Jonathan Jimenez

5/4/2020

Professor Assiss

Final Paper

I have decided to use the Auto data frame from the ISLR package that has 392 vehicles and 9 variables. The 9 variables that are included in this data frame:

- Mpg(data.frame): miles per gallon
- Cylinders(numeric): Number of cylinders between 4 and 8
- Displacement(numeric): Engine displacement (cu. inches)
- Horsepower(numeric): Engine horsepower
- Weight(numeric): Vehicle weight (lbs.)
- Acceleration(numeric): Time to accelerate from 0 to 60 mph (sec.)
- Year(numeric): Model year (modulo 100)
- Origin(numeric): Origin of car (1. American, 2. European, 3. Japanese)
- Name(factor): Vehicle name

I will be using the Auto data frame to predict Horsepower using the features displacement, and acceleration. To see if these features have a correlation with horsepower, I plotted each individual feature with horsepower so I can visualize the data(Fig1,Fig2). As you can see below, these 2 features show a linear correlation with horsepower, which supports the idea of being able to make an accurate linear regression model that predicts horsepower based on the acceleration, and displacement. Although the graphs do show correlation, that does not mean the features are a good choice to solve this particular problem. To find out if these features can help predict Horsepower, statistical values will need to be evaluated as well as testing the model to see if it can accurate represent the data points.

Fig1:

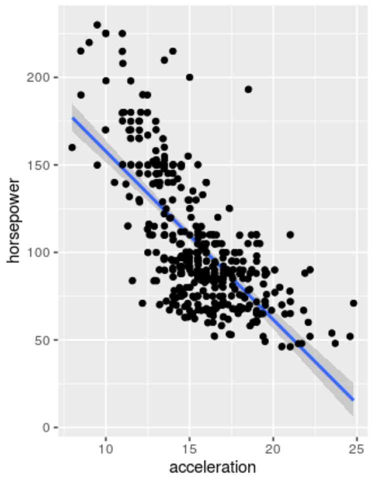
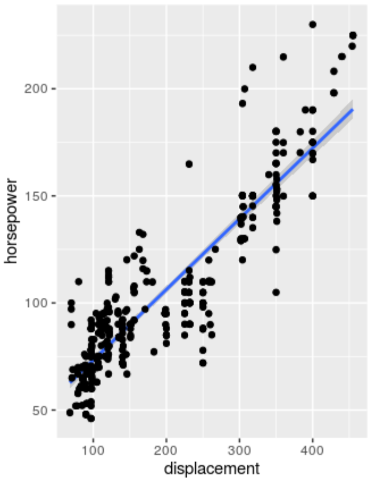


Fig2:



Interpret Statistical Values:

Call: lm(formula = horsepower ~ acceleration + displacement, data = Auto)

	Estimate	Std. Error	T value	Pr(>|t|)
(Intercept)	113.383963	5.926778	19.13	<2e-16
Acceleration	-3.987096	0.312461	-12.76	<2e-16
Displacement	0.272875	0.008238	33.12	<2e-16

I fit a multi-linear regression model to show the statistical values to interpret if the features are useful. The least square regression line can be calculated using the formula Horsepower= 113.38 – (3.99*Acceleration) + (0.27*Displacement). This formula represents an estimate of the best linear approximation to the true relationship between Horsepower and the included features. Along with these values, the test statistics are also high which represents support towards rejecting the Null hypotheses. The Standard Error is the amount of error that will usually occur but acceleration and displacement have low standard error which again supports the rejection of the Null hypothesis since it shows that there isn't a lot of error from the fitted model to the data point.

Confidence Interval:

	2.5%	97.5
(Intercept)	101.7676925	125.0002340
Acceleration	-4.5995085	-3.3746826
Displacement	0.2567292	0.2890209

Using the linear regression model, I create a 95% confidence interval to see the upper and lower limit of our features within the model. None of the 95% confidence interval's overlap 0 which means this further supports the 2 features since the model needs the 2 features for at least 95% of the model. Now to make sure the response is also following the model, a prediction function was used to determine what the 95% confidence interval is for that point prediction.

Collinearity

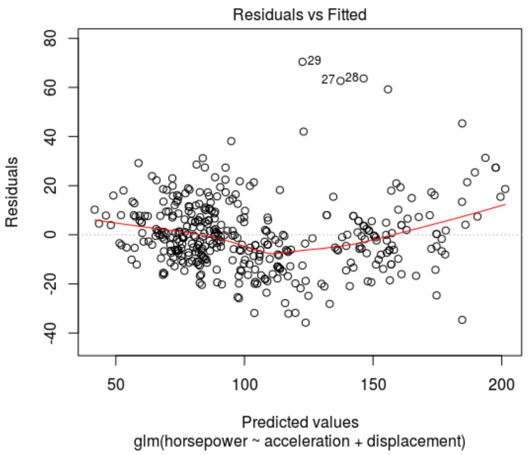
Acceleration	Displacement
1.419888	1.419888

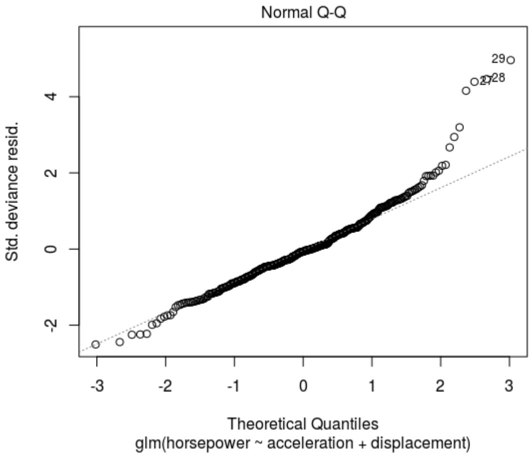
Features can become inflated due to linear dependence on other features. To make sure that the features aren't inflated, a test is used to determine the amount of linear dependency each feature has in the model. Above is the result of the test and since none of these features showed a result higher than 5 that means there is minimal linear dependency in this model.

Testing Null Hypotheses

Now I will test the Null hypotheses, $H_0:B_1=B_2=0$. Since the p-value for the Null hypothesis is $p<2.2e-16$, we reject the Null hypotheses, supporting a relationship between Horsepower and at least one other feature in the model. Now looking at the individual features, all the features have a p-value<0.1 which means we will reject the Null hypotheses for $H_0:B_1=0$ and $H_0:B_2=0$ since it shows that the chance of an extreme data point based on the observed data points is small.

Examining Fit:





The Residuals vs Fitted graph is used to visually represent the residual, which is the distance from the data point to the fitted model. In this graph, it shows that there is a lack of fit since there are a lot of residual data points scattered throughout the graph.

The Normal Q-Q graph is used to visually represent the distribution of the data points in the data set. Since the plotted points follow the straight line, it shows there is a good amount of distribution but towards the ends, the distribution start to vary especially towards the higher valued distributions.

Testing Data

In order to see if our model can predict values, I used Last One Out Cross Validation to test my linear regression model. The testing error was 206.82% and the training error was 206.8181% which are values that are too significantly high to even considered this model accurate. This shows that the model was overfitted and doesn't represent the data points well. This linear regression model can't predict future values of Horsepower using the features acceleration and displacement.

Conclusion

In conclusion, I believe there is not enough data to support this model since the testing data was significantly inaccurate, when trying to test the model. Though this model may have statistical values that represent an accurate model, the Number of iterations it took to model this fit was 2 which is too small to show a valid interpretation of the data points in the data frame. To improve this linear regression model, more features could be introduced to see if another feature has a better linear relationship with Horsepower than the current features. Along with adding more features, more observations could also help improve the linear regression model since there was only 392 observations.