claims_year <dbl>	N_claims_history <int>
1	59.59726	39.65479	4	1	2	0	222.52	0	0
2	60.60000	40.65753	4	1	2	0	213.78	0	0
3	61.60000	41.65753	4	2	2	0	214.84	0	0
4	62.60000	42.65753	4	2	2	0	216.99	0	0
5	61.73836	41.79589	4	2	2	0	213.70	0	0
6	62.73836	42.79589	4	2	2	0	215.83	0	0

6 rows | 1-10 of 20 columns

Figure 15: data4 glimpse part 1

Description: df [6 × 20]

Type_risk <int>	Area <int>	Second_driver <int>	age_car <dbl>	Power <int>	Cylinder_capacity <int>	Value_vehicle <dbl>	N_doors <int>	Type_fuel <chr>	Length <dbl>
1	0	0	11	80	599	7068	0	P	4.252007
1	0	0	12	80	599	7068	0	P	4.252007
1	0	0	13	80	599	7068	0	P	4.252007
1	0	0	14	80	599	7068	0	P	4.252007
1	0	0	13	80	599	7068	0	P	4.252007
1	0	0	14	80	599	7068	0	P	4.252007

6 rows | 11-20 of 20 columns

Figure 16: data4 glimpse part 2

Description: df [6 × 20]

Area <int>	Second_driver <int>	age_car <dbl>	Power <int>	Cylinder_capacity <int>	Value_vehicle <dbl>	N_doors <int>	Type_fuel <chr>	Length <dbl>	Weight <int>
0	0	11	80	599	7068	0	P	4.252007	190
0	0	12	80	599	7068	0	P	4.252007	190
0	0	13	80	599	7068	0	P	4.252007	190
0	0	14	80	599	7068	0	P	4.252007	190
0	0	13	80	599	7068	0	P	4.252007	190
0	0	14	80	599	7068	0	P	4.252007	190

6 rows | 12-21 of 20 columns

Figure 17: data4 glimpse part 3

5.4 Detailed description of the predictors

The predictors are:

1. **brithday_diff:** The age of the driver at the time of paying the premium in number of years.
2. **drive_age:** The amount of time the driver has been driving in number of years.
3. **Seniority:** How long the insured person has been in the relationship with the insurance company.
4. **Policies_in_force:** The total number of active policies for the insured person between Date_last_renewal to Date_next_renewal.
5. **Max_policies:** The maximum number of insurance policies a person can hold at any time, which is decided by the insurance company.
6. **Lapse:** The number of insurance policies that can expire if the insured person does not renewal the policy before the Date_next_renewal.
7. **Cost_claims_year:** The cost of claims for the insurance company during the annual insurance policy i.e. between Date_last_renewal and Date_next_renewal.
8. **N_claims_history:** The total number of claims submitted during the entire span of insurance policy.
9. **Type_risk:** The type of risk for policy, specifically each value represents a different type of vehicle: 1 for motorbikes, 2 for vans, 3 for passenger cars, and 4 for agricultural vehicles.
10. **Area:** The area of the insured person stays, 1 means a urban area with more than 30000 people, and 0 means rural areas.
11. **Second_driver:** The number of drivers included in the insurance policy for the particular vehicle, 0 means only 1 driver, and 1 means more than 1 driver.
12. **age_car:** The age of the vehicle in the policy since the registration until the payment of the premium in number of years.
13. **Power:** The horse power of the vehicles in the policy.
14. **Cylinder_capacity:** It is a technical term used to describe the volume of all the cylinders of the engine of the vehicle.
15. **Value_vehicle:** The market value of the vehicle on the last day of calendar year 2019.
16. **N_doors:** The number of doors on the vehicle.
17. **Type_fuel:** The type of fuel the vehicle uses, P means petrol and D means diesel.
18. **Length:** The length of the vehicle in meters.
19. **Weight:** The weight of the vehicle in kilograms.

6 Acknowledgement

Report Written by: Jiaying (Cindy) Liu, Marta Gonczar, Haoran (Alex) Yu
Code: Haoran (Alex) Yu, Malek Sibai, Mingyu (Oscar) Qin