Regression Model Course Project

Junaid Khan

May 10, 2018

```
library(ggplot2)
library(datasets)
data("mtcars")
```

Summary

Motor Trend, a magazine about the automobile industry, is interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). They 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. Manual transmissions achieve a higher value of MPG compared to automatic transmission. This increase is approximately 2.1 MPG when switching from an automatic transmission to a manual one, with the weight, horsepower and discplacement held constant.

Exploratory Analysis

We will first summarise the mtcars dataset and have a brief overview

```
summary(mtcars)
```

```
##       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
```

```
str(mtcars)
```

```
## 'data.frame':    32 obs. of  11 variables:
```

```
## $ mpg : num 21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
## $ cyl : num 6 6 4 6 8 6 8 4 4 6 ...
## $ disp: num 160 160 108 258 360 ...
## $ hp  : num 110 110 93 110 175 105 245 62 95 123 ...
## $ drat: num 3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
## $ wt  : num 2.62 2.88 2.32 3.21 3.44 ...
## $ qsec: num 16.5 17 18.6 19.4 17 ...
## $ vs  : num 0 0 1 1 0 1 0 1 1 1 ...
## $ am  : num 1 1 1 0 0 0 0 0 0 0 ...
## $ gear: num 4 4 4 3 3 3 3 4 4 4 ...
## $ carb: num 4 4 1 1 2 1 4 2 2 4 ...
```

Finding the correlation between MPG and other variables

```
##        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
```

So we observe that **MPG is significantly correlated with with "cyl" , "disp","hp", "wt"**

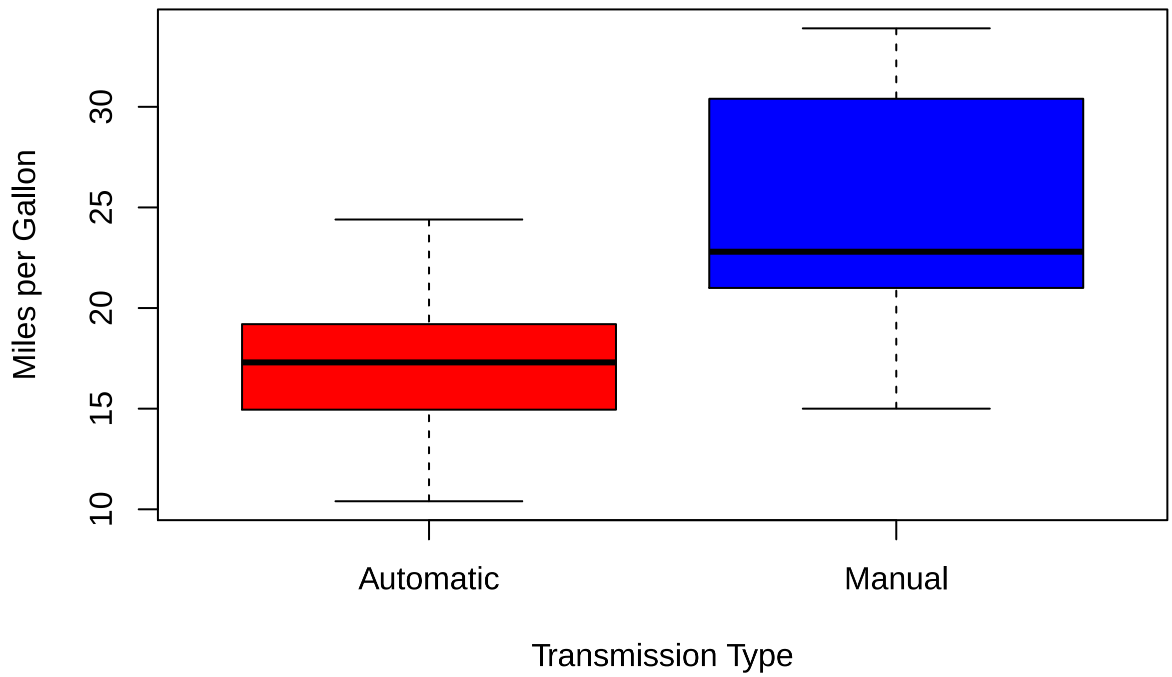
The "am" variable stands for Transmission, denoting 0 = automatic , 1 = manual.

Factoring the "am"" Variable

```
mtcars$am <- factor(mtcars$am, levels = c(0,1),labels = c("Automatic","Manual"))
```

Boxplot Depicting the Overall Effect of Transmission on Mpg

```
boxplot(mpg ~ am,mtcars, col = c('red','blue'),xlab="Transmission Type",ylab="Miles per Gallon")
```



From the above plot its evident that **Manual Transmission is better for MPG** We will now prove it by regression modelling :

Performing a linear regression with MPG as the predictor and Transmission as regressor, we will observe the coefficient

```
coef(summary(lm(mpg ~ factor(am),mtcars)))
```

```
##                     Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)        17.147368   1.124603 15.247492 1.133983e-15
## factor(am)Manual    7.244939   1.764422  4.106127 2.850207e-04
```

The coefficient Table tells us that **Manual transmission yields on average 7 MPG more than Automatic**

```
summary(lm(mpg ~ factor(am),mtcars))
```

```
##
## 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)Manual      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 0.0003, so we will not reject the hypothesis.

Linear Models

The linear dependencies suggests to analyse the linear models as:

```
fit1 <- lm(mpg ~ am,data = mtcars)
fit2 <- lm(mpg ~ am + wt,data = mtcars)
fit3 <- lm(mpg ~ am + wt + hp,data = mtcars)
fit4 <- lm(mpg ~ am + wt + hp + disp,data = mtcars)
fit5 <- lm(mpg ~ ., data = mtcars)
```

We now perform ANOVA test to compare the models

```
anova(fit1,fit2,fit3,fit4,fit5)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + wt
## Model 3: mpg ~ am + wt + hp
## Model 4: mpg ~ am + wt + hp + disp
## Model 5: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     29 278.32  1    442.58 63.0133 9.325e-08 ***
## 3     28 180.29  1     98.03 13.9571  0.001219 **
## 4     27 179.91  1      0.38  0.0546  0.817510
## 5     21 147.49  6     32.41  0.7692  0.602559
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We start with the variable "mpg" as a function of the vairable "am" add one variable after another and do the ANOVA routine to find the simplest modell that explains significantly the change in "mpg". We see that adding the variables "wt" and "hp" significantly improve the model, so it's the modell "fit3" which we use further.

Summary of model fit3

```
summary(fit3)
```

```
##
## Call:
## lm(formula = mpg ~ am + wt + hp, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4221 -1.7924 -0.3788  1.2249  5.5317
```

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.002875   2.642659  12.867 2.82e-13 ***
## amManual     2.083710   1.376420   1.514 0.141268
## wt          -2.878575   0.904971  -3.181 0.003574 **
## hp          -0.037479   0.009605  -3.902 0.000546 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.538 on 28 degrees of freedom
## Multiple R-squared:  0.8399, Adjusted R-squared:  0.8227
## F-statistic: 48.96 on 3 and 28 DF,  p-value: 2.908e-11
```

Conclusion

Is an automatic or manual transmission better for MPG?

It appears that manual transmission cars are better for MPG compared to automatic cars. However when modeled with confounding variables like displacement, HP and weight, the difference is not as significant as it seems in the beginning: a big part of the difference is explained by other variables.

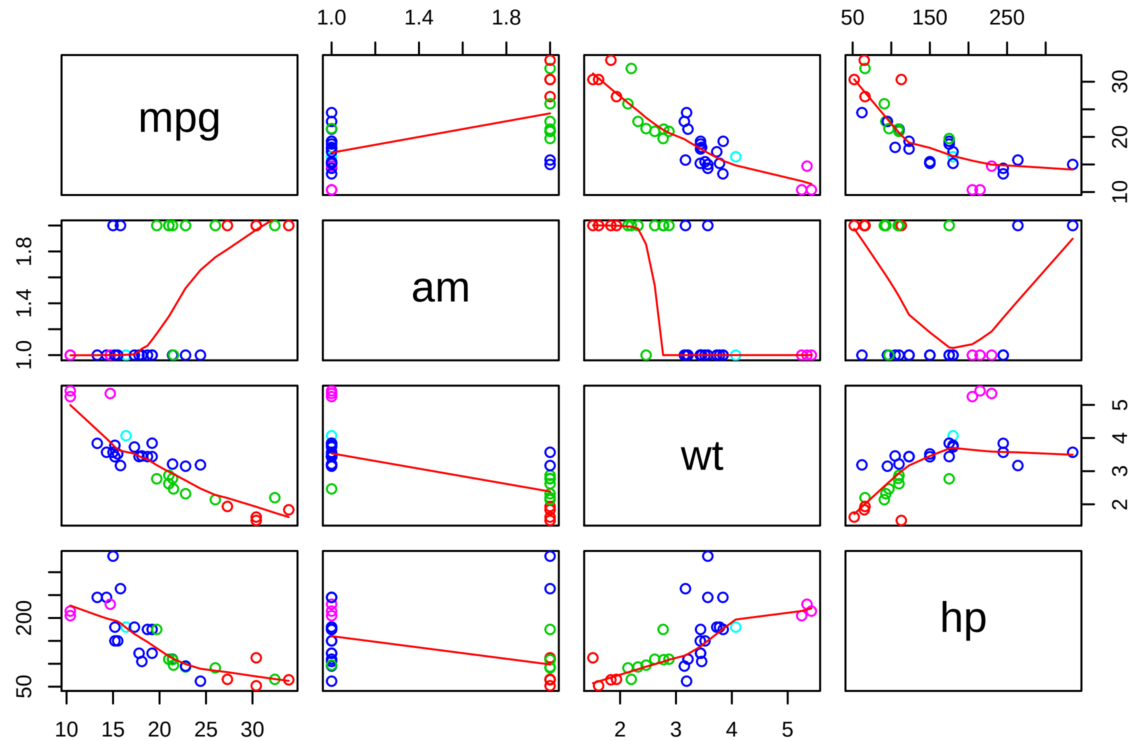
Quantify the MPG difference between automatic and manual transmissions

Analysis shows that when only transmission was used in the model manual cars have an mpg increase of 7.245. However, when variables wt and hp are included, the manual car advantage drops to 2.084 with other variables contributing, sometimes more (e.g. weight) to the effect.

Appendix

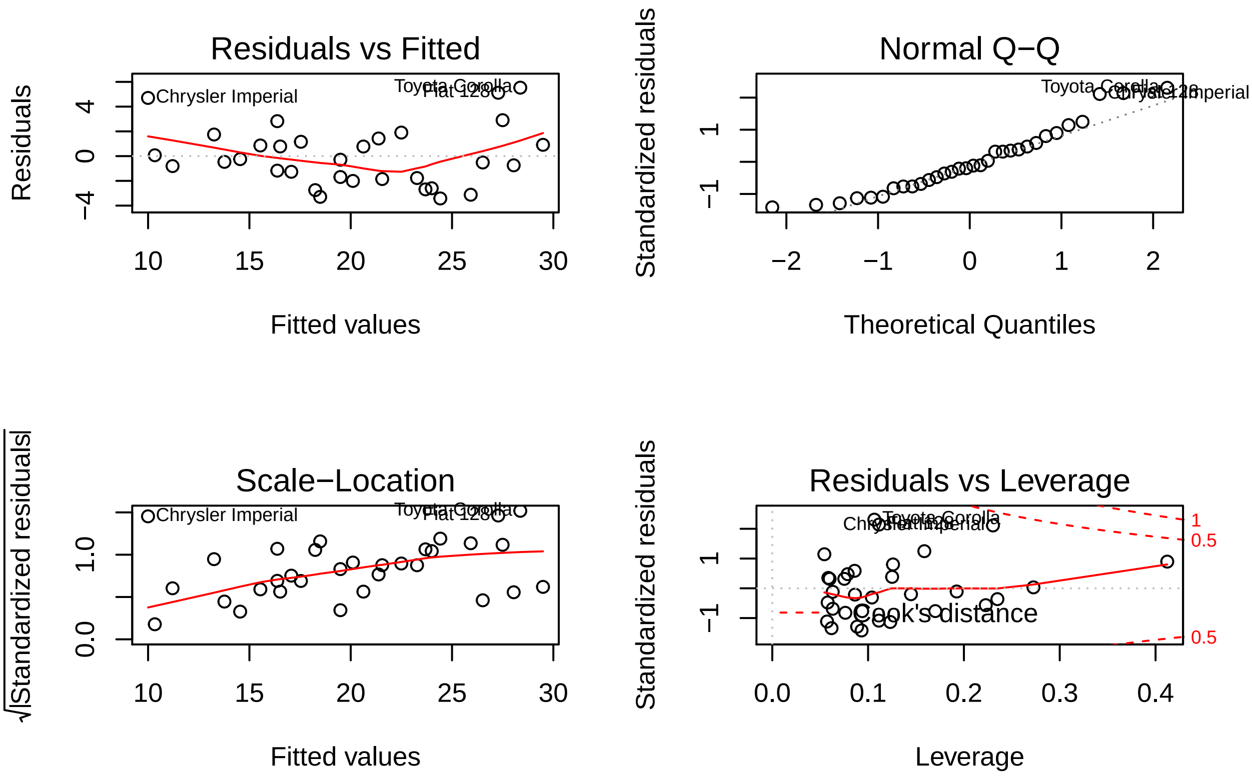
Correlation of the variables of the Model "fit3"

```r
mtcars_vars <- mtcars[, c(1, 9, 6, 4)]
mar.orig <- par()$mar  # save the original values
par(mar = c(1, 1, 1, 1))  # set your new values
pairs(mtcars_vars, panel = panel.smooth, col = 9 + mtcars$wt)
```



Residual plot of the Model "fit3"

```
par(mfrow=c(2,2))
plot(fit3)
```


Residuals vs Fitted
Residuals
Fitted values
Chrysler Imperial
Toyota Corolla
Fiat 128
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Residuals vs Leverage
Leverage
Cook's distance
0.5
1