Regression Models Course Project

Executive Summary

This is a project for the Regression Models course, part of the John Hopkins University Data Science Specialization on Coursera. Using the `mtcars` dataset, this project answers the following questions:
* Is an automatic or manual transmission better for MPG?
* Quantify the MPG difference between automatic and manual transmissions.

Data Description

The mtcars 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). It is stored as a data frame with 32 observations on 11 (numeric) variables.
[, 1] mpg Miles/(US) gallon
[, 2] cyl Number of cylinders
[, 3] disp Displacement (cu.in.)
[, 4] hp Gross horsepower
[, 5] drat Rear axle ratio
[, 6] wt Weight (1000 lbs)
[, 7] qsec 1/4 mile time
[, 8] vs Engine (0 = V-shaped, 1 = straight)
[, 9] am Transmission (0 = automatic, 1 = manual)
[,10] gear Number of forward gears
[,11] carb Number of carburetors

Data Processing

```
# Load data and packages
library(ggplot2)
library(MASS)
data(mtcars)

# Convert categorical variables to factors
mtcars$cyl <- factor(mtcars$cyl)
mtcars$vs <- factor(mtcars$vs)
mtcars$am <- factor(mtcars$am, labels=c('Automatic','Manual'))
mtcars$gear <- factor(mtcars$gear)
mtcars$carb <- factor(mtcars$carb)

# Examine data
str(mtcars)
```

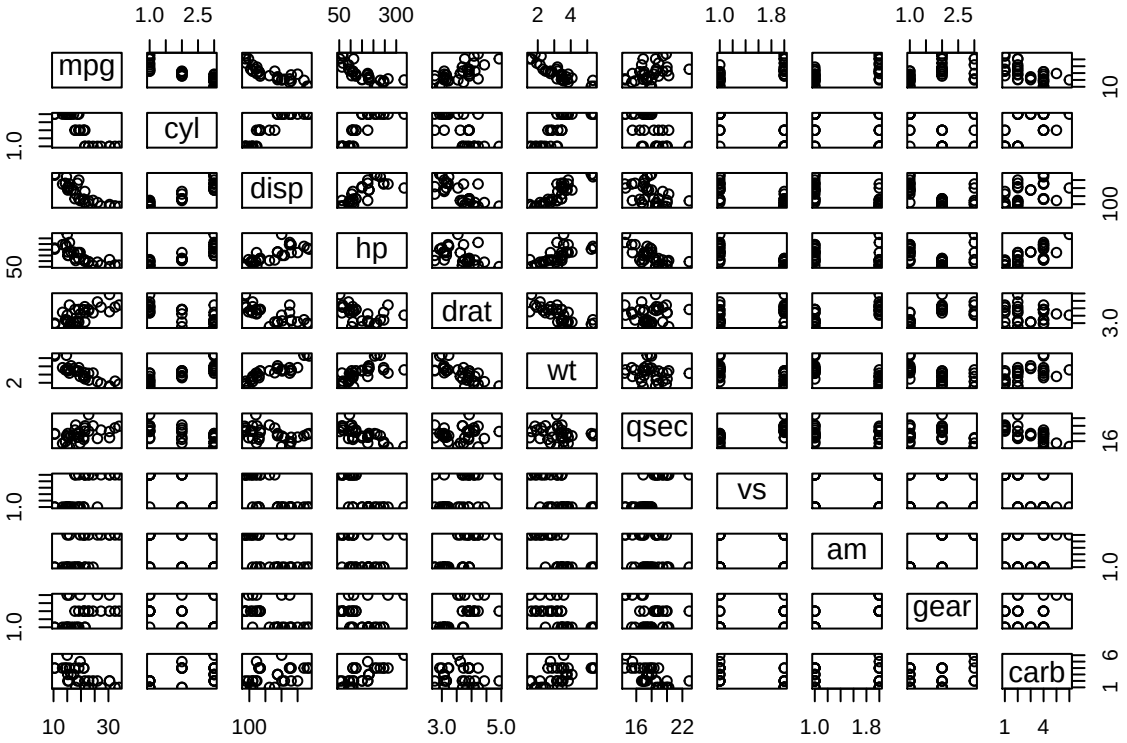
```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : Factor w/ 3 levels "4","6","8": 2 2 1 2 3 2 3 1 1 2 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
```

```
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : Factor w/ 2 levels "0","1": 1 1 2 2 1 2 1 2 2 2 ...
##  $ am  : Factor w/ 2 levels "Automatic","Manual": 2 2 2 1 1 1 1 1 1 1 ...
##  $ gear: Factor w/ 3 levels "3","4","5": 2 2 2 1 1 1 1 2 2 2 ...
##  $ carb: Factor w/ 6 levels "1","2","3","4",..: 4 4 1 1 2 1 4 2 2 4 ...
```

Exploratory Data Analysis

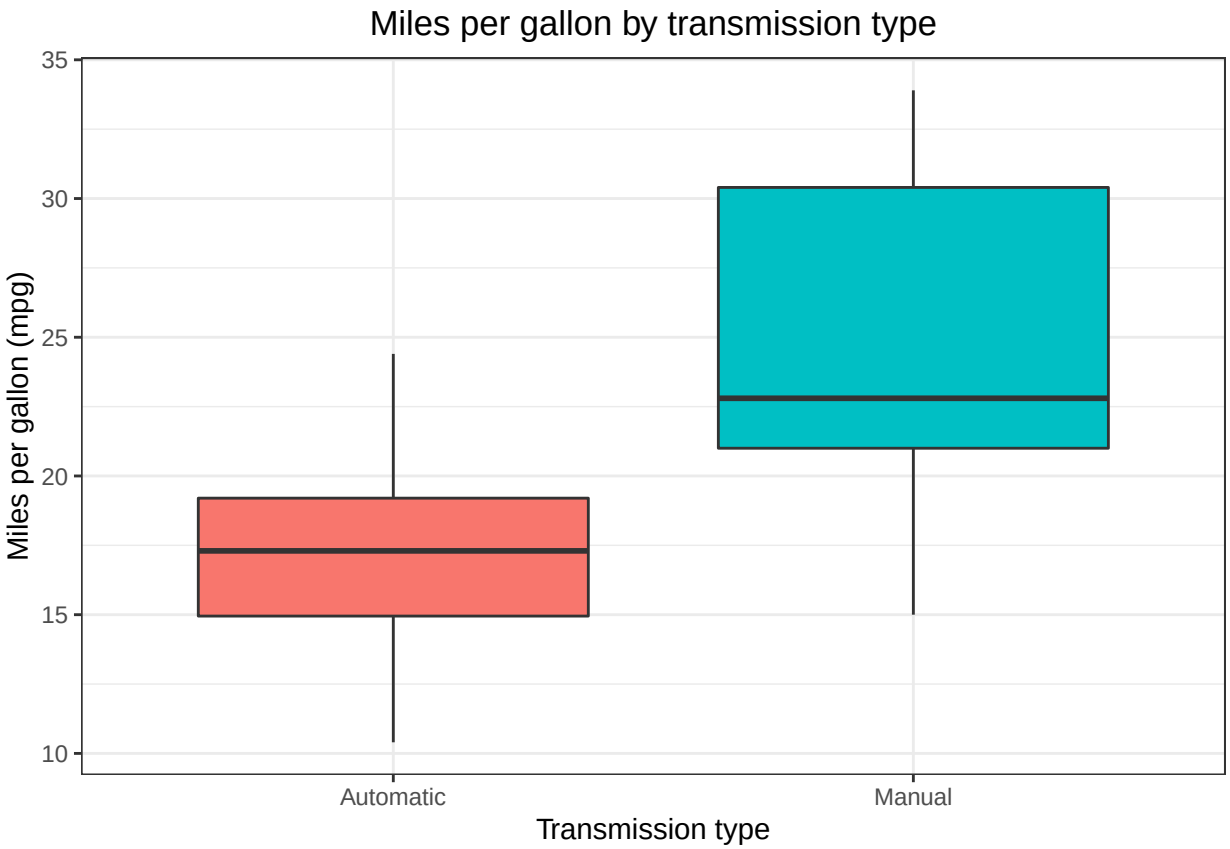
First, let's briefly examine any relationships between the variables

```
pairs(mpg ~ ., data = mtcars)
```



Let's also get a sense of the mean and spread of MPG by transmission type

```
ggplot(mtcars, aes(x = factor(am), y = mpg)) +
  geom_boxplot(aes(fill = factor(am)), show.legend = FALSE) +
  labs(x = "Transmission type",
       y = "Miles per gallon (mpg)",
       title = "Miles per gallon by transmission type"
       ) +
  scale_x_discrete(labels = c("Automatic","Manual")) +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5))
```



On first glance, it looks like manual transmission tends to have higher fuel efficiency (mpg). We will formally test this hypothesis in a regression analysis.

Regression Analysis

Model fitting and selection

We first fit a simple linear regression model to test the relationship between transmission type and mpg.

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

```
##
## Call:
## lm(formula = mpg ~ 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 ***
## amManual       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
```

The R-squared estimate shows that only about 36% of the variation in mpg is explained by the model, indicating that we need to control for other variables in the model.

Next, we will fit a multivariable regression model with all variables of the dataset included.

```
multifit <- lm(mpg ~ ., data = mtcars)
```

Then, we perform a stepwise variable selection using the `stepAIC()` function to determine which variables should be included in the final model. The smaller the AIC value, the better the model fit. See here for more information on the AIC test.

```
bestfit <- stepAIC(multifit, direction = "both")
```

```
summary(bestfit)
```

```
##
## Call:
## lm(formula = mpg ~ cyl + hp + wt + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9387 -1.2560 -0.4013  1.1253  5.0513
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.70832    2.60489  12.940 7.73e-13 ***
## cyl6        -3.03134    1.40728  -2.154  0.04068 *
## cyl8        -2.16368    2.28425  -0.947  0.35225
## hp          -0.03211    0.01369  -2.345  0.02693 *
## wt          -2.49683    0.88559  -2.819  0.00908 **
## amManual     1.80921    1.39630   1.296  0.20646
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.41 on 26 degrees of freedom
## Multiple R-squared:  0.8659, Adjusted R-squared:  0.8401
## F-statistic: 33.57 on 5 and 26 DF,  p-value: 1.506e-10
```

We see that the final model consists of `cyl`, `hp`, and `wt` as covariates, with `mpg` as outcome and `am` as predictor. In addition, about 87% of variance in mpg is now explained by the model.

Inference

1. The difference in the average mpg between auto and manual transmission is about 1.81, where manual has 1.81 greater mpg than auto transmission.

2. However, the p-value for manual transmission versus auto transmission (reference) is about 0.206, greater than the standard threshold of 0.05. Hence, we cannot conclusively say that manual transmission results in better mpg, compared to auto transmission.

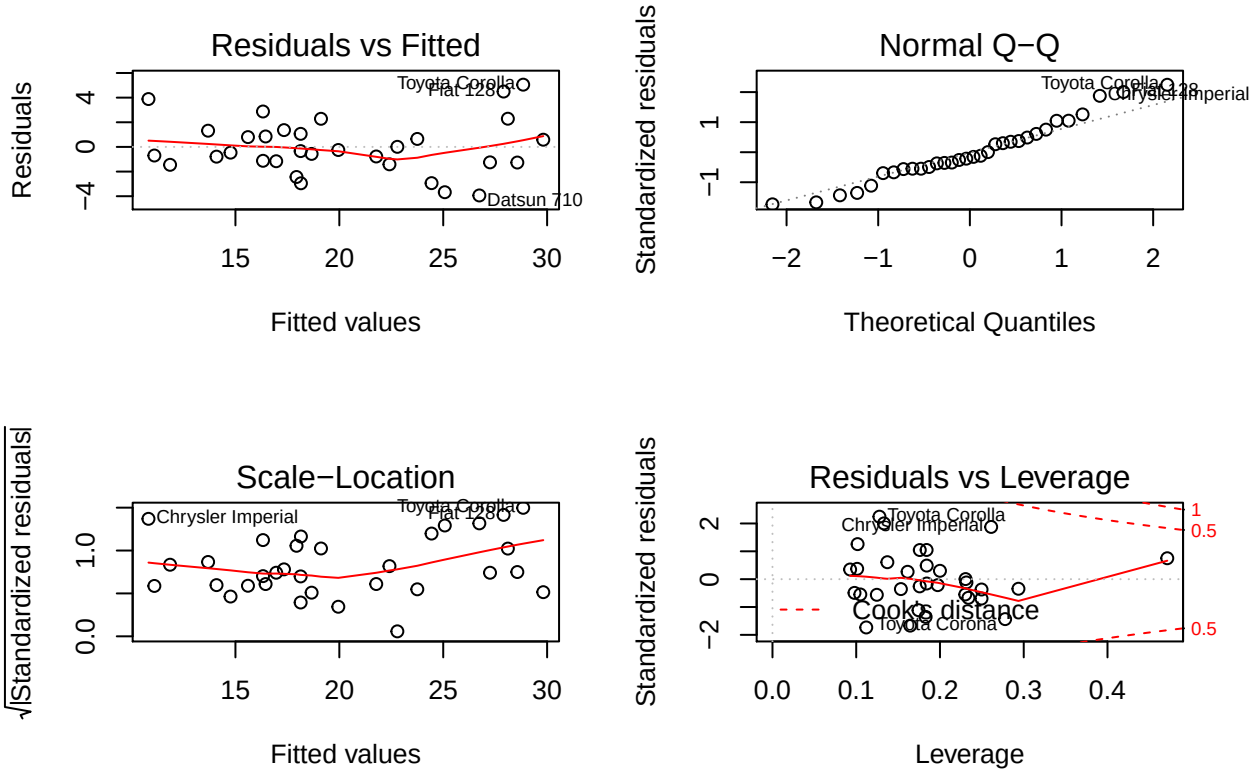
Conclusions

1. Manual transmission has 1.81 more mpg, on average, than auto transmission.

2. However, this difference is not statistically significant. We cannot conclude that either type of transmission has a better mpg than the other.

Appendix - Diagnostics

```
# Plot residuals
par(mfrow = c(2, 2))
plot(bestfit)
```




```
# Find 5 most influential points on slope coefficients
influence <- dfbetas(bestfit)
head(sort(influence[,6], decreasing = TRUE), 5)
```

```
##     Toyota Corona          Fiat 128 Chrysler Imperial    Toyota Corolla
##        0.73054020        0.42920432        0.35074579        0.28853987
##        Camaro Z28
##        0.08398495
```