Week 8: Linear Regression I

POP88162 Introduction to Quantitative Research Methods

Tom Paskhalis

Department of Political Science, Trinity College Dublin

Topics for Today

- Slope of the regression line
- Linear model
- Bivariate linear regression model
- Ordinary least squared method
- Testing regression coefficients

Today's Plan

Regression Line → Regression Model → Estimation → Statistical Inference

Previously…

Review: Significance Test

- **Significance test** is used to summarize the evidence about a hypothesis.
- It does so by comparing the our estimates with those predicted by a null hypothesis.
- 5 components of a significance test:
  - **Assumptions**: scale of measurement, randomization, population distribution, sample size
  - **Hypotheses**: null H_0 and alternative H_a hypothesis
  - **Test statistic**: compares estimate to those under H_0
  - **P-value**: weight of evidence against H_0, smaller P indicate stronger evidence
  - **Conclusion**: decision to *reject* or *fail to reject* H_0.

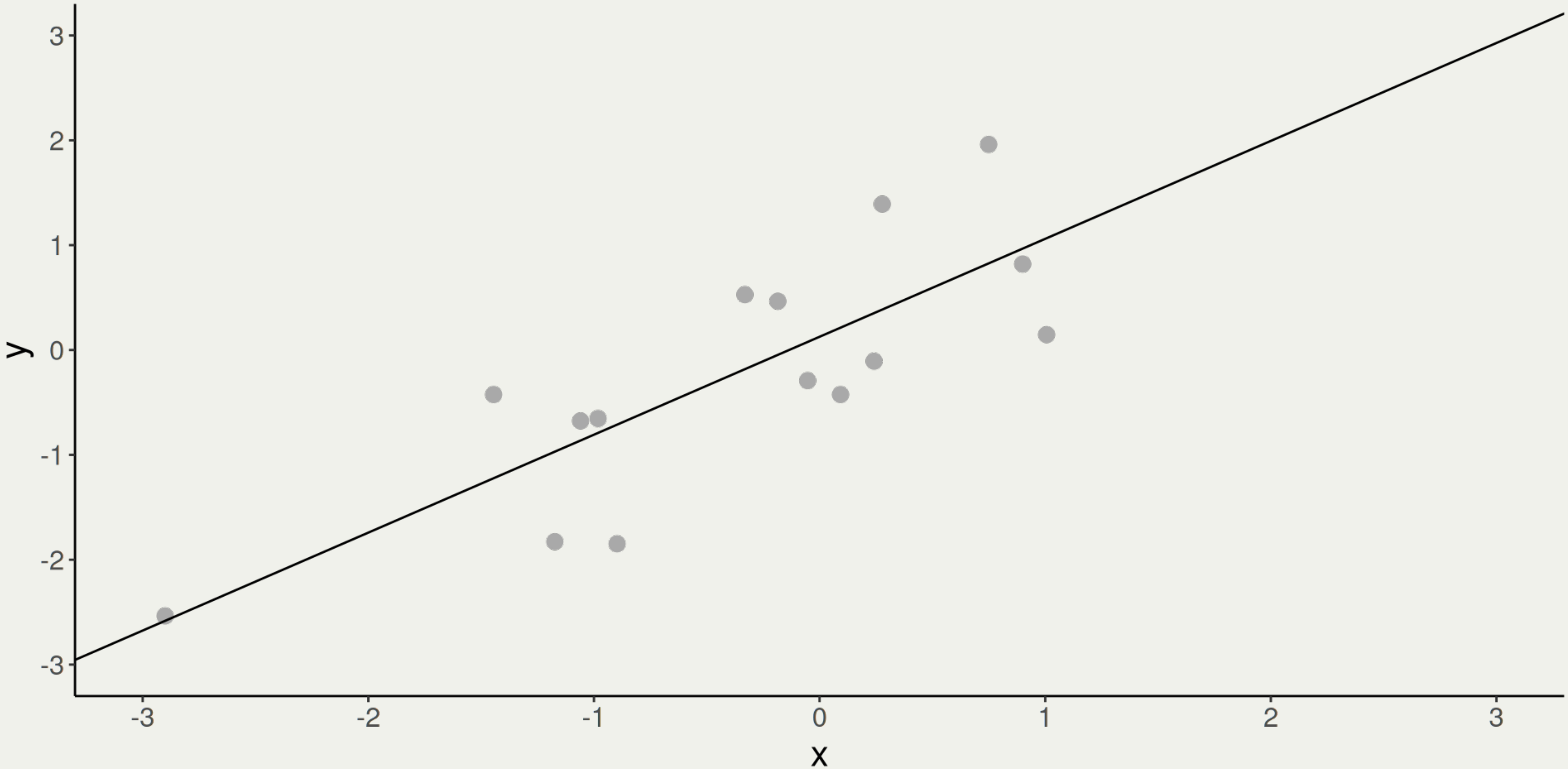
Review: Statistical Tests

		Dependent Variable	
		Nominal/ Ordinal	**Interval**
Independent Variable	**Nominal/ Ordinal**	χ^2 (chi-squared) test	Mean comparison test
	Interval	Logistic Regression	Linear Regression

Regression Line

Using a Line to Predict

Using a Line to Predict



Linear Relationship

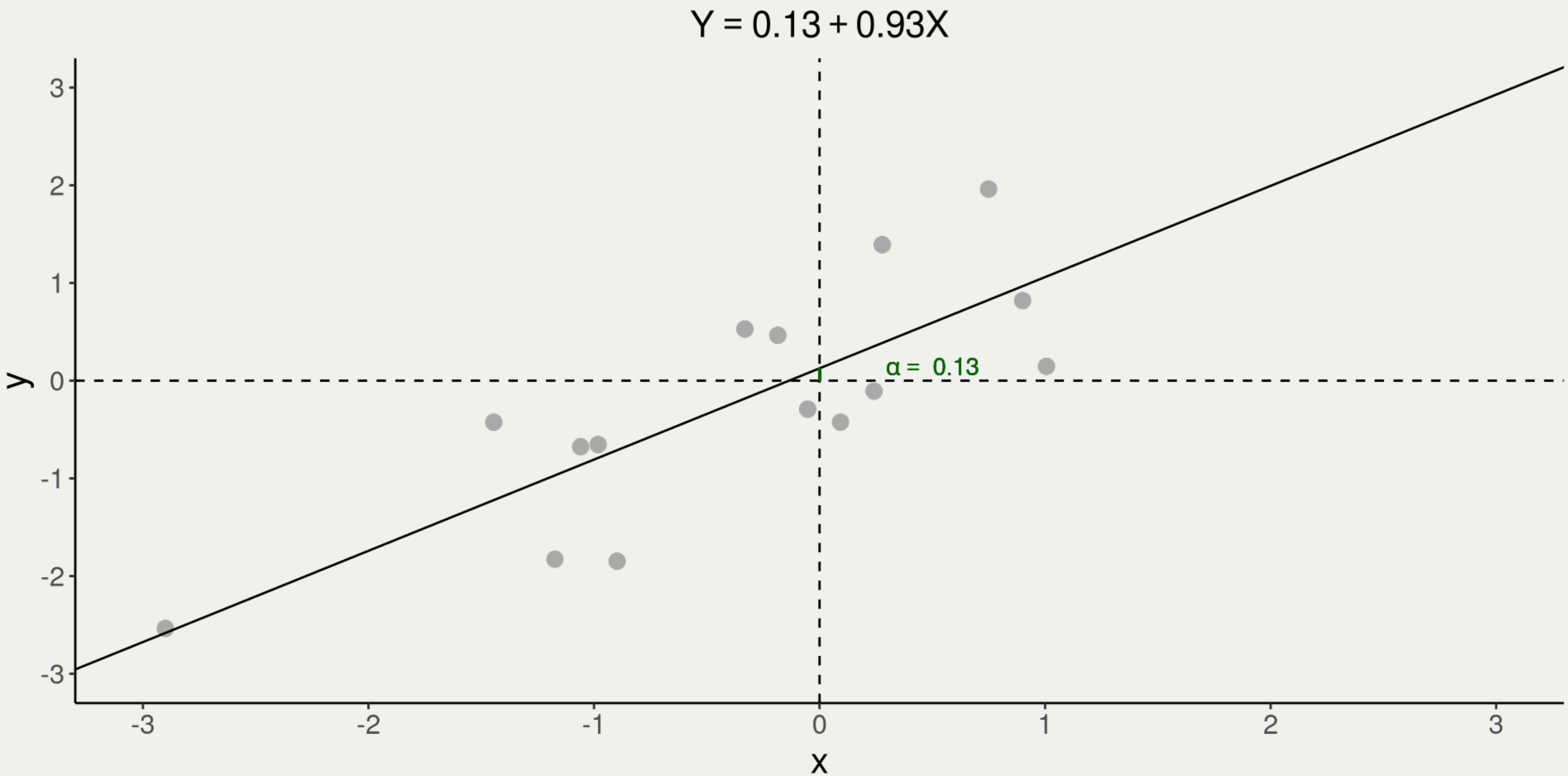
- The simplest way to describe the relationship between two continuous variables is with a line.
- A straight line can be represented by this formula:

$$Y = \alpha + \beta X$$

- α is the **intercept**, the expected value of Y when $X = 0$
- β is the **slope**, the change in expected value of Y when X increases by one unit.

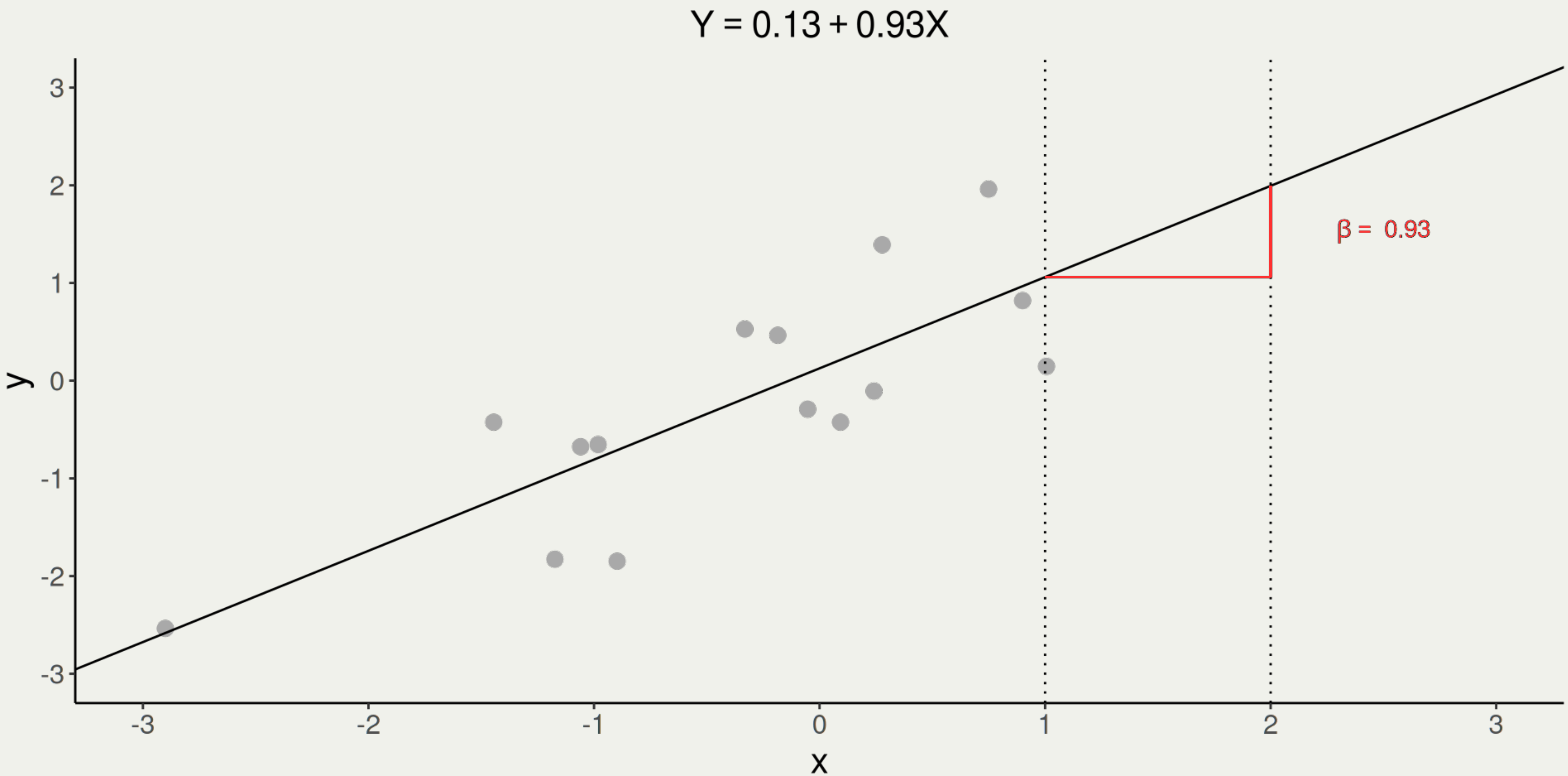
Intercept

Y takes the value of 0.13 when $X = 0$.



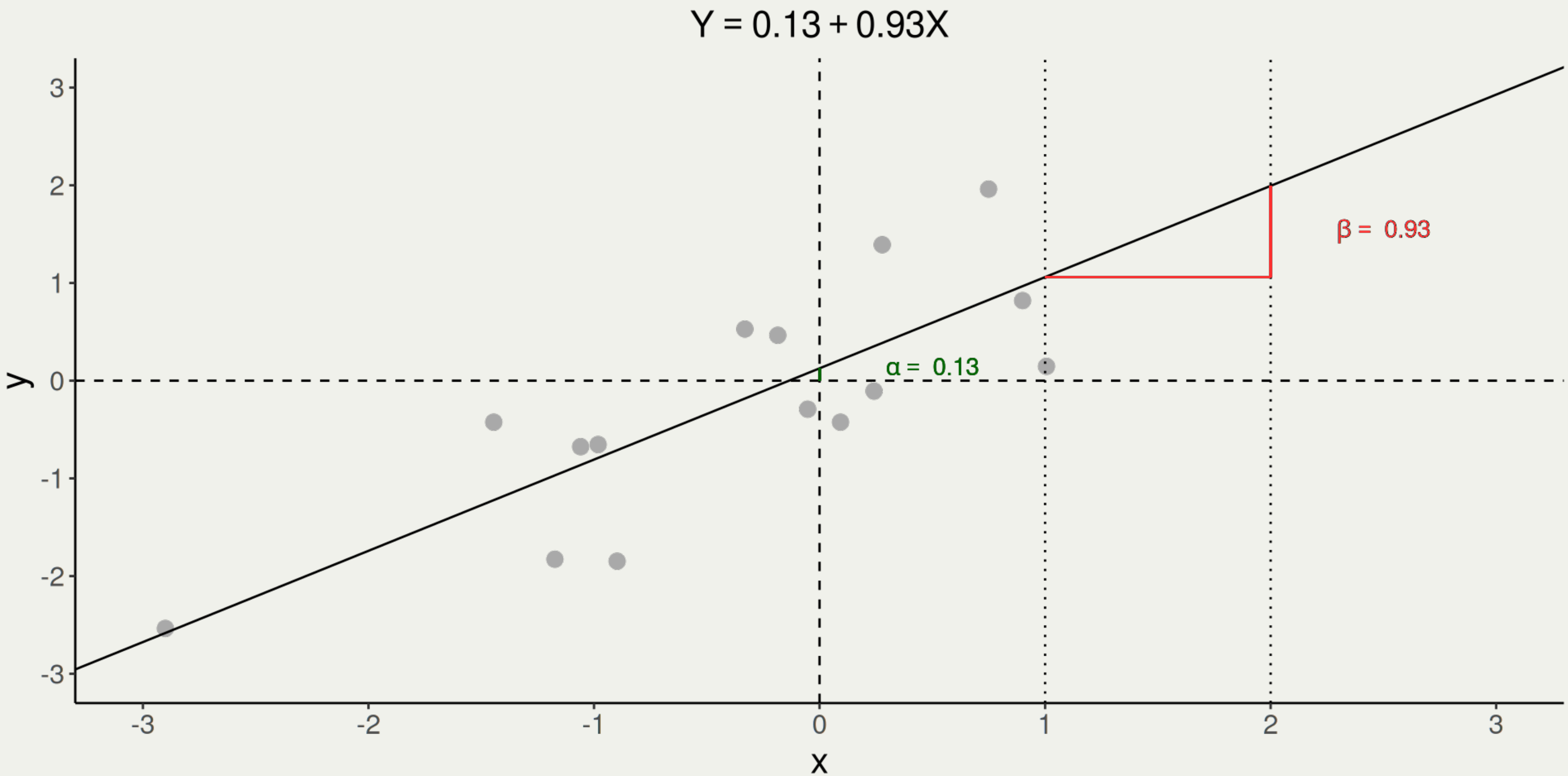
Slope

A one unit increase in X is associated, on average, with a 0.93 increase in Y.

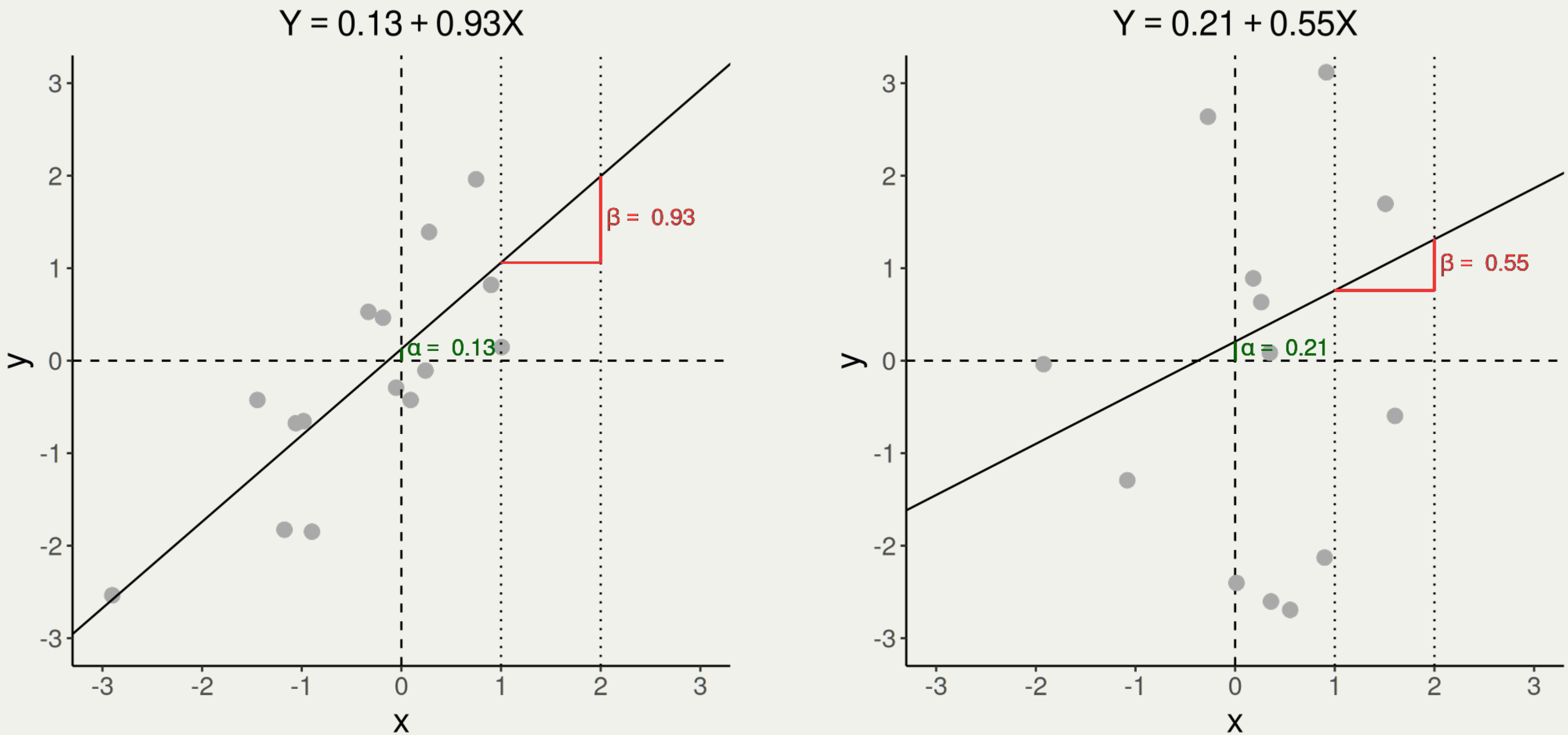


Regression Line

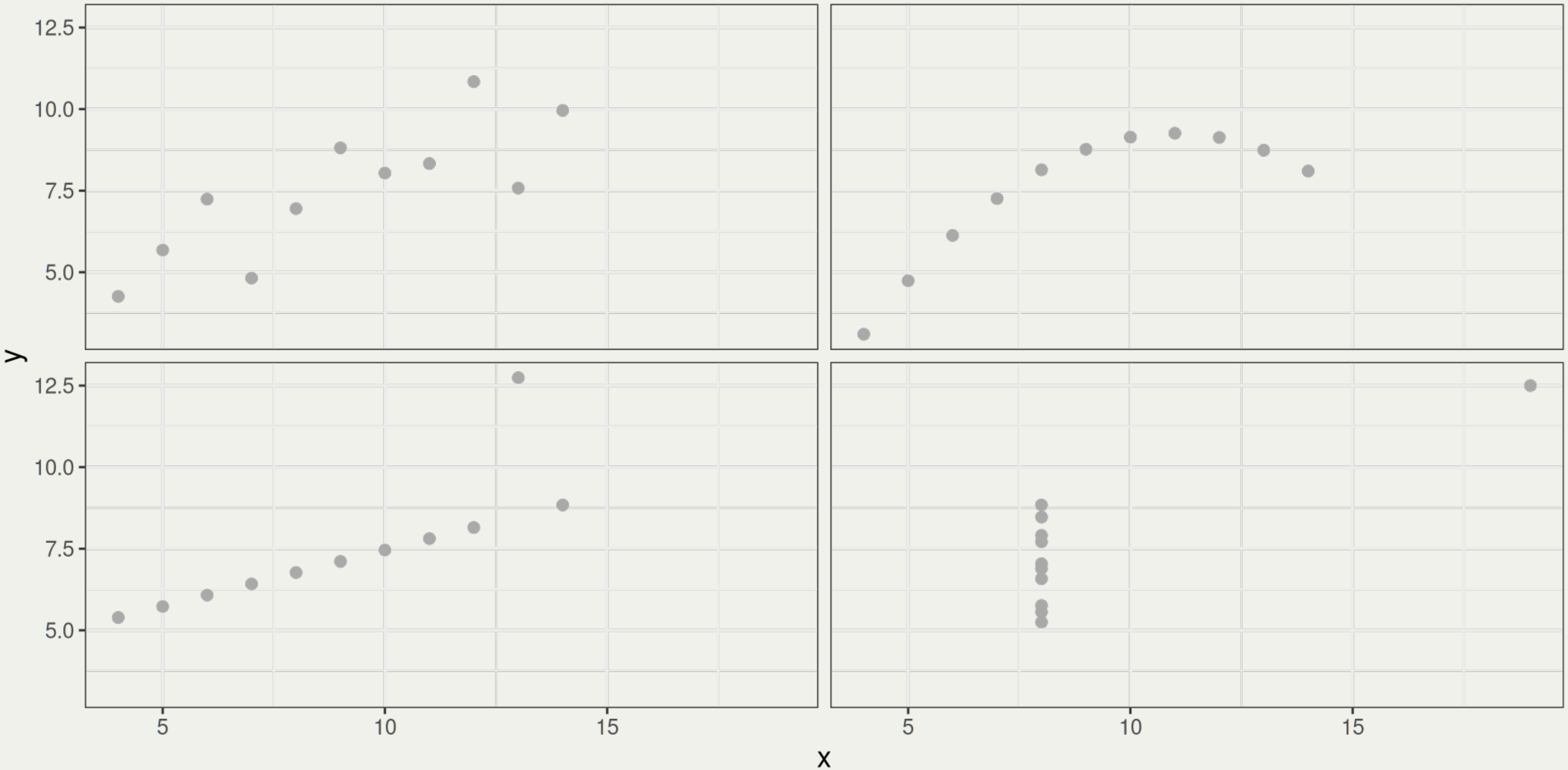
The value of Y can be calculated as 0.13 plus 0.93 times the value of X.



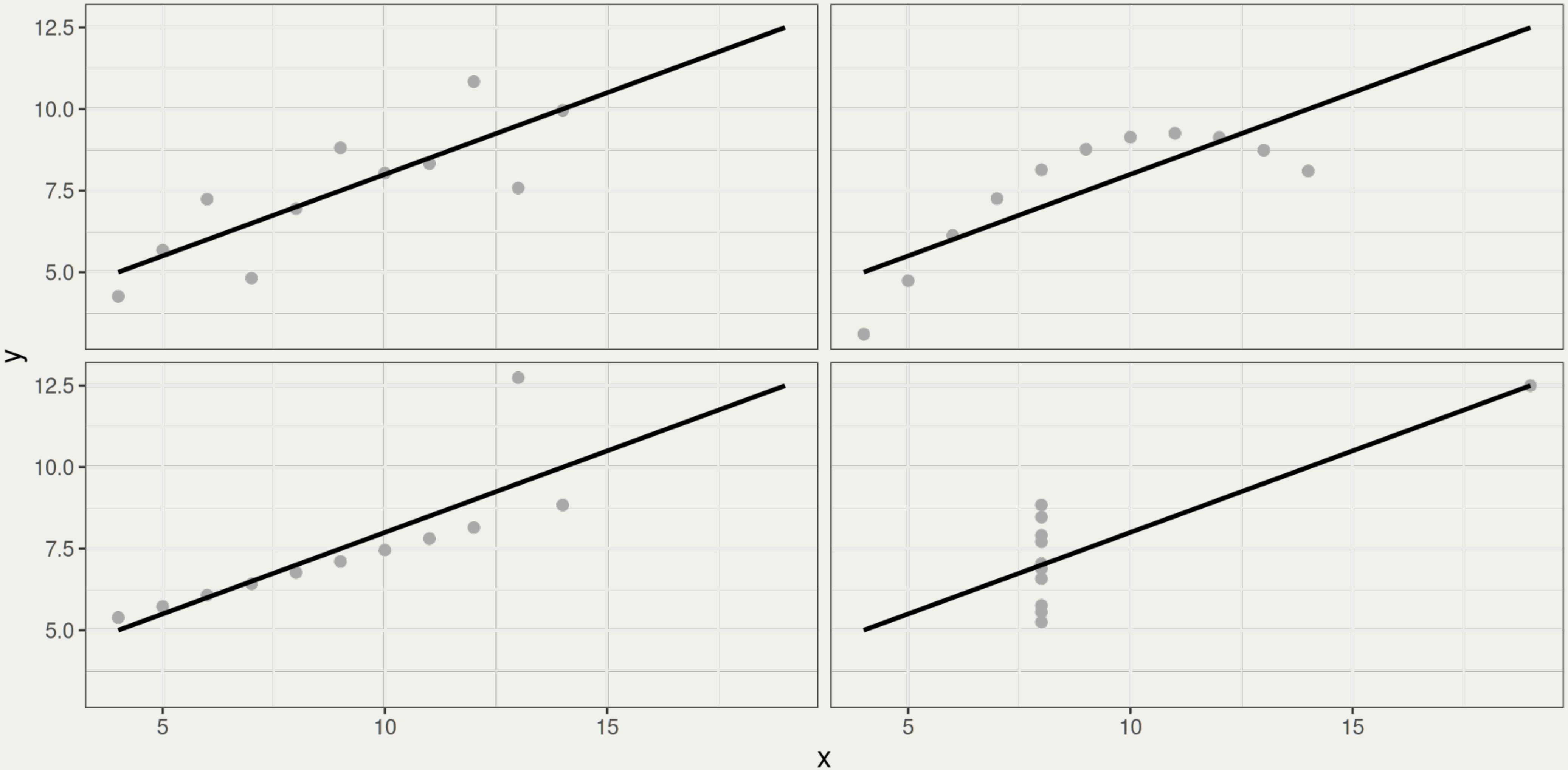
Varieties of Linear Relationships



Anscombe's Quartet



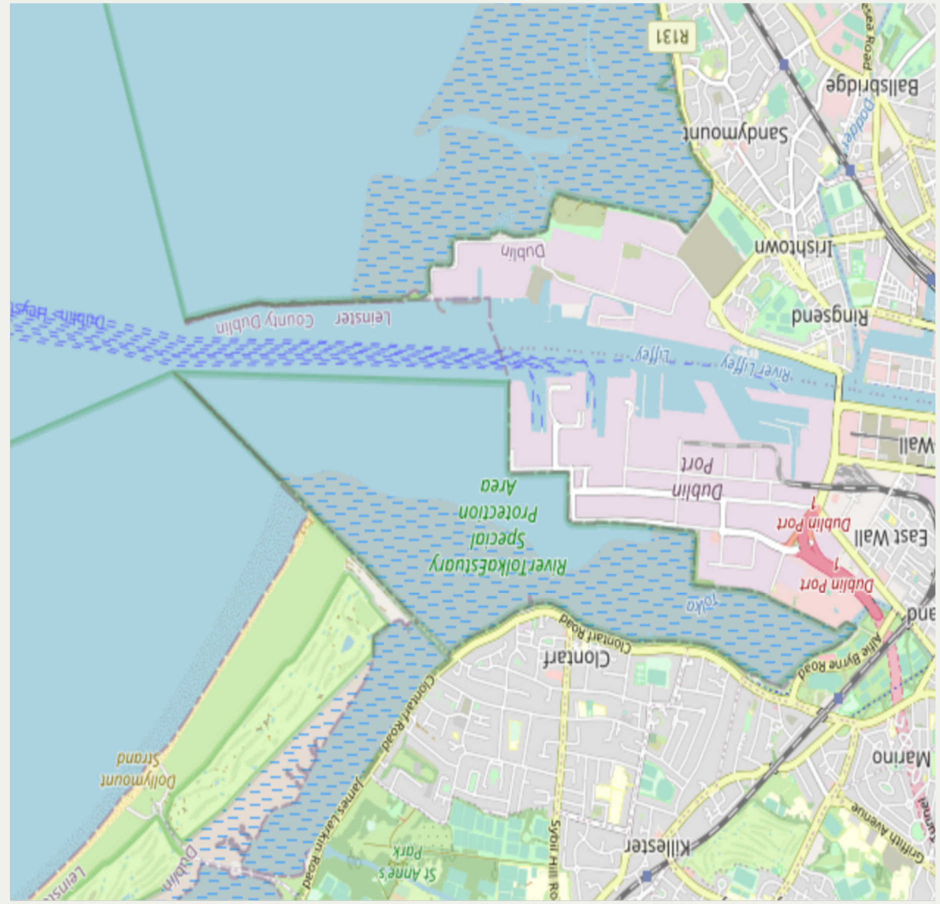
Linearity Assumption



Linear Regression Model

Model

A *simplified* description of an object.



Statistical Model

A *simplified* description of relationships between variables.

Winning Election = Party + Incumbency + Campaign Spending

> All models are wrong, but some are useful.
>
> George Box

Linear Regression Model

- We can express linear relationship (association) between two continuous variables with **bivariate linear regression model**:

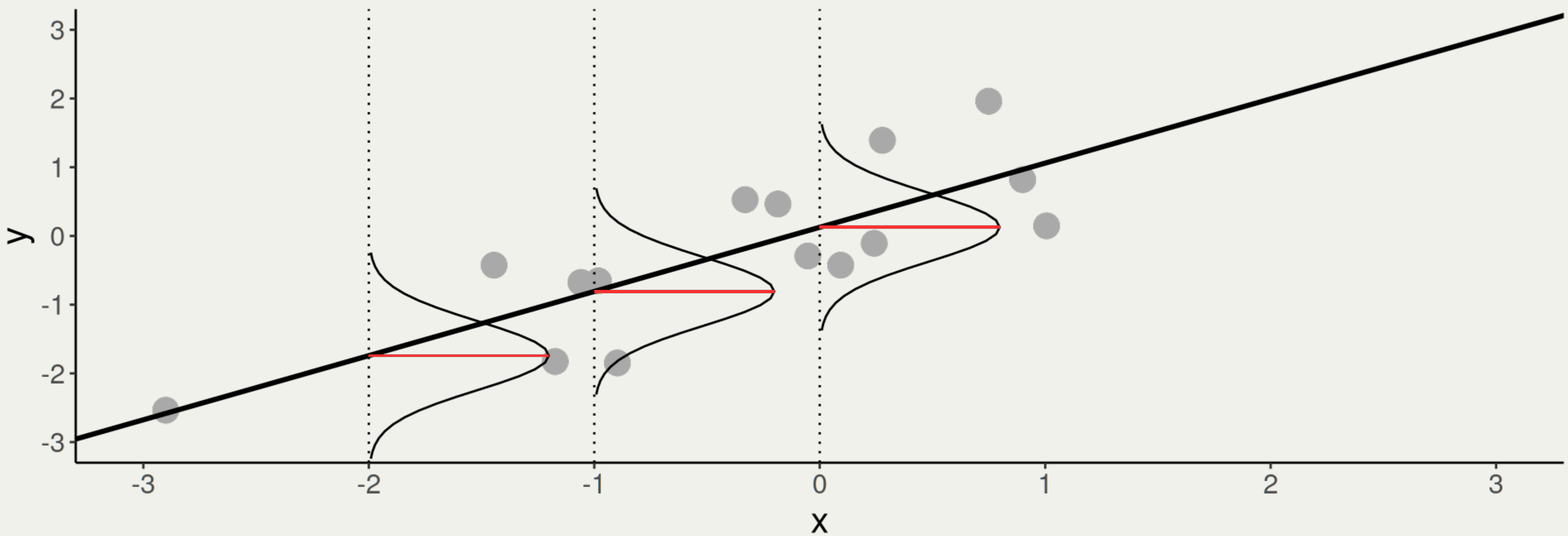
$$Y_i = \alpha + \beta X_i + \epsilon_i$$

where:

- Observations $i = 1, \ldots, n$
- Y is the dependent variable
- X is the independent variable
- α is the **intercept** or **constant**
- β is the **slope**
- ϵ_i is the **error term**

Error Term

- The relationship between real world variables is (almost) never *deterministic*.
- Note that other than indexing, the key difference between an equation describing a line and an equation describing linear regression model is ϵ (pronounced epsilon) or **error term**.
- Error term is assumed to be normally distributed and has a mean of 0 and variance σ^2.



Parameters of Regression

- Standard bivariate regression model:

$$Y_i = \alpha + \beta X_i + \epsilon_i$$

has three population parameters:

1. **intercept** α
2. **slope** β
3. **error variance** σ_ϵ^2

- α and β can both be referred to as *regression coefficients*.
- In essence, α and β describe the best straight line to summarise the association, and σ_ϵ^2 describes the variation of the data around that line.

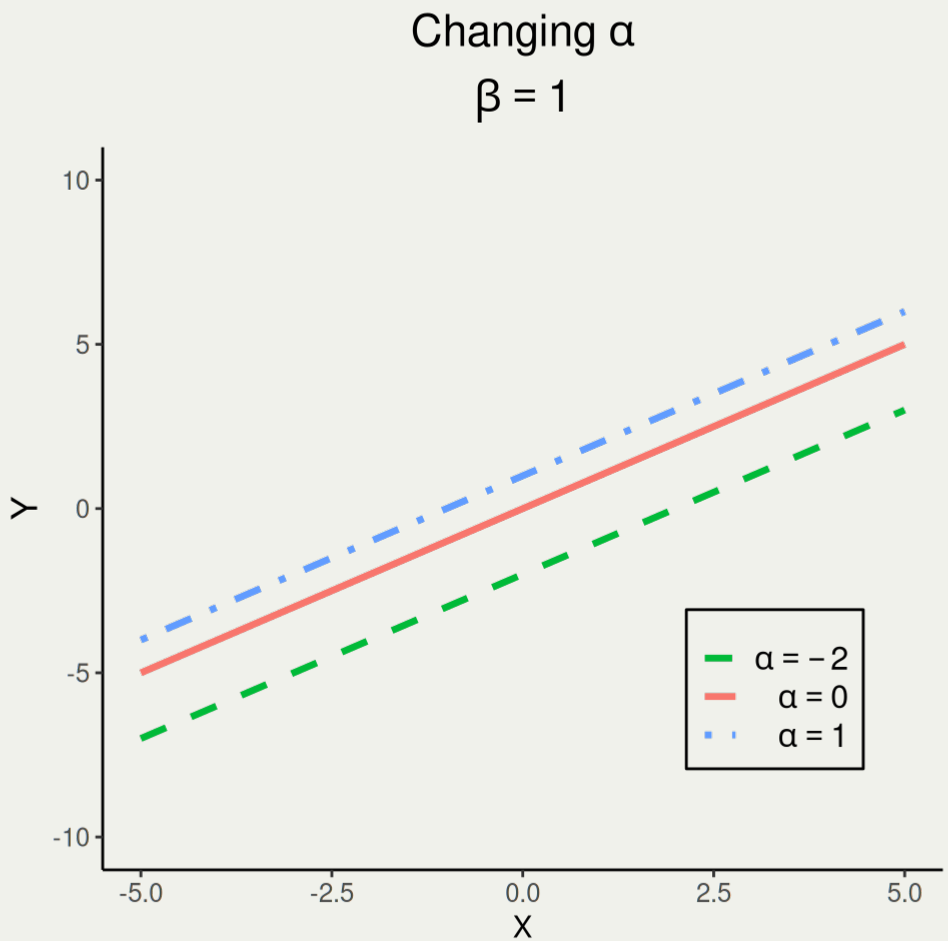
Parameter Estimates of Regression

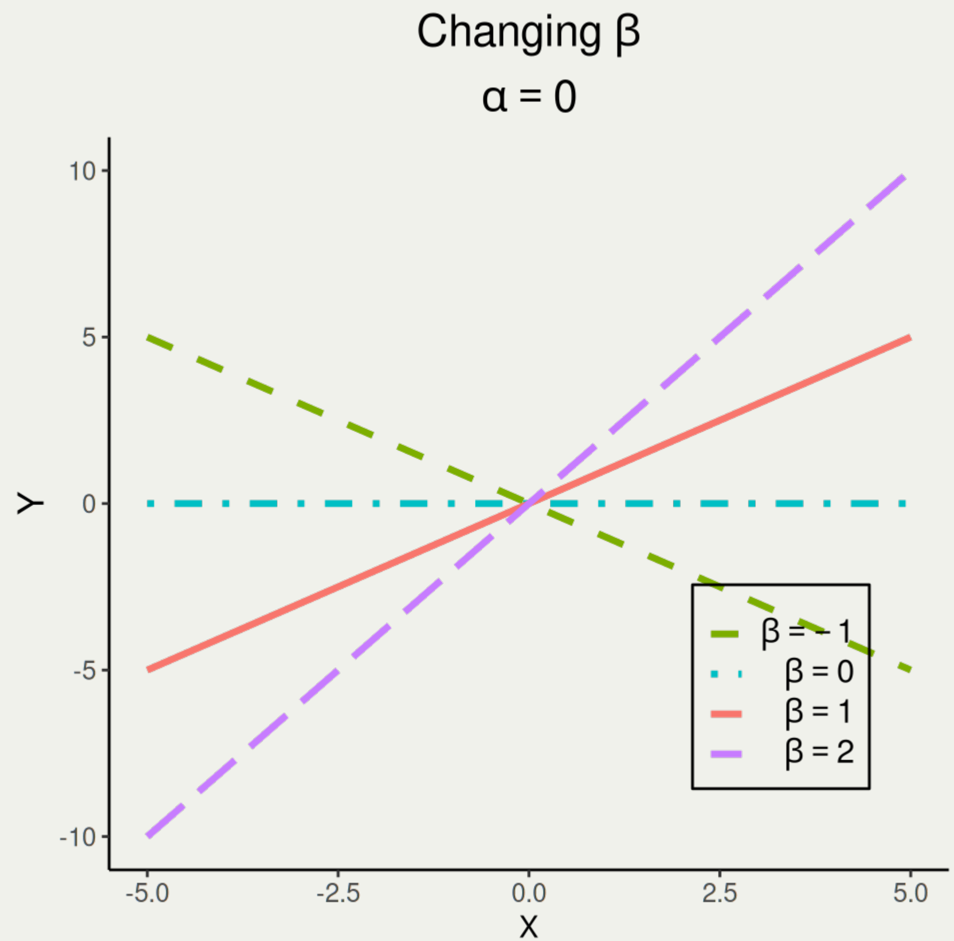
- We would like to know the population parameters.
- But (as usually) we do not have access to the entire population.
- Thus, we must *estimate* parameters of linear regression model from a sample.
- We can denote estimated linear regression model as:

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$$

- Population vs data:
  - $\alpha, \beta, \sigma^2 \rightarrow$ population parameter values;
  - $\hat{\alpha}, \hat{\beta}, \widehat{\sigma_\epsilon}^2 \rightarrow$ estimated parameter values.

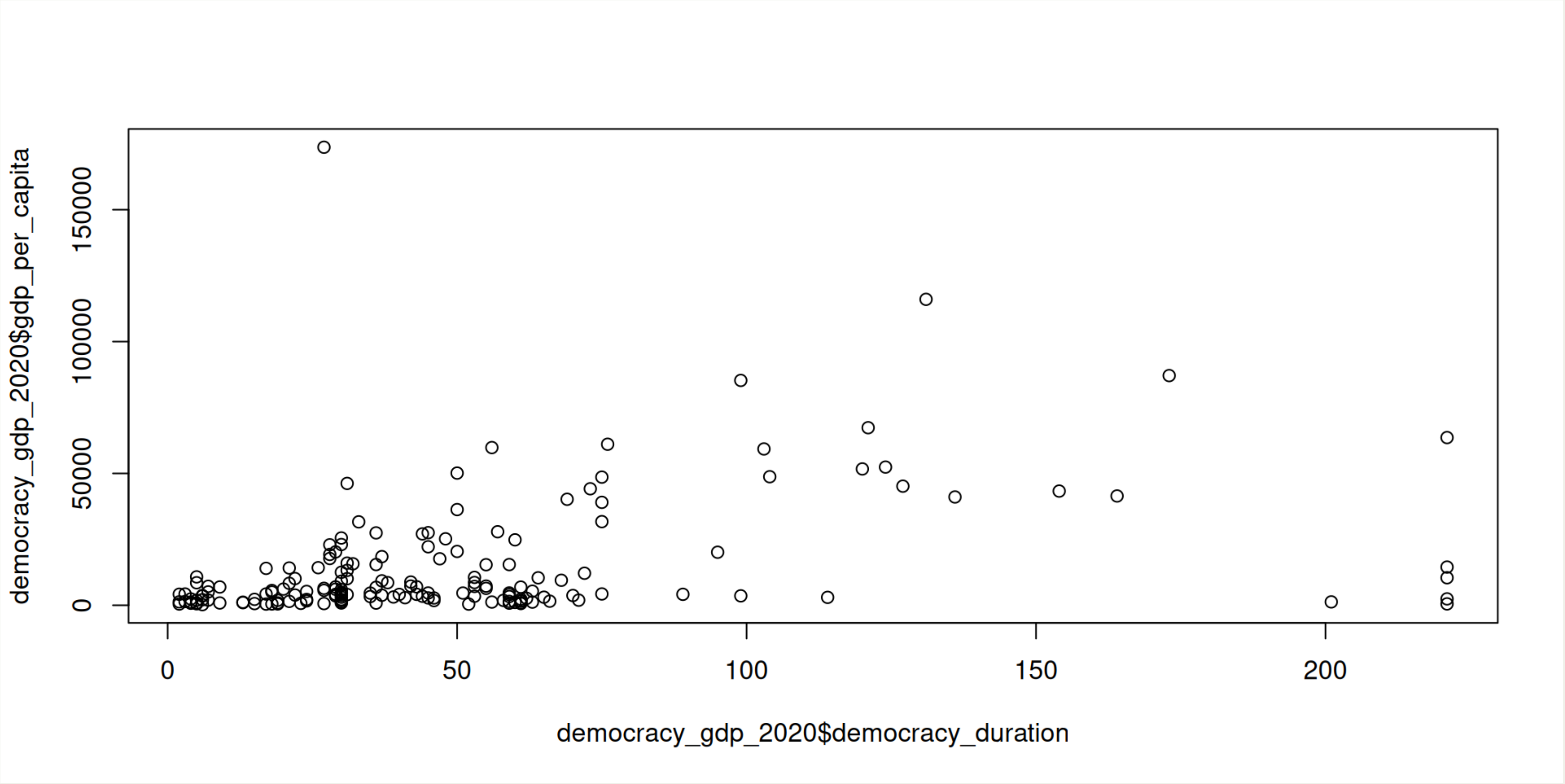
Varying Parameters





Example: Regime Longevity and GDP in 2020

```r
democracy_gdp_2020 <- read.csv("../data/democracy_gdp_2020.csv")
plot(democracy_gdp_2020$democracy_duration, democracy_gdp_2020$gdp_per_capita)
```



Example: Linear Regression Model

- In this example the population regression model would be:

$$GDP_i = \alpha + \beta Longevity_i + \epsilon_i$$

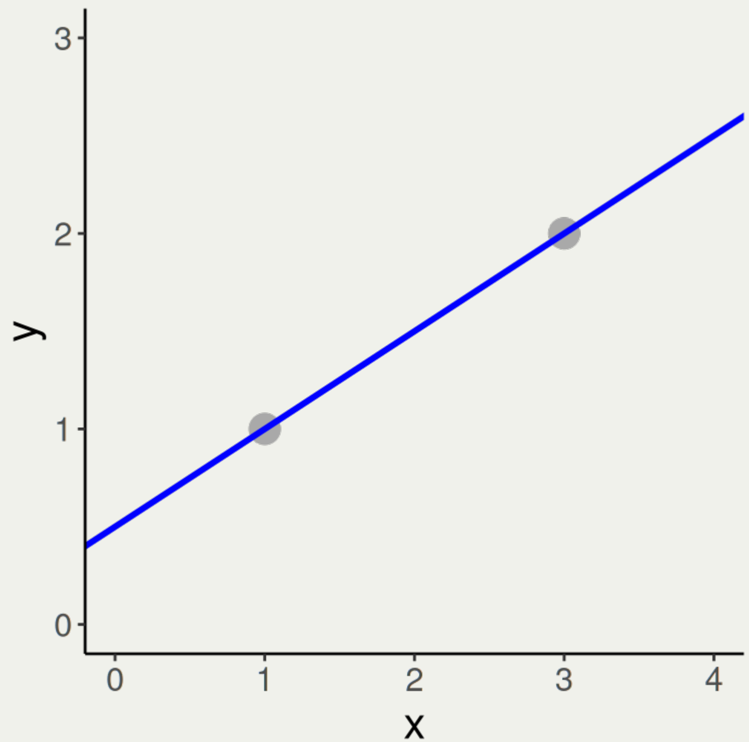
- As much as we would like to know the true population parameters, we are only able to calculate their estimates.
- Thus, our estimated model is:

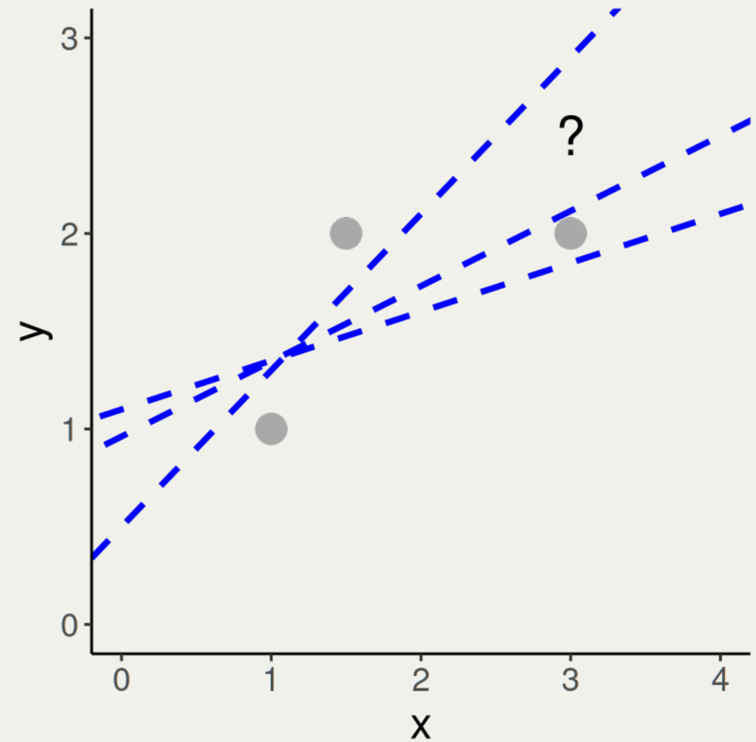
$$\widehat{GDP}_i = \hat{\alpha} + \hat{\beta} Longevity_i$$

Estimation of Regression Model

Drawing a Line

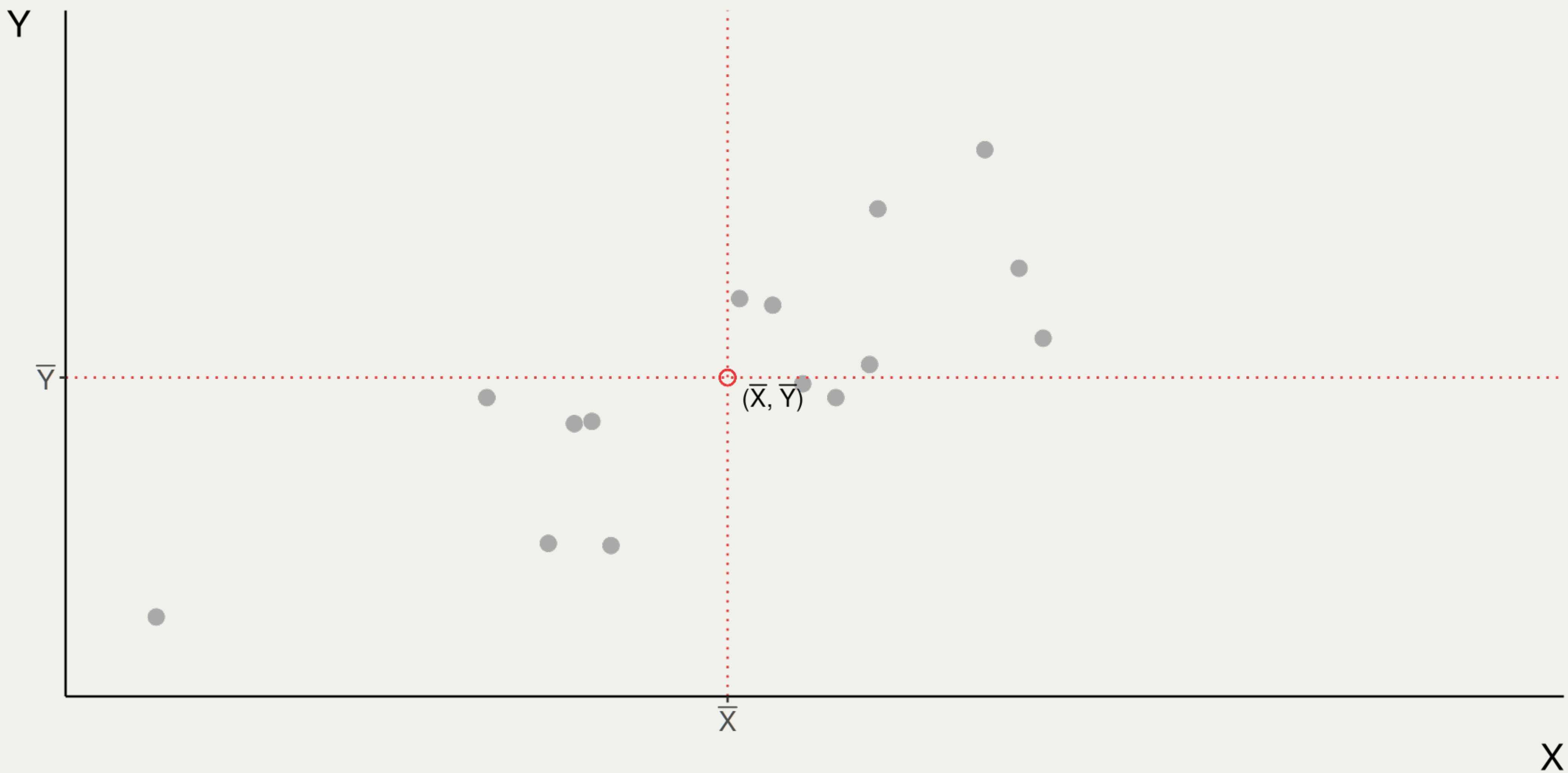
- We know from basic school geometry:
  - How to draw a straight line through two points
  - But how do we draw a line through more points?





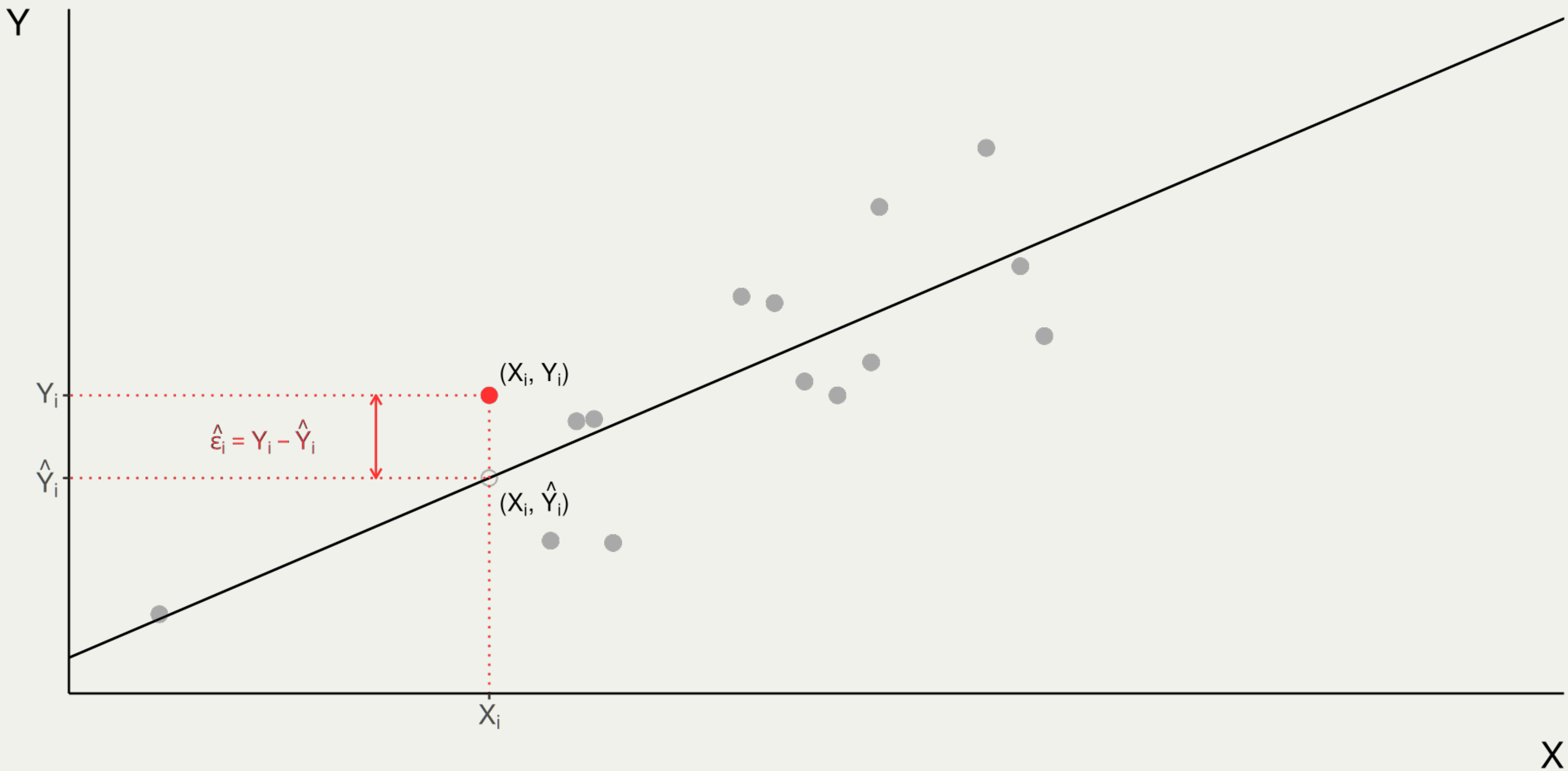
Pivotal Point

There are infinitely many lines that go through $(\bar{X}, \bar{Y})$.



Residual (Error)

Residual $\hat{\epsilon}_i$ is the difference (vertical distance) between *observed* Y_i and *predicted* $\hat{Y}_i$.



Residuals

- Residual $\hat{\epsilon}_i$ is just one error for an i-th observation.
- To calculate the size of overall error we could sum them up:

$$\sum_{i=1}^{n} \hat{\epsilon}_i = \sum_{i=1}^{n} Y_i - \hat{Y}_i$$

- But since residuals cancel each other out, any line going through $(\bar{X}, \bar{Y})$ has:

$$\sum_{i=1}^{n} \hat{\epsilon}_i = 0$$

Residuals Continued

- Recall our discussion of variance calculation.
- We have two solutions to this problem:
  - Summing absolute values of residuals:

$$\sum_{i=1}^{n} |\hat{\epsilon}_i| = \sum_{i=1}^{n} |Y_i - \hat{Y}_i|$$

  - Summing squared residuals:

$$\sum_{i=1}^{n} \hat{\epsilon}_i^{\,2} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

- As with variance, for technical reasons squared residuals are easier to work with.

Ordinary Least Squares (OLS)

- The most common method of estimating parameters of the linear regression model is the **ordinary least squares (OLS)** method.
- The line that best fits the data has the smallest **sum of squared errors (SSE)**.
- More formally a line that minimises the following expression is chosen:

$$SSE = \sum_{i=1}^{n} \hat{\epsilon}_i^{\,2} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - (\hat{\alpha} + \hat{\beta} X_i))^2$$

- SSE is also often called the **residual sum of squares (RSS)**.

OLS Continued

- We will use R to estimate the parameters using OLS method.
- But it can also be calculated using these formulas:

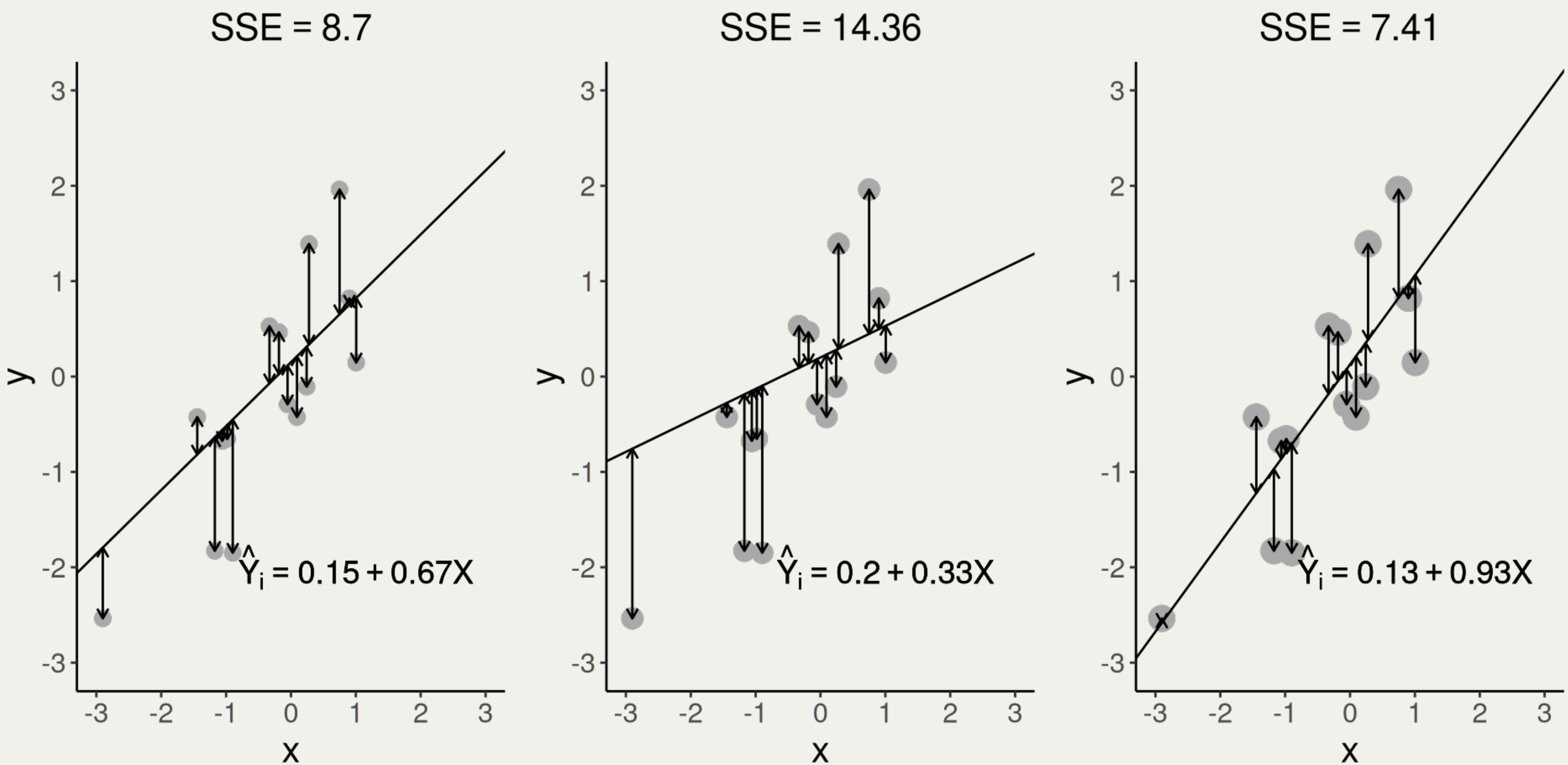
$$\hat{\beta} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})}$$

and

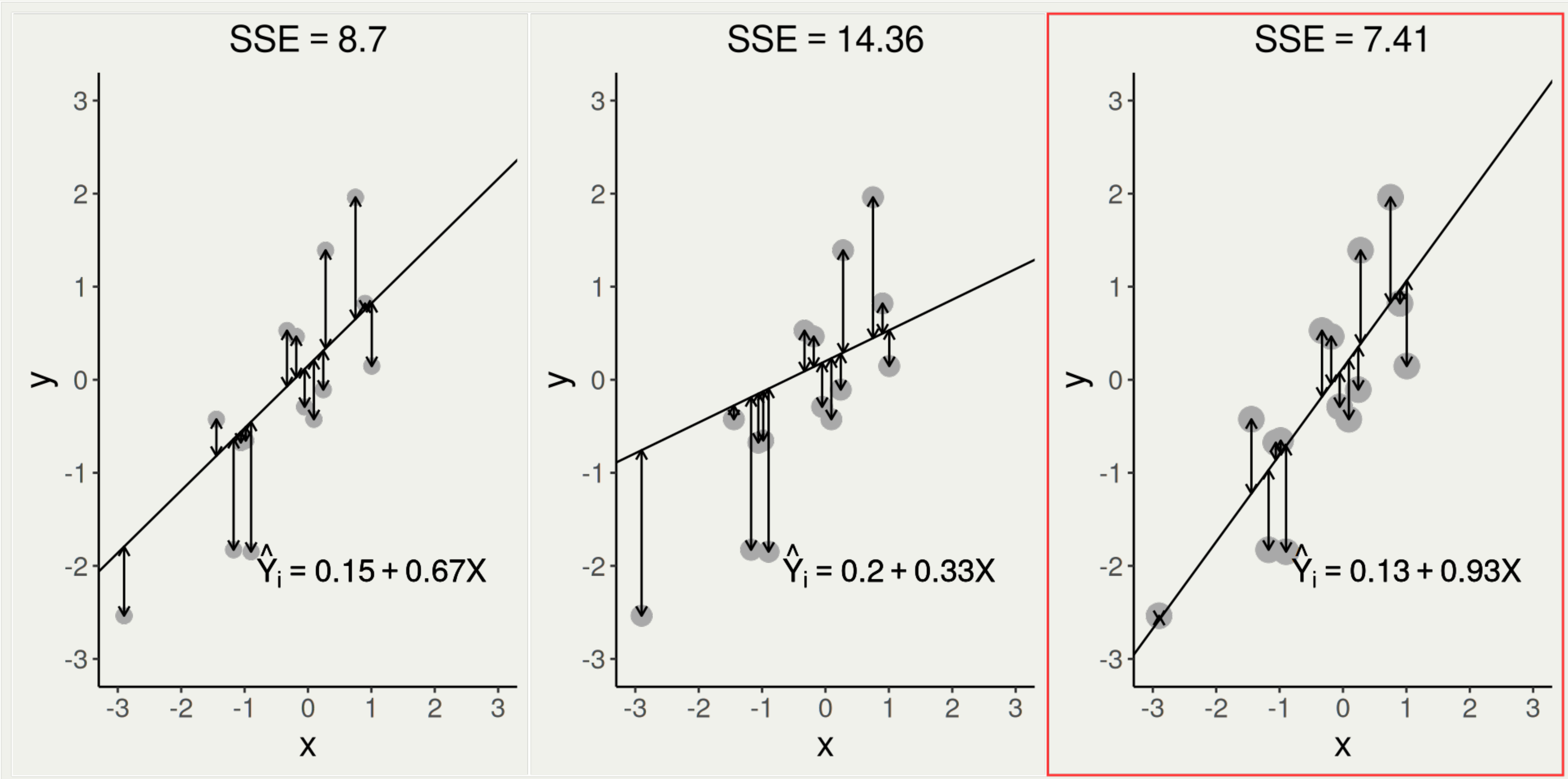
$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$$

- Note the similarity of the numerator of the former to the formula for covariance.

OLS Minimises SSE



OLS Minimises SSE



Estimand vs Estimate vs Estimator

A parameter can also be called an **estimand** (something that is estimated).

X (Twitter)

Example: Ordinary Least Squares

- Now let's estimate our $Longevity \rightarrow GDP$ model in R:

$$\widehat{GDP}_i = \hat{\alpha} + \hat{\beta} Longevity_i$$

```r
# Note the formula syntax: Y ~ X
lm(gdp_per_capita ~ democracy_duration, data = democracy_gdp_2020)
```

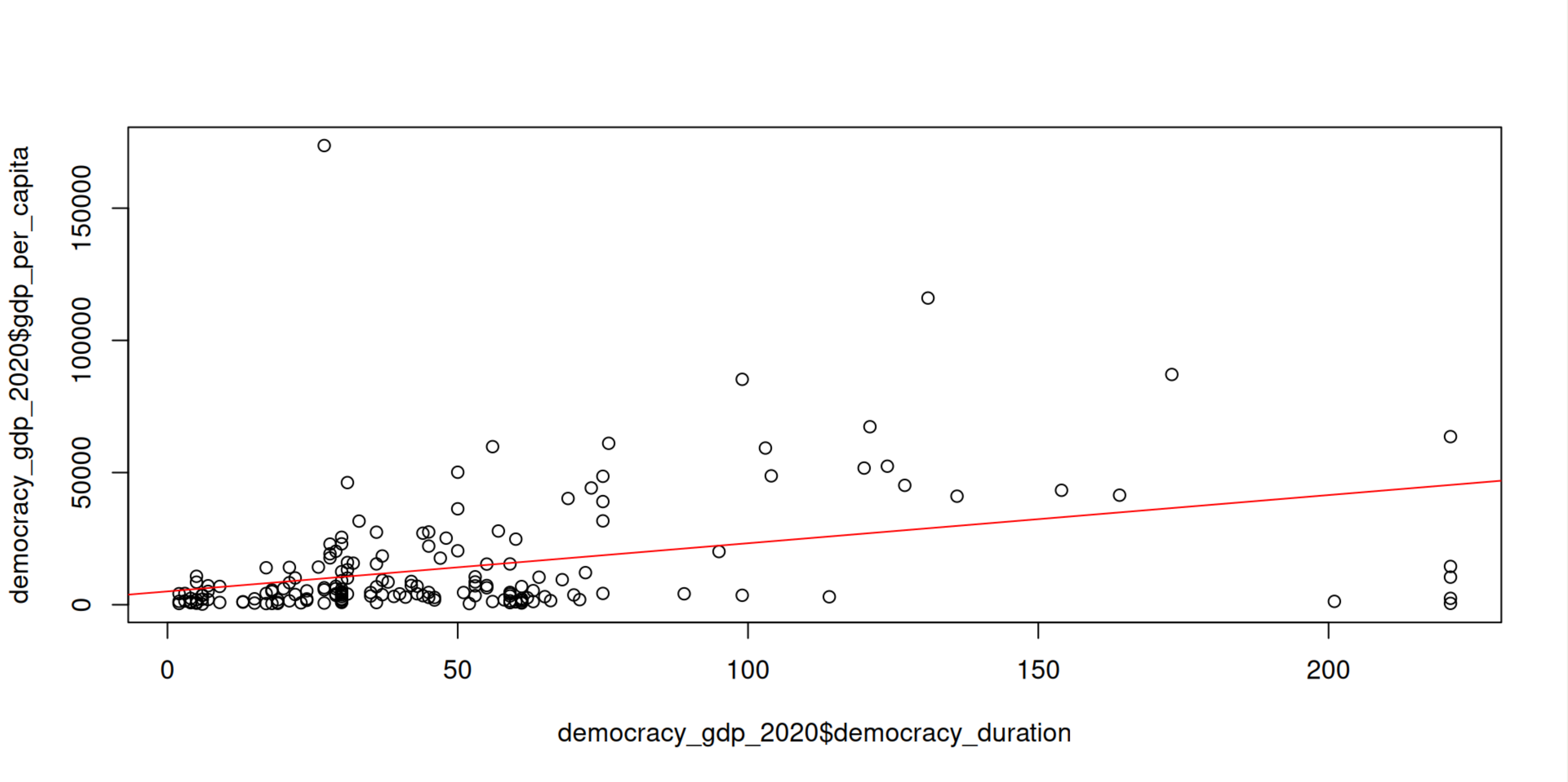
```
Call:
lm(formula = gdp_per_capita ~ democracy_duration, data = democracy_gdp_2020)

Coefficients:
       (Intercept)  democracy_duration
            5051.4               182.2
```

- In other words our OLS estimate of this model is:

$$\widehat{GDP}_i = 5051.4 + 182.2 \times Longevity_i$$

Example: Fitted Regression

```
plot(democracy_gdp_2020$democracy_duration, democracy_gdp_2020$gdp_per_capita)
abline(lm(gdp_per_capita ~ democracy_duration, data = democracy_gdp_2020), col = "red")
```



Example: Interpreting OLS Estimates

- Let's interpret our fitted model:

$$\widehat{GDP}_i = 5051.4 + 182.2 \times Longevity_i$$

  - $\hat{\alpha} = 5051.4$ - the expected GDP per capita for a state where a political regime lasted 0 years is 5051.4 USD.
  - $\hat{\beta} = 182.2$ - each additional year of political regime's longevity, on average, is associated with a 182.2 USD increase in GDP per capita.

Statistical Inference for Regression

Hypothesis Testing

- Null hypothesis: $H_0 : \beta = 0$ - in the population the expected GDP per capita is not associated with that state's political regime longevity.
- Alternative hypothesis: $H_a : \beta \neq 0$ - in the population the expected GDP per capita is associated with that state's political regime longevity.
- In other words, we want to:
  - quantify the sampling uncertainty associated with β;
  - use $\hat{\beta}$ to test hypotheses such as $\beta = 0$;
  - construct a confidence interval for $\hat{\beta}$.

t Test

- The test statistic for a single regression coefficient is:

$$t = \frac{\hat{\beta} - \beta_{H_0}}{\hat{\sigma}_{\hat{\beta}}}$$

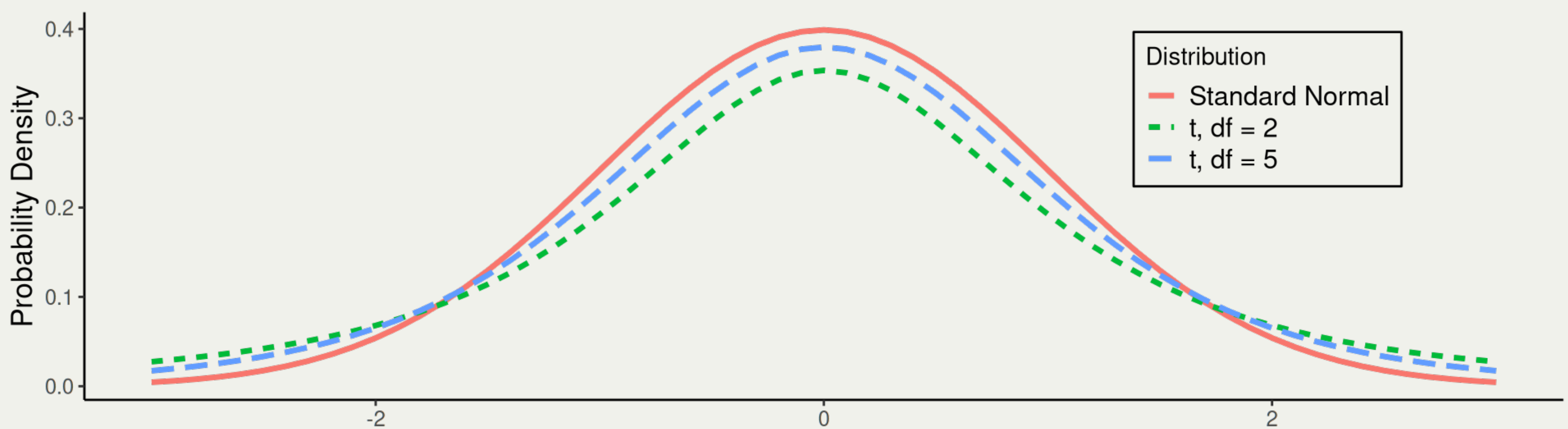
  where:

- $\hat{\beta}$ - is the estimated slope (coefficient)
- β_{H_0} - is the slope under H_0
- $\hat{\sigma}_{\hat{\beta}}$ - is the *standard error* of $\hat{\beta}$

Note that in the very common case the null hypothesis is $\beta_{H_0} = 0$ the t-statistic simplifies to $t = \frac{\hat{\beta}}{\hat{\sigma}_{\hat{\beta}}}$

Sampling Distribution of OLS Estimator

- When n is small (< 30), t follows a t-distribution with $n - 2$ degrees of freedom.
- When n is large (> 30) the Central Limit Theorem implies that t will follow the standard normal distribution.
- Most regression packages always use the t distribution as the normal distribution is only correct for large sample sizes



Example: *t* Test in R

```r
lm_fit <- lm(gdp_per_capita ~ democracy_duration, data = democracy_gdp_2020)
summary(lm_fit) # Use `summary()` function to get a more detailed output
```

```
Call:
lm(formula = gdp_per_capita ~ democracy_duration, data = democracy_gdp_2020)

Residuals:
   Min     1Q Median     3Q    Max 
-44806  -8756  -4944   4820 163717 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)         5051.44    2370.78   2.131   0.0345 *  
democracy_duration   182.22      35.15   5.185 5.99e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20900 on 173 degrees of freedom
  (20 observations deleted due to missingness)
Multiple R-squared:  0.1345,	Adjusted R-squared:  0.1295 
F-statistic: 26.88 on 1 and 173 DF,  p-value: 5.995e-07
```

Example: Working Out t Test

- While the full R output contains most details, let's see how t test was done here:

$$t = \frac{\hat{\beta} - \beta_{H_0}}{\hat{\sigma}_{\hat{\beta}}} = \frac{182.22 - 0}{35.15} \approx 5.184$$

- As for large sample sizes t-distribution approximates standard normal:

```r
(1 - pnorm(5.184)) * 2
```

```
[1] 2.171769e-07
```

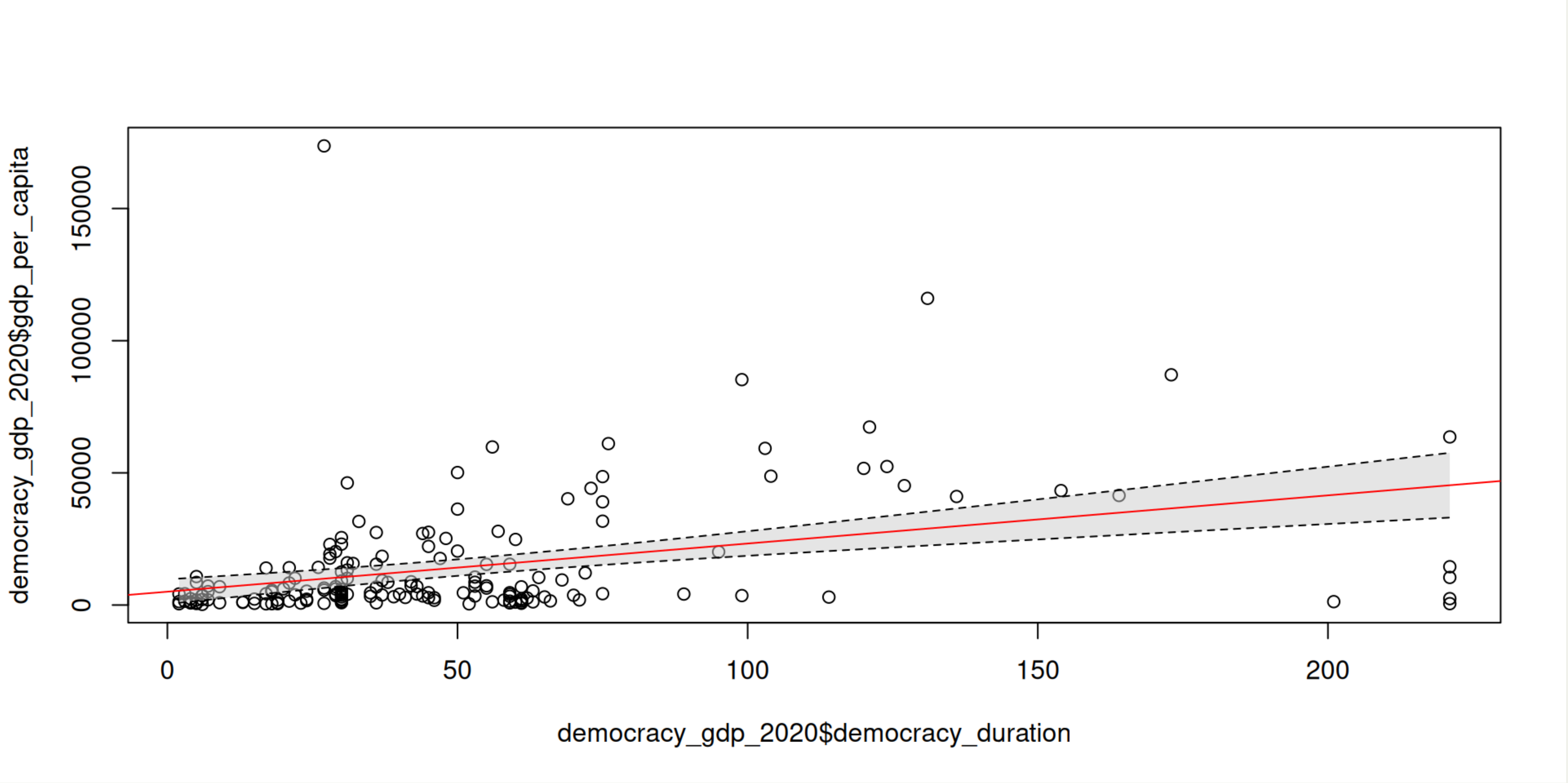
What Conclusion Do We Make?

- The probability of observing this difference under the null hypothesis is ≈ 0.000000599
- Thus, we can reject the null hypothesis of no association in the population between regime longevity and GDP at 0.001%-level.
- In other words, it is very unlikely that we would observe this test-statistic if the null hypothesis were true.

Confidence Intervals for Regression Coefficients

- As with other estimated parameters we can calculate **confidence intervals** for $\hat{\beta}$
  - 95% Confidence interval : $\hat{\beta} \pm 1.96\hat{\sigma}_{\hat{\beta}}$
  - 99% Confidence interval : $\hat{\beta} \pm 2.58\hat{\sigma}_{\hat{\beta}}$
- For our regression model the 95% confidence interval is:
  - Lower bound: $182.22 - 1.96 \times 35.15 = 113.326$
  - Upper bound: $182.22 + 1.96 \times 35.15 = 251.114$

Example: Confidence Intervals



Next

- Workshop:
  - RQ Presentations I
- Next week:
  - Linear regression II