Regression Models Project

Analyzing the relationship between variables in the **mtcars** data set from R.

Allan R Brewer Cappellin

June 19, 2015

Executive Summary

In this project we will analyze the **mtcars** data set from R. We will do a Exploratory Data Analysis of it, fit various models to explain the interaction between variables, respond to the questions of weather Manual or Automatic transmission is better for MPG, quantify the difference and do residual and uncertainty plots and analysis. This analysis will be done using the tools learned in the Regression Models Course from Coursera.

From the liner regression fitted and tested in this project we can say that transmission type is not associated significantly with MPG ($\alpha = 0.05$) and for this reason it can not be said if manual or automatism has an effect on MPG. The variable was considered as a factor with two other predictor which were significant and with a coefficient different than zero at a 5% significance.

Exploratory Data Analysis and Summary

With the summary information in the Appendix #2 we will create a correlation matrix of all the variables to understand their relationship, the correlation matrix can be seen in the same appendix section. We will use this information to determine the ones that should be incorporated to the model.

From the correlations matrix we can define the variables to incorporate in our models. Each variable added to a model will be tested to determine their significance in the outcome via an ANOVA. The following conclusions are drawn from the plot: Since **Number of Cylinder**, **Displacement** and **HP** are so highly correlated between them we will only include **No of Cylinders** which has the highest correlation to **MPG**, **Drat** will be included since it has a relative high correlation with **MPG** and low correlation with the rest of the variables, **Weight**, **Qsec** and **VS** will not be incorporated since we consider they do not have much correlation with **MPG**, **AM** will obviously be included since it is the variable in question and **Number of Gears** and **Carburetors** will not be incorporated since they do not correlate as much with the outcome variable in question. Even though leaving uncorrelated variables out of the model can bias the results of the coefficients we will proceed with this approach since the work needed to study every possible model and their results is to time consuming and unnecessary for this project.

Linear Models

We will create a few linear models and do an ANOVA to determine the variables to incorporate in the final model. The first one is the model with **mpg** as the outcome and the variable in question, **am**, as the predictor. After this we will incorporate the variables mentioned above and determine which are to be present in the final model using the ANOVA. The variables **Number of Cylinders** will be kept as a numeric variable since it can take other values outside from the ones on the data set.

Analysis of Variance between models

Now we will compare the analysis of variance between the models and determine which will be the variables that should be present in the final model. This analysis will determine if the incorporation of variables between models is significant in determining the outcome variable, in this case **MPG**.

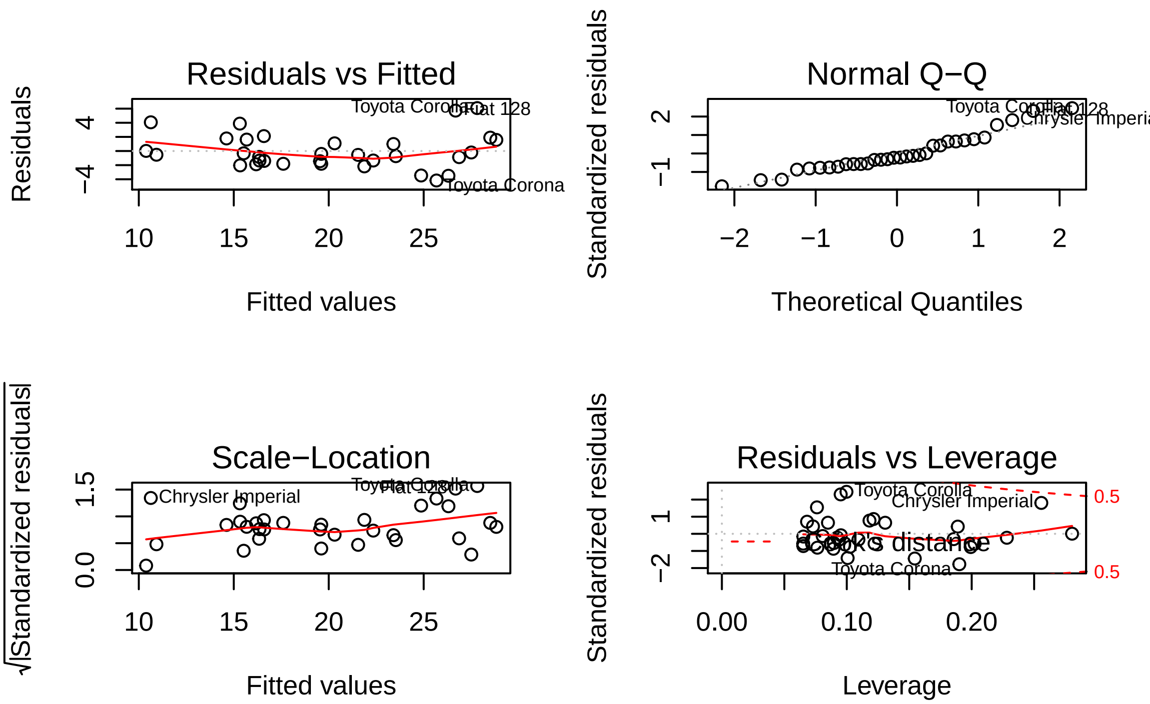
Using the P-Values reported in the appendix under ANOVAs and a 5% significance we can determine that including **am**, **cyl** and **wt** is statistically significant but including the variable **drat** is not correct. The final model which will be analyzed is: **lm(mpg ~ am + cyl + wt)**

Final Model

From the information retrieved in the ANOVAs we can now analyze the information in the final linear model. This model is formed by the dependent variable **mpg** and the independent variables **am**, **cyl** and **wt**. The result of the linear models, the coefficients are shown in the following figure.

```
##
## Call:
## lm(formula = mpg ~ am + cyl + wt, data = x)
##
## Coefficients:
## (Intercept)          am1          cyl           wt
##     39.4179       0.1765      -1.5102      -3.1251
```

The coefficients for the model selected are considered different to zero with a 5% significance for the **cyl** and **wt** variables, this information can be seen in the Appendix #3. In the case of the **am** variable the null hypothesis can not be rejected with the same significance value of 5% and so the coefficient is to be considered equal to zero. With this in mind we can say that transmission type does not have an effect on MPG in the cars of the data set.



Finally from this plots we can conclude that the regression model complies with the assumption that that the residual follow a random Normal Distribution with mean zero and that there are no outliers that produce any kind of variation inflation or leverage.

Conclusions

In this project a multiple regression was conducted to examine the predictors of mpg in cars. Multiple variables were incorporated into different models and compared to determine the correct one. The final model has three predictors: Number of Cylinders, Weight and Transmission Type. Together they accounted for 81% of the variance in MPG. The all the variables Number of Cylinders and Weight were significant predictors of MPG. **cyl** $\beta_2 = -1.5102$ and **wt** $\beta_3 = -3.1251$ were the strongest predictors and Negatively associated with MPG, whereas **am** $\beta_1 = 0.1765$ was not considered statistically significant since the null hypothesis could not be rejected. Inference done with a degree of uncertainty defined by an $\alpha = 0.05$ in every Hypothesis Test.

Appendix 1: R code

In this first section of the appendix you can find all of the R code that is not shown in the main body of the document.

R Code #1 : Load Data R Code

```
data(mtcars)
x <- mtcars; rm(mtcars)
```

R Code #2 : Summary R Code

```
xsum <- summary(x)
```

R Code #3 : Correlation Matrix

```
xcor <- cor(x)
```

R Code #4 : Linear Models R Code

```
x$am <- as.factor(x$am)
model_am <- lm(mpg ~ am , data = x)
model2 <- lm(mpg ~ am + cyl, data = x)
model3 <- lm(mpg ~ am + cyl + wt, data = x)
model4 <- lm(mpg ~ am + cyl + drat, data = x)
model5 <- lm(mpg ~ am + cyl + wt + drat, data = x)
```

R Code #6 : ANOVAs R Code

```
anova1 <- anova(model_am, model2, model3)
anova2 <- anova(model_am, model2, model4)
anova3 <- anova(model_am, model2, model3, model5)
```

R Code #6 : Residual and Analysis Plots R Code

```
par(mfrow=c(2,2))
plot2 <- plot(model3)
```

Appendix 2: Exploratory Data Analysis

Data Summary

```
##       mpg             cyl             disp             hp       
##  Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0  
##  1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5  
##  Median :19.20   Median :6.000   Median :196.3   Median :123.0  
##  Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7  
##  3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0  
##  Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0  
##       drat             wt             qsec             vs        
##  Min.   :2.760   Min.   :1.513   Min.   :14.50   Min.   :0.0000  
##  1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0.0000  
##  Median :3.695   Median :3.325   Median :17.71   Median :0.0000  
##  Mean   :3.597   Mean   :3.217   Mean   :17.85   Mean   :0.4375  
##  3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90   3rd Qu.:1.0000  
```

```
##  Max.   :4.930   Max.   :5.424   Max.   :22.90   Max.   :1.0000
##        am              gear            carb
##  Min.   :0.0000   Min.   :3.000   Min.   :1.000
##  1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000
##  Median :0.0000   Median :4.000   Median :2.000
##  Mean   :0.4062   Mean   :3.688   Mean   :2.812
##  3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000
##  Max.   :1.0000   Max.   :5.000   Max.   :8.000
```

Correlation Matrix

```
##             mpg        cyl       disp         hp        drat         wt
## mpg   1.0000000 -0.8521620 -0.8475514 -0.7761684  0.68117191 -0.8676594
## cyl  -0.8521620  1.0000000  0.9020329  0.8324475 -0.69993811  0.7824958
## disp -0.8475514  0.9020329  1.0000000  0.7909486 -0.71021393  0.8879799
## hp   -0.7761684  0.8324475  0.7909486  1.0000000 -0.44875912  0.6587479
## drat  0.6811719 -0.6999381 -0.7102139 -0.4487591  1.00000000 -0.7124406
## wt   -0.8676594  0.7824958  0.8879799  0.6587479 -0.71244065  1.0000000
## qsec  0.4186840 -0.5912421 -0.4336979 -0.7082234  0.09120476 -0.1747159
## vs    0.6640389 -0.8108118 -0.7104159 -0.7230967  0.44027846 -0.5549157
## am    0.5998324 -0.5226070 -0.5912270 -0.2432043  0.71271113 -0.6924953
## gear  0.4802848 -0.4926866 -0.5555692 -0.1257043  0.69961013 -0.5832870
## carb -0.5509251  0.5269883  0.3949769  0.7498125 -0.09078980  0.4276059
##             qsec         vs          am       gear        carb
## mpg   0.41868403  0.6640389  0.59983243  0.4802848 -0.55092507
## cyl  -0.59124207 -0.8108118 -0.52260705 -0.4926866  0.52698829
## disp -0.43369788 -0.7104159 -0.59122704 -0.5555692  0.39497686
## hp   -0.70822339 -0.7230967 -0.24320426 -0.1257043  0.74981247
## drat  0.09120476  0.4402785  0.71271113  0.6996101 -0.09078980
## wt   -0.17471588 -0.5549157 -0.69249526 -0.5832870  0.42760594
## qsec  1.00000000  0.7445354 -0.22986086 -0.2126822 -0.65624923
## vs    0.74453544  1.0000000  0.16834512  0.2060233 -0.56960714
## am   -0.22986086  0.1683451  1.00000000  0.7940588  0.05753435
## gear -0.21268223  0.2060233  0.79405876  1.0000000  0.27407284
## carb -0.65624923 -0.5696071  0.05753435  0.2740728  1.00000000
```

Appendix 3: ANOVAS and Model Analysis

ANOVA Results for Model Comparison

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + cyl
## Model 3: mpg ~ am + cyl + wt
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     29 271.36  1    449.53 65.884 7.751e-09 ***
## 3     28 191.05  1     80.32 11.771  0.001886 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
```

```
## Model 2: mpg ~ am + cyl
## Model 3: mpg ~ am + cyl + drat
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     29 271.36  1    449.53 46.4518 2.101e-07 ***
## 3     28 270.97  1      0.39  0.0407    0.8415
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + cyl
## Model 3: mpg ~ am + cyl + wt
## Model 4: mpg ~ am + cyl + wt + drat
##   Res.Df    RSS Df Sum of Sq       F   Pr(>F)
## 1     30 720.90
## 2     29 271.36  1    449.53 63.5481 1.44e-08 ***
## 3     28 191.05  1     80.32 11.3537 0.002281 **
## 4     27 191.00  1      0.05  0.0072 0.932807
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Final Linear Model

```
##
## Call:
## lm(formula = mpg ~ am + cyl + wt, data = x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1735 -1.5340 -0.5386  1.5864  6.0812
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  39.4179     2.6415  14.923 7.42e-15 ***
## am1           0.1765     1.3045   0.135  0.89334
## cyl          -1.5102     0.4223  -3.576  0.00129 **
## wt           -3.1251     0.9109  -3.431  0.00189 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.612 on 28 degrees of freedom
## Multiple R-squared:  0.8303, Adjusted R-squared:  0.8122
## F-statistic: 45.68 on 3 and 28 DF,  p-value: 6.51e-11
```