# 

# 

# 

# **Determinants of Automobile Paid Losses**

## 

## 

## **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 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. Actuarians 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 baggages (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]](#footnote-1) | Sample size | Mean | Median | Standard Deviation | Min | Max | Number of missing obs |
| 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.
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:

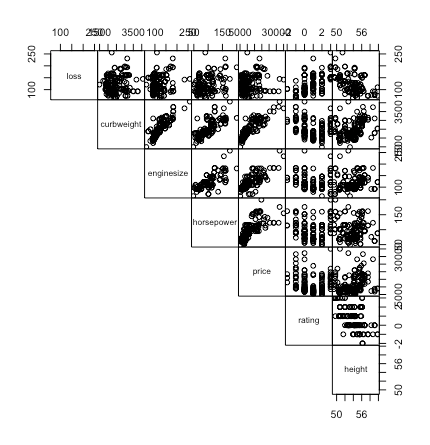
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Normalized  losses | Curb  weight | Engine  size | Horsepower | Price | Symboling | Height |
| Normalized  losses | 1.00 | 0.12 | 0.17 | 0.30 | 0.20 | 0.53 | -0.43 |
| Curb  weight | 0.12 | 1.00 | 0.87 | 0.78 | 0.89 | -0.25 | 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”
* .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.

[[2]](#footnote-2)

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 = βo + β1\*X1 + β2\*X2 + β3\*X3 + β4\*X4 + β5\*X5 + β6\*X6 + β7\*X1\*X4 + Ɛ

In words, the model is expressed as:

E(Paid losses) = βo + β1\*CurbWeight + β2\*EngineSize + β3\*Horsepower + β4\*Price + β5\*Symboling + β6\*Height + β7\*CurbWeight\*Price + Ɛ

**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 (), coefficient of determination adjusted for degrees of freedom (), 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, , depends on another independent variable, . 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 severeness of the multicollinearity. To capture the relationships among several variables, we used the Variance Inflation Factor (VIF). It is defined as = , for *j* = 1, 2, …, k.

## 

## **Result**

### **Linear Regression Model Fit**

Below presents the R² and adjusted R² values for the three linear regression models we selected before to fit our data:

|  |  |  |
| --- | --- | --- |
|  | R² | Ra² |
| 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 R² and adjusted R². R² is calculated by regression sum of squares divided by total sum of squares as follows:

R² = RSS/TSS = 1 - ESS/TSS

Adjusted R² is calculated by:

Ra² = 1- {ESS/[n-(k+1)]}/ [TSS/(n-1)]

Ra² is a better measure of goodness of fit than R² because R² never decrease when an explanatory variable is added to the model. If adding a completely nonsense variable, R² will stay the same, but Ra², even if having same or even smaller ESS, but smaller df (trade-off between adding a variable and smaller df), Ra² can be smaller.

For the above three linear regression models, Model III is clearly the best fit because its R² and Ra² are both significantly larger than those of Model I and Model II. Larger R² value indicates better fit of the regression model for the data. The R² of the linear regression Model III is 0.47 and the adjusted R² 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 | | | | | |
| 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:

Ŷi= 282.3 + 0.034\*Curb weight + 0.16\*Engine size - 0.17\*Horsepower + 12.23\*Rating +   
 0.0097\*Price - 5.35\*Height - 0.0000026\*Curb weight\*Price

The estimates for β are presented in the following table:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |
| 282.3 | 0.034 | 0.16 | -0.17 | 12.23 | 0.0097 | -5.35 | -0.0000026 |

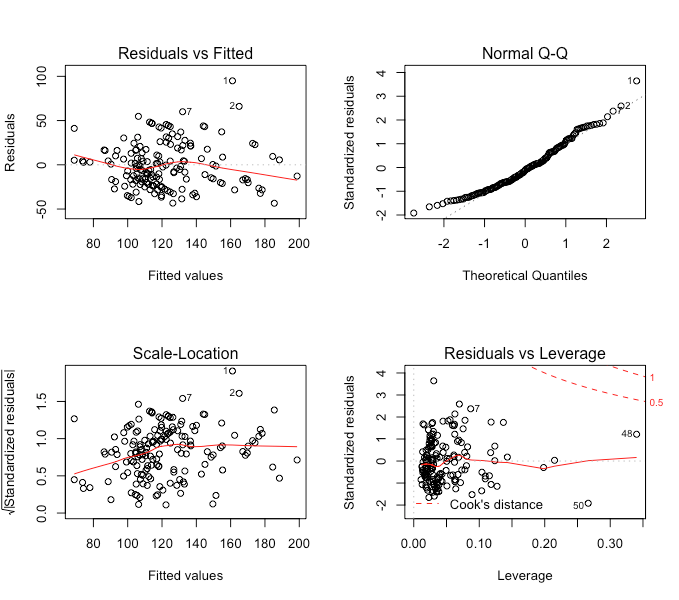
From the linear regression model analysis, we can also conclude that risk rating, price, height, and the interaction term of the curb-weight and price, are strongly statistically significant variables.

### **Analysis of Variance Table**

Below presents the ANOVA Table for the linear regression model:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Analysis of Variance | | | | | | |
| 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 |

**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:

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 heteroskedasticity, 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)[[3]](#footnote-3) |
| Curb Weight | 10.0401 |
| Engine Size | 4.6404 |
| Horsepower | 3.5702 |
| Price | 5.2715 |
| Symboling | 1.4261 |
| Height | 1.8011 |

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.

## **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 is needed to be conducted.

Statement

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

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

1. Figure 1.1 Summary Statistics [↑](#footnote-ref-1)
2. Figure 1.2 Correlation Matrix Table [↑](#footnote-ref-2)
3. Figure 1.4 VIF Table [↑](#footnote-ref-3)