Regression Models Course Project

Instructions

You work for Motor Trend, a magazine about the automobile industry. Looking at a data set of a collection of cars, they are 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:

- "Is an automatic or manual transmission better for MPG"
- "Quantify the MPG difference between automatic and manual transmissions"

Load and check data

```
rm(list = ls())
cardata <- mtcars
head(cardata)
```

```
##                   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
```

```
summary(cardata) # no NA's
```

```
##       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
```

"Is an automatic or manual transmission better for MPG?"

Plotting

```
library(ggplot2)
ggplot(data = cardata, aes(y = mpg, x = factor(am), fill = factor(am)))+
  geom_boxplot()+
  scale_fill_discrete(name="Tranmission\nType")+
  xlab("Transmission type 0 = auto and 1 = manual")+
  ylab("Miles / Gallon")+
  ggtitle("Figure 1 - is automatic or manual transmission better for MPG?")
```

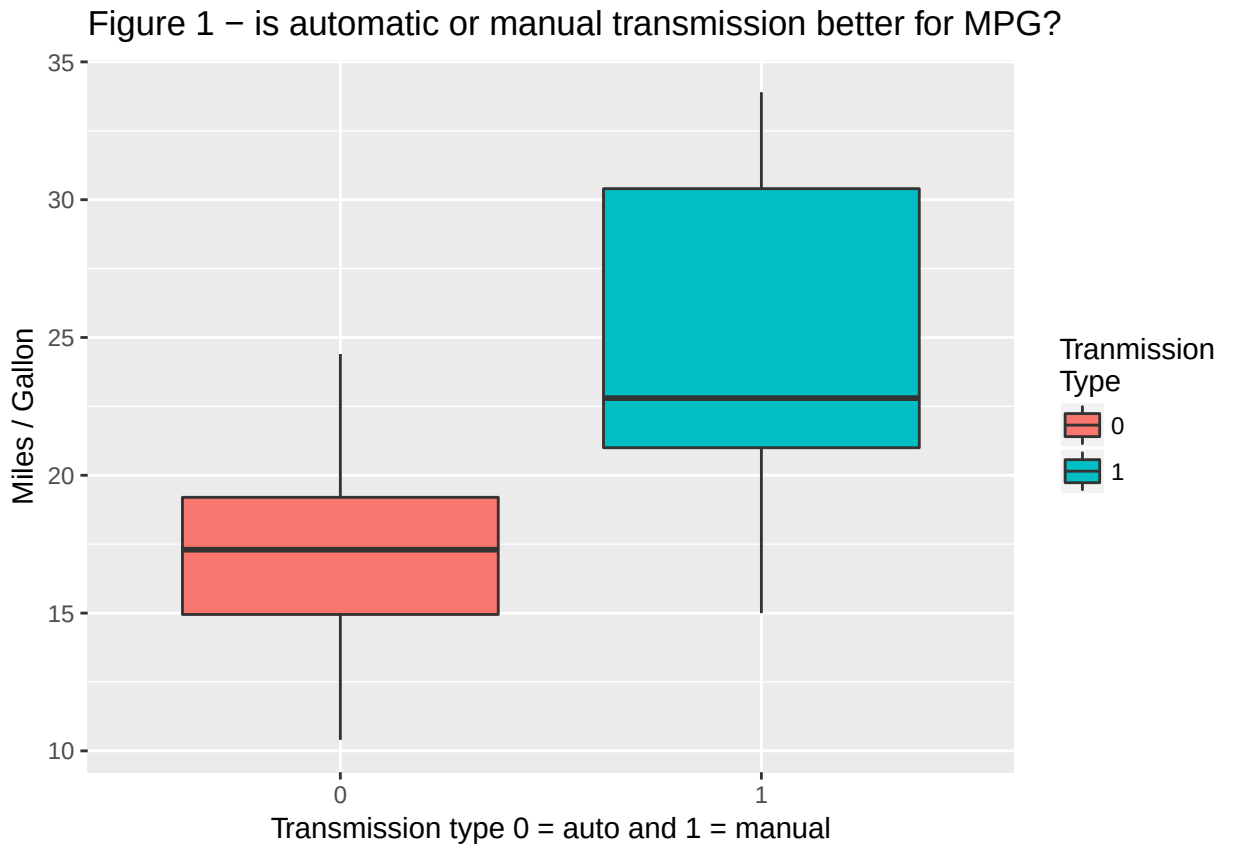


Figure 1 indicates that manual transmission is better for MPG than automatic transmission. We could check if this diference is significant.

Checking the significance of the observed MPG diferences

Let's consider:
- Null hypothesis: the MPG of the automatic and manual transmission has no difference.
- Alternative hypothesis: there difference of the MPG for the automatic and manual transmissions is significant.

The hypothesis can be test with a t-test

```
t.res <- t.test(mpg ~ factor(am), data = cardata)
t.res
```

```
##
```

```
##  Welch Two Sample t-test
##
## data:  mpg by factor(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 is smaller than 0.05, so the null hypothesis is rejected and it can be said that there is a significant difference between for the MPG of the automatic and manual cars. The mean MPG for cars with automatic transmission is 17.15, while for cars with manual transmission is 24.39 miles per gallon.

"Quantify the MPG difference between automatic and manual transmissions"

Regression Models

```
linReg <- lm(mpg ~ factor(am), data = cardata)
summary(linReg)
```

```
##
## Call:
## lm(formula = mpg ~ factor(am), data = cardata)
##
## 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
```

Using a linear model to quantify the MPG difference between automatic and manual transmissions we obtain that automatic cars have 17.15 average MPG, whereas manual ones have 7.24 MPG higher. Additionally, despite the low p-values, the adj-R-squared is 0.33, which corresponds to a model that is accounting for only ~33 % of the MPG variance, a very poor model.

Improving the model

We can try to improve the model by using all the variables.

```
fullLinReg <- lm(mpg ~. , data = cardata)
summary(fullLinReg)
```

```
##
## Call:
```

```
## lm(formula = mpg ~ ., data = cardata)
##
## 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
```

```
# plot(residuals(fullLinReg))
```

Using all the variables in the model we got an adj-R-Squared of 0.81, however, p-values are all greater than 0.05 which indicates no significance of the variables to the model. We must use only significant variables in the model, which can be done with the step function.

Selecting significant variables

```
stepmodel <- step(lm(mpg ~. , data = cardata), trace = 0, steps = 1000)
summary(stepmodel)
```

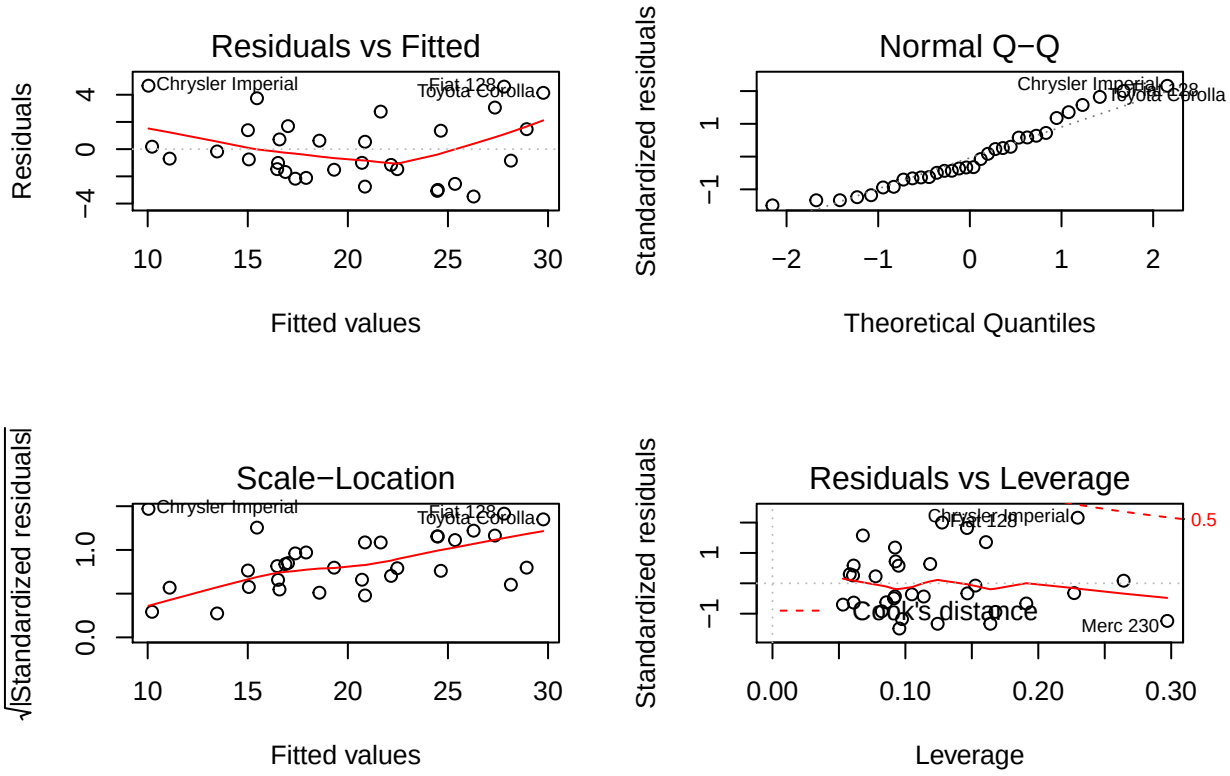
```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = cardata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4811 -1.5555 -0.7257  1.4110  4.6610
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.6178     6.9596   1.382 0.177915
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## am            2.9358     1.4109   2.081 0.046716 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 2.459 on 28 degrees of freedom
## Multiple R-squared:  0.8497, Adjusted R-squared:  0.8336
## F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

With the selected significant variables (wt, qsec and am) the model accounted to the explanation of ~ 83% of the MPF variance. Am coef indicates that manual transmission cars have 2.9358 more mpg than automatic transmission cars on average when all other variables are constant.

Analysis and Diagonostics of Residuals

```
# Residual plot
par(mfrow = c(2,2))
plot(stepmodel)
```



- In the Residuals vs Fitted values plot no pattern can be recognized. The points are scattered around zero.
- Normal Q-Q plot most part of the data points fall on the normal distribution line.
- Scale-Location indicates constant variance with the scattered points pattern.
- Residuals vs Leverage shows that all values are in the 0.5 range indicating that no outliers are present in the dataset.

Conclusion

MPG analysis showed that manual transmission cars are better in terms of fuel consumption when compared to automatic cars with and average MPG of 24.39 for manual cars, 7.24 MPG higher than the average of automatic cars.