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 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: efficient $\rightarrow$ min $V[\hat{\mu}_x]$

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

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

③ CLT: central limit th$^{\underline{m}}$

as $N \to \infty$, $\bar{X} \sim \mathcal{N}\left(\mu_x, \frac{\sigma_x^2}{N}\right)$

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

Sample average is unbiased

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

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

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

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

unbiased!

Variance of $\bar{X}$:

$$V[\bar{X}] = V\left[\frac{1}{N}\sum_{i=1}^{N} X_i\right] = \frac{1}{N^2} V[X_1 + X_2 + X_3 + \cdots + X_N] \quad \text{iid}$$

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

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

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

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

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

Sample Variance

How to estimate $V[X] = E[X - E[X]] = E[X - \mu_x]$?

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

$E[\hat{\sigma}_x^2] = \sigma_x^2 (1 - \frac{1}{N})$

bias

For unbiased estimator, need $\frac{1}{N-1}$ instead of $\frac{1}{N}$: $\hat{\sigma}_x^2 = \frac{1}{N-1} \sum_{i=1}^{N} (X_i - \bar{X})^2$ ②

degrees of freedom

Estimator ① is biased but consistent and efficient

Estimator ② is unbiased and consistent <u>but</u> inefficient

tradeoff bet. bias and variance

t-Tests

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

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

in large samples, can form $p = 2(1 - \Phi[t])$

for 2-sided test of null $\theta = \theta_0$

$\Rightarrow$ CI $\hat{\theta} \pm 1.96\, SE[\hat{\theta}]$

One sample

$\bar{X}$ to test null that $\mu_x = 0$

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

Two sample

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

$\bar{X}_w \sim N(\mu_w, \sigma_w^2/N_w)$, $\bar{X}_B \sim N(\mu_B, \sigma_B^2/N_B)$

$\bar{X}_w - \bar{X}_B \sim N\left(\mu_w - \mu_B, \frac{\sigma_w^2}{N_w} + \frac{\sigma_B^2}{N_B}\right)$

$V[\bar{X}_w - \bar{X}_B] = V[\bar{X}_w] + V[\bar{X}_B] + \cancel{2\,\mathrm{cov}} =$

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

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

OLS Estimator

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

Minimize mean squared error $\rightarrow$ error $U = Y - \hat{Y} = Y - b_0 - b_1 X$

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

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

In sample $U_i = Y_i - \hat{Y}_i = Y_i - \hat{b}_0 - \hat{b}_1 X_i$ is called the residual

Optimal: $\beta_1 = \dfrac{\mathrm{cov}(X, Y)}{V[X]}$, $\hat{\beta}_1 = \dfrac{\frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2}$

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

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

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

Gauss - Markov Theorem

- fixed (non-random) X_i
- 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. $\text{cov}(U_i, U_j) = 0$ for $i \neq j$

Theorem: under 1-3, OLS is BLUE.

Equivalencies

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

$Y_i = \beta_0 + U_i \rightarrow \hat{\beta}_0 = \bar{Y}$ consistent & unbiased estimators of μ_Y

$Y_i = \beta_0 + \beta_1 \text{BLACK}_i + U_i \rightarrow$

$\hat{\beta}_0 = \bar{Y}_W$

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

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