Last lecture

Sec. 11.4 Properties on $r_{X,Y}$ $\left\langle \begin{array}{l} r_{X,Y} = r_{Y,X} \\ r_{X,Y} \in [-1,1] \end{array}\right.$

Sec 11.5 the Error in Regression $MSE = (1-r^2)\sigma_Y^2$

Sec 12.1 Simple linear regression model ← "mathematical model"

(Least square linear regression) ← { define "good" / give best soln

Today.

Sec. 12.2 Estimated slope in simple linear regression model

~~Sec 12.3~~

Next week:

Monday – Review

Wednesday – Q & A

Friday – Final, during lecture time. (almost same rule as MT)

During MT review "what to do before & during exam?"

Sec. 12.2 Distr. of estimated slope

Input → $\boxed{}$ → $\boxed{\text{signal}}$ + error noise → output

$Y_i = \boxed{\beta_0 + \beta_1 x_i} + \varepsilon_i$ $i = 1, 2, 3, \cdots, n$

minimizing this difference

$\hat{Y}_i = \boxed{\hat{\beta}_0 + \hat{\beta}_1 x_i}$

From Least squares $\left(\begin{array}{l} Y = aX + b \\ \hat{Y} = \hat{a}X + \hat{b} \end{array} \right)$ $a - \beta_1$, $b - \beta_0$

$\hat{\beta}_1$ (as $\hat{a}$ in least square)

$= \dfrac{\mathbb{E} D_X D_Y}{\sigma_X^2}$ (given by least square)

Use empirical distr. given by sample pts $\{(x_i, Y_i) : i = 1, 2, \cdots, n\}$

→ Unif $\left(\{(x_i, Y_i) : i = 1, 2, \cdots, n\}\right)$

In particular, the marginal of x is Unif$(\{x_i : i = 1, 2, \cdots, n\})$

Under the empirical distr. $\mathbb{E}X = \frac{1}{n}\sum_{i=1}^n x_i =: \bar{x}$

$\sigma_X^2 = \mathbb{E}(X - \mathbb{E}X)^2$

$= \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2$

$\mathbb{E} D_X D_Y = \mathbb{E}(X - \mathbb{E}X)(Y - \mathbb{E}Y)$

$= \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})$

Note: x_i's are constants

Y_i's are i.i.d. Normals

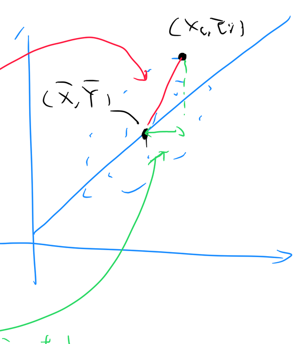
estimated slope

$\hat{\beta}_1 = \dfrac{\mathbb{E} D_X D_Y}{\sigma_X^2} = \dfrac{\cancel{\frac{1}{n}}\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\cancel{\frac{1}{n}}\sum_{i=1}^n (x_i - \bar{x})^2}$

$= \dfrac{\sum_{i=1}^n \left(\boxed{(x_i - \bar{x})^2} \cdot \dfrac{Y_i - \bar{Y}}{x_i - \bar{x}}\right)}{\sum_{i=1}^n \boxed{(x_i - \bar{x})^2}}$ slope

weights

weighted average of the slopes, with weights equal to — squared horizontal distance from $\bar{x}$



		5 pts
1 professional	×3	4
2 amateur	×1	2 3

$\dfrac{3 \times 4 + 2 + 3}{3 + 1 + 1} = \dfrac{3}{3+1+1} \times 4 + \dfrac{1}{3+1+1} \times 2 + \dfrac{1}{3+1+1} \times 3$

$n = 3$

X	Y
(1,	4)
(2,	6)
(3,	5)

$\hat{\beta}_1 = \dfrac{\sum_{i=1}^n (x_i - \bar{x}) \cdot (Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \sum_{i=1}^n \dfrac{(x_i - \bar{x})}{S_x} \cdot (Y_i - \bar{Y}) = \sum_{i=1}^n x_i^* (Y_i - \bar{Y})$

↓ linear combination of $Y_1, Y_2, \cdots, Y_n$ (indpt normals)

⇓ $\hat{\beta}_1$ is also a Normal.

define $S_x := \sum_{i=1}^n (x_i - \bar{x})^2$ } known constants

$x_i^* = \dfrac{x_i - \bar{x}}{S_x}$

① Expection of $\hat{\beta}_1$ $Y_i = \underbrace{\beta_0 + \beta_1 x_i}_{\text{Const.}} + \underbrace{\varepsilon_i}_{N(0,\sigma^2)} \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$

$\hat{\beta}_1 = \dfrac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$

$\mathbb{E} Y_i = \beta_0 + \beta_1 x_i$

$\mathbb{E}\bar{Y} = \mathbb{E}\frac{1}{n}\sum_{i=1}^n Y_i = \frac{1}{n}\sum_{i=1}^n \mathbb{E}Y_i$

$= \frac{1}{n}\sum_{i=1}^n (\beta_0 + \beta_1 x_i) = \beta_0 + \beta_1 \bar{x}$

$\mathbb{E}(Y_i - \bar{Y}) = (\beta_0 + \beta_1 x_i) - (\beta_0 + \beta_1 \bar{x}) = \beta_1 (x_i - \bar{x})$

$\mathbb{E}\hat{\beta}_1 = \mathbb{E}\dfrac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$

$= \dfrac{\sum_{i=1}^n (x_i - \bar{x})\beta_1 (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \beta_1$ – unbiased estimator

② Var of $\hat{\beta}_1$.

Recall $S_x = \sum_{i=1}^n (x_i - \bar{x})^2$ is const.

$\mathrm{Var}(\hat{\beta}_1) = \mathrm{Var}\left(\dfrac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{S_x}\right)$

$= \dfrac{1}{S_x^2} \cdot \mathrm{Var}\left(\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})\right)$

$\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})$ sum of deviations $= \sum_{i=1}^n x_i - n\bar{x}$

$= \underbrace{\sum_{i=1}^n (x_i - \bar{x})Y_i}_{\text{sum of indpt normals}} - \underbrace{\left(\sum_{i=1}^n (x_i - \bar{x})\right)}_{=0}\bar{Y}$ $= n\bar{x} - n\bar{x} = 0$

$= \dfrac{1}{S_x^2}\mathrm{Var}\left(\sum_{i=1}^n (x_i - \bar{x})Y_i\right)$

(Additivity of indpt. R.V.'s) $= \dfrac{1}{S_x^2}\sum_{i=1}^n \boxed{\mathrm{Var}((x_i - \bar{x})Y_i)}$ $= (x_i - \bar{x})^2 \mathrm{Var}(Y_i)$

$= \dfrac{1}{S_x^2}\underbrace{\left(\sum_{i=1}^n (x_i - \bar{x})^2\right)}_{S_x}\overbrace{\sigma^2}^{\mathrm{Var}(Y_i)}$

$= \dfrac{\sigma^2}{S_x}$.

Note on $\mathrm{Var}(\hat{\beta}_1)$: σ^2 unknown parameter

$S_x = \sum_{i=1}^n (x_i - \bar{x})^2$ known fixed number   will be larger when # data pts gets larger ⇒ SE will be smaller

estimate $\mathrm{Var}(\hat{\beta}_1)$: need an estimation to σ^2 – variance of error / residual

$\hat{\sigma}^2 = MSE$ is empirical variance of error / residual

$SD(\hat{\beta}_1) = \dfrac{\sigma}{\sqrt{S_x}}$

$SE(\hat{\beta}_1) = \dfrac{\hat{\sigma}}{\sqrt{S_x}}$ $(S_x = \sum_{i=1}^n (x_i - \bar{x})^2)$

When the SD of $\hat{\beta}_1$ is estimated from data, we call it the SE (standard error)

Standardized slope – statistic that helps us give CI / perform test.

$T = \dfrac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)}$

Claim (w/o. proof) $T \sim N(0,1)$ when sample size n is large

In this case, 95% CI ?

$\mathbb{P}(-2 \le T \le 2) = 95\%$

$\mathbb{P}\left(-2 \le \dfrac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} \le 2\right) = 95\%$

$\mathbb{P}(\hat{\beta}_1 - 2SE \le \beta_1 \le \hat{\beta}_1 + 2SE) = 95\%$

⇒ 95% CI for β_1 is given by $\hat{\beta}_1 \pm 2SE(\hat{\beta}_1)$.

Test: (if X & Y are linearly associated ?)

$H_0: \beta_1 = 0$

$H_a: \beta_1 \neq 0$ $(H_a: \beta_1 > 0)$ → Cannot use 95% CI

Quick way: use 95% CI we obtained

if $0 \in$ 95% CI, accept H_0

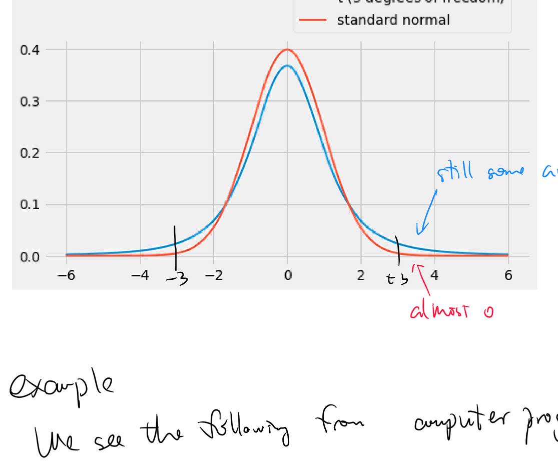
o/w. reject H_0

Test statistic: $T = \dfrac{\hat{\beta}_1 - \beta_1}{SE}$

Under H_0, and n large: $T \sim N(0,1)$

What would happen when n is not large enough ?

For sample size n $T \sim t_{n-2}$ ← t distr. with degree of freedom $n-2$



example

We see the following from computer program.

(x, y) $n = 232$

$\left\{ \begin{array}{lll} \text{slope} & = 1 & - \hat{\beta}_1 \\ \text{intercept} & = 13 & - \hat{\beta}_0 \\ r & = 0.6 & \text{correlation coefficient} \\ p & = 2 \times 10^{-24} & \text{p-value for the test } \left\langle \begin{array}{l} H_0: \beta_1 = 0 \\ H_a: \beta_1 \neq 0 \end{array}\right. \\ \text{se\_slope} & = 0.1 & SE(\hat{\beta}_1) \end{array}\right.$

1) 95% CI of $\hat{\beta}_1$ slope ± 2. se_slope $= [0.8, 1.2]$

2) SE of residual ? ($\hat{\sigma}$)

obtainable from programing

$\underset{\hat{\sigma}^2}{MSE} = (1 - r^2)\sigma_Y^2$ i.e. SD of all y_i's

$\Rightarrow \hat{\sigma} = \sqrt{1 - r^2}\,\sigma_Y = \sqrt{1 - 0.6^2} \cdot \sigma_Y$

Alt. sol'n.

$SE(\hat{\beta}_1) = \dfrac{\hat{\sigma}}{\sqrt{S_x}}$ ← want

se_slope

$\sum_{i=1}^n (x_i - \bar{x})^2 = n\sigma_x^2$ Var of all x_i's. $\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2$

$\hat{\sigma} = \text{se\_slope} \cdot \sqrt{n}\,\sigma_x$