Coursera: Regression Models - Course Project

Monday, November 17, 2014

Executive Summary

In this report, we explore the relationship between a set of variables and miles per gallon (MPG) using a data set `mtcars`. Two questions were analyzed: 1. Is an automatic or manual transmission better for MPG? 2. Quantify the MPG difference between automatic and manual transmissions? A regression model was created using multivariate linear regression, correlation and nested likelihood methods. Using the model, we conclude that manual transmission is better for mileage `mpg` by a factor of 1.8 compared to automatic transmission.

Data Processing

A number of variables (cyl, vs, gear, carb, am) were transformed into factor variables as shown below.

```
data(mtcars)
mtcars$cyl <- factor(mtcars$cyl); mtcars$vs <- factor(mtcars$vs)
mtcars$gear <- factor(mtcars$gear); mtcars$carb <- factor(mtcars$carb)
mtcars$am <- factor(mtcars$am,labels=c('Automatic','Manual'))
```

Exploratory Data Analysis

Using the `pairs` plot (appendix figure 1), we see that most of the variables are influencing `mpg` in one way or another. Since we are interested in the relationship between `mpg` and transmission `am`, we use a boxplot (appendix figure 2). The plot shows that `mpg` is higher when the transmission `am` is manual.

Regression Model

Approach

We build a regression model to predict `mpg`. We found from figure 1 that all the variables in the dataset influence the miles per gallon outcome `mpg`. However, this does not mean that we should include all the variables in the model . First, we check the correlation between `mpg` and the rest of the predictors. Notice that `mpg` is strongly correlated to weight `wt`, number of cylinders `cyl` and displacement `disp`.

```
data(mtcars); sort(cor(mtcars)[1,])
```

```
##         wt        cyl       disp         hp       carb       qsec
## -0.8676594 -0.8521620 -0.8475514 -0.7761684 -0.5509251  0.4186840
##       gear         am         vs       drat        mpg
##  0.4802848  0.5998324  0.6640389  0.6811719  1.0000000
```

However, we are still not sure that we should include all these predictors in the model. So we check the overall dataset's correlation matrix. Here we notice that some variables are very strongly related with each other. For example, the correlation between number of cylinders `cyl` and displacement `disp` is very high. So we should consider using only one of these variables in our model.

```r
data(mtcars); cor(mtcars)
```

Build the Best Model

For our first model, we include all the predictors. Then we perfom stepwise model selection to get the best model. The best model consists of the variables `cyl`, `wt` and `hp` as confounding variables and `am` as the independent variable. It explains 84% of the variance.

```r
data(mtcars); mtcars$cyl <- factor(mtcars$cyl); mtcars$vs <- factor(mtcars$vs)
mtcars$gear <- factor(mtcars$gear); mtcars$carb <- factor(mtcars$carb);
mtcars$am <- factor(mtcars$am,labels=c('Automatic','Manual'))
#fit the first model with all the predictors
firstModelFit <- lm(mpg ~ ., data = mtcars)
#fit the best model
bestModel <- step(firstModelFit, direction = "both")
summary(bestModel)
```

```r
summary(bestModel)$coefficients
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 33.70832390 2.60488618 12.940421 7.733392e-13
## cyl6        -3.03134449 1.40728351 -2.154040 4.068272e-02
## cyl8        -2.16367532 2.28425172 -0.947214 3.522509e-01
## hp          -0.03210943 0.01369257 -2.345025 2.693461e-02
## wt          -2.49682942 0.88558779 -2.819404 9.081408e-03
## amManual     1.80921138 1.39630450  1.295714 2.064597e-01
```

Compare the Best Model to the Base Model

Our base model consists of just the transmission variable `am`. We compare this base model to the best model to make sure that the predictors included in the best model are truly significant. The model comparison results show that the predictors in the best model are significant.

```r
baseModel <- lm(mpg ~ am, data = mtcars)
anova(baseModel, bestModel)
```

```r
anova(baseModel, bestModel)$"Pr(>F)"
```

```
## [1]           NA 1.688435e-08
```

Diagnostics (Figure 4)

1. The Residuals Vs Fitted values plot shows that the residuals do not follow a pattern. So most of the variance has been explained by the model.
2. The Normal QQ plot shows that the residuals satisfy the assumption of normality.
3. The Scale-Location plot shows the standardized residuals (rescaled with a mean of zero and a variance of one) plotted against the fitted values. The trend line is relatively flat, indicating a constant variance and evidence against homoskedasticity.
4. The last plot shows the standardized residuals against leverage. Here, the trend line stays close to the horizontal line. No points have a large Cook's distance (> 0.5). So we conclude that no observations have undue leverage.

t-test

The t-test was performed to understand whether the means of the manual and automatic transmissions are significatively different. Based on the p-value of the test, we conclude that the means are in fact, different.

```
t.test(mpg ~ am, data = mtcars)$p.value
```

```
## [1] 0.001373638
```

Conclusion

1. Cars with manual transmission get more mileage `mpg` by a factor of 1.8 (adjusted by `hp`, `cyl`, and `wt`) when compared to cars with automatic transmission.
2. Mileage `mpg` will decrease by 2.5 (adjusted by `hp`, `cyl`, and `am`) for every 1000 lb increase in weight `wt`.
3. If number of cylinders, `cyl` increases from 4 to 6 to 8, `mpg` will decrease by a factor of 3 and 2.1 respectively (adjusted by `hp`, `wt`, and `am`).

Appendix

Figure 1: Pairs Plot

```
par(mfrow = c(1, 1)); par (mar=c(0,0,0,0))
pairs (mpg ~ ., mtcars)
```

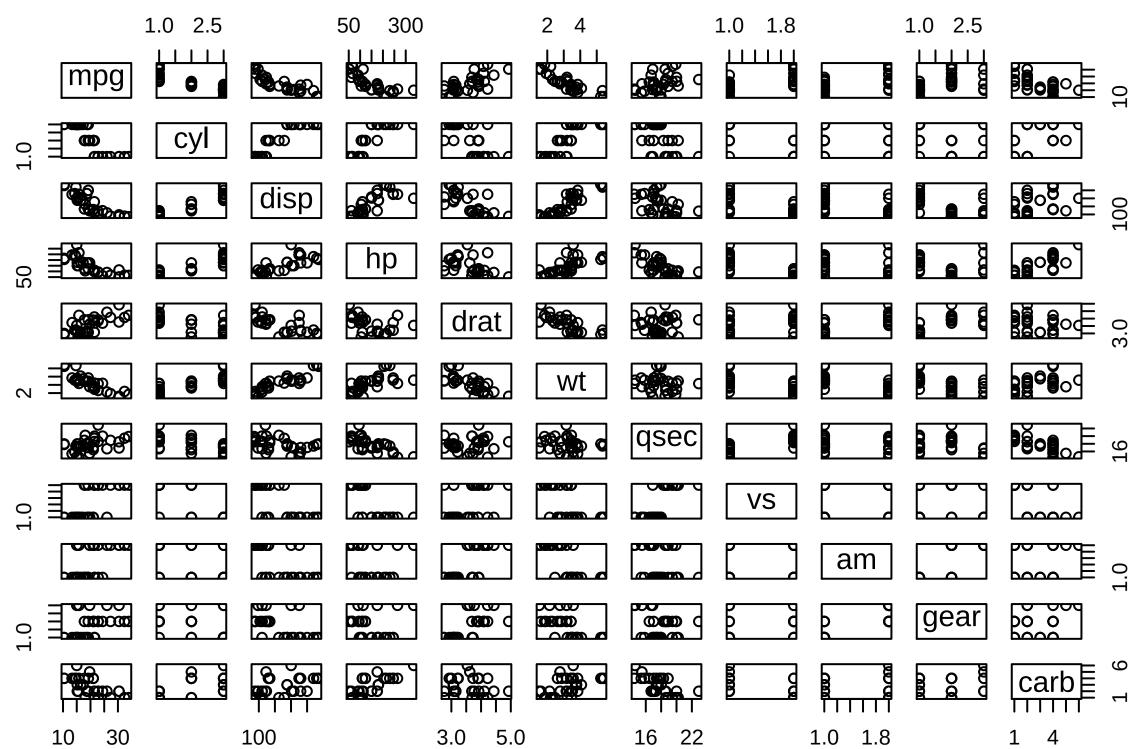


Figure 2: Histogram

```
par(mfrow = c(1, 3))
# Histogram with Normal Curve
hist(mtcars$mpg, breaks=10, col="blue", xlab="Miles/Gallon", main="Histogram of Miles per Gallon")
# Density Plot
#density(mtcars$mpg)
plot(d, xlab = "MPG", main ="MPG Density Plot")
boxplot(mpg ~ am, data = mtcars, xlab = "Transmission", ylab = "Miles per Gallon", main = "MPG by Transm
```

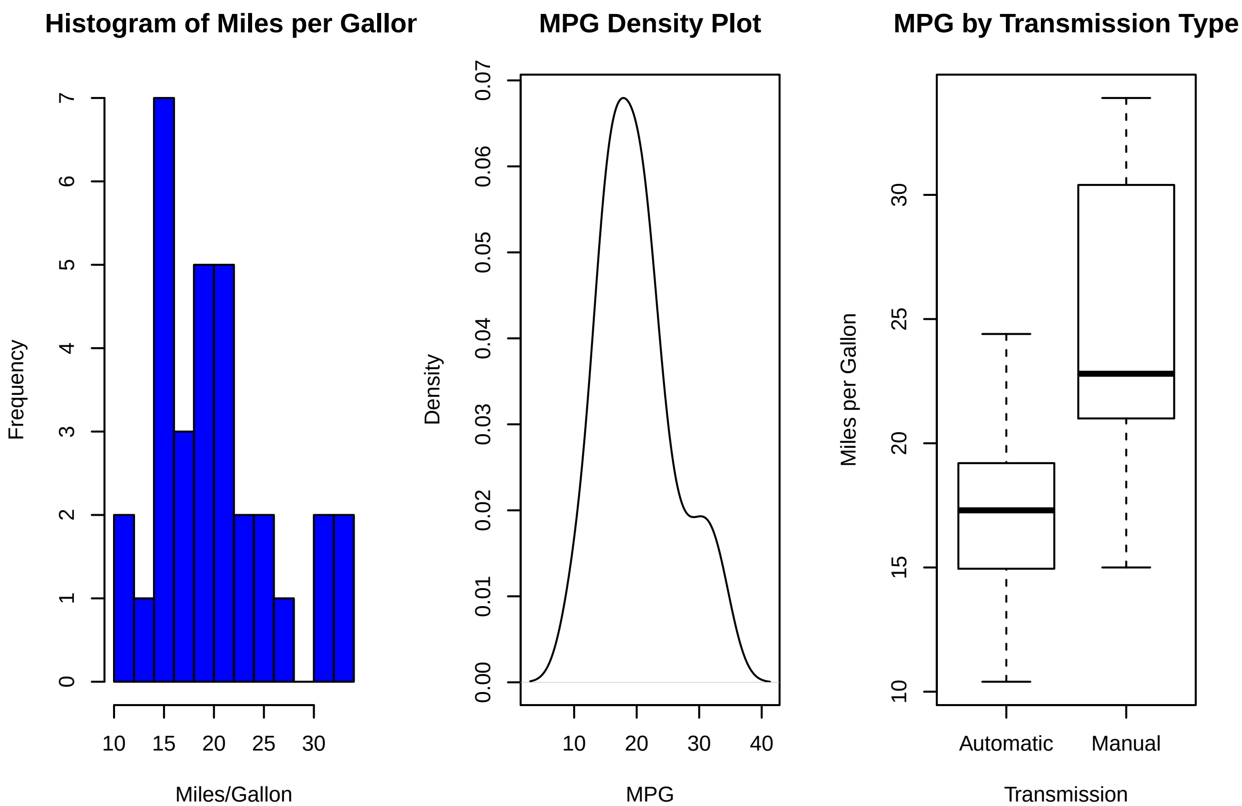
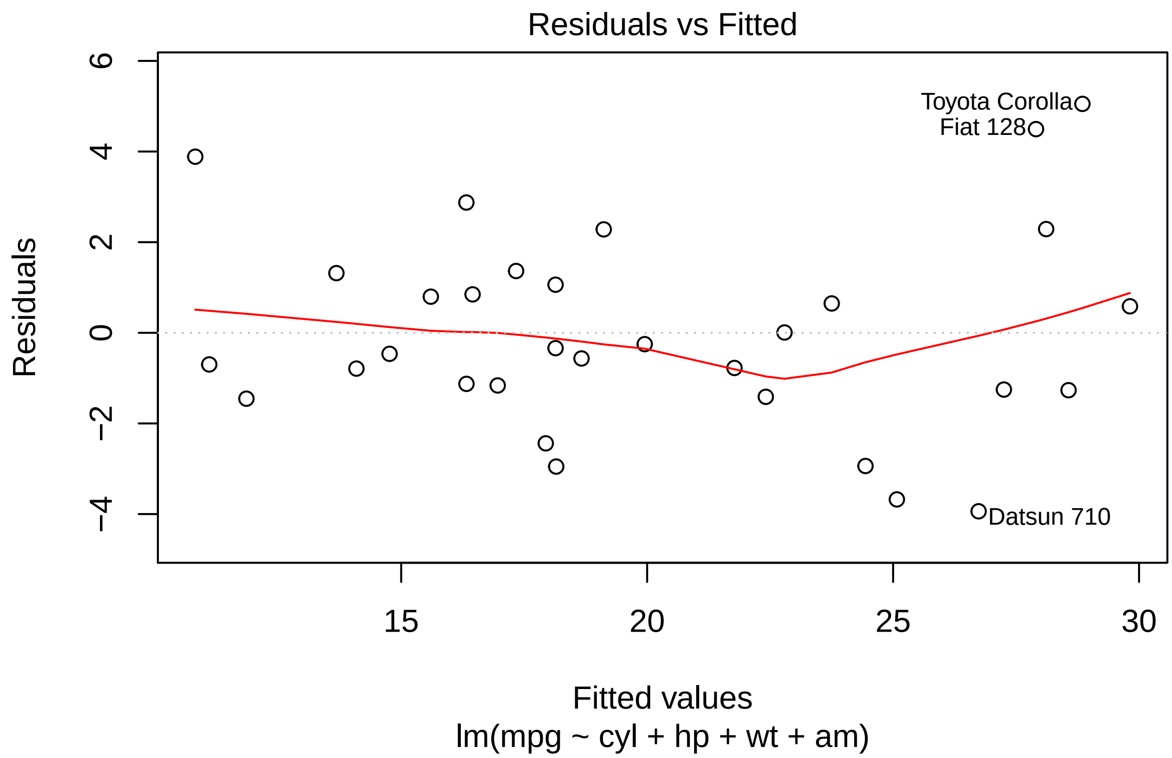
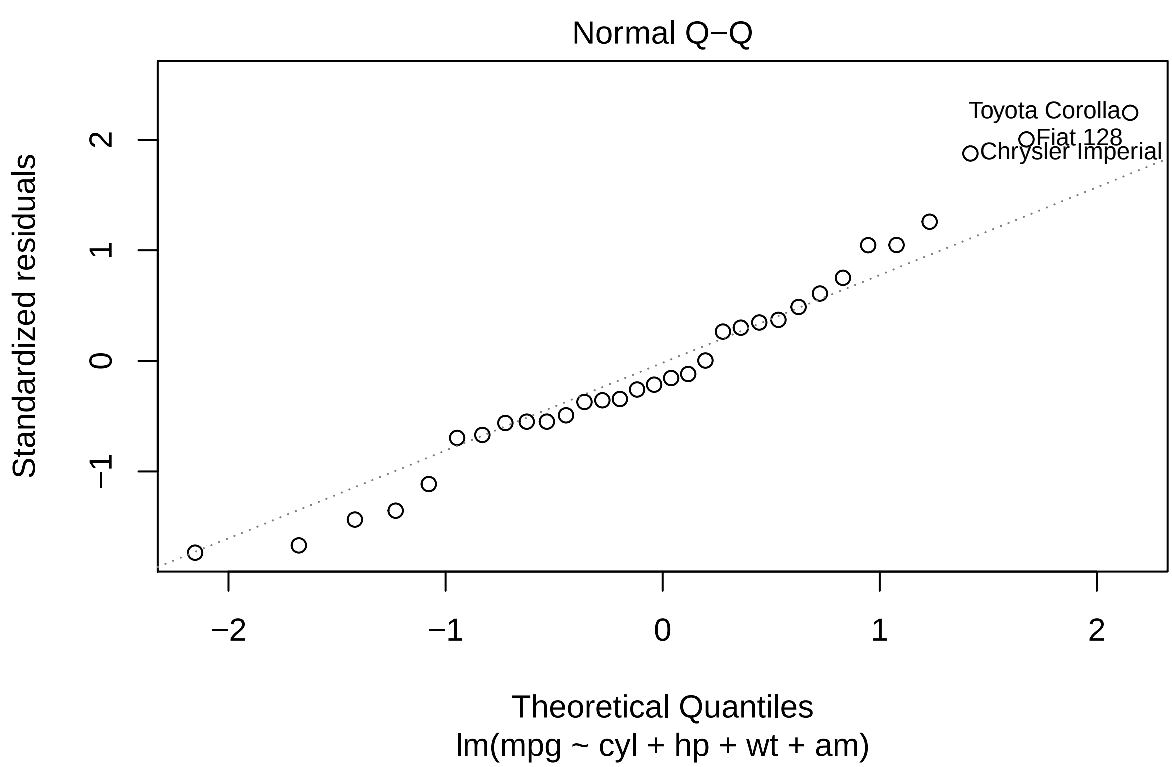
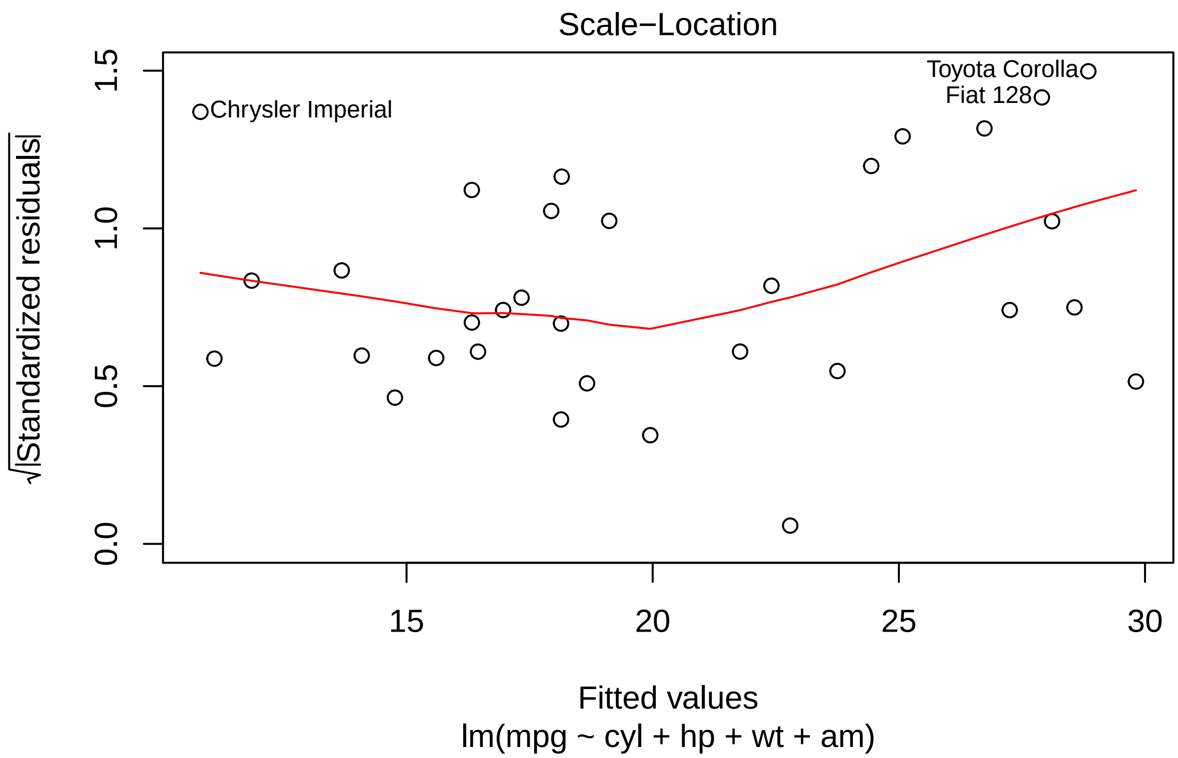


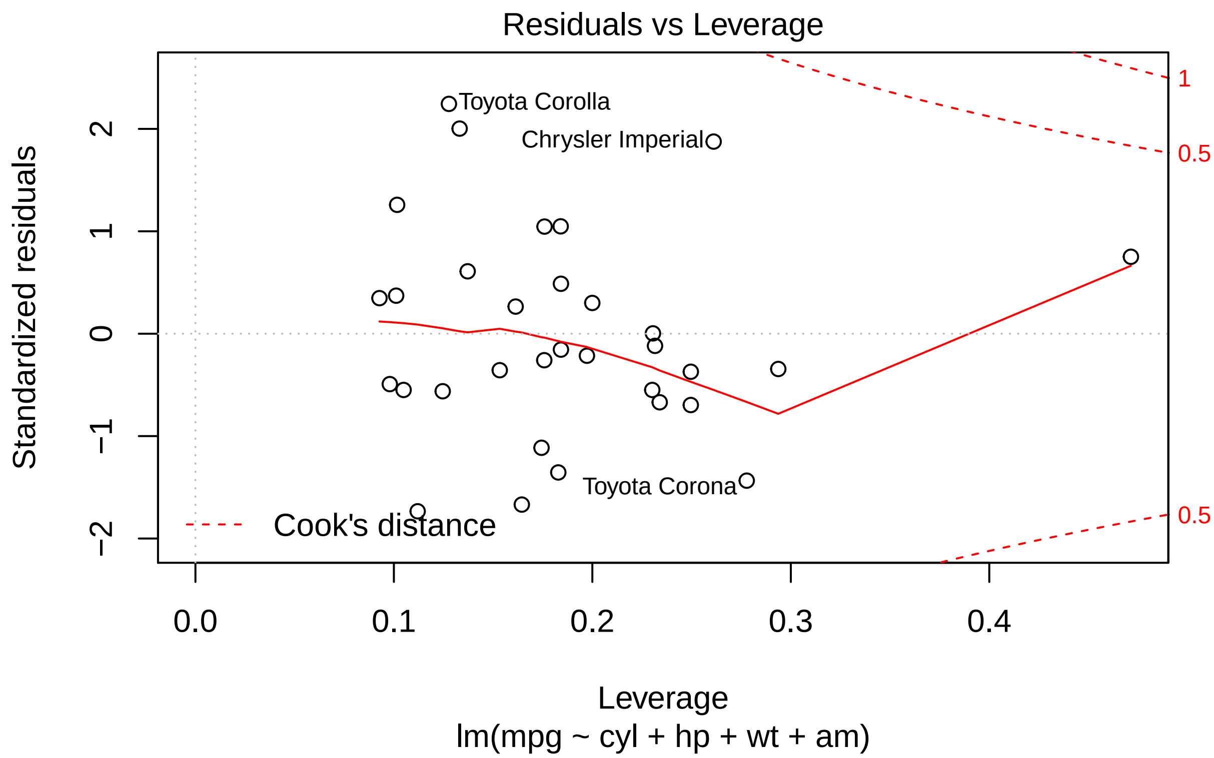
Figure 4: Diagnostics

```
plot.new(); plot(bestModel)
```


Residuals vs Fitted
Toyota Corolla
Fiat 128
Datsun 710
Residuals
Fitted values
lm(mpg ~ cyl + hp + wt + am)
15
20
25
30
6
4
2
0
−2
−4


Normal Q−Q
Toyota Corolla
Fiat 128
Chrysler Imperial
Standardized residuals
Theoretical Quantiles
lm(mpg ~ cyl + hp + wt + am)
−2
−1
0
1
2


Scale−Location
Toyota Corolla
Fiat 128
Chrysler Imperial
√|Standardized residuals|
0.0
0.5
1.0
1.5
15
20
25
30
Fitted values
lm(mpg ~ cyl + hp + wt + am)


Residuals vs Leverage
Standardized residuals
Leverage
lm(mpg ~ cyl + hp + wt + am)
Toyota Corolla
Chrysler Imperial
Toyota Corona
Cook's distance
1
0.5
0.5
2
1
0
−1
−2
0.0
0.1
0.2
0.3
0.4