Regression Models Course Project. MTCARS data set

Vamshidhar Pandrapagada

March 18, 2017

Executive Summary

As part of this project , we review the data set of a collection of cars we are interested in and explore the relationship between a set of variables and miles per gallon (MPG) (outcome). During the intial analysis it is evident that (i.e a model fit between Transmission and mpg), manual vehicles have 7.24 more MPG than automatic vehicles. But after fitting multiple regression models, we have deduced the following conclusions:
1) Even though intial observation suggested that Manual vehicles have higher mileage than Automatic, MPG as an outcome is dependent on other features in the car.
2) This is done by rejecting the NULL Hypothesis and accepting the alternate hypothesis by including other variables in the model.
3) Automatic/Manual trasmission provides only the correlation. Causation, as indicated is dependent on other predictors.

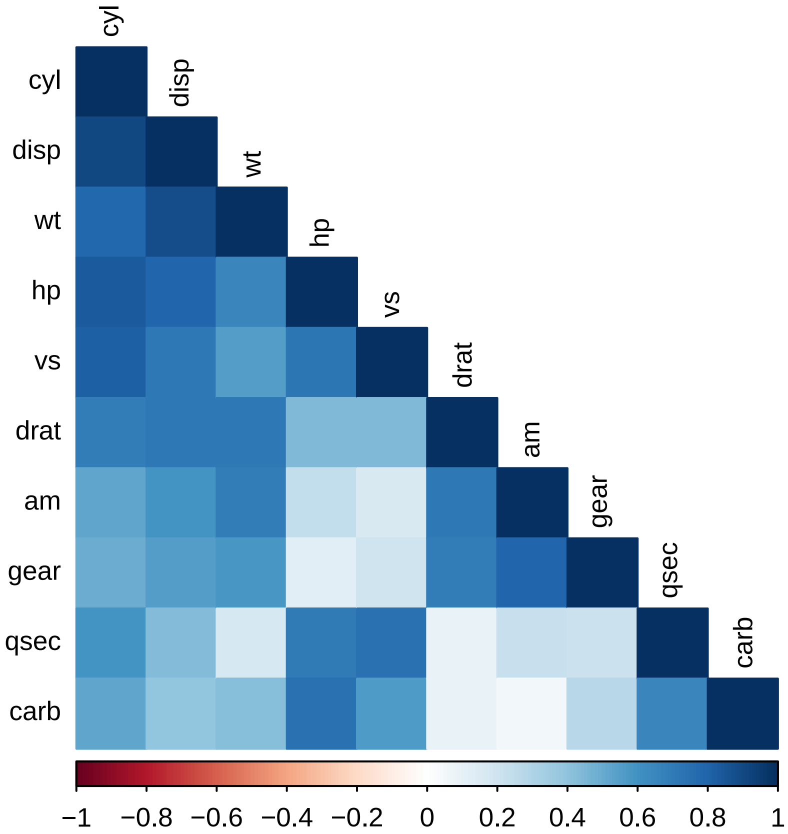
Load the required libraries and data sets (Code is Hidden)

Exploratory Data Analysis

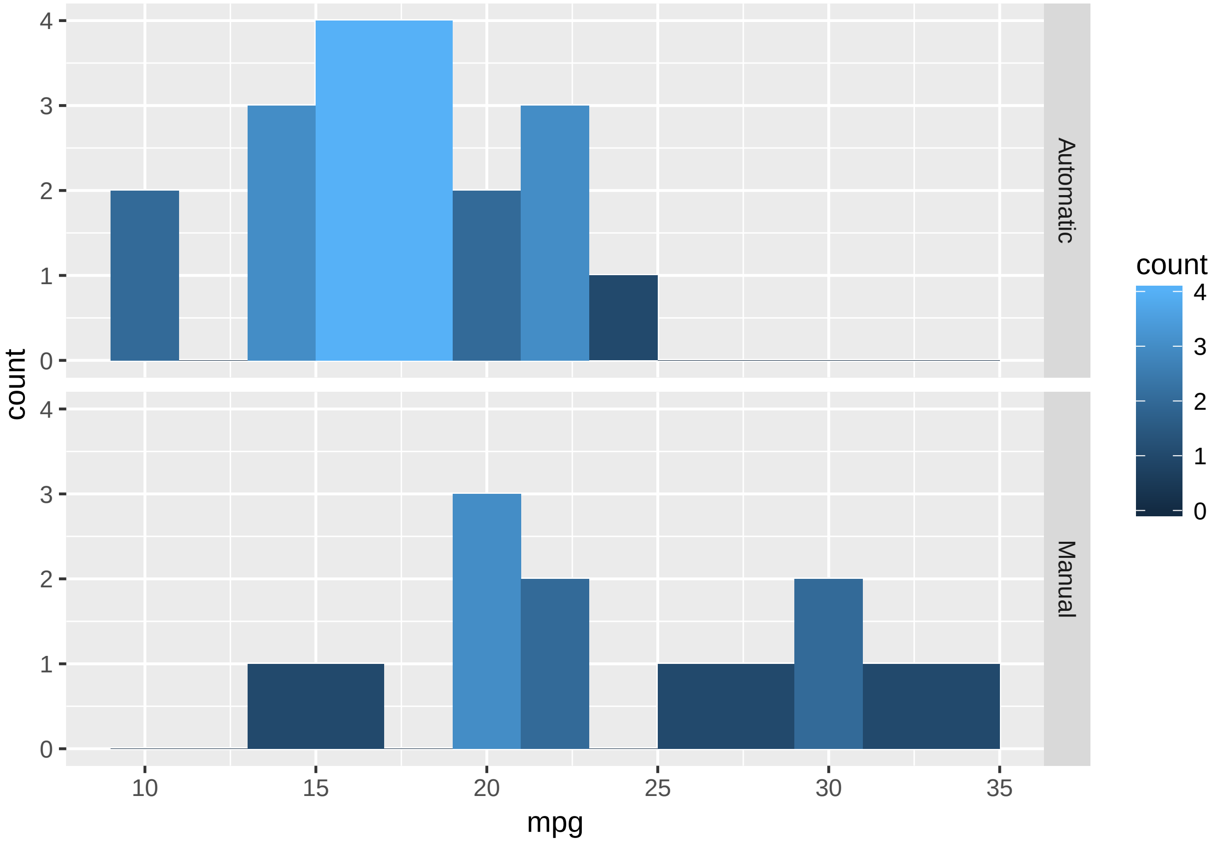
As a best practice, it is always advised to do EDA before building any prediction model.

1) Let's assess the correlation between all the variables before assessing their impact on mpg.

```
mtcars2 <- mtcars
row.names(mtcars2) <-NULL
mtcars2$am <-as.factor(mtcars2$am)
cars_features <- mtcars[,-1]
row.names(cars_features) <- NULL
corr_matrix<-abs(cor(cars_features))
corrplot(corr_matrix, order = "FPC", method = "color", type = "lower",tl.cex = 0.8, tl.col = rgb(0, 0, 0
```



As we see, the correlation does not exceed 0.8 between the predictors. This does not seem to be a major problem for now. Lets now examine a histogram and check if Automatic vs Manual trasmission has any key role to play on mileage.



The above plots clearly show an impact that the trasmission mode has on MPG. MPG is relatively high when transmission mode is Manual. Can we now answer the first question. Is an automatic or manual transmission better for MPG? Looks like YES is an answer.

But is this just a correlation? Or Transmission mode is also a cause (Causation) for low mileage? Let's examine more by fitting a couple of regression models.

T-Test, NULL Hypothesis and and Statistical inference

Step 1: Convert the predictors to factor variables (if they are not continous). ##Code Hidden

Step 2: Conduct a T-test on mpg and transmission mode to validate the significance of p-value. Validate the test by actually fitting regression model on these two variables

```
fit1<- lm(mpg ~ am, data=mtcars2)
fit1$coef
```

```
## (Intercept)         am1
##   17.147368    7.244939
```

```
t_test <-t.test(mpg ~ am, data = mtcars)
t_test
```

```
##
##  Welch Two Sample t-test
##
## data:  mpg by am
## t = -3.7671, df = 18.332, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.280194  -3.209684
## sample estimates:
## mean in group 0 mean in group 1
##        17.14737        24.39231
```

p-value from the T-test 0.001374 is statistically significant. This establishes a strong case to reject NULL HYPOTHESIS. Both from T-Test and the regression result, we can observe that Manual Transmission vehicles (with value 1) have 7.24 more MPG on an average. The adjusted R squared value of 0.3385 was able to explain only 33% variation in the model

Again, is this still true when we regress the model using other predictors?

Step 3: Fit a multivariate linear regression model using all the predictors with mpg as outcome to obtain a better model. Choose a model based by AIC using a stepwise algorithm

```r
fit2 <- lm(mpg ~. , data = mtcars2)
stepFull <- step(fit2)
```

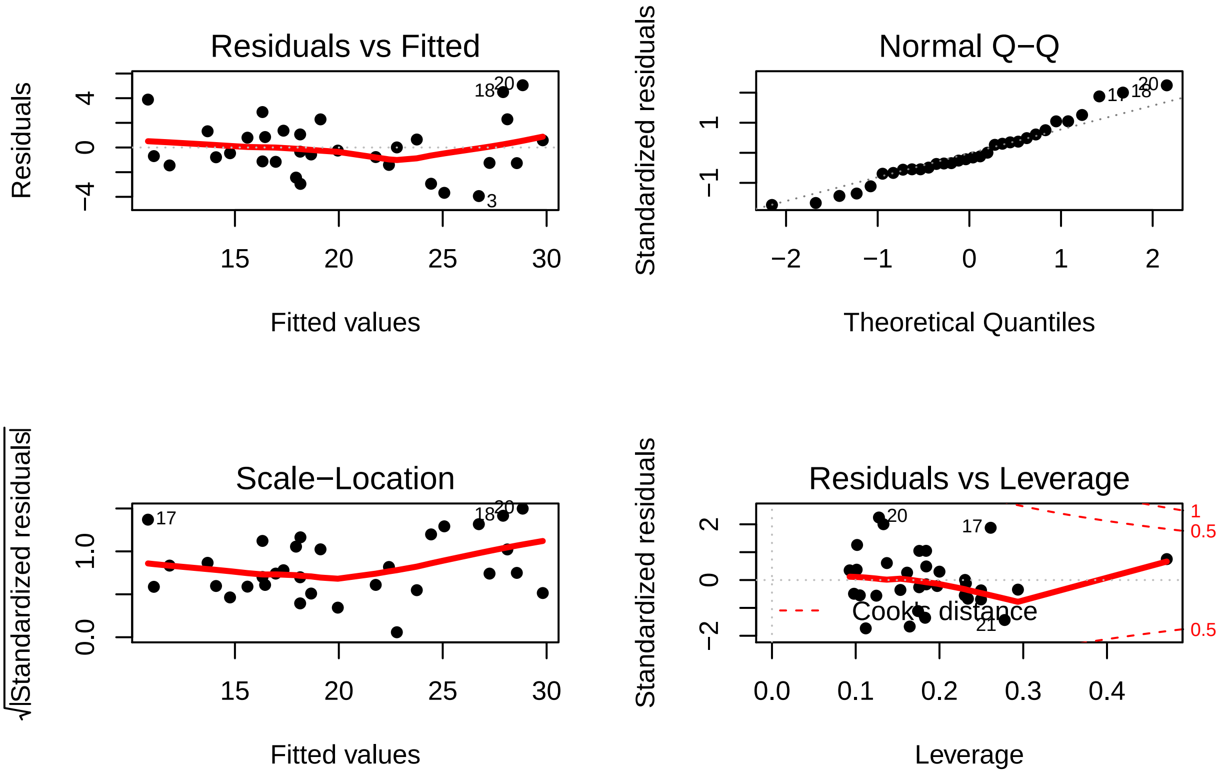
The model with all the predictors is now able to explain the model very well. Lets go through each of the sections in the summary.

Residuals vs Fitted Analysis

Residual is equal to the actual value minus the predicted value. Note that from the stepFull model above (results hidden due to page limitation), The max residual error of 5.0513 suggests that this model under-predicted mpg by nearly 5.05 miles for atleast one observation. %50 percent of the errors fall between 1Q (first Quartile) and 3Q (3rd Quartile) values. Most of the prediction were between 1.25 miles over the true value and 1.12 miles under the true value. Also

Residual Plot

```r
#plot(predict(fit2),resid(fit2), pch=10)
#abline(h = 0, lwd=3)
par(mfrow = c(2, 2))
plot(stepFull, pch=16, lwd=3)
```



The residual plot shows almost constant variance in residuals. This can be done by checking whether the residuals are not larger on a average for a range of fitted values and smaller in a different range.

There is no pattern observed. This is a positive sign indicating that the underlying model is linear in terms of the features involved.

Multiple R Squared Value

This tells us that how well the model was able to explain the values of the dependent variable. Closer the value to 1.0, better the model is. In the fullModel case, the value is close to 0.86, which means the model nearly explains 86% of the variation in dependent variable.

T- Value

The T-Value for the transmission mode (AM) variable is:
1) **4.106** when the model is fit only using AM as predictor. So this coefficient is 4 standard errors from zero, which is faily a good indicator that this co-effecient is not likely to be zero. Higher the T-Value, the more like we should include this feature in our linear model with Non-zero coeffecient.
2) **1.296** when the model is fit using all the predictors. This coeffecient is only 1.2 standard errors from zero which indicates that this co-effecient is likely to be zero. So when all the predictors are involved,the confidence level has really gone down reducing the linear relationship with this feature.

Conclusion

Even though intial observation suggested that Manual vehicles have higher mileage than Automatic, MPG as an outcome is dependent on other features in the car. This is done by rejecting the NULL Hypothesis and accepting the alternate hypothesis by including other variables in the model. Automatic/Manual trasmission provides only the correlation. Causation, as indicated is dependent on other predictors.