Developing_Data_Product_Project_Intro

Alfred Homere Ngandam Mfomdoum

October 16, 2016

This is an introduction of the data, chooosen from the Regression Models course project to buld our Shiny method.

Executive Summary

Motor Trend is an automobile trend magazine that is interested in exploring the relationship between a set of variables and miles per gallon (MPG) outcome. In this project, we will analyze the mtcars dataset from the 1974 Motor Trend US magazine to answer the following questions:

- Is an automatic or manual transmission better for miles per gallon (MPG)?
- How different is the MPG 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. Manual transmissions achieve a higher value of MPG compared to automatic transmission. This increase is approximately 1.8 MPG when switching from an automatic transmission to a manual one, with all else held constant.In this analysis we are attempting to find out whether a manual or automatic transmission is better for miles per gallon (mpg). This was done using a statistical analysis to quantify how different mpg is for cars using manual and automatic transmissions. The summarise the findings, we note that manual transmissions on average do give 1.55 miles per gallon more than automatic transmission; however this is taking into account he confounding variables of weight and cylinders. We then start by loading the data

```r
library (datasets)
data(mtcars)
head(mtcars)
```

```
##                    mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

```r
knitr::opts_chunk$set(echo = TRUE)
```

The dataset has 32 observations of 11 variables. We will do a quick analysis on the variables to gain some insight on the distribution of mpg and the two modes of transmission. * Firstly vs and am should be modelled as categorical variables

```r
mtcars$vs <- as.factor(mtcars$vs)
mtcars$am <- as.factor(mtcars$am)
knitr::opts_chunk$set(echo = TRUE)
```

- Then we can summarize them:

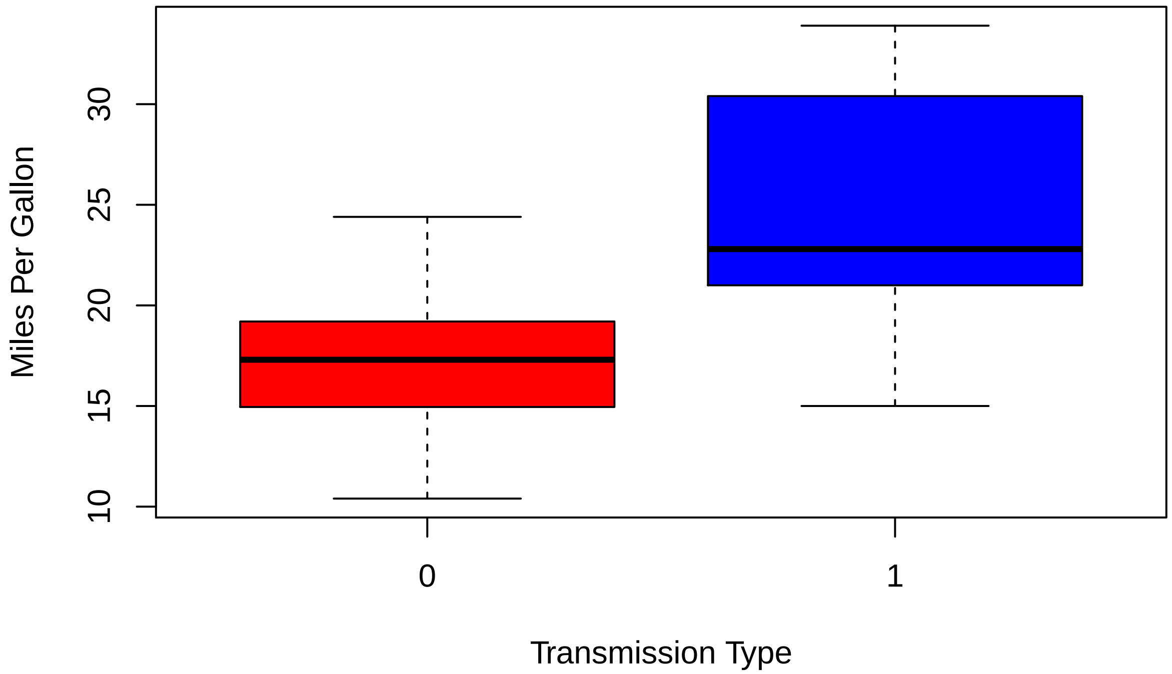
```
summary(mtcars$mpg)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   10.40   15.42   19.20   20.09   22.80   33.90
```

```
knitr::opts_chunk$set(echo = TRUE)
```

- After that they are represented trough the plot function

```
boxplot(mpg ~ am, data = mtcars, col = (c("red","blue")), ylab = "Miles Per Gallon", xlab = "Transmissi
```



```
knitr::opts_chunk$set(echo = TRUE)
```

We focus mainly on the relationship between the variables mpg and am in the boxplot and we can see there is distinguisable difference between the gas mileage for each type of transmission. However we must now try to fit on a linear regression model.

Simple linear regression model

We will use mpg as the dependent variable and am as the independent variable to fit a linear regression, where Beta1 is the group mean for automatic and Beta0 is the intercept.

```r
fit_simple <- lm(mpg ~ factor(am), data=mtcars)
summary(fit_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
```

```r
knitr::opts_chunk$set(echo = TRUE)
```

We can see that the adjusted R squared value is only 0.338 which means that only 33.8% of the regression variance can be explained by our model. However we must not forget there are several other predictor variables that we must take into account to see if any play a bigger role in the model.

Multivariable Regression Model

The first model we will run is a linear regression model against mpg for each variable. This gives us insight into variables with coefficient significance as well as an initial attempt at explaining mpg. Additionally, we will also look at the correlation of variables with mpg to help us choose an appropriate model.

```r
data(mtcars)
fit_multi <- lm(mpg ~ . ,data=mtcars)
summary(fit_multi)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4506 -1.6044 -0.1196  1.2193  4.6271
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.30337   18.71788   0.657   0.5181
## cyl         -0.11144    1.04502  -0.107   0.9161
## disp         0.01334    0.01786   0.747   0.4635
## hp          -0.02148    0.02177  -0.987   0.3350
## drat         0.78711    1.63537   0.481   0.6353
## wt          -3.71530    1.89441  -1.961   0.0633 .
## qsec         0.82104    0.73084   1.123   0.2739
## vs           0.31776    2.10451   0.151   0.8814
## am           2.52023    2.05665   1.225   0.2340
## gear         0.65541    1.49326   0.439   0.6652
## carb        -0.19942    0.82875  -0.241   0.8122
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.65 on 21 degrees of freedom
## Multiple R-squared:  0.869,  Adjusted R-squared:  0.8066
## F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07
```

```
knitr::opts_chunk$set(echo = TRUE)
```

- We compute the correlation model to appreciate it

```
cor(mtcars)[1,]
```

```
##        mpg        cyl       disp         hp       drat         wt
##  1.0000000 -0.8521620 -0.8475514 -0.7761684  0.6811719 -0.8676594
##       qsec         vs         am       gear       carb
##  0.4186840  0.6640389  0.5998324  0.4802848 -0.5509251
```

```
knitr::opts_chunk$set(echo = TRUE)
```

From the above output from 'fit_multi' and the 'cor' functions, we can see cyl,wt,hp,disp show strong correlations and significance for the model. Hence we choose those variables plus am for a linear model. This gives us the following model below:

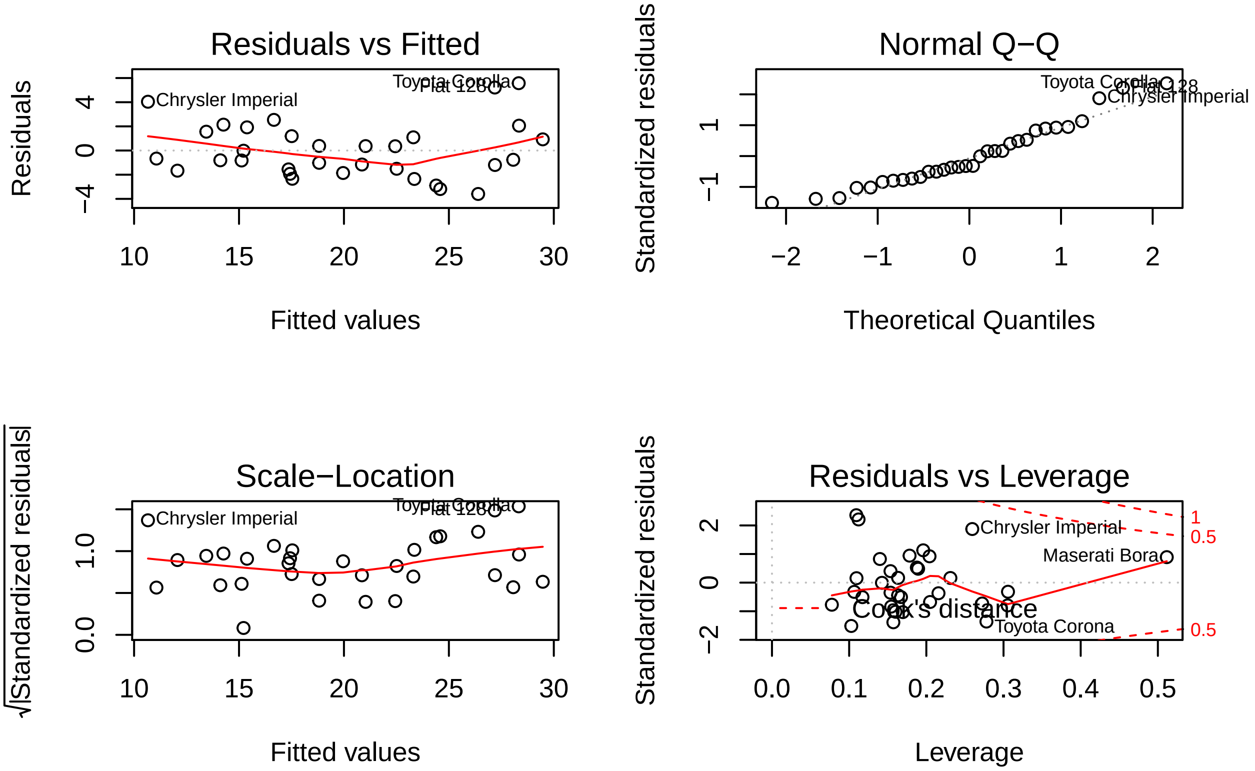
```
fit_final <- lm(mpg ~ wt+hp+disp+cyl+am, data = mtcars)
fit_final
```

```
##
## Call:
## lm(formula = mpg ~ wt + hp + disp + cyl + am, data = mtcars)
##
## Coefficients:
## (Intercept)           wt           hp         disp          cyl
##    38.20280     -3.30262     -0.02796      0.01226     -1.10638
##          am
##     1.55649
```

```
knitr::opts_chunk$set(echo = TRUE)
```

Additionally, we also plot the residuals to examine any heteroskedacity between the fitted and residual values; as well as to check for any non-normality.

```r
par(mfrow = c(2, 2))
plot(fit_final)
```



```r
knitr::opts_chunk$set(echo = TRUE)
```

Fitted graph we can see that our residuals are homosekdastic where they approximately have the sae variance and also we can see they are normally distributed using the quantile plot. Furthermore, using the final multivariable regression model put together we can see the multiple R squared value is much higher at 0.855, where 85.5% of the regression variance can be explained by the chosen variables. We can thus conclude that ‘wt’ and ‘cyl’are confounding variables in the relationship between ’am and ’mpg’ and that manual transmission cars on average have 1.55 miles per gallon more than automatic cars.

Conclusions

Following observations are made from the above plots.

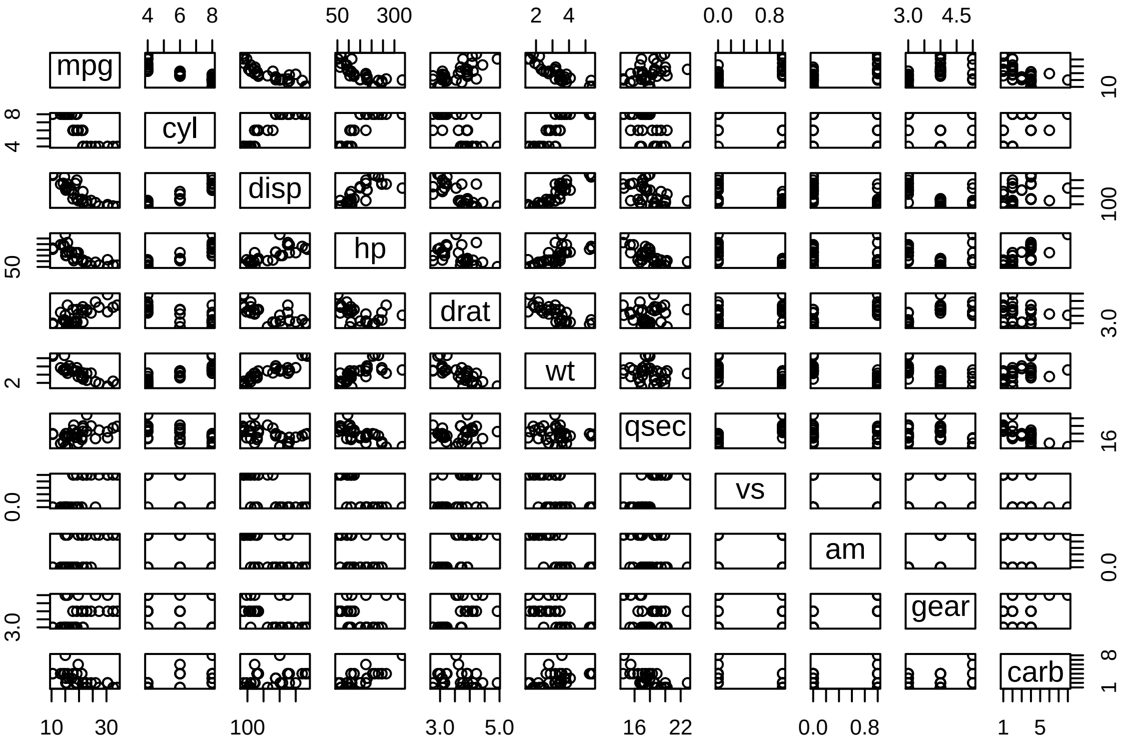
- The points in the Residuals vs. Fitted plot are randomly scattered on the plot that verifies the independence condition.
- The Normal Q-Q plot consists of the points which mostly fall on the line indicating that the residuals are normally distributed.
- The Scale-Location plot consists of points scattered in a constant band pattern, indicating constant variance.

- There are some distinct points of interest (outliers or leverage points) in the top right of the plots that may indicate values of increased leverage of outliers.

Appendix

Pairs plots for the "mtcars" dataset

```
pairs(mpg ~ ., data = mtcars)
```



```
pairs
```

```
## function (x, ...)
## UseMethod("pairs")
## <bytecode: 0x0000000007084fd8>
## <environment: namespace:graphics>
```

```
knitr::opts_chunk$set(echo = TRUE)
```