Predictors of MPG

Executive Summary

There were two objective for this assignment:

- Assess whether an automatic or manual transmission is better for MPG
- Quantify the difference in MPG between automatic and manual transmissions

After importing and cleaning the data, I performed an initial t-test to assess if there was a difference in MPG between automatic and manual transmissions, and there was. I then created a series of models to determine if the observed differences could be (partially) explained by other variables. I concluded that, when holding weight and horsepower constant, the transmission type does not appear to significantly influence MPG. However, when ignoring all other variables, automatic transmission vehicles have a significantly higher MPG on average (Auto avg = 24.4, Manual avg = 17.1; p = .001).

Recoding

The `am`, `vs`, and `cyl` variables were recoded as factors.

```
dat <- mtcars
dat$am <- as.factor(mapvalues(dat$am, c(0,1), c("Automatic", "Manual")))
dat$vs <- as.factor(mapvalues(dat$vs, c(0,1), c("V-shaped", "Straight")))
dat$cyl <- as.factor(dat$cyl)
```

Hypothesis Test

A two sample t-test could be used to assess if there is a significant difference between MPG of manual and automatic vehicles.

```
t.test(dat$mpg[dat$am == "Automatic"], dat$mpg[dat$am == "Manual"])
```

```
##
##  Welch Two Sample t-test
##
## data:  dat$mpg[dat$am == "Automatic"] and dat$mpg[dat$am == "Manual"]
## 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
```

The results of the t-test show vehicles with an automatic transmission have a significantly higher MPG (Auto avg = 24.4, Manual avg = 17.1; p = .001). However, other variables could potentially explain/modify the observed difference.

Model Comparison

I created 3 linear models to assess the relationship between MPG and transmission type:

```
mod1 <- lm(mpg ~ am + wt, data = dat)
```

```
mod2 <- lm(mpg ~ am + wt + hp, data = dat)
```

```
mod3 <- lm(mpg ~ am + wt + hp + cyl, data = dat)
```

```
anova(mod1, mod2, mod3)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am + wt
## Model 2: mpg ~ am + wt + hp
## Model 3: mpg ~ am + wt + hp + cyl
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     29 278.32
## 2     28 180.29  1    98.029 16.8762 0.0003525 ***
## 3     26 151.03  2    29.265  2.5191 0.0999982 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Th ANOVA printout shows that the 2nd model is significantly better than the 1st, but the 3rd model isn't much better than the 2nd So, I'll use the 2nd model. I also chose the 2nd model because the 3rd model splits the subgroups by cylinder size, which lead to very small sample sizes:

```
table(dat$am, dat$cyl)
```

```
##
##              4  6  8
##   Automatic  3  4 12
##   Manual     8  3  2
```

Referring back to the 2nd model, you can see that the `amManual` is not significant; this suggests that **there is not a meaningful difference in MPG between automatic and manual transmission vehicles when holding weight and horsepower constant**.

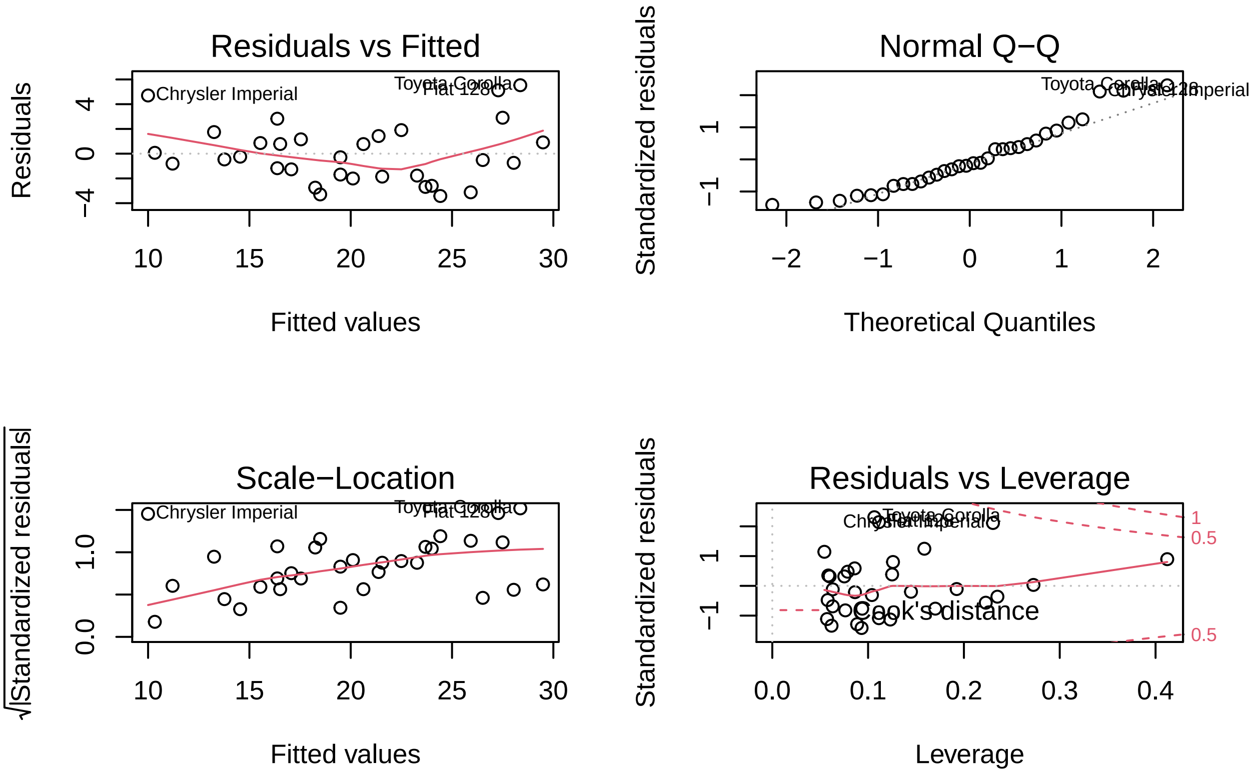
```
summary(mod2)
```

```
##
## Call:
## lm(formula = mpg ~ am + wt + hp, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4221 -1.7924 -0.3788  1.2249  5.5317
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.002875   2.642659  12.867 2.82e-13 ***
## amManual     2.083710   1.376420   1.514 0.141268
## wt          -2.878575   0.904971  -3.181 0.003574 **
## hp          -0.037479   0.009605  -3.902 0.000546 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 2.538 on 28 degrees of freedom
## Multiple R-squared:  0.8399, Adjusted R-squared:  0.8227
## F-statistic: 48.96 on 3 and 28 DF,  p-value: 2.908e-11
```

Diagnostics

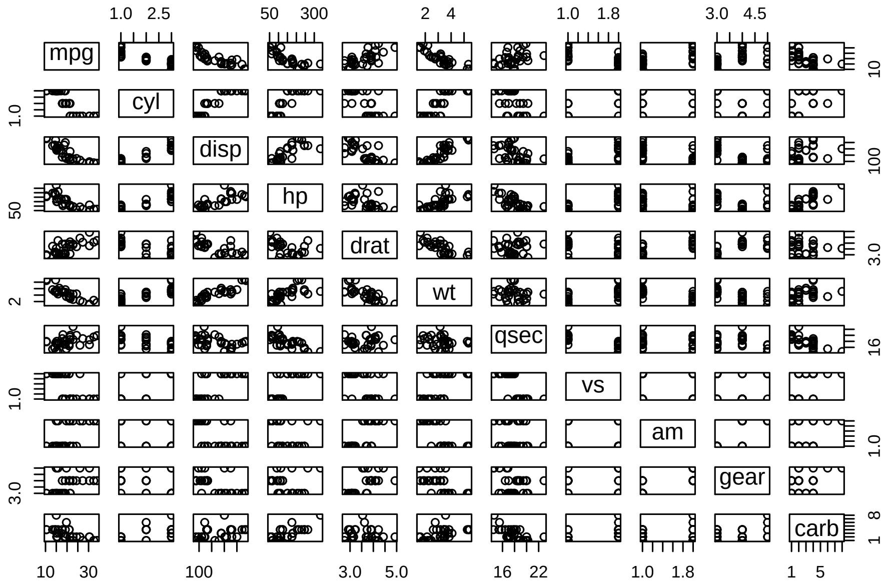
```r
par(mfrow = c(2, 2))
plot(mod2)
```



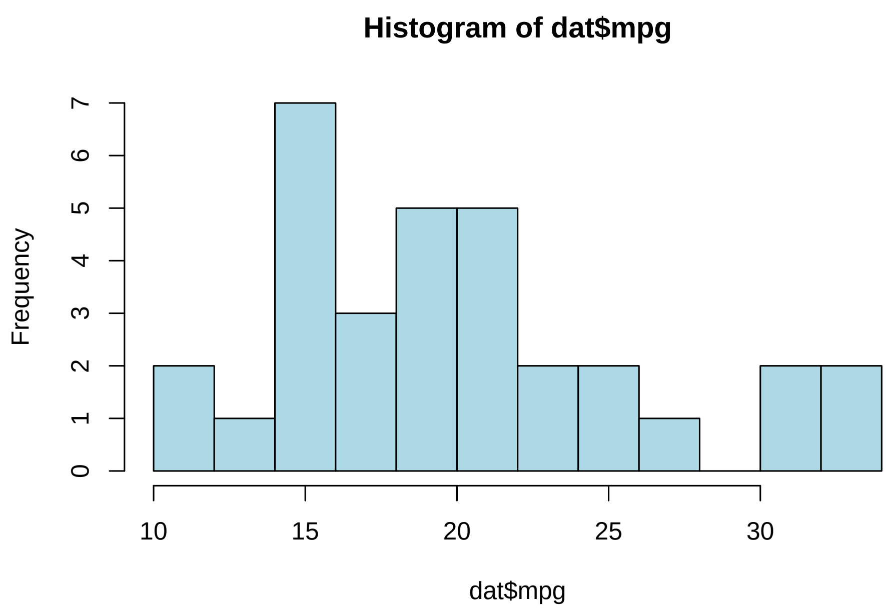
The diagnostics for this model are fairly good. The normal Q-Q plot has several points that do not align with the theoretical quantiles, but these points do not appear to have much leverage.

Supporting Figures

```r
plot(dat)
```



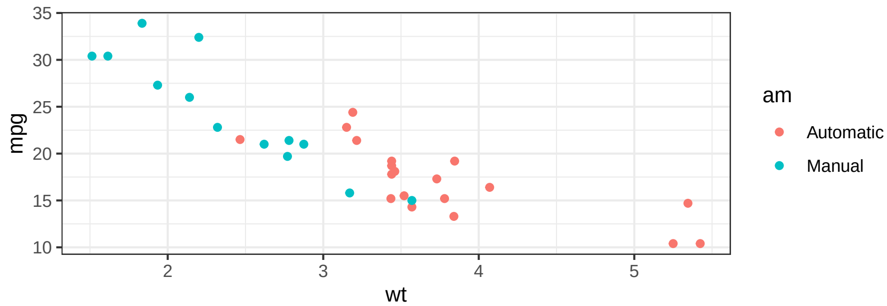

```r
hist(dat$mpg, col = "lightblue", breaks = 10)
```

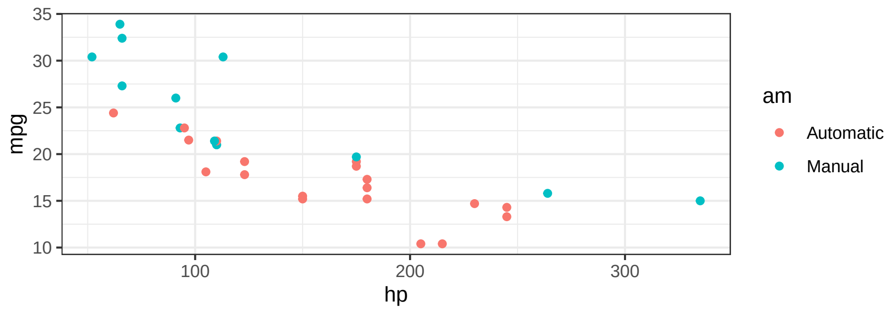



```r
p1 <-
ggplot(dat, aes(wt, mpg, col= am))+
  geom_point() +
  theme_bw()

p2 <-
ggplot(dat, aes(hp, mpg, col= am))+
  geom_point() +
  theme_bw()

gridExtra::grid.arrange(p1, p2)
```






```
ggplot(dat, aes(x = am, y = mpg, fill = cyl)) +
  geom_boxplot() +
  theme_bw()
```