Comparing MPG of cars for automatic and manual transmission

Executive summary

An analysis is performed to examine the effect of transmission type on the mpg of a car. The difference is calculated along with the effect of other variables. The analysis is performed using simple linear regression as well as multiple linear regression. In both the scenarios it can be seen that cars with manual transmission have higher MPG than cars with automatic transmissions.This is true when comparing MPG with transmission type. On considering the effect of the variables wt and qsec the conclusion still holds. The visualizations are attached under the appendix.

Analysis

Simple Linear regression

```
data(mtcars); n <- length(mtcars$mpg); alpha <- 0.05
fit <- lm(mpg ~ am, data = mtcars)
coef(summary(fit))
```

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

beta0 = intercept coeff = mean MPG for cars with automatic transmission
beta1 = mean increase in MPG for cars with manual transmission
beta0 + beta1 = mean MPG for cars with manual transmission

Calculating 95 % confidence interval for beta1

```
pe <- coef(summary(fit))["am", "Estimate"]; se <- coef(summary(fit))["am", "Std. Error"]
tstat <- qt(1 - alpha/2, n - 2)  # n - 2 for model with intercept and slope
pe + c(-1, 1) * (se * tstat)
```

```
## [1]  3.64151 10.84837
```

The p-value of 2.850207^{-4} for beta1 is small and the CI does not include zero.We can reject null in favor of the alternative hypothesis that there is a significant difference in MPG between the two groups at alpha = 0.05.

Multiple Linear regression

The bestglm package is used to select wt, qsec and am in an automated manner as they yield the highest R-squared value.

```
bestfit <- lm(mpg ~ wt + qsec + am, data = mtcars)
coef(summary(bestfit))
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  9.617781  6.9595930  1.381946 1.779152e-01
## wt          -3.916504  0.7112016 -5.506882 6.952711e-06
## qsec         1.225886  0.2886696  4.246676 2.161737e-04
## am           2.935837  1.4109045  2.080819 4.671551e-02
```

The output above is used to calculate the 95 % C.I. for beta3 (am)

```
pe <- coef(summary(bestfit))["am", "Estimate"]; se <- coef(summary(bestfit))["am", "Std.
Error"]
tstat <- qt(1 - alpha/2, n - 2)  # n - 2 for model with intercept and slope
pe + c(-1, 1) * (se * tstat)
```

```
## [1] 0.05438576 5.81728862
```

The p-value of 0.0467 for beta3 is small and the CI does not include zero, so we reject null hypothesis in favor of the alternative hypothesis that there is a significant difference in MPG between the two groups at alpha = 0.05.

Testing using Nested model

```
fit1 <- lm(mpg ~ wt, data = mtcars); fit2 <- update(fit1, mpg ~ wt + qsec);
fit3 <- update(fit2, mpg ~ wt + qsec + am)
anova(fit1, fit2, fit3)
```

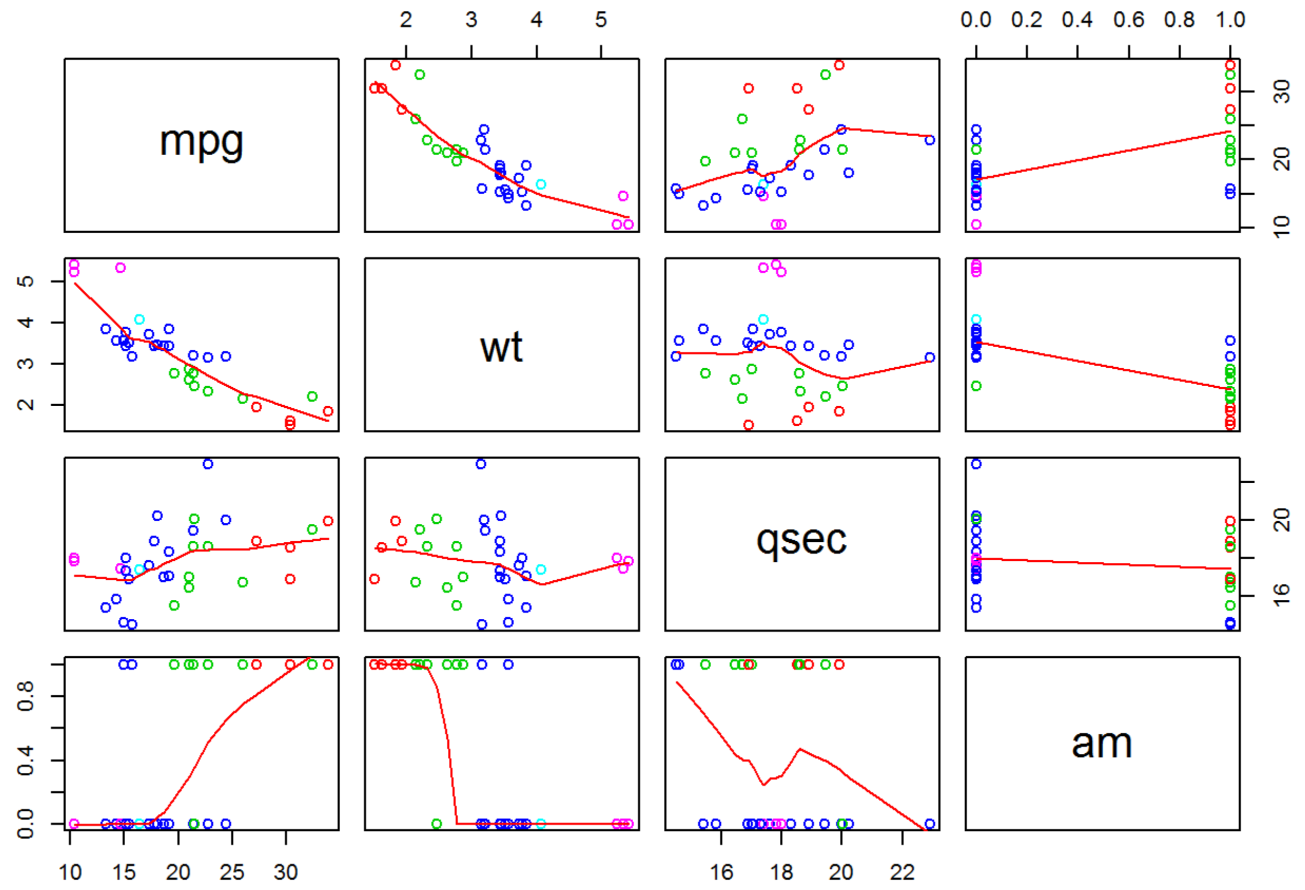
```
## Analysis of Variance Table
##
## Model 1: mpg ~ wt
## Model 2: mpg ~ wt + qsec
## Model 3: mpg ~ wt + qsec + am
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 278.32
## 2     29 195.46  1    82.858 13.7048 0.0009286 ***
## 3     28 169.29  1    26.178  4.3298 0.0467155 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This confirms that all three variables are significant

Appendix - Visualizations and exploratory analysis

Correlations between the variables above in a trellis plot

```
mtcars_vars <- mtcars[, c(1, 6, 7, 9)]
mar.orig <- par()$mar
par(mar = c(1, 1, 1, 1))
pairs(mtcars_vars, panel = panel.smooth, col = 9 + mtcars$wt)
```



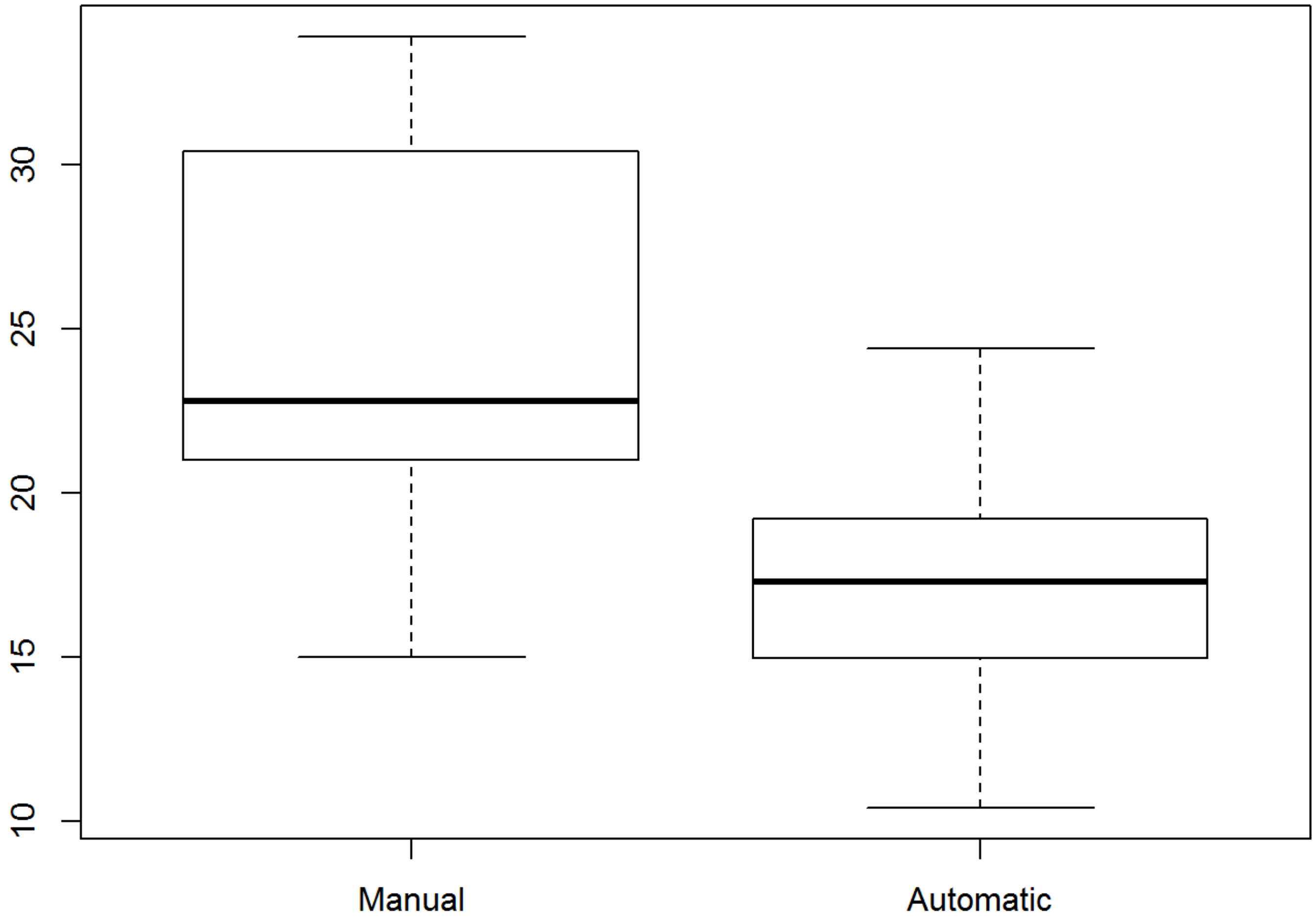
```
par(mar = mar.orig)
cor(mtcars_vars)
```

```
##             mpg         wt       qsec         am
## mpg   1.0000000 -0.8676594  0.4186840  0.5998324
## wt   -0.8676594  1.0000000 -0.1747159 -0.6924953
## qsec  0.4186840 -0.1747159  1.0000000 -0.2298609
## am    0.5998324 -0.6924953 -0.2298609  1.0000000
```

Checking for Homogeneity of variance

Box plots for automatic and manual transmisions

```
mtcars_vars <- mtcars[, c(1, 6, 7, 9)]
mar.orig <- par()$mar  # save the original values
par(mar = c(2, 2, 2, 2))  # set your new values
boxplot(mtcars_vars[mtcars_vars$am == 1, ]$mpg, mtcars_vars[mtcars_vars$am ==
                                                0, ]$mpg, names = c("Manual", "Automati
c"))
```



```
par(mar = mar.orig)
```

Standard deviations of MPG for automatic and manual transmissions

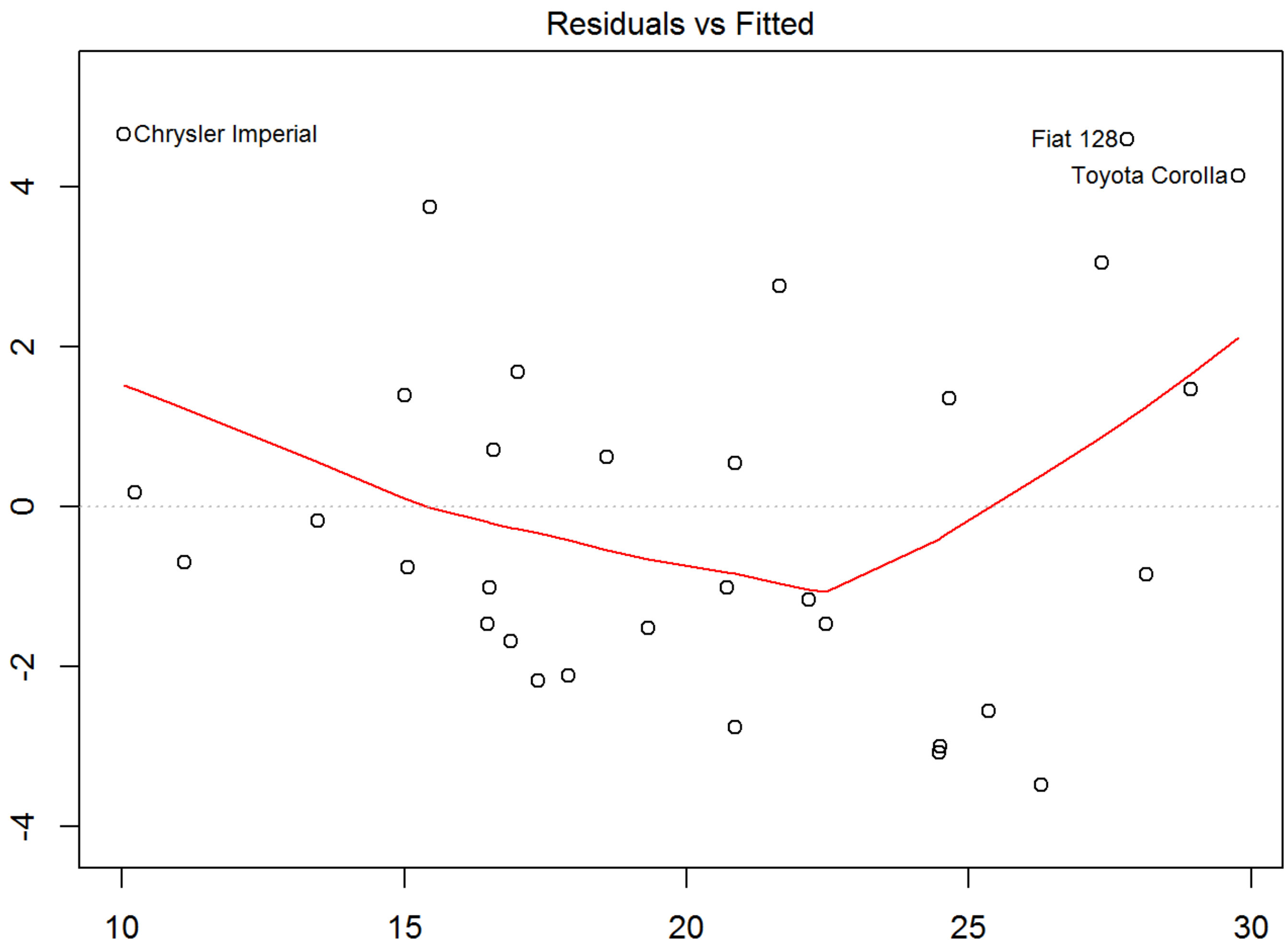
```
by(mtcars_vars$mpg, mtcars_vars$am, sd)
```

```
## mtcars_vars$am: 0
## [1] 3.833966
## --------------------------------------------------------
## mtcars_vars$am: 1
## [1] 6.166504
```

The above chart and test show that the assumption of homogeneity of variance might not be completely true

Residuals plot

```
mar.orig <- par()$mar
par(mar = c(2, 2, 2, 2))
plot(bestfit, which = c(1:1))
```



```
par(mar = mar.orig)
```

The residuals plot is slightly curved above. Models like Chrysler Imperial, Fiat 128 and Toyota Corolla have a significant effect on this plot.