# 강의교안 이용 안내

- 본 강의교안의 저작권은 **이윤환**과 **한빛아카데미(주)**에 있습니다.
- 이 자료를 무단으로 전제하거나 배포할 경우 저작권법 136조에 의거하여 벌금에 처할 수 있고 이를 병과(倂科)할 수도 있습니다.

한빛아카데미 Hanbit Academy, Inc.


CHAPTER 09

# 상관과 회귀

한빛아카데미 Hanbit Academy, Inc.

# Contents

## 9.1 상관계수

- 두 집단 간의 관계를 표현하는 공분산과 상관계수
- 상관계수가 나타내 는 의미

## 9.2 회귀분석

- 독립변수와 종속변수, 그리고 인과관계
- 단순선형회귀분석의 과정
- 회귀분석의 가정 확인

# 01. 상관계수

: 두 변수 간 관계의 정도

1. 두 집단 간의 관계를 표현하는 공분산과 상관계수에 대해 학습한다.
2. 상관계수가 나타내 는 의미를 학습한다.

# 상관계수

- **공분산**
  - 두 확률변수 사이의 관계를 선형관계로 나타낼 때 두 변수 사이 상관의 정도를 나타내며 다음과 같이 구합니다.

$$Cov(X,\ Y) = E[(X-E(X))(Y-E(Y))]$$
$$= E[(X-\mu_X)(Y-\mu_Y)], \quad E(X)=\mu_X,\ E(Y)=\mu_Y$$

  - 두 확률변수 $X, Y$의 공분산은 $Cov(X,Y)$로 표기하고, 공분산이 갖는 값에 따라 두 확률변수의 관계를 확인할 수 있습니다.
    - $Cov(X,Y) > 0$ : 두 확률변수 $X, Y$ 의 변화가 같은 방향임을 나타냅니다. 즉 $X$ 증가하면 $Y$ 도 증가하고, 반대로 한 변수가 감소하면 같이 감소합니다.
    - $Cov(X,Y) > 0$ : 두 확률변수 $X, Y$ 의 변화가 같은 방향임을 나타냅니다. 즉 $X$ 증가하면 $Y$ 도 증가하고, 반대로 한 변수가 감소하면 같이 감소합니다.
    - $Cov(X,Y) = 0$ : 두 확률변수 간에 어떠한 (선형) 관계가 없음을 나타냅니다.

# 상관계수

- **상관계수**
  - 두 확률변수 $X, Y$의 공분산을 각 확률변수의 표준편차의 곱으로 나눈 값을 (모)상관계수라 하고, 기호로 $\rho_{XY}$ (혹은 $\rho$)로 나타냅니다.

$$\rho_{XY} = \frac{Cov(X,\ Y)}{\sigma_X \sigma_Y} = \frac{E[(X - E(X))(Y - E(Y))]}{\sigma_X \sigma_Y}$$

  - (모)상관계수는 -1부터 1사이의 값을 가집니다.
    - 공분산의 경우 자료의 단위에 따라 값의 크기가 일정하지 않아 비교하기 힘듦니다.
    - 공분산의 성질을 그대로 이어 받아 두 변수 간의 변화의 방향이 같으면 양수, 반대이면 음수를 갖습니다.
  - (모)상관계수는 모집단의 특성 중에 하나로 일반적으로 알 수 없으며, 두 확률변수로부터 추출한 표본의 특성을 통해 구하는 (피어슨의) 표본상관계수를 이용하여 추정합니다.
    - 표본상관계수를 구하기 위해 먼저 표본공분산을 구해봅시다.

# 상관계수

- 표본공분산
  - 두 확률변수 X, Y로 부터 추출한 n개의 표본 쌍 $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$에서 확률변수 X로 부터 추출한 표본 $x_1, x_2, \dots, x_n$의 평균을 $\bar{x}$, 표준편차를 $s_x$, 확률변수 Y로 부터 추출한 표본 $y_1, y_2, \dots, y_n$의 평균을 $\bar{y}$, 표준편차를 $s_y$라 하면, 표본공분산 $cov(x, y)$는 다음과 같이 두 표본의 편차의 곱을 모두 합하고 이를 자료의 개수(표본 쌍의 개수) – 1 로 나누어 구합니다.

$$cov(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

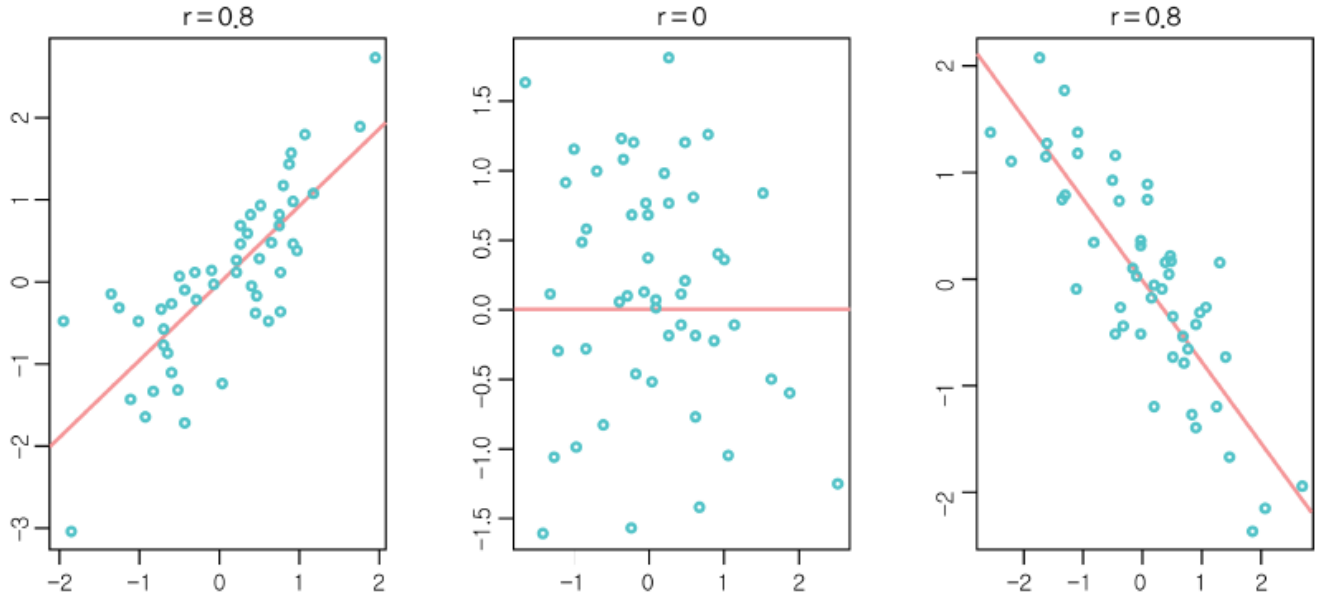
- 표본상관계수
  - 표본공분산을 각 표본의 표준편차의 곱으로 나누어 구합니다.

$$r = \frac{cov(x, y)}{s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)\sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}\sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}}$$

$$= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

# 상관계수

- 표본상관계수는 모상관계수와 동일한 성질을 가져
  - -1 혹은 1에 가까울수록 강한 상관을 나타내고,
  - 0에 가까이 갈수록 약한 상관을 나타냅니다.
  - 양수일 경우 두 변수의 값의 변화는 같은 방향으로 진행되고, 음수일 경우 값의 변화는 서로 반대가 됩니다.



# 상관계수

## 예제 9-1 아버지와 아들 키의 공분산과 상관계수

준비파일 | 02.correlation.R

- 8장에서 준비한 아버지와 아들의 키 자료로부터 아버지와 아들의 키의 공분산과 상관계수를 구해봅니다.

```
6.  f.mean <- mean(hf.son$Father)
7.  s.mean <- mean(hf.son$Height)
8.  cov.num <- sum( (hf.son$Father-f.mean) * (hf.son$Height - s.mean) )
9.  (cov.xy <- cov.num / (nrow(hf.son) - 1))
10. cov(hf.son$Father, hf.son$Height)
11.
12. (r.xy <- cov.xy / (sd(hf.son$Father) * sd(hf.son$Height)))
13. cor(hf.son$Father, hf.son$Height)
```

# 상관계수

- 표본공분산 구하기
  - 6, 7줄 : 표본공분산 계산을 위해 아버지의 키(hf.son$Father)의 평균을 변수 f.mean에, 아들의 키(hf.son$Height)의 평균을 변수 s.mean에 저장합니다.
  - 8줄 : 두 변수의 편차의 곱을 전부 합한 값을 변수 cov.num에 저장합니다.
  - 9줄 : 위에서 구한 편차 곱의 합을 (자료의 개수-1)로 나누고 변수 cov.xy에 저장한 후 출력합니다.
  - 10줄 : R에서는 **cov()** 함수를 이용하여 표본공분산을 구합니다.
    9줄에서 직접 구한 표본공분산과 비교하면 동일한 값임을 확인할 수 있습니다.

```
> (cov.xy <- cov.num / (nrow(hf.son) - 1))
[1] 2.368441

> cov(hf.son$Father, hf.son$Height)
[1] 2.368441
```

# 상관계수

- 표본 상관계수를 구합니다.
  - 12줄 : (9줄 혹은 10줄에서 구한) 표본공분산을 두 변수의 표본표준편차의 곱으로 나누어 표본상관계수를 구하고, 이를 변수 cov.xy에 저장한 후 출력합니다.
  - 13줄 : R의 표본상관계수 함수는 **cor()**입니다. 12줄에서 직접 구한 표본상관계수와 비교하면 동일한 값임을 확인할 수 있습니다

```
> (r.xy <- cov.xy / (sd(hf.son$Father) * sd(hf.son$Height)))
[1] 0.3913174

> cor(hf.son$Father, hf.son$Height)
[1] 0.3913174
```

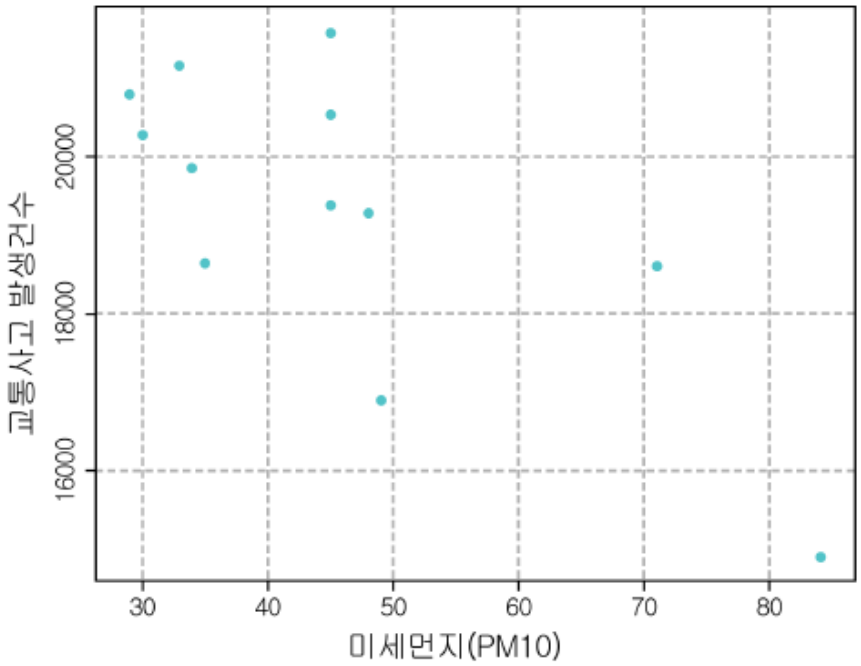
# 02. 회귀분석

## : 대표적인 통계 모형

1. 독립변수와 종속변수를 구별하고 인과관계에 대해 학습한다.
2. 통계적 모형 구축의 예로 단순선형회귀분석의 과정을 학습한다.
3. 회귀분석의 가정을 만족하는지 확인하는 방법에 대해 학습한다.

# 독립변수와 종속변수

- **인과관계**
  - 원인과 결과 관계를 뜻하는 인과관계는 상관관계처럼 계산을 통해 구하는 것이 아닌, 주의 깊은 자료의 관찰을 통해 얻을 수 있는 관계입니다.
  - 자료에 대한 깊은 통찰이 없다면 잘못된 인과관계를 도출할 수 있습니다.
  - 다음의 도표를 살펴봅시다.



# 독립변수와 종속변수

- 앞선 도표는 미세먼지(PM-10) 농도에 따라 교통사고 발생이 어떤 연관이 있는지 알아보고자 작성해본 도표입니다.
- 전반적으로 미세먼지 농도가 짙어질 수록 교통사고 발생은 줄어드는 경향을 보이고 있습니다.
- 이를 통해 미세먼지가 증가하면 교통사고 건수가 줄어든다고 할 수 있을까요?

## 관찰연구

- 사회현상은 관찰연구를 통해 연구하는 경우가 많습니다. 이 경우에는 실험을 통제할 수 없음을 인정하고, 사전지식과 사회에 대한 깊은 통찰력을 가져야 합니다.
- 다음을 고민해 봅시다.
  - 두 변수의 연관성
  - 원인과 결과에 대한 고민
  - 제3의 요인

# 독립변수와 종속변수

- 사례 (위키피디아 참조)
  - 아이스크림 판매량이 증가할수록 익사사고 발생이 증가하였다. 즉 익사사고 발생을 억제하기 위해 아이스크림의 판매를 금지해야 한다(제3의 요인 : 계절).
  - 불을 켜고 자는 어린이의 경우, 나이가 들어 근시가 될 경우가 많다. 즉 근시를 예방하기 위해 어릴 때부터 잠을 잘 때 불을 켜지 말아야 한다(제3의 요인 : 부모의 근시).
  - 국가 부채가 GDP의 90% 이상이 될 경우 국가의 성장률이 느려진다. 즉 높은 국가 부채는 국가의 성장을 느리게 한다(뒤바뀐 인과관계)
  - 사과의 수입이 증가할수록 이혼률이 증가한다. 즉이혼률을 낮추기 위해 사과 수입을 금지한다(인과관계를 확인할 수 없는 두 변수)
- 변수 간의 관계에서
  - 다른 변수에 의해 영향을 받아 그 값이 결정되는 변수를 종속변수,
  - 영향을 미치는 변수를 독립변수라고 합니다.

# 단순선형회귀분석

- **단순선형회귀모형**
  - 두 확률변수 X, Y에서 X가 독립변수이고, Y가 종속변수일 경우 독립변수 X의 개별값 $x_1, x_2, \dots, x_n$ 에 대응하는 종속변수 Y의 관찰값 $y_1, y_2, \dots, y_n$에 대해 다음과 같은 모형을 단순선형회귀모형이라고 합니다.

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad i = 1, 2, \dots, n, \epsilon_i \sim N(0, \sigma^2)$$

  - 회귀계수
    - 위의 식에서 두 상수 $\beta_0, \beta_1$을 (모집단)회귀계수라 하는데, 이는 각각 직선의 방정식에서 절편과 기울기의 역할을 합니다.
    - 두 상수는 미지의 모수로, 표본으로부터 추정을 통해 구합니다.
      - 추정된 회귀계수를 이용하여 구한 식으로 나타나는 직선을 추정된 회귀직선이라고 합니다.

# 단순선형회귀분석

- 회귀계수의 추정
  - 앞선 회귀식을 $\epsilon_i$ 에 대해 정리하면,

$$\epsilon_i = y_i - \beta_0 - \beta_1 x_i, \qquad \epsilon_i \sim N(0, \sigma^2)$$

이 되고, $\epsilon_i$를 오차 혹은 오차항이라고 합니다.

회귀계수의 추정은 오차들의 제곱합을 이용하여 구합니다.

$$S = \sum_{i=1}^{n} {\epsilon_i}^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

  - 최소제곱법과 최소제곱추정량
    - 회귀계수추정의 한 방법으로 오차들의 제곱합을 최소로 하는 $\beta_0, \beta_1$의 추정량인 $b_0, b_1$를 구하는 최소제곱법을 통해 구한 추정량을 최소제곱추정량이라 합니다.

$$b_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \qquad b_0 = \bar{y} - b_1\bar{x} = \frac{1}{n}\sum_{i=1}^{n} y_i - b_1 \frac{1}{n}\sum_{i=1}^{n} x_i$$

# 단순선형회귀분석

- 추정된 회귀직선
  - 회귀계수에 대한 추정량 $b_0, b_1$ 과 종속변수 Y의 예측값을 $\hat{y}$이라 하면, 추정된 회귀직선은 다음과 같습니다.

$$\hat{y} = b_0 + b_1 x$$

    - 또한 $b_0 = \bar{y} - b_1 \bar{x}$ 이므로 $\hat{y} = b_0 + b_1 x = \bar{y} - b_1 \bar{x} + b_1 x = \bar{y} + b_1 (x - \bar{x})$
  - 추정된 회귀직선을 통해 독립변수가 가질 수 있는 값에 대응하는 종속변수의 값을 추측할 수 있습니다.

# 단순선형회귀분석

## 예제 9-2 아버지와 아들 키 자료로부터 회귀계수 추정

준비파일 | 04.lse.R

- 아버지와 아들의 키 자료로부터 회귀계수를 추정하고, 이로부터 추정된 회귀직선을 구해 봅시다.

```
6. mean.x <- mean(hf.son$Father)
7. mean.y <- mean(hf.son$Height)

8. sxy <- sum((hf.son$Father - mean.x)*(hf.son$Height - mean.y))
9. sxx <- sum((hf.son$Father - mean.x)^2)

10.( b1 <- sxy / sxx )
11.( b0 <- mean.y - b1 * mean.x )
```

# 단순선형회귀분석

- 아버지와 아들의 키의 각각의 평균을 구합니다.
  - 6줄 : 아버지의 키의 평균을 구해 변수 mean.x 에 저장합니다($\bar{x}$).
  - 7줄 : 아들의 키의 평균을 구해 변수 mean.y 에 저장합니다 ($\bar{y}$).
- 편차들을 계산합니다.
  - 9줄 : 아버지의 키의 편차와 아들의 키의 편차들의 곱의 합을 변수 sxy에 저장합니다($S_{xy}$).
  - 10줄 : 아버지의 키의 편차제곱합을 구해 변수 sxx에 저장합니다($S_{xx}$).
- 회귀계수의 추정치를 구합니다.
  - 12줄 : 추정량 $b_1$을 구하기 위해 9줄에서 구한 sxy를 10줄에서 구한 sxx로 나눈 값을 변수 b1에 저장하고 출력합니다.
  - 13줄 : 추정량 $b_0$를 구하기 위해 아들의 키의 평균에서 12줄에서 구한 추정량 $b_1$의 추정치가 저장된 변수 b1과 아버지의 키의 곱을 뺀 값을 변수 b0에 저장하고 출력합니다.

# 단순선형회귀분석

```
> ( b1 <- sxy / sxx )
[1] 0.4477479

> ( b0 <- mean.y - b1 * mean.x )
[1] 38.25891
```

- 여기서 구한 회귀계수의 추정치를 통해 추정된 회귀직선의 식은 다음과 같습니다.

$$\hat{y} = 38.259 + 0.448x$$

- **R의 함수를 이용하여 회귀계수를 추정해 봅시다.**
  - R에서는 선형모형을 의미하는 lm() 함수를 통해 회귀계수를 추정할 수 있습니다.

```
17.lm(Height ~ Father, data=hf.son)
```

  - 17줄 : R에서 종속변수와 독립변수를 나타내는 수식은 '종속변수 ~ 독립변수'로 나타내고, lm() 함수는 이를 첫 번째 전달인자로 사용하여 회귀계수를 구한 결과를 반환합니다.

# 단순선형회귀분석

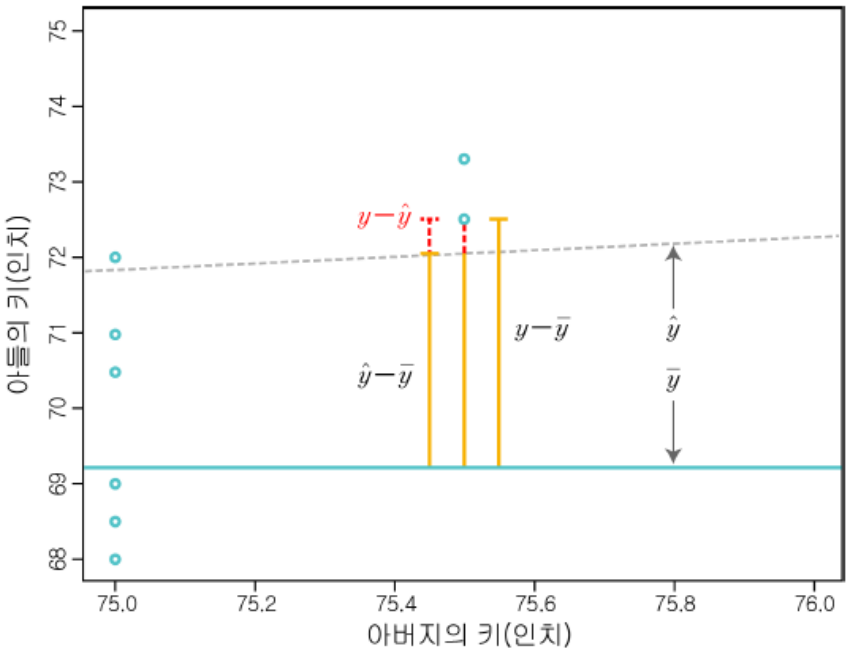
```
> lm(Height ~ Father, data=hf.son)

Call:
lm(formula = Height ~ Father, data = hf.son)

Coefficients:
(Intercept)       Father
    38.2589       0.4477
```

- 출력 결과에서 (Intercept)는 추정량 $b_0$의 추정치를 나타내고, 독립변수의 이름인 Father는 추정량 $b_1$의 추정치를 나타냅니다.
- 앞서 12, 13줄에서 최소제곱추정량인 식을 직접 구한 결과와 같음을 확인할 수 있습니다.

# 단순선형회귀분석

- **모형의 유의성**
  - 일원분산분석에서와 같이 회귀에서도 전체 변동량의 구성을 분석해봅시다.
  - 총편차
    - 종속변수의 추정치와 실제 값의 차이 : $y - \bar{y}$
    - 총편차는 다음과 같이 분해할 수 있습니다.

$$y - \bar{y} = (y - \hat{y}) + (\hat{y} - \bar{y})$$



# 단순선형회귀분석

- 앞서 총편차의 제곱합으로 전체 변동량을 구하고 SST로 표기합시다.
- 전체 변동량은 다음과 같이 두 부분의 제곱합으로 분해할 수 있습니다.

$$\sum_{i=1}^{n}(y_i-\bar{y})^2 = \underbrace{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}_{②} + \underbrace{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}_{①}$$

  - ① 에서 $\hat{y}_i-\bar{y}$ 는 추정값($\hat{y}_i$)과 전체의 평균($\bar{y}$)과의 차이를 나타내며, 이들의 제곱합인 ① 을 회귀제곱합이라 부르고 SSR로 나타냅시다.
  - ② 에서 $y_i-\hat{y}_i$ 는 관찰값($y_i$)과 추정값($\hat{y}_i$) 과의 차이를 나타내며, 이를 $e_i$로 표기하고 잔차라고 합니다. 잔차들의 제곱합인 ②를 오차제곱합이라 부르고 SSE로 나타냅시다.
- 오차제곱합
  - 앞서 오차항은 서로 독립이고, 평균이 0, 분산이 $\sigma^2$인 정규분포를 따르는 것으로 가정하였습니다.
  - 여기서 오차항의 분산 $\sigma^2$은 미지의 모수로서 추정의 대상이 됩니다. 오차항의 분산 $\sigma^2$의 추정량으로, 오차제곱합을 자유도로 나눈 평균제곱오차(MSE)를 사용합니다.

# 단순선형회귀분석

- 오차제곱합의 자유도는 (표본의 개수 - 회귀계수)의 개수로 구하고, 단순선형회귀분석에서는 회귀계수의 개수가 2이므로 (n-2)가 오차제곱합의 자유도가 됩니다.
- 오차항의 분산 $\sigma^2$의 추정량을 $\widehat{\sigma^2}$으로 나타낼 때,

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

- 추정값의 표준오차
  - 오차항의 표준편차에 대한 추정에서 위에서 구한 평균제곱오차의 제곱근($\sqrt{MSE}$)을 추정량으로 사용하고, 이를 독립변수를 통해 종속변수를 추정할 때 추정값의 표준오차라고 합니다.
  - 이 표준오차가 작은 모형이 좋은 모형이 됩니다.

# 단순선형회귀분석

- 회귀모형의 유의성 검정
  - 회귀모형이 타당한지를 검정합니다.
  - 가설수립
    - 영가설 : 종속변수와 독립변수 간 선형관계가 없다. ($H_0: \beta_1 = 0$)
    - 대안가설 : 종속변수와 독립변수 간 선형관계가 있다. ($H_1: \beta_1 \neq 0$)
  - 검정통계량
    - 추정값의 표준오차가 큰지 작은지를 나타내기 위해 다른 값과 비교하는데, 이때 앞서 전체 변동량을 구성하는 두 요소 중 하나인 회귀제곱합을 이용해 비교합니다.
    - 회귀제곱합은 자유도로 (회귀계수의 개수-1)을 가지며, 단순선형회귀분석의 경우 회귀계수는 $\beta_0, \beta_1$의 두 개이므로 1이 됩니다.
    - 회귀제곱합을 그들의 자유도로 나눈 것을 회귀의 평균제곱합(MSR)이라고 하며, 이를 MSE와 비교하여 회귀모형의 유의성을 검정합니다.

# 단순선형회귀분석

| 요인 | 제곱합 | 자유도 | 평균제곱합 | F(df1, df2) |
|---|---|---|---|---|
| 회귀 | SSR | 1<br>(회귀계수의 수−1) | $MSR=\frac{SSR}{1}$ | $\frac{MSR}{MSE}$ |
| 잔차 | SSE | $n-2$<br>($n$−회귀계수의 수) | $MSE=\frac{SSE}{n-2}$ | |
| 합 | SST | $n-1$ | | |

- 이상의 과정을 분산분석에 학습한 것과 유사한 분산분석표로 요약해 볼 수 있으며, 검정통계량은 $\frac{MSR}{MSE} \sim F(df_1, df_2)$ 입니다.

# 단순선형회귀분석

## 예제 9-3 회귀모형의 유의성 검정

준비파일 | 06.regression.R

- 아버지와 아들의 키 자료를 이용하여 회귀계수 추정을 통해 구축된 회귀모형의 유의성을 검정합시다.
  - R의 함수를 이용하여 분산분석표를 작성하고, 이로부터 얻어지는 검정통계량으로 회귀모형의 유의성을 검정합니다.

```
6. out <- lm(Height ~ Father, data=hf.son)
7. anova(out)
```

  - 회귀모형을 구축합니다.
    - 6줄 : 앞서 회귀계수의 추정에서 구한 것과 같이 회귀모형을 구축하고 결과를 out에 저장합니다.

# 단순선형회귀분석

- 회귀의 분산분석표를 출력하고 검정통계량을 구합니다.
  - 7줄 : anova() 함수는 구축된 회귀모형으로부터 분산분석표를 만들어줍니다. 전달 인자로 6줄에서 구한 변수 out(R로 구한 회귀모형을 담고 있는 변수)을 사용합니다.

```
> anova(out)
Analysis of Variance Table
Response: Height
           Df  Sum Sq Mean Sq F value    Pr(>F)
Father      1  492.06  492.06  83.719 < 2.2e-16 ***
Residuals 463 2721.28    5.88
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

# 단순선형회귀분석

- 검정통계량은 자유도가 1과 463인 F-분포에서 83.719로 나옵니다.
  - 유의확률로 계산된 값 '< 2.2e-16'은 $2.2 * 10^{-16}$보다 작음을 나타내며, 이값은 거의 0에 가까울 정도로 작은 값입니다.
  - 유의확률 옆에 세 개의 '*'가 표시되는데, 이것이 의미하는 바는 출력물의 맨 아랫줄에 나옵니다. '***'인 경우 유의수준0.001에서 유의함을 나타내며. 당연히 이는 0.001보다 큰 유의수준(0.01, 0.05 등) 에서도 유의합니다.

▫ 결정계수($R^2$) : 모형의 성능

- 모형이 얼마나 효율적인지 나타내는 통계량입니다.
- 결정계수는 0부터 1 사이의 값을 가지며, 1에 가까울 수록 회귀모형의 성능이 좋은 것으로 봅니다.
  - 모형의 성능은 전체 변동 중에서 회귀모형에 의해 설명되는 변동의 비율로 평가합니다.

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

# 단순선형회귀분석

- 회귀계수의 유의성 검정
  - 독립변수가 종속변수에 대해 선형관계로 나타낼 수 있는지를 검정합니다.
  - 종속변수와 선형관계를 나타내는 회귀계수 $\beta_1$에 대한 추정을 실시해봅시다.
  - 가설 수립
    - 영가설 : 독립변수는 종속변수와 선형관계를 갖지 않는다.
    - 영가설 : 독립변수는 종속변수와 선형관계를 갖는다.
    - 양의 관계, 음의 관계를 갖는지는 한쪽검정을 통해 실시 할 수 있습니다.
  - 검정통계량 : 검정통계량은 영가설 하에서 자유도가 n-2인 t-분포를 따릅니다.

$$t = \frac{b_1 - \beta_1}{\sqrt{\frac{SSE}{n-2}}/\sqrt{S_{xx}}} = \frac{b_1 - \beta_1}{\sqrt{\frac{MSE}{S_{xx}}}} \sim t(n-2)$$

# 단순선형회귀분석

## 예제 9-4 회귀계수의 유의성 검정

준비파일 | 06.regression.R

- 아버지와 아들의 키 자료를 이용하여 구축된 결과를 담고 있는 변수 out에는 좀 더 많은 정보들이 함께 들어가 있습니다.
- 이 정보들은 전체 회귀분석 과정의 결과물을 담고 있습니다.
- 이를 확인하기 위해 만능 함수 summary()를 사용합니다.
  - 함수 summary()는 주어진 전달인자에 맞춰 각종 결과를 요약해 줍니다.
  - 데이터 프레임의 각 변수들을 요약할 때도 summary()를 사용하였습니다.

```
9. summary(out)
```

- 9줄 : 회귀모형 구축의 정보를 요약해 줍니다.

# 단순선형회귀분석

```
> summary(out)
Call:
lm(formula = Height ~ Father, data = hf.son)

Residuals:
    Min      1Q  Median      3Q     Max
-9.3774 -1.4968  0.0181  1.6375  9.3987

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 38.25891    3.38663   11.30   <2e-16 ***        ①
Father       0.44775    0.04894    9.15   <2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.424 on 463 degrees of freedom
Multiple R-squared:  0.1531, Adjusted R-squared:  0.1513   ②
F-statistic: 83.72 on 1 and 463 DF,  p-value: < 2.2e-16    ③
```

# 단순선형회귀분석

① 각 회귀계수의 추정과 검정에 해당합니다.

- 위의 예에서 아버지의 키에 대한 회귀계수 $\beta_1$의 추정값 0.44775에 대해 검정을 실시한 결과, 검정통계량(t)이 9.15, 유의확률이 '< 2.2e-16'보다 작은, 즉 거의 0이라고 볼 수 있습니다. 따라서 유의수준 0.001에서도 유의함을 나타냅니다.

② 결정계수 $R^2$을 출력합니다.

- 결정계수는 앞서 학습한 전체 변동량 중 회귀에 의한 변동량의 비율을 그대로 사용하는 Multiple R-squared와 수정결정계수(Adjusted R-squared)를 출력해줍니다.
- 단순선형회귀분석에서 결정계수 $R^2$은 앞서 살펴본 (표본)상관계수 r의 제곱과 같습니다.

③ 모형의 유의성을 분산분석표 없이 검정통계량과 유의확률만으로 표시하여, 연구자가 판단하도록 하고 있습니다.

# 단순선형회귀분석

- 평균반응의 구간추정
  - 회귀계수를 이용하여 독립변수의 특정한 값 $x_0$에 대한 예측값은 종속변수의 평균에 대한 예측값이 되며, 이는 다음과 같이 나타냅니다.

$$\hat{E}(y|x = x_0) = b_0 + b_1 x$$

  - 앞서 구한 아버지와 아들의 키에서 추정된 두 회귀계수를 이용하여 아버지의 키가 74.5인치일 때 아들의 키의 평균을 예측하면 다음과 같습니다.

$$\hat{E}(y|x = 74.5) = 38.259 + 0.448 \times 74.5 \approx 71.635$$

  - 종속변수의 예측값에 대한 신뢰구간을 구해봅시다.
    - 평균반응에 대한 표준오차는 다음과 같이 알려져 있습니다.

$$\sqrt{MSE\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}$$

# 단순선형회귀분석

- 이로부터 평균반응의 표본분포는 MSE의 자유도를 따르는 t-분포를 따르는 것으로 알려져 있으며, 다음과 같습니다.

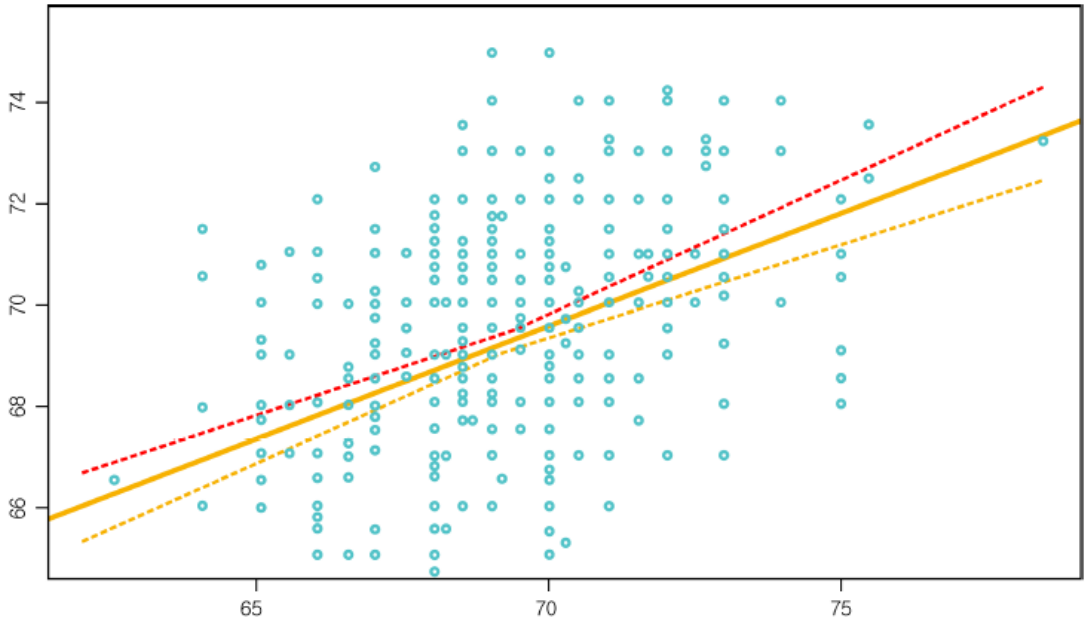
$$t = \frac{(b_0 + b_1 x_0) - \mu_0}{\sqrt{MSE(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}})}} \sim t(n-2), \qquad \mu_0 = \beta_0 + \beta_1 x_0$$

- 또한 100(1-$\alpha$)% 신뢰구간은 다음과 같이 구할 수 있습니다.

- $(b_0 + b_1 x_0) \pm t_{\alpha/2}\sqrt{MSE(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}})}$

- 이런 복잡한 계산을 통해 전체 표본으로부터 관찰된 독립변수에 따라 신뢰구간을 다음과 와 같이 그려볼 수 있습니다.

# 단순선형회귀분석



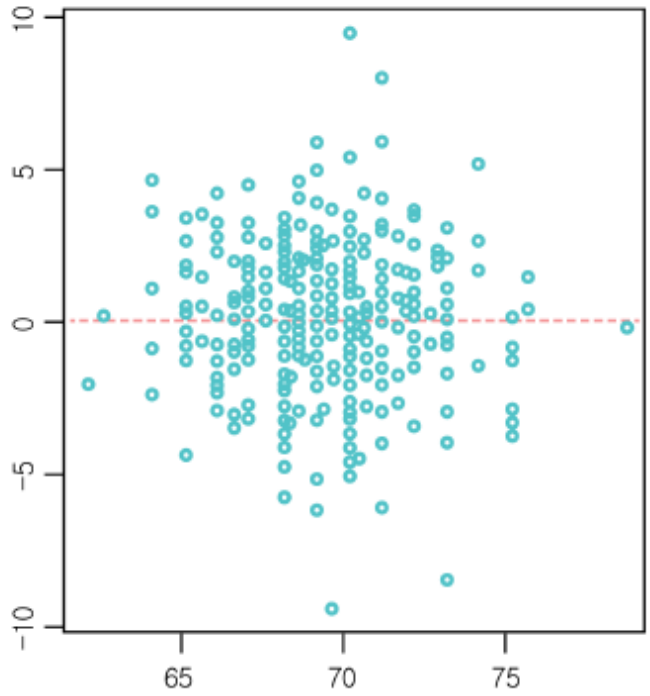

- 점선으로 표시된 선이 아버지의 키로부터 추정한 아들의 키의 평균에 대한 95% 신뢰구간입니다.
- 아버지의 키의 평균에서 그 폭이 가장 좁고, 점점 멀어질수록 폭이 넓어집니다.
- 만일 모형을 만드는 데 사용한 표본 값이 범위를 벗어난다면, 아들의 키의 평균에 대한 신뢰구간은 더 넓어져 정확한 예측을 할 수 없게 됩니다.

# 회귀분석의 가정 확인

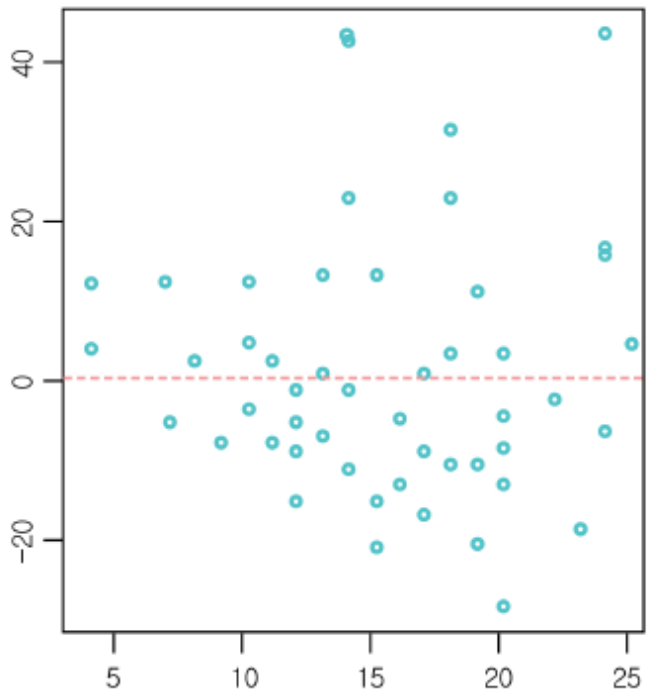
- **오차에 대한 가정**
  - 종속변수와 독립변수 간의 연관의 관계를 회귀모형(직선식)으로 추정하여 나타내고, 이를 이용해 종속변수에 대한 예측을 실시할 수 있습니다.
  - 회귀분석을 통해 모형을 구축하기 위해서는 기본적인 가정들을 만족해야 합니다.
  - 회귀분석을 위한 가정 : 오차($\varepsilon_i$)에 대한 가정
    - 독립성 : 오차들은 서로 독립이다.
    - 동일분산성 : 오차들의 분산은 $\sigma^2$으로 모두 동일하다.
    - 정규성 : 오차는 평균이 0이고 분산이 $\sigma^2$ 인 정규분포를 따른다.
  - 오차는 모집단에서의 변동을 나타내는 것으로 우리가 관찰할 수 없는 자료입니다. 그렇기에 오차의 추정량인 잔차($e_i$)를 이용하여 위의 가정에 대해 분석합니다.

# 회귀분석의 가정 확인

- 잔차와 독립변수와의 산점도
  - 비교를 위해 R의 내장자료인 cars 자료를 이용하여 자동차의 속도와 제동거리에 대한 회귀모형을 사용하였습니다.




(a) 아버지와 아들의 키




(b) 자동차의 속도와 제동거리

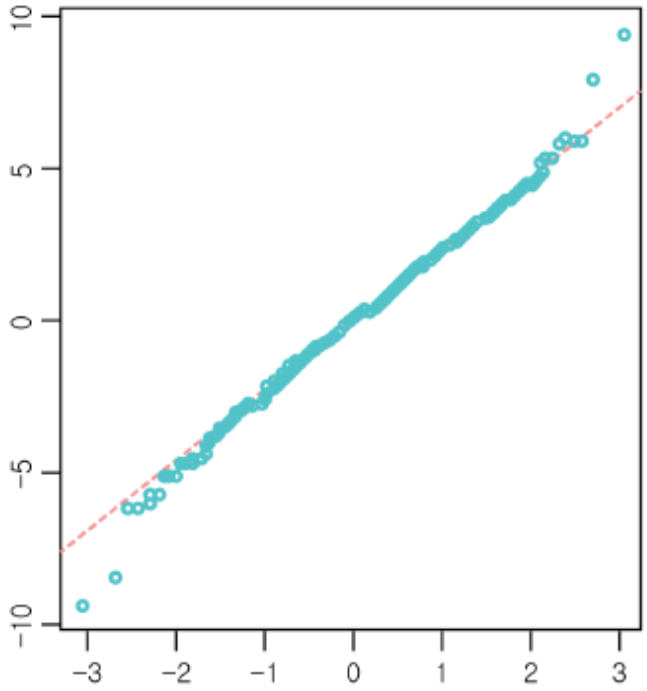
# 회귀분석의 가정 확인

- x축에는 각 분석의 독립변수의 값, y축에는 잔차로 하는 산점도를 나타냅니다.
- 위의 도표에서 좌측의 그림은 0 주변으로 특정한 패턴 없이(랜덤하게) 잔차들이 많이 몰려있으나 우측의 경우는 점점 퍼지는 형태의 산점도를 보이고 있습니다.
  - 잔차들이 서로 등분산일 경우 좌측의 그림처럼 패턴없이 0주변에 랜덤하게 몰려있게 됩니다.
- 잔차와 독립변수와의 산점도만 그렸으나, 잔차와 예측값과의 산점도를 함께 그려보는 것을 추천합니다. (예측값은 R에서 predict(out)으로 알 수 있습니다.)

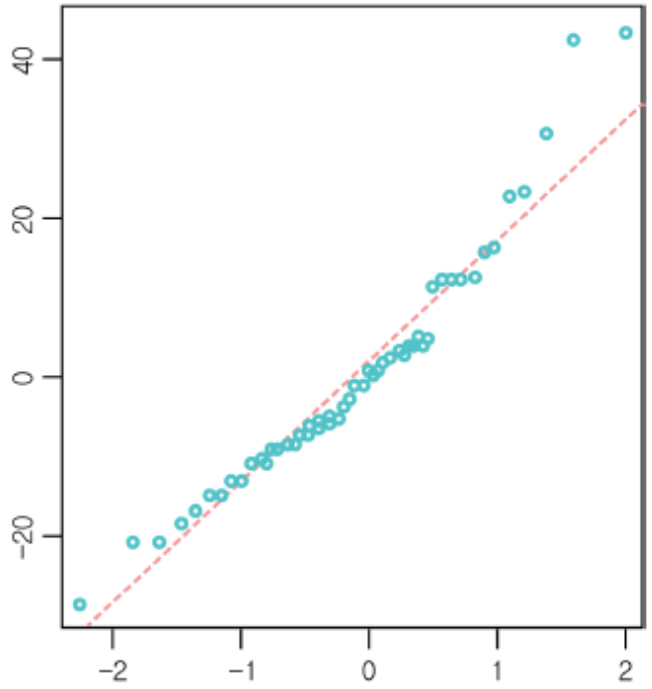
## 잔차의 정규확률 그림

- 정규확률그림(normal Q-Q plot)은 x축으로는 이론적인 정규분포의 값을, y축으로는 자료의 값을 갖는 산점도입니다.
- 만일 자료가 정규분포를 따른다면, 정규분포 적합선 위에 자료가 패턴 없이 많이 분포합니다.

# 회귀분석의 가정 확인




(a)




(b)

- (a)는 붉은 점선으로 나타나는 정규분포의 적합선 위에 잔차가 고르게 분포해 있는 경향이 강합니다.
- (b)는 -1이하에서 적합선 위로 나타나다가 -1부터 1 사이에 적합선 아래로 나타나고, 1이상에서 적합선 위로 나타나는 형태를 보이고 있으, ±1 이후의 값들이 적합선에서 너무 멀리 떨어져 있어 정규분포가 의심스러운 상황입니다.

# Q & A

수고하셨습니다.