Regression Model Project

Aruna

5/31/2020

Overview

In the project we will work on Motor Trend, a magazine about the automobile industry. Looking at a data set of a collection of cars, and exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). We are particularly interested in the following two questions:

1. Is an automatic or manual transmission better for MPG
2. Quantify the MPG difference between automatic and manual transmissions

Using simple linear regression analysis, we determine that there is a signficant difference between the mean MPG for automatic and manual transmission cars.

We can that Manual Transmission provides better MPG

Regression Analysis

Now calculate mean MPG values for cars with Automatic and Manual transmission

```
aggregate(mtcars$mpg,by=list(mtcars$am.label),FUN=mean)
```

```
##     Group.1        x
## 1 Automatic 17.14737
## 2    Manual 24.39231
```

We can see again that Manual transmission yields on average 7 MPG more than Automatic

Simple Linear Regression Test

```
T_simple <- lm(mpg ~ factor(am), data=mtcars)
summary(T_simple)
```

```
## 
## Call:
## lm(formula = mpg ~ factor(am), data = mtcars)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -9.3923 -3.0923 -0.2974  3.2439  9.5077 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   17.147      1.125  15.247 1.13e-15 ***
## factor(am)1    7.245      1.764   4.106 0.000285 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 4.902 on 30 degrees of freedom
## Multiple R-squared:  0.3598, Adjusted R-squared:  0.3385 
## F-statistic: 16.86 on 1 and 30 DF,  p-value: 0.000285
```

The p-value is less than 0.003, so we will not reject the hypothesis.

Let’s perform an ANOVA

```
T_variance_analysis <- aov(mpg ~ ., data = mtcars)
summary(T_variance_analysis)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)    
## cyl          1  817.7   817.7 102.591 2.3e-08 ***
## disp         1   37.6    37.6   4.717 0.04525 *  
## hp           1    9.4     9.4   1.176 0.29430    
## drat         1   16.5    16.5   2.066 0.16988    
## wt           1   77.5    77.5   9.720 0.00663 ** 
## qsec         1    3.9     3.9   0.495 0.49161    
## vs           1    0.1     0.1   0.016 0.90006    
## am           1   14.5    14.5   1.816 0.19657    
## gear         2    2.3     1.2   0.145 0.86578    
## carb         5   19.0     3.8   0.477 0.78789    
## Residuals   16  127.5     8.0                    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the above Analysis of Variance, we can look for p-values of less than .5. This gives us cyl, disp, and wt to consider in addition to transmission type (am)

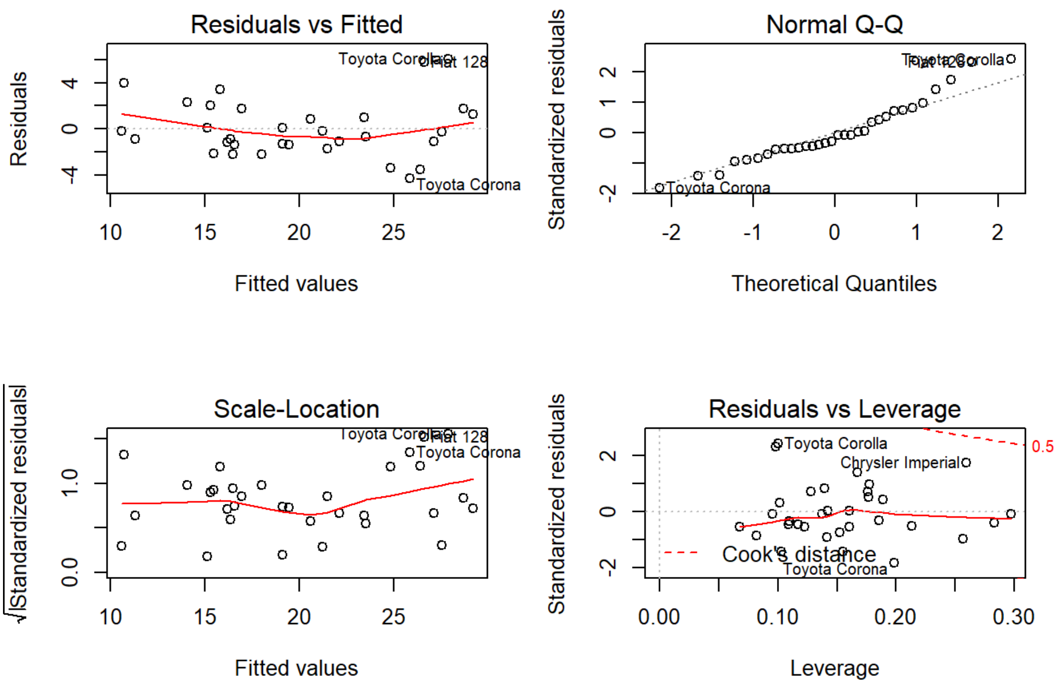
```
T_multivar <- lm(mpg ~ cyl + disp + wt + am, data = mtcars)
summary(T_multivar)
```

```
## 
## Call:
## lm(formula = mpg ~ cyl + disp + wt + am, data = mtcars)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -4.318 -1.362 -0.479  1.354  6.059 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 40.898313   3.601540  11.356 8.68e-12 ***
## cyl         -1.784173   0.618192  -2.886  0.00758 ** 
## disp         0.007404   0.012081   0.613  0.54509    
## wt          -3.583425   1.186504  -3.020  0.00547 ** 
## am           0.129066   1.321512   0.098  0.92292    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 2.642 on 27 degrees of freedom
## Multiple R-squared:  0.8327, Adjusted R-squared:  0.8079 
## F-statistic: 33.59 on 4 and 27 DF,  p-value: 4.038e-10
```

This Multivariable Regression test now gives us an R-squared value of over .83, suggesting that 83% or more of variance can be explained by the multivariable model. P-values for cyl (number of cylinders) and weight are below 0.5, suggesting that these are confounding variables in the relation between car Transmission Type and Miles per Gallon.

Residual Plot and Analysis

```
par(mfrow = c(2, 2))
plot(T_multivar)
```



The “Residuals vs Fitted” plot here shows us that the residuals are homoscedastic. We can also see that they are normally distributed, with the exception of a few outliers.