Lecture 2 - 19 Oct 2023

- **DESCRIPTIVE PROPERTIES** of the estimated linear model

1) the mean of the response at the observed locations $(x_1, \dots, x_n)$ is equal to the mean of the predicted values at those locations

$$\frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n} \hat{y}_i \qquad (\bar{y} = \bar{\hat{y}})$$

$$\bar{\hat{y}} = \frac{1}{n}\sum_{i=1}^{n}\hat{y}_i = \frac{1}{n}\sum_{i=1}^{n}(\hat{\beta}_1 + \hat{\beta}_2 x_i) = \frac{1}{\cancel{n}}\cancel{n}\hat{\beta}_1 + \hat{\beta}_2\bar{x} = \hat{\beta}_1 + \hat{\beta}_2\bar{x}$$

$$= \underbrace{\bar{y} - \hat{\beta}_2\bar{x}}_{\hat{\beta}_1} + \hat{\beta}_2\bar{x} = \bar{y}$$

2) the estimated regression line passes for the point $(\bar{x}, \bar{y})$

i.e. $\bar{y} = \hat{\beta}_1 + \hat{\beta}_2\bar{x}$

compute $\hat{y}$ at $\bar{x}$: $\quad \hat{y} = \hat{\beta}_1 + \hat{\beta}_2\bar{x} = \bar{y} - \cancel{\hat{\beta}_2\bar{x}} + \cancel{\hat{\beta}_2\bar{x}} = \bar{y}$

3) the sample mean of the residuals is equal to zero

i.e. $\bar{e} = \frac{1}{n}\sum_{i=1}^{n} e_i = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) = 0$

$$\bar{e} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) = \bar{y} - \hat{\beta}_1 - \hat{\beta}_2\bar{x} = \bar{y} - \bar{y} + \hat{\beta}_2\bar{x} - \hat{\beta}_2\bar{x} = 0$$

- **INFERENTIAL PROPERTIES** of the estimated linear model

(now we need those assumptions on the errors)

$\hat{\beta}_1$ and $\hat{\beta}_2$ are the estimates (they are not random variables, they are numbers).

we study the properties of the corresponding ESTIMATORS $\hat{\beta}_1(\underline{Y})$, $\hat{\beta}_2(\underline{Y})$

(the random variable is $\underline{Y} = (Y_1, \dots, Y_n)$, and the estimators $\hat{\beta}_1(\underline{Y})$, $\hat{\beta}_2(\underline{Y})$ are a transformation of $\underline{Y}$.)

→ in general, I will omit the dependence from $\underline{Y}$ when it is clear from the context if we are talking about the estimate or the estimator ←

$$\hat{\beta}_1(\underline{Y}) = \bar{Y} - \hat{\beta}_2(\underline{Y})\bar{x}$$

$$\hat{\beta}_2(\underline{Y}) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

all these are ESTIMATORS

we can compute $\mathbb{E}[\hat{\beta}_1]$, $\mathbb{E}[\hat{\beta}_2]$, $var(\hat{\beta}_1)$, $var(\hat{\beta}_2)$

MEAN and VARIANCE of the estimators $\hat{\beta}_1$ and $\hat{\beta}_2$

- Let's start with $\hat{\beta}_2(\underline{Y})$

$$\hat{\beta}_2 = \frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\left\{\sum_{i=1}^{n}(x_i - \bar{x})Y_i - \cancel{\bar{Y}\underbrace{\sum_{i=1}^{n}(x_i - \bar{x})}_{=0}}\right\}$$

$=0$ since $\sum_{i=1}^{n}(x_i - \bar{x}) = \sum_{i=1}^{n} x_i - n\bar{x} = n\bar{x} - n\bar{x} = 0$

$$= \sum_{i=1}^{n}\underbrace{\frac{(x_i - \bar{x})}{\left[\sum_{h=1}^{n}(x_h - \bar{x})^2\right]}}_{w_i} \cdot Y_i$$

$$= \sum_{i=1}^{n} w_i \cdot Y_i \qquad \text{is a linear combination of } Y_1, \dots, Y_n$$

$$\mathbb{E}[\hat{\beta}_2] = \mathbb{E}\left[\sum_{i=1}^{n} w_i Y_i\right] = \sum_{i=1}^{n} w_i\,\mathbb{E}[Y_i] = \sum_{i=1}^{n} w_i(\beta_1 + \beta_2 x_i) =$$

$$= \beta_1\underbrace{\sum_{i=1}^{n} w_i}_{A} + \beta_2\underbrace{\sum_{i=1}^{n} w_i x_i}_{B} = \beta_2$$

$$A = \sum_{i=1}^{n} w_i = \sum_{i=1}^{n}\frac{x_i - \bar{x}}{\sum_{h=0}^{n}(x_h - \bar{x})^2} = 0$$

$$B = \sum_{i=1}^{n} w_i x_i = \sum_{i=1}^{n}\frac{x_i(x_i - \bar{x})}{\sum_{h=0}^{n}(x_h - \bar{x})^2} = \frac{\sum_{i=1}^{n} x_i^2 - \bar{x}\sum_{i=1}^{n} x_i}{\sum_{h=0}^{n}(x_h - \bar{x})^2} = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{\sum_{h=0}^{n}(x_h - \bar{x})^2} = 1$$

$$var(\hat{\beta}_2) = var\left(\sum_{i=1}^{n} w_i Y_i\right) \overset{\text{since } cov(Y_i, Y_k) = 0}{=} \sum_{i=1}^{n} var(w_i Y_i) = \sum_{i=1}^{n} w_i^2\,var(Y_i) = \sum_{i=1}^{n} w_i^2 \cdot \sigma^2$$

$$= \sigma^2\frac{\cancel{\sum_{i=1}^{n}(x_i - \bar{x})^2}}{\left[\sum_{h=1}^{n}(x_i - \bar{x})^2\right]^{\cancel{2}}} = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- Turning now to $\hat{\beta}_1(\underline{Y})$

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2\bar{x} = \frac{1}{n}\sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} w_i Y_i \cdot \bar{x} = \sum_{i=1}^{n} Y_i\underbrace{\left(\frac{1}{n} - w_i\bar{x}\right)}_{v_i} = \sum_{i=1}^{n} v_i Y_i$$

linear combination

$$\mathbb{E}[\hat{\beta}_1] = \mathbb{E}\left[\sum_{i=1}^{n} v_i Y_i\right] = \sum_{i=1}^{n} v_i\,\mathbb{E}[Y_i] = \sum_{i=1}^{n} v_i(\beta_1 + \beta_2 x_i) = \beta_1\underbrace{\sum_{i=1}^{n} v_i}_{C} + \beta_2\underbrace{\sum_{i=1}^{n} v_i x_i}_{D} = \beta_1$$

$$C = \sum_{i=1}^{n} v_i = \sum_{i=1}^{n}\left(\frac{1}{n} - w_i\bar{x}\right) = 1 - \bar{x}\cancel{\sum_{i=1}^{n} w_i} = 1$$

$$D = \sum_{i=1}^{n} v_i x_i = \sum_{i=1}^{n}\left(\frac{1}{n} - w_i\bar{x}\right)x_i = \bar{x} - \bar{x}\sum_{i=1}^{n} w_i x_i = 0$$

$$var(\hat{\beta}_1) = var\left(\sum_{i=1}^{n} v_i Y_i\right) = \sum_{i=1}^{n} var(v_i Y_i) = \sum_{i=1}^{n} v_i^2\,var(Y_i) = \sum_{i=1}^{n} v_i^2\sigma^2 =$$

$$= \sigma^2 \cdot \sum_{i=1}^{n}\left(\frac{1}{n} - w_i\bar{x}\right)^2 = \sigma^2\sum_{i=1}^{n}\left(\frac{1}{n^2} + w_i^2\bar{x}^2 - \frac{2}{n}\bar{x}w_i\right) =$$

$$= \sigma^2\left(\frac{1}{n} + \bar{x}^2\sum_{i=1}^{n} w_i^2 - \cancel{\frac{2}{n}\bar{x}\sum_{i=1}^{n} w_i}\right) =$$

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

REMARKS

- $\hat{\beta}_1$ and $\hat{\beta}_2$ are UNBIASED estimators (i.e. $\mathbb{E}[\hat{\beta}_1] = \beta_1$; $\mathbb{E}[\hat{\beta}_2] = \beta_2$)
- if σ^2 increases ⇒ $var(\hat{\beta}_1)$ and $var(\hat{\beta}_2)$ increase (fixing the rest)
- if $\Sigma(x_i - \bar{x})^2$ increases ⇒ $var(\hat{\beta}_1)$ and $var(\hat{\beta}_2)$ decrease

  ↓

  I can estimate the line better if the x_i's are well spread out
- $var(\hat{\beta}_1)$ and $var(\hat{\beta}_2)$ depend on σ^2 (unknown) → can we estimate it? ⊛

ESTIMATE of σ^2

Recall that σ^2 is the (common) variance of the errors $\varepsilon_1, \dots, \varepsilon_n$. However, these are not observable quantities.

The corresponding sample quantities (observable) are the RESIDUALS $e_i = y_i - \hat{y}_i$, $i = 1, \dots, n$.

(note: they are not an estimate of the errors)

Idea to estimate σ^2: we estimate it using the sample variance of the residuals, i.e

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(e_i - \bar{e})^2$$

we have seen that $\bar{e} = 0 \Rightarrow \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2$

We can consider the corresponding estimator $\hat{\sigma}^2(\underline{Y})$ to study its properties

$$\hat{\sigma}^2(\underline{Y}) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i)^2$$

It can be shown that $\mathbb{E}[\hat{\sigma}^2] = \frac{n-2}{n}\sigma^2$ it is a BIASED estimator of σ^2

- if n is large, the bias is small

  indeed, $\lim_{n \to +\infty}\mathbb{E}[\hat{\sigma}^2] = \sigma^2$ asymptotically unbiased
- we can define an unbiased version $S^2 = \frac{n}{n-2}\hat{\sigma}(Y)^2 = \frac{n}{n-2} \cdot \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i)^2$

$$= \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i)^2$$

⊛ once we compute the estimate of σ^2, s^2,

we can plug it into $var(\hat{\beta}_1)$ and $var(\hat{\beta}_2)$ to obtain an estimate of these quantities

$$\widehat{var}(\hat{\beta}_1) = s^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$$

$$\widehat{var}(\hat{\beta}_2) = \frac{s^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$