Regression Models Course Project

This document is written by Hyunsik Shim.

Overview - Problem Statement

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:

- “Is an automatic or manual transmission better for MPG”
- "Quantify the MPG difference between automatic and manual transmissions"

Dataset - Motor Trend Car Road Tests

The data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models).

- `mpg` Miles/(US) gallon `cyl` Number of cylinders `disp` Displacement (cu.in.)
- `hp` Gross horsepower `drat` Rear axle ratio `wt` Weight (lb/1000)
- `qsec` 1/4 mile time `vs` V/S `am` Transmission (0 = automatic, 1 = manual)
- `gear` Number of forward gears `carb` Number of carburetors

Result and Executive Summary

You can show this summary from refering the following appendix which include simple linear regression,multivariate linear regression and model comparison(ANOVA).

From simple linear regression model, we can know the following:

1. The expression, MPG = 17.15 + 7.24 am explains 36% of the variation of MPG in the data.
2. The model is not so good.

By carefully compensating for weight and horsepower in our limited data set from multivariate linear regression, we can make the following statements:

1. In comparing automatic to manual transmission MPG, the observed differences are most likely due to vehicle weight and horsepower, rather than transmission type.
2. Vehicle weight and horsepower explain the variation in MPG, with much less than 1% probability that the differences are due to random chance. The expression, MPG =

34.00288 + 2.08371 am- 2.87858 wt - 0.03748 hp explains 84% of the variation of MPG in the data.

From model comparison, Model 2(result of multivariate linear regression) is more significant than Model 1 (resilt of simple linear regression).

Appendix - Data Analysis

A.1 Exploratoty Data Analysis

```
library("UsingR")
library("ggplot2")
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 ...
```

- All Data are number.

```
mtcars$am <- as.factor(mtcars$am)
levels(mtcars$am)
```

```
## [1] "0" "1"
```

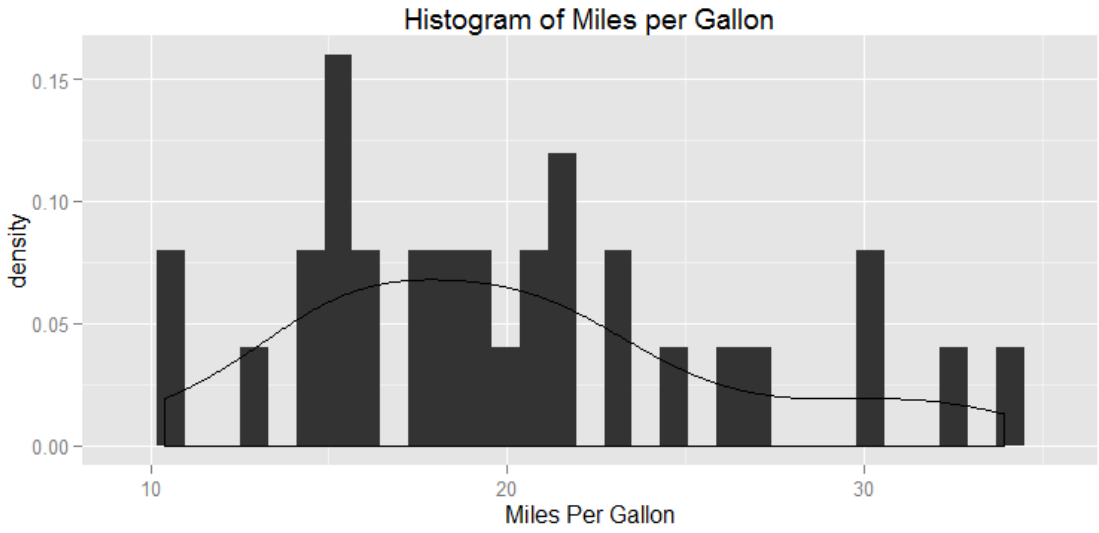
```
levels(mtcars$am) <- c("Automatic", "Manual")
```

- We transfom transmission(`am`) from number to factor "Automatic" and "Manual" as a level

Histogram and density of Miles per Gallon

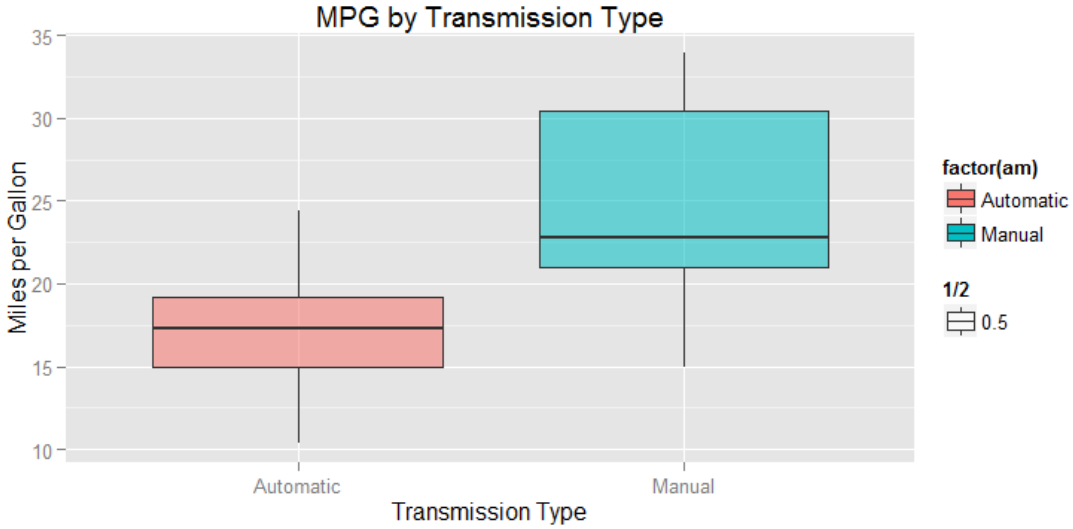
```
# Histogram with Normal Curve
p <- ggplot(mtcars, aes(x = mpg))
p <- p + geom_histogram(aes(y=..density..,binwidth=1.5))
p <- p + geom_density(fill=NA, colour="black")
p <- p+ xlab("Miles Per Gallon") + ggtitle("Histogram of Miles per Gallon")
p
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust thi
s.
```



Box plot of Miles per Gallon

```
p <- ggplot(mtcars, aes(factor(am), mpg))
p <- p + geom_boxplot(aes(fill = factor(am), alpha = 1/2))
p <- p + ggtitle("MPG by Transmission Type")
p + xlab("Transmission Type")+ylab("Miles per Gallon")
```



- We can find no outlier in Box plot
- We considered only one variable am to explorempg The boxplot shows that there is a difference in the MPG by transmission type.
- Manual transmission seems to have more miles per gallon than automatic transmission.

A.2 T-Test

```
result<- aggregate(mpg~am, data = mtcars, mean)
result
```

```
##          am   mpg
## 1 Automatic 17.15
## 2    Manual 24.39
```

- The mean of manaul transmission is (MPG) higher than automaic transmission.
- So, we need a verificaiton if the mean is different using T-test.

t-Test

- T- test is tried as 2 types when the assumption is normal distribution and when it is not.
- $H_0: \mu_{Auto} = \mu_{Manual}$ vs $H_1:$ not H_0

```
autoData <- mtcars[mtcars$am == "Automatic",]
manualData <- mtcars[mtcars$am == "Manual",]
t.test(autoData$mpg, manualData$mpg, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  autoData$mpg and manualData$mpg
## t = -4.106, df = 30, p-value = 0.000285
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -10.848  -3.642
## sample estimates:
## mean of x mean of y
##     17.15     24.39
```

```
t.test(autoData$mpg, manualData$mpg)
```

```
##
##  Welch Two Sample t-test
##
## data:  autoData$mpg and manualData$mpg
## t = -3.767, df = 18.33, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.28  -3.21
## sample estimates:
## mean of x mean of y
##     17.15     24.39
```

- We can know thay both significance level are less than 0.05.
- There, Reject H_0!, that is, the null hypothesis is rejected and vehicle fuel efficiency in accordance with the manual or automatic transmission is hard to look like

A.3 Simple Linear Regression

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

```
##
## Call:
## lm(formula = mpg ~ am, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.392 -3.092 -0.297  3.244  9.508
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    17.15       1.12   15.25  1.1e-15 ***
## am              7.24       1.76    4.11  0.00029 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.9 on 30 degrees of freedom
## Multiple R-squared:  0.36,   Adjusted R-squared:  0.338
## F-statistic: 16.9 on 1 and 30 DF,  p-value: 0.000285
```

Interpretting Coefficient

- The Intercept 17.1474 is the average miles per gallon for automatic transmission, the Slope 7.2449 is the increased miles per gallon for manual transmision.
- r-squred value 0.3598 means our model only explains 35.98% of the total variance.
- With a p-value of 2.8502 × 10-4, we reject the null hypothesis and claim that there is a signficiant difference in the mean MPG between manual transmission cars and automatic transmission cars.

A.4 Multivariate Linear Regression

Correlation

- To choose the appropriate covariates for the regression model, we did covariate adjustment and multiple models to prob the effects. Before that, let's look at the corrlation for mpg variable of our dataset mtcars

```
a<-cor(mtcars)
sort(a[1,])
```

```
##      wt     cyl    disp      hp    carb    qsec    gear      am      vs
## -0.8677 -0.8522 -0.8476 -0.7762 -0.5509  0.4187  0.4803  0.5998  0.6640
##    drat     mpg
##  0.6812  1.0000
```

- In addition to `am`, we see that `wt`, `cyl`, `disp`, and `hp` are highly correlated with our dependent variable `mpg`. As such, they may be good candidates to include in our model. However, after we look at the correlation matrix, we see that `cyl` is highly correlated with `hp`. and `disp` is highly correlated with `wt`, and they are both corrlated with each other. Since predictors should not exhibit collinearity, we may should not have `cyl` and `disp` in in our model.

```
bestfit <- lm(mpg~am + wt + hp, data = mtcars)
summary(bestfit)
```

```
##
## Call:
## lm(formula = mpg ~ am + wt + hp, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.422 -1.792 -0.379  1.225  5.532
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.00288    2.64266   12.87  2.8e-13 ***
## am           2.08371    1.37642    1.51  0.14127
## wt          -2.87858    0.90497   -3.18  0.00357 **
## hp          -0.03748    0.00961   -3.90  0.00055 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.54 on 28 degrees of freedom
## Multiple R-squared:  0.84,   Adjusted R-squared:  0.823
## F-statistic:   49 on 3 and 28 DF,  p-value: 2.91e-11
```

- All varaibles except `am` are significant.
- R-squred value 0.8399 means our model only explains 83.99% of the total variance.

Model comparison (ANOVA)

```
anova(fit, bestfit)
```

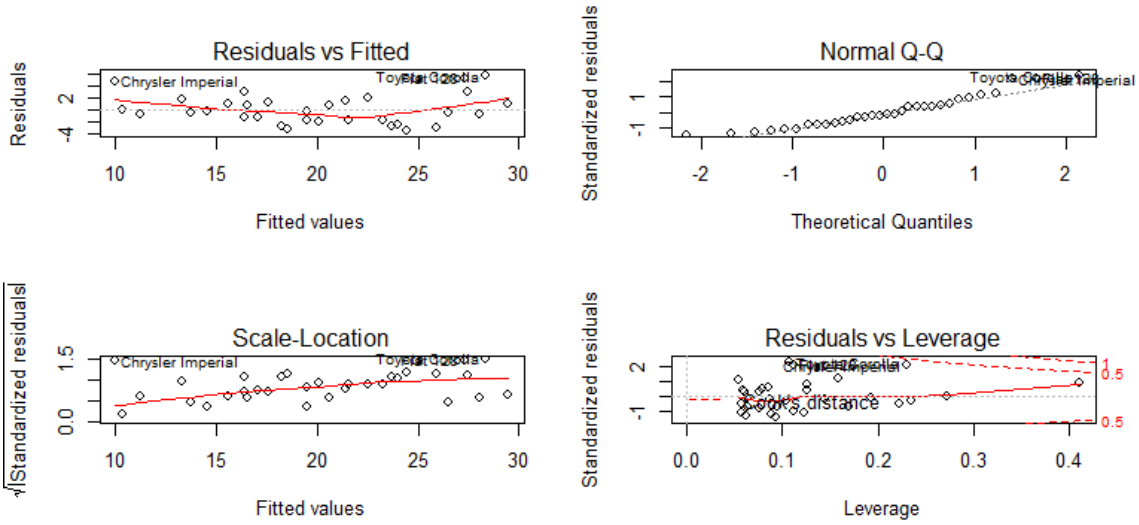
```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + wt + hp
##   Res.Df RSS Df Sum of Sq  F  Pr(>F)
## 1     30 721
## 2     28 180  2       541 42 3.7e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- We compare two model `fit`and `bestfit` using ANOVA.

- P-value is **3.745e-09** and `bestfitmode` is significant.

Residual analysis

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



- We can find normality and homoscedasticity from graph.
- Toyota Corolla, Fiat 128, & Honda Civic, with manual transmissions, are outliers with this formula, due to their very light weight & low hp.