James Taylor

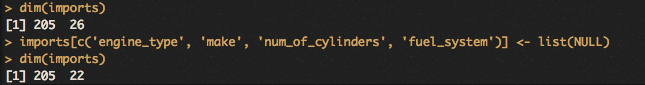
Exercise 4

Linear Regression

**Introduction**

The dependent variable for the imports-85 dataset is the price pf the vehicle. There are 25 independent variables, including make, fuel-type, horsepower, fuel consumption, number of doors and more. I expect the regression to predict the price of the vehicle quite well. Many of the variables will need to be transformed so the regression equation can use them. I believe the R2 value will be around .75, but if the make is transformed to be included in the regression I think the R2 will be closer to .8 value.

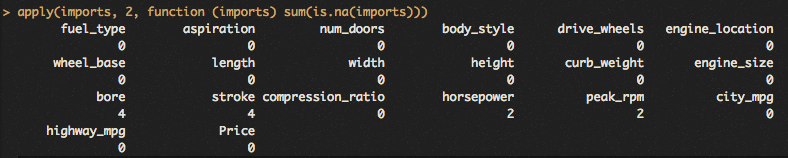
**Data Processing**



Above is the command removing unwanted variables. The dimensions of the data frame decreased from 26 variables to 22.

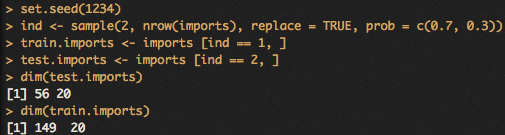
../../Desktop/2.png

Above is some additional preprocessing, removing two more variables. These variables are created and uses by actuarians to determine insurance risk. This is not something the purchaser would know, so it should not be in the model that is meant to predict how much a customer is willing to pay for particular car.



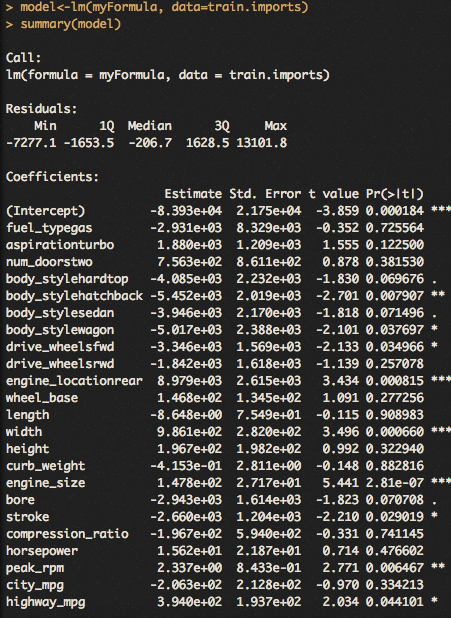
Above is the output for missing values for each variable that is left. I did not deal with missing values because the model that was used omits missing values on its own, solving the issue. The lm function also handles categorical values automatically, so there is no need to convert factors into zeros and ones.

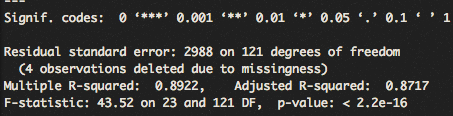
**Training and Test data**



Above are the commands separating the dataset into a training and test set. This is important for validating the accuracy of the model. The commands randomly assigns each observation to either test or train sample. The parameters specify that the train sample will be 70% and test sample 30% of the original dataset.

**Build Model**





Above is the summary of the linear model built using the training set of observations. The model is a linear regression, so it looks for linear relationships between the independent variables and the dependent one. Each value in the independent variables may influence the dependent one depending on the data and math. The ‘Estimate’ column is meant to show the influence that one value has on the dependent variable, so a one unit increase in horsepower correspond to a $15 dollar increase in price using the above model. Categorical variables were automatically handled by the model as Boolean values. For example, for drive wheels variable, each observation has a true or false for front-wheel driving and rear-wheel driving. When one is true, the other is false and doesn’t effect the dependent variable. The residuals is the difference between the model’s predicted value and the real value in the observed data. The intercept is the value that is added or subtracted once all the independent variables are considered. The p-value is the statistical significance attributed to a variable’s influence on the dependent variable. The overall accuracy can be seen in the R2 value, which is the percentage of the variation in the dependent variable the model can explain for. This models explains for 89% which is quite good.

**Evaluate Model on Test Set**

../../Desktop/7.png



The above figure shows the difference in the prediction and the actual value for the test dataset. The vertical distance between the point and the line is the residual value, or how far the model was off from the observed/true value.

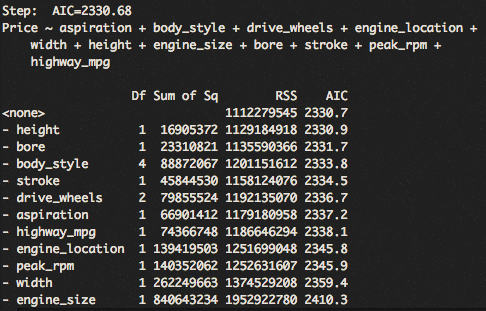
**Residuals Plots**



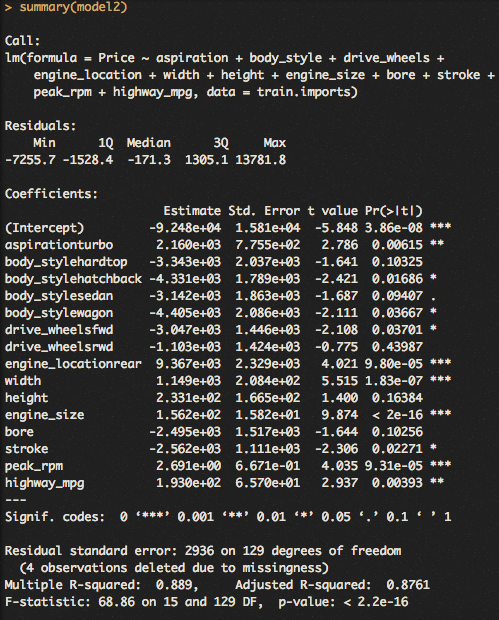
The above figure shows the residuals across the fitted values. The red line represents the average of the residuals. The points vertical difference from the grey dashed line shows the residuals. As the plot indicates, the model overestimates values below 10,000, and underestimates values between 10,000 and 30,000.

**Minimal adequate mode**

**../../Desktop/8.png**

****

Above is the minimal adequate model. Originally, it is the same linear model as above, but with statistically non-significant variables removed from it. This is helpful because it can greatly simplify a model without sacrificing much accuracy.

****

The variables fuel type, number of doors, wheel base, length, curb weight, compression ratio, horsepower, and city mpg were all removed for the minimal adequate model. The coefficients are how the variable effects the dependent variable, all else equal. The intercept is what is added or subtracted to the predicted dependent value after all variables and their coefficients have been calculated for. This model’s R2 is only 1% less than the original model, all while having eight less variables in it.

**How to use model for new car**

The way to use this model to predict a new car is like any equation. For example, looking at the top the first variable of the model is if it aspirated by a turbo, if true, the estimated value sits at $2,160. Next the body style, but here no matter what body style the vehicle is, one would subtract from $2,160, by how much depends on which body style the new vehicle actually is. While working with numerical variables, like peak rpm’s, one neess to multiply the peak rpm’s pf the new vehicle by the estimated coefficient for that variable which is 2.69 units. That means for every rpm increase, the value of the car is expected to increase by two dollars. Sum all of these for each variable, then add or subtract the intercept and you should arrive at the final estimated price of this new car.

**Summary**

A linear regression is very capable of predicating the dependent variable in this dataset. Based on how residual plot on the test sample, it predicts the values quite accurately.

I was most challenged by splitting the data randomly into test and training data. Its definitely common practice, I just wasn’t aware of the process. It was nice not needing to add an extra variable in the dataset to divide them up, which is what my assumption was before practicing in the earlier portion.