Executive Summary

Let's take a look and investigate the dataset mtcars. According to the R description, the data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973-74 models).

This is a data frame with 32 observations on 11 variables. See variables description via ?mtcars in R.

We have to analyze this data in the following context: "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)."

The following questions must be answered from this data:

-Is an automatic or manual transmission better for MPG?

-Quantify the MPG difference between automatic and manual transmissions?

Analisys

Let's investigate the distribution and see Quantile-Quantile plot to verify normality (see Figure 1).

In the graph (see Figure 1) we can see that there are many points away from the 1st and 3rd quantile, suggesting that this distribution should not be normal. We need to use other methods to investigate this data.

Let's see the means:

```
library(dplyr)
summarise(group_by(mtcars, am), mean(mpg), sd(mpg), qty=n())
```

At first sight, the mpg mean for automatics are lower than those that are manual. So automatics seem to be more inefficient but we cannot conclude that yet.

```
summary(lm(data=mtcars, formula=mpg ~ factor(am)))
```

If we try to create a linear regression model like above, we can see above that the coefficients are the means of both automatic and manual measured mpg. We can see that this regression can explain about 35.98% of the variance of the data. See Figure 2 for residuals information.

The residuals are in the range [-10, 10], as we can see in the y-axis. There is a residual standard error of 4.90.

Now let's switch to the full model, evaluate all the 11 variables through a linear regression to see if there is improvement. Let's see what happens when the try a regression similar to the former, selecting more variables:

```
library(MASS)
lm_full<-lm(mpg~ ., data=mtcars)
step <- stepAIC(lm_full, direction="both")
step$anova
```

This will select the following variables for us (wt, qsec and am), and will be able to explain about 84.97% of the variance of the data, which is higher than 35.98% from the previous one. Let's stick with this one and investigate further.

The residuals now are in Figure 3.

```
lm(data=mtcars, formula=mpg ~ wt + qsec -am +factor(am))
```

The new interval in the y-axis are about [-4,5] which is much lower than the previous one of [-10, 10]. This model tells us that an manual car can run in average more 2.935 miles per galon than a automatic car (this information is in am line in the summary above). There is a residual standard error of 2.459 (lower than the previous one of 4.90).

However, we have a very huge standard error in the intercept (factor(am)0) of about 6.96 (the slope (factor(am)1) is 2.935 and the error in the slope is 1.41, high as well). Aside from that, if we compare F-Statistic value, we can verify that the latter model is a more relevant (has much higher F-statistic) than the previous one. This tells us that the difference between the groups is more relevant when we use the augmented model. The last model has some high p-values for some of the variables, which means that we added some noise to the model.

In order to elaborate a little more about the slopes of the am variable, let's check the 95% confidence interval of the slopes from the latter model:

```
confint(lm(data=mtcars, formula=mpg ~ wt + qsec -am +factor(am)))
```

```
##                    2.5 %    97.5 %
## (Intercept) -4.63829946 23.873860
## wt          -5.37333423 -2.459673
## qsec         0.63457320  1.817199
## factor(am)1  0.04573031  5.825944
```

With this information, we will choose the latter model to conclude our analysis.

Conclusion

In the light of the analysis done so far, is the data about the transmission sufficient do quantify the difference between the groups and tell which one is better? If we take a look at the standard error of the base variable (intercept, which is factor(am)0 - manual cars), we'll see a number about 6.959, which are quite huge in the latter model. Let's switch to the othe related variable, factor(am)1 and conclude the analysis.

Taking into the account the 95% confidence interval calculated above, we can see that the manual cars are more effective than automatic cars in about [0.04, 5.825] mpg - which is the 95% confidence interval of the factor(am)1 (manual cars), with an average of 2.936 mpg better (the best fit in the linear regression model).

Appendix

Figure 1:

```
qqnorm(mtcars$mpg)
qqline(mtcars$mpg) ## plots a line that goes from 0.25 to 0.75 quantile (2nd and 4th quantile)
```

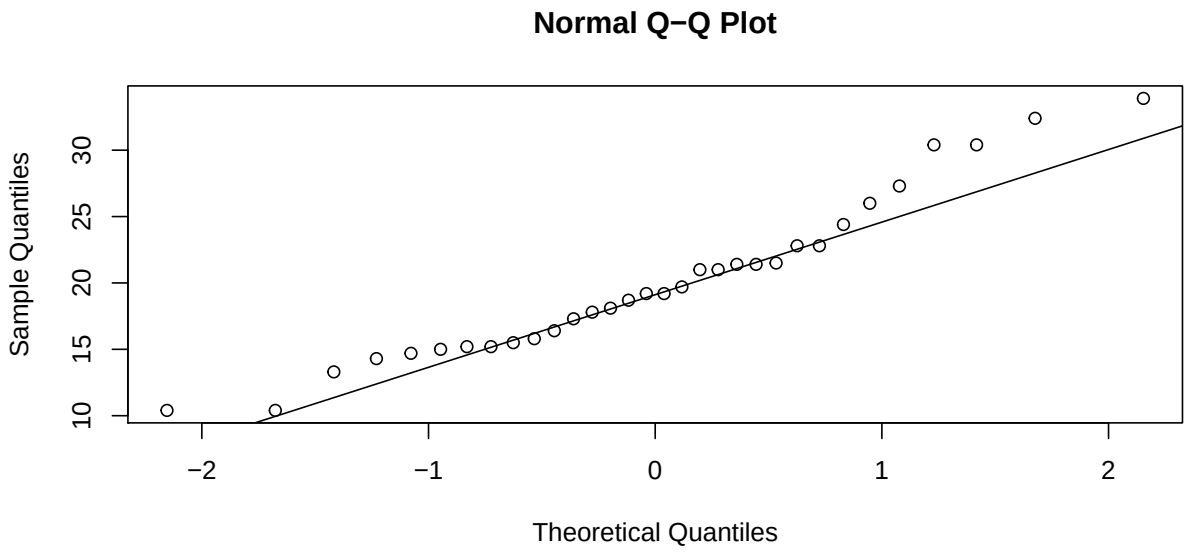


Figure 2:

```
plot(resid(lm(data=mtcars, formula=mpg ~ factor(am))),
     ylab="Residuals", xlab="mpg", main="Residuals x mpg")
```

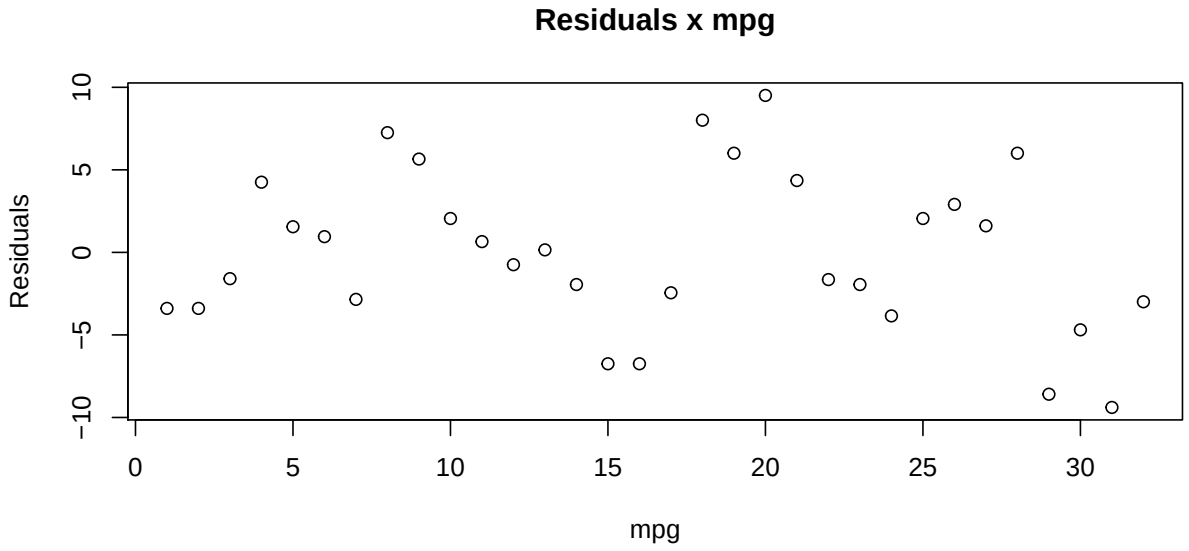
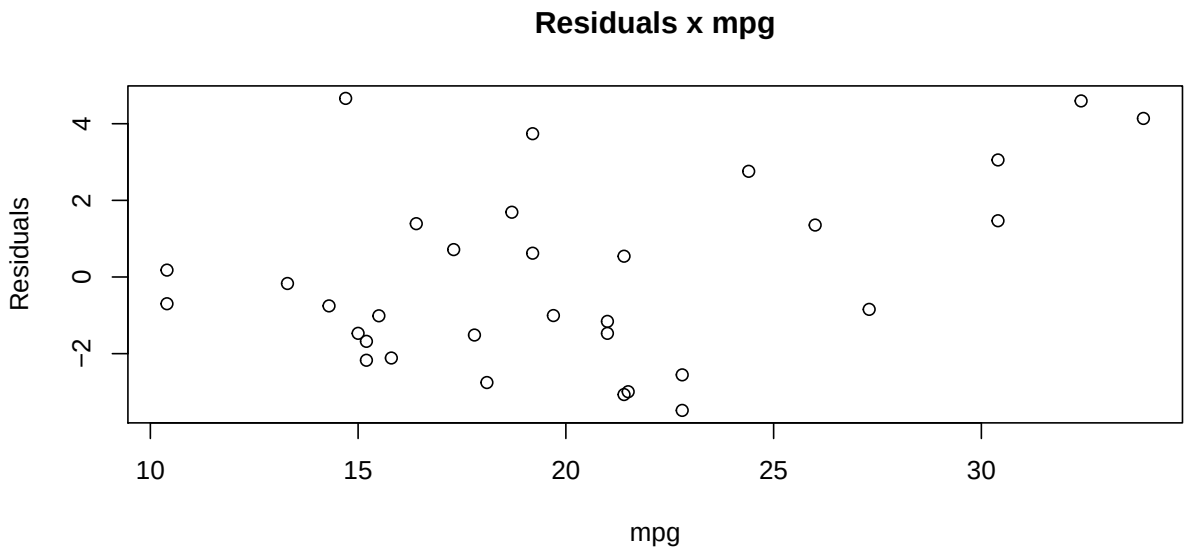


Figure 3:

```
plot(mtcars$mpg, resid(lm(data=mtcars, formula=mpg~ wt + qsec -am + factor(am))),
     ylab="Residuals", xlab="mpg", main="Residuals x mpg")
```


Residuals x mpg
Residuals
mpg
−2
0
2
4
10
15
20
25
30