Mtcars MPG Analysis

Author: James Lee

May 07, 2018

Contents

Main Analysis

Please check appendix for exploratory analysis and diagnostics graphs.

Executive Summary

Analysis suggests that while transmission initially seemed significant in explaining miles per gallon, removing the linear effects of horsepower and weight suggested that that might not be the case. After removing outliers, we see that transmission's p-value is even further increased, i.e. they are further away from being significant. Therefore, we conclude that there is no strong evidence to suggest that transmission is statistically significant in explaining mpg after having removed the linear effects of weigth and horsepower. On the other hand there is a strong reason to beleive that weight and horsepower have strong correlation.

To answer the question while ignoring the drawbacks of the model and lack of significance of the transmission variable, manual vehicles are expected to get.26895 boost in mpg.

Linear Regression

We will first fit just mpg vs transmission as a factor, we will later add more values.

```r
fit <- lm(mpg~factor(am), mtcars)
```

Looking at the summary, using factor automatic as a base we see that manual's slope coefficient is clearly statistically significant with a very low p value.

Model Selection Using Nested Model Testing

We might be leaving bias into our coefficient by not fitting other variables that may be relevant. I fitted all the variables to check their significance. Looks like horsepower and weight are closest to being significant, we will try out one variable at a time to try to remove bias from the transmission regressor as much as possible without overfitting.

```
fit.hp <- lm(mpg~hp+factor(am),mtcars)
```

Summary and anova both show p-values that suggest including hp might be statistically significant.

We add weight

```
fit.hp.wt <- lm(mpg~hp+wt+factor(am),mtcars)
```

Anova suggest that including weight might be statistically significant. Surprisingly, looking at the summary(fit.hp.wt), transmission is no longer statistically significant. This suggest that when you remove the linear effects of weight and horsepower, transmission's linear effect is not significant enough.

Diagnostics

We will look at the standard diagnostics plots to check that our model is good.

```
par(mfrow=c(2,2))
plot(fit.hp.wt)
```

Please see Appendix for the graphs. Looks like we might have some outliers, chrysler, toyota and fiat. However, overall the residuals show no signifcant pattern. Residuals that are high in value suggest that we might have a right skewed data, as in we have outliers that are showing very high mpg than our model would suggest. We confirm this in a normal Q-Q plot which shows chracteristics of a right-skewness. Looking at cook's distance, we don't see any points that are very high in leverage. Overall, we have a fairly good fit although there is some concerns of right skewness.

We also adjust for outliers. To summarize, I dropped outliers that were identified by hatvalues and dfbetas. They are shown below.

```
drop.outliers<- c('Chrysler Imperial','Maserati Bora','Fiat 128' ,'Toyota Corolla')
mtcars[rownames(mtcars) %in% drop.outliers,]
```

```
##                    mpg cyl  disp  hp drat    wt  qsec vs am gear carb
## Chrysler Imperial 14.7   8 440.0 230 3.23 5.345 17.42  0  0    3    4
## Fiat 128          32.4   4  78.7  66 4.08 2.200 19.47  1  1    4    1
## Toyota Corolla    33.9   4  71.1  65 4.22 1.835 19.90  1  1    4    1
## Maserati Bora     15.0   8 301.0 335 3.54 3.570 14.60  0  1    5    8
```

```
mtcars2 <- mtcars[ !(rownames(mtcars) %in% drop.outliers), ]
fit.hp.wt2 <- lm(mpg~hp+wt+factor(am),mtcars2)
```

We see that 3 of the outliers we dropped were manual cars. It is very possible that Fiat and Corrolla were skewing the transmission variable by themselves. Maserati Bora's mpg however is very small but it has a very high horsepower and weight, which reduces the mpg. So, it might be a case that it has a good mpg relative to its high weight and horsepower.

Answers to the questions

First, I don't belive that the company should be looking at the transmissions as a variable. As I explained, it is very likely that the outliers were skewing the leverage that transmission variable exerted onto mpg. This was clearly shown by how the variable became statistically insignifcant after having removed the linear effects of horsepower and weight. Also another point to make is that removing the outliers decreased the significance of the transmission even further to a p-value of .8. There is no statistical evidence to think that transmission is significant.

So I do not suggest using the transmission variable. However, for the completeness of the analysis I will use my original fit wihtout the outliers removed to answer the questions. Let's revisit the model

```
fit.hp.wt2
```

```
##
## Call:
## lm(formula = mpg ~ hp + wt + factor(am), data = mtcars2)
##
## Coefficients:
## (Intercept)           hp           wt  factor(am)1
##    35.91660     -0.03382     -3.62474      0.26895
```

We can see that while manual transmission has no statistical signifiance after having removed the linear effects of weight and horsepower, slope coefficient is still higher for manual vehicles. Manual vehicles get .26895 boost in mpg as shown in the coefficient. So yes, to answer the question while ignoring the drawbacks of the model, manual transmission is expected to get .26895 boost in mpg.

Appendix

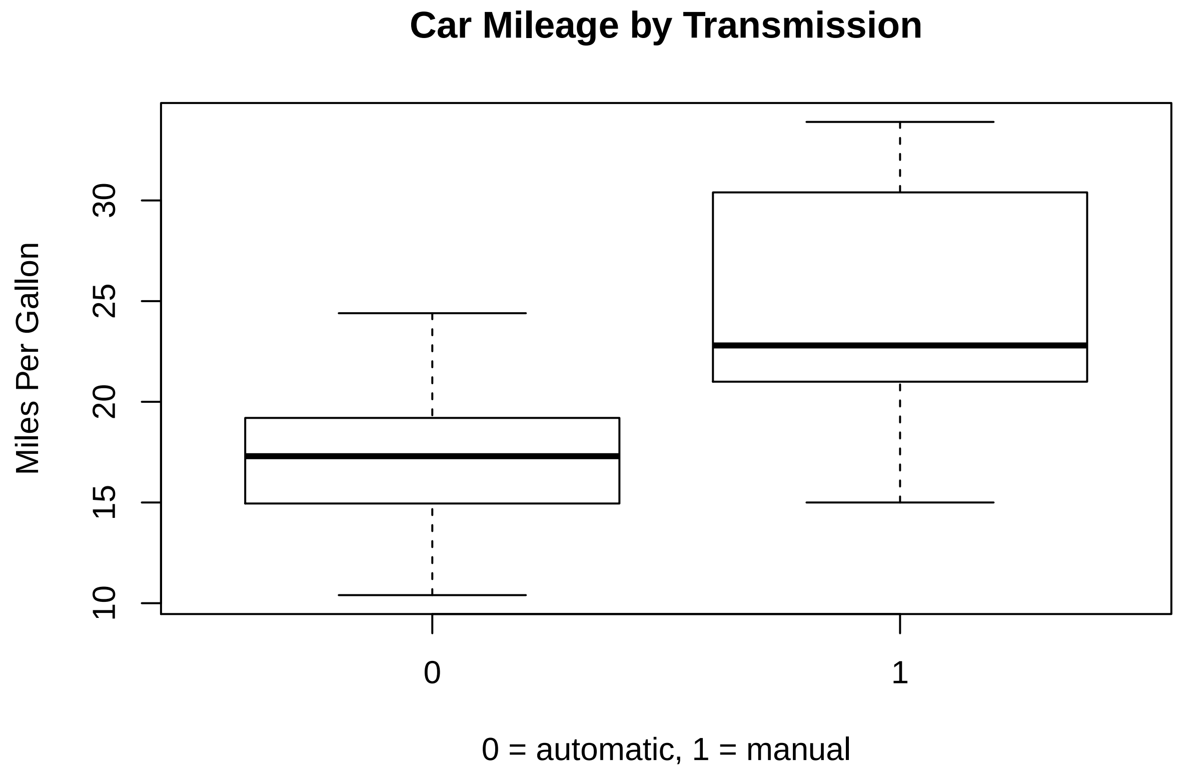
Exploratory Analysis

Before I begin the analysis, I will first look at the details of the data and plot miles per gallon box plot, split by transmission.

```
head(mtcars,1)
```

```
##           mpg cyl disp  hp drat   wt  qsec vs am gear carb
## Mazda RX4  21   6  160 110  3.9 2.62 16.46  0  1    4    4
```

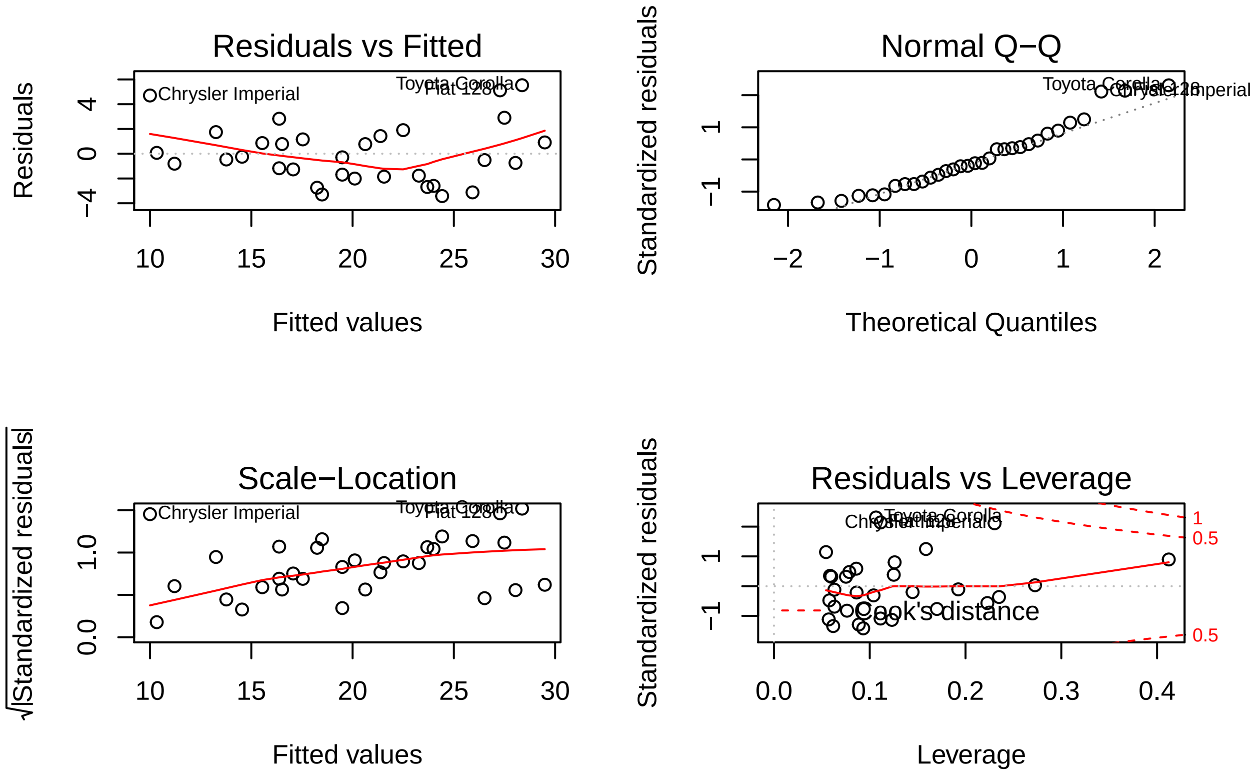
```
boxplot(mpg~am, data=mtcars, main = "Car Mileage by Transmission", xlab = "0 = automatic, 1 = manual", 
```



We see that manual clearly has a lot more miles per gallon in the picture. We will test significance using a linear model. Another point to note is that the data is tricky in a way that manual is 1 and automatic is 0, which was contrary to what I would have thought.

Diagnostics

```r
par(mfrow=c(2,2))
plot(fit.hp.wt)
```


Residuals vs Fitted
Residuals
Fitted values
−4
0
4
10
15
20
25
30
Chrysler Imperial
Toyota Corolla
Fiat 128
Normal Q−Q
Standardized residuals
Theoretical Quantiles
−1
1
−2
−1
0
1
2
Toyota Corolla
Fiat 128
Chrysler Imperial
Scale−Location
√|Standardized residuals|
Fitted values
0.0
1.0
10
15
20
25
30
Chrysler Imperial
Toyota Corolla
Fiat 128
Residuals vs Leverage
Standardized residuals
Leverage
−1
1
0.0
0.1
0.2
0.3
0.4
Toyota Corolla
Chrysler Imperial
Fiat 128
Cook's distance
1
0.5
0.5