Module 1: Statistical Learning and the Role of Econometrics

Econometrics II

Max Heinze (mheinze@wu.ac.at)
Department of Economics, WU Vienna

Sannah Tijani (stijani@wu.ac.at)
Department of Economics, WU Vienna

October 16, 2025

Statistical Learning

Prediction

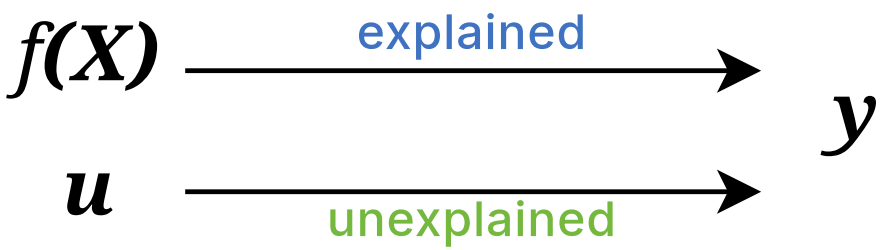
Inference

Models

Finding Relationships

We are interested in finding a **relationship** between the samples

- $\boldsymbol{y} \in \mathbb{R}^N$, the **dependent**, and
- $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_K) \in \mathbb{R}^{N \times K}$, the **independent variables**.



We can write this **relationship** as

$$\boldsymbol{y} = f(\boldsymbol{X}) + \boldsymbol{u},$$

where $f(\cdot)$ is an unknown function that represents information that $\boldsymbol{X}$ provides about $\boldsymbol{y}$. All other relevant information is contained in the error term $\boldsymbol{u}$.

Notes on Terminology and Notation

As you know, there are many names for the **dependent** variable, such as:

- outcome,
- response, or
- explained variable.

Likewise, we know a multitude of alternative terms for the **independent** variables, such as:

- explanatory variables, or
- predictors.

You also might come across different ways of denoting the **error term**, such as

$$u, \qquad e, \qquad \varepsilon.$$

We will use u in the materials of this course, but you can choose whichever you prefer.

Why Statistical Learning?

You may ask yourself, **"what are we doing this for?"** This is a **very good question** (and you should ask these types of questions very often), and there are **two answers** to it.

Prediction

We want to learn about Y beyond our sample $\boldsymbol{y}$.

Example: We know that a congestion tax reduces asthma in young children. There is a proposal to introduce a congestion tax, and we want to predict how large the health benefits are.

Inference

We want to learn more about f, the relation between Y and X.

Example: We observe that after the introduction of a carbon tax, carbon emissions declined. We want to find out whether there is a causal relationship between the two or whether emissions had declined anyway.

Statistical Learning

Prediction

Inference

Models

The Role of Econometrics

What Does "Prediction" Mean?

When we **predict**, we use $\boldsymbol{X}$ and an estimate of f, $\hat{f}$, to obtain new values of Y.

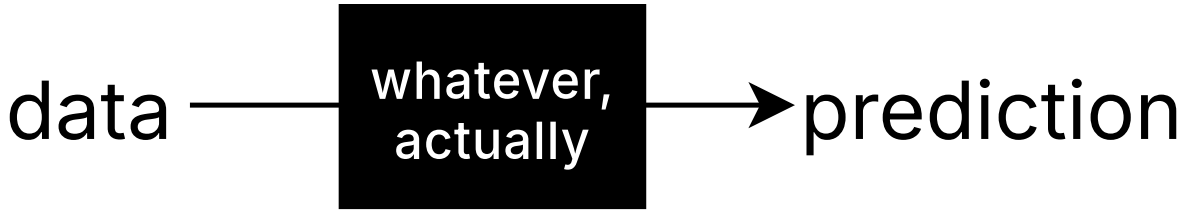
- We can obtain an estimate of $\boldsymbol{y}$, which we call $\hat{\boldsymbol{y}}$. This is called an **in-sample prediction**.
- With **new data** $\tilde{\boldsymbol{X}}$, we can obtain an **out-of-sample** prediction.

In the **Kaggle Competition**, you will get a training dataset $\boldsymbol{X}$ and $\boldsymbol{y}$, which you will use to estimate $\hat{f}$. You can then predict $\hat{\boldsymbol{y}}$ and check the predictions against $\boldsymbol{y}$.

There is a second part of the data, the test dataset, of which you get only $\tilde{\boldsymbol{X}}$, and we will keep the $\tilde{\boldsymbol{y}}$. You will try to get good **out-of-sample** predictions, and in the end we will reveal who fared the best.

Which f to Choose?

For prediction, we **do not care** about what our $f(\cdot)$ looks like. As long as we get useful predictions, we can treat $f(\cdot)$ as a **black box**.

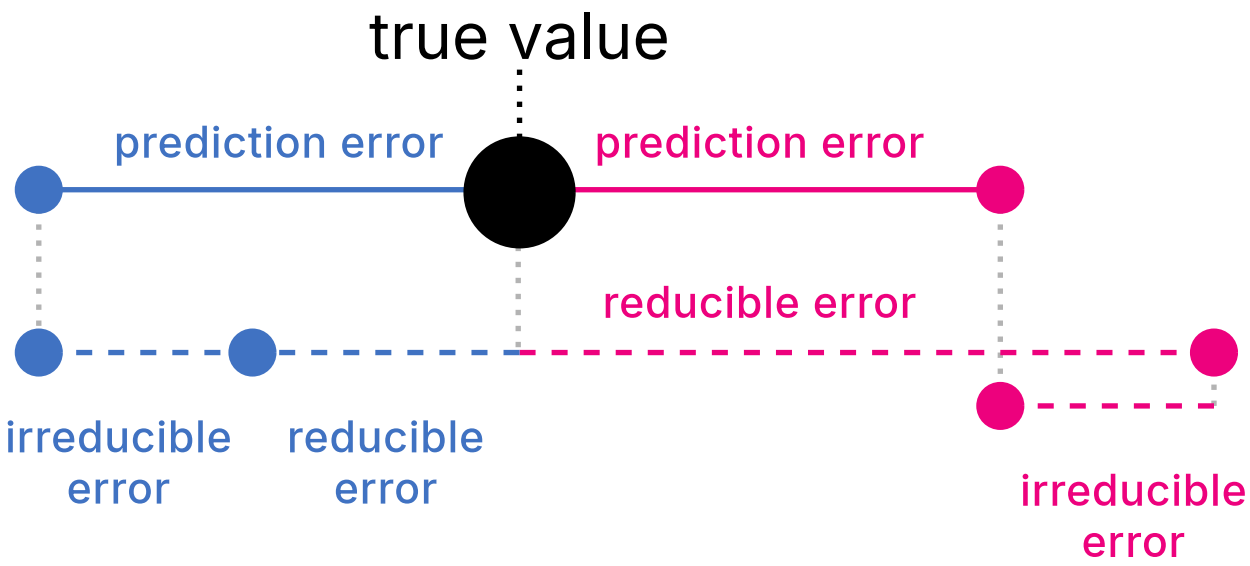


In 2010, Paul the octopus correctly predicted the outcome of all FIFA Mens' World Cup games in which the German national team played, plus the final.

Other examples include Spotify's and Youtube's recommendation algorithms, or Large Language Models like ChatGPT.

Prediction Accuracy

The accuracy of our prediction depends on the sum of **two types of errors**:



- The **reducible error** stems from imperfectly estimating f.
- The **irreducible error** is contained in the error term, it consists of elements that cannot be explained by $\boldsymbol{X}$.

Error Decomposition

Let us look at the **mean squared prediction error**.

$$\begin{aligned}\mathrm{E}\left((\hat{\boldsymbol{y}} - \boldsymbol{y})^2\right) &= \mathrm{E}\left(\left(f(\boldsymbol{X}) + \boldsymbol{u} - \hat{f}(\boldsymbol{X})\right)^2\right) \\ &= \mathrm{E}\left(\left(f(\boldsymbol{X}) - \hat{f}(\boldsymbol{X})\right)^2\right) + \mathrm{Var}(\boldsymbol{u}).\end{aligned}$$

We have **decomposed** the mean squared error into a **reducible** and an **irreducible** part.
We can now split the reducible error once more.

$$= \mathrm{Bias}\left(\hat{f}(\boldsymbol{X})\right)^2 + \mathrm{Var}\left(\hat{f}(\boldsymbol{X})\right) + \mathrm{Var}(\boldsymbol{u}).$$

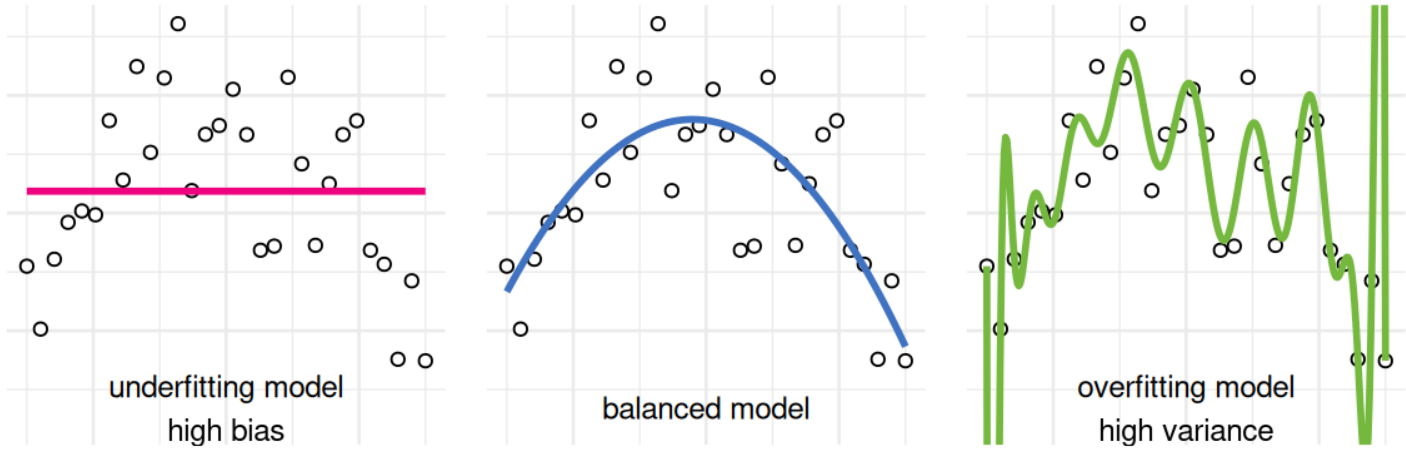
The **reducible error** consists of the **squared bias** of $\hat{f}$ and its variance.

- show decomposition #1
- show decomposition #2

Overfitting and Underfitting

We want to minimize the **reducible error** as far as possible by **balancing bias and variance**. However, we want to **avoid** trying to reduce the **irreducible error**.

When we try to minimize the irreducible error, we will overfit our model.



Why is it **bad** to **overfit**? We call it an "irreducible" error for a reason. We can fit something that matches the data in the sample arbitrarily close. But this will lead to **poor out-of-sample performance**.

Occam's Razor

Occam's (or Ockham's, or Ocham's) **Razor**, also called the **principle of parsimony**, is a simple rule:

Of two competing theories, choose the simpler one.

- This relates to our notion of **overfitting**.
- Of course, we should not omit valuable information (in the worst case, our results will be severely biased), but we should **not include unnecessary information**, either.

We can use this principle to inform our notion of which model is "better" than the other.

- We know that there are ways to determine **how well a model fits** the data.
- However, a very complicated model will fit the data well, and we have no idea when we **start to explain random noise**.
- So instead of blatantly choosing the best-fitting model, we should **think about out-of-sample performance** and, if in doubt, choose the **simpler, more general model**.

Statistical Learning

Prediction

Inference

Models

The Role of Econometrics

The Linear Model

Inference vs. Prediction

In a sense, **prediction** and **inference** are opposite approaches. Before, we cared only about the fitted value and treated $f(\cdot)$ as a black box; **now, we care only about** $f(\cdot)$ (or, more precisely, our estimate $\hat{f}(\cdot)$).

With knowledge about $\hat{f}(\cdot)$, we can answer questions like these:

- Are X and Y correlated?
- What happens *if* we increase X by one?
- Did this increase *cause* a change in Y?
- *How* does $\hat{f}$ map X to Y?

Inference: Which Questions Can We Ask?

How much of the **gender pay gap** is caused by **discrimination**?

Is a **long life** correlated with **olive oil consumption**?

Will your **Econometrics II grade** improve if you spend time **studying** for the exam?

Was the use of **facial masks** related to **Covid-19 prevalence**? If so, in which direction?

Do **malaria nets** reduce the number of people **infected** by the disease?

Are **croplands** less fertile if they lie downstream of a **gold mine**?

Do more generous **unemployment benefits** prompt people to **work less**?

Is **wealth** correlated with **happiness**?

Does an **Economics degree** make people more likely to **comment** on issues they have zero expertise about?

Parallel Universes

Causal inference is easy under one assumption: We can **switch** between **two states** of the world. Consider this:

> Will your **Econometrics II grade** improve if you spend time **studying** for the exam?

Say I want to answer this question. I now only need to do **two things**:

(1) At the end of the course, ask you how much you studied and record your exam results.

(2) **Switch** to a **world** where none of you studied, ask again, and compare.

It should be **apparent** that this is **not possible**. We call this the **Fundamental Problem of Causal Inference**. The existence of this problem is the reason that you have to take this course.

Correlation is not Enough

We need to deal with the **Fundamental Problem of Causal Inference** in some way if we want to perform causal inference. Just looking at the data and checking correlations (which is what we do with a naive regression) is not enough:

Do **malaria nets** reduce the number of people **infected** by the disease?

It is thinkable that people who install malaria nets are richer, more health-conscious, or both, than people who do not install malaria nets. This may cause part of the correlation.

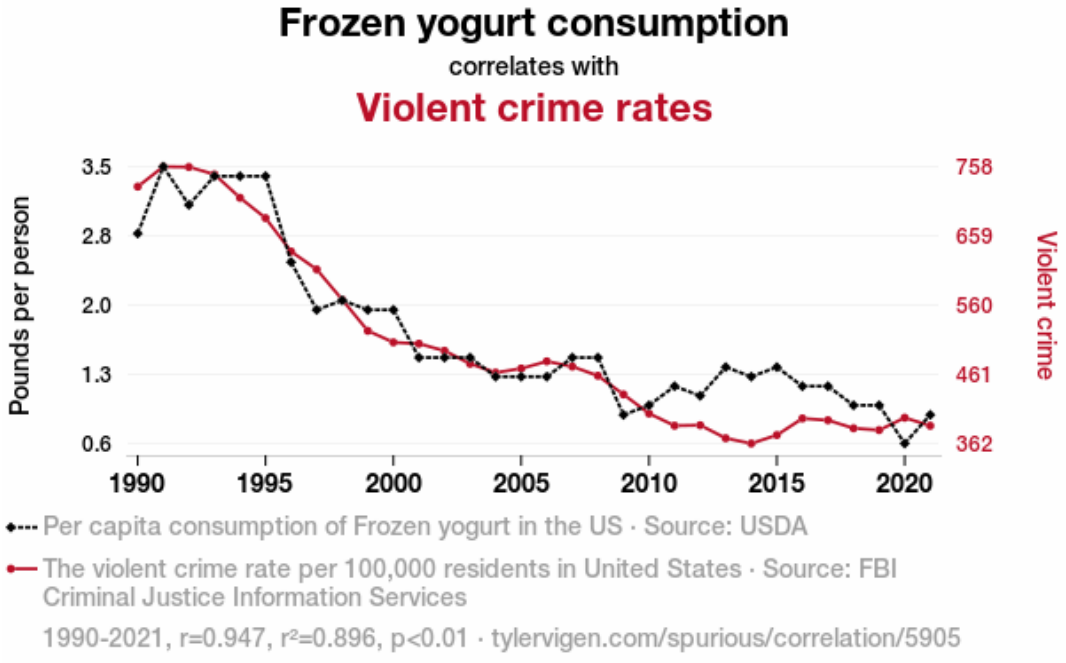
Does an **Economics degree** make people more likely to **comment** on issues they have zero expertise about?

We might observe this behavior more often in economists than in the general population. Even so, it may be caused by the fact that most economists are men, and not by their degree.

Correlation vs. Causality

You likely have heard this sentence before:

Correlation does not mean causality.



(Vigen, 2024)

You may also have seen examples like the one of the left, e.g. from Tyler Vigen's site tylervigen.com/spurious-correlations.

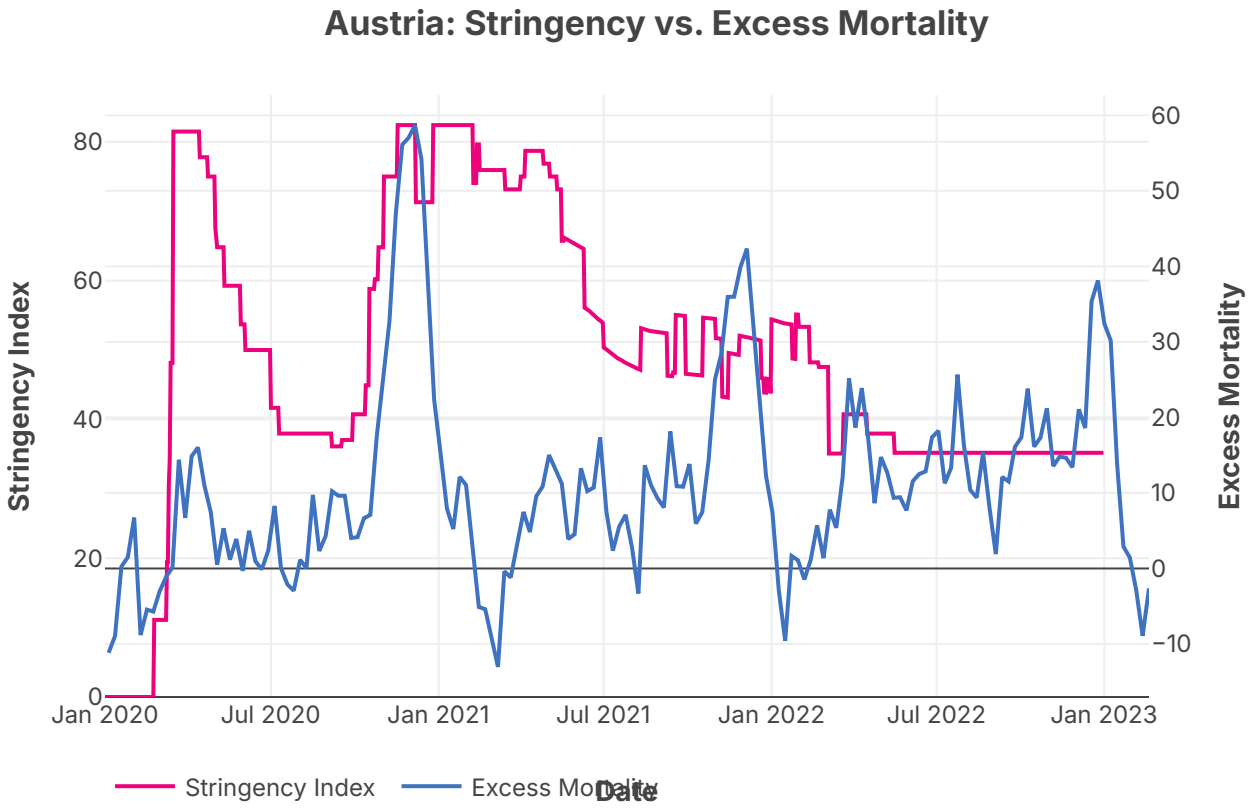
In this course, we will investigate **why** correlation does not necessarily imply causality, and how we can **deal with this** when performing causal inference.

Correlation vs. Causality

Correlation does not mean causality.

But have you also thought about this?

***No correlation* does not mean *no causality*.**



The chart on the left shows the **stringency** of containment measures and the **excess mortality** during the Covid-19 pandemic in Austria. The two time series are only **weakly correlated**.

Different effects could be at play **at the same time**. One hypothesis: Containment measures reduce mortality, but high mortality prompts stricter containment.

Statistical Learning

Prediction

Inference

Models

The Role of Econometrics

The Linear Model

Appendix

Is This a Bad Model?

My **workplace** is up here.

I tried to use this map to **bike** home. It was **utterly useless**, and it also didn't tell me that I was constantly biking uphill.

I live in the **10th District**.

All Models are Beautiful

As we know,

All models are wrong, but some are useful.

—Partly coined by Box (1976)

The model (i.e., subway map) on the previous slide is useful for navigating the subway. Of course, it is useless when you use a bike. Models are an **approximation of reality** that are specific to a certain context and allow us to learn specific things.

To learn about the true f, we need a **model** that suits our purpose and the data at hand. We can characterize models, e.g., like this:

- **Parametric** (containing a finite number of parameters) vs. **non-parametric** models,
- **Supervised** (fitted using information on $\boldsymbol{y}$) vs. **unsupervised** models,
- **Regression** (quantitative $\boldsymbol{y}$) vs. **classification** (qualitative $\boldsymbol{y}$).

Parametric and Non-Parametric Models

Parametric models are models that impose a certain **parametric** structure on f. In case of a **linear model**, the dependent is a linear combination of $\boldsymbol{X}$, with parameters $\boldsymbol{\beta} \in \mathbb{R}^{K+1}$:

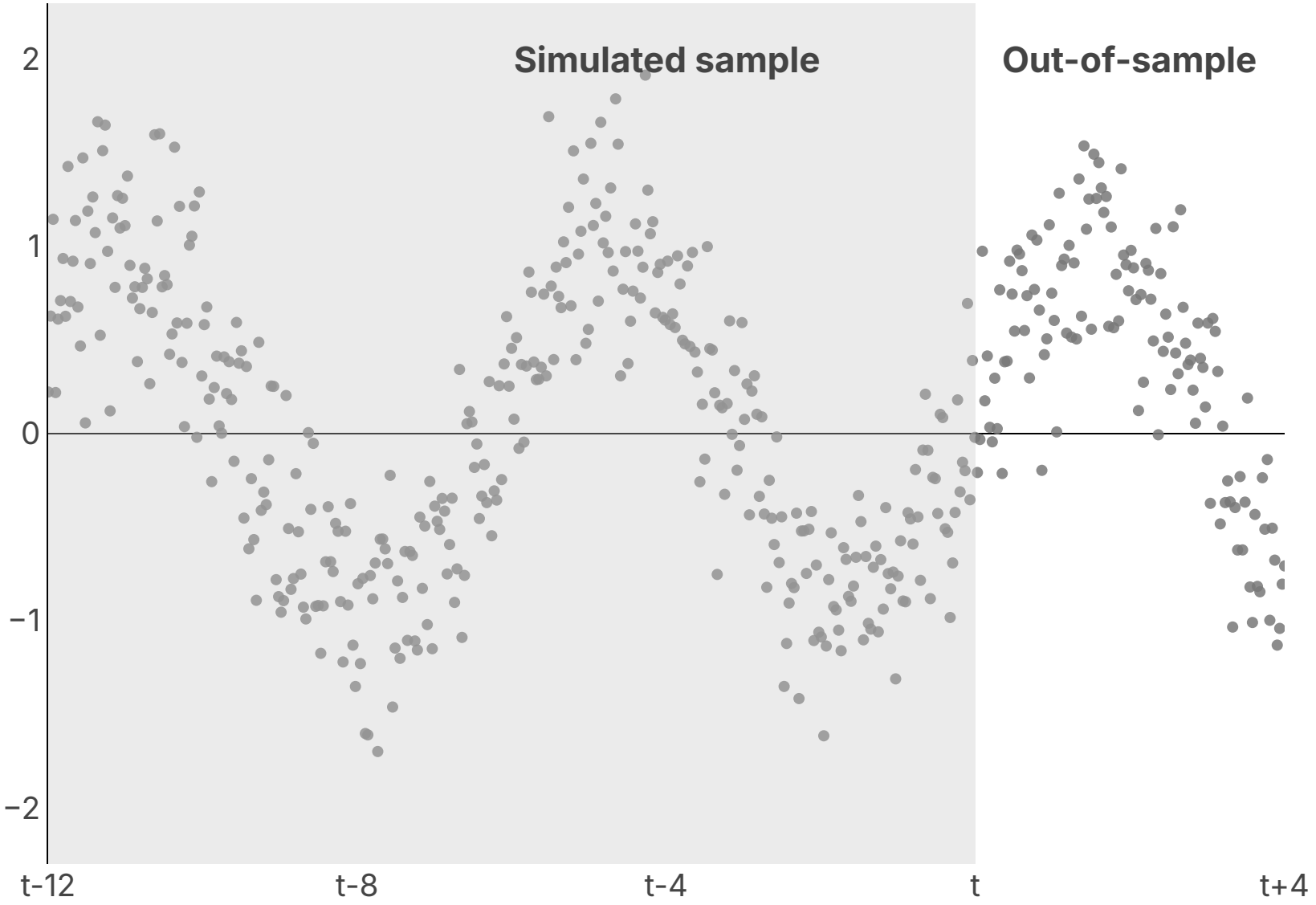
$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}$$

- The **advantage** of this is that we only need to estimate $K + 1$ parameters, which is easy to do and easy to interpret. Linear models are also less prone to overfitting.
- A **disadvantage** is their lacking flexibility, meaning that $\hat{f}$ may deviate too far from f.

Non-parametric models do not impose a structure on f a priori. Rather, we fit f to be as cloase as possible to the data under certain constraints.

- **Advantages** include flexibility to mirror many possible forms of f, a better fit, and less reliance on model building.
- **Disadvantages** are the high data requirements, difficulty to interpret f, and that the approach is more prone to overfitting.

Parametric vs. Non-Parametric Fit

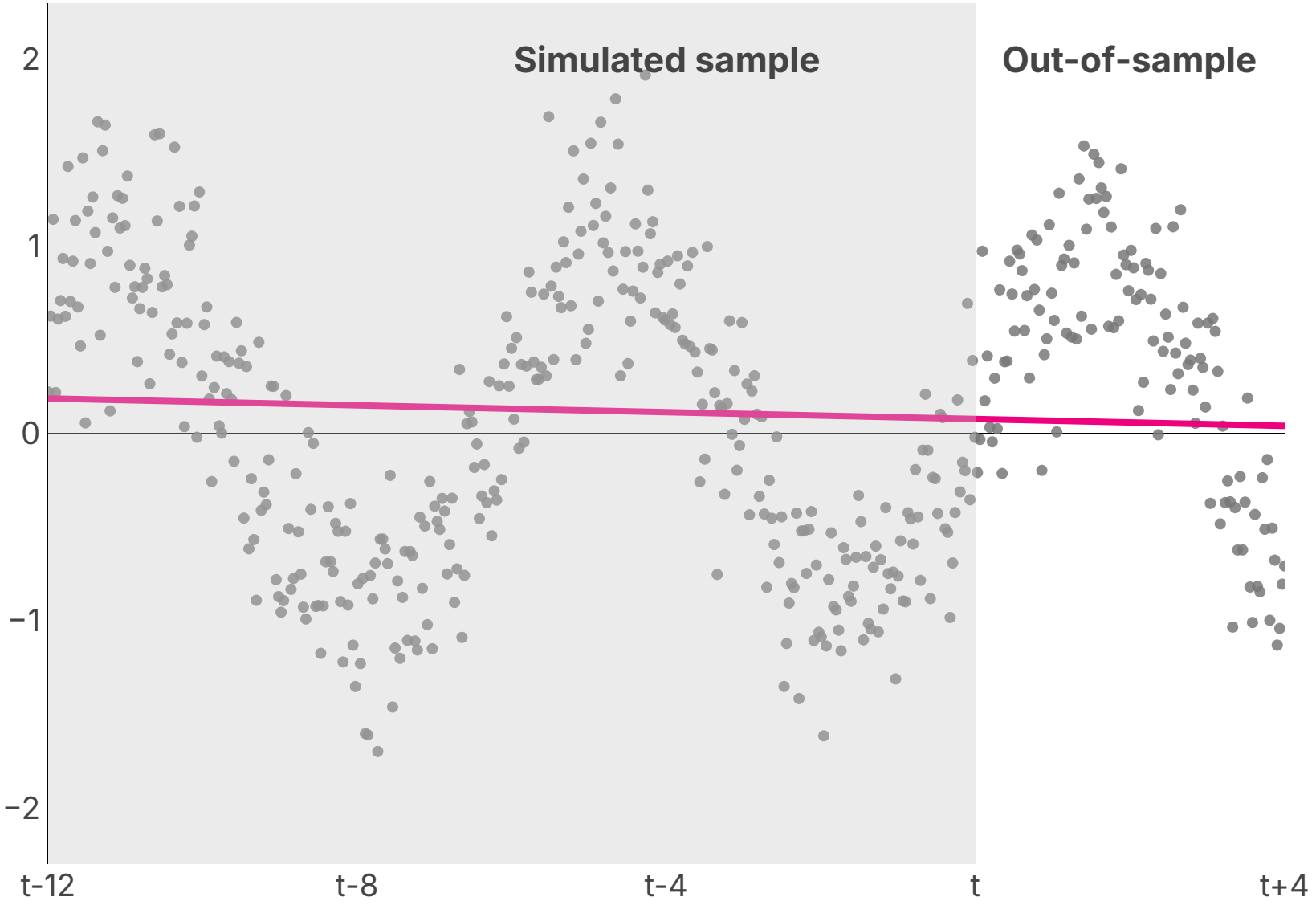


We simulate some data from

$$Y = \sin(X)$$

and plot them. We can now compare how well different models fit.

Parametric vs. Non-Parametric Fit

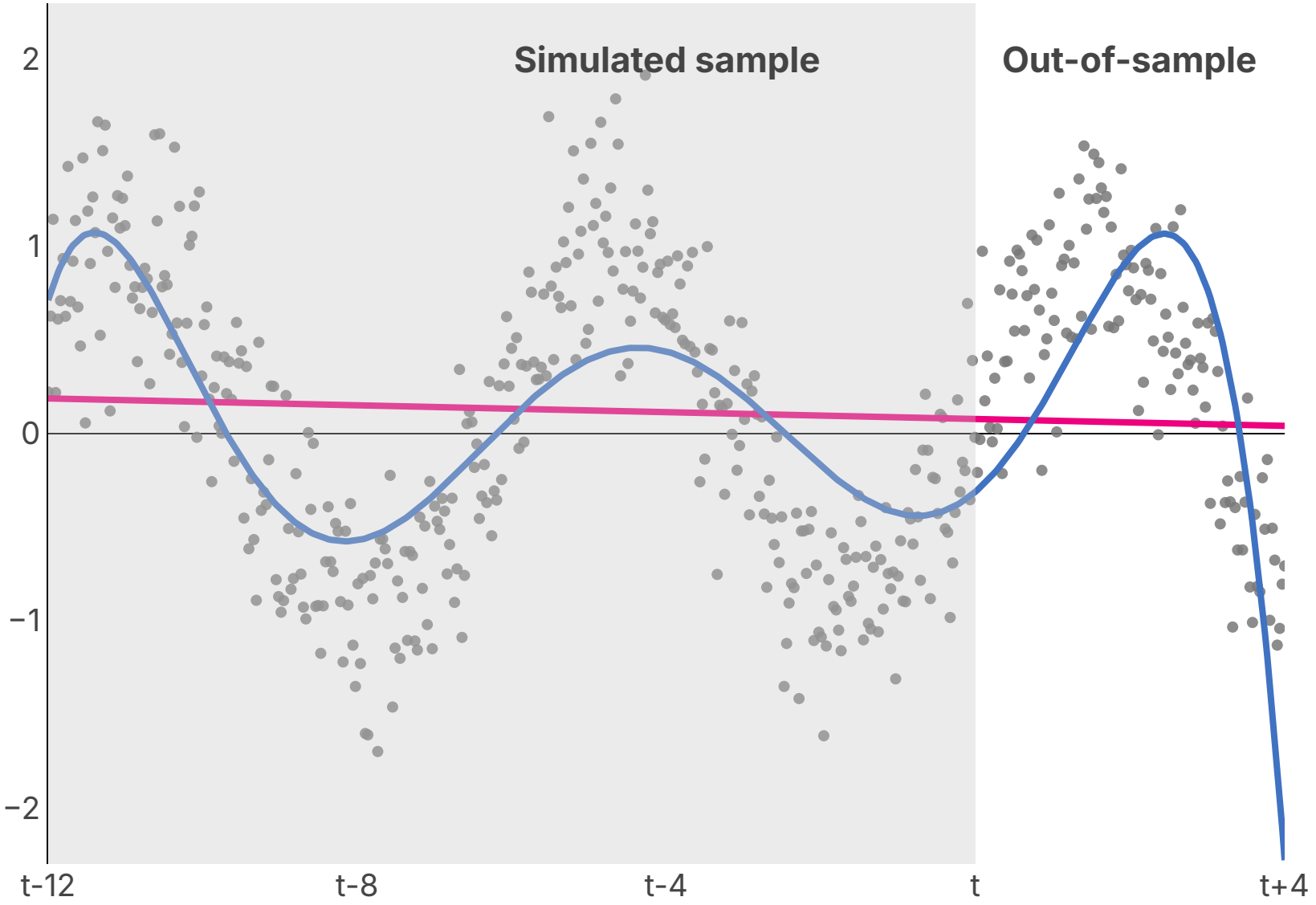


We start by fitting a **straight line**, i.e., the following linear model:

$$\boldsymbol{y} = \beta_0 + \boldsymbol{x}\beta_1.$$

We can see that the fit is far from perfect **in-sample**. **Out of sample**, it has comparable accuracy. The one parameter is easy to interpret, but we are missing out on important information.

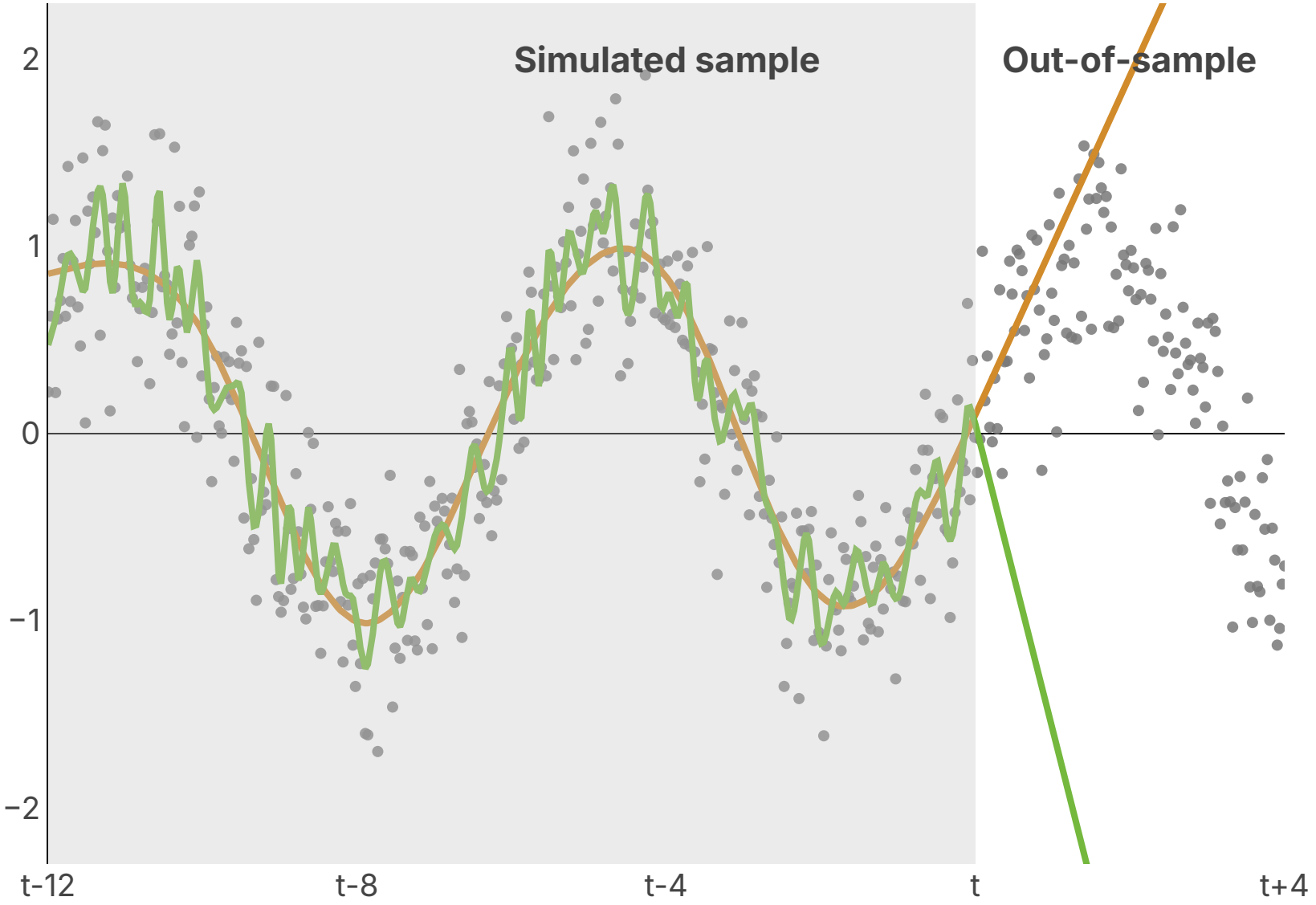
Parametric vs. Non-Parametric Fit



Next, we still fit a linear and parametric model, but we increase the number of a parameters by using a **sixth-order polynomial**.

We can see that the model fit improves, but we run into problems **out-of-sample**. The fit is bad there and gets infinitely worse.

Parametric vs. Non-Parametric Fit



Finally, we try out some **non-parametric models**. A **spline** is a piecewise-defined polynomial function. We are fitting one with **6 degrees of freedom** and one with **100 degrees of freedom**.

Both fit very well in-sample, and both do not perform perfectly out-of-sample. In the **latter case**, we are **blatantly overfitting**.

Supervised and Unsupervised Learning

Supervised Learning includes **everything we did so far**. We have **data on y** on which we can **train** our model, e.g.

- We run a **regression** of income on education. Our dataset includes income information as well as education information.
- We try to **classify** images in whether they picture a turtle or not. We have a dataset of images that includes information on whether the image shows a turtle or not.

Unsupervised Learning is when we train a model on **large amounts of data**, **without any labeling**.

Initially, an unsupervised model may have difficulty telling whether this image pictures a turtle.

Think

Have you ever been asked to tell a machine whether an image contains a traffic light?

Unsupervised Learning

Another example of **Unsupervised Learning** are **Large Language Models**. Since ChatGPT was released in late 2022, they have been pretty well known, but their development and public availability predates ChatGPT.

The first **Generative Pre-Trained Transformer (GPT)** was introduced by OpenAI in mid-2018. The model was called GPT-1, had 117 million parameters, and was trained on 7,000 unpublished books.

The model is able to complete a text prompt with meaningful sentences, but noticeably lacks context awareness.

```
This is our Econometrics II course. In this course, we will be able to calculate the trajectory of the nuclear bomb."

"what do you mean, calculate the trajectory?" "this is a simple calculation of the time it would take the bomb to detonate."

"we have three minutes," the technician said. "it could go either way."
```

Unsupervised Learning

The successor model, GPT-2, was published in early 2019. It was trained on text from documents and webpages that were upvoted on Reddit, and contained 1.5 billion parameters.

```
This is our Econometrics II course. In this course, we take a look at some of the statistical principles and statistics used in Econometrics.

We will take a look at a number of different statistics, both in terms of number of data items and of quality of data.
```

This model is more context-aware, but you can still easily see that it is generating words rather than meaning.

In 2020, OpenAI released GPT-3, which is no longer publicly available; and in 2022, it released GPT-3.5 These models have 175 billion parameters:

Unsupervised Learning

In 2020, OpenAI released GPT-3, which is no longer publicly available; and in 2022, it released GPT-3.5. These models have 175 billion parameters:

```
This is our Econometrics II course. In this course, we will learn advanced statistical
and econometric methods for analyzing economic data. This course will build upon the
foundational knowledge acquired in Econometrics I and will delve deeper into topics
such as panel data analysis, time series analysis, and instrumental variables. We will
also explore advanced topics such as endogeneity, selection bias, and nonlinear
models.
```

Of course, it does not know our syllabus, but this is a pretty reasonable guess for what a course entitled “Econometrics II” could be about.

Current models GPT-4o, GPT-4.5 and GPT-4.1 are rumored to have between 200 billion and 1 trillion parameters. The current product by Chinese competitor DeepSeek, DeepSeek V3, has 671 billion parameters. Claude 4 by Anthropic likely has a comparable number of parameters.

Regression and Classification

A **regression problem** is a problem with a **quantitative** dependent variable (e.g., height, econometrics grade, carbon emissions, ...).

In contrast, we refer to a problem with a **qualitative** dependent variable as a **classification problem**.

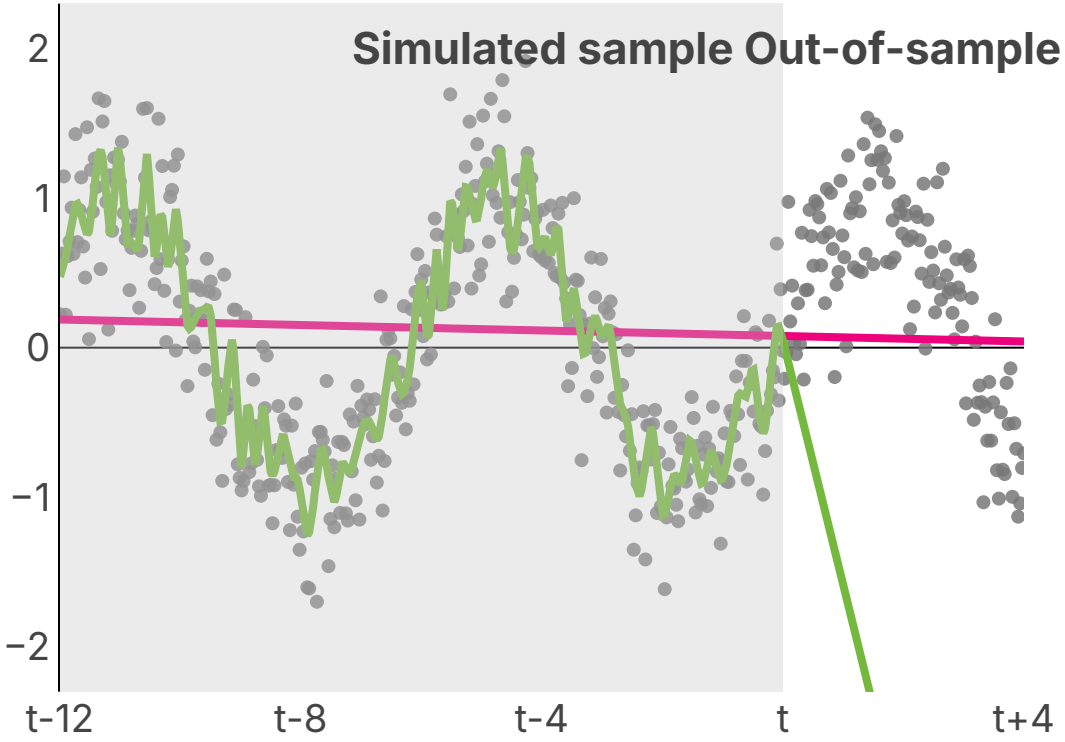
The distinction between the two is not always perfectly clear:

- **Linear regression** is undoubtedly a regression method.
- However, we can use linear regression with a **binary dependent variable**, yielding a Linear Probability Model (LPM). This can be viewed as a form of classification; or you could argue that class probabilities are still numbers and this is still regression.
- In any case, what matters for distinction is the **dependent variable**, while whether **explanatory variables** are quantitative or qualitative is generally less important.

The Interpretability-Flexibility Tradeoff

Different methods have different degrees of **flexibility**, i.e. they can only produce a narrow range of possible shapes of f. For example, linear regression is rather inflexible.

The **benefit** of choosing a flexible approach is evident. However, there is an important downside: **More flexible** methods yield results that are **less easy to interpret**.



The **linear fit** from before is relatively easy to interpret. We have a relationship that is goverened by one parameter:

$$\boldsymbol{y} = \boldsymbol{x}\beta + \boldsymbol{u}.$$

In contrast, the f we get from the **100-df spline** is extremely complicated, and it is difficult for us to understand how predictors relate to the Y values.

How to Choose a Model?

We choose a model and estimation method depending on the issue of interest, and the available data. Central questions we may ask ourselves include the following.

- What is the **goal** of our analysis?
  - How easy to interpret should our estimate $\hat{f}$ be?
  - Do we need to generate accurate predictions?
- What does our **data** look like?
  - How much data do we have (observations N, and covariates K)?
  - Are we dealing with a regression or classification problem?
- In what ways can we **test** our model?
- How much **time** (personal and computer) and money do we have to spare?

Prediction

Inference

Models

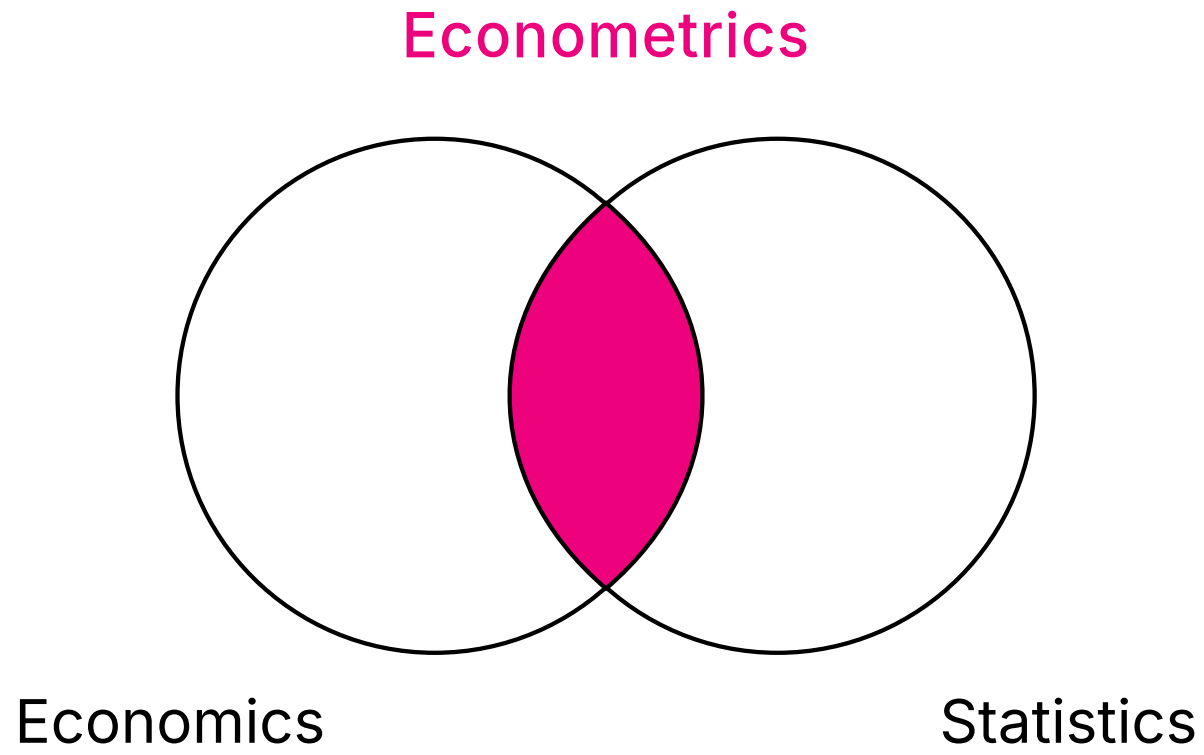
The Role of Econometrics

The Linear Model

Appendix

Economics, Statistics, Econometrics

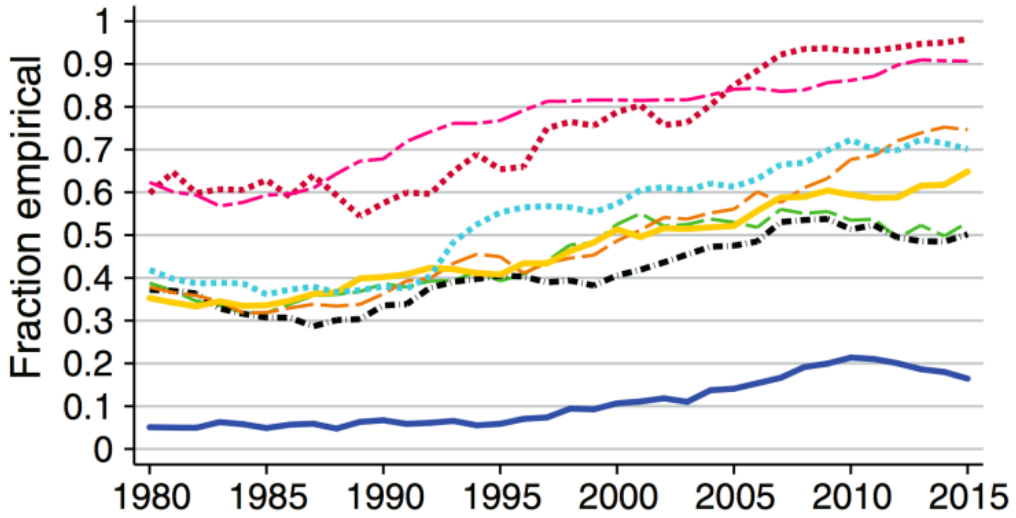
Econometrics seeks to **apply and develop statistical methods** to **learn about economic phenomena** using **empirical data**.



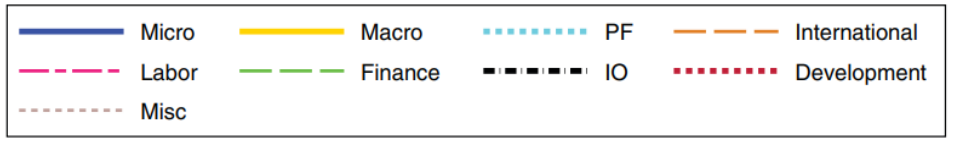
An Empirical Shift in Economics

Econometrics plays an important role in an **empirical shift in economic research**, away from pure theory (Angrist et al., 2017; Hamermesh, 2013). Today, economic theories are routinely confronted with real-world data.

> "Experience has shown that each [...] of statistics, economic theory, and mathematics, is a necessary [...] condition for a real understanding of the quantitative relations in modern economic life." — Ragnar Frisch (1933)



Weighted share of empirical publications in various economic fields (Angrist et al., 2017).



The Credibility Revolution

Econometric methods are **constantly developing**. There is **no one-size-fits-all approach** that fits any kind of data and research question. Econometrics has seen **considerable challenges and developments** since its inception. Important milestones concern

- uncertainty around model choice (e.g. Leamer, 1983; Steel, 2020),
- better research designs (e.g. Angrist & Pischke, 2010),
- randomised experiments (see Athey & Imbens, 2017),
- more flexible methods (Athey & Imbens, 2019).

You can be sure that the methods we learn today will **evolve and change** within the next years, during your carreer, and beyond. This gives you an opportunity to go into econometric research if you choose this career path, but it also means that you have to keep up with new developments.

Inference

Models

The Role of Econometrics

The Linear Model

Appendix

Why Is the Linear Model so Popular?

Consider how to transform the following **economic model** into an **econometric model**:

$$\text{wage} \approx f(\text{education}, \text{experience}).$$

A sensible choice might be the following **linear regression model**:

$$\mathbf{wage} = \mathbf{education}\,\beta_1 + \mathbf{experience}\,\beta_2 + \boldsymbol{u}.$$

Why is a linear model a **sensible choice** and why do we choose them **so often**?

- They are easy to interpret,
- parsimonious (meaning as simple and minimalistic as reasonably possible),
- and can be easily extended.

Goals of Econometrics

The linear model's popularity is not surprising, given the classical tasks:

- testing a theory — Does class size affect grades?,
- evaluating a policy — What are impacts of an oil embargo?,
- forecasting the future — How quickly do stocks go up?

The central task is arguably distilling a **causal effect** from *observational data*, since experimental data is rare.

When forecasting, economic theory can provide us with valuable **structural information**.

Linear Algebra and the Linear Model

The linear model is an essential building block, and **linear algebra** gives us a very convenient way of expressing and dealing with these models. Let

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u},$$

where the $N \times 1$ vector $\mathbf{y}$ holds the dependent variable for all N observations, and the $N \times K$ matrix $\mathbf{X}$ contains all K explanatory variables.

That is,

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1K} \\ x_{21} & x_{22} & \cdots & x_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{NK} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_N \end{pmatrix}.$$

The OLS Estimator

The **ordinary least squares** (OLS) estimator minimises the **sum of squared residuals**, which is given by $\hat{\boldsymbol{u}}'\hat{\boldsymbol{u}}$ (i.e. $\sum_{i=1}^{N} \hat{u}_i^2$). To find the estimate $\hat{\boldsymbol{\beta}}_{OLS}$, we

(1) re-express the **sum of squared residuals**,

(2) find an **extreme value** via the partial derivative ($\frac{\partial \hat{\boldsymbol{u}}'\hat{\boldsymbol{u}}}{\partial \hat{\boldsymbol{\beta}}} = 0$),

(3) check whether we found a **minimum** via the second partial derivative.

$$\begin{aligned}\hat{\boldsymbol{u}}'\hat{\boldsymbol{u}} &= (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \\ &= \boldsymbol{y}'\boldsymbol{y} - 2\hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y} + \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}.\end{aligned}$$

$$\frac{\partial \hat{\boldsymbol{u}}'\hat{\boldsymbol{u}}}{\partial \hat{\boldsymbol{\beta}}} = -2\boldsymbol{X}'\boldsymbol{y} + 2\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}, \qquad \frac{\partial^2 \hat{\boldsymbol{u}}'\hat{\boldsymbol{u}}}{\partial^2 \hat{\boldsymbol{\beta}}} = 2\boldsymbol{X}'\boldsymbol{X}.$$

The estimator $\hat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ is directly available and a minimum. • derivation

References

Angrist, J. D., Azoulay, P., Ellison, G., Hill, R., & Lu, S. F. (2017). Economic research evolves: Fields and styles. *American Economic Review*, *107*(5), 293–297. https://doi.org/10.1257/aer.p20171117

Angrist, J. D., & Pischke, J.-S. (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of Economic Perspectives*, *24*(2), 3–30. https://doi.org/10.1257/jep.24.2.3

Athey, S., & Imbens, G. W. (2017). The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives*, *31*(2), 3–32. https://doi.org/10.1257/jep.31.2.3

Athey, S., & Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics*, *11*(1), 685–725. https://doi.org/10.1146/annurev-economics-080217-053433

Box, G. E. P. (1976). Science and statistics. *Journal of the American Statistical Association*, *71*(356), 791–799. https://doi.org/10.1080/01621459.1976.10480949

Hamermesh, D. S. (2013). Six decades of top economics publishing: Who and how? *Journal of Economic Literature*, *51*(1), 162–172. https://doi.org/10.1257/jel.51.1.162

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning*. Springer US. https://doi.org/10.1007/978-1-0716-1418-1

Leamer, E. E. (1983). Let's take the con out of econometrics. *American Economic Review*, *73*(1), 31–43. https://www.jstor.org/stable/1803924

Steel, M. F. J. (2020). Model averaging and its use in economics. *Journal of Economic Literature*, *58*(3), 644–719. https://doi.org/10.1257/jel.20191385

Vigen, T. (2024). *Frozen yogurt consumption correlates with violent crime rates*. https://www.tylervigen.com/spurious/correlation/5905_frozen-yogurt-consumption_correlates-with_violent-crime-rates

Models

The Role of Econometrics

The Linear Model

Appendix

Reducible and Irreducible Error – Decomposition

We have $\boldsymbol{y} = f(\boldsymbol{X}) + \boldsymbol{u}$, $\hat{\boldsymbol{y}} = \hat{f}(\boldsymbol{X})$, and $\mathrm{E}(\boldsymbol{u}) = 0$, Recall that $\mathrm{Var}(\boldsymbol{u}) = \mathrm{E}\big((\boldsymbol{u} - \mathrm{E}(\boldsymbol{u}))^2\big)$.

$$
\begin{aligned}
\mathrm{E}\big((\boldsymbol{y} - \hat{\boldsymbol{y}})^2\big) &= \mathrm{E}\big((f(\boldsymbol{X}) + \boldsymbol{u} - \hat{f}(\boldsymbol{X}))^2\big) \\
&= \mathrm{E}\big(((f(\boldsymbol{X}) - \hat{f}(\boldsymbol{X})) + \boldsymbol{u})^2\big) \\
&= \mathrm{E}\big((f(\boldsymbol{X}) - \hat{f}(\boldsymbol{X}))^2 + 2\boldsymbol{u}(f(\boldsymbol{X}) - \hat{f}(\boldsymbol{X})) + \boldsymbol{u}^2\big) \\
&= \mathrm{E}\big((f(\boldsymbol{X}) - \hat{f}(\boldsymbol{X}))^2\big) + \mathrm{E}\big(2\boldsymbol{u}(f(\boldsymbol{X}) - \hat{f}(\boldsymbol{X}))\big) + \mathrm{E}\big(\boldsymbol{u}^2\big) \\
&= \mathrm{E}\big((f(\boldsymbol{X}) - \hat{f}(\boldsymbol{X}))^2\big) + 0 + \mathrm{E}\big(\boldsymbol{u}^2\big) \\
&= \mathrm{E}\big((f(\boldsymbol{X}) - \hat{f}(\boldsymbol{X}))^2\big) + \mathrm{Var}(\boldsymbol{u}).
\end{aligned}
$$

• go back

Bias and Variance – Decomposition

We use the shorthands $f = f(\boldsymbol{X})$ and $\hat{f} = \hat{f}(\boldsymbol{X})$. Recall that $\mathrm{Bias}(\hat{f}) = \mathrm{E}(\hat{f}) - f$.

$$
\begin{aligned}
\mathrm{E}\big((\boldsymbol{y} - \hat{\boldsymbol{y}})^2\big) &= \mathrm{E}\big((f - \hat{f})^2\big) + \mathrm{Var}(\boldsymbol{u}) \\
&= \mathrm{E}\big((f - \mathrm{E}(\hat{f}) + \mathrm{E}(\hat{f}) - \hat{f})^2\big) + \mathrm{Var}(\boldsymbol{u}) \\
&= \mathrm{E}\big(((f - \mathrm{E}(\hat{f})) + (\mathrm{E}(\hat{f}) - \hat{f}))^2\big) + \mathrm{Var}(\boldsymbol{u}) \\
&= \mathrm{E}\big((f - \mathrm{E}(\hat{f}))^2\big) + 2\,\mathrm{E}\big((f - \mathrm{E}(\hat{f}))(\mathrm{E}(\hat{f}) - \hat{f})\big) + \mathrm{E}\big((\mathrm{E}(\hat{f}) - \hat{f})^2\big) + \mathrm{Var}(\boldsymbol{u}) \\
&= (f - \mathrm{E}(\hat{f}))^2 + 0 + \mathrm{E}\big((\mathrm{E}(\hat{f}) - \hat{f})^2\big) + \mathrm{Var}(\boldsymbol{u}) \\
&= \mathrm{Bias}(\hat{f})^2 + \mathrm{Var}(\hat{f}) + \mathrm{Var}(\boldsymbol{u}).
\end{aligned}
$$

• go back

OLS estimator – Derivartion

We have $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \hat{\boldsymbol{u}}$, which lets us re-express the sum of squared residuals as

$$\begin{aligned}\hat{\boldsymbol{u}}'\hat{\boldsymbol{u}} &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = (\boldsymbol{y}' - \boldsymbol{\beta}'\boldsymbol{X}')(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\ &= \boldsymbol{y}'\boldsymbol{y} - \boldsymbol{y}'\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} \\ &= \boldsymbol{y}'\boldsymbol{y} - 2\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta},\end{aligned}$$

where we use the fact that for a scalar $\alpha = \alpha'$ to simplify $\boldsymbol{y}'\boldsymbol{X}\boldsymbol{\beta} = (\boldsymbol{y}'\boldsymbol{X}\boldsymbol{\beta})' = \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{y}$. Next, we set the first derivative $\frac{\partial \hat{\boldsymbol{u}}'\hat{\boldsymbol{u}}}{\partial \boldsymbol{\beta}} = -2\boldsymbol{X}'\boldsymbol{y} + 2\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}$ to zero

$$\begin{aligned}-2\boldsymbol{X}'\boldsymbol{y} + 2\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} &= 0 \\ \boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} &= \boldsymbol{X}'\boldsymbol{y} \\ \boldsymbol{\beta} &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y}.\end{aligned}$$

The second partial derivative $2\boldsymbol{X}'\boldsymbol{X}$ is positive (definite) as long as it is invertible.

• go back