Simple Linear Regression

EC 339

Marcio Santetti

Fall 2022

Motivation

On notation

In our course, we will adopt the following **notation** for a regression model:

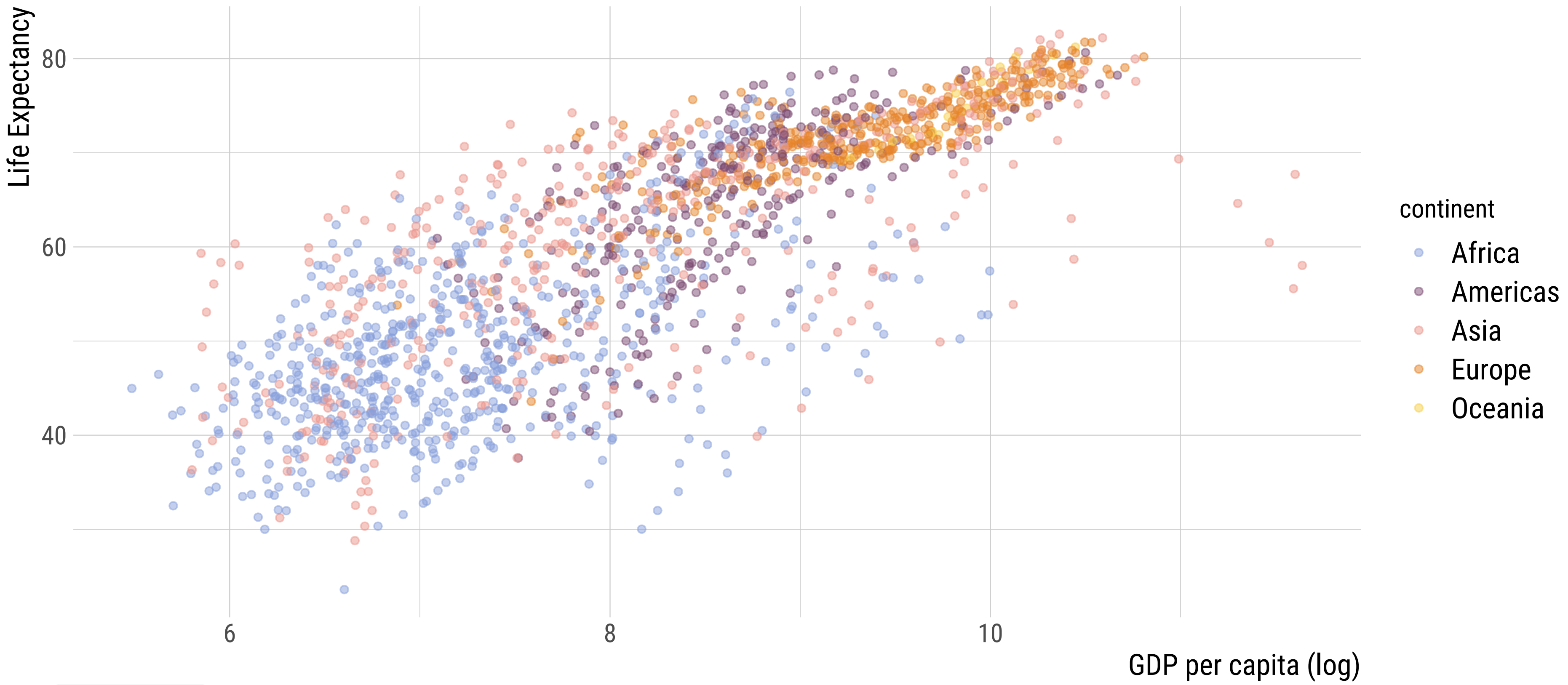
$$y_i = \beta_0 + \beta_1 x_{1i} + u_i$$

- where:
  - y_i: **dependent variable**'s value for the i^{th} individual;
  - x_i: **independent variable**'s value for the i^{th} individual;
  - β_0: **intercept** term;
  - β_1: **slope** coefficient;
  - u_i: **residual/error** term (the i^{th} individual's **random** deviation from the population parameters).

Motivating regression models

Data are fuzzy

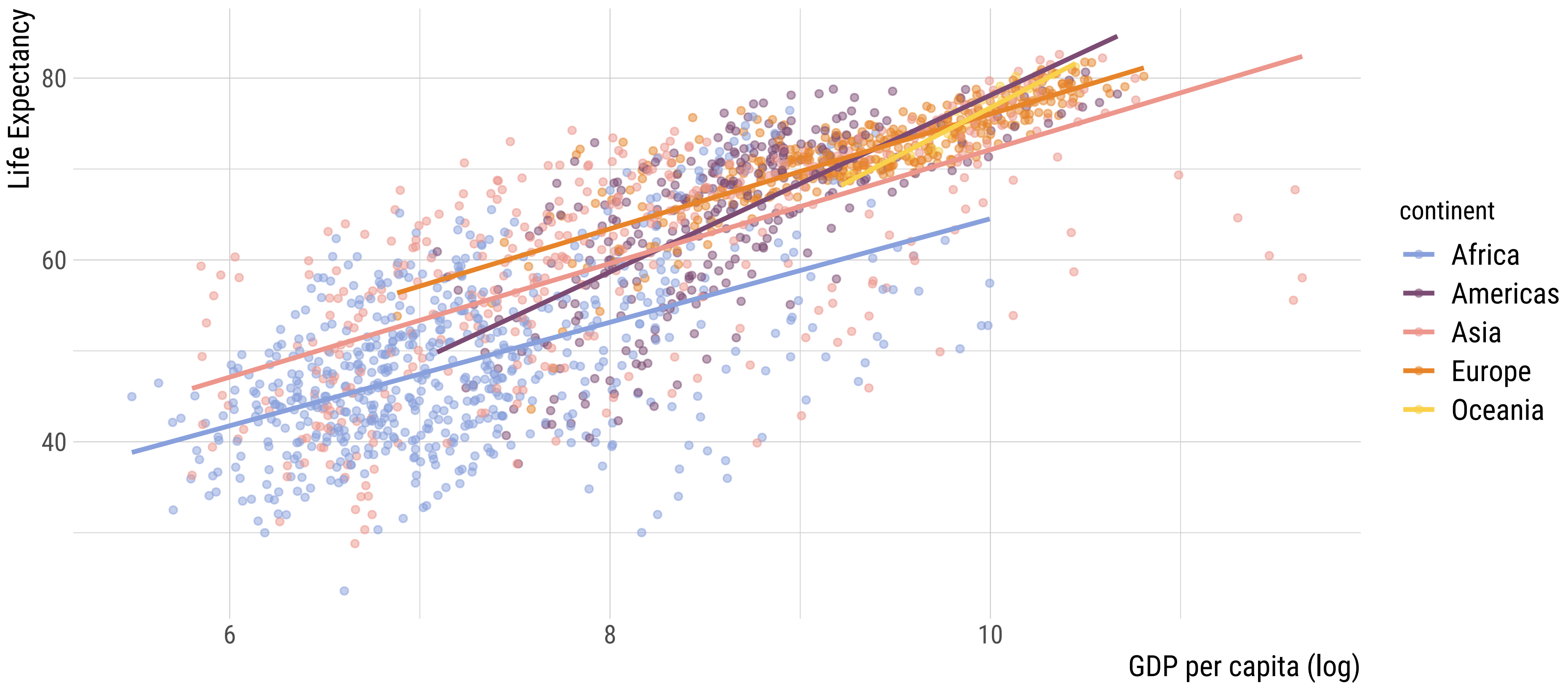
Life expectancy *vs.* GDP per capita (1952—2007):*



[*]: Data from `Gapminder`.

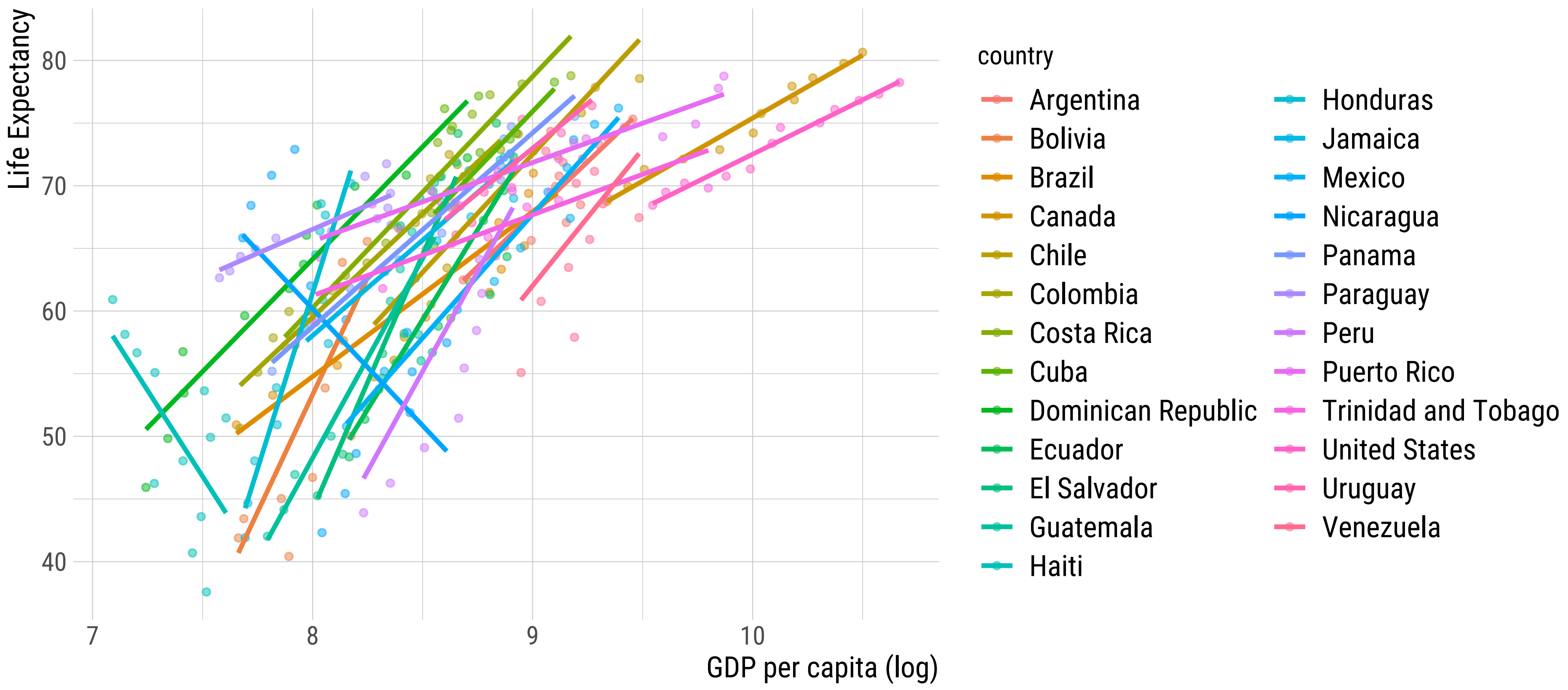
Data are fuzzy

Now, including **regression lines**:



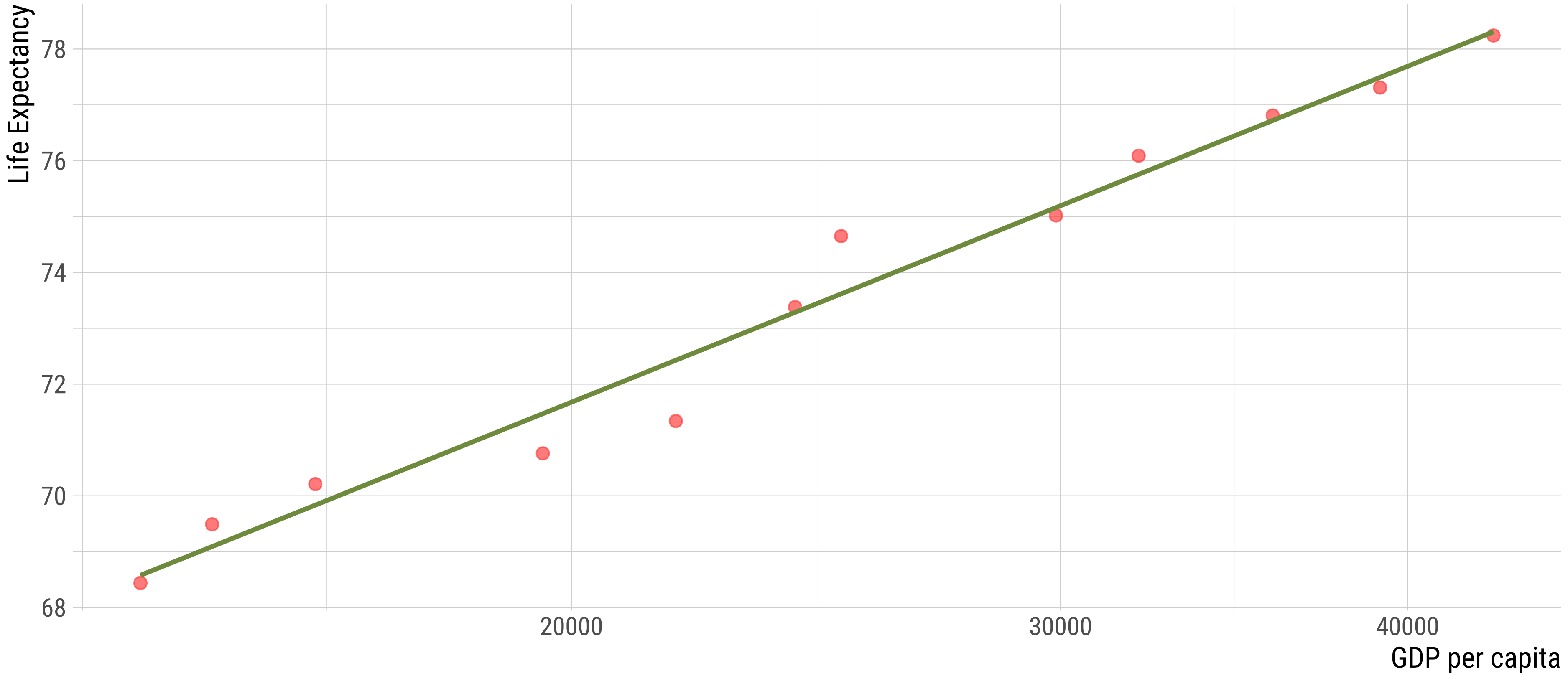
Data are fuzzy

Narrowing down to the Americas:



Data are fuzzy

Now, for the US…



Which method to use?

Ordinary Least Squares (OLS)

The **Ordinary Least Squares (OLS) Estimator**:

- OLS **minimizes** the *squared distance* between the data points and the regression line it generates.
- This way, we are **minimizing** *error* (*ignorance*) about our data and the relationship we are trying to better understand.
- In addition, it is **easy** to estimate and interpret.

Ordinary Least Squares (OLS)

The **Ordinary Least Squares (OLS) Estimator**:

$$\text{SSR} = \textstyle\sum_{i=1}^{n} u_i^2 \quad \text{where} \quad u_i = y_i - \hat{y}_i$$

- Why **squaring** these residuals?
- Bigger errors, bigger **penalties**.

$$\min_{\hat{\beta}_0, \hat{\beta}_1} \text{SSR}$$

$$\min_{\hat{\beta}_0, \hat{\beta}_1} (y_i - \hat{y}_i)^2$$

$$\min_{\hat{\beta}_0, \hat{\beta}_1} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2$$

Ordinary Least Squares (OLS)

The **Ordinary Least Squares (OLS) Estimator**:

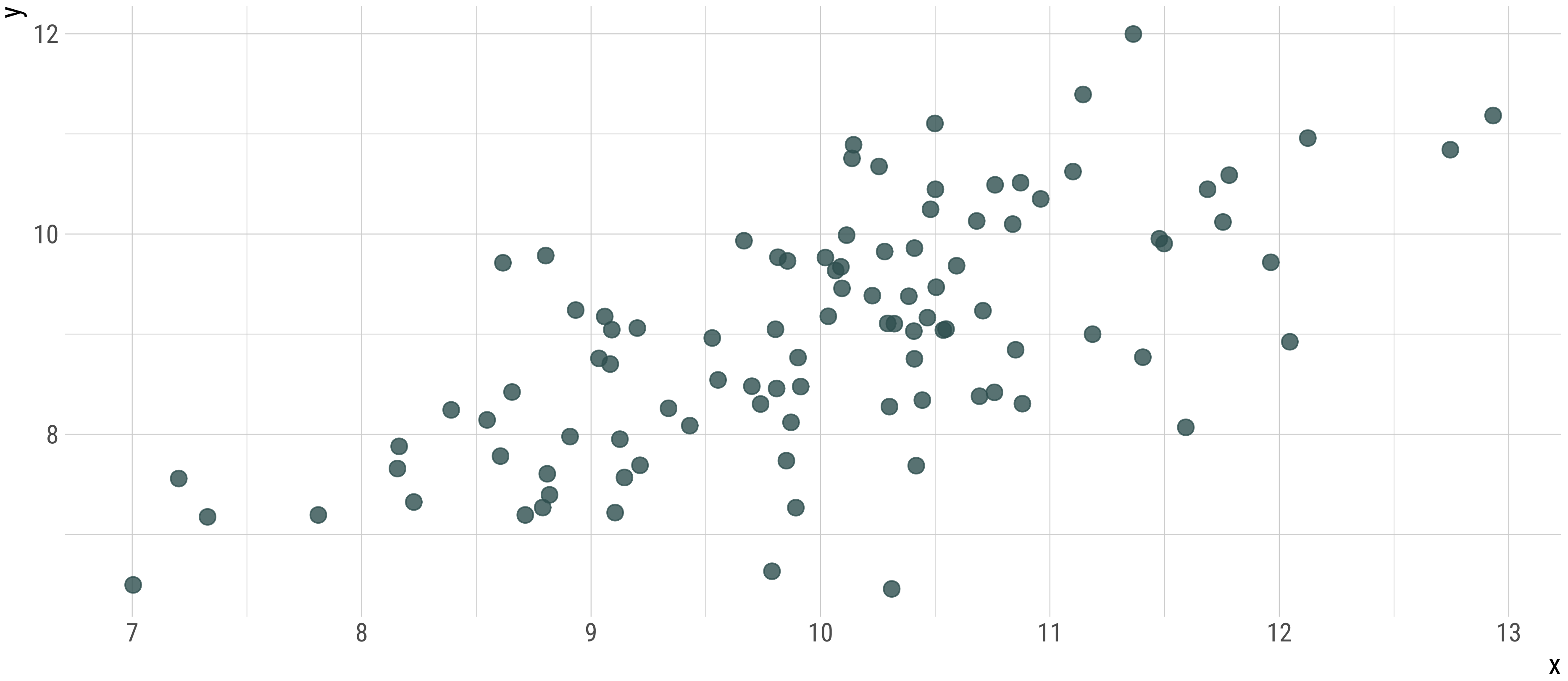
- **Slope coefficient**:

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sum_i (x_i - \overline{x})^2} = \frac{Cov(x, y)}{Var(x)}$$

- **Intercept coefficient**:

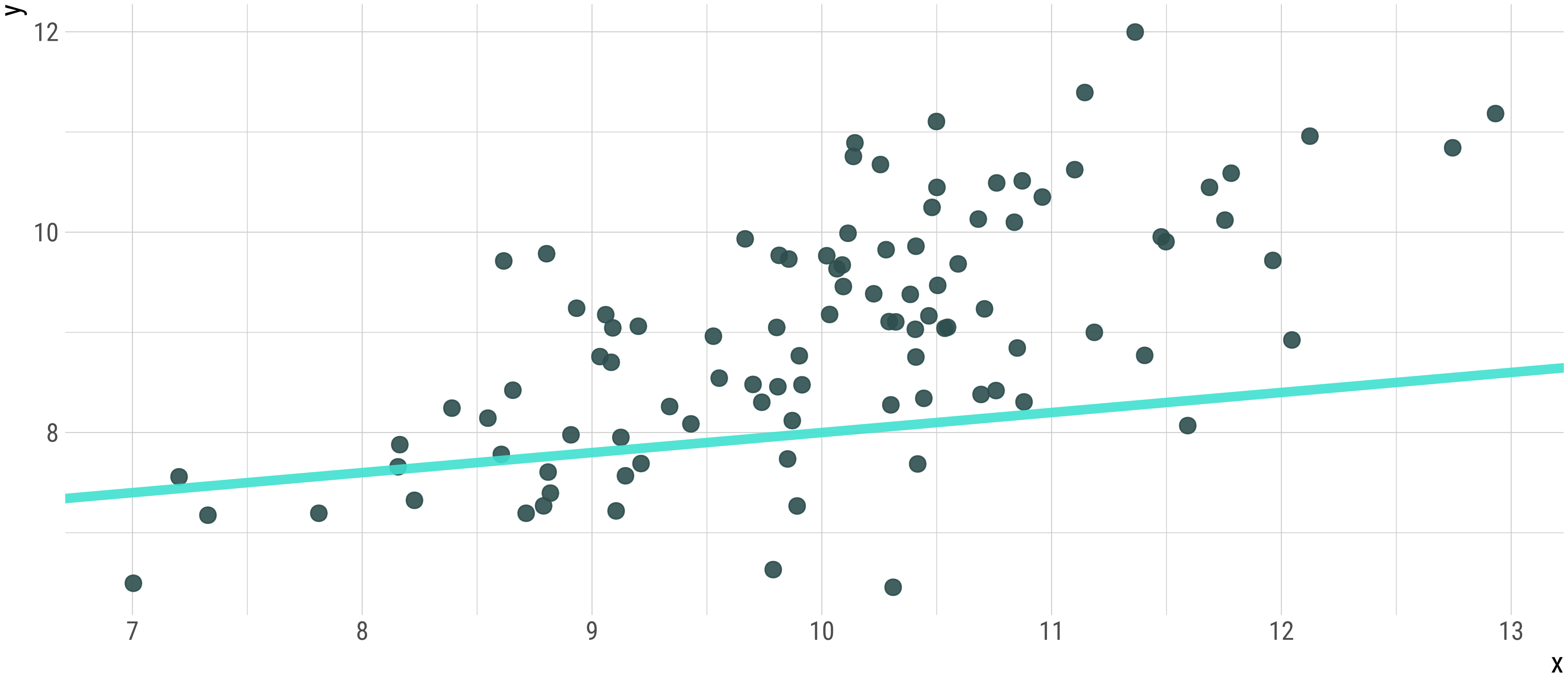
$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

"Best" regression lines



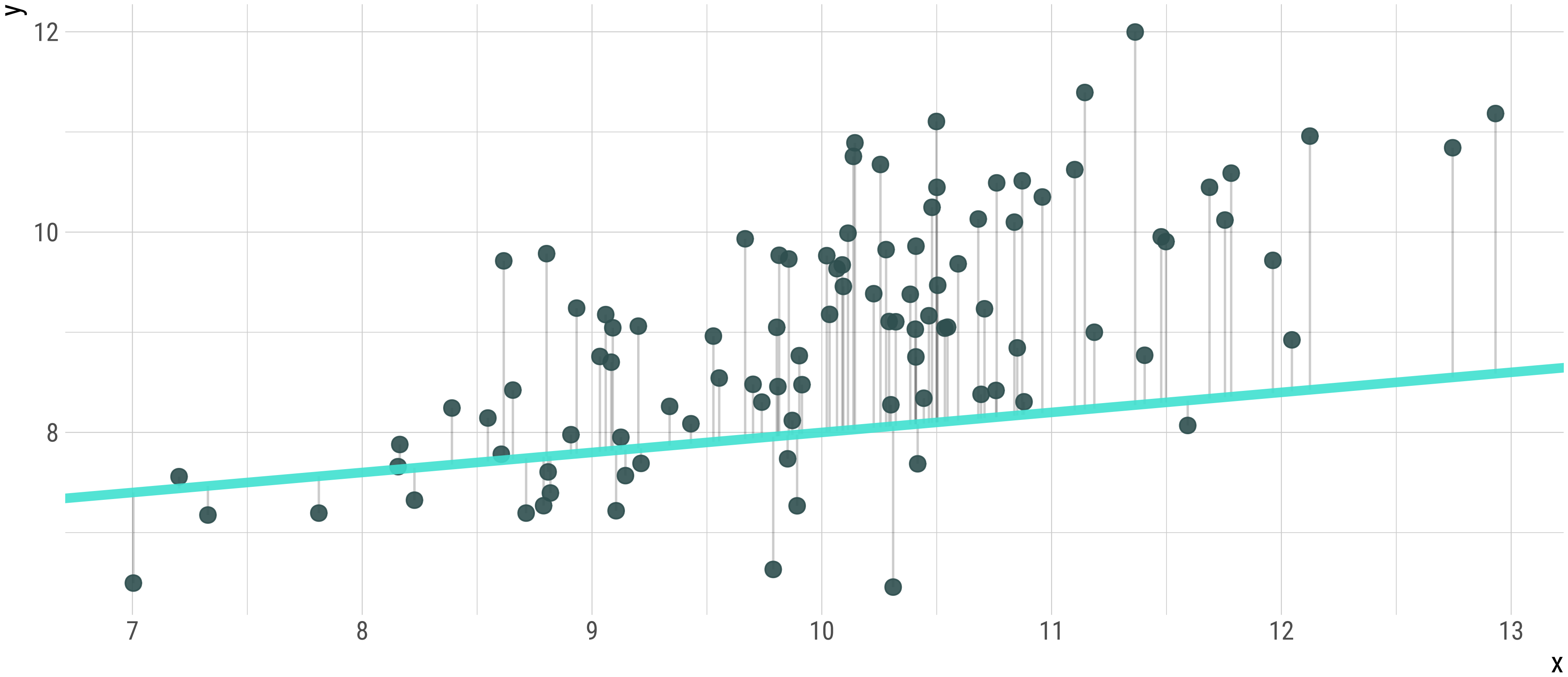
"Best" regression lines

For any line – $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$



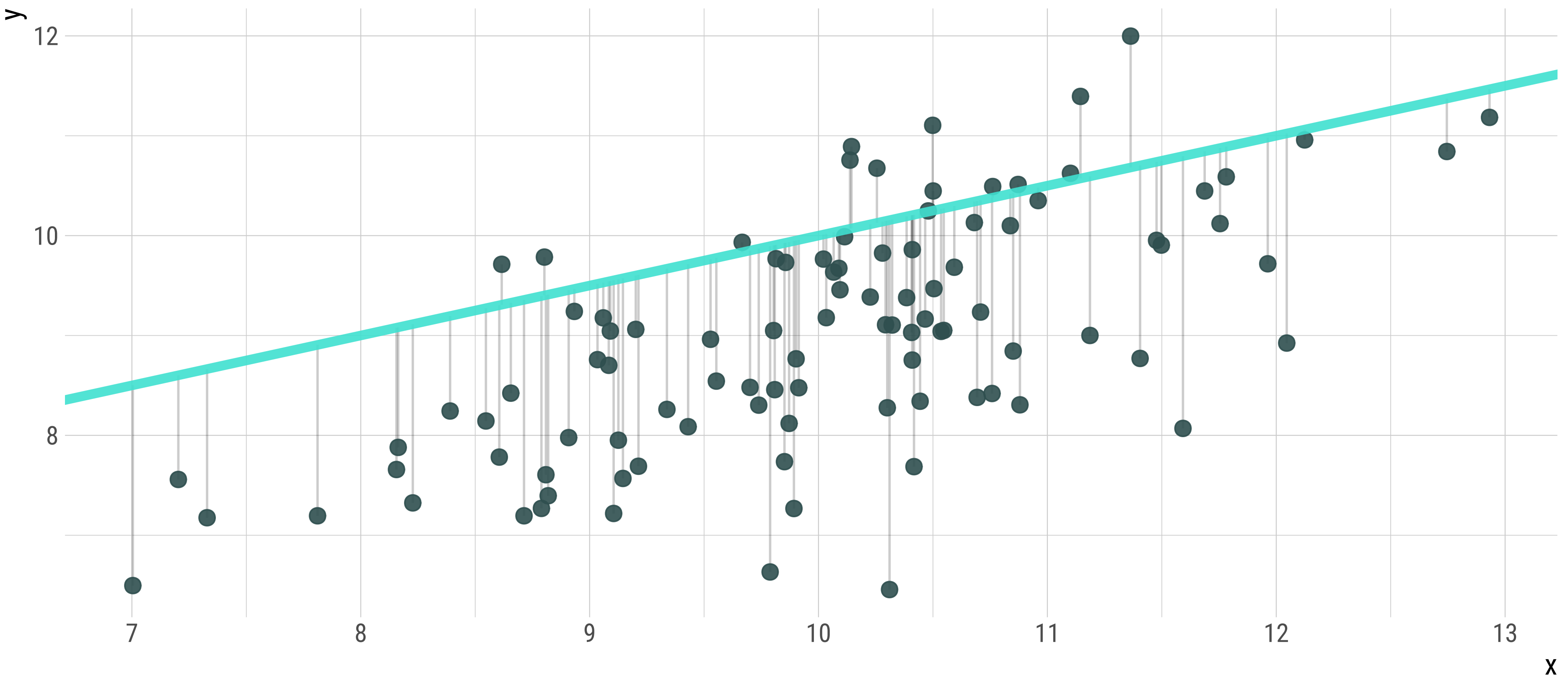
"Best" regression lines

For any line — $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ —, we can calculate residuals: $u_i = y_i - \hat{y}_i$



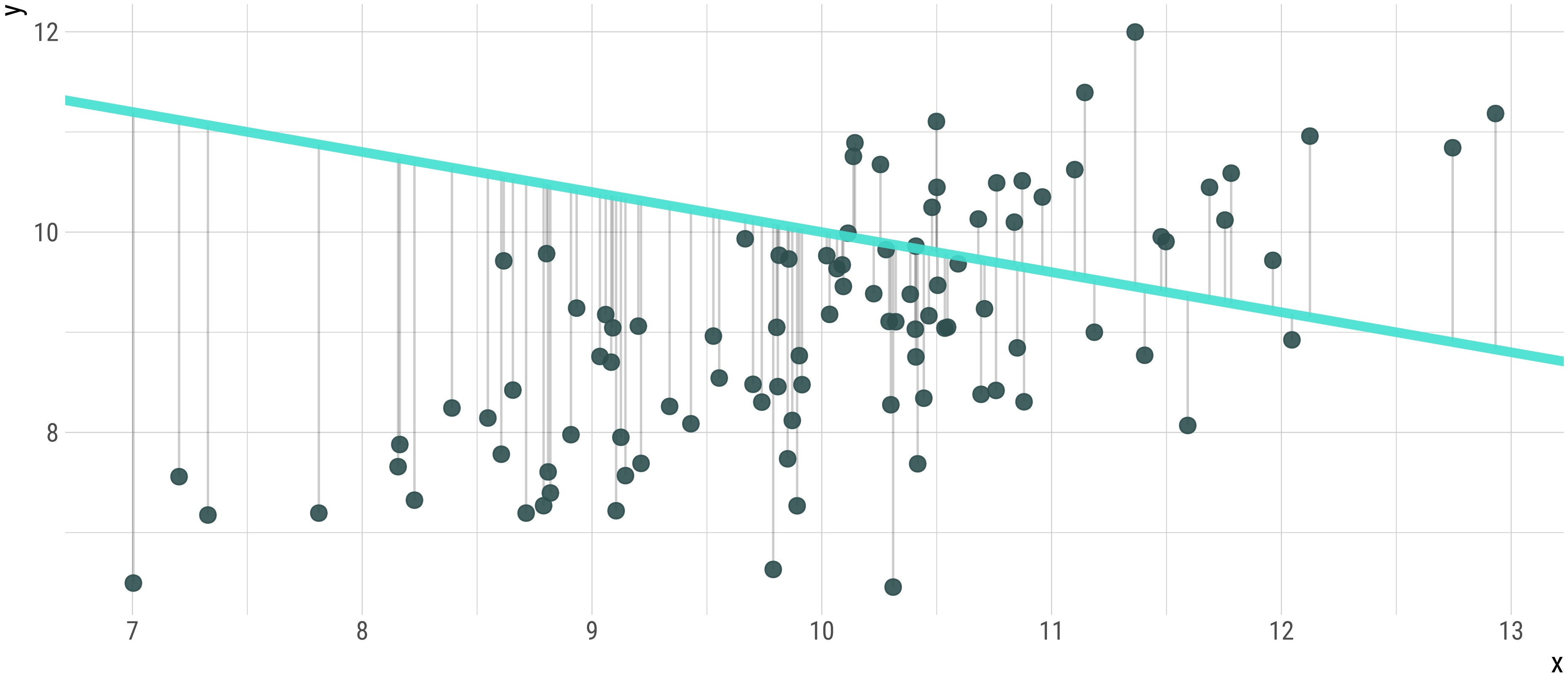
"Best" regression lines

For any line — $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ —, we can calculate residuals: $u_i = y_i - \hat{y}_i$



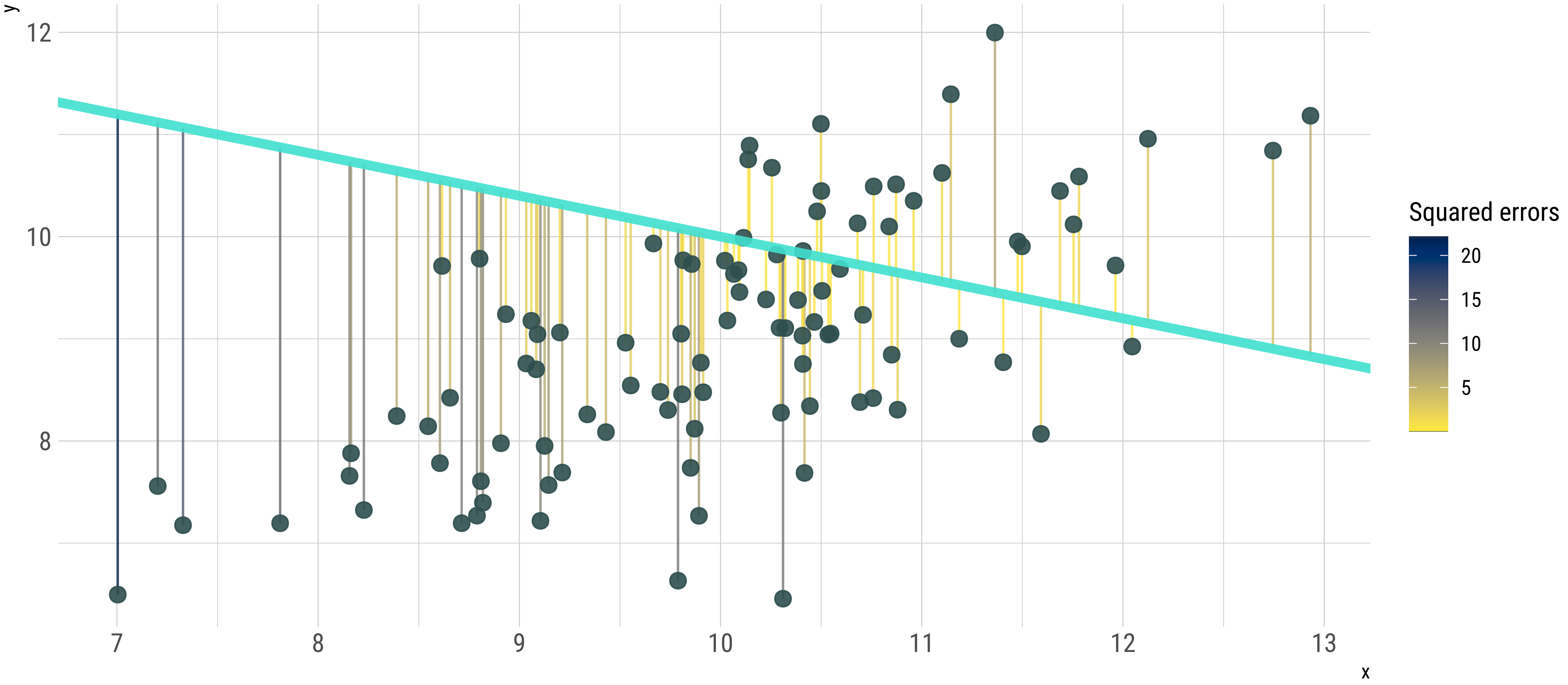
"Best" regression lines

For any line — $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ —, we can calculate residuals: $u_i = y_i - \hat{y}_i$



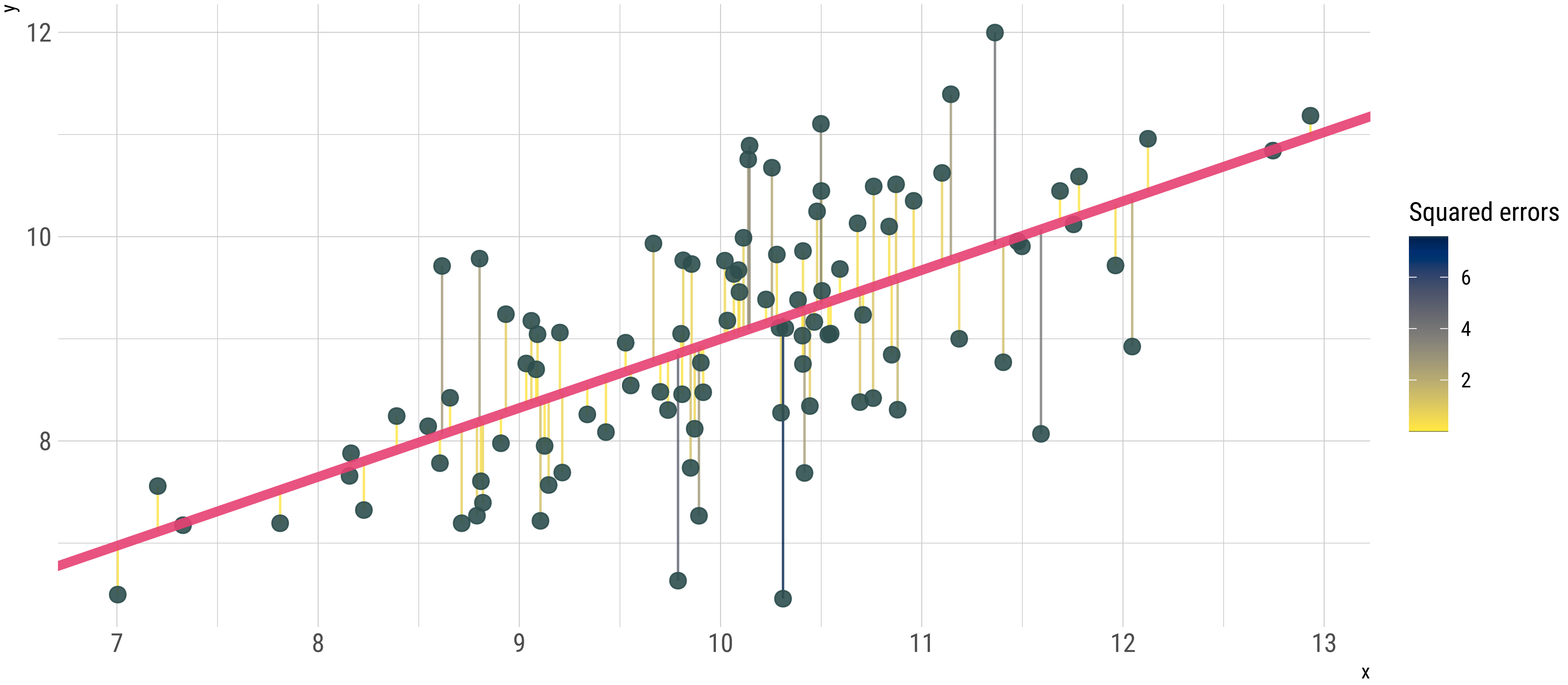
"Best" regression lines

SSR squares the errors $\left(\sum u_i^2\right)$: bigger errors get bigger penalties.



"Best" regression lines

The OLS estimate is the combination of $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize SSR.



Interpretation

Interpreting OLS coefficients

- **Slope** coefficient: the change (increase/decrease) in the dependent variable (y) generated by a 1-unit increase in the independent variable (x).
- **Intercept** term: the value of the dependent variable (y) when $x = 0$.

Example:

- Interpret the following estimated regression models:

$$\widehat{wage}_i = 10 + 2.65\ educ_i$$

$$\widehat{sleep}_i = 6.5 - 0.65\ kids_i$$

Next time: Simple regression in practice