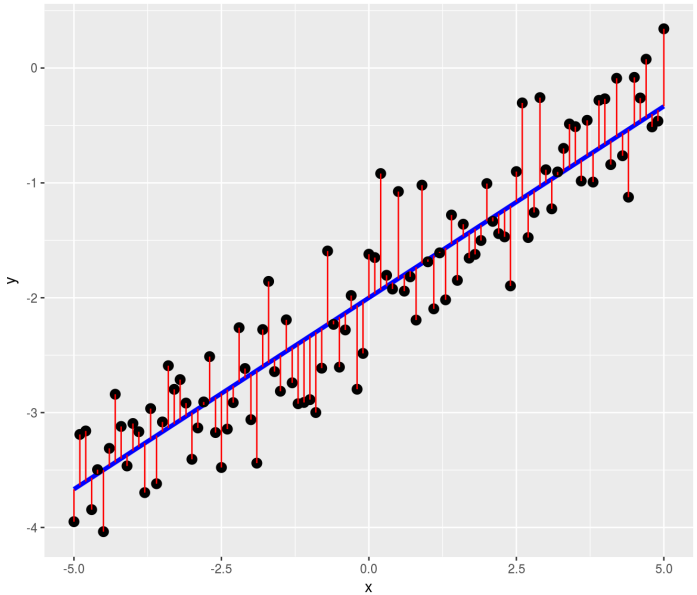
세상을 이해하는 통계학의 렌즈

REGRESSION

2020학년도 2학기

OLS Regression



$$\arg\min_{a,\ b} \frac{1}{n} \sum_{i=1}^{n} (y_i - (ax_i + b))^2$$

$$y = a^* x + b^*$$

OLS Regression

$$a^* = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$b^* = \bar{y} - a^*\bar{x}$$

$\bar{x}$: x의 평균
$\bar{y}$: y의 평균

$$f(\mathbf{x}) = a^*x + b^*$$

OLS Regression

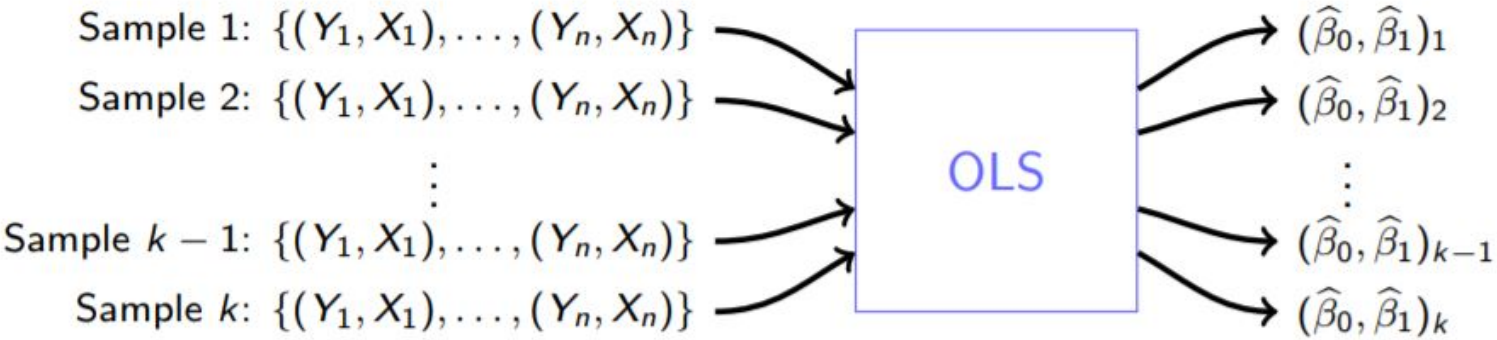
$$a^* = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = r_{XY} \cdot \frac{SD_y}{SD_x}$$

$\bar{x}$: x의 평균
$\bar{y}$: y의 평균

$$f(\mathbf{x}) = a^* x + b^*$$

Sampling distribution of the OLS estimator

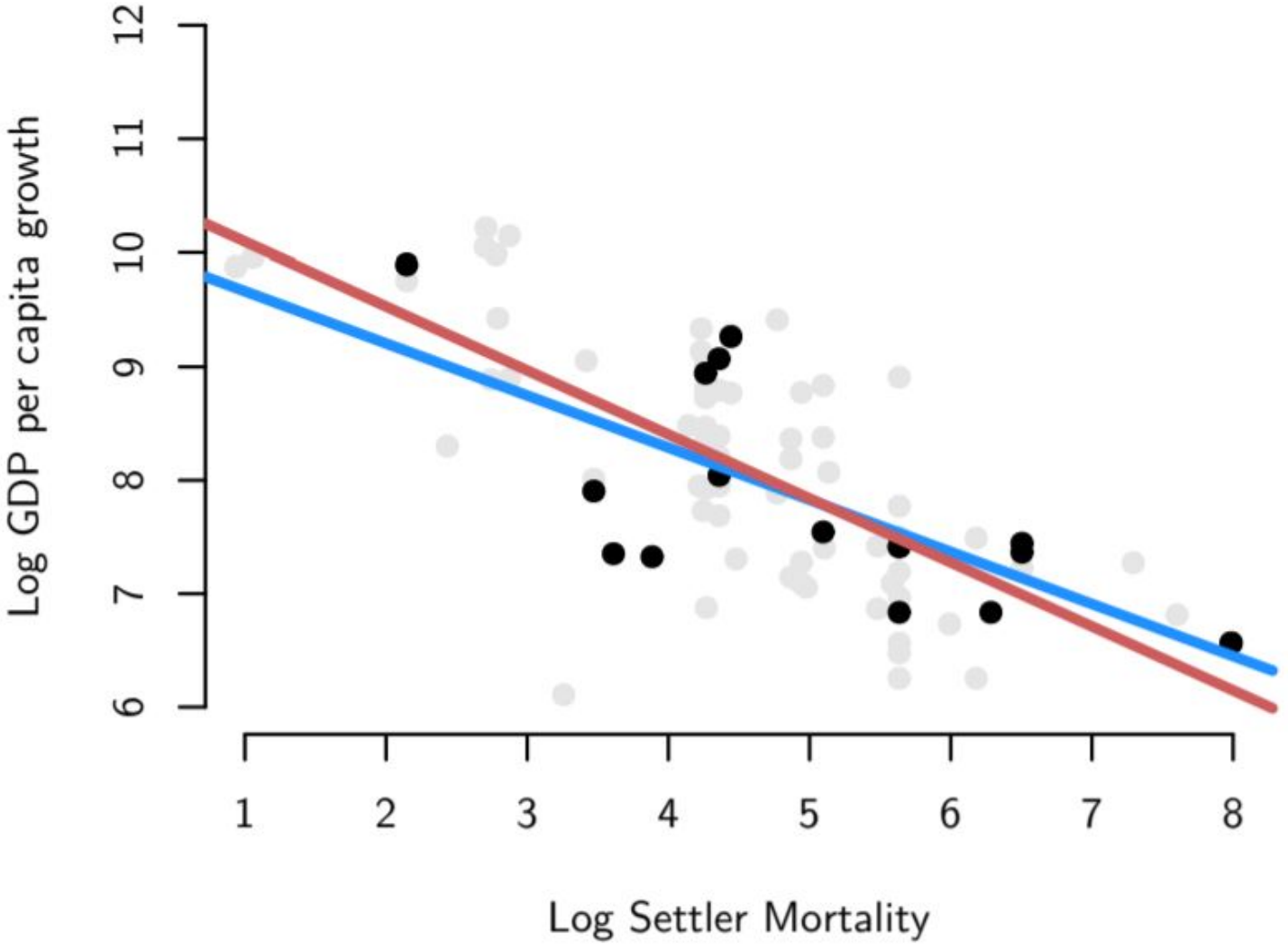
- Remember: OLS is an estimator—it's a machine that we plug samples into and we get out estimates.



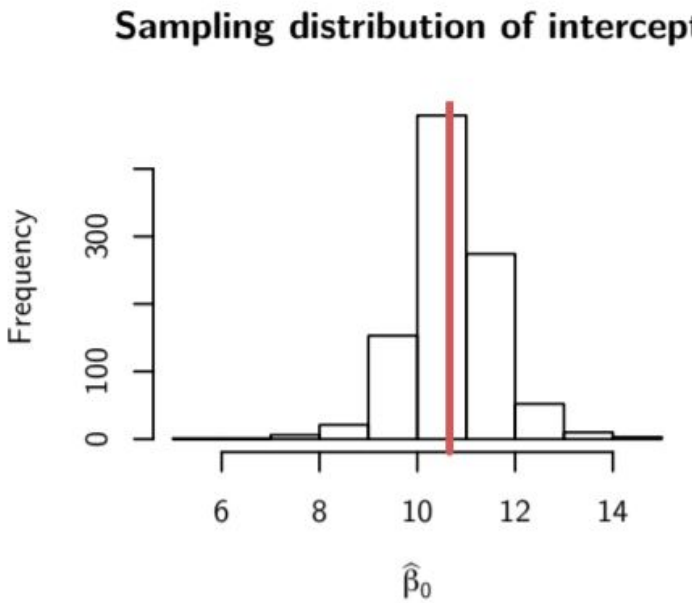
- Just like the sample mean, sample difference in means, or the sample variance
- It has a sampling distribution, with a sampling variance/standard error, etc.

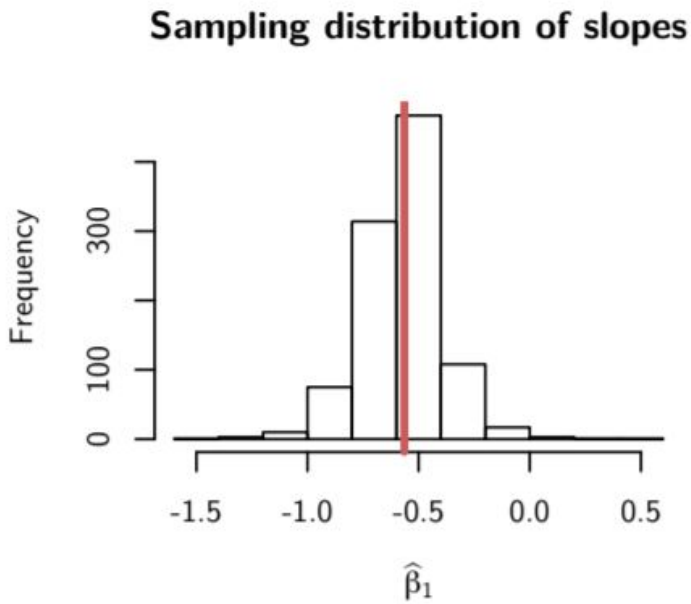
$$(\hat{\beta}_1, \hat{\beta}_0) = (a^*, b^*)$$

Sampling distribution of the OLS estimator



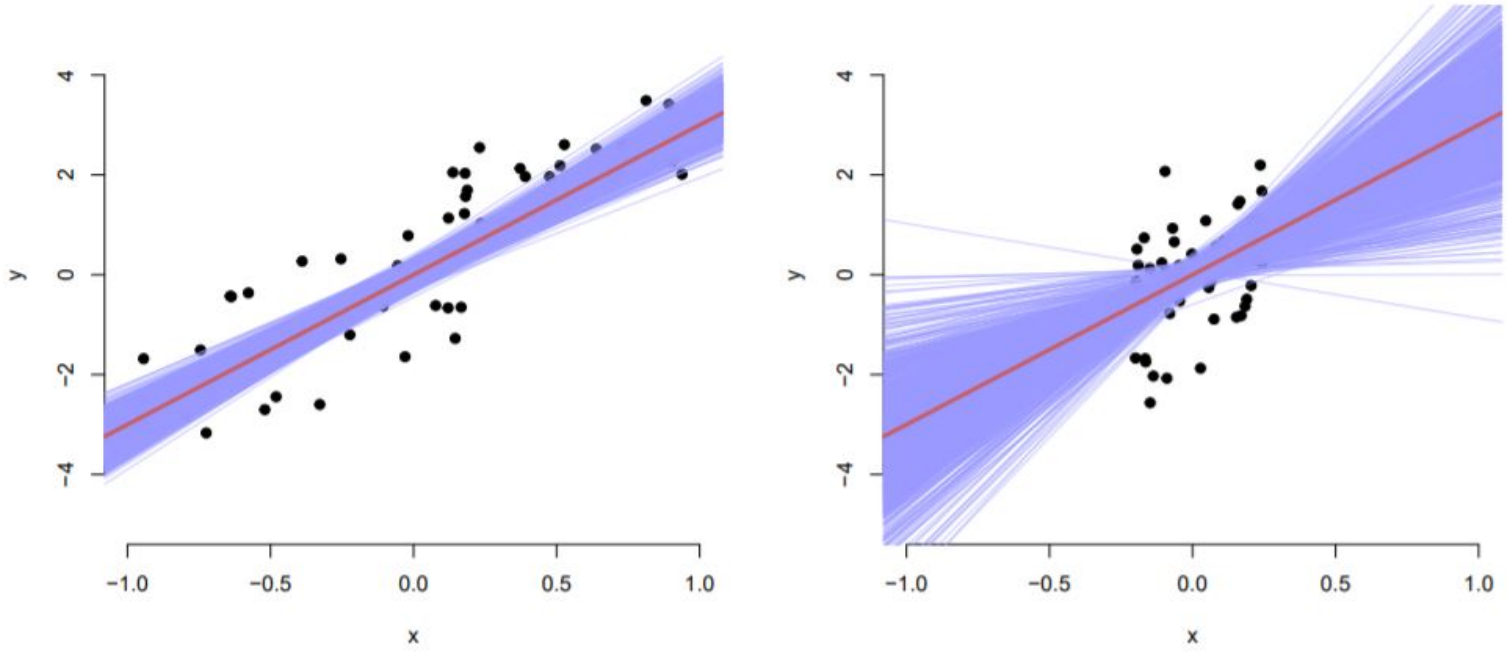
Sampling distribution of the OLS estimator





$$(\hat{\beta}_1, \hat{\beta}_0) = (a^*, b^*)$$

Sampling distribution of the OLS estimator



Sampling distribution of the OLS estimator

$$SE(a^*) = \frac{\sqrt{\frac{\sum_{i=1}^{n}(y_i - a^* x_i - b*)}{n-2}}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

$$SE(b^*) = \sqrt{\frac{\sum_{i=1}^{n}(y_i - a^* x_i - b^*)}{n-2}} \cdot \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Test statistic

- Under the null of $H_0 : \beta_1 = c$, we can use the following familiar test statistic:

$$T = \frac{\widehat{\beta}_1 - c}{\widehat{SE}[\widehat{\beta}_1]}$$

- Under the null hypothesis:
  - large samples: $T \sim \mathcal{N}(0, 1)$
  - any size sample with normal errors: $T \sim t_{n-2}$
  - conservative to use t_{n-2} anyways since t_{n-2} is approximately normal in large samples.
- Thus, under the null, we know the distribution of T and can use that to formulate a rejection region and calculate p-values.
- By default, R shows you the test statistic for $\beta_1 = 0$ and uses the t distribution.

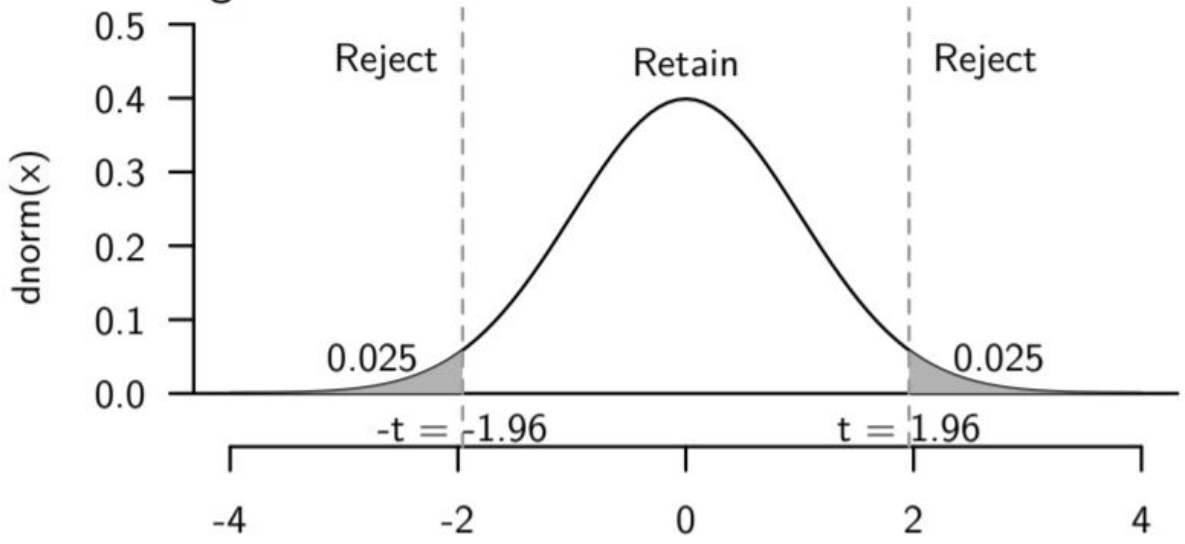
$$(\hat{\beta}_1, \hat{\beta}_0) = (a^*, b^*)$$

Rejection region

- Choose a level of the test, α, and find rejection regions that correspond to that value under the null distribution:

$$\mathbb{P}(-t_{\alpha/2,n-2} < T < t_{\alpha/2,n-2}) = 1 - \alpha$$

- This is exactly the same as with sample means and sample differences in means, except that the degrees of freedom on the t distribution have changed.



p-value

- The interpretation of the p-value is the same: the probability of seeing a test statistic at least this extreme if the null hypothesis were true
- Mathematically:

$$\mathbb{P}\left(\left|\frac{\widehat{\beta}_1 - c}{\widehat{SE}[\widehat{\beta}_1]}\right| \geq |T_{obs}|\right)$$

- If the p-value is less than α we would reject the null at the α level.

$$(\hat{\beta}_1, \hat{\beta}_0) = (a^*, b^*)$$