Regression Models Course Project

Thuyen Ho

February 27, 2016

Executive summary

In this project, we will analyze the `mtcars` data set and explore the relationship between a set of variables and miles per gallon (MPG) which will be our outcome.

Main objects:

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

Data processing

First, we load the dataset `mtcars` and explore summary about dataset.

```r
data(mtcars)
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 : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ 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  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

```r
head(mtcars)
```

```
##                    mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

As you see above, the dataset evaluation shows that it has 11 variables and 32 samples . the variables `vs`, `am`, `gear` and `card` are numeric variables. Those ones must be factor variables, so that we perform the necessary data transformations by factoring the necessary variables and look at the data.

```r
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

Initially, we plot the relationships between all the variables of the dataset (see Figure 1 in the appendix). From the plot, we notice that variables like `cyl`, `disp`, `hp`, `drat`, `wt`, `vs` and `am` seem to have some strong correlation with `mpg`. But we will use linear models to quantify that in the regression analysis section.

In this analysis, we are interested in the effects of car transmission type on mpg (see Figure 2 in the appendix). So, we look at the distribution of mpg for each level of am (Automatic or Manual) by plotting box plot. This plot clearly depicts that manual transmissions tend to have higher MPG. This data is further analyzed and discussed in regression analysis section by fitting a linear model.

Regression analysis

Our initial model includes all variables as predictors of mpg. Then we perform step-wise regression/model selection algorithm on the following initial model.

```r
initModel <- lm(mpg ~ ., data = mtcars)
bestModel <- step(initModel, direction = "both")
```

The best model obtained from the above computations shows that variables, `cyl`, `wt`, `hp` and `am`.

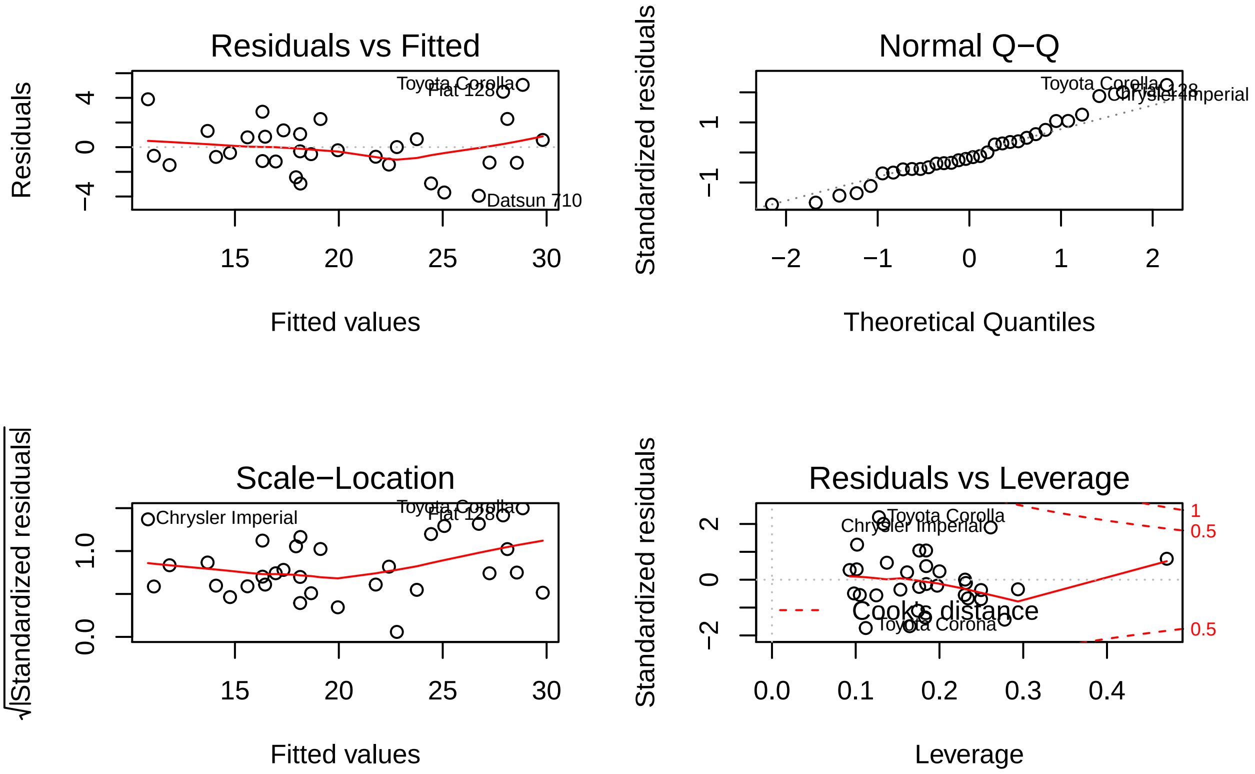
```r
summary(bestModel)
```

```
##
## 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
```

The adjusted R-squared value of 0.84, so that means 0.84% percent of total variability that is explained by the regression model above.

Residuals and diagnostic

```r
par(mfrow = c(2, 2)); plot(bestModel)
```



There is a bit of a curve to the residual plot, so that it departs slightly from normality. The residuals for the Chrysler Imperial, Fiat 128, and Toyota Corolla are called out because they exert some influence on the shape of the curve.

```r
leverage <- hatvalues(bestModel)
tail(sort(leverage),3)
```

```
##       Toyota Corona Lincoln Continental       Maserati Bora
##           0.2777872           0.2936819           0.4713671
```

```r
influence <- dfbetas(bestModel)
tail(sort(influence), 3)
```

```
## [1] 0.5436814 0.7305402 0.9389082
```

Conclusions

Cars with manual transmission get 1.8 more miles per gallon compared to cars with Automatic transmission. (1.8 adjusted for hp, cyl, and wt).

Appendix

Figure 1. Relationships between variables.

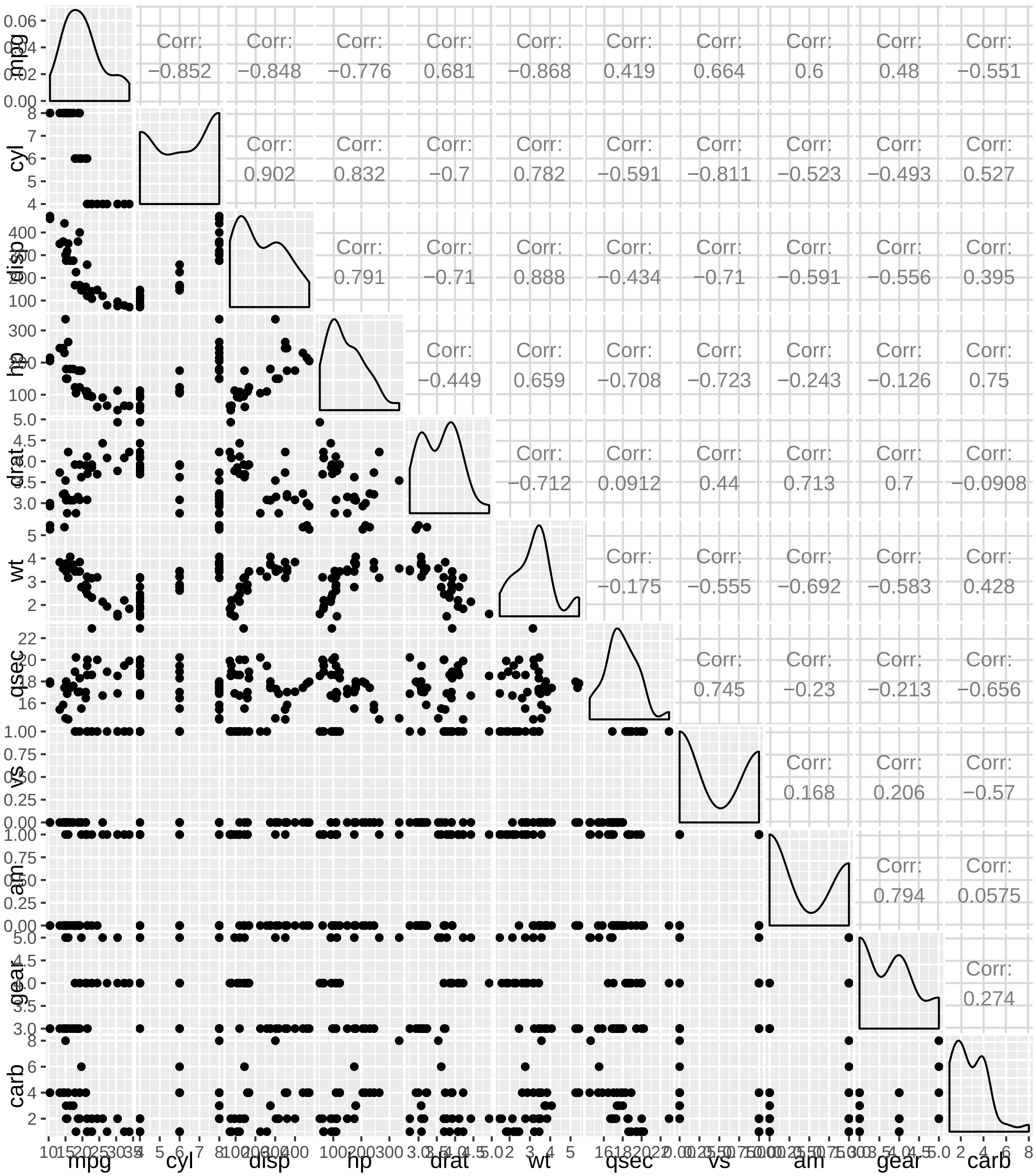
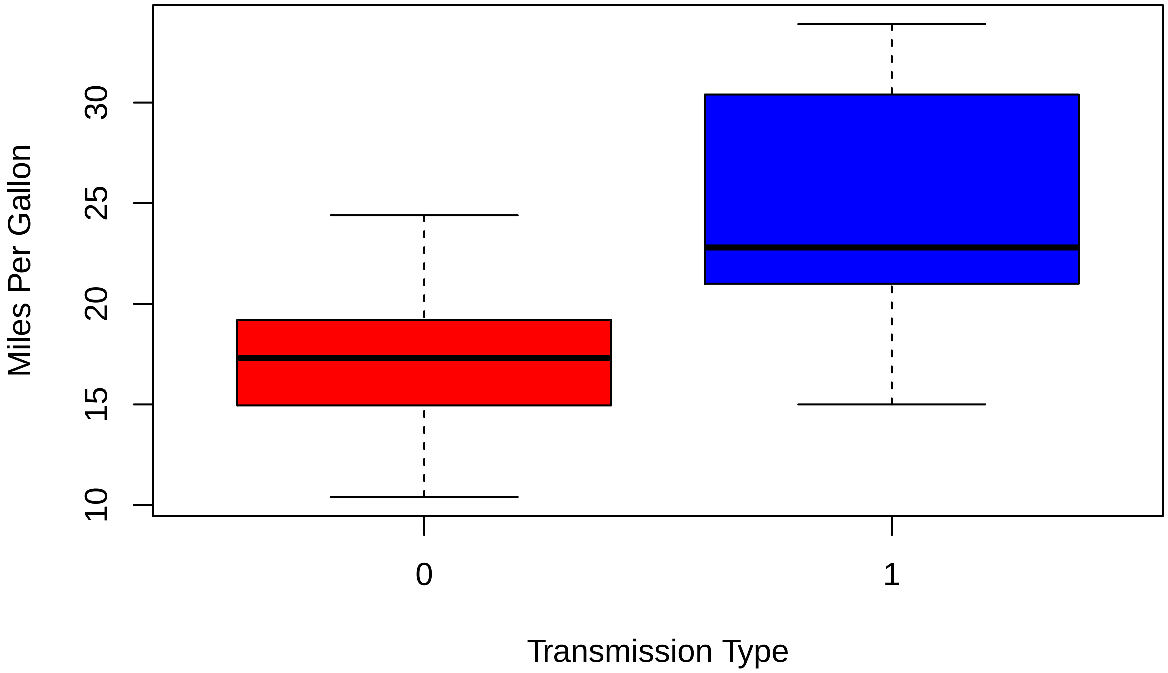


Figure 2. MPG per transmission type.


Miles Per Gallon
Transmission Type