Determinants of Paid Losses

Authors: Caspar Chen, Raja Fara Athirah Raja Hamzah, & Jenna Shen

Table of Contents

Abstract

This paper reviews determinant factors of automobile paid losses for collision from several aspects. This paper asks and answers questions such as whether physical characteristics of automobiles can explain the paid losses. Summary statistics and correlation matrix are provided with the original dataset taken from Machine Learning Repository of UCI. Multiple linear regression model serves as the primary framework for this analysis. Methods such as ANOVA, residual analysis, and multicollinearity are used to reach conclusion and further research suggestions. The results about linear regression model fit and variable selection will be discussed and explained.

Purpose

The research aims at studying the determinants of automobile paid losses for collision accidents. Paid loss is how much amount of money the insurer companies pay to the insured according to the insurance contract, after applying deductibles and limits. The research mainly focuses on the determinant factors of relevant to the physical characteristics of automobiles such as curb weight, horsepower, engine size, as well as the settle-down selling price and risk rating of automobiles. The research results can be useful for insurer companies in pricing their insurance products because they can include these characteristics of automobiles into their risk rating and pricing process. Meanwhile, with more automobiles on the road today, the collision accident exposures are continuously increasing for the insurers. Therefore, our research can also help insureds predict how much amount of losses their insurer companies can pay them if their automobiles involving into collision accidents.

Background

According to Casualty Actuarial Society (CAS), Property Casualty Insurers Association of America (PCI) and Society of Actuary (SOA), in the second half of 2013, insurers began to notice an increase in property damage liability and collision losses in the personal automobile line (2018, pp.2). An increase of 2.6% of collision frequency and 8.2% in collision severity observed in 2014 to 2016 can be directly linked to the increase of number of people driving (Insurance Information Institute [III], 2016). With this alarming increasing trend in the auto insurance

industry, accurate and sufficient premium charges are of high priority to ensure insurers remain solvent and will be able to payout the incurred losses to insureds. A common practice amongst insurers is using demographic factors, which includes age and gender, driving habits and types of coverage purchased by insureds as some of the rating factors in determining the expected loss of an insurance policy and further extended to determine the premiums for the automobile insurance (Allstate Insurance Company, 2017). In our research, we would like to explore other potential rating factors, specifically the characteristics of automobile, that can provide better estimation of insurance premiums.

Several researches had been done by other researchers that are indirectly related to the topic of our paper. According to Duan (2018), insurance company should consider using auto burden index into the traditional rate making model. Auto burden index is a method of classifying rates into different risk factors. In another study conducted by Heller and Styczyski (2016), they found that low-income and medium-income drivers are charged higher premiums, a common pricing practices among major carriers. We are unable to find any studies that is similar to our topic of interest, which is a good thing because we are doing a research on a new topic and hopefully our study can be good reference to others in the future.

Data

Data Description

Original dataset is archived from Machine Learning Repository provided by University of California Irvine's. The dataset was donated in May 1987. It is composed of two parts: data regarding the amount of losses is provided by the Insurance Collision Report by Insurance Institute for Highway Safety, whereas data regarding the physical characteristics of automobiles are provided by the 1985 Ward's Automotive Yearbook. Automobiles in this study varies in their manufacturing company, which includes Alfa-Romeo, Audi, BMW, Honda, Porsche, Toyota, Volkswagen, and others. This automobile dataset consists of entities that give the specification of an auto in terms of various characteristics. The original dataset contains 26 variables, and 205 observations in total. We intend to explore this dataset to study the factors that may influence the amount of loss in automobiles. Relevant factors provided by this dataset includes explanatory variables such as fuel-type, number of doors, wheel-base, length and height of automobiles, curb-weight, engine size, horsepower.

We have a cross-sectional data of N = 164 after excluding missing observations. Normalized losses are chosen to be the dependent variable. This variable is defined as the average loss payment per insured vehicle year. This value is normalized for all autos within a particular size classification and represents the average loss per car per year. The six explanatory variables are curb weight, engine size, horsepower, price, risk symboling, and height. The main explanatory variable is risk symboling, which is a categorical variable, which corresponds to the degree to which the auto is more risky than its price indicates. Cars are

initially assigned a risk factor symbol associated with its price. Then, if it is more risky (or less), this symbol is adjusted by moving it up (or down) the scale. Actuaries call this process "symboling". A value of +3 indicates that the auto is risky, -3 that it is relatively safer. The variable curb weight is defined as the weight of an automobile without any occupants or baggage (lb.). The engine size is measured by displacement, expressed in liters (L). The horsepower is defined as the power of an engine measured in terms of 550 foot-pounds per second (745.70 watts). The price is defined as the settle-down price of the automobiles (dollar). The variable height of automobiles is in inch. Below is the summary statistics. After checking the data, there are forty-one missing observations for normalized losses. There are two missing observations for horsepower. There are four missing observations for settled-down price. These two explanatory variables, which have missing observations overlap with our dependent variable Y (normalized losses) that has missing observations. Therefore, the database left with 164 effective observations for later research.

Summary Statistics

Below presents the table for the summary statistics of both independent and dependent variables including sample size, mean, median, standard deviation, minimum value, maximum value, and number of missing observations (which the entire rows have been deleted).

Summary Statistics[1]	Sample size	Mean	Median	Standard Deviation	Min	Max	Number of missing observation
Normalized Losses (Y)	164	122.00	115.00	35.44	65.00	256.00	41
Curb weight (X1)	164	2458.27	2367.50	475.09	1488.00	4066.00	0
Engine size (X2)	164	117.96	109.00	30.90	61.00	258.00	0
Horsepower (X3)	164	96.21	91.00	30.41	48.00	200.00	2
Price (X4)	164	11466.52	9268.50	5803.49	5118.00	35056.00	4
Symboling (X5)	164	0.79	1.00	1.23	-2.00	3.00	0
Height (X6)	164	53.77	54.10	2.34	49.40	59.80	0

From the summary statistics table of independent and dependent variables, we made the following observations:

1. Data of dependent variable is right skewed and its standard deviation is relatively large given the size of its mean and median.
2. For independent variables, data for curb weight, engine size, horsepower, price are right skewed, while the data for the risk symboling and height are left skewed.
3. The standard deviation for price of automobiles is large, which makes sense because the range for price is also very large because the dataset covering automobiles that have cheaper brands as well as luxury brands.

[1] Figure 1.1 Summary Statistics

4. For the risk symboling data, the minimum value is -2, while the maximum value is 3. This indicates that automobiles covered in this dataset have the most risky automobiles, which are indicated by 3, but does not contain the safest ones which should be indicated by -3.

Correlation

Below presents a table for correlation coefficient values between each pair of two variables:

Correlation Table[2]	Normalized losses	Curb weight	Engine size	Horsepower	Price	Symboling	Height
Normalized losses	1.00	0.12	0.17	0.30	0.20	0.53	-0.43
Curb weight	0.12	1.00	0.87	0.78	0.89	-0.25	0.36
Engine size	0.17	0.87	1.00	0.77	0.81	-0.17	0.17
Horsepower	0.30	0.78	0.77	1.00	0.76	0.02	0.01
Price	0.20	0.89	0.81	0.76	1.00	-0.14	0.22
Symboling	0.53	-0.25	-0.17	0.02	-0.14	1.00	-0.52
Height	-0.43	0.36	0.17	0.01	0.22	-0.52	1.00

For the correlation coefficient, correlation is an effect size to describe the strength of the correlation based on the guide that Evans (1996) suggests for the absolute value of r:

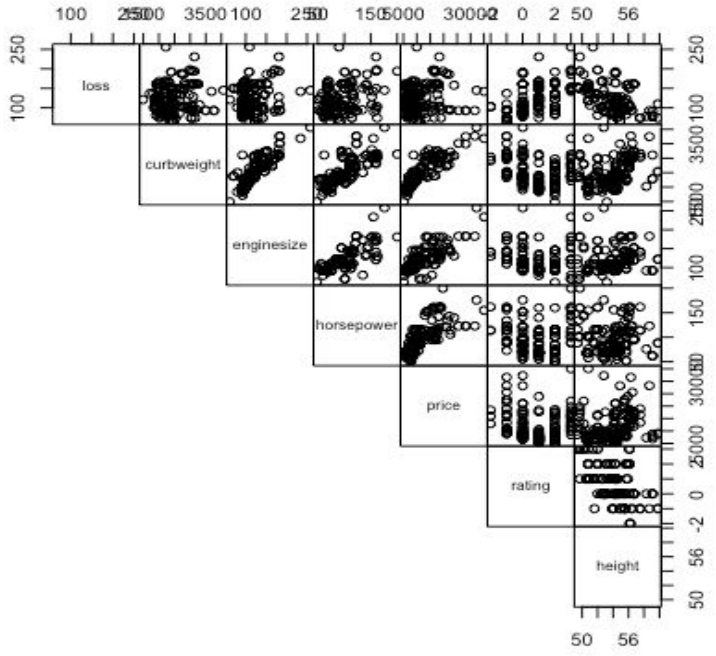
- .00-.19 “very weak”
- .20-.39 “weak”

[2] Figure 1.2 Correlation Table

- .40-.59 “moderate”
- .60-.79 “strong”
- .80-1.0 “very strong”

Therefore, according to the correlation coefficient values presented in the above table, independent variables including engine size, horsepower, and price have weak correlations with dependent variable, the paid losses, while curb-weight, risk symboling, and height have moderate correlations with paid losses.

Along with the correlation table, a correlation matrix table is also presented below. For the plot, each grid shows the correlation plot between two variables. Y axis is provided at the very left, and the x axis is provided at the very bottom.



[3]

[3] Figure 1.3 Correlation Matrix Table

According to the correlation matrix table above, we mainly focuses on the first grid line to explore the correlation between different independent variables with the dependent variable. Curb-weight, engine size, horsepower, and risk rating have positive correlations with paid losses. The plot with paid losses and rating appears in that way because the risk symboling is a categorical variable. Height is the only one that has negative correlation with paid losses. This can be interpreted in a way that when the height of automobiles increases, the paid losses will decreases.

Method

The preliminary regression model is as follows:

$$Y = \beta o + \beta 1*X1 + \beta 2*X2 + \beta 3*X3 + \beta 4*X4 + \beta 5*X5 + \beta 6*X6 + \beta 7*X1*X4 + \mathcal{E}$$

In words, the model is expressed as:

$$E(\text{Paid losses}) = \beta o + \beta 1*\text{Curb-Weight} + \beta 2*\text{EngineSize} + \beta 3*\text{Horsepower} + \beta 4*\text{Price} + \beta 5*\text{Symboling} + \beta 6*\text{Height} + \beta 7*\text{Curb-Weight}*\text{Price} + \mathcal{E}$$

Linear Regression Model

We performed a number of regression models in search of the best model to fit our data. The purpose of this is to find values of residual standard deviation (s), coefficient of determination (R^2), coefficient of determination adjusted for degrees of freedom (Ra^2), estimates of β and statistical significance of each independent variables.

In Model I, we only include 5 independent variables which are curb weight, engine size, horsepower, rating and price in exploring the linear relationship between these variables and the dependent variable which is the paid losses. The intuition behind this is that these variables are the most relevant factors that stand out as to the quality and risk-levels of the automobiles which can directly affect the amount of paid losses when automobiles are involved in collision.

In Model II, we added another independent variable, height, to improve the goodness of fit of our data. Among all the independent variables, height is the only variable with a negative correlation with paid losses.

For Model III, we added an interaction term to see if the expected Y per unit change in an independent variable, x_1, depends on another independent variable, x_2. After testing for multiple combinations of interaction terms, we settled for the interaction between curb weight and price.

Analysis of Variance (ANOVA)

The purpose of the ANOVA table is to keep track of the sources of variability. We are particularly interested in mean square error (MSE) and the F-ratio to perform the F-test to make statements about the statistical significance of the independent variables.

Residual Analysis

In linear regression, a residual is a response minus the corresponding fitted value under the model. If the model is an adequate representation of the data, residuals should closely approximate random errors. We plotted 4 different graphs to help us analyze the residuals. We used *residual vs fitted graph* and *scale-location graph* to test for heteroskedasticity. Then, we performed a qq-plot to see if our data is normally distributed. Finally we have the *residuals vs leverage graph* which uses the Cook's distance to find outliers and high leverage points that we may have in our data. For the Cook's Distance graph, we observed several points that could potentially be leverage points according to the concept from matrix algebra. Therefore, we used the formula derived from matrix algebra to check the high leverage points for multiple linear regression.

Multicollinearity

Multicollinearity occurs when one explanatory variable is, or nearly is, a linear combination of the other explanatory variables. Here, we would like to test for the multicollinearity. To capture the relationships among several variables, we used the Variance Inflation Factor (VIF). It is defined as $VIF_j = \frac{1}{1 - R_j^2}$, for j = 1, 2, ..., k.

Results

Linear Regression Model Fit

Below presents the R^2 and adjusted R^2 values for the three linear regression models we selected before to fit our data:

[4]	R^2	Ra^2
Model I	0.37	0.35
Model II	0.42	0.40
Model III	0.47	0.44

We select our linear regression model mainly based on R^2 and adjusted R^2. R^2 is calculated by regression sum of squares divided by total sum of squares as follows:

$R^2 = RSS/TSS = 1 - ESS/TSS$

Adjusted R^2 is calculated by:

$Ra^2 = 1- \{ESS/[n-(k+1)]\}/ [TSS/(n-1)]$

Ra^2 is a better measure of goodness of fit than R^2 because R^2 never decrease when an explanatory variable is added to the model. If adding a completely nonsense variable, R^2 will stay the same, but Ra^2, even if having same or even smaller ESS, but smaller degree of freedom (trade-off between adding a variable and smaller degree of freedom), Ra^2 can be smaller.

For the above three linear regression models, Model III is clearly the best fit because its R^2 and Ra^2 are both significantly larger than those of Model I and Model II. Larger R^2 value indicates better fit of the regression model for the data. The R^2 of the linear regression Model

[4] Figure 1.4 Linear Regression Model Fit Table

III is 0.47 and the adjusted R^2 is 0.44, which indicates that 44% of the total errors can be explained by the linear regression model, or the explanatory variables of the regression model.

Below presents the table with coefficient estimates for β of the linear regression model, along with the standard error of the estimates, and t values for the t-test:

Linear Regression for Model III[5]					
Coefficients:					
	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	2.82E+02	6.08E+01	4.64	7.32E-06	***
curbweight	3.42E-02	1.51E-02	2.266	0.024839	*
enginesize	1.63E-01	1.55E-01	1.048	0.296265	
horsepower	-1.70E-01	1.46E-01	-1.169	0.244278	
rating	1.22E+01	2.02E+00	6.055	1.01E-08	***
price	9.76E-03	2.64E-03	3.703	0.000295	***
height	-5.35E+00	1.21E+00	-4.434	1.74E-05	***
curbweight*price	-2.644e-06	7.54E-07	-3.509	0.000588	***

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

Linear regression table provides the estimates for β and the estimate equation for the linear regression model is presented as below:

$$\hat{Y}_i = 282.3 + 0.034*\text{Curb weight} + 0.16*\text{Engine size} - 0.17*\text{Horsepower} + 12.23*\text{Rating} + 0.0097*\text{Price} - 5.35*\text{Height} - 0.0000026*\text{Curb weight}*\text{Price}$$

The estimates for β are presented in the following table:

[5] Figure 1.5 Linear Regression of Model III Results Table

[6] b_0	b_1	b_2	b_3	b_4	b_5	b_6	b_7
282.3	0.034	0.16	-0.17	12.23	0.0097	-5.35	-0.0000026

From the linear regression model analysis, we can also conclude that risk rating, price, height, and the interaction term of the curb-weight and price, are strongly statistically significant variables.

Analysis of Variance Table

Below presents the ANOVA Table for the linear regression model:

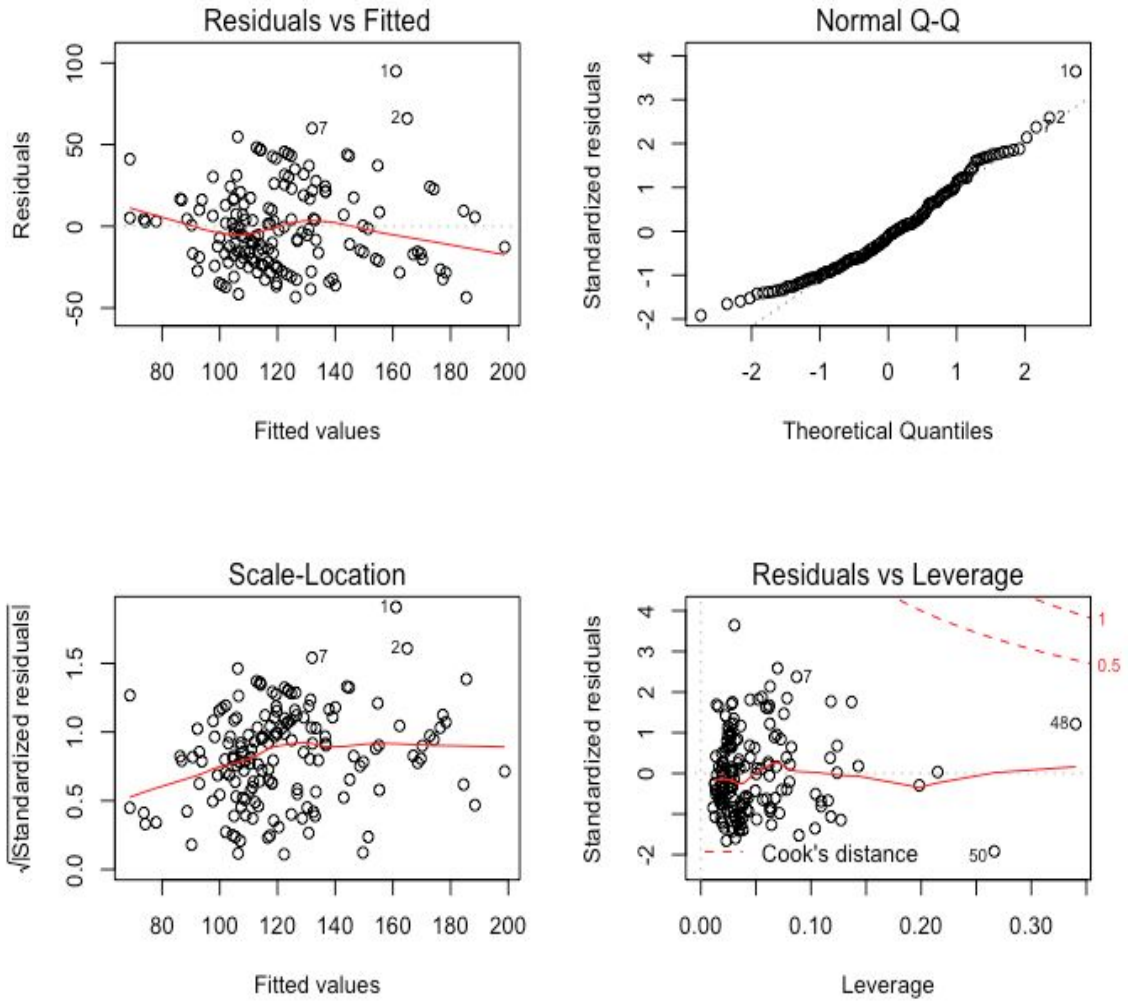
Analysis of Variance[7]						
Response: loss						
	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
curbweight	1	2943	2943	4.2042	0.0419973	*
enginesize	1	3326	3326	4.7509	0.0307816	*
horsepower	1	18595	18595	26.5627	7.62E-07	***
rating	1	49940	49940	71.3372	1.97E-14	***
price	1	1708	1708	2.4405	0.1202621	
height	1	10411	10411	14.8723	0.0001679	***
curbweight*price	1	8621	8621	12.3142	0.0005877	***
Residuals	156	109208	700			

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

[6] Figure 1.6 Linear Regression Coefficients Table
[7] Figure 1.7 ANOVA Table

Residual Analysis

For the residual analysis, we conduct four plots to test the goodness of fit of our linear regression model for the data. Below presents the four plots:



8

[8] Figure 1.8 Residual Analysis Plots

The residuals versus fitted values plot. The scale versus location plot is the standardized residuals versus fitted values plot. These two plots show that our data is heteroscedasticity, which means that variance of estimate of Y and variance of estimate of epsilon are both non-constant.

The normal QQ-plot shows that our data follows normal distribution in the center, but there are some skewness at the two ending points.

The residuals versus leverage plot shows that within the 0.5 Cook's Distance criteria, our data does not have any outliers or high leverage points.

Multicollinearity

When checking for multicollinearity using Variance Inflation Factor (VIF), the larger R-squared, the larger VIF the explanatory variable will have.

Explanatory Variables	Variance Inflation Factor (VIF)[9]
Curb Weight	11.96
Engine Size	5.35
Horsepower	4.57
Price	54.52
Symboling	1.43
Height	1.86
Curb-weight*Price	67.74

Analyzing the magnitude of multicollinearity by considering the size of the VIF. A rule of thumb is that if VIF >10, then multicollinearity is high and severe. From the table, we can tell Curb-weight has a severe multicollinearity.

[9] Figure 1.9 VIF Table

Conclusion

Our research indicates that our data fits a linear regression model moderately. We determined that height and risk rating are the most statistically significant explanatory variables in our model. Other explanatory variables included in our model includes curb weight, engine size, horsepower, and price. We also included an interaction term in our model to get a better fit which is the interaction between curb weight and price.

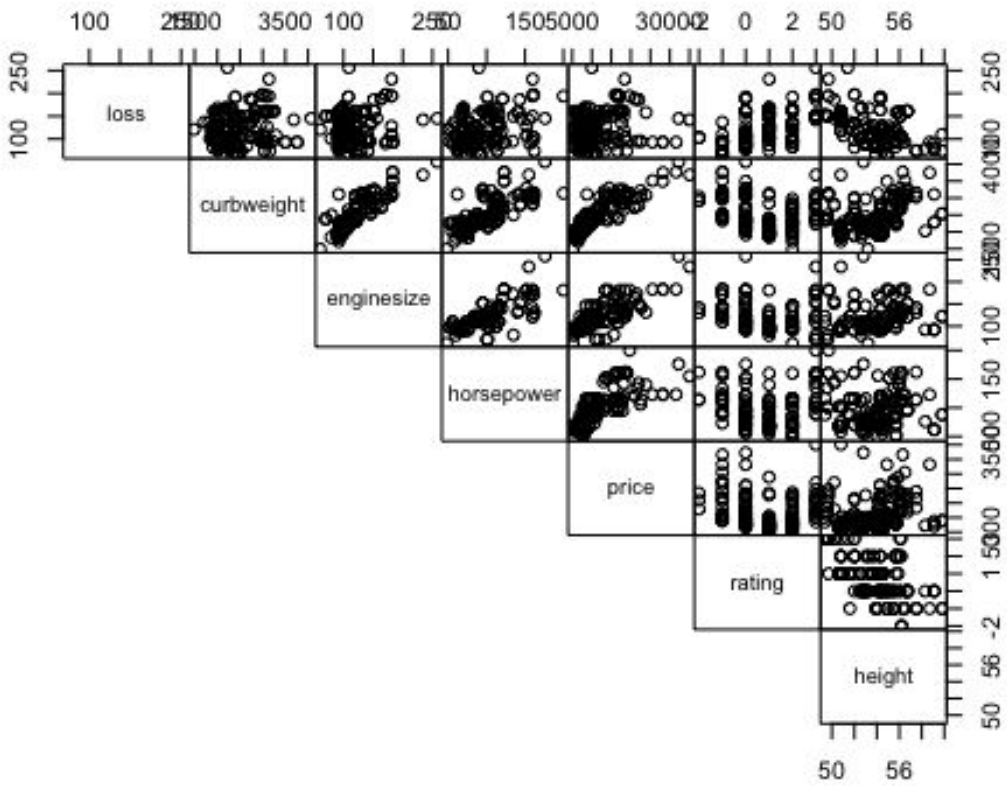
In order to obtain a better fit of the regression model, we believe that more significant variables should be included such as drivers' information which refers to ages, driving years, marital status and others. However, these variables are not readily available in this dataset. Another possible implication for the moderate fit of a linear regression model is that there might be a better model to fit our data such as a quadratic trend and other related models. Therefore, further research need to be conducted.

Appendix A: R code

```
rm(list=ls())
#load data
Auto<-read.csv("AutomobileData.csv")

#Y: Automobile normalized losses
loss<-Auto$normalized.losses
curbweight <-Auto$curb.weight
enginesize <-Auto$engine.size
horsepower <-Auto$horsepower
price <-Auto$price
rating <-Auto$symboling
height <- Auto$height

#summary statistics of dependent and explanatory variable
library(psych)

## Warning: package 'psych' was built under R version 3.4.4

describe(Auto)

##                   vars   n     mean      sd median  trimmed     mad    min
## normalized.losses    1 164   122.00   35.44  115.0   119.51   35.58   65.0
## curb.weight          2 164  2458.27  475.09 2367.5  2418.92  505.57 1488.0
## engine.size          3 164   117.96   30.90  109.0   113.78   19.27   61.0
## horsepower           4 164    96.21   30.41   91.0    92.76   32.62   48.0
## price                5 164 11466.52 5803.49 9268.5 10571.14 3591.60 5118.0
## symboling            6 164     0.79    1.23    1.0     0.77    1.48   -2.0
## height               7 164    53.77    2.34   54.1    53.73    2.37   49.4
##                     max   range skew kurtosis     se
## normalized.losses   256.0   191.0 0.75     0.43   2.77
## curb.weight        4066.0  2578.0 0.79     0.17  37.10
## engine.size         258.0   197.0 1.39     2.62   2.41
## horsepower          200.0   152.0 0.88     0.22   2.37
## price             35056.0 29938.0 1.56     2.50 453.18
## symboling             3.0     5.0 0.10    -0.64   0.10
## height               59.8    10.4 0.14    -0.43   0.18

#relationship among the variables
cor(Auto)

##                   normalized.losses curb.weight engine.size horsepower
## normalized.losses         1.0000000   0.1198927   0.1673650 0.29577197
## curb.weight               0.1198927   1.0000000   0.8678938 0.78431981
## engine.size               0.1673650   0.8678938   1.0000000 0.76861770
## horsepower                0.2957720   0.7843198   0.7686177 1.00000000
```

```
## price                       0.2032542   0.8917505   0.8075661 0.75903994
## symboling                   0.5286667  -0.2462814  -0.1693416 0.01630120
## height                     -0.4323348   0.3586628   0.1745565 0.01120426
##                          price  symboling      height
## normalized.losses  0.2032542  0.5286667 -0.43233484
## curb.weight        0.8917505 -0.2462814  0.35866281
## engine.size        0.8075661 -0.1693416  0.17455654
## horsepower         0.7590399  0.0163012  0.01120426
## price              1.0000000 -0.1440779  0.22493642
## symboling         -0.1440779  1.0000000 -0.51641976
## height             0.2249364 -0.5164198  1.00000000
```

```
#pairwise plots
pairs(cbind(loss, curbweight, enginesize, horsepower, price, rating, height),
lower.panel = NULL , gap = 0)
```




```
#Linear Regression Model I
Model.1 <- lm(loss~curbweight+enginesize+horsepower+rating+price)
summary(Model.1)
```

```
##
## Call:
```

```
## lm(formula = loss ~ curbweight + enginesize + horsepower + rating +
##     price)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -55.819 -18.128  -5.415  12.264 101.475
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 90.7609868 19.0247404   4.771 4.15e-06 ***
## curbweight  -0.0103468  0.0133811  -0.773    0.441
## enginesize   0.0753350  0.1524120   0.494    0.622
## horsepower   0.2178946  0.1323099   1.647    0.102
## rating      15.4000621  2.0156410   7.640 1.96e-12 ***
## price        0.0012748  0.0008787   1.451    0.149
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28.49 on 158 degrees of freedom
## Multiple R-squared:  0.3737, Adjusted R-squared:  0.3539
## F-statistic: 18.85 on 5 and 158 DF,  p-value: 1.127e-14
```

```
#Linear Regression Model II (addind an additional variable HEIGHT)
Model.2 <- lm(loss~curbweight+enginesize+horsepower+rating+price+height)
summary(Model.2)
```

```
##
## Call:
## lm(formula = loss ~ curbweight + enginesize + horsepower + rating +
##     price + height)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -59.832 -18.550  -5.598  16.531 100.205
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.128e+02  6.235e+01   5.016 1.41e-06 ***
## curbweight   1.298e-02  1.431e-02   0.907 0.365770
## enginesize  -3.626e-02  1.496e-01  -0.242 0.808765
## horsepower   6.924e-02  1.333e-01   0.519 0.604291
## rating       1.248e+01  2.090e+00   5.972 1.51e-08 ***
## price        9.699e-04  8.489e-04   1.143 0.254964
## height      -4.576e+00  1.229e+00  -3.725 0.000272 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.4 on 157 degrees of freedom
```

```
## Multiple R-squared:  0.4245, Adjusted R-squared:  0.4025
## F-statistic:  19.3 on 6 and 157 DF,  p-value: < 2.2e-16
```

```
#Linear Regression Model III (adding an interaction term)
Model.3 <-
lm(loss~curbweight+enginesize+horsepower+rating+price+height+curbweight*price
)
summary(Model.3)
```

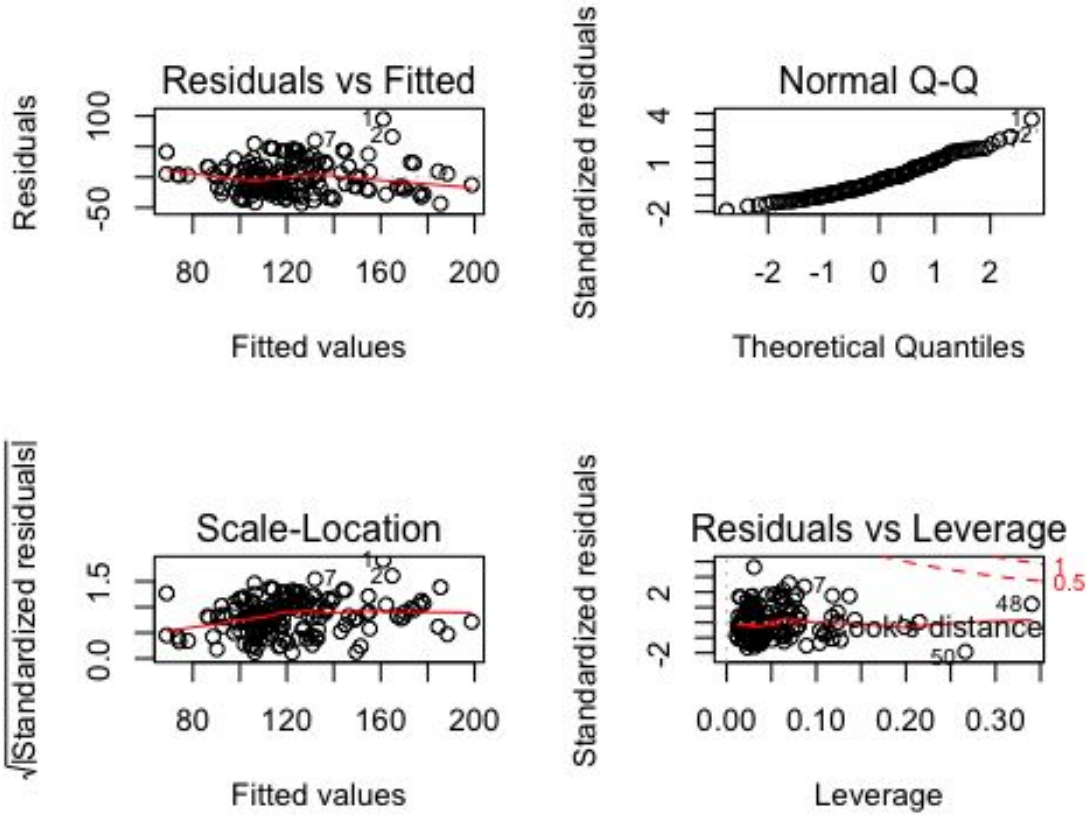
```
##
## Call:
## lm(formula = loss ~ curbweight + enginesize + horsepower + rating +
##     price + height + curbweight * price)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -43.459 -20.005  -3.044  16.894  94.930
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2.823e+02  6.084e+01   4.640 7.32e-06 ***
## curbweight        3.418e-02  1.508e-02   2.266 0.024839 *
## enginesize        1.626e-01  1.552e-01   1.048 0.296265
## horsepower       -1.704e-01  1.458e-01  -1.169 0.244278
## rating            1.223e+01  2.020e+00   6.055 1.01e-08 ***
## price             9.764e-03  2.637e-03   3.703 0.000295 ***
## height           -5.352e+00  1.207e+00  -4.434 1.74e-05 ***
## curbweight:price -2.644e-06  7.535e-07  -3.509 0.000588 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26.46 on 156 degrees of freedom
## Multiple R-squared:  0.4666, Adjusted R-squared:  0.4427
## F-statistic:  19.5 on 7 and 156 DF,  p-value: < 2.2e-16
```

```
#ANOVA table
anova(Model.3)
```

```
## Analysis of Variance Table
##
## Response: loss
##                   Df Sum Sq Mean Sq F value    Pr(>F)
## curbweight         1   2943    2943  4.2042 0.0419973 *
## enginesize         1   3326    3326  4.7509 0.0307816 *
## horsepower         1  18595   18595 26.5627 7.620e-07 ***
## rating             1  49940   49940 71.3372 1.968e-14 ***
## price              1   1708    1708  2.4405 0.1202621
## height             1  10411   10411 14.8723 0.0001679 ***
```

```
## curbweight:price   1   8621    8621 12.3142 0.0005877 ***
## Residuals        156 109208     700
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#model plot
par(mfrow=c(2,2))
plot(Model.3)
```




```
#Variance Inflation Factor (VIF)
library(faraway)
```

```
##
## Attaching package: 'faraway'
```

```
## The following object is masked from 'package:psych':
##
##     logit
```

```
autodata <- data.frame(curbweight, enginesize, horsepower,price, rating,
height, curbweight*price)
vif(autodata)
```

```
##         curbweight         enginesize         horsepower
##           11.956393           5.353476           4.574562
##               price             rating             height
##           54.522153           1.427994           1.863551
## curbweight...price
##           67.750998
```

References

Automobile Dataset (1987). *Machine Learning Repository*. Retrieved from https://archive.ics.uci.edu/ml/datasets/Automobile

Casualty of Actuary, Property Casualty Insurers Association of America & Society of Actuary (2018). Auto Loss Cost Trends Report. Retrieved from https://www.casact.org/cms/files/AutoLossCost1Q_20180130_1.pdf

Duan, et al. (2018). A Logistic Regression Based Auto Insurance Rate-Making Model Designed for the Insurance Rate Reform, *International Journal of Financial Studies, 6.*

Heller, D. & Styczynski, M. (2016). Major Auto Insurers Raise Rates Based on Economic Factors, Low- and Moderate-income Drivers Charged Higher Premiums. Retrieved from http://consumerfed.org/wp-content/uploads/2016/06/6-27-16-Auto-Insurance-and-Economic-Status_Report.pdf

Insurance Information Institute (2016). More Accidents, Larger Claims Drive Costs Higher. Retrieved from https://www.iii.org/sites/default/files/docs/pdf/auto_rates_wp_092716-62.pdf

Statements

We certify that the work is original and that we complete it on our own.

We agree that our report can be shown to future students as examples.