PhD Course on Econometrics (Economic Modelling), Part I

Version: 29th Apr 2019[1]

Dr Povilas Lastauskas (PhD Cantab)

CEFER
Bank of Lithuania

Spring Semester 2019

[1] Errors and omissions please report to P.Lastauskas@cantab.net, www.lastauskas.com.

What Are the Goals of this Part?

- **Mainly revise** some basic econometric concepts, tools and methods that you will employ throughout your PhD studies.
- Provide a very selected introduction to some areas in empirical economics.
  - Emphasis is on **causal inference** using observational data (as opposed to experimental data).
- By the end of this part: you should have consolidated knowledge in a number of econometric concepts and methods, also be comfortable with applied work using econometric techniques, be able to follow some applied literature, and, most importantly, be able to translate an economic problem into an econometric one.
  - Emphasis on **exogeneity** and **endogeneity** as well as causal inference.
  - Due to heterogeneity in background knowledge, technicalities are kept to minimum for PhD classes; nevertheless, the course is *conceptually rigorous*.

What Are the Main Topics?

- According to Angrist and Pischke (2008), the most important items in an applied econometrician's toolkit are:

1. **Regression models** designed to control for variables that may mask the causal effects of interest;
2. **Instrumental variables methods** for the analysis of real and natural experiments;
3. Diff-in-diff, matching, time series methods for dynamic causal effects as well as panel methods covered in later parts of the course.

In addition, we will cover:

- **Very concise overview** of the start-of-the-art methods, including a few ideas from the **machine-learning** field.

Structure of This Part

- Theory + Problem Set + Paper Replication
  - The first part of the meeting is spent on the theoretical notions, some proofs, and examples.
  - The second part of the meeting is devoted to (**a part of**) the problem set: **your participation is therefore crucial**!
- There will also be a homework for the research paper to be read and replicated.

Literature

- **Main texts**:
  - Angrist and Pischke, Mostly Harmless Econometrics: An Empiricist's Companion (Princeton University Press, 2008).
  - Stachurski, A Primer in Econometric Theory (MIT Press, 2016).
- Intermediate econometrics:
  - Stock, J. H. and M. W. Watson: *Introduction to Econometrics,* Third Edition, Pearson Education, 2014.
  - Wooldridge, Jeffrey M.: *Introductory Econometrics: A Modern Approach*, Fifth Edition, Cengage Learning, 2013.
- More advanced micro and macro-econometrics, respectively:
  - Cameron, A. C. and P. Trivedi, *Microeconometrics: Methods and Applications*, Cambridge University Press, 2005.
  - Pesaran, M. Hashem: *Time Series and Panel Data Econometrics*. Oxford University Press, 2015.

Literature

- Other readings include:
  - Acemoglu, Daron, Simon Johnson and James A. Robinson (2001). "The Colonial Origins of Comparative Development: An Empirical Investigation". *American Economic Review*. 91 (5): 1369-1401. `http://economics.mit.edu/faculty/acemoglu/data/ajr2001`
  - Angrist, Joshua and Alan Krueger (1991). "Does Compulsory School Attendance Affect Schooling and Earnings?", *Quarterly Journal of Economics*, 106 (4): 979-1014.
  - Angrist, Joshua (1990). "Lifetime Earnings and the Vietnam Draft Lottery: Evidence from Social Security Administrative Records". *American Economic Review*. 80 (3): 313-336.
  - Card, David and Alan B. Krueger (1994). "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania". *American Economic Review*. 84 (4): 772-793.
  - Gobillon, L. and T. Magnac (2016). "Regional Policy Evaluation: Interactive Fixed Effects and Synthetic Controls". *The Review of Economics and Statistics*, 98(3): 535-551.

Exam

- **Eighty percent of success is showing up** – *Woody Allen*
- More details on exam will be covered in the second part on Microeconometrics by Dr Swapnil Singh.

Outline

Statistics Vocabulary

- Before moving to the theory, we need to be precise about a few statistical concepts.

Definition

A **sample** is a collection of random variables (say $X_1, X_2, \ldots, X_n$) from the population. If the random variables constituting the sample are independently and identically distributed, then the sample is said to be **random**.

- What is a statistic?

Definition

A **statistic** is a function of the sample (say $\hat{\theta} = f(X_1, X_2, \ldots, X_n)$. Therefore, a **statistic** is a **random variable**.

Statistics Vocabulary

Examples

Some useful statistics are: sample mean, sample variance, and sample covariance (when we jointly sample two variables X and Y).

Definition

(**Estimator** and **Estimate**) An **estimator** is a statistic we use to assign a value to an unknown parameter. An **estimate** is the actual value assumed by the estimator. In other words, the estimator is a random variable, and thus has a distribution; an estimate is a particular realised value of the estimator.

- Estimators are described by a set of behavioural properties.

Statistics Vocabulary

- Estimator is said to be **unbiased** for a function θ if it equals θ in expectation:
$$\mathbb{E}_\theta f\left(X_1, X_2, \ldots, X_n\right) = \mathbb{E}_\theta \hat{\theta} = \theta.$$
- An estimator $\hat{\theta}$ is said to be **consistent** if it converges to θ in probability, i.e. if for all $\epsilon > 0$,
$$\lim_{n \to \infty} \Pr\left(\left|\hat{\theta} - \theta\right| > \epsilon\right) = 0.$$
  - This is an asymptotic property when sample size tends to infinity.
- An **efficient** estimator is the minimum variance unbiased estimator.
  - For some estimators, they can attain efficiency asymptotically and are thus called **asymptotically efficient estimators**.

Properties of Expectations

Proposition

By Law of Iterated Expectations, $\mathbb{E}_X\mathbb{E}(Y \mid X) = \mathbb{E}(Y)$.

Proof.

$\mathbb{E}_X\mathbb{E}(Y \mid X) = \mathbb{E}_X\left(\sum_y yP(Y = y \mid X = x)\right) = \sum_x \left(\sum_y yP(Y = y \mid X = x)\right) P(X = x)$.
Rewrite

$$\sum_x \left(\sum_y yP(Y = y \mid X = x)\right) P(X = x) = \sum_x \sum_y yP(Y = y \mid X = x) P(X = x).$$

Note that $P(Y = y \mid X = x) P(X = x) = P(X = x \mid Y = y) P(Y = y)$ *since, recall, you can express* $P(Y \cap X)$ *both ways. Then,*

$$\begin{aligned}\sum_x \sum_y yP(Y = y \mid X = x) P(X = x) &= \sum_x \sum_y yP(X = x \mid Y = y) P(Y = y) \\ &= \sum_y yP(Y = y)\left(\sum_x P(X = x \mid Y = y)\right),\end{aligned}$$

Notice that $\sum_x P(X = x \mid Y = y) = 1$. *Hence,* $\mathbb{E}_X\mathbb{E}(Y \mid X) = \sum_y yP(Y = y) = \mathbb{E}(Y)$.

Properties of Expectations

Proposition

By Law of Iterated Expectations, $\mathbb{E}_X \mathbb{E}(Y \mid X) = \mathbb{E}(Y)$.

Proof.

(Alternative)

$$
\begin{aligned}
& E[E[Y|X]] = \int E[Y|X] f(X) dX \\
= & \int \left(\int Y f(Y|X) dY \right) f(X) dX = \int \int Y f(Y|X) f(X) dY dX \\
= & \int \int Y f(X,Y) dY dX = \int Y \left(\int f(X,Y) dX \right) dY \\
= & \int Y f(Y) dY = E[Y].
\end{aligned}
$$

□

Properties of Expectations

- For constants a and b,
  - $\mathbb{E}(aX) = a\mathbb{E}X$, multiplicative constants come outside expectations.
  - $\mathbb{E}(aX + b) = a\mathbb{E}X + b$, additive constants come outside the expectation.
  - $\mathbb{E}(aX + bY) = a\mathbb{E}X + b\mathbb{E}Y$, expectations pass through sums.
  - $Var(X) = \mathbb{E}\left[(X - \mathbb{E}X)^2\right]$ and $Cov(X, Y) = \mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)]$.

Convergence in Probability

- Most estimators are functions of sample means, so focus on what happens to these means in large samples.
- We define asymptotic arguments with respect to the sample size n.
- Let x_n by a sequence random variable (a set of random variables $\{x_1, x_2, \ldots, x_n\}$) indexed by its sample size n.

Definition

The random variable x_n ***converges in probability*** *to a constant* c *if* $\lim_{n\to\infty} \Pr(|x_n - c| > \varepsilon) = 0$ *for any* $\varepsilon > 0$

- The probability that x takes values far from c disappears as $n \to \infty$.
- If x_n converges in probability to c, we say that plim $x_n = c$.

Convergence in Mean Square

- It is often hard to verify convergence in probability, so we usually use a special case.

Definition (Convergence in Mean Square)

*If x_n has mean μ_n and variance σ_n^2 such that with limits c and 0, respectively, then x_n **converges in mean square** to c and plim $x_n = c$*

- Easier to check that the mean and variance have limits.
- Note that mean square convergence implies convergence in probability, but not vice versa.

Definition

*An estimator $\hat{\theta}_n$ of a parameter θ is a **consistent** estimator of θ if and only if*

$$\text{plim } \hat{\theta}_n = \theta.$$

Consistency

- In fact, the example of the sample mean generalises:

Remark

For any function $g(x)$, if $\mathbb{E}[g(x)]$ and $\text{Var}[g(x)]$ are finite constants, then

$$\text{plim } \frac{1}{n}\sum_{i=1}^{n} g(x_i) = \mathbb{E}[g(x)].$$

The Law of Large Numbers and Slutsky's Theorem

Theorem (*Law of Large Numbers*)

If x_i $i = 1, \ldots, n$ *is a random (i.i.d.) sample from a distribution with finite mean* $\mathbb{E}[x_i] = \mu < \infty$*, then*

$$plim\ \bar{x}_n = \mu.$$

Theorem (Slutsky's Theorem)

For a continuous function $g(x_n)$ *that is not a function of* n*,*

$$plim\ g(x_n) = g(plim\ x_n).$$

- These results are very powerful and allow us to show that **estimators** (functions of the data x_n) **are consistent**.

Properties of plims

- Probability limits have a number of useful properties: If x_n and y_n are RVs with plim $x_n = c$ and plim $y_n = d$, then:

1. plim $(x_n + y_n) = c + d$,
2. plim $x_n y_n = cd$,
3. plim $x_n / y_n = c/d$ as long as $d \neq 0$.

Remark

If $\mathbf{W}_n$ is a matrix whose elements are RVs, and if plim $\mathbf{W}_n = \mathbf{\Omega}$

$$\text{plim } \mathbf{W}_n^{-1} = \mathbf{\Omega}^{-1}.$$

If $\mathbf{X}_n$ and $\mathbf{Y}_n$ are random matrices with plim $\mathbf{X}_n = \mathbf{A}$ and plim $\mathbf{Y}_n = \mathbf{B}$

$$\text{plim } \mathbf{X}_n \mathbf{Y}_n = \mathbf{AB}.$$

Convergence in Distribution

- We use the *plim* to analyze whether estimators are *consistent.*
- In order to make inference (for instance, is a coefficient $= 0$?), we need to know the *distribution* of the estimator.

Definition (Convergence in Distribution)

x_n converges in distribution to a random variable x with CDF $F(x)$ if $\lim_{n\to\infty} |F_n(x_n) - F(x)| = 0$ over the whole support of $F(x)$. We denote this as

$$x_n \xrightarrow{d} x.$$

Rules for Convergence in Distribution

- Analogously to the rules for plims, if $x_n \xrightarrow{d} x$ and plim $y_n = c$, then

$$x_n y_n \xrightarrow{d} cx,$$
$$x_n + y_n \xrightarrow{d} x + c,$$
$$x_n / y_n \xrightarrow{d} x/c \quad \text{if } c \neq 0.$$

- If $x_n \xrightarrow{d} x$ and $g(x_n)$ is a continuous function, then

$$g(x_n) \xrightarrow{d} g(x).$$

- If $y_n \xrightarrow{d} y$ and plim $(x_n - y_n) = 0$, then $x_n \xrightarrow{d} y$.
- If $\mathbf{x}_n \xrightarrow{d} \mathbf{x}$ then $\mathbf{c}'\mathbf{x}_n \xrightarrow{d} \mathbf{c}'\mathbf{x}$.

Asymptotic Normality and the Central Limit Theorem

- In principle, if plim $\hat{\theta}_n = \theta$, then $\hat{\theta}_n \xrightarrow{d} \theta$ and the limiting distribution of $\hat{\theta}_n$ is a spike.
- Of course, we don't think that in any given sample this is a reasonable thing to assume.
- Instead, to get more reasonable statistical properties of the estimator, we use a **stabilising transformation**.

Asymptotic Normality and the Central Limit Theorem

Univariate Central Limit Theorem

If $x_1, x_2, \ldots, x_n$ are a random sample from a distribution with mean $\mu < \infty$ and variance $\sigma^2 < \infty$ and $\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$, then

$$\sqrt{n}\left(\bar{x}_n - \mu\right) \xrightarrow{d} N\left[0, \sigma^2\right].$$

Theorem (Multivariate Central Limit Theorem)

If $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ are a random sample from a multivariate distribution with finite mean vector $\boldsymbol{\mu}$ and finite covariance matrix $\mathbf{Q}$, and $\bar{\mathbf{x}}_n = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i$, then

$$\sqrt{n}\left(\bar{\mathbf{x}}_n - \boldsymbol{\mu}\right) \xrightarrow{d} N\left[\mathbf{0}, \mathbf{Q}\right].$$

Outline

Conditional Expectation Function

Definition

The Conditional Expectation Function (CEF) for a dependent variable Y_i, given covariates X_i, is the expectation – or the population average – of Y_i with X_i held fixed.

- Denote by

$$\mathbb{E}\left[Y_i \mid X_i\right] = \begin{cases} \int Y_i f_Y\left(Y_i \mid X_i = x\right) dY_i, & \text{continuous,} \\ \sum Y_i P\left(Y_i \mid X_i = x\right), & \text{discrete.} \end{cases}$$

- Hence, CEF is random as it is a function of X_i which is random.
- Further, it is a *population* concept which the researcher attempts to uncover using the sample CEF.

Properties of Conditional Expectation Function

Claim

CEF decomposition property.

- By decomposition property,

$$Y_i = \mathbb{E}[Y_i \mid X_i] + \varepsilon_i,$$

where ε_i is mean independent of X_i, $\mathbb{E}[\varepsilon_i \mid X_i] = 0$. Notice that this follows from

$$\mathbb{E}[\varepsilon_i \mid X_i] = \mathbb{E}[Y_i \mid X_i] - \mathbb{E}[[\mathbb{E}[Y_i \mid X_i]] \mid X_i] = 0.$$

- Importantly, we can demonstrate that this property produces a result that ε_i is uncorrelated with *any* function of X_i, i.e., let $h(X_i)$ be such a function of X_i, then

$$\mathbb{E}[h(X_i)\,\varepsilon_i] = \mathbb{E}[\mathbb{E}[(h(X_i)\,\varepsilon_i) \mid X_i]] = \mathbb{E}[h(X_i)\,\mathbb{E}(\varepsilon_i \mid X_i)] = 0.$$

- We employed the Law of Iterated Expectations – refer to this Slide.

Conditional Expectation Function

Claim

Best predictor property.

- $\mathbb{E}[Y_i \mid X_i]$ is the Best (minimum mean squared error, MMSE) predictor of Y_i in that it minimises the function $\mathbb{E}(Y_i - h(X_i))^2$, or

$$\mathbb{E}[Y_i \mid X_i] = \arg\min_{h(X_i)} \mathbb{E}\left[(Y_i - h(X_i))^2\right],$$

  where $h(X_i)$ is any function of X_i.
- The proof follows immediately by subtracting and adding $\mathbb{E}[Y_i \mid X_i]$ inside the brackets, so that

$$\mathbb{E}(Y_i - h(X_i))^2 = \mathbb{E}\Big([Y_i - \mathbb{E}[Y_i \mid X_i]]^2 + 2[Y_i - \mathbb{E}[Y_i \mid X_i]] \times [\mathbb{E}[Y_i \mid X_i] - h(X_i)] + [\mathbb{E}[Y_i \mid X_i] - h(X_i)]^2\Big).$$

- The first term does not involve $h(X_i)$, the second one is zero from the decomposition property, and the last one is minimised at zero, yielding a result that $h(X_i)$ is CEF.

Regression

- We have not specified $h(X_i)$ so far and not really linked CEF to *regression*.

Definition

The population regression of Y on X is given by $\mathbb{E}[Y \mid X]$. The disturbance or error term of the regression is defined by $\varepsilon = Y - \mathbb{E}[Y \mid X]$. X is called the regressor or explanatory variable. Y is called the regressand or explained variable.

- It, therefore, follows that we can write the population regression as

$$Y = \mathbb{E}[Y \mid X] + \varepsilon.$$

- If we interpret $\mathbb{E}[Y \mid X]$ as a prediction of Y for given X, then ε is the prediction error.
- In the simple linear regression model, we assume that

$$\mathbb{E}[Y \mid X] = \alpha + \beta X, \quad \mathbb{E}(\varepsilon|X) = 0.$$

Why Regression?

- The following are important *reasons* to concentrate on the regression:

1. Regression solves the population least squares problem and is, therefore, the best linear predictor of Y_i given X_i;
2. If the CEF is linear, regression is exactly it;
3. Regression gives the best linear approximation to the CEF.

- The first claim is true by definition. The second one follows from the orthogonality, $\mathbb{E}(\varepsilon|X) = 0$. The third one can be proved by showing that $\hat{\beta}X$ is the minimum mean square error linear approximation to $\mathbb{E}[Y_i \mid X_i]$, i.e., $\beta = \arg\min \mathbb{E}\left(\mathbb{E}[Y_i \mid X_i] - \hat{\beta}X_i\right)^2$.

Orthogonality

- Let $\mathbf{x}$ and $\mathbf{z}$ be vectors in $\mathbb{R}^N$.
- If $\langle \mathbf{x}, \mathbf{z} \rangle = 0$, then we call $\mathbf{x}$ and $\mathbf{z}$ orthogonal.
- Write $\mathbf{x} \perp \mathbf{z}$.
- In $\mathbb{R}^2$, **orthogonal means perpendicular**.
- Note that if S is a linear subspace, then we say that $\mathbf{x}$ is orthogonal to S if $\mathbf{x} \perp \mathbf{z}$ for all $\mathbf{z} \in S$, and denote by $\mathbf{x} \perp S$.

Orthogonality

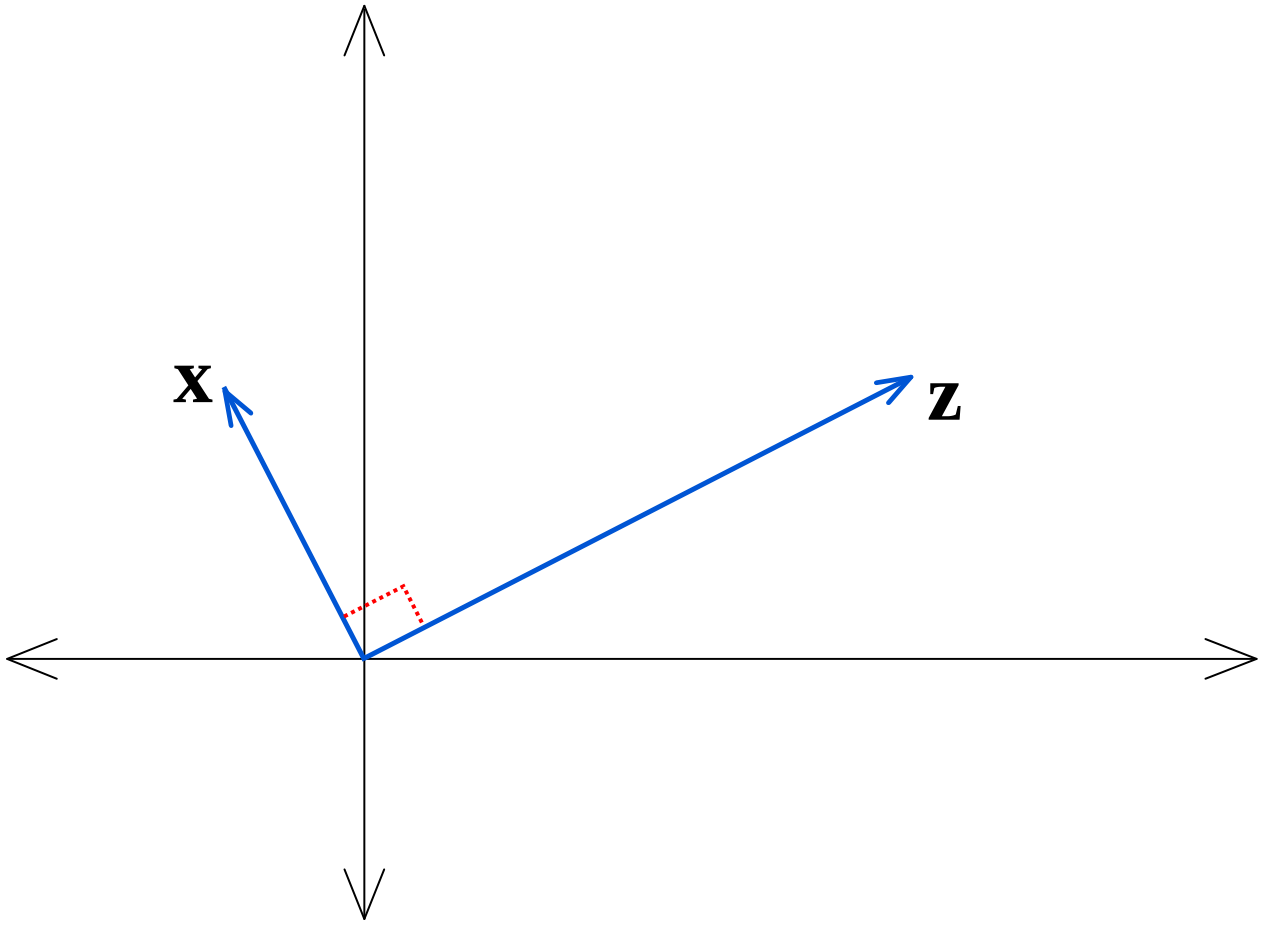


Figure: $\mathbf{x} \perp \mathbf{z}$

Problem

- Suppose you are asked to solve the following problem:

Problem

Given $\mathbf{y} \in \mathbb{R}^N$ and subspace S, find closest element of S to $\mathbf{y}$.

- To formalise a bit, you are confronted with

$$\mathbf{y} = \arg\min_{\hat{\mathbf{y}} \in S} \|\mathbf{y} - \hat{\mathbf{y}}\|. \tag{1}$$

- Existence and uniqueness of solution are not immediately obvious.
- Orthogonal projection theorem: $\hat{\mathbf{y}}$ always exists and is unique. It also provides a useful characterisation of solution.

Solution

Orthogonal Projection Theorem I

Let $\mathbf{y} \in \mathbb{R}^N$ *and let* S *be any nonempty linear subspace of* $\mathbb{R}^N$.

The following statements are true:

1. The optimization problem (1) has exactly one solution.
2. $\hat{\mathbf{y}} \in \mathbb{R}^N$ solves (1) if and only if $\hat{\mathbf{y}} \in S$ and $\mathbf{y} - \hat{\mathbf{y}} \perp S$

- The unique solution $\hat{\mathbf{y}}$ is called the orthogonal projection of $\mathbf{y}$ onto S.

Orthogonal Projection

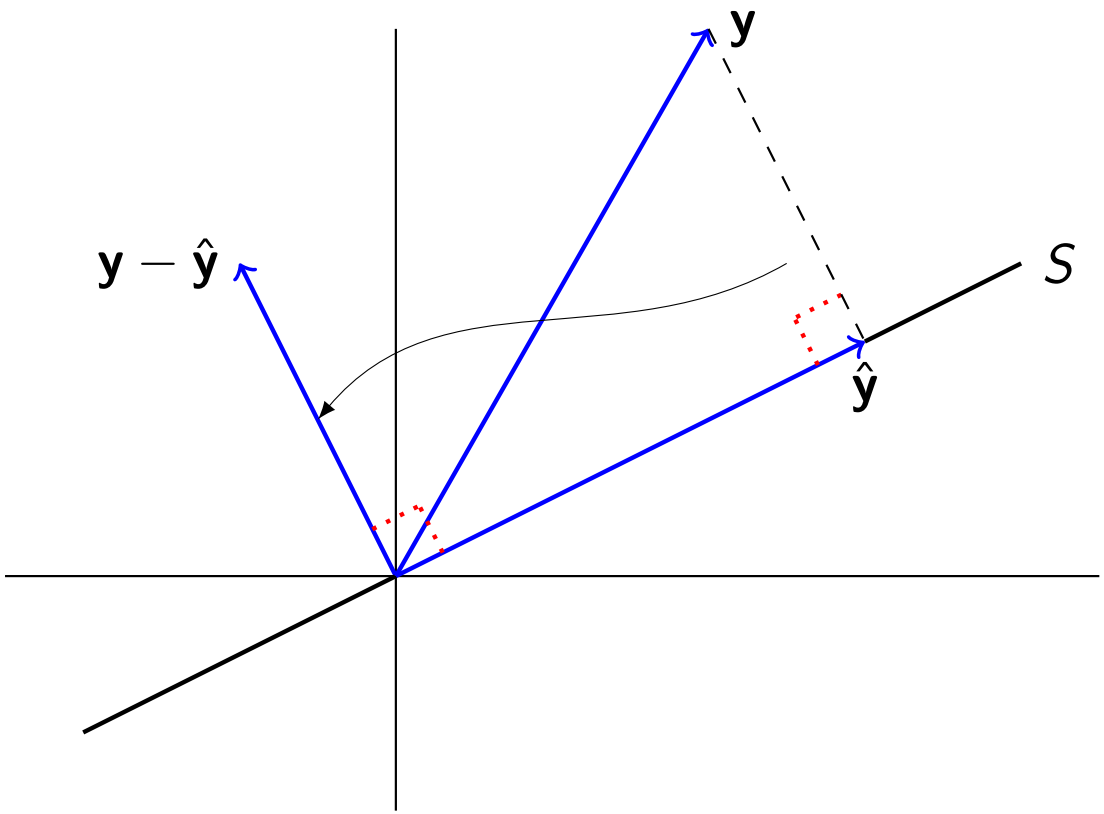


Figure: Orhogonal projection

Projections

- Holding subspace S fixed, we have a functional relationship

$$\mathbf{y} \mapsto \text{ its orthogonal projection } \hat{\mathbf{y}} \in S.$$

  This is a well-defined function from $\mathbb{R}^N$ to $\mathbb{R}^N$.
- The function is typically denoted by the projection operator, $\mathbf{P}$.
- We let $\mathbf{Py}$ represent $\hat{\mathbf{y}}$.
- $\mathbf{P}$ is called the orthogonal projection mapping onto S and we write

$$\mathbf{P} = \text{proj}S.$$

Orthogonal Projection Theorem II

Let S be any linear subspace of $\mathbb{R}^N$ and let $\mathbf{P} = projS$. The following statements are true:

1. $\mathbf{P}$ is a linear function.

Moreover, for any $\mathbf{y} \in \mathbb{R}^N$, we have

1. $\mathbf{Py} \in S$,
2. $\mathbf{y} - \mathbf{Py} \perp S$,
3. $\|\mathbf{y}\|^2 = \|\mathbf{Py}\|^2 + \|\mathbf{y} - \mathbf{Py}\|^2$,
4. $\|\mathbf{Py}\| \leq \|\mathbf{y}\|$,
5. $\mathbf{Py} = \mathbf{y}$ if and only if $\mathbf{y} \in S$,
6. $\mathbf{Py} = \mathbf{0}$ if and only if $\mathbf{y} \in S^{\perp}$.

Projections

- Project $\mathbf{y}$ onto S, where S is a linear subspace of $\mathbb{R}^N$.
- Closest point to $\mathbf{y}$ in S is $\hat{\mathbf{y}} = \mathbf{P}\mathbf{y}$ here $\mathbf{P} = \text{proj}S$.
- Unless $\mathbf{y}$ was already in S, some error $\mathbf{y} - \mathbf{P}\mathbf{y}$ remains.
- Introduce operator $\mathbf{M}$ that takes $\mathbf{y} \in \mathbb{R}^N$ and returns the residual projection

$$\mathbf{M} = \mathbf{I} - \mathbf{P}, \tag{2}$$

  where $\mathbf{I}$ is the identity mapping onto $\mathbb{R}^N$.
- For any $\mathbf{y}$ we have $\mathbf{M}\mathbf{y} = \mathbf{I}\mathbf{y} - \mathbf{P}\mathbf{y} = \mathbf{y} - \mathbf{P}\mathbf{y}$. (In regression analysis $\mathbf{M}$ shows up as a matrix called the 'annihilator' or the 'residual maker').
- We refer to $\mathbf{M}$ as the residual projection.

Projections

Fact

Let S be a linear subspace of $\mathbb{R}^N$, let $\mathbf{P} = projS$, and let $\mathbf{M}$ be the residual projection as defined in (2). The following statements are true:

1. $\mathbf{M} = \text{proj}S^{\perp}$,
2. $\mathbf{y} = \mathbf{P}\mathbf{y} + \mathbf{M}\mathbf{y}$ for any $\mathbf{y} \in \mathbb{R}^N$,
3. $\mathbf{P}\mathbf{y} \perp \mathbf{M}\mathbf{y}$ for any $\mathbf{y} \in \mathbb{R}^N$,
4. $\mathbf{M}\mathbf{y} = \mathbf{0}$ if and only if $\mathbf{y} \in S$.

Projections

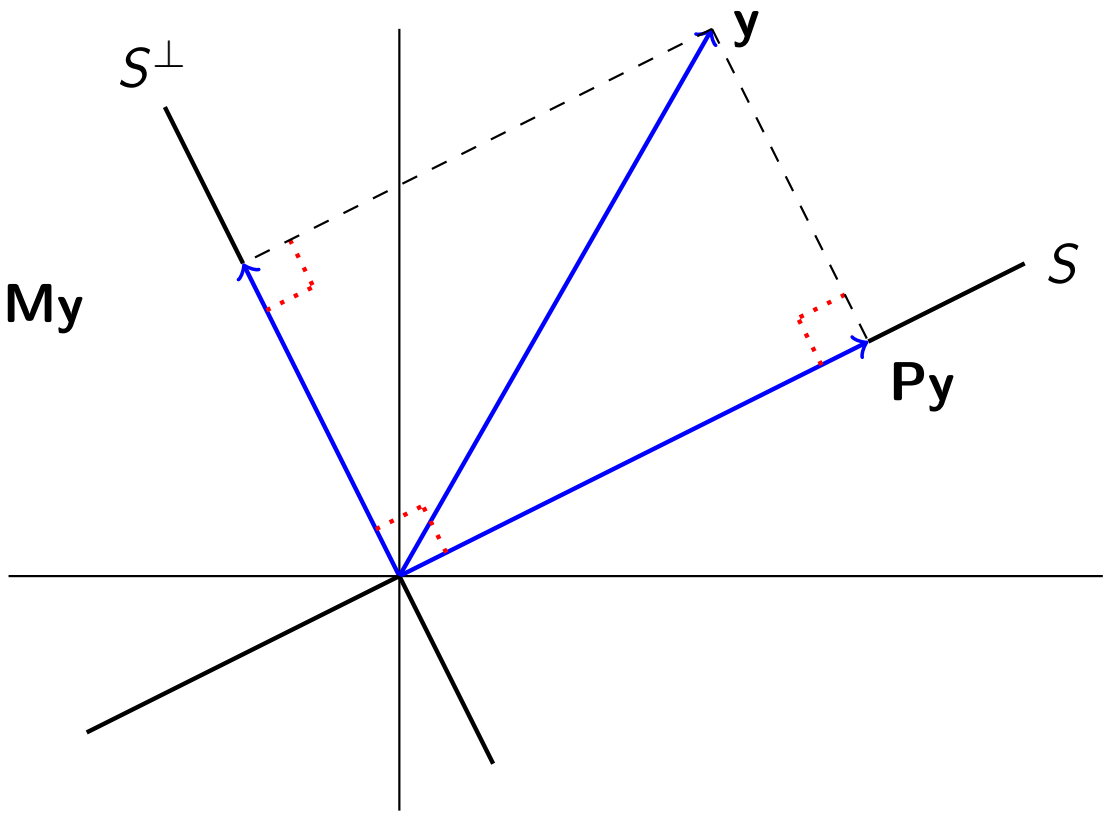


Figure: The residual projection

Multiple Regression

- After having covered basic geometric concepts, let us come back to algebraic interpretation of the regression (some concepts from geometry will soon prove very useful to talk about partitioning of the regression).
- Start with the **three-variable classical linear regression model**, which is a straightforward extension of the two-variable model.
- Suppose we have a data set $(x_{11}, x_{21}, y_1), \ldots, (x_{1n}, x_{2n}, y_n)$, which is a realisation of a sample $(X_{11}, X_{21}, Y_1), \ldots, (X_{1n}, X_{2n}, Y_n)$ from some population distribution $F_{X_1, X_2, Y}$.

Multiple Regression

- This leads to a simple extension of a bivariate regression to a multiple regression:

Definition

The population regression of Y on X_1 and X_2 is given by $\mathbb{E}[Y \mid X_1, X_2]$. The disturbance or error term of the regression is defined by $\varepsilon = Y - \mathbb{E}[Y \mid X_1, X_2]$.

- The model can be rewritten as $Y = \mathbb{E}[Y \mid X_1, X_2] + \varepsilon$ in which $\mathbb{E}[Y \mid X_1, X_2]$ can be interpreted as the best prediction of Y given (X_1, X_2), and ε as the corresponding prediction error.

Multiple Regression With Matrices

- Consider the OLS model for N individuals and $k < N$ regressors. We have:

$$\begin{aligned} y_1 &= \beta_1 + \beta_2 x_{12} + \beta_3 x_{13} + \ldots + \beta_k x_{1k} + \varepsilon_1 \\ y_2 &= \beta_1 + \beta_2 x_{22} + \beta_3 x_{23} + \ldots + \beta_k x_{2k} + \varepsilon_1 \\ \vdots &= \vdots \\ y_N &= \beta_1 + \beta_2 x_{1N} + \beta_3 x_{1N} + \ldots + \beta_k x_{1N} + \varepsilon_N \end{aligned}$$

- or, in matrix notation:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} = \begin{pmatrix} 1 & x_{12} & x_{13} & \cdots & x_{1k} \\ 1 & x_{22} & x_{23} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N2} & x_{N3} & \cdots & x_{Nk} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{pmatrix},$$

$$\underset{N\times 1}{\mathbf{y}} = \underset{N\times k}{\mathbf{X}} \underset{k\times 1}{\beta} + \underset{N\times 1}{\varepsilon}.$$

Multiple Regression With Matrices

- The least squares problem is:

$$\begin{aligned}\min_{\beta} f(\beta) &= \sum_{i=1}^{N} \varepsilon_i^2 = \sum_{i=1}^{N} (y_i - \mathbf{x}_i'\beta)^2 \\ &= \hat{\varepsilon}'\hat{\varepsilon} = (\mathbf{y}-\mathbf{X}\beta)'(\mathbf{y}-\mathbf{X}\beta) \\ &= \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}\beta - \beta'\mathbf{X}'\mathbf{y} + \beta'\mathbf{X}'\mathbf{X}\beta.\end{aligned}$$

- Differentiate wrt β to find the optimum $\hat{\beta}$: $\partial f\left(\hat{\beta}\right)/\partial\beta = 0$:

$$\begin{aligned}0 &= -\frac{\partial \mathbf{y}'\mathbf{X}\hat{\beta}}{\partial\beta} - \frac{\partial \hat{\beta}'\mathbf{X}'\mathbf{y}}{\partial\beta} + \frac{\partial \hat{\beta}'\mathbf{X}'\mathbf{X}\hat{\beta}}{\partial\beta} \\ &= -\mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\hat{\beta} \\ &= -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\hat{\beta}.\end{aligned}$$

OLS With Matrices

- Need to solve
$$\mathbf{X}'\mathbf{X}\hat{\beta} = \mathbf{X}'\mathbf{y}.$$
- If $(\mathbf{X}'\mathbf{X})$ is invertible then we can pre-multiply both sides by $(\mathbf{X}'\mathbf{X})^{-1}$ to get
$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$
- When is $(\mathbf{X}'\mathbf{X})$ invertible?
  - Need $(\mathbf{X}'\mathbf{X})$ to have full rank:
$$rank\left(\mathbf{X}'\mathbf{X}\right) = rank\left(\mathbf{X}\right) = k.$$
  - This is also known as **no (perfect) multicollinearity**.

Anatomy of Regression

- There is a useful link between simple and multivariate regressions.
- In a bivariate regression, the slope coefficient is $\beta = \frac{\mathrm{Cov}(Y_i, X_i)}{\mathrm{Var}(X_i)}$ and the intercept is $\alpha = \mathbb{E}(Y_i) - \beta\mathbb{E}(X_i)$.
- In a multiple regression, by analogy, the kth slope coefficient is

$$\beta_k = \frac{\mathrm{Cov}\left(Y_i, \tilde{X}_{ki}\right)}{\mathrm{Var}\left(\tilde{X}_{ki}\right)}, \tag{3}$$

  where $\tilde{X}_{ki}$ is the **residual from a regression of X_{ki} on all other covariates**.
- This is the **Frisch–Waugh–Lovell** type of argument: each coefficient in a multiple regression is the bivariate slope coefficient for the corresponding regressor, after **partialling out** other variables in the model. Refer to refer to this Slide

Anatomy of Regression

- One of the most prevalent challenge in applied research, the **omitted variable bias**, then follows immediately; suppose we omitted X_{2i} from the above regression.
- The result is

$$\frac{\mathrm{Cov}\left(Y_i,\, X_{1i}\right)}{\mathrm{Var}\left(X_{1i}\right)} = \beta_1 + \beta_2 \frac{\mathrm{Cov}\left(X_{1i},\, X_{2i}\right)}{\mathrm{Var}\left(X_{1i}\right)} \overset{\text{generally}}{\neq} \beta_1.$$

- Another important aspect — a variance in the context of the multiple regression. Using a variance for the parameter in a multiple regression,

$$\begin{aligned} \mathrm{Var}\left(\beta_k|X\right) &= \mathrm{Var}\left(\frac{\mathrm{Cov}\left(Y_i, \tilde{X}_{ki}\right)}{\mathrm{Var}\left(\tilde{X}_{ki}\right)}|X\right) \\ &= \frac{\mathrm{Var}\left(\mathrm{Cov}\left(Y_i, \tilde{X}_{ki}\right)|X\right)}{\left[\mathrm{Var}\left(\tilde{X}_{ki}\right)\right]^2} \overset{IID}{=} \frac{\sigma^2}{TSS_k(X)\left(1-R_k^2(X)\right)}, \end{aligned}$$

where $\mathrm{Var}\left(Y_i|X\right) = \sigma^2$ for all i, $TSS_k\left(X\right)$ is the total sum of squares of X_k, $\sum_i \left(X_{ki} - \bar{X}_k\right)^2$, and $R_k^2\left(X\right)$ comes from regressing X_k on all other regressors.

The Frisch-Waugh-Lovell Theorem in Matrix Notation

- Let us come back to the **Frisch-Waugh-Lovell** argument, and employ linear algebra to enhance rigour.
- Imagine a regression involving two sets of variables $\mathbf{X}_1$ and $\mathbf{X}_2$

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon = \begin{pmatrix} \mathbf{X}_1 & \mathbf{X}_2 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \varepsilon = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \varepsilon.$$

- Say we are only interested in β_2. Do we have to solve for the whole $\beta = \begin{pmatrix} \beta_1' & \beta_2' \end{pmatrix}'$ vector?
- OLS solves the **normal equations**:

$$\begin{aligned} (\mathbf{X}'\mathbf{X})\,\hat{\beta} &= \mathbf{X}'\mathbf{y}, \\ \left(\begin{pmatrix} \mathbf{X}_1' \\ \mathbf{X}_2' \end{pmatrix} \begin{pmatrix} \mathbf{X}_1 & \mathbf{X}_2 \end{pmatrix} \right) \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} &= \begin{pmatrix} \mathbf{X}_1' \\ \mathbf{X}_2' \end{pmatrix} \mathbf{y}, \\ \begin{pmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_2 \\ \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2 \end{pmatrix} \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} &= \begin{pmatrix} \mathbf{X}_1'\mathbf{y} \\ \mathbf{X}_2'\mathbf{y} \end{pmatrix}. \end{aligned}$$

The Frisch-Waugh-Lovell Theorem

- Start with the solution of the first set of normal equations for $\hat{\beta}_1$:

$$\begin{aligned}
\mathbf{X}_1'\mathbf{X}_1\hat{\beta}_1 + \mathbf{X}_1'\mathbf{X}_2\hat{\beta}_2 &= \mathbf{X}_1'\mathbf{y}, \\
\mathbf{X}_1'\mathbf{X}_1\hat{\beta}_1 &= \mathbf{X}_1'\mathbf{y} - \mathbf{X}_1'\mathbf{X}_2\hat{\beta}_2, \\
\hat{\beta}_1 &= \left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\mathbf{y} - \left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\mathbf{X}_2\hat{\beta}_2 \\
&= \left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\left(\mathbf{y} - \mathbf{X}_2\hat{\beta}_2\right).
\end{aligned}$$

Theorem (Orthogonal Partitioned Regression)

If $\mathbf{X}_1$ and $\mathbf{X}_2$ are orthogonal, then the coefficients can be obtained by separately regressing $\mathbf{y}$ on $\mathbf{X}_1$ and regressing $\mathbf{y}$ on $\mathbf{X}_2$

Proof.

If $\mathbf{X}_1$ and $\mathbf{X}_2$ are orthogonal, then $\mathbf{X}_1'\mathbf{X}_2 = 0$ (by definition) and so the solution to the normal equations above is simply $\hat{\beta}_1 = \left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\mathbf{y}$, the OLS coefficients from a regression of $\mathbf{y}$ on $\mathbf{X}_1$. ($\hat{\beta}_2$ follows by analogy). □

The Frisch-Waugh-Lovell Theorem

- If $\mathbf{X}_1$ and $\mathbf{X}_2$ aren't orthogonal, we can still obtain $\hat{\beta}_1$ and $\hat{\beta}_2$ separately:

Theorem (Frisch-Waugh-Lovell Theorem)

In the regression of $\mathbf{y}$ *on 2 sets of variables* $\mathbf{X}_1$ *and* $\mathbf{X}_2$*, the coefficients* $\hat{\beta}_2$ *can be obtained by regressing the* ***residuals*** *from a regression of* $\mathbf{y}$ *on* $\mathbf{X}_1$ *on the* ***residuals*** *from regressions of each column of* $\mathbf{X}_2$ *on* $\mathbf{X}_1$*.*

Proof.

The second set of normal equations is

$$\mathbf{X}_2'\mathbf{X}_1\hat{\beta}_1 + \mathbf{X}_2'\mathbf{X}_2\hat{\beta}_2 = \mathbf{X}_2'\mathbf{y}.$$

Inserting the solution for $\hat{\beta}_1$

$$\mathbf{X}_2'\mathbf{X}_1\left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\mathbf{y} - \mathbf{X}_2'\mathbf{X}_1\left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\mathbf{X}_2\hat{\beta}_2 + \mathbf{X}_2'\mathbf{X}_2\hat{\beta}_2 = \mathbf{X}_2'\mathbf{y}.$$

The Frisch-Waugh-Lovell Theorem

Proof.

Rearranging

$$\mathbf{X}_2' \left[\mathbf{I} - \mathbf{X}_1 \left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1} \mathbf{X}_1'\right] \mathbf{X}_2 \hat{\beta}_2 = \mathbf{X}_2' \underbrace{\left[\mathbf{I} - \mathbf{X}_1 \left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1} \mathbf{X}_1'\right] \mathbf{y}}_{\text{residuals in regression of } \mathbf{y} \text{ on } \mathbf{X}_1 \text{ only}},$$

$$\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2\hat{\beta}_2 = \mathbf{X}_2'\mathbf{M}_1\mathbf{y},$$

$$\hat{\beta}_2 = \left(\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2\right)^{-1} \mathbf{X}_2'\mathbf{M}_1\mathbf{y}.$$

□

- To sum it up, we can find $\hat{\beta}_2$ through a series of regressions
  - regress $\mathbf{y}$ on $\mathbf{X}_1$ and store the residuals $\mathbf{M}_1\mathbf{y}$,
  - regress each column of $\mathbf{X}_2$ on $\mathbf{X}_1$ and store all the residuals in $\mathbf{M}_1\mathbf{X}_2$,
  - regress $\mathbf{M}_1\mathbf{y}$ on $\mathbf{M}_1\mathbf{X}_2$.

Multiple Regression: Algebra and Geometry

- We shall repeat some arguments and link **algebraic** and **geometric** concepts together.
- Consider a multiple regression (written in a scalar format):

$$y = x_1\hat{\beta}_1 + x_2\hat{\beta}_2 + \ldots + x_k\hat{\beta}_k,$$

  where $\boldsymbol{x}$ and $\hat{\boldsymbol{\beta}}$ are k-dimensional vectors.
- Define the product of two vectors to be $\boldsymbol{x}\cdot\hat{\boldsymbol{\beta}} = \sum x_i\hat{\beta}_i$ or a dot product (inner product).
- Recall that two vectors are **orthogonal** if their dot product equals zero, meaning graphically that the vectors are perpendicular.

Multiple Regression: Algebra and Geometry

- Going to more statistical interpretation, let $\boldsymbol{x}$ and $\boldsymbol{y}$ be two random variables, each with mean zero, and N elements.
  - We can then construct a vector $\boldsymbol{x} = (x_1, x_2, \ldots, x_N)$ and a similar vector $\boldsymbol{y} = (y_1, y_2, \ldots, y_N)$.
  - When we take their **dot product**, we calculate: $\boldsymbol{x} \cdot \boldsymbol{y} = \sum x_i y_i = (N-1)\,\hat{Cov}\,(x, y)$ and $\boldsymbol{x} \cdot \boldsymbol{x} = \sum x_i x_i = (N-1)\,\hat{Var}\,(x)$.
- Geometrically, two vectors in N-dimensional space, the angle θ between them must always satisfy: $\cos\theta = \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{\sqrt{\boldsymbol{x} \cdot \boldsymbol{x}}\sqrt{\boldsymbol{y} \cdot \boldsymbol{y}}} = \hat{Corr}\,(x, y)$.
  - The cosine of two rays is one if they point in exactly the same direction, zero if they are perpendicular, negative one if they point in exactly opposite directions – **exactly the same as with correlations**.

Multiple Regression: Algebra and Geometry

- Let's construct a system of equations with a vector $\boldsymbol{y} = (y_1, y_2, \ldots, y_N)'_{N\times 1}$ containing all of the y values, also parameter vector $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_K)'_{K\times 1}$ and matrix

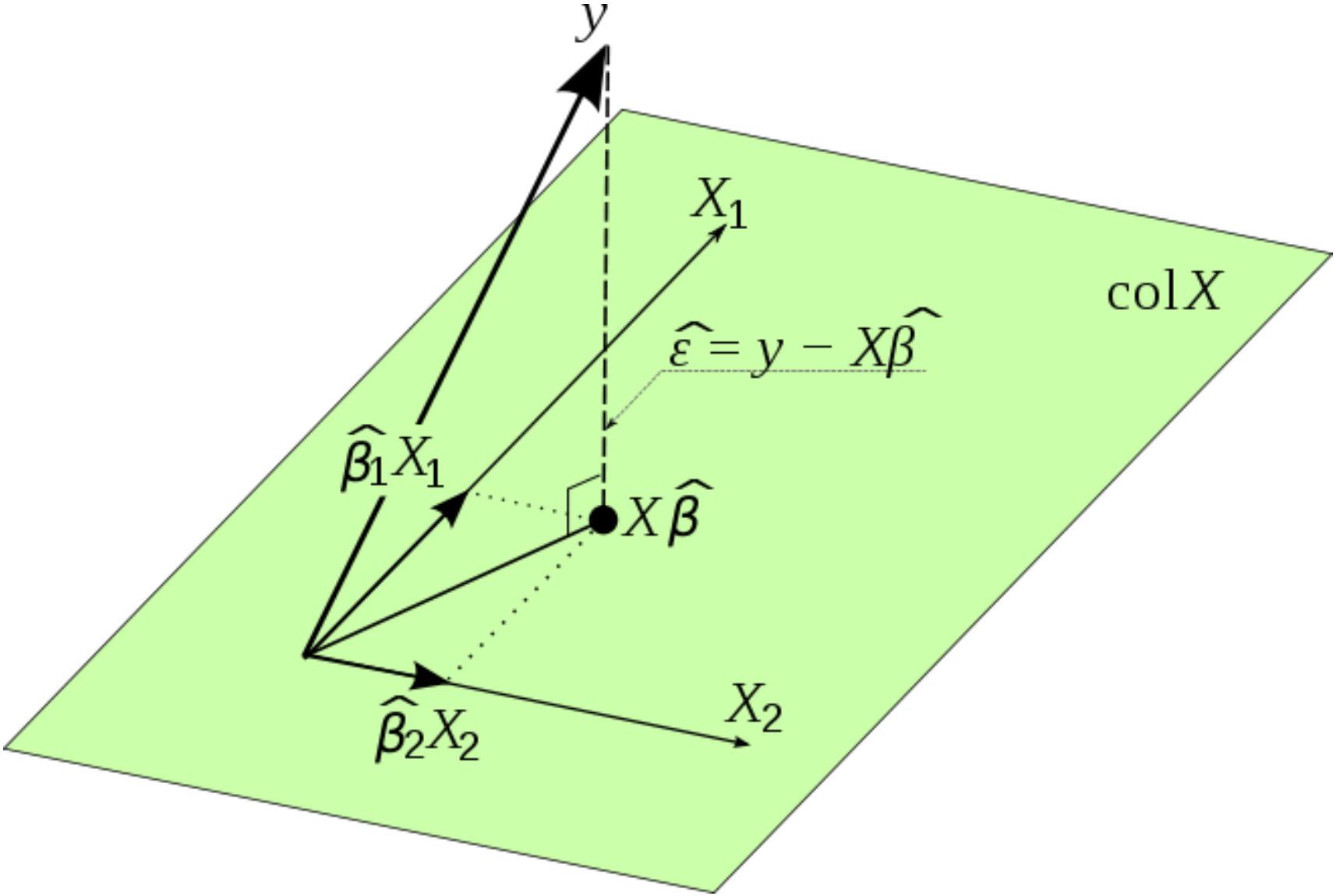
$$X = \begin{bmatrix} 1 & x_{12} & \cdots & x_{1K} \\ 1 & x_{22} & \cdots & x_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N2} & \cdots & x_{NK} \end{bmatrix}_{N\times K},$$

  and a vector $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N)'_{N\times 1}$ containing all of the "unobservable" determinants of the outcome $\boldsymbol{y}$.
- The system of equations can be represented as: $\boldsymbol{y}_{N\times 1} = \boldsymbol{X}_{N\times K}\boldsymbol{\beta}_{K\times 1} + \boldsymbol{\varepsilon}_{N\times 1}$.
- Its econometric model is $\hat{\boldsymbol{\varepsilon}}_{N\times 1} = \boldsymbol{y}_{N\times 1} - \boldsymbol{X}_{N\times K}\hat{\boldsymbol{\beta}}_{K\times 1}$, where residuals $\hat{\boldsymbol{\varepsilon}}_{N\times 1}$ are obtained once $\hat{\boldsymbol{\beta}}_{K\times 1}$ is estimated.
- We want residuals to be such that their size (or **norm**) of the vector $\hat{\boldsymbol{\varepsilon}}$ is minimised, i.e., $\min \|\hat{\boldsymbol{\varepsilon}}\| = \min \sqrt{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}$ or, more familiarly, $\min \|\hat{\boldsymbol{\varepsilon}}\|^2 = \min \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}$.

Multiple Regression: Algebra and Geometry

- We want to minimise, as covered before, the following expression: $\min_{\hat{\beta}} \hat{\varepsilon}'\hat{\varepsilon} = \left(\boldsymbol{y} - \boldsymbol{X}\hat{\beta}\right)' \left(\boldsymbol{y} - \boldsymbol{X}\hat{\beta}\right) = \boldsymbol{y}'\boldsymbol{y} - \boldsymbol{y}'\boldsymbol{X}\hat{\beta} - \hat{\beta}'\boldsymbol{X}'\boldsymbol{y} + \hat{\beta}'\boldsymbol{X}'\boldsymbol{X}\hat{\beta}$.
- The optimal $\hat{\beta}^{OLS}$ yields $\hat{\beta}^{OLS} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y}$.
- Notice that $\hat{\varepsilon} = \boldsymbol{y} - \boldsymbol{X}\hat{\beta} = \boldsymbol{y} - \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y} = \left(\boldsymbol{I} - \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\right)\boldsymbol{y}$.
- We can decompose $\boldsymbol{y}$ into two components:
  - the **orthogonal projection** onto the K dimensional space spanned by $\boldsymbol{X}$, $\boldsymbol{X}\hat{\beta}$, and
  - the component that is the orthogonal projection onto the $n - K$ subspace that is orthogonal to the **span** of $\boldsymbol{X}$, $\hat{\varepsilon}$.
- Since $\hat{\beta}$ is chosen to make $\hat{\varepsilon}$ as short as possible, $\hat{\varepsilon}$ will be orthogonal to the space spanned by $\boldsymbol{X}$ as in this space, $\boldsymbol{X}'\hat{\varepsilon} = 0$. **The FOCs that define the least squares estimator imply that this is so**.

Geometry of OLS



Projection Matrices

- We have that $\boldsymbol{X}\hat{\beta}$ is the projection of $\boldsymbol{y}$ on the span of $\boldsymbol{X}$ or, $\boldsymbol{X}\hat{\beta} = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{y}$.
- Then, $\hat{\varepsilon}$ is the projection of $\boldsymbol{y}$ off the space spanned by $\boldsymbol{X}$, in other words, onto the space that is orthogonal to the span of $\boldsymbol{X}$: $\hat{\varepsilon} = \left(\boldsymbol{I} - \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\right)\boldsymbol{y} = \left(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}}\right)\boldsymbol{y} = \boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{y}$.
- Therefore, $\boldsymbol{y} = \boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{y} + \boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{y} = \left(\boldsymbol{P}_{\boldsymbol{X}} + \boldsymbol{M}_{\boldsymbol{X}}\right)\boldsymbol{y}$.
- Note that both $\boldsymbol{P}_{\boldsymbol{X}}$ and $\boldsymbol{M}_{\boldsymbol{X}}$ are **symmetric** and **idempotent** ($\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{P}_{\boldsymbol{X}} = \boldsymbol{P}_{\boldsymbol{X}}$ and $\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{M}_{\boldsymbol{X}} = \boldsymbol{M}_{\boldsymbol{X}}$).

Projection Matrices and Coefficient of Determination, R^2

- To determine the **goodness-of-fit,** also use the expression $\boldsymbol{y}'\boldsymbol{y} = \left(\hat{\beta}'\boldsymbol{X}' + \hat{\varepsilon}'\right)\left(\boldsymbol{X}\hat{\beta} + \hat{\varepsilon}\right) = \hat{\beta}'\boldsymbol{X}'\boldsymbol{X}\hat{\beta} + \hat{\beta}'\boldsymbol{X}'\hat{\varepsilon} + \hat{\varepsilon}'\boldsymbol{X}\hat{\beta} + \hat{\varepsilon}'\hat{\varepsilon} = \hat{\beta}'\boldsymbol{X}'\boldsymbol{X}\hat{\beta} + \hat{\varepsilon}'\hat{\varepsilon}$ since $\boldsymbol{X}'\hat{\varepsilon} = \boldsymbol{0}$.
- Then uncentred R_U^2 is defined as $R_U^2 = 1 - \hat{\varepsilon}'\hat{\varepsilon}/\boldsymbol{y}'\boldsymbol{y} = \hat{\beta}'\boldsymbol{X}'\boldsymbol{X}\hat{\beta}/\boldsymbol{y}'\boldsymbol{y} = \|\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{y}\|^2 / \|\boldsymbol{y}\|^2 = \cos^2\theta$, where θ is the angle between $\boldsymbol{y}$ and the span of $\boldsymbol{X}$.
- For a more usual **centred coefficient of determination**, introduce the n-vector $\boldsymbol{i} = (1, 1, \ldots, 1)'$ which we can use in forming $\boldsymbol{M}_{\boldsymbol{i}} = I_n - \boldsymbol{i}\left(\boldsymbol{i}'\boldsymbol{i}\right)^{-1}\boldsymbol{i}' = I_n - \boldsymbol{i}\boldsymbol{i}'/n$.
- $\boldsymbol{M}_{\boldsymbol{i}}\boldsymbol{y}$ gives the vector of deviations from the mean: thus, $R_C^2 = 1 - \hat{\varepsilon}'\hat{\varepsilon}/\boldsymbol{y}'\boldsymbol{M}_{\boldsymbol{i}}\boldsymbol{y} = 1 - ESS/TSS$.
  - Recalling that we construct residuals to average to zero (when a constant is included), $\boldsymbol{M}_{\boldsymbol{i}}\hat{\varepsilon} = \hat{\varepsilon}$.

Outline

Gauss-Markov Theorem

- Turning to the justification of OLS, we shall invoke the G-M theorem; it can be proved under **three main assumptions**:

1. **Orthogonality**: the errors and the regressors are uncorrelated, $\mathbb{E}(\varepsilon_i|X) = 0$.
2. **Homoskedasticity**: the errors ε_i have a constant conditional variance, $\mathbb{E}\left(\varepsilon_i^2|X\right) = \sigma^2$ for all i.
3. **Not** (**serially** or **cross-sectionally**) **correlated errors**, $\mathbb{E}(\varepsilon_i\varepsilon_j|X) = 0$ for $i \neq j$.

- Note that Assumption (1) implies that $\mathbb{E}(\varepsilon_i) = 0$, which is a result of the **Law of Iterated Expectations**, $\mathbb{E}_X\mathbb{E}(\varepsilon_i|X) = \mathbb{E}(\varepsilon_i) = 0$.
- **Correct functional form, linearity of an estimator, and sufficient variation in X (full column rank) are also implicit to the theorem**.

Gauss-Markov Theorem

Theorem

Under assumptions 1-3, the OLS estimator has the least variance in the class of linear unbiased estimators, namely, it is the ***best linear unbiased estimator*** *(BLUE).*

Proof.

By linearity we mean that an estimator of β_j is a linear combination $\widehat{\beta}_j = w_{1j}y_1 + \cdots + w_{nj}y_n = \sum_i w_{ij}y_i$, in which the weights w_{ij} are not allowed to depend on the underlying coefficients β_j, since those are not observable, but are allowed to depend on the values X_{ij}, since these data are observable. Denote by $\widehat{\beta}_j$ an OLS estimator. Continued...

□

Proof (cont.)

Proof.

Suppose there exists another set of weights, call them $\{\omega_i\}_{i=1}^n$, not corresponding to the OLS, which makes the linear unbiased estimator's variance smaller. Then, under a simple population model $y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, we know that any linear estimator can be described as
$\tilde{\beta}_1 = \sum_i \omega_i y_i = \beta_0 \sum_i \omega_i + \beta_1 \sum_i \omega_i X_i + \sum_i \omega_i \varepsilon_i$.
Unbiasedness requires $\mathbb{E}\left(\tilde{\beta}_1 \mid X\right) = \beta_1$ or $\sum_i \omega_i = 0$ and $\sum_i \omega_i X_i = 1$ (since, after conditioning on X, $\mathbb{E}\left(\sum_i \omega_i \varepsilon_i | X\right) = \sum_i \omega_i \mathbb{E}\left(\varepsilon_i | X\right) = 0$).
We need to find a variance for $\tilde{\beta}_1$, which, under $\sum_i \omega_i = 0$ and $\sum_i \omega_i X_i = 1$, is simply obtainable from $\tilde{\beta}_1 = \beta_1 + \sum_i \omega_i \varepsilon_i$, i.e.,

$$\mathrm{Var}\left(\tilde{\beta}_1 | X\right) = \mathbb{E}\left(\left(\tilde{\beta}_1 - \beta_1\right)^2 \mid X\right) = \mathbb{E}\left(\left(\sum_i \omega_i \varepsilon_i\right)^2 \mid X\right) = \sigma^2 \left(\sum_i \omega_i^2\right),$$

where we used the uncorrelatedness assumption, $\mathbb{E}\left(\varepsilon_i \varepsilon_j | X\right) = 0$ for $i \neq j$, and the homoskedasticity assumption, $\mathbb{E}\left(\varepsilon_i^2 | X\right) = \sigma^2$ for all i (we have also treated the weights as given, after conditioning on X).

□

Proof (cont.)

Proof.

The last "trick" is to subtract and add a term $\frac{x_i-\bar{x}}{\sum_i (x_i-\bar{x})^2}$, i.e.,

$$\begin{aligned}
\sum\left(\left(\tilde{\beta}_1-\beta_1\right)^2 \mid X\right) &= \sigma^2\left(\sum_i\left(\omega_i-\frac{x_i-\bar{x}}{\sum_i (x_i-\bar{x})^2}+\frac{x_i-\bar{x}}{\sum_i (x_i-\bar{x})^2}\right)^2\right) \\
&= \sigma^2\sum_i\left(\omega_i-\frac{x_i-\bar{x}}{\sum_i (x_i-\bar{x})^2}\right)^2+\sigma^2\sum_i\left(\frac{x_i-\bar{x}}{\sum_i (x_i-\bar{x})^2}\right)^2 \\
&\quad +2\sigma^2\sum_i\left(\omega_i-\frac{x_i-\bar{x}}{\sum_i (x_i-\bar{x})^2}\right)\frac{x_i-\bar{x}}{\sum_i (x_i-\bar{x})^2} \\
&= \sigma^2\sum_i\left(\omega_i-\frac{x_i-\bar{x}}{\sum_i (x_i-\bar{x})^2}\right)^2+\sigma^2\sum_i\left(\frac{x_i-\bar{x}}{\sum_i (x_i-\bar{x})^2}\right)^2 \\
&= \sigma^2\sum_i\left(\omega_i-\frac{x_i-\bar{x}}{\sum_i (x_i-\bar{x})^2}\right)^2+\operatorname{Var}\left(\widehat{\beta}_1|X\right)\geq \operatorname{Var}\left(\widehat{\beta}_1|X\right).
\end{aligned}$$

Only for the case $\omega_i=\frac{x_i-\bar{x}}{\sum_i (x_i-\bar{x})^2}$, $\operatorname{Var}\left(\tilde{\beta}_1|X\right)=\operatorname{Var}\left(\widehat{\beta}_1|X\right)$, but then we are back to the OLS weights, i.e., $\omega_i=w_i$.

□

Use of Matrix Algebra

- The same result can be invoked in a somewhat more elegant (but conceptually identical) way, employing linear algebra.
- By linearity we mean that the estimator $\hat{\beta} = \mathbf{Ay}$ is a linear combination of $\mathbf{y}$. Suppose $\hat{\beta}$ is **any** linear conditionally unbiased estimator of β. Then

$$Var\left(\hat{\beta}|\mathbf{X}\right) = \mathbf{A}\,Var\left(\mathbf{y}|\mathbf{X}\right)\mathbf{A}' = \sigma^2\mathbf{AA}'.$$

- To prove BLUEness of the estimator, we will focus on **efficiency** (linearity is $\hat{\beta} = \mathbf{Ay}$ and unbiasedness holds by assumption $\mathbb{E}\left(\mathbf{Ay}|\mathbf{X}\right) = \beta$).
- Note that, since $\mathbf{y} = \mathbf{X}\beta + \varepsilon$, we have $\mathbb{E}\left(\mathbf{Ay}|\mathbf{X}\right) = \mathbb{E}\left(\mathbf{AX}\beta + \mathbf{A}\varepsilon|\mathbf{X}\right) = \beta$. As for the error term, $\mathbb{E}\left(\mathbf{A}\varepsilon|\mathbf{X}\right) = 0$, so it must be true that $\mathbf{AX} = \mathbf{I}$.
- Using *add and subtract* trick, it follows that $\mathbf{AA}' = \left(\mathbf{A} - \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}' + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right)\left(\mathbf{A}' - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1} + \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\right)$ (recall rules of matrix transposition).

Use of Matrix Algebra

- As in previous proof, introduce weights, equal to matrices $\mathbf{W} \equiv \mathbf{A} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. This helps to re-express $\mathbf{AA}'$ as

$$\begin{aligned}\mathbf{AA}' &= \left(\mathbf{W} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)\left(\mathbf{W}' + \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\right) \\ &= \mathbf{WW}' + \mathbf{WX}(\mathbf{X}'\mathbf{X})^{-1} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}' + (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1} \\ &= \mathbf{WW}' + (\mathbf{X}'\mathbf{X})^{-1}.\end{aligned}$$

- This is because

$$\begin{aligned}\mathbf{WX}(\mathbf{X}'\mathbf{X})^{-1} &= A\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1} - (\mathbf{X}'\mathbf{X})^{-1} = \mathbf{0},\end{aligned}$$

  using the fact $\mathbf{AX} = \mathbf{I}$. Similarly, $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}' = \mathbf{0}$.

- It follows that $\mathbf{AA}' - (\mathbf{X}'\mathbf{X})^{-1} = \mathbf{WW}'$ is a **positive semi-definite matrix**. OLS is BLUE.

OLS Asymptotics: Consistency

- We need to modify the assumptions we made when studying OLS in finite samples slightly. We assume:
  1. $(\mathbf{x}_i, \varepsilon_i)$ $i = 1, \ldots, n$ is a sequence of *independent* observations,
  2. plim $\frac{\mathbf{X}'\mathbf{X}}{n} = \mathbf{Q}$, a non-singular matrix.
- Rewrite the OLS estimate $\hat{\beta}$ as

$$\hat{\beta} = \beta + \left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1} \left(\frac{\mathbf{X}'\varepsilon}{n}\right).$$

- Asymptotically:

$$\text{plim } \hat{\beta} = \beta + \mathbf{Q}^{-1}\text{plim } \left(\frac{\mathbf{X}'\varepsilon}{n}\right).$$

OLS Asymptotics: Consistency

- So we need to find plim $\left(\frac{\mathbf{X}'\boldsymbol{\varepsilon}}{n}\right)$. Notice we can write this as

$$\frac{1}{n}\mathbf{X}'\boldsymbol{\varepsilon} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\varepsilon_i = \frac{1}{n}\sum_{i=1}^{n}\mathbf{w}_i = \bar{\mathbf{w}}.$$

- So that plim $\hat{\beta}=\beta + \mathbf{Q}^{-1}\operatorname{plim}\bar{\mathbf{w}}$.
- To find plim $\bar{\mathbf{w}}$ we need to see what happens to $\mathbb{E}[\bar{\mathbf{w}}]$ and $\operatorname{Var}[\bar{\mathbf{w}}]$ as $n \to \infty$,

$$\mathbb{E}[\mathbf{w}_i] = \mathbb{E}_x[\mathbb{E}[\mathbf{w}_i|\mathbf{x}_i]] = \mathbb{E}_x\left[\mathbf{x}_i \underbrace{\mathbb{E}[\varepsilon_i|\mathbf{x}_i]}_{=0 \text{ (exogeneity assumption)}}\right] = 0,$$

hence,

$$\mathbb{E}[\bar{\mathbf{w}}] = 0 \quad (<\infty).$$

OLS Asymptotics: Consistency

- Turning to $\text{Var}[\bar{\mathbf{w}}]$:

$$\text{Var}[\bar{\mathbf{w}}] = \mathbb{E}[\text{Var}[\bar{\mathbf{w}}|\mathbf{X}]] + \underbrace{\text{Var}[\mathbb{E}[\bar{\mathbf{w}}|\mathbf{X}]]}_{=0\ (\mathbb{E}[\varepsilon_i|\mathbf{X}]=0)}$$

$$\text{Var}[\bar{\mathbf{w}}|\mathbf{X}] = \mathbb{E}\left[\bar{\mathbf{w}}\bar{\mathbf{w}}'|\mathbf{X}\right] = \frac{1}{n}\mathbf{X}'\mathbb{E}\left[\varepsilon\varepsilon'|\mathbf{X}\right]\mathbf{X}\frac{1}{n}$$

$$= \frac{1}{n}\mathbf{X}'\sigma^2\mathbf{I}\mathbf{X}\frac{1}{n} = \left(\frac{\sigma^2}{n}\right)\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right),$$

$$\text{Var}[\bar{\mathbf{w}}] = \left(\frac{\sigma^2}{n}\right)\mathbb{E}\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right),$$

$$\lim_{n\to\infty} \text{Var}[\bar{\mathbf{w}}] = 0 \times \mathbf{Q} = \mathbf{0}.$$

- Therefore, $\lim_{n\to\infty}\mathbb{E}[\bar{\mathbf{w}}] = 0 < \infty$ and $\lim_{n\to\infty}\text{Var}[\bar{\mathbf{w}}] = 0$ so $\bar{\mathbf{w}}$ converges in mean square to 0 and

$$\text{plim}\ \bar{\mathbf{w}} = 0.$$

- **Putting this all together**:

$$\text{plim}\ \hat{\beta} = \beta + \mathbf{Q}^{-1}\cdot\mathbf{0} = \beta.$$

OLS Asymptotic Distribution

- We want to **relax the normality assumption**, but we still want to know the distribution of $\hat{\beta}$ (at least in large enough samples).
- To do this, we will use the **central limit theorem** (**CLT**), which requires us to assume that the observations are *independent.*
- We will focus on

$$\sqrt{n}\left(\hat{\beta}-\beta\right)=\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1}\left(\frac{1}{\sqrt{n}}\right)\mathbf{X}'\varepsilon.$$

- Using the rules of convergence in distribution, if this has a **limiting distribution**, it's the same as that of

$$\left[\text{plim}\ \left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1}\right]\left(\frac{1}{\sqrt{n}}\right)\mathbf{X}'\varepsilon=\mathbf{Q}^{-1}\left(\frac{1}{\sqrt{n}}\right)\mathbf{X}'\varepsilon.$$

OLS Asymptotic Distribution

- Let's find the **limiting distribution** of

$$\left(\frac{1}{\sqrt{n}}\right)\mathbf{X}'\boldsymbol{\varepsilon} = \sqrt{n}\bar{\mathbf{w}}.$$

- $\mathbb{E}[\mathbf{w}_i] = 0$, what about $\text{Var}[\mathbf{w}_i]$?
- $\mathbf{w}_i = \mathbf{x}_i\varepsilon_i$ so

$$\text{Var}[\mathbf{x}_i\varepsilon_i] = \mathbb{E}\left[\mathbf{x}_i\varepsilon_i^2\mathbf{x}_i'\right] = \sigma^2\mathbb{E}[\mathbf{x}_i\mathbf{x}_i'] = \sigma^2\mathbf{Q}.$$

- Applying the **central limit theorem**:

$$\left(\frac{1}{\sqrt{n}}\right)\mathbf{X}'\boldsymbol{\varepsilon} \xrightarrow{d} N\left[0, \sigma^2\mathbf{Q}\right].$$

OLS Asymptotic Distribution

- Putting pieces together:

$$\mathbf{Q}^{-1}\left(\frac{1}{\sqrt{n}}\right)\mathbf{X}'\varepsilon \xrightarrow{d} N\left[\mathbf{Q}^{-1}\mathbf{0}, \mathbf{Q}^{-1}\left(\sigma^2\mathbf{Q}\right)\mathbf{Q}^{-1}\right],$$
$$\sqrt{n}\left(\hat{\beta}-\beta\right) \xrightarrow{d} N\left[0, \sigma^2\mathbf{Q}^{-1}\right].$$

Theorem

If ε_i are i.i.d. with mean 0 and variance σ^2, then

$$\hat{\beta} \overset{a}{\sim} N\left[\beta, \frac{\sigma^2}{n}\mathbf{Q}^{-1}\right].$$

- We estimate $(1/n)\,\mathbf{Q}^{-1}$ with $(\mathbf{X}'\mathbf{X})^{-1}$ and σ^2 with $\hat{\varepsilon}'\hat{\varepsilon}/(n-k)$.

OLS Asymptotic Distribution

- So, we have an (asymptotically) normal distribution for $\hat{\beta}$!
  - But it's not because we assumed the ε_i are normally distributed.
  - It is coming from the **central limit theorem**.

What Makes OLS Causal?

- The classical regression model assumes that $\mathbf{X}$ is non-random.
- This comes from agricultural field experiments, where $\mathbf{X}$ is fully controlled by the experimenter. For instance, regressors may correspond to various amounts of fertilisers of different types applied to experimental fields.
- Typically, there would be many experimental fields that are exposed to the same experimental conditions (same regressors).
  - The fields with the same regressors would differ because of uncontrollable factors such as small fluctuations in the quality of soil across experimental fields (captured by ε_i).
- Unfortunately, the same setup cannot be applied to many economics problems. What can be done nevertheless?

Correlation vs Causation

- Conditional on regressors, ε_i is IID.
- From econometrics perspective, ε_i is thought of as the cumulative effect on y of the causal factors **not** included in the regression.
  - Much of the econometric theory assumes **random sampling** (at least at some level): think of CLT and how to prove the properties of estimators.
  - Sometimes, the assumption is questionable.
- Sometimes the data covers literally all cross-sectional units, such as, for example, all the EU members or US states. How can we think about such data as coming from random sampling?
  - At times it might be helpful to think of an imaginary experiment where a supreme being draws at random destinies for different cross-sectional units, so that the available data is a random sample from the pool of possible fates...

Causal Mechanism

- It is clear that two problems – **prediction** and **causal interpretation** – are not necessarily consistent.
  - **Minimising a prediction error**, for instance, using machine learning tools, **places no weight on mechanisms that give rise to the data generating process** (as long as the prediction error can be made lower).
  - E.g. inflation and cumulative rainfall in Scotland (David Hendry). The two have no causal link but one can nevertheless have predictive power for another (the source of which can be linked to the stochastic properties which we ignore for the time being).
  - Many examples from economics: X are usually stochastic and correlated with ε_i. OLS **cannot** describe causal effects **unless** one can establish other channels, unrelated to ε_i but linked to X. If prediction was the only goal, however, this strategy to find indirect channels would be hard to justify as it just wastes efficiency (**classical tradeoff between unbiasedness and efficiency**).
- **Can we formally link regression to causal inference using observational data?**

Outline

Program Evaluation

- Research agenda for applied economics requires to address these major issues:

1. **Causal relationship;**
2. **Ideal experiment;**
3. **Identification strategy;**
4. **Statistical inference.**

- Reality manifests in many ways which complicate empirical enquiry, especially when questions about ***causality*** are raised.
- Drawing from Angrist and Pischke (2008), such questions hinge on the experimentalist paradigm and potential outcomes.
- This new paradigm in econometrics requires some new vocabulary...

Program Evaluation

- Suppose we analyse a variable Y_i, described by

$$\text{Potential outcome} = \begin{cases} Y_{1i} & \text{Outcome if } i \text{ is treated;} \\ Y_{0i} & \text{Outcome if } i \text{ is not treated.} \end{cases}$$

- In effect, Y_{1i} (***potential*** outcome ***under treatment***) and Y_{0i} (***potential*** outcome ***without treatment***) are outcomes in alternative states of the world.
- Nonexistence of parallel worlds leads to inability to measure treatment effects at the individual level.
- However, notice that the ***observed*** outcome Y_i can be expressed in terms of potential ones:

$$Y_i = \begin{cases} Y_{1i}, & \text{if } D_i = 1, \\ Y_{0i}, & \text{if } D_i = 0, \end{cases}$$

or $Y_i = Y_{0i} + (Y_{1i} - Y_{0i})\, D_i$.

Program Evaluation

- One hope is, instead of individual ones, to analyse average treatment effects.

Question

So what about a comparison of the difference in means for treated and untreated?

Program Evaluation

- We obtain

$$E[Y_i \mid D_i = 1] - E[Y_i \mid D_i = 0] = E[Y_{1i} \mid D_i = 1] - E[Y_{0i} \mid D_i = 0],$$

and subtracting and adding $E[Y_{0i} \mid D_i = 1]$ produces

$$\begin{aligned} & E[Y_i \mid D_i = 1] - E[Y_i \mid D_i = 0] \\ = & \underbrace{E[Y_{1i} \mid D_i = 1] - E[Y_{0i} \mid D_i = 1]}_{\text{ATT}} \\ + & \underbrace{E[Y_{0i} \mid D_i = 1] - E[Y_{0i} \mid D_i = 0]}_{\text{selection bias}}. \end{aligned} \tag{4}$$

- The result is decomposed into two components: ***average treatment effect on treated*** *(ATT)* and ***selection bias***. The latter is simply the difference in average outcome under non-treatment between those who were treated and those who were not.

Problem

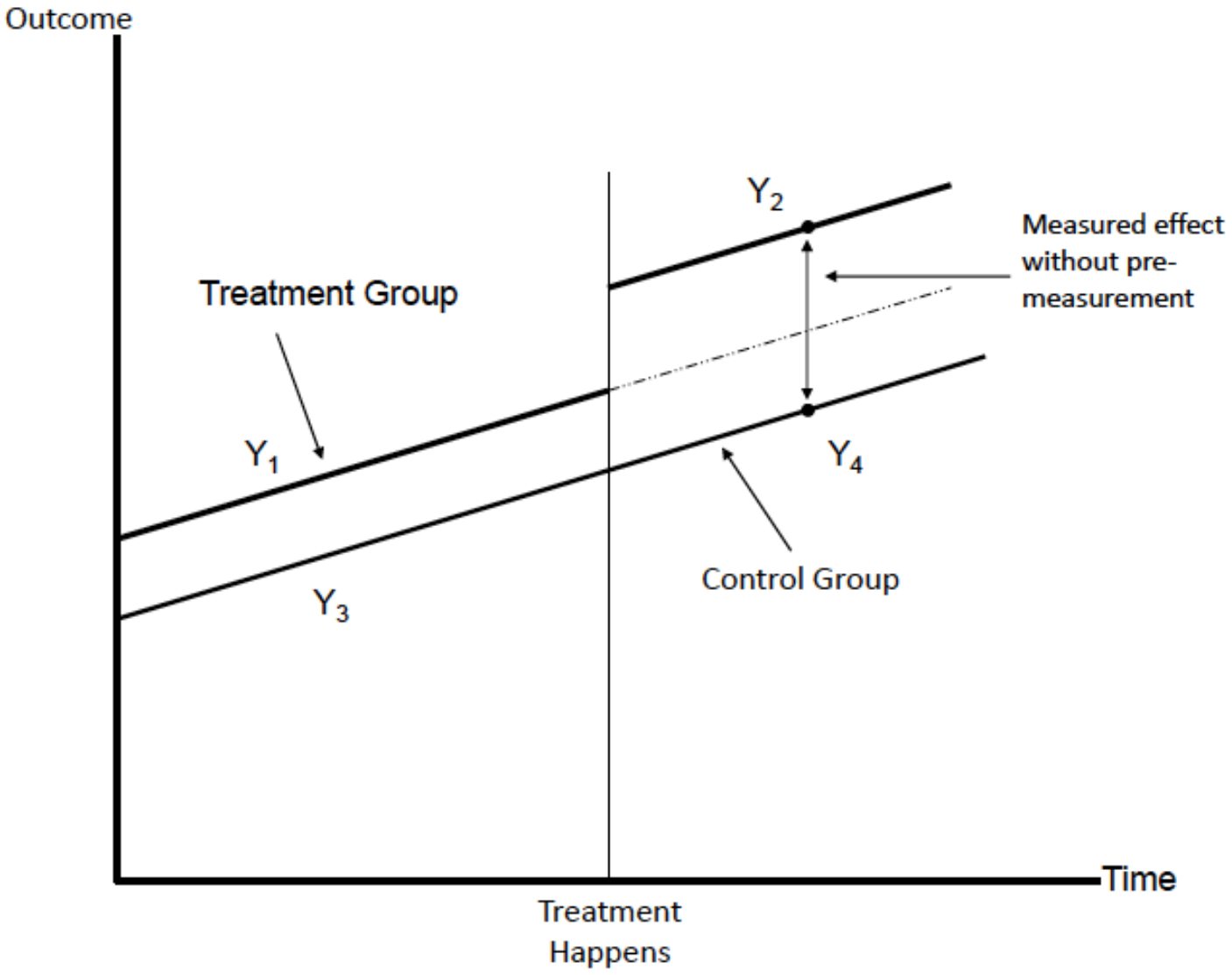
Actual treatment status D_i is not independent of potential outcomes.

- One of the solutions is ***randomly assign treatment*** to individuals in the population since D_i becomes independent of potential outcomes, and so the selection bias disappears.
- Under this,

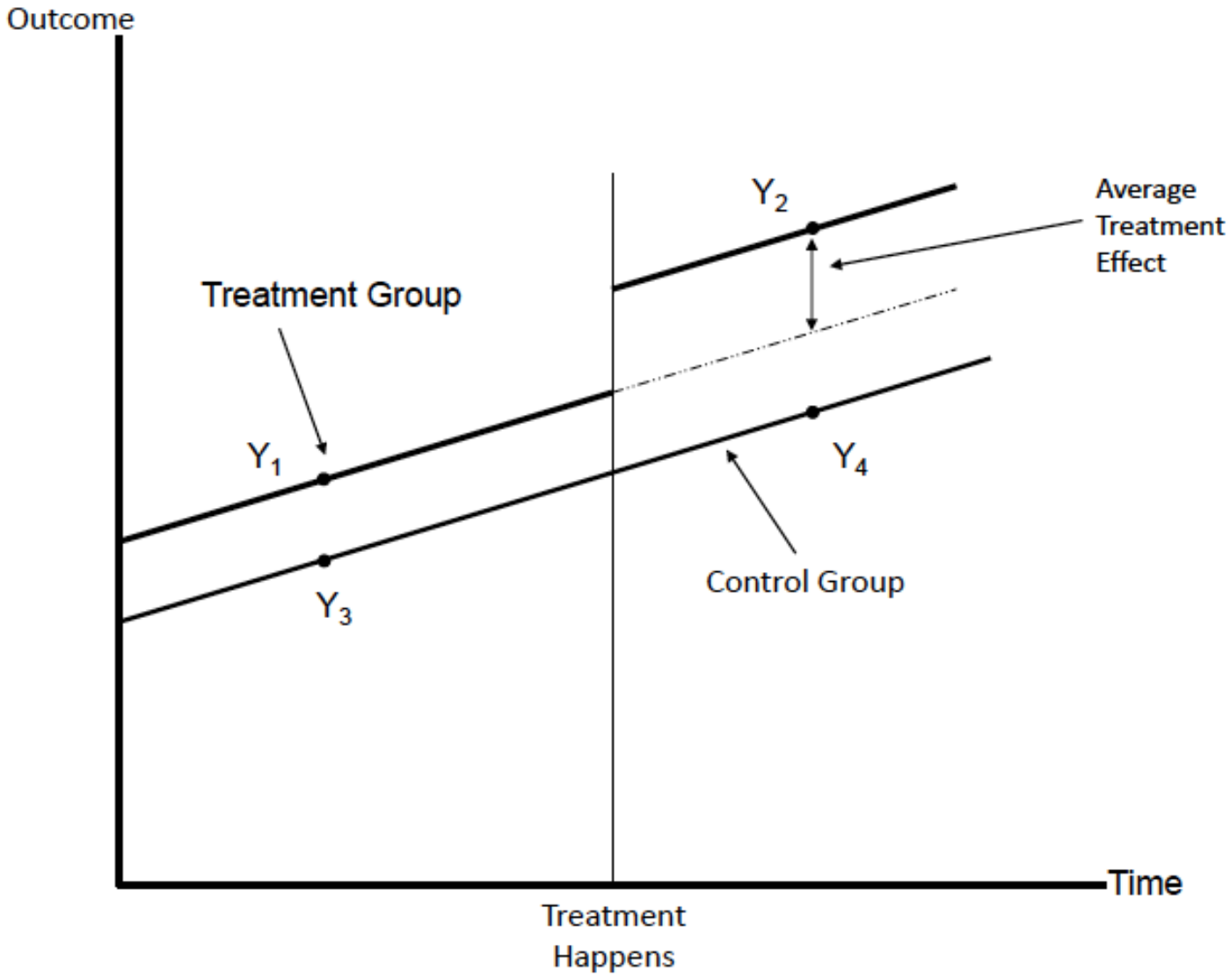
$$\begin{aligned} & E\left[Y_i \mid D_i = 1\right] - E\left[Y_i \mid D_i = 0\right] \\ = & E\left[Y_{1i} \mid D_i = 1\right] - E\left[Y_{0i} \mid D_i = 1\right], \end{aligned}$$

  since $E\left[Y_{0i} \mid D_i = 1\right] - E\left[Y_{0i} \mid D_i = 0\right] = 0$.
- This enables to infer ATT simply from the difference in means. This is because ATT coincides with the **average treatment effect** (ATE) in the entire population:
  $E\left[Y_{1i} \mid D_i = 1\right] - E\left[Y_{0i} \mid D_i = 1\right] = E\left[Y_{1i}\right] - E\left[Y_{0i}\right] = \mathrm{ATE}.$

Program Evaluation



Program Evaluation



Program Evaluation

- **Under random assignment**, we can obtain ATT and ATE by running ordinary least squares (OLS) regressions,

$$Y_i = \alpha + \rho D_i + \varepsilon_i.$$

- Suppose that the treatment effect is the **same** for everybody, such that $Y_{1i} - Y_{0i} = \rho$.
- Using $Y_i = Y_{0i} + (Y_{1i} - Y_{0i}) D_i$, we can re-express it as $Y_i = EY_{0i} + (Y_{1i} - Y_{0i}) D_i + (Y_{0i} - EY_{0i}) = \alpha + \rho D_i + \varepsilon_i$, where $\alpha \equiv EY_{0i}$ and $\varepsilon_i \equiv Y_{0i} - EY_{0i}$.
- It then immediately follows that

$$\begin{aligned} E[Y_i \mid D_i = 1] - E[Y_i \mid D_i = 0] = \rho \\ + E[\varepsilon_i \mid D_i = 1] - E[\varepsilon_i \mid D_i = 0]. \end{aligned}$$

Claim

The ***selection bias*** amounts to non-zero correlation between the regression error term ε_i and the regressor D_i: what we called ***endogeneity bias*** is known as ***selection bias*** here.

- The correlation reflects the difference in potential outcomes (under no treatment) between those who get treated and those who don't.
- To see this, note that
  $E[\varepsilon_i \mid D_i = 1] - E[\varepsilon_i \mid D_i = 0] = E[Y_{0i} \mid D_i = 1] - E[Y_{0i} \mid D_i = 0]$.
- **Clearly, if D_i is randomly assigned, there is no selection bias, so that a regression of observed outcomes Y_i on actual treatment status D_i estimates the causal effect**.

Selection Bias and Controls

- One of the potential solutions to the selection bias is to control for the observables (***if selection happened to be on observables***).
  - There is, of course, a possibility of ***selection over unobservables***.
- Regarding selection over observables, you might think ***the more controls the better***.
- Sadly **no**, there are 2 kinds of ways this can go wrong: **bad controls**.

1. Putting variables that are themselves affected by the variable of interest on the right hand side of the regression.
2. Putting bad proxies for unmeasured variables on the right hand side of the regression.

Bad Control

- Imagine the following exercise:
  - We are trying to estimate the effect of going to college (university) on earnings.
  - People can work in 2 occupations: Blue Collar and White Collar.
  - Clearly occupation is correlated with both college and earnings, so should it be a control?
  - Problem: College affects both occupational choice, and earnings.
- Let's use our potential outcomes framework to study this:

$$y_i = C_i y_{1i} + (1 - C_i) y_{0i},$$
$$W_i = C_i W_{1i} + (1 - C_i) W_{0i},$$

where $C_i = 1$ if go to college (0 otherwise) and $W_i = 1$ if white collar (0 if blue collar) and $y_{1i}, y_{0i}, W_{1i}, W_{0i}$ are the potential outcomes.

Bad Control

- Let's assume C_i is randomly assigned, so that it is independent of all the potential outcomes: $\{y_{1i}, y_{0i}, W_{1i}, W_{0i}\} \perp C_i$.
- So, we can estimate the effect of college on earnings and occupational choice as follows:

$$\mathbb{E}[y_i|C_i = 1] - \mathbb{E}[y_i|C_i = 0] = \mathbb{E}[y_{1i} - y_{0i}],$$
$$\mathbb{E}[W_i|C_i = 1] - \mathbb{E}[W_i|C_i = 0] = \mathbb{E}[W_{1i} - W_{0i}].$$

- The problem is that the **comparison of earnings conditional on W_i is *not* the causal effect of college conditional on occupation because of a selection problem**.
- College changes the composition of people in each occupation: **college affects white collar earnings, but it also affects *who* becomes a white collar worker**.

Bad Control

- Imagine regressing y_i on C_i in the subsample of white collar workers:

$$\begin{aligned}&\mathbb{E}\left[y_i|W_i=1,C_i=1\right]-\mathbb{E}\left[y_i|W_i=1,C_i=0\right],\\&\quad=\mathbb{E}\left[y_{1i}|W_i=1,C_i=1\right]-\mathbb{E}\left[y_{0i}|W_i=1,C_i=0\right].\end{aligned}$$

- Since C_i is randomly assigned and independent of the potential outcomes:

$$\begin{aligned}&\mathbb{E}\left[y_{1i}|W_i=1,C_i=1\right]-\mathbb{E}\left[y_{0i}|W_i=1,C_i=0\right]\\&\quad=\mathbb{E}\left[y_{1i}|W_{1i}=1\right]-\mathbb{E}\left[y_{0i}|W_{0i}=1\right]\\&\quad=\underbrace{\mathbb{E}\left[y_{1i}-y_{0i}|W_{1i}=1\right]}_{\text{causal effect}}+\underbrace{\mathbb{E}\left[y_{0i}|W_{1i}=1\right]-\mathbb{E}\left[y_{0i}|W_{0i}=1\right]}_{\text{selection bias}}.\end{aligned}$$

Bad Proxies

- Again, let's take a concrete example: Let's say you want to measure the effect of **schooling S_i on earnings y_i**:

$$y_i = \alpha + \rho S_i + \gamma a_i + \varepsilon_i.$$

- If we don't control at all for ability a_i we have omitted variable bias:

$$\hat{\rho} = \rho + \frac{\text{Cov}(S, a)}{\text{Var}(S)}\gamma.$$

- Imagine we had the scores on an IQ test at age 14, before people make any schooling choices (assume everyone completes 8th grade): a_{ei}.
  - then controlling for a_{ei} fixes the problem: $\mathbb{E}[S_i \varepsilon_i] = \mathbb{E}[a_{ei} \varepsilon_i] = 0$.

Bad Proxies

- These kinds of measures are **very hard to come by** though.
- Imagine instead that you had test scores on a test employers use to screen applicants a_{Ii}.
  - The problem is that this ability measure is measured after schooling choices have been made.
  - If the measure is affected by schooling, then we have a problem:

$$a_{Ii} = \pi_0 + \pi_1 S_i + \pi_2 a_i.$$

- Substituting out a_i we see that

$$y_i = \left(\alpha - \gamma\frac{\pi_0}{\pi_2}\right) + \left(\rho - \gamma\frac{\pi_1}{\pi_2}\right) S_i + \frac{\gamma}{\pi_2} a_{Ii} + \varepsilon_i.$$

Bad Proxies

- What can be done?
- In this example $\gamma > 0$, $\pi_1 > 0$, and $\pi_2 > 0$ so $\rho - \gamma \frac{\pi_1}{\pi_2} < \rho$.
- We can regress a_{Ii} on S_i to get a sense of how large π_1 is likely to be. If π_1 is small, maybe not too much of a problem,
- Also, note that:
  - regression without ability measure **overestimates** ρ,
  - regression controlling for a_{Ii} **underestimates** ρ,
  - so we can put **bounds** on ρ.

Outline

Exogeneity and Endogeneity

- From applied perspective, the most important thing that OLS has to assume is exogeneity:

$$\mathbb{E}(\varepsilon|\mathbf{X}) = 0.$$

  - Implies that the errors ε and the right-hand side variables $\mathbf{X}$ are uncorrelated.
- Some of the many possible ways this might be violated include

1. **Omitted Variable Bias**:
   - Right-hand side variables are correlated with some other, omitted variable that directly affects the outcome y.
2. **Endogenous right-hand side variables**:
   - Right-hand side variables are actually the outcome of some other variable we don't include.
3. **Measurement Error**.

Instrumental Variables

- Generally, there are two ways to deal with **endogeneity**:
  - Include more variables. They do not need to be exactly equal to the omitted variables. However, conditional on their values, there must be no correlation between the regressors and the error of the regression. Such variables are called **proxy** variables.
  - Alternatively, use **instrumental variables** (IV) estimation and its more general version – **two stage least squares** (TSLS) or **generalised IV** or **generalised method of moments** estimation.

Instrumental Variables

- Consider a simple regression model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \tag{5}$$

where $\mathrm{Cov}(x_i, \varepsilon_i) \neq 0$ even in large samples. In such a case, OLS yields inconsistent estimates, since $\mathrm{plim}\hat{\beta}_1 = \beta_1 + \frac{\mathrm{Cov}(x_i, \varepsilon_i)}{\mathrm{Var}(x_i)} \neq \beta_1$.

- Suppose that there exists an instrument z_i such that

$$\begin{aligned} \mathrm{Cov}(z_i, \varepsilon_i) &= 0, \quad \text{IV Exogeneity}, \\ \mathrm{Cov}(z_i, x_i) &\neq 0, \quad \text{IV Relevance}. \end{aligned}$$

- Given these conditions are met (i.e., an IV exists), use (5) to arrive at

$$\mathrm{Cov}(z_i, y_i) = \beta_1 \mathrm{Cov}(z_i, x_i) \text{ and } \beta_1^{IV} = \frac{\mathrm{Cov}(z_i, y_i)}{\mathrm{Cov}(z_i, x_i)}.$$

- Use a sample analogue to get an estimator: $\hat{\beta}_1^{IV} = \frac{\sum_i (z_i - \bar{z})(y_i - \bar{y})}{\sum_i (z_i - \bar{z})(x_i - \bar{x})}$.

Two-stage Least Squares

- Consider a so-called Mincerian regression

$$\log(wage_i) = \beta_0 + \beta_1 educ_i + \beta_2 exper_i + \beta_3 exper_i^2 + \varepsilon_i. \quad (6)$$

- **Education** is an **endogenous variable**. Suppose you have access to two IVs, z_1 – mother's education and z_2 – father's education.
- Technically, we deal with an overidentified model because we have 2 IVs for 1 endogenous variable. In such a case, there exist the so-called **overidentifying restrictions**.
- Instead of using either z_1 or z_2 as an instrument, we can choose the linear combination that is most highly correlated with education. It is given by the **reduced form regression** (first stage).
- Then the coefficients in the original regression can be estimated by the IV procedure that uses fitted value from the first stage as an instrument for education. This procedure is known as **the two stage least squares** (TSLS).

Two-stage Least Squares and Exogeneity

- If there is no endogeneity, TSLS is **not necessary and costly**.
  - If you suspect that x_1 is endogenous, you can run the first stage (reduced form) on all instruments and exogenous variables (summarised in $\boldsymbol{z}$) from the original equation, defined by $y_i = \boldsymbol{x}'\beta + \varepsilon$:
  $$x_1 = \boldsymbol{z}'\pi + e.$$
  - If x_1 is exogenous, then $\mathrm{Cov}(\varepsilon, e) = 0$. Otherwise, $\varepsilon = \delta e + u$, where $\delta \neq 0$.
  - You can, therefore, run $y_i = \boldsymbol{x}'\beta + \delta e + u$, and test $H_0 : \delta = 0$.
  - The partialling-out approach (Frisch-Waugh theorem – refer to the previous slides) implies that the OLS estimate of β_1 in regression $y_i = \boldsymbol{x}'\beta + \delta e + u$ must be identical to $\hat{\beta}_{TSLS}$.
- This gives a **useful interpretation** of TSLS: adding $\hat{e}$ to the original equation clears up the endogeneity of x_1.

Projection Matrices and IV

- The power of the use of projection matrices, however, is best seen in more complex environments, such as dealing with **instrumental variable estimation**.
- Suppose $E\left(\boldsymbol{X}'\varepsilon\right) \neq 0$ due to simultaneity, omitted variables or errors-in-variables.
- Then consider some matrix $\boldsymbol{Z}$ which is formed of variables uncorrelated with ε.
- This matrix defines a projection matrix $\boldsymbol{P_Z} = \boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}$, so that anything that is projected onto the space spanned by $\boldsymbol{Z}$ will be uncorrelated with ε by the definition of $\boldsymbol{Z}$.
- Then transform the original model $\boldsymbol{y} = \boldsymbol{X}\beta + \varepsilon$ into

$$\boldsymbol{P_Z y} = \boldsymbol{P_Z X}\beta + \boldsymbol{P_Z}\varepsilon,$$

and observe that $\boldsymbol{P_Z X} = \boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{X}$ is the fitted value from a regression of $\boldsymbol{X}$ on $\boldsymbol{Z}$ ("**first stage**") and
$E\left(\left(\boldsymbol{P_Z X}\right)'\boldsymbol{P_Z}\varepsilon\right) = E\left(\boldsymbol{X}'\boldsymbol{P_Z}\varepsilon\right) = 0$.

Projection Matrices and IV

- This is the ***generalised instrumental variables estimator***, defined as

$$\begin{aligned}\hat{\beta}_{IV} &= \left(X'P_ZX\right)^{-1}X'P_Zy \\ &= \left(X'P_ZX\right)^{-1}X'P_Z\left(X\beta+\varepsilon\right) \\ &= \beta+\left(X'P_ZX\right)^{-1}X'P_Z\varepsilon,\end{aligned}$$

  and the bias given by

$$\left(\left(X'Z\right)\left(Z'Z\right)^{-1}\left(Z'X\right)\right)^{-1}\left(X'Z\right)\left(Z'Z\right)^{-1}Z'\varepsilon.$$

- However, dividing each term by N, and applying LLN, we can demonstrate that all terms go to finite matrices while $\left(Z'\varepsilon\right)/N \xrightarrow{p} 0$, stemming from $E\left(Z'\varepsilon\right)=0$.
- Hence, $\hat{\beta}_{IV} \xrightarrow{p} \beta$; similarly, CLT can be invoked by scaling $\hat{\beta}_{IV}-\beta$ by $\sqrt{N}$.
- Hence, IV estimator is **consistent**, **asymptotically normally distributed** but **biased** in general, since even though $E\left(X'P_Z\varepsilon\right)=0$, $E\left(\left(X'P_ZX\right)^{-1}X'P_Z\varepsilon\right)$ may not be zero, since $\left(X'P_ZX\right)^{-1}$ and $X'P_Z\varepsilon$ are not independent.

Two-stage Least Squares and Tests

- To test whether the IVs conditions are satisfied, we can resort to
  - **First stage regression** (**reduced form**), and test whether instruments are sufficiently strongly correlated with an endogenous variable.
  - A useful rule of thumb is to view instruments as weak if $F < 10$ (Stock and Yogo).
  - To (partially) test whether IVs are exogenous, we require **more IVs than endogenous variables**.
    - Run TSLS, obtain the residuals $\hat{\varepsilon}$ (not the residuals from the second stage of TSLS!)
    - Regress $\hat{\varepsilon}$ on all instruments and all exogenous regressors. Obtain R^2.
    - Under the null hypothesis that all instruments are exogenous, nR^2 is asymptotically distributed as χ^2_q, where q is the number of the overidentifying restrictions (i.e., a difference between a number of IVs and endogenous variables).
    - Rejection of the null makes you conclude that some IVs are endogenous.

More Details on Durbin-Wu-Hausman

- The Durbin-Wu-Hausman test uses the following logic:
  1. If there really aren't any endogenous RHS variables, then both $\hat{\beta}_{OLS}$ and $\hat{\beta}_{IV}$ are **consistent**.
  2. If there is at least one endogenous RHS variable, then $\hat{\beta}_{IV}$ is **consistent**, but $\hat{\beta}_{OLS}$ **isn't**.
  3. Therefore we can use the distance $\mathbf{d} = \hat{\beta}_{IV} - \hat{\beta}_{OLS}$ to look for endogeneity. If it is "too far" from 0, we conclude we have an endogeneity problem.
- Note that of course, for this to work, we need to assume that $\hat{\beta}_{IV}$ *is* consistent, i.e. that the **instruments are valid**!

More Details on Durbin-Wu-Hausman

- So, to test whether we need to bother doing IV, we have the following hypotheses:
  - H_0 : plim $\mathbf{d} = \mathbf{0}$ (OLS is consistent),
  - H_1 : plim $\mathbf{d} \neq \mathbf{0}$ (OLS is inconsistent).
- The **Wald statistic** for this test is

$$H = \mathbf{d}' \left[\widehat{\text{aVar}}\left(\mathbf{d}\right)\right]^{-1} \mathbf{d}.$$

- So we need

$$\begin{aligned}\text{aVar}\left(\mathbf{d}\right) = \text{aVar}\left(\hat{\beta}_{IV} - \hat{\beta}_{OLS}\right) &= \text{aVar}\left(\hat{\beta}_{IV}\right) \\ &+ \text{aVar}\left(\hat{\beta}_{OLS}\right) - \text{aCov}\left(\hat{\beta}_{IV}, \hat{\beta}_{OLS}\right) - \text{aCov}\left(\hat{\beta}_{OLS}, \hat{\beta}_{IV}\right).\end{aligned}$$

- **Hausman** showed that $\text{Cov}\left(\hat{\beta}_{IV}, \hat{\beta}_{IV}\right) = \text{Var}\left(\hat{\beta}_{OLS}\right)$, so this becomes

$$\text{aVar}\left(\mathbf{d}\right) = \text{aVar}\left(\hat{\beta}_{IV}\right) - \text{aVar}\left(\hat{\beta}_{OLS}\right).$$

More Details on Durbin-Wu-Hausman

- So now we have our Wald statistic, the Hausman statistic:

$$H = \frac{\mathbf{d}' \left[\left(\hat{\mathbf{X}}' \hat{\mathbf{X}} \right)^{-1} - \left(\mathbf{X}' \mathbf{X} \right)^{-1} \right]^{-1} \mathbf{d}}{s^2}.$$

- Since $\mathbf{X}$ and $\mathbf{Z}$ have $k_1$variables in common, we are really only testing the endogeneity of the remaining k_2 variables.
- So H has an F distribution with k_2 and $n - k - k_2$ degrees of freedom.

Equivalent to the Hausman Test

- An **alternative way** to do this is to run the regression:

$$\mathbf{y} = \mathbf{X}\beta + \hat{\mathbf{X}}_2\gamma + \boldsymbol{\varepsilon}^*,$$

  where $\hat{\mathbf{X}}_2$ are the fitted values in a regression of the endogenous part of $\mathbf{X}$, $\mathbf{X}_2$ on $\mathbf{Z}$.
- The **F test for the joint significance of γ in this regression is equivalent to the Hausman test** (Wu, 1973).

Overidentification Test

- We'd like to have a way of testing the exogeneity assumption: plim $(\mathbf{Z}'\varepsilon/n) = 0$.
  - Why not try and base this test on its sample analog $\mathbf{Z}'\hat{\varepsilon}/n$?
  - When the number of instruments is the same as the number of endogenous variables, **we can't do this**, the assumption is true mechanically:

$$\begin{aligned}\mathbf{Z}'\hat{\varepsilon} &= \mathbf{Z}'\left(\mathbf{y} - \mathbf{X}\hat{\beta}_{\mathbf{IV}}\right) \\ &= \mathbf{Z}'\left(\mathbf{y} - \mathbf{X}\left(\mathbf{Z}'\mathbf{X}\right)^{-1}\mathbf{Z}'\mathbf{y}\right) \\ &= \mathbf{Z}'\mathbf{y} - \mathbf{Z}'\mathbf{X}\left(\mathbf{Z}'\mathbf{X}\right)^{-1}\mathbf{Z}'\mathbf{y} = \mathbf{0}.\end{aligned}$$

Overidentification Test

- When we have more instruments than endogenous variables ($l > k_2$) then $\mathbf{Z}'\hat{\varepsilon}/n$ won't be exactly 0.
- 2SLS tries to make $\mathbf{Z}'\varepsilon$ as close to 0 as possible, but it's not exactly 0.
- How close to 0 it gets is a test of $l - k_2$ of the assumptions in $\mathbf{Z}'\varepsilon = 0$,
- So we base our test on how close to 0 it manages to get:
  - H_0 : plim $\mathbf{Z}'\varepsilon/n = 0$,
  - H_1 : plim $\mathbf{Z}'\varepsilon/n \neq 0$.

Overidentification Test

- We'll base our test on the sample moment

$$\bar{\mathbf{m}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{z}_i \hat{\varepsilon}_{2SLS,i} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{z}_i \left(y_i - \mathbf{x}_i' \hat{\beta}_{2SLS} \right).$$

- And create the **Wald statistic**

$$W = \bar{\mathbf{m}}' \left[\text{Var}\left(\bar{\mathbf{m}} \right) \right]^{-1} \bar{\mathbf{m}},$$

  which asymptotically has distribution $\chi^2 \left[l - k_2 \right]$.
- Of course, we don't know the variance of $\bar{\mathbf{m}}$ so we replace it with an estimator to get the feasible statistic:

$$W = \left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{z}_i \hat{\varepsilon}_{2SLS,i} \right)' \left(\frac{1}{n^2}\sum_{i=1}^{n} \hat{\varepsilon}_{2SLS,i}^2 \mathbf{z}_i \mathbf{z}_i' \right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{z}_i \hat{\varepsilon}_{2SLS,i} \right).$$

Endogeneity in the Policy Evaluation Framework

- Return to our **potential outcomes framework**, and assume $y_{si} = f_i(s)$ where y_{si} is wages, s is schooling and $f_i(\cdot)$ is an individual-specific function that links the two, which we are trying to estimate.
- Let's assume that

$$f_i(s) = \alpha + \rho s + \eta_i.$$

  and note that ρ is the same for everybody.
- Let $\eta_i = A_i'\gamma + \nu_i$ where A_i is (unobserved) ability.
- Finally, let's assume that A_i is the only reason η_i and s_i are correlated: $\mathbb{E}[s_i \nu_i] = 0$,

$$y_i = \alpha + \rho s_i + A_i'\gamma + \nu_i.$$

Endogeneity in the Policy Evaluation Framework

- If we could observe A_i we'd be fine. We'd regress y_i on s_i and A_i and calculate $\hat{\rho}$ which has plim $\hat{\rho} = \rho$
- A good instrumental variable(s) allows us to estimate ρ even without observing A_i
- Recall that a good instrumental variable z_i requires:

1. **Relevance:** $\text{Cov}(s_i, z_i) \neq 0$
2. **Exclusion Restriction:** $\text{Cov}(\eta_i, z_i) = 0$
   - 2.1 No direct effect: z_i doesn't belong on the RHS directly.
   - 2.2 No indirect effect: z_i is not correlated with relevant omitted variables.

Endogeneity in the Policy Evaluation Framework

- With these two assumptions we can write:

$$\begin{aligned}\text{Cov}(y_i, z_i) &= \text{Cov}(\alpha + \rho s_i + \eta_i, z_i) \\ &= \rho \underbrace{\text{Cov}(s_i, z_i)}_{\neq 0 \text{ (relevance)}} + \underbrace{\text{Cov}(\eta_i, z_i)}_{=0 \text{ (exclusion)}}.\end{aligned}$$

- So we have identified ρ!

$$\rho = \frac{\text{Cov}(y_i, z_i)}{\text{Cov}(s_i, z_i)} = \frac{\text{Cov}(y_i, z_i)/\text{Var}(z_i)}{\text{Cov}(s_i, z_i)/\text{Var}(z_i)}.$$

- **The coefficient ρ is the ratio of the population coefficient in a regression of y_i on z_i (we call this the *reduced form*) to the population coefficient in a regression of s_i on z_i (we call this the *first stage*).**

Outline

Complexity Reduction

- So far we focused on causality, assuming that
  - Functional form is correct;
  - Degrees of freedom are sufficient ($n >> k$);
  - Asymptotic properties approximate well small sample behaviour of estimators.
- We will talk about two issues:
  - What if some (or all) of these assumptions are violated?
  - What is the difference between prediction and causal inference?

Complexity Reduction

- Turn attention to finite sample properties.
- **Ridge regression** – a popular method of estimation in both econometrics and machine learning.
- Ridge regression connects to ideas at the heart of finite sample theory:
  - complexity,
  - prior knowledge,
  - bias–variance trade-off.

Complexity Reduction

- Start with the OLS setting with the standard assumptions covered before.
- Recall the usual OLS estimator:

$$\hat{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{b} \in \mathbb{R}^K} \sum_{i=1}^{n} \left(y_i - \boldsymbol{x}_i' \boldsymbol{b} \right)^2 .$$

- OLS estimator minimises the empirical risk under quadratic loss when the hypothesis space is the set of linear functions.
- Under our assumptions, the OLS estimator:
  - unbiased for β,
  - has the lowest variance among all linear unbiased estimators of β.

Complexity Reduction

- More natural way to evaluate estimators is to consider their mean squared error:

$$MSE\left(\hat{\beta}, \beta\right) = \mathbb{E}\left\|\hat{\beta} - \beta\right\|^2.$$

- Alternatively,

$$MSE\left(\hat{\beta}, \beta\right) = \underbrace{\mathbb{E}\left\|\hat{\beta} - \mathbb{E}\hat{\beta}\right\|^2}_{\text{variance}} + \left\|\underbrace{\mathbb{E}\hat{\beta} - \beta}_{\text{bias}}\right\|^2.$$

- Minimisation of MSE involves a trade-off between:
  1. variance term,
  2. bias term.

Ridge Regression

- There exists a biased linear estimator with lower mean squared error than $\hat{\boldsymbol{\beta}}$.
- The estimator is the solution to the modified least squares problem (ℓ_2 penalty):

$$\min_{\boldsymbol{b}\in\mathbb{R}^K} \left\{ \sum_{i=1}^{n} \left(y_i - \boldsymbol{x}_i'\boldsymbol{b}\right)^2 + \lambda \left\|\boldsymbol{b}\right\|^2 \right\}, \tag{7}$$

  where $\lambda \geq 0$ is called the **regularisation (penalty) parameter.**
- Minimising the empirical risk plus a term that penalises large values of $\left\|\boldsymbol{b}\right\|$.
- The solution to equation (7) is

$$\hat{\boldsymbol{\beta}}_\lambda = \left(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}'\mathbf{y}.$$

Ridge Regression

- The estimator $\hat{\beta}_\lambda$ is called the **ridge regression estimator**.
- Note:
  - $\hat{\beta}_\lambda$ is the OLS estimator when $\lambda = 0$, and
  - $\hat{\beta}_\lambda$ is biased whenever $\lambda > 0$.

Theorem

Under the OLS assumptions (as covered before), there exists a $\lambda > 0$ such that

$$MSE\left(\hat{\beta}_\lambda, \beta\right) < MSE\left(\hat{\beta}, \beta\right).$$

Proof.

See Hoerl and Kennard (1970). □

Ridge Regression

- Traditional view of **ridge regression**:
  - OLS assumptions valid, however, instances where $\mathbf{X}'\mathbf{X}$ is almost singular due to strong correlation between regressors,
  - in this case, inverting $\mathbf{X}'\mathbf{X}$ is numerically unstable,
  - stabilise the inversion by adding some positive value of λ.
- Alternative view of ridge regression:
  - standard **OLS assumptions are implausible**,
  - obtaining the regression function f (an infinite dimensional object) with a finite amount of data is ill-posed,
  - **regularisation term in ridge regression manages complexity of the candidate functions used to approximate the regression function**.

Ridge Regression

- Unlike least squares, which generates only **one set of coefficient estimates**, ridge regression will produce a different set of coefficient estimates, $\hat{\beta}_\lambda$, for each value of λ. Selecting a good value for λ is critical.
- A standard approach to determining λ is **cross-validation**.
  - Divide the available set of observations into two parts, a **training set** and a **validation set** or **hold-out set**. The model is fit on the training set, and the fitted model is used to predict the responses for the observations in the validation set. The resulting validation set error rate—typically assessed using MSE in the case of a quantitative response—provides an **estimate of the test error rate**.
  - Cross-validation is often used for parameter tuning by doing cross-validation for several (or many) possible values of a parameter and choosing the parameter value that gives the lowest cross-validation average error.

LASSO Alternative

- An alternative to the ridge regression is LASSO regression (Tibshirani, 1996).
  - One obvious disadvantage of ridge regression – it will include **all predictors in the final model** (unless $\lambda = \infty$).
- The Lasso has become popular in the recent literature, partly because of its ability to **nearly achieve the risk (mean-squared-error) of infeasible optimal selection** in classical regression. It also sets some parameters **exactly to zero**.
- The Lasso minimises the sum of squared errors subject to an ℓ_1 penalty:

$$\hat{\beta}^{Lasso} = \arg \min_{\boldsymbol{b} \in \mathbb{R}^K} \left\{ \sum_{i=1}^{n} \left(y_i - \boldsymbol{x}_i' \boldsymbol{b} \right)^2 + \lambda \left\| \boldsymbol{b} \right\| \right\},$$

where, as before, $\lambda \geq 0$ is called the **regularisation (penalty) parameter.**

LASSO Alternative

- One feature of the Lasso estimator $\hat{\beta}^{Lasso}$ is that it simultaneously performs **selection** (it yields sparse models) and **shrinkage**.
  - That is, some of the individual coefficient estimates will equal zero, so that $\hat{\beta}_k^{Lasso} = 0$ for some k.
- Let $\hat{\mathbf{S}}$ be a selector matrix which selects the coefficients not set to zero, so that $\mathbf{X}\hat{\mathbf{S}}$ are the regressors "selected" by the Lasso.
- The **OLS post-Lasso estimator** is least-squares performed on these variables, and can be written as:

$$\hat{\beta}^{Lasso} = \left(\hat{\mathbf{S}}'\mathbf{X}'\mathbf{X}\hat{\mathbf{S}}\right)^{-1}\hat{\mathbf{S}}'\mathbf{X}'\mathbf{y}.$$

LASSO Alternative

- As demonstrated by Hansen (2013), the performance of Lasso estimation depends critically upon the values of the coefficients.
- When the true coefficients satisfy a **strong sparsity condition** (there are many coefficients truly equal to zero), then Lasso estimation can indeed have low MSE.
- However, in other contexts the estimator need not have low MSE, and **can actually perform worse than simple OLS**.

Machine Learning + Causality?

- Victor Chernozhukov (MIT) (`http://www.mit.edu/~vchern/`) and his co-authors have initiated a growing literature on causality inference using tools from machine learning.
- Consider merging regression with the program evaluation notation:

$$Y_i = D_i\alpha + \sum_{j=1}^{k} X_{ij}\beta_j + \varepsilon_i, \tag{8}$$

where D_i is treatment and α is the **treatment effect**, main objective, whereas a rich set of controls is $\{X_{ij}\}$ and β_j are **nuisance parameters**.

- Note that $\{X_{ij}\}$ can also contain transformation of "raw" controls in an effort to make models more flexible (functional form consideration).

Machine Learning + Causality? Yes, but be careful!

- If you try to do model selection in model (8) directly, you are doing it **wrong**. Have to do **additional selection** to make it right.
- Post-double selection procedure by Belloni, Chernozhukov, Hansen (2013):
  - Include X_i if it is a significant predictor of Y_i as judged by a conservative test (t-test, LASSO, etc).
  - Include X_i if it is a significant predictor of D_i as judged by a conservative test (t-test, LASSO, etc). Note that in the IV models, you must include X_i if it a significant predictor of Z_i.
  - Refit the model after selection, use standard confidence intervals.

Proposition

Double selection works in low-dimensional setting and in high-dimensional approximately sparse settings.

Intuition of Double Selection

- **Double Selection** — the selection among the controls X_i that predict either D_i **or** Y_i – creates **robustness**. It finds controls whose omission would lead to a "large" omitted variable bias, and includes them in the regression.
- In effect the procedure is a model selection version of **Frisch-Waugh-Lovell partialling-put procedure** for estimating linear regression.
- When conducted on published papers, there are positive and negative results:
  - Abortion rates on crime in the U.S., Donohue and Levitt (2001): result disappears after double selection.
  - Acemoglu et al (2014) results on democracy causing growth and Acemoglu et al (2001) results on institutions causing growth remain.

Model Selection

- Belloni and Chernozhukov (2013, 2014) find that LASSO provides high-quality model selection (technically, under approximate sparsity and restricted isometry conditions, LASSO and Root-LASSO find parsimonious models of approximately optimal size and the OLS can approximate the regression functions at the nearly optimal rates in the root mean squared error).
- This finding also holds for endogenous models, see Chernozhukov, Hansen, Spindler (2015).

Summary

- In this short course, we focused on the tools that are most often invoked by the applied economist. In particular, we have covered: **regression models, program evaluation, instrumental variables methods** and techniques from the field of **machine learning**.
- We have also touched a few statistical properties of estimators along with asymptotics and hypothesis testing in the context of endogeneity.
- We have necessarily abstracted from a number of topics, and other identification strategies.
  - For instance, synthetic control methodology, matching estimators (to be covered by Swapnil), causal graphs, many more machine learning tools applied to treatment effects, etc.
  - Refer to Gobillon and Magnac (2016) for a nice connection of and comparison between synthetic control and fixed effects methodologies, after covering panel data component of the course.
  - Good sources of further study include Imbens and Rubin (2015, Cambridge University Press), Morgan (2014, Cambridge University Press), Lee (2016, Oxford University Press) among many other contributions in this fast growing field.

Outline

Properties of Expectations

- Conditional density $f(Y|X) = f(X,Y)/f(X)$.
- Conditional expectation $E[Y|X] = \int Yf(Y|X)\,dY$.
- Properties of conditional expectations
  1. $E[E[Y|X_1,X_2]|X_1] = E[Y|X_1]$,
  2. $E[E[Y|X_1]|X_1,X_2] = E[Y|X_1]$,
  3. $E[h(X)\,Y|X] = h(X)\,E[Y|X]$.

Statistics Vocabulary

Definition

The rth moment (about the origin) of a random variable X, denoted by μ'_r, is the expected value of X^r; symbolically,

$$\mu'_r = \mathbb{E}\left(X^r\right) = \sum_x x^r f\left(x\right)$$

for $r = 0, 1, 2, \ldots$ when X is discrete (use integrals for continuous variables, i.e, $\mu'_r = \int x^r f\left(x\right) dx$).

- When $r = 1$, we have $\mu'_1 = \mathbb{E}\left(X^1\right) = \mathbb{E}\left(X\right) = \mu$, which is just the expected value of the random variable X. In other words, μ'_1 is the mean of the distribution of X, or simply the mean of X.

Statistics Vocabulary

Definition

The rth central moment of a random variable X, also known as the rth moment about the mean of random variable X, denoted by μ_r, is the expected value of $(X - \mu)^r$, where μ is the mean; symbolically,

$$\mu_r = \mathbb{E}\,(X - \mu)^r = \sum_x (x - \mu)^r\, f\,(x)$$

for $r = 0, 1, 2, \ldots$ when X is discrete. Note that $\mu_0 = 1$ and $\mu_1 = 0$ for any random variable for which μ exists.

- You can see that the second central moment, μ_2, is the covariance between X and X, or the variance of the distribution of X, or simply the variance of X.

Matrices

- A **matrix** is a rectangular array of numbers.
- An $r \times c$ matrix has r rows and c columns.

$$\mathbf{A} = [a_{ij}] = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1c} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2c} \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3c} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{r1} & a_{r2} & a_{r3} & \cdots & a_{rc} \end{bmatrix}.$$

- A matrix is *square* if $r = c$.

Vectors

- A **row vector** is a $1 \times c$ matrix

$$\mathbf{x} = (x_1, x_2, x_3, \ldots, x_c)\,.$$

- A **column vector** is an $r \times 1$ matrix

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_r \end{pmatrix}\,.$$

Square Matrices

- A square matrix is *symmetric* if $a_{ij} = a_{ji}\ \forall i, j$.
- The *main* or *principal* diagonal of a square matrix is $diag\,(A) = (a_{11}, a_{22}, a_{33}, \ldots, a_{cc})$.
- A square matrix is *diagonal* if $a_{ij} = 0\ \forall i \neq j$

$$\mathbf{A} = \begin{bmatrix} a_{11} & 0 & 0 & \cdots & 0 \\ 0 & a_{22} & 0 & \cdots & 0 \\ 0 & 0 & a_{33} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & a_{cc} \end{bmatrix}.$$

- The $n \times n$ **identity matrix** $\mathbf{I_n}$ is a square, diagonal matrix with $diag\,(\mathbf{I_n}) = (1, 1, 1, \ldots, 1)$.
- The *trace* of a matrix is the sum of its diagonal elements $tr\,(\mathbf{A}) = \sum_{i=1}^{n} a_{ii}$.

Transposing Matrices

- The **transpose** of a matrix flips the rows and columns of the matrix.
- We denote the transpose of $\mathbf{A}$ by $\mathbf{A}' \equiv [a_{ji}]$.
- If $\mathbf{A}$ is $r \times c$ then $\mathbf{A}'$ is $c \times r$.
- If $\mathbf{A}$ is symmetric, then $\mathbf{A} = \mathbf{A}'$.
- By definition $(\mathbf{A}')' = \mathbf{A}$.
- If $\mathbf{a}$ is a column vector, then $\mathbf{a}'$ is a row vector.
- $(\mathbf{A} + \mathbf{B})' = \mathbf{A}' + \mathbf{B}'$.

Multiplying Matrices

- To multiply two matrices $\mathbf{A}_{n\times m}$ and $\mathbf{B}_{m\times p}$, the number of *columns* in $\mathbf{A}$ must equal the number of *rows* in $\mathbf{B}$.
- Multiplying gives $\mathbf{AB} = \left[ab_{ij} = \sum_{k=1}^{p} a_{ik} \times b_{kj}\right]$.
- That is, the ij^{th} element of $\mathbf{A}$ is the product of the i^{th} row of $\mathbf{A}$ and the j^{th} column of $\mathbf{B}$:

$$(a_{i1}, a_{i2}, \cdots, a_{im}) \cdot \begin{pmatrix} b_{1j} \\ b_{2j} \\ b_{3j} \\ \vdots \\ b_{mj} \end{pmatrix} = \sum_{k=1}^{p} a_{ik} b_{kj} = ab_{ij}.$$

- In general, $\mathbf{AB} \neq \mathbf{BA}$.

Properties of Matrix Products

- $\mathbf{IA} = \mathbf{A}$ and $\mathbf{AI} = \mathbf{A}$ where $\mathbf{I}$ are appropriate identity matrices.
- $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$.
- $\mathbf{ABC} = (\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC})$.
- $\mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC}$; $(\mathbf{A} + \mathbf{B})\mathbf{C} = \mathbf{AC} + \mathbf{BC}$.
- for $\mathbf{A}_{n\times m}$, $\mathbf{A}'\mathbf{A}_{m\times m}$ and $\mathbf{AA}'_{n\times n}$ are symmetric, but not equal.
- For example, for $\mathbf{x}_{n\times 1}$

$$\mathbf{x}'\mathbf{x}_{1\times 1} = (x_1, x_2, \cdots, x_n) \cdot \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \sum_{i=1}^{n} x_i^2,$$

$$\mathbf{x}\mathbf{x}'_{n\times n} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \cdot (x_1, x_2, \cdots, x_n) = \begin{bmatrix} x_1^2 & x_1x_2 & \cdots & x_1x_n \\ x_2x_1 & x_2^2 & \cdots & x_2x_n \\ \vdots & \vdots & \ddots & \vdots \\ x_nx_1 & x_nx_2 & \cdots & x_n^2 \end{bmatrix}.$$

Properties of Matrix Products

- If $\mathbf{A}_{n\times n}$ is diagonal, then $\mathbf{x}'_{1\times n}\mathbf{A}\mathbf{x}_{n\times 1} = \sum_{i=1}^{n} a_{ii}x_i^2$ is a weighted sum of squares.
- $tr\left(\mathbf{A}'\mathbf{A}\right) = tr\left(\mathbf{A}\mathbf{A}'\right).$
- a (square) matrix $\mathbf{B}$ is **idempotent** if $\mathbf{B} = \mathbf{B}^2$.
- If $\mathbf{B}$ is symmetric and idempotent, then $\mathbf{B}'\mathbf{B} = \mathbf{B}$.

The Identity Vector and the Centering Matrix

- The *identity vector* is a column vector of ones

$$\iota_{n\times 1} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}.$$

- So $\iota'\iota = n$ and $\iota'\mathbf{x}_{n\times 1} = \mathbf{x}_{n\times 1}\iota = \sum_{i=1}^{n} x_i$ and $(\iota'\iota)^{-1}\iota'\mathbf{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$.
- $(\iota'\iota)^{-1}\iota\iota'$ is an $n \times n$ matrix with all elements equal to $1/n$.
- Since $\mathbf{x} = \mathbf{I}\mathbf{x}$,

$$\begin{bmatrix} x_1 - \bar{x} \\ x_2 - \bar{x} \\ \vdots \\ x_n - \bar{x} \end{bmatrix} = \left[\mathbf{x} - (\iota'\iota)^{-1}\iota\iota'\mathbf{x}\right] = \left[\mathbf{I} - (\iota'\iota)^{-1}\iota\iota'\right]\mathbf{x} = \mathbf{M}^0\mathbf{x},$$

where $\mathbf{M}^0$ is the centering matrix.

The Centering Matrix

- $\mathbf{M}^0$ has diagonal elements $(1 - 1/n)$ and off diagonal elements $-1/n$ so it is symmetric and idempotent
- We also find that

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = (\mathbf{M}^0\mathbf{x})' (\mathbf{M}^0\mathbf{x}) = \mathbf{x}'\mathbf{M}^{0\prime}\mathbf{M}^0\mathbf{x} = \mathbf{x}'\mathbf{M}^0\mathbf{x}.$$

- And we can combine 2 vectors $\mathbf{x}$ and $\mathbf{y}$ to arrive at

$$\begin{bmatrix} \sum_{i=1}^{n} (x_i - \bar{x})^2 & \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \\ \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) & \sum_{i=1}^{n} (y_i - \bar{y})^2 \end{bmatrix} = \begin{bmatrix} \mathbf{x}'\mathbf{M}^0\mathbf{x} & \mathbf{x}'\mathbf{M}^0\mathbf{y} \\ \mathbf{y}'\mathbf{M}^0\mathbf{x} & \mathbf{y}'\mathbf{M}^0\mathbf{y} \end{bmatrix}.$$

Matrix Rank

- The **column rank** of a matrix is the *maximum number of linearly independent columns* of the matrix.
- A column is linearly independent of the other columns if we can't write it as a *linear* combination of the other columns, e.g if

$$\mathbf{a} = \begin{pmatrix} 1 \\ 2 \\ 2 \end{pmatrix}, \mathbf{b} = \begin{pmatrix} 1 \\ 0 \\ 4 \end{pmatrix}, \mathbf{c} = \begin{pmatrix} 3 \\ 4 \\ 8 \end{pmatrix},$$

  then $\mathbf{c} = 2\mathbf{a} + \mathbf{b}$ and so $\mathbf{c}$ is not linearly independent of $\mathbf{a}$ and $\mathbf{b}$.
- Similarly, the **row rank** of a matrix is the *maximum number of linearly independent rows* of the matrix.
- It turns out the row and column rank of a matrix are equal, so we just talk about the **rank** of a matrix $\mathbf{A}$: $rank\,(\mathbf{A})$.

Matrix Rank

- For a matrix $\mathbf{A}_{r\times c}$ $rank\left(\mathbf{A}\right) \leq \min\left\{r, c\right\}$.
- If $rank\left(\mathbf{A}\right) = \min\left\{r, c\right\}$ then $\mathbf{A}$ has *full rank.*
- If $rank\left(A\right) < \min\left\{r, c\right\}$ then $\mathbf{A}$ is *rank deficient.*
- Furthermore, $rank\left(\mathbf{A}\right) = rank\left(\mathbf{A}'\mathbf{A}\right) = rank\left(\mathbf{A}\mathbf{A}'\right)$.

Matrix Determinant

- The **minor** $\mathbf{A}_{ij}$ of a matrix $\mathbf{A}$ is the matrix you get by deleting the i^{th} row and the j^{th} column from $\mathbf{A}$, e.g.,

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & 1 \\ 3 & 1 & 0 \\ 0 & 2 & 1 \end{pmatrix}, \mathbf{A}_{11} = \begin{pmatrix} 1 & 0 \\ 2 & 1 \end{pmatrix} \text{ \& } \mathbf{A}_{23} = \begin{pmatrix} 1 & 2 \\ 0 & 2 \end{pmatrix}.$$

- The **co-factor** of a matrix minor $\mathbf{A}_{ij}$ is defined as $c_{ij} = (-1)^{i+j} |\mathbf{A}_{ij}|$ where $|\mathbf{A}|$ is the determinant of a matrix, defined as...

Matrix Determinant

- The **determinant** of a matrix $\mathbf{A}_{n\times n}$ is

$$|\mathbf{A}| = \sum_{j=1}^{n} a_{ij} c_{ij} = \sum_{j=1}^{n} a_{ij} (-1)^{i+j} |\mathbf{A}_{ij}|$$

  for any row i (or column) of the matrix $\mathbf{A}$.
- Note that for a 2×2 matrix $\mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, $\mathbf{A}_{11} = d$, $\mathbf{A}_{12} = c$, $\mathbf{A}_{21} = b$ and $\mathbf{A}_{22} = a$. $|\mathbf{A}| = ad - bc$.
- Note that the determinant of a matrix is a *scalar.*

Inverse of a Matrix

- An $n \times n$ matrix $\mathbf{A}$ has an **inverse**, denoted by $\mathbf{A}^{-1}$ if and only if $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}_n$ and $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}_n$.
- Think of the scalar case: $a \times a^{-1} = 1 \leftrightarrow a^{-1} = \frac{1}{a}$.
- However, for matrices it's usually not true that $a_{ij}^{-1} = \frac{1}{a_{ij}}$.
- A matrix that has no inverse is *singular, or non-invertible.*
- Turns out that for an $n \times n$ matrix $\mathbf{A}$:

$$\mathbf{A} \text{ is singular } \iff |\mathbf{A}| = 0 \iff rank\,(\mathbf{A}) < n.$$

Properties of the Matrix Inverse

- If a matrix has an inverse, it is unique.
- $(\alpha \mathbf{A})^{-1} = (1/\alpha)\,\mathbf{A}^{-1}$ for scalar $\alpha \neq 0$ and $\mathbf{A}^{-1}$ exists.
- $\left(\mathbf{A}^{-1}\right)^{-1} = \mathbf{A}$.
- $(\mathbf{A}')^{-1} = \left(\mathbf{A}^{-1}\right)'$.
- $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$.

Matrix Calculus

- Recall, that if x is a scalar and we have a function $y = f(x)$

$$\frac{\partial f(x)}{\partial x} = \frac{\partial y}{\partial x}.$$

- When, instead $y = f(x_1, x_2, \ldots, x_n) = f(\mathbf{x})$, the *vector* of partial derivatives, the **gradient vector** is

$$\mathbf{g}(\mathbf{x}) = \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \begin{bmatrix} \partial y/\partial x_1 \\ \partial y/\partial x_2 \\ \vdots \\ \partial y/\partial x_n \end{bmatrix}.$$

Matrix Calculus

- The second derivatives matrix, or the **Hessian matrix**

$$\mathbf{H} = \begin{bmatrix} \partial^2 y/\partial x_1^2 & \partial^2 y/\partial x_1 \partial x_2 & \cdots & \partial^2 y/\partial x_1 \partial x_n \\ \partial^2 y/\partial x_2 \partial x_1 & \partial^2 y/\partial x_2^2 & \cdots & \partial^2 y/\partial x_2 \partial x_n \\ \vdots & \vdots & \ddots & \vdots \\ \partial^2 y/\partial x_n \partial x_1 & \partial^2 y/\partial x_n \partial x_2 & \cdots & \partial^2 y/\partial x_n^2 \end{bmatrix}$$

  is a square, symmetric matrix. Note that

$$\begin{aligned} \mathbf{H} &= \left[\frac{\partial \left(\partial y/\partial \mathbf{x} \right)}{\partial x_1} \frac{\partial \left(\partial y/\partial \mathbf{x} \right)}{\partial x_2} \cdots \frac{\partial \left(\partial y/\partial \mathbf{x} \right)}{\partial x_n} \right] \\ &= \frac{\partial \left(\partial y/\partial \mathbf{x} \right)}{\partial \left(x_1, x_2, \ldots, x_n \right)} = \frac{\partial \left(\partial y/\partial \mathbf{x} \right)}{\partial \mathbf{x}'} = \frac{\partial^2 y}{\partial \mathbf{x} \partial \mathbf{x}'}. \end{aligned}$$

Matrix Calculus

- In particular, consider a *linear* function

$$y = f(x_1, x_2, \ldots, x_n) = a_1 x_1 + a_2 x_2 + \ldots + a_n x_n,$$

which we can write as

$$y = \mathbf{a}'\mathbf{x} = \mathbf{x}'\mathbf{a} = \sum_{i=1}^{n} a_i x_i.$$

- Then we have that

$$\mathbf{g}(\mathbf{x}) = \frac{\partial \mathbf{a}'\mathbf{x}}{\partial \mathbf{x}} = \mathbf{a}.$$

Matrix Calculus

- Consider the set of linear functions

$$\mathbf{y} = \mathbf{A}\mathbf{x},$$

  so that $y_i = \mathbf{a}_i'\mathbf{x}$ where $\mathbf{a}_i'$ is the i^{th} row of $\mathbf{A}$.
- Then $\partial y_i/\partial \mathbf{x} = \mathbf{a}_i$, the transpose of the i^{th} row of $\mathbf{A}$.
- This means that

$$\begin{bmatrix} \partial y_1/\partial \mathbf{x} \\ \partial y_2/\partial \mathbf{x} \\ \vdots \\ \partial y_n/\partial \mathbf{x} \end{bmatrix} = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n],$$

$$\frac{\partial \mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = \mathbf{A}'.$$

OLS and ML

- Recall that the least squares objective function is

$$s(\beta) = \sum_{t=1}^{n} (y_t - x_t'\beta)^2 .$$

- Let the error term ε is normally distributed; then, the model is

$$y = X\beta_0 + \varepsilon,$$

$$\begin{aligned} \varepsilon &\sim N(0, \sigma_0^2 I_n), \text{ so} \\ f(\varepsilon) &= \prod_{t=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\varepsilon_t^2}{2\sigma^2}\right). \end{aligned}$$

OLS and ML

- The joint density for y can be constructed using a change of variables.
- We have $\varepsilon = y - X\beta$, so $\frac{\partial \varepsilon}{\partial y'} = I_n$ and $|\frac{\partial \varepsilon}{\partial y'}| = 1$, so

$$f(y) = \prod_{t=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_t - x_t'\beta)^2}{2\sigma^2}\right).$$

- Taking logs,

$$\ln L(\beta, \sigma) = -n \ln \sqrt{2\pi} - n \ln \sigma - \sum_{t=1}^{n} \frac{(y_t - x_t'\beta)^2}{2\sigma^2}.$$

- Maximizing the log-likelihood function with respect to β and σ gives the **maximum likelihood** (ML) estimator.

OLS and ML

- It turns out that ML estimators are **asymptotically efficient**.
- Note that the first order conditions for the MLE of β_0 are the same as the first order conditions that define the OLS estimator (up to a constant), so the OLS estimator of β is also the ML estimator.
- *Under the classical assumptions with normality, the OLS estimator $\hat{\beta}$ is asymptotically efficient.*
- Generally, with nonnormal errors it will be necessary to use nonlinear estimation methods to achieve asymptotically efficient estimation.

Bibliography

Angrist, J. D. and J.-S. Pischke (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Stachurski, J. (2016). *A Primer in Econometric Theory*. Online access with DDA: Askews. MIT Press.