통계 분석

2019년 2학기

강봉주

회귀 분석

회귀 분석

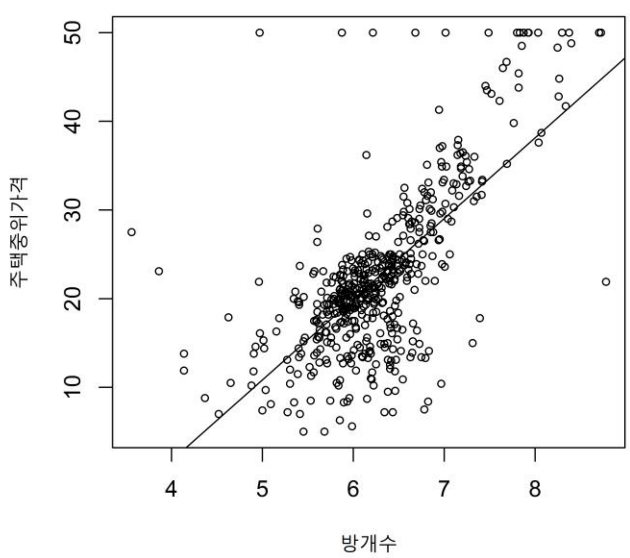
[개요]

- 선형회귀분석((linear regression analysis)는 1개의 종속변수(dependent variable, response variable, target variable)에 대하여 여러 개의 독립변수(independent variable, explanatory variable, predictor variable, input variable)와의 관련성을 선형으로 모형화하는 기법
- 1개의 독립변수만 있는 경우를 단순선형회귀(simple linear regression)

회귀 분석

[개요]

- $Y = \beta_0 + \beta_1 x + \epsilon$
- β_0는 절편(intercept), β_1는 기울기 모수(parameter)
- 모수에 대한 1차식



회귀 분석

[모수의 추정]

- $\{(x_i, y_i) | i = 1, \dots, n\}$: 표본
- $y_i \approx \beta_0 + \beta_1 x_i$가 되도록 모수를 추정
- $r_i = y_i - (\beta_0 + \beta_1 x_i)$: 잔차(residual)
- 잔차 제곱합이 최소가 되도록

$$\min_{\beta_0, \beta_1} \sum_i r_i^2 = \min_{\beta_0, \beta_1} \sum_i \left(y_i - (\beta_0 + \beta_1 x_i)\right)^2$$

회귀 분석

[모수의 추정]

- $\min\limits_{\beta_0,\beta_1} \sum_i r_i^2 = \min\limits_{\beta_0,\beta_1} \sum_i \big(y_i - (\beta_0 + \beta_1 x_i)\big)^2 = \min\limits_{\beta_0,\beta_1} (y - \mathrm{X}\beta)^T (y - \mathrm{X}\beta)$
- β_0, β_1에 대하여 미분한 후 정리
- 정규(수직)방정식(normal equation): $y - \mathrm{X}\beta \perp range(\mathrm{X})$

정규 방정식: $\begin{cases} n\beta_0 + \sum_i x_i \beta_1 = \sum_i y_i \\ \sum_i x_i \beta_0 + \sum_i x_i^2 \beta_1 = \sum_i x_i y_i \end{cases} \rightarrow \mathrm{X}^\mathrm{T}\mathrm{X}\beta = \mathrm{X}^\mathrm{T} y \rightarrow \mathrm{X}^\mathrm{T}(y - \mathrm{X}\beta) = 0$

정규 행렬: $\mathrm{X}^\mathrm{T}\mathrm{X} = \begin{pmatrix} n & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{pmatrix}$

회귀 분석

[모수의 추정]

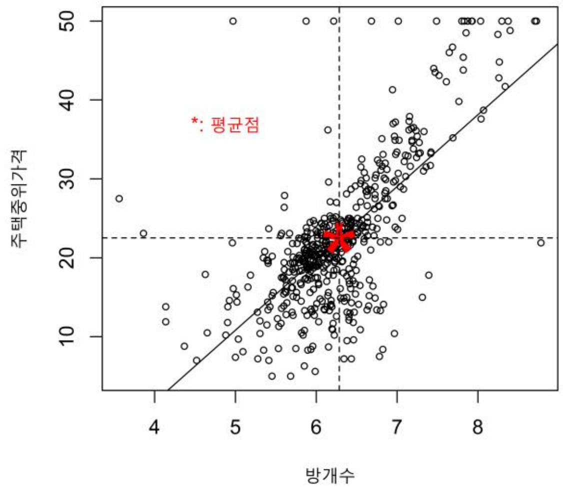
- $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$
- $\hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$

$$\mathrm{X}^{\mathrm{T}}\mathrm{X}\beta = \mathrm{X}^{\mathrm{T}}y$$

회귀 분석

[모수의 추정]

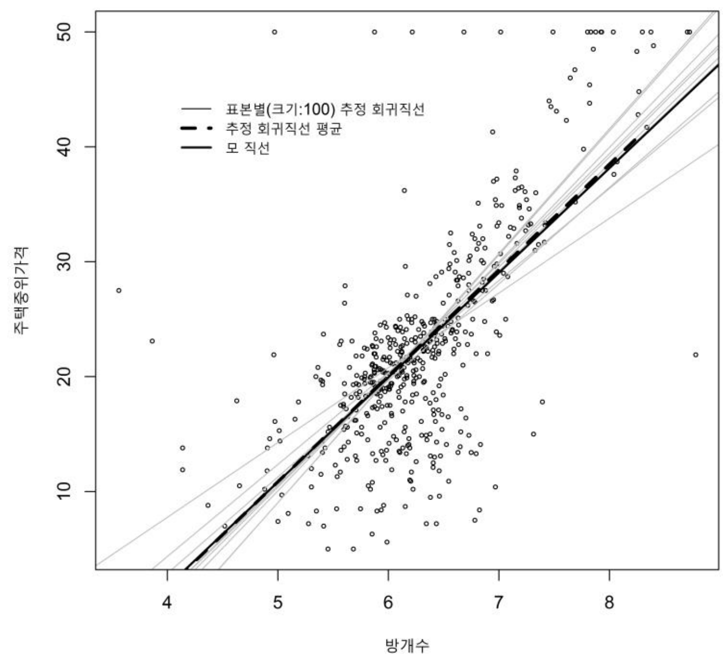
- $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$
- $\hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$
- $\hat{y}_i = \bar{y} + \hat{\beta}_1 (x_i - \bar{x}) \Rightarrow \hat{y}_i - \bar{y} = \hat{\beta}_1 (x_i - \bar{x})$
- 추정된 직선은 반드시 $(\bar{x}, \bar{y})$를 통과하는 기울기가 $\hat{\beta}_1$인 직선



회귀 분석

[추정된 회귀 직선의 의미]

- 표본의 변화가 발생한 경우에 추정된 회귀 직선이 어떻게 변할까?
- 추정된 회귀 직선의 분포가 존재



회귀 분석

[회귀직선의 분포에 대한 가정]

- 오차와 목표 변수만 확률 변수
- 모형에서 추정하고 하는 것은 $E(Y_i|x_i), Y_i$ 을 추정하는 것이 아님

순번	가정	비고
1	$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$	모형식
2	$E(\epsilon_i) = 0$	따라서 $E(Y_i|x_i) = \beta_0 + \beta_1 x_i$ 이므로 모형식의 선형성 가정
3	$Var(\epsilon_i) = \sigma^2$	오차분산의 등분산성 가정
4	$\epsilon_1, \ldots, \epsilon_n$은 서로 독립	오차의 독립성 가정

$$\epsilon_1, \ldots, \epsilon_n \overset{iid}{\sim} (0, \sigma^2)$$

회귀 분석

[오차 분산의 추정]

- 회귀식에서 마지막 남은 모수에 대한 추정
- $r_i = Y_i - \hat{Y}_i$: (표본) 잔차(오차의 추정량)
- $\hat{\sigma}^2 = \frac{\sum_i (Y_i - \hat{Y}_i)^2}{n-2} = \frac{SSE}{n-2} = \frac{SSE}{n-p} = MSE, p$: 추정되는 모수의 개수
- $\hat{\sigma} = \sqrt{MSE}$: 잔차 표준편차 또는 잔차 표준오차
- 추정 오차 분산의 계산식

$$\frac{\sum_i \left(Y_i - \hat{Y}_i\right)^2}{n-2} = \frac{\sum_i \left(Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)\right)^2}{n-2}$$

$$= \frac{\sum_i \left(Y_i - \left(\bar{Y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i\right)\right)^2}{n-2}$$

$$= \frac{1}{n-2}\left(\sum_i (Y_i - \bar{Y})^2 - \frac{\left(\sum_i (x_i - \bar{x})(Y_i - \bar{Y})\right)^2}{\sum_i (x_i - \bar{x})^2}\right)$$

회귀 분석

[오차 분산의 추정]

예제)

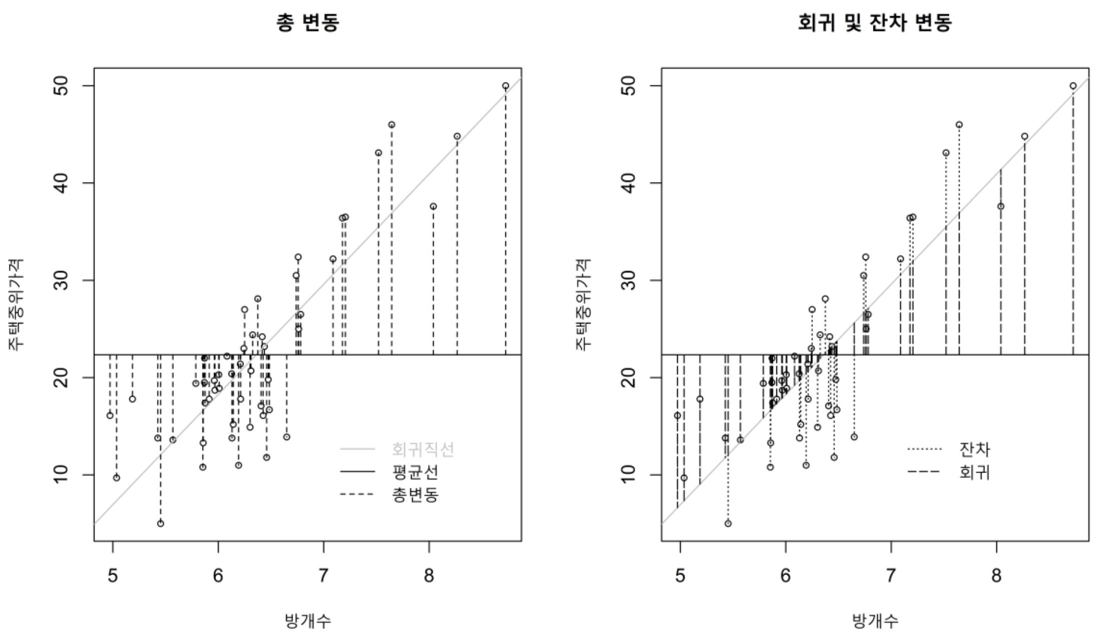
1) [HOUSING] 데이터에서 방의 개수와 주택 중위값(목표변수)간의 표본의 크기가 100인 회귀분석을 실시하고 난 후 오차 분산을 추정하세요.

검증은 lm() 함수의 Residual standard error 값으로 확인

회귀 분석

[회귀직선 간의 비교: 변동의 분해로부터]

- $Y_i = \beta_0 + \epsilon_i$: 절편만 있는 모델(영 모델: null model), $\hat{\beta}_0 = \bar{Y}$
- $\sum_{\mathrm{i}}(Y_i - \bar{Y})^2$: 총 변동
- $Y_i - \bar{Y} = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y})$: 편차의 분해(잔차와 회귀직선)



회귀 분석

[회귀직선 간의 비교: 변동의 분해로부터]

- $\sum_i (Y_i - \bar{Y})^2 = \sum_i \left(Y_i - \hat{Y}_i\right)^2 + \sum_i \left(\hat{Y}_i - \bar{Y}\right)^2$
- SST = SSE + SSR
- 작 적합된 직선이면 잔차에 의한 변동이 작아야 함
- $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$: 결정계수(coefficient of determination)
- $SSR = \sum_i \left(\hat{Y}_i - \bar{Y}\right)^2 = \hat{\beta}_1^2 \sum_i (x_i - \bar{x})^2 = \frac{\left(\sum_i (x_i - \bar{x})(Y_i - \bar{Y})\right)^2}{\sum_i (x_i - \bar{x})^2}$
- $R^2 = \frac{\frac{\left(\sum_i (x_i - \bar{x})(Y_i - \bar{Y})\right)^2}{\sum_i (x_i - \bar{x})^2}}{\sum_i (Y_i - \bar{Y})^2} = \frac{\left(\sum_i (x_i - \bar{x})(Y_i - \bar{Y})\right)^2}{\sum_i (x_i - \bar{x})^2 \sum_i (Y_i - \bar{Y})^2} = (R)^2$: 표본 상관계수의 제곱

회귀 분석

[회귀직선 간의 비교: 변동의 분해로부터]

예제)

1) 앞의 예제에서 결정계수 값을 구하세요.

회귀 분석

[모수의 추론]

- 추론의 대상: 회귀 계수, 회귀직선
- $\epsilon_i \sim N(0, \sigma^2), i = 1, \dots, n$: 분포에 대한 가정
- $\hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_i (x_i - \bar{x})^2}$ 의 분포는?
- $\sum_i (x_i - \bar{x})^2 \overset{let}{=} c$ (상수값)

$$E(Y_i) = E(\beta_0 + \beta_1 x_i + \epsilon_i) = \beta_0 + \beta_1 x_i$$

$$Var(Y_i) = Var(\beta_0 + \beta_1 x_i + \epsilon_i) = Var(\epsilon_i) = \sigma^2$$

$$Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$

회귀 분석

[모수의 추론]

$$E(\hat{\beta}_1) = \frac{1}{c}\ E\left(\sum_i (x_i - \bar{x})Y_i\right) = \frac{1}{c}\sum_i (x_i - \bar{x})\ E(Y_i)$$

$$= \frac{1}{c}\sum_i (x_i - \bar{x})\ (\beta_0 + \beta_1 x_i) = \frac{1}{c}\beta_1 c = \beta_1$$

$$Var(\hat{\beta}_1) = \frac{1}{c^2}\sum_i (x_i - \bar{x})^2 Var(Y_i) = \frac{1}{c^2} c\sigma^2 = \frac{\sigma^2}{c} = \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2}$$

회귀 분석

[모수의 추론]

- $\frac{\widehat{\beta}_1 - \mathrm{E}(\widehat{\beta}_1)}{\sqrt{\mathrm{V}ar(\widehat{\beta}_1)}} = \frac{\widehat{\beta}_1 - \beta_1}{\sigma\sqrt{\frac{1}{\sum_i (x_i - \bar{x})^2}}} \sim N(0,1)$: 오차의 분산을 아는 경우
- $T = \frac{\widehat{\beta}_1 - \beta_1}{\frac{\widehat{\sigma}}{\sqrt{\sum_i (x_i - \bar{x})^2}}} \sim t(n-2)$: 오차 분산을 모르는 경우
- $\left(\hat{\beta}_1 - t_{\frac{\alpha}{2}}(n-2)\frac{\widehat{\sigma}}{\sqrt{\sum_i (x_i - \bar{x})^2}},\ \bar{x} + t_{\frac{\alpha}{2}}(n-2)\frac{\widehat{\sigma}}{\sqrt{\sum_i (x_i - \bar{x})^2}}\right)$: 신뢰 구간

회귀 분석

[모수의 추론]

예제)

1) 앞의 예에서 100개의 표본을 구성한 후 기울기에 대한 95% 신뢰 구간을 구해보세요.

회귀 분석

[모수의 추론]

- 회귀 계수에 대한 가설 검증

- $t_0 = \frac{\hat{\beta}_1 - \beta_1}{\frac{\hat{\sigma}}{\sqrt{\sum_i (x_i - \bar{x})^2}}} \sim t(n-2)$: 오차 분산을 모르는 경우

대립 가설	P값	영가설 기각
$H_1: \beta_1 > 0$	$p_0 = \Pr(T \geq t_0)$	$t_0 \geq t_{\alpha}(n-2)$
$H_1: \beta_1 < 0$	$p_0 = \Pr(T \geq t_0)$	$t_0 \geq t_{1-\alpha}(n-2)$
$H_1: \beta_1 \neq 0$	$p_0 = \Pr(\lvert T \rvert \geq \lvert t_0 \rvert)$	$\lvert t_0 \rvert \geq t_{\frac{\alpha}{2}}(n-2)$

회귀 분석

[모수의 추론]

예제)

1) 앞의 예에서 100개의 표본을 구성한 후 기울기에 대한 대립가설 $H_1: \beta_1 \neq 0$ 에 대하여 5% 유의수준으로 검증하세요.

회귀 분석

[모 회귀직선의 유 의미성에 대한 추론]

- 변동의 분해 기법 이용(분산분석표 이용): 일반적인 접근 방법
- $\sum_i (Y_i - \bar{Y})^2 = \sum_i \left(Y_i - \hat{Y}_i\right)^2 + \sum_i \left(\hat{Y}_i - \bar{Y}\right)^2$
- SST = SSE +SSR
- $H_0\colon \beta_1 = 0 \Rightarrow F = \frac{\frac{SSR}{1}}{\frac{SSE}{n-2}} = \frac{MSR}{MSE} \sim F(1, n-2)$

요인	제곱합	자유도	평균제곱	F값	P값
회귀	SSR	1	$MSR = \frac{SSR}{1}$	$f_0 = \frac{MSR}{MSE}$	$\Pr(F \geq f_0)$
잔차	SSE	$n-2$	$MSE = \frac{SSE}{n-2}$		
계	SST	$n-1$			

회귀 분석

[모 회귀직선의 유 의미성에 대한 추론]

예제)

1) 앞의 예에서 100개의 표본을 구성한 후 모 회귀직선의 유의미성에 대하여 5% 유의 수준으로 검증하세요. 단, 분산 분석표를 이용하세요.

회귀 분석

[모 회귀직선의 신뢰 구간]

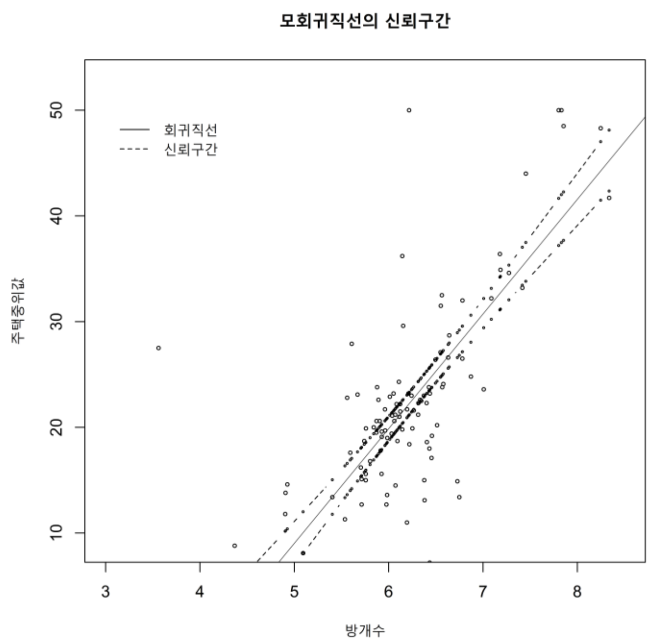
- $\mathrm{E}\left(\hat{Y}_i\right) = \mathrm{E}\left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right) = \beta_0 + \beta_1 x_i$
- $\mathrm{var}\left(\hat{Y}_i\right) = \sigma^2 \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{c}\right)$: 교재 참조
- $\dfrac{\hat{\beta}_0 + \hat{\beta}_1 x_i - (\beta_0 + \beta_1 x_i)}{\hat{\sigma}\sqrt{\left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{c}\right)}} \sim t(n-2)$
- $100(1-\alpha)\%$ 신뢰구간

$$\left(\hat{\beta}_0 + \hat{\beta}_1 x_i - t_{\frac{\alpha}{2}}(n-2)\hat{\sigma}\sqrt{\left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{c}\right)}, \qquad \hat{\beta}_0 + \hat{\beta}_1 x_i + t_{\frac{\alpha}{2}}(n-2)\hat{\sigma}\sqrt{\left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{c}\right)}\right)$$

회귀 분석

[모 회귀직선의 신뢰 구간]

- $\mathrm{E}(\hat{Y}_i) = \mathrm{E}(\hat{\beta}_0 + \hat{\beta}_1 x_i) = \beta_0 + \beta_1 x_i$
- $\mathrm{var}(\hat{Y}_i) = \sigma^2 \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{c}\right)$



- 독립변수의 평균값에서 멀어질 수록 분산이 커짐
- 독립변수의 평균값에서 분산이 제일 작음

회귀 분석

[잔차 분석]

- 회귀 분석의 각 종 가정에 대한 분석
- 잔차를 통하여 분석
- $\sum_{\mathrm{i}} \hat{r}_i = \sum_i \left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\right) = \sum_i \left(y_i - (\bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i)\right) = 0$
- $E(r_i) = E\left(Y_i - \hat{Y}_i\right) = \beta_0 + \beta_1 x_i - (\beta_0 + \beta_1 x_i) = 0$
- $Var(r_i) = \sigma^2 \left(1 - \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{c}\right)\right) = \sigma^2 \left(\frac{n-1}{n} - \frac{(x_i - \bar{x})^2}{c}\right), c = \sum_i (x_i - \bar{x})^2$

✓ 잔차의 합은 0이다. 즉, 서로 간의 관련성이 있다.

✓ 잔차의 분산은 오차의 분산보다 항상 작다.

✓ 잔차의 분산은 독립변수의 평균에서 멀어질수록 작아진다. 즉 잔차의 위치에 따라 분산이 일정하지 않다.

회귀 분석

[잔차 분석]

- $Z_i = \frac{\epsilon_i - E(\epsilon_i)}{\sqrt{Var(\epsilon_i)}} = \frac{\epsilon_i}{\sigma} = \frac{Y_i - (\beta_0 + \beta_1 x_i)}{\sigma} \sim N(0,1)$: 관측되지 않는 값

- $T_i = \frac{Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)}{\sqrt{\widehat{Var(r_i)}}} = \frac{r_i}{\widehat{\sigma_i}}$: 관측되는 버전으로 대치
- 내부 스튜던트화 잔차(internally Studentized residual), 표준화 잔차(standardized residual)

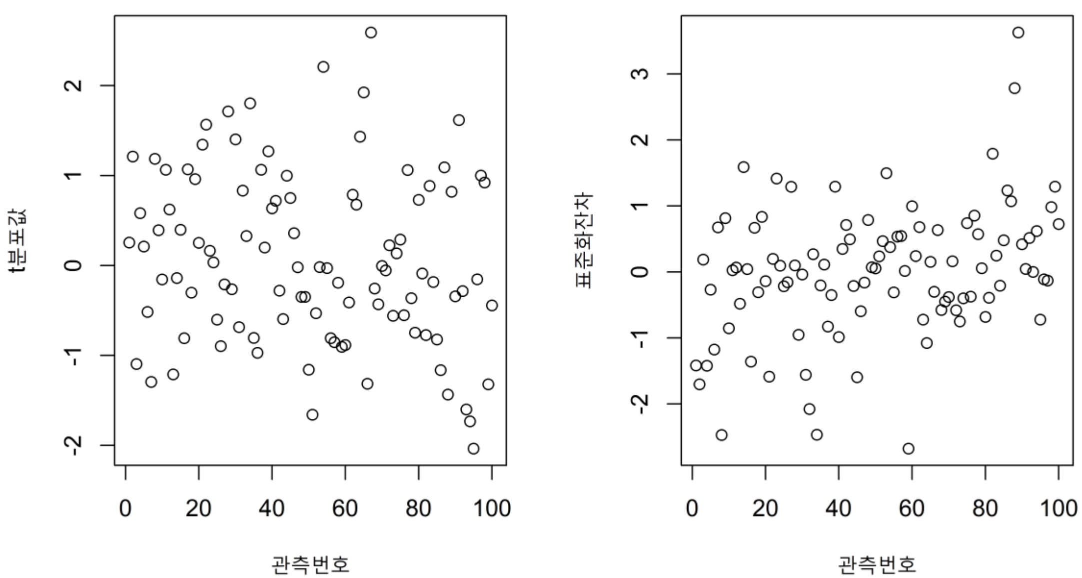
- $\hat{\sigma}_i = \hat{\sigma}\sqrt{\left(1 - \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{c}\right)\right)}$

✓ 표준화 잔차는 잔차의 위치에 무관하게 분산이 일정

회귀 분석

[잔차 분석]

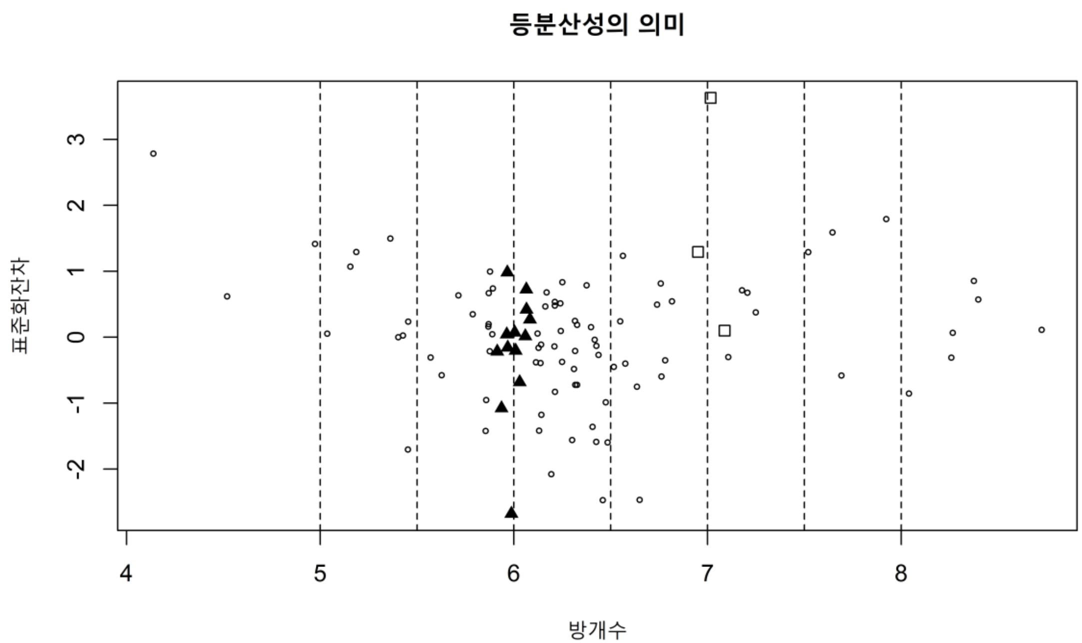
- 독립성에 대한 검증
- 관측 번호에 따른 오차들이 서로 독립: 어떠한 연관성도 없어야 함



회귀 분석

[잔차 분석]

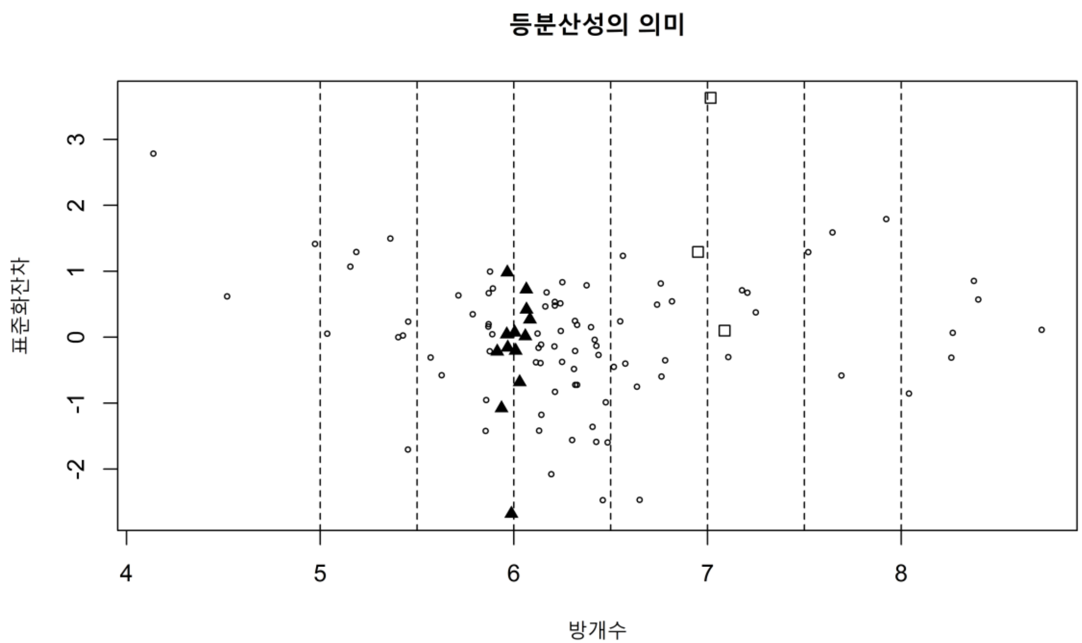
- 등분산성에 대한 검증
- 관측 번호에 따른 오차의 분산들이 일정
- 독립변수의 각 관측값에 관계 없이 오차의 분산들이 일정



회귀 분석

[잔차 분석]

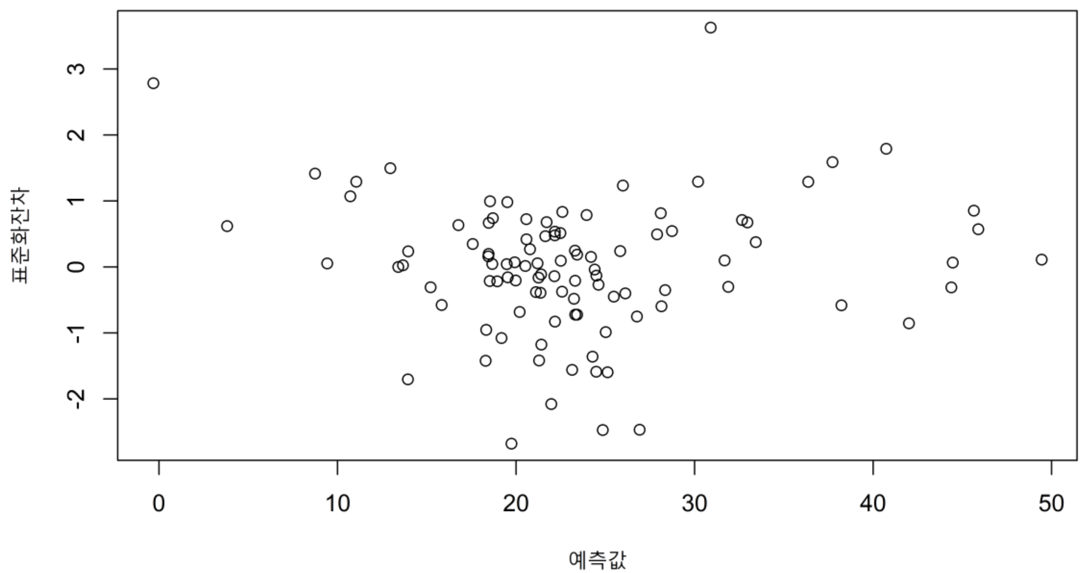
- 등분산성에 대한 검증
- 관측 번호에 따른 오차의 분산들이 일정
- 독립변수의 각 관측값에 관계 없이 오차의 분산들이 일정



회귀 분석

[잔차 분석]

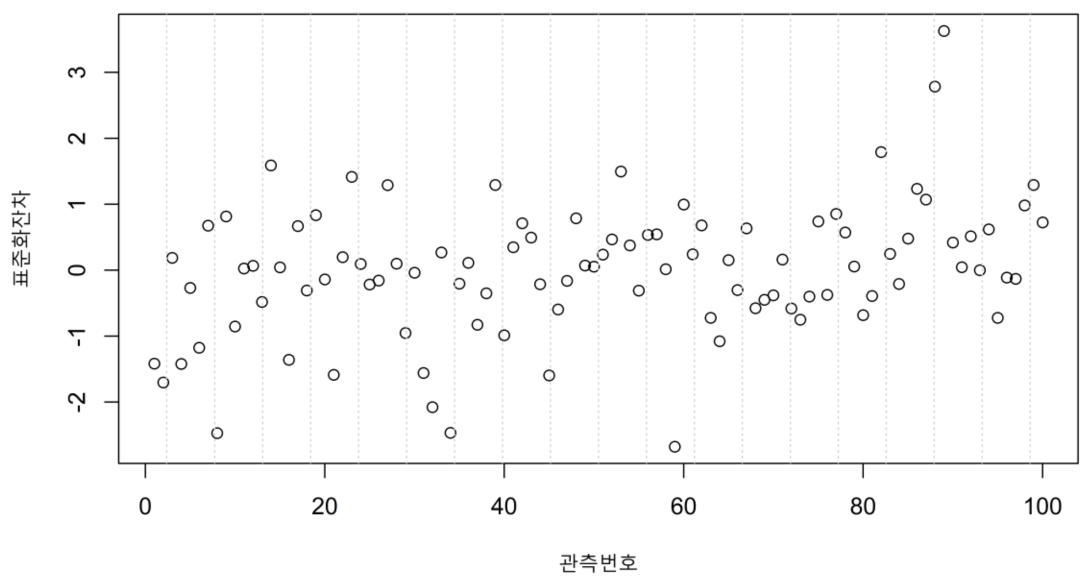
- 등분산성에 대한 검증
- 관측 번호에 따른 오차의 분산들이 일정
- 독립변수의 각 관측값에 관계 없이 오차의 분산들이 일정



회귀 분석

[잔차 분석]

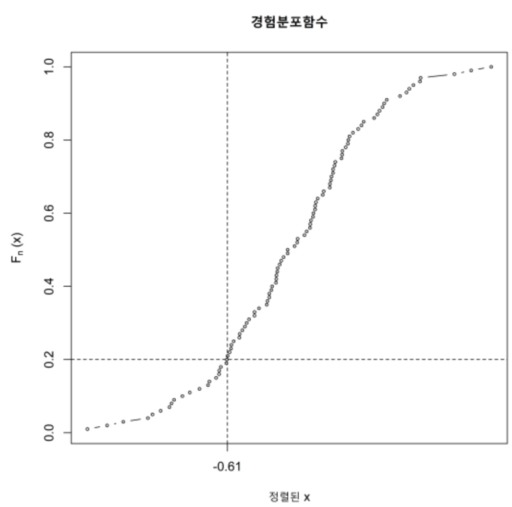
- 등분산성에 대한 검증
- 관측 번호에 따른 오차의 분산들이 일정
- 독립변수의 각 관측값에 관계 없이 오차의 분산들이 일정



회귀 분석

[잔차 분석]

- 정규성에 대한 검증: 데이터의 건수가 많으면 그다지 문제가 안됨
- 경험 분포의 분포 함수
- $F_n(x) = \dfrac{x\text{보다 작은 데이터 개수}}{n} = \frac{1}{n}\sum_i 1_{X_i \le x}$



회귀 분석

[잔차 분석]

- 분위수-분위수 그림(QQ plot)