Regression models course project

Executive Summary

This assignment involves analysing the mtcars dataset to investigate and explore the relationship between a set of variables and miles per gallon (MPG) and to answer the following questions - "Is an automatic or manual transmission better for MPG" and "What is the quantified MPG difference between automatic and manual transmissions"

My analysis would involve the following steps in sequence- 1. Loading the dataset and performing basic EDA. 2. Comparing the MPG values for automatic and manual transmission by doing a t-test. 3. Building linear models. 4. Residuals and diagnostics.

Importing Dataset and doing basic EDA

Lets load mtcars dataset and perform some EDA.

```
df <- mtcars
str(df)
```

```
## '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 ...
```

```
head(df)
```

```
##                   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
```

So we have 11 variables including mpg and 32 observations and all variables are of numeric class, variable am represents transmission (0 = automatic, 1 = manual). More details can be found from mtcars documentation.

Comparing mpg values for automatic and manual trasmission groups

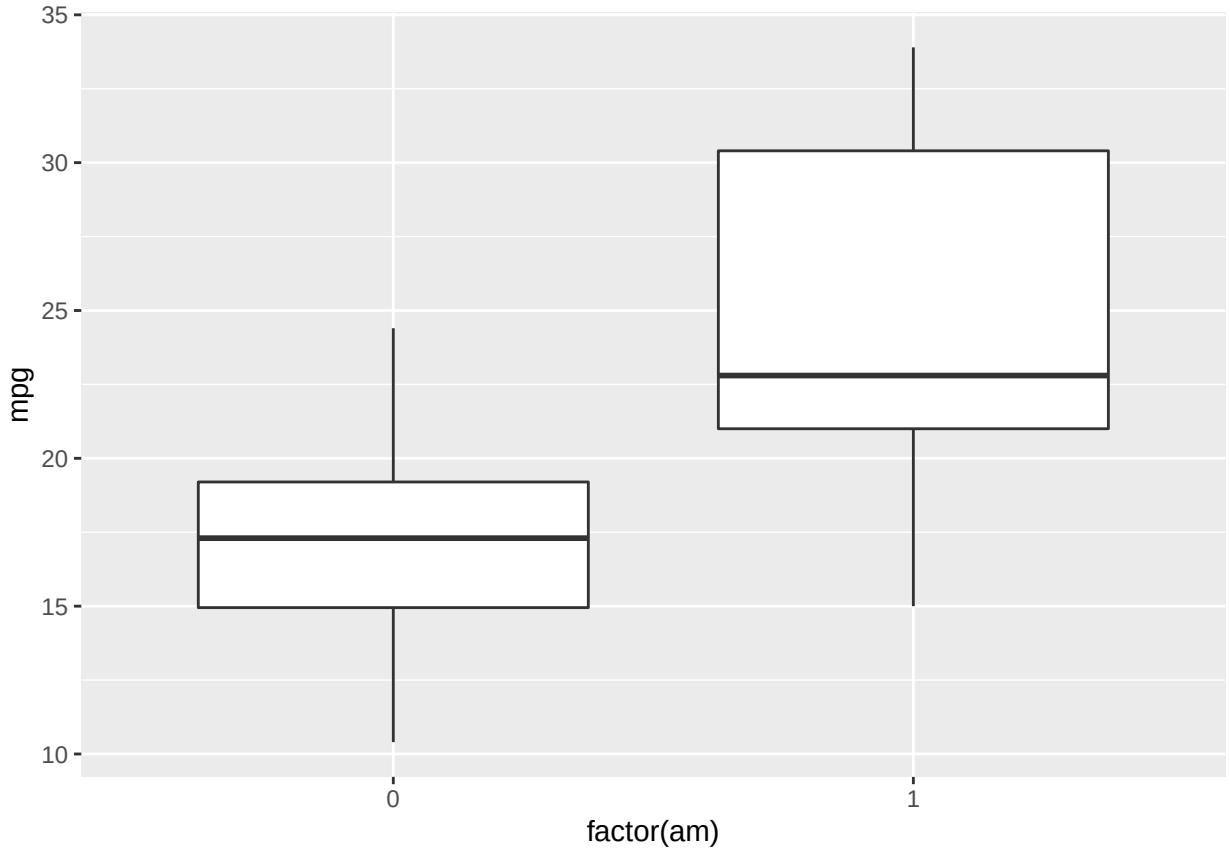
Lets now do a t-test to compare if there is any significant difference between mpg values for automatic and manual transmission. Our null hypothesis is that the difference in mean in zero.

```
t.test(df[df$am == 0,]$mpg, df[df$am == 1,]$mpg)
```

```
##
##  Welch Two Sample t-test
##
## data:  df[df$am == 0, ]$mpg and df[df$am == 1, ]$mpg
## 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 of x mean of y
##  17.14737  24.39231
```

From the t-test since $p < 0.05$ we conclude that difference in mpg values for automatic and manual transmission is significant. The mean mpg for automatic transmission is 17.14 miles/gallon and for manual is about 24.39 miles/gallon. Hence, mpg for manual transmission is about 7.24 miles/gallon higher. We can see the plot of mpg versus transmission(am) as below.

```
library(ggplot2)
ggplot(df, aes(x = factor(am), y = mpg)) + geom_boxplot()
```



Building linear regression models

So far it seems like mpg values for manual transmission is higher than automatic transmission. But we need to build linear models to model the actual relationship and include other variables.

So first lets build a linear model with all variables included.

```
fit <- lm(mpg ~ ., data = df)
summary(fit)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = df)
##
## 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
```

We can see that not a single coefficient has $p < 0.05$ hence, none of the them are significant. So now what we can do is to select variables using stepwise algorithm by using step function. I have hid the results to maintain report length short. However you can cross check the results.

```
stepwise_fit <- step(fit)
summary(stepwise_fit)
```

The final model with the least AIC includes the following variables - wt, qsec, am. The summary of the model shows that all coefficients have $p < 0.05$ and the model explains about 84% of the variance. Moreover the coefficients for qsec and am are positive and for wt negative. Correspondingly, we can interpret that mpg increases from automatic to manual transmission hence, manual transmission has higher mpg.

Residuals and Diagnostics

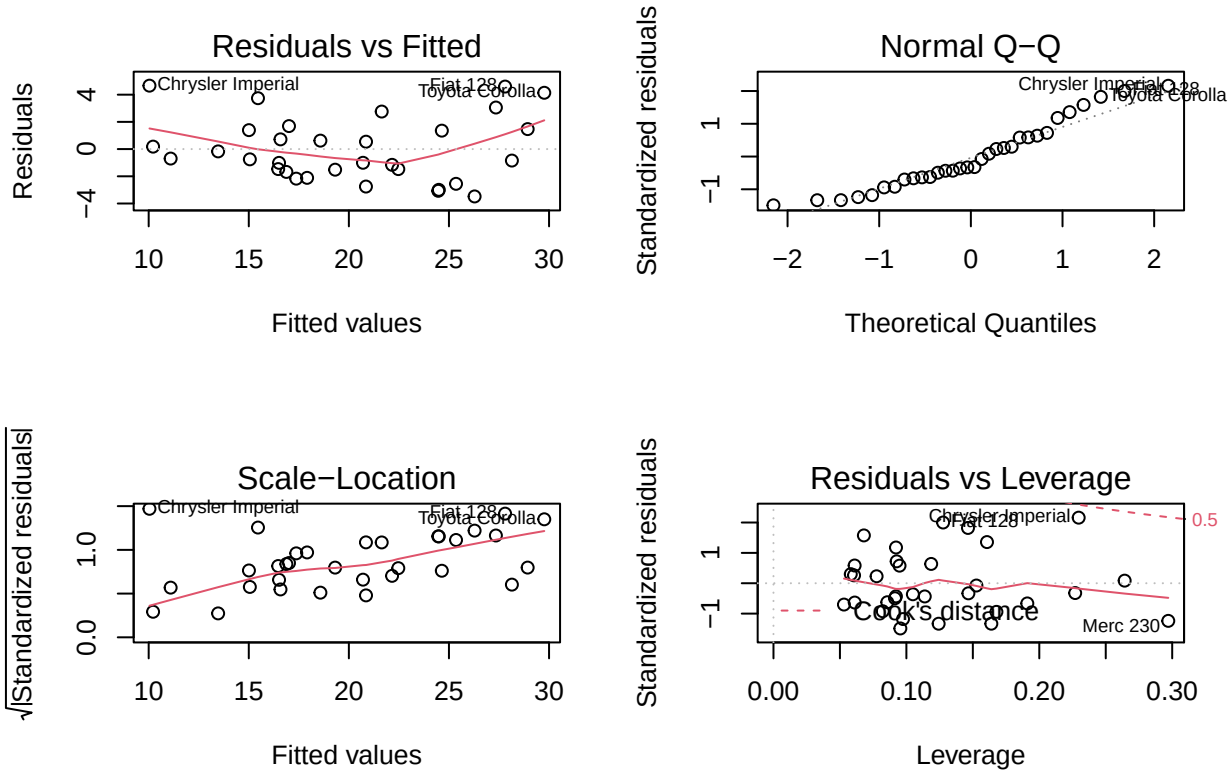
Now we shall do some residual analysis and diagnostics. First let us look at influence measures for our data points with our final model. I have hid the results to maintain report length short.

```
final_fit <- lm(mpg ~ wt + qsec + am, data = df)
influence.measures(final_fit)
```

We don't see any huge deviations in any of the influence measures like hat values, dfb, dffit hence, we conclude that we don't have and outliers/ entry errors and don't need to exclude any data points.

Now lets look at the residual plots for our data.

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



The residual plots seem to be ok and we don't see any pattern hence, we can conclude residuals are random as expected.

Conclusion

We can conclude by answering our initial questions by saying that manual transmission has higher mpg than automatic transmission by 7.24 miles/gallon. Therefore, manual transmission is better for higher mpg. Our assumptions include that the data points are iid. We can do normality testing on our data points to check for that.