Multivariate statistics – Regression and regression diagnostics

Suckwon Hong

Ulsan National Institute of Science and Technology, 50 UNIST-gil Ulsan, Republic of Korea
amoeba94@unist.ac.kr

Question 1. Dealing with missing values and outlier

1.1. Missing values

A variable *Horsepower* contains six missing values. We assigned mean value of the variable to the missing ones.

1.2. Outliers

We employed two approaches to handle outliers existing in continuous variables (i.e., *mpg, cylinders, displacement, horsepower, weight, acceleration, model.year)*. First, for each continuous variable, we replace data points above *(75% quantile +1.5*interquartile)* or *(25% quantile – 1.5*interquartile)* with the value of *(70% quantile)* and *(30% quantile)*, respectively. Second, we made simple linear regression model, and calculated Cook's distance. According to the rule of thumb, we checked whether there exist data points with the value above *0.03*. As a result, we eliminated four data points from the original.

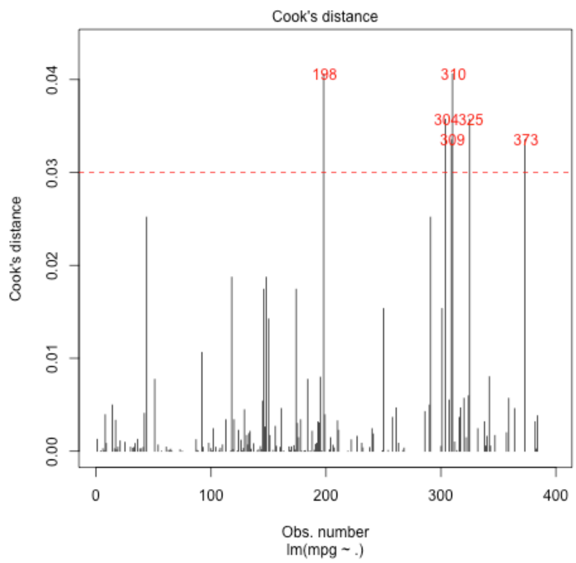


Figure 1. Cook's distance

Question 2. Testing assumptions for a linear regression

2.1. Normality assumption

Three approaches are employed to test the normality assumption: histogram, qqplot, Shapiro-Wilk test. Histogram and qqplot for residuals show that residuals generally follow a normal distribution. After, we conducted Shapiro-Wilk normality test. However, the result showed that residuals do not follow a normal distribution *(p-value < 2.2e-16)*.

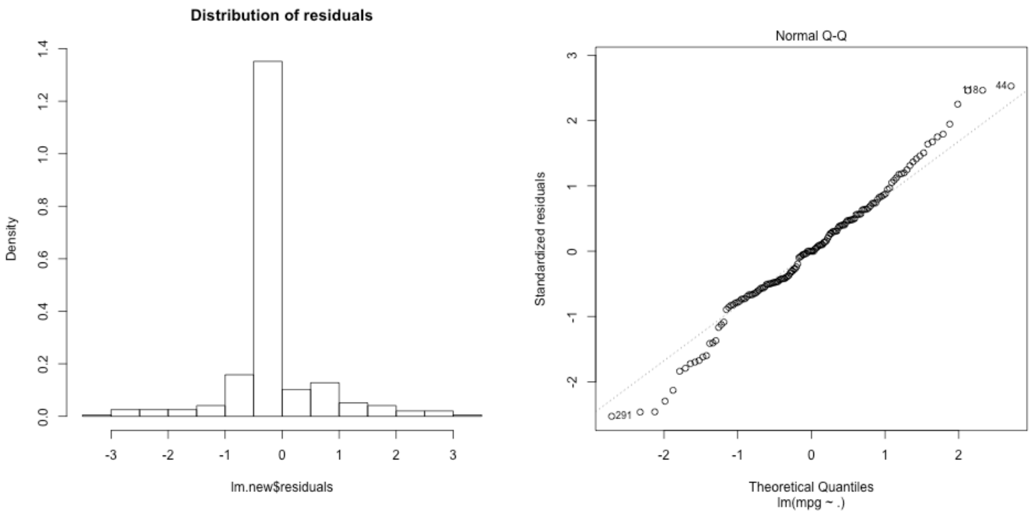
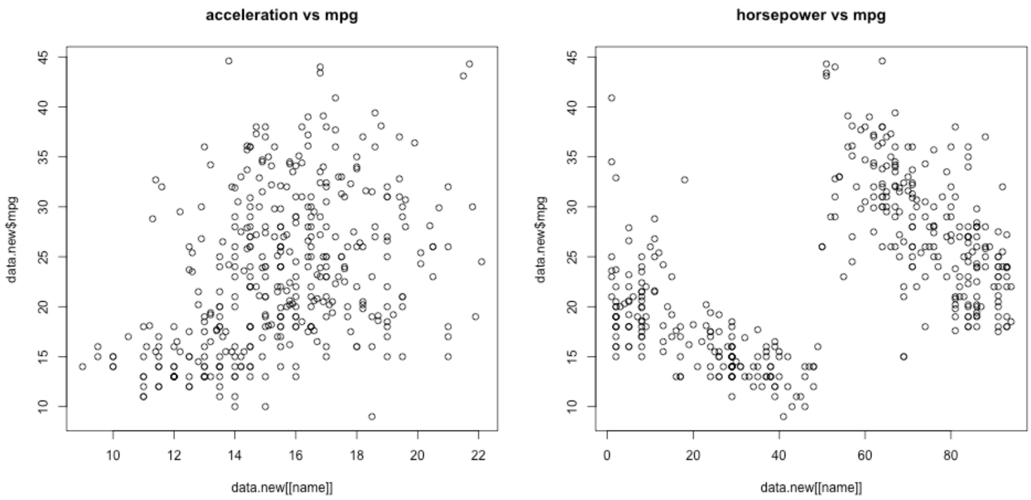


Figure 2. Histogram and qqplot for continuous variables

2.2. Linearity assumption

To test a linearity assumption, we drew scatterplots between dependent variable and different independent variables. It seems that the variable *horsepower* does not meet an assumption, so we omit the variable in running a regression.



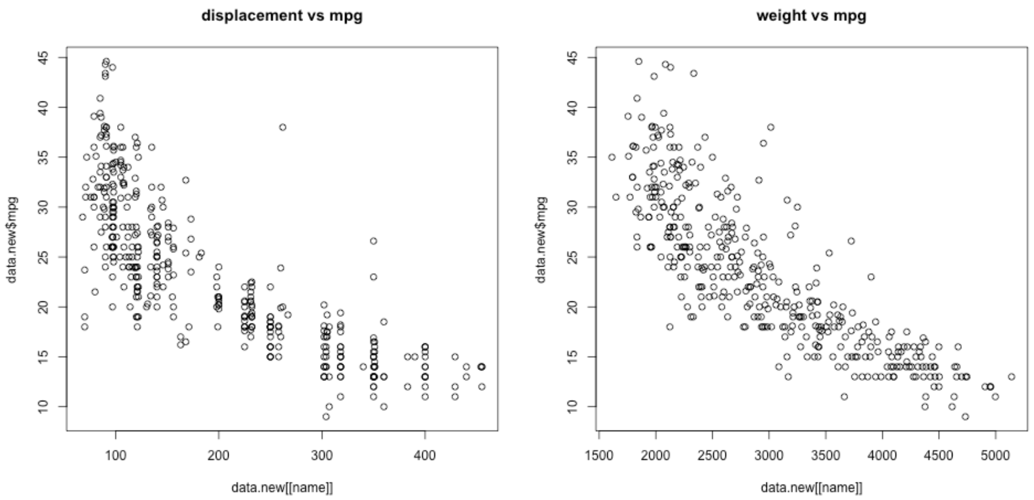


Figure 3. Scatterplots vs dependent variable

2.3. Homoscedasticity assumption

We drew residual plot for the last assumption (i.e. homoscedasticity). Since we observed no pattern from the plot, the data seem to meet the assumption.

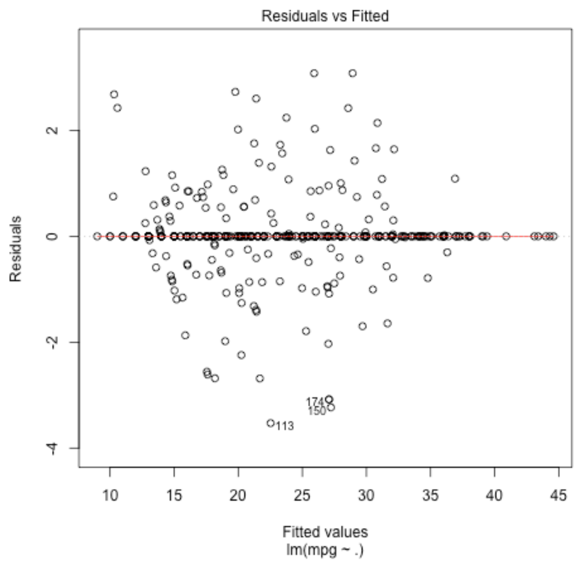


Figure 4. Residuals vs fitted plot

Question 3. Evaluate multicollinearity

We computed correlation between continuous variables and variance inflation factor (VIF) to check an existence of multicollinearity. Since we found that variables *weight* and *displacement* are highly correlated from the correlation matrix, we computed VIF. The VIF of the model was 93.3245, and it is found that multiple values of the variable *car.name* are highly correlated. Thus, we omit the variable.

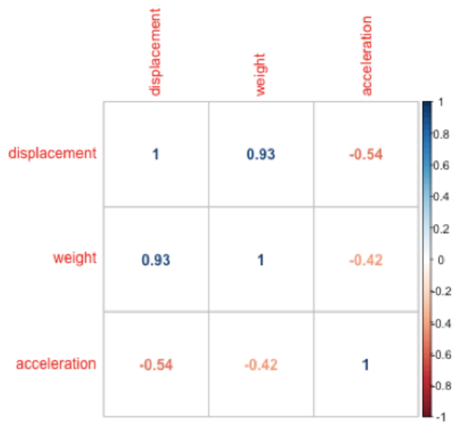


Figure 5. Correlation matrix

Question 4. Regression results

As a result, we run regression on *mpg* using variables of *cylinders, displacement, weight, acceleration, model.year, origin*. We tested the assumptions again, and it seems that the data meet the assumptions. Additionally, the value of VIF was 7.5330 implying that there is no multicollinearity problem.

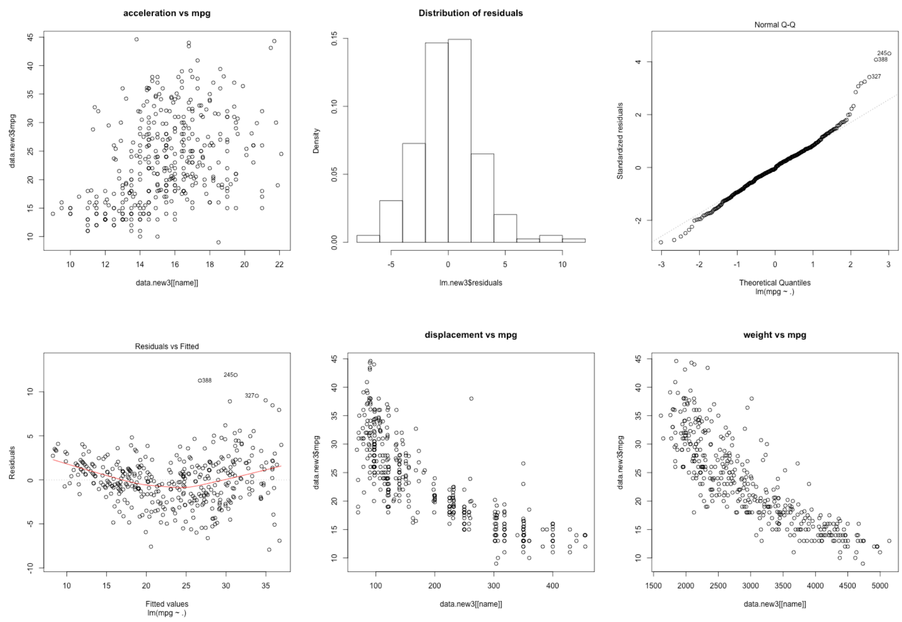


Figure 6. Regression diagnostics

The value of R-squared of our final regression model was 0.8673, which means that the independent variables explained about 87% of total variance of dependent variable *mpg.* Furthermore, we concluded that the model is significantly significant because the p-value of the model was small enough. Most values of categorical variables (i.e., *cylinders, model.year, origin)* were shown to be statistically significant while only one continuous variable (i.e., *weight)* was significant. From the coefficients, we can say that for every additional unit of *weight,* we can expect *mpg* to decrease by average of 0.006.

```
Call:
lm(formula = mpg ~ ., data = data.new3)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.9205 -1.6678 -0.1363  1.5740 11.9114 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)  28.4268767  2.1001197  13.536  < 2e-16 ***
cylinders4    6.8947847  1.5480193   4.454 1.12e-05 ***
cylinders5    7.4197881  2.3561718   3.149 0.001771 ** 
cylinders6    4.8748593  1.7074968   2.855 0.004547 ** 
cylinders8    6.7294035  1.9744644   3.408 0.000726 ***
displacement  0.0059664  0.0065765   0.907 0.364873    
weight       -0.0061118  0.0005379 -11.363  < 2e-16 ***
acceleration  0.0913918  0.0775128   1.179 0.239134    
model.year71  1.6854487  0.7810048   2.158 0.031567 *  
model.year72  0.1737705  0.7921370   0.219 0.826483    
model.year73 -0.0163790  0.7112849  -0.023 0.981641    
model.year74  2.1902900  0.8146049   2.689 0.007496 ** 
model.year75  1.8900314  0.7925717   2.385 0.017597 *  
model.year76  2.4481421  0.7708549   3.176 0.001619 ** 
model.year77  3.8036186  0.7969196   4.773 2.62e-06 ***
model.year78  3.6821787  0.7618101   4.833 1.97e-06 ***
model.year79  5.7138255  0.8000996   7.141 4.93e-12 ***
model.year80  8.9162413  0.8367824  10.655  < 2e-16 ***
model.year81  7.5835506  0.8162609   9.291  < 2e-16 ***
model.year82  8.7312281  0.8243937  10.591  < 2e-16 ***
origin2       1.8148616  0.5158109   3.518 0.000488 ***
origin3       1.9194970  0.4894114   3.922 0.000105 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.865 on 370 degrees of freedom
Multiple R-squared:  0.8673,	Adjusted R-squared:  0.8597 
F-statistic: 115.1 on 21 and 370 DF,  p-value: < 2.2e-16
```

Figure 7. Regression results

R Code.

```r
### Multivaraite statistics - assignment 2
rm(list = ls())
## install packages
install.packages("car")
install.packages("dplyr")
install.packages("corrplot")
install.packages("fmsb")
library(car)
library(dplyr)
library(corrplot)
library(fmsb)
## load data
data <- read.csv("mpg.csv",header = TRUE)
## data manipulation
data$origin <- as.factor(data$origin)
data[["model.year"]]<-as.factor(data[["model.year"]])
data[["cylinders"]]<-as.factor(data[["cylinders"]])
data[["horsepower"]]<-as.numeric(data[["horsepower"]])
str(data)
## handling missing data
data[data=="?"] <- NA
na_count <-sapply(data, function(x) sum(length(which(is.na(x)))))
data[is.na(data)] <- mean(data$horsepower, na.rm = TRUE) # mean implementation

## evaluate outliers
# using IQR
for(name in names(data)){
  if((name == "car.name") || (name=="origin")||(name=="cylinders")||(name=="model.year")){
  }
  else{
  x <- data[[name]]
  qnt <- quantile(x, probs=c(.25, .75), na.rm = T)
  caps <- quantile(x, probs=c(.3, .7), na.rm = T)
  H <- 1.5 * IQR(x, na.rm = T)
  x[x < (qnt[1] - H)] <- caps[1]
  x[x > (qnt[2] + H)] <- caps[2]
  data[[name]]<-x
  }
}
```

```r
# using cook's distance
lm.pre <- lm(mpg~., data = data)
cooksd<- cooks.distance(lm.pre)
png(filename = "cookd.png")
plot(lm.pre, which=4, labels.id = "")
abline(h=0.03, lty=2, col="red")
text(x=1:length(cooksd)+1, y=cooksd, labels=ifelse(cooksd>0.03,names(cooksd),""), col="red")
dev.off()

data$cooksd<-cooksd
data[data=="NaN"]<- 0
data$cooky <- ifelse(data$cooksd<0.03,"keep","no")
data.new <- subset(data, cooky=="keep")
data.new <- data.new[-c(10,11)]
lm.new <- lm(mpg~., data = data.new)

## test regression assumptions
#normality using histogram
png(filename = "histogram.png")
hist(lm.new$residuals, main = "Distribution of residuals", freq = FALSE)
dev.off()
#normality using qqplot
png(filename = "qqplot.png")
plot(lm.new, which = 2)
dev.off()
#normality using Shaprio-Wilk W test
print(shapiro.test(lm.new$residuals))
#linearity using scatterplot
for(name in names(data.new)){
  if((name == "car.name") || (name=="origin")||(name=="cylinders")||(name=='mpg')||(name=="model.year")){
  }
  else{
  png(filename = paste("scatterplot",name,".png"))
  plot(data.new[[name]],data.new$mpg, main=paste(name,"vs mpg"))
  dev.off()
  }
}
data.new2 <- subset(data.new, select = -c(horsepower))
lm.new2 <- lm(mpg~., data = data.new2)
```

```r
#Homoscedasticity using residual plot
png(filename = "residual vs fitted.png")
plot(lm.new2,which=1)
dev.off()

## evaluate multicollinearity
#correlation matrix
cor<-cor(select(data.new2, displacement, weight, acceleration)) #only continuous
cor <- as.matrix(cor)
png(filename = "correlation matrix.png")
corrplot(as.matrix(cor), method = "number")
dev.off()

fmsb::VIF(lm.new2)
ld.vars <- attributes(alias(lm.new2)$Complete)
data.new3 <- subset(data.new2,select = -c(car.name))

## run final regression
lm.new3 <- lm(mpg~., data = data.new3)
# test assumptions again
png(filename = "histogram.png")
hist(lm.new3$residuals, main = "Distribution of residuals", freq = FALSE)
dev.off()
png(filename = "qqplot.png")
plot(lm.new3, which = 2)
dev.off()
print(shapiro.test(lm.new3$residuals))

for(name in names(data.new3)){
  if((name == "car.name") || (name=="origin")||(name=="cylinders")||(name=='mpg')||(name=="model.year")){
  }
  else{
    png(filename = paste("scatterplot",name,".png"))
    plot(data.new3[[name]],data.new3$mpg, main=paste(name,"vs mpg"))
    dev.off()
  }
}
png(filename = "residual vs fitted.png")
plot(lm.new3,which=1)
dev.off()
fmsb::VIF(lm.new3)
summary(lm.new3)
```