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

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

Properties:

① BLUE: $\bar{X}$ is the best linear unbiased estimator

linear: $\hat{\mu}_x = a_1 X_1 + a_2 X_2 + \cdots + a_N X_N$

unbiased: $E[\hat{\mu}_x] = \mu_x$

best: min. variance (efficient)

$a_i = \frac{1}{N}$

② Law of large numbers: $\bar{X} \not\to \mu_x$ (consistent)

③ Central limit th$^{\underline{m}}$: as $N \to \infty$, $\bar{X} \to \mathcal{N}(\mu_x, \frac{1}{N}\sigma_x^2)$

$\frac{\sqrt{N}}{\sigma_x}(\bar{X} - \mu_x) \sim \mathcal{N}(0,1)$ in large samples

Sample average is <u>unbiased</u>:

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

i.i.d

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

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

Variance of $\bar{X}$:

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

i.i.d

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

i.i.d

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

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

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

Sample Variance

→ Estimate $V[X]$

→ Can we use $\hat{\sigma}_X^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2$? ①

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

→ For unbiased estimator, need $\frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2$ ②

degrees of freedom

→ Estimator ① is biased but efficient and consistent

→ Estimator ② is unbiased and consistent <u>but</u> not efficient

tradeoff between bias and variance

t - Tests

estimator, null

$$\rightarrow t = \frac{\hat{\theta} - \theta_0}{SE[\hat{\theta}]}$$

std error of $\hat{\theta}$

$\leftarrow$ CLT: as $N \to \infty$, $t \sim \mathcal{N}(0,1)$

$\rightarrow p = 2(1 - \Phi[t])$ for a 2-sided test of $\hat{\theta} = \theta_0$

One sample test

$\rightarrow$ test that $\bar{X} = 0$

$$\rightarrow t = \frac{\bar{X}}{\sigma_X / \sqrt{N}}$$

Two sample test

$\rightarrow \mu_W, \mu_B, \sigma_W^2, \sigma_B^2$

$\rightarrow$ CLT: as $N \to \infty$, $\bar{X}_W \sim \mathcal{N}(\mu_W, \sigma_W^2/N_W)$, $\bar{X}_B \sim \mathcal{N}(\mu_B, \sigma_B^2/N_B)$

$$\rightarrow \bar{X}_W - \bar{X}_B \sim \mathcal{N}\left(\mu_W - \mu_B, \frac{\sigma_W^2}{N_W} + \frac{\sigma_B^2}{N_B}\right)$$

$$\rightarrow SE[\bar{X}_W - \bar{X}_B] = \sqrt{\frac{\hat{\sigma}_W^2}{N_W} + \frac{\hat{\sigma}_B^2}{N_B}}$$

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

$$CI: \bar{X}_W - \bar{X}_B \pm 1.96\, SE[\cdot]$$

$$V[\bar{X}_W - \bar{X}_B] = V[\bar{X}_W] + V[\bar{X}_B] - 2\,\cancel{cov(\bar{X}_W, \bar{X}_B)} = \frac{\sigma_W^2}{N_W} + \frac{\sigma_B^2}{N_B}$$

b/c independent

OLS Estimator

→ two variables, Y and X

→ $\hat{Y} = b_0 + b_1 X$ "best fit"

→ minimizes mean squared error $U = Y - \hat{Y} = Y - b_0 - b_1 X$

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

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

or mean squared residual $\hat{U}_i = Y_i - \hat{Y}_i$

→ solutions: $\beta_0, \beta_1, \hat{\beta}_0, \hat{\beta}_1$

→ model: $Y_i = \beta_0 + \beta_1 X_i + U_i$

→ $$\beta_1 = \frac{\mathrm{cov}(X, Y)}{V[X]}$$

$$\hat{\beta}_1 = \frac{\frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

↑

$\hat{\beta}_1 \not\to \beta_1$

Gauss Markov Theorem

- Fixed X_i (non-random)
- Three 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 (homoskedasticity)
  3. $cov(U_i, U_j) = 0$ for $i \neq j$ (zero covariance)
- Theorem: Under assumptions 1-3, OLS is BLUE.

Equivalencies

1. Estimating means
2. t-tests (1- and 2-sample)
3. Regression

$Y_i = \beta_0 + U_i \Rightarrow \hat{\beta}_0 = \bar{Y}_i$ both unbiased & consistent for μ_Y

$Y_i = \beta_0 + \beta_1 BLACK_i + U_i \Rightarrow$

$$\hat{\beta}_0 = \bar{Y}_W$$
$$\hat{\beta}_0 + \hat{\beta}_1 = \bar{Y}_B$$
$$\hat{\beta}_1 = \bar{Y}_W - \bar{Y}_B$$