Lecture 39 Sampling Distribution For Coefficients In Simple Linear Regression

BIO210 Biostatistics

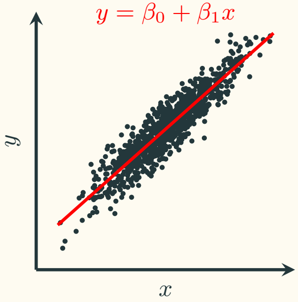
Xi Chen

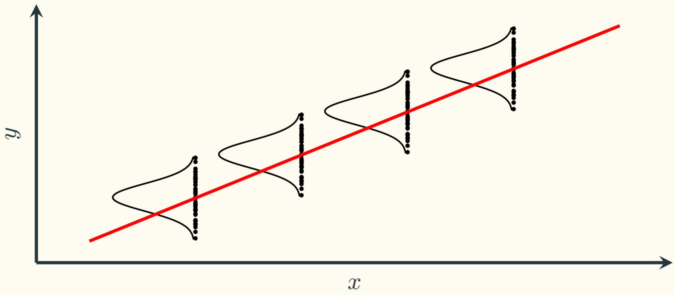
Spring, 2022

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: $E[\boldsymbol{Y}|\boldsymbol{X}] = \mu_{y|x} = \beta_0 + \beta_1 x$

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

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

The Distribution of ϵ

According to the **LINE** assumptions: $\boldsymbol{\epsilon} \sim \mathcal{N}(?, ?)$

$$\boldsymbol{Y}|_{\boldsymbol{X}=x} = \beta_0 + \beta_1 x + (\boldsymbol{\epsilon}|_{\boldsymbol{X}=x}) \Rightarrow \boldsymbol{\epsilon}|_{\boldsymbol{X}=x} = (\boldsymbol{Y}|_{\boldsymbol{X}=x}) - (\beta_0 + \beta_1 x)$$
$$\Rightarrow E[\boldsymbol{\epsilon}|_{\boldsymbol{X}=x}] = E[(\boldsymbol{Y}|_{\boldsymbol{X}=x}) - (\beta_0 + \beta_1 x)] = E[(\boldsymbol{Y}|_{\boldsymbol{X}=x})] - E[(\beta_0 + \beta_1 x)]$$
$$\Rightarrow E[\boldsymbol{\epsilon}|_{\boldsymbol{X}=x}] = \mu_{y|x} - \mu_{y|x} = 0$$

$$var(\boldsymbol{\epsilon}|_{\boldsymbol{X}=x}) = var[(\boldsymbol{Y}|_{\boldsymbol{X}=x}) - (\beta_0 + \beta_1 x)] = var[(\boldsymbol{Y}|_{\boldsymbol{X}=x}) - \mu_{y|x}]$$
$$\Rightarrow var(\boldsymbol{\epsilon}|_{\boldsymbol{X}=x}) = var[\boldsymbol{Y}|_{\boldsymbol{X}=x}] = \sigma^2_{y|x}$$

$\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2_{y|x})$, $\sigma^2_{y|x}$ is called the “common error variance”.

Sampling Distribution of The Coefficients in OLS

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

$\xrightarrow{\text{take a sample}}$

OLS regression line:
$\hat{\mu}_{y|x} = \hat{\beta}_0 + \hat{\beta}_1 x$

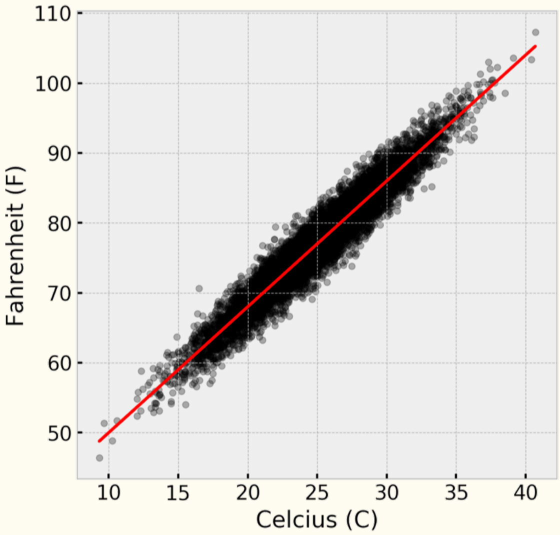
$\hat{\mu}_{y|x}$, $\hat{\beta}_0$, $\hat{\beta}_1$ have nice distributions

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

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

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

Sampling Distribution of The Coefficients in OLS - Example



Population regression line:

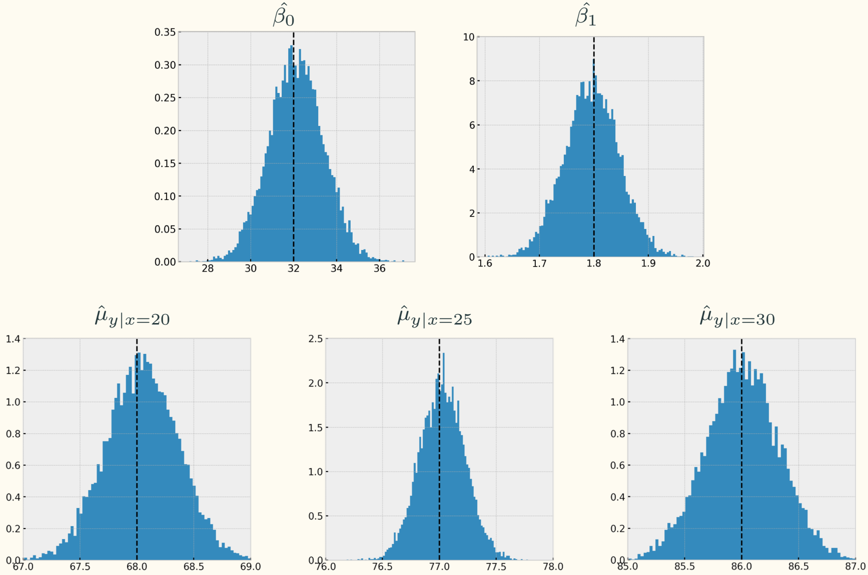
$F = \beta_0 + \beta_1 \cdot C$

$\beta_0 = 32$

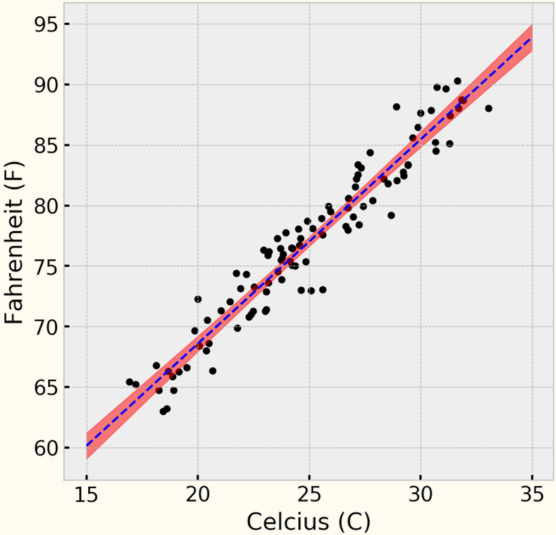
$\beta_1 = 1.8$

$\sigma^2_{y|x} = 4$

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 $E[F|C]$

What Is $\sigma^2_{y|x}$?

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

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

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

In reality, we rarely know $\sigma^2_{y|x}$, what is the best estimate for $\sigma^2_{y|x}$?

$s_y^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1}$? good estimate for the variance of the entire population of y, not for $\sigma^2_{y|x}$

We denote the best estimate for $\sigma^2_{y|x}$ as $s^2_{y|x}$. Since $\sigma^2_{y|x} = var(\boldsymbol{\epsilon}|x)$, intuitively, we should use:

$$s^2_{y|x} = MSE = \frac{SS_E}{n-2} = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2}$$

When using $s^2_{y|x}$ to estimate $\sigma^2_{y|x}$, we introduce some error, those distributions become $\boldsymbol{t}_{n-2}$

Is There A Linear Relationship Between x And y ?

H_0: no linear relationship

H_1: some linear relationship

$$\begin{cases} \text{Use Pearson's } r: \begin{array}{l} H_0: \rho = 0 \\ H_1: \rho \neq 0 \end{array} \quad \dfrac{r}{\sqrt{(1-r^2)/(n-2)}} \sim \boldsymbol{t}_{n-2} \\[2em] \text{Use Regression slope}: \begin{array}{l} H_0: \beta_1 = 0 \\ H_1: \beta_1 \neq 0 \end{array} \quad \dfrac{\hat{\beta}_1 - \beta_1}{\sqrt{\dfrac{MSE}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}} \sim \boldsymbol{t}_{n-2} \\[2em] \text{Use var.}: \begin{array}{l} H_0: \text{most var. is NOT explained by the regression} \\ H_1: \text{most var. is explained by the regression} \end{array} \end{cases}$$

$$\frac{MSR}{MSE} \sim \boldsymbol{F}_{1,n-2}$$