출석수업 과제물(평가결과물) 표지(온라인제출용)

**교과목명 : 회귀모형**

**학 번 : 202135-367895**

**성 명 : 김태정**

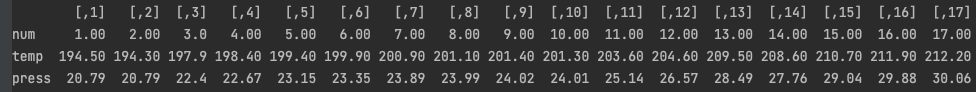
**강 의 실 : 지역대학 호**

**연 락 처 : 010-4172-4516**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**1번 문제**

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



먼저 데이터를 가져옵니다.

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

데이터에서 press는 지침에 따라서 로그스케일로 변환해줍니다.

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

텍스트, 별, 옥외설치물, 밤하늘이(가) 표시된 사진

자동 생성된 설명

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

텍스트이(가) 표시된 사진

자동 생성된 설명

회귀모형을 적합시킨 결과 β₀(절편)은 -42.16이고 β₁(기울기)는 0.89입니다. 이후 기울기가 0이 아닌지, 즉 H₀:β₁=0인지 판단해야 합니다. 판단 방법은 t value를 기각역에 비교하거나 p value가 0.05보다 낮은지 확인해야합니다. 해당 summay에서 0.05보다 작은게 확인되므로 H₀를 기각시켜서 기울기는 0이 아니라고 판단합니다. 즉 이 회귀직선은 기울기가 0이 아니며 회귀식은 유의합니다.

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

텍스트이(가) 표시된 사진

자동 생성된 설명

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

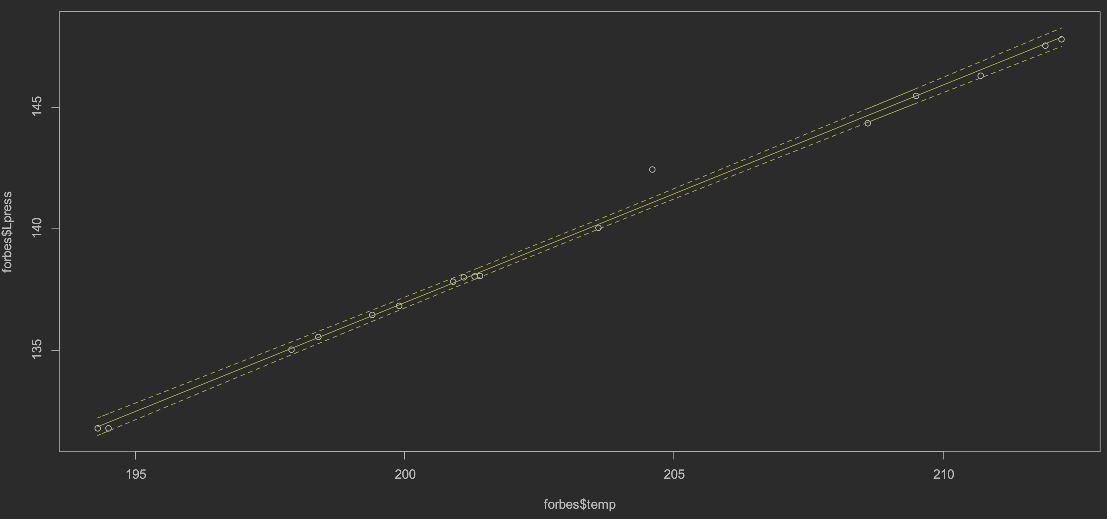
*plot*(forbes$temp, forbes.lm$residuals, pch = 19)  
*abline*(h = 0, lty = 2)

텍스트이(가) 표시된 사진

자동 생성된 설명

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.

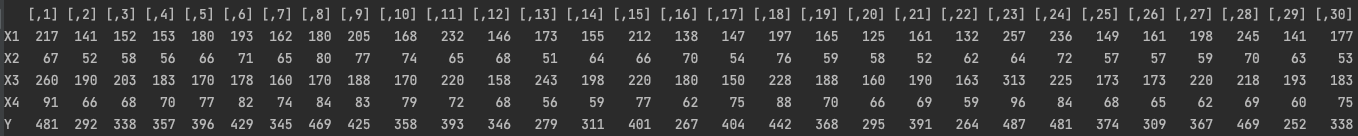
p.x <- *data.frame*(temp = forbes$temp)  
pc <- *predict*(forbes.lm, int = 'c', newdata = p.x)  
pred.x <- p.x$temp  
*plot*(forbes$temp, forbes$Lpress, ylim = *range*(forbes$Lpress, pc))  
*matlines*(pred.x, pc, lty = *c*(1, 2, 2), col = 'BLUE')



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

**2번 문제**

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



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

*cor*(health)

텍스트이(가) 표시된 사진

자동 생성된 설명

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]

텍스트이(가) 표시된 사진

자동 생성된 설명

변수 X4에서 X1을 추가한 것의 차이는 26745입니다.

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

텍스트이(가) 표시된 사진

자동 생성된 설명

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]

텍스트이(가) 표시된 사진

자동 생성된 설명

모든 변수를 다 쓰는 것 역시 기여도가 떨어집니다.

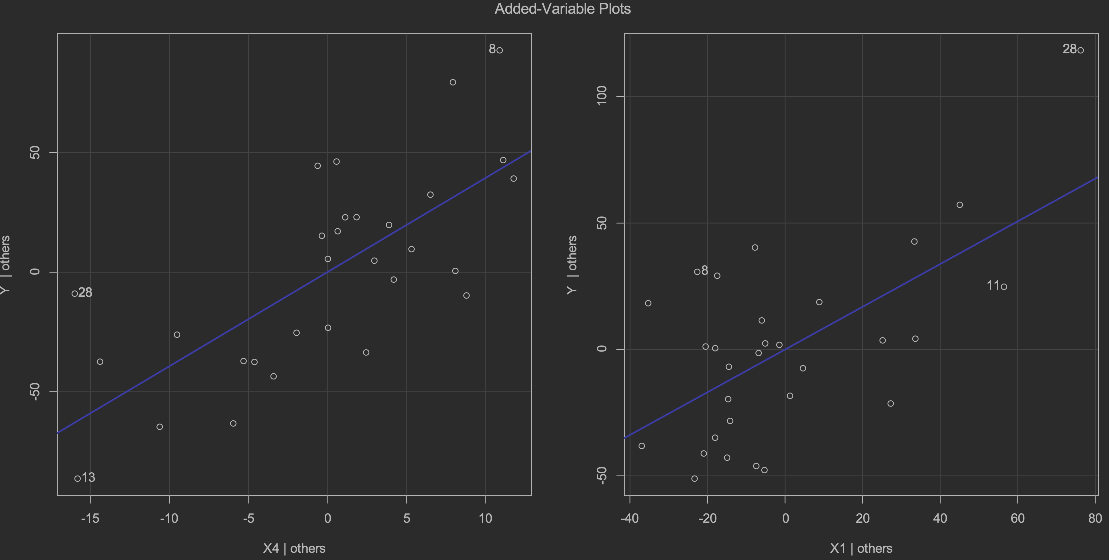
anova0 <- *anova*(h1.lm, h2.lm)  
anova0  
anova0$RSS[1] - anova0$RSS[2]

텍스트이(가) 표시된 사진

자동 생성된 설명

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

텍스트이(가) 표시된 사진

자동 생성된 설명

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))

텍스트, 모니터이(가) 표시된 사진

자동 생성된 설명

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