Project for Regression Models at Coursera

Xiang Jiang

August 23, 2014

Executive summary

This project aims at answering the two questions:

- Is an automatic or manual transmission better for MPG?
- Quantify the MPG difference between automatic and manual transmissions.

We conclude that:

- Manual transmission is better than automatic for MPG.
- Compared with automatic transmissions, manual transmittions have an additional 2.94mpg.

Exploratory Analysis

We begin by inspecting the distribution of mpg for different types of transmission, please find figure 1 in appendix.

We also inspect the difference of mean and standard deviation for mpg with different transmission types. As shown below, we can see the mean and standard deviation for mpg with different transmission types are significantly different.

```r
data(mtcars)
summary(mtcars$mpg[mtcars$am==0])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    10.4    15.0    17.3    17.1    19.2    24.4
```

```r
sd(mtcars$mpg[mtcars$am==0])
```

```
## [1] 3.834
```

```r
summary(mtcars$mpg[mtcars$am==1])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    15.0    21.0    22.8    24.4    30.4    33.9
```

```r
sd(mtcars$mpg[mtcars$am==1])
```

```
## [1] 6.167
```

Model Selection & Fitting

Here we choose models by AIC in a stepwise algorithm.

For transmission type 0 (atomatic), we get the model.

```r
lm1 <- step(lm(mpg ~ ., data=mtcars))
```

Residual standard error for this model is:

```r
summary(lm1)$sigma
```

```
## [1] 2.459
```

The coefficients suggest that, compared with automatic transmissions, manual transmittions have an additional 2.94mpg.

To identify whether *wt* and *qsec* have significant influences to the model. We build models with and without *wt* and *qsec* respectively. We, then, use residual standard error to measure the performance of different models

The second model is:

```r
lm2 <- lm(mpg ~ am + wt, data=mtcars)
```

Residual standard error for this model is:

```r
summary(lm2)$sigma
```

```
## [1] 3.098
```

The third model is:

```r
lm3 <- lm(mpg ~ am + qsec, data=mtcars)
```

Residual standard error for this model is:

```r
summary(lm3)$sigma
```

```
## [1] 3.487
```

The fourth model is:

```r
lm4 <- lm(mpg ~ am, data=mtcars)
```

Residual standard error for this model is:

```r
summary(lm4)$sigma
```

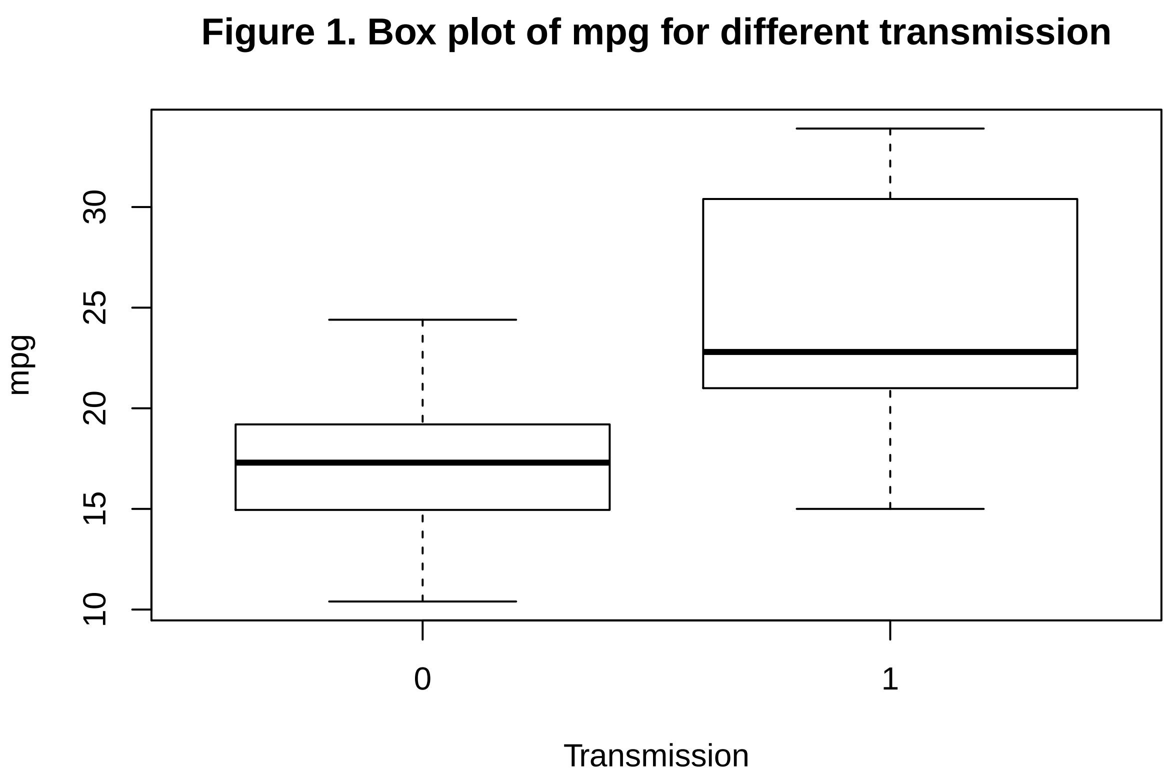
```
## [1] 4.902
```

As we can see, the first model (lm(formula = mpg ~ wt + qsec + am, data = mtcars)) has the smallest residual standard error.

We visualize residuals in figure 2 of appendix, from these figures, we can see, there is no significant relationship between residuals and fitted values.

Appendix

```
boxplot(mtcars$mpg ~ mtcars$am,main='Figure 1. Box plot of mpg for different transmission', xlab='Transm
```




```
par(mfrow = c(2, 2))
plot(lm1)
```

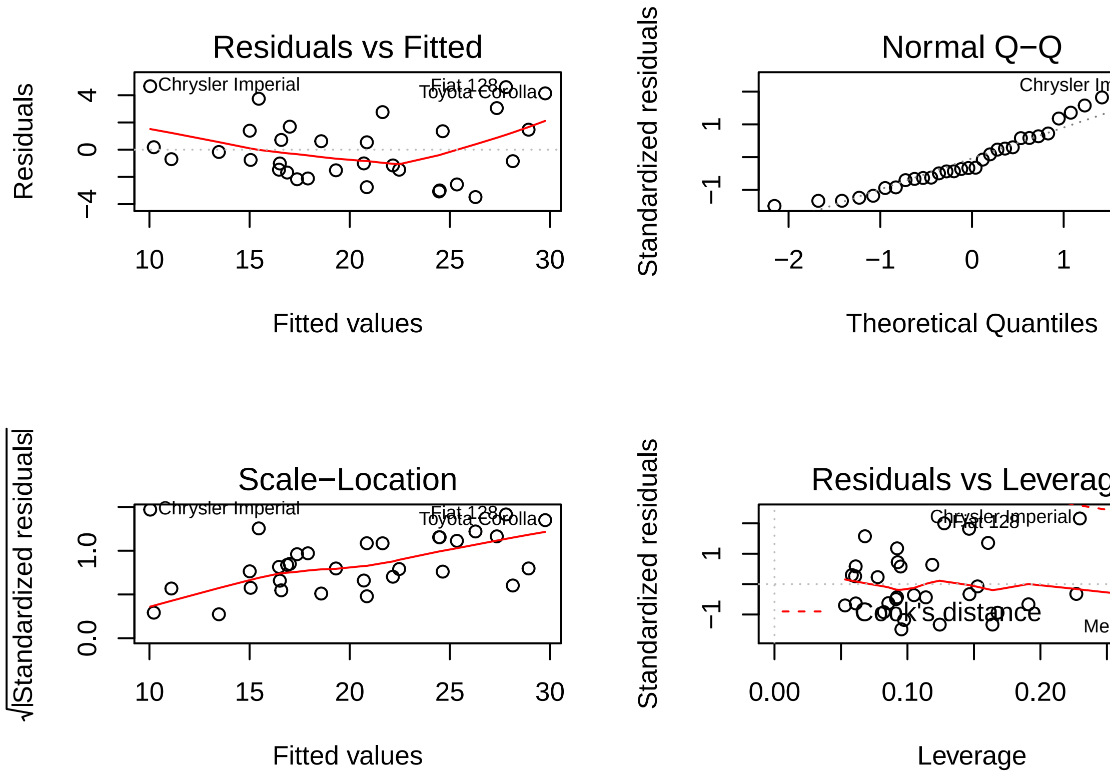


Figure 2. Residual Plot