L2. Correction: $Var(e_i) = \sigma^2 [1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{S_{xx}}]$

- For a given data $(x_1, y_1), \cdots, (x_n, y_n)$

Fit a SLR:

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

where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the OLSE based on $(x_1, y_1), \cdots, (x_n, y_n)$

The residual e_i is the estimation bias of $\boxed{\text{in sample}}$ y_i

$$e_i = y_i - \hat{y}_i$$

- $$Var(e_i) = Var(y_i - \hat{y}_i) = Var[(\beta_0 + \beta_1 x_i + \varepsilon_i) - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]$$

$$= Var[(\beta_0 + \beta_1 x_i) + \varepsilon_i - (\bar{y} - \hat{\beta}_1 \bar{x}) - \hat{\beta}_1 x_i)]$$

$$= Var(\varepsilon_i) + Var(\bar{y}) + (x_i - \bar{x})^2 Var(\hat{\beta}_1)$$

$$- 2cov(\varepsilon_i, \bar{y})$$

$$- 2(x_i - \bar{x}) cov(\varepsilon_i, \hat{\beta}_1)$$

$$+ 2(x_i - \bar{x}) cov(\bar{y}, \hat{\beta}_1)$$

- $\mathrm{cov}(\bar{Y}, \hat{\beta}_1) = \mathrm{cov}\left(\sum \frac{Y_i}{n}, \sum k_j y_j\right) = \sum_i \sum_j \mathrm{cov}\left(\frac{y_i}{n}, k_j y_j\right)$

$$= \sum_i \sum_j \frac{k_j}{n} \mathrm{cov}(y_i, y_j)$$

$$\boxed{\begin{array}{l} \text{Because } \mathrm{cov}(y_i, y_j) = 0 \text{ for } i \neq j \\ \qquad \mathrm{cov}(y_j, y_j) = \mathrm{var}(y_i) = \sigma^2 \end{array}}$$

$$= \sum_j \frac{k_j}{n} \sigma^2 = \frac{\sigma^2}{n} \underbrace{\sum k_j}_{= 0} = 0$$

- $\mathrm{cov}(\bar{Y}, \varepsilon_i) = \mathrm{cov}\left(\sum_{j=1}^{n} \frac{y_j}{n}, \varepsilon_i\right)$

$$= \sum_{j=1}^{n} \frac{1}{n} \mathrm{cov}(y_j, \varepsilon_i)$$

$$\boxed{\begin{array}{l} \text{when } i \neq j, \quad y_j = \beta_0 + \beta_1 x_j + \varepsilon_j, \quad \mathrm{cov}(y_j, \varepsilon_i) = 0 \\ \text{when } i = j, \quad \mathrm{cov}(y_j, \varepsilon_i) = \mathrm{cov}(\beta_0 + \beta_1 x_i + \varepsilon_i, \varepsilon_i) = \sigma^2 \end{array}}$$

$$= \frac{1}{n} \mathrm{cov}(y_i, \varepsilon_i) = \frac{\sigma^2}{n}$$

- $cov(\varepsilon_i, \hat{\beta}_1)$ similarly

$$= cov(\varepsilon_i, \sum_j k_j y_j)$$

$$= k_i \, cov(\varepsilon_i, y_i)$$

$$= k_i \sigma^2 = \frac{(x_i - \bar{x})}{\sum (x_j - \bar{x})^2} \sigma^2 = \frac{(x_i - \bar{x})}{S_{xx}} \sigma^2$$

so $Var(e_i) = Var(\bar{Y}) + (x_i - \bar{x})^2 Var(\hat{\beta}_1) + Var(\varepsilon_i)$

$$- 2(x_i - \bar{x}) cov(\varepsilon_i, \hat{\beta}_1) - 2 cov(\bar{Y}, \varepsilon_i)$$

$$= \frac{\sigma^2}{n} + (x_i - \bar{x})^2 \cdot \frac{\sigma^2}{S_{xx}} + \sigma^2$$

$$- \frac{2(x_i - \bar{x})^2}{S_{xx}} \sigma^2 - 2\frac{\sigma^2}{n}$$

$$= \sigma^2 - \frac{\sigma^2}{n} - \frac{(x_i - \bar{x})^2}{S_{xx}} \sigma^2$$

$$\boxed{= \sigma^2 \left[1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{S_{xx}} \right]}$$

In L3, when giving a new X_0, and look at the $\boxed{\text{out of sample}}$ prediction bias:

$$e_0 = y_0 - \hat{y}_0$$

$$\begin{aligned} Var(e_0) &= Var(y_0 - \hat{y}_0) \\ &= Var(\varepsilon_0) + Var(\bar{Y}) + (X_0 - \bar{X}) Var(\hat{\beta}_1) \\ &\quad - 2\, cov(\varepsilon_0, \bar{Y}) \\ &\quad - 2\, (X_0 - \bar{X})\, cov(\varepsilon_0, \hat{\beta}_1) \\ &\quad + 2\, (X_i - \bar{X})\, cov(\bar{Y}, \hat{\beta}_1) \end{aligned}$$

since ε_0 is out of the sample, so ε_0 is uncorrelated with ε_i for all $i = 1, \cdots, n$, therefore

- $cov(\varepsilon_0, \bar{Y}) = 0$
- $cov(\varepsilon_0, \hat{\beta}_1) = 0$
- $cov(\bar{Y}, \hat{\beta}_1) = 0$ is proved the same as before.

$$Var(e_0) = Var(\bar{y}) + (x_0 - \bar{x})^2 \, Var(\hat{\beta}_1) + Var(\varepsilon_0)$$

$$= \frac{\sigma^2}{n} + (x_0 - \bar{x})^2 \frac{\sigma^2}{S_{xx}} + \sigma^2$$

$$= \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right]$$

- when the new x_0 is far away from $\bar{x}$, variance of the prediction bias is large. $\rightarrow$ unstable prediction