Lecture Note 2: Means, t-tests, and Regressions

Sample average: $\hat{\mu}_X = \bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i = \frac{1}{N}(X_1 + X_2 + \cdots + X_N)$

Properties:

① <u>BLUE</u>: $\bar{X}$ is the <u>best</u> linear <u>unbiased</u> estimator

linear: $\hat{\mu}_X = a_1 X_1 + a_2 X_2 + \cdots + a_N X_N$ $\left\}\Sigma_i a_i = 1\right.$ $\Big\}$ optimal $a_i = \frac{1}{N}$

unbiased: $E[\hat{\mu}_X] = \mu_X$

best: $\min V[\hat{\mu}_X]$

② Law of large numbers: $\bar{X} \xrightarrow{p} \mu_X$ (<u>consistent</u>)

③ CLT: as $N \to \infty$, $\bar{X} \sim N\left(\mu_X, \frac{\sigma_X^2}{N}\right)$

so in "large" samples, $\frac{\sqrt{N}}{\sigma_X}(\bar{X} - \mu_X) \sim N(0,1)$

$\bar{X}$ is unbiased:

$$E[\bar{X}] = E\left[\frac{1}{N}\sum_{i=1}^{N} X_i\right] = \frac{1}{N} E[X_1 + X_2 + \cdots + X_N]$$

$$= \frac{1}{N}\sum_{i=1}^{N} E[X_i]$$

i $\boxed{\text{id}}$

$$= \frac{1}{N}\sum_{i=1}^{N} \mu_X$$

$$= \cancel{\frac{1}{N}} \cancel{N} \mu_X$$

unbiased!

Variance of $\bar{X}$:

$$V[\bar{X}] = V\left[\frac{1}{N}\sum_{i=1}^{N} X_i\right] = \frac{1}{N^2} V[X_1 + X_2 + \cdots X_N]$$

iid

$$= \frac{1}{N^2}\sum_{i=1}^{N} V[X_i]$$

iid

$$= \frac{1}{N^2}\sum_{i=1}^{N} \sigma_X^2$$

$$= \frac{1}{N^{\cancel{2}}} \cancel{N} \sigma_X^2$$

$$= \frac{\sigma_X^2}{N}$$

$$V[aX] = a^2 V[X]$$

$$V[aX + bY] = a^2 V[X] + b^2 V[Y] + 2ab\,\mathrm{cov}(X, Y)$$

Sample Variance

How to estimate $V[X] = E[(X - E[X])^2]$

How about $\hat{\sigma}_X^2 = \frac{1}{N} \sum_{i=1}^{N} (X_i - \bar{X})^2$ 1

$$E[\hat{\sigma}_X^2] = \sigma_X^2 \underbrace{\left(1 - \frac{1}{N}\right)}_{\text{bias}}$$

For unbiased, need $\frac{1}{N-1}$ instead of $\frac{1}{N}$

2 degees of freedom

Estimator 1 is biased but consistent and efficient

Estimator 2 is unbiased and consistent but inefficient

tradeoff bet. bias and variance

t-test

$$t = \frac{\hat{\theta} - \theta_0}{SE[\hat{\theta}]}$$ θ_0: null

$\Rightarrow$ CLT: as $N \to \infty$, $t \sim N(0,1)$

One sample: use $\bar{X}$ to test $\mu_x = 0$

$$V[\bar{X}] = \frac{\sigma_x^2}{N} \to SE[\bar{X}] = \frac{\sigma_x}{\sqrt{N}} \to t = \frac{\bar{X}}{\hat{\sigma}_x / \sqrt{N}}$$

Two sample: $\bar{X}_w, \bar{X}_B, \mu_w, \mu_B, \sigma_w^2, \sigma_B^2$

$$\bar{X}_w \sim N\left(\mu_w, \frac{\sigma_w^2}{N_w}\right)$$

$$\bar{X}_B \sim N\left(\mu_B, \frac{\sigma_B^2}{N_B}\right)$$

$$t = \frac{\bar{X}_w - \bar{X}_B}{SE[\bar{X}_w - \bar{X}_B]}$$

$$SE = \sqrt{V[\bar{X}_w - \bar{X}_B]} = \sqrt{V[\bar{X}_w] + V[\bar{X}_B] - 2\operatorname{cov}(\bar{X}_w, \bar{X}_B)}$$

$$= \frac{\bar{X}_w - \bar{X}_B}{\sqrt{\frac{\sigma_w^2}{N_w} + \frac{\sigma_B^2}{N_B}}} = \frac{\bar{X}_w - \bar{X}_B}{\sqrt{SE_w^2 + SE_B^2}}$$

OLS Estimator

$X, Y, \quad \hat{Y} = b_0 + b_1 X$ "best fit"

Min mean sq error → error $\quad U = Y - \hat{Y} = Y - b_0 - b_1 X$

$$\min_{b_0, b_1} E[(Y - b_0 - b_1 X)^2]$$ population

$$\min_{\hat{b}_0, \hat{b}_1} \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{b}_0 - \hat{b}_1 X_i)^2$$ sample

In sample, $U_i = Y_i - \hat{Y}_i$ is called residual

Optima: $\beta_1 = \frac{cov(X,Y)}{V[X]}$, $\hat{\beta}_1 = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_i (X_i - \bar{X})^2}$

Can view $\hat{\beta}_1$ as an estimator for β_1 in the model:

$$Y_i = \beta_0 + \beta_1 X_i + U_i$$

$$\hat{\beta}_1 \xrightarrow{P} \beta_1$$

Gauss-Markov Thm

- fixed (non-random) X_i
- Assumptions about $Y_i = \beta_0 + \beta_1 X_i + U_i$

    1. $E[U_i] = 0$ for all i
    2. $V(U_i) = \sigma^2$ for all i
    3. $cov(U_i, U_j) = 0$ for $i \neq j$

Theorem: OLS is BLUE.