

Lecture 39 Sampling Distribution For Coefficients In Simple Linear Regression

BIO210 Biostatistics

Xi Chen

Fall, 2023

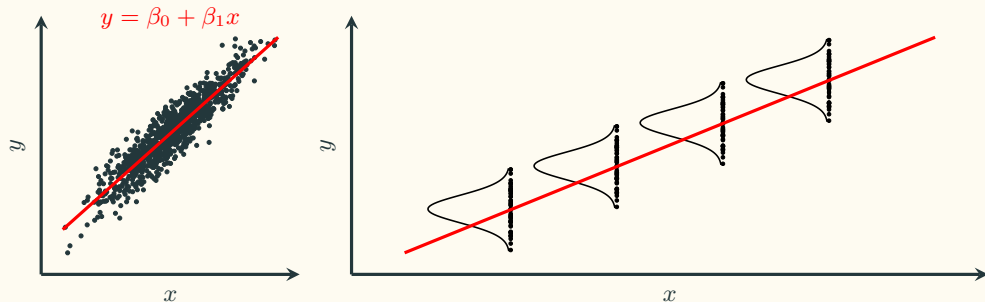
School of Life Sciences

Southern University of Science and Technology



南方科技大学生命科学学院
SUSTech · SCHOOL OF
LIFE SCIENCES

Summary of Simple Linear Regression Using OLS



Population regression line: $\mathbb{E}[Y|X] = \mu_{y|x} = \beta_0 + \beta_1 x$

Take a sample to make estimate β_0 and β_1 using OLS:

$$\hat{y} = \hat{\mu}_{y|x} = \hat{\beta}_0 + \hat{\beta}_1 x, \text{ where } \hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Sampling Distribution of The Coefficients in OLS

$$\begin{array}{ccc} \text{Population regression line:} & \xrightarrow{\text{take a sample}} & \text{OLS regression line:} \\ \mathbb{E}[\mathbf{Y}|\mathbf{X}] = \mu_{y|x} = \beta_0 + \beta_1 x & & \hat{\mu}_{y|x} = \hat{\beta}_0 + \hat{\beta}_1 x \end{array}$$

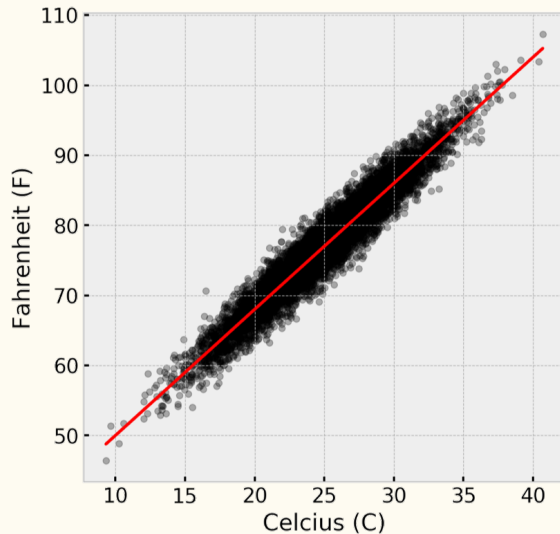
$\hat{\mu}_{y|x}$, $\hat{\beta}_0$, $\hat{\beta}_1$ have nice distributions

$$\hat{\beta}_0 \sim \mathcal{N} \left(\beta_0, \frac{\sigma_{\epsilon}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \cdot \frac{\sum_{i=1}^n x_i^2}{n} \right)$$

$$\hat{\beta}_1 \sim \mathcal{N} \left(\beta_1, \frac{\sigma_{\epsilon}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right)$$

$$\hat{\mu}_{y|x} \sim \mathcal{N} \left(\mu_{y|x}, \sigma_{\epsilon}^2 \cdot \left[\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right] \right)$$

Sampling Distribution of The Coefficients in OLS - Example



Population regression line:

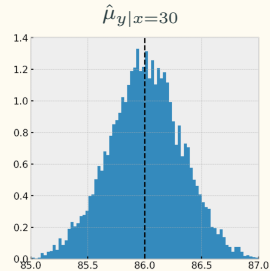
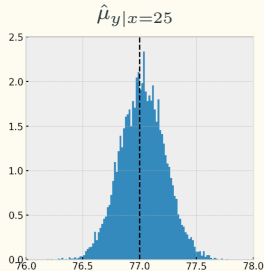
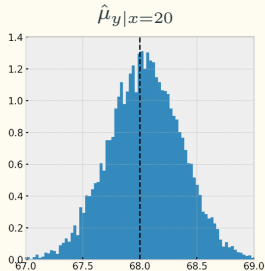
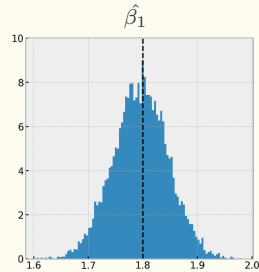
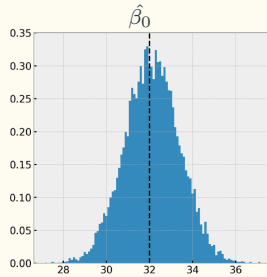
$$F = \beta_0 + \beta_1 \cdot C$$

$$\beta_0 = 32$$

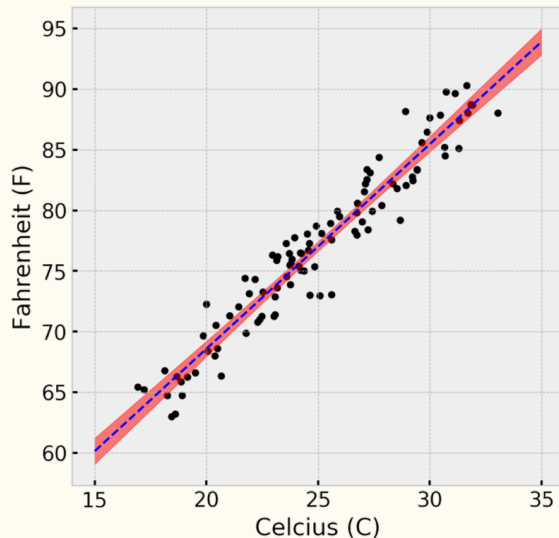
$$\beta_1 = 1.8$$

$$\sigma_{\epsilon}^2 = 4$$

Sampling Distribution of The Coefficients in OLS - Example



95% Confidence Interval for $\hat{\mu}_{y|x}$



$$F = 34.85 + 1.69 \cdot C$$

95% confidence interval of $\mathbb{E}[F|C]$

What Is σ_{ϵ}^2 ?

$$\hat{\beta}_0 \sim \mathcal{N}\left(\beta_0, \frac{\sigma_{\epsilon}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \cdot \frac{\sum_{i=1}^n x_i^2}{n}\right)$$

$$\hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma_{\epsilon}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)$$

$$\hat{\mu}_{y|x} \sim \mathcal{N}\left(\mu_{y|x}, \sigma_{\epsilon}^2 \cdot \left[\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right]\right)$$

In reality, we rarely know σ_{ϵ}^2 , what is the best estimate for σ_{ϵ}^2 ?

$$s_y^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n - 1} ?$$

good estimate for the variance of the entire population of y, not for σ_{ϵ}^2

We denote the best estimate for σ_{ϵ}^2 as s_{ϵ}^2 . Since $\sigma_{\epsilon}^2 = \text{Var}(\epsilon|x)$, intuitively, we should use:

$$s_{\epsilon}^2 = MSE = \frac{SSE}{n - 2} = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - 2}$$

When using s_{ϵ}^2 to estimate σ_{ϵ}^2 , we introduce some error, those distributions become t_{n-2}

Is There A Linear Relationship Between x And y ?

$$\begin{array}{l} H_0: \text{no linear relationship} \\ H_1: \text{some linear relationship} \end{array} \left\{ \begin{array}{l} \text{Use Pearson's } r : \begin{array}{l} H_0 : \rho = 0 \\ H_1 : \rho \neq 0 \end{array} \quad \frac{r}{\sqrt{(1-r^2)/(n-2)}} \sim t_{n-2} \\ \\ \text{Use Regression slope : } \begin{array}{l} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{array} \quad \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{MSE}{\sum_{i=1}^n (x_i - \bar{x})^2}}} \sim t_{n-2} \\ \\ \text{Use var. : } \begin{array}{l} H_0 : \text{most var. is NOT explained by the regression} \\ H_1 : \text{most var. is explained by the regression} \end{array} \quad \frac{MSR}{MSE} \sim \mathcal{F}(1, n-2) \end{array} \right.$$