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:

```r
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.1.2

##
## Attaching package: 'dplyr'
##
## The following object is masked from 'package:stats':
##
##     filter
##
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

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

```
## Source: local data frame [2 x 4]
##
##   am mean(mpg)  sd(mpg) qty
## 1  0  17.14737 3.833966  19
## 2  1  24.39231 6.166504  13
```

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.

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

```
##
## 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
```

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:

```r
library(MASS)
```

```
##
## Attaching package: 'MASS'
##
## The following object is masked from 'package:dplyr':
##
##     select
```

```r
lm_full<-lm(mpg~ ., data=mtcars)
step <- stepAIC(lm_full, direction="both")
```

```
## Start:  AIC=70.9
## mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##
##        Df Sum of Sq    RSS    AIC
## - cyl   1    0.0799 147.57 68.915
## - vs    1    0.1601 147.66 68.932
## - carb  1    0.4067 147.90 68.986
## - gear  1    1.3531 148.85 69.190
## - drat  1    1.6270 149.12 69.249
```

```
## - disp  1    3.9167 151.41 69.736
## - hp    1    6.8399 154.33 70.348
## - qsec  1    8.8641 156.36 70.765
## <none>              147.49 70.898
## - am    1   10.5467 158.04 71.108
## - wt    1   27.0144 174.51 74.280
##
## Step:  AIC=68.92
## mpg ~ disp + hp + drat + wt + qsec + vs + am + gear + carb
##
##        Df Sum of Sq    RSS    AIC
## - vs    1    0.2685 147.84 66.973
## - carb  1    0.5201 148.09 67.028
## - gear  1    1.8211 149.40 67.308
## - drat  1    1.9826 149.56 67.342
## - disp  1    3.9009 151.47 67.750
## - hp    1    7.3632 154.94 68.473
## <none>              147.57 68.915
## - qsec  1   10.0933 157.67 69.032
## - am    1   11.8359 159.41 69.384
## + cyl   1    0.0799 147.49 70.898
## - wt    1   27.0280 174.60 72.297
##
## Step:  AIC=66.97
## mpg ~ disp + hp + drat + wt + qsec + am + gear + carb
##
##        Df Sum of Sq    RSS    AIC
## - carb  1    0.6855 148.53 65.121
## - gear  1    2.1437 149.99 65.434
## - drat  1    2.2139 150.06 65.449
## - disp  1    3.6467 151.49 65.753
## - hp    1    7.1060 154.95 66.475
## <none>              147.84 66.973
## - am    1   11.5694 159.41 67.384
## - qsec  1   15.6830 163.53 68.200
## + vs    1    0.2685 147.57 68.915
## + cyl   1    0.1883 147.66 68.932
## - wt    1   27.3799 175.22 70.410
##
## Step:  AIC=65.12
## mpg ~ disp + hp + drat + wt + qsec + am + gear
##
##        Df Sum of Sq    RSS    AIC
## - gear  1     1.565 150.09 63.457
## - drat  1     1.932 150.46 63.535
## <none>              148.53 65.121
## - disp  1    10.110 158.64 65.229
## - am    1    12.323 160.85 65.672
## - hp    1    14.826 163.35 66.166
## + carb  1     0.685 147.84 66.973
## + vs    1     0.434 148.09 67.028
## + cyl   1     0.414 148.11 67.032
## - qsec  1    26.408 174.94 68.358
## - wt    1    69.127 217.66 75.350
```

```
##
## Step:  AIC=63.46
## mpg ~ disp + hp + drat + wt + qsec + am
##
##        Df Sum of Sq    RSS    AIC
## - drat  1     3.345 153.44 62.162
## - disp  1     8.545 158.64 63.229
## <none>              150.09 63.457
## - hp    1    13.285 163.38 64.171
## + gear  1     1.565 148.53 65.121
## + cyl   1     1.003 149.09 65.242
## + vs    1     0.645 149.45 65.319
## + carb  1     0.107 149.99 65.434
## - am    1    20.036 170.13 65.466
## - qsec  1    25.574 175.67 66.491
## - wt    1    67.572 217.66 73.351
##
## Step:  AIC=62.16
## mpg ~ disp + hp + wt + qsec + am
##
##        Df Sum of Sq    RSS    AIC
## - disp  1     6.629 160.07 61.515
## <none>              153.44 62.162
## - hp    1    12.572 166.01 62.682
## + drat  1     3.345 150.09 63.457
## + gear  1     2.977 150.46 63.535
## + cyl   1     2.447 150.99 63.648
## + vs    1     1.121 152.32 63.927
## + carb  1     0.011 153.43 64.160
## - qsec  1    26.470 179.91 65.255
## - am    1    32.198 185.63 66.258
## - wt    1    69.043 222.48 72.051
##
## Step:  AIC=61.52
## mpg ~ hp + wt + qsec + am
##
##        Df Sum of Sq    RSS    AIC
## - hp    1     9.219 169.29 61.307
## <none>              160.07 61.515
## + disp  1     6.629 153.44 62.162
## + carb  1     3.227 156.84 62.864
## + drat  1     1.428 158.64 63.229
## - qsec  1    20.225 180.29 63.323
## + cyl   1     0.249 159.82 63.465
## + vs    1     0.249 159.82 63.466
## + gear  1     0.171 159.90 63.481
## - am    1    25.993 186.06 64.331
## - wt    1    78.494 238.56 72.284
##
## Step:  AIC=61.31
## mpg ~ wt + qsec + am
##
##        Df Sum of Sq    RSS    AIC
## <none>              169.29 61.307
```

```
## + hp    1     9.219 160.07 61.515
## + carb  1     8.036 161.25 61.751
## + disp  1     3.276 166.01 62.682
## + cyl   1     1.501 167.78 63.022
## + drat  1     1.400 167.89 63.042
## + gear  1     0.123 169.16 63.284
## + vs    1     0.000 169.29 63.307
## - am    1    26.178 195.46 63.908
## - qsec  1   109.034 278.32 75.217
## - wt    1   183.347 352.63 82.790
```

```
step$anova
```

```
## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##
## Final Model:
## mpg ~ wt + qsec + am
##
##
##     Step Df   Deviance Resid. Df Resid. Dev      AIC
## 1                             21   147.4944 70.89774
## 2  - cyl  1 0.07987121        22   147.5743 68.91507
## 3   - vs  1 0.26852280        23   147.8428 66.97324
## 4 - carb  1 0.68546077        24   148.5283 65.12126
## 5 - gear  1 1.56497053        25   150.0933 63.45667
## 6 - drat  1 3.34455117        26   153.4378 62.16190
## 7 - disp  1 6.62865369        27   160.0665 61.51530
## 8   - hp  1 9.21946935        28   169.2859 61.30730
```

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:

```r
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:

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

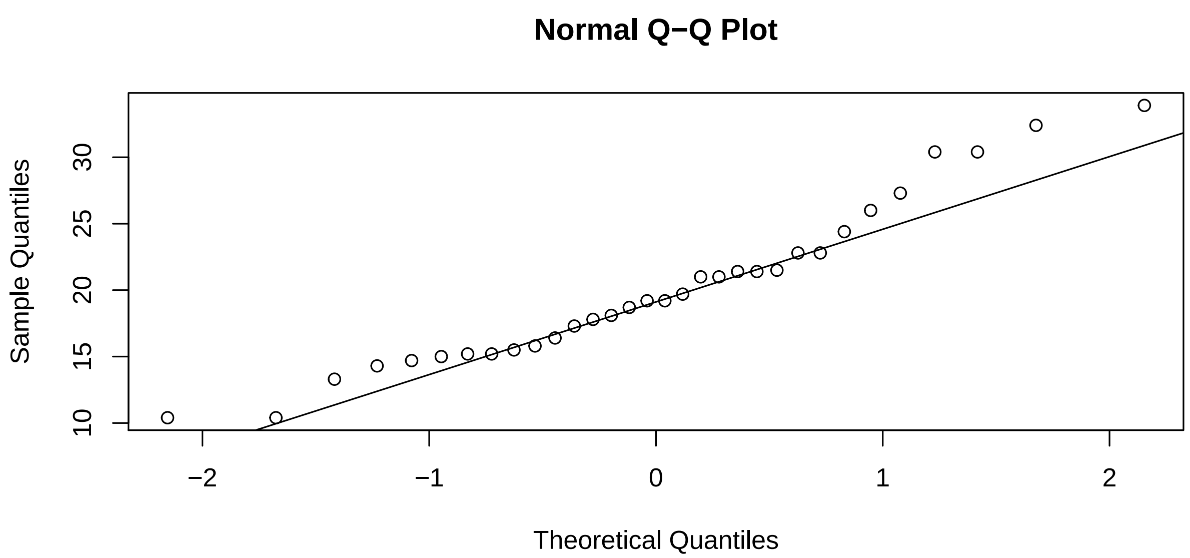

Figure 2:

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

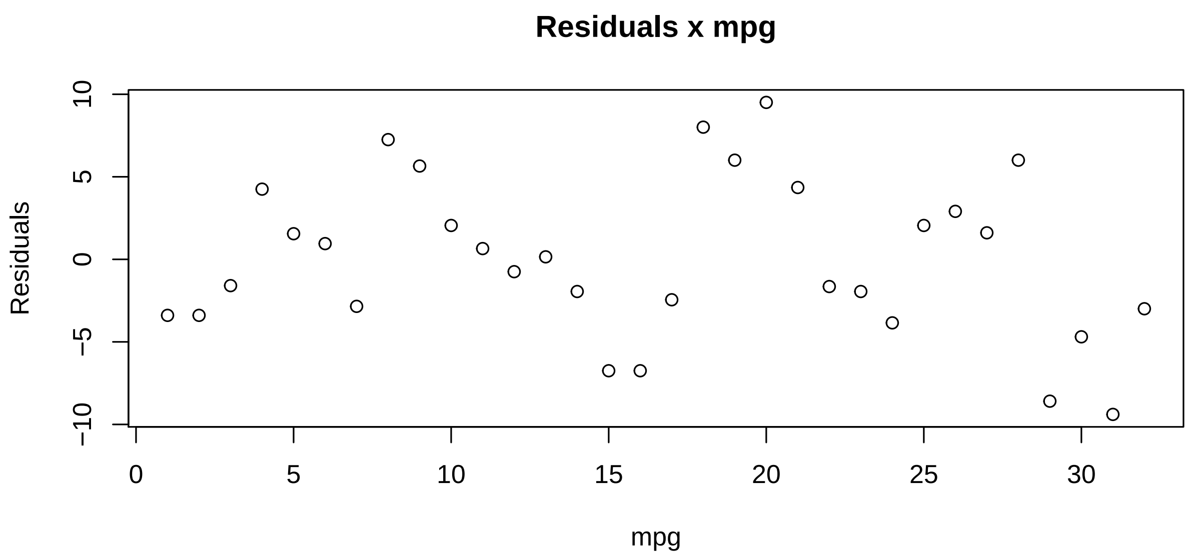

Figure 3:

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