단순 선형 회귀

- 단순 회귀 모형 실습
  - 대학생 90명의 키와 몸무게 데이터를 이용해, 회귀 모델을 생성하고,
  - 회귀 계수, 회귀 계수의 신뢰구간, 잔차, 잔차 제곱의 합,
  - 새로운 학생 키로 몸무게 예측, 모델 평가 해보기

실습 순서
1. 데이터 셋 읽어오기
2. 회귀 모델 생성
3. 회귀 계수 구하기
4. 회귀 계수 값 검증하기
5. 잔차 구하기
6. 잔차 제곱 합 구하기
7. 회귀 계수 신뢰 구간 구하기
8. 새로운 학생 키로 몸무게 예측하기
9. 모델 평가 하기

단순 선형 회귀

- 데이터 읽어오기
  - 대학생 90명의 키와 몸무게 데이터 셋 메모리에 로딩하기

```
PSDS_PATH <- file.path('.', 'source')

# 대학생 92명의 키와 몸무게 데이터 읽기
std90 <- read.table(file.path(PSDS_PATH, "data", "student90.csv"),
                    sep = ",",
                    stringsAsFactors = FALSE,
                    header = TRUE,
                    na.strings = "")
nrow(std90)
#[1] 90
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
```

단순 선형 회귀

- 회귀 모델 생성
  - 대학생 90명의 키와 몸무게 데이터 셋을 이용한 회귀 모델 생성하기
    - ✓ R의 lm() 함수 이용
    - ✓ 학생의 몸무게(kg) = β_0 + β_1 × 학생의 키(cm)

      절편 계수

      ※ 절편과 계수를 회귀 계수

```
(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
```

단순 선형 회귀

- 회귀 계수
  - 대학생 90명의 키와 몸무게 데이터에서 몸무게(kg)~키(cm) 생성된 회귀 모델에서 회귀 계수 구하기
    - ✓ R의 coef(model) 함수 이용
    - ✓ 학생의 몸무게(kg) = 32.66 + 0.225 * 학생의 키(cm)

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

단순 선형 회귀

- 적합(예측)된 값 구하기
  - 대학생 90명의 키와 몸무게 데이터에서 몸무게(kg)~키(cm) 생성된 회귀 모델에서 적합(예측)된 값 구하기
    - ✓ R의 fitted(model) 함수 이용
    - ✓ 대학생 90명의 키와 몸무게 데이터의 1~4번째 데이터의 적합(예측)된 값과 1~4번째 학생의 키를 회귀 식을 이용해 계산한 값과 비교
    - ✓ 학생의 몸무게(kg) = 32.66 + 0.225 * 학생의 키(cm) ← 실제학생의 키(cm)

```
# 대학생 90명 데이터의 1~4번째 적합(예측)된 값 확인하기 : fitted(model)
fitted(m)[1:4]
#        1        2        3        4
# 77.14319 70.85270 70.85270 77.14319

# 학생의 몸무게(kg) = 32.66 + 0.224 * 학생의 키(cm)
((32.6604144) + (0.2246605) * (std90$height_cm[1:4]))
#        1        2        3        4
# 77.14319 70.85270 70.85270 77.14319
```

같은 값

단순 선형 회귀

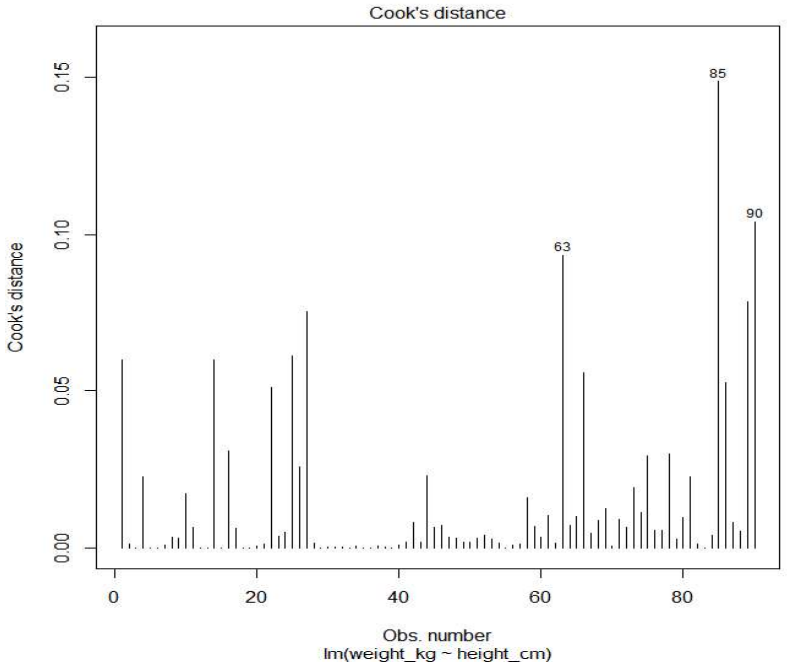
- 이상값 진단
  - 단순 선형 회귀모형에서 잔차분석은 오차 가정을 진단하는 과정이고 추가로 모형진단에서 주의해야 할 것은 이상값(outlier) 탐색이다.
  - 단순선형회귀모형에서 이상값이란 선형관계 및 오차 범위를 벗어난 값을 말한다.
  - 이상값이 발생하는 원인은 다양하고,
  - 가장 단순한 원인으로 입력 오류를 들 수 있고,
  - 만약 이상값이 존재한다면 단순 선형 회귀모형에서 계수 추정이 통계적으로 적절하지 않으며, 잔차도 마찬가지이다.
  - Cook's distance는 잔차와 영향값(influential points)으로 고안된 통계량으로,
  - 회귀모형에서 이상값과 영향값(influential observation)을 탐색하는 데 유용한 통계적 기법이다.
  - 이상치 검출에서는 잔차, 특히 외면 스튜던트화 잔차(Externally Studentized Residual)를 사용하고,
  - R의 car::outlierTest() 함수를 사용해 쉽게 구할 수 있다.

단순 선형 회귀

- 대학생 90명의 키와 몸무게 데이터에서 몸무게(kg)~키(cm) 생성된 회귀 모델에서 이상값(outlier) 진단 구하기

  ✓ R의 cooks.distance(model) 함수 이용하여 Cook's distance 값을 구한다.

  ✓ 이 값을 그림으로 나타내기 위해 plot() 함수를 수행하면 다음과 같다.




```
# 키와 몸림무게의 이상값 그리
plot(m, which = 4)
```

✓ Cook's distance 그림을 보면 3개 값이 이상값으로 의심되지만 수치적인 탐색을 이용하여 이상값을 결정할 필요가 있다.

단순 선형 회귀

- 이상값 진단(계속)

  - ✓ R의 cooks.distance(model) 함수 이용하여 Cook's distance 값을 구한다.
  - ✓ F-분포(분모 자유도=2, 분자 자유도=88)의 분위수(50%)가 수치적 탐색의 지표이고 분모 자유도와 분자 자유도의 합은 자료수 90개,
  - ✓ 즉 Cook distance 값이 F-분포의 50% 분위수 이상이면 이상값으로 간주한다.

```
# 이상값 진단
x_cooks.d <- cooks.distance(m)
x_cooks.d[1:4]
#                1             2             3             4
#5.992961e-02 1.202838e-03 2.314356e-05 2.277257e-02

NROW(x_cooks.d)
#[1] 90

x_cooks.d[which(x_cooks.d>qf(0.5, df1 = 2, df2 = 88))]
#named numeric(0)
```

단순 선형 회귀

- 이상값 진단(계속)

  ✓ R의 car::outlierTest(model) 함수 이용하여 본페로니(Bonferroni) p가 0.05 보다 작은 경우 이상치인 것으로 판단한다.

```
install.packages("car")
library(car)
outlierTest(m)
#No Studentized residuals with Bonferonni p < 0.05
#Largest |rstudent|:
#       rstudent unadjusted p-value  Bonferonni p
# 90 2.709609             0.0081125        0.73013
```

  ✓ 실행 결과, 본페로니 p(=0.73) > 0.05 이므로 이상치가 검출되지 않음을 알 수 있다.

단순 선형 회귀

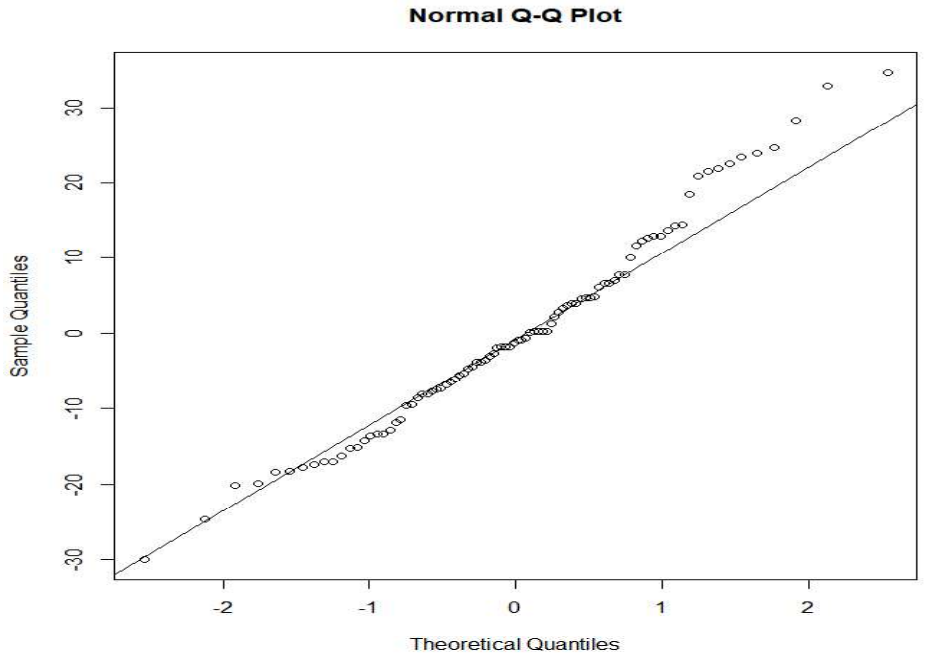
- 잔차
  - 대학생 90명의 키와 몸무게 데이터에서 몸무게(kg)~키(cm) 생성된 회귀 모델에서 잔차(residual) 구하기
    - ✓ R의 residual(model) 함수 이용한다.
    - ✓ 대학생 90명의 키와 몸무게 데이터의 1~4번째 데이터의 잔차 값과 1~4번째 학생의 적합된 값을 이용해 계산한 값과 실제 값을 비교
    - ✓ 실제 데이터 값 = 적합 된 값 + 잔차

```
# 대학생 90명 데이터의 1~4번째 잔차 구하기 : residuals(model)
residuals(m)[1:4]
#          1          2          3          4
# 20.8568064  6.1473004 -0.8526996 12.8568064

# 실제 데이터 값 = 적합된 값 + 잔차
# 대학생 90명 데이터의 1 ~ 4번째 실제 몸무게
std90$weight_kg[1:4]
#  98 77 70 90

# 대학생 90명 데이터의 1 ~ 4번째 적합된 값 + 잔차
fitted(m)[1:4] + residuals(m)[1:4]
#   1  2  3  4
# 98 77 70 90
```

단순 선형 회귀

- 잔차 분석
  - 대학생 90명의 키와 몸무게 데이터에서 몸무게(kg)~키(cm) 생성된 회귀 모델에서 잔차(residual) 분석 하기
    - ✓ R에서 제공하는 Q-Q plot도를 이용하여 잔차의 정규성 확인

```
# Q-Q plot
qqnorm(residuals(m))
qqline(residuals(m))
```



단순 선형 회귀

- 잔차 분석(계속)

  ✓ R에서 제공하는 샤피로 윌크 검정(Shapiro-Wilk Test)을 이용하여 잔차의 정규성 확인하기

```
# 샤피로 윌크 검정: 일변수 자료에 대해 수치적으로 정규성을 검정하는 기법
shapiro.test(residuals(m))
#
#Shapiro-Wilk normality test
#
#data:  residuals(m)
#W = 0.98121, p-value = 0.2189
```

  ✓ R에서 제공하는 샤피로 윌크 검정(Shapiro-Wilk Test)을 이용하여 잔차의 정규성 확인

  ✓ 귀무 가설 H_0: 잔차가 정규분포를 따른다.

  ✓ 대립 가설 H_1: 잔차가 정규분포를 따르지 않는다.

  - 샤피로 윌크 검정 결과, p-value(=0.2189) > 0.05 이므로 데이터가 정규 분포를 따른다는 귀무가설을 기각할 수 없다.

단순 선형 회귀

- 회귀 계수의 신뢰구간
  - 대학생 90명의 키와 몸무게 데이터에서 몸무게(kg)~키(cm) 생성된 회귀 모델에서 회귀 계수의 신뢰구간 구하기
    - ✓ R의 confint(model) 함수 이용한다.
    - ✓ 단순 선형 회귀에서 절편과 기울기는 정규 분포를 따른다.
    - ✓ 따라서, t 분포를 사용한 95%의 신뢰구간을 confit(model)을 사용해 구할 수 있다.

```
# 회귀 계수의 95% 신뢰구간 구하기 : confint(model)
confint(m, level = 0.95)
#                   2.5 %      97.5 %
# (Intercept)  4.68512548  60.6357032
# height_cm    0.05911794   0.3902031
```

단순 선형 회귀

- 신뢰구간
  - ✓ R의 predict() 함수의 옵션 interval="confidence" 을 선택하고,
  - ✓ 기본적으로 유의수준은 95%이고, 옵션 level=0.99를 사용하면 유의수준은 99%를 계산할 수 있다.

```
# 신뢰구간 구하기
m_conf <- predict(m, level = 0.95, interval = "confidence")
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
```

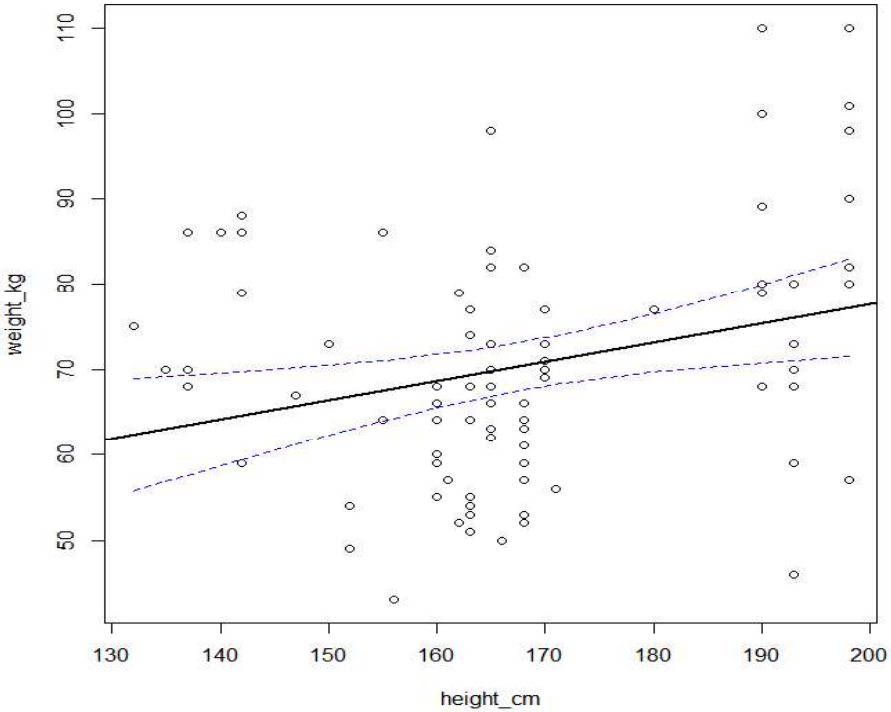
단순 선형 회귀

- 신뢰구간(계속)

  ✓ 주어진 키에 대한 평균 몸무게의 95% 신뢰구간(파란 점선)과 함께 산포도,

  ✓ 추정된 평균 몸무게(실선)를 그린 예시

```
# 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)
```



단순 선형 회귀

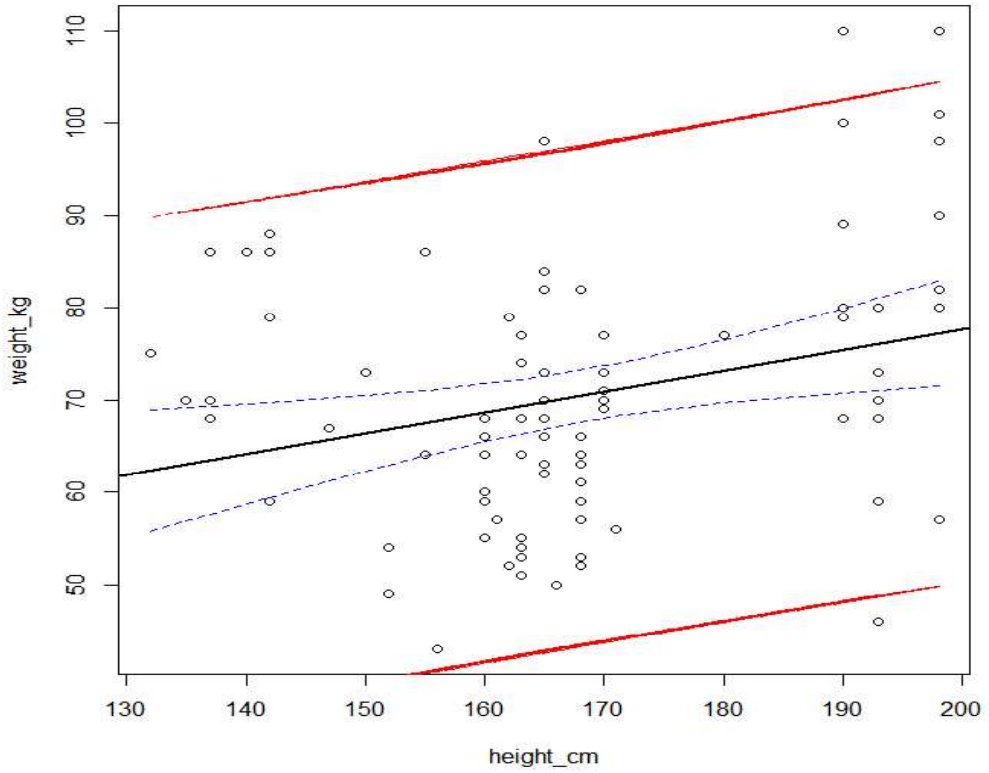
- 예측구간
  - 대학생 90명의 키와 몸무게 데이터에서 몸무게(kg)~키(cm) 생성된 회귀 모델에서 예측구간 구하기
    - ✓ R의 predict() 함수의 옵션 interval="confidence" 을 선택하고,
    - ✓ 기본적으로 유의수준은 95%이고, 옵션 level=0.99를 사용하면 유의수준은 99%를 계산 할 수 있다.

```
# 키와 몸무게의 예측구간
m_pred <- predict(m, level = 0.95, interval = "predict")
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
```

단순 선형 회귀

- 예측구간(계속)

```
# 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)
```



단순 선형 회귀

- 잔차 제곱의 합
  - 대학생 90명의 키와 몸무게 데이터에서 몸무게(kg)~키(cm) 생성된 회귀 모델에서 잔차 제곱의 합 구하기
    - ✓ R의 deviance(model) 함수 이용

$$\sum e_i^2 = \sum (Y_i - \hat{Y}_i)^2$$

```
# 잔체 제곱 합 구하기 : deviance(model)
deviance(m)
#  15899.88
```

단순 선형 회귀

- 예측하기
  - 대학생 90명의 키와 몸무게 데이터에서 몸무게(kg)~키(cm) 생성된 회귀 모델에서 새로운 학생 키(cm)로 몸무게 예측 하기
    - ✓ R의 predict() 함수 이용
    - ✓ 새로운 학생의 키가 175cm 라면, 이 학생의 예상되는 몸무게 구하기

```
# 새로운 학생의 키가 175cm 라면, 예상되는 몸무게 구하기
predict(m, newdata = data.frame(height_cm=175), interval = "confidence")
#       fit         lwr         upr
# 71.976   68.93945   75.01255
```

    - ✓ 회귀 계수(절편과 기울기)의 신뢰 구간을 고려하기 위해 type="confidence"를 지정
    - ✓ fit은 예측값의 점 추정치, lwr과 upr은 각각 신뢰 구간의 하한과 상한 값이다.
    - ✓ 예측결과, 새로운 학생의 몸무게는 약 72 kg인 것으로 예측된다.

단순 선형 회귀

- 모델 평가
  - 대학생 90명의 키와 몸무게 데이터에서 몸무게(kg)~키(cm) 생성된 회귀 모델에서 회귀 모델 평가하기

✓ R의 summary(model) 함수 이용

✓ 회귀 계수(Coefficients)에서는 모델의 계수와 이 계수들의 통계적 유의성을 알려준다.

✓ 몸무게(kg) = 32.66 + 0.225 * 학생의 키(cm)

✓ F 통계량(F-statistic)은 모델이 통계적으로 얼마나 의미가 있는지를 알려준다.

✓ F 통계량=7.274, p-value는 0.008 이다.

✓ 귀무가설 H_0: 계수(또는 절편)이 0 이다.

✓ 대립가설 H_1: 계수(또는 절편)이 0 아니다.

```
summary(m)
# Call:
# lm(formula = weight_kg ~ height_cm, data = s90)
#
# 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
```

단순 선형 회귀

```
summary(m)
# Call:
# lm(formula = weight_kg ~ height_cm, data = s90)
#
# 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
```

- ✓ Pr(>|t|)열은 t 분포를 사용하여 각 변수가 얼마나 유의한지를 판단할 수 있는 p-value를 알려준다.
- ✓ 수정 결정 계수(Adjusted R-squred)는 모델이 데이터의 분산을 얼마나 설명하는지를 알려준다.

단순 선형 회귀

- 회귀 모델 평가 결과

  ✓ 절편과 계수는 통계적으로 유의(절편과 계수의 p-값 < 0.05)

  ✓ 추정값의 95% 신뢰구간에 0이 포함되어 있지 않다.

  ✓ 그러나, 결정계수가 0.076으로 종속변수와 독립변수의 선형관계가 매우 낮다.

```
summary(m)
# Call:
# lm(formula = weight_kg ~ height_cm, data = s90)
#
# 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
```

단순 선형 회귀

- 분산 분석 및 모델간의 비교
  - 대학생 90명의 키와 몸무게 데이터에서 몸무게(kg)~키(cm) 생성된 회귀 모델에서 회귀 모델 평가하기
    - ✓ R의 anova() 함수를 이용한 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
```

    - ✓ F 통계량(F-statistic)은 모델이 통계적으로 얼마나 의미가 있는지를 알려준다.
    - ✓ p-value는 0.008 이다.
    - ✓ F 통계량은 '$H_0: \beta_1 = 0$', '$H_1: \beta_1 \neq 0$'에 대한 가설 검정 결과이다.

단순 선형 회귀

- 분산 분석 및 모델간의 비교(계속)
  - 대학생 90명의 키와 몸무게 데이터에서 몸무게(kg)~키(cm) 생성된 회귀 모델, 축소 모델인 몸무게(kg)~1 의 두 모델 비교하기
    - ✓ 축소 모델은 원래 사용한 모델보다 설명 변수를 줄인 모델로 키(cm)를 제거하고, 몸무게(kg)를 상수값으로 예측하는 경우

```
(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
```

단순 선형 회귀

- 분산 분석 및 모델간의 비교(계속)

```
# 두 모델 비교 결과
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 **
#  ---
#  Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

✓ F 통계량(F-statistic)은 모델이 통계적으로 얼마나 의미가 있는지를 알려준다.

✓ F 통계량은 7.274으로 낮게 나타나고,

✓ p-value는 0.008 이다.

✓ 두 모델 간에는 유의한 차이가 있다고 결론을 내림(즉, 키(cm)열이 유의미한 설명 변수임을 뜻함)

단순 선형 회귀

- RMSE, MAE를 이용한 모델간의 비교

```
rmse(m_a, std90)    # root-mean-squared-error
#[1] 13.29155
rmse(m_b, std90)    # root-mean-squared-error
#[1] 13.82996

mae(m_a, std90)     # mean absolute error
#[1] 10.45572
mae(m_b, std90)     # mean absolute error
#[1] 10.66296
```

모델	RMSE	MAE
m_a	13.29155	10.45572
m_b	13.82996	10.66296

✓ m_a 모델의 RMSE 값과 MAE 값이 작게 나와 더 우수하다고 할 수 있다.