5. Prediction

new x_0 to predict the y_0

$$E(y_0) = \beta_0 + \beta_1 x_0$$ [redacted]

$$var(y_0) = \sigma^2$$ [redacted]

$$y_0 \perp \{y_i : i = 1, 2, \cdots, n\}$$ [redacted]

5.1 Point prediction 点(单值)预测

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$ [redacted]

$\hat{y}_0$ 是对 y_0 的预测，却是对 $E(y_0)$ 的无偏估计 [redacted]

Proof

$$E(\hat{y}_0) = E(\hat{\beta}_0) + E(\hat{\beta}_1) \cdot x_0 = \beta_0 + \beta_1 x_0 = E(y_0)$$

故 $\hat{y}_0$ 是对 $E(y_0)$ 的无偏估计

Remark：估计 $E(y_0)$，而没有估计 y_0

Remark：y_0 是一个随机变量

$E(y_0)$ 是一个未知的参数

5.2 Prediction Interval 区间预测

对于给定的显著性水平 α （例 $\alpha = 0.05$），找一个区间 (T_1, T_2)

使 $P(T_1 < y_0 < T_2) = 1 - \alpha$

5.2.1 y_0 Prediction Inteval : PI

As we know

$$y_0 \sim N(\beta_0+\beta_1 x_0,\ \sigma^2)$$ [redacted]

Remember:

$$\hat{y}_0 = \sum_{i=1}^{n}\left(\frac{1}{n}+\frac{(x_i-\bar{x})(x_0-\bar{x})}{L_{xx}}\right)y_i$$ [redacted]

So:

$$var(\hat{y}_0) = \left(\frac{1}{n}+\frac{(x_0-\bar{x})^2}{L_{xx}}\right)\sigma^2$$ [redacted]

Therefore

$$\hat{y}_0 \sim N\left(\beta_0+\beta_1 x_0,\ \left(\frac{1}{n}+\frac{(x_0-\bar{x})^2}{L_{xx}}\right)\sigma^2\right)$$ [redacted]

记 $h_{00} = \frac{1}{n}+\frac{(x_0-\bar{x})^2}{L_{xx}}$

独立性： $y_0 \perp y_1, y_2, \cdots, y_n$

$\hat{y}_0 = \text{linear function}\ (y_1, y_2, \cdots, y_n)$

$\Rightarrow y_0 \perp \hat{y}_0$

二者是独立的，正态分布的样本，且

$$\begin{cases} E(y_0-\hat{y}_0) = 0 \\ var(y_0-\hat{y}_0) = \sigma^2 + h_{00}\sigma^2 \end{cases}$$ [redacted]

从而：

☆ $$y_0-\hat{y}_0 \sim N(0,\ (1+h_{00})\sigma^2)$$ [redacted]

① t分布

☆ $$t = \frac{y_0-\hat{y}_0}{\sqrt{1+h_{00}}\cdot\hat{\sigma}} \sim t(n-2)$$ [redacted]

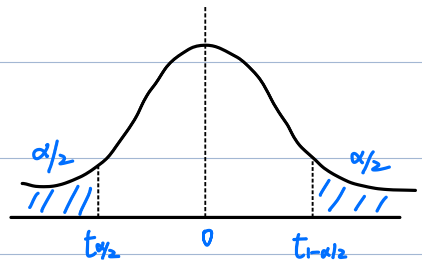
Proof: $$\frac{y_0-\hat{y}_0}{\sqrt{1+h_{00}}\cdot\sigma} \sim N(0,1)$$

$$\frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2)$$

$$\Rightarrow \quad \frac{y_0-\hat{y}_0}{\sqrt{1+h_{00}}\cdot\sigma} \Big/ \sqrt{\frac{(n-2)\hat{\sigma}^2}{\sigma^2}\Big/(n-2)} \sim t(n-2)$$

即 $$t = \frac{y_0-\hat{y}_0}{\sqrt{1+h_{00}}\cdot\hat{\sigma}} \sim t(n-2)$$

② PI (y_0 的预测区间



$$P\left(\left|\frac{y_0-\hat{y}_0}{\sqrt{1+h_{00}}\cdot\hat{\sigma}}\right| \leqslant t_{1-\alpha/2}(n-2)\right) = 1-\alpha$$

PI $$\left[\hat{y}_0 - t_{1-\alpha/2}(n-1)\cdot\sqrt{1+h_{00}}\cdot\hat{\sigma},\quad \hat{y}_0 + t_{1-\alpha/2}(n-1)\cdot\sqrt{1+h_{00}}\cdot\hat{\sigma}\right]$$

↑ y_0 的预测区间

5.2.2 $E(y_0)$ Confidence Interval : CI

① $E(y_0)$ 参数

$$E(y_0) = \beta_0 + \beta_1 x_0$$

是一个参数（未知的常数）

② t分布

$$\hat{y}_0 \sim N(\beta_0+\beta_1 x_0,\ h_{00}\sigma^2)$$

$$\hat{y}_0 - E(y_0) \sim N(0,\ h_{00}\sigma^2)$$

$$\frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2)$$

☆ $$t = \frac{\hat{y}_0 - E(y_0)}{\sqrt{h_{00}}\,\hat{\sigma}} \sim t(n-2)$$ [redacted]

③ CI

$$P\left(\left|\frac{\hat{y}_0 - E(y_0)}{\sqrt{h_{00}}\,\hat{\sigma}}\right| \leq t_{1-\alpha/2}(n-2)\right) = 1-\alpha$$ [redacted]

CI $$\left[\hat{y}_0 - t_{1-\alpha/2}(n-2)\cdot\sqrt{h_{00}}\,\hat{\sigma},\ \hat{y}_0 + t_{1-\alpha/2}(n-2)\cdot\sqrt{h_{00}}\,\hat{\sigma}\right]$$ [redacted]

$E(y_0)$ 的置信区间

5.3 Relationship of PI, CI

y_0 PI $$\left[\hat{y}_0 - t_{1-\alpha/2}(n-1)\cdot\sqrt{1+h_{00}}\cdot\hat{\sigma},\ \hat{y}_0 + t_{1-\alpha/2}(n-1)\cdot\sqrt{1+h_{00}}\cdot\hat{\sigma}\right]$$ [redacted]

$E(y_0)$ CI $$\left[\hat{y}_0 - t_{1-\alpha/2}(n-2)\cdot\sqrt{h_{00}}\,\hat{\sigma},\ \hat{y}_0 + t_{1-\alpha/2}(n-2)\cdot\sqrt{h_{00}}\,\hat{\sigma}\right]$$ [redacted]

$\text{len}(CI) < \text{len}(PI)$: CI 更精确

And $n \longrightarrow \infty$ $h_{00} \longrightarrow 0$

$CI \rightarrow \hat{y}_0$ $PI \rightarrow [\hat{y}_0 \pm t_{1-\alpha/2}(n-2)\,\hat{\sigma}]$