Regression Models Course Project

## 1. Problem explanation

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

## 2. Data Analysis

#### 2.1 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)

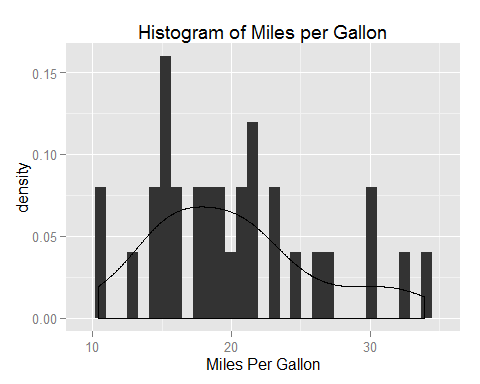
## [1] "0" "1"

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

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

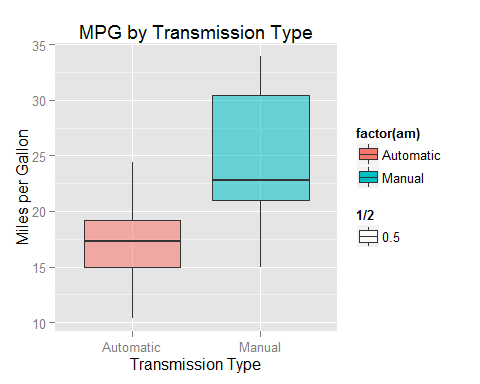
##### Histogram and density of Miles per Gallon

# 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



##### 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 + labs(title = "MPG by Transmission Type")   
p + labs(x = "Transmission Type", y = "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.

#### 2.2 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 2 types when the assumption is normal distribution and when it is not.

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 !, that is, the null hypothesis is rejected and vehicle fuel efficiency in accordance with the manual or automatic transmission is hard to look like

#### 2.3 Simple Linear Regression

data(mtcars)   
cor(mtcars)

## mpg cyl disp hp drat wt qsec vs  
## mpg 1.0000 -0.8522 -0.8476 -0.7762 0.68117 -0.8677 0.4187 0.6640  
## cyl -0.8522 1.0000 0.9020 0.8324 -0.69994 0.7825 -0.5912 -0.8108  
## disp -0.8476 0.9020 1.0000 0.7909 -0.71021 0.8880 -0.4337 -0.7104  
## hp -0.7762 0.8324 0.7909 1.0000 -0.44876 0.6587 -0.7082 -0.7231  
## drat 0.6812 -0.6999 -0.7102 -0.4488 1.00000 -0.7124 0.0912 0.4403  
## wt -0.8677 0.7825 0.8880 0.6587 -0.71244 1.0000 -0.1747 -0.5549  
## qsec 0.4187 -0.5912 -0.4337 -0.7082 0.09120 -0.1747 1.0000 0.7445  
## vs 0.6640 -0.8108 -0.7104 -0.7231 0.44028 -0.5549 0.7445 1.0000  
## am 0.5998 -0.5226 -0.5912 -0.2432 0.71271 -0.6925 -0.2299 0.1683  
## gear 0.4803 -0.4927 -0.5556 -0.1257 0.69961 -0.5833 -0.2127 0.2060  
## carb -0.5509 0.5270 0.3950 0.7498 -0.09079 0.4276 -0.6562 -0.5696  
## am gear carb  
## mpg 0.59983 0.4803 -0.55093  
## cyl -0.52261 -0.4927 0.52699  
## disp -0.59123 -0.5556 0.39498  
## hp -0.24320 -0.1257 0.74981  
## drat 0.71271 0.6996 -0.09079  
## wt -0.69250 -0.5833 0.42761  
## qsec -0.22986 -0.2127 -0.65625  
## vs 0.16835 0.2060 -0.56961  
## am 1.00000 0.7941 0.05753  
## gear 0.79406 1.0000 0.27407  
## carb 0.05753 0.2741 1.00000

sort(cor(mtcars)[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

* factor로 변경해준 am변수를 초기화 하기위해 데이터를 새로 불러들임
* correlation matrix는 수치형 변수만 계산가능하기 때문
* mpg에 대한 correlation 값을 보기 위해 sort()와 변수선택을 실행
* am을 비롯한 wt, cyl, disp, hp 변수가 mpg와 상관관계가 높은 것으로 나타남
* 여기서 cyl과 disp변수의 상관관계가 굉장히 높은 것으로 나타난것을 확인
* collinearity(공선성)을 고려하여 두 변수는 제거
* 또한 wt와 hp는 높을수록 연비가 하락할것이라는 상식적인 배경을 바탕으로 분석 실시

단순선형회귀 실시

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

* 위의 검정에서 차이를 보았기 때문에 이 모형( mpg ~ am )에서 정보를 더이상 얻긴 어려움
* 다만 회귀계수가 모두 유의함
* 따라서 자동변속기 차량은 평균적으로 17.147MPG의 연비를 가지고 있고 수동변속기 차량은 7.245MPG만큼의 연비효율이 있음
* 또한 이 모형은 R-squared 값에 의해 35.98%정도 설명력을 가지고 있다고 할 수 있음

#### 6.4 다중회귀분석

종속변수를 연비, 독립변수를 무게, 마력, 변속기종류 3개로 두고 다중회귀분석을 실시

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

* am을 제외한 나머지 변수는 모두 유의하다는 것을 확인
* R-squared 값이 0.8399로 모형이 83.99%의 설명력을 가지고 있다고 할 수 있음

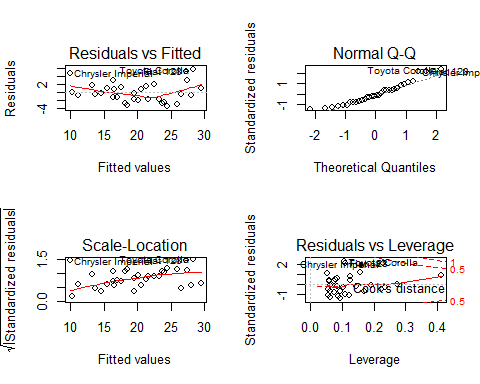
모델비교(ANOVA)

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

* 단순선형회귀식인 fit과 다중회귀식인 bestfit이 차이가 있는지 ANOVA를 통해 비교
* p-value가 **3.745e-09**로 bestfit모형이 유의함을 확인 > 잔차분석

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

 - 결과를 통해 정규성 및 등분산성 확인 - "Chrysler Imperial", "Toyota Corolla", "Fiat 128"은 이상점으로 필요시 데이터 정제