단순 회귀 모형 실습

대전 1반 서현택, 김지원

1. 데이터 셋 읽어오기

1) 코드

```
# 1. 데이터 셋 읽어오기
PSDS_PATH <- file.path('C:/Users/User/Desktop/sht3898/Data_Science/빅데이터인턴십/R')
PSDS_PATH

std90 <- read.table(file.path(PSDS_PATH, "student90.csv"),
                    sep=",",
                    stringsAsFactors = FALSE,
                    header = TRUE,
                    na.strings = "")
```

2) 결과

```
> head(std90)
  no sex weight_kg height_cm
1  1   m        98       198
2  2   m        77       170
3  3   m        70       170
4  4   m        90       198
5  5   m        71       170
6  6   m        70       165
```

2. 회귀 모델 생성

1) 코드

```
# 2. 회귀 모델 생성
m <- lm(weight_kg~height_cm, data=std90)
```

2) 생성 결과

```
> # 2. 회귀 모델 생성
> (m <- lm(weight_kg~height_cm, data=std90))

Call:
lm(formula = weight_kg ~ height_cm, data = std90)

Coefficients:
(Intercept)    height_cm
    32.6604       0.2247
```

3. 회귀 계수 구하기

1) 코드

```
# 3. 회귀 계수 구하기
coef(m)
# height_cm: 0.2247, 잔차: 32.6604
```

2) 결과

```
> # 3. 회귀 계수 구하기
> coef(m)
(Intercept)   height_cm
 32.6604144   0.2246605
```

* 학생의 몸무게(kg) = 32.6604 + 0.2247*학생의 키(cm)

4. 회귀 계수 값 검증하기

1) 코드

```
# 4. 회귀 계수 값 검증하기
# 이상값 구하기
plot(m, which=4)

# 이상값 진단
x_cooks.d <- cooks.distance(m)
x_cooks.d[1:4]

NROW(x_cooks.d)

x_cooks.d[which(x_cooks.d>qf(0.5, df1 = 2, df2 = 88))]
summary(m)

# install.packages("car")
# library(car)
outlierTest(m)
```

2) Cook's distance 결과

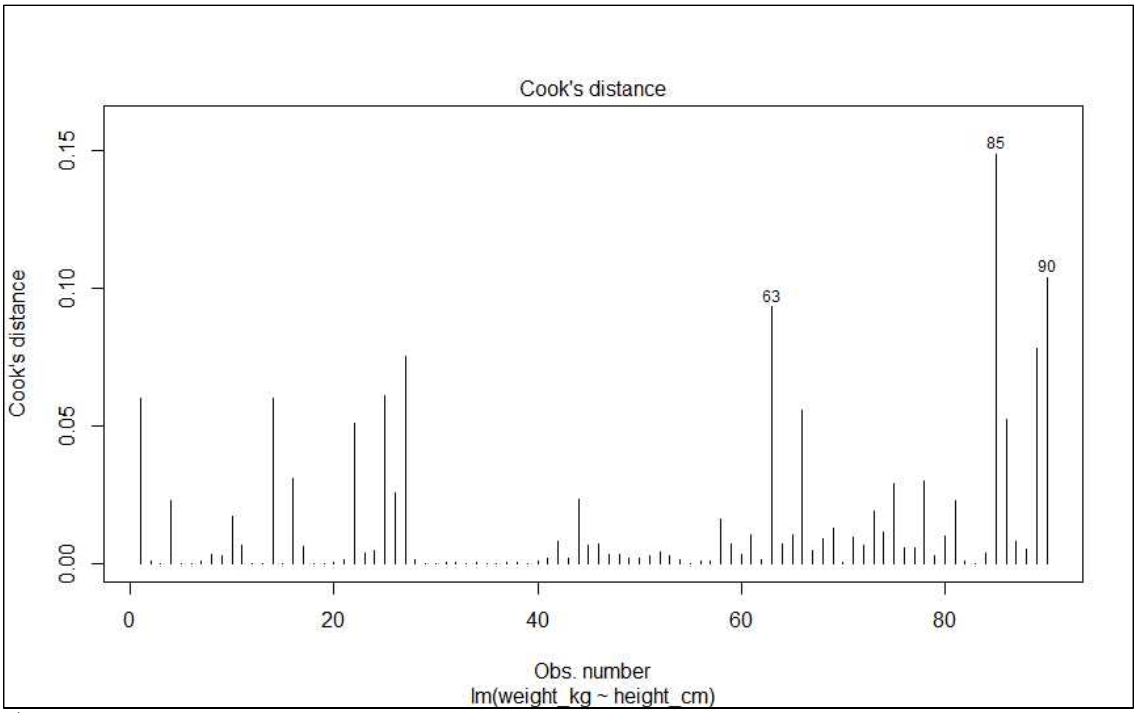
* 3개의 이상치 존재(63, 85, 90)

3) 본페로니 검사

```
> outlierTest(m)
No Studentized residuals with Bonferroni p < 0.05
Largest |rstudent|:
   rstudent unadjusted p-value Bonferroni p
90 2.709609          0.0081125      0.73013
```

* 검사 결과 본페로니 p(=0.73) > 0.05 이므로 이상치가 검출되지 않음을 알 수 있다.

5. 잔차 구하기



1) 코드

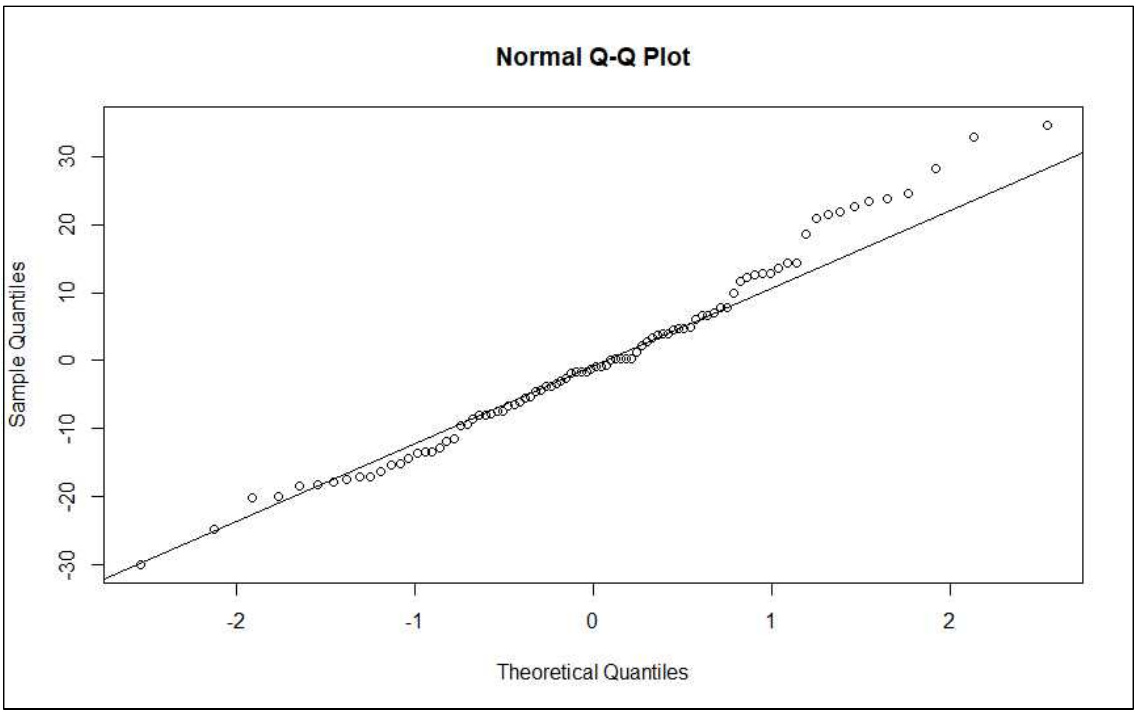
```
# 5. 잔차 구하기
# 잔차
residuals(m)[1:4]

std90$weight_kg[1:4]
fitted(m)[1:4] + residuals(m)[1:4]

# 잔차의 정규성 확인
qqnorm(residuals(m))
qqline(residuals(m))
shapiro.test(residuals(m))
```

2) 잔차 결과 확인

```
> std90$weight_kg[1:4]
[1] 98 77 70 90
> fitted(m)[1:4] + residuals(m)[1:4]
 1  2  3  4
98 77 70 90
```

* 원 값과 회귀식+잔차 값의 같음을 확인

3) 잔차의 정규성 확인




```
> # 잔차의 정규성 확인
> qqnorm(residuals(m))
> qqline(residuals(m))
> shapiro.test(residuals(m))

        Shapiro-Wilk normality test

data:  residuals(m)
W = 0.98121, p-value = 0.2189
```

* qqplot 그림에서도 정규성의 형태를 띄게 나왔고, shapiro-test에서도 p-value > 0.05 이므로 정규성을 따른다는 결론을 도출함.

6. 잔차 제곱 합 구하기

1) 코드

```
# 6. 잔차 제곱 합 구하기
deviance(model)
```

2) 결과

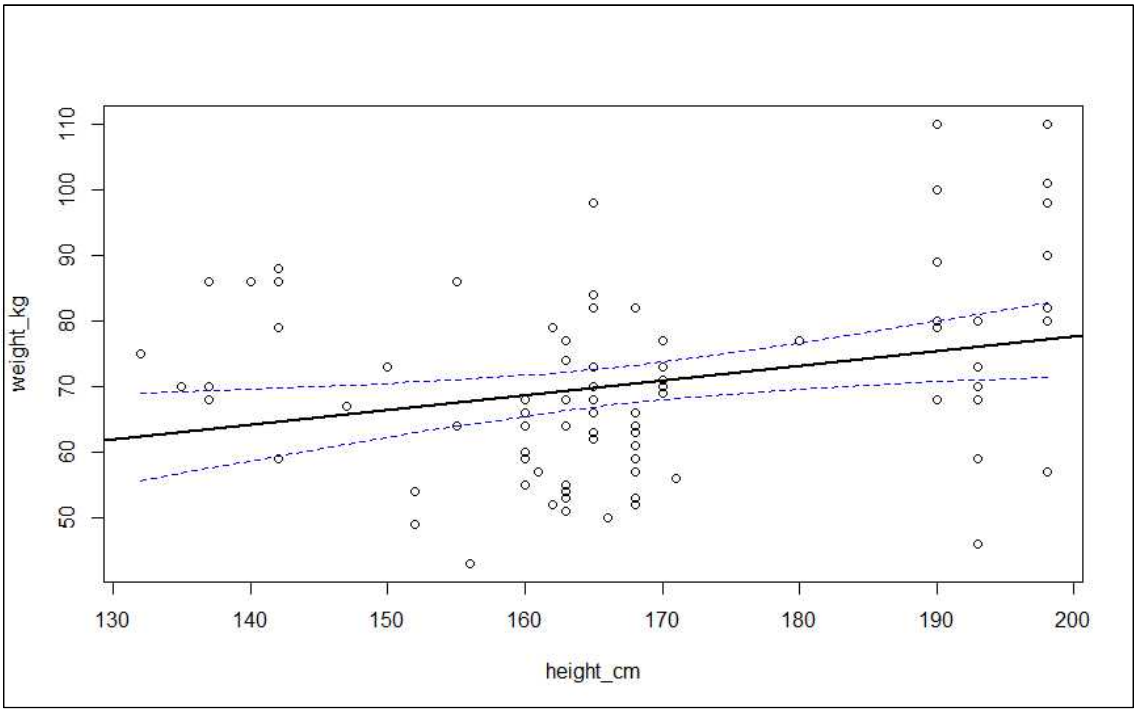
```
> # 6. 잔차 제곱 합 구하기
> deviance(model)
[1] 1234764
```

7. 회귀 계수 신뢰 구간 구하기

1) 코드

```
# 7. 회귀 계수 신뢰 구간 구하기
# 회귀 계수의 95% 신뢰구간 구하기기
confint(m, level=0.95)
m_conf <- predict(m, level = 0.95, interval = "confidence")
head(m_conf)

# 키와 몸무게 산포도, 추정된 평균 몸무게, 신뢰구간
plot(weight_kg~height_cm, data=std90)
lwr <- m_conf[,2]
upr <- m_conf[,3]
sx <- sort(std90$height_cm, index.return=TRUE)
abline(coef(m), lwd=2)
lines(sx$x, lwr[sx$ix], col="blue", lty=2)
lines(sx$x, upr[sx$ix], col="blue", lty=2)
```

2) plot



3) 신뢰구간

```
> head(m_conf)
       fit      lwr      upr
1 77.14319 71.45341 82.83298
2 70.85270 68.02003 73.68536
3 70.85270 68.02003 73.68536
4 77.14319 71.45341 82.83298
5 70.85270 68.02003 73.68536
6 69.72940 66.86626 72.59253
```

8. 새로운 학생 키로 몸무게 예측하기

1) 코드

```
# 8. 새로운 학생 키로 몸무게 예측하기
m_pred <- predict(m, level = 0.95, interval = "predict")
head(m_pred)

# 키와 몸무게 산포도, 예측구간
p_lwr <- m_pred[,2]
p_upr <- m_pred[,3]
lines(std90$height_cm, p_lwr, col="red", lty=2)
lines(std90$height_cm, p_upr, col="red", lty=2)

predict(m, newdata=data.frame(height_cm=175), interval = "confidence")
```

2) 175cm인 학생의 예측 결과

```
> m_pred <- predict(m, level = 0.95, interval = "predict")
경고메시지(들):
In predict.lm(m, level = 0.95, interval = "predict") :
  현재 데이터를 이용한 예측은 _future_response을 의미합니다

> head(m_pred)
       fit      lwr       upr
1 77.14319 49.83131 104.45507
2 70.85270 43.99029  97.71511
3 70.85270 43.99029  97.71511
4 77.14319 49.83131 104.45507
5 70.85270 43.99029  97.71511
6 69.72940 42.86376  96.59504
> # 키와 몸무게 산포도, 예측구간
> p_lwr <- m_pred[,2]
> p_upr <- m_pred[,3]
> lines(std90$height_cm, p_lwr, col="red", lty=2)
> lines(std90$height_cm, p_upr, col="red", lty=2)
> predict(m, newdata=data.frame(height_cm=175), interval = "confidence")
     fit      lwr      upr
1 71.976 68.93945 75.01255
>
```

* 71.976kg로 예측되었고 신뢰구간은 68.93945부터 75.01255까지이다.

9. 모델 평가하기

```
> summary(m)

Call:
lm(formula = weight_kg ~ height_cm, data = std90)

Residuals:
    Min      1Q  Median      3Q     Max
-30.020  -8.460  -1.066   6.918  34.654

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  32.6604    14.0771   2.320  0.02265 *
height_cm     0.2247     0.0833   2.697  0.00838 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.44 on 88 degrees of freedom
Multiple R-squared:  0.07635,   Adjusted R-squared:  0.06585
F-statistic: 7.274 on 1 and 88 DF,  p-value: 0.008385

> # F 통계량 구하기
> anova(m)
Analysis of Variance Table

Response: weight_kg
           Df  Sum Sq Mean Sq F value   Pr(>F)
height_cm   1  1314.2 1314.22  7.2737 0.008385 **
Residuals  88 15899.9  180.68
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> (m_a <- lm(weight_kg~height_cm, data = std90))

Call:
lm(formula = weight_kg ~ height_cm, data = std90)

Coefficients:
(Intercept)    height_cm
    32.6604       0.2247

> (m_b <- lm(weight_kg~1, data=std90))

Call:
lm(formula = weight_kg ~ 1, data = std90)

Coefficients:
(Intercept)
      70.43

> anova(m_a, m_b)
Analysis of Variance Table

Model 1: weight_kg ~ height_cm
Model 2: weight_kg ~ 1
  Res.Df   RSS Df Sum of Sq      F   Pr(>F)
1     88 15900
2     89 17214 -1   -1314.2 7.2737 0.008385 **
```

* 분석결과 모든 변수의 p value < 0 이기 때문에 변수들은 유의수준 0.05에서 유의하다 볼 수 있다.
* 분산분석 결과 height_cm의 p value < 0 이므로 모델이 유의하다 할 수 있다.
* 축소 모델과 비교 결과 p value < 0 이므로 키가 유의미한 설명 변수임을 알 수 있다.