출석수업 과제물(평가결과물) 표지(온라인제출용)

교과목명 : 회귀모형

학 번 : 202135-367895

성 명 : 김태정

강 의 실 : 지역대학 호

연 락 처 : 010-4172-4516

---

1 번 문제

```
forbes <- read.table('./reg2020/forbes.txt', header=T)
t(forbes)
```

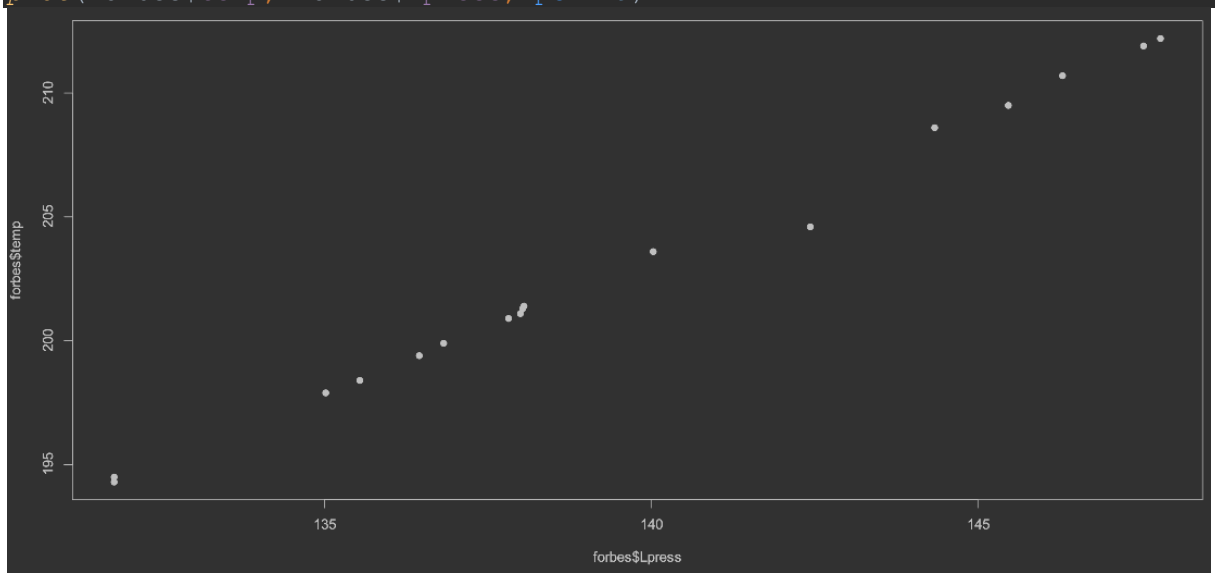
	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]	[,11]	[,12]	[,13]	[,14]	[,15]	[,16]	[,17]
num	1.00	2.00	3.0	4.00	5.00	6.00	7.00	8.00	9.00	10.00	11.00	12.00	13.00	14.00	15.00	16.00	17.00
temp	194.50	194.30	197.9	198.40	199.40	199.90	200.90	201.10	201.40	201.30	203.60	204.60	209.50	208.60	210.70	211.90	212.20
press	20.79	20.79	22.4	22.67	23.15	23.35	23.89	23.99	24.02	24.01	25.14	26.57	28.49	27.76	29.04	29.88	30.06

먼저 데이터를 가져옵니다.

```
forbes$Lpress <- 100 * log10(forbes$press)
```

데이터에서 press는 지침에 따라서 로그스케일로 변환해줍니다.

```
plot(forbes$temp, forbes$Lpress, pch=19)
```



그 다음 먼저 plot으로 산포도를 출력해봅니다. 출력한 산포도를 보면 선형관계가 존재함을 시각적으로 확인할 수 있습니다. 따라서 회귀모형을 적합해 볼 수 있습니다.

```
forbes.lm <- lm(forbes$Lpress ~ forbes$temp)
forbes.lm.summary <- summary(forbes.lm)
forbes.lm.summary
```

```
Call:
lm(formula = forbes$Lpress ~ forbes$temp)

Residuals:
     Min       1Q   Median       3Q      Max
-0.31974 -0.14707 -0.06890  0.01877  1.35994

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -42.16418    3.34136  -12.62 2.17e-09 ***
forbes$temp   0.89562    0.01646   54.42  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3792 on 15 degrees of freedom
Multiple R-squared:  0.995, Adjusted R-squared:  0.9946
F-statistic:  2962 on 1 and 15 DF,  p-value: < 2.2e-16
```

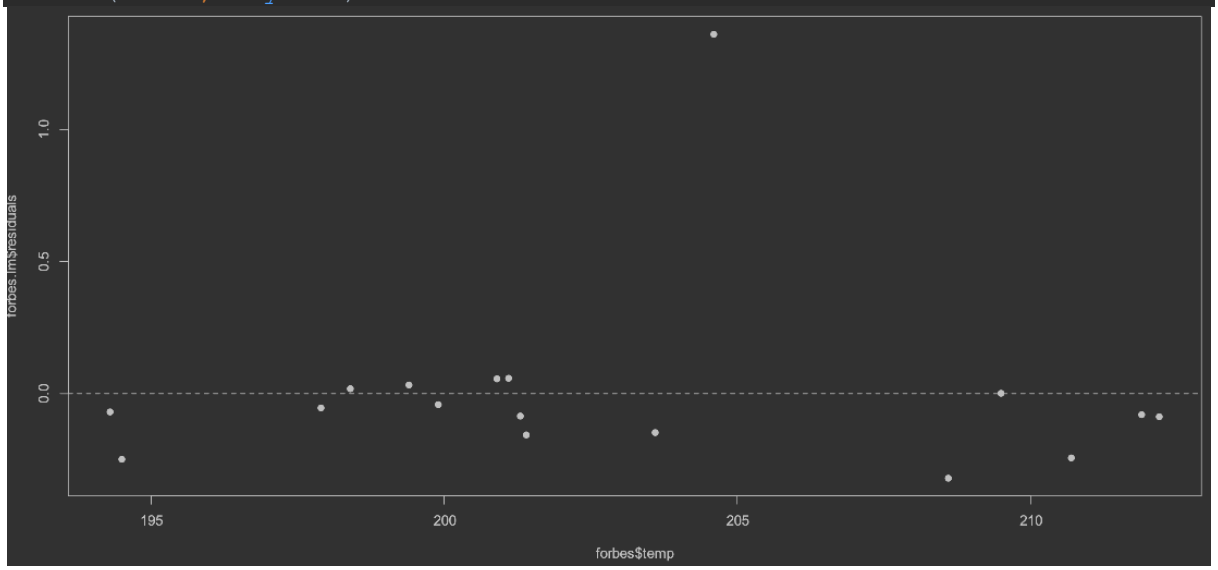
회귀모형을 적합시킨 결과 β_0(절편)은 -42.16이고 β_1(기울기)는 0.89입니다. 이후 기울기가 0이 아닌지, 즉 $H_0: \beta_1=0$인지 판단해야 합니다. 판단 방법은 t value를 기각역에 비교하거나 p value가 0.05보다 낮은지 확인해야합니다. 해당 summay에서 0.05보다 작은게 확인되므로 H_0를 기각시켜서 기울기는 0이 아니라고 판단합니다. 즉 이 회귀직선은 기울기가 0이 아니며 회귀식은 유의합니다.

```
cbind(forbes, forbes.lm$residuals, forbes.lm$fitted.values)
```

	num	temp	press	Lpress	forbes.lm$residuals	forbes.lm$fitted.values
1	1	194.5	20.79	131.7854	-0.24802254	132.0335
2	2	194.3	20.79	131.7854	-0.06889899	131.8543
3	3	197.9	22.40	135.0248	-0.05377004	135.0786
4	4	198.4	22.67	135.5452	0.01877126	135.5264
5	5	199.4	23.15	136.4551	0.03310101	136.4220
6	6	199.9	23.35	136.8287	-0.04111891	136.8698
7	7	200.9	23.89	137.8216	0.05618981	137.7654
8	8	201.1	23.99	138.0030	0.05847608	137.9445
9	9	201.4	24.02	138.0573	-0.15593374	138.2132
10	10	201.3	24.01	138.0392	-0.08445627	138.1237
11	11	203.6	25.14	140.0365	-0.14706580	140.1836
12	12	204.6	26.57	142.4392	1.35994454	141.0792
13	13	209.5	28.49	145.4692	0.00150698	145.4677
14	14	208.6	27.76	144.3419	-0.31973578	144.6617
15	15	210.7	29.04	146.2997	-0.24281806	146.5425
16	16	211.9	29.88	147.5381	-0.07916126	147.6172
17	17	212.2	30.06	147.7989	-0.08700828	147.8859

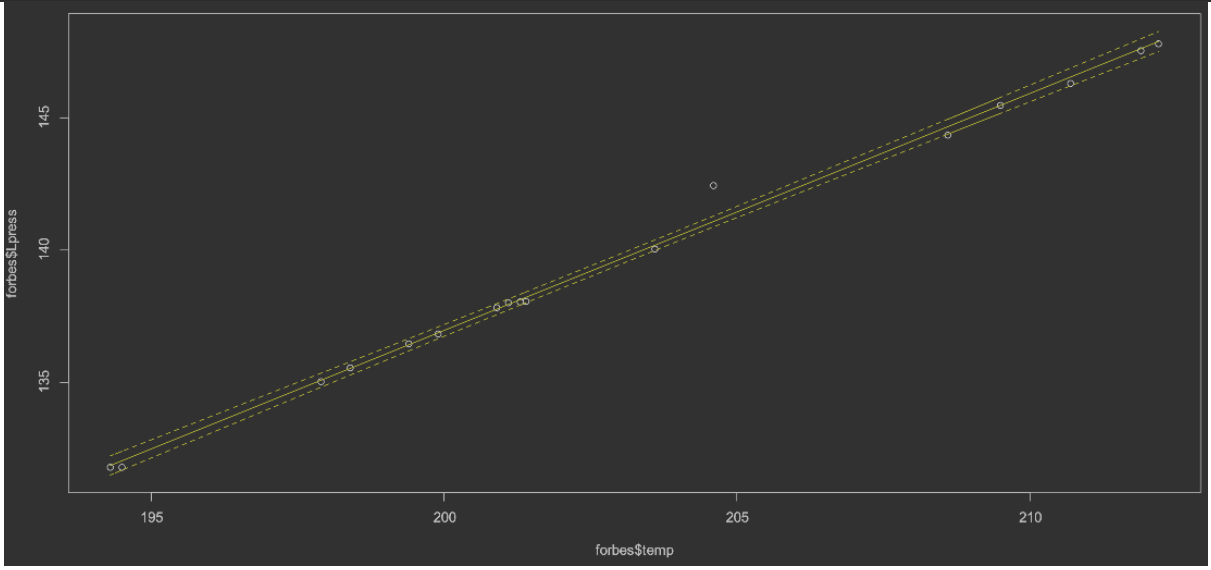
그 후 잔차와 추정값을 확인해보고 직관적으로 잘 적합됐는지 확인할 수 있습니다. 잔차(residuals)가 0근처이며 fitted.values(예측 값)이 선형관계를 보이는 것으로 보아 어느정도 회귀에 적합됐음을 확인할 수 있습니다.

```r
plot(forbes$temp, forbes.lm$residuals, pch = 19)
abline(h = 0, lty = 2)
```



그 후 이를 plot을 사용해서 시각적으로 잔차그림을 보여줄 수 있습니다. 잔차그림을 통해서 보면 잔차의 값은 0 을중심으로 일정 범위 안에 있는걸 확인할 수 있으며 X가 증가함에도 직선관계인 것을 보아 1차 회귀식에 더 맞는 형태임을 확인할 수 있습니다.

```r
p.x <- data.frame(temp = forbes$temp)
pc <- predict(forbes.lm, int = 'c', newdata = p.x)
pred.x <- p.x$temp
plot(forbes$temp, forbes$Lpress, ylim = range(forbes$Lpress, pc))
matlines(pred.x, pc, lty = c(1, 2, 2), col = 'BLUE')
```



신뢰대를 그려면 값들이 점추정값 기준으로 신뢰대 안에 대채로 잘 들어가 있음을 알 수 있습니다.

2 번 문제

```
health <- read.table('./reg2020/health.txt', header = T)
health <- health[, -1]
t(health)
```

```
   [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13] [,14] [,15] [,16] [,17] [,18] [,19] [,20] [,21] [,22] [,23] [,24] [,25] [,26] [,27] [,28] [,29] [,30]
X1  217  141  152  153  180  193  162  180  205  168  232  146  173  155  212  138  147  197  165  125  161  132  257  236  149  161  198  245  141  177
X2   67   52   58   56   66   71   65   80   77   74   65   68   51   64   66   70   54   76   59   58   52   62   64   72   57   57   59   70   63   53
X3  260  190  203  183  170  178  160  170  188  170  220  158  243  198  220  180  150  228  188  160  190  163  313  225  173  173  220  218  193  183
X4   91   66   68   70   77   82   74   84   83   79   72   68   56   59   77   62   75   88   70   66   69   59   96   84   68   65   62   69   60   75
Y   481  292  338  357  396  429  345  469  425  358  393  346  279  311  401  267  404  442  368  295  391  264  487  481  374  309  367  469  252  338
```

해당 값들은 health.txt에 있으므로 이를 가져옵니다. ID는 제거 해야하기 때문에 -1로 빼주면 health값들을 확인할 수 있습니다

```
cor(health)
```

```
          X1        X2        X3        X4         Y
X1 1.0000000 0.41992379 0.73655862 0.6431189 0.7980913
X2 0.4199238 1.00000000 0.06033638 0.5388523 0.5011203
X3 0.7365586 0.06033638 1.00000000 0.4001311 0.4446888
X4 0.6431189 0.53885228 0.40013111 1.0000000 0.8480748
Y  0.7980913 0.50112029 0.44468884 0.8480748 1.0000000
```

상관관계행렬을 뽑아보면은 Y와 연결성이 높은 항목은 X4와 X1임을 직감적으로 알 수 있습니다.

```
h0.lm <- lm(Y ~ X1, data = health)
h1.lm <- lm(Y ~ X4, data = health)
h2.lm <- lm(Y ~ X4 + X1, data = health)
h3.lm <- lm(Y ~ X4 + X1 + X2, data = health)
h4.lm <- lm(Y ~ X4 + X1 + X2 + X3, data = health)
```

각 값들의 비교를 위해서 변수의 값을 조절해가면서 회귀모형을 뽑아봅니다. 그 이후 변수의 선택을 하기위해서 분산분석을 시행합니다.

```
anova1 <- anova(h0.lm, h2.lm)
anova1
anova1$RSS[1] - anova1$RSS[2]
```

```
Analysis of Variance Table

Model 1: Y ~ X1
Model 2: Y ~ X4 + X1
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1     28 50795
2     27 24049  1     26746 30.027 8.419e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
[1] 26745.56
```

변수 X4에서 X1을 추가한 것의 차이는 26745입니다.

```
anova2 <- anova(h2.lm, h3.lm)
anova2
anova2$RSS[1] - anova2$RSS[2]
```

```
Analysis of Variance Table

Model 1: Y ~ X4 + X1
Model 2: Y ~ X4 + X1 + X2
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1     27 24050
2     26 24018  1    31.284 0.0339 0.8554
[1] 31.28424
```

변수에 X2를 추가한것의 차이는 31로 숫자가 적을수록 영향을 적게 미치기에 X2는 기여도가 떨어지는 것을 확인할 수 있습니다.

```
anova3 <- anova(h3.lm, h4.lm)
anova3
anova3$RSS[1] - anova3$RSS[2]
```

```
Analysis of Variance Table

Model 1: Y ~ X4 + X1 + X2
Model 2: Y ~ X4 + X1 + X2 + X3
  Res.Df   RSS Df Sum of Sq      F  Pr(>F)
1     26 24018
2     25 20551  1    3466.8 4.2173 0.05061 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
[1] 3466.843
```

모든 변수를 다 쓰는 것 역시 기여도가 떨어집니다.

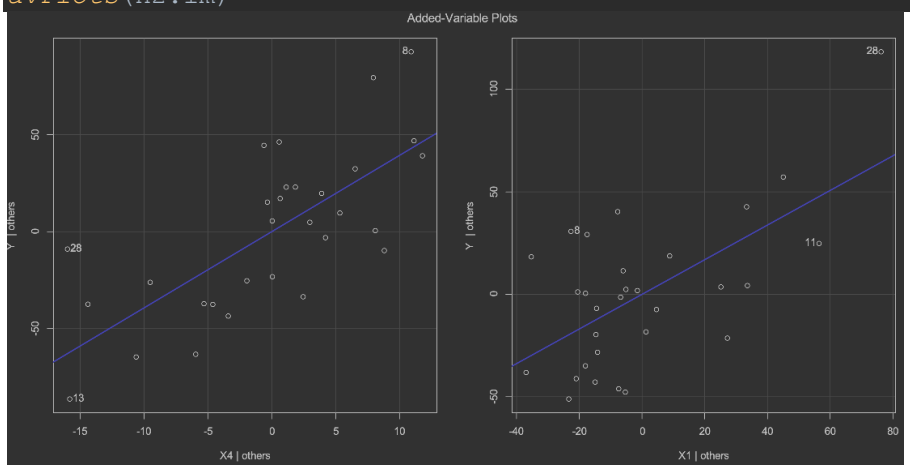
```
anova0 <- anova(h1.lm, h2.lm)
anova0
anova0$RSS[1] - anova0$RSS[2]
```

```
Analysis of Variance Table

Model 1: Y ~ X4
Model 2: Y ~ X4 + X1
  Res.Df   RSS Df Sum of Sq      F   Pr(>F)
1     28 39283
2     27 24049  1     15233 17.102 0.000309 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
[1] 15233.44
```

가장 기여도가 큰 X4에서 X1을 추가할 때의 p value를 보면 0.05를 넘어서므로 둘의 분산은 다르다고 할 수 있고 따라서 귀무가설을 기각할 수 있습니다. 이는 X4와 X1모두 회귀식에 기여를 하고 있음을 나타냅니다.

```
library(car)
avPlots(h2.lm)
```



추가변수 그림을 보면 데이터들이 회귀선에 시각적으로 모여있음을 확인할 수 있습니다.

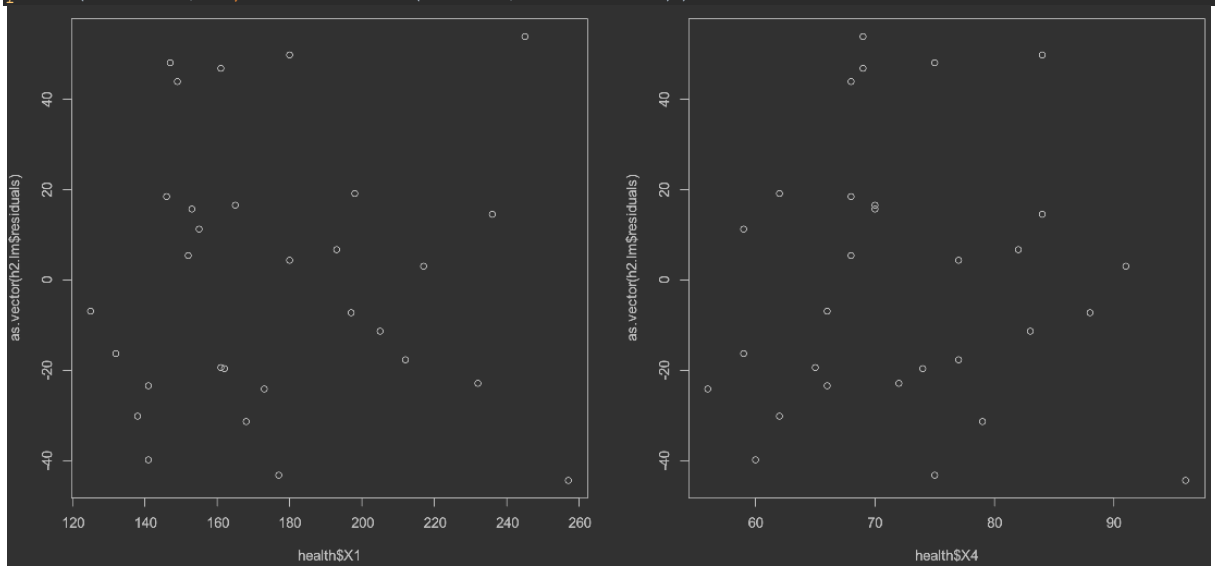
```
health.lm.anova <- anova(h2.lm)
health.lm.anova
```

```
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value    Pr(>F)
X4         1 100629  100629 112.975 3.776e-11 ***
X1         1  15233   15233  17.102  0.000309 ***
Residuals 27  24049     891
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

해당식의 분산분석표를 뽑아보면 p value가 매우 작아서 중회귀모형이 유의함을 확인할 수 있습니다.

```
par(mfrow = c(1, 2))
plot(health$X1, as.vector(h2.lm$residuals))
plot(health$X4, as.vector(h2.lm$residuals))
```



Health의 X1과 X2의 잔차 산점도를 보면 특이한 이상점으 없는걸 확인할 수 있습니다.