Lecture Notes on Econometrics

Zhentao Shi

October 6, 2020

This PDF document is compiled as an overview of the contents to be covered in Econ5121A. Don't print this document! All chapters will be revised as the course progresses.

Contents

1 Probability **5**
- 1.1 Introduction . . . 5
- 1.2 Axiomatic Probability . . . 5
- 1.3 Expected Value . . . 8
- 1.4 Multivariate Random Variable . . . 10
- 1.5 Summary . . . 11

2 Regression, Projection and Causality **13**
- 2.1 Conditional Expectation . . . 13
- 2.2 Linear Projection . . . 14
- 2.3 Causality . . . 16
- 2.4 Summary . . . 20

3 Least Squares: Linear Algebra **21**
- 3.1 Estimator . . . 21
- 3.2 Subvector . . . 25
- 3.3 Goodness of Fit . . . 26
- 3.4 Summary . . . 28
- 3.5 Appendix . . . 28

4 Least Squares: Finite Sample Theory **30**
- 4.1 Maximum Likelihood . . . 30
- 4.2 Likelihood Estimation for Regression . . . 33
- 4.3 Finite Sample Distribution . . . 34
- 4.4 Mean and Variance . . . 35
- 4.5 Gauss-Markov Theorem . . . 37
- 4.6 Summary . . . 38
- 4.7 Appendix . . . 39

5 Asymptotic Theory **40**
- 5.1 Introduction . . . 40
- 5.2 Asymptotic Properties of OLS . . . 47

6 Hypothesis Testing **51**
- 6.1 Hypothesis Testing . . . 51
- 6.2 Confidence Interval . . . 54
- 6.3 Bayesian Credible Set . . . 54
- 6.4 Application in OLS . . . 55

7 Panel Data **60**
7.1 Panel Data 60

8 Endogeneity **64**
8.1 Introduction 64
8.2 Examples 66

9 Generalized Method of Moments **68**
9.1 Introduction 68
9.2 GMM in Linear Model 68

Chapter 1

Probability

For the convenience of online teaching in the fall semester of 2020, the layout is modified with wide margins and line space for note taking.

1.1 Introduction

With the advent of big data, computer scientists have come up with a plethora of new algorithms that are aimed at revealing patterns from data. *Machine learning* and *artificial intelligence* become buzz words that attract public attention. They defeated best human Go players, automated manufacturers, powered self-driving vehicles, recognized human faces, and recommended online purchases. Some of these industrial successes are based on statistical theory, and statistical theory is based on probability theory. Although this probabilistic approach is not the only perspective to understand the behavior of machine learning and artificial intelligence, it offers one of the most promising paradigms to rationalize existing algorithms and engineer new ones.

Economics has been an empirical social science since Adam Smith (1723–1790). Many numerical observations and anecdotes were scattered in his *Wealth of Nations* published in 1776. Ragnar Frisch (1895–1973) and Jan Tinbergen (1903--1994), two pioneer econometricians, were awarded in 1969 the first Nobel Prize in economics. Econometrics provides quantitative insights about economic data. It flourishes in real-world management practices, from households and firms up to governance at the global level. Today, the big data revolution is pumping fresh energy into research and exercises of econometric methods. The mathematical foundation of econometric theory is built on probability theory as well.

1.2 Axiomatic Probability

Human beings are awed by uncertainty in daily life. In the old days, Egyptians consulted oracles, Hebrews inquired prophets, and Chinese counted on diviners to interpret tortoise shell or bone cracks. Fortunetellers are abundant in today's Hong Kong.

Probability theory is a philosophy about uncertainty. Over centuries, mathematicians strove to contribute to the understanding of randomness. As measure theory matured in the early 20th century, Andrey Kolmogorov (1903-1987) built the edifice of modern probability theory in his monograph published in 1933. The formal mathematical language is a system that allows rigorous explorations which have made fruitful advancements, and is now widely accepted in academic and industrial research.

In this lecture, we will briefly introduce the axiomatic probability theory along with familiar results covered in undergraduate *probability and statistics*. This lecture note is at the level

- Hansen (2020): Introduction to Econometrics, or
- Stachurski (2016): A Primer in Econometric Theory, or
- Casella and Berger (2002): Statistical Inference (second edition)

Interested readers may want to read this textbook for more examples.

1.2.1 Probability Space

A *sample space* Ω is a collection of all possible outcomes. It is a set of things. An *event* A is a subset of Ω. It is something of interest on the sample space. A *σ-field*, denoted by $\mathcal{F}$, is a collection of events such that

1. $\emptyset \in \mathcal{F}$;
2. if an event $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$;
3. if $A_i \in \mathcal{F}$ for $i \in \mathbb{N}$, then $\bigcup_{i\in\mathbb{N}} A_i \in \mathcal{F}$.

Implications: (a) Since $\Omega = \emptyset^c \in \mathcal{F}$, we have $\Omega \in \mathcal{F}$. (b) If $A_i \in \mathcal{F}$ for $i \in \mathbb{N}$, then $A_i^c \in \mathcal{F}$ for $i \in \mathbb{N}$. Thus, if $\bigcup_{i\in\mathbb{N}} A_i^c \in \mathcal{F}$, then $\bigcap_{i\in\mathbb{N}} A_i = (\bigcup_{i\in\mathbb{N}} A_i^c)^c \in \mathcal{F}$.

Remark 1.1. Intuitively, a σ-field is a pool which is closed for countable sets to conduct union, difference, and intersection operations. These are algebraic operations of sets. σ-field is also called σ-algebra.

Example 1.1. Let $\Omega = \{1, 2, 3, 4, 5, 6\}$. Some examples of *$\sigma$-fields* include

- $\mathcal{F}_1 = \{\emptyset, \{1, 2, 3\}, \{4, 5, 6\}, \Omega\}$;
- $\mathcal{F}_2 = \{\emptyset, \{1, 3\}, \{2, 4, 5, 6\}, \Omega\}$.
- Counterexample: $\mathcal{F}_3 = \{\emptyset, \{1, 2\}, \{4, 6\}, \Omega\}$ is not a *σ-field* since $\{1, 2, 4, 6\} = \{1, 2\} \bigcup \{4, 6\}$ does not belong to $\mathcal{F}_3$.

The σ-field can be viewed as a well-organized structure built on the ground of the sample space. The pair $(\Omega, \mathcal{F})$ is called a *measure space*.

Let $\mathcal{G} = \{B_1, B_2, \ldots\}$ be an arbitrary collection of sets, not necessarily a σ-field. We say $\mathcal{F}$ is the smallest σ-field generated by $\mathcal{G}$ if $\mathcal{G} \subseteq \mathcal{F}$, and $\mathcal{F} \subseteq \tilde{\mathcal{F}}$ for any $\tilde{\mathcal{F}}$ such that $\mathcal{G} \subseteq \tilde{\mathcal{F}}$. A *Borel σ-field* $\mathcal{R}$ is the smallest σ-field generated by the open sets on the real line $\mathbb{R}$.

Example 1.2. Let $\Omega = \{1, 2, 3, 4, 5, 6\}$ and $A = \{\{1\}, \{1, 3\}\}$. Then the smallest σ-field generated by A is

$$\sigma(A) = \{\emptyset, \{1\}, \{1, 3\}, \{3\}, \{2, 4, 5, 6\}, \{2, 3, 4, 5, 6\}, \{1, 2, 4, 5, 6\}, \Omega\}.$$

A function $\mu : (\Omega, \mathcal{F}) \mapsto [0, \infty]$ is called a *measure* if it satisfies

1. (positiveness) $\mu(A) \geq 0$ for all $A \in \mathcal{F}$;

2. (countable additivity) if $A_i \in \mathcal{F}$, $i \in \mathbb{N}$, are mutually disjoint, then

$$\mu\left(\bigcup_{i\in\mathbb{N}} A_i\right) = \sum_{i\in\mathbb{N}} \mu\left(A_i\right).$$

Measure can be understand as weight or length. In particular, we call μ a *probability measure* if $\mu(\Omega) = 1$. A probability measure is often denoted as P. The triple $(\Omega, \mathcal{F}, P)$ is called a *probability space*.

So far we have answered the question: "What is a mathematically well-defined probability?", but we have not yet answered "How to assign the probability?" There are two major schools of thinking on probability assignment. One is *frequentist*, who considers probability as the average chance of occurrence if a large number of experiments are carried out. The other is *Bayesian*, who deems probability as a subjective brief. The principles of these two schools are largely incompatible, while each school has merits and difficulties, which will be elaborated when discussing hypothesis testing.

1.2.2 Random Variable

The terminology *random variable* is a historic relic which belies its modern definition of a deterministic mapping. It is a link between two measurable spaces such that any event in the σ-field installed on the range can be traced back to an event in the σ-field installed on the domain.

Formally, a function $X : \Omega \mapsto \mathbb{R}$ is $(\Omega, \mathcal{F}) \setminus (\mathbb{R}, \mathcal{R})$ *measurable* if

$$X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{F}$$

for any $B \in \mathcal{R}$. *Random variable* is an alternative, and somewhat romantic, name for a measurable function. The *σ-field generated by the random variable X* is defined as $\sigma(X) = \left\{X^{-1}(B) : B \in \mathcal{R}\right\}$.

We say a measurable is a *discrete random variable* if the set $\{X(\omega) : \omega \in \Omega\}$ is finite or countable. We say it is a *continuous random variable* if the set $\{X(\omega) : \omega \in \Omega\}$ is uncountable.

A measurable function connects two measurable spaces. No probability is involved in its definition yet. While if a probability measure P is installed on $(\Omega, \mathcal{F})$, the measurable function X will induce a probability measure on $(\mathbb{R}, \mathcal{R})$. It is easy to verify that $P_X : (\mathbb{R}, \mathcal{R}) \mapsto [0, 1]$ is also a probability measure if defined as

$$P_X(B) = P\left(X^{-1}(B)\right)$$

for any $B \in \mathcal{R}$. This P_X is called the probability measure *induced* by the measurable function X. The induced probability measure P_X is an offspring of the parent probability measure P though the channel of X.

1.2.3 Distribution Function

We go back to some terms that we have learned in a undergraduate probability course. A *(cumulative) distribution function* $F : \mathbb{R} \mapsto [0, 1]$ is defined as

$$F(x) = P(X \leq x) = P(\{X \leq x\}) = P(\{\omega \in \Omega : X(\omega) \leq x\}).$$

It is often abbreviated as CDF, and it has the following properties.

(i) $\lim_{x\to-\infty} F(x) = 0$,

(ii) $\lim_{x\to\infty} F(x) = 1$,

(iii) non-decreasing,

(iv) right-continuous $\lim_{y\to x^+} F(y) = F(x)$.

Exercise 1.1. Draw the CDF of a binary distribution; that is, $X = 1$ with probability $p \in (0,1)$ and $X = 0$ with probability $1 - p$.

For continuous distribution, if there exists a function f such that for all x,

$$F(x) = \int_{-\infty}^{x} f(y)\,\mathrm{d}y,$$

then f is called the *probability density function* of X, often abbreviated as PDF. It is easy to show that $f(x) \geq 0$ and $\int_a^b f(x)\,dx = F(b) - F(a)$.

Example 1.3. We have learned many parametric distributions like the binary distribution, the Poisson distribution, the uniform distribution, the exponential distribution, the normal distribution, χ^2, t, F distributions and so on. They are parametric distributions, meaning that the CDF or PDF can be completely characterized by very few parameters.

1.3 Expected Value

1.3.1 Integration

Integration is one of the most fundamental operations in mathematical analysis. We have studied Riemann's integral in the undergraduate calculus. Riemann's integral is intuitive, but Lebesgue integral is a more general approach to defining integration. Lebesgue integral is constructed by the following steps. X is called a *simple function* on a measurable space $(\Omega, \mathcal{F})$ if $X = \sum_i a_i \cdot 1\{A_i\}$ and this summation is finite, where $a_i \in \mathbb{R}$ and $\{A_i \in \mathcal{F}\}_{i\in\mathbb{N}}$ is a partition of Ω. A simple function is measurable.

1. Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. The integral of the simple function X with respect to μ is

$$\int X\mathrm{d}\mu = \sum_i a_i \mu(A_i).$$

   Unlike the Rieman integral, this definition of integration does not partition the domain into splines of equal length. Instead, it tracks the distinctive values of the function and the corresponding measure.

2. Let X be a non-negative measurable function. The integral of X with respect to μ is

$$\int X\mathrm{d}\mu = \sup\left\{\int Y\mathrm{d}\mu : 0 \leq Y \leq X,\ Y \text{ is simple}\right\}.$$

3. Let X be a measurable function. Define $X^+ = \max\{X, 0\}$ and $X^- = -\min\{X, 0\}$. Both X^+ and X^- are non-negative functions. The integral of X with respect to μ is

$$\int X\mathrm{d}\mu = \int X^+\mathrm{d}\mu - \int X^-\mathrm{d}\mu.$$

The Step 1 above defines the integral of a simple function. Step 2 defines the integral of a non-negative function as the approximation of steps functions from below. Step 3 defines the integral of a general function as the difference of the integral of two non-negative parts.

Remark 1.2. The integrand that highlights the difference between the Lebesgue integral and Riemann integral is the Dirichelet function on the unit interval $1\{x \in \mathbb{Q} \cap [0,1]\}$. It is not Riemann-integrable whereas its Lebesgue integral. is well defined and $\int 1\{x \in \mathbb{Q} \cap [0,1]\} dx = 0$.

If the measure μ is a probability measure P, then the integral $\int X \mathrm{d}P$ is called the *expected value,* or *expectation,* of X. We often use the notation $E[X]$, instead of $\int X \mathrm{d}P$, for convenience.

Expectation provides the average of a random variable, despite that we cannot foresee the realization of a random variable in a particular trial (otherwise the study of uncertainty is trivial). In the frequentist's view, the expectation is the average outcome if we carry out a large number of independent trials.

If we know the probability mass function of a discrete random variable, its expectation is calculated as $E[X] = \sum_x xP(X = x)$, which is the integral of a simple function. If a continuous random variable has a PDF $f(x)$, its expectation can be computed as $E[X] = \int xf(x)\,\mathrm{d}x$. These two expressions are unified as $E[X] = \int X \mathrm{d}P$ by the Lebesgue integral.

1.3.2 Properties of Expectations

Here are some properties of mathematical expectations.

- The probability of an event A is the expectation of an indicator function. $E[1\{A\}] = 1 \times P(A) + 0 \times P(A^c) = P(A)$.
- $E[X^r]$ is call the r-moment of X. The *mean* of a random variable is the first moment $\mu = E[X]$, and the second *centered* moment is called the *variance* $\mathrm{var}[X] = E\left[(X-\mu)^2\right]$. The third centered moment $E\left[(X-\mu)^3\right]$, called *skewness*, is a measurement of the symmetry of a random variable, and the fourth centered moment $E\left[(X-\mu)^4\right]$, called *kurtosis*, is a measurement of the tail thickness.
- Moments do not always exist. For example, the mean of the Cauchy distribution does not exist, and the variance of the $t(2)$ distribution does not exist.
- $E[\cdot]$ is a linear operation. If $\phi(\cdot)$ is a linear function, then $E[\phi(X)] = \phi(E[X])$.
- *Jensen's inequality* is an important fact. A function $\varphi(\cdot)$ is convex if $\varphi(ax_1 + (1-a)x_2) \leq a\varphi(x_1) + (1-a)\varphi(x_2)$ for all x_1, x_2 in the domain and $a \in [0,1]$. For instance, x^2 is a convex function. Jensen's inequality says that if $\varphi(\cdot)$ is a convex function, then $\varphi(E[X]) \leq E[\varphi(X)]$.
- *Markov inequality* is another simple but important fact. If $E[|X|^r]$ exists, then $P(|X| > \epsilon) \leq E[|X|^r]/\epsilon^r$ for all $r \geq 1$. *Chebyshev inequality* $P(|X| > \epsilon) \leq E\left[X^2\right]/\epsilon^2$ is a special case of the Markov inequality when $r = 2$.
- The distribution of a random variable is completely characterized by its CDF or PDF. A moment is a function of the distribution. To back out the underlying distribution from moments, we need to know the moment-generating function (mgf) $M_X(t) = E[e^{tX}]$ for $t \in \mathbb{R}$ whenever the expectation exists. The rth moment can be computed from mgf as

$$E[X^r] = \frac{\mathrm{d}^r M_X(t)}{\mathrm{d}t^r}\Big|_{t=0}.$$

Just like moments, mgf does not always exist.

1.4 Multivariate Random Variable

A bivariate random variable is a measurable function $X : \Omega \mapsto \mathbb{R}^2$, and more generally a multivariate random variable is a measurable function $X : \Omega \mapsto \mathbb{R}^n$. We can define the *joint CDF* as $F(x_1, \ldots, x_n) = P(X_1 \leq x_1, \ldots, X_n \leq x_n)$. Joint PDF is defined similarly.

It is sufficient to introduce the joint distribution, conditional distribution and marginal distribution in the simple bivariate case, and these definitions can be extended to multivariate distributions. Suppose a bivariate random variable (X, Y) has a joint density $f(\cdot, \cdot)$. The *conditional density* can be roughly written as $f(y|x) = f(x, y)/f(x)$ if we do not formally deal with the case $f(x) = 0$. The *marginal density* $f(y) = \int f(x, y)\, dx$ integrates out the coordinate that is not interested.

1.4.1 Conditional Probability and Bayes' Theorem

In a probability space $(\Omega, \mathcal{F}, P)$, for two events $A_1, A_2 \in \mathcal{F}$ the *conditional probability* is

$$P(A_1|A_2) = \frac{P(A_1 A_2)}{P(A_2)}$$

if $P(A_2) > 0$. In the definition of conditional probability, A_2 plays the role of the outcome space so that $P(A_1A_2)$ is standardized by the total mass $P(A_2)$. If $P(A_2) = 0$, the conditional probability can still be valid in some cases, but we need to introduce the *dominance* between two measures, which we do not elaborate here.

Since A_1 and A_2 are symmetric, we also have $P(A_1A_2) = P(A_2|A_1)P(A_1)$. It implies

$$P(A_1|A_2) = \frac{P(A_2|A_1)\, P(A_1)}{P(A_2)}$$

This formula is the *Bayes' Theorem*.

1.4.2 Independence

We say two events A_1 and A_2 are *independent* if $P(A_1A_2) = P(A_1)P(A_2)$. If $P(A_2) \neq 0$, it is equivalent to $P(A_1|A_2) = P(A_1)$. In words, knowing A_2 does not change the probability of A_1.

Regarding the independence of two random variables, X and Y are *independent* if $P(X \in B_1, Y \in B_2) = P(X \in B_1)\, P(Y \in B_2)$ for any two Borel sets B_1 and B_2.

If X and Y are independent, then $E[XY] = E[X]E[Y]$. The expectation of their product is the product of their expectations.

1.4.3 Law of Iterated Expectations

Given a probability space $(\Omega, \mathcal{F}, P)$, a *sub σ-algebra* $\mathcal{G} \subseteq \mathcal{F}$ and a $\mathcal{F}$-measurable function Y with $E|Y| < \infty$, the *conditional expectation* $E[Y|\mathcal{G}]$ is defined as a $\mathcal{G}$-measurable function such that

$$\int_A Y \mathrm{d}P = \int_A E[Y|\mathcal{G}]\, \mathrm{d}P$$

for all $A \in \mathcal{G}$. Here $\mathcal{G}$ is a coarse σ-field and $\mathcal{F}$ is a finer σ-field.

Taking $A = \Omega$, we have $E[Y] = \int Y \mathrm{d}P = \int E[Y|\mathcal{G}] \mathrm{d}P = E[E[Y|\mathcal{G}]]$. The *law of iterated expectation*

$$E[Y] = E[E[Y|\mathcal{G}]]$$

is a trivial fact which follows this definition of the conditional expectation. In the bivariate case, if the conditional density exists, the conditional expectation can be computed as $E[Y|X] = \int y f(y|X) \mathrm{d}y$, where the conditioning variable $E[\cdot|X] = E[\cdot|\sigma(X)]$ is a concise notation for the smallest σ-field generated by X. The law of iterated expectation implies $E[E[Y|X]] = E[Y]$.

Below are some 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. $E[h(X) Y|X] = h(X) E[Y|X]$.

1.5 Summary

If it is your first encounter of measure theory, the new definitions here may seem overwhelmingly abstract. A natural question is that: "I earned high grade in my undergraduate probability and statistics; do I really need the fancy mathematics in this lecture to do well in econometrics?" The answer is yes and no. *No* is in the sense that if you want to use econometric methods, instead of grasp the underlying theory, then the axiomatic probability does not add much to your weaponry. You can be an excellent economist or applied econometrician without knowing measure theoretic probability. *Yes* is in the sense that without measure theory, we cannot even formally define conditional expectation, which will be the subject of our next lecture and is a core concept of econometrics. Moreover, the pillars of asymptotic theory — law of large numbers and central limit theorem — can only be made accurate with this foundation. If you are aspired to work on econometric theory, you will meet and use measure theory so often in your future study and finally it becomes part of your muscle memory.

In this course, we try to keep a balance manner. On the one hand, many econometrics topics can be presented with elementary mathematics. Whenever possible, econometrics should reach wider audience with a plain appearance, instead of intimidating people by arcane languages. On the other hand, we introduce these concepts in this lecture and will invoke them in the discussion of asymptotic theory later. Your investment in advanced mathematics will not be wasted in vain.

Historical notes: Measure theory was established in the early 20th century by a constellation of French/German mathematicians, represented by Émile Borel, Henri Lebesgue, Johann Radon, etc. Generations of Russian mathematicians such as Andrey Markov and Andrey Kolmogorov made fundamental contributions in mathematizing seemingly abstract concepts of uncertainty and randomness. Their names are immortalized by the Borel set, the Lebesgue integral, the Radon measure, Markov chain, Kolmogorov's zero–one law and many other terminologies named after them.

Fascinating questions about probability attracted great economists. Francis Edgeworth (1845–1926) wrote extensively on probability and statistics. John Maynard Keynes (1883–1946) published *A Treatise on Probability* in 1921 which mixed probability and philospophy, although this piece of work was not as influential as his *General Theory of Employment, Interest and Money* in 1936 which later revolutionized economics.

Today, the technology of collecting data and the processing data is unbelievably cheaper than that 100 years ago. Unfortunately, the cost of learning mathematics and developing mathematics

has not been significantly lowered over one century. Only a small handful of talents, like you, enjoy the privilege and luxury to appreciate the ideas of these great minds.

Further reading: Doob (1996) summarized the development of axiomatic probability in the first half of the 20th century.

Zhentao Shi. Sep 12, 2020.

Doob, J. L. (1996). The development of rigor in mathematical probability (1900–1950). *The American mathematical monthly 103*(7), 586–595.

Chapter 2

Regression, Projection and Causality

Notation: In this note, y is a scale random variable, and $x = (x_1, \ldots, x_K)'$ is a $K \times 1$ random vector. Throughout this course, a vector is a *column* vector, i.e. a one-column matrix.

2.1 Conditional Expectation

Machine learning is a big basket that contains the regression models. We motivate the conditional expectation model from the perspective of prediction. We view a regression as *supervised learning*. Supervised learning uses a function of x, say, $g(x)$, to predict y. x cannot perfectly predict y; otherwise their relationship is deterministic. The prediction error $y - g(x)$ depends on the choice of g. There are numerous possible choices of g. Which one is the best? Notice that this question is not concerned about the underlying data generating process (DGP) of the joint distribution of (y, x). We want to find a general rule to achieve accurate prediction of y given x, no matter how this pair of variables is generated.

To answer this question, we need to decide a criterion to compare different g. Such a criterion is called the *loss function* $L(y, g(x))$. A particularly convenient one is the *quadratic loss*, defined as

$$L(y, g(x)) = (y - g(x))^2.$$

Since the data are random, $L(y, g(x))$ is also random. "Random" means uncertainty: sometimes *this* happens, and sometimes *that* happens. To get rid of the uncertainty, we average the loss function with respect to the joint distribution of (y, x) as $R(y, g(x)) = E[L(y, g(x))]$, which is called *risk*. Risk is a deterministic quality. For the quadratic loss function, the corresponding risk is

$$R(y, g(x)) = E\left[(y - g(x))^2\right],$$

is called the *mean squared error* (MSE). MSE is the most widely used risk measure, although there exist many alternative measures, for example the *mean absolute error* (MAE) $E[|y - g(x)|]$. The popularity of MSE comes from its convenience for analysis in closed-form, which MAE does not enjoy due to its nondifferentiability. This is similar to the choice of utility functions in economics. There are only a few functional forms for the utility, for example CRRA, CARA, and so on. They are popular because they lead to close-form solutions that are easy to handle. Now our quest is narrowed to: What is the optimal choice of g if we minimize the MSE?

Proposition 2.1. *The conditional mean function (CEF)* $m(x) = E[y|x] = \int y f(y|x)\, \mathrm{d}y$ *minimizes MSE.*

Before we prove the above proposition, we first discuss some properties of the conditional mean function. Obviously

$$y = m(x) + (y - m(x)) = m(x) + \epsilon,$$

where $\epsilon := y - m(x)$ is called the *regression error*. This equation holds for (y, x) following any joint distribution, as long as $E[y|x]$ exists. The error term ϵ satisfies these properties:

- $E[\epsilon|x] = E[y - m(x)|x] = E[y|x] - m(x) = 0$,
- $E[\epsilon] = E[E[\epsilon|x]] = E[0] = 0$,
- For any function $h(x)$, we have

$$E[h(x)\epsilon] = E[E[h(x)\epsilon|x]] = E[h(x)E[\epsilon|x]] = 0. \tag{2.1}$$

The last property implies that ϵ is uncorrelated with any function of x. In particular, when h is the identity function $h(x) = x$, we have $E[x\epsilon] = \operatorname{cov}(x, \epsilon) = 0$.

Proof of Proposition 2.1. The optimality of the CEF can be confirmed by "guess-and-verify." For an arbitrary $g(x)$, the MSE can be decomposed into three terms

$$\begin{aligned} & E\left[(y - g(x))^2\right] \\ = \; & E\left[(y - m(x) + m(x) - g(x))^2\right] \\ = \; & E\left[(y - m(x))^2\right] + 2E[(y - m(x))(m(x) - g(x))] + E\left[(m(x) - g(x))^2\right]. \end{aligned}$$

The first term is irrelevant to $g(x)$. The second term

$$2E[(y - m(x))(m(x) - g(x))] = 2E[\epsilon(m(x) - g(x))] = 0$$

by invoking (2.1) with $h(x) = m(x) - g(x)$. The second term is again irrelevant of $g(x)$. The third term, obviously, is minimized at $g(x) = m(x)$. □

Our perspective so far deviates from many econometric textbooks that assume that the dependent variable y is generated as $g(x) + \epsilon$ for some unknown function $g(\cdot)$ and error term ϵ such that $E[\epsilon|x] = 0$. Instead, we take a predictive approach regardless the DGP. What we observe are y and x and we are solely interested in seeking a function $g(x)$ to predict y as accurately as possible under the MSE criterion.

2.2 Linear Projection

The CEF $m(x)$ is the function that minimizes the MSE. However, $m(x) = E[y|x]$ is a complex function of x, for it depends on the joint distribution of (y, x), which is mostly unknown in practice. Now let us make the prediction task even simpler. How about we minimize the MSE within all linear functions in the form of $h(x) = h(x; b) = x'b$ for $b \in \mathbb{R}^K$? The minimization problem is

$$\min_{b \in \mathbb{R}^K} E\left[\left(y - x'b\right)^2\right]. \tag{2.2}$$

Take the first-order condition of the MSE

$$\frac{\partial}{\partial b}E\left[\left(y-x'b\right)^2\right]=E\left[\frac{\partial}{\partial b}\left(y-x'b\right)^2\right]=-2E\left[x\left(y-x'b\right)\right],$$

where the first equality holds if $E\left[\left(y-x'b\right)^2\right]<\infty$ so that the expectation and partial differentiation is interchangeable, and the second equality hods by the chain rule and the linearity of expectation. Set the first order condition to 0 and we solve

$$\beta=\arg\min_{b\in\mathbb{R}^K}E\left[\left(y-x'b\right)^2\right]$$

in the closed-form

$$\beta=\left(E\left[xx'\right]\right)^{-1}E\left[xy\right]$$

if $E\left[xx'\right]$ is invertible. Notice here that b is an arbitrary K-vector, while β is the optimizer. The function $x'\beta$ is called the *best linear projection* (BLP) of y on x, and the vector β is called the *linear projection coefficient*.

Remark 2.1. The linear function is not as restrictive as one might thought. It can be used to produce some nonlinear (in random variables) effect if we re-define x. For example, if

$$y=x_1\beta_1+x_2\beta_2+x_1^2\beta_3+e,$$

then $\frac{\partial}{\partial x_1}m\left(x_1,x_2\right)=\beta_1+2x_1\beta_3$, which is nonlinear in x_1, while it is still linear in the parameter $\beta=(\beta_1,\beta_2,\beta_3)$ if we define a set of new regressors as $(\tilde{x}_1,\tilde{x}_2,\tilde{x}_3)=\left(x_1,x_2,x_1^2\right)$.

Remark 2.2. If (y,x) is jointly normal in the form

$$\begin{pmatrix}y\\x\end{pmatrix}\sim\mathrm{N}\left(\begin{pmatrix}\mu_y\\\mu_x\end{pmatrix},\begin{pmatrix}\sigma_y^2&\rho\sigma_y\sigma_x\\\rho\sigma_y\sigma_x&\sigma_x^2\end{pmatrix}\right)$$

where ρ is the correlation coefficient, then

$$E\left[y|x\right]=\mu_y+\rho\frac{\sigma_y}{\sigma_x}\left(x-\mu_x\right)=\left(\mu_y-\rho\frac{\sigma_y}{\sigma_x}\mu_x\right)+\rho\frac{\sigma_y}{\sigma_x}x,$$

is a liner function of x. In this example, the CEF is linear.

Remark 2.3. Even though in general $m\left(x\right)\neq x'\beta$, the linear form $x'\beta$ is still useful in approximating $m\left(x\right)$. That is, $\beta=\arg\min_{b\in\mathbb{R}^K}E\left[\left(m(x)-x'b\right)^2\right]$.

Proof. The first-order condition gives $\frac{\partial}{\partial b}E\left[\left(m(x)-x'b\right)^2\right]=-2E[x(m(x)-x'b)]=0$. Rearrange the terms and obtain $E[x\cdot m(x)]=E[xx']b$. When $E[xx']$ is invertible, we solve

$$\left(E\left[xx'\right]\right)^{-1}E[x\cdot m(x)]=\left(E\left[xx'\right]\right)^{-1}E[E[xy|x]]=\left(E\left[xx'\right]\right)^{-1}E[xy]=\beta.$$

Thus β is also the best linear approximation to $m\left(x\right)$ under MSE. □

We may rewrite the linear regression model, or the *linear projection model,* as

$$\begin{aligned}y&=x'\beta+e\\E[xe]&=0,\end{aligned}$$

where $e=y-x'\beta$ is called the *linear projection error*, to be distinguished from $\epsilon=y-m(x)$.

Exercise 2.1. Show (a) $E\left[xe\right]=0$. (b) If x contains a constant, then $E\left[e\right]=0$.

2.2.1 Omitted Variable Bias

We write the *long regression* as

$$y = x_1'\beta_1 + x_2'\beta_2 + \beta_3 + e_\beta,$$

and the *short regression* as

$$y = x_1'\gamma_1 + \gamma_2 + e_\gamma,$$

where e_β and e_γ are the projection errors, respectively. If β_1 in the long regression is the parameter of interest, omitting x_2 as in the short regression will render *omitted variable bias* (meaning $\gamma_1 \neq \beta_1$) unless x_1 and x_2 are uncorrelated.

We first demean all the variables in the two regressions, which is equivalent as if we project out the effect of the constant. The long regression becomes

$$\tilde{y} = \tilde{x}_1'\beta_1 + \tilde{x}_2'\beta_2 + \tilde{e}_\beta,$$

and the short regression becomes

$$\tilde{y} = \tilde{x}_1'\gamma_1 + \tilde{e}_\gamma,$$

where *tilde* denotes the demeaned variable.

Exercise. Show $\tilde{e}_\beta = e_\beta$ and $\tilde{e}_\gamma = e_\gamma$.

After demeaning, the cross-moment equals to the covariance. The short regression coefficient

$$\begin{aligned}
\gamma_1 &= \left(E\left[\tilde{x}_1\tilde{x}_1'\right]\right)^{-1} E\left[\tilde{x}_1\tilde{y}\right] \\
&= \left(E\left[\tilde{x}_1\tilde{x}_1'\right]\right)^{-1} E\left[\tilde{x}_1\left(\tilde{x}_1'\beta_1 + \tilde{x}_2'\beta_2 + \tilde{e}_\beta\right)\right] \\
&= \left(E\left[\tilde{x}_1\tilde{x}_1'\right]\right)^{-1} E\left[\tilde{x}_1\tilde{x}_1'\right]\beta_1 + \left(E\left[\tilde{x}_1\tilde{x}_1'\right]\right)^{-1} E\left[\tilde{x}_1\tilde{x}_2'\right]\beta_2 \\
&= \beta_1 + \left(E\left[\tilde{x}_1\tilde{x}_1'\right]\right)^{-1} E\left[\tilde{x}_1\tilde{x}_2'\right]\beta_2,
\end{aligned}$$

where the third line holds as $E\left[\tilde{x}_1\tilde{e}_\beta\right] = 0$. Therefore, $\gamma_1 = \beta_1$ if and only if $E\left[\tilde{x}_1\tilde{x}_2'\right]\beta_2 = 0$, which demands either $E\left[\tilde{x}_1\tilde{x}_2'\right] = 0$ or $\beta_2 = 0$.

Exercise 2.2. Show that $E\left[\left(y - x_1'\beta_1 - x_2'\beta_2 - \beta_3\right)^2\right] \leq E\left[\left(y - x_1'\gamma_1 - \gamma_2\right)^2\right]$.

Obviously we prefer to run the long regression to attain β_1 if possible, for it is a more general model than the short regression and achieves no larger variance in the projection error. However, sometimes x_2 is unobservable so the long regression is unavailable. This example of omitted variable bias is ubiquitous in applied econometrics. Ideally we would like to directly observe some regressors but in reality we do not have them at hand. We should be aware of the potential consequence when the data are not as ideal as we have wished. When only the short regression is available, in some cases we are able to sign the bias, meaning that we can argue whether γ_1 is bigger or smaller than β_1 based on our knowledge.

2.3 Causality

2.3.1 Structure and Identification

Unlike physical laws such as Einstein's mass–energy equivalence $E = mc^2$ and Newton's universal gravitation $F = Gm_1m_2/r^2$, economic phenomena can rarely be summarized in such a minimalistic

style. When using experiments to verify physical laws, scientists often manage to come up with smart design in which signal-to-noise ratio is so high that small disturbances are kept at a negligible level. On the contrary, economic laws do not fit a laboratory for experimentation. What is worse, the subjects in economic studies — human beings — are heterogeneous and with many features that are hard to control. People from distinctive cultural and family backgrounds respond to the same issue differently and researchers can do little to homogenize them. The signal-to-noise ratios in economic laws are often significantly lower than those of physical laws, mainly due to the lack of laboratory setting and the heterogeneous nature of the subjects.

Educational return and the demand-supply system are two classical topics in econometrics. A person's incomes is determined by too many random factors in the academic and career path that is impossible to exhaustively observe and control. The observable prices and quantities are outcomes of equilibrium so the demand and supply affect each other.

Generations of thinkers have been debating the definitions of causality. In economics, an accepted definition is *structural causality*. Structural causality is a thought experiment. It assumes that there is a DGP that produces the observational data. If we can use data to recover the DGP or some features of the DGP, then we have learned causality or some implications of causality.

A key issue to resolve before looking at the realized sample is *identification*. We say a model or DGP is *identified* if the each possible parameter of the model under consideration generates distinctive features of the observable data. A model is *under-identified* if more than one parameter in the model can generate exact the same features of the observable data. In other words, a model is under-identified if from the observable data we cannot trace back to a unique parameter in the model. A correctly specified model is the prerequisite for any discussion of identification. In reality, all models are wrong. Thus when talking about identification, we are indulged in an imaginary world. If in such a thought experiment we still cannot unique distinguish the true parameter of the data generating process, then identification fails. We cannot determine what is the true model no matter how large the sample is.

2.3.2 Treatment Effect

We narrow down to the framework of the relationship between y and x. One question of particular interest is *treatment effect*. The treatment effect is how much y will change if we change a variable of interest, say d, by one unit while keeping all other variables (including the unobservable variables) the same. The Latin phrase *ceteris paribus* means "keep all other things constant."

Example 2.1. During the 2020 covid-19 pandemic, Hong Kong's unemployment rate rose to a high-level and consumption collapsed. In order to boost the economy, some Hong Kong residents were qualified in receiving 10,000 HKD cash allowance from the government. We are interested in learning how much does the 10,000 HKD allowance increase people's consumption. For an individual, we imagine two parallel worlds: one with the cash allowance and one without. The difference of the consumption in the world with the allowance, denoted $Y(1)$, and that in the world without the allowance, denoted $Y(0)$, is the treatment effect of that particular person. This thought experiment is called the *potential outcome framework*.

However, in reality one and only one scenario happens, which echos the saying of ancient Greek philosopher Heraclitus (553 BC--475 BC) "You cannot step into the same river twice." The individual treatment effect is not operational (*operational* means it can be computed from data at the population level), because one and only one outcome is realized. With many people available, we can define *average treatment effect* (ATE) as

$$ATE = E[Y(1) - Y(0)] = E[Y(1)] - E[Y(0)].$$

Notice that $E[Y(1)]$ and $E[Y(0)]$ are still not operational before we observe a companion variable

$$D = 1\{\text{treatment received}\}.$$

Once each individual's treatment status is observable, $E[Y(1)|D=1]$ and $E[Y(0)|D=0]$ are operational from the data.

If the two potential outcomes $(Y(1), Y(0))$ are independent of the assignment D, then $E[Y(1)] = E[Y(1)|D=1]$ and $E[Y(0)] = E[Y(0)|D=0]$ so that ATE can be estimated from the data in an operational way as

$$ATE = E[Y(1)|D=1] - E[Y(0)|D=0].$$

Therefore, to evaluate ATE ideally we would like use a lottery to randomly decide that some people receive the treatment (*treatment group*, with $D=1$) and the others do not (*control group*, with $D=0$).

When we have other control variables, we can also define a finer treatment effect conditional on x:

$$ATE(x) = E[Y(1)|x] - E[Y(0)|x].$$

ATE is the average effect in the population of individuals when we hypothetical give them the treatment, keeping all other factors x constant. If conditioning on x, the treatment D is independent of $(Y(1), Y(0))$, then ATE becomes operational:

$$ATE(x) = E[Y(1)|D=1,x] - E[Y(0)|D=0,x]$$

The important condition $((Y(1), Y(0)) \perp D)|x$ is called the *conditional independence assumption* (CIA).

Example 2.2. CIA is more plausible than full independence. Consider the example $Y(1) = x + u(1)$, $Y(0) = x + u(0)$ and $D = 1\{x + u_d \geq 0\}$. If $((u(0), u(1)) \perp u_d)|x$, then CIA is satisfied. Nevertheless $(Y(1), Y(0))$ and D are statistically dependent, since x is involved in all random variables.

2.3.3 ATE and CEF

In the previous section the treatment D is binary. Now we consider a continuous treatment D. Suppose the DGP, or the structural model, is $Y = h(D, x, u)$ where D and x are observable and u is unobservable. It is natural to define ATE with the continuous treatment (Hansen's book Chapter 2.30 calls it *average causal effect*) as

$$ATE(d,x) = E\left[\lim_{\Delta\to 0}\frac{h(d+\Delta, x, u) - h(d, x, u)}{\Delta}\right] = E\left[\frac{\partial}{\partial d}h(d, x, u)\right],$$

where the continuous differentiability of $h(d, x, u)$ at d is implicitly assumed. Unlike the binary treatment case, here d explicitly shows up in $ATE(d, x)$ because the effect can vary at different values of d. ATE here is the average effect in the population of individuals if we hypothetical move D a tiny bit around d, keeping all other factors x constant.

In the previous sections, we focused on the CEF $m(d, x)$, where d is added to x as an additional variable of interest. We did not intend to model the underlying economic mechanism $h(D, x, u)$, which may be very complex. Can we learn the $ATE(d, x)$ which bears the structural causal

interpretation, from the mechanical $m(d,x)$ which merely cares about best prediction? The answer is positive under CIA: $(u \perp D)|x$.

$$\begin{aligned}\frac{\partial}{\partial d}m(d,x) &= \frac{\partial}{\partial d}E[y|d,x] = \frac{\partial}{\partial d}E[h(d,x,u)|d,x] = \frac{\partial}{\partial d}\int h(d,x,u)\,f(u|d,x)\,du \\ &= \int \frac{\partial}{\partial d}[h(d,x,u)\,f(u|d,x)]\,du \\ &= \int \left[\frac{\partial}{\partial d}h(d,x,u)\right] f(u|d,x)\,du + \int h(d,x,u)\left[\frac{\partial}{\partial d}f(u|d,x)\right] du,\end{aligned}$$

where the second line implicitly assumes interchangeability between the integral and the partial derivative. Under CIA, $\frac{\partial}{\partial d}f(u|d,x) = 0$ and the second term drops out. Thus

$$\frac{\partial}{\partial d}m(d,x) = \int \left[\frac{\partial}{\partial d}h(d,x,u)\right] f(u|d,x)\,du = E\left[\frac{\partial}{\partial d}h(d,x,u)\right] = ATE(d,x).$$

This is an important result. It says that if CIA holds, we can learn the causal effect of d on y by the partial derivative of CEF conditional on x. In particular, if we further assume a linear CEF $m(d,x) = \beta_d d + \beta_x' x$, then the causal effect is the coefficient β_d.

CIA is the key condition that links the CEF and the causal effect. CIA is not an innocuous assumption. In applications, our causal results are credible only when we can convincing defend CIA.

Exercise 2.3. Let factories' output be a Cobb-Douglas function $Y = AK^\alpha L^\beta$, where the capital level K and labor L as well as the output Y is observable, while the "technology" A is unobservable. Take logarithm on both sides of the equation:

$$y = u + \alpha k + \beta l \tag{2.3}$$

where $y = \log Y$, $u = \log A$, $k = \log K$ and $l = \log L$. Suppose $\begin{pmatrix} u \\ k \\ l \end{pmatrix} \sim N\left(\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 & 0.5 & 0 \\ 0.5 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\right)$ and $\alpha = \beta = 1/2$ make the true DGP. Here u and k are correlated, because factories of larger scale can afford robots to facilitate automation.

1. What is the partial derivative of CEF when we use k as a treatment variable for a fixed labor level l? (Hint: the CEF is a linear function thanks to the joint normality.)

2. Does it coincide with $\alpha = 1/2$, the coefficient in the causal model (2.3)? (Hint: No, because CIA is violated.)

Sometimes applied researchers assume by brute force that $y = m(d,x) + u$ is the DGP and $E[u|d,x] = 0$, where d is the variable of interest and x is the vector of other control variables. Under these assumptions,

$$ATE(d,x) = E\left[\frac{\partial}{\partial d}(m(d,x) + u)\,|d,x\right] = \frac{\partial m(d,x)}{\partial d} + \frac{\partial}{\partial d}E[u|d,x] = \frac{\partial m(d,x)}{\partial d},$$

where the second equality holds if $\frac{\partial}{\partial d}E[u|d,x] = E\left[\frac{\partial}{\partial d}u|d,x\right]$. At a first glance, it seems that the mean independence assumption $E[u|d,x] = 0$, which is weaker than CIA, implies the equivalence between $ATE(d,x)$ and $\partial m(d,x)/\partial d$ here. However, such slight weakening is achieved by a very

strong assumption that the DGP $h(d, x, u)$ follows the additive separable form $m(d, x)+u$. Without economic theory to defend the choice of the assumed DGP $y = m(d, x) + u$, this is at best the *reduced-form* approach.

The *structural approach* here models the economic mechanism, guided by economic theory. The *reduced-form approach* is convenient and can document stylized facts when suitable economic theory is not immediately available. There are constant debates about the pros and cons of the two approaches; see *Journal of Economic Perspectives* Vol. 24, No. 2 Spring 2010. In macroeconomics, the so-called Phillips curve, attributed to A.W. Phillips about the negative correlation between inflation and unemployment, is a stylized fact learned from the reduced-form approach. The Lucas critique (Lucas, 1976) exposed its lack of microfoundation and advocated modeling deep parameters that are invariant to policy changes. The latter is a structural approach. Ironically, more than 40 years has passed since the Lucas critique, equations with little microfoundation still dominate the analytical apparatus of central bankers.

2.4 Summary

In this lecture, we cover the conditional mean function and causality. When we are faced with a pair of random variable (y, x) drawn from some joint distribution, the CEF is the best predictor. When we go further into the structural causality about some treatment d to the dependent variable y, under CIA we can find equivalence between ATE and the partial derivative of CEF. All analyses are conducted in population. We have not touched the sample yet.

Historical notes: Regressions and conditional expectations are concepts from statistics and they are imported to econometrics in early time. Researchers at the Cowles Commission (now Cowles Foundation for Research in Economics) — Jacob Marschak (1898–1977), Tjalling Koopmans (1910–1985, Nobel Prize 1975), Trygve Haavelmo (1911–1999, Nobel Prize 1989) and their colleagues — were trailblazers of the econometric structural approach.

The potential outcome framework is not peculiar to economics. It is widely used in other fields such as biostatistics and medical studies. It was initiated by Jerzy Neyman (1894–1981) and extended by Donald B. Rubin (1943–), Professor of Statistics at Tsinghua University.

Further reading: Lewbel (2019) offers a comprehensive summary of identification in econometrics. Accounting is an applied field with many claimed causal inference drawn from simple regressions; it is encouraging to hear Gow et al. (2016) to reflect causality in their practices.

`Zhentao Shi. Sep 17, 2020`

Gow, I. D., D. F. Larcker, and P. C. Reiss (2016). Causal inference in accounting research. *Journal of Accounting Research 54*(2), 477–523.

Lewbel, A. (2019). The identification zoo: Meanings of identification in econometrics. *Journal of Economic Literature 57*(4), 835–903.

Lucas, R. E. (1976). Econometric policy evaluation: A critique. In *Carnegie-Rochester conference series on public policy*, Volume 1, pp. 19–46.

Chapter 3

Least Squares: Linear Algebra

Notation: y_i is a scalar, and $x_i = (x_{i1}, \ldots, x_{iK})'$ is a $K \times 1$ vector. $Y = (y_1, \ldots, y_n)'$ is an $n \times 1$ vector, and

$$X = \begin{bmatrix} x_1' \\ x_2' \\ \vdots \\ x_n' \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1K} \\ x_{21} & x_{22} & \cdots & x_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{22} & \cdots & x_{nK} \end{bmatrix}$$

is an $n \times K$ matrix. I_n is an $n \times n$ identity matrix.

Ordinary least squares (OLS) is the most basic estimation technique in econometrics. It is simple and transparent. Understanding it thoroughly paves the way to study more sophisticated linear estimators. Moreover, many nonlinear estimators resemble the behavior of linear estimators in a neighborhood of the true value. In this lecture, we learn a series of facts from the linear algebra operation.

To manipulate Leopold Kronecker's famous saying "God made the integers; all else is the works of man", I would say "Gauss made OLS; all else is the works of applied researchers." Popularity of OLS goes far beyond our dismal science. But be aware that OLS is a pure statistical or supervised machine learning method which reveals correlation instead of causality. Rather, economic theory hypothesizes causality while data are collected to test the theory or quantify the effect.

3.1 Estimator

As we have learned from the linear project model, the projection coefficient β in the regression

$$y = x'\beta + e$$

can be written as

$$\beta = \left(E\left[xx'\right]\right)^{-1} E\left[xy\right]. \tag{3.1}$$

We draw a pair of (y, x) from the joint distribution, and we mark it as (y_i, x_i) for $i = 1, \ldots, n$ repeated experiments. We possess a *sample* $(y_i, x_i)_{i=1}^n$.

Remark 3.1. Is (y_i, x_i) random or deterministic? Before we make the observation, they are treated as random variables whose realized values are uncertain. (y_i, x_i) is treated as random when we talk about statistical properties — statistical properties of a fixed number is meaningless. After we make the observation, they become deterministic values which cannot vary anymore.

Remark 3.2. In reality, we have at hand fixed numbers (more recently, words, photos, audio clips, video clips, etc., which can all be represented in digital formats with 0 and 1) to feed into a computational operation, and the operation will return one or some numbers. All statistical interpretation about these numbers are drawn from the probabilistic thought experiments. A *thought experiment* is an academic jargon for a *story* in plain language. Under the axiomatic approach of probability theory, such stories are mathematical consistent and coherent. But mathematics is a tautological system, not science. The scientific value of a probability model depends on how close it is to the *truth* or implications of the truth. In this course, we suppose that the data are generated from some mechanism, which is taken as the truth. In the linear regression model for example, the joint distribution of (y, x) is the truth, while we are interested in the linear projection coefficient β, which is an implication of the truth as in (3.1).

The sample mean is a natural estimator of the population mean. Replace the population mean $E[\cdot]$ in (3.1) by the sample mean $\frac{1}{n}\sum_{i=1}^{n}\cdot$, and the resulting estimator is

$$\begin{aligned}\widehat{\beta} &= \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} x_i y_i \\ &= \left(\frac{X'X}{n}\right)^{-1} \frac{X'y}{n} = \left(X'X\right)^{-1} X'y\end{aligned}$$

if $X'X$ is invertible. This is one way to motivate the OLS estimator.

Alternatively, we can derive the OLS estimator from minimizing the sum of squared residuals $\sum_{i=1}^{n}\left(y_i - x_i'b\right)^2$, or equivalently

$$Q(b) = \frac{1}{2n}\sum_{i=1}^{n}\left(y_i - x_i'b\right)^2 = \frac{1}{2n}(Y - Xb)'(Y - Xb) = \frac{1}{2n}\|Y - Xb\|^2,$$

where the factor $\frac{1}{2n}$ is nonrandom and does not change the minimizer, and $\|\cdot\|$ is the Euclidean norm of a vector. Solve the first-order condition

$$\frac{\partial}{\partial b}Q(b) = \begin{bmatrix} \partial Q(b)/\partial b_1 \\ \partial Q(b)/\partial b_2 \\ \vdots \\ \partial Q(b)/\partial b_K \end{bmatrix} = -\frac{1}{n}X'(Y - Xb) = 0.$$

This necessary condition for optimality gives exactly the same $\widehat{\beta} = (X'X)^{-1}X'y$. Moreover, the second-order condition

$$\frac{\partial^2}{\partial b \partial b'}Q(b) = \begin{bmatrix} \frac{\partial^2}{\partial b_1^2}Q(b) & \frac{\partial^2}{\partial b_2 \partial b_2}Q(b) & \cdots & \frac{\partial^2}{\partial b_K \partial b_1}Q(b) \\ \frac{\partial^2}{\partial b_1 \partial b_2}Q(b) & \frac{\partial^2}{\partial b_2^2}Q(b) & \cdots & \frac{\partial^2}{\partial b_K \partial b_2}Q(b) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2}{\partial b_1 \partial b_K}Q(b) & \frac{\partial^2}{\partial b_2 \partial b_K}Q(b) & \cdots & \frac{\partial^2}{\partial b_K^2}Q(b) \end{bmatrix} = \frac{1}{n}X'X$$

shows that $Q(b)$ is convex in b due to the positive semi-definite matrix $X'X/n$. (The function $Q(b)$ is strictly convex in b if $X'X/n$ is positive definite.)

Remark 3.3. In the derivation of OLS we presume that the K columns in X are *linearly independent*, which means there is no $K \times 1$ vector b such that $b \neq 0_K$ and $Xb = 0_n$. Linear independence

of the columns implies $n \geq K$ and the invertibility of $X'X/n$. Linear independence is violated when some regressors are *perfectly collinear*, for example when we use dummy variables to indicate categorical variables and put all these categories into the regression. Modern econometrics software automatically detects and reports perfect collinearity. What is treacherous is *nearly collinear*, meaning that the minimal eigenvalue of $X'X/n$ is close to 0, though not exactly equal to 0. We will talk about the consequence of near collinearity in the chapter of asymptotic theory.

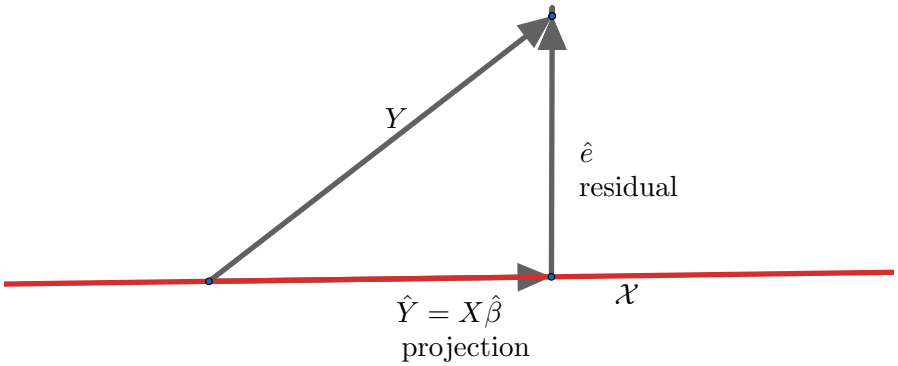
Here are some definitions and properties of the OLS estimator.

- Fitted value: $\widehat{Y} = X\widehat{\beta}$.
- Projection matrix: $P_X = X\left(X'X\right)^{-1}X$; Residual maker matrix: $M_X = I_n - P_X$.
- $P_X X = X$; $X'P_X = X'$.
- $M_X X = 0_{n\times K}$; $X'M_X = 0_{K\times n}$.
- $P_X M_X = M_X P_X = 0_{n\times n}$.
- If $AA = A$, we call it an *idempotent* matrix. Both P_X and M_X are idempotent. All eigenvalues of an idempotent matrix must be either 1 or 0.
- $\text{rank}\left(P_X\right) = K$, and $\text{rank}\left(M_X\right) = n - K$ (See the Appendix of this chapter).
- Residual: $\widehat{e} = Y - \widehat{Y} = Y - X\widehat{\beta} = Y - X(X'X)^{-1}X'Y = (I_n - P_X)Y = M_X Y = M_X\left(X\beta + e\right) = M_X e$. Notice $\widehat{e}$ and e are two different objects.
- $X'\widehat{e} = X'M_X e = 0_K$.
- $\sum_{i=1}^n \widehat{e}_i = 0$ if x_i contains a constant.
  (Because $X'\widehat{e} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ \heartsuit & \heartsuit & \cdots & \heartsuit \\ \cdots & \cdots & \ddots & \vdots \\ \heartsuit & \heartsuit & \cdots & \heartsuit \end{bmatrix}\begin{bmatrix} \widehat{e}_1 \\ \widehat{e}_2 \\ \vdots \\ \widehat{e}_n \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$ and the the first row implies $\sum_{i=1}^n \widehat{e}_i = 0$. "$\heartsuit$" indicates the entries irrelevant to our purpose.)

The operation of OLS bears a natural geometric interpretation. Notice $\mathcal{X} = \left\{Xb : b \in \mathbb{R}^K\right\}$ is the linear space spanned by the K columns of $X = [X_{\cdot 1}, \ldots, X_{\cdot K}]$, which is of K-dimension if the columns are linearly independent. The OLS estimator is the minimizer of $\min_{b\in\mathbb{R}^K} \|Y - Xb\|$ (Square the Euclidean norm or not does not change the minimizer because a^2 is a monotonic transformation for $a \geq 0$). In other words, $X\widehat{\beta}$ is the point in $\mathcal{X}$ such that it is the closest to the vector Y in terms of the Euclidean norm.

The relationship $Y = X\widehat{\beta} + \widehat{e}$ decomposes Y into two orthogonal vectors $X\widehat{\beta}$ and $\widehat{e}$ as $\left\langle X\widehat{\beta}, \widehat{e}\right\rangle = \widehat{\beta}X'\widehat{e} = 0'_K$, where $\langle\cdot,\cdot\rangle$ is the *inner product* of two vectors. Therefore $X\widehat{\beta}$ is the *projection* of Y onto $\mathcal{X}$, and $\widehat{e}$ is the corresponding *projection residuals*. The Pythagorean theorem implies

$$\|Y\|^2 = \|X\widehat{\beta}\|^2 + \|\widehat{e}\|^2 .$$



Example 3.1. Here is a simple simulated example to demonstrate the properties of OLS. Given $(x_{1i}, x_{2i}, x_{3i}, e_i)' \sim N(0_4, I_4)$, the dependent variable y_i is generated from

$$y_i = 0.5 + 2 \cdot x_{1i} - 1 \cdot x_{2i} + e_i$$

The researcher does not know x_{3i} is redundant, and he regresses y_i on $(1, x_{1i}, x_{2i}, x_{3i})$.

```
library(magrittr); set.seed(2020-9-23)
n = 20 # sample size
K = 4  # number of paramters
b0 = as.matrix( c(0.5, 2, -1, 0) ) # the true coefficient
X = cbind(1, matrix( rnorm(n * (K-1)), nrow = n ) )  # the regressor matrix
e = rnorm(n) # the error term
Y = X %*% b0 + e # generate the dependent variable
bhat = solve(t(X) %*% X, t(X) %*% Y ) %>% as.vector() %>% print()

## [1]  0.3151672  1.9546647 -0.8520387  0.1508770
```

The estimated coefficient $\widehat{\beta}$ is (0.315, 1.955, -0.852, 0.151). It is close to the true value, but not very accurate due to the small sample size.

```
ehat = Y - X %*% bhat
as.vector( t(X) %*% ehat ) %>% print()

## [1]  1.665335e-15 -1.776357e-15 -3.996803e-15 -2.366163e-15
```

```r
MX = diag(n) - X %*% solve( crossprod(X) ) %*% t(X)
data.frame(e = e, ehat = ehat, MXY = MX%*%Y, MXe = MX%*%e ) %>% head()

##              e       ehat        MXY        MXe
## 1  0.11468775  0.2195704  0.2195704  0.2195704
## 2 -1.09300952 -0.7358326 -0.7358326 -0.7358326
## 3  1.06084816  0.7873848  0.7873848  0.7873848
## 4 -0.93399293 -0.5797384 -0.5797384 -0.5797384
## 5  0.05697917  0.3604994  0.3604994  0.3604994
## 6  0.03431877  0.1489134  0.1489134  0.1489134

cat("The mean of the residual is ", mean(ehat), ".\n")

## The mean of the residual is  7.215094e-17 .

cat("The mean of the true error term is", mean(e), ".")

## The mean of the true error term is -0.1582708 .
```

3.2 Subvector

The Frish-Waugh-Lovell (FWL) theorem is an algebraic fact about the formula of a subvector of the OLS estimator. To derive the FWL theorem we need to use the inverse of partitioned matrix. For a positive definite symmetric matrix $A = \begin{pmatrix} A_{11} & A_{12} \\ A_{12}' & A_{22} \end{pmatrix}$, the inverse can be written as

$$A^{-1} = \begin{pmatrix} \left(A_{11} - A_{12}A_{22}^{-1}A_{12}'\right)^{-1} & -\left(A_{11} - A_{12}A_{22}^{-1}A_{12}'\right)^{-1} A_{12}A_{22}^{-1} \\ -A_{22}^{-1}A_{12}'\left(A_{11} - A_{12}A_{22}^{-1}A_{12}'\right)^{-1} & \left(A_{22} - A_{12}'A_{11}^{-1}A_{12}\right)^{-1} \end{pmatrix}.$$

In our context of OLS estimator, let $X = \begin{pmatrix} X_1 & X_2 \end{pmatrix}$

$$\begin{aligned} \begin{pmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{pmatrix} = \widehat{\beta} &= (X'X)^{-1}X'Y \\ &= \left(\begin{pmatrix} X_1' \\ X_2' \end{pmatrix} \begin{pmatrix} X_1 & X_2 \end{pmatrix} \right)^{-1} \begin{pmatrix} X_1'Y \\ X_2'Y \end{pmatrix} \\ &= \begin{pmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{pmatrix}^{-1} \begin{pmatrix} X_1'Y \\ X_2'Y \end{pmatrix} \\ &= \begin{pmatrix} \left(X_1'M_{X_2}'X_1\right)^{-1} & -\left(X_1'M_{X_2}'X_1\right)^{-1} X_1'X_2 \left(X_2'X_2\right)^{-1} \\ \heartsuit & \heartsuit \end{pmatrix} \begin{pmatrix} X_1'Y \\ X_2'Y \end{pmatrix}. \end{aligned}$$

The subvector

$$\begin{aligned}\widehat{\beta}_1 &= \left(X_1' M_{X_2}' X_1\right)^{-1} X_1' Y - \left(X_1' M_{X_2}' X_1\right)^{-1} X_1' X_2 \left(X_2' X_2\right)^{-1} X_2' Y \\ &= \left(X_1' M_{X_2}' X_1\right)^{-1} X_1' Y - \left(X_1' M_{X_2}' X_1\right)^{-1} X_1' P_{X_2} Y \\ &= \left(X_1' M_{X_2}' X_1\right)^{-1} \left(X_1' Y - X_1' P_{X_2} Y\right) \\ &= \left(X_1' M_{X_2}' X_1\right)^{-1} X_1' M_{X_2} Y.\end{aligned}$$

Notice that $\widehat{\beta}_1$ can be obtained by the following:

1. Regress Y on X_2, obtain the residual $\tilde{Y}$;
2. Regress X_1 on X_2, obtain the residual $\tilde{X}_1$;
3. Regress $\tilde{Y}$ on $\tilde{X}_1$, obtain OLS estimates $\widehat{\beta}_1$.

Similar derivation can also be carried out in the population linear projection. See Hansen (2020) [E] Chapter 2.22-23.

```
X1 = X[,1:2];X2 = X[,3:4]
PX2 = X2 %*% solve( t(X2) %*% X2) %*% t(X2)
MX2 = diag(rep(1,n)) - PX2

bhat1 <- (solve(t(X1)%*% MX2 %*% X1, t(X1) %*% MX2 %*% Y )) %>%
  as.vector() %>% print()

## [1] 0.3151672 1.9546647

ehat1 = MX2 %*% Y - MX2 %*% X1 %*% bhat1
data.frame(ehat = ehat, ehat1 = ehat1) %>% head() %>% print()

##         ehat      ehat1
## 1  0.2195704  0.2195704
## 2 -0.7358326 -0.7358326
## 3  0.7873848  0.7873848
## 4 -0.5797384 -0.5797384
## 5  0.3604994  0.3604994
## 6  0.1489134  0.1489134
```

3.3 Goodness of Fit

Consider the regression with the intercept $Y = X_1\beta_1 + \beta_2 + e$. The OLS estimator gives

$$Y = \widehat{Y} + \widehat{e} = \left(X_1\widehat{\beta}_1 + \widehat{\beta}_2\right) + \widehat{e}. \tag{3.2}$$

Applying the FWL theorem with $X_2 = \iota$, where ι (Greek letter, iota) is an $n \times 1$ vector of 1's. Then $M_{X_2} = M_\iota = I_n - \frac{1}{n}\iota\iota'$. Notice M_ι is the *demeaner* in that $M_\iota z = z - \bar{z}$. It subtract the vector mean $\bar{z} = \frac{1}{n}\sum_{i=1}^n z_i$ from the original vector z. The above three-step procedure becomes

1. Regress Y on ι, and the residual is $M_\iota Y$;
2. Regress X_1 on ι, and the residual is $M_\iota X_1$;
3. Regress $M_\iota Y$ on $M_\iota X_1$, and the OLS estimates is exactly the same as $\widehat{\beta}_1$ in (3.2).

The last step gives the decomposition

$$M_\iota Y = M_\iota X_1 \widehat{\beta}_1 + \tilde{e}, \tag{3.3}$$

and the Pythagorean theorem implies

$$\|M_\iota Y\|^2 = \|M_\iota X_1 \widehat{\beta}_1\|^2 + \|\widehat{e}\|^2 .$$

Exercise 3.1. Show that $\widehat{e}$ in (3.2) is exactly the same as $\tilde{e}$ in (3.3).

R-squared is a popular measure of goodness-of-fit in the linear regression. The (in-sample) R-squared

$$R^2 = \frac{\|M_\iota X_1 \widehat{\beta}_1\|^2}{\|M_\iota Y\|^2} = 1 - \frac{\|\tilde{e}\|^2}{\|M_\iota Y\|^2}.$$

is well defined only when a constant is included in the regressors.

Exercise 3.2. Show

$$R^2 = \frac{\widehat{Y}' M_\iota \widehat{Y}}{Y' M_\iota Y} = \frac{\sum_{i=1}^n (\widehat{y}_i - \overline{y})^2}{\sum_{i=1}^n (y_i - \overline{y})^2}$$

as in the decomposition (3.2). In other words, it is the ratio between the sample variance of $\widehat{Y}$ and the sample variance of Y.

The magnitude of R-squared varies in different contexts. In macro models with the lagged dependent variables, it is not unusually to observe R-squared larger than 90%. In cross sectional regressions it is often below 20%.

Exercise 3.3. Consider a short regression "regress y_i on x_{1i}" and a long regression "regress y_i on (x_{1i}, x_{2i})". Given the same dataset (Y, X_1, X_2), show that the R-squared from the short regression is no larger than that from the long regression. In other words, we can always (weakly) increase R^2 by adding more regressors.

Conventionally we consider the regressions when the number of regressors K is much smaller the sample size n. In the era of big data, it can happen that we have more potential regressors than the sample size.

Exercise 3.4. Show $R^2 = 1$ when $K \geq n$. (When $K > n$, the matrix $X'X$ must be rank deficient. We can generalize the definition OLS fitting as any vector that minimizes $\|Y - Xb\|^2$ though the minimizer is not unique.

```
n = 5; K = 6;
Y = rnorm(n)
X = matrix( rnorm(n*K), n)
summary( lm(Y~X) )
```

```
##
## Call:
## lm(formula = Y ~ X)
##
## Residuals:
## ALL 5 residuals are 0: no residual degrees of freedom!
##
## Coefficients: (2 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.2229         NA      NA       NA
## X1           -0.6422         NA      NA       NA
## X2            0.1170         NA      NA       NA
## X3            1.1844         NA      NA       NA
## X4            0.5883         NA      NA       NA
## X5                NA         NA      NA       NA
## X6                NA         NA      NA       NA
##
## Residual standard error: NaN on 0 degrees of freedom
## Multiple R-squared:      1,Adjusted R-squared:    NaN
## F-statistic:   NaN on 4 and 0 DF,  p-value: NA
```

With a new dataset $(Y^{\text{new}}, X^{\text{new}})$, the *out-of-sample* (OOS) R-squared is

$$OOS\ R^2 = \frac{\widehat{\beta}' X^{\text{new}\prime} M_\iota X^{\text{new}} \widehat{\beta}}{Y^{\text{new}\prime} M_\iota Y^{\text{new}}}.$$

OOS R-squred measures the goodness of fit in a new dataset given the coefficient estimated from the original data. In financial market shorter-term predictive models, a person may become a billion if he can systematically achieve 2% OOS R-squared.

3.4 Summary

The linear algebraic properties holds in finite sample no matter the data are taken as fixed numbers or random variables. The Gauss Markov theorem holds under two crucial assumptions: linear CEF and homoskedasticity.

Historical notes: Carl Friedrich Gauss (1777–1855) claimed he had come up with the operation of OLS in 1795. With only three data points at hand, Gauss successfully applied his method to predict the location of the dwarf planet Ceres in 1801. While Gauss did not publish the work on OLS until 1809, Adrien-Marie Legendre (1752–1833) presented this method in 1805. Today people tend to attribute OLS to Gauss, assuming that a giant like Gauss had no need to tell a lie to steal Legendre's discovery.

3.5 Appendix

Let A be any $n \times K$ generic real matrix. *Singular value decomposition* (SVD) factorizes $A = USV'$, where U is an $n \times n$ real unitary matrix (A real unitary matrix is invertible and $U'U = UU' = I$,

which implies $U^{-1} = U'$), $S = \begin{bmatrix} S_1 \\ 0_{(n-K)\times K} \end{bmatrix}$ is an $n \times K$ rectangular diagonal matrix with S_1 a $K \times K$ diagonal matrix of non-negative real elements (called *singular values*), and V is a $K \times K$ real unitary matrix.

We apply SVD to the projection matrix $P_X = X\left(X'X\right)^{-1}X$, where X is an $n \times K$ data matrix with K linearly independent columns. Substitute $X = USV'$ into P_X:

$$\begin{aligned} P_X &= USV'\left(VS'U'USV'\right)^{-1}VS'U' = USV'\left(VS'SV'\right)^{-1}VS'U' \\ &= USV'V'^{-1}\left(S'S\right)^{-1}V^{-1}VS'U' = US\left(S'S\right)^{-1}S'U' \\ &= U\begin{bmatrix} S_1 \\ 0 \end{bmatrix} S_1^{-1}S_1^{-1}\begin{bmatrix} S_1 & 0 \end{bmatrix}U' = U\begin{bmatrix} I_K & 0_{K\times(n-K)} \\ 0_{(n-K)\times K} & 0_{(n-K)\times(n-K)} \end{bmatrix}U' \\ &= U\,\mathrm{diag}\left(\iota_K, 0_{n-K}\right)U'. \end{aligned}$$

All real symmetric matrices are diagonalizable, and the the last expression is the diagonalization of P_X. The projection matrix P_X has K repeated eigenvalues of 1 and $(n-K)$ repeated eigenvalues of 0, and obviously $\mathrm{rank}(P_X) = K$.

Two generic square matrices A and B are *similar* if there exists an invertible matrix Q such that $A = Q^{-1}BQ$. By this definition, P_X is similar to the diagonal matrix $\mathrm{diag}\left(\iota_K, 0_{n-K}\right)$, and $M_X = I_n - P_X$ is similar to $\mathrm{diag}\left(0_K, \iota_{n-K}\right)$ because

$$\begin{aligned} U'M_XU &= U'\left(I_n - P_X\right)U = U'U - U'P_XU \\ &= I_n - \mathrm{diag}\left(\iota_K, 0_{n-K}\right) = \mathrm{diag}\left(0_K, \iota_{n-K}\right). \end{aligned}$$

It implies that $\mathrm{rank}(M_X) = n - K$.

Both P_X and M_X are symmetric idempotent matrices. For a general idempotent matrices C which does not have to be symmetric,

- C is diagonalizable (See Horn and Johnson (1985, p.148)).

This fact immediately implies that

- All eigenvalues of C are either 0 and 1;
- $\mathrm{rank}\left(C\right) = \mathrm{trace}\left(C\right)$.

`Zhentao Shi. Oct 4.`

Horn, R. A. and C. R. Johnson (1985). *Matrix analysis*. Cambridge University Press.

Chapter 4

Least Squares: Finite Sample Theory

4.1 Maximum Likelihood

There are very few *principles* in statistics, and maximum likelihood is one of them. In this chapter, we first give an introduction of the maximum likelihood estimation. Consider a random sample of $Z = (z_1, z_2, \ldots, z_n)$ drawn from a parametric distribution with density $f_z(z_i; \theta)$, where z_i is either a scalar random variable or a random vector. A parametric distribution is completely characterized by a finite-dimensional parameter θ. We know that θ belongs to a parameter space Θ. We use the data to estimate θ.

The log-likelihood of observing the entire sample Z is

$$L_n(\theta; Z) := \log\left(\prod_{i=1}^{n} f_z(z_i; \theta)\right) = \sum_{i=1}^{n} \log f_z(z_i; \theta). \tag{4.1}$$

In reality the sample Z is given and for each $\theta \in \Theta$ we can evaluate $L(\theta; Z)$. The maximum likelihood estimator is

$$\widehat{\theta}_{MLE} := \arg\max_{\theta \in \Theta} L_n(\theta; Z).$$

Why maximizing the log-likelihood function is desirable? An intuitive explanation is that $\widehat{\theta}_{MLE}$ makes observing Z the "most likely" in the entire parametric space.

A more formal justification requires an explicitly defined distance. Suppose that the true parameter value that generates the data is θ_0, so that the true distribution is $f_z(z_i; \theta_0)$. Any generic point $\theta \in \Theta$ produce $f_z(z_i; \theta)$. To measure their difference, we introduce the *Kullback-Leibler distance*, or Kullback-Leibler *divergence*, defined as

$$\begin{aligned} D_f(\theta, \theta_0) = D(f_z(z_i; \theta), f_z(z_i; \theta_0)) &:= E_{\theta_0}\left[\log \frac{f_z(z_i; \theta_0)}{f_z(z_i; \theta)}\right] \\ &= E_{\theta_0}[\log f_z(z_i; \theta_0)] - E_{\theta_0}[\log f_z(z_i; \theta)]. \end{aligned}$$

We call it a "distance" because it is non-negative. To see this, notice that $-\log(\cdot)$ is strictly convex and then by Jensen's inequality

$$\begin{aligned} E_{\theta_0}\left[\log \frac{f_z(z_i; \theta_0)}{f_z(z_i; \theta)}\right] &= E_{\theta_0}\left[-\log \frac{f_z(z_i; \theta)}{f_z(z_i; \theta_0)}\right] \geq -\log\left(E_{\theta_0}\left[\frac{f_z(z_i; \theta)}{f_z(z_i; \theta_0)}\right]\right) \\ &= -\log\left(\int \frac{f_z(z_i; \theta)}{f_z(z_i; \theta_0)} f_z(z_i; \theta_0)\, dz_i\right) = -\log\left(\int f_z(z_i; \theta)\, dz_i\right) \\ &= -\log 1 = 0, \end{aligned}$$

where $\int f_z(z_i;\theta)\,dz_i = 1$ for any pdf. The equality holds if and only if $f_z(z_i;\theta) = f_z(z_i;\theta_0)$ almost everywhere. Furthermore, if there is a one-to-one mapping between θ and $f_z(z_i;\theta)$ on Θ (identification), then $\theta_0 = \arg\min_{\theta\in\Theta} D_f(\theta,\theta_0)$ is the unique solution.

In information theory, $-E_{\theta_0}[\log f_z(z_i;\theta_0)]$ is the *entropy* of the continuous distribution of $f_z(z_i;\theta_0)$. The Kullback-Leibler can be interpreted as the gap between $-E_{\theta_0}[\log f_z(z_i;\theta)]$, which is a function of θ, and the entropy at the true value θ_0, which is the maximal chaos a thermodynamic system can achieve.

Example 4.1. Consider the Gaussian location model $z_i \sim N(\mu,1)$, where μ is the unknown parameter to be estimated. The likelihood of observing z_i is $f_z(z_i;\mu) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}(z_i-\mu)^2\right)$. The likelihood of observing the sample Z is

$$f_Z(\mu;Z) = \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}(x_i-\mu)^2\right)$$

and the log-likelihood is

$$L_n = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(z_i-\mu)^2.$$

The (averaged) log-likelihood function for n observations is

$$\ell_n(\mu) = -\frac{1}{2}\log(2\pi) - \frac{1}{2n}\sum_{i=1}^{n}(z_i-\mu)^2.$$

We work with the averaged log-likelihood ℓ_n, instead of the (raw) log-likelihood L_n, to make it directly comparable with its population counterpart

$$\begin{aligned}
E_{\mu_0}[\ell_n(\mu)] &= -\frac{1}{2}\log(2\pi) - \frac{1}{2}E\left[(z_i-\mu)^2\right] \\
&= -\frac{1}{2}\log(2\pi) - \frac{1}{2}E\left[((z_i-\mu_0)+(\mu_0-\mu))^2\right] \\
&= -\frac{1}{2}\log(2\pi) - \frac{1}{2}E\left[(z_i-\mu_0)^2\right] - E[z_i-\mu_0](\mu_0-\mu) - \frac{1}{2}(\mu_0-\mu)^2 \\
&= -\frac{1}{2}\log(2\pi) - \frac{1}{2} - \frac{1}{2}(\mu-\mu_0)^2.
\end{aligned}$$

Obviously, $\ell_n(\mu)$ is maximized at $\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i$ while $E_{\mu_0}[\ell_n(\mu)]$ is maximized at $\mu = \mu_0$. The Kullback-Leibler distance in this example is

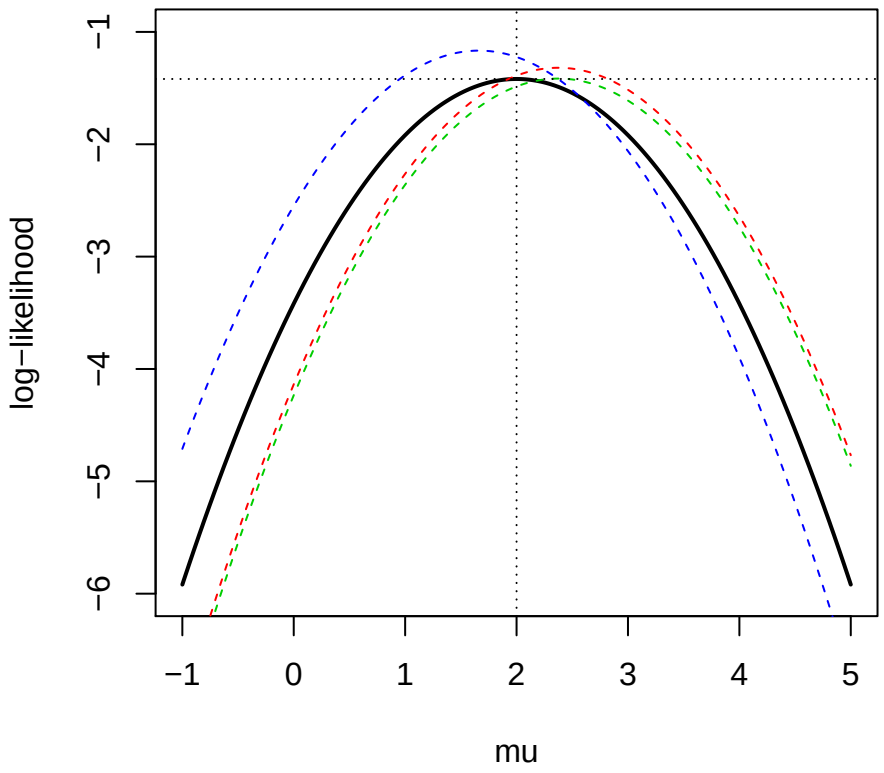
$$\begin{aligned}
D(\mu,\mu_0) &= E_{\mu_0}[\ell_n(\mu_0)] - E_{\mu_0}[\ell_n(\mu)] \\
&= -E_{\mu_0}[\ell_n(\mu)] - \text{entropy}(\mu_0) \\
&= \frac{1}{2}(\mu-\mu_0)^2,
\end{aligned}$$

where $-E_{\mu_0}[\ell_n(\mu_0)] = \frac{1}{2}(\log(2\pi)+1)$ is the entropy of the normal distribution with unit variance.

We use the following code to demonstrate the population log-likelihood $E[\ell_n(\mu)]$ when $\mu_0 = 2$ and the 3 sample realizations when $n = 4$.

```r
set.seed(2020-10-7)
mu0 <- 2; gamma0 <- 1
# population likelihood function
L <- function(mu) {
  ell = -0.5 * log(2*pi*gamma0) - 0.5 / gamma0 *( 1 + (mu - mu0)^2 )
  return(ell) }
# sample likelihood function
Ln <- function(mu) {
  elln = -0.5 * log(2*pi*gamma0) - 0.5 / gamma0 * mean( (z - mu)^2 )
  return(elln) }

mu_base = mu0 + seq(-3, 3, by = 0.01)
# draw sample log-likelihood graph
n = 4
lnz = matrix(0, length(mu_base), 3)
for (rr in 1:3){
  z <- rnorm(n, mu0, sqrt(gamma0) )
  lnz[,rr] <- plyr::laply(.data = mu_base, .fun = Ln)
}
# draw the graph
matplot(x = mu_base, y = cbind( L(mu_base), lnz),
        type = "l", lty = c(1, rep(2,3)),
        lwd =  c(2,rep(1,3)), col = 1:4, ylim = c(-6, -1),
        xlab = "mu", ylab = "log-likelihood")
abline(v = mu0, lty = 3)
abline(h = L(mu0), lty = 3)
```



4.2 Likelihood Estimation for Regression

Notation: y_i is a scalar, and $x_i = (x_{i1}, \ldots, x_{iK})'$ is a $K \times 1$ vector. Y is an $n \times 1$ vector, and X is an $n \times K$ matrix.

We continue with properties of OLS. Noticing that OLS coincides with the maximum likelihood estimator if the error term follows a normal distribution, we derive its finite-sample exact distribution which can be used for statistical inference. The Gauss-Markov theorem justifies the optimality of OLS under the classical assumptions.

In this chapter we employ the classical statistical framework under restrictive distributional assumption

$$y_i | x_i \sim N\left(x_i'\beta, \gamma\right), \tag{4.2}$$

where $\gamma = \sigma^2$ to ease the differentiation. This assumption is equivalent to $e_i | x_i = \left(y_i - x_i'\beta\right) | x_i \sim N\left(0, \gamma\right)$. Because the distribution of e_i is invariant to x_i, the error term $e_i \sim N\left(0, \gamma\right)$ and is statistically independent of x_i. This is a very strong assumption.

The likelihood of observing a pair (y_i, x_i) is

$$\begin{aligned} f_{yx}\left(y_i, x_i\right) &= f_{y|x}\left(y_i | x_i\right) f_x\left(x\right) \\ &= \frac{1}{\sqrt{2\pi\gamma}} \exp\left(-\frac{1}{2\gamma}\left(y_i - x_i'\beta\right)^2\right) \times f_x\left(x\right), \end{aligned}$$

where f_{yx} is the joint pdf, $f_{y|x}$ is the conditional pdf and f_x is the marginal pdf of x, and the second equality holds under the assumption (4.2). The likelihood a random sample $(y_i, x_i)_{i=1}^n$ is

$$\begin{aligned}\prod_{i=1}^{n} f_{yx}(y_i, x_i) &= \prod_{i=1}^{n} f_{y|x}(y_i|x_i) f_x(x) \\ &= \prod_{i=1}^{n} f_{y|x}(y_i|x_i) \times \prod_{i=1}^{n} f_x(x) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\gamma}} \exp\left(-\frac{1}{2\gamma}\left(y_i - x_i'\beta\right)^2\right) \times \prod_{i=1}^{n} f_x(x).\end{aligned}$$

The parameters of interest (β, γ) are irrelevant to the second term $\prod_{i=1}^n f_x(x)$ for they appear only in the conditional likelihood

$$\prod_{i=1}^{n} f_{y|x}(y_i|x_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\gamma}} \exp\left(-\frac{1}{2\gamma}\left(y_i - x_i'\beta\right)^2\right).$$

We focus on the conditional likelihood. To facilitate derivation, we work with the (averaged) conditional log-likelihood function

$$\ell_n(\beta, \gamma) = -\frac{1}{2}\log 2\pi - \frac{1}{2}\log\gamma - \frac{1}{2n\gamma}\sum_{i=1}^{n}\left(y_i - x_i'\beta\right)^2,$$

for $\log(\cdot)$ is a monotonic transformation that does not change the maximizer. The maximum likelihood estimator $\widehat{\beta}_{MLE}$ can be found using the FOC:

$$\frac{\partial}{\partial\beta}\ell_n(\beta, \gamma) = \frac{1}{n\gamma}\sum_{i=1}^{n} x_i\left(y_i - x_i'\beta\right) = 0.$$

Rearranging the above equation in matrix form $X'X\widehat{\beta}_{MLE} = X'Y$, we explicitly solve

$$\widehat{\beta}_{MLE} = (X'X)^{-1}X'Y$$

when $X'X$ is invertible. The maximum likelihood estimator (MLE) coincides with the OLS estimator. Similarly, the other FOC with respect to γ gives $\widehat{\gamma}_{\text{MLE}} = \widehat{e}'\widehat{e}/n$.

4.3 Finite Sample Distribution

We can show the finite-sample exact distribution of $\widehat{\beta}$ assuming the error term follows a Gaussian distribution. *Finite sample distribution* means that the distribution holds for any n; it is in contrast to *asymptotic distribution*, which is a large sample approximation to the finite sample distribution. We first review some properties of a generic jointly normal random vector.

Fact 4.1. *Let $z \sim N(\mu, \Omega)$ be an $l \times 1$ random vector with a positive definite variance-covariance matrix Ω. Let A be an $m \times l$ non-random matrix where $m \leq l$. Then $Az \sim N(A\mu, A\Omega A')$.*

Fact 4.2. *If $z \sim N(0, 1)$, $w \sim \chi^2(d)$ and z and w are independent. Then $\frac{z}{\sqrt{w/d}} \sim t(d)$.*

The OLS estimator

$$\widehat{\beta} = \left(X'X\right)^{-1} X'Y = \left(X'X\right)^{-1} X' \left(X'\beta + e\right) = \beta + \left(X'X\right)^{-1} X'e,$$

and its conditional distribution can be written as

$$\begin{aligned}
\widehat{\beta}|X &= \beta + \left(X'X\right)^{-1} X'e|X \\
&\sim \beta + \left(X'X\right)^{-1} X' \cdot N\left(0_n, \sigma^2 \cdot I_n\right) \\
&\sim N\left(\beta, \sigma^2 \left(X'X\right)^{-1} X'X \left(X'X\right)^{-1}\right) \sim N\left(\beta, \sigma^2 \left(X'X\right)^{-1}\right)
\end{aligned}$$

by Fact 4.1. The k-th element of the vector coefficient

$$\widehat{\beta}_k|X = \eta_k'\widehat{\beta}|X \sim N\left(\beta_k, \sigma^2 \eta_k' \left(X'X\right)^{-1} \eta_k\right) \sim N\left(\beta_k, \sigma^2 \left(X'X\right)^{-1}_{kk}\right),$$

where $\eta_k = \left(1\left\{l = k\right\}\right)_{l=1,\ldots,K}$ is the selector of the k-th element.

In reality, σ^2 is an unknown parameter, and

$$s^2 = \widehat{e}'\widehat{e}/\left(n - K\right) = e'M_X e/\left(n - K\right)$$

is an unbiased estimator of σ^2. Consider the t-statistic

$$\begin{aligned}
T_k &= \frac{\widehat{\beta}_k - \beta_k}{\sqrt{s^2 \left[\left(X'X\right)^{-1}\right]_{kk}}} = \frac{\widehat{\beta}_k - \beta_k}{\sqrt{\sigma^2 \left[\left(X'X\right)^{-1}\right]_{kk}}} \cdot \frac{\sqrt{\sigma^2}}{\sqrt{s^2}} \\
&= \frac{\left(\widehat{\beta}_k - \beta_k\right) / \sqrt{\sigma^2 \left[\left(X'X\right)^{-1}\right]_{kk}}}{\sqrt{\frac{e'}{\sigma} M_X \frac{e}{\sigma} / \left(n - K\right)}}.
\end{aligned}$$

The numerator follows a standard normal, and the denominator follows $\frac{1}{n-K}\chi^2\left(n - K\right)$. Moreover, the numerator and the denominator are statistically independent (See Section 4.7). As a result, we conclude $T_k \sim t\left(n - K\right)$ by Fact 4.2. This finite sample distribution allows us to conduct statistical inference.

4.4 Mean and Variance

Now we relax the normality assumption and statistical independence. Instead, we represent the regression model as $Y = X\beta + e$ and

$$\begin{aligned}
E[e|X] &= 0_n \\
\text{var}\left[e|X\right] &= E\left[ee'|X\right] = \sigma^2 I_n.
\end{aligned}$$

where the first condition is the *mean independence* assumption, and the second condition is the *homoskedasticity* assumption. These assumptions are about the first and second *moments* of e_i conditional on x_i. Unlike the normality assumption, they do not restrict the distribution of e_i.

- Unbiasedness:

$$\begin{aligned}
E\left[\widehat{\beta}|X\right] &= E\left[\left(X'X\right)^{-1} XY|X\right] = E\left[\left(X'X\right)^{-1} X\left(X'\beta + e\right)|X\right] \\
&= \beta + \left(X'X\right)^{-1} XE\left[e|X\right] = \beta.
\end{aligned}$$

  Unbiasedness does not rely on homoskedasticity.

- Variance:

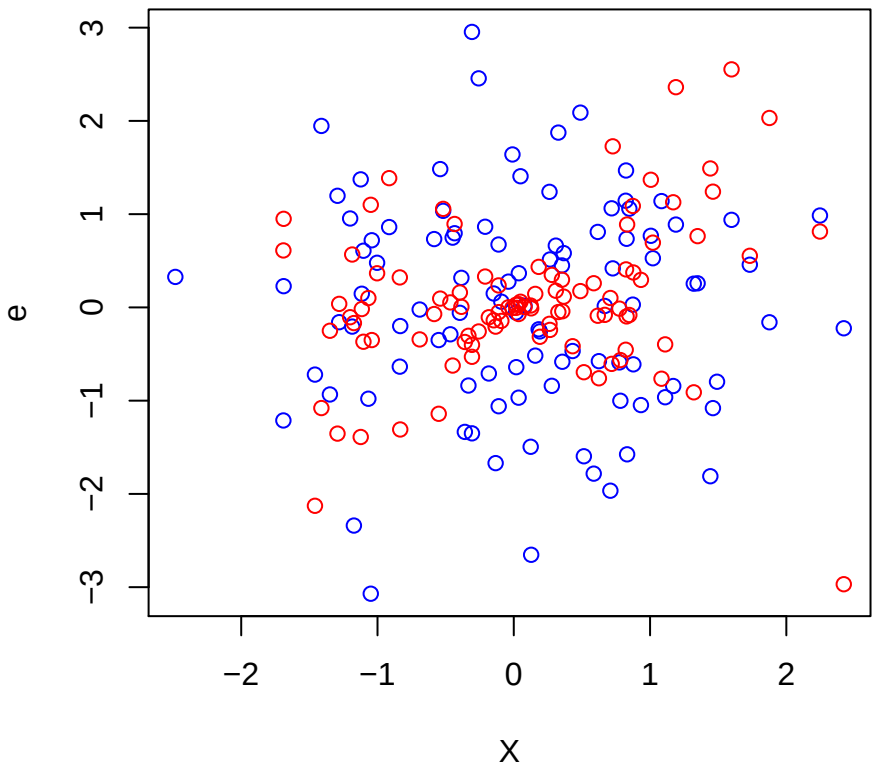
$$\begin{aligned}\text{var}\left[\widehat{\beta}|X\right] &= E\left[\left(\widehat{\beta}-E\widehat{\beta}\right)\left(\widehat{\beta}-E\widehat{\beta}\right)'|X\right] \\ &= E\left[\left(\widehat{\beta}-\beta\right)\left(\widehat{\beta}-\beta\right)'|X\right] \\ &= E\left[\left(X'X\right)^{-1}X'ee'X\left(X'X\right)^{-1}|X\right] \\ &= \left(X'X\right)^{-1}X'E\left[ee'|X\right]X\left(X'X\right)^{-1}\end{aligned}$$

  where the second equality holds as $E\left[\widehat{\beta}\right] = E\left[E\left[\widehat{\beta}|X\right]\right] = \beta$. Under the assumption of homoskedasticity, it can be simplified as

$$\text{var}\left[\widehat{\beta}|X\right] = \left(X'X\right)^{-1}X'\left(\sigma^2 I_n\right)X\left(X'X\right)^{-1} = \sigma^2\left(X'X\right)^{-1}.$$

Example 4.2. (Heteroskedasticity) If $e_i = x_i u_i$, where x_i is a scalar random variable, u_i is statistically independent of x_i, $E[u_i] = 0$ and $E\left[u_i^2\right] = \sigma^2$. Then $E[e_i|x_i] = 0$ but $E\left[e_i^2|x_i\right] = \sigma^2 x_i^2$ is a function of x_i. We say e_i^2 is a heteroskedastic error.

```
n = 100; X = rnorm(n)
e1 = rnorm(n)
plot( y = e1, x = X, col = "blue", ylab = "e")
e2 = X * rnorm(n)
points( y = e2, x = X, col = "red")
```



It is important to notice that independently and identically distributed sample (iid) (y_i, x_i) does not imply homoskedasticity. Homoskedasticity or heterskedasticity is about the relationship between $(x_i, e_i = y_i - \beta x)$, whereas iid is about the relationship between (y_i, x_i) and (y_j, x_j) for $i \neq j$.

4.5 Gauss-Markov Theorem

Gauss-Markov theorem is concerned about the optimality of OLS. It justifies OLS as the efficient estimator among all linear unbiased ones. *Efficient* here means that it enjoys the smallest variance in a family of estimators.

We have shown that OLS is unbiased in that $E\left[\widehat{\beta}\right] = \beta$. There are numerous linearly unbiased estimators. For example, $(Z'X)^{-1} Z'y$ for $z_i = x_i^2$ is unbiased because $E\left[(Z'X)^{-1} Z'y\right] = E\left[(Z'X)^{-1} Z'(X\beta + e)\right] = \beta$. We cannot say OLS is better than those other unbiased estimators because they are equally good in this aspect. Thus, we move to the second order property of variance: an estimator is better if its variance is smaller.

Fact 4.3. *For two generic random vectors X and Y of the same size, we say X's variance is smaller or equal to Y's variance if $(\Omega_Y - \Omega_X)$ is a positive semi-definite matrix. The comparison is defined*

this way because for any non-zero constant vector c, the variance of the linear combination of X

$$\text{var}\left(c'X\right) = c'\Omega_X c \leq c'\Omega_Y c = \text{var}\left(c'Y\right)$$

is no bigger than the same linear combination of Y .

Let $\tilde{\beta} = A'y$ be a generic linear estimator, where A is any $n \times K$ functions of X. As

$$E\left[A'y|X\right] = E\left[A'\left(X\beta + e\right)|X\right] = A'X\beta.$$

So the linearity and unbiasedness of $\tilde{\beta}$ implies $A'X = I_n$. Moreover, the variance

$$\text{var}\left(A'y|X\right) = E\left[\left(A'y - \beta\right)\left(A'y - \beta\right)'|X\right] = E\left[A'ee'A|X\right] = \sigma^2 A'A.$$

Let $C = A - X\left(X'X\right)^{-1}$.

$$\begin{aligned} A'A - \left(X'X\right)^{-1} &= \left(C + X\left(X'X\right)^{-1}\right)'\left(C + X\left(X'X\right)^{-1}\right) - \left(X'X\right)^{-1} \\ &= C'C + \left(X'X\right)^{-1}X'C + C'X\left(X'X\right)^{-1} \\ &= C'C, \end{aligned}$$

where the last equality follows as

$$\left(X'X\right)^{-1}X'C = \left(X'X\right)^{-1}X'\left(A - X\left(X'X\right)^{-1}\right) = \left(X'X\right)^{-1} - \left(X'X\right)^{-1} = 0.$$

Therefore $A'A - \left(X'X\right)^{-1}$ is a positive semi-definite matrix. The variance of any $\tilde{\beta}$ is no smaller than the OLS estimator $\widehat{\beta}$. The above derivation shows OLS achieves the smallest variance among all linear unbiased estimators.

Homoskedasticity is a restrictive assumption. Under homoskedasticity, $\text{var}\left[\widehat{\beta}\right] = \sigma^2\left(X'X\right)^{-1}$. Popular estimator of σ^2 is the sample mean of the residuals $\widehat{\sigma}^2 = \frac{1}{n}\widehat{e}'\widehat{e}$ or the unbiased one $s^2 = \frac{1}{n-K}\widehat{e}'\widehat{e}$. Under heteroskedasticity, Gauss-Markov theorem does not apply.

4.6 Summary

The linear algebraic properties holds in finite sample no matter the data are taken as fixed numbers or random variables. The exact distribution under the normality assumption of the error term is the classical statistical results. The Gauss Markov theorem holds under two crucial assumptions: linear CEF and homoskedasticity.

Historical notes: MLE was promulgated and popularized by Ronald Fisher (1890–1962). He was a major contributor of the frequentist approach which dominates mathematical statistics today, and he sharply criticized the Bayesian approach. Fisher collected the iris flower dataset of 150 observations in his biological study in 1936, which can be displayed in R by typing `iris`. Fisher invented the many concepts in classical mathematical statistics, such as sufficient statistic, ancillary statistic, completeness, and exponential family, etc.

Further reading: Phillips (1983) offers a comprehensive treatment of exact small sample theory in econometrics. After that, theoretical studies in econometrics swiftly shifted to large sample theory, which we will introduce in the next chapter.

4.7 Appendix

$Y = (y_1, \ldots, y_n)$ consists of n iid observations. We say $T(Y)$ is a sufficient statistic for a parameter θ if the conditional probability $f(Y|T(Y))$ does not depend on θ. For example, for $y_i \sim N(\mu, \sigma^2)$ with known σ^2 and unknown μ, We verify that the sample mean $\bar{y} = n^{-1}\sum_{i=1}^n y_i$ is a sufficient statistic for μ. Notice that the joint density of Y is

$$\begin{aligned} f(Y) &= (2\pi\sigma)^{-n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^n (y_i - \mu)^2\right) \\ &= \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \bar{y})^2\right) \exp\left(-\frac{n}{2\sigma^2}(\bar{y} - \mu)^2\right). \end{aligned}$$

Because $\bar{y} \sim N(\mu, \sigma^2/n)$, the marginal density is

$$f(\bar{y}) = \left(2\pi\sigma^2/n\right)^{-1/2} \exp\left(-\frac{1}{2\sigma^2/n}(\bar{y} - \mu)^2\right).$$

The conditional density is

$$f(Y|\bar{y}) = \frac{f(Y)}{f(\bar{y})} = \frac{\left(2\pi\sigma^2\right)^{-n/2}}{\left(2\pi\sigma^2/n\right)^{-1/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \bar{y})^2\right)$$

is independent of μ, and thus $\bar{y}$ is a sufficient statistic for μ.

In the meantime, the sample standard deviation $s^2 = \frac{1}{n-1}\sum_{i=1}^n (y_i - \bar{y})$ is an *ancillary statistic* for μ , because the distribution of s^2 does not depend on μ.

Basu's theorem says that a *complete* sufficient statistic is statistically independent from any ancillary statistic. For a normal distribution with unknown mean and known variance, the sample mean $\bar{y}$ is the sufficient statistic and the sample standard deviation s^2 is an ancillary statistic.

A parametric distribution indexed by θ is a member of the *exponential family* is its PDF can be written as

$$f(Y|\theta) = h(Y)\, g(\theta) \exp\left(\eta(\theta)' T(Y)\right),$$

where $g(\theta)$ and $\eta(\theta)$ are functions depend, only on θ and $h(Y)$ and $T(Y)$ are functions depend only on Y. The normal distribution with known σ^2 and unknown μ belongs to the exponential family in view of the decomposition

$$\begin{aligned} f(Y) &= \left(\sqrt{2\pi}\sigma\right)^{-n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \mu)^2\right) \\ &= \underbrace{\exp\left(-\sum_{i=1}^n \frac{y_i^2}{2\sigma^2}\right)}_{h(Y)} \cdot \underbrace{\left(\sqrt{2\pi}\sigma\right)^{-n} \exp\left(-\frac{n}{2\sigma^2}\mu^2\right)}_{g(\theta)} \cdot \underbrace{\exp\left(\frac{\mu n}{2\sigma^2}\bar{y}\right)}_{\exp\left(\eta(\theta)' T(Y)\right)}. \end{aligned}$$

The exponential family is a class of distributions with the special functional form which is convenient for deriving sufficient statistics as well as other desirable properties in classical mathematical statistics.

`Zhentao Shi. Oct 6.`

Phillips, P. C. (1983). Exact small sample theory in the simultaneous equations model. *Handbook of econometrics 1*, 449–516.

Chapter 5

Asymptotic Theory

5.1 Introduction

Our universe, though enormous, consists of fewer than 10^{82} atoms, which is a finite number. However, mathematical ideas are not bounded by the secular reality. Asymptotic theory is concerned about the behavior of statistics when the sample size is arbitrarily large, up to infinity. It is an approximation technique to simplify complicated finite-sample analysis. It is the cornerstone of modern econometrics that seeks general conditions beyond the classical ones.

However, in reality we always have a finite sample, and in most cases it is difficult to increase the sample size for more accurate evaluation of the parameter of interest. In this sense, both the classical parametric approach and the asymptotic approach deviate from the reality, and it is difficult to judge which one is better. One may argue that the data size increases in the era of big data. However, the size of the model is also growing. The way I view the prevalence of asymptotic theory is its mathematical amenability. The required mathematical level to work with asymptotic theory, seemingly advance though, is lower than that for the finite sample theory. The underlying economics theory predicts that for substitute goods, if the price of one good is lowered, the demand increases.

5.1.1 Modes of Convergence

Before we talk about convergence of random variables, we first review what is convergence for a non-random sequence. Let $z_1, z_2, \ldots$ be an infinite sequence of non-random variables. Convergence of this non-random sequence means that for any $\varepsilon > 0$, there exists an $N(\varepsilon)$ such that for all $n > N(\varepsilon)$, we have $|z_n - z| < \varepsilon$. We say z is the limit of z_n, and write as $z_n \to z$.

Instead of a deterministic sequence, we are interested in the convergence of a sequence of random variables. Since a random variable is "random" thanks to the induced probability measure by the measurable function, we must be clear what *convergence* means. Several modes of convergence are often encountered.

- Convergence almost surely*
- *Convergence in probability*: for any $\varepsilon > 0$, as $n \to \infty$ the probability $P(\omega : |z_n(\omega) - z| < \varepsilon) \to 1$ (or equivalently $P(\omega : |z_n(\omega) - z| \geq \varepsilon) \to 0$. Denoted as $z_n \xrightarrow{p} z$. The limit z can be either a random variable or a non-random constant.
- *Squared-mean convergence*: $\lim_{n\to\infty} E\left[(z_n - z)^2\right] = 0$. Denoted as $z_n \xrightarrow{m.s.} z$.

Example 5.1. (z_n) is a sequence of binary random variables: $z_n = \sqrt{n}$ with probability $1/n$, and $z_n = 0$ with probability $1 - 1/n$. Then $z_n \xrightarrow{p} 0$ but $z_n \overset{m.s.}{\nrightarrow} 0$. For any $\varepsilon > 0$, we have $P(\omega : |z_n(\omega) - 0| < \varepsilon) = P(\omega : z_n(\omega) = 0) = 1 - 1/n \to 1$. Thus $z_n \xrightarrow{p} 0$. On the other hand, $E\left[(z_n - 0)^2\right] = n \cdot 1/n + 0 \cdot (1 - 1/n) = 1 \nrightarrow 0$. Thus $z_n \overset{m.s.}{\nrightarrow} 0$.

Convergence in probability does not count what happens on a subset in the sample space of small probability. Squared-mean convergence deals with the average over the entire probability space. If a random variable can take a wild value, with small probability though, it may blow away the squared-mean convergence. On the contrary, such irregularity does not undermine convergence in probability.

Both convergence in probability and squared-mean convergence are about convergence of random variables to a target random variable or constant. Instead, convergence in distribution is the convergence of CDF, but the random variable.

- Convergence in distribution: $x_n \xrightarrow{d} x$ if $F(x_n) \to F(x)$ for each x on which $F(x)$ is continuous.

Example 5.2. Convergence in distribution is about *pointwise* convergence of CDF, not the random variables themselves.Let $x \sim N(0,1)$. If $z_n = x + 1/n$, then $z_n \xrightarrow{p} x$ and of course $z_n \xrightarrow{d} x$. However, if $z_n = -x + 1/n$, or $z_n = y + 1/n$ where $y \sim N(0,1)$ is independent of x, then $z_n \xrightarrow{d} x$ but $z_n \overset{p}{\nrightarrow} x$.

Example 5.3. (z_n) is a sequence of binary random variables: $z_n = \sqrt{n}$ with probability $1/n$, and $z_n = 0$ with probability $1 - 1/n$. Then $z_n \xrightarrow{d} z = 0$. Because

$$F(z_n) = \begin{cases} 0 & z_n < 0 \\ 1 - 1/n & 0 \le z_n \le \sqrt{n} \\ 1 & z_n \ge \sqrt{n} \end{cases}.$$

$F(z) = \begin{cases} 0 & z < 0 \\ 1 & z \ge 0 \end{cases}$. It is obvious that $F(z_n)$ converges to $F(z)$ pointwise on the set where $F(z)$ is continuous.

Squared-mean convergence implies convergence in probability. Convergence in probability implies convergence in distribution. Cramer-Wold device handles convergence in distribution for random vectors. We say a sequence of K-dimensional random vectors (X_n) converge in distribution to X if $\lambda' X_n \xrightarrow{d} \lambda' X$ for any $\lambda \in \mathbb{R}^K$.

5.1.2 Law of Large Numbers

(Weak) law of large numbers (LLN) is a collection of statements about convergence in probability of the sample average to its population counterpart. The basic form of LLN is:

$$\frac{1}{n}\sum_{i=1}^{n}(z_i - E[z_i]) \xrightarrow{p} 0$$

as $n \to \infty$. Various versions of LLN work under different assumptions about the distributions and dependence of the random variables.

- Chebyshev LLN: if $(z_1, \ldots, z_n)$ is a sample of i.i.d. observations, $E[z_1] = \mu$, and $\sigma^2 = \text{var}[z_1] < \infty$ exists, then $\frac{1}{n}\sum_{i=1}^n z_i - \mu \xrightarrow{p} 0$.

Chebyshev LLN utilizes *Chebyshev inequality*.

- *Chebyshev inequality*: for any random variable x , we have $P(|x| > \varepsilon) \leq E\left[x^2\right]/\varepsilon^2$ for any $\varepsilon > 0$, if $E\left[x^2\right] < \infty$.

Chebyshev inequality is a special case of *Markov inequality*.

- *Markov inequality*: $P(|x| > \varepsilon) \leq E[|x|^r]/\varepsilon^r$ for $r \geq 1$ and any $\varepsilon > 0$, if $E[|x|^r] < \infty$.

It is easy to verify Markov inequality.

$$\begin{aligned} E[|x|^r] &= \int_{|x|>\varepsilon} |x|^r \, dF_X + \int_{|x|\leq\varepsilon} |x|^r \, dF_X \\ &\geq \int_{|x|>\varepsilon} |x|^r \, dF_X \geq \varepsilon^r \int_{|x|>\varepsilon} dF_X = \varepsilon^r P(|x| > \varepsilon). \end{aligned}$$

Consider a partial sum $S_n = \sum_{i=1}^n x_i$, where $\mu_i = E[x_i]$ and $\sigma_i^2 = \text{var}[x_i]$. We apply the Chebyshev inequality to the sample mean $\overline{x} - \bar{\mu} = n^{-1}(S_n - E[S_n])$.

$$\begin{aligned} P(|\bar{x} - \bar{\mu}| \geq \varepsilon) &= P(|S_n - E[S_n]| \geq n\varepsilon) \\ &\leq (n\varepsilon)^{-2} E\left[\sum_{i=1}^n (x_i - \mu_i)^2\right] \\ &= (n\varepsilon)^{-2} \text{var}\left(\sum_{i=1}^n x_i\right) \\ &= (n\varepsilon)^{-2} \left[\sum_{i=1}^n \text{var}(x_i) + \sum_{i=1}^n \sum_{j\neq i} \text{cov}(x_i, x_j)\right]. \end{aligned}$$

From the above derivation, convergence in probability holds as long as the right-hand side shrinks to 0 as $n \to \infty$. Actually, the convergence can be maintained under much more general conditions than just under the i.i.d. assumption. The random variables in the sample do not have to be identically distributed, and they do not have to be independent either.

Another useful LLN is *Kolmogorov LLN*. Since its derivation requires advanced knowledge of probability theory, we state the result without proof.

- Kolmogorov LLN: if $(z_1, \ldots, z_n)$ is a sample of i.i.d. observations and $E[z_1] = \mu$ exists, then $\frac{1}{n}\sum_{i=1}^n z_i - \mu \xrightarrow{p} 0$.

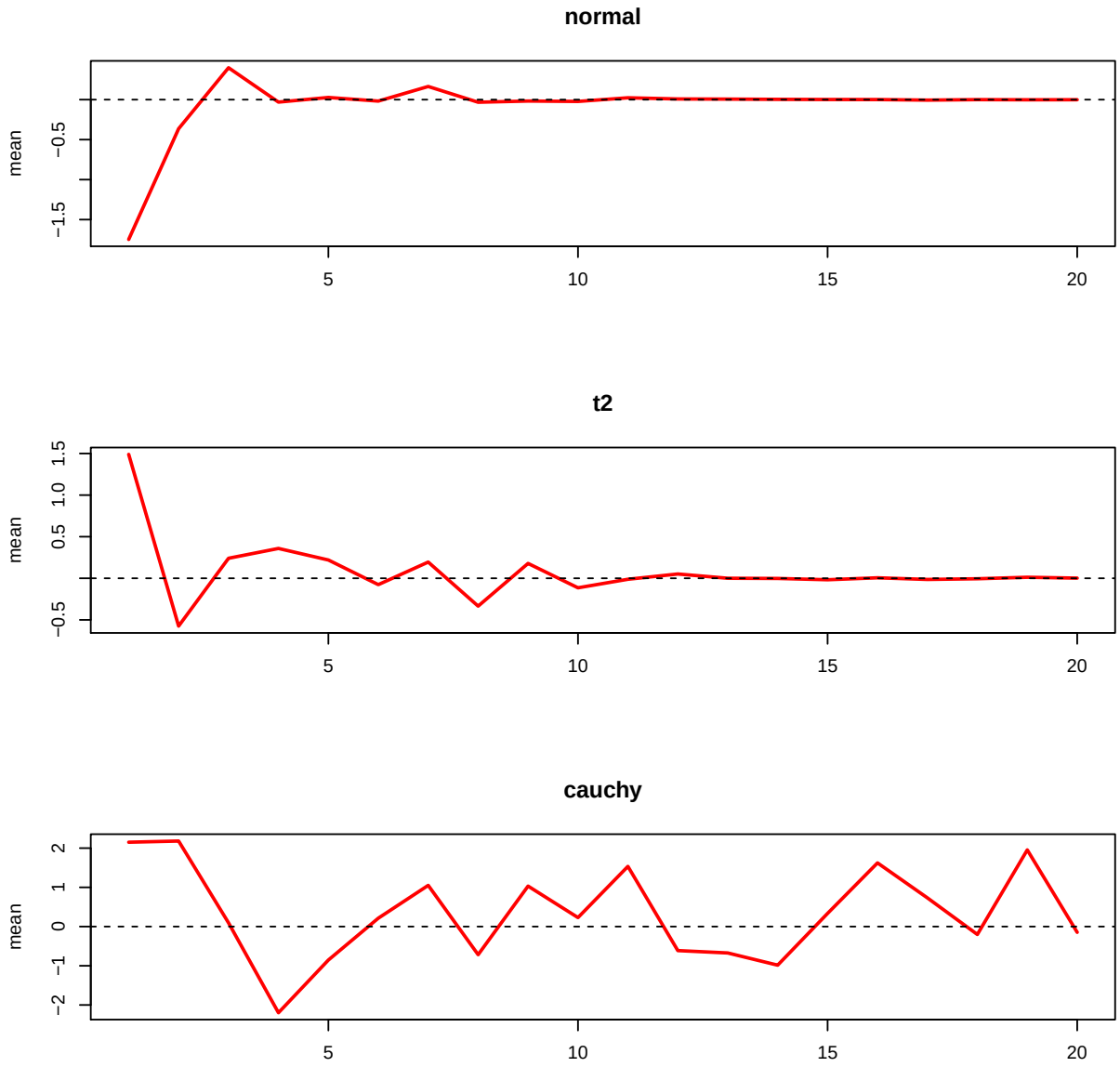
Compared to Chebyshev LLN, Kolmogorov LLN only requires the existence of the population mean, but not any higher moments. On the other hand, i.i.d. is essential for Kolmogorov LLN.

```
sample.mean = function( n, distribution ){
# get sample mean for a given distribution
  if (distribution == "normal"){ y = rnorm( n ) }
  else if (distribution == "t2") {y = rt(n, 2) }
  else if (distribution == "cauchy") {y = rcauchy(n) }
```

```
  return( mean(y) )
}

LLN.plot = function(distribution){
# draw the sample mean graph
  ybar = rep(0, length(NN) )
  for ( i in 1:length(NN)){
    n = NN[i]; ybar[i] = sample.mean(n, distribution)
  }
  plot(ybar, type = "l", col = "red", ylab = "mean", xlab = "",
  lwd = 2, main = distribution)
  abline(h = 0, lty = 2)
  return(ybar)
}

# calculation
NN = 2^(1:20); set.seed(888); par(mfrow = c(3,1))
l1 = LLN.plot("normal"); l2 = LLN.plot("t2"); l3 = LLN.plot("cauchy")
```



5.1.3 Central Limit Theorem

The central limit theorem (CLT) is a collection of probability results about the convergence in distribution to a stable law, usually the normal distribution. The basic form of the CLT is: for a sample $(z_1, \ldots, z_n)$ of *zero-mean* random variables,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} z_i \xrightarrow{d} N\left(0, \sigma^2\right). \tag{5.1}$$

Various versions of CLT work under different assumptions about the random variables.

Lindeberg-Levy CLT is the simplest CLT.

- If the sample is i.i.d., $E[x_1] = 0$ and var $\left[x_1^2\right] = \sigma^2 < \infty$, then CLT holds.

Lindeberg-Levy CLT is easy to verify by the characteristic function. For any random variable x, the function $\varphi_x(t) = E[\exp(ixt)]$ is called its *characteristic function*. The characteristic function fully describes a distribution, just like PDF or CDF. For example, the characteristic function of $N(\mu, \sigma^2)$ is $\exp\left(it\mu - \frac{1}{2}\sigma^2 t^2\right)$.

Here is a very heuristic argument. If $E\left[|x|^k\right] < \infty$ for a positive integer k, then

$$\varphi_X(t) = 1 + itE[X] + \frac{(it)^2}{2}E\left[X^2\right] + \ldots \frac{(it)^k}{k!}E\left[X^k\right] + o\left(t^k\right).$$

Under the assumption of Lindeberg-Levy CLT,

$$\varphi_{X_i/\sqrt{n}}(t) = 1 - \frac{t^2}{2n}\sigma^2 + o\left(\frac{t^2}{n}\right)$$

for all i, and by independence we have

$$\begin{aligned}\varphi_{\frac{1}{\sqrt{n}}\sum_{i=1}^n x_i}(t) = \prod_{i=1}^n \varphi_{x_i/\sqrt{n}}(t) &= \left(1 + i \cdot 0 - \frac{t^2}{2n}\sigma^2 + o\left(\frac{t^2}{n}\right)\right)^n \\ &\to \exp\left(-\frac{\sigma^2}{2}t^2\right),\end{aligned}$$

where the limit is exactly the characteristic function of $N(0, \sigma^2)$.

- Lindeberg-Feller CLT: i.n.i.d., and *Lindeberg condition*: for any fixed $\varepsilon > 0$,

$$\frac{1}{s_n^2}\sum_{i=1}^n E\left[x_i^2 \cdot \mathbf{1}\{|x_i| \geq \varepsilon s_n\}\right] \to 0$$

  where $s_n = \left(\sum_{i=1}^n \sigma_i^2\right)^{1/2}$.

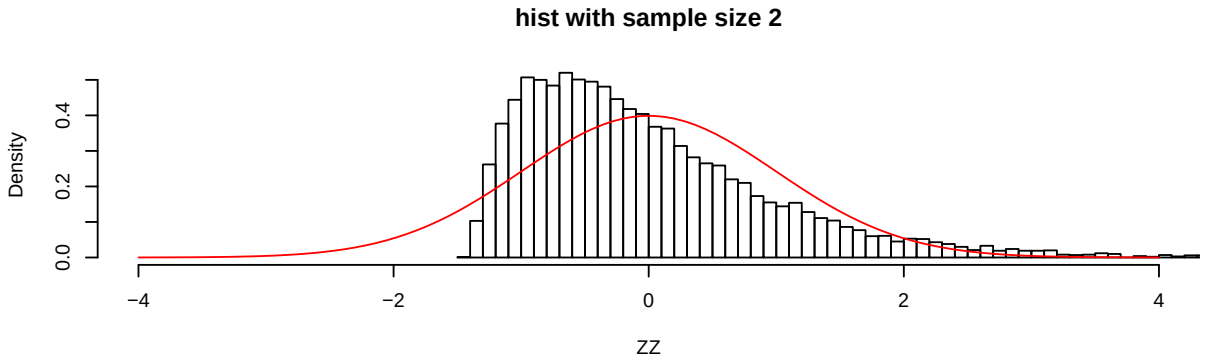
- Lyapunov CLT: i.n.i.d.: $\max_{i \leq n} E\left[|x_i|^3\right] < C < \infty$.

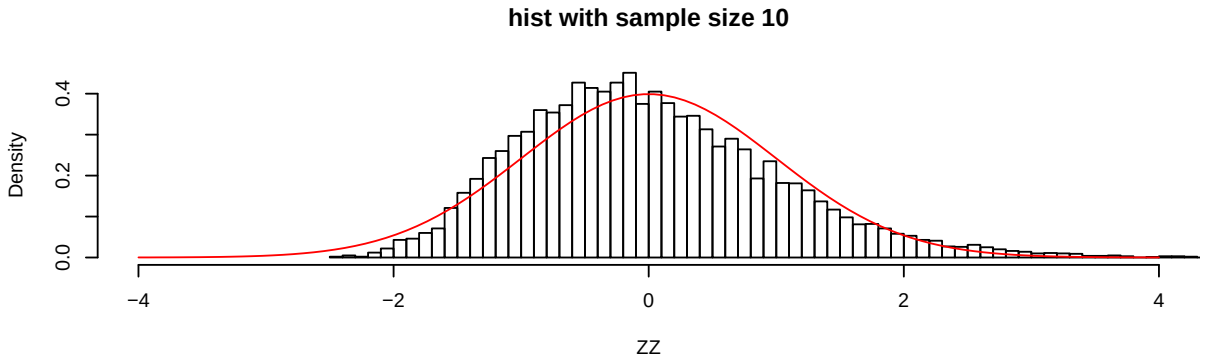
This is a simulated example.

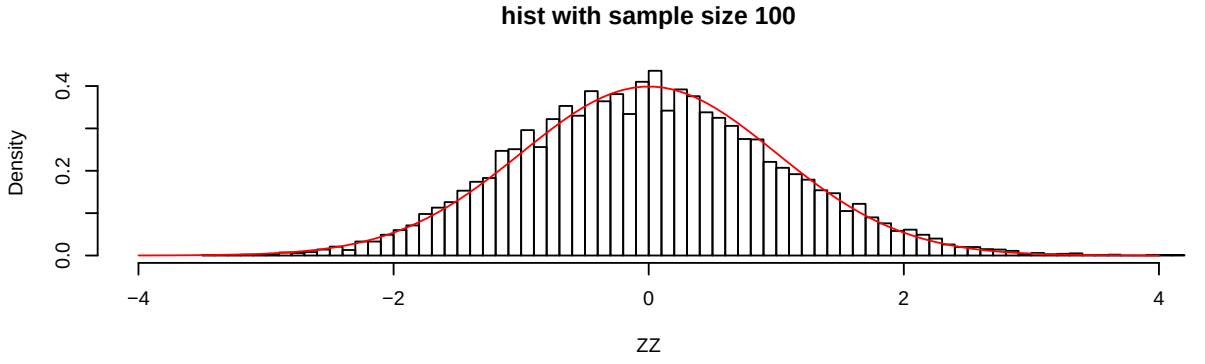
```
Z_fun = function(n, distribution){
  if (distribution == "normal"){
      z = sqrt(n) * mean(rnorm(n))
        } else if (distribution == "chisq2") {
      df = 2;
      x = rchisq(n,2)
      z = sqrt(n) * ( mean(x) - df ) / sqrt(2*df)
      }
  return (z)
}
CLT_plot = function(n, distribution){
  Rep = 10000
  ZZ = rep(0, Rep)
  for (i in 1:Rep) {ZZ[i] = Z_fun(n, distribution)}
```

```
  xbase = seq(-4.0, 4.0, length.out = 100)
  hist( ZZ, breaks = 100, freq = FALSE,
    xlim = c( min(xbase), max(xbase) ),
    main = paste0("hist with sample size ", n) )
  lines(x = xbase, y = dnorm(xbase), col = "red")
  return (ZZ)
}

par(mfrow = c(3,1))
phist = CLT_plot(2, "chisq2")
phist = CLT_plot(10, "chisq2")
phist = CLT_plot(100, "chisq2")
```







5.1.4 Tools for Transformations

In their original forms, LLN deals with the sample mean, and CLT handles the scaled (by $\sqrt{n}$) and/or standardized (by standard deviation) sample mean. However, most of the econometric estimators of interest are functions of sample means. Therefore, we need tools to handle transformations.

- Small op: $x_n = o_p(r_n)$ if $x_n/r_n \xrightarrow{p} 0$.
- Big Op: $x_n = O_p(r_n)$ if for any $\varepsilon > 0$, there exists a $c > 0$ such that $P(|x_n|/r_n > c) < \varepsilon$.
- Continuous mapping theorem 1: If $x_n \xrightarrow{p} a$ and $f(\cdot)$ is continuous at a, then $f(x_n) \xrightarrow{p} f(a)$.
- Continuous mapping theorem 2: If $x_n \xrightarrow{d} x$ and $f(\cdot)$ is continuous almost surely on the support of x, then $f(x_n) \xrightarrow{d} f(x)$.
- Slutsky's Theorem: If $x_n \xrightarrow{d} x$ and $y_n \xrightarrow{p} a$, then
  - $x_n + y_n \xrightarrow{d} x + a$
  - $x_n y_n \xrightarrow{d} ax$
  - $x_n/y_n \xrightarrow{d} x/a$ if $a \neq 0$.
- Delta method: if $\sqrt{n}\left(\widehat{\theta} - \theta_0\right) \xrightarrow{d} N(0, \Omega)$, and $f(\cdot)$ is continuously differentiable at θ_0, then

$$\sqrt{n}\left(f\left(\widehat{\theta}\right) - f(\theta_0)\right) \xrightarrow{d} N\left(0, \frac{\partial f}{\partial \theta'}(\theta_0)\,\Omega\left(\frac{\partial f}{\partial \theta}(\theta_0)\right)'\right).$$

Proof. Proof: take a Taylor expansion of $f\left(\widehat{\theta}\right)$ around $f(\theta_0)$. □

5.2 Asymptotic Properties of OLS

We apply large sample theory to study the OLS estimator $\widehat{\beta} = (X'X)^{-1}X'Y$.

5.2.1 Consistency

We say $\widehat{\beta}$ is *consistent* if $\widehat{\beta} \xrightarrow{p} \beta$ as $n \to \infty$. To verify consistency, we write

$$\widehat{\beta} - \beta = \left(X'X\right)^{-1}X'e = \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1}\frac{1}{n}\sum_{i=1}^{n} x_i e_i. \tag{5.2}$$

The first term

$$\widehat{Q} = \frac{1}{n}\sum_{i=1}^{n} x_i x_i' \xrightarrow{p} Q = E\left[x_i x_i'\right],$$

and the second term

$$\frac{1}{n}\sum_{i=1}^{n} x_i e_i \xrightarrow{p} 0.$$

No matter whether $(y_i, x_i)_{i=1}^n$ is an i.i.d., i.n.i.d., or dependent sample, as long as the convergence in probability holds for the above two expressions and Q is an invertible matrix, we have $\widehat{\beta} - \beta \xrightarrow{p} Q^{-1}0 = 0$ by the continuous mapping theorem. In other words, $\widehat{\beta}$ is a consistent estimator of β.

5.2.2 Asymptotic Normality

In finite sample, $\widehat{\beta}$ is a random variable. We have shown the distribution of $\widehat{\beta}$ under normality in the previous lecture. Without the restrictive normality assumption, how can we characterize the randomness of the OLS estimator?

We know from the previous section that $\hat{\beta} - \beta \xrightarrow{p} 0$ degenerates to a constant. To study its distribution, we must scale it up by a proper multiplier so that in the limit it neither degenerates nor explodes. The suitable scaling factor is $\sqrt{n}$:

$$\sqrt{n}\left(\widehat{\beta} - \beta\right) = \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1} \frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i e_i.$$

Since $E[x_i e_i] = 0$, we apply a CLT to obtain

$$n^{-1/2}\sum_{i=1}^{n} x_i e_i \xrightarrow{d} N(0, \Sigma)$$

where $\Sigma = E\left[x_i x_i' e_i^2\right]$. By the continuous mapping theorem,

$$\sqrt{n}\left(\widehat{\beta} - \beta\right) \xrightarrow{d} Q^{-1} \times N(0, \Sigma) \sim N(0, \Omega)$$

where $\Omega = Q^{-1}\Sigma Q^{-1}$ is called the *asymptotic variance*. This is the *asymptotic normality* of the OLS estimator.

Up to now we have derived the asymptotic distribution of $\widehat{\beta}$. However, to make it feasible, we still have to estimator the asymptotic variance Ω. If $\widehat{\Sigma}$ is a consistent estimator of Σ, then $\widehat{\Omega} = \widehat{Q}^{-1}\widehat{\Sigma}\widehat{Q}^{-1}$ is a consistent estimator of Ω. (Of course, there are other ways to estimate the asymptotic variance.) A feasible version about the distribution of $\widehat{\beta}$ is

$$\widehat{\Omega}^{-1/2}\sqrt{n}\left(\widehat{\beta} - \beta\right) \xrightarrow{d} N(0, I_K).$$

In the special case of homoskedasticity, $\Sigma = E\left[x_i x_i' e_i^2\right] = E\left[x_i x_i' E\left[e_i^2 | X\right]\right] = \sigma^2 E\left[x_i x_i'\right] = \sigma^2 Q$ so that

$$n^{-1/2}\sum_{i=1}^{n} x_i e_i \xrightarrow{d} N\left(0, \sigma^2 Q\right).$$

Again by the continuous mapping theorem,

$$\sqrt{n}\left(\widehat{\beta} - \beta\right) \xrightarrow{d} Q^{-1} \times N\left(0, \sigma^2 Q\right) \sim N\left(0, \sigma^2 Q^{-1}\right).$$

We can estimate the unknown parameter σ^2 by either $\widehat{\sigma}^2 = \widehat{e}'\widehat{e}/(n-K)$ or $\widehat{\sigma}^2 = \widehat{e}'\widehat{e}/n$.

5.2.3 Estimation of the Variance

We show all elements of $\Sigma = E\left[x_i x_i' e_i^2\right]$ is finite. That is, $\|\Sigma\|_\infty := \max_{i,j \le K} |\sigma_{ij}| < \infty$. Let $z_i = x_i e_i$, so $\Sigma = E[z_i z_i']$. Because of the Cachy-Schwarz inequality,

$$\|\Sigma\|_\infty = \max_{k=1,\ldots,K} E\left[z_{ik}^2\right].$$

For each k, $E\left[z_{ik}^2\right] = E\left[x_{ik}^2 e_i^2\right] \le \left(E\left[x_{ik}^4\right] E\left[e_i^4\right]\right)^{1/2}$.

For the estimation of variance, if the error is homoskedastic,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\widehat{e}_i^2 &= \frac{1}{n}\sum_{i=1}^{n}\left(e_i + x_i'\left(\widehat{\beta}-\beta\right)\right)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}e_i^2 + \left(\frac{2}{n}\sum_{i=1}^{n}e_i x_i\right)'\left(\widehat{\beta}-\beta\right) + \frac{1}{n}\sum_{i=1}^{n}e_i^2\left(\widehat{\beta}-\beta\right)' x_i x_i'\left(\widehat{\beta}-\beta\right).\end{aligned}$$

The second term

$$\left(\frac{2}{n}\sum_{i=1}^{n}e_i x_i\right)'\left(\widehat{\beta}-\beta\right) = o_p(1)\,o_p(1) = o_p(1).$$

The third term

$$\left(\widehat{\beta}-\beta\right)\left(\frac{1}{n}\sum_{i=1}^{n}e_i^2 x_i x_i'\right)\left(\widehat{\beta}-\beta\right) = o_p(1)\,O_p(1)\,o_p(1) = o_p(1).$$

As $\frac{1}{n}\sum_{i=1}^{n}\widehat{e}_i^2 = \frac{1}{n}\sum_{i=1}^{n}\widehat{e}_i^2 + o_p(1)$ and $\frac{1}{n}\sum_{i=1}^{n}e_i^2 = \sigma_e^2 + o_p(1)$, we have $\frac{1}{n}\sum_{i=1}^{n}\widehat{e}_i^2 = \sigma_e^2 + o_p(1)$. In other words, $\frac{1}{n}\sum_{i=1}^{n}\widehat{e}_i^2 \xrightarrow{p} \sigma_e^2$.

For general heteroskedasticity,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\widehat{e}_i^2 &= \frac{1}{n}\sum_{i=1}^{n}x_i x_i'\left(e_i + x_i'\left(\widehat{\beta}-\beta\right)\right)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}x_i x_i' e_i^2 + \frac{1}{n}\sum_{i=1}^{n}x_i x_i e_i x_i'\left(\widehat{\beta}-\beta\right) + \frac{1}{n}\sum_{i=1}^{n}x_i x_i'\left(\left(\widehat{\beta}-\beta\right)' x_i\right)^2.\end{aligned}$$

The third term is bounded by

$$\begin{aligned}&\operatorname{trace}\left(\frac{1}{n}\sum_{i=1}^{n}x_i x_i'\left(\left(\widehat{\beta}-\beta\right)' x_i\right)^2\right) \\ &\le K\max_k \frac{1}{n}\sum_{i=1}^{n}x_{ik}^2\left[\left(\widehat{\beta}-\beta\right)' x_i\right]^2 \\ &\le K\left\|\widehat{\beta}-\beta\right\|_2^2 \max_k \frac{1}{n}\sum_{i=1}^{n}x_{ik}^2\left\|x_i\right\|_2^2 \\ &\le K\left\|\widehat{\beta}-\beta\right\|_2^2 \frac{1}{n}\sum_{i=1}^{n}\left\|x_i\right\|_2^2\left\|x_i\right\|_2^2 \\ &= K\left\|\widehat{\beta}-\beta\right\|_2^2 \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{k=1}^{K}x_{ik}^2\right)^2 \\ &\le K\left\|\widehat{\beta}-\beta\right\|_2^2 K\sum_{k=1}^{K}\frac{1}{n}\sum_{i=1}^{n}x_{ik}^4 = o_p(1)\,O_p(1) = o_p(1).\end{aligned}$$

where the third inequality follows by $(a_1+\cdots+a_K)^2 \le K\left(a_1^2+\cdots+a_K^2\right)$. The second term is

bounded by

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^{n} x_{ik}x_{ik'}e_i x_i'\left(\widehat{\beta}-\beta\right)\right| \\
&\leq \max_k \left|\widehat{\beta}_k-\beta_k\right| K \max_{k,k',k''}\left|\frac{1}{n}\sum_{i=1}^{n} e_i x_{ik}x_{ik'}x_{ik''}\right| \\
&\leq \left\|\widehat{\beta}-\beta\right\|_2 \left(\frac{1}{n}\sum_{i=1}^{n} e_i^4\right)^{1/4} K \max_{k,k',k''}\left(\frac{1}{n}\sum_{i=1}^{n}\left(x_{ik}x_{ik'}x_{ik''}\right)^{4/3}\right)^{3/4} \\
&\leq \left\|\widehat{\beta}-\beta\right\|_2 K \max_k \left(\frac{1}{n}\sum_{i=1}^{n} x_{ik}^4\right)^{3/4} = o_p(1)\, O_p(1)
\end{aligned}$$

where the second and the third inequality hold by the Holder's inequality.

Zhentao Shi. October 6, 2020

Chapter 6

Hypothesis Testing

Notation: $\mathbf{X}$ denotes a random variable or random vector. $\mathbf{x}$ is its realization.

6.1 Hypothesis Testing

- A *hypothesis* is a statement about the parameter space Θ.
- The *null hypothesis* Θ_0 is a subset of Θ of interest, ideally suggested by scientific theory.
- The *alternative hypothesis* $\Theta_1 = \Theta \backslash \Theta_0$ is the complement of Θ_0.
- *Hypothesis testing* is a decision, based on the observed evidence, to accept the null hypothesis or to reject it.
- If Θ_0 is a singleton, we call it a *simple hypothesis*; otherwise we call it a *composite hypothesis*. For example, if the parameter space $\Theta = \mathbb{R}$, then $\Theta_0 = \{0\}$ (or equivalently $\theta_0 = 0$) is a simple hypothesis, while $\Theta_0 = (-\infty, 0]$ (or equivalently $\theta_0 \leq 0$) is a composite hypothesis.
- A *test function* is a mapping
$$\phi_\theta : \mathcal{X}^n \mapsto \{0, 1\},$$
where $\mathcal{X}$ is the sample space. The null hypothesis is accepted if $\phi_\theta(\mathbf{X} = \mathbf{x}) = 0$, or rejected if $\phi_\theta(\mathbf{X} = \mathbf{x}) = 1$. Notice that the test function depends on the hypothesized parameter value θ.
- The *acceptance region* is defined as $A_\phi = \{\mathbf{x} \in \mathcal{X}^n : \phi_\theta(\mathbf{x}) = 0\}$, and the *rejection region* is $R_\phi = \{\mathbf{x} \in \mathcal{X}^n : \phi_\theta(\mathbf{x}) = 1\}$.
- The *power function* of a test ϕ_θ is
$$\beta_\phi(\theta) = P(\{\phi_\theta(\mathbf{X}) = 1\}) = E(\phi_\theta(\mathbf{X})).$$
The power function measures, at a given point θ, the probability that the test function rejects the null.
- The *power* of a test for some $\theta \in \Theta_1$ is the value of $\beta_\phi(\theta)$. The *size* of the test is $\sup_{\theta \in \Theta_0} \beta_\phi(\theta)$. Notice that the definition of power depends on a θ in the alternative hypothesis Θ_1, whereas that of size is independent of θ due to the supremum over the set of null Θ_0.

- The *level* of a test is any value $\alpha \in (0,1)$ such that $\alpha \geq \sup_{\theta \in \Theta_0} \beta_\phi(\theta)$, which is often used when it is difficult to attain the exact supremum. A test of size α is also of level α or bigger; while a test of level α must have size smaller or equal to α.

```
| decision     |  reject $H_{1}$  | reject $H_{0}$
|--------------|------------------| ---------------
| $H_{0}$ true |     correct      | Type I error
| $H_{0}$ false| Type II error    |   correct
```

- size = P(reject H_0 when H_0 is true)
- power = P(reject H_0 when H_0 is false)
- The *probability of committing Type I error* is $\beta_\phi(\theta)$ for some $\theta \in \Theta_0$.
- The *probability of committing Type II error* is $1 - \beta_\phi(\theta)$ for $\theta \in \Theta_1$.

The philosophy on hypothesis testing has been debated for centuries. At present the prevailing framework in statistics textbooks is the frequentist perspective. A frequentist views the parameter as a fixed constant, and they keep a conservative attitude about the Type I error. Only if overwhelming evidence is demonstrated shall a researcher reject the null. Under the philosophy of protecting the null hypothesis, a desirable test should have a small level. Conventionally we take $\alpha = 0.01$, 0.05 or 0.1. There can be many tests of correct size.

Example A trivial test function, $\phi_\theta(\mathbf{X}) = 1\{0 \leq U \leq \alpha\}$ for all $\theta \in \Theta$, where U is a random variable from a uniform distribution on $[0,1]$, has correct size α but no power. We say a test is *unbiased* if $\beta_\phi(\theta) > \alpha$ for all $\theta \in \Theta_1$. The trivial test mentioned here is not an unbiased one. On the other extreme, the trivial test function $\phi_\theta(\mathbf{X}) = 1$ for all θ has the biggest power but incorrect size.

Usually, we design a test by proposing a test statistic $T_n : \mathcal{X}^n \times \Theta \mapsto \mathbb{R}^+$ and a critical value $c_{1-\alpha}$. Given T_n and $c_{1-\alpha}$, we write the test function as

$$\phi_\theta(\mathbf{X}) = 1\{T_n(\mathbf{X}, \theta) > c_{1-\alpha}\}.$$

To ensure such a $\phi(\mathbf{x})$ has correct size, we need to figure out the distribution of T_n under the null hypothesis (called the *null distribution*), and choose a critical value $c_{1-\alpha}$ according to the null distribution and the desirable size or level α.

The concept of *level* is useful if we do not have sufficient information to derive the exact size of a test.

Example If $(X_{1i}, X_{2i})_{i=1}^n$ are randomly drawn from some unknown joint distribution, but we know the marginal distribution is $X_{ji} \sim N(\theta_j, 1)$, for $j = 1, 2$. In order to test the joint hypothesis $\theta_1 = \theta_2 = 0$, we can construct a test function

$$\phi_{\theta_1=\theta_2=0}(\mathbf{X}_1, \mathbf{X}_2) = 1\left\{\left\{\sqrt{n}\left|\overline{X}_1\right| \geq z_{1-\alpha/4}\right\} \cup \left\{\sqrt{n}\left|\overline{X}_2\right| \geq z_{1-\alpha/4}\right\}\right\},$$

where $z_{1-\alpha/4}$ is the $(1-\alpha/4)$-th quantile of the standard normal distribution. The level of this test is

$$\begin{aligned} P(\phi_{\theta_1=\theta_2=0}(\mathbf{X}_1, \mathbf{X}_2)) &\leq P\left(\sqrt{n}\left|\overline{X}_1\right| \geq z_{1-\alpha/4}\right) + P\left(\sqrt{n}\left|\overline{X}_2\right| \geq z_{1-\alpha/4}\right) \\ &= \alpha/2 + \alpha/2 = \alpha. \end{aligned}$$

where the inequality follows by the *Bonferroni inequality*

$$P(A \cup B) \leq P(A) + P(B).$$

(The seemingly trivial Bonferroni inequality is useful in many proofs of probability results.) Therefore, the level of $\phi(\mathbf{X}_1, \mathbf{X}_2)$ is α, but the exact size is unknown without the knowledge of the joint distribution. (Even if we know the correlation of X_{1i} and X_{2i}, putting two marginally normal distributions together does not make a jointly normal vector in general.)

Denote the class of test functions of level α as $\Psi_\alpha = \left\{\phi : \sup_{\theta \in \Theta_0} \beta_\phi(\theta) \leq \alpha\right\}$. A *uniformly most powerful test* $\phi^* \in \Psi_\alpha$ is a test function such that, for every $\phi \in \Psi_\alpha$,

$$\beta_{\phi^*}(\theta) \geq \beta_\phi(\theta)$$

uniformly over $\theta \in \Theta_1$.

Example Suppose a random sample of size 6 is generated from

$$(X_1, \ldots, X_6) \sim \text{i.i.d.} N(\theta, 1),$$

where θ is unknown. We want to infer the population mean of the normal distribution. The null hypothesis is H_0: $\theta \leq 0$ and the alternative is H_1: $\theta > 0$. All tests in

$$\Psi = \left\{1\left\{\bar{X} \geq c/\sqrt{6}\right\} : c \geq 1.64\right\}$$

has the correct level. Since $\bar{X} = N(\theta, 1/6)$, the power function for those in Ψ is

$$\begin{aligned}
\beta_\phi(\theta) &= P\left(\bar{X} \geq \frac{c}{\sqrt{6}}\right) \\
&= P\left(\frac{\bar{X} - \theta}{1/\sqrt{6}} \geq \frac{\frac{c}{\sqrt{6}} - \theta}{1/\sqrt{6}}\right) \\
&= P\left(N \geq c - \sqrt{6}\theta\right) \\
&= 1 - \Phi\left(c - \sqrt{6}\theta\right)
\end{aligned}$$

where $N = \frac{\bar{X} - \theta}{1/\sqrt{6}}$ follows the standard normal, and Φ is the cdf of standard normal. It is clear that $\beta_\phi(\theta)$ is monotonically decreasing in c. Thus the test function

$$\phi_{\theta=0}(\mathbf{X}) = 1\left\{\bar{X} \geq 1.64/\sqrt{6}\right\}$$

is the most powerful test in Ψ, as $c = 1.64$ is the lower bound that Ψ allows.

Another commonly used indicator in hypothesis testing is p-value:

$$\sup_{\theta \in \Theta_0} P(T_n(\mathbf{x}, \theta) \leq T_n(\mathbf{X}, \theta)).$$

In the above expression, $T_n(\mathbf{x}, \theta)$ is the realized value of the test statistic T_n, while $T_n(\mathbf{X}, \theta)$ is the random variable generated by $\mathbf{X}$ under the null $\theta \in \Theta_0$. The interpretation of the p-value is tricky. p-value is the probability that we observe $T_n(\mathbf{X}, \theta)$ being greater than the realized $T_n(\mathbf{x}, \theta)$ if the null hypothesis is true. p-value is *not* the probability that the null hypothesis is true. Under the

frequentist perspective, the null hypothesis is either true or false, with certainty. The randomness of a test comes only from sampling, not from the hypothesis. It measures whether the dataset is consistent with the null hypothesis, or whether the evidence from the data is compatible with the null hypothesis. p-value is closely related to the corresponding test. When p-value is smaller than the specified test size α, the test rejects the null.

So far we have been talking about hypothesis testing in finite sample. The discussion and terminologies can be carried over to the asymptotic world when $n \to \infty$. If we denote the power function as $\beta_{n,\phi}(\theta)$, in which we make its dependence on the sample size n explicit, the test is of asymptotic size α if $\limsup_{n\to\infty} \beta_{n,\phi}(\theta) \leq \alpha$ for all $\theta \in \Theta_0$. A test is *consistent* if $\beta_{n,\phi}(\theta) \to 1$ for all $\theta \in \Theta_1$.

6.2 Confidence Interval

An *interval estimate* is a function $C : \mathcal{X}^n \mapsto \{\Theta' : \Theta' \subseteq \Theta\}$ that maps a point in the sample space to a subset of the parameter space. The *coverage probability* of an *interval estimator* $C(\mathbf{X})$ is defined as $P_\theta(\theta \in C(\mathbf{X}))$. When θ is of one dimension, we usually call the interval estimator *confidence interval*. When θ is of multiple dimensions, we call the it *confidence region* and it of course includes the one-dimensional θ as a special case. The coverage probability is the frequency that the interval estimator captures the true parameter that generates the sample (From the frequentist perspective, the parameter is fixed while the confidence region is random). It is *not* the probability that θ is inside the given confidence interval (From the Bayesian perspective, the parameter is random while the confidence region is fixed conditional on $\mathbf{X}$.)

Exercise: Suppose a random sample of size 6 is generated from

$$(X_1, \ldots, X_6) \sim \text{i.i.d. } N(\theta, 1).$$

Find the coverage probability of the random interval is

$$\left[\bar{X} - 1.96/\sqrt{6}, \bar{X} + 1.96/\sqrt{6}\right].$$

Hypothesis testing and confidence region are closely related. Sometimes it is difficult to directly construct the confidence region, but easy to test a hypothesis. One way to construct confidence region is by *inverting a test*. Suppose ϕ_θ is a test of size α. If $C(\mathbf{X})$ is constructed as

$$C(\mathbf{X}) = \{\theta \in \Theta : \phi_\theta(\mathbf{X}) = 0\}.$$

For any $\theta \in \Theta_0$, its coverage probability

$$P(\theta \in C(\mathbf{X})) = P(\{\phi_\theta(\mathbf{X}) = 0\}) = 1 - P(\{\phi_\theta(\mathbf{X}) = 1\}) = 1 - \beta_\phi(\theta) \geq 1 - \alpha$$

where the last inequality follows as $\beta_\phi(\theta) \leq \alpha$. If Θ_0 is a singleton, the equality holds.

6.3 Bayesian Credible Set

The Bayesian framework offers a coherent and natural language for statistical decision. However, the major criticism against Bayesian statistics is the arbitrariness of the choice of the prior.

In the Bayesian framework, both the data $\mathbf{X}_n$ and the parameter θ are random variables. Before she observes the data, she holds a *prior distribution* π about θ. After observing the data,

she updates the prior distribution to a *posterior distribution* $p(\theta|\mathbf{X}_n)$. The *Bayes Theorem* connects the prior and the posterior as

$$p(\theta|\mathbf{X}_n) \propto f(\mathbf{X}_n|\theta)\pi(\theta)$$

where $f(\mathbf{X}_n|\theta)$ is the likelihood function.

Here is a classical example to illustrate the Bayesian approach of statistical inference. Suppose we have an iid sample $\mathbf{X}_n = (X_1, \ldots, X_n)$ drawn from a normal distribution with unknown θ and known σ. If a researcher's prior distribution $\theta \sim N(\theta_0, \sigma_0^2)$, her posterior distribution is, by some routine calculation, also a normal distribution

$$p(\theta|\mathbf{x}) \sim N\left(\tilde{\theta}, \tilde{\sigma}^2\right),$$

where $\tilde{\theta} = \frac{\sigma^2}{n\sigma_0^2+\sigma^2}\theta_0 + \frac{n\sigma_0^2}{n\sigma_0^2+\sigma^2}\bar{x}$ and $\tilde{\sigma}^2 = \frac{\sigma_0^2\sigma^2}{n\sigma_0^2+\sigma^2}$. Thus the Bayesian credible set is

$$\left(\tilde{\theta} - z_{1-\alpha/2}\cdot\tilde{\sigma}, \tilde{\theta} + z_{1-\alpha/2}\cdot\tilde{\sigma}\right).$$

This posterior distribution depends on θ_0 and σ_0^2 from the prior. When the sample size is sufficiently large the posterior can be approximated by $N(\bar{x}, \sigma^2/n)$, where the prior information is overwhelmed by the information accumulated from the data.

In contrast, a frequentist will estimate $\hat{\theta} = \bar{x} \sim N(\theta, \sigma^2/n)$. Her confidence interval is

$$\left(\bar{x} - z_{1-\alpha/2}\cdot\sigma/\sqrt{n}, \bar{x} - z_{1-\alpha/2}\cdot\sigma/\sqrt{n}\right).$$

The Bayesian credible set and the frequentist confidence interval are different for any finite n, but they coincide when hen $n \to \infty$.

6.4 Application in OLS

6.4.1 Wald Test

Suppose the OLS estimator $\widehat{\beta}$ is asymptotic normal, i.e.

$$\sqrt{n}\left(\widehat{\beta} - \beta\right) \xrightarrow{d} N\left(0, \Omega\right)$$

where Ω is a $K \times K$ positive definite covariance matrix and R is a $q \times K$ constant matrix, then $R\sqrt{n}\left(\widehat{\beta} - \beta\right) \xrightarrow{d} N\left(0, R\Omega R'\right)$. Moreover, if $\text{rank}\left(R\right) = q$, then

$$n\left(\widehat{\beta} - \beta\right)' R'\left(R\Omega R'\right)^{-1} R\left(\widehat{\beta} - \beta\right) \xrightarrow{d} \chi_q^2.$$

Now we intend to test the null hypothesis $R\beta = r$. Under the null, the Wald statistic

$$W_n = n\left(R\widehat{\beta} - r\right)'\left(R\widehat{\Omega}R'\right)^{-1}\left(R\widehat{\beta} - r\right) \xrightarrow{d} \chi_q^2$$

where $\widehat{\Omega}$ is a consistent estimator of Ω.

Example (Single test) In a linear regression

$$y = x_i'\beta + e_i = \sum_{k=1}^{5}\beta_k x_{ik} + e_i.$$

$$E\left[e_i x_i\right] = \mathbf{0}_5,$$

where y is wage and

$$x = \left(\text{edu}, \text{age}, \text{experience}, \text{experience}^2, 1\right)'.$$

To test whether *education* affects *wage*, we specify the null hypothesis $\beta_1 = 0$. Let $R = (1, 0, 0, 0, 0)$ and $r = 0$.

$$\sqrt{n}\widehat{\beta}_1 = \sqrt{n}\left(\widehat{\beta}_1 - \beta_1\right) = \sqrt{n}R\left(\widehat{\beta} - \beta\right) \xrightarrow{d} N\left(0, R\Omega R'\right) \sim N\left(0, \Omega_{11}\right), \tag{6.1}$$

where Ω_{11} is the $(1,1)$ (scalar) element of Ω. Under $H_0 : R\beta = (1,0,0,0,0)\begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{pmatrix} = \beta_1 = 0$, we have

$$\sqrt{n}R\left(\widehat{\beta} - \beta\right) = \sqrt{n}\widehat{\beta}_1 \xrightarrow{d} N\left(0, \Omega_{11}\right)$$

Therefore,

$$\sqrt{n}\frac{\widehat{\beta}_1}{\widehat{\Omega}_{11}^{1/2}} = \sqrt{\frac{\Omega_{11}}{\widehat{\Omega}_{11}}}\sqrt{n}\frac{\widehat{\beta}_1}{\sqrt{\Omega_{11}}}$$

If $\widehat{\Omega} \xrightarrow{p} \Omega$, then $\left(\Omega_{11}/\widehat{\Omega}_{11}\right)^{1/2} \xrightarrow{p} 1$ by the continuous mapping theorem. As $\sqrt{n}\widehat{\beta}_1/\Omega_{11}^{1/2} \xrightarrow{d} N(0,1)$, we conclude $\sqrt{n}\widehat{\beta}_1/\widehat{\Omega}_{11}^{1/2} \xrightarrow{d} N\left(0,1\right)$.

The above example is a test about a single coefficient, and the test statistic is essentially a t-statistic. The following example gives a test about a joint hypothesis.

Example (Joint test) We want to simultaneously test $\beta_1 = 1$ and $\beta_3 + \beta_4 = 2$ in the above example. The null hypothesis can be expressed in the general form $R\beta = r$, where the restriction matrix R is

$$R = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \end{pmatrix}$$

and $r = (1, 2)'$. Once we figure out R, it is routine to construct the test.

These two examples are linear restrictions. In order to test a nonlinear regression, we need the so-called *delta method*.

Delta method If $\sqrt{n}\left(\widehat{\theta} - \theta_0\right) \xrightarrow{d} N\left(0, \Omega_{K\times K}\right)$, and $f : \mathbb{R}^K \mapsto \mathbb{R}^q$ is a continuously differentiable function for some $q \leq K$, then

$$\sqrt{n}\left(f\left(\widehat{\theta}\right) - f\left(\theta_0\right)\right) \xrightarrow{d} N\left(0, \frac{\partial f}{\partial \theta}\left(\theta_0\right)\Omega\frac{\partial f}{\partial \theta}\left(\theta_0\right)'\right).$$

This result can be easily shown by a mean-value expansion

$$f(\hat{\theta}) - f(\theta_0) = \frac{\partial f(\tilde{\theta})}{\partial \theta}(\hat{\theta} - \theta_0)$$

where $\tilde{\theta}$ lies on the line segment connecting $\hat{\theta}$ and θ_0. Multiply both sides by $\sqrt{n}$ and notice $\tilde{\theta} \xrightarrow{p} \theta_0$, by Slutsky theorem we have $\sqrt{n}(f(\hat{\theta}) - f(\theta_0)) \xrightarrow{d} \frac{\partial f}{\partial \theta}\left(\theta_0\right) N(0, \Omega)$.

In the example of linear regression, the optimal experience level can be found by setting to zero the first order condition with respective to experience, $\beta_3 + 2\beta_4 \text{experience}^* = 0$. We test the hypothesis that the optimal experience level is 20 years; in other words,

$$\text{experience}^* = -\frac{\beta_3}{2\beta_4} = 20.$$

This is a nonlinear hypothesis. If $q \leq K$ where q is the number of restrictions, we have

$$n\left(f\left(\widehat{\theta}\right) - f\left(\theta_0\right)\right)' \left(\frac{\partial f}{\partial \theta}\left(\theta_0\right) \Omega \frac{\partial f}{\partial \theta}\left(\theta_0\right)'\right)^{-1} \left(f\left(\widehat{\theta}\right) - f\left(\theta_0\right)\right) \xrightarrow{d} \chi_q^2,$$

where in this example, $\theta = \beta$, $f(\beta) = -\beta_3/(2\beta_4)$. The gradient

$$\frac{\partial f}{\partial \beta}(\beta) = \left(0, 0, -\frac{1}{2\beta_4}, \frac{\beta_3}{2\beta_4^2}, 0\right)$$

Since $\widehat{\beta} \xrightarrow{p} \beta_0$, by the continuous mapping theorem, if $\beta_{0,4} \neq 0$, we have $\frac{\partial}{\partial \beta} f\left(\widehat{\beta}\right) \xrightarrow{p} \frac{\partial}{\partial \beta} f\left(\beta_0\right)$. Therefore, the (nonlinear) Wald test is

$$W_n = n\left(f\left(\widehat{\beta}\right) - 20\right)' \left(\frac{\partial f}{\partial \beta}\left(\widehat{\beta}\right) \widehat{\Omega} \frac{\partial f}{\partial \beta}\left(\widehat{\beta}\right)'\right)^{-1} \left(f\left(\widehat{\beta}\right) - 20\right) \xrightarrow{d} \chi_1^2.$$

This is a valid test with correct asymptotic size.

However, we can equivalently state the null hypothesis as $\beta_3 + 40\beta_4 = 0$ and we can construct a Wald statistic accordingly. Asymptotically equivalent though, in general a linear hypothesis is preferred to a nonlinear one, due to the approximation error in the delta method under the null and more importantly the invalidity of the Taylor expansion under the alternative. It also highlights the problem of Wald test being *variant* to re-parametrization.

6.4.2 Lagrangian Multiplier Test*

Restricted least square

$$\min_{\beta} (y - X\beta)'(y - X\beta) \text{ s.t. } R\beta = r.$$

Turn it into an unrestricted problem

$$L(\beta, \lambda) = \frac{1}{2n}(y - X\beta)'(y - X\beta) + \lambda'(R\beta - r).$$

The first-order condition

$$\begin{aligned}
\frac{\partial}{\partial \beta} L &= -\frac{1}{n} X'\left(y - X\tilde{\beta}\right) + \tilde{\lambda} R = -\frac{1}{n} X'e + \frac{1}{n} X'X\left(\tilde{\beta} - \beta^*\right) + R'\tilde{\lambda} = 0. \\
\frac{\partial}{\partial \lambda} L &= R\tilde{\beta} - r = R\left(\tilde{\beta} - \beta^*\right) = 0
\end{aligned}$$

Combine these two equations into a linear system,

$$\begin{pmatrix} \widehat{Q} & R' \\ R & 0 \end{pmatrix} \begin{pmatrix} \tilde{\beta} - \beta^* \\ \tilde{\lambda} \end{pmatrix} = \begin{pmatrix} \frac{1}{n} X'e \\ 0 \end{pmatrix},$$

where $\hat{Q} = X'X/n$.

Thus we can explicitly express the estimator as

$$\begin{aligned}\begin{pmatrix}\tilde{\beta}-\beta^* \\ \tilde{\lambda}\end{pmatrix} &= \begin{pmatrix}\widehat{Q} & R' \\ R & 0\end{pmatrix}^{-1}\begin{pmatrix}\frac{1}{n}X'e \\ 0\end{pmatrix} \\ &= \begin{pmatrix}\widehat{Q}^{-1}-\widehat{Q}^{-1}R'\left(R\widehat{Q}^{-1}R'\right)^{-1}R\widehat{Q}^{-1} & \widehat{Q}^{-1}R'\left(R\widehat{Q}^{-1}R'\right)^{-1} \\ \left(R\widehat{Q}^{-1}R'\right)^{-1}R\widehat{Q}^{-1} & (R'Q^{-1}R)^{-1}\end{pmatrix}\begin{pmatrix}\frac{1}{n}X'e \\ 0\end{pmatrix}.\end{aligned}$$

We conclude that

$$\sqrt{n}\tilde{\lambda} = \left(R\widehat{Q}^{-1}R'\right)^{-1}R\widehat{Q}^{-1}\frac{1}{\sqrt{n}}X'e \xrightarrow{d} N\left(0,\left(RQ^{-1}R'\right)^{-1}RQ^{-1}\Omega Q^{-1}R'\left(RQ^{-1}R'\right)^{-1}\right).$$

Let $W = \left(RQ^{-1}R'\right)^{-1}RQ^{-1}\Omega Q^{-1}R'\left(RQ^{-1}R'\right)^{-1}$, we have

$$n\tilde{\lambda}'W^{-1}\tilde{\lambda} \xrightarrow{d} \chi^2_q.$$

If homoskedastic, then $W = \sigma^2\left(RQ^{-1}R'\right)^{-1}RQ^{-1}QQ^{-1}R'\left(RQ^{-1}R'\right)^{-1} = \sigma^2\left(RQ^{-1}R'\right)^{-1}$. Replace W with the estimated $\hat{W}$,

$$\begin{aligned}\frac{n\tilde{\lambda}'R\hat{Q}^{-1}R'\tilde{\lambda}}{\hat{\sigma}^2} &= \frac{1}{n\hat{\sigma}^2}\left(y-X\tilde{\beta}\right)'X\hat{Q}^{-1}R'(R\hat{Q}^{-1}R')^{-1}R\hat{Q}^{-1}X'\left(y-X\tilde{\beta}\right) \\ &= \frac{1}{n\hat{\sigma}^2}\left(y-X\tilde{\beta}\right)'P_{X\hat{Q}^{-1}R'}\left(y-X\tilde{\beta}\right).\end{aligned}$$

6.4.3 Likelihood-Ratio test*

For likelihood ratio test, the starting point can be a criterion function $L(\beta) = (y-X\beta)'(y-X\beta)$. It does not have to be the likelihood function.

$$\begin{aligned}L\left(\tilde{\beta}\right) - L\left(\widehat{\beta}\right) &= \frac{\partial L}{\partial\beta}\left(\widehat{\beta}\right) + \frac{1}{2}\left(\tilde{\beta}-\widehat{\beta}\right)'\frac{\partial L}{\partial\beta\partial\beta}\left(\dot{\beta}\right)\left(\tilde{\beta}-\widehat{\beta}\right) \\ &= 0 + \frac{1}{2}\left(\tilde{\beta}-\widehat{\beta}\right)'\widehat{Q}\left(\tilde{\beta}-\widehat{\beta}\right).\end{aligned}$$

From the derivation of LM test, we have

$$\begin{aligned}\sqrt{n}\left(\tilde{\beta}-\beta^*\right) &= \left(\widehat{Q}^{-1}-\widehat{Q}^{-1}R'\left(R\widehat{Q}^{-1}R'\right)^{-1}R\widehat{Q}^{-1}\right)\frac{1}{\sqrt{n}}X'e \\ &= \frac{1}{\sqrt{n}}\left(X'X\right)X'e - \widehat{Q}^{-1}R'\left(R\widehat{Q}^{-1}R'\right)^{-1}R\widehat{Q}^{-1}\frac{1}{\sqrt{n}}X'e \\ &= \sqrt{n}\left(\widehat{\beta}-\beta^*\right) - \widehat{Q}^{-1}R'\left(R\widehat{Q}^{-1}R'\right)^{-1}R\widehat{Q}^{-1}\frac{1}{\sqrt{n}}X'e\end{aligned}$$

Therefore

$$\sqrt{n}\left(\tilde{\beta}-\widehat{\beta}\right) = -\widehat{Q}^{-1}R'\left(R\widehat{Q}^{-1}R'\right)^{-1}R\widehat{Q}^{-1}\frac{1}{\sqrt{n}}X'e$$

and

$$\begin{aligned}&n\left(\tilde{\beta}-\widehat{\beta}\right)'\widehat{Q}\left(\tilde{\beta}-\widehat{\beta}\right) \\ &= \frac{1}{\sqrt{n}}e'X\widehat{Q}^{-1}R'\left(R\widehat{Q}^{-1}R'\right)^{-1}R\widehat{Q}^{-1}\widehat{Q}\widehat{Q}^{-1}R'\left(R\widehat{Q}^{-1}R'\right)^{-1}R\widehat{Q}^{-1}\frac{1}{\sqrt{n}}X'e \\ &= \frac{1}{\sqrt{n}}e'X\widehat{Q}^{-1}R'\left(R\widehat{Q}^{-1}R'\right)^{-1}R\widehat{Q}^{-1}\frac{1}{\sqrt{n}}X'e\end{aligned}$$

In general, it is a quadratic form of normal distributions. If homoskedastic, then

$$\left(R\widehat{Q}^{-1}R'\right)^{-1/2} R\widehat{Q}^{-1}\frac{1}{\sqrt{n}}X'e$$

has variance

$$\sigma^2\left(RQ^{-1}R'\right)^{-1/2} RQ^{-1}QQ^{-1}R'\left(RQ^{-1}R'\right)^{-1/2} = \sigma^2 I_q.$$

We can view the optimization of the log-likelihood as a two-step optimization with the inner step $\sigma = \sigma(\beta)$. By the envelop theorem, when we take derivative with respect to β, we can ignore the indirect effect of $\partial\sigma(\beta)/\partial\beta$.

Zhentao Shi. October 6, 2020

Chapter 7

Panel Data

7.1 Panel Data

Economists mostly work with observational data. The data generation process is out of the researchers' control. If we only have a cross sectional dataset at hand, it is difficult to control heterogeneity among the individuals. On the other hand, panel data offers a chance to control heterogeneity of some particular forms.

A panel dataset tracks the same individuals across time $t = 1, \ldots, T$. We assume the observations are independent across $i = 1, \ldots, n$, while we allow some form of dependence within a group across $t = 1, \ldots, T$ for the same i. We maintain the linear equation

$$y_{it} = \beta_1 + x_{it}\beta_2 + u_{it},\ i = 1, \ldots, n; t = 1, \ldots, T \tag{7.1}$$

where $u_{it} = \alpha_i + \epsilon_{it}$ is called the *composite error*. Note that α_i is the time-invariant unobserved heterogeneity, while ϵ_{it} varies across individuals and time periods.

The most important techniques of panel data estimation are the fixed effect regression and the random effect regression. The asymptotic distributions of both estimators can be derived from knowledge about the OLS regression. In this sense, panel data estimation becomes applied examples of the theory that we have covered in this course. It highlights the fundamental role of theory in econometrics.

7.1.1 Fixed Effect

OLS is consistent for the linear projection model. Since α_i is unobservable, it is absorbed into the composite error $u_{it} = \alpha_i + \epsilon_{it}$. If $\operatorname{cov}(\alpha_i, x_{it}) = 0$, the OLS is consistent; otherwise the consistency breaks down. The fixed effect model allows α_i and x_{it} to be arbitrarily correlated. The trick to regain consistency is to eliminate $\alpha_i, i = 1, \ldots, n$. The rest of this section develops the consistency and asymptotic distribution of the *within estimator*, the default fixed-effect (FE) estimator. The within estimator transforms the data by subtracting all the observable variables by the corresponding group means. Averaging the T equations of the original regression for the same i, we have

$$\overline{y}_i = \beta_1 + \overline{x}_i\beta_2 + \overline{u}_{it} = \beta_1 + \overline{x}_i\beta_2 + \alpha_i + \overline{\epsilon}_{it}. \tag{7.2}$$

where $\overline{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{it}$. Subtracting the averaged equation from the original equation gives

$$\tilde{y}_{it} = \tilde{x}_{it}\beta_2 + \tilde{\epsilon}_{it} \tag{7.3}$$

where $\tilde{y}_{it} = y_{it} - \overline{y}_i$. We then run OLS with the demeaned data, and obtain the within estimator

$$\widehat{\beta}_2^{FE} = \left(\tilde{X}'\tilde{X}\right)^{-1} \tilde{X}'\tilde{y},$$

where $\tilde{y} = (y_{it})_{i,t}$ stacks all the nT observations into a vector, and similarly defined is $\tilde{X}$ as an $nT \times K$ matrix, where K is the dimension of β_2.

We know that OLS would be consistent if $E\left[\tilde{\epsilon}_{it}|\tilde{x}_{it}\right] = 0$. Below we provide a sufficient condition, which is often called *strict exogeneity*.

Assumption FE.1 $E\left[\epsilon_{it}|\alpha_i, \mathbf{x}_i\right] = 0$ where $\mathbf{x}_i = (x_{i1}, \ldots, x_{iT})$.

Its strictness is relative to the contemporary exogeneity $E\left[\epsilon_{it}|\alpha_i, x_{it}\right] = 0$. FE.1 is more restrictive as it assumes that the error ϵ_{it} is mean independent of the past, present and future explanatory variables.

When we talk about the consistency in panel data, typically we are considering $n \to \infty$ while T stays fixed. This asymptotic framework is appropriate for panel datasets with many individuals but only a few time periods.

Proposition If FE.1 is satisfied, then $\widehat{\beta}_2^{FE}$ is consistent.

The variance estimation for the FE estimator is a little bit tricky. We assume a homoskedasitcity condition to simplify the calculation. Violation of this assumption changes the form of the asymptotic variance, but does not jeopardize the asymptotic normality.

Assumption FE.2 $\text{var}\left(\epsilon_i|\alpha_i, \mathbf{x}_i\right) = \sigma_\epsilon^2 I_T$ for all i.

Under FE.1 and FE.2, $\widehat{\sigma}_\epsilon^2 = \frac{1}{n(T-1)} \sum_{i=1}^n \sum_{t=1}^T \widehat{\tilde{\epsilon}}_{it}^2$ is a consistent estimator of σ_ϵ^2, where $\widehat{\tilde{\epsilon}} = \tilde{y}_{it} - \tilde{x}_{it}\widehat{\beta}_2^{FE}$. Note that the denominator is $n\left(T-1\right)$, not nT. The necessity of adjusting the degree of freedom can be easily seen from the FWL theorem: the FE estimator for the slope coefficient is numerical the same as its counterpart in the full regression with a dummy variable for each cross sectional unit.

If FE.1 and FE.2 are satisfied, then

$$\left(\widehat{\sigma}_\epsilon^2 \left(\tilde{X}'\tilde{X}\right)^{-1}\right)^{-1/2} \left(\widehat{\beta}_2^{FE} - \beta_2^0\right) \xrightarrow{d} N\left(0, I_K\right).$$

Proof. Let $M_\iota = I_T - \frac{1}{T}\iota_T\iota_T'$ be the within-group demeaner, and $M = I_n \otimes M_\iota$. The FE estimator can be explicitly written as

$$\widehat{\beta}_2^{FE} = \left(\tilde{X}'\tilde{X}\right)^{-1} \tilde{X}'\tilde{Y} = \left(X'MX\right)^{-1} X'MY.$$

So

$$\sqrt{nT}\left(\widehat{\beta}_2^{FE} - \beta_2^0\right) = \left(\frac{X'MX}{nT}\right)^{-1} \frac{X'M\epsilon}{\sqrt{nT}} = \left(\frac{\tilde{X}'\tilde{X}}{nT}\right)^{-1} \frac{\tilde{X}'\epsilon}{\sqrt{nT}}$$

Since

$$\text{var}\left(\frac{\tilde{X}'\epsilon}{\sqrt{nT}}|X\right) = \frac{1}{nT}E\left(X'M\epsilon\epsilon'MX|X\right) = \frac{1}{nT}X'ME\left(\epsilon\epsilon'|X\right)MX = \left(\frac{\tilde{X}'\tilde{X}}{nT}\right)\sigma^2,$$

We apply a law of large numbers and conclude

$$\left(\tilde{X}'\tilde{X}\right)^{1/2}\left(\widehat{\beta}_2^{FE} - \beta_2^0\right) \xrightarrow{d} N\left(0, \sigma_\epsilon^2 I_K\right).$$

For simplicity, suppose we can direct observe $\tilde{\epsilon}_{it}$. Then

$$\begin{aligned}\frac{1}{n(T-1)}E\left[\sum_{i=1}^{n}\sum_{t=1}^{T}\tilde{\epsilon}_{it}^2\right] &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{T-1}E\left[\epsilon_i' M_\iota \epsilon_i\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{T-1}\mathrm{tr}\left(E\left[M_\iota E\left[\epsilon_i\epsilon_i'|\mathbf{x}_i\right]\right]\right) \\ &= \frac{\sigma_\epsilon^2}{n}\sum_{i=1}^{n}\frac{1}{T-1}\mathrm{tr}\left(M_\iota\right) = \sigma_\epsilon^2.\end{aligned}$$

Although in reality we only observe $\widehat{\tilde{\epsilon}}_{it}$, we can show that the estimation error between $\widehat{\tilde{\epsilon}}_{it}$ and $\tilde{\epsilon}_{it}$ is negligible. Thus by the law of large numbers

$$\widehat{\sigma}_\epsilon^2 = \frac{1}{n(T-1)}\sum_{i=1}^{n}\sum_{t=1}^{T}\widehat{\tilde{\epsilon}}_{it}^2 \xrightarrow{d} \frac{1}{n(T-1)}E\left[\sum_{i=1}^{n}\sum_{t=1}^{T}\tilde{\epsilon}_{it}^2\right] = \sigma_\epsilon^2$$

is a consistent estimator of the variance. The stated conclusion follows. □

We implicitly assume some regularity conditions that allow us to invoke a law of large numbers and a central limit theorem. We ignore those technical details here.

It is important to notice that the within-group demean in FE eliminates all time-invariant explanatory variables, including the intercept. Therefore from FE we cannot obtain the coefficient estimates of these time-invariant variables.

7.1.2 Random Effect

The random effect estimator pursues efficiency at a knife-edge special case $\mathrm{cov}(\alpha_i, x_{it}) = 0$. As mentioned above, FE is consistent when α_i and x_{it} are uncorrelated. However, an inspection of the covariance matrix reveals that OLS is inefficient.

The starting point is again the original model, while we assume

Assumption RE.1 $E[\epsilon_{it}|\alpha_i, \mathbf{x}_i] = 0$ and $E[\alpha_i|\mathbf{x}_i] = 0$.

RE.1 obviously implies $\mathrm{cov}(\alpha_i, x_{it}) = 0$, so

$$S = \mathrm{var}(u_i|\mathbf{x}_i) = \sigma_\alpha^2 \mathbf{1}_T \mathbf{1}_T' + \sigma_\epsilon^2 I_T, \text{ for all } i = 1, \ldots, n.$$

Because the covariance matrix is not a scalar multiplication of the identity matrix, OLS is inefficient.

As mentioned before, FE estimation kills all time-invariant regressors. In contrast, RE allows time-invariant explanatory variables. Let us rewrite the original equation as

$$y_{it} = w_{it}\boldsymbol{\beta} + u_{it},$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2')'$ and $w_{it} = (1, x_{it})$ are $K+1$ vectors, i.e., $\boldsymbol{\beta}$ is the parameter including the intercept, and w_{it} is the explanatory variables including the constant. Had we known S, the GLS estimator would be

$$\widehat{\boldsymbol{\beta}}^{RE} = \left(\sum_{i=1}^{n}\mathbf{w}_i' S^{-1}\mathbf{w}_i\right)^{-1}\sum_{i=1}^{n}\mathbf{w}_i' S^{-1}\mathbf{y}_i = \left(W'\mathbf{S}^{-1}W\right)^{-1}W'\mathbf{S}^{-1}y$$

where $\mathbf{S} = I_T \otimes S$. ("$\otimes$" denotes the Kronecker product.) In practice, σ_α^2 and σ_ϵ^2 in S are unknown, so we seek consistent estimators. Again, we impose a simplifying assumption parallel to FE.2.

Assumption RE.2 $\text{var}(\epsilon_i|\mathbf{x}_i, \alpha_i) = \sigma_\epsilon^2 I_T$ and $\text{var}(\alpha_i|\mathbf{x}_i) = \sigma_\alpha^2$.

Under this assumption, we can consistently estimate the variances from the residuals $\widehat{u}_{it} = y_{it} - x_{it}\widehat{\boldsymbol{\beta}}^{RE}$. That is

$$\widehat{\sigma}_u^2 = \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\widehat{u}_{it}^2$$

$$\widehat{\sigma}_\epsilon^2 = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{T(T-1)}\sum_{t=1}^{T}\sum_{r=1}^{T}\sum_{r\neq t}\widehat{u}_{it}\widehat{u}_{ir}.$$

We claim the asymptotic normality without proof. If RE.1 and RE.2 are satisfied, then

$$\left(\widehat{\sigma}_u^2\left(W'\widehat{\mathbf{S}}^{-1}W\right)^{-1}\right)^{-1/2}\left(\widehat{\boldsymbol{\beta}}^{RE} - \boldsymbol{\beta}_0\right) \xrightarrow{d} N(0, I_{K+1})$$

where $\widehat{\mathbf{S}}$ is a consistent estimator of $\mathbf{S}$.

The complicated formula of the RE estimator is not important because it is be automatically handled by an econometric package. What is important is the conceptual difference of FE and RE on their treatment of the unobservable individual heterogeneity.

Chapter 8

Endogeneity

8.1 Introduction

In microeconomic analysis, exogenous variables are the factors determined outside of the economic system under consideration, and endogenous variables are those decided within the economic system. The terms "endogenous" and "exogenous" in microeconomics will be carried over into multiple-equation econometric models. While in a single-equation regression model

$$y_i = x_i'\beta + e_i \tag{8.1}$$

is only part of the equation system. To make it simple, in the single-equation model we say an x_{ik} is *endogenous,* or is an *endogenous variable*, if $\operatorname{cov}(x_{ik}, \epsilon_i) \neq 0$; otherwise x_{ik} is an *exogenous variable.*

Empirical works using linear regressions are routinely challenged by questions about endogeneity. Such questions plague economic seminars and referee reports. To defend empirical strategies in quantitative economic studies, it is important to understand the source of potential endogeneity and thoroughly discuss attempts for resolving endogeneity.

Endogeneity usually implies difficulty in identifying the parameter of interest with only (y_i, x_i). Identification is critical for the interpretation of empirical economic research. We say a parameter is *identified* if the mapping between the parameter in the model and the distribution of the observed variable is one-to-one; otherwise we say the parameter is *under-identified*, or the parameter is failed to be identified. This is an abstract definition, and let us discuss it in more family linear regression context.

Example 8.1 (Identification failure due to collinearity)**.** The linear projection model implies the moment equation

$$\mathbb{E}\left[x_i x_i'\right]\beta = \mathbb{E}\left[x_i y_i\right]. \tag{8.2}$$

If $E\left[x_i x_i'\right]$ is of full rank, then $\beta = \left(\mathbb{E}\left[x_i x_i'\right]\right)^{-1}\mathbb{E}\left[x_i y_i\right]$ is a function of the quantities of the population moment and it is identified. On the contrary, if some x_k's are perfect collinear so that $\mathbb{E}\left[x_i x_i'\right]$ is rank deficient, there are multiple β that satisfies the k-equation system (8.2). Identification fails. □

Example 8.2 (Identification failure due to endogeneity)**.** Suppose x_i is a scalar random variable,

$$\begin{pmatrix} x_i \\ e_i \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \sigma_{xe} \\ \sigma_{xe} & 1 \end{pmatrix}\right)$$

follows a joint normal distribution, and the dependent variable y_i is generated from (8.1). The joint normal assumption implies that the conditional mean

$$\mathbb{E}[y_i|x_i] = \beta x_i + \mathbb{E}[e_i|x_i] = (\beta + \sigma_{xe})\, x_i$$

coincides with the linear projection model, and $\beta + \sigma_{xe}$ is the linear projection coefficient. From the observable random variable (y_i, x_i), we can only learn $\beta + \sigma_{xe}$. As we cannot learn σ_{xe} from the data due to the unobservable e_i, there is no way to recover β. This is exactly the *omitted variable bias* that we have discussed earlier in this course. The gap lies between the available data (y_i, x_i) and the identification of the model. In the special case that we assume $\sigma_{xe} = 0$, the endogeneity vanishes and β is identified.

The linear projection model is so far the most general model in this course that justifies OLS. OLS is consistent for the linear projection coefficient. By the definition of the linear projection model, $\mathbb{E}[x_i e_i] = 0$ so there is no room for endogeneity in the linear projection model. In other words, if we talk about endogeneity, we must not be working with the linear projection model, and the coefficients we pursue are not the linear projection coefficients. □

In econometrics we are often interested in a model with economic interpretation. The common practice in empirical research assumes that the observed data are generated from a parsimonious model, and the next step is to estimate the unknown parameters in the model. Since it is often possible to name some factors not included in the regressors but they are correlated with the included regressors and in the mean time also affects y_i, endogeneity becomes a fundamental problem.

To resolve endogeneity, we seek extra variables or data structure that may guarantee the identification of the model. The most often used methods are (i) fixed effect model (ii) instrumental variables. The fixed effect model requires that multiple observations, often across time, are collected for each individual i. Moreover, the source of endogeneity is time invariant and enters the model additively in the form

$$y_{it} = x_{it}'\beta + u_{it},$$

where $u_{it} = \alpha_i + \epsilon_{it}$ is the composite error. The panel data approach extends (y_i, x_i) to $(y_{it}, x_{it})_{i=1}^T$ if data are available along the time dimension.

The instrumental variable approach extends (y_i, x_i) to (y_i, x_i, z_i), where the extra random variable z_i is called the *instrument variable*. It is assumed that z_i is orthogonal to the error e_i . Therefore, along with the model it adds an extra variable z_i.

Before closing this section, we stress that either the panel data approach or the instrumental variable approach entails extra information beyond (y_i, x_i). Without such extra data, there is no way to resolve the identification failure. Just as the linear project model is available for any joint distribution of (y_i, x_i) with existence of suitable moments, from a pure statistical point of view a linear IV model is an artifact depends only on the choice of (y_i, x_i, z_i) without referencing to any economics. In essence, the linear IV model seeks a linear combination $y_i - \beta x_i$ that is orthogonal to the linear space spanned by z_i.

(**New added paragraph**) There are two requirements for valid IVs: orthogonality and relevance. Orthogonality entails that the model is correctly specified, so that we can focus on relevance. If relevance is violated, meaning that the IVs are not correlated with the endogenous variable, then multiple parameters can generate the observable data. Identification, as in the standard definition in econometrics, breaks down.

8.2 Examples

As econometricians mostly work with non-experimental data, we cannot overstate the importance of the endogeneity problem. We go over a few examples.

Example 8.3 (Dynamic Panel Model)**.** We know that the first-difference (FD) estimator is consistent for (static) panel data model. Nevertheless, the FD estimator encounters difficulty in a dynamic panel model

$$y_{it} = \beta_1 + \beta_2 y_{it-1} + \beta_3 x_{it} + \alpha_i + \epsilon_{it},$$

even if we assume

$$\mathbb{E}\left[\epsilon_{it} | \alpha_i, x_{i1}, \ldots, x_{iT}, y_{it-1}, y_{it-2}, \ldots, y_{i0}\right] = 0. \tag{8.3}$$

When taking difference of the above equation for periods t and $t-1$, we have

$$\left(y_{it} - y_{it-1}\right) = \beta_2 \left(y_{it-1} - y_{it-2}\right) + \beta_3 \left(x_{it} - x_{it-1}\right) + \left(\epsilon_{it} - \epsilon_{it-1}\right).$$

Under (8.3), $\mathbb{E}\left[\left(x_{it} - x_{it-1}\right)\left(\epsilon_{it} - \epsilon_{it-1}\right)\right] = 0$, but

$$\mathbb{E}\left[\left(y_{it-1} - y_{it-2}\right)\left(\epsilon_{it} - \epsilon_{it-1}\right)\right] = -\mathbb{E}\left[y_{it-1}\epsilon_{it-1}\right] = -\mathbb{E}\left[\epsilon_{it-1}^2\right] \neq 0. \quad \square$$

Example 8.4 (Classical Measurement Error)**.** Endogeneity also emerges when an explanatory variables is not directly observable but is replaced by a measurement with error. Suppose the true linear model is

$$y_i = \beta_1 + \beta_2 x_i^* + u_i, \tag{8.4}$$

with $\mathbb{E}\left[u_i | x_i^*\right] = 0$. We cannot observe x_i^* but we observe x_i, a measurement of x_i^*, and they are linked by

$$x_i = x_i^* + v_i$$

with $\mathbb{E}\left[v_i | x_i^*, u_i\right] = 0$. Such a formulation of the measurement error is called the *classical measurement error*. Substitute out the unobservable x_i^* in (8.4),

$$y_i = \beta_1 + \beta_2 \left(x_i - v_i\right) + u_i = \beta_1 + \beta_2 x_i + e_i \tag{8.5}$$

where $e_i = u_i - \beta_2 v_i$. The correlation

$$\mathbb{E}\left[x_i e_i\right] = \mathbb{E}\left[\left(x_i^* + v_i\right)\left(u_i - \beta_2 v_i\right)\right] = -\beta_2 \mathbb{E}\left[v_i^2\right] \neq 0.$$

OLS (8.5) would not deliver a consistent estimator. $\square$

Next, we give two examples of equation systems, one from microeconomics and the other from macroeconomics.

Example 8.5 (Demand-Supply System)**.** Let p_i and q_i be a good's log-price and log-quantity on the i-th market, and they are iid across markets. We are interested in the demand curve

$$p_i = \alpha_d - \beta_d q_i + e_{di} \tag{8.6}$$

for some $\beta_d \geq 0$ and the supply curve

$$p_i = \alpha_s + \beta_s q_i + e_{si} \tag{8.7}$$

for some $\beta_s \geq 0$. We use a simple linear specification so that the coefficient β_d can be interpreted as demand elasticity and β_s as supply elasticity. Undergraduate microeconomics teaches the deterministic form but we add an error term to cope with the data. Can we learn the elasticities by regression p_i on q_i?

The two equations can be written in a matrix form

$$\begin{pmatrix} 1 & \beta_d \\ 1 & -\beta_s \end{pmatrix} \begin{pmatrix} p_i \\ q_i \end{pmatrix} = \begin{pmatrix} \alpha_d \\ \alpha_s \end{pmatrix} + \begin{pmatrix} e_{di} \\ e_{si} \end{pmatrix}. \tag{8.8}$$

Microeconomic terminology calls (p_i, q_i) endogenous variables and (e_{di}, e_{si}) exogenous variables. (8.8) is a *structural equation* because it is motivated from economic theory so that the coefficients bear economic meaning. If we rule out the trivial case $\beta_d = \beta_s = 0$, we can solve

$$\begin{pmatrix} p_i \\ q_i \end{pmatrix} = \frac{1}{\beta_s + \beta_d} \begin{pmatrix} \beta_s & \beta_d \\ 1 & -1 \end{pmatrix} \left[\begin{pmatrix} \alpha_d \\ \alpha_s \end{pmatrix} + \begin{pmatrix} e_{di} \\ e_{si} \end{pmatrix} \right]. \tag{8.9}$$

This equation (8.9) is called the *reduced form*—the endogenous variables are expressed as explicit functions of the parameters and the exogenous variables. In particular,

$$q_i = \left(\alpha_d + e_{di} - \alpha_s - e_{si}\right) / \left(\beta_s + \beta_d\right)$$

so that the log-price is correlated with both e_{si} and e_{di}. As q_i is endogenous (in the econometric sense) in either (8.6) or (8.7), neither the demand elasticity nor the supply elasticity is identified with (p_i, q_i). Indeed, as

$$q_i = \left(\beta_s \alpha_d + \beta_d \alpha_s + \beta_s e_{di} + \beta_d e_{si}\right) / \left(\beta_s + \beta_d\right)$$

from (8.9), the linear projection coefficient of p_i on q_i is

$$\frac{\operatorname{cov}\left(p_i, q_i\right)}{\operatorname{var}\left(q_i\right)} = \frac{\beta_s \sigma_d^2 - \beta_d \sigma_s^2 + \left(\beta_d - \beta_s\right) \sigma_{sd}}{\beta_d^2 \sigma_d^2 + \beta_d \sigma_s^2 + 2\beta_d \beta_s \sigma_{sd}},$$

where $\sigma_d^2 = \operatorname{var}\left(e_{di}\right)$, $\sigma_s^2 = \operatorname{var}\left(e_{si}\right)$ and $\sigma_{sd} = \operatorname{cov}\left(e_{di}, e_{si}\right)$.

This is a classical example of the demand-supply system. The structural parameter cannot be directly identified because the observed (p_i, q_i) is the outcome of an equilibrium—the crossing of the demand curve and the supply curve. To identify the demand curve, we will need an instrument that shifts the supply curve only; and vice versa. □

Example 8.6 (Keynesian-Type Macro Equations)**.** This is a model borrowed from Hayashi (2000, p.193) but originated from Haavelmo (1943). An econometrician is interested in learning β_2, the marginal propensity of consumption, in the Keynesian-type equation

$$C_i = \beta_1 + \beta_2 Y_i + u_i \tag{8.10}$$

where C_i is household consumption, Y_i is the GNP, and u_i is the unobservable error. However, Y_i and C_i are connected by an accounting equality (with no error)

$$Y_i = C_i + I_i,$$

where I_i is investment. We assume $\mathbb{E}\left[u_i | I_i\right] = 0$ as investment is determined in advance. OLS (8.10) will be inconsistent because in the reduced-form $Y_i = \frac{1}{1-\beta_2}\left(\beta_1 + u_i + I_i\right)$ implies $\mathbb{E}\left[Y_i u_i\right] = \mathbb{E}\left[u_i^2\right] / \left(1 - \beta_2\right) \neq 0$. □

Chapter 9

Generalized Method of Moments

9.1 Introduction

Generalized method of moments (GMM) is an estimation principle that extends *method of moments*. It seeks the parameter value that minimizes a quadratic form of the moments. It is particularly useful in estimating structural models in which moment conditions can be derived from economic theory. GMM emerges as one of the most popular estimators in modern econometrics, and it includes conventional methods like the two-stage least squares (2SLS) and the three-stage least square as special cases.

9.2 GMM in Linear Model

In this section we discuss GMM in a linear single structural equation. A structural equation is a model of economic interest. Consider the following linear structural model

$$y_i = x_{1i}'\beta_1 + z_{1i}'\beta_2 + \epsilon_i, \tag{9.1}$$

where x_{1i} is a k_1-dimensional endogenous explanatory variables, z_{1i} is a k_2-dimensional exogenous explanatory variables with the intercept included. In addition, we have z_{2i}, a k_3-dimensional excluded exogenous variables. Let $K = k_1 + k_2$ and $L = k_2 + k_3$. Denote $x_i = (x_{1i}', z_{1i}')'$ as a K-dimensional explanatory variable, and $z_i = (z_{1i}', z_{2i}')$ as an L-dimensional exogenous vector.

In the context of endogeneity, we can call the exogenous variable *instrument variables*, or simply *instruments*. Let $\beta = (\beta_1', \beta_2')'$ be a K-dimensional parameter of interest. From now on, we rewrite (9.1) as

$$y_i = x_i'\beta + \epsilon_i, \tag{9.2}$$

and we have a vector of instruments z_i.

Before estimating any structural econometric model, we must check identification. A model is *identified* if there is a one-to-one mapping between the distribution of the observed variables and the parameters. In other words, in an identified model any two parameter values β and $\tilde{\beta}$, $\beta \neq \tilde{\beta}$, cannot generate the same distribution for the observable data. In the context of (9.2), identification requires that the true value β_0 is the only value on the parameters space that satisfies the moment condition

$$\mathbb{E}\left[z_i\left(y_i - x_i'\beta\right)\right] = 0_L. \tag{9.3}$$

The rank condition is sufficient and necessary for identification.

Assumption (Rank condition)**.** $\operatorname{rank}\left(\mathbb{E}\left[z_i x_i'\right]\right) = K$.

Note that $\mathbb{E}\left[x_i' z_i\right]$ is a $K \times L$ matrix. The rank condition implies the *order condition* $L \geq K$, which says that the number of excluded instruments must be no fewer than the number of endogenous variables.

Theorem. *The parameter in (9.3) is identified if and only if the rank condition holds.*

Proof. (The "if" direction). For any $\tilde{\beta}$ such that $\tilde{\beta} \neq \beta_0$,

$$\begin{aligned}\mathbb{E}\left[z_i\left(y_i - x_i'\tilde{\beta}\right)\right] &= \mathbb{E}\left[z_i\left(y_i - x_i'\beta_0\right)\right] + \mathbb{E}\left[z_i x_i'\right]\left(\beta_0 - \tilde{\beta}\right) \\ &= 0_L + \mathbb{E}\left[z_i x_i'\right]\left(\beta_0 - \tilde{\beta}\right).\end{aligned}$$

Because $\operatorname{rank}\left(\mathbb{E}\left[z_i x_i'\right]\right) = K$, we would have $\mathbb{E}\left[z_i x_i'\right]\left(\beta_0 - \tilde{\beta}\right) = 0_L$ if and only if $\beta_0 - \tilde{\beta} = 0_K$, which violates $\tilde{\beta} \neq \beta_0$. Therefore β_0 is the unique value that satisfies (9.3).

(The "only if" direction is left as an exercise. Hint: By contrapositiveness, if the rank condition fails, then the model is not identified. We can easily prove the claim by making an example.) □

Because identification is a prerequisite for structural estimation, from now on we always assume that the model is identified. When it is just-identified ($L = K$), by (9.3) we can express the parameter as

$$\beta = \left(\mathbb{E}\left[z_i x_i'\right]\right)^{-1}\mathbb{E}\left[z_i y_i\right]. \tag{9.4}$$

It follows by the principle of method of moments that

$$\hat{\beta} = \left(\frac{Z'X}{n}\right)^{-1}\frac{Z'y}{n} = \left(Z'X\right)^{-1}Z'y,$$

which is exactly the 2SLS when $L = K$.

In the rest of this section, we focus on the over-identified case ($L > K$). When $L > K$, (9.3) involves more equations than the number of parameters, so that directly taking the inverse as in (9.4) is inapplicable.

In order to express β explicitly, we define a criterion function

$$Q\left(\beta\right) = \mathbb{E}\left[z_i\left(y_i - x_i\beta\right)\right]' W \mathbb{E}\left[z_i\left(y_i - x_i\beta\right)\right],$$

where W is an arbitrary $L \times L$ positive-definite symmetric matrix. Because of the quadratic form, $Q\left(\beta\right) \geq 0$ for all β. Identification indicates that $Q\left(\beta\right) = 0$ if and only if $\beta = \beta_0$. Therefore we conclude

$$\beta_0 = \arg\min_{\beta} Q\left(\beta\right).$$

Since $Q\left(\beta\right)$ is a smooth function of β, the minimizer β_0 can be characterized by the first-order condition

$$0_K = \frac{\partial}{\partial\beta}Q\left(\beta_0\right) = -\mathbb{E}\left[x_i z_i'\right] W \mathbb{E}\left[z_i\left(y_i - x_i\beta_0\right)\right]$$

Rearranging the above equation, we have

$$\mathbb{E}\left[x_i z_i'\right] W \mathbb{E}\left[z_i x_i'\right]\beta_0 = \mathbb{E}\left[x_i z_i'\right] W \mathbb{E}\left[z_i y_i\right].$$

Denote $\Sigma = \mathbb{E}\left[z_i x_i'\right]$. Under the rank condition, $\Sigma' W \Sigma$ is invertible so that we can solve

$$\beta_0 = \left(\Sigma' W \Sigma\right)^{-1}\Sigma' W \mathbb{E}\left[z_i y_i\right].$$

In practice, we use the sample moments to replace the corresponding population moments. The GMM estimator mimics its population formula.

$$\begin{aligned}\widehat{\beta} &= \left(\frac{1}{n}\sum x_i z_i' W \frac{1}{n}\sum z_i x_i'\right)^{-1} \frac{1}{n}\sum x_i z_i' W \frac{1}{n}\sum z_i y_i \\ &= \left(\frac{X'Z}{n} W \frac{Z'X}{n}\right)^{-1} \frac{X'Z}{n} W \frac{Z'y}{n} \\ &= \left(X'ZWZ'X\right)^{-1} X'ZWZ'y.\end{aligned}$$

Exercise. The same GMM estimator $\hat{\beta}$ can be obtained by minimizing

$$\left[\frac{1}{n}\sum_{i=1}^{n} z_i\left(y_i - x_i\beta\right)\right]' W \left[\frac{1}{n}\sum_{i=1}^{n} z_i\left(y_i - x_i\beta\right)\right] = \frac{\left(y - X\beta\right)' Z}{n} W \frac{Z'\left(y - X\beta\right)}{n},$$

or more concisely,

$$\hat{\beta} = \arg\min_{\beta} \left(y - X\beta\right)' ZWZ' \left(y - X\beta\right).$$

Now we check the asymptotic properties of $\widehat{\beta}$. A few assumptions are in order.

Assumption (A.1)**.** *$Z'X/n \xrightarrow{\text{p}} \Sigma$ and $Z'\epsilon/n \xrightarrow{\text{p}} 0_L$.*

A.1 assumes that we can apply a law of large numbers, so that that the sample moments $Z'X/n$ and $Z'\epsilon/n$ converge in probability to their population counterparts.

Theorem. *Under A.1, $\widehat{\beta}$ is consistent.*

Proof. The step is similar to the consistency proof of OLS.

$$\begin{aligned}\widehat{\beta} &= \left(X'ZWZ'X\right)^{-1} X'ZWZ' \left(X'\beta_0 + \epsilon\right) \\ &= \beta_0 + \left(\frac{X'Z}{n} W \frac{Z'X}{n}\right)^{-1} \frac{X'Z}{n} W \frac{Z'\epsilon}{n} \\ &\xrightarrow{\text{p}} \beta_0 + \left(\Sigma' W \Sigma\right)^{-1} \Sigma' W 0 = \beta_0.\end{aligned}$$

□

To check asymptotic normality, we assume that a central limit theorem can be applied.

Assumption (A.2)**.** $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} z_i'\epsilon_i \Rightarrow N\left(0_L, \Omega\right)$*, where* $\Omega = \mathbb{E}\left[z_i' z_i \epsilon_i^2\right]$.

Theorem (Asymptotic Normality)**.** *Under A.1 and A.2,*

$$\sqrt{n}\left(\widehat{\beta} - \beta_0\right) \Rightarrow N\left(0_K, \left(\Sigma' W \Sigma\right)^{-1} \Sigma' W \Omega W \Sigma \left(\Sigma' W \Sigma\right)^{-1}\right). \tag{9.5}$$

Proof. Multiply $\widehat{\beta} - \beta_0$ by the scaling factor $\sqrt{n}$,

$$\begin{aligned}\sqrt{n}\left(\widehat{\beta} - \beta_0\right) &= \left(\frac{X'Z}{n} W \frac{Z'X}{n}\right)^{-1} \frac{X'Z}{n} W \frac{Z'\epsilon}{\sqrt{n}} \\ &= \left(\frac{X'Z}{n} W \frac{Z'X}{n}\right)^{-1} \frac{X'Z}{n} W \frac{1}{\sqrt{n}}\sum_{i=1}^{n} z_i'\epsilon_i.\end{aligned}$$

The conclusion follows as

$$\frac{X'Z}{n} W \frac{Z'X}{n} \xrightarrow{\text{p}} \Sigma' W \Sigma$$

and

$$\frac{X'Z}{n} W \frac{1}{\sqrt{n}}\sum z_i'\epsilon_i \Rightarrow \Sigma' W \times N\left(0, \Omega\right).$$

□

It is clear from (9.5) that the GMM estimator's asymptotic variance depends on the choice of W. A natural question follows: can we optimally choose a W to make the asymptotic variance as small as possible? Here we claim the result without a proof.

Claim. The choice $W = \Omega^{-1}$ makes $\widehat{\beta}$ an asymptotically efficient estimator, under which the asymptotic variance is

$$\left(\Sigma'\Omega^{-1}\Sigma\right)^{-1}\Sigma'\Omega^{-1}\Omega\Omega^{-1}\Sigma\left(\Sigma'\Omega^{-1}\Sigma\right)^{-1} = \left(\Sigma'\Omega^{-1}\Sigma\right)^{-1}.$$

In practice, Ω is unknown but can be estimated. Hansen (1982) suggests the following procedure, which is known as the *two-step GMM*.

1. Choose any valid W, say $W = I_L$, to get a consistent (but inefficient in general) estimator $\hat{\beta}$. Save the residual $\widehat{\epsilon}_i = y_i - x_i'\widehat{\beta}$ and estimate the variance matrix $\widehat{\Omega} = \frac{1}{n}\sum z_i z_i'\widehat{\epsilon}_i^2$.

2. Set $W = \widehat{\Omega}^{-1}$ and obtain a second estimator

$$\widehat{\beta} = \left(X'Z\widehat{\Omega}^{-1}Z'X\right)^{-1}X'Z\widehat{\Omega}^{-1}Z'y.$$

   This second estimator is asymptotic efficient.

If we further assume conditional homoskedasticity, then $\Omega = \mathbb{E}\left[z_i z_i'\epsilon_i^2\right] = \mathbb{E}\left[z_i z_i'\mathbb{E}\left[\epsilon_i^2|z_i\right]\right] = \sigma^2\mathbb{E}\left[z_i z_i'\right]$. Therefore in the first-step of the two-step GMM we can estimate the variance of the error term by $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n \widehat{\epsilon}_i^2$ and the variance matrix by $\widehat{\Omega} = \widehat{\sigma}^2\frac{1}{n}\sum_{i=1}^n z_i z_i' = \widehat{\sigma}^2 Z'Z/n$. When we plug this $W = \widehat{\Omega}^{-1}$ into the GMM estimator,

$$\begin{aligned}\widehat{\beta} &= \left(X'Z\left(\widehat{\sigma}^2\frac{Z'Z}{n}\right)^{-1}Z'X\right)^{-1}X'Z\left(\widehat{\sigma}^2\frac{Z'Z}{n}\right)^{-1}Z'y \\ &= \left(X'Z\left(Z'Z\right)^{-1}Z'X\right)^{-1}X'Z\left(Z'Z\right)^{-1}Z'y.\end{aligned}$$

This is exactly the same expression of 2SLS for $L > K$. Therefore, 2SLS can be viewed as a special case of GMM with $W = (Z'Z/n)^{-1}$. Under conditional homoskedasticity, 2SLS is the efficient estimator; otherwise 2SLS is inefficient.

Remark. 2SLS gets its name because it can be obtained using two steps: first regress X on all instruments Z, and then regress y on the fitted value along with the included exogenous variables. However, 2SLS can actually be obtained by one step using the above equation. It is a special case of GMM.

9.2.1 GMM in Nonlinear Model*

The principle of GMM can be used in models where the parameter enters the moment conditions nonlinearly. Let $g_i(\beta) = g(w_i, \beta) \mapsto \mathbb{R}^L$ be a function of the data w_i and the parameter β. If economic theory implies $\mathbb{E}\left[g_i(\beta)\right] = 0$, we can write the GMM population criterion function as

$$Q(\beta) = \mathbb{E}\left[g_i(\beta)\right]' W \mathbb{E}\left[g_i(\beta)\right]$$

Example. Nonlinear models nest the linear model as a special case. For the linear IV model in the previous section, the data is $w_i = (y_i, x_i, z_i)$, and the moment function is $g(w_i, \beta) = z_i'(y_i - x_i\beta)$.

In practice we use the sample moments to mimic the population moments in the criterion function

$$Q_n(\beta) = \left(\frac{1}{n}\sum_{i=1}^{n} g_i(\beta)\right)' W \left(\frac{1}{n}\sum_{i=1}^{n} g_i(\beta)\right).$$

The GMM estimator is defined as

$$\hat{\beta} = \arg\min_{\beta} Q_n(\beta).$$

In these nonlinear models, a closed-form solution is in general unavailable, while the asymptotic properties can still be established. We state these asymptotic properties without proofs.

Theorem. *If the model is identified, and*

$$\mathbb{P}\left[\sup_{\beta}\left|\frac{1}{n}\sum_{i=1}^{n} g_i(\beta) - \mathbb{E}\left[g_i(\beta)\right]\right| > \varepsilon\right] \to 0$$

for any constant $\varepsilon > 0$, then $\hat{\beta} \xrightarrow{\text{p}} \beta$. If in addition $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g_i(\beta_0) \Rightarrow N(0, \Omega)$, then

$$\sqrt{n}\left(\hat{\beta} - \beta_0\right) \Rightarrow N\left(0, \left(\Sigma' W \Sigma\right)^{-1}\left(\Sigma' W \Omega W \Sigma\right)\left(\Sigma' W \Sigma\right)^{-1}\right)$$

where $\Sigma = \mathbb{E}\left[\frac{\partial}{\partial \beta'} g_i(\beta_0)\right]$ and $\Omega = \mathbb{E}\left[g_i(\beta_0)\, g_i(\beta_0)'\right]$. If we choose $W = \Omega^{-1}$, then the GMM estimator is efficient, and the asymptotic variance becomes $\left(\Sigma' \Omega^{-1} \Sigma\right)^{-1}$.

Remark. The list of assumptions in the above statement is incomplete. We only lay out the key conditions but neglect some technical details.

$Q_n(\beta)$ measures how close are the moments to zeros. It can serve as a test statistic with proper formulation. Under the null hypothesis $\mathbb{E}\left[g_i(\beta)\right] = 0_L$, this so-called "$J$-test" checks whether a moment condition is violated. The test statistic is

$$\begin{aligned} J\left(\widehat{\beta}\right) &= n\left(\frac{1}{n}\sum_{i=1}^{n} g_i\left(\widehat{\beta}\right)\right)' \widehat{\Omega}^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} g_i\left(\widehat{\beta}\right)\right) \\ &= \left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g_i\left(\widehat{\beta}\right)\right)' \widehat{\Omega}^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g_i\left(\widehat{\beta}\right)\right) \end{aligned}$$

where $\widehat{\Omega}$ is a consistent estimator of Ω, and $\widehat{\beta}$ is an efficient estimator, for example, the second $\widehat{\beta}$ from the two-step GMM. This statistics converges in distribution to a chi-square random variable with degree of freedom $L - K$. That is, under the null,

$$J\left(\widehat{\beta}\right) \Rightarrow \chi^2(L - K)$$

If the null hypothesis is false, then the test statistic tends to be large, and it is more likely to reject the null.

Zhentao Shi. October 6, 2020