Linear Regression

8/25/25

Recap:

$$y_i = \underbrace{\beta_0 + \beta_1 x_i}_{\text{signal}} + \underbrace{\varepsilon_i}_{\text{noise}} \qquad i = 1, \ldots, n$$

Using the LS principle, we tried to find estimates $(\hat{\beta}_0, \hat{\beta}_1)$ of the 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 the closed form:

random $\searrow$

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

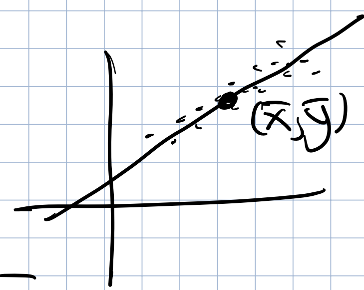
random $\searrow$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$\uparrow$ came from

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$$

$$\hookrightarrow \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$



Properties of LS Ests:

Are these estimators any good?

Gauss-Markov Theorem:

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

i. unbiased for β_0 & β_1 respectively,

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

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

&

ii. the LS estimators are BLUE

= "best linear unbiased estimators" //

Focus: on $\hat{\beta}_1$ for now

linear in y_i

(linear) $$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \underset{\text{claim}}{=} \sum_{i=1}^{n} k_i y_i$$

(best) $$Var(\hat{\beta}_1) \leq Var(\tilde{\beta}_1)$$ where $\tilde{\beta}_1$ is some other linear unbiased estimator.

How do I know this then is true?

Still focus on $\hat{\beta}_1$ for now:

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

Want to show

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

Def:

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

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

Lemma:

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

why?

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

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

want to show $= 0$.

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

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

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

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

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

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

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

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

$$= \sum_{i=1}^{n} \underbrace{\left[\frac{(x_i - \bar{x})}{SSX}\right]}_{= k_i} y_i = \sum_{i=1}^{n} k_i y_i$$

↑ linearity in y_i

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

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

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

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

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

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

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

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

$$= \beta_0 \underbrace{\left(\sum_{i=1}^{n} k_i \right)}_{\substack{=0 \\ Ⓐ}} + \beta_1 \underbrace{\left(\sum_{i=1}^{n} k_i x_i \right)}_{\substack{=1 \\ Ⓑ}}$$

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

$$= \frac{1}{SSX} (0) = 0$$

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

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

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

$\sum x_i^2 - 2\bar{x} x_i + \bar{x}^2$

$\sum x_i^2 - 2\bar{x} \sum x_i + n\bar{x}^2$

$\sum x_i^2 - 2n\bar{x}^2 + n\bar{x}^2$

$\boxed{\sum x_i^2 - n\bar{x}^2}$

$\sum x_i^2 - x_i \bar{x}$

$\sum x_i^2 - \bar{x} \sum x_i$

$\boxed{\sum x_i^2 - n\bar{x}^2}$

$$= \frac{SSx}{SSx} = \frac{\sum_i x_i^2 - n\bar{x}^2}{\sum_i x_i^2 - n\bar{x}^2} = 1$$

$\Rightarrow E(\hat{\beta}_1) = \beta_0 (0) + \beta_1 (1) = \beta_1$ $\square$

Your turn:

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

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

but

$\sum_{i=1}^{n} x_i^2 \neq n\bar{x}^2$

Part 2 of GM Thm:

Goal: If I consider some other linear unbiased estimator $\tilde{\beta}_1$ then

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

How can I show this?

Sketch:

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

by covariance assumption

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

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

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

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

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

$$\left. = \frac{1}{(SSX)^2}(SSX) = \frac{1}{SSX} \right]$$

Consider other linear unbiased estimators $\tilde{\beta}_1$:

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

where
$\tilde{k}_i = k_i + d_i$
$\Leftrightarrow d_i = \tilde{k}_i - k_i$

$d_i \neq 0 \quad \forall i$

$\forall$ = "for all"

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

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

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

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

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

if ≥ 0 then we're done

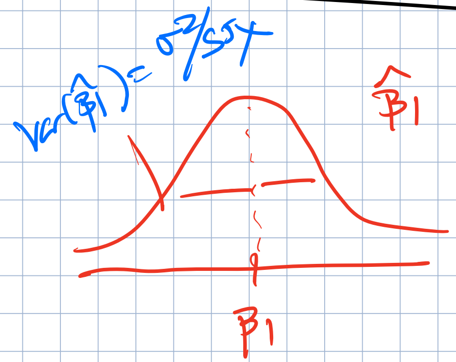
$$= \sigma^2 \left[\underline{\Sigma_i k_i^2} + 2 \overbrace{\underbrace{\boxed{\Sigma_i d_i k_i}}_{\text{Ⓐ} \; = 0} + \underbrace{\Sigma_i d_i^2}_{\text{Ⓑ} \; > 0}} \right)$$

Ⓐ $\boxed{\sum_i d_i k_i = 0}$ try it. (Exercise)

Ⓑ $\sum_i d_i^2 > 0$ (sum of qtys that are 0 or +)

[repeat for $\hat{\beta}_0$.]

Slope:



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

$$Var(\hat{\beta}_1) = \sigma^2 \sum_{i=1}^{n} k_i^2 = \sigma^2/SSX$$

If (iv):

$$\hat{\beta}_1 \sim N(\beta_1, \sigma^2/SSX) \quad \text{(classical)}$$

$$? \; (\beta_1, \sigma^2/SSX) \quad \text{(i - iii. only)}$$

Int:

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

$$Var(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{SSX} \right)$$

see notes / textbook

If (iv), we know

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

$$? \; \left(\beta_0, \sigma^2 \left[\frac{1}{n} + \frac{\bar{X}^2}{SSX}\right]\right) \quad \text{(i - iii. only)}$$

Fitted Values:

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

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

$$\hat{y}_i \sim N\Big(\underset{?}{\underline{\qquad}}, \underset{?}{\underline{\qquad}}\Big)$$

Think about it for next time!