Linear Regression

Recap:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \dots, n$$

Using the LS principle, we tried to find estimates $(\hat{\beta}_0, \hat{\beta}_1)$ of the true parameters (β_0, β_1).

Specifically the LS estimators are:

$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{\beta_0, \beta_1}{\operatorname{argmin}} \, Q(\beta_0, \beta_1)$$

$$= \underset{\beta_0, \beta_1}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

We found a closed form:

(random) $$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \overset{\text{HW}}{=} \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

(random) $$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$ (b/c $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$ $\hookrightarrow$ $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$)

Properties of LS Estimators:

Are these ests any good?

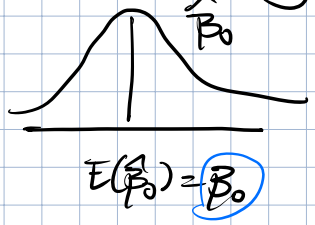
Gauss – Markov Theorem:

Under the SLR model assumptions, the LS estimates $\hat{\beta}_0$ & $\hat{\beta}_1$ are:

i. unbiased for β_0 & β_1 respectively,

i.e. $E(\hat{\beta}_0) = \beta_0$

$E(\hat{\beta}_1) = \beta_1$



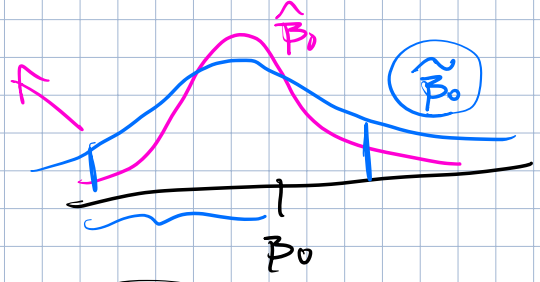
ii. the LS estimators are BLUE:

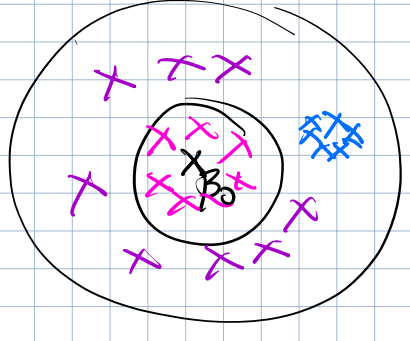
"best linear unbiased estimators"

i.e. ← some other linear unbiased est

$Var(\hat{\beta}_0) \leq Var(\tilde{\beta}_0)$

$Var(\hat{\beta}_1) \leq Var(\tilde{\beta}_1)$





How do I know this?

i. How can I show $E(\hat{\beta}_1) = \beta_1$?

WTS:

$$E(\hat{\beta}_1) = E\left(\frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} \right) = \beta_1$$

Def: $SSX = \sum_{i=1}^{n} (x_i - \bar{x})^2$ (denom)

$SSXY = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$

Lemma:

claim: $SSXY = \sum_{i=1}^{n} (x_i - \bar{x}) y_i$

Why?

$$SSXY = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} \left[(x_i - \bar{x}) y_i - (x_i - \bar{x}) \bar{y} \right]$$

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

↳ WTS $= 0$

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

$$= \bar{y} \left(\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x} \right)$$

$$= \bar{y} \left(n\bar{x} - n\bar{x} \right)$$

$$= \bar{y}(0) = 0$$

$\sum_{i=1}^{n} a c_i = a \sum_{i=1}^{n} c_i$

$\bar{x} = \frac{\sum_i x_i}{n}$

$n\bar{x} = \sum_i x_i$

$$\hat{\beta}_1 = \frac{SSXY}{SSX} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})y_i}{SSX}$$

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

$$= \sum_{i=1}^{n}\underbrace{\frac{(x_i-\bar{x})}{SSX}}_{=k_i} y_i = \sum_{i=1}^{n} k_i y_i \quad \text{(linearity)}$$

$$E(\hat{\beta}_1) = E\left(\sum_{i=1}^{n} k_i y_i\right) = \sum_{i=1}^{n} E(k_i y_i)$$

$$= \sum_{i=1}^{n} k_i E(y_i)$$

$$= \sum_{i=1}^{n} k_i E(\beta_0 + \beta_1 x_i + \varepsilon_i)$$

$$= \sum_{i=1}^{n} k_i \left[E(\beta_0) + E(\beta_1 x_i) + E(\varepsilon_i)\right]$$

$$= \sum_{i=1}^{n} k_i \left[\beta_0 + \beta_1 x_i + 0\right]$$

$$= \sum_{i=1}^{n} k_i \beta_0 + \sum_{i=1}^{n} \beta_1 k_i x_i$$

$$= \beta_0\left(\sum_{i=1}^{n} k_i\right) + \beta_1\left(\sum_{i=1}^{n} k_i x_i\right) = \beta_1$$

$\sum k_i = 0$ (A)

$\sum k_i x_i = 1$ (B)

(A) $\sum_{i=1}^{n} k_i = \sum_{i=1}^{n} \frac{x_i - \bar{x}}{SSX} = \frac{1}{SSX} \sum_{i=1}^{n} (x_i - \bar{x}) = \frac{1}{SSX}(0) = 0$ ✓

(B) $\sum_{i=1}^{n} k_i x_i = \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{SSX}\right) x_i$

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

$= SSX$

$\frac{SSX}{SSX} = 1.$

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

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

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

$= \bar{x} \left[\sum_{i=1}^{n} (x_i - \bar{x}) \right]$

$= \bar{x} (\Sigma_i x_i - \Sigma_i \bar{x})$

$= \bar{x} (n\bar{x} - n\bar{x})$

$= \bar{x} (n\bar{x} - n\bar{x}) = 0$

That's the slope unbiasedness.

Y'all try the intercept.

Part 2 of G-M Thm:

How can I show that

$$Var(\hat{\beta}_1) \leq Var(\tilde{\beta}_1)$$

for some other linear unbiased est $\tilde{\beta}_1$?

Sketch:

$$Var(\hat{\beta}_1) = Var\left(\sum_{i=1}^{n} k_i y_i\right)$$

by covariance assumption.

$$= \sum_{i=1}^{n} Var(k_i y_i)$$

$$= \sum_{i=1}^{n} k_i^2 Var(y_i)$$

$$= \sum_{i=1}^{n} k_i^2 \sigma^2$$

$$= \sigma^2 \sum_{i=1}^{n} k_i^2 = \frac{\sigma^2}{SSX}$$

Aside

$Var(W+Z) = Var(W) + Var(Z) + 2Cov(W,Z)$

If $Cov(W,Z) = 0 \Rightarrow$

$Var(W+Z) = Var(W) + Var(Z)$

What is $\sum_i k_i^2$?

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

$$= \frac{SSX}{(SSX)^2} = \frac{1}{SSX}$$

I want to show $\mathrm{Var}(\tilde{\beta}_1) \geq \sigma^2/SSX = \sigma^2 \sum_{i=1}^{n} k_i^2$

Write

$$\tilde{\beta}_1 = \sum_{i=1}^{n} \tilde{k}_i y_i = \sum_{i=1}^{n} (k_i + d_i) y_i$$

where

$d_i = \tilde{k}_i - k_i$

for at least one

$i = 1, \ldots, n$

$$\mathrm{Var}(\tilde{\beta}_1) = \mathrm{Var}\left(\sum_{i=1}^{n} (k_i + d_i) y_i\right)$$

$$= \sum_{i=1}^{n} \mathrm{Var}((k_i + d_i) y_i)$$

$$= \sum_{i=1}^{n} (k_i + d_i)^2 \mathrm{Var}(y_i)$$

$$= \sigma^2 \sum_{i=1}^{n} (k_i + d_i)^2$$

$$= \sigma^2 \left[\sum_{i=1}^{n} (k_i^2 + 2k_i d_i + d_i^2)\right]$$

$$= \sigma^2 \left[\Sigma_i k_i^2 + \underbrace{2\Sigma_i k_i d_i}_{\text{(A)}} + \underbrace{\Sigma_i d_i^2}_{\text{(B)}}\right]$$

$\Sigma_i d_i^2 > 0$

(A) $\Sigma_i k_i d_i = 0$ (try to show)

(B) $\Sigma_i d_i^2 > 0$ (easy to see)

not all $d_i = 0 \Rightarrow$ one $d_i^2 > 0 \Rightarrow \Sigma_i d_i^2 > 0$

Sampling Distribution of $\hat{\beta}_0$ & $\hat{\beta}_1$.

Slope: $E(\hat{\beta}_1) = \beta_1$

$Var(\hat{\beta}_1) = \sigma^2 / SSX$] under (i–iii)

If I assume (iv) $\Leftrightarrow$ normality of errors:

$\hat{\beta}_1 \sim N(\beta_1, \sigma^2/SSX)$ with (iv)

$\overset{?}{\sim} (\beta_1, \sigma^2/SSX)$ without (iv).

Intercept:

$E(\hat{\beta}_0) = \beta_0$

$Var(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{SSX} \right)$ (see Cody notes/ textbook)

Under Normality assumption:

$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{SSX}\right)\right)$

(or) $\overset{?}{\sim} \left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{SSX}\right)\right)$ (without iv.)

Consider the fitted values:

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

What is the distribution of $\hat{y}_i$?

(under Normality is ok)

$$\hat{y}_i \sim N(\quad , \quad)$$