Regression Models Course Project

Composed by: Akihiro Hayashi (21/03/2015)

Executive Summary

To answer the following questions:

1. "Is an automatic or manual transmission better for MPG"
2. "Quantify the MPG difference between automatic and manual transmissions"

We use "mtcars" datasets to do EDA, correlation analysis, multivariate regression and residual analysis for answering the questions.

Results: From our model, mpg of a manual car is "7.25" higher than an automatic car. But it's NOT significant because it's confounded by weight and horsepower.

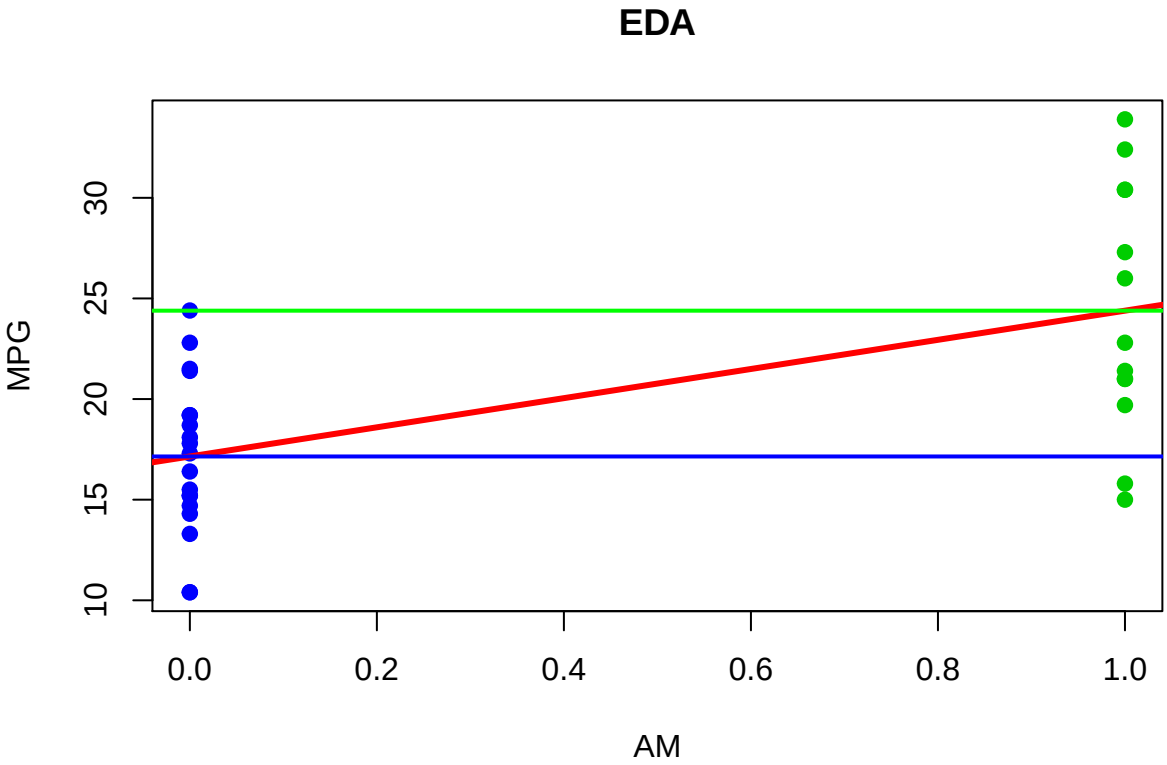
1. Exploratory Data Analysis

First, we take a look at this dataset.

```
data(mtcars)
?mtcars
head(mtcars)
```

According to the questions, we have interest in the relation between mpg and the variable called "am". Then, we make a plot to take a quick look.

```
plot(mtcars$am, mtcars$mpg, col = 3 + (mtcars$am == 0), pch = 19, main = "EDA", xlab = "AM", ylab = "MP
abline(lm(mtcars$mpg ~ mtcars$am), lwd = 3, col = "red")
abline(h = mean(mtcars$mpg[mtcars$am == 0]), lwd = 2, col = "blue")
abline(h = mean(mtcars$mpg[mtcars$am == 1]), lwd = 2, col = "green")
```



It seems when we use cars with manual transmissions (green) have higher mpg than automatic transmissions (blue). So we do an unadjusted estimate.

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

```
##
## 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 ***
## am             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
```

We rejected the hypothesis that "type of transmissions doesn't affect mpg" at a 95% significant level (p-value is lower than 0.001). According to the coefficient of "am", we can say that the mpg of a manual car is significantly 7.25 higher than an automatic car.

BUT, we still have to examine whether there is any confounder which may mislead our judgement. So we calculate the correlations between mpg and other variables and pick some variables with relatively higher correlation value.

2. Correlation Analysis

```
sort.cor <- sort(abs(cor(mtcars)[1, ]), decreasing = TRUE)
sort.cor
```

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

According to correlation analysis, we can include (wt, cyl, disp, hp, am) in our model. And then, we check correlation again to see whether thers is any collinearity between them.

```
cor(mtcars)[c(6, 2, 3, 4, 9), c(6, 2, 3, 4, 9)]
```

```
##              wt        cyl       disp         hp         am
## wt    1.0000000  0.7824958  0.8879799  0.6587479 -0.6924953
## cyl   0.7824958  1.0000000  0.9020329  0.8324475 -0.5226070
## disp  0.8879799  0.9020329  1.0000000  0.7909486 -0.5912270
## hp    0.6587479  0.8324475  0.7909486  1.0000000 -0.2432043
## am   -0.6924953 -0.5226070 -0.5912270 -0.2432043  1.0000000
```

We find out wt has high collinearity with disp and cyl, so we remove them and only include (wt, hp) as confounders in our model. Now we use anova to examinate our decision.

3. Multivariate Regression

```
fit1 <- lm(mpg ~ am, data = mtcars)
fit2 <- lm(mpg ~ am + wt, data = mtcars)
fit3 <- lm(mpg ~ am + wt + hp, data = mtcars)
anova(fit1, fit2, fit3)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + wt
## Model 3: mpg ~ am + wt + hp
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     29 278.32  1    442.58 68.734 5.071e-09 ***
## 3     28 180.29  1     98.03 15.224 0.0005464 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Obviously, our model can SIGNIFICANTLY explain the variation of mpg. Then we summarize our model.

```
summary(fit3)
```

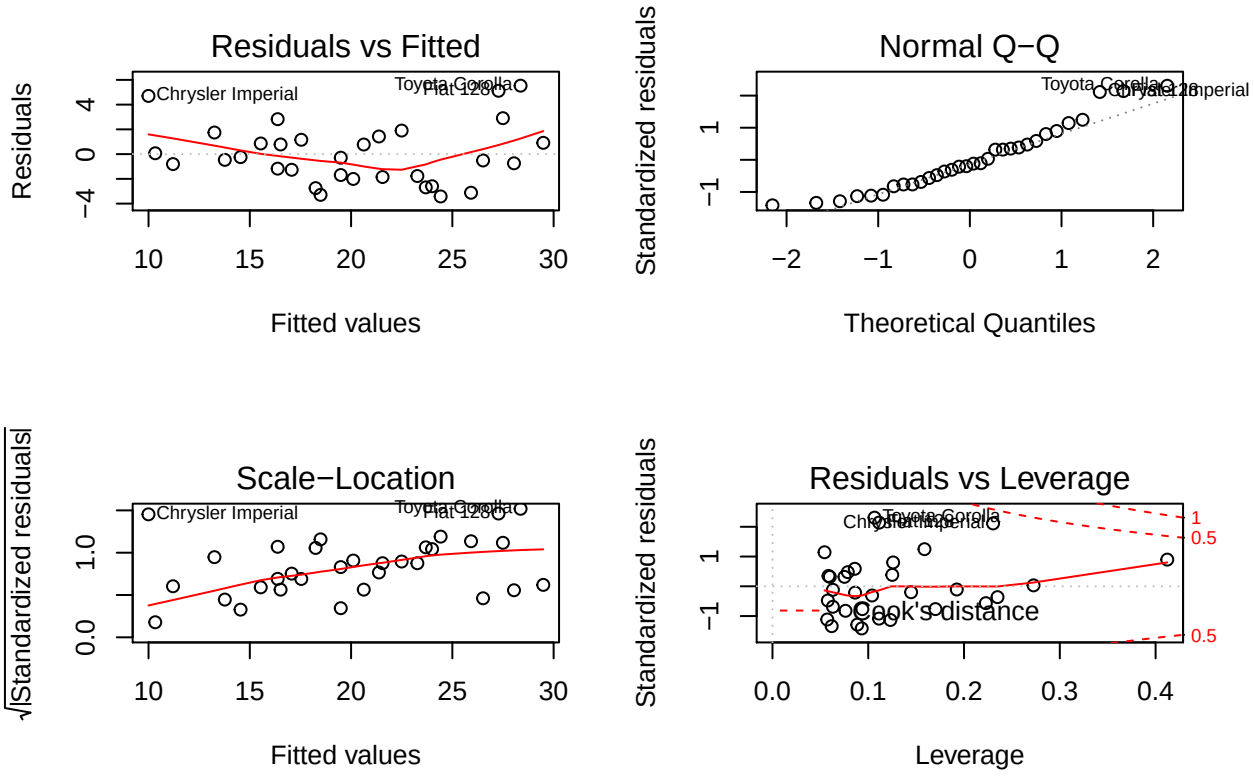
```
##
## Call:
## lm(formula = mpg ~ am + wt + hp, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4221 -1.7924 -0.3788  1.2249  5.5317
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.002875   2.642659  12.867 2.82e-13 ***
## am           2.083710   1.376420   1.514 0.141268
## wt          -2.878575   0.904971  -3.181 0.003574 **
## hp          -0.037479   0.009605  -3.902 0.000546 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.538 on 28 degrees of freedom
## Multiple R-squared:  0.8399, Adjusted R-squared:  0.8227
## F-statistic: 48.96 on 3 and 28 DF,  p-value: 2.908e-11
```

At the beginning, we rejected "type of transmissions doesn't affect mpg" at a 95% significant level. But in our final model, we failed to reject it. To explain the coefficient of "am", it means that mpg of manual cars is 7.25 higher than a automatic car. But it's NOT obvious and it is confounded by "wt" and "hp". Weight and gross horsepower do have ACTUAL effect on mpg (mostly weight), as confounding factors. Per unit increase in weight may decrease 2.88 mpg. Per unit increase in horsepower may decrease 0.04 mpg.

The R-squared value of our final model is 84%. And the adjucted R-squared value is 82.3%. Our model can explain 82.3% variation of mpg, seems like a good model. At last, we do residual analysis.

4. Residual Analysis

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



From the Normal Q-Q plot, the distribution of the residuals nearly looks like normal. And from the right-down plot, no obvious pattern in our residuals. We can use this model to predict mpg.