2023년 1학기 회귀모형 출석수업 과제물

홍원표_202135-368864

2023-04-08

2023년 1학기 회귀모형 출석수업 과제물

이름 : 홍원표

학번 : 202135-368864

연락처 : 010-5343-4341

1 번. 연습문제 1장 1번(p. 39) 자료를 이용하여 1.7 분석사례와 같이 분석하고, 설명하시오.

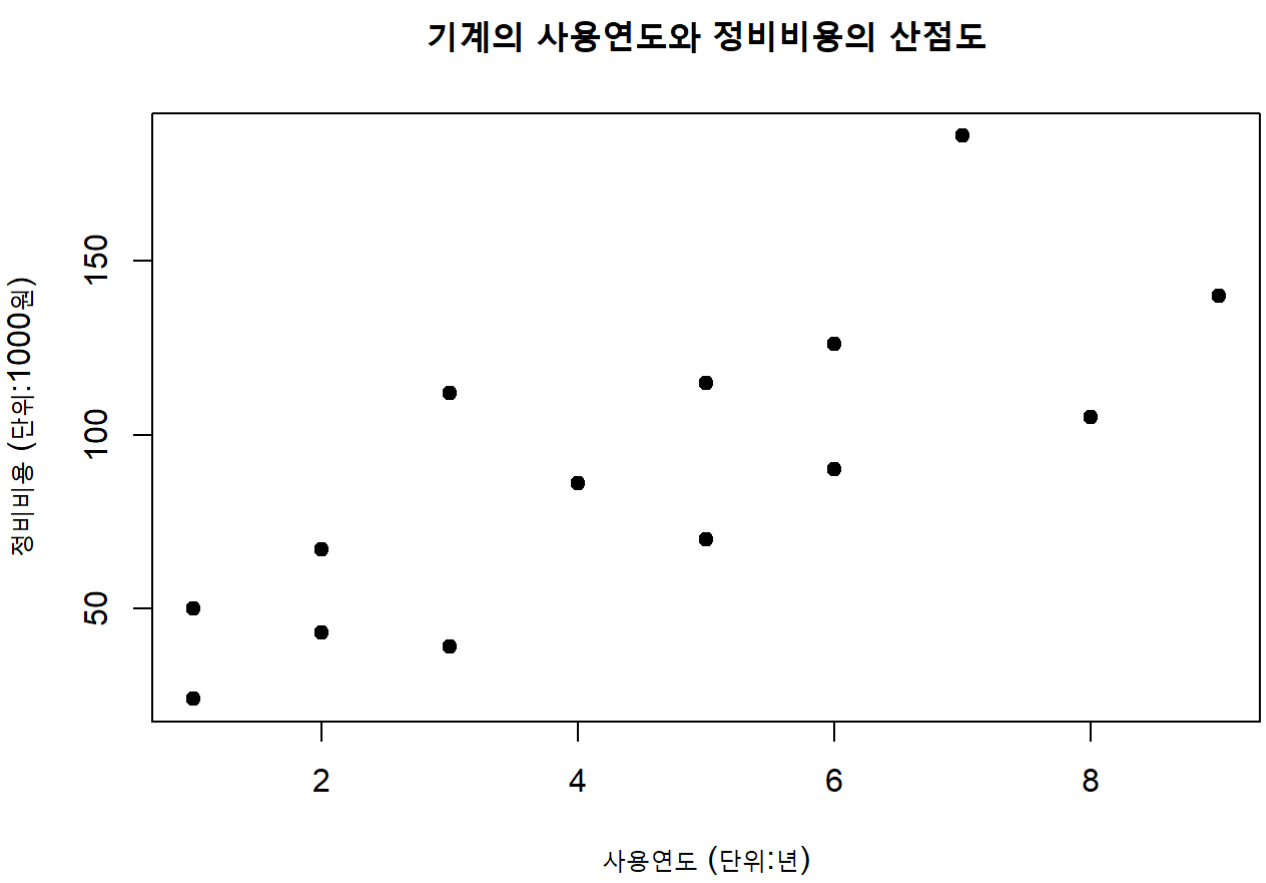
1) 교재의 데이터 입력 및 확인

```
# 교재의 데이터 입력 및 확인
X = c(3, 1, 5, 8, 1, 4, 2, 6, 9, 3, 5, 7, 2, 6)
Y = c(39, 24, 115, 105, 50, 86, 67, 90, 140, 112, 70, 186, 43, 126)
# X, Y 두 변수의 마지막 3개의 값을 확인
tail(cbind(X, Y), 3)
```

```
##       X   Y
## [12,] 7 186
## [13,] 2  43
## [14,] 6 126
```

2) 산점도를 그려본다.

```
# 산점도를 그린다.
plot(X, Y, pch=19, main="기계의 사용연도와 정비비용의 산점도",
     xlab = "사용연도 (단위:년)", ylab = "정비비용 (단위:1000원)")
```



- 사용연도와 정비비용 변수간에는 선형상관계가 존재하는 것으로 보여진다.

3) 회귀모형 적합 및 요약 정보 확인

```
# 회귀모형 적합 및 요약 정보 확인
machine.lm = lm(Y~X)
summary(machine.lm)
```

```
##
## Call:
## lm(formula = Y ~ X)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -33.204 -20.383  -4.748  13.957  61.433
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   29.107     15.969   1.823 0.093341 .
## X             13.637      3.149   4.330 0.000978 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29.11 on 12 degrees of freedom
## Multiple R-squared:  0.6098, Adjusted R-squared:  0.5773
## F-statistic: 18.75 on 1 and 12 DF,  p-value: 0.0009779
```

적합된 모형의 요약정보에서 회귀계수의 추정값의 절편 $b_0 = 29.107$ 이고 기울기 $b_1 = 13.637$인 단순회귀 방정식은 $\hat{Y} = 29.107 + 13.637 \times X$가 된다.

결정계수 $R^2 = 0.6098$로서 총변동중에서 약 61%가 회귀방정식으로 설명되는 회귀변동이 차지하고 있다는 것을 타나낸다.

회귀계수 X의 F-값은 4.330이고 p-값은 0.000978 으로 매우 작으므로 귀무가설 $H_0 : \beta_1 = 0$ 을 기각하고 회귀계수 β_1 이 유의하다는 것을 알 수 있다.

4) 분산분석표 확인

```
anova(machine.lm)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## X          1  15887 15887.2  18.753 0.0009779 ***
## Residuals 12  10166   847.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

분산분석표에서도 검정통계량 $F_0 = 18.753$로 크고 p-value는 0.0009779 로 매우 작기 때문에 적합된 회귀선이 유의하다는 것을 알 수 있다.

5) 잔차와 추정값 보기

```
# 회귀모형 적합 결과(machine.lm)의 변수 확인
names(machine.lm)
```

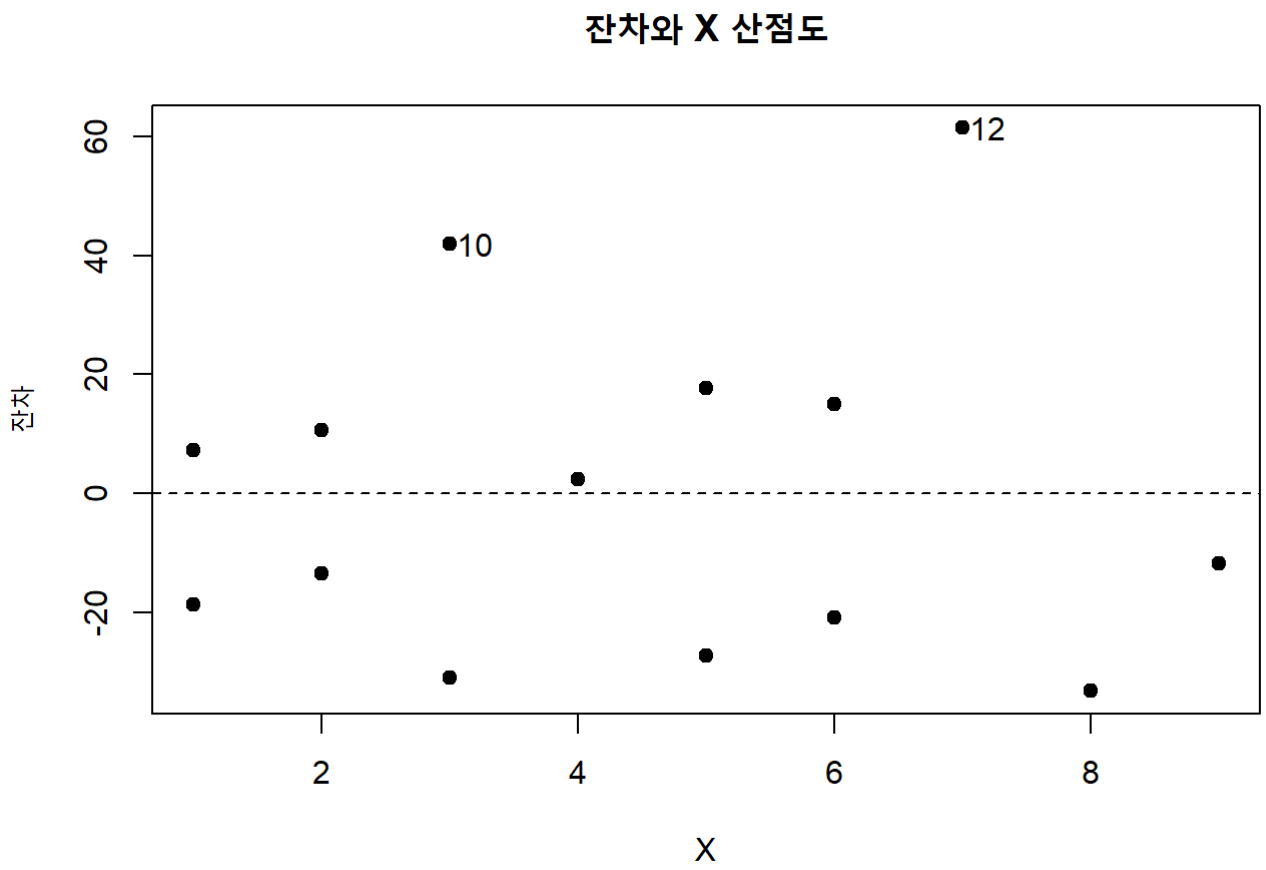
```
##  [1] "coefficients"  "residuals"     "effects"       "rank"
##  [5] "fitted.values" "assign"        "qr"            "df.residual"
##  [9] "xlevels"       "call"          "terms"         "model"
```

```
# X와 Y와 잔차 및 추정값을 합쳐서 보기
cbind(X, Y, resid(machine.lm), fitted(machine.lm))
```

```
##    X   Y
## 1  3  39 -31.018395  70.01839
## 2  1  24 -18.744147  42.74415
## 3  5 115  17.707358  97.29264
## 4  8 105 -33.204013 138.20401
## 5  1  50   7.255853  42.74415
## 6  4  86   2.344482  83.65552
## 7  2  67  10.618729  56.38127
## 8  6  90 -20.929766 110.92977
## 9  9 140 -11.841137 151.84114
## 10 3 112  41.981605  70.01839
## 11 5  70 -27.292642  97.29264
## 12 7 186  61.433110 124.56689
## 13 2  43 -13.381271  56.38127
## 14 6 126  15.070234 110.92977
```

6) 잔차와 독립변수 X 에 대한 산점도를 그려본다.

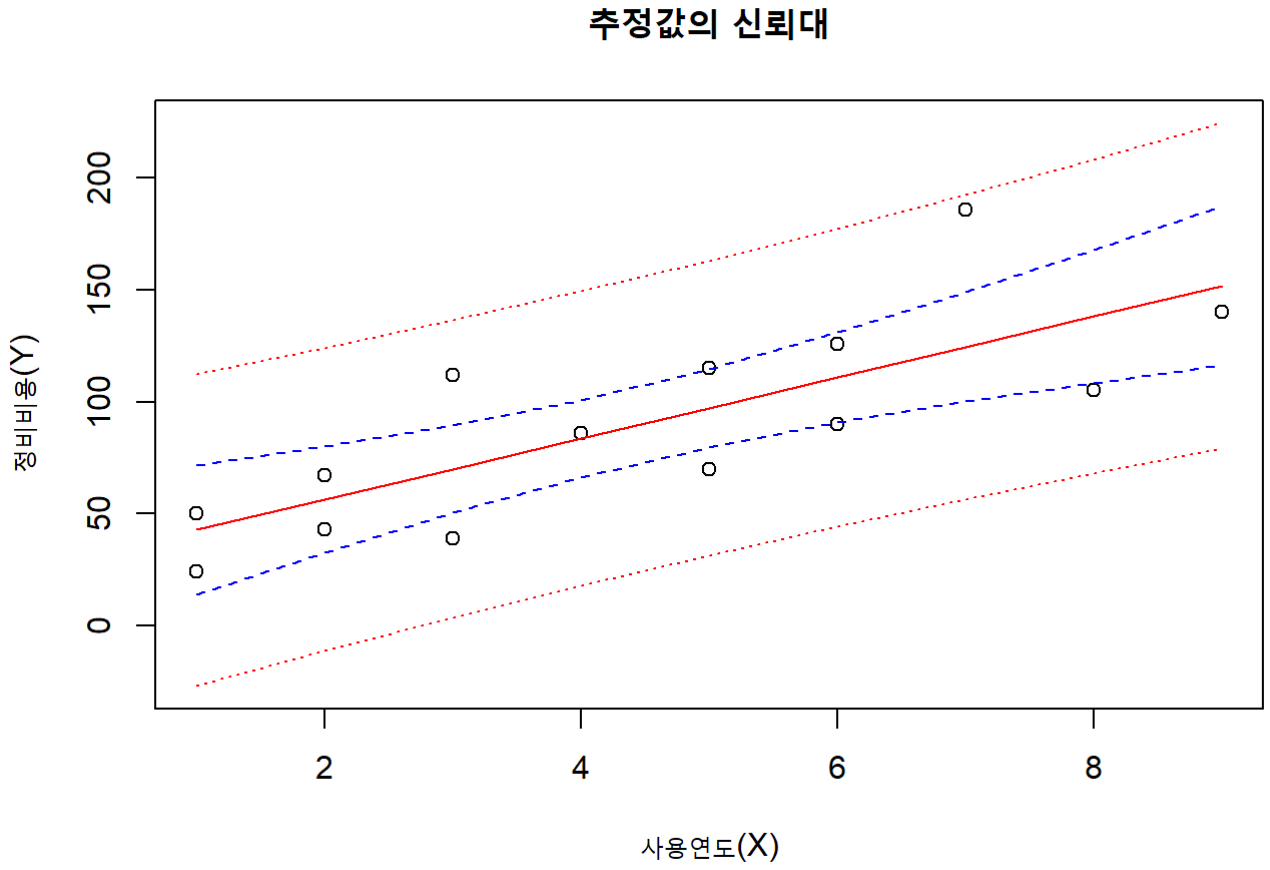
```
# 잔차를 독립변수 X에 대해 산점도를 그려본다.
plot(X, resid(machine.lm), pch=19, main="잔차와 X 산점도", xlab="X", ylab="잔차" )
for( i in 1:length(machine.lm$resid))
{
  if(abs(machine.lm$resid[i]) > 40)
    text(X[i]+0.2, machine.lm$resid[i], as.character(i))
}
# 잔차가 0인 라인 타입 2번 선을 그린다.

abline(h=0, lty=2)
```



잔차 중 10번과 12번의 측정값의 잔차는 다른 측정값이 비해 잔차의 크기가 큰 값을 갖고 있지만 0을 중심으로 일정한 범위내에 있기 때문에 회귀의 기본 가정을 만족한다고 할 수 있다.

7) 추정값의 신뢰대 그리기

```
# X의 범위와 갯수에 맞게 생성 range(X) -> 1 ~ 9
machine.frame = data.frame(X=1:9)
pc = predict(machine.lm, int="c", newdata=machine.frame)
pp = predict(machine.lm, int="p", newdata=machine.frame)
plot(X, Y, ylim=range(pp), main="추정값의 신뢰대", xlab="사용연도(X)", ylab="정비비용(Y)")
matlines(machine.frame$X, pc, lty=c(1,2,2), col="BLUE")
matlines(machine.frame$X, pp, lty=c(1,3,3), col="RED")
```



2번. 연습문제 2장 3번(p. 87) 자료를 이용하여 2.8 분석사례와 같이 분석하고, 설명하시오.

1) 교재의 데이터 입력

```
Y = c(2.8, 3.9, 3.9, 4.4, 3.1, 3.1, 3.5, 3.6, 3.0, 3.3)
X1 = c(10, 24, 25, 28, 15, 18, 22, 22, 12, 15)
X2 = c(27, 26, 28, 26, 30, 24, 27, 25, 27, 25)
X3 = c(64, 72, 80, 88, 81, 45, 46, 69, 54, 39)
# 마지막 데이터 3개를 출력해서 갯수가 맞게 입력된것인지 확인
tail(cbind(Y, X1, X2, X3), 3)
```

```
##          Y X1 X2 X3
##  [8,] 3.6 22 25 69
##  [9,] 3.0 12 27 54
## [10,] 3.3 15 25 39
```

```
res = as.data.frame(cbind(Y, X1, X2, X3))
```

2) 기술통계량 및 상관계수 보기

```
# 자료 요약 보기
summary(res)
```

```
##        Y               X1              X2              X3
##  Min.   :2.800   Min.   :10.0   Min.   :24.00   Min.   :39.0
##  1st Qu.:3.100   1st Qu.:15.0   1st Qu.:25.25   1st Qu.:48.0
##  Median :3.400   Median :20.0   Median :26.50   Median :66.5
##  Mean   :3.460   Mean   :19.1   Mean   :26.50   Mean   :63.8
##  3rd Qu.:3.825   3rd Qu.:23.5   3rd Qu.:27.00   3rd Qu.:78.0
##  Max.   :4.400   Max.   :28.0   Max.   :30.00   Max.   :88.0
```

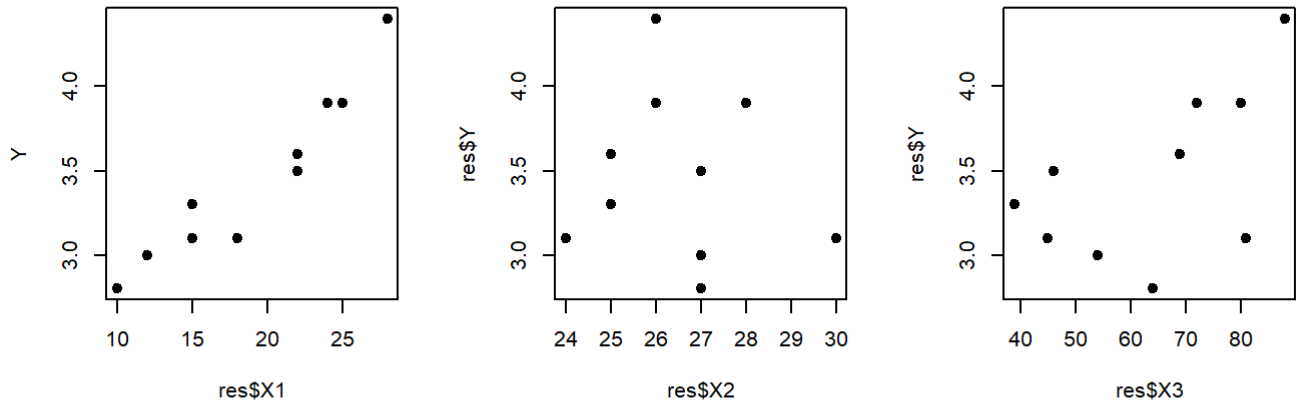
```
# 상관계수 보기
cor(res)
```

```
##             Y         X1         X2        X3
## Y   1.0000000  0.9493267 -0.1172336 0.5557581
## X1  0.9493267  1.0000000 -0.1567513 0.4721047
## X2 -0.1172336 -0.1567513  1.0000000 0.4971997
## X3  0.5557581  0.4721047  0.4971997 1.0000000
```

상관계수를 보면 반응변수 Y에 상관계수가 높은 설명변수는 X1으로 0.95 값을 가지고 다른 설명변수 X2와 X3는는 반응변수 Y와의 상관계수는 그리 높지 않게 나타나고 있지 않다.

3) 산점도 그리기

```
par(mfrow=c(1,3), pty="s")
plot(res$X1, Y, pch = 19)
plot(res$X2, res$Y, pch = 19)
plot(res$X3, res$Y, pch = 19)
```



반응변수 Y와 각 설명변수 X1, X2, X3의 산점도를 각각 그려보면 Y와 X1의 분포가 선형관계가 강하게 있고 X3도 어느정도의 선형관계가 예상된다.

4) 회귀모형 적합하기

```
res.lm = lm(Y ~ X1+X2+X3, data=res)
summary(res.lm)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3, data = res)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.23490 -0.07744 -0.02166  0.08840  0.23442
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.409213   1.125954   2.140  0.07618 .
## X1           0.069788   0.012640   5.521  0.00149 **
## X2          -0.024767   0.044830  -0.552  0.60060
## X3           0.005864   0.005052   1.161  0.28978
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.172 on 6 degrees of freedom
## Multiple R-squared:  0.9202, Adjusted R-squared:  0.8803
## F-statistic: 23.05 on 3 and 6 DF,  p-value: 0.001079
```

5-1) 추정된 회귀방정식

```
# 추정된 회귀방정식
str = paste0("분산분석표 이용: hat_", names(res.lm$model)[1], " = ")
str = paste0( str, round(coef(res.lm)["(Intercept)"], 3))
for(i in 2:length(coef(res.lm)))
{
  str = paste0( str, " + ", round(coef(res.lm)[names(res.lm$model)[i]], 3),
                "*", names(res.lm$model)[i])
}
print(str)
```

```
## [1] "분산분석표 이용: hat_Y = 2.409 + 0.07*X1 + -0.025*X2 + 0.006*X3"
```

5-2) 추정된 회귀방정식

```
# 추정된 회귀방정식을 최소제곱법으로 β를 행렬방정식으로 계산
mX=as.matrix(cbind(1, X1, X2, X3))
mY=as.matrix(Y)
mβ=solve(t(mX)%*%mX)%*%t(mX)%*%mY

str = paste0(" 최소제곱법 이용 : hat_Y = ", round(mβ[1], 3))
for(i in 2:length(rownames(mβ)))
{
  str = paste0( str, " + ", round(mβ[i], 3), "*", rownames(mβ)[i])
}
print(str)
```

```
## [1] " 최소제곱법 이용 : hat_Y = 2.409 + 0.07*X1 + -0.025*X2 + 0.006*X3"
```

```
paste0("결정계수 R.squared =", round(summary(res.lm)$r.squared, 3), "으로",
round(summary(res.lm)$r.squared*100,1), "% 설명력이 있다.")
```

```
## [1] "결정계수 R.squared =0.92으로92% 설명력이 있다."
```

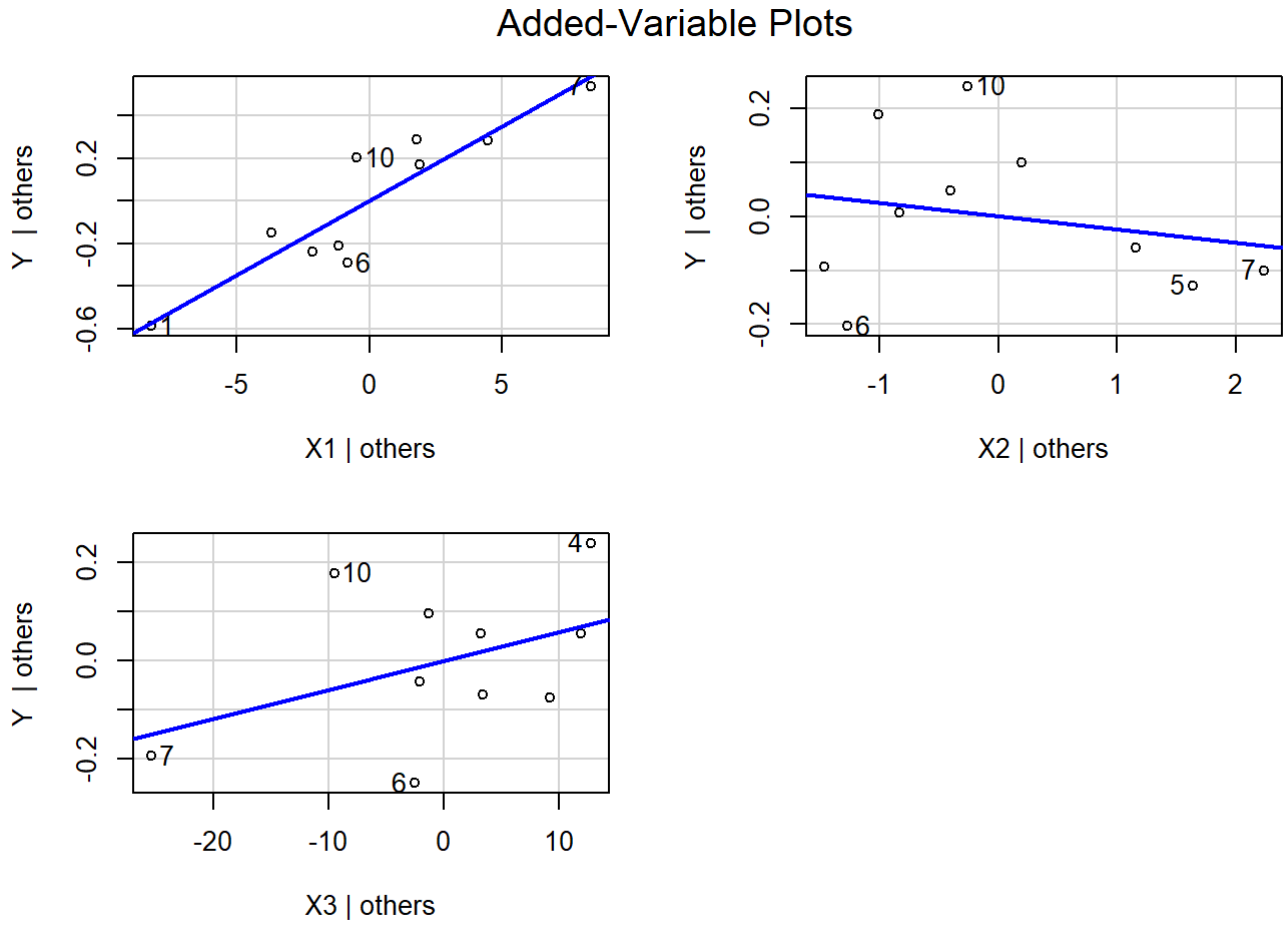
설명변수 X1의 p-value = 0.00149, X2의 p-value = 0.60060, X3의 p-value = 0.28978이고 유의 수준 $\alpha = 0.05$ 를 기준으로 할 때, X1은 Y를 설명하는데 유의하지만 변수 X2와 X3는 유의하지 못하다.

6) 추가변수그림 그려보기

```
library(car)
```

```
## Loading required package: carData
```

```
avPlots(res.lm)
```



추가 그림을 그려보면 X1이 Y을 설명하는데 유의한 변수임을 알 수 있다.

7-1) 분산분석표 구하기

```
anova(res.lm)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## X1         1 2.00432 2.00432 67.7386 0.0001738 ***
## X2         1 0.00227 0.00227  0.0768 0.7909535
## X3         1 0.03988 0.03988  1.3477 0.2897756
## Residuals  6 0.17753 0.02959
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

7-2) 분산분석표 정리하여 만들기기

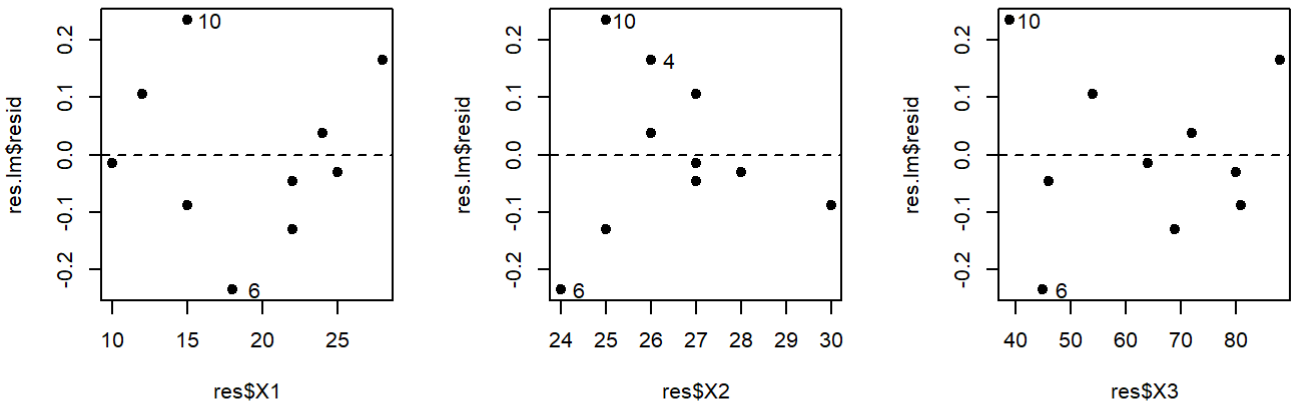
```
# 분산분석표를 정리하기 위한 행렬 생성
AVT = matrix(c(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0), nrow=3)
colnames(AVT)=colnames(anova(res.lm))
rownames(AVT)=c("Regression", "Residuals", "Total")

# 회귀계수들은 합한 후 계산하기 때문에 자유도와 제곱합을 합해서 저장한다
for(i in 1:(length(rownames(summary(res.lm)$coefficients))-1))
{
  AVT["Regression", "Df"]=AVT["Regression", "Df"]+as.double(anova(res.lm)[i,"Df"])
  AVT["Regression", "Sum Sq"]=
  AVT["Regression", "Sum Sq"]+as.double(anova(res.lm)[i,"Sum Sq"])
}
# 분산분석 데이터에서 잔차에 필요한 정보를 행렬에 넣는다.
AVT["Residuals", "Df"]=as.double(anova(res.lm)["Residuals","Df"])
AVT["Residuals", "Sum Sq"]=as.double(anova(res.lm)["Residuals","Sum Sq"])
# 전체 자유도와 평균제곱합을 계산한다.
AVT["Total", "Df"]=AVT["Regression", "Df"]+AVT["Residuals", "Df"]
AVT["Total", "Sum Sq"]=AVT["Regression", "Sum Sq"]+AVT["Residuals", "Sum Sq"]
AVT["Regression", "Mean Sq"] = AVT["Regression", "Sum Sq"] / AVT["Regression", "Df"]
AVT["Residuals", "Mean Sq"] = AVT["Residuals", "Sum Sq"] / AVT["Residuals", "Df"]
AVT["Regression", "F value"] =
  round(AVT["Regression", "Mean Sq"] / AVT["Residuals", "Mean Sq"], 1)
AVT["Regression", "Pr(>F)"] =
  round( (1 - pf( AVT["Regression", "F value"],
                  AVT["Regression", "Df"],
                  AVT["Residuals", "Df"])), 6)
print(AVT)
```

```
##            Df    Sum Sq    Mean Sq F value   Pr(>F)
## Regression  3 2.0464661 0.68215537    23.1 0.001073
## Residuals   6 0.1775339 0.02958898     0.0 0.000000
## Total       9 2.2240000 0.00000000     0.0 0.000000
```

anova(res.lm)로 표시된 분산분석표를 보면 설명변수의 순서와 상관없이 잔차의 제곱합 및 잔차제곱의 평균값은 변동이 없으나 회귀계수는 설명변수의 순서에 따라 값이 변경된다. 하지만 정리된 분산분석표에서는 회귀계수들의 값들이 합하여 표시되기 때문에 영향이 없다.

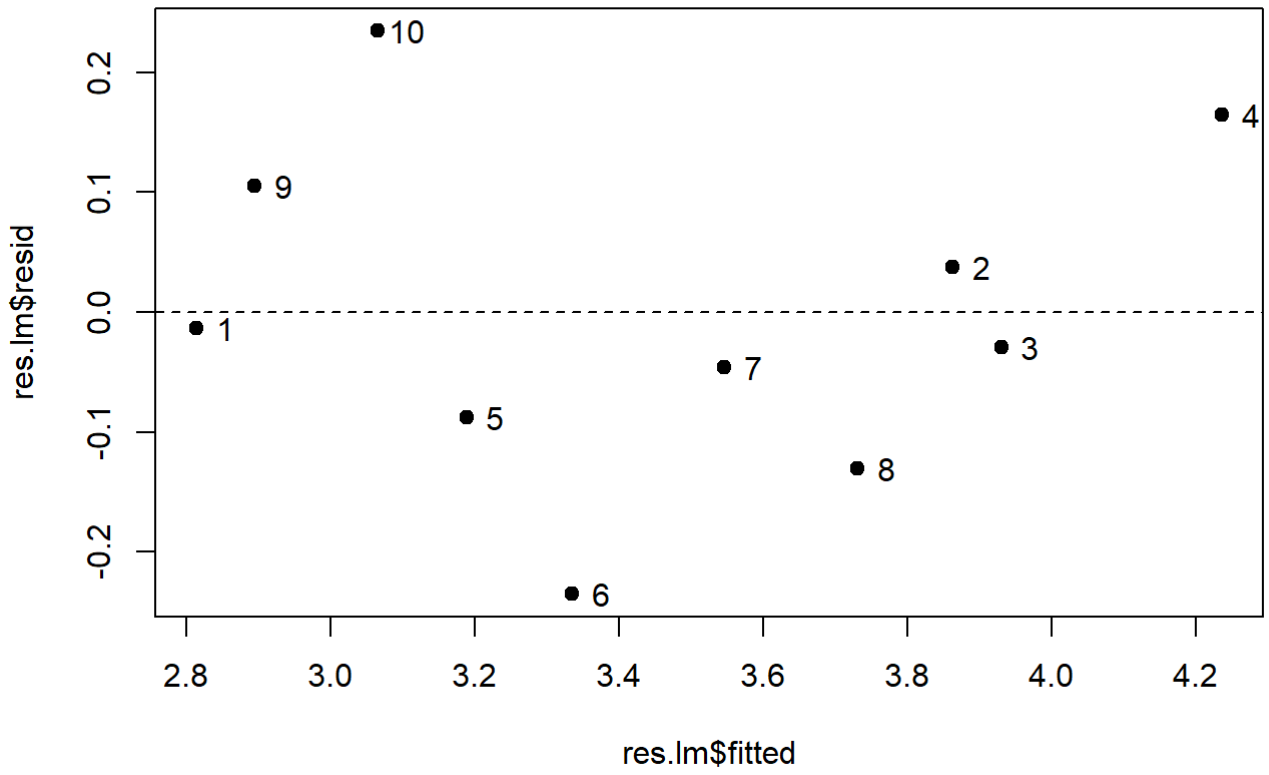
8-1) 잔차 산점도(독립변수와 잔차)

```
# 독립변수와 잔차의 산점도를 그리고 각 케이스의 번호 출력
par(mfrow=c(1,3), pty="s")
plot(res$X1, res.lm$resid, pch=19)
for( i in 1:length(res.lm$resid))
{
  if(res.lm$resid[i] > quantile(res.lm$resid, c(0.1, 0.8))[2] ||
     res.lm$resid[i] < quantile(res.lm$resid, c(0.1, 0.8))[1])
    text(res$X1[i]+1.5, res.lm$resid[i], as.character(i))
}
abline(h=0, lty=2)

plot(res$X2, res.lm$resid, pch=19)
for( i in 1:length(res.lm$resid))
{
  if(res.lm$resid[i] > quantile(res.lm$resid, c(0.1, 0.8))[2] ||
     res.lm$resid[i] < quantile(res.lm$resid, c(0.1, 0.8))[1])
    text(res$X2[i]+0.4, res.lm$resid[i], as.character(i))
}
abline(h=0, lty=2)

plot(res$X3, res.lm$resid, pch=19)
for( i in 1:length(res.lm$resid))
{
  if(res.lm$resid[i] > quantile(res.lm$resid, c(0.1, 0.8))[2] ||
     res.lm$resid[i] < quantile(res.lm$resid, c(0.1, 0.8))[1])
    text(res$X3[i]+3.5, res.lm$resid[i], as.character(i))
}
abline(h=0, lty=2)
```



8-2) 잔차 산점도 -2 : (추정값, 잔차)

```
plot(res.lm$fitted, res.lm$resid, pch=19)
for( i in 1:length(res.lm$resid))
{
    text(res.lm$fitted[i]+0.04, res.lm$resid[i], as.character(i))
}
abline(h=0, lty=2)
```



추정값의 잔차는 일정범위 안에 들어가기 때문에 회귀에 대한 기본가정을 만족한다고 할수 있다.

2023년 회귀모형 출석수업 과제물 끝
