2019 학년도 1학기
DATA MINING
HW3

과목명	데이터마이닝
담당교수명	송종우 교수님
제출일	2019.04.03
학번	182STG27
이름	임지연

Ⅰ. Description

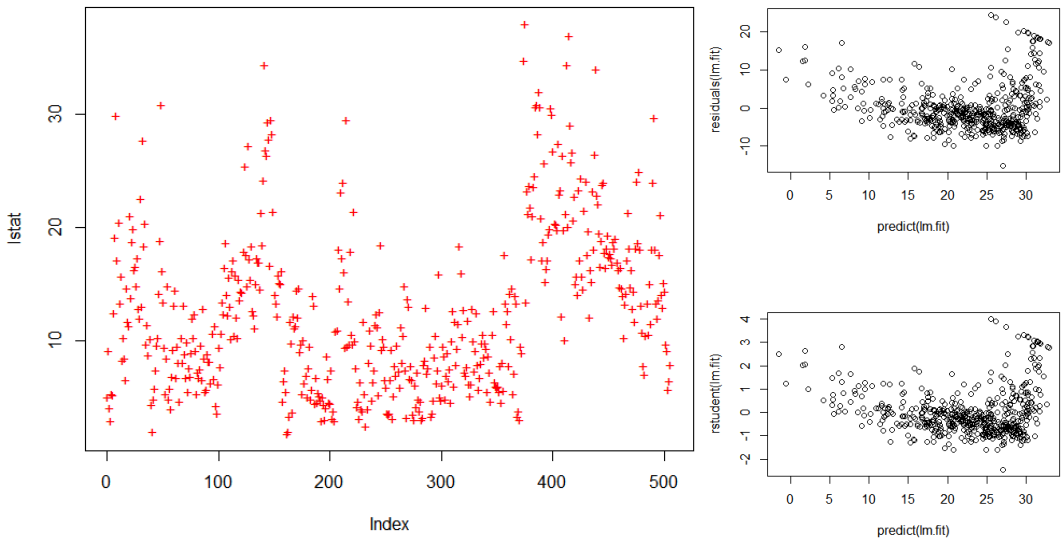
Linear Regression 은 설명변수와 종속변수간의 선형관계가 있는지를 알기 위하여 실시하는 방법이다. 따라서 R 에서 기본적으로 사용되는 lm 함수를 이용하여 데이터가 선형함수에 적합한지를 알 수 있으며 RSS가 최소가 되는 직선을 가장 최적의 모델로 선정한다. 기본적인 lm 적합을 알기 위해 Lab 에 나와있는 예제 코드를 실행해본 후, Example 8,9,13,14 문제를 풀어보며 실제 데이터에 적합해보는 연습을 한다.

Ⅱ. Implementation

Question 1.

Lab : Linear Regression 의 코드를 실행해보고 감상문을 써라

R에서 lm함수를 사용하여 회귀분석을 복습할 수 있었다. lm , predict , plot , summary 함수를 통해 데이터를 적합해보고 회귀진단을 실시하는 일련의 과정을 시행해 볼 수 있었다. 더하여 hatvalues, poly, contrasts 함수에 대해서 새롭게 알게 되었다. 또한 다중공선성을 진단하는 vif(다중공선성)에 대해 다시 한번 생각해보며 R로도 실행해 볼 수 있었다. 보통 vif 가 10 보다 크면 해당 다중공선성이 있다고 판단하고 변수들 사이의 어떤 변수끼리의 관계가 있는지를 살펴본 후 변수들 일부를 빼거나 새로운 변수를 만들거나 할 수 있다. 그 외에 제시되어 있던 코드를 실행해 본 그래프를 아래에 첨부했다.



Question 2.

3.7 Applied - Example 8,9,13,14 풀어라

[Example 8]

This question involves the use of simple linear regression on the Auto data set.

(a) Use the lm() function to perform a simple linear regression with mpg as the response and horsepower as the predictor. Use the summary() function to print the results. Comment on the output.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	39.936	0.717	55.66	0
horsepower	-0.158	0.006	-24.489	0
				Multiple R-squared = 0.6059

For example:

i. Is there a relationship between the predictor and the response?

▶ Multiple R-squared 값이 0.6059 로 종속변수와 설명변수간의 상관관계가 있다.

ii. How strong is the relationship between the predictor and the response?

▶ p-value 값이 0에 가깝기 때문에, 둘 사이에 강한 선형관계가 있다.

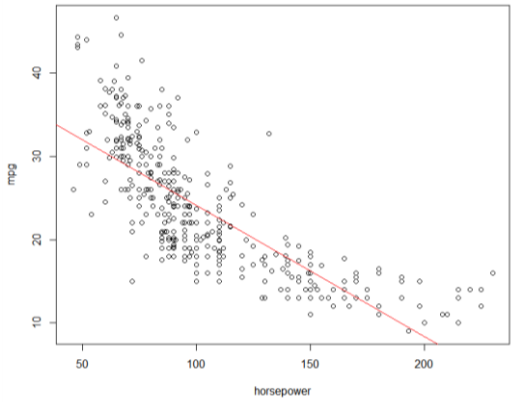
iii. Is the relationship between the predictor and the response positive or negative?

▶ 설명변수 horsepower 의 회귀계수가 -0.158로 둘 사이에 음의 상관관계가 있다.

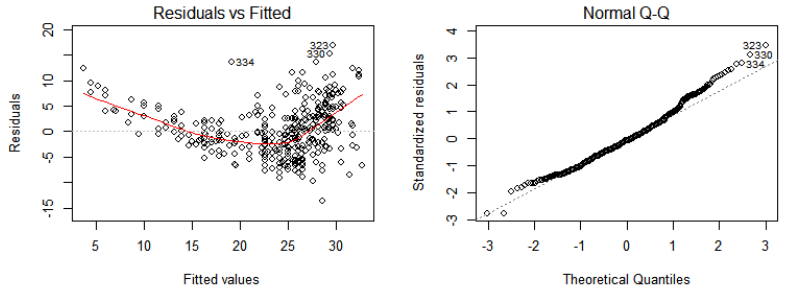
iv. What is the predicted mpg associated with a horsepower of 98? What are the associated 95% confidence and prediction intervals?

	Fit	Lower	Upper
Confidence	24.47	23.97	24.96
Prediction	24.47	14.81	34.12

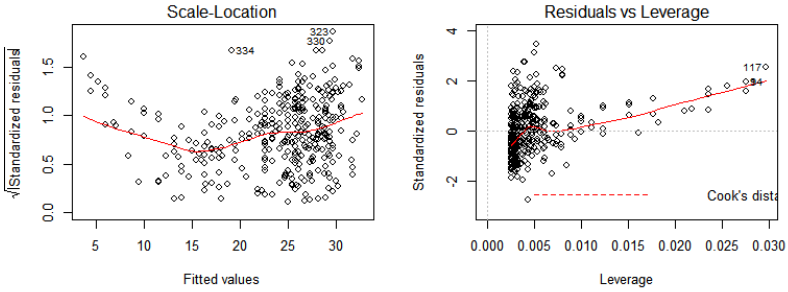
(b) Plot the response and the predictor. Use the abline() function to display the least squares regression line.



(c) Use the plot() function to produce diagnostic plots of the least squares regression fit. Comment on any problems you see with the fit.



▶ residuals VS fitted 그래프를 살펴봤을 때, 2차함수 형태를 가진다. 따라서 비선형관계임을 알 수 있고, 선형회귀분석을 시행하기에 무리가 있다.

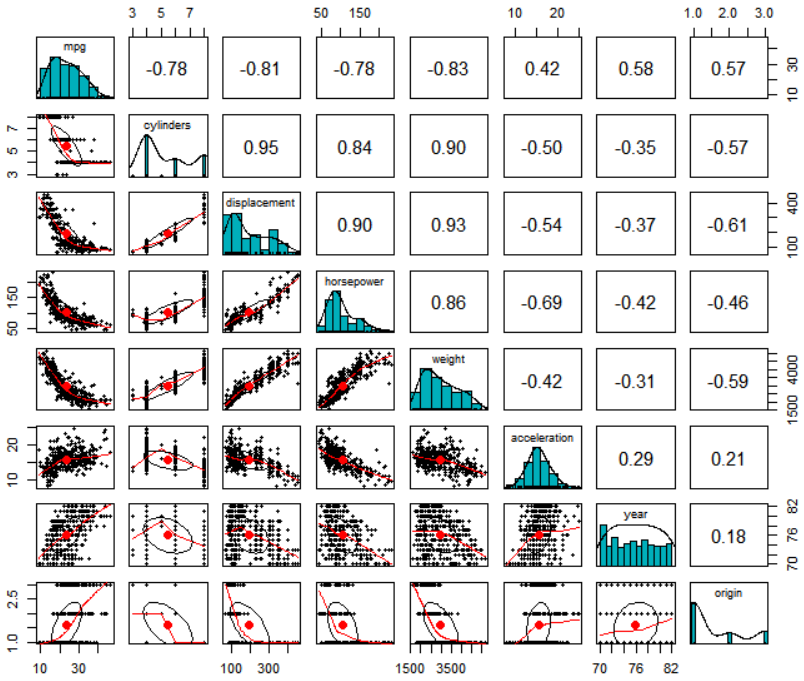


▶ residuals VS fitted 그래프를 살펴봤을 때, obs 1170 점이 high leverage point 라는 것을 알 수 있다.

[Example 9]

9. This question involves the use of multiple linear regression on the Auto data set.

(a) Produce a scatterplot matrix which includes all of the variables in the data set.



(b) Compute the matrix of correlations between the variables using the function cor(). You will need to exclude the name variable, which is qualitative.

	mpg	cylinders	displacement	horsepower	weight	acceleration	year	origin
mpg	1	-0.78	-0.81	-0.78	-0.83	0.42	0.58	0.57
cylinders	-0.78	1	0.95	0.84	0.9	-0.5	-0.35	-0.57
displacement	-0.81	0.95	1	0.9	**0.93**	-0.54	-0.37	**-0.61**
horsepower	-0.78	0.84	0.9	1	0.86	-0.69	-0.42	-0.46
weight	-0.83	0.9	0.93	0.86	1	-0.42	-0.31	-0.59
acceleration	0.42	-0.5	-0.54	-0.69	-0.42	1	0.29	0.21
year	0.58	-0.35	-0.37	-0.42	-0.31	0.29	1	**0.18**
origin	0.57	-0.57	-0.61	-0.46	-0.59	0.21	0.18	1

(c) Use the lm() function to perform a multiple linear regression with mpg as the response and all other variables except name as the predictors. Use the summary() function to print the results. Comment on the output. For instance:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-17.218	4.644	-3.707	0
cylinders	-0.493	0.323	-1.526	0.128
displacement	**0.02**	**0.008**	**2.647**	**0.008**
horsepower	-0.017	0.014	-1.23	0.22
weight	**-0.006**	**0.001**	**-9.929**	**0**
acceleration	0.081	0.099	0.815	0.415
year	**0.751**	**0.051**	**14.729**	**0**
origin	**1.426**	**0.278**	**5.127**	**0**
			Multiple R-squared:	0.8215

i. Is there a relationship between the predictors and the response?

▶ Multiple R-squared 값이 0. 8215 로 종속변수와 설명변수들간의 상관관계가 있다.

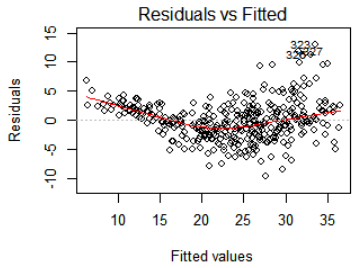
ii. Which predictors appear to have a statistically significant relationship to the response?

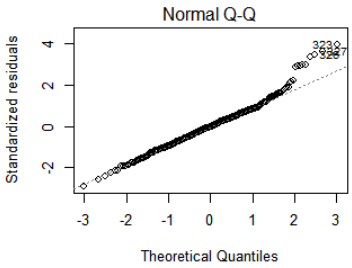
▶ p-value 값이 0에 가까운 displacement, weight, year, origin 변수가 가장 유의하며 종속변수와 유의한 상관관계가 있다는 결론을 얻을 수 있다.

iii. What does the coefficient for the year variable suggest?

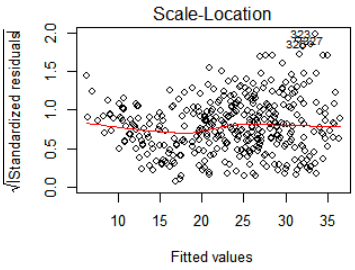
▶ year 변수의 회귀계수가 0.751로, 다른 변수는 고정한 상태로 year 변수가 한 단위 증가할 때 종속변수인 mpg 는 0.751만큼 증가한다. 즉, 0.15만큼 감소한다.

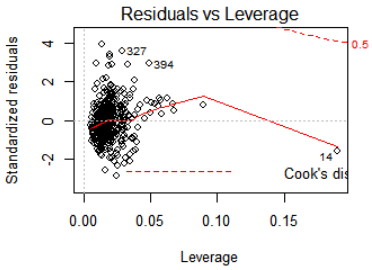
(d) Use the plot() function to produce diagnostic plots of the linear regression fit. Comment on any problems you see with the fit. Do the residual plots suggest any unusually large outliers? Does the leverage plot identify any observations with unusually high leverage?





▶ residuals VS fitted 그래프를 살펴봤을 때, 2차함수 형태를 가진다. 따라서 비선형관계임을 알 수 있고, 선형회귀분석을 시행하기에 무리가 있다.





▶ Residuals vs Fitted 그래프로부터 비선형성을 알 수 있고, Residuals vs Leverage 그래프로부터 Cook's ditance 기준에서 14 observation 이 high leverage point 이다.

(e) Use the * and : symbols to fit linear regression models with interaction effects. Do any interactions appear to be statistically significant?

▶ 앞서 displacement, weight, year, origin 변수가 가장 유의한 것으로 판단되었기 때문에 이 4가지 변수만을 사용하여 상호작용효과를 판단해 보았다. 4개 변수끼리 상호작용을 넣어 몇몇 모델을 시행해 본 결과 대부분의 상호작용 효과 항이 유의하였고, 그 중 R-squared 값이 가장 높았던 아래와 같은 모델을 선택했다.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-107.6	12.904	-8.339	0
displacement	0	0.005	-0.088	0.93
origin	0.912	0.255	3.579	0
year	1.962	0.172	11.436	0
weight	0.026	0.005	5.722	0
year:weight	0	0	-7.214	0
Multiple R-squared: 0.8397				

(f) Try a few different transformations of the variables, such as $\log(X), \sqrt{X}, X^2$. Comment on your findings.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-22.772	3.82	-5.962	0
poly(displacement, 2)1	-15.437	9.868	-1.564	0.119
poly(displacement, 2)2	27.589	3.621	7.62	0
weight	-0.005	0.001	-9.761	0
year	0.809	0.047	17.304	0
origin	0.367	0.274	1.342	0.18
Multiple R-squared: 0.8419				

[Example 13]

In this exercise you will create some simulated data and will fit simple linear regression models to it. Make sure to use set.seed(1) prior to starting part (a) to ensure consistent results.

(a) Using the rnorm() function, create a vector, x, containing 100 observations drawn from a *N(0, 1)* distribution. This represents a feature, X.

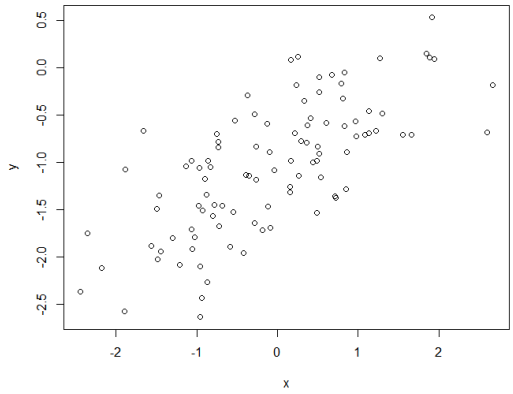
(b) Using the rnorm() function, create a vector, eps, containing 100 observations drawn from a *N(0, 0.25)* distribution i.e. a normal distribution with mean zero and variance 0.25.

(c) Using x and eps, generate a vector y according to the model ——————————$Y = -1 + 0.5X + \varepsilon$. (3.39)

What is the length of the vector y? What are the values of β_0 and β_1 in this linear model?

▶ y 의 길이는 100 이고, 여기서 β_0=-1, β_1=0.5 값을 갖는다.

(d) Create a scatterplot displaying the relationship between x and y. Comment on what you observe.



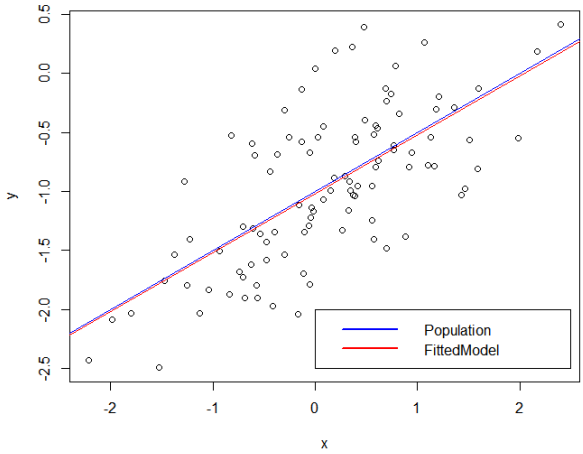
▶ x,y 는 서로 양의 관계를 갖는 듯 보인다.

(e) Fit a least squares linear model to predict y using x. Comment on the model obtained. How do $\hat{\beta}_0$ and $\hat{\beta}_1$ compare to β_0 and β_1?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.019	0.048	-21.01	0
x	0.499	0.054	9.273	0
Multiple R-squared: 0.4674				

▶ 각각 $\hat{\beta}_0 = -1.019$, $\hat{\beta}_1 = 0.499$ 로, $\beta_0 = -1$, β_1=0.5 와 매우 가까운 값을 갖는다.

(f) Display the least squares line on the scatterplot obtained in (d). Draw the population regression line on the plot, in a different color. Use the legend() command to create an appropriate legend.



(g) Now fit a polynomial regression model that predicts y using x and x2. Is there evidence that the quadratic term improves the model fit? Explain your answer.

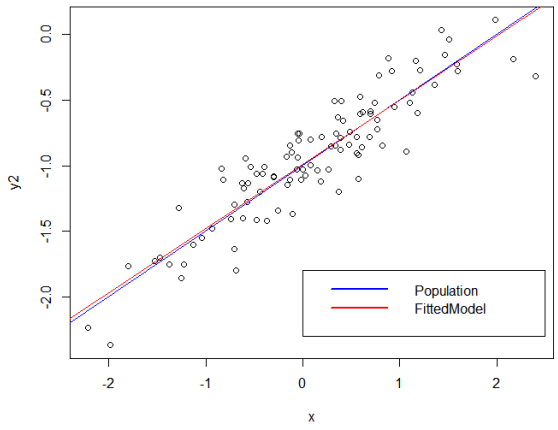
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.019	0.048	-21.01	0
x	0.499	0.054	9.273	0
Multiple R-squared: 0.4674				

Model 1: y ~ x				
Model 2: y ~ x + x^2				
	Res.Df	RSS	Df	Sum of Sq
1	98	22.709		
2	98	22.709	0	0

▶ 2 차항을 포함한 모델이 1 차항만을 포함한 모델보다 더 성능이 좋다고 할 수 없다. 왜냐하면 RSS 값이 감소하지 않았기 때문이다. 따라서 더 간단한 모델인 1 차항만 포함된 모델을 사용하는 것이 좋다.

(h) Repeat (a)–(f) after modifying the data generation process in such a way that there is less noise in the data. The model (3.39) should remain the same. You can do this by decreasing the variance of the normal distribution used to generate the error term ε in (b). Describe your results.

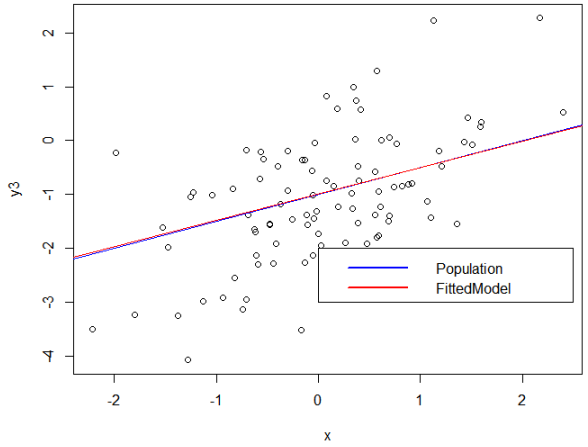
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.988	0.02	-49.285	0
x	0.489	0.022	21.945	0
Multiple R-squared: 0.8309				



▶ 데이터의 분산이 eps 값이 0.5 일 때보다 감소한 것이 눈으로 확인된다. Coef 추정치는 거의 같지만 RSE, R-squared 값은 향상된 것을 알 수 있다.

(i) Repeat (a)–(f) after modifying the data generation process in such a way that there is more noise in the data. The model (3.39) should remain the same. You can do this by increasing the variance of the normal distribution used to generate the error term ε in (b). Describe your results.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.211	0.109	-11.106	0
x	0.604	0.121	4.986	0
Multiple R-squared: 0.2024				



▶ 데이터의 분산이 eps 값이 0.5 일 때보다 증가한 것이 눈으로 확인된다. Coef 추정치는 거의 같지만 RSE, R-squared 값은 감소된 것을 알 수 있다.

(j) What are the confidence intervals for β0 and β1 based on the original data set, the noisier data set, and the less noisy data set? Comment on your results.

	The less noisy data set		Original data set		The noiser data set	
	2.5%	97.5%	2.5%	97.5%	2.5%	97.5%
(Intercept)	-1.028	-0.949	-1.115	-0.923	-1.427	-0.995
X	0.445	0.533	0.393	0.606	0.364	0.844

▶ noise 가 작은 데이터셋에서 noise 가 큰 데이터셋으로 갈수록 신뢰구간의 넓이가 점점 넓어지는 것을 알 수 있다.

[Example 14]

This problem focuses on the collinearity problem.

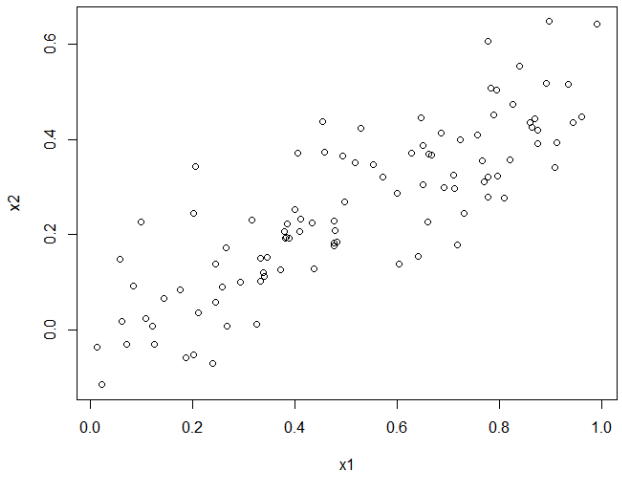
(a) Perform the following commands in R:

```
> set .seed (1)
> x1=runif (100)
> x2 =0.5* x1+rnorm (100) /10
> y=2+2* x1 +0.3* x2+rnorm (100)
```

The last line corresponds to creating a linear model in which y is a function of x1 and x2. Write out the form of the linear model. What are the regression coefficients?

▶ $\beta_0 = 2,\ \beta_1 = 2,\ \beta_2 = 0.3$ 의 형태이다.

(b) What is the correlation between x1 and x2? Create a scatterplot displaying the relationship between the variables.



▶ Correlation = 0.835 로 x1, x2 간의 선형관계가 높은 것으로 나타난다.

(c) Using this data, fit a least squares regression to predict y using x1 and x2. Describe the results obtained. What are $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$? How do these relate to the true β0, β1, and β2? Can you reject the null hypothesis H0 : β1 = 0? How about the null hypothesis H0 : β2 = 0?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.13	0.232	9.188	0
x1	1.44	0.721	1.996	0.049
x2	1.01	1.134	0.891	0.375

▶ β_0의 추정량 b0 의 p-value = 0.049 로, β_0 =0 이라는 귀무가설을 기각할 수 있고, 따라서 y 와의 선형관계가 있음을 알 수 있다. β_1의 추정량 b1 의 p-value = 0.375 로 $\beta_1 = 0$ 이라는 귀무가설을 기각할 수 없다.

(d) Now fit a least squares regression to predict y using only x1. Comment on your results. Can you reject the null hypothesis H0 : β1 = 0?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.112	0.231	9.155	0
x1	1.976	0.396	4.986	0

▶ β_0의 추정량 b0 의 p-value = 0 으로, β_0 =0 이라는 귀무가설을 기각할 수 있고, 따라서 y 와의 선형관계가 있음을 알 수 있다.

(e) Now fit a least squares regression to predict y using only x2. Comment on your results. Can you reject the null hypothesis H0 : β1 = 0?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.39	0.195	12.261	0
x2	2.9	0.633	4.58	0

▶ β_1의 추정량 b1 의 p-value = 0 으로, β_1 =0 이라는 귀무가설을 기각할 수 있고, 따라서 y 와의 선형관계가 있음을 알 수 있다.

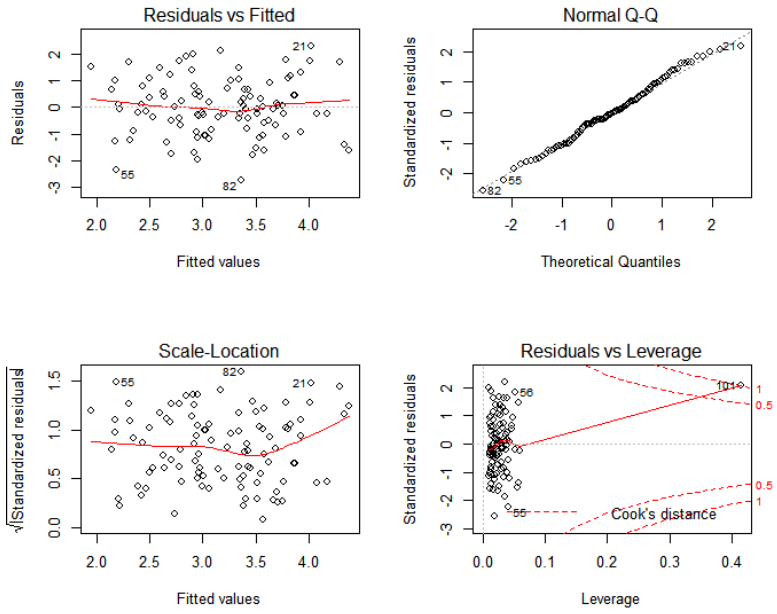
(f) Do the results obtained in (c)–(e) contradict each other? Explain your answer.

▶ 각각의 변수가 독립적으로 모형에 포함되면 유의하지만 모두 포함된 모형에서는 매우 유의하지는 않게 나온다.

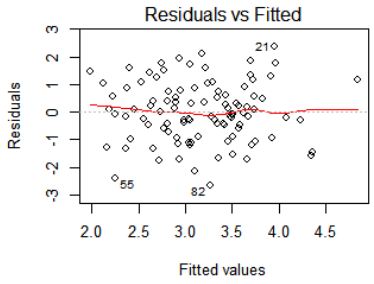
(g) Now suppose we obtain one additional observation, which was unfortunately mismeasured.

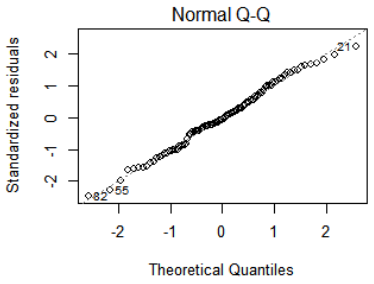
> x1=c(x1 , 0.1) > x2=c(x2 , 0.8) > y=c(y,6)

Re-fit the linear models from (c) to (e) using this new data. What effect does this new observation have on the each of the models? In each model, is this observation an outlier? A high-leverage point? Both? Explain your answers.

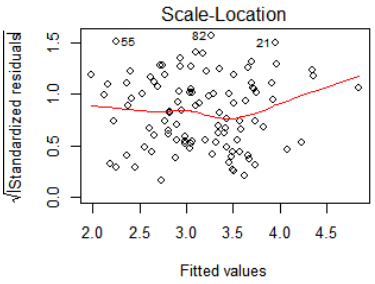


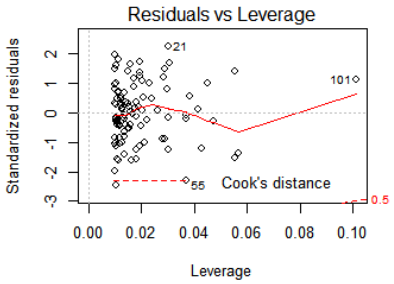
▶ residual vs leverage plot 을 보면 알 수 있듯이 obs 101 은 hight leverage point 를 나타낸다. 심각한 문제가 발생할 수 있다.

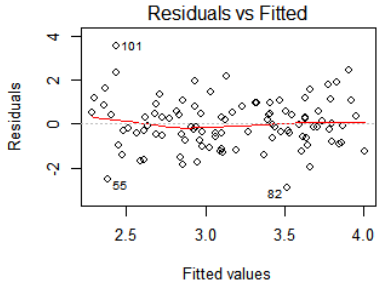


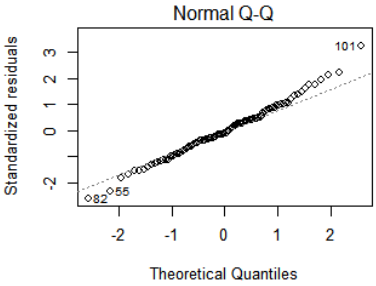


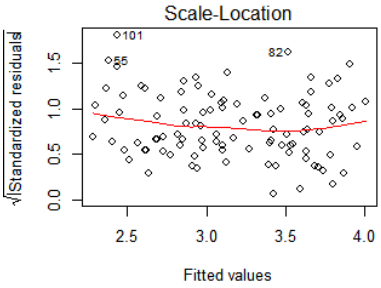
▶ residual vs leverage plot 을 보면 알 수 있듯이 obs 101 은 hight leverage point 를 나타낸다. 하지만 x 의 범위 내에 들어오기 때문에 심각한 문제가 발생하지 않는다.

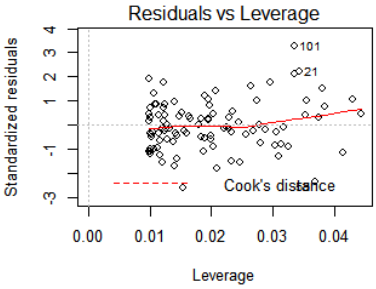












▶ residual vs leverage plot 을 보면 알 수 있듯이 obs 101 은 hight leverage point 를 나타낸다. 하지만 회귀선 내에있기 때문에 심각한 문제는 발생하지 않는다.

III. Discussion

이번 과제를 통해 R 에서 lm 함수를 적합하며 회귀분석의 전반적인 내용을 공부할 수 있었다. 보편적으로 선형회귀분석은 y 와 각각의 x변수들과의 대략적인 관계를 알기 위하여 산점도를 그린 후, 대략적인 형태를 살펴본다. 그 후 step 함수를 사용하여 가장 설명력을 좋게 해주는 모델을 만들기 위한 설명변수를 고른다. 다음으로는 해당 모델에 대한 회귀진단을 하기 위하여 잔차그래프 등을 그려보며 모형의 적합성 및 안정성 등을 평가하면 최종 모형이 선택된다. 이번 과제에서는 step

함수를 사용하여 가장 적절한 모형을 선택하는 과정은 포함되어 있지 않았다. 선형회귀분석의 가장 큰 장점은 직관적인 해석이 가능하다는 것이다. 만약, 여러 교호작용이 포함된 모델에서는 해석하기가 힘들기 때문에 우리는 적절한 교호효과를 포함하여 설명력을 높일 수 있으면서 간단한 모델이 선호한다. 그 후 최종 모델이 여러 가정을 만족시키는지까지 반드시 시행해야만 한다.

Ⅳ. Appendix – R code

```
## 8
#a
options(scipen = 100)
library(ISLR)
library(tidyverse)
data(Auto)
lm.fit= lm(mpg ~ horsepower, data =Auto )
summary(lm.fit)
predict(lm.fit, Auto %>% filter(horsepower == 98))
predict(lm.fit, Auto %>% filter(horsepower == 98), interval = "confidence")
predict(lm.fit, Auto %>% filter(horsepower == 98), interval = "prediction")
#b
par(mfrow = c(1,1))
attach(Auto)
plot(horsepower, mpg)
abline(lm.fit, col = "red")
#c
par(mfrow=c(2,2))
plot(lm.fit)
## 9
#a
library(psych)
pairs.panels(Auto,
             method = "pearson", # correlation method
             hist.col = "#00AFBB",
             density = TRUE,   # show density plots
             ellipses = TRUE # show correlation ellipses)
#b#c
```

```
lm.fit = lm(mpg ~. -name, Auto)
summary(lm.fit)
#d
par(mfrow = c(2,2))
plot(lm.fit)
#e
step( lm(mpg ~ displacement*weight*year*origin) ,direction = "both")
lm.fit1 = lm( mpg~displacement+weight+year:origin, data = Auto)
lm.fit2 = lm( mpg~displacement+weight+year*origin, data = Auto)
lm.fit3 = lm( mpg~displacement+origin+year*weight, data = Auto)
lm.fit4 = lm( mpg~origin+weight+year* displacement, data = Auto)
lm.fit5 = lm( mpg~year+weight+origin* displacement, data = Auto)
summary(lm.fit1)
summary(lm.fit2)
summary(lm.fit3)
summary(lm.fit4)
summary(lm.fit5)
#f
lm.fit1 <- lm(mpg~poly(displacement,3)+weight+year+origin, data=Auto)
lm.fit2 <- lm(mpg~displacement+weight^2+year+origin, data=Auto)
lm.fit3 <- lm(mpg~displacement+log(year)+origin, data=Auto)
summary(lm.fit1)
summary(lm.fit2)
summary(lm.fit3)
##13
#a
set.seed(1)
x = rnorm(100)
#b
eps = rnorm(100, sd = 0.25^0.5)
#c
y = -1 + 0.5*x + eps
length(y)
#d
par(mfrow= c(1,1))
```

```
plot(x,y)
#e
lm.fit = lm(y ~ x)
#f
plot(x,y)
abline(-1, 0.5, col="blue")
abline(lm.fit, col="red")
legend(x = c(0,2.5),y = c(-2.5,-2),
       legend = c("Population", "FittedModel"),
       col = c("blue","red"), lwd=2 )
#g
lm.fit1 <- lm(y~x+x^2)
summary(lm.fit1)
anova(lm.fit, lm.fit1)
#h
eps2 <- rnorm(100, sd=0.2)  # 0.5 -> 0.2로 감소함
y2 = -1 + 0.5*x + eps2
lm.fit2 <- lm(y2 ~ x)
summary(lm.fit2)
plot(x, y2)
abline(-1, 0.5, col="blue")
abline(lm.fit2, col="red")
legend(x = c(0,2.5),y = c(-2.3,-1.8),
       legend = c("Population", "FittedModel"),
       col = c("blue","red"), lwd=2 )
#i
eps3 <- rnorm(100, sd=1)  # 0.5 -> 0.2로 감소함
y3 = -1 + 0.5*x + eps3
lm.fit3 <- lm(y3 ~ x)
summary(lm.fit3)
plot(x, y3)
abline(-1, 0.5, col="blue")
abline(lm.fit3, col="red")
legend(x = c(0,2.5),y = c(-3,-2),
       legend = c("Population", "FittedModel"),
```

```
          col = c("blue","red"), lwd=2 )
#j
confint(lm.fit)
confint(lm.fit2)
confint(lm.fit3)
## 14
#a
set.seed(1)
x1 <- runif(100)
x2 <- 0.5*x1 + rnorm(100)/10
y <- 2 + 2*x1 + 0.3*x2 + rnorm(100)
#b
cor(x1,x2); plot(x1,x2)
#c
fit.lm <- lm(y~x1+x2)
summary(fit.m)
#d
fit.lm1 <- lm(y~x1)
summary(fit.lm1)
#e
fit.lm2 <- lm(y~x2)
summary(fit.lm2)
#g
x1 = c(x1, 0.1)
x2 = c(x2, 0.8)
y = c(y, 6)
par(mfrow=c(2,2))
fit.lm <- lm(y~x1+x2)
summary(fit.lm)
plot(fit.lm)
fit.lm1 <- lm(y~x2)
summary(fit.lm1)
plot(fit.lm1)
fit.lm2 <- lm(y~x1); summary(fit.lm2)
plot(fit.lm2)
```