출석수업 과제물(평가결과물) 표지(온라인제출용)

교과목명 : 회귀모형

학 번 : 202135-368864

성 명 : 홍 원 표

강 의 실 : 경기(성남) 지역대학 호

연 락 처 : 010-5343-4341

- 이하 과제 작성

04 월 23 일 회귀모형 출석수업 과제물

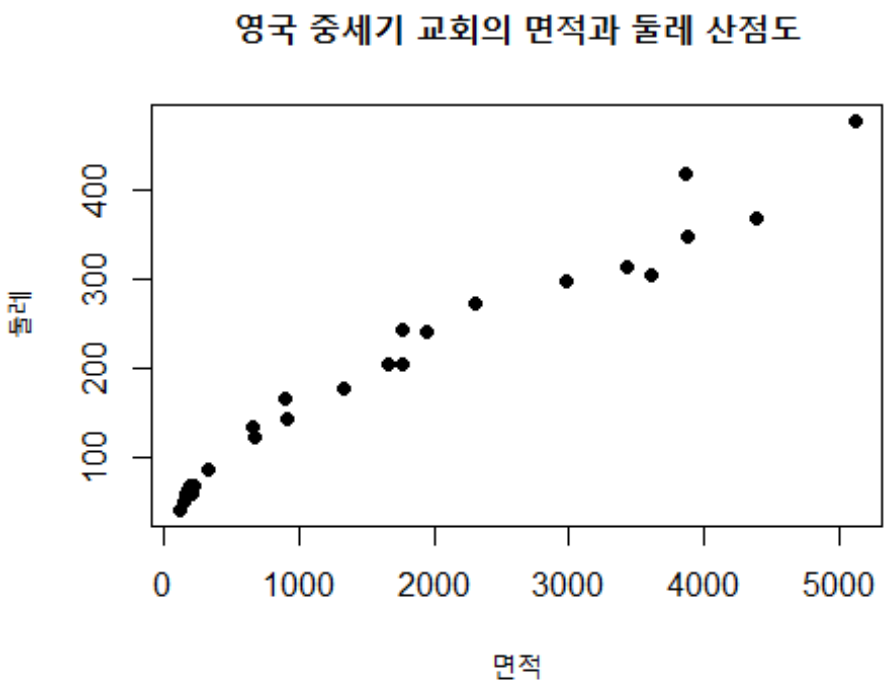
1 번.연습문제 6 장 4 번 (p. 187) 교회 자료를 이용하여 교재 1.7 분석사례와 같이 분석하고, 설명하시오

(단, 둘레를 독립변수로, 면적의 제곱근을 반응변수로 하기 바람)

```
# 데이터 로드
churches = read.csv("./data/p187.csv", header = T)
head(churches, 3)
```

```
##   no size area
## 1  1  348 3883
## 2  2  369 4392
## 3  3  143  914
```

```
# 오브젝트의 변수를 바로 사용하기 위해 오브젝트 등록
attach(churches)
# 산점도를 그린다.
plot(area, size, pch=19, main="영국 중세기 교회의 면적과 둘레 산점도", xlab="면적",
 ylab="둘레")
```




```
# 회귀모형 적합
churches.lm = lm(area~size, data = churches)
summary(churches.lm)

##
## Call:
## lm(formula = area ~ size, data = churches)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -524.19 -139.99   15.64  165.37  606.19
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -674.5399   115.4252  -5.844  5.9e-06 ***
## size          12.0877     0.4947  24.434  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 307.8 on 23 degrees of freedom
## Multiple R-squared:  0.9629, Adjusted R-squared:  0.9613
## F-statistic:   597 on 1 and 23 DF,  p-value: < 2.2e-16
```

회귀적합 결과에서 회귀계수의 추정값은 절편 $b_0 = -674.5399$이고 기울기 $b_1 = 12.0877$의 단순회귀방정식은 $\widehat{area} = -674.5399 + 12.0877 \times size$ 가 된다. 기울기 $t -$

$값 = 24.434$이고 $p - 값 = 2 \times 10^{-16}$이 매우 작으므로 $H_0: \beta_1 = 0$이라는 귀무가설을 기각한다.
결정계수 $R^2 = 0.9629$로서 총변동 중에서 96.29%가 회귀방정식으로 설명되는 회귀변동이 차지하고 있다는 것을 나타낸다. $F - 값 = 597$이고, 이에 대한 $P - 값 = 2.2 \times 10^{-16}$으로서 적합된 회귀직선이 유의하다는 것을 알 수 있다.

```
# 분산분석표 구하기
anova(churches.lm)
```

```
## Analysis of Variance Table
##
## Response: area
##           Df   Sum Sq  Mean Sq F value    Pr(>F)
## size       1 56578697 56578697  597.04 < 2.2e-16 ***
## Residuals 23  2179607    94766
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

분산분석표에서 보면 검정통계량 $F_0 = 597.04$이고 이에 대한 $p - 값 = 2.2 \times 10^{-16}$이 매우 작으므로 적합된 회귀선이 유의하다는 것을 알 수 있다.

```
# 잔차 및 추정값 보기
# 회귀모형 적합 결과(churches.lm)의 변수 확인
names(churches.lm)
```

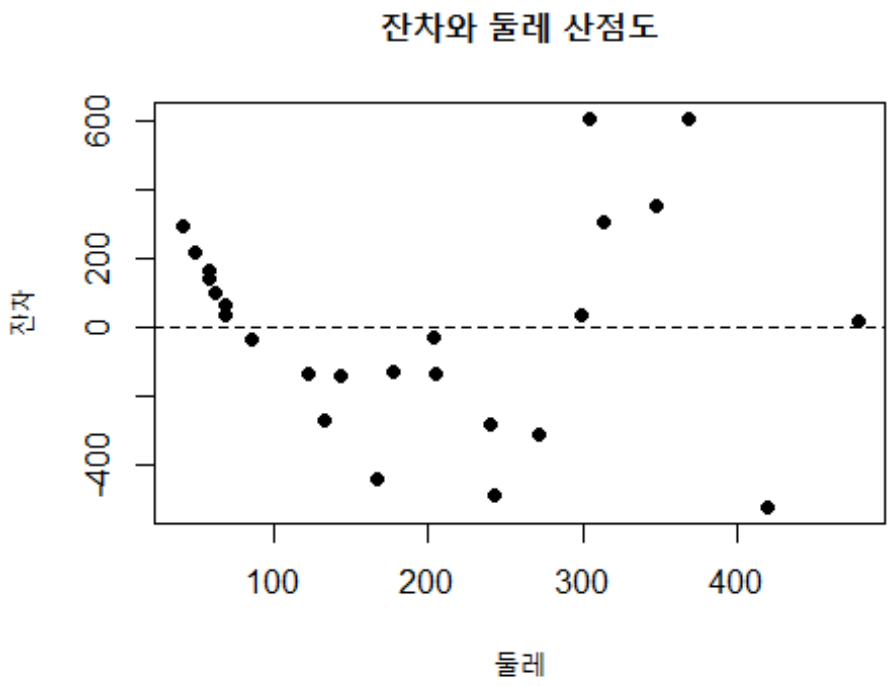
```
##  [1] "coefficients"  "residuals"     "effects"       "rank"
##  [5] "fitted.values" "assign"        "qr"            "df.residual"
##  [9] "xlevels"       "call"          "terms"         "model"
```

```
# churches 와 잔차와 추정값을 합쳐서 보기
cbind(churches, churches.lm$resid, churches.lm$fitted)
```

```
##    no size area churches.lm$resid churches.lm$fitted
## 1   1  348 3883         351.03544         3531.96456
## 2   2  369 4392         606.19466         3785.80534
## 3   3  143  914        -139.99500         1053.99500
## 4   4  205 1666        -137.42969         1803.42969
## 5   5  305 3616         603.80467         3012.19533
## 6   6  419 3866        -524.18816         4390.18816
## 7   7  243 1774        -488.76064         2262.76064
## 8   8  240 1946        -280.49767         2226.49767
## 9   9  272 2300        -313.30267         2613.30267
## 10 10  299 2975          35.33061         2939.66939
```

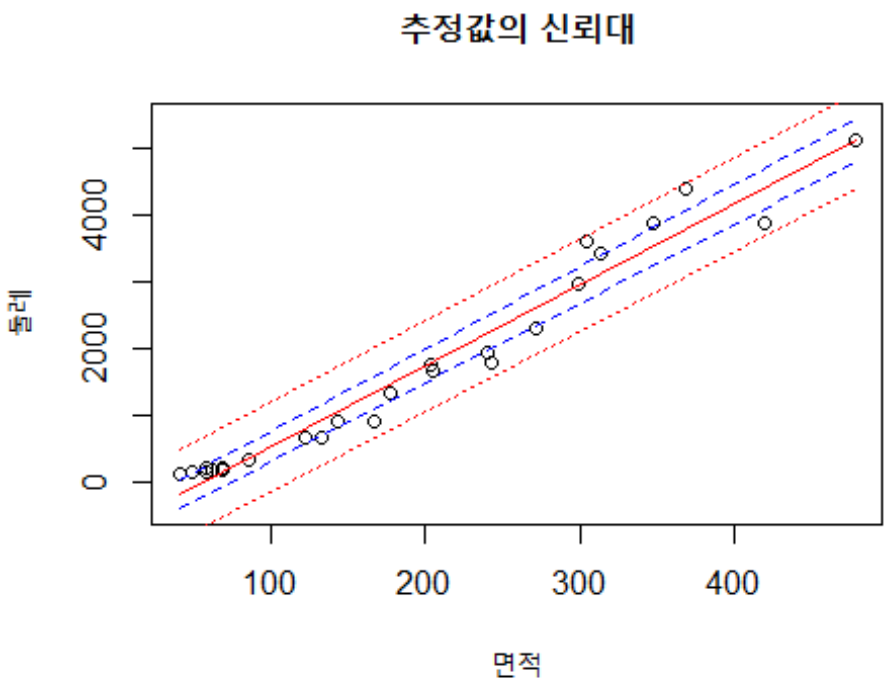
```
## 11 11  478 5119          15.64011         5103.35989
## 12 12  133  660        -273.11844          933.11844
## 13 13  167  904        -440.09875         1344.09875
## 14 14  314 3427         306.01576         3120.98424
## 15 15  204 1761         -30.34204         1791.34204
## 16 16  177 1337        -127.97532         1464.97532
## 17 17   59  204         165.36814           38.63186
## 18 18   69  222          62.49157          159.50843
## 19 19   50  146         216.15704          -70.15704
## 20 20   69  192          32.49157          159.50843
## 21 21   63  186          99.01751           86.98249
## 22 22   58  169         142.45579           26.54421
## 23 23   86  331         -33.99859          364.99859
## 24 24   41  113         291.94595         -178.94595
## 25 25  123  674        -138.24187          812.24187
```

```r
# 잔차를 독립변수 size 에 대해 산점도를 그려본다.
plot(size, churches.lm$resid, pch=19, main="잔차와 둘레 산점도", xlab="둘레",
ylab="잔차")

# 잔차가 0 인 라인 타입 2 번 선을 그린다.
abline(h=0, lty=2)
```



잔차는 0 을 중심으로 일정한 범위 내에 있으므로 회귀에 대한 기본 가정을 만족한다고 할 수 있으나, X 가 증가함에 따라 곡선 관계를 보여주고 있다. 따라서 2 차곡선회귀식 $\hat{Y} = b_0 + b_1 X + b_2 X^2$을 구해보는 것도 의미가 있으리라고 생각된다.

```
# 추정값의 신뢰대 그리기
churches.frame = data.frame(size=range(churches$size))
pc = predict(churches.lm, int="c", newdata=churches.frame)
pp = predict(churches.lm, int="p", newdata=churches.frame)
plot(churches$size, churches$area, ylim = range(churches$area, pc), main="추정값
의 신뢰대", xlab="면적", ylab="둘레")
matlines(churches.frame$size, pc, lty=c(1,2,2), col="BLUE")
matlines(churches.frame$size, pp, lty=c(1,3,3), col="RED")
```



2 번. 연습문제 3 장 1 번(p. 114) 자료를 이용하여 교재 2.8 분석사례와 같이 분석하고, 설명하시오

(변수 EVAP 를 반응변수로, 나머지 변수를 독립변수로 하기 바람)

MAXST: 토양 내 최고온도

MINST: 토양 내 최저온도

AVST: 토양 내 평균온도

MAXAT: 최고기온

MINAT: 최저기온

AVAT: 평균기온

EVAP: 증발되는 수분의 양

```
# 데이터 로드
climate = read.csv("./data/p114.csv", header = T)
head(climate, 3)
```

```
##   DAY MAXST MINST AVST MAXAT MINAT AVAT EVAP
## 1   6    84    65  147    85    59  151   30
## 2   7    84    65  149    86    61  159   34
## 3   8    79    66  142    83    64  152   33
```

```
# 첫번째 열인 날짜를 제외하고 각 변수들의 기술통계량을 본다.
summary(climate[,-1])
```

```
##      MAXST           MINST           AVST           MAXAT      
##  Min.   :73.00   Min.   :65.00   Min.   :131.0   Min.   :77.00  
##  1st Qu.:81.00   1st Qu.:67.00   1st Qu.:147.0   1st Qu.:84.00  
##  Median :84.00   Median :69.00   Median :161.0   Median :88.00  
##  Mean   :83.92   Mean   :69.12   Mean   :160.6   Mean   :87.72  
##  3rd Qu.:88.00   3rd Qu.:72.00   3rd Qu.:171.0   3rd Qu.:92.00  
##  Max.   :93.00   Max.   :74.00   Max.   :188.0   Max.   :94.00  
##      MINAT           AVAT           EVAP      
##  Min.   :59.00   Min.   :147.0   Min.   : 4.00  
##  1st Qu.:67.00   1st Qu.:159.0   1st Qu.:23.00  
##  Median :69.00   Median :177.0   Median :33.00  
##  Mean   :69.12   Mean   :180.8   Mean   :31.28  
##  3rd Qu.:72.00   3rd Qu.:201.0   3rd Qu.:43.00  
##  Max.   :76.00   Max.   :211.0   Max.   :54.00
```

```
# 첫번째 열인 날짜를 제외하고 각 변수들의 상관계수를 살펴 본다.
cor(climate[,-1])
```

```
##           MAXST     MINST      AVST     MAXAT     MINAT      AVAT      EVAP
## MAXST 1.0000000 0.7469553 0.9486608 0.9268580 0.5048854 0.8250186 0.8933048
## MINST 0.7469553 1.0000000 0.8706342 0.7842173 0.8470404 0.8544593 0.6059399
## AVST  0.9486608 0.8706342 1.0000000 0.9282693 0.6834957 0.8928071 0.8173403
## MAXAT 0.9268580 0.7842173 0.9282693 1.0000000 0.6256655 0.9094757 0.8509974
```

```
## MINAT 0.5048854 0.8470404 0.6834957 0.6256655 1.0000000 0.8307017 0.4544024
## AVAT  0.8250186 0.8544593 0.8928071 0.9094757 0.8307017 1.0000000 0.7675541
## EVAP  0.8933048 0.6059399 0.8173403 0.8509974 0.4544024 0.7675541 1.0000000
```

종속변수 '증발되는 수분의 양'은 *최고온도*와 *토양 내 최고온도*와 *평균온도* 독립변수 들과 상관 계수가 높다는 것도 알 수 있다.

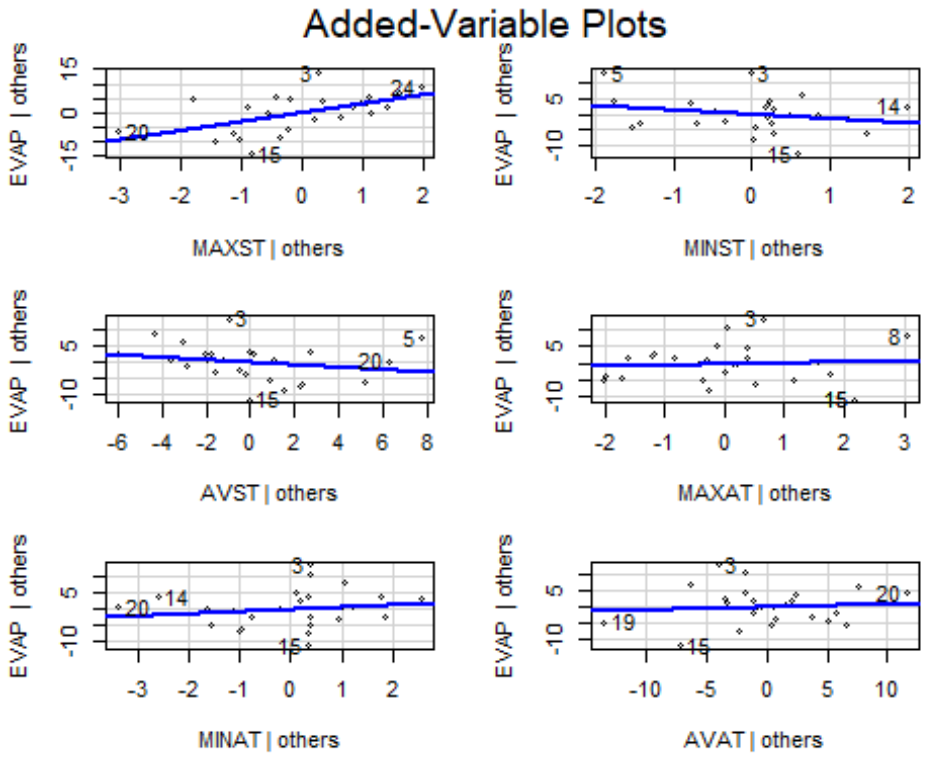
```
# 회귀모형 적합하기
climate.lm = lm (EVAP~MAXST+MINST+AVST+MAXAT+MINAT+AVAT, data = climate)
summary(climate.lm)
```

```
##
## Call:
## lm(formula = EVAP ~ MAXST + MINST + AVST + MAXAT + MINAT + AVAT,
##     data = climate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.6796  -3.9117   0.0074   2.8489  13.0390
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -164.85909   92.38388  -1.785   0.0912 .
## MAXST          3.13716    1.13321   2.768   0.0127 *
## MINST         -1.47499    1.38625  -1.064   0.3014
## AVST          -0.40671    0.41303  -0.985   0.3378
## MAXAT          0.40732    1.03947   0.392   0.6998
## MINAT          0.70419    0.95919   0.734   0.4723
## AVAT           0.08692    0.25522   0.341   0.7374
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.523 on 18 degrees of freedom
## Multiple R-squared:  0.839,  Adjusted R-squared:  0.7854
## F-statistic: 15.64 on 6 and 18 DF,  p-value: 2.917e-06
```

추정된 회귀방정식은 $\hat{Y} = -164.86 + 3.14 \times MAXST - 1.47 \times MINST - 4.1 \times AVST + 0.41 \times MAXAT + 0.70 \times MINAT + 0.09 \times AVAT$이고, 이 모형에 대한 결정계수 $R^2 = 0.839$로서 중회귀모형이 종속변수 EVAP 의 총변동을 83.9% 정도 설명하고 있다는 것을 나타낸다. 또한 추정값의 표준오차 $\sqrt{MSE} = 6.523$로서 σ의 추정치가 6.523임을 알 수 있다.

```
# 추가 변수 그림
library(car)
```

```
## 필요한 패키지를 로딩중입니다: carData

avPlots(climate.lm)
```



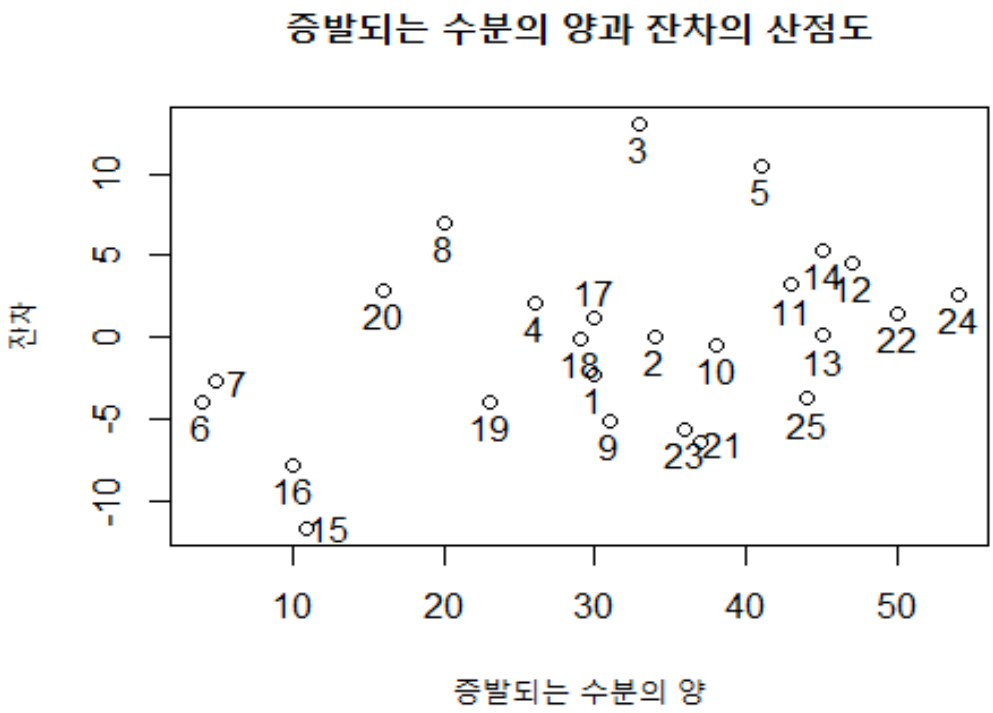

```
# 분산분석표
anova(climate.lm)

## Analysis of Variance Table
##
## Response: EVAP
##           Df Sum Sq Mean Sq F value    Pr(>F)
## MAXST      1 3797.7  3797.7 89.2422 2.128e-08 ***
## MINST      1   40.5    40.5  0.9512    0.3423
## AVST       1    6.8     6.8  0.1601    0.6938
## MAXAT      1   61.3    61.3  1.4394    0.2458
## MINAT      1   81.9    81.9  1.9243    0.1823
## AVAT       1    4.9     4.9  0.1160    0.7374
## Residuals 18  766.0    42.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

분산분석표에 의한 F-검정

요인	자유도	제곱합	평균제곱	F_0
회귀	6	3993.05	665.51	15.64
잔차	18	766.0	42.56	
계	24	4759.04		

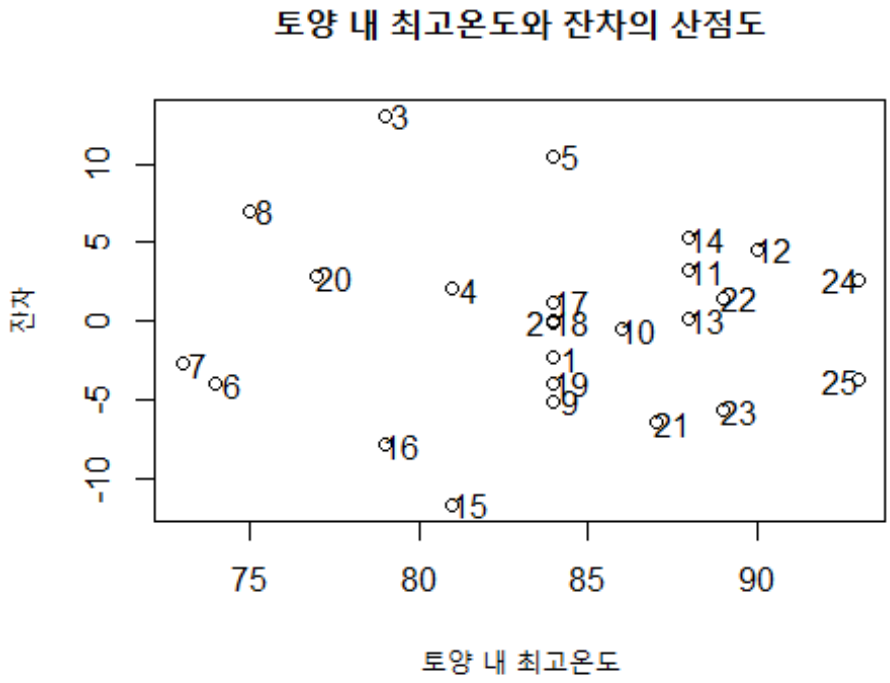
여기서 $\textit{회귀제곱합} = 3797.7 + 40.5 + 6.8 + 61.3 + 81.9 + 4.9 + 766.0 = 4759.04$이다. $F - \textit{값} = 15.64$에 대한 유의 확률이 0.00000292 로 매우 작아서 중회귀모형이 매우 유의함을 알 수 있다. 또한 오차분산 σ^2의 추정치 $MSE = 42.6$임을 알 수 있다.

```r
plot(climate$EVAP, climate.lm$resid, main="증발되는 수분의 양과 잔차의 산점도", xlab="증발되는 수분의 양", ylab="잔차")
for (i in 1:length(climate$EVAP))
{
  if ( i == 7 )
    text(climate$EVAP[i]+1.5, climate.lm$resid[i], as.character(i))
  else if (i == 15)
    text(climate$EVAP[i]+1.5, climate.lm$resid[i], as.character(i))
  else if ( i == 17)
    text(climate$EVAP[i], climate.lm$resid[i]+1.5, as.character(i))
  else if (i == 21)
    text(climate$EVAP[i]+1.5, climate.lm$resid[i], as.character(i))
  else
    text(climate$EVAP[i], climate.lm$resid[i]-1.5, as.character(i))
}
```



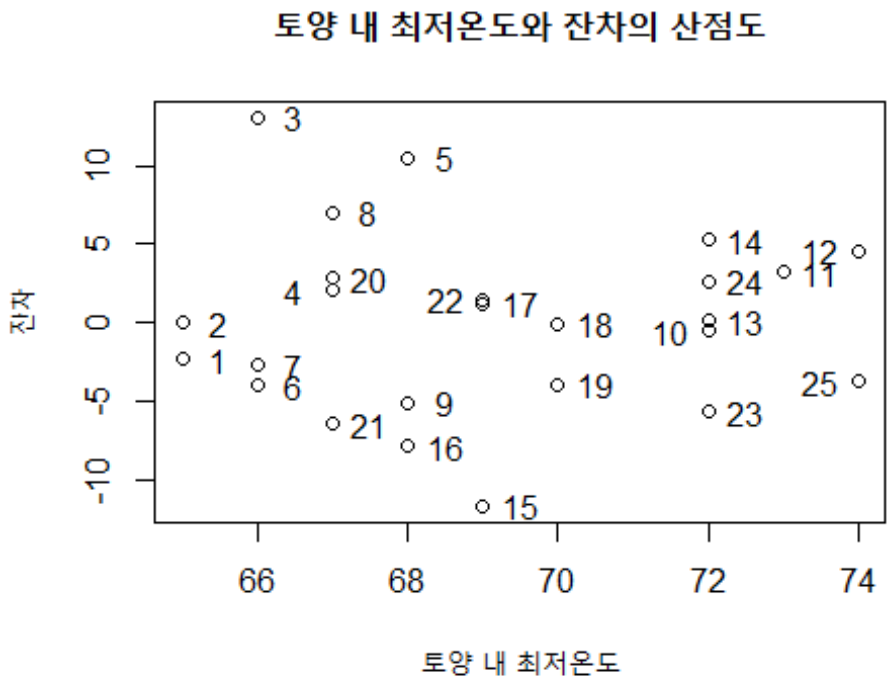
증발되는 수분의 양과 잔차의 산점도는 등분산성 모양으로 잔차들이 고르게 분포되어 있는것 같다.

```r
plot(climate$MAXST, climate.lm$resid, main="토양 내 최고온도와 잔차의 산점도", xlab="토양 내 최고온도", ylab="잔차")
for (i in 1:length(climate$MAXST))
{
  if (i==2 || i == 24 || i == 25)
    text(climate$MAXST[i]-0.5, climate.lm$resid[i], as.character(i))
  else
    text(climate$MAXST[i]+0.5, climate.lm$resid[i], as.character(i))
}
```



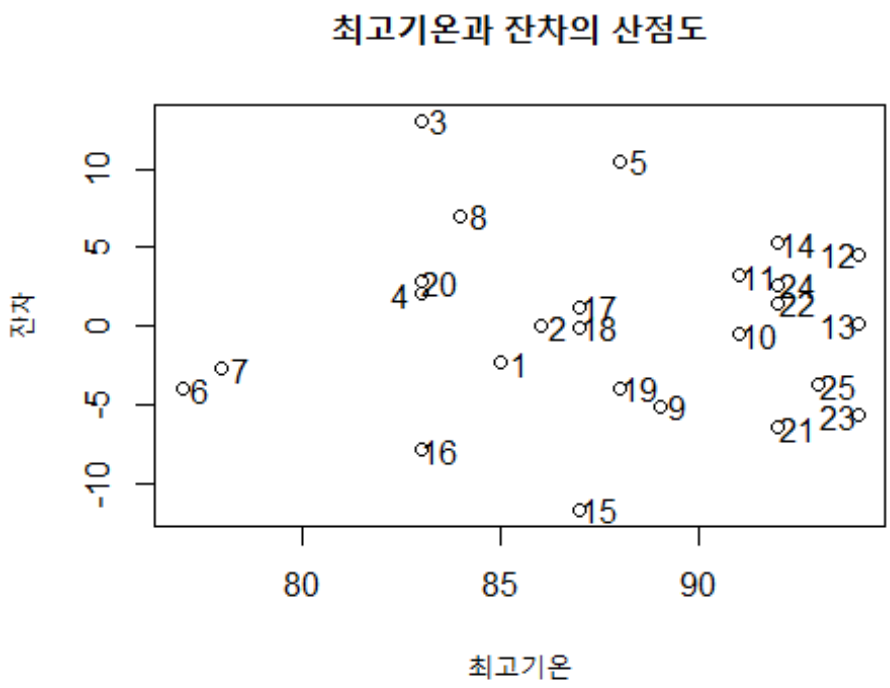
토양 내 최고온도와 잔차의 산점도는 등분산성 모양으로 잔차들이 고르게 분포되어 있는것 같다.

```r
plot(climate$MINST, climate.lm$resid, main="토양 내 최저온도와 잔차의 산점도", xlab="토양 내 최저온도", ylab="잔차")
for (i in 1:length(climate$MINST))
{
  if (i == 12 || i == 25 || i == 10 || i == 4 || i == 22)
    text(climate$MINST[i]-0.5, climate.lm$resid[i], as.character(i))
  else
    text(climate$MINST[i]+0.5, climate.lm$resid[i], as.character(i))

}
```



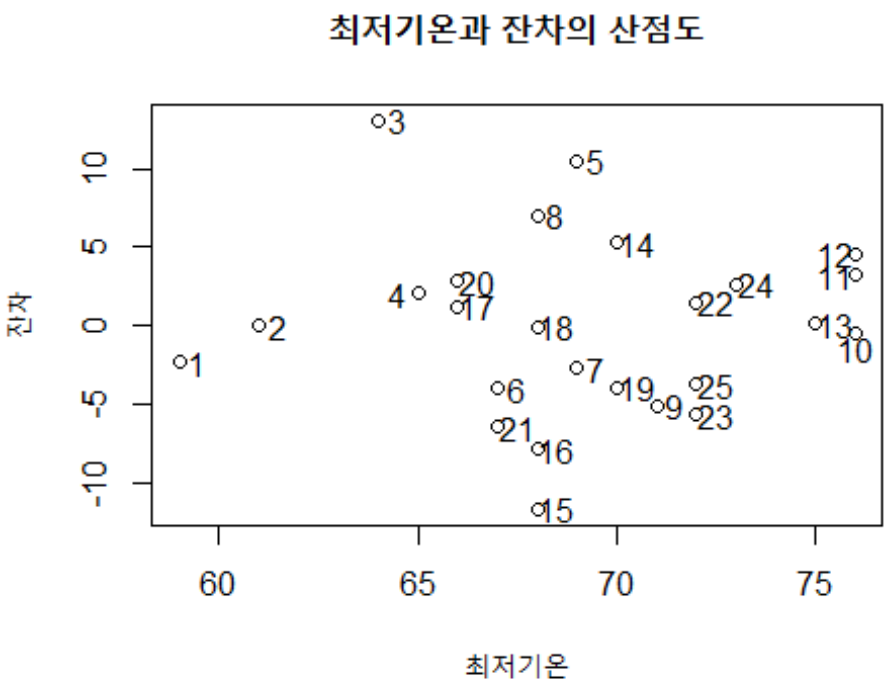
토양 내 최저온도와 잔차의 산점도는 등분산성 모양으로 잔차들이 고르게 분포되어 있는것 같다.

```
plot(climate$MAXAT, climate.lm$resid, main="최고기온과 잔차의 산점도", xlab="최고기온", ylab="잔차")
for (i in 1:length(climate$MAXAT))
{
  if (i == 12 || i == 13 || i == 23 || i == 4 )
    text(climate$MAXAT[i]-0.5, climate.lm$resid[i], as.character(i))
  else
    text(climate$MAXAT[i]+0.5, climate.lm$resid[i], as.character(i))
}
```



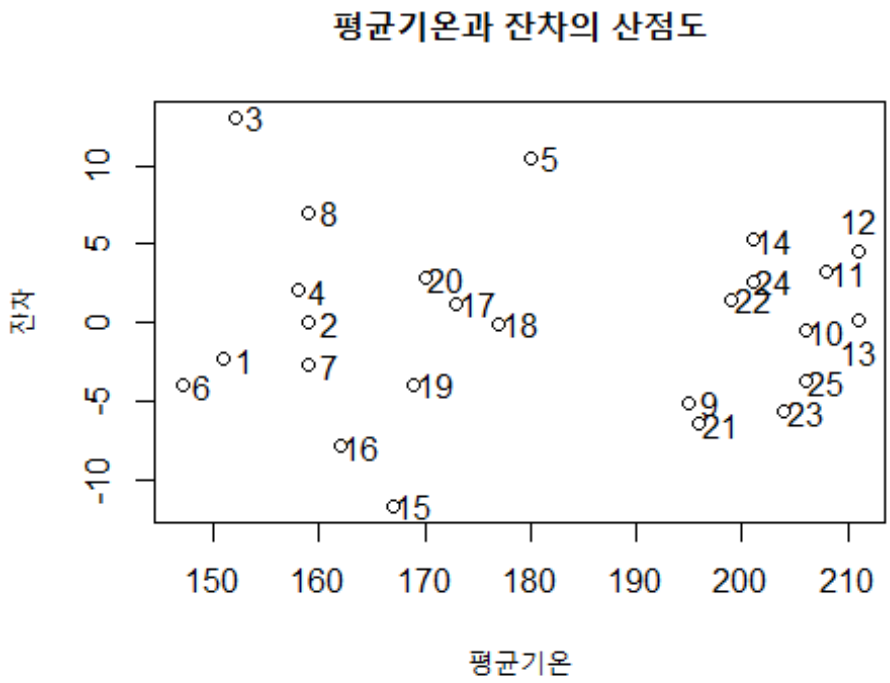
최고기온과 잔차의 산점도는 3 번과 15 번케이스를 제외하면 잔차들의 분포가 이분산성 모양으로 볼수도 있겠지만 등분산성 모양에 더 가깝다고 볼수 있다.

```
plot(climate$MINAT, climate.lm$resid, main="최저기온과 잔차의 산점도", xlab="최저기온", ylab="잔차")
for (i in 1:length(climate$MINAT))
{
  if (i == 12 || i == 11 || i == 4 )
    text(climate$MINAT[i]-0.5, climate.lm$resid[i], as.character(i))

  else if(i==10)
      text(climate$MINAT[i], climate.lm$resid[i]-1.0,as.character(i))
  else
    text(climate$MINAT[i]+0.5, climate.lm$resid[i],as.character(i))

}
```



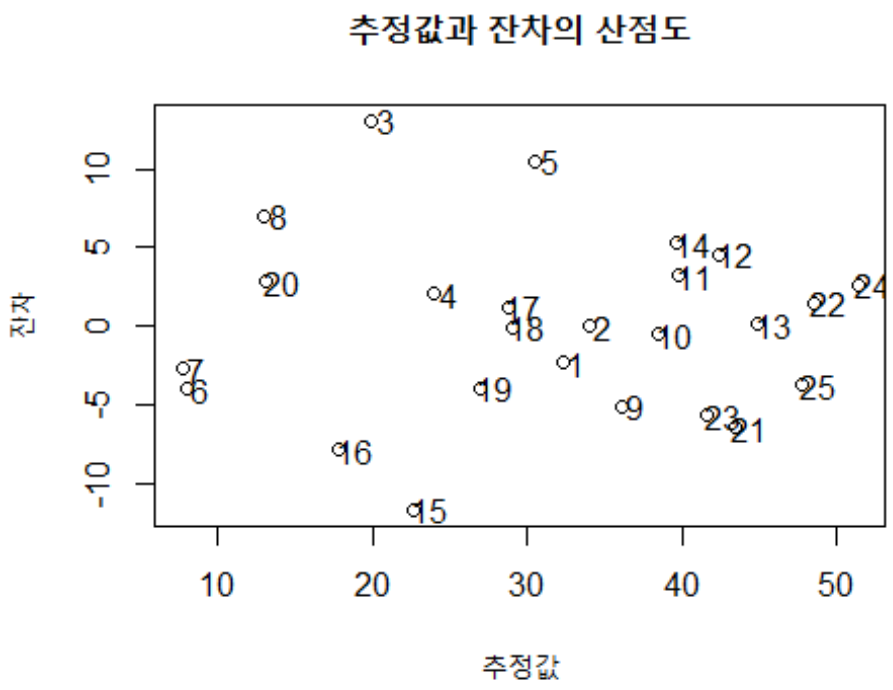
최저기온과 잔차의 산점도는 이분산성 모양으로 볼수도 있겠지만 등분산성 모양에 더 가깝다고 볼수 있다.

```r
plot(climate$AVAT, climate.lm$resid, main="평균기온과 잔차의 산점도", xlab="평균기온", ylab="잔차")
for (i in 1:length(climate$AVAT))
{
  if (i == 12 )
    text(climate$AVAT[i], climate.lm$resid[i]+2.0, as.character(i))
  else if(i==13)
    text(climate$AVAT[i], climate.lm$resid[i]-2.0, as.character(i))
  else
    text(climate$AVAT[i]+2.0, climate.lm$resid[i], as.character(i))
}
```



평균기온과 잔차의 산점도는 3 번과 15 번케이스를 제외하면 등분산성 모양으로 잔차들이 고르게 분포되어 있는것 같다.

```r
plot(climate.lm$fitted, climate.lm$resid, main="추정값과 잔차의 산점도", xlab="추정값", ylab="잔차")

for (i in 1:length(climate.lm$resid))
{
  if (i == 120 )
    text(climate.lm$fitted[i], climate.lm$resid[i], as.character(i))
  else if(i==130)
    text(climate.lm$fitted[i], climate.lm$resid[i], as.character(i))
  else
    text(climate.lm$fitted[i]+1.0, climate.lm$resid[i], as.character(i))
}
```



추정값과 잔차의 산점도를 보면 3 번과 15 번 케이스를 제외하면 어떤 뚜련한 현상은 나타나고 않고 있다. 따라서 3 번과 15 번 케이스에 대한 면밀한 조사를 거쳐 특이점으로 판명되면 이 두 케이스를 제외하고 다시 분석에 들어가는 것이 좋을 수도 있다는 생각이 든다.

-- 회귀모형 출석수업 과제물 끝 --