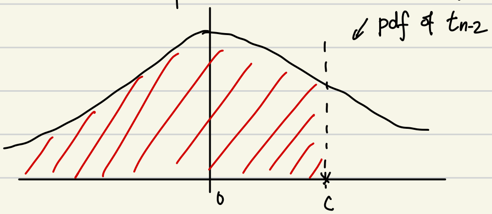
Lecture 4 Stat 331

Last class

$$\frac{\hat{\beta}_1-\beta_1}{SE(\hat{\beta}_1)} \sim t_{n-2} \quad , \quad SE(\hat{\beta}_1)=\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$$



Suppose $T \sim t_{n-2}$, $P(T \le c) = 0.975 \Leftrightarrow F(c) = 0.975$

Quantile function is the inverse of a CDF s.t. $F^{-1}(0.975) = c$

$\therefore$ c the 0.975 quantile of t_{n-2} dist^n.

$$c = t_{0.975,\, n-2}$$

(c : the value for which the CDF is 0.975).

We wish to use the sampling dist^n of $\hat{\beta}_1$ (LS estimator) to:

(1) Construct confidence intervals

(2) Conduct hypothesis test to determine if $\beta_1 = 0$.

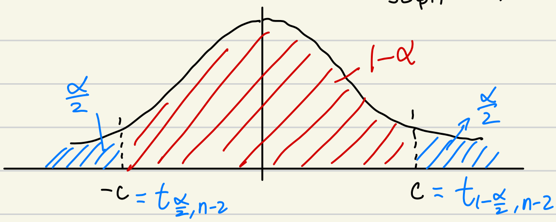
Confidence Intervals (CI) :

We want to find the $100(1-\alpha)\%$ CI.

α: significance level (usually 0.05)

Given $\frac{\hat{\beta}_1-\beta_1}{SE(\hat{\beta}_1)} \sim t_{n-2}$: We want to find

$$P\left(-c \le \frac{\hat{\beta}_1-\beta_1}{SE(\hat{\beta}_1)} \le c\right) = 1-\alpha$$



C: $1-\frac{\alpha}{2}$ quantile of t_{n-2} $\left(t_{1-\frac{\alpha}{2},n-2} = -t_{\frac{\alpha}{2},n-2}\right)$

The $100(1-\alpha)\%$ CI is $\hat{\beta}_1 \pm c\,SE(\hat{\beta}_1)$

$$\left(\hat{\beta}_1 - c\,SE(\hat{\beta}_1)\ ,\ \hat{\beta}_1 + c\,SE(\hat{\beta}_1)\right)$$

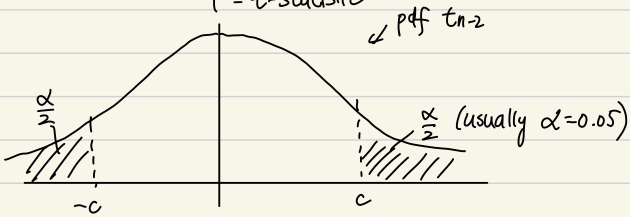
(c: critical value)

Hypothesis test: Suppose that we want to test whether or not $\beta_1 = 0$

We can write the hypothesis as follows:

$$\underbrace{H_0 : \beta_1 = 0}_{\text{null hypothesis}} \quad \text{vs.} \quad \underbrace{H_a : \beta_1 \neq 0}_{\text{alternative hypothesis}}$$

If H_0 is true, $\underbrace{\hat{\beta}_1 / SE(\hat{\beta}_1)}_{T = t\text{-statistic}} \sim t_{n-2}$



Decision point: If $|t| > c$, reject $H_0 : \beta_1 = 0$.

Given a random sample, observing $|t|$ in the tails of the t_{n-2} dist$^{\underline{n}}$ indicates strong evidence to reject H_0.

p-value $= P(|T| \geq |t|) = 2P(T \geq |t|)$

if $|t| = c$, then $2P(T \geq c) = \alpha$

if $|t| > c$, then $2P(T \geq |t|) < \alpha$

(c is the $1 - \frac{\alpha}{2}$ quantile of t_{n-2})

In words, p-value is the probability of observing a test-statistic that is at least as extreme as the one observed from our data under H_0.

Remarks: (1) In t-test, we do not reject H_0 if $|t| \leq c$

$|t| \leq c$ is same as $|\hat{\beta}_1 / SE(\hat{\beta}_1)| \leq c$

$\Rightarrow -c \leq \hat{\beta}_1 / SE(\hat{\beta}_1) \leq c$

$\Rightarrow \underbrace{\hat{\beta}_1 - c\,SE(\hat{\beta}_1)} \leq 0 \leq \underbrace{\hat{\beta}_1 + c\,SE(\hat{\beta}_1)}$

$\rightarrow 100(1-\alpha)\%$ CI

$\therefore$ We do not reject H_0 if the $100(1-\alpha)\%$ CI contains 0 (hypothesized null value)

(2) Not rejecting $H_0 \neq$ accept H_0.

Prediction using SLR.

Recall : Suppose that we want to calculate mean response at a particular value of x: $x = x_p$.

From SLR we have

$$\mu_p = E(Y \mid x_p) = \beta_0 + \beta_1 x_p$$

We can estimate $\hat{\mu}_p = \hat{\beta}_0 + \hat{\beta}_1 x_p$

① Construct $100(1-\alpha)\%$ CI for mean response:

We must first determine the sampling distⁿ of $\hat{\mu}_p$. As a R.V.

① $E(\hat{\mu}_p) = E(\hat{\beta}_0 + \hat{\beta}_1 x_p) = E(\hat{\beta}_0) + x_p E(\hat{\beta}_1)$

$= \beta_0 + \beta_1 x_p = \mu_p$.

(unbiased).

② $Var(\hat{\mu}_p) = Var(\hat{\beta}_0 + \hat{\beta}_1 x_p)$ plug in $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

$= Var[\bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_p]$

$= Var(\bar{y} + \hat{\beta}_1 (x_p - \bar{x})]$

$= Var(\bar{y}) + (x_p - \bar{x})^2 Var(\hat{\beta}_1) + 2(x_p - \bar{x}) \overbrace{Cov(\bar{y}, \hat{\beta}_1)}^{=0}$

$= \sigma^2/n + (x_p - \bar{x})^2 \sigma^2 / S_{xx}$

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

Since $\hat{\mu}_p = \sum_{i=1}^{n} a_i Y_i$ $\therefore$ $\hat{\mu}_p \sim N\left(\mu_p, \sigma^2 \left(\frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}}\right)\right)$.

$\therefore$ The sampling distⁿ is given by

$$\frac{\hat{\mu}_p - \mu_p}{SE(\hat{\mu}_p)} \sim t_{n-2}, \quad SE(\hat{\mu}_p) = \sqrt{\hat{\sigma}^2 \left(\frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}} \right)}$$

$\therefore$ $100(1-\alpha)\%$ CI is given by $\hat{\mu}_p \pm c\,SE(\hat{\mu}_p)$; where c is $1-\frac{\alpha}{2}$ quantile of t_{n-2} distⁿ.

② Prediction of response for a new observation

Suppose that we want to make prediction for a new observation (not part of our sample) with explanatory variable $x = x_0$.

The (true) response variable follows $Y_0 = \beta_0 + \beta_1 x_0 + \varepsilon_0$, ε_0 is the error term for this new observation. ($\varepsilon_0 \sim N(0, \sigma^2)$).

Naturally, we can replace β_0 and β_1 with LS estimates to predict response.

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

The prediction error is given by $y_0 - \hat{y}_0$

Some properties of $Y_0 - \hat{Y}_0$ (as a R.V.)

① $E(Y_0 - \hat{Y}_0) = 0$

Pf:
$$\begin{aligned} E(Y_0 - \hat{Y}_0) &= E(Y_0) - E(\hat{Y}_0) \\ &= \beta_0 + \beta_1 x_0 - \beta_0 - \beta_1 x_0 \\ &= 0 \quad \blacksquare \end{aligned}$$

②
$$\begin{aligned} Var(Y_0 - \hat{Y}_0) &= Var(\beta_0 + \beta_1 x_0 + \varepsilon_0 - \hat{\beta}_0 - \hat{\beta}_1 x_0) \\ &= Var(\varepsilon_0 - \hat{\beta}_0 - \hat{\beta}_1 x_0) \end{aligned}$$

Because $\hat{\beta}_0$ and $\hat{\beta}_1$ are functions of our analyzed data, the new random error ε_0 is not related to the data, $\therefore$ ε_0 is indpt of $\hat{\beta}_0, \hat{\beta}_1$

$$\begin{aligned} &= Var(\varepsilon_0) + Var(\hat{\beta}_0 + \hat{\beta}_1 x_0) \\ &= \sigma^2 + \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right) \\ &= \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right]. \end{aligned}$$

$\therefore$ We have,

$$\frac{Y_0 - \hat{Y}_0}{SE(Y_0 - \hat{Y}_0)} \sim t_{n-2} \quad \text{where } SE(Y_0 - \hat{Y}_0) = \sqrt{\hat{\sigma}^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right]}$$

Intuition for variance of $Y_0 - \hat{Y}_0$: There are two sources of uncertainty

① Uncertainty associated with parameter estimators $\hat{\beta}_0$ and $\hat{\beta}_1$

② Uncertainty associated with random error of the new observation.

$\therefore$ The $100(1-\alpha)\%$ prediction interval (PI) for response is

$$\hat{y}_0 \pm c\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}},$$ where c is $1-\frac{\alpha}{2}$ quantile of t_{n-2}.

Example Lung function data (Kahn, 2005)

Studies the lung function in children and teens. We want to study the association between lung function (measured as FEV) and age (in years).

Q. Suppose we want to estimate mean response for a child who is 8 yrs old. $(n=655)$, $\hat{\beta}_0 = 0.4705$, $\hat{\beta}_1 = 0.2187$, $\bar{x} = 9.9237$, $S_{xx} = 5722.1832$

$SS(Res) = 224.8002$.

A: $\hat{\mu} = 0.4705 + 0.2187(8)$; $c = t_{0.975, 653} = 1.9636$

$\therefore$ 95% CI is given by

$$\hat{\mu} \pm 1.9636 \, \underbrace{SE(\hat{\mu})}_{\hat{\sigma}\sqrt{1/n+\frac{(8-\bar{x})^2}{S_{xx}}}} \qquad \hat{\sigma}^2 = SS(Res)/653.$$

$\therefore$ We get 95% CI $= (2.1665, 2.2740)$

Q: Calculate a 95% PI for $x_0 = 8$

A: $SE(Y_0 - \hat{Y}_0) = \hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(8-\bar{x})^2}{S_{xx}}} = 0.5874$

$\therefore$ 95% PI is given by $(1.0669, 3.3736)$

Remark: 95% PI is much wider than 95% CI (PI has more uncertainty).