The Classical Linear Regression Model (CLRM)

EC 339

Marcio Santetti
Spring 2023

Motivation

OLS works, but it needs assumptions

- The goal when using OLS is to obtain **unbiased**, **efficient**, and **consistent** estimators.
- Moreover, we want to be able to do **hypothesis testing**.
- All these properties are made possible through **7 assumptions**.
- This set of assumptions is known as the **Classical Linear Regression Model** (CLRM).

The Classical Assumptions

The set of Classical Assumptions

1. The regression model is **linear**, **correctly specified**, and has an **additive** stochastic error term.

2. The stochastic error term (u_i) has a **zero** population mean.

3. All explanatory variables (x_i) are **uncorrelated** with the error term.

4. Observations of the error term are **uncorrelated** with each other.

5. The error term has a **constant variance**.

6. No explanatory variable is a **perfect linear function** of any other explanatory variable.

7. The error term is **normally distributed**.

Assumption 1

> "*The regression model is **linear**, **correctly specified**, and has an **additive** stochastic error term.*"

- *Linear* means linear in **parameters** (β_i);
- *Correctly specified* means that it has the correct **functional form** and **no** omitted variables.
- And an **additive** error term implies **no** other form in which u_i appears in a model.

- **Examples**:

$$y_i = \beta_0\beta_1 x_{1i} + \beta_2 x_{2i} + u_i$$

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} u_i$$

$$y_i = \beta_0 + log(\beta_1) x_{1i} + \beta_2 x_{2i} + u_i$$

Assumption 1

One of the main reasons for a *violation* of CLRM Assumption I is an **incorrectly specified** model.

- This may happen due to
  - Incorrect **functional form** (data visualization matters!);
  - **Omitted** variables (leading to omitted variables bias).

A regression's error term may sometimes be a **black box**.

- Recall that any potentially omitted variable(s) lie(s) there!

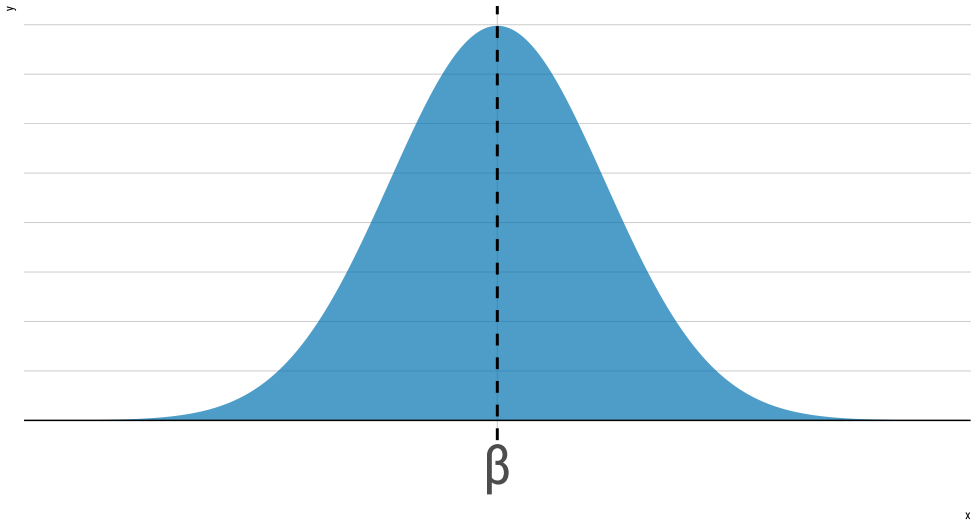
Therefore, our models must have a **theoretical** motivation.

What is bias?

An estimator is **biased** if its expected value is different from the *true* population parameter.

When considering our slope coefficients ($\hat{\beta}_i$), we expect that they, on average, are close to the **"true"** population parameter, β_{pop}.

Unbiased: $\mathbb{E}\left[\hat{\beta}_{OLS}\right] = \beta_{pop}$



Biased: $\mathbb{E}\left[\hat{\beta}_{OLS}\right] \neq \beta_{pop}$

Assumption 2

> *"The stochastic error term (u_i) has a **zero** population mean."*

- Values of the stochastic error term are defined by **pure chance**.
- It follows a probability **distribution** centered around zero.
- Also known as the **exogeneity** assumption.

From standard Microeconomic theory, recall:

- Factors that influence the **demand** for a given good:
    - Price of the good itself, price of substitutes, preferences...

Assumption 2

> *"The stochastic error term (u_i) has a **zero** population mean."*

In practice, what is the difference between $\mathbb{E}[u \mid x] = 0$ and $\mathbb{E}[u \mid x] \neq 0$?

Assumption 3

> *"All explanatory variables (x_i) are **uncorrelated** with the error term."*

- Observed values of the independent variable are determined **independently** of the values contained in the error term
- $Cor(x_i, u_i) \neq 0 \implies$ **violation** of CLRM Assumption III.
- A possible reason: a variable correlated with some x_i being **omitted** from the model.

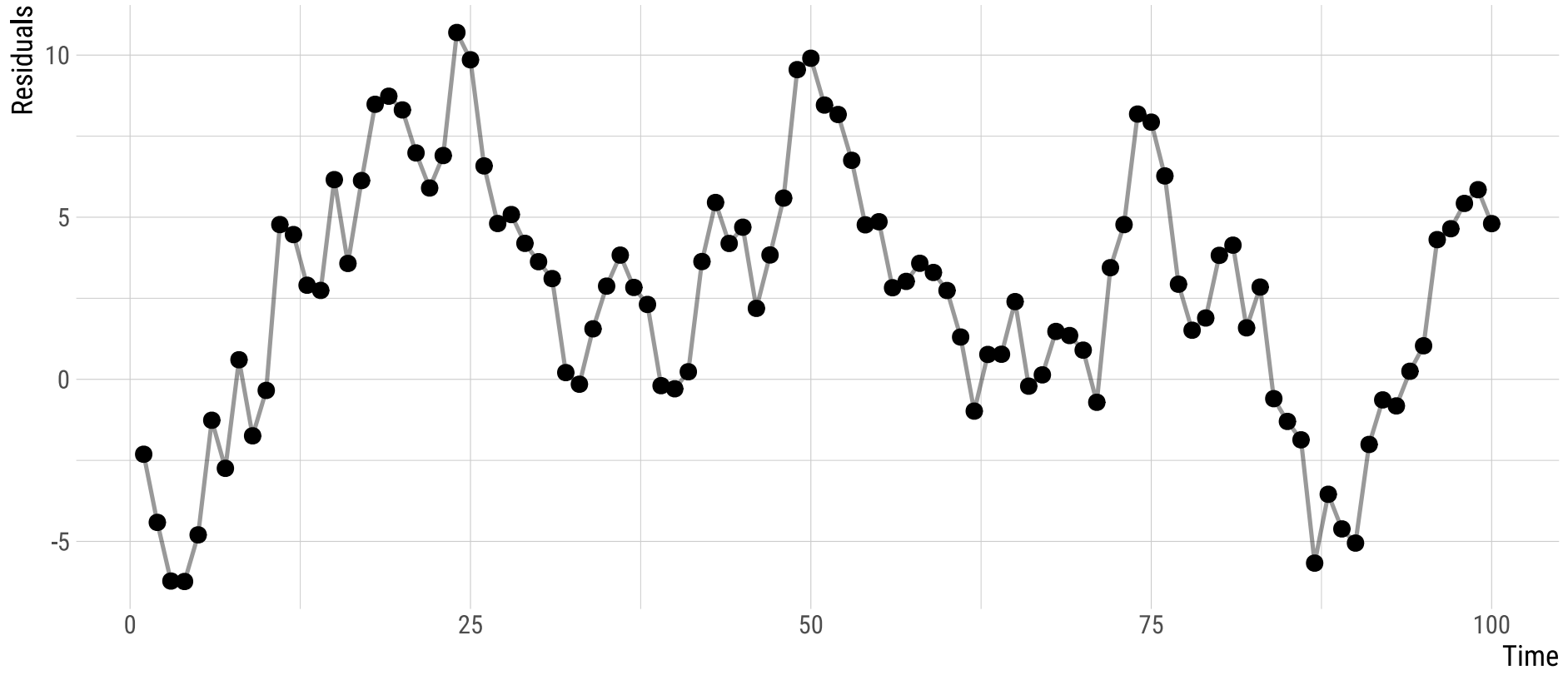
Assumption 4

> *"Observations of the error term are **uncorrelated** with each other."*

- Also known as **autocorrelation**.
- Common in **time-series** data.
- Occurs when the model's disturbances are correlated **over time**, i.e., $Cor(u_t, u_j) \neq 0$ for $t \neq j$.

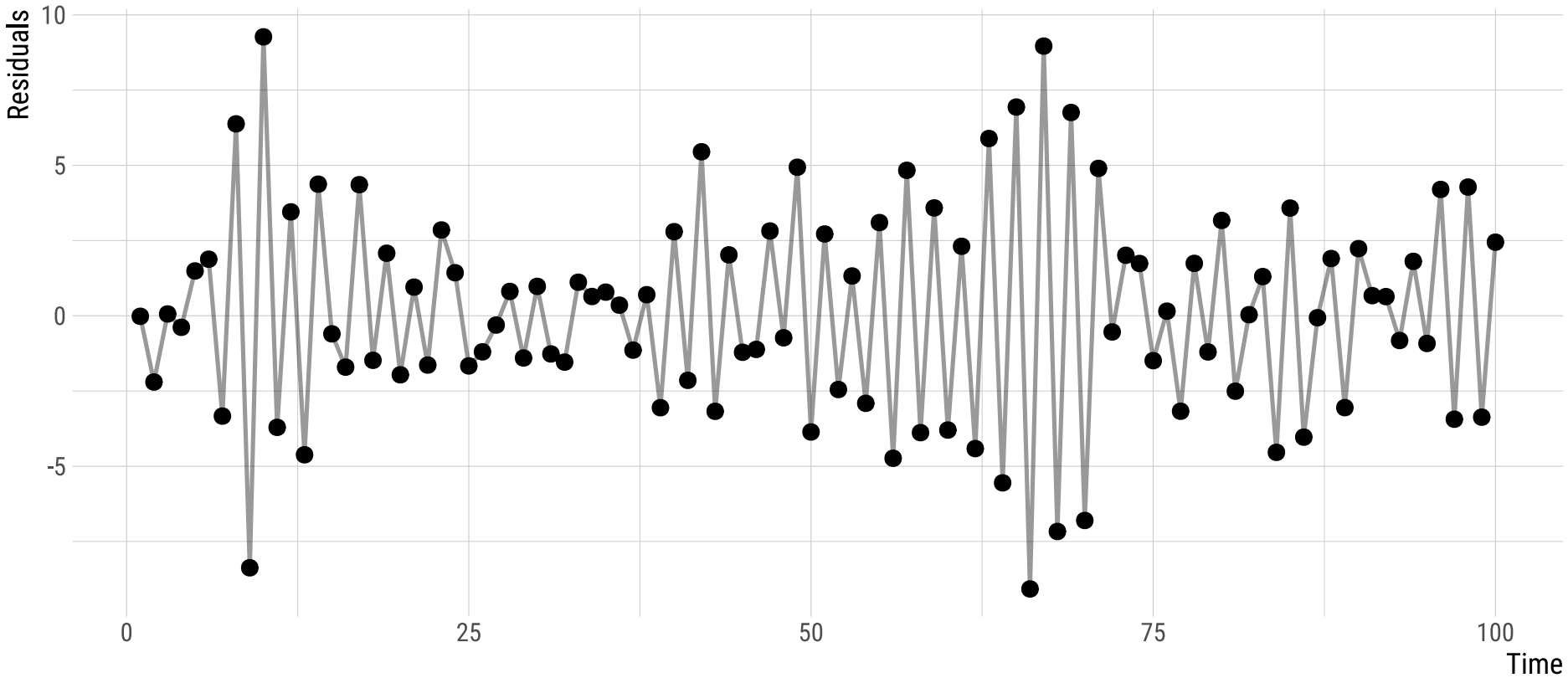
Assumption 4

Behavior of u_t over time (positive serial correlation)



Assumption 4

Behavior of u_t over time (negative serial correlation)



Assumption 5

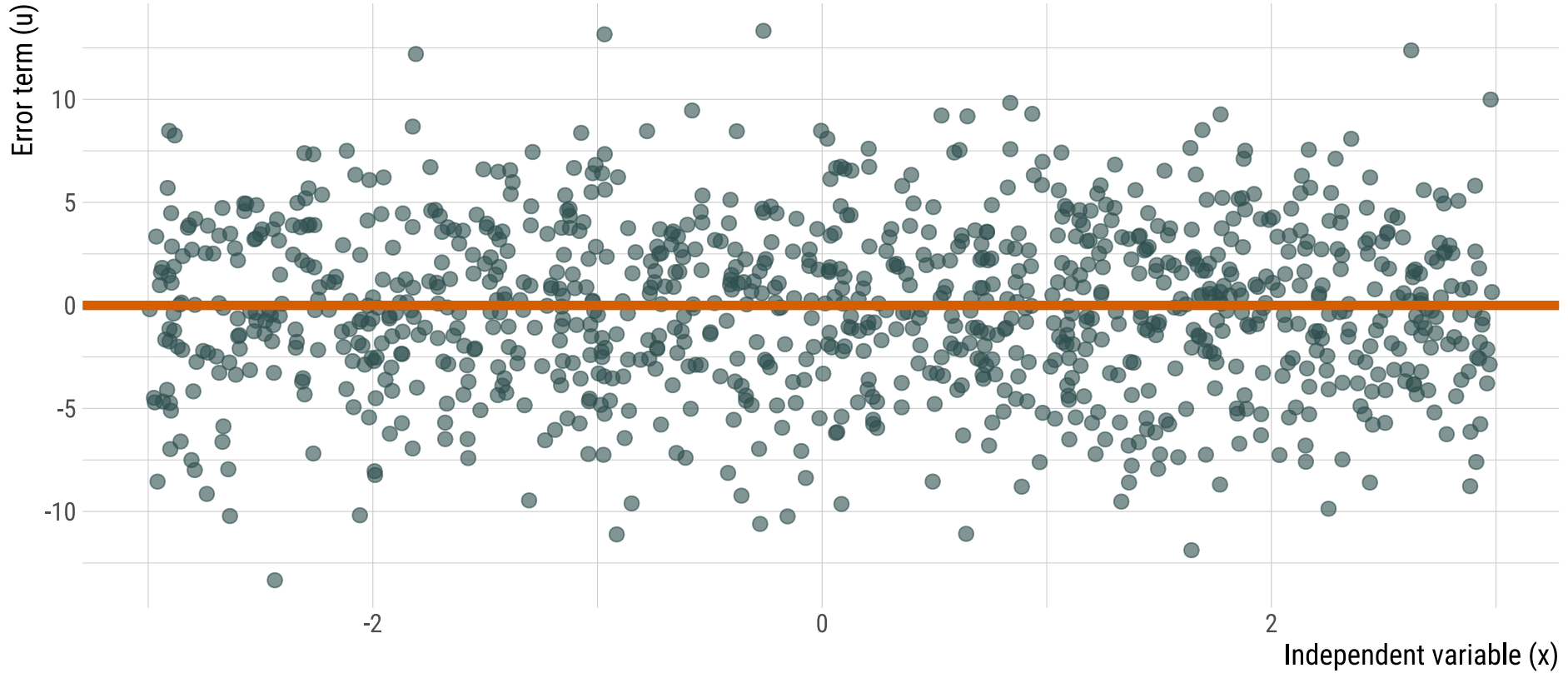
> *"The error term has a **constant variance**."*

- Also known as the **homoskedasticity** assumption.
- If violated, we have **heteroskedasticity**.
- Extremely **common** in cross-section data sets (also in financial time-series data).

- This assumption implies that the error term has the **same variance** for each value of the independent variable.
  - $Var(u|x) = \sigma^2$

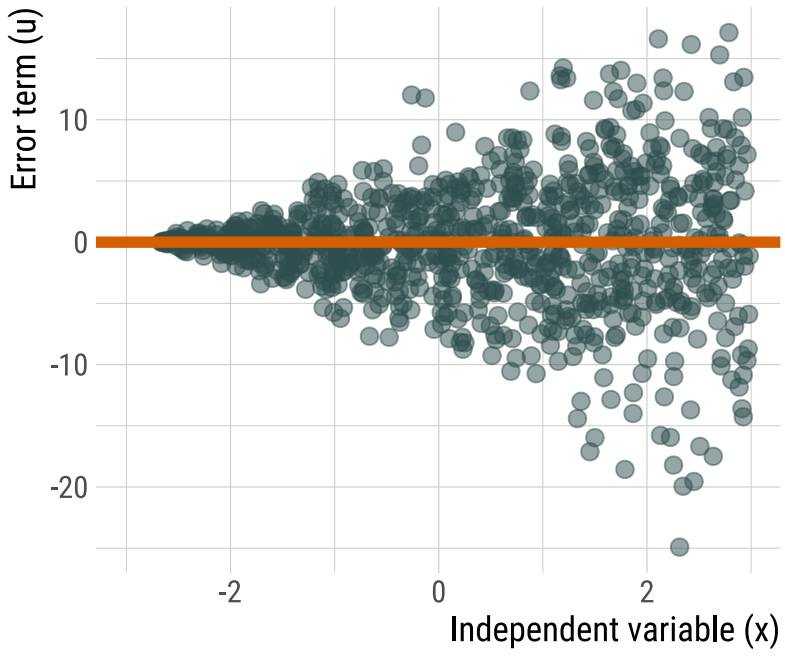
Assumption 5

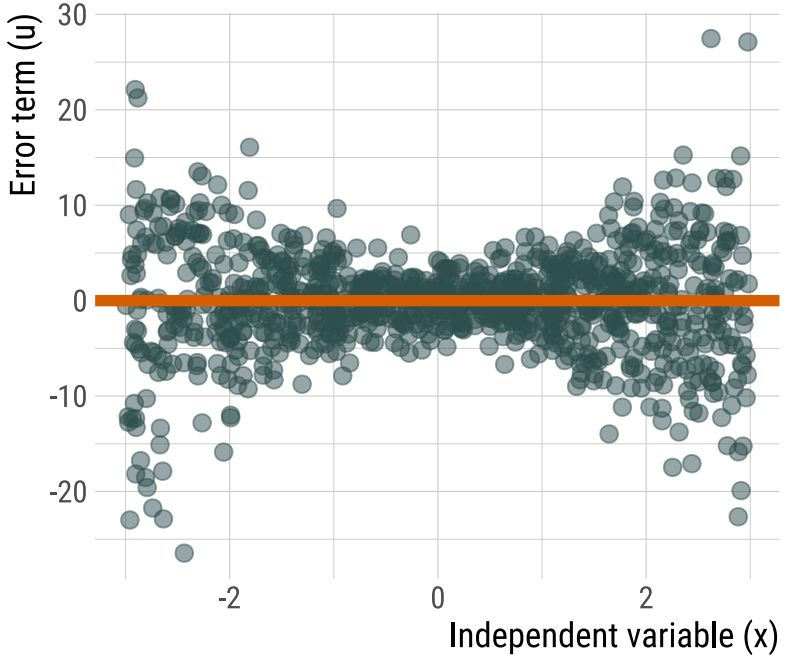
- **Homoskedastic** residuals:



Assumption 5

- **Heteroskedastic** residuals:





Assumption 6

> *"No explanatory variable is a **perfect linear function** of any other explanatory variable."*

- Also known as the **no perfect multicollinearity** assumption.
- Only completely **violated** if an independent variable x_i is a **deterministic** function of another variable x_j, for $i \neq j$

Examples:

- $x_3 = x_1 - 1,000$
- $x_2 = 50 + x_1$

Assumption 7

> *"The error term is **normally distributed**."*

- Summarized by $u_i \sim \mathcal{N}(0, \sigma^2)$.

OLS **still works** without this assumption!

But crucial for **hypothesis testing and inference**.

The Gauss-Markov theorem

The Gauss-Markov theorem

From CLRM Assumptions **I through VI**, we guarantee that OLS is **BLUE**.

We will learn how to deal with the most common **violations** of CLRM Assumption after the Midterm exam.

Next time: CLRM in practice