Simple Linear Regression

Applied Regression in R

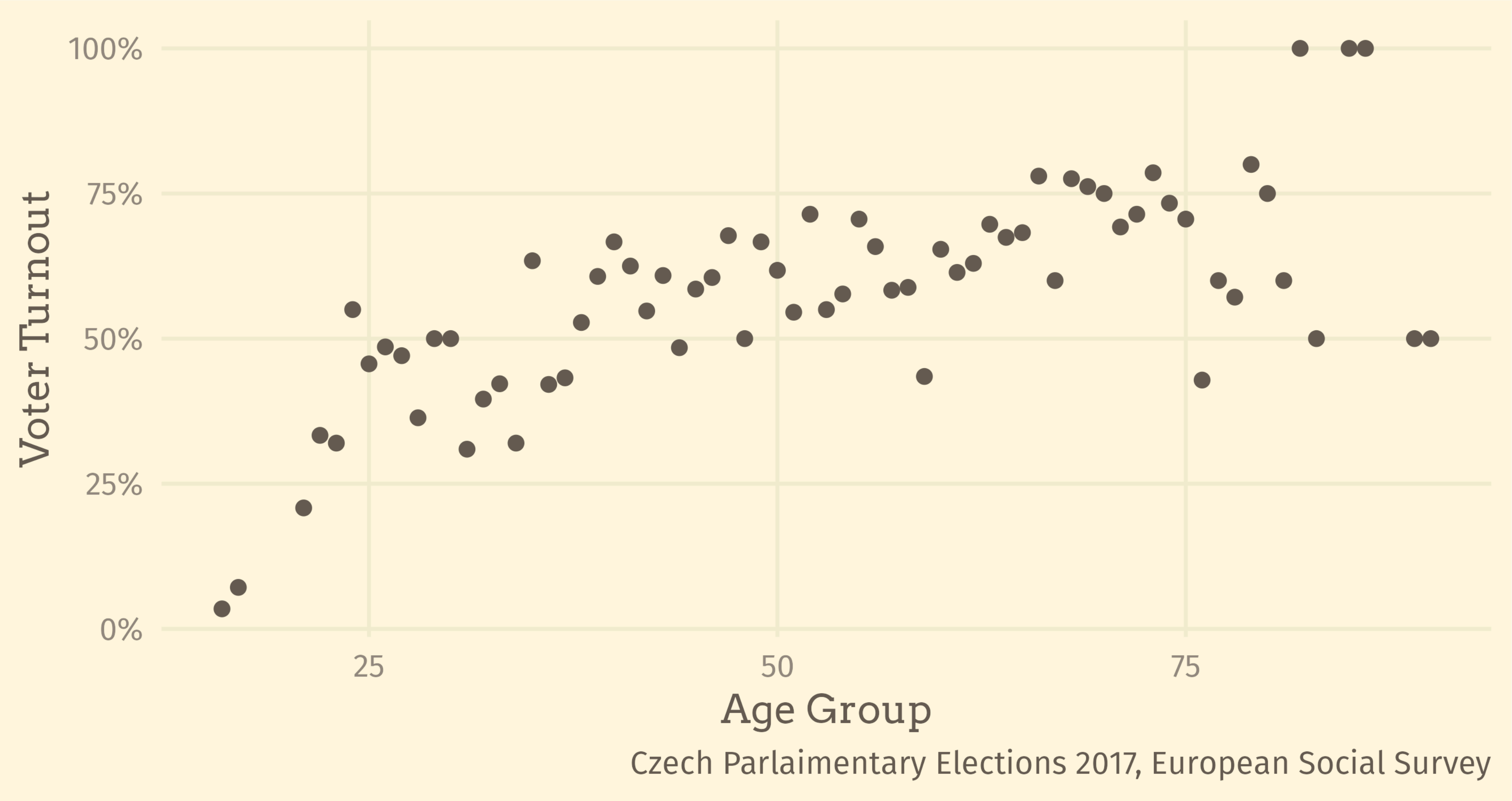
Aleš Vomáčka

11. 11. 2025

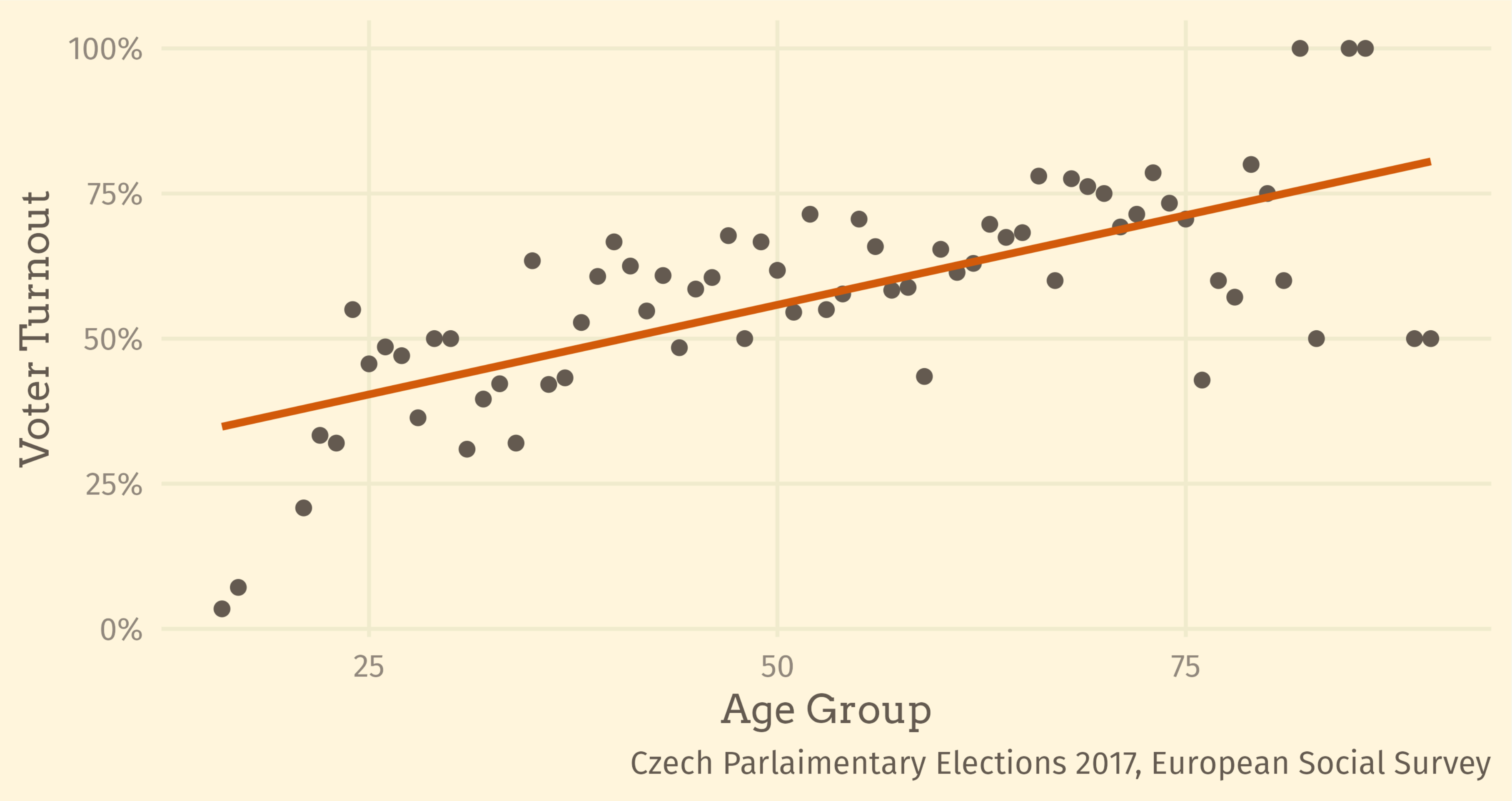
Faculty of Arts, Charles University

What is Linear regression?

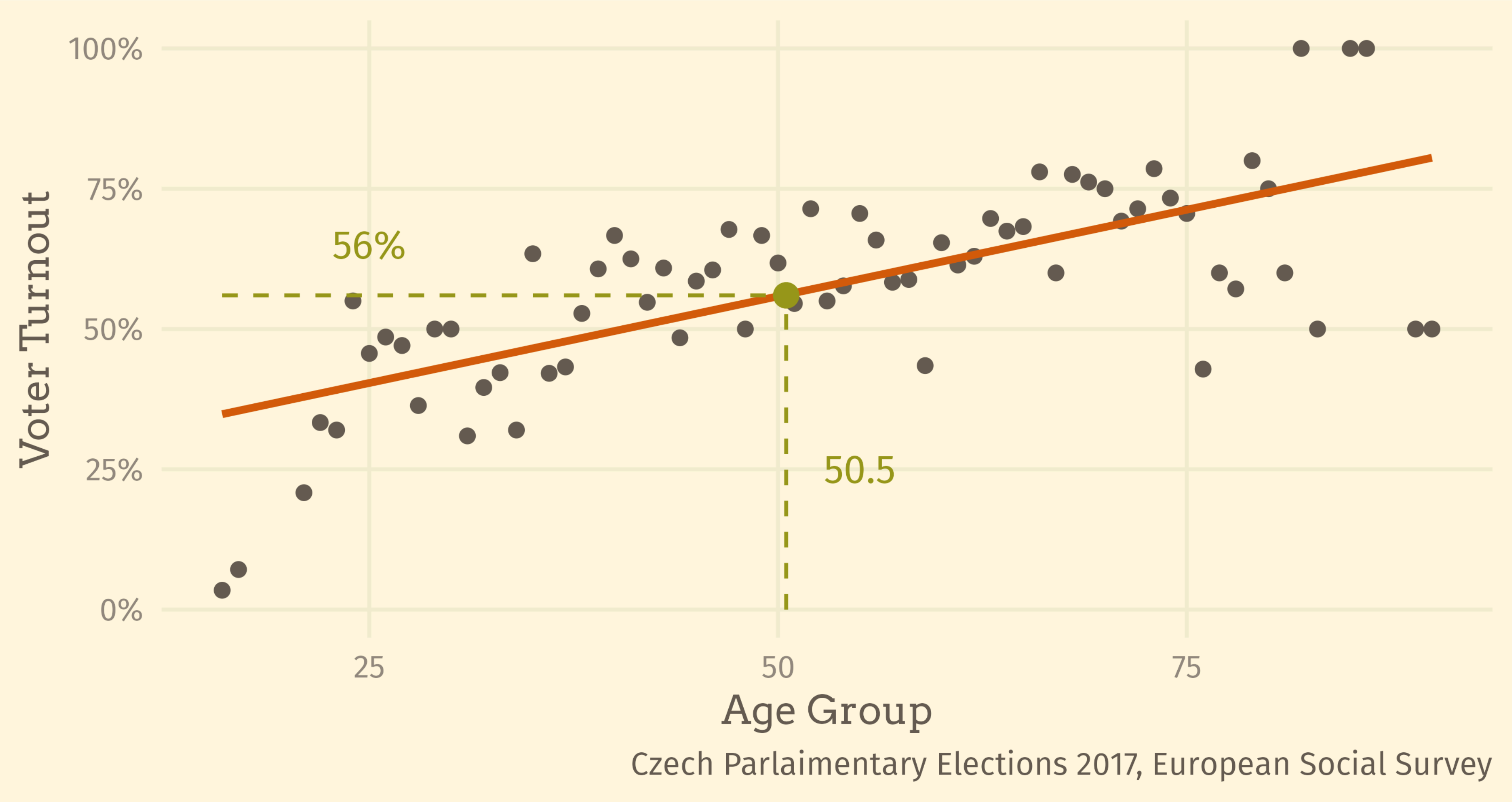
Voter Turnout and Age



Voter Turnout and Age



Voter Turnout and Age



Linear regression is the science of drawing lines

(And curves. And planes. We'll get to that)

Questions?

How to Draw a Line

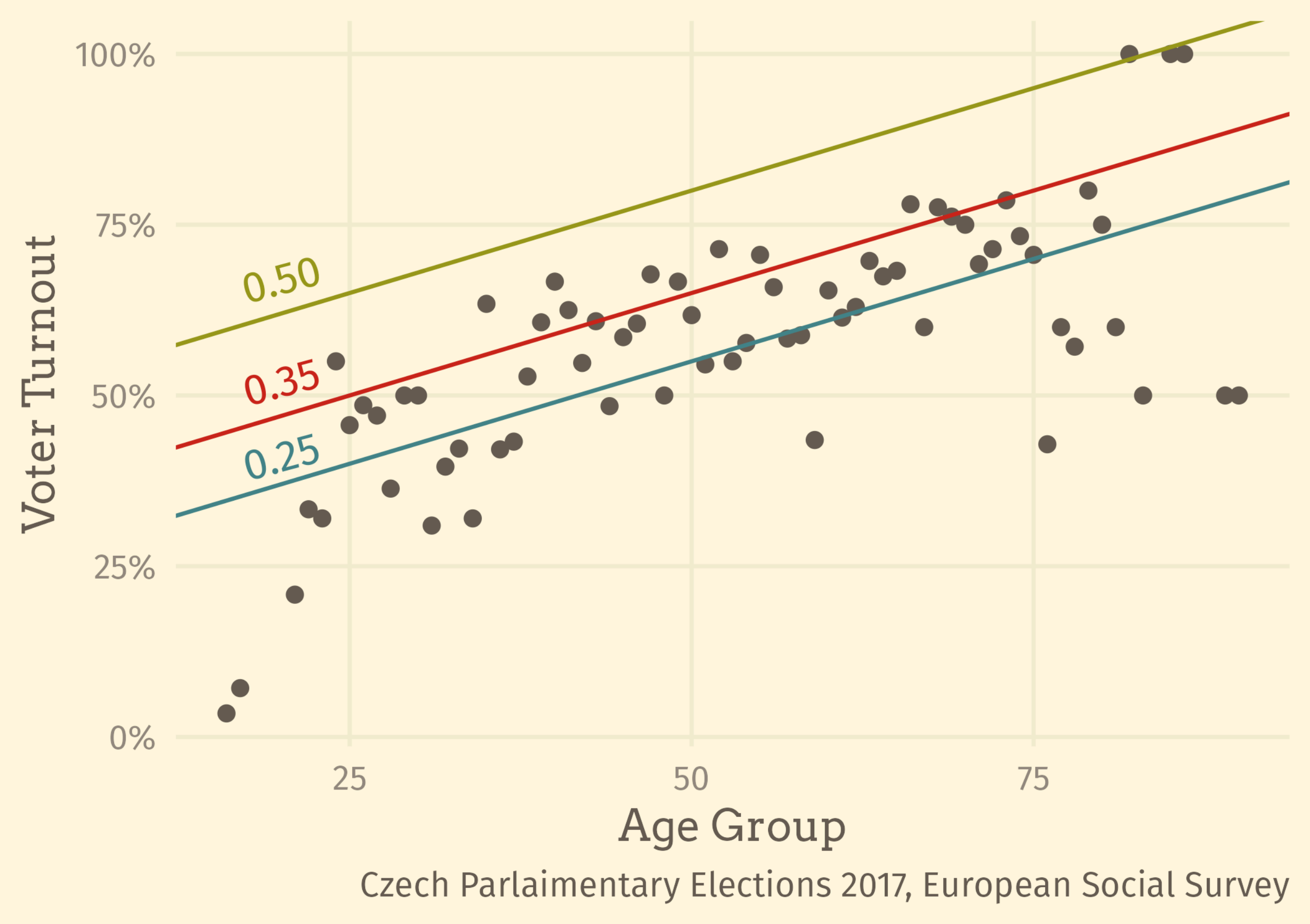
How to Draw a Line

Line is defined using **two parameters** (coefficients, betas, β):

- Intercept (β_0)
- Slope (β_1)

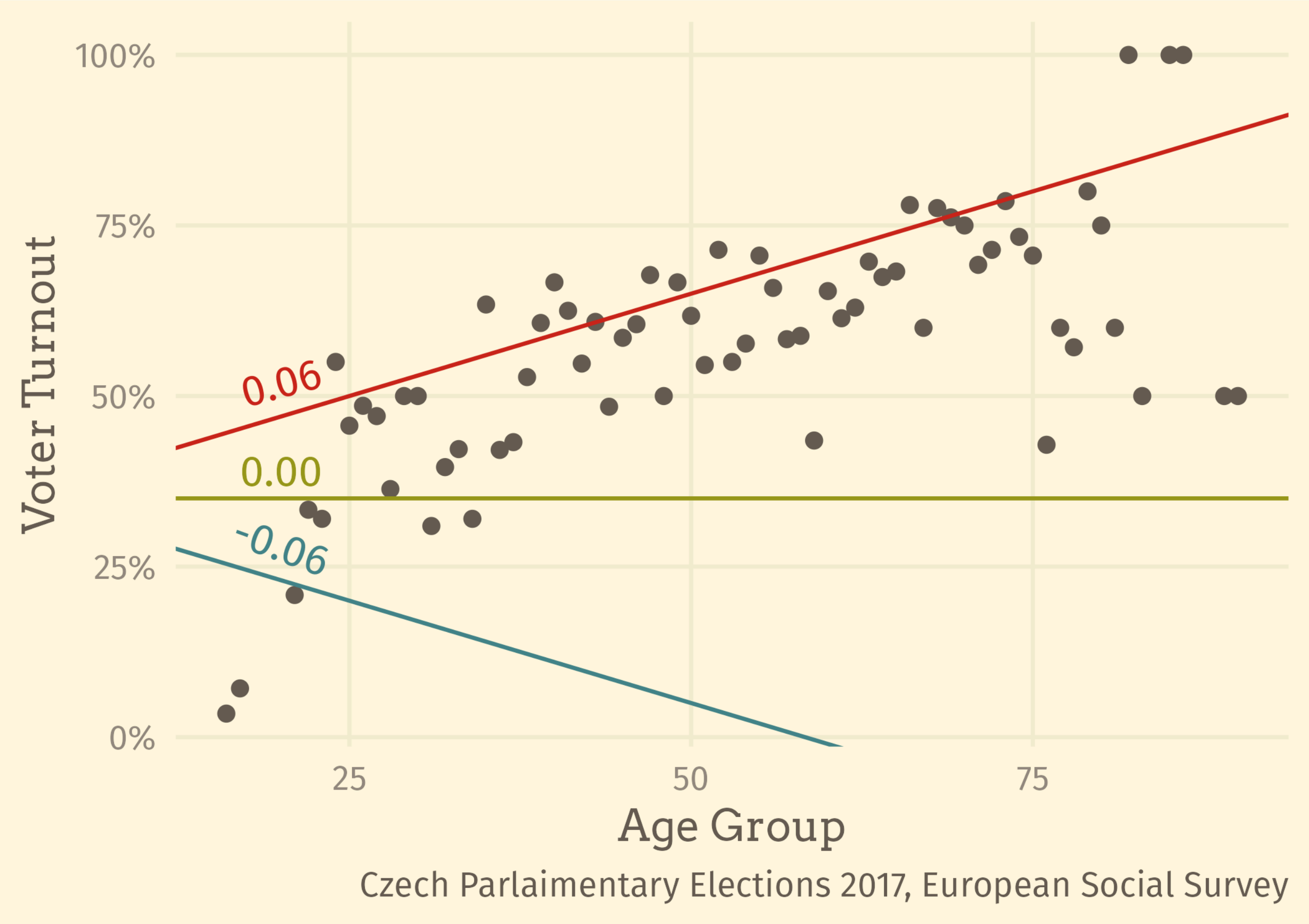
Intercept (β_0)

- **Starting point** of the line.
- The value on the vertical axis, when the horizontal line is zero.
- Moves the line up and down.



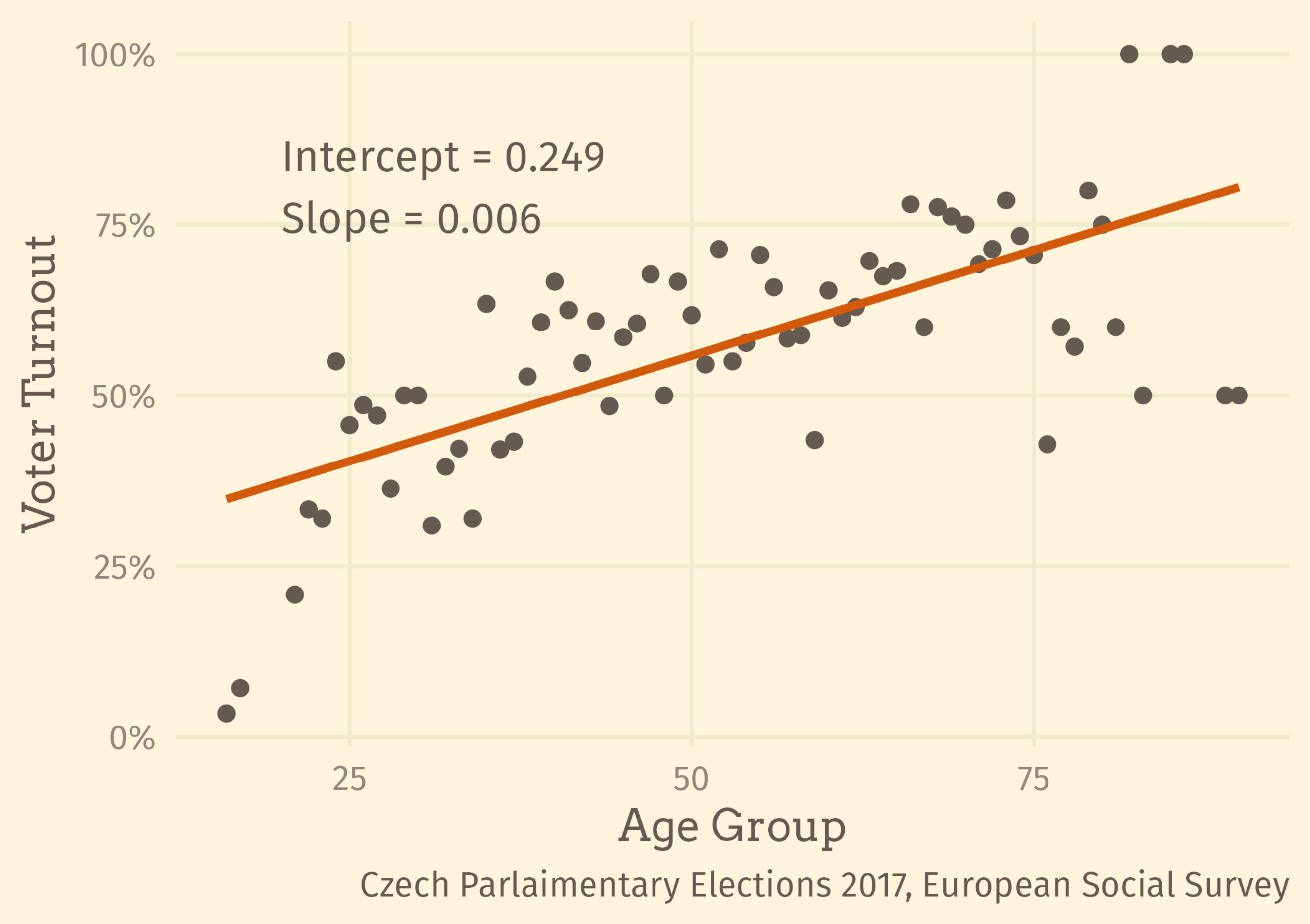
Slopes (β_1)

- **Trend** of the line.
- Change on the vertical axis given 1 unit change on the horizntal one.
- „Aims“ the line.



Intercept and Slope

The goal of linear regression is to find the **combination of intercept and slope** that best fits the data.



Linear Regression Formally

Linear regression formula:

$$\text{Expected value(Turnout)} = 0.249 + 0.006 \cdot \text{age}$$

More generally:

$$E(Y) = \beta_0 + \beta_1 \cdot X$$

Linear Regression Formally

Sometimes you can also see:

$$y_i = \beta_0 + \beta_1 \cdot x_i + \varepsilon_i$$

Where $\beta_0 + \beta_1 \cdot x_i$ is the **expected value** and ε is the **error** term (difference between true and expected value).

Working with the Regression Formula

We can use our regression formula to answer all kinds of questions.

$$\text{Expected value}(\text{Turnout}) = 0.249 + 0.006 \cdot \text{age}$$

1. What is the expected turnout for people who are 50 years old?

Working with the Regression Formula

We can use our regression formula to answer all kinds of questions.

$$\text{Expected value}(\text{Turnout}) = 0.249 + 0.006 \cdot \text{age}$$

1. What is the expected turnout for people who are 50 years old?

$$\text{Expected value}(\text{Turnout}) = 0.249 + 0.006 \cdot 50 = 0.558$$

Working with the Regression Formula

We can use our regression formula to answer all kinds of questions.

$$\text{Expected value}(\text{Turnout}) = 0.249 + 0.006 \cdot \text{age}$$

2. What is the expected difference in voter turnout between people who are 50 and 51 years old?

Working with the Regression Formula

We can use our regression formula to answer all kinds of questions.

$$\text{Expected value}(\text{Turnout}) = 0.249 + 0.006 \cdot \text{age}$$

2. What is the expected difference in voter turnout between people who are 50 and 51 years old?

$$0.006 = 0.6\%$$

Working with the Regression Formula

We can use our regression formula to answer all kinds of questions.

$$\text{Expected value}(\text{Turnout}) = 0.249 + 0.006 \cdot \text{age}$$

3. What is the expected difference in voter turnout between people who are 50 and 60 years old?

Working with the Regression Formula

We can use our regression formula to answer all kinds of questions.

$$\text{Expected value}(\text{Turnout}) = 0.249 + 0.006 \cdot \text{age}$$

3. What is the expected difference in voter turnout between people who are 50 and 60 years old?

$$0.006 \cdot 10 = 6\%$$

Linear regression is defined by intercept and slope

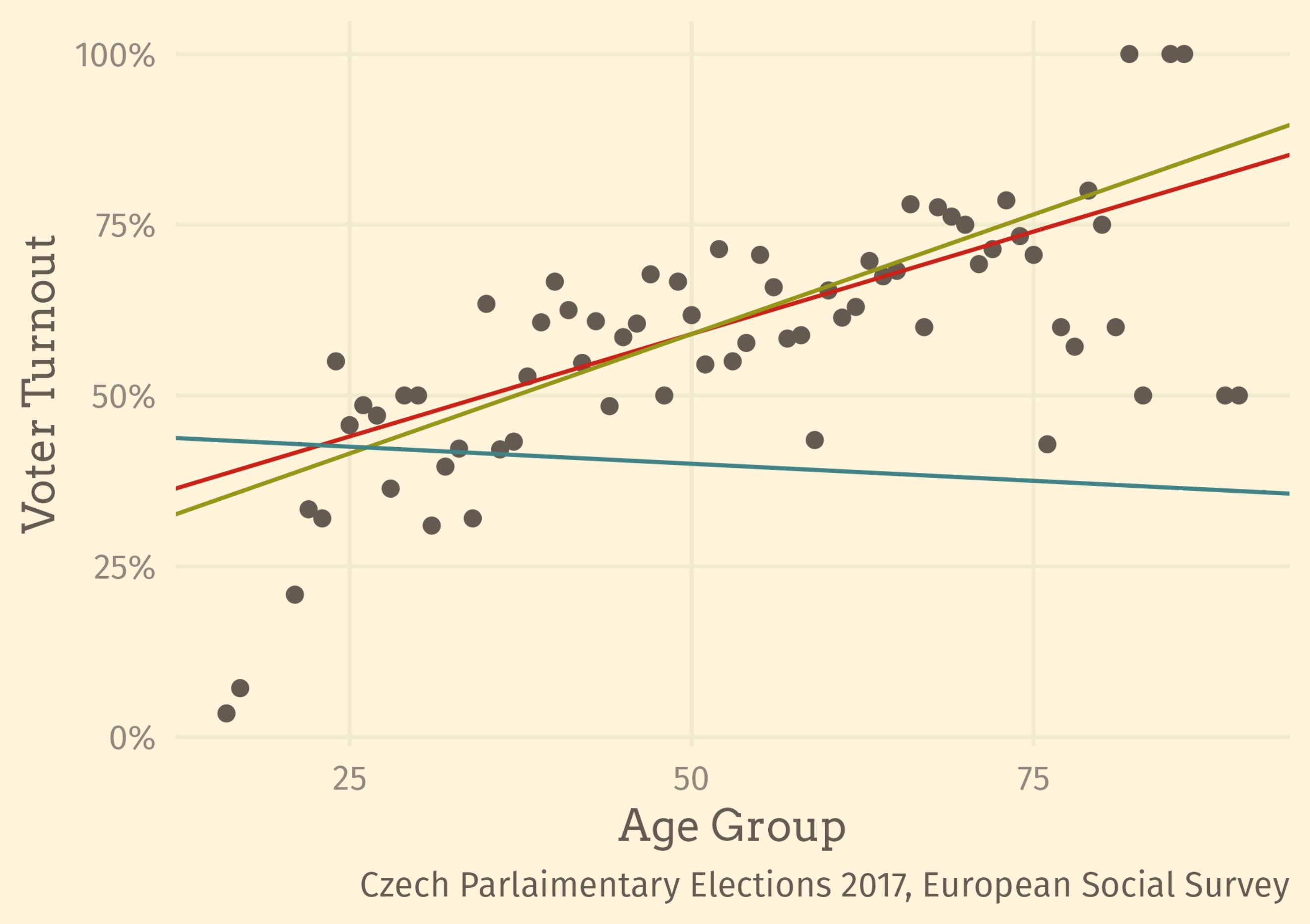
Questions?

Finding the Best Line

Finding the Best Line

We know how to draw and read regression lines.

But how can we **pick the best** one?



Finding the Best Line

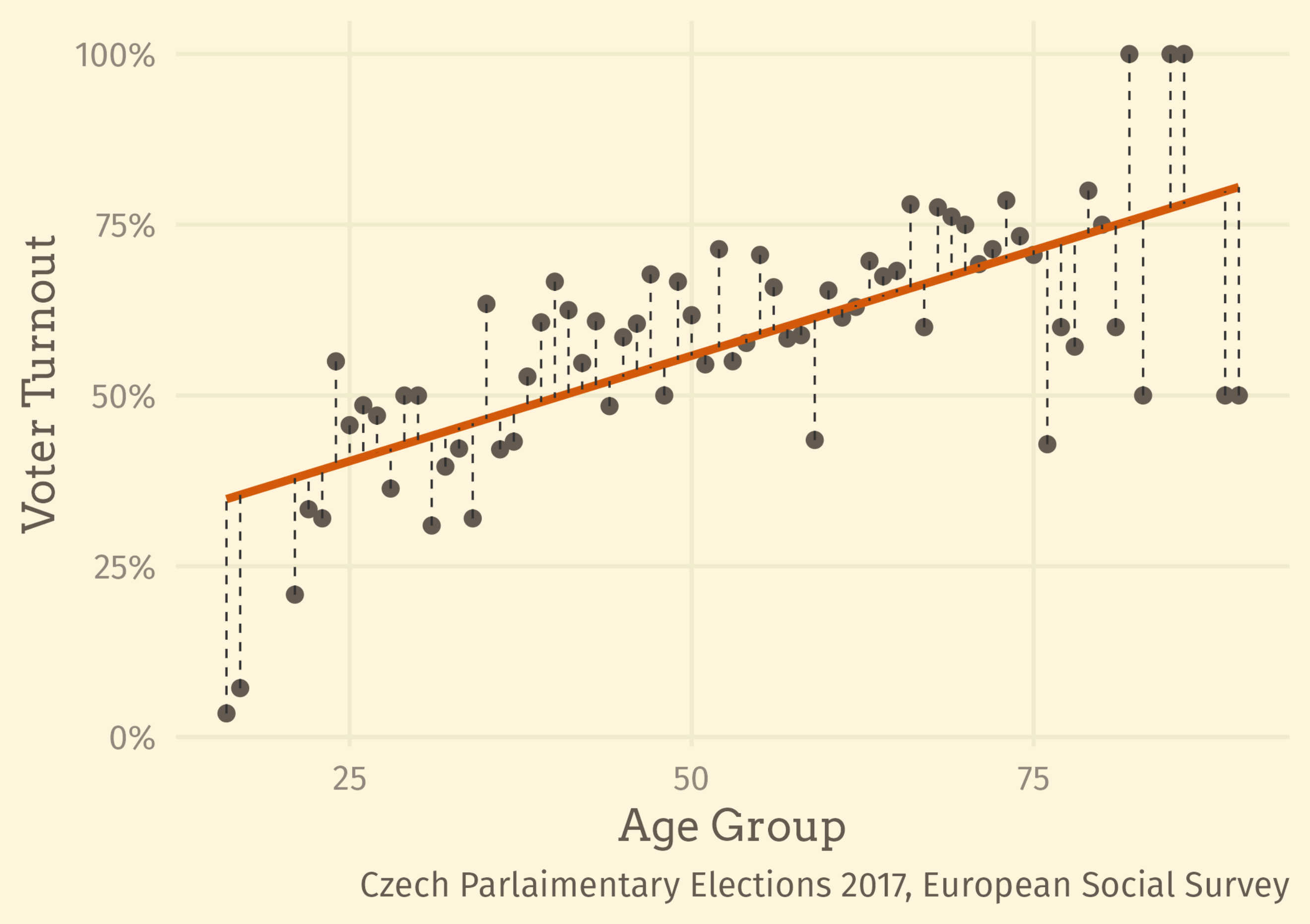
There are many ways to find the „best“ line (many **estimators**).

None is actually the universally best one.

For linear regression, by far the most common is the **Ordinary Least Squares** (OLS) estimator.

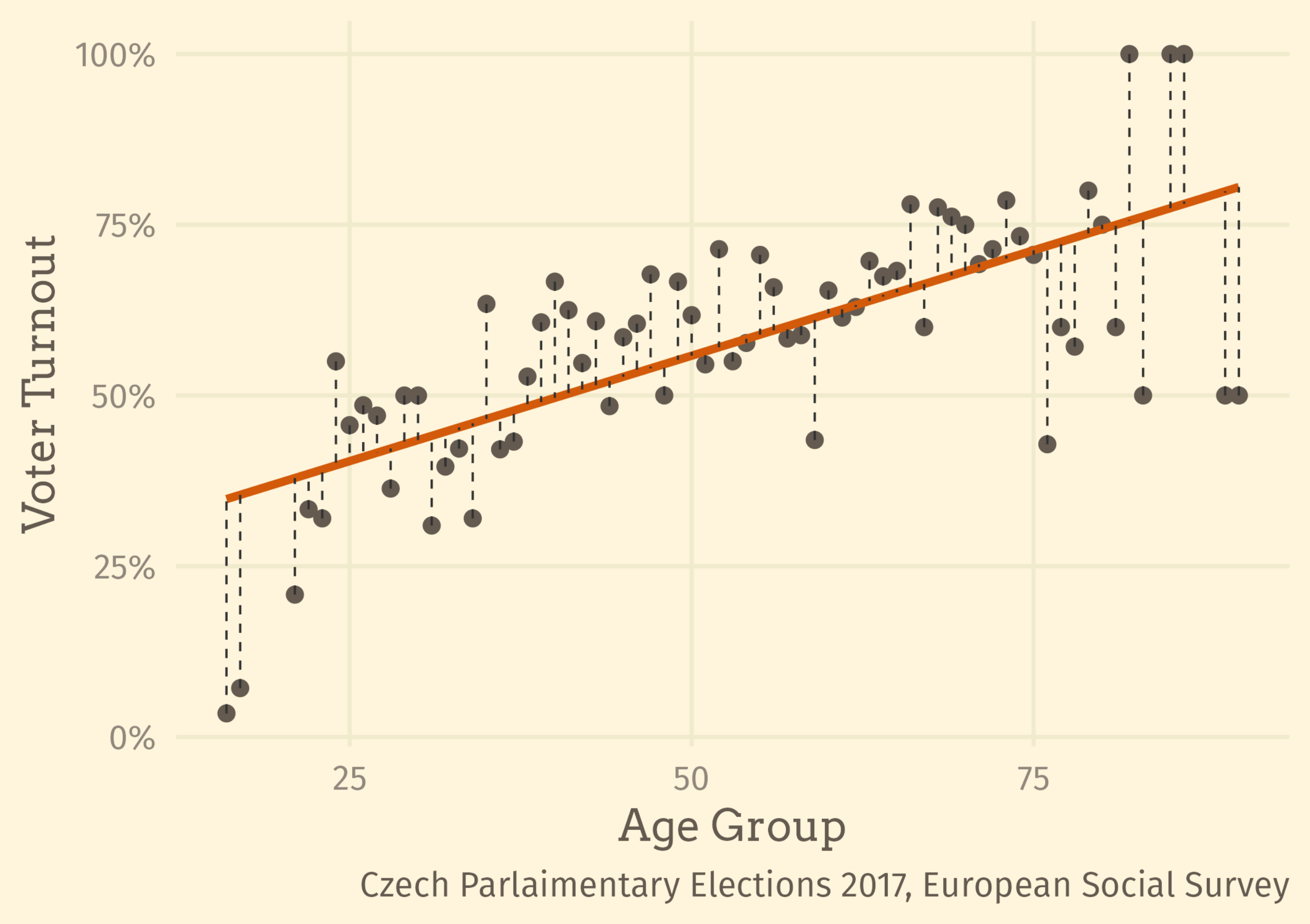
Ordinary Least Squares

1. Compute difference between each point and the regression line (residuals)
2. Square them (residuals^2)
3. Sum the squared residuals



Ordinary Least Squares

The best line is the one that **minimizes** the sum of squared residuals.



Why Minimize Squared Residuals?

Why Minimize Squared Residuals?

1. We are **minimizing the variance** of differences between expected and observed values (minimizing prediction error).

Why Minimize Squared Residuals?

1. We are **minimizing the variance** of differences between expected and observed values (minimizing prediction error).
2. The regression line represents **expected arithmetic mean** of the dependent variable (assuming good fit).

Why Minimize Squared Residuals?

1. We are **minimizing the variance** of differences between expected and observed values (minimizing prediction error).
2. The regression line represents **expected arithmetic mean** of the dependent variable (assuming good fit).
3. (It's **easy to compute**, even without computers).

Other Estimators

There are many other ways to fit regression models

- Maximum Likelihood (logistic regression)
- Weighted Least Squares (survey weights)
- Penalized Least Squares (prediction models, e.g. ridge regression)

We won't be talking about them.

Questions?

Main Takeaways

- Simple linear regression is about **drawing line** that describes relationshop between the dependent and the independent variables.
- Regression line is defined by its **intercept and slope**.
- Values sfor intercept and slope are computed by **Ordinary Least Squares** estimator, which minimizes sum of squared residuals.

InteRmezzo!