통계학의 이해 II

회귀계수(절편)에 대한 통계적 추론

절편에 대한 통계적 추론

- 회귀계수 중 절편에 해당하는 β_0의 중심축량과 구간추정에 대해 알아본다.

절편 $\boldsymbol{\beta_0}$에 대한 추론

- $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$의 역할
  - β_0: $x = 0$일 때 $E(Y)$의 값
  - 최소제곱법 추정의 추정과정:

$$\frac{\partial D}{\partial b_0} = -2 \sum (y_i - b_0 - b_1 x_i) = 0$$

    - β_0가 없는 모형에서의 잔차합은 0이 되지 않을 수 있음
    - β_0의 포함여부($\beta_0 = 0$)에 대한 추론은 일반적으로 하지 않음
  - 설명변수가 0인 상황이 주요한 경우에 해석

◇ $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{x}$의 통계적 성질

$$\hat{\beta}_0 = \bar{Y} - \frac{S_{xY}}{S_{xx}}\bar{x} = \sum\left\{\frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{S_{xx}}\right\}Y_i \quad \Rightarrow Y\text{들의 선형결합}$$

✓ $E(\hat{\beta}_0) = E(\bar{Y} - \hat{\beta}_1\bar{x}) = \frac{1}{n}\sum E(Y_i) - E(\hat{\beta}_1)\bar{x} = \beta_0 + \beta_1\bar{x} - \beta_1\bar{x} = \beta_0$

✓ $Var(\hat{\beta}_0) = \sum\left\{\frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{S_{xx}}\right\}^2 Var(Y_i)$

$$= \sigma^2\sum\left\{\frac{1}{n^2} - 2\frac{(x_i - \bar{x})\bar{x}}{nS_{xx}} + \frac{(x_i - \bar{x})^2\bar{x}^2}{S_{xx}^2}\right\}$$

$$= \sigma^2\left\{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}^2}\right\}$$

- $\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left\{\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right\}\right)$

- 중심축량

$$\frac{\hat{\beta}_0-\beta_0}{\sqrt{MSE}\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}}} \sim t_{n-2}$$

- $100(1-\alpha)\%$ 신뢰구간

$$\hat{\beta}_0 \pm t_{\frac{\alpha}{2},n-2}\sqrt{MSE}\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}}$$

◇ 예제】 올림픽 육상 100m 우승기록

- 남자자료: 연도 x, 기록 y
  - $n = 28, \bar{x} = 1958.43, S_{xx} = 37514.86,\ MSE = 0.0529$
  - $\hat{\beta}_0 = 34.988$
  - SE 추정값:

$$\sqrt{0.0529}\sqrt{\frac{1}{28} + \frac{1958.43^2}{37514.86}} = 2.325$$

  - 95% 신뢰구간:$34.988 \pm 2.056 \times 2.325 = (30.209, 39.768)$

절편에 대한 통계적 추론

- 회귀계수 중 절편에 해당하는 β_0의 중심축량과 구간추정에 대해 알아본다.

✓ 중심축량:
$$\frac{\hat{\beta}_0 - \beta_0}{\sqrt{MSE}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}} \sim t_{n-2}$$

✓ $100(1-\alpha)\%$ 신뢰구간
$$\hat{\beta}_0 \pm t_{\frac{\alpha}{2},n-2}\sqrt{MSE}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}$$

통계학의 이해 II

예측값 평균에 대한 통계적 추론

예측값 평균에 대한 통계적 추론

- 예측값의 평균, $E(Y) = \beta_0 + \beta_1 x$,를 추론하기 위한 중심축량과 예측구간을 알아본다.

반응변수의 기댓값 $E(Y_k)$에 대한 추론

- $E(Y_k) = \beta_0 + \beta_1 x_k \;\Rightarrow$ 점추정량: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_k$
- 점추정량의 성질

$$\hat{Y}_k = \hat{\beta}_0 + \hat{\beta}_1 x_k = \bar{Y} + (x_k - \bar{x})\hat{\beta}_1 = \bar{Y} + (x_k - \bar{x})\frac{S_{xY}}{S_{xx}}$$

$$= \sum \left\{ \frac{1}{n} + (x_k - \bar{x}) \frac{(x_i - \bar{x})}{S_{xx}} \right\} Y_i \;\Rightarrow Y\text{들의 선형결합}$$

- $E(\hat{Y}_k) = E(\hat{\beta}_0 + \hat{\beta}_1 x_k) = \beta_0 + \beta_1 x_k$
- $Var(\hat{Y}_k) = \sigma^2 \sum \left\{ \frac{1}{n} - (x_k - \bar{x}) \frac{(x_i - \bar{x})}{S_{xx}} \right\}^2 = \sigma^2 \left\{ \frac{1}{n} + \frac{(x_k - \bar{x})^2}{S_{xx}} \right\}$
  - x_k가 $\bar{x}$ 에서 멀어질수록 분산이 커짐

- $\hat{Y}_k \sim N\left(\beta_0 + \beta_1 x_k, \sigma^2 \left\{\frac{1}{n} + \frac{(x_k - \bar{x})^2}{S_{xx}}\right\}\right)$

- 중심축량

$$\frac{\hat{Y}_k - E(Y_k)}{\sqrt{MSE}\sqrt{\frac{1}{n} + \frac{(x_k - \bar{x})^2}{S_{xx}}}} \sim t_{n-2}$$

- $100(1-\alpha)\%$ 예측구간

$$\hat{Y}_k \pm t_{\frac{\alpha}{2}, n-2}\sqrt{MSE}\sqrt{\frac{1}{n} + \frac{(x_k - \bar{x})^2}{S_{xx}}}$$

◇ 예제】 올림픽 육상 100m 우승기록

- 남자자료: 연도 x, 기록 y
  - $\hat{y} = 34.988 - 0.0126x$
  - 2024년 우승기록 평균 예측값: $34.988 - 0.0126 \times 2024 = 9.487$
- 2024년 우승기록 평균의 예측구간
  - $n = 28, \bar{x} = 1958.43,\ S_{xx} = 37514.86, MSE = 0.0529$

$$\sqrt{0.0529}\sqrt{\frac{1}{28} + \frac{(2024 - 1958.43)^2}{37514.86}} = 0.0891$$

  - 95% 예측구간 $9.487 \pm 2.056 \times 0.0891 = (9.303, 9.670)$

예측값 평균에 대한 통계적 추론

- 예측값의 평균에 대한 중심축량과 예측구간을 알아본다.

  - 중심축량

$$\frac{\hat{Y}_k - E(Y_k)}{\sqrt{MSE}\sqrt{\frac{1}{n} + \frac{(x_k - \bar{x})^2}{S_{xx}}}} \sim t_{n-2}$$

  - $100(1-\alpha)\%$ 신뢰구간

$$\hat{Y}_k \pm t_{\frac{\alpha}{2},n-2}\sqrt{MSE}\sqrt{\frac{1}{n} + \frac{(x_k - \bar{x})^2}{S_{xx}}}$$

통계학의 이해 II

새로운 관측값에 대한 예측

새로운 관측값 대한 예측

- 새로운 설명변수에 대한 예측값에 대한 추정과 예측구간을 알아본다.

◇ 새로운 x_*에 대한 예측값 Y_*의 추론

- $Y_* = \beta_0 + \beta_1 x_* + \varepsilon_*$ $\Leftarrow$ 예측값: $\hat{Y}_* = \hat{\beta}_0 + \hat{\beta}_1 x_*$

◇ 예측오차 $\hat{Y}_* - Y_*$에 대한 추론

- 예측값 $\hat{Y}_* = \hat{\beta}_0 + \hat{\beta}_1 x_*$와 Y_* 모두 확률변수
- $\hat{Y}_*$는 기존 Y들의 선형결합이고 Y_*는 새로운 변수 $\Rightarrow$ $\hat{Y}_*$와 Y_*는 독립
- $E(\hat{Y}_*) = \beta_0 + \beta_1 x_*$, $E(Y_*) = \beta_0 + \beta_1 x_*$ $\Rightarrow$ $E(\hat{Y}_* - Y_*) = 0$
- $Var(\hat{Y}_*) = \sigma^2 \left\{ \frac{1}{n} + \frac{(x_* - \bar{x})^2}{S_{xx}} \right\}$
- $Var(\hat{Y}_* - Y_*) = Var(\hat{Y}_*) + Var(Y_*) = \sigma^2 \left\{ 1 + \frac{1}{n} + \frac{(x_* - \bar{x})^2}{S_{xx}} \right\}$

☑ $\hat{Y}_* - Y_* \sim N\left(0, \sigma^2\left\{1 + \frac{1}{n} + \frac{(x_k - \bar{x})^2}{S_{xx}}\right\}\right)$

$$\Rightarrow \frac{\hat{Y}_* - Y_*}{\sqrt{MSE}\sqrt{1 + \frac{1}{n} + \frac{(x_k - \bar{x})^2}{S_{xx}}}} \sim t_{n-2}$$

☑ Y_*에 대한 $100(1-\alpha)\%$ 예측구간

$$\hat{Y}_* \pm t_{\frac{\alpha}{2},n-2}\sqrt{MSE}\sqrt{1 + \frac{1}{n} + \frac{(x_k - \bar{x})^2}{S_{xx}}}$$

◇ 예제】 올림픽 육상 100m 우승기록

- 남자자료: 연도 x, 기록 y
  - $\hat{y} = 34.988 - 0.0126x$
  - 2024년 우승기록 예측값: $34.988 - 0.0126 \times 2024 = 9.487$
- 2024년 우승기록에 대한 예측구간
  - $n = 28, \bar{x} = 1958.43,\ S_{xx} = 37514.86, MSE = 0.0529$

$$\sqrt{0.0529}\sqrt{1 + \frac{1}{28} + \frac{(2024 - 1958.43)^2}{37514.86}} = 0.2466$$

  - 95% 예측구간 $9.487 \pm 2.056 \times 0.2466 = (8.980, 9.994)$

새로운 관측값 대한 예측

- 새로운 설명변수에 대한 예측값에 대한 추정과 예측구간을 알아본다.

  - $$\frac{\hat{Y}_* - Y_*}{\sqrt{MSE}\sqrt{1 + \frac{1}{n} + \frac{(x_k - \overline{x})^2}{S_{xx}}}} \sim t_{n-2}$$

  - Y_*에 대한 $100(1-\alpha)\%$ 예측구간

$$\hat{Y}_* \pm t_{\frac{\alpha}{2},n-2}\sqrt{MSE}\sqrt{1 + \frac{1}{n} + \frac{(x_k - \overline{x})^2}{S_{xx}}}$$

통계학의 이해 II

잔차검진

잔차검진(Residual Diagnostics)

- 분석에 사용된 회귀모형의 적절성과 통계적 추론의 가정을 만족하는지를 확인하는 방법에 대해 알아본다.

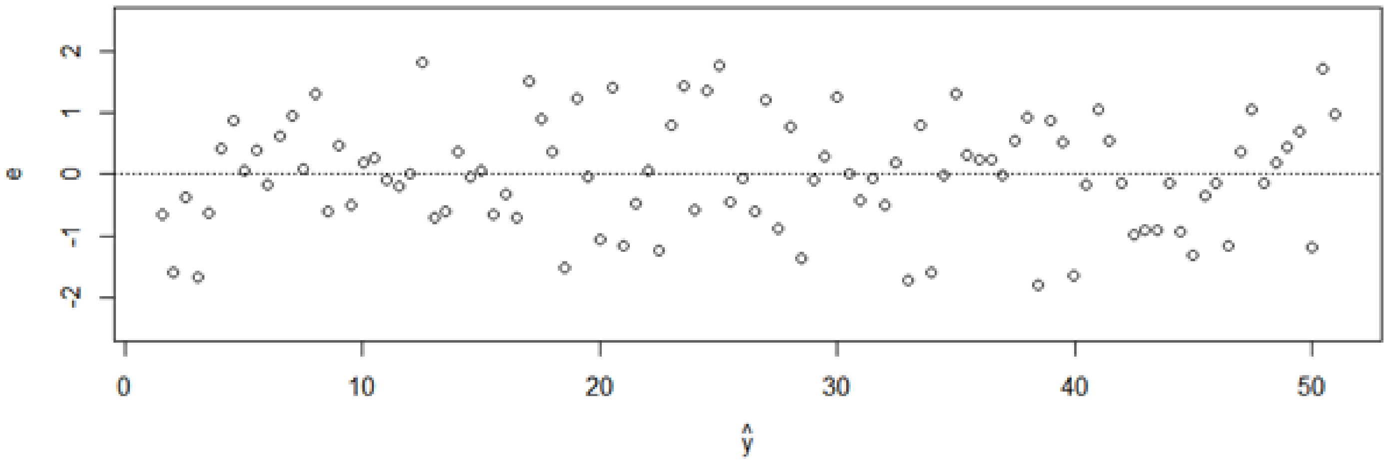
오차항의 가정

- $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n \sim iid\ N(0, \sigma^2)$
  - 정규성
  - 등분산성
  - 독립성

- 잔차(residual): 관측값과 예측값의 차이
  - $e_i = y_i - \hat{y}_i,\qquad e_i = Y_i - \hat{Y}_i$
  - 잔차가 특정한 패턴을 가진다면 모형(설명되는 부분)에 추가해야 할 요소가 남아 있음을 의미
  - 잔차가 오차항의 가정을 심각하게 위반하면 통계적 추론에 문제 발생

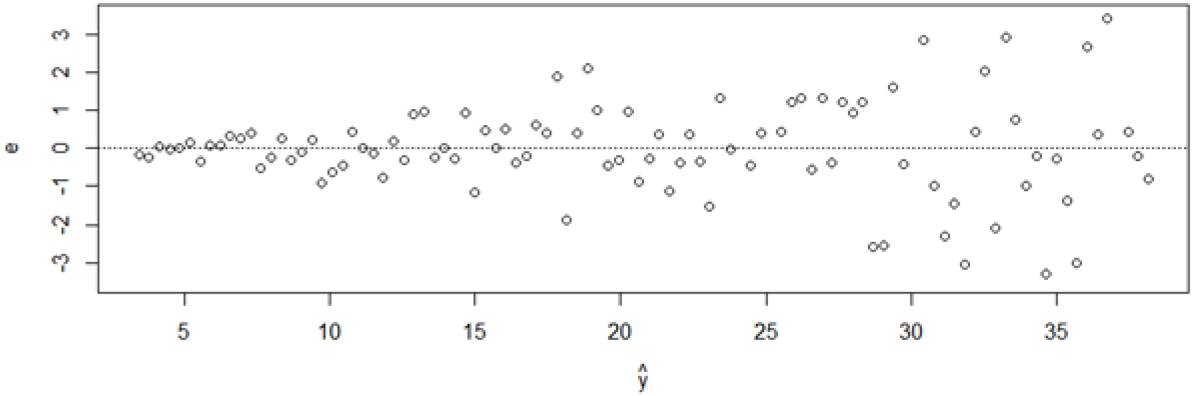
잔차그림(Residual plot)

- $(\hat{y}, e)$의 산점도
- 정상적인 잔차그림: 0을 중심으로 $\hat{y}$ 값에 관계없이 일정 범위 내에서 특정한 패턴을 가지지 않게 분포됨



◇ 대표적인 비정상적 잔차그림

✓ $\hat{y}$가 커지면서 e의 폭이 커짐

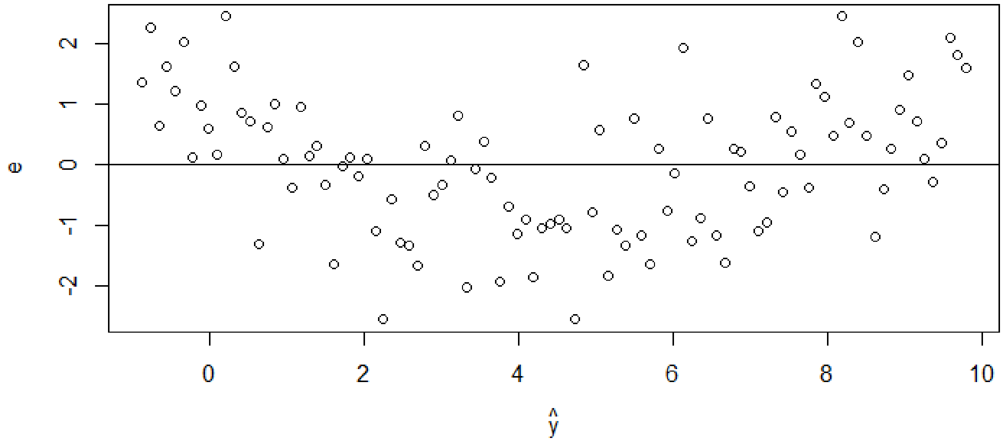


⇒ 등분산성을 만족하지 않음

✓ 대안: 반응변수의 변환

○ 예제】 $Y_i^* = \log(Y_i)$, $\quad Y_i^* = \beta_0 + \beta_1 x_i + \varepsilon_i$, $\quad \varepsilon_i \sim iid\ N(0, \sigma^2)$

☑ $\hat{y}$가 커지면서 e가 하강(상승)하다가 상승(하강)함



⇒ 설명변수의 제곱항이 생략되어 있을 가능성이 큼

☑ 대안: 제곱항 추가, 변수변환

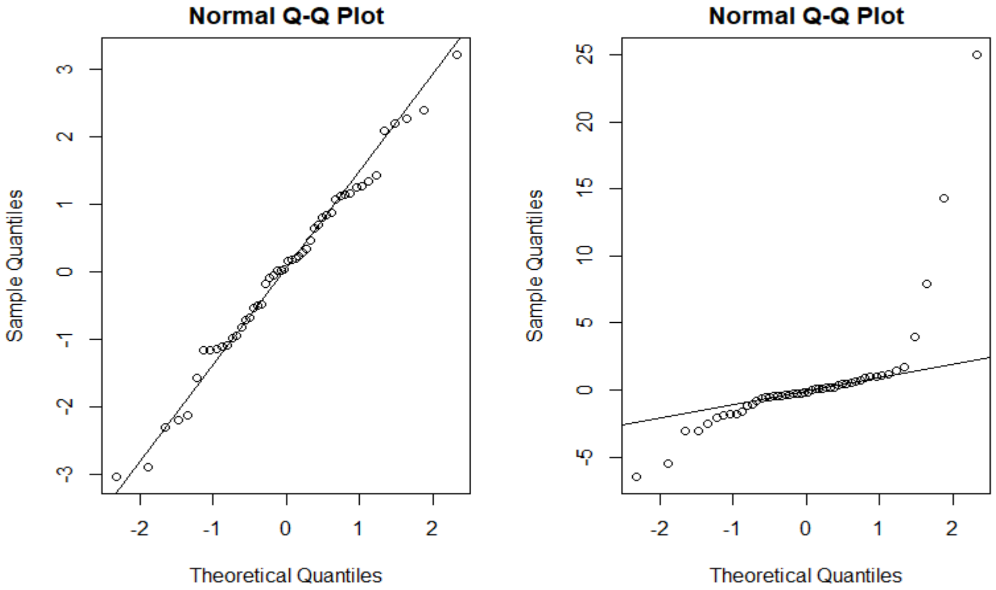
○ 예제】 $Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i, \quad \varepsilon_i \sim iid\ N(0, \sigma^2)$

등분산성 검정

- 등분산성 가정 하에서 σ^2를 MSE로 추정
- 잔차그림을 통해 확인할 수 있음
- Breusch-Pagan 검정

◇ 정규성 검정

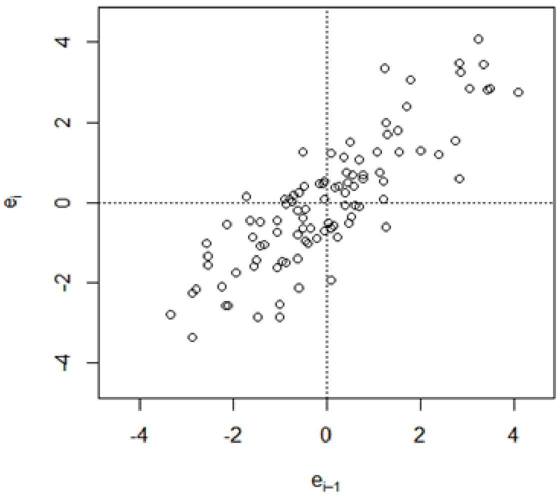
☑ 히스토그램, Q-Q plot, …



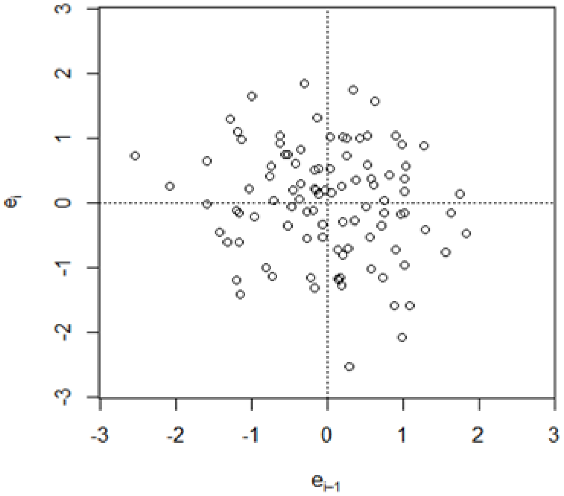
☑ Shapiro-Wilk검정, Jarque-Bera 검정, …

◇ 독립성 검정

- 자료가 시간 순으로 관측된 경우(시계열자료): (e_{t-k}, e_t)의 산점도



양의 자기상관관계 존재



자기상관관계 없음

- Durbin-Watson 검정, ACF, …

잔차검진(Residual Diagnostics)

- 분석에 사용된 회귀모형의 적절성과 통계적 추론의 가정을 만족하는지를 확인하는 방법에 대해 알아본다.
  - 등분산성: 잔차그림, Breusch-Pagan 검정
  - 정규성: 히스토그램, Q-Q plot, Shapiro-Wilk검정, Jarque-Bera 검정
  - 독립성: Durbin-Watson검정, ACF

통계학의 이해 II

강의정리 및 실습

절편에 대한 통계적 추론

- 회귀계수 중 절편에 해당하는 β_0의 중심축량과 구간추정에 대해 알아본다.

✓ 중심축량:

$$\frac{\hat{\beta}_0 - \beta_0}{\sqrt{MSE}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}} \sim t_{n-2}$$

✓ $100(1-\alpha)\%$ 신뢰구간

$$\hat{\beta}_0 \pm t_{\frac{\alpha}{2},n-2}\sqrt{MSE}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}$$

예측값 평균에 대한 통계적 추론

- 예측값의 평균에 대한 중심축량과 예측구간을 알아본다.
  - 중심축량

$$\frac{\hat{Y}_k - E(Y_k)}{\sqrt{MSE}\sqrt{\frac{1}{n} + \frac{(x_k - \bar{x})^2}{S_{xx}}}} \sim t_{n-2}$$

  - $100(1-\alpha)\%$ 신뢰구간

$$\hat{Y}_k \pm t_{\frac{\alpha}{2},n-2}\sqrt{MSE}\sqrt{\frac{1}{n} + \frac{(x_k - \bar{x})^2}{S_{xx}}}$$

새로운 관측값 대한 예측

- 새로운 설명변수에 대한 예측값에 대한 추정과 예측구간을 알아본다.

  - $$\frac{\hat{Y}_* - Y_*}{\sqrt{MSE}\sqrt{1 + \frac{1}{n} + \frac{(x_k - \overline{x})^2}{S_{xx}}}} \sim t_{n-2}$$

  - Y_*에 대한 $100(1-\alpha)\%$ 예측구간

$$\hat{Y}_* \pm t_{\frac{\alpha}{2},n-2}\sqrt{MSE}\sqrt{1 + \frac{1}{n} + \frac{(x_k - \overline{x})^2}{S_{xx}}}$$

잔차검진(Residual Diagnostics)

- 분석에 사용된 회귀모형의 적절성과 통계적 추론의 가정을 만족하는지를 확인하는 방법에 대해 알아본다.
  - 등분산성: 잔차그림, Breusch-Pagan 검정
  - 정규성: 히스토그램, Q-Q plot, Shapiro-Wilk검정, Jarque-Bera 검정
  - 독립성: Durbin-Watson검정, ACF

R 실습

- 선형회귀모형 추론: lm
- 적합값, 예측값, 잔차: predict, residuals
- 잔차검진: plot(잔차그림), shapiro.test, ncvTest, durbinWatsonTest

과제

- 올림픽 100m 우승기록(1900년부터 2004년까지 자료)

  - 남녀별로 나누어 단순선형회귀분석을 하여 잔차검진을 실시하여라.
    - Breusch-Pagan, Shapiro-Wilk, Durbin-Watson 검정
    - 잔차그림, Q-Q plot, ACF

  - 추정된 회귀식을 이용하여 2152, 2156, 2160년 올림픽의 남녀별 우승기록을 예측하여라.

  - Andrew Tatem에 의하면 예측구간을 고려하면, 빠르면 2064년, 늦어도 2788년에는 역전 될 것이라고 주장하였는데 그 근거를 알아보아라.