회귀분석

정의

- 개요

1. 한 변수와 다른 변수와의 관계를 통해 Y의 기댓값을 추정하는 것
2. E(Y) =μ(x)라는 어떤 함수의 정의를 통해 알려진 관측값 $x_1, \dots, x_n$에 대하여 확률변수 Y의 반응값을 관측한다.

1) 이 때, n개의 관측된 쌍 (x_1, y_1), (x_2, y_2) ,…, (x_n, y_n)에 대하여

2) Y_i=$\alpha + \beta(x_i - \bar{x}) + e_i$의 선형함수를 정의하자.

(1) 이 때, e_i는 $\mathrm{N}(0, \sigma^2)$을 따르는 확률변수이고, $\alpha + \beta(x_i - \bar{x})$는 위치 이동 모수이므로

(2) Y_i~$\mathrm{N}[\alpha + \beta(x_i - \bar{x}), \sigma^2]$을 따른다.

3) 이 때, Y의 우도함수는 $\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{[y_i - \alpha - \beta(x_i - \bar{x})]^2}{2\sigma^2}\right) = \left[\frac{1}{\sqrt{2\pi}\sigma}\right]^{\frac{n}{2}} \exp\left(-\frac{\sum y_i - \alpha - \beta(x_i - \bar{x})^2}{2{\theta_0}^2}\right)$

정의

- 개요

3) 이 때, Y의 우도함수는 $\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{[y_i-\alpha-\beta(x_i-\bar{x})]^2}{2\sigma^2}\right) = \left[\frac{1}{\sqrt{2\pi}\sigma}\right]^{\frac{n}{2}} \exp\left(-\frac{\sum[y_i-\alpha-\beta(x_i-\bar{x})]^2}{2{\theta_0}^2}\right)$

(1) 이를 이용하여$\mu = \alpha + \beta(x_i - \bar{x})$의 최댓값 μ_{mle}를 추정하면

(2) $\mathrm{l}(\alpha, \beta, \sigma^2) = \frac{n}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}\sum[y_i - \alpha - \beta(x_i - \bar{x})]^2$

- $\frac{\partial \mathrm{l}(\alpha,\beta,\sigma^2)}{\partial \alpha} = 2\sum[y_i - \alpha - \beta(x_i - \bar{x})] \cdot (-1) = 0$
- $\frac{\partial \mathrm{l}(\alpha,\beta,\sigma^2)}{\partial \beta} = 2\sum[y_i - \alpha - \beta(x_i - \bar{x})] \cdot [-(x_i - \bar{x})] = 0$
- $\frac{\partial \mathrm{l}(\alpha,\beta,\sigma^2)}{\partial \sigma^2} = \frac{n}{2\sigma^2} - \frac{-\sum[y_i-\alpha-\beta(x_i-\bar{x})]^2}{2(\sigma^2)^2}=0$

(3) 각각의 파라미터에 대해 정리하면

- $\hat{\alpha} = \bar{Y} = \frac{\sum y_i}{n}$
- $\hat{\beta} = \frac{\sum y_i(x_i-\bar{x})}{\sum(x_i-\bar{x})^2}$
- $\sigma^2 = \frac{1}{n}\sum[y_i - \alpha - \beta(x_i - \bar{x})]^2$
- 한편 $E[Y_i] = \hat{Y} = \alpha + \beta(x_i - \bar{x})$ 일 때$(y_i - \hat{Y}) = y_i - \alpha - \beta(x_i - \bar{x}) = e_i$이므로 $\frac{1}{n}\sum {e_i}^2 = \sigma^2$ 이다.

정의

- 파라미터 α, β의 분포 추정

1. $\hat{\alpha}$는 iid이고, 확률변수 Y와 관련된 선형함수이다. 이 때

1) $\mathrm{E}[\hat{\alpha}]$ = $\frac{\sum E(y_i)}{n}$ 에서

(1) $E(y_i)$ = $\alpha + \beta(x_i - \bar{x})$ 에서 $\frac{1}{n}[\sum \alpha + \beta(x_i - \bar{x})]$ = $\frac{1}{n}\mathrm{n}\alpha$ + $\sum \beta(x_i - \bar{x})$ = α

(2) 따라서 $\frac{\sum y_i}{n} = \bar{Y}$는 $\hat{\alpha}$에 대한 불편추정량이다.

2) $\mathrm{var}[\hat{\alpha}]$ = $\frac{\sum var(y_i)}{n^2} + \sum 0$ = $\frac{\mathrm{n}\sigma^2}{n^2}$ = $\frac{\sigma^2}{\mathrm{n}}$

(1)) $\hat{\alpha}$ ~ $\mathrm{N}(\alpha, \frac{\sigma^2}{\mathrm{n}})$ 이다.

정의

- 파라미터 α, β의 분포 추정

2. $\hat{\beta}$는 iid이고, 확률변수 Y와 관련된 선형함수이다. 이 때

1) $\mathrm{E}[\hat{\beta}] = \frac{\sum y_i(x_i - \bar{x})}{\sum(x_i - \bar{x})^2}$에서

(1) $E(y_i) = \alpha + \beta(x_i - \bar{x})$ 에서 $\frac{\sum[\alpha + \beta(x_i - \bar{x})](x_i - \bar{x})}{\sum(x_i - \bar{x})^2} = \frac{\alpha\sum(x_i - \bar{x}) + \beta\sum(x_i - \bar{x})^2}{\sum(x_i - \bar{x})^2} = \beta$

(2) 따라서$\frac{\sum y_i(x_i - \bar{x})}{\sum(x_i - \bar{x})^2}$는 $\hat{\alpha}$에 대한 불편추정량이다.

2) $\mathrm{var}[\hat{\beta}] = \left[\frac{\sum(x_i - \bar{x})}{\sum(x_i - \bar{x})^2}\right]^2 var(y_i) = \left[\frac{\sum(x_i - \bar{x})}{\sum(x_i - \bar{x})^2}\right]^2 \sigma^2 = \frac{\sigma^2}{\sum(x_i - \bar{x})^2}$

3) 따라서 $\mathrm{N}\left(\frac{\sum y_i(x_i - \bar{x})}{\sum(x_i - \bar{x})^2}, \frac{\sigma^2}{\sum(x_i - \bar{x})^2}\right)$를 따른다.

정의

- σ^2의 분포 추정

1. Q = $\sum[y_i - \alpha - \beta(x_i - \bar{x})]^2$에서, 이는 2차형식으로 볼 수 있다. 따라서

1) Q = $Q_1 + Q_2 + Q_3$로 분해하면

2) $\sum\{(\hat{\alpha} - \alpha) + \left(\hat{\beta} - \beta\right)(x_i - \bar{x}) + [y_i - \alpha - \beta(x_i - \bar{x})]\}^2$

(1) $\mathrm{n}(\hat{\alpha} - \alpha) + \left(\hat{\beta} - \beta\right)\sum(x_i - \bar{x})^2$ + $\mathrm{n}\sigma^2$ 이다.

(2) 이 때, $\frac{Q_3}{\sigma^2} \sim x^2[r - r_1 - r_2]$ 이므로, 이를 이용하여 σ^2의 분포를 추정 가능하다.

정의

- σ^2의 분포 추정

1. Q = $\sum[y_i - \alpha - \beta(x_i - \bar{x})]^2$에서, 이는 2차형식으로 볼 수 있다. 따라서

1) Q = $Q_1 + Q_2 + Q_3$로 분해하면

2) $\sum\{(\hat{\alpha} - \alpha) + \left(\hat{\beta} - \beta\right)(x_i - \bar{x}) + [y_i - \alpha - \beta(x_i - \bar{x})]\}^2$

(1) $\mathrm{n}(\hat{\alpha} - \alpha) + \left(\hat{\beta} - \beta\right)\sum(x_i - \bar{x})^2$ $+\mathrm{n}\hat{\sigma}^2$ 이다.

(2) 이 때, $\frac{Q_3}{\sigma^2} \sim x^2[r - r_1 - r_2]$ 이므로, 이를 이용하여 σ^2의 분포를 추정 가능하다.

정의

- 파라미터의 신뢰구간 추정

1. Q = $\sum[y_i - \alpha - \beta(x_i - \bar{x})]^2$ 은

1) $\mathrm{E}[y_i - \alpha - \beta(x_i - \bar{x})] = E[e_i]$ = 0

2) $\mathrm{var}[\mathrm{y_i} - \alpha - \beta(\mathrm{x_i} - \bar{\mathrm{x}})] = \mathrm{var}[e_i] = \sigma^2$

3) 이는 곧 $\mathrm{N}(0, \sigma^2)$을 따른다는 것을 알 수 있다.

2. $\frac{y_i-\alpha-\beta(x_i-\bar{x})}{\sigma}$ 는 CLT에 따라 N(0,1)을 따른다.

1) 따라서, 그 2차형식 $\left[\frac{y_i-\alpha-\beta(x_i-\bar{x})}{\sigma}\right]^2 \sim x^2(1)$ 이고, $\sum[\frac{y_i-\alpha-\beta(x_i-\bar{x})}{\sigma^2}]^2 \sim x^2(n)$ 을 따른다.

정의

- 파라미터의 신뢰구간 추정

3. 한편, $\frac{Q_1}{\sigma^2} = \frac{\mathrm{n}(\hat{\alpha}-\alpha)^2}{\sigma^2} \sim x^2(1)$ 이고, $\frac{Q_2}{\sigma^2} = \left(\hat{\beta}-\beta\right)\sum(x_i-\bar{x})^2 \sim x^2(1)$ 이므로

1) $Q_3 = \frac{\mathrm{n}\hat{\sigma}^2}{\sigma^2} = x^2(n-1-1=n-2)$을 따른다.

2) 이 때,

(1) $T_1 = \frac{\sqrt{n}(\hat{\alpha}-\alpha)/\sigma}{\sqrt{\frac{\mathrm{n}\hat{\sigma}^2}{\sigma^2}/(n-2)}} = \frac{(\hat{\alpha}-\alpha)}{\sqrt{\hat{\sigma}^2/(n-2)}} \sim T(n-2)$ 이고

(2) $T_2 = \frac{\frac{\sqrt{(\hat{\beta}-\beta)\sum(x_i-\bar{x})^2}}{\sigma}}{\sqrt{\frac{\mathrm{n}\hat{\sigma}^2}{\sigma^2}/(n-2)}} = \frac{(\hat{\beta}-\beta)}{\sqrt{\frac{\mathrm{n}\hat{\sigma}^2}{(n-2)\sum(x_i-\bar{x})^2}}} \sim T(n-2)$ 이므로

(3) 이를 이용하여 각 파라미터에 대한 신뢰구간을 정의할 수 있다.