상관분석 vs 단순 선형 회귀분석

- 상관 분석은 두 양적 자료의 선형관계(또는 연관) 유무를 통계적 관점으로 다룰 수 있지만 두 자료의 선형 관계식이 제시되지는 않는다.

- 상관관계가 두 변수 사이의 전체적인 관련 강도를 측정하는 것이라면, 회귀는 관계 자체를 정량화하는 방법이라는 점에서 차이가 있음

- 단순 선형 회귀모형은 두 자료의 선형식과 설명할 수 없는 오차항(확률변수)으로 만들어진 모형이고,

- 예를 들어, 키와 몸무게를 단순 선형 회귀모형에 적용하면, 키가 170cm일 때 몸무게 값을 파악할 수 있다.

선형 회귀의 기본 가정

- 선형 회귀의 기본 가정

  - 회귀 모형은 i번째 관측 값을 뜻하는 변수들이 $(X_{i1}, X_{i2}, \dots, X_{ip}, Y_i)$ 형태로 주어졌을 때 종속 변수 Y_i와 p개의 독립 변수 $X_{i1}, X_{i2}, \dots, X_{ip}$를 다음과 같은 선형 식으로 표현한다.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \varepsilon_i$$

ε : Epsilon(엡실론)

  - 위 식에서 $\beta_0, \beta_1, \dots, \beta_p$는 회귀 모델의 계수며, ε_i 는 오차(error)

  - 예를 들어, 선형 모델이 유용한 경우는 자동차 제동 거리와 브레이크를 밟기 전의 주행 속도 간의 관계, 아버지의 키와 아들 키의 관계, 학생들의 키와 몸무게의 관계 등을 들 수 있다.

선형 회귀의 기본 가정

- 다음 식으로 표현되는 선형 회귀는 일반적으로 다음과 같은 내용을 가정

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \varepsilon_i$$

  - 종속 변수와 독립 변수들 간에 위 식과 같은 선형성이 성립한다.
  - 독립 변수 X_{ij}는 정확히 측정된 값으로 확률적으로 변하는 값이 아닌 고정된 값이다.
  - 오차 ε_i는 평균이 0, 분산이 σ^2 인 정규 분포를 따르며 모든 i에 대해 평균과 분산이 일정하다. 또, 서로 다른 i, j에 대해 ε_i, ε_j는 독립이다.
  - 독립 변수 간에는 다중 공선성(multicollinarity)이 적어야 한다.
- 다중 공선성은 회귀 모델에서 변수 간의 상관관계가 커서 한 변수를 다른 변수들의 선형 조합으로 예측할 수 있는 경우를 뜻한다.
- 다중 공선성이 존재하면 계수 $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$의 추정이 어려워진다.
- 예를 들어, $X_{i1} = aX_{i2} + bX_{i3}$이 성립한다면 선형 회귀 식에서 X_{i1} 변수의 사용이 무의미해지기 때문이다.

선형 회귀의 기본 가정

- **선형성에 대한 가정**
  - $Y_i = \beta_0 + \beta_1 X_i^2 + \varepsilon_i$와 같은 형태는 X_i^2항 때문에 비선형이라고 생각하는 경우가 종종 있다.
  - 그러나 선형 회귀에서 $X_{i1}, X_{i2}, \ldots, X_{ip}$는 고정된 값으로 가정
  - 따라서 X_{ij}에 상관없이 선형 회귀에서의 '선형성'은 파라미터 $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$의 선형조합을 의미한다.
  - 따라서 다음 둘은 선형 회귀에 해당한다.

$$Y_i = \beta_0 + \beta_1 X_{i2} + \varepsilon_i$$

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_{i1}^2 + \beta_3 X_{i2} X_{i3} + \beta_4 X_{i4}^2 + \beta_5 X_{i5} + \varepsilon_i$$

  - 그러나 다음과 같이 계수 $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$에 대해 비선형적인 모델은 선형 회귀에 해당하지 않는다.

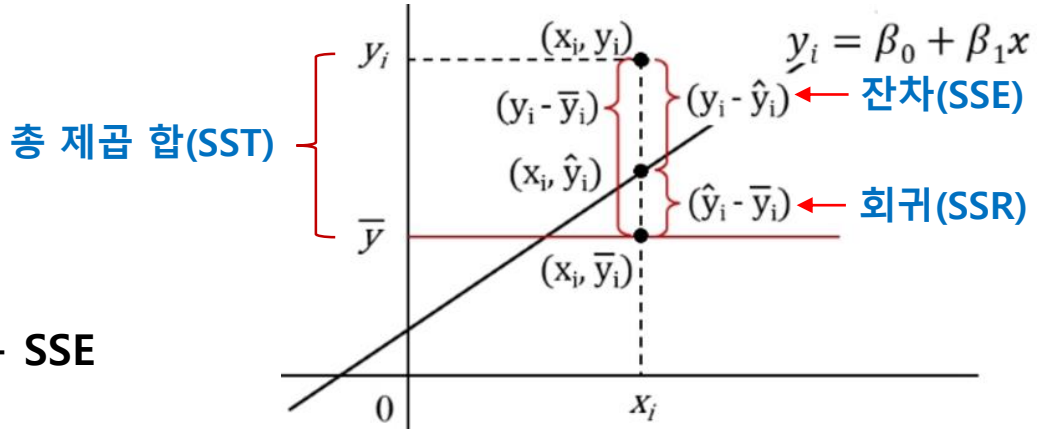
$$Y_i = \beta_0 + \beta_0 \beta_1 X_{i1} + \beta_1 X_{i2} + \varepsilon_i$$

$$Y_i = \beta_0 + \beta_0^2 X_1 + \varepsilon_i$$

선형 회귀

- 회귀직선 또는 최소제곱 직선
  - 최소 제곱법이란 제곱의 합이 최소가 되도록 값을 정하는 방법으로, 회귀 모형에서는 오차의 제곱 합(SSE) $\sum \varepsilon^2$이 최소가 되도록 회귀 계수를 정한다.

$$SST = \sum (Y_i - \overline{Y})^2 \qquad SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad SSR = \sum (\hat{Y}_i - \overline{Y})^2$$



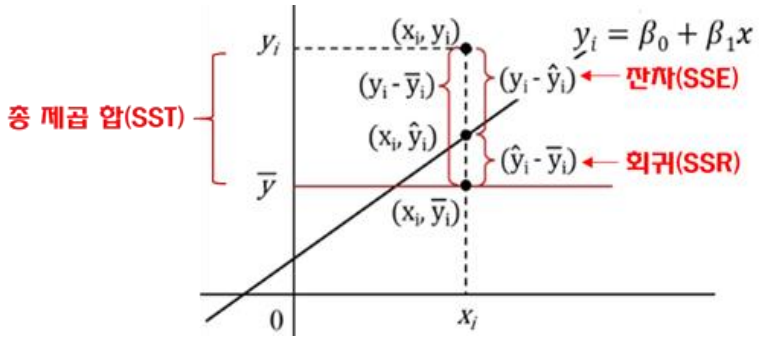
SST = SSR + SSE

※ SST(Total Sum of Squares) : 총 제곱 합
SSR(Sum of Squares due to Regression) : 회귀 제곱 합
SSE(Sum of Squares to residual Error) : 잔차(오차) 제곱 합

선형 회귀

- 결정 계수
  - 결정 계수(R-squared; R^2로 표기)는 다음과 같이 정의함

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$



$$R^2 = \frac{\text{회귀선에 의해 설명되는 제곱합}}{\text{총제곱합}}$$

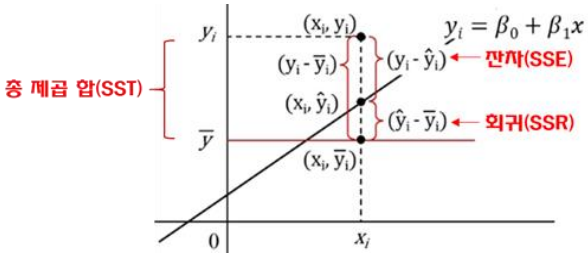
$$= 1 - \frac{\text{회귀선에 의해 설명되지 않는 제곱합}}{\text{총제곱합}}$$

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \quad (0 \le R^2 \le 1)$$

- 결정계수(R^2)는 총제곱합 중에서 회귀선으로 설명되는 제곱합의 비율임
- 결정계수(R^2)가 1에 가까울수록 설명력이 높고 바람직한 회귀선이 됨

선형 회귀

- 결정 계수



  - 결정 계수(R-squared; R^2로 표기)는 다음과 같이 정의함

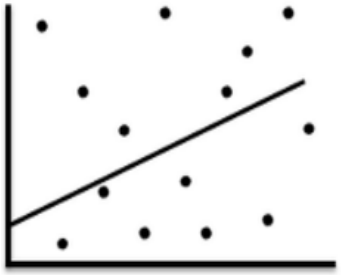
$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

$$SST = \sum (Y_i - \bar{Y})^2 \qquad SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad SSR = \sum (\hat{Y}_i - \bar{Y})^2$$

  - 위 식에서 Y_i는 i번째 종속 변수의 값, $\bar{Y}$는 종속변수 Y_i의 평균, $\hat{Y}_i$는 i번째 종속 변수의 값을 선형 회귀 모델로 예측한 값이다.
  - SST는 관측된 Y_i값이 Y_i들의 평균 $\bar{Y}$로부터 얼마 떨어져 있는지를 뜻한다.
  - SSR은 추정치 $\hat{Y}_i$가 평균 $\bar{Y}$로부터 얼마나 떨어져 있는지를 뜻한다.
  - 이 둘의 비율인 R^2 은 Y_i의 총 변동에 대비해 회귀 모델이 얼마나 그 변동성을 설명하는지를 알려준다.

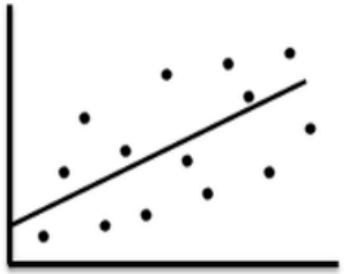
선형 회귀

- 결정 계수
  - R^2의 범위는 $0 \leq R^2 \leq 1$을 만족하며, 1에 가까울수록 회귀 모델이 데이터를 더 잘 설명한다고 할 수 있다.



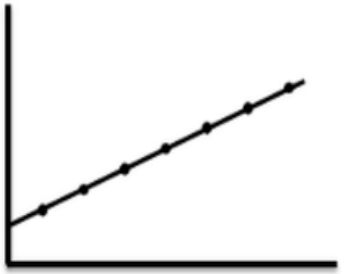
$R^2 = 0$

믿을게 못 된다



$R^2 = 0.5$

어느 정도 믿을 만 하다



$R^2 = 1$

믿을 만 하다

선형 회귀

- 결정계수와 수정 결정계수
  - 결정 계수(R^2)은 독립 변수가 늘어나면 값이 커지는 성질이 있으므로, 서로 다른 개수의 독립 변수를 사용하는 모델 간의 비교에 사용하기에는 적합하지 않다.
  - 따라서 R^2을 자유도로 나눈 수정 결정계수(Adjusted R-squared; ${R_{adj}}^2$으로 표기)를 더 많이 사용한다.

$$\mathbf{R_{adj}}^2 = 1 - \frac{\frac{\mathbf{SSE}}{(\mathbf{n-k-1})}}{\frac{\mathbf{SST}}{(\mathbf{n-1})}}$$

  - 위 식에서 SSE는 잔차 제곱 합, SST는 총 제곱 합이고, n은 데이터의 수, k는 독립변수의 수

선형 회귀

- 예측
  - 선형회귀 모델 생성 lm() 함수를 통해 모델을 만들고 나면 새로운 데이터에 대한 예측 값은 predict() 함수로 구할 수 있다.
  - predict() 함수는 일반 함수(generic function)로 여러 가지 방식으로 모델을 만들었을 때 해당 모델로부터 새로운 데이터에 대한 예측 값을 구하는 데 사용할 수 있다.
  - predict() 함수는 인자로 주어진 모델에 따라 내부적으로 predict.glm(), predict.lm(), predict.nls() 등의 함수를 부르게 되는데, 선형 회귀의 경우 predict.lm()이 호출된다.

선형 회귀

- **선형 회귀 분석의 분산분석표**

단순회귀분석의 분산분석표	• 객관적으로 도출된 회귀식이 통계적으로 유의한가를 평가하는 방법 • 회귀선의 설명력(R^2)이 아무리 높아도 통계적으로 유의하지 않으면 일반화하여 사용하기 어려움 • 분산분석에서와 같은 방법으로 회귀식의 통계적 유의성을 검정함

원 천	제곱합(SS)	자유도(df)	평균제곱(MS)	검정통계량 F
회 귀	$SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$	$(k+1)-1$	$MSR = \frac{SSR}{(k+1)-1}$	$\frac{MSR}{MSE}$
잔 차	$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	$n-(k+1)$	$MSE = \frac{SSE}{n-(k+1)}$	
총(합계)	$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$	$n-1$		여기서, n: 관측치의 수 k: 독립변수의 수

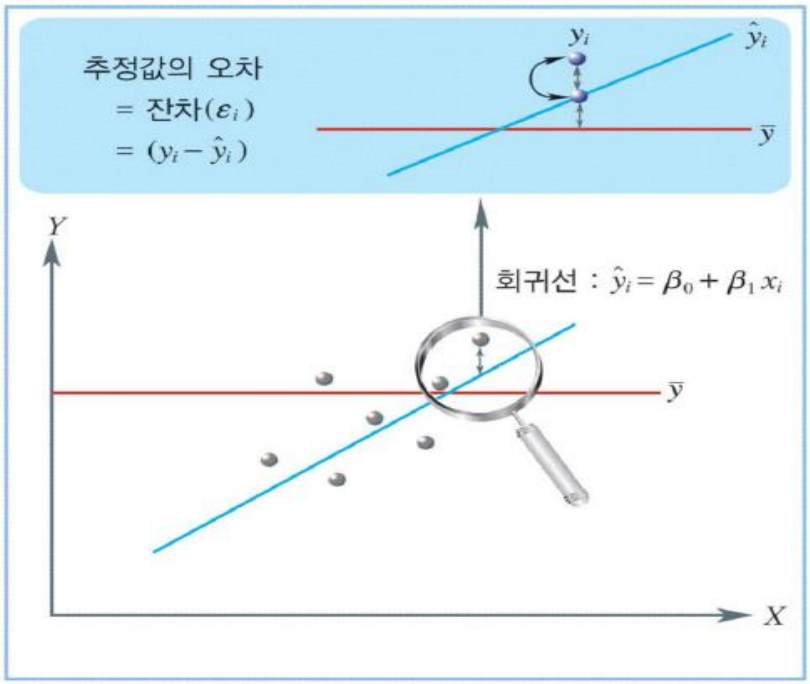
(출처: Research Methodology, 이영훈)

선형 회귀

- **추정값의 표준오차**

추정값의 표준오차

- 실제 관측치(y_i)값과 추정된 회귀선의 예측값($\hat{y}_i$)과의 차이, 즉 오차 혹은 잔차(ε_i)의 표준편차를 말함
- 잔차는 음의 값과 양의 값이 있기 때문에 서로 상충되지 않도록 잔차의 합($\sum_{i=1}^{n}\varepsilon_i$) 대신 잔차제곱합($\sum_{i=1}^{n}\varepsilon_i^2$)을 사용함



잔차제곱의 합(sum of squared error)

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}[y_i - (\beta_0 + \beta_1 x_i)]^2$$

- $\hat{y}_i$을 구하는 과정에서 β_0와 β_1을 사용하기 때문에 잔차제곱합(SSE)은 2개의 자유도를 잃게 되어 잔차제곱합(SSE)의 자유도는 $n-2$가 됨
- 잔차제곱합(SSE)을 자유도인 $n-2$로 나누어주면 잔차평균제곱(MSE)이 됨

잔차평균제곱(mean of squared error)

$$MSE = \frac{SSE}{n-2} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}$$

(출처: Research Methodology, 이영훈)

선형 회귀

- 모델 평가
  - 평균 제곱근 오차(Root Mean Square Error; RMSE)는 모델에 의해 예측된 값과 실제 환경에서 관찰되는 값의 차이를 다룰 때 많이 사용하는 측도이며 모델의 정밀도(precision)를 표현하는데 적합하다.

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$$

여기서, y_i는 관측값, $\hat{y}_i$는 추정값

  - 평균 절대 오차(Mean Absolute Error; MAE)는 두 연속형 변수 간의 차이를 측정하는 통계량이다.

$$\text{MAE} = \frac{\sum_{i=1}^{n}|y_i - x_i|}{n}$$

여기서, y_i는 추정값, x_i는 관측값

  - RMSE와 MAE는 0에 가까운 값일수록 정확도가 높은 것으로 해석한다.

선형 회귀

- 포뮬러(Formula) 해석하기

  Y1 + Y2 + ... + Yn ~ X1 + X2 + ... + Xm

  - 포뮬러이 정확한 의미는 사용하는 함수에 따라 다르나 일반적으로 "(Y1, Y2, ..., Yn)의 순서쌍을 (X1, X2, ..., Xm)의 순서쌍으로 모델링한다"고 볼 수 있다.
  - 이때 상수항은 임시적으로 허용된다.
  - Y ~ X1은 선형회귀에서 Y = a*X1 + b를 의미한다.

연산자	예	의미
+	Y ~ X1+X2	• Y를 X1, X2로 모델링. • 상수항은 암시적으로 허용 • 선형회귀에 이 포뮬러를 사용하면 Y = a*X1 + b*X2 + c를 의미
-	Y~X1-X2	• Y를 X1로 모델링하되 X2는 제외 • 특히 선형 회귀에서 Y~X1+X2-1은 Y를 X1과 X2로 모델링하되 상수항은 제외한다는 의미 • Y = a*X1 + b*X2를 의미
|	Y~X1|X2	• X2의 값에 따라 데이터를 그룹으로 묶은 후 각 그룹별로 Y~X1을 적용
:	Y~X1:X2	• Y를 X1과 X2의 상호 작용(interaction)에 따라 모델링 • 상호작용은 Y=a*X1*X2+b와 같이 X1과 X2기 동시에 Y 값에 영향을 주는 상황을 말함
*	Y~X1*X2	• Y~X1+X2+X1:X2의 축약 형 표현
위에 정리한 내용은 일반적인 해석이며 정확한 사용은 각 패키지와 함수 도움말 참고		

단순 선형 회귀

- 단순 선형 회귀(Simple Linear Regression)는 종속 변수 Y_i를 하나의 독립 변수 X_i로 설명
- 두 개 이상의 독립 변수로 설명하는 경우는 중선형 회귀(Multiple Linear Regression) 또는 다중 선형 회귀라고도 한다.
- 단순 선형 회귀 모델은 다음과 같이 표현됨

$$\mathbf{Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i}$$

  - 위 식에서 β_0는 절편, β_1은 독립 변수 X_i 의 계수이며,
  - 이들을 회귀 계수(Regression Coefficient)라 부르고, ε_i는 오차 error를 나타낸다.

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) = $\beta_0 + \beta_1 \times$ 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

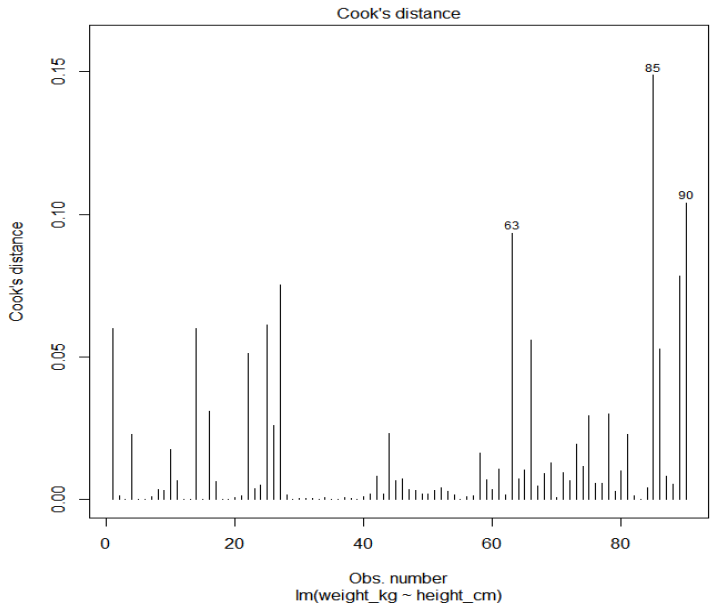
- 이상값 진단
  - 단순 선형 회귀모형에서 잔차분석은 오차 가정을 진단하는 과정이고 추가로 모형진단에서 주의해야 할 것은 이상값(outlier) 탐색이다.
  - 단순선형회귀모형에서 이상값이란 선형관계 및 오차 범위를 벗어난 값을 말한다.
  - 이상값이 발생하는 원인은 다양하고,
  - 가장 단순한 원인으로 입력 오류를 들 수 있고,
  - 만약 이상값이 존재한다면 단순 선형 회귀모형에서 계수 추정이 통계적으로 적절하지 않으며, 잔차도 마찬가지이다.
  - Cook's distance는 잔차와 영향값(influential points)으로 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

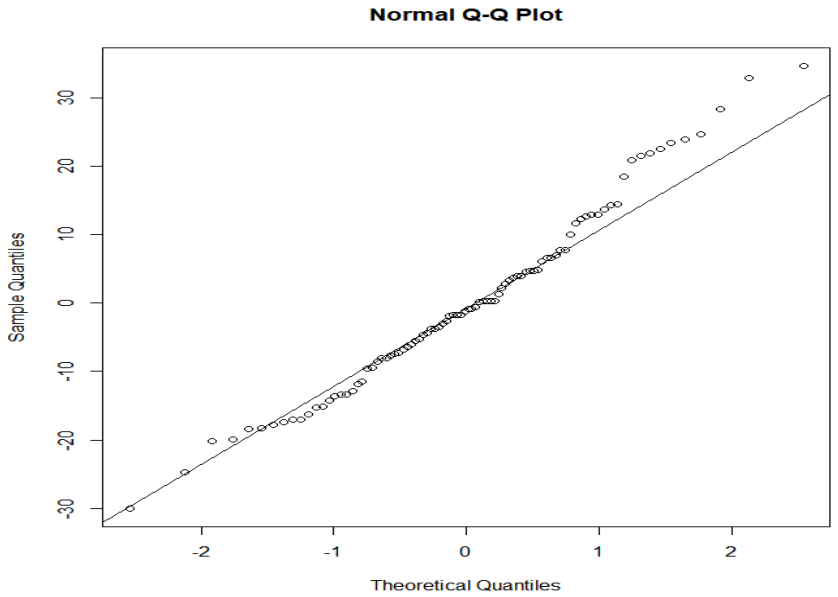
- 잔차
  - 대학생 90명의 키와 몸무게 데이터에서 몸무게(kg)~키(cm) 생성된 회귀 모델에서 잔차(residual) 구하기
    - ✓ R의 residual(model) 함수 이용한다.
    - ✓ 대학생 90명의 키와 몸무게 데이터의 1~4번째 데이터의 잔차 값과 1~4번째 학생의 적합된 값을 이용해 계산한 값과 실제 값을 비교
    - ✓ 실제 데이터 값 = 적합 된 값 + 잔차

```
# 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

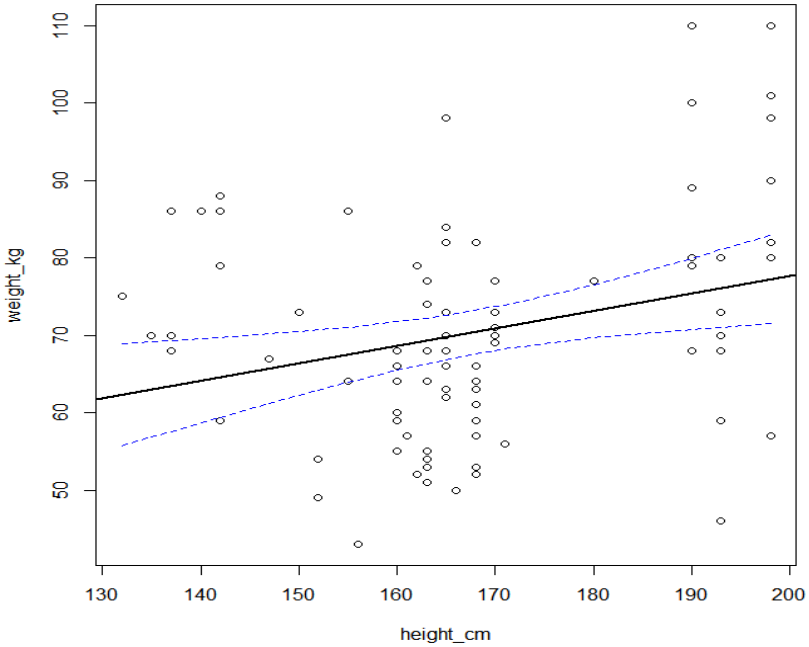
- 신뢰구간
  - ✓ R의 predict() 함수의 옵션 interval="confidence" 을 선택하고,
  - ✓ 기본적으로 유의수준은 95%이고, 옵션 level=0.99를 사용하면 유의수준은 99%를 계산할 수 있다.

```
# 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

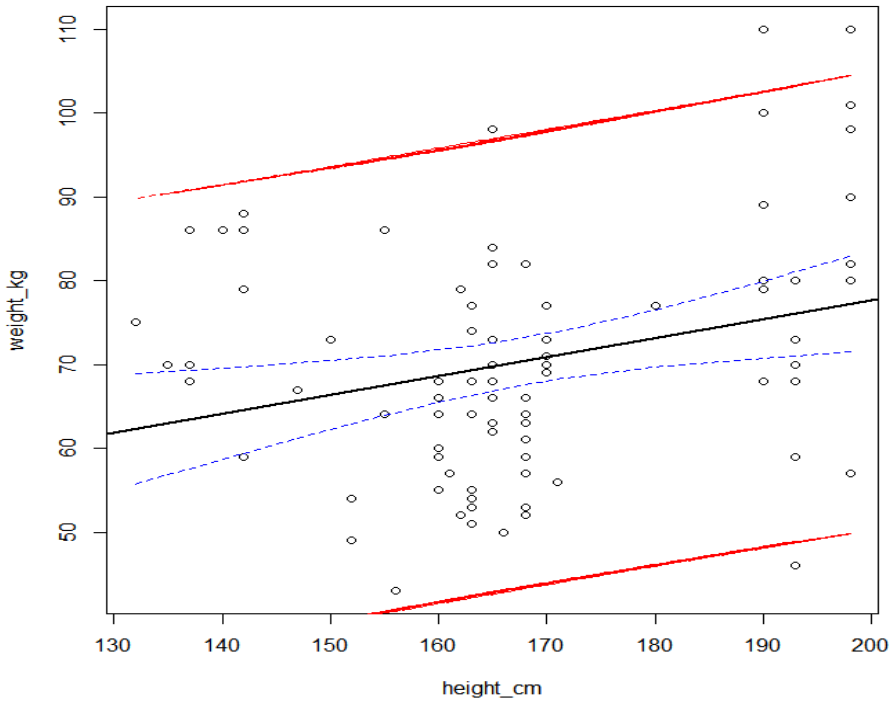
- 예측구간
  - 대학생 90명의 키와 몸무게 데이터에서 몸무게(kg)~키(cm) 생성된 회귀 모델에서 예측구간 구하기
    - ✓ R의 predict() 함수의 옵션 interval="confidence" 을 선택하고,
    - ✓ 기본적으로 유의수준은 95%이고, 옵션 level=0.99를 사용하면 유의수준은 99%를 계산 할 수 있다.

```
# 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.

다중 선형 회귀

- (예제) R의 trees 데이터를 이용해 다중 선형 회귀 수행하기

  ✓ trees 데이터에는 벚나무 31개 각각에 대해 나무의 지름(Girth), 나무의 키(Height), 목재의 부피(Volume) 3개의 숫자형 변수로 구성되어 있다.

```
str(trees)
#'data.frame':          31 obs. of  3 variables:
#$ Girth : num  8.3 8.6 8.8 10.5 10.7 10.8 11 11 11.1 11.2 ...
#$ Height: num  70 65 63 72 81 83 66 75 80 75 ...
#$ Volume: num  10.3 10.3 10.2 16.4 18.8 19.7 15.6 18.2 22.6 19.9 ...

summary(trees)
#     Girth           Height       Volume
#Min.   : 8.30   Min.   :63   Min.   :10.20
#1st Qu.:11.05   1st Qu.:72   1st Qu.:19.40
#Median :12.90   Median :76   Median :24.20
#Mean   :13.25   Mean   :76   Mean   :30.17
#3rd Qu.:15.25   3rd Qu.:80   3rd Qu.:37.30
#Max.   :20.60   Max.   :87   Max.   :77.00
```

  ✓ 가장 먼저 해야 할 일은 설명변수와 반응변수를 정하는 일이다.

  ✓ 목재상이 관심을 두는 변수는 목재의 부피(Volume)일 것이다.

  ✓ 부피는 벚나무의 지름이 클수록, 키가 클수록 클 것이다.

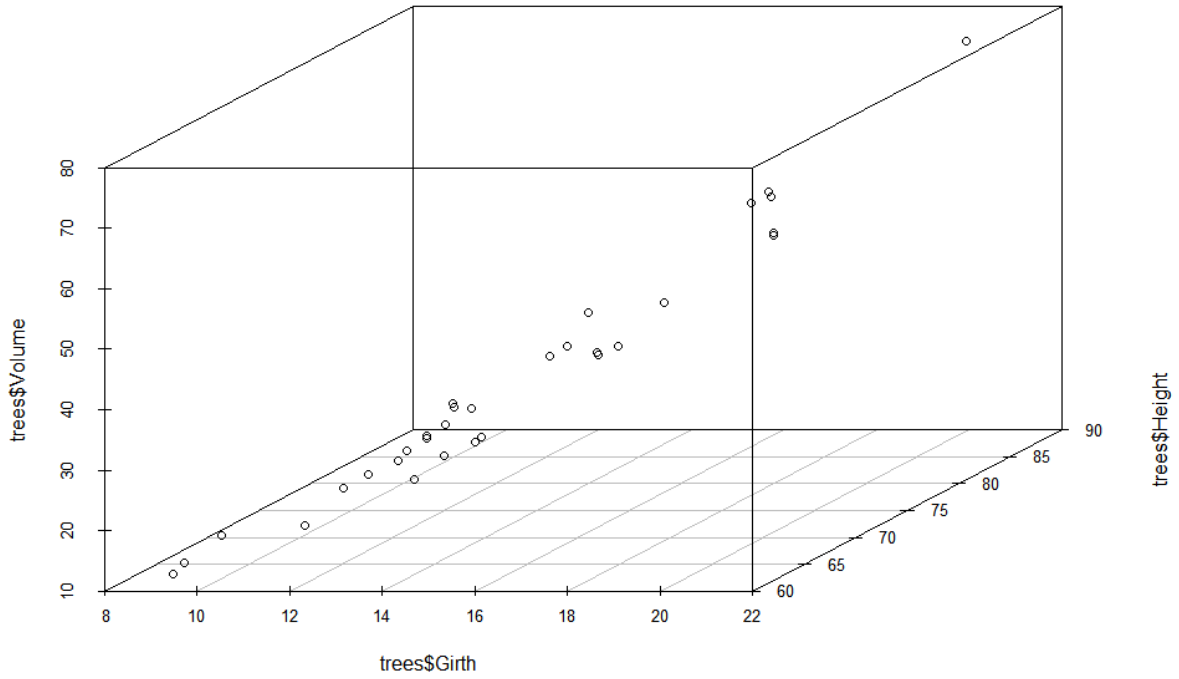
  ✓ 목재상은 지름과 키를 가지고 부피를 예측하는 일에 관심이 있을 것이다.

  ✓ 따라서 지름(Girth), 나무의 키(Height)를 설명변수, 부피(Volume)를 반응 변수로 한다.

다중 선형 회귀

- scatterplot3d() 함수로 trees 데이터의 분포를 확인하기

```
library(scatterplot3d)
scatterplot3d(trees$Girth, trees$Height, trees$Volume)
```



다중 선형 회귀

- 다중 선형 회귀 모델 생성하기

```
m <- lm(Volume ~ Girth + Height, data = trees)
m
#
#Call:
#  lm(formula = Volume ~ Girth + Height, data = trees)
#
#Coefficients:
#  (Intercept)        Girth       Height
#      -57.9877       4.7082       0.3393
```

✓ 수행 결과 부피(Volume)와 나무의 지름(Girth), 나무의 키(Height) 간의 관계는 다음과 같이 구해졌음

$$\mathbf{Volume} = -57.9877 + 4.7082 * \boldsymbol{Girth} + 0.3393 * \boldsymbol{Height} + \varepsilon_i$$

✓ 위 식에서 -57.9877는 절편,

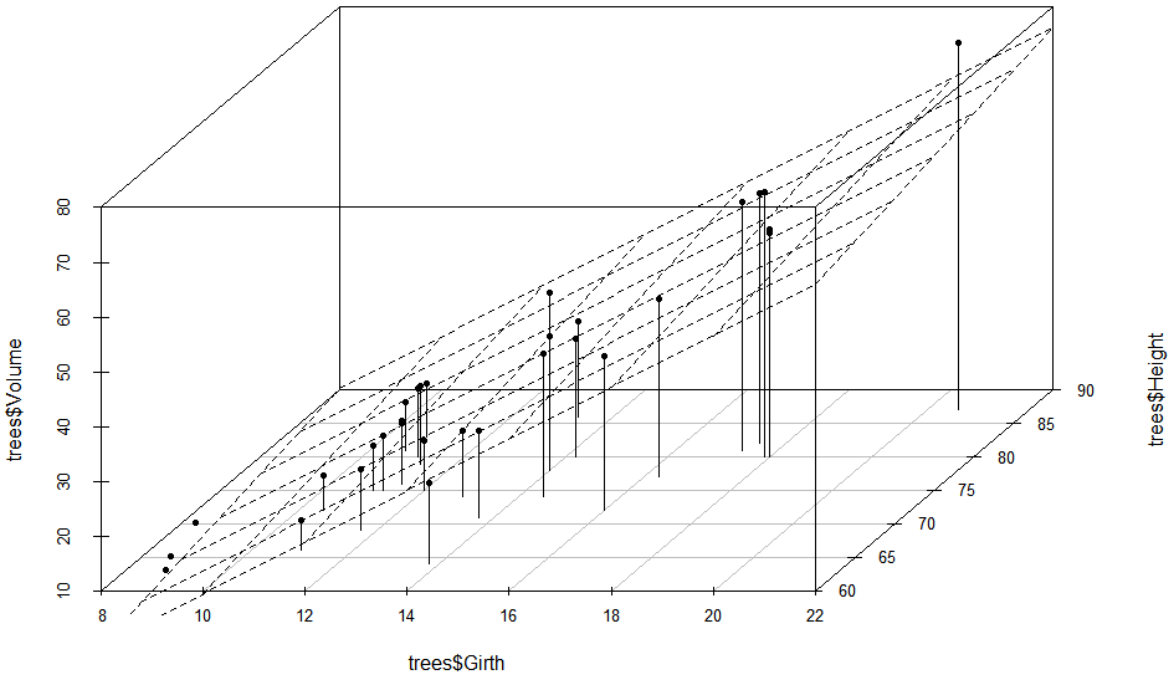
✓ -57.9877, 4.7082, 0.3393는 독립 변수(Girth, Height)의 계수(이들을 회귀계수라고 부름)이며,

✓ ε_i는 오차(error)를 나타냄

다중 선형 회귀

- trees 데이터와 회귀 모델을 중첩하여 시각화하기

```
s <- scatterplot3d(trees$Girth, trees$Height, trees$Volume, pch = 20, type = 'h',
                   angle = 55)
s$plane3d(m)
```



다중 선형 회귀

- 벗나무 세 그루의 지름과 키를 측정하여 나무의 부피를 예상하기

```
(n.data <- data.frame(Girth=c(8.5, 13.0, 19.0), Height=c(72, 86, 85)))
(n.y <- predict(m, newdata = n.data))
#         1         2         3
#6.457794 32.394034 60.303746

-57.9877 + 4.7082*8.5 + 0.3393*72
#[1] 6.4616
-57.9877 + 4.7082*13.0 + 0.3393*86
#[1] 32.3987
-57.9877 + 4.7082*19.0 + 0.3393*85
#[1] 60.3086
```

$$6.457794 = -57.9877 + 4.7082 * 8.5 + 0.3393 * 72$$

$$32.3987 = -57.9877 + 4.7082 * 13.0 + 0.3393 * 86$$

$$60.3086 = -57.9877 + 4.7082 * 19.0 + 0.3393 * 85$$

다중 선형 회귀

- 벗나무 세 그루의 지름과 키를 측정하여 나무의 부피를 예상한 결과 시각화

```
s <- scatterplot3d(c(8.5, 13.0, 19.0), c(72, 86, 85), n.y, pch = 20, type = 'h',
                   color = 'red', angle = 55)
s$plane3d(m)
```