Lecture 39 Sampling Distribution For Coefficients In Simple Linear Regression

BIO210 Biostatistics

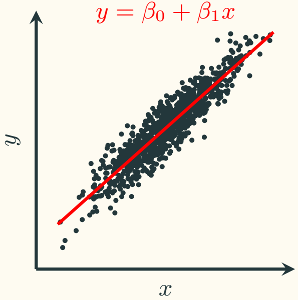
Xi Chen

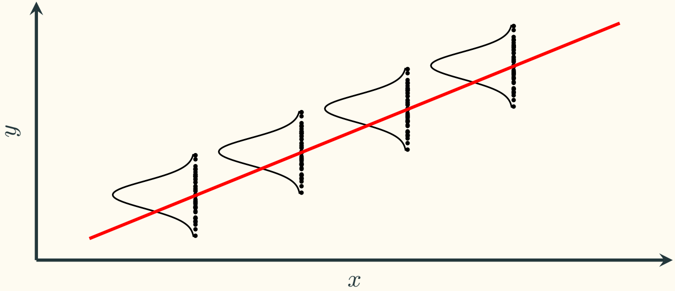
Fall, 2024

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

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

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

take a sample →

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

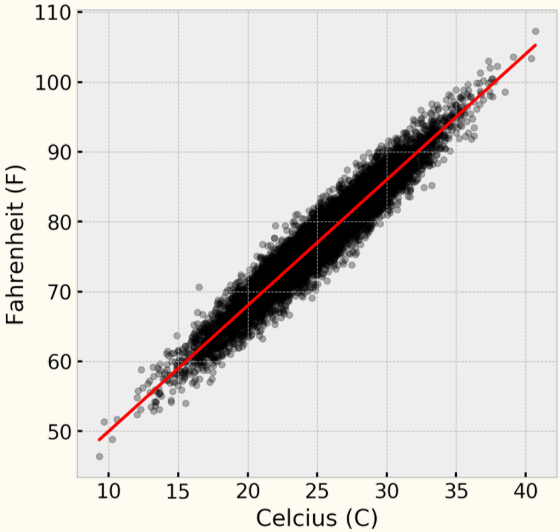
$\hat{\boldsymbol{\mu}}_{\boldsymbol{y|x}}, \hat{\boldsymbol{\beta}}_{\boldsymbol{0}}, \hat{\boldsymbol{\beta}}_{\boldsymbol{1}}$ have nice distributions

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{0}} \sim \mathcal{N}\left(\beta_0, \frac{\sigma_{\boldsymbol{\epsilon}}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \cdot \frac{\sum_{i=1}^n x_i^2}{n}\right)$$

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

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

Sampling Distribution of The Coefficients in OLS - Example



Population regression line:

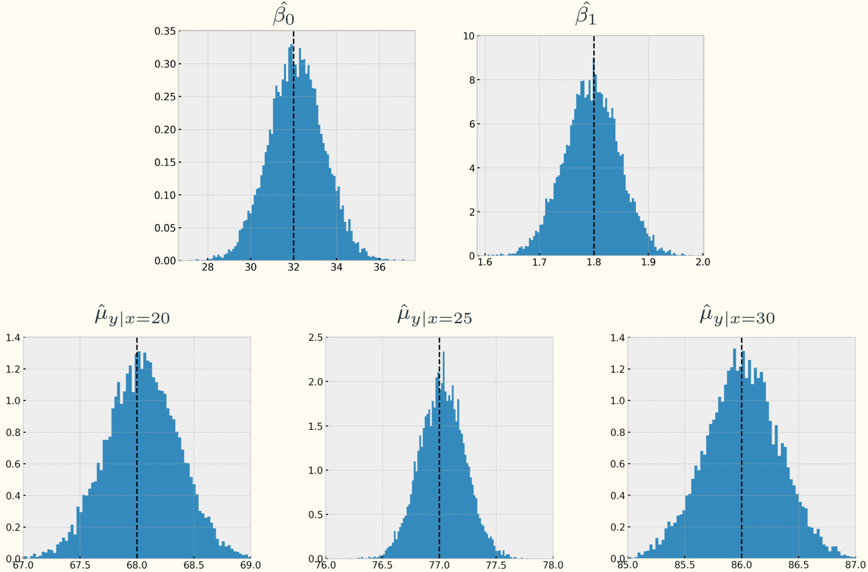
$F = \beta_0 + \beta_1 \cdot C$

$\beta_0 = 32$

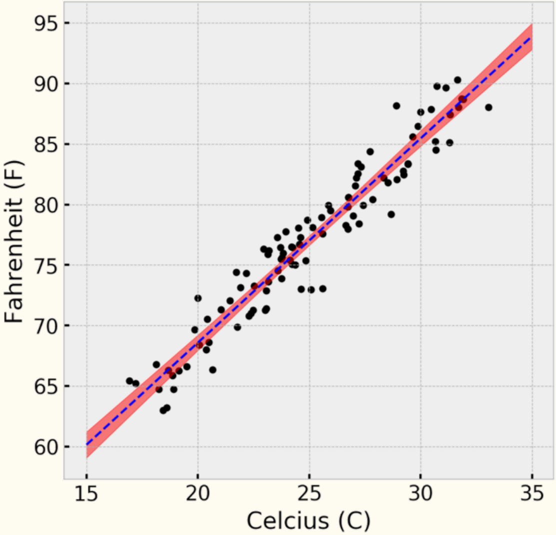
$\beta_1 = 1.8$

$\sigma_\epsilon^2 = 4$

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 $\sigma_\epsilon^2 = \mathbb{V}\text{ar}\,(\epsilon|x)$, 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 ?

H_0: no linear relationship

H_1: some linear relationship

- Use Pearson's r : $H_0 : \rho = 0$, $H_1 : \rho \neq 0$

$$\frac{r}{\sqrt{(1-r^2)/(n-2)}} \sim \boldsymbol{t}_{n-2}$$

- Use Regression slope : $H_0 : \beta_1 = 0$, $H_1 : \beta_1 \neq 0$

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\dfrac{MSE}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}} \sim \boldsymbol{t}_{n-2}$$

- Use var. : H_0 : most var. is NOT explained by the regression, H_1 : most var. is explained by the regression

$$\frac{MSR}{MSE} \sim \boldsymbol{\mathcal{F}}(1, n-2)$$