통계학의 이해 II

회귀모형의 형태

회귀모형의 형태

- 수치변수들 간의 관계를 간단히 알아보는 방법을 알아본다.
- 수치변수들 간 인과관계를 설명하기 위한 대표적인 통계모형인 회귀모형을 소개한다.

예제】 올림픽 육상 100m 우승기록

- Andrew Tatem 등이 2004년 9월 Nature에 논문 발표
- 1900~2004년까지의 남자와 여자의 육상 100m 우승 기록을 분석
- 2008년 베이징 올림픽 남자의 우승기록은 9.73±0.144(9.586, 9.874), 여자는 10.57±0.232(10.338, 10.802)로 예측
  - 실제 기록: 남자 9.69, 여자 10.75
- 2156년 올림픽에서 여성 우승 기록이 남성기록보다 빠를 것으로 예측
  - 남성 우승기록 8.098초, 여성 우승기록 8.079초
  - 통계적 오차(예측구간)를 고려하면, 빠르면 2064년, 늦어도 2788년에는 역전 될 것이라고 주장

다변량 자료(Multivariate Data)

- 어떤 대상에 대해 여러 가지 변수들을 관측(측정)한 자료들의 집합
  - 예】 신체검사: 연령, 성별, 신장, 체중, 시력, 혈액형, ...

- 자료의 형태

관측값	변수1	변수2	…	변수p
1	x_{11}	x_{21}	$\cdots$	x_{1p}
2	x_{21}	x_{22}	$\cdots$	x_{2p}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
n	x_{n1}	x_{n2}	$\cdots$	x_{np}

  - 관측값 간에는 관련성이 없음: 독립적인 관측값

- 주요관심사: 변수들 간의 관계
  - ☑ 변수들 간 관계가 있는가?
  - ☑ 있다면 어떤 관계가 있는가?
  - ☑ 관계가 어느 정도 되는가?

산점도, 상관분석으로 가능

  - ☑ 관계를 식으로 표시할 수 있는가?
  - ☑ 관계식을 유도할 수 있는가?

산점도, 상관분석?

  - ☑ 유도된 관계식을 통해 다른 값을 예측할 수 있는가?

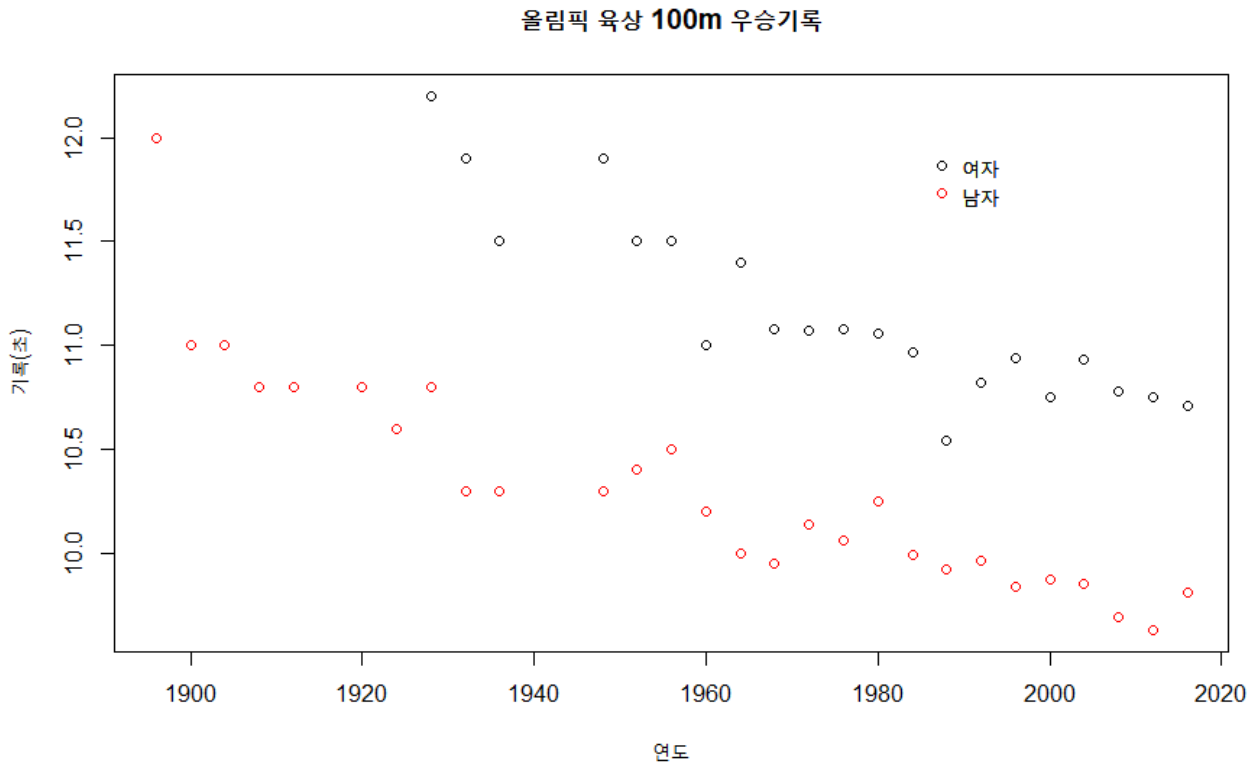
- 분석 목적이 관계유도 및 예측인 대표적인 모형이 회귀모형

산점도 & 상관분석

1896년~2016년까지 올림픽 육상 100m 우승기록

연도	남자	여자	연도	남자	여자	연도	남자	여자	연도	남자	여자
1896	12	-	1928	10.8	12.2	1964	10.0	11.4	1992	9.96	10.82
1900	11	-	1932	10.3	11.9	1968	9.9	11.0	1996	9.84	10.94
1904	11	-	1936	10.3	11.5	1972	10.14	11.07	2000	9.87	10.75
1908	10.8	-	1948	10.3	11.9	1976	10.06	11.08	2004	9.85	10.93
1912	10.8	-	1952	10.4	11.5	1980	10.25	11.06	2008	9.69	10.78
1920	10.8	-	1956	10.5	11.5	1984	9.99	10.97	2012	9.63	10.75
1924	10.6	-	1960	10.2	11.0	1988	9.92	10.54	2016	9.81	10.71

산점도(Scatter plot)



상관분석(Analysis of Correlation)

- 상관계수: 두 변수 간의 직선(선형) 관계의 정도

$$R_{XY} = \frac{1}{n-1}\sum\left(\frac{X_i - \bar{X}}{S_X}\right)\left(\frac{Y_i - \bar{Y}}{S_Y}\right) = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}}$$

- $S_{XY} = \sum(X_i - \bar{X})\,(Y_i - \bar{Y})$, $S_{XX} = \sum(X_i - \bar{X})^2$, $S_{YY} = \sum(Y_i - \bar{Y})^2$
- 두 변수가 모두 정규분포를 따르면, H_0: $\rho = 0$에 대한 검정통계량

$$T = \frac{\sqrt{n-2}R}{\sqrt{1-R^2}} \sim t_{n-2}$$

- 이상점이 있는 경우 Spearman의 순위상관, Kendall의 tau 등 대체상관분석을 할 수 있음

예제】 올림픽 육상 100m 우승기록

남자자료: 연도 x, 기록 y

- $n = 28, \bar{x} = 1958.43, \bar{y} = 10.313$
- $\sum x_i^2 = 107429904,\ \sum y_i^2 = 2985.27,\ \sum x_i y_i = 565043.2$
- $S_{xx} = 37514.86, S_{yy} = 7.330,\ S_{xy} = -472.67$
- $r = \dfrac{-472.67}{\sqrt{(37514.86)(7.330)}} = -0.901$
- $t = \dfrac{\sqrt{28-2}(-0.901)}{\sqrt{1-(-0.901)^2}} = -10.615 < -2.47 = t_{0.01,26}$

회귀모형(Regression Model)

변수들 간의 인과 관계 유도

$$\mathbf{X} \quad \rightarrow \quad f \quad \rightarrow \quad Y$$

Input System Output

- 입력변수 X: 설명(explanatory)변수, 독립(independent)변수
  - 양적변수: 공변량(covariate), 질적변수: 요인(factor)
- 출력변수 Y: 반응(response)변수, 종속 (dependent)변수
- 예】 광고비와 판매량, 공부량과 시험성적,
  (비료량, 평균강수량, 평균기온, 평균일조량)과 수확량
- 동일한 입력변수 X에 대해 출력변수 Y는 다른 값을 가질 수 있음

선형회귀모형(Linear Regression Model)

관계식 가정

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

- β: 회귀계수(regression coefficients)
- 관계식은 회귀계수 β에 대해 선형
  - 선형과 비선형의 구분은?
- ε: 오차(error)
  - 모형으로 설명이 안되는 부분
  - 오차에 특정 패턴이 있으면 모형화 할 수 있는 부분이 남아 있음
  - 통계적 추론을 위한 가정: $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n \sim iid\ N(0, \sigma^2)$

회귀모형의 형태

- 수치변수들 간의 관계를 나타내는 방법을 복습한다.
  - ☑ 산점도
  - ☑ 직선관계를 나타내는 상관계수와 상관분석
- 수치변수들 간 인과관계를 설명하기 위한 대표적인 통계모형인 회귀모형을 소개한다.
  - ☑ 선형회귀모형

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

통계학의 이해 II

단순선형회귀에서의 모수추정

단순선형회귀모형에서의 모수추정

- 설명변수가 하나인 회귀모형에서 관측값과 회귀선과의 거리를 어떻게 표시하는지 알아본다.

- 최소제곱법을 이용한 회귀모수를 추정하는 방법을 알아본다.

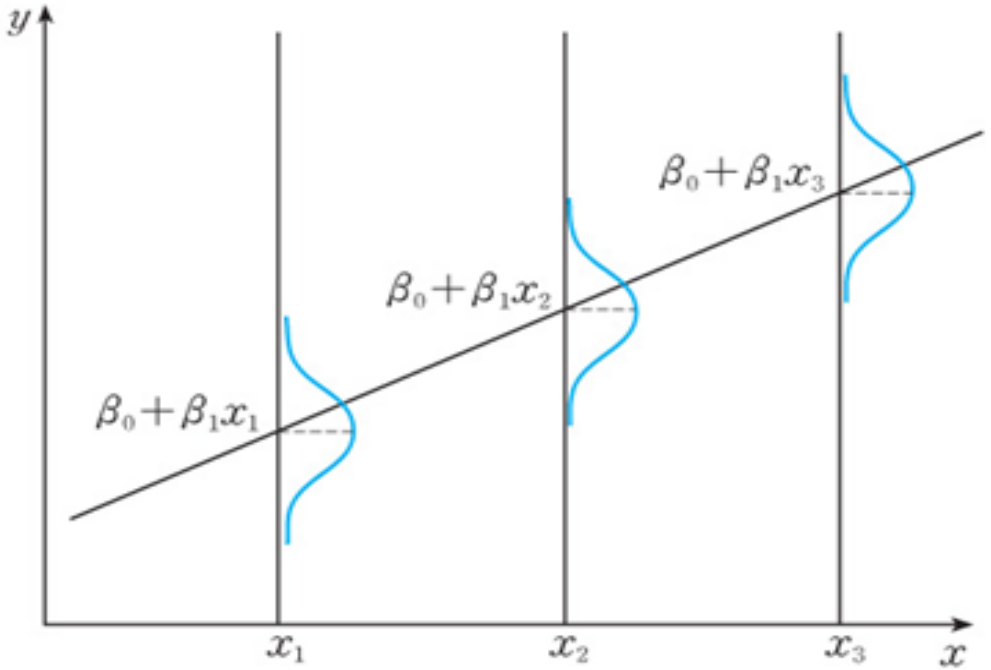
단순선형회귀모형(Simple Linear Regression Model)

설명변수가 하나인 선형회귀모형

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad \varepsilon_i \sim iid\ N(0, \sigma^2)$$

- 설명변수는 조절 가능한 상수로 가정
  - 광고비에 따른 판매량: 광고비는 회사에서 결정 가능
  - 일조량에 따른 수확량: 관측된 값으로 주어진 값으로 처리
- 설명변수가 여러 개인 경우: (다)중회귀모형(multiple regression model)
- 미지의 모수 절편 β_0, 기울기 β_1를 추정
  - β_1은 x를 한 단위 증가시킬 때 Y의 평균증가량
  - $\beta_1 = 0$이면 x가 Y에 영향을 주지 않는다는 것을 의미

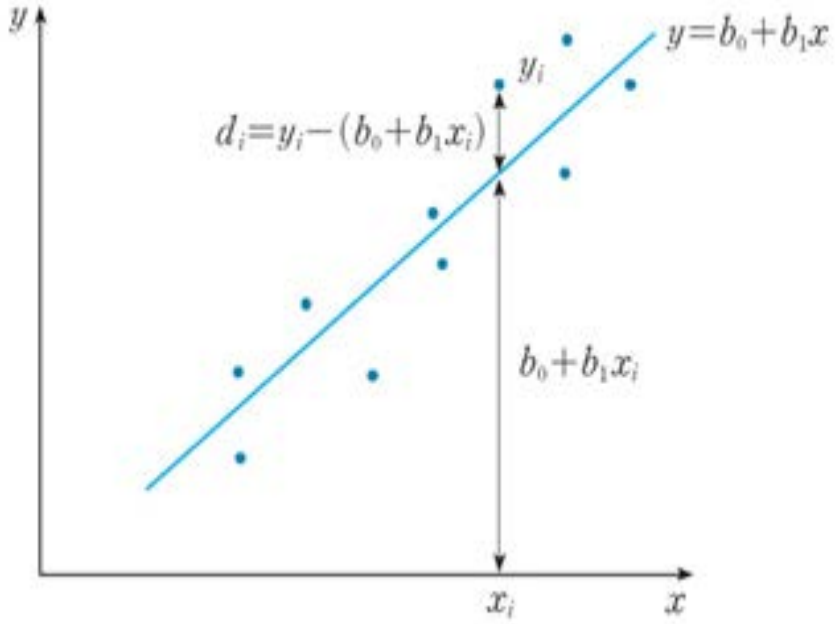
$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,\ \ \varepsilon_i \sim iid\ N(0, \sigma^2) \ \Rightarrow Y_i \sim\ N(\beta_0 + \beta_1 x_i, \sigma^2)$$



모수추정

(β_0, β_1)의 추정

- (b_0, b_1): (β_0, β_1)의 추정값
- $d_i = y_i - (b_0 + b_1 x_i)$



- 추정된 직선이 좋은 직선인가 아닌가에 대한 기준 설정이 필요
  - 관측값 y_i와 추정된 직선에서의 $(b_0 + b_1 x_i)$의 거리에 대한 정의 필요
  - 최소절대편차법(Least Absolute Deviation method)

$$D_1(b_0, b_1) = \sum |d_i| = \sum |y_i - b_0 - b_1 x_i|$$

  - 최소제곱법(Least Squares method)

$$D_2(b_0, b_1) = \sum d_i^{\,2} = \sum (y_i - b_0 - b_1 x_i)^2$$

    - 장점: b_0, b_1에 대해 미분가능하여 최소로 만드는 b_0, b_1를 쉽게 찾을 수 있음

최소제곱법

$$D(b_0, b_1) = \sum d_i^{\,2} = \sum (y_i - b_0 - b_1 x_i)^2$$

- $\frac{\partial D}{\partial b_0} = -2\sum (y_i - b_0 - b_1 x_i) = 0 \;\rightarrow \sum y_i - n b_0 - b_1 \sum x_i = 0$
- $\frac{\partial D}{\partial b_1} = -2\sum x_i (y_i - b_0 - b_1 x_i) = 0 \;\rightarrow \sum x_i y_i - b_0 \sum x_i - b_1 \sum x_i^{\,2} = 0$
- $b_1 = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^{\,2} - n\bar{x}^2} = \frac{S_{xy}}{S_{xx}} \quad \Rightarrow \hat{\beta}_1$
- $b_0 = \bar{y} - b_1 \bar{x} \quad \Rightarrow \hat{\beta}_0$

- 최소제곱추정값: $\hat{\beta}_1 = S_{xy}/S_{xx}$, $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$ $\Longrightarrow$ 실제 분석에 사용

  - ☑ 최소제곱추정량: $\hat{\beta}_1 = S_{xY}/S_{xx}$, $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{x}$

    $\Longrightarrow$ 통계적 추론(분포, 기댓값)할 때 사용

적합값(Fitted value)

- 적합회귀직선(추정회귀직선): $\hat{\beta}_0 + \hat{\beta}_1 x$
- 적합값(예측값, predicted value):

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \qquad \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

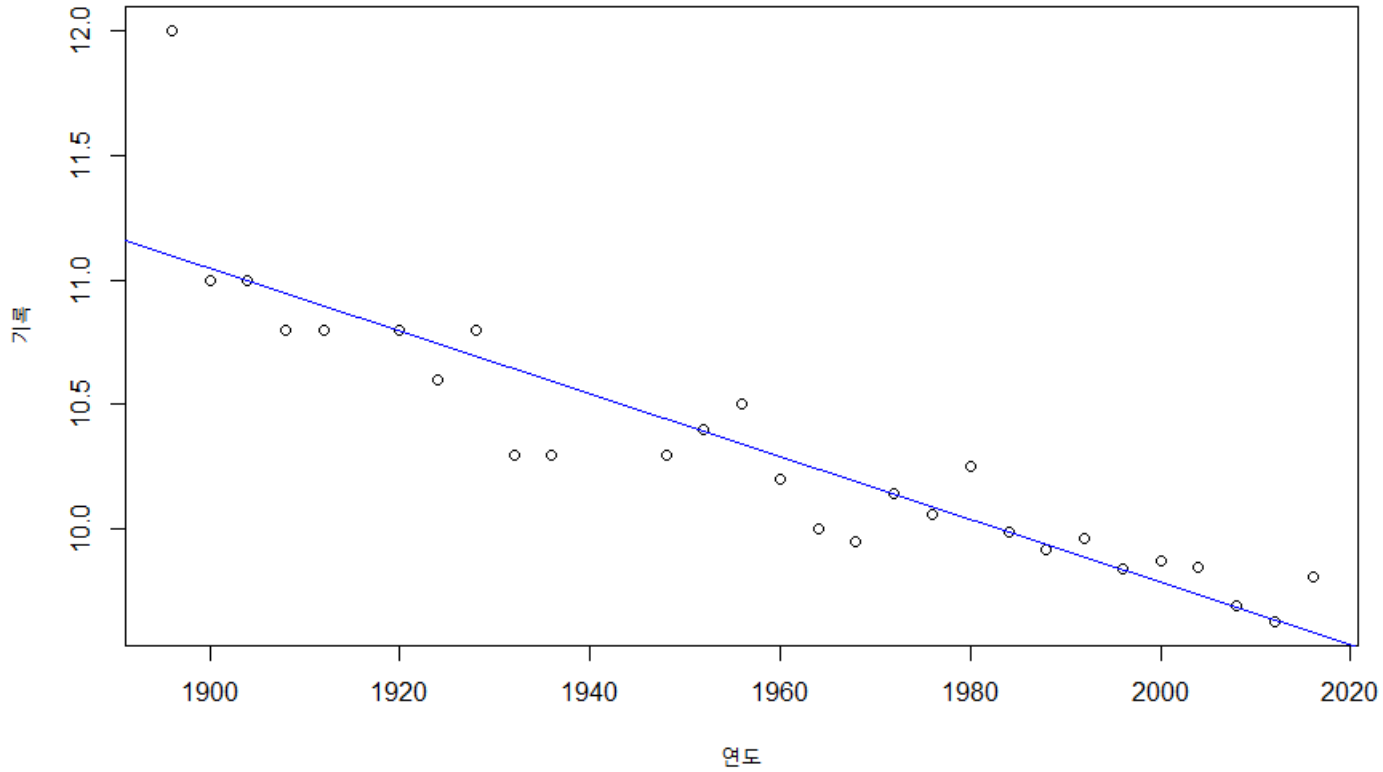
잔차(residual): 관측값과 예측값의 차이

- $e_i = y_i - \hat{y}_i, \qquad e_i = Y_i - \hat{Y}_i$
- 최소제곱추정량을 유도과정
  - $-2\sum\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0 \quad \Rightarrow \sum e_i = 0$
  - $-2\sum x_i\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0 \quad \Rightarrow \sum x_i e_i = 0$

예제】 올림픽 육상 100m 우승기록

- 남자자료: 연도 x, 기록 y
  - $n = 28, \bar{x} = 1958.43, \bar{y} = 10.313$
  - $S_{xx} = 37514.86, S_{yy} = 7.330,\ S_{xy} = -472.67$
  - $\hat{\beta}_1 = \dfrac{-472.67}{37514.86} = -0.0126$
  - $\hat{\beta}_0 = 10.313 - (-0.0126)1958.43 = 34.99$
  - $y_t = 34.99 - 0.0126x_t$

    $\Rightarrow$ 매 경기 때마다 평균 $0.0126 \times 4 = 0.0504$초 정도 단축됨


12.0
11.5
11.0
10.5
10.0
기록
1900
1920
1940
1960
1980
2000
2020
연도

단순선형회귀모형에서의 모수추정

- 설명변수가 하나인 회귀모형에서 관측값과 회귀선과의 거리를 어떻게 표시하는지 알아본다.
  - ☑ $D_1(b_0, b_1) = \sum |d_i| = \sum |y_i - b_0 - b_1 x_i|$
  - ☑ $D_2(b_0, b_1) = \sum {d_i}^2 = \sum (y_i - b_0 - b_1 x_i)^2$
- 최소제곱법을 이용한 회귀모수를 추정하는 방법을 알아본다.
  - ☑ 최소제곱추정량: $\hat{\beta}_1 = S_{xY}/S_{xx}$, $\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{x}$

통계학의 이해 II

회귀추론을 위한 기본이론

회귀추론을 위한 기본이론

- 회귀모형의 모수 또는 예측값을 추론을 위한 기본 통계이론을 정리한다.

회귀모형식 가정

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad \varepsilon_i \sim iid\ N(0, \sigma^2)$$

- 최소제곱법에 의한 모수 추정에서는 특별히 오차항의 가정을 사용하지 않음
- 모수 추정량 또는 예측값의 성질을 유도하기 위해 오차항의 가정 필요
- $\varepsilon_i \sim N(0, \sigma^2) \implies Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$

표집분포

- $Y_i \sim N(\mu_i, \sigma^2)$이고 서로 독립이면 $\sum a_i Y_i \sim N(\sum a_i \mu_i, \sigma^2 \sum {a_i}^2)$
- 표준화: $\frac{\sum a_i Y_i - \sum a_i \mu_i}{\sqrt{\sigma^2 \sum {a_i}^2}} \sim N(0,1)$
- σ^2의 추정량은?
- $\frac{\sum a_i Y_i - \sum a_i \mu_i}{\sqrt{\hat{\sigma}^2 \sum {a_i}^2}} \sim t_\nu$,
  - 자유도 ν는?

복습】 σ^2의 추정

☑ $Y_1, Y_2, \cdots, Y_n \sim iid\ N(\mu, \sigma^2) \Rightarrow Y_i - \mu = \varepsilon_i,\ \varepsilon_i \sim iid\ N(0, \sigma^2)$

$$\sigma^2 = Var(\varepsilon_i) = E(\varepsilon_i^2) \Rightarrow S^2 = \frac{1}{n-1}\sum(Y_i - \bar{Y})^2$$

☑ $X_1, \cdots, X_m \sim iid\ N(\mu_1, \sigma^2),\ Y_1, \cdots, Y_n \sim iid\ N(\mu_2, \sigma^2)$

$\Rightarrow X_i - \mu_1 = \varepsilon_{iX},\ Y_i - \mu_2 = \varepsilon_{iY},\ \varepsilon_{iX}, \varepsilon_{iY} \sim iid\ N(0, \sigma^2)$

$$S_p^2 = \frac{\sum(X_i - \bar{X})^2 + \sum(Y_i - \bar{Y})^2}{m + n - 2}$$

☑ $Y_{ij} \sim N(\mu_i, \sigma^2),\ i = 1, \ldots, k,\ j = 1, \ldots, n_i \Rightarrow Y_{ij} - \mu_i = \varepsilon_i$

$$MSE = \frac{\sum\sum(Y_{ij} - \bar{Y}_i)^2}{N - k}$$

회귀모형

$$Y_i - (\beta_0 + \beta_1 x_i) = \varepsilon_i, \qquad \varepsilon_i \sim iid\ N(0, \sigma^2)$$

☑ 분자: $\sum \left(Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\right)^2 = \sum e_i^2 = SSE$

☑ 자유도(degree of freedom)?

○ $\sum e_i = 0,\ \sum x_i e_i = 0$

$\Rightarrow$ n 개의 잔차 중 $n-2$만 자유롭게 가질 수 있음

○ 자유도 = 자료의 수 – 해당 통계량에 포함된 추정량의 수

$$MSE = \frac{1}{n-2} \sum \left(Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\right)^2$$

☑ $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$이고 서로 독립이면

$$\frac{\sum a_i Y_i - \sum a_i(\beta_0 + \beta_1 x_i)}{MSE\sqrt{\sum a_i{}^2}} \sim t_{n-2}$$

MSE 계산

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = \bar{Y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i = \bar{Y} + \hat{\beta}_1 (x_i - \bar{x})$$

$$\Rightarrow \hat{Y}_i - \bar{Y} = \hat{\beta}_1 (x_i - \bar{x})$$

$$S_{YY} = \sum (Y_i - \bar{Y})^2 = \sum \left(Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y}\right)^2$$

$$= \sum \left(Y_i - \hat{Y}_i\right)^2 + \sum \left(\hat{Y}_i - \bar{Y}\right)^2$$

$$= SSE + \hat{\beta}_1^2 \sum (x_i - \bar{x})^2 = SSE + S_{xY}^2 / S_{xx}$$

$$\Rightarrow MSE = \frac{1}{n-2} \left(S_{YY} - S_{xY}^2 / S_{xx}\right)$$

회귀추론을 위한 기본이론

- 회귀모형의 모수 또는 예측값을 추론을 위한 기본 통계이론을 정리한다.

☑ σ^2의 추정량:

$$MSE = \frac{1}{n-2}\sum\left(Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)\right)^2$$

☑ $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$이고 서로 독립이면

$$\frac{\sum a_i Y_i - \sum a_i(\beta_0 + \beta_1 x_i)}{MSE\sqrt{\sum a_i{}^2}} \sim t_{n-2}$$

통계학의 이해 II

회귀계수(기울기)에 대한 통계적 추론

기울기에 대한 통계적 추론

- 회귀계수 중 기울기에 해당하는 β_1의 중심축량, 구간추정, 가설검정에 대해 알아본다.

기울기 β_1에 대한 추론

$\hat{\beta}_1 = S_{xY}/S_{xx}$의 통계적 성질

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(Y_i - \bar{Y})}{S_{xx}} = \frac{\sum(x_i - \bar{x})Y_i}{S_{xx}} \quad \Rightarrow Y\text{들의 선형결합}$$

- $E(\hat{\beta}_1) = \frac{1}{S_{xx}}\sum(x_i - \bar{x})E(Y_i) = \frac{1}{S_{xx}}\sum(x_i - \bar{x})(\beta_0 + \beta_1 x_i) = \beta_1$
- $Var(\hat{\beta}_1) = \frac{1}{S_{xx}^2}\sum(x_i - \bar{x})^2 Var(Y_i) = \frac{\sigma^2}{S_{xx}}$
- $\hat{\beta}_1 = S_{xY}/S_{xx} \sim N(\beta_1, \sigma^2/S_{xx})$

중심축량:

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{MSE/S_{xx}}} \sim t_{n-2}$$

MSE 계산

- $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = \bar{Y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i = \bar{Y} + \hat{\beta}_1 (x_i - \bar{x})$

  $\Rightarrow \hat{Y}_i - \bar{Y} = \hat{\beta}_1 (x_i - \bar{x})$

- $S_{YY} = \sum (Y_i - \bar{Y})^2 = \sum \left(Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y}\right)^2$

  $= \sum \left(Y_i - \hat{Y}_i\right)^2 + \sum \left(\hat{Y}_i - \bar{Y}\right)^2$

  $= SSE + \hat{\beta}_1^2 \sum (x_i - \bar{x})^2 = SSE + S_{xY}^2 / S_{xx}$

  $\Rightarrow MSE = \frac{1}{n-2} \left(S_{YY} - S_{xY}^2 / S_{xx}\right)$

구간추정

$100(1-\alpha)\%$ 신뢰구간

$$P\left(-t_{\frac{\alpha}{2},n-2} \le \frac{\hat{\beta}_1 - \beta_1}{\sqrt{MSE/S_{xx}}} \le t_{\frac{\alpha}{2},n-2}\right) = 1-\alpha$$

$$= P\left(\hat{\beta}_1 - t_{\frac{\alpha}{2},n-2}\sqrt{\frac{MSE}{S_{xx}}} \le \beta_1 \le \hat{\beta}_1 + t_{\frac{\alpha}{2},n-2}\sqrt{\frac{MSE}{S_{xx}}}\right)$$

가설검정

- 가설: $H_0: \beta_1 = \beta_1^*$ vs $H_0: \beta_1 \neq \beta_1^*$ (>, < 가능)
- 검정통계량

$$T_0 = \frac{\hat{\beta}_1 - \beta_1^*}{\sqrt{MSE/S_{xx}}} \sim t_{\frac{\alpha}{2}, n-2}$$

- 기각역: $|T_0| \geq t_{\frac{\alpha}{2}, n-2}$
- 설명변수가 반응변수에 영향을 주는지 여부를 $\beta_1 = 0$인지 아닌지로 확인

예제】 올림픽 육상 100m 우승기록

- 남자자료: 연도 x, 기록 y
  - $n = 28, \bar{x} = 1958.43, \bar{y} = 10.313$
  - $S_{xx} = 37514.86, S_{yy} = 7.330,\ S_{xy} = -472.67$
  - $\hat{\beta}_1 = -0.0126$
  - $MSE = \frac{1}{28-2}\left(7.330 - \frac{(-472.67)^2}{37514.9}\right) = 0.0529$
  - 95% 신뢰구간: $-0.0126 \pm 2.056\sqrt{\frac{0.0529}{37514.9}} = (-0.01504, -0.0102)$
  - 검정통계값: $t_o = \frac{-0.0126}{\sqrt{\frac{0.0529}{37514.9}}} = -10.62$

기울기에 대한 통계적 추론

- 회귀계수 중 기울기에 해당하는 β_1의 중심축량, 구간추정, 가설검정에 대해 알아본다.

  - 중심축량: $\dfrac{\hat{\beta}_1 - \beta_1}{\sqrt{MSE/S_{xx}}} \sim t_{n-2}$

  - $100(1-\alpha)\%$ 신뢰구간: $\left(\hat{\beta}_1 - t_{\frac{\alpha}{2},n-2}\sqrt{\dfrac{MSE}{S_{xx}}}, \hat{\beta}_1 + t_{\frac{\alpha}{2},n-2}\sqrt{\dfrac{MSE}{S_{xx}}} \right)$

  - 검정통계량: $T_0 = \dfrac{\hat{\beta}_1 - \beta_1^*}{\sqrt{MSE/S_{xx}}} \sim t_{\frac{\alpha}{2},n-2}$

통계학의 이해 II

강의정리 및 실습

회귀모형의 형태

- 수치변수들 간의 관계를 나타내는 방법을 복습한다.
  - ☑ 산점도
  - ☑ 직선관계를 나타내는 상관계수와 상관분석
- 수치변수들 간 인과관계를 설명하기 위한 대표적인 통계모형인 회귀모형을 소개한다.
  - ☑ 선형회귀모형

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

단순선형회귀모형에서의 모수추정

- 설명변수가 하나인 회귀모형에서 관측값과 회귀선과의 거리를 어떻게 표시하는지 알아본다.
  - ☑ $D_1(b_0, b_1) = \sum |d_i| = \sum |y_i - b_0 - b_1 x_i|$
  - ☑ $D_2(b_0, b_1) = \sum {d_i}^2 = \sum (y_i - b_0 - b_1 x_i)^2$
- 최소제곱법을 이용한 회귀모수를 추정하는 방법을 알아본다.
  - ☑ 최소제곱추정량: $\hat{\beta}_1 = S_{xY}/S_{xx}$, $\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{x}$

회귀추론을 위한 기본이론

- 회귀모형의 모수 또는 예측값을 추론을 위한 기본 통계이론을 정리한다.

☑ σ^2의 추정량:

$$MSE = \frac{1}{n-2}\sum\left(Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)\right)^2$$

☑ $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$이고 서로 독립이면

$$\frac{\sum a_i Y_i - \sum a_i(\beta_0 + \beta_1 x_i)}{MSE\sqrt{\sum a_i{}^2}} \sim t_{n-2}$$

기울기에 대한 통계적 추론

- 회귀계수 중 기울기에 해당하는 β_1의 중심축량, 구간추정, 가설검정에 대해 알아본다.

  - 중심축량: $\dfrac{\hat{\beta}_1 - \beta_1}{\sqrt{MSE/S_{xx}}} \sim t_{n-2}$

  - $100(1-\alpha)\%$ 신뢰구간: $\left(\hat{\beta}_1 - t_{\frac{\alpha}{2},n-2}\sqrt{\dfrac{MSE}{S_{xx}}}, \hat{\beta}_1 + t_{\frac{\alpha}{2},n-2}\sqrt{\dfrac{MSE}{S_{xx}}} \right)$

  - 검정통계량: $T_0 = \dfrac{\hat{\beta}_1 - \beta_1^*}{\sqrt{MSE/S_{xx}}} \sim t_{\frac{\alpha}{2},n-2}$

R 실습

- 상관분석: cor, cor.test
- 선형회귀모형 추론: lm
- 적합값, 예측값, 잔차: predict, residuals
- 그림: plot

과제

✧ 올림픽 100m 우승기록(1900년부터 2004년까지 자료)

- 남녀별로 나누어 단순선형회귀분석을 하여 한 그림에 산점도와 추정회귀선을 표시하여라.
- 추정된 회귀결과를 이용하여 남녀별로 2008년부터 2016년까지의 실제값과 예측값을 비교하여라.
- 각 회귀분석결과를 이용하여 MSE을 유도하여라.
- 남녀별 회귀계수(기울기)에 대한 95% 신뢰구간을 구하여라.