Regression Session

Executive summary

The impact on transmission class (automatic vs manual) on vehicle miles per gallon (MPG) was analyzed, along with several other data that may explain variation in MPG rating. It was found that, while vehicle weight was the principal determinant of MPG, transmission class had the second largest impact of the measured vehicle dimensions, and an automatic transmission decreases the MPG of a vehicle beyond our 95% confidence interval.

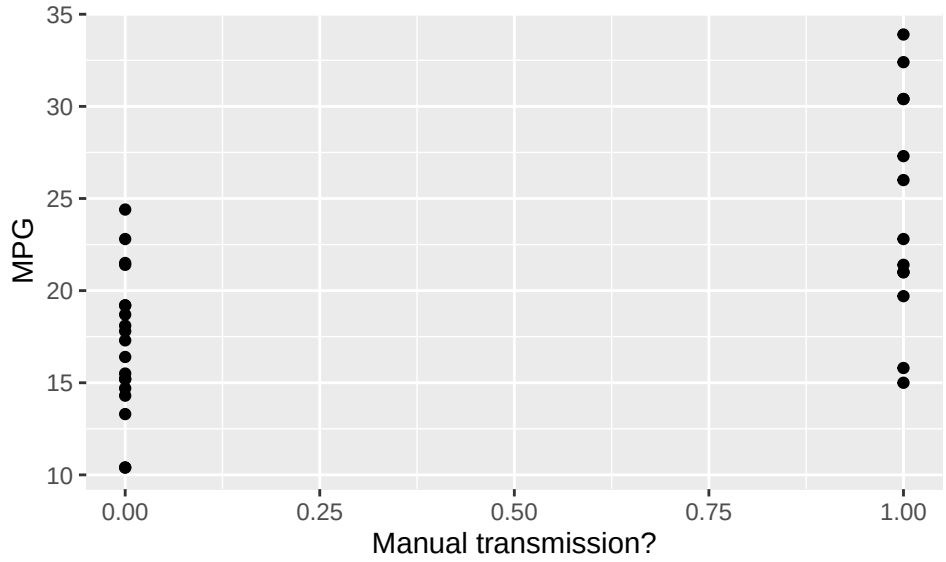
Exploratory Data Analysis

Let's have a look at the data.

```
library(ggplot2)
data(mtcars)
#mtcars$am <- factor(mtcars$am)
colnames(mtcars)
```

```
##  [1] "mpg"  "cyl"  "disp" "hp"   "drat" "wt"   "qsec" "vs"   "am"   "gear"
## [11] "carb"
```

```
ggplot(mtcars, aes(y=mtcars$mpg, x=mtcars$am, group=mtcars$am)) +
        geom_point() +
        ylab('MPG') +
        xlab('Manual transmission?')
```



It seems that the manual transmission cars (x=1) have a higher MPG on average. We will follow up with some summary stats and do a regression analysis.

```
## [1] "Automatic transmission mean mpg"
```

```
## [1] 17.14737
```

```
## [1] "Manual transmission mean mpg"
```

```
## [1] 24.39231
```

```
##
##  Welch Two Sample t-test
##
## data:  mtcars$mpg[mtcars$am == 0] and mtcars$mpg[mtcars$am == 1]
## 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 t-test indicates that the null hypothesis remains outside of the 95% confidence interval, thus we infer that the mpg ratings for automatic and manual cars are truly drawn from separate distributions.

Can we model this difference?

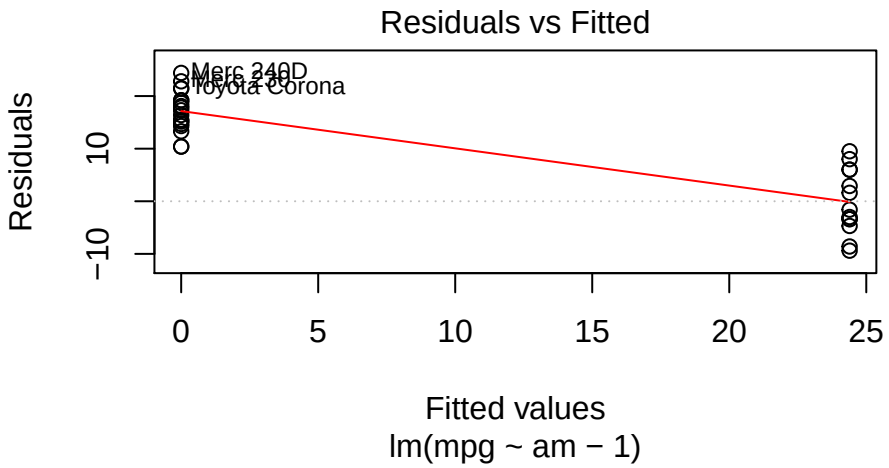
Preliminary linear model fitting

Now we'll have a look at a very simple univariate linear model, using only transmission class as a predictor:

```
model <- lm(mpg ~ am-1, data=mtcars)
summary(model)
```

```
##
## Call:
## lm(formula = mpg ~ am - 1, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.392  2.583 13.800 17.875 24.400
##
## Coefficients:
##    Estimate Std. Error t value Pr(>|t|)
## am   24.392      3.956   6.166 7.67e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.26 on 31 degrees of freedom
## Multiple R-squared:  0.5508, Adjusted R-squared:  0.5363
## F-statistic: 38.01 on 1 and 31 DF,  p-value: 7.666e-07
```

```
plot(model, which = 1)
```



It seems that the residuals are substantially higher for the automatic transmission cars (dummy var = 0) than for the manuals. Perhaps there is another variable that will explain the variation that the model is missing.

A better model?

Let's include all of the varibles in the model and scrutinze their regression coeffecients.

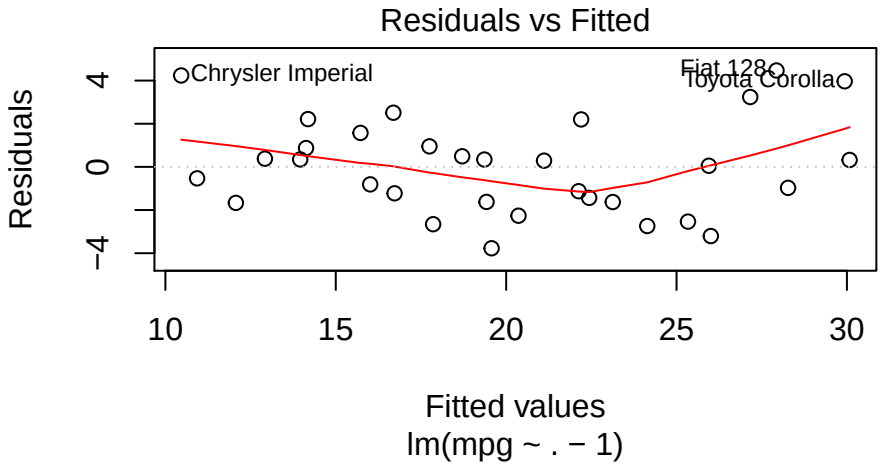
Are there any other factors that we can account for in the dataset that influence MPG? We will employ a linear model to fit the data, because this is a simple model that is readily interpreted post hoc.

```
model <- lm(mpg ~.-1, data=mtcars)
summary(model)
```

```
##
## Call:
## lm(formula = mpg ~ . - 1, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7721 -1.6249  0.1699  1.1068  4.4666
##
## Coefficients:
##      Estimate Std. Error t value Pr(>|t|)
## cyl   0.35083    0.76292   0.460   0.6501
## disp  0.01354    0.01762   0.768   0.4504
## hp   -0.02055    0.02144  -0.958   0.3483
## drat  1.24158    1.46277   0.849   0.4051
## wt   -3.82613    1.86238  -2.054   0.0520 .
## qsec  1.19140    0.45942   2.593   0.0166 *
## vs    0.18972    2.06825   0.092   0.9277
## am    2.83222    1.97513   1.434   0.1656
## gear  1.05426    1.34669   0.783   0.4421
## carb -0.26321    0.81236  -0.324   0.7490
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.616 on 22 degrees of freedom
## Multiple R-squared:  0.9893, Adjusted R-squared:  0.9844
## F-statistic:   203 on 10 and 22 DF,  p-value: < 2.2e-16
```

```
plot(model, which = 1)
```



Model interpretation

Looking at the regression coefficients, we see that vehicle weight has the greatest effect on the MPG rating, but the transmission class has the second greatest weight.

This makes intuitive sense, given that heavier things require more gas to push.

This model seems pretty good- the residuals appear more or less randomly distributed.

The R^2 is also much closer to 1 than the previous model, indicating a better fit to the data.

From the residual coefficient for transmission class:

```
exp(2.83222)
```

```
## [1] 16.98312
```

```
exp(1.97513)
```

```
## [1] 7.207557
```

Thus, in our model, having a manual transmission correlates with ~17 additional MPG for the car with a standard error in this estimate of 7.2 MPG.