Regression Models Course Project

Kevin Huang

November 5, 2017

Executive Summary

In this project, we use model selection and linear regression to estimate the relationship between the transmission variable(am) and other independent variables, such as Weight(wt), Number of cylinders(cyl), Gross horsepower(hp), to figure out how the transmission will impact on MPG.

We have concluded the following:
1. *Manual* transmission has better MPG compare to *Automatic* transmission when we only use transmission along in the model. However, when we add in other variables, transmission has lower effect in terms of MPG.
2. MPG will increase by 1.8 when the car is *manual* tranmission.

Instrutions

You work for Motor Trend, a magazine about the automobile industry. Looking at a data set of a collection of cars, they are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome).
They are particularly interested in the following two questions:
1. "Is an automatic or manual transmission better for MPG"
2. "Quantify the MPG difference between automatic and manual transmissions"

Data Description

The data set *mtcars* contains a data frame with 32 observations on 11 variables.

[, 1] mpg Miles/(US) gallon
[, 2] cyl Number of cylinders
[, 3] disp Displacement (cu.in.)
[, 4] hp Gross horsepower
[, 5] drat Rear axle ratio
[, 6] wt Weight (1000 lbs)
[, 7] qsec 1/4 mile time
[, 8] vs V/S
[, 9] am Transmission (0 = automatic, 1 = manual)
[,10] gear Number of forward gears
[,11] carb Number of carburetors

Data Processing, Transformation, and Exploratory data analysis

We load the data into R, and convert some of the variables to factors.

```
data(mtcars)
df <- mtcars
names(df)
```

```
##  [1] "mpg"  "cyl"  "disp" "hp"   "drat" "wt"   "qsec" "vs"   "am"   "gear"
## [11] "carb"
```

```r
dim(df)
```

```
## [1] 32 11
```

```r
head(df)
```

```
##                    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
```

```r
str(df)
```

```
## '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
df$cyl <- as.factor(df$cyl)
df$vs <- as.factor(df$vs)
df$am <- as.factor(df$am)
df$gear <- as.factor(df$gear)
df$carb <- as.factor(df$carb)
summary(df)
```

```
##       mpg         cyl         disp              hp             drat
##  Min.   :10.40   4:11   Min.   : 71.1   Min.   : 52.0   Min.   :2.760
##  1st Qu.:15.43   6: 7   1st Qu.:120.8   1st Qu.: 96.5   1st Qu.:3.080
##  Median :19.20   8:14   Median :196.3   Median :123.0   Median :3.695
##  Mean   :20.09          Mean   :230.7   Mean   :146.7   Mean   :3.597
##  3rd Qu.:22.80          3rd Qu.:326.0   3rd Qu.:180.0   3rd Qu.:3.920
##  Max.   :33.90          Max.   :472.0   Max.   :335.0   Max.   :4.930
##        wt             qsec        vs     am     gear   carb
##  Min.   :1.513   Min.   :14.50   0:18   0:19   3:15   1: 7
##  1st Qu.:2.581   1st Qu.:16.89   1:14   1:13   4:12   2:10
##  Median :3.325   Median :17.71                 5: 5   3: 3
##  Mean   :3.217   Mean   :17.85                        4:10
##  3rd Qu.:3.610   3rd Qu.:18.90                        6: 1
##  Max.   :5.424   Max.   :22.90                        8: 1
```

Inference

Before we do the model selection, we perform a t-test to test if there is significant difference in mean *mpg* between *automatic* and *manaual* transmission.

```
result <- t.test(mpg ~ am, df)
```

The result of the t-test shows the p-value is 0.001; therefore, it is significantly different in the mean of *automatic* and *manual*. The mean *mpg* for *automatic* transmission is 17.147, and the mean *mpg* for *manual* transmission is 24.392.

We can also see the boxplot in appendix 1 comparing the means *mpg* for *manual* and *automatic* transmission.

Regression Analysis and Model Selection

Simple Linear Regression

First off, we fit a initial model which *MPG* as outcome and *am* as the only predictor

```
iniModel <- lm(mpg ~ am, df)
summary(iniModel)
```

```
##
## Call:
## lm(formula = mpg ~ am, data = df)
##
## 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 ***
## am1            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
```

The p-value is almost 0, whcih tells us the variable *am* is significant. However, the R-squares is 0.36, which means only 36% of the variance is explianed by this model. Therefore, we will have to fit other models that includes significant variables to explian the variance.

Model Selection

By looking at pairs plot (appendix 2) and the correlations between *MPG* and variables (appendix 3), I choose *cyl*, *disp*, *hp*, *wt* along with *am* to fit more models because they are highly correlated to *MPG*. I will be using nested model testing to find the significant variables.

```
fit1 <- lm(mpg ~ am + wt, df)
fit2 <- lm(mpg ~ am + wt + cyl, df)
fit3 <- lm(mpg ~ am + wt + cyl + disp, df)
```

```r
fit4 <- lm(mpg ~ am + wt + cyl + disp + hp, df)
anova.test.1 <- anova(iniModel,fit1, fit2, fit3, fit4)

fit4_rm_disp <- lm(mpg ~ am + wt + cyl + hp, df)
anova.test.2 <- anova(iniModel,fit1, fit2, fit4_rm_disp)

bestModel <- lm(mpg ~ am + wt + cyl + hp, df)
```

I put *am* as first predictors and followed by the order of correlation to *MPG*. Then I run the anova analysis (appendix 4). From model 3 and model 4, we can see that by adding *disp* does not have much impact to the model, but model 5 is significant, meaning adding *hp* is significant. Therefore, I remove disp and run anova again (appendix 5).

The anova analysis suggests that the 4th model, removed disp, is significant. We have the best model, *mpg ~ am + wt + cyl + hp*.

Multiple Linear Regression

```r
best <- summary(bestModel)
best
```

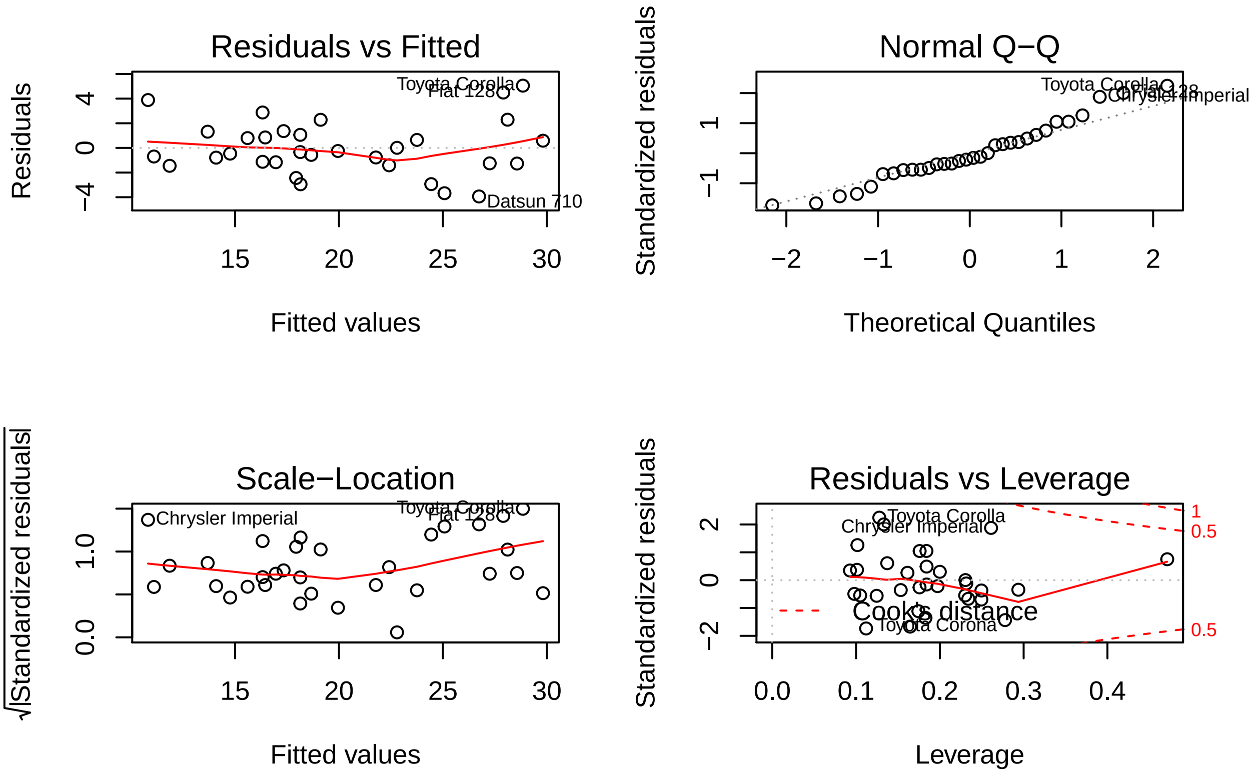
```
##
## Call:
## lm(formula = mpg ~ am + wt + cyl + hp, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9387 -1.2560 -0.4013  1.1253  5.0513
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.70832    2.60489  12.940 7.73e-13 ***
## am1          1.80921    1.39630   1.296  0.20646
## wt          -2.49683    0.88559  -2.819  0.00908 **
## 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 *
## ---
## 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 p-values are small, the R-square is 0.866, the coefficient for *am* is 1.81 which suggests that driving *manual* transmission car will increase *mpg* by 1.81 than *automatic*.

Residual and Diagnostics

We will use the residual plot to diagnosis if there is any outlier that affect the model.

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



From the above plot, we have a few observations,
1. The *Residuals vs Fitted* plot does not appear a linear relationship.
2. The *QQ plot* appears a little bit of tail, but it is normal overall. Therefore, it suggests that the residuals are normally distributed. 3. The *Scale Location* plot appears a horizontal line, again, it suggests the residuals are spread equally along the ranges of predictors, meaning homoscedasticity.
4. The *Residual vs Leverage* plot appears there is no influtial case in the model.

Conclusion

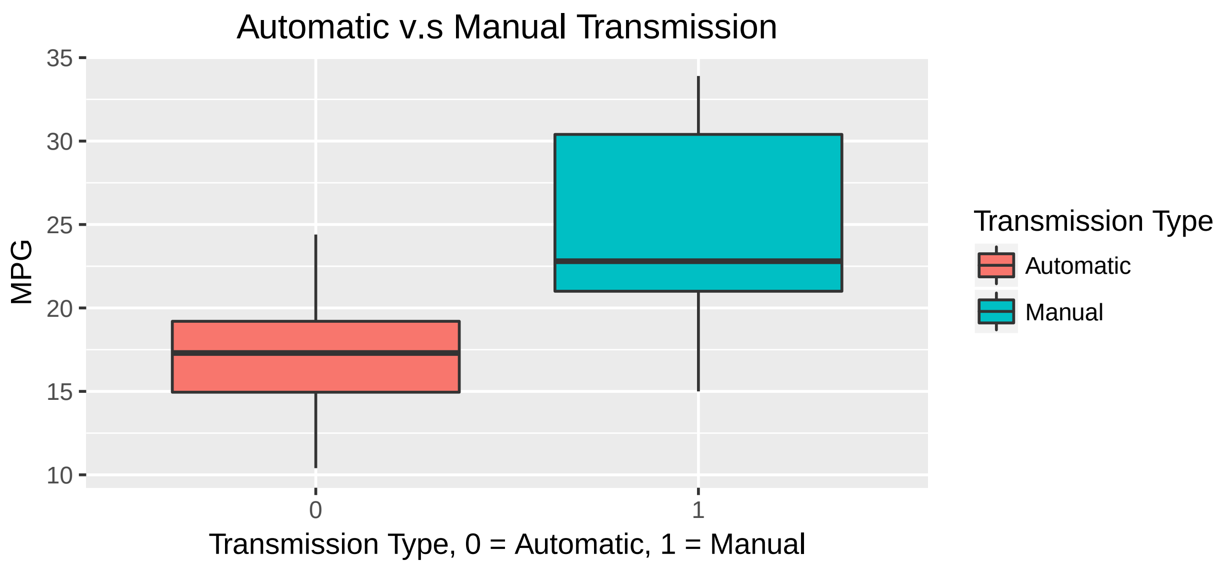
1. After the model selection, our best model is *lm(mpg ~ am + wt + cyl + hp, df)*.
2. After performing the residuals analysis, we can be sure our model is correct since the residuals are normally distributed, and there is no influtial case in the model.
3. *MPG* will increase 1.8 when choosing *manual* transmission over *automatic*.

Appendix

Appendix 1

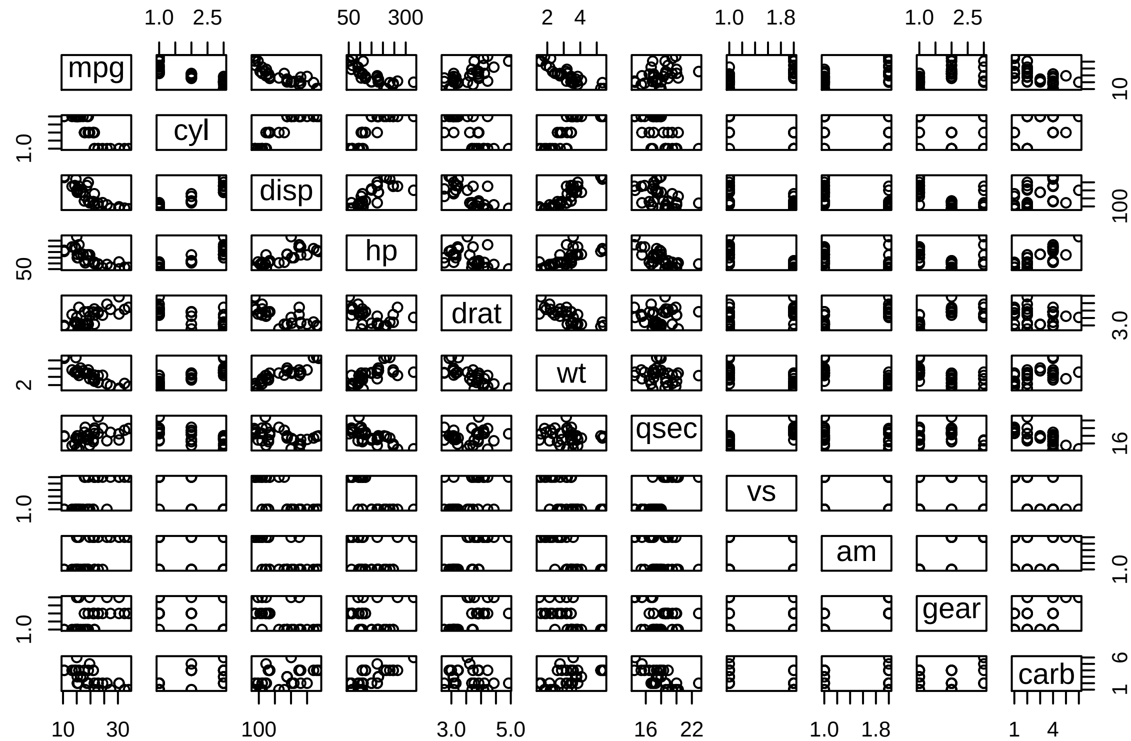
```
library(ggplot2)
ggplot(df, aes(x = am, y = mpg, group = am)) +
    geom_boxplot(aes(fill = am)) +
    ggtitle("Automatic v.s Manual Transmission") +
```

```r
    xlab("Transmission Type, 0 = Automatic, 1 = Manual") + ylab("MPG") +
    theme(plot.title = element_text(hjust = 0.5)) +
    scale_fill_discrete(name="Transmission Type",
                        breaks=c("0", "1"),
                        labels=c("Automatic", "Manual"))
```



Appendix 2

```r
pairs(df)
```



Appendix 3

```
suppressPackageStartupMessages(library(Hmisc))
cor <- rcorr(as.matrix(df))
cor$r[1,]
```

```
##        mpg        cyl       disp         hp       drat         wt
##  1.0000000 -0.8521619 -0.8475513 -0.7761683  0.6811719 -0.8676594
##       qsec         vs         am       gear       carb
##  0.4186840  0.6640389  0.5998324  0.4802848 -0.5509251
```

Appendix 4

```
anova.test.1
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + wt
## Model 3: mpg ~ am + wt + cyl
## Model 4: mpg ~ am + wt + cyl + disp
## Model 5: mpg ~ am + wt + cyl + disp + hp
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
```

```
## 2     29 278.32  1    442.58 73.5623 6.452e-09 ***
## 3     27 182.97  2     95.35  7.9244   0.00216 **
## 4     26 182.87  1      0.10  0.0165   0.89895
## 5     25 150.41  1     32.46  5.3954   0.02862 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Appendix 5

```
anova.test.2
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + wt
## Model 3: mpg ~ am + wt + cyl
## Model 4: mpg ~ am + wt + cyl + hp
##   Res.Df    RSS Df Sum of Sq       F   Pr(>F)
## 1     30 720.90
## 2     29 278.32  1    442.58 76.1924 3.32e-09 ***
## 3     27 182.97  2     95.35  8.2077 0.001725 **
## 4     26 151.03  1     31.94  5.4991 0.026935 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```