Regression Models: Peer Assessment

Introduction

Regression Models: Peer Assessment
Author: Tim Westran
Purpose: Determine if manual or automatic transmissions are more efficient
Data set used: The mtcars data set.

Executive Summary

The basic goal of this assignment is to utilize the `mtcars` dataset, which contains fuel economy metrics for 32 vehicles manufactured between 1973 and 1974, to determine if an automatic or manual transmission is better for fuel economy.

Questions answered: 1. Is an automatic or manual transmission better for MPG? 2. Quantify the MPG difference between automatic and manual transmissions

People may assume that a vehicle with a manual transmission will have higher fuel economy than a vehicle with an automatic transmission. We set out to prove or disprove the hypothesis that manual drive vehicles are more efficient.

From the analysis we conduct, we will see that the transmission type does not have a significant impact on vehicle fuel ecomony.

Data Processing

Load libraries and set options

First, we want to load libraries and set options.

```r
## Ensure we have loaded the libraries we need.
library("datasets")
library("plyr")
library("ggplot2")

## load the data set
data(mtcars)

## disable scientific notation
options(scipen = 999)
```

Data cleaning

Now, we want to perform some data cleaning.

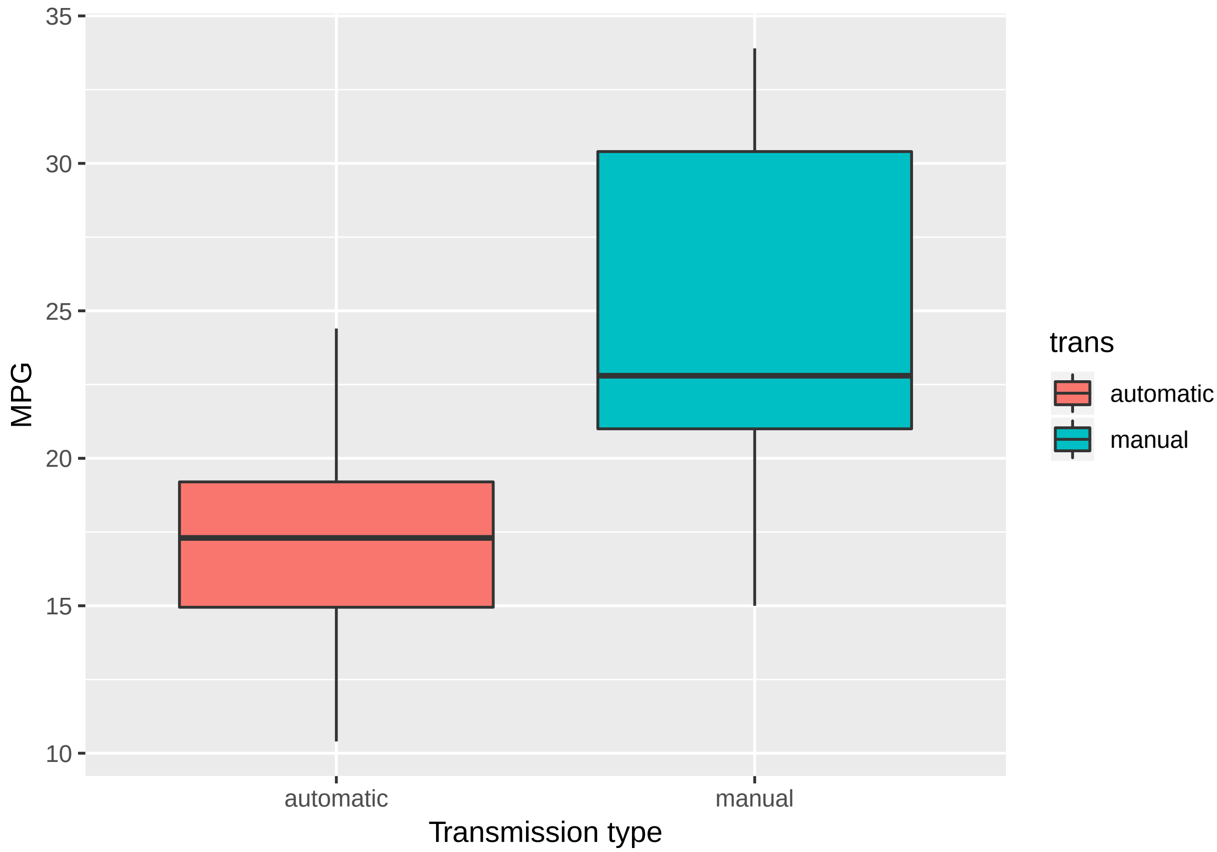
```r
## convert variables to factors.
mtcars$cyl <- factor(mtcars$cyl)
mtcars$vs <- factor(mtcars$vs)
mtcars$am <- factor(mtcars$am)
mtcars$gear <- factor(mtcars$gear)
mtcars$carb <- factor(mtcars$carb)

## Convert the binary variable, transmission type.
trans <- revalue(mtcars$am, c('0'="automatic", '1'="manual"))
```

```r
levels(mtcars$am) <- c("Automatic", "Manual")
levels(mtcars$vs) <- c("V Engine", "Straight Engine")
```

Data analysis

Now, some rudimentary data analysis.

```r
## Let's build a boxplot showing transmission type versus fuel economy
ggplot(mtcars, aes(x=trans, y=mpg, fill=trans)) +
    geom_boxplot() +
    xlab("Transmission type") +
    ylab("MPG")
```



From the box plot, we see a large difference in fuel economy betweeen vehicles with automatic transmissions and vehicles with manual transmissions. However, this correlation could be due to other factors.

Now, let's see how much of the variability can be explained away by transmission type.

```r
## Create a linear model, comparing mpg to transmission type.
fit1 <- lm(mpg ~ am, data=mtcars)
```

From the summary of our first linear model (in appendix), we see that the coefficients for both the intercept and the transmission type are significant. However, we also see that accounting for only transmission type only explains away 35.9798943% of the variation in fuel economy.

Based on the analysis documented in the appendix, several variables seem to have high correlation with MPG. We will create a new model using all variables, then select a subset of these variables which yield the highest correlation.

```r
## First, let's create a new model with all variables.
fit2 <- lm(mpg ~ ., data=mtcars)

## Now, let's trim the model to capture the most correlated variables.
fit3 <- step(fit2, direction="both", trace=0)
```

The optimal model contains four variables: Transmission type (am), vehicle weight (weight), horsepower (hp), and the number of cylinders (cyl). These 4 variables explain 86.5879872% of variation in fuel economy. It logically makes sense that vehicle weight and engine size would contribute significantly to the overall fuel economy. indeed, we see that the type of transmission has a minimal impact on fuel economy.

Final results

We reject the null hypothesis, that manual transmission vehicles are more efficient than automatic transmission vehicles. From the analysis performed, we see that vehicle weight has a strong correlation with mileage. A 1000 lb increase in vehicle weight will result in a decrease of 2.4968294 MPG. Similarly, an increase of one horsepower will result in a decrease of 0.0321094 MPG. Finally, we see that engine size has a strong correlation with mileage, as moving from a 4 to 6 cylinder engine decreases fuel economy 3.0313445 MPG, and moving from a 6 to an 8 cylinder engine decreases fuel economy by 2.1636753 MPG.

Appendix

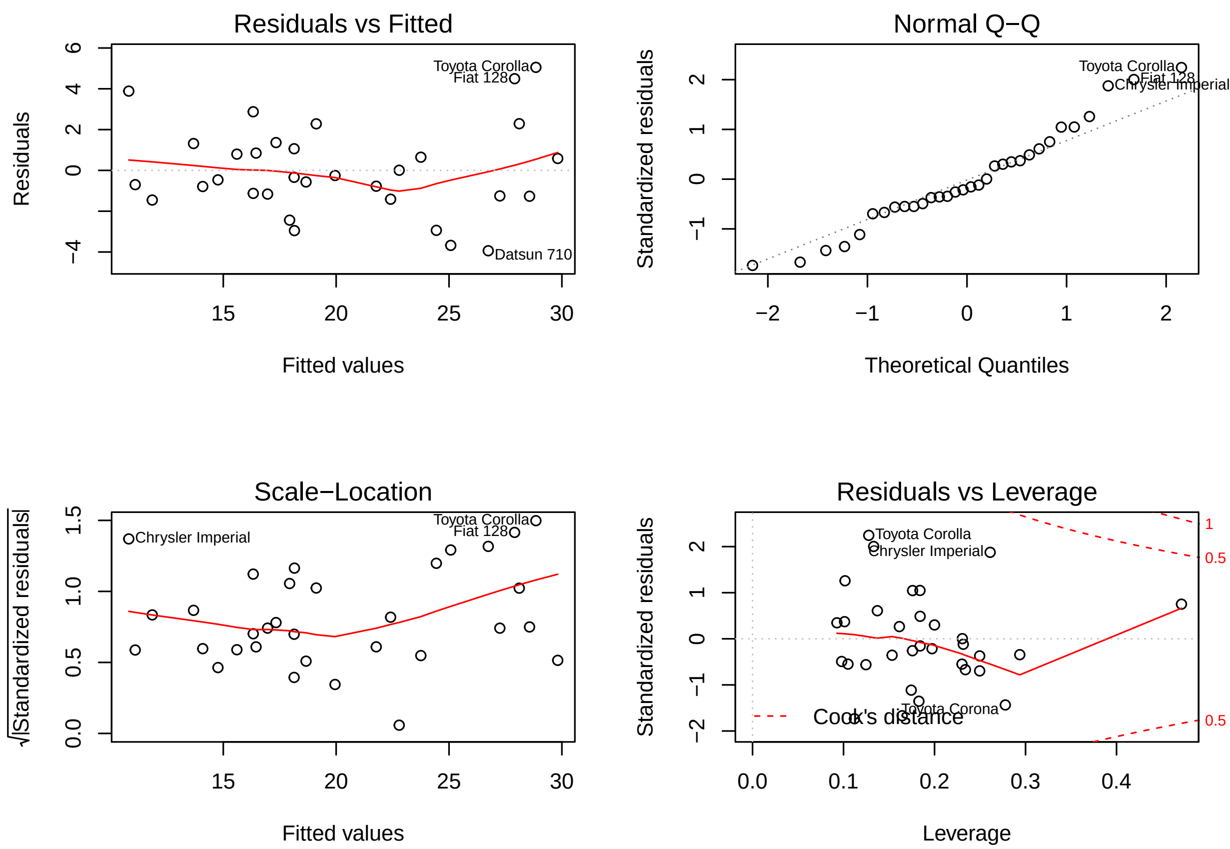
Shows a summary of the most correlated variables.

```r
summary(fit3)
```

```
##
## 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 0.000000000000773 ***
## 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: 0.0000000001506
```

Looking at the most correlated variables

```r
par(mfrow = c(2,2))
plot(fit3)
```




```r
## Show summary statistics on this first linear model.
summary(fit1)
```

```
##
## 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 0.00000000000000113 ***
## 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
```