# Advanced Econometrics[1]

Part I. Methods of Econometrics

Part II. Econometric Analysis of Cross-Sectional Dependence Models

Tadao Hoshino (星野匡郎)[2]

September 1, 2022

[1]Updated irregularly. Please check the course page occasionally for the latest version.

[2]School of Political Science and Economics, Waseda University. 1-6-1 Nishi-waseda, Shinjuku-ku, Tokyo 169-8050, Japan. Email: thoshino@waseda.jp. If you find any typos and errors please let me know.

# Contents

# Part I

# Methods of Econometrics

# Chapter 1

# Consistency and Asymptotic Normality

## 1.1 Convergence of real sequences

Real numbers are numbers that have points on the number line $\mathbb{R}$. Let $a$ be a real number and $(a_n)_{n=1}^{\infty}$ be a sequence of real numbers:

$$(a_n)_{n=1}^{\infty} = a_1,\ a_2,\ a_3, \ldots$$

For example, a sequence

$$a_1 = 1,\ a_2 = 1.4,\ a_3 = 1.41,\ \cdots,\ a_n = 1.4142\ldots, \cdots$$

gets closer and closer to $\alpha = \sqrt{2}$ $(= 1.414213\ldots)$ as $n$ increases. The number $\alpha$ is called the **limit** of the sequence $(a_n)_{n=1}^{\infty}$. We say that the sequence $(a_n)_{n=1}^{\infty}$ **converges** to $\alpha$, and write

$$\lim_{n\to\infty} a_n = \alpha$$

$$\text{or equivalently } \ a_n \to \alpha \ \ (n \to \infty).$$

For other examples, the limit of the sequence $a_1 = 1,\ a_2 = 1/2,\ a_3 = 1/3,\ \cdots,\ a_n = 1/n, \cdots$ is clearly $\lim_{n\to\infty} a_n = 0$. It is well-known that the sequence

$$a_1 = (1 + (1/1))^1,\ a_2 = (1 + (1/2))^2,\ \cdots,\ a_n = (1 + (1/n))^n, \cdots$$

converges to the Napier's constant: $\lim_{n\to\infty} a_n = e$ $(= 2.718\cdots)$. On the other hand, the limit of the sequence $a_1 = 1,\ a_2 = 2,\ a_3 = 3,\ \cdots,\ a_n = n, \cdots$ is infinite: $\lim_{n\to\infty} a_n = \infty$. When the limit of a sequence is infinite, we say that the sequence is a **divergent**. There are sequences that are neither convergent nor divergent. For example, the sequence

$$a_1 = -1,\ a_2 = 1,\ a_3 = -1,\ \cdots,\ a_n = (-1)^n, \cdots$$

alternates between 1 and $-1$.

The statement "the sequence $(a_n)_{n=1}^{\infty}$ gets closer to $\alpha$ as $n$ increases" is rather informal and mathematically unclear. A formal definition of convergence is as follows.

**Definition 1.1.1** *We say that a sequence of real numbers $(a_n)_{n=1}^{\infty}$ converges to a limit $\alpha$ if for any given positive number $\kappa > 0$, there exists a natural number $n(\kappa)$ such that*

$$|a_n - \alpha| < \kappa \ \text{ for all } n \text{ satisfying } n(\kappa) \le n. \tag{1.1.1}$$

The interpretation of (1.1.1) is as follows. Choose an arbitrary small $\kappa$, say $\kappa < 10^{-100}$ or even smaller. Even for such a small value of $\kappa$, if we set $n(\kappa)$ sufficiently large, the distance between the terms of the sequence after $a_{n(\kappa)}$ and $\alpha$ can be smaller than $\kappa$ (see Figure 1.1).

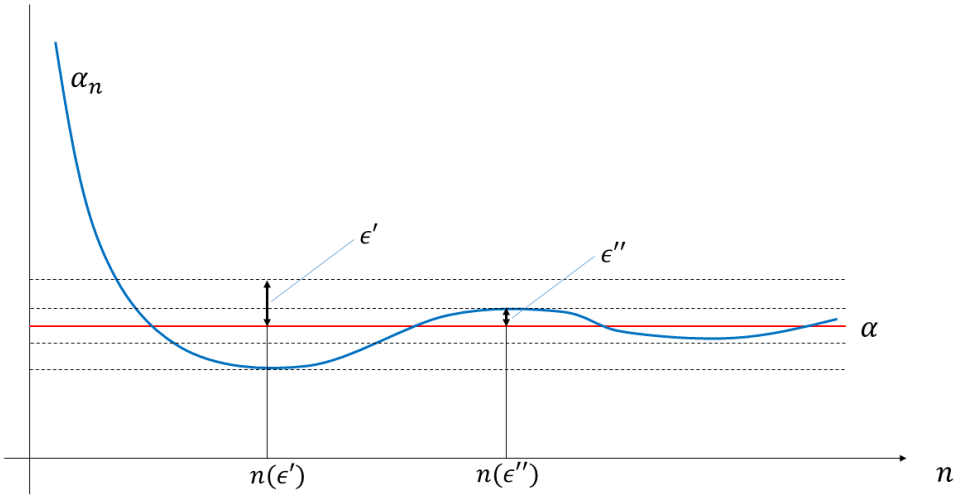



Figure 1.1: Convergence of a real sequence

**Exercise 1.1.2** What are the limits of the following sequences?

1. $a_n = n^2 - n$
2. $a_n = (-1)^n/\sqrt{n}$
3. $a_n = \sqrt{n+1} - \sqrt{n}$

**Exercise 1.1.3** Prove that the infinite decimal $0.999\ldots$ is equal to one using the notion of convergence and limit.

**Exercise 1.1.4** Let $S_n = \sum_{k=1}^n a_k$ and $\overline{S}_n = \sum_{k=1}^n |a_k|$. Show that if $\overline{S}_n$ converges to a finite number, so does $S_n$. (This is called **absolute convergence**.)

Note that $a_k \to 0$ as $k \to \infty$ does not ensure the convergence of $S_n = \sum_{k=1}^n a_k$ at all. For example, as is well known, $\sum_{k=1}^n k^{-1}$ diverges to infinity.[1]

[1] Proof: Observe that

$$\exp(1 + 2^{-1} + 3^{-1} + \cdots + n^{-1}) = \exp(1)\exp(2^{-1})\exp(3^{-1})\cdots\exp(n^{-1}) \geq (1+1)(1+2^{-1})(1+3^{-1})\cdots(1+n^{-1})$$
$$= 2 \cdot \frac{3}{2} \cdot \frac{4}{3} \cdots \frac{n+1}{n} = n+1,$$

where we have used the fact $\exp(x) \geq 1 + x$. This implies that $1 + 2^{-1} + 3^{-1} + \cdots + n^{-1} \geq \log(n+1)$, where the right-hand side diverges

## 1.2 Convergence in probability

Now let $Z_n \in \mathbb{R}$ be a sequence of "random variables". Recall that random variables are not numbers, but they are functions (rules) that assign some real numbers to random events. Thus, whether the sequence of random variables $Z_n$ converges to a limit or not is itself a stochastic event, and hence the definition 1.1.1 cannot be applied directly. However, since a probability is a real number, we can still consider the convergence of a sequence of probabilities in the usual sense. Then, we introduce the notion of "convergence in probability".

**Definition 1.2.1** *For any given positive number $\kappa > 0$, if*

$$\lim_{n\to\infty} \Pr(|Z_n - \mu| \leq \kappa) = 1$$

*holds, we say that $Z_n$ **converges in probability** to $\mu$ as $n \to \infty$, and write*

$$\operatorname*{plim}_{n\to\infty} Z_n = \mu$$
$$\text{or equivalently } Z_n \xrightarrow{p} \mu \ \ (n \to \infty).$$

The above definition can be read as follows: $a_n \equiv \Pr(|Z_n - \mu| \leq \kappa)$, which is a real number, converges to one for any $\kappa$ in the sense of Definition 1.1.1. It is important to note that $\mu$ can be either a random variable or a constant number in this definition.

When $Z_n$ is a sequence of vector-valued random variables, one can use an appropriate norm $||\cdot||$, such as the Euclidean norm, in place of the absolute value $|\cdot|$.[2] Note that the choice of $\kappa$ is arbitrary, and thus it can be, for example, $\kappa < 10^{-100}$ or even smaller as long as it is larger than 0. This definition says that the probability of observing the event $\{|Z_n - \mu| \leq \kappa\}$ approaches to one as $n$ increases even for such very small $\kappa$.

Note that a constant is a special case of a random variable. Therefore, when a sequence of constants converges $a_n \to \alpha$, it trivially holds that $a_n \xrightarrow{p} \alpha$. Further, for a sequence of random variables that converges in probability $Z_n \xrightarrow{p} \mu$, we also have $a_n Z_n \xrightarrow{p} \alpha\mu$ because

$$\begin{aligned} |a_n Z_n - \alpha\mu| &\leq |a_n Z_n - a_n\mu| + |a_n\mu - \alpha\mu| \\ &= |a_n| \cdot |Z_n - \mu| + |\mu| \cdot |a_n - \alpha| \xrightarrow{p} 0, \end{aligned}$$

where the inequality follows from the triangle inequality.

**Lemma 1.2.2 (Markov's inequality)** *For any random variable $Z_n$ and positive constant $\kappa > 0$,*

$$\Pr(|Z_n| > \kappa) \leq \frac{\mathbb{E}|Z_n|}{\kappa}. \tag{1.2.1}$$

**Proof.** First, observe the following equality:

$$\begin{aligned} \Pr(|Z_n| > \kappa) &= \mathbb{E}[\mathbf{1}(|Z_n| > \kappa)] \\ &= \mathbb{E}\left[\mathbf{1}\left(\frac{|Z_n|}{\kappa} > 1\right)\right] \end{aligned}$$

as $n$ tends to infinity.

[2] Throughout, when not explicitly mentioned, $||\cdot||$ stands for the Euclidean norm (i.e., the square root of the sum of squared elements).

Since $\frac{|Z_n|}{\kappa}$ is non-negative, we have $\mathbf{1}\left(\frac{|Z_n|}{\kappa} > 1\right) \leq \frac{|Z_n|}{\kappa}$. Hence,

$$\mathbb{E}\left[\mathbf{1}\left(\frac{|Z_n|}{\kappa} > 1\right)\right] \leq \frac{\mathbb{E}|Z_n|}{\kappa}.$$

This implies the desired result. ■

As a corollary of Markov's inequality, we can obtain the following useful inequality:

$$\Pr(|Z_n - \mathbb{E}Z_n| > \kappa) = \Pr((Z_n - \mathbb{E}Z_n)^2 > \kappa^2) \leq \frac{\mathrm{Var}(Z_n)}{\kappa^2}, \tag{1.2.2}$$

which is known as **Chebyshev's inequality**. Note that these inequalities do not make any assumptions about the probability distribution of $Z_n$; that is, they are distribution-free inequalities. When $\mathbb{E}|Z_n|$ and $\mathrm{Var}(Z_n)$ do not exist, such that $\mathbb{E}|Z_n| = \infty$ and $\mathrm{Var}(Z_n) = \infty$, the inequalities (1.2.1) and (1.2.2) hold trivially.

Here, let $X \in \mathbb{R}$ be a random variable, and $\overline{X}_n$ be the sample mean of $X$, where $n$ denotes the sample size. Chebyshev's inequality implies that, if $\mathrm{Var}(\overline{X}_n) \to 0$ as $n \to \infty$,

$$\Pr(|\overline{X}_n - \mathbb{E}\overline{X}_n| > \kappa) \leq \frac{\mathrm{Var}(\overline{X}_n)}{\kappa^2} \to 0.$$

In other words, since

$$\Pr(|\overline{X}_n - \mathbb{E}\overline{X}_n| \leq \kappa) = 1 - \underbrace{\Pr(|\overline{X}_n - \mathbb{E}\overline{X}_n| > \kappa)}_{\to 0},$$

the sample mean converges to its population mean in probability:

$$\overline{X}_n \xrightarrow{p} \mathbb{E}\overline{X}_n.$$

This result is known as the **weak law of large numbers** (WLLN).[3] As shown above, an easy-to-check sufficient condition for WLLN to hold is that $\mathrm{Var}(\overline{X}_n)$ converges to zero as $n$ increases. In the case when the data $\{X_1, \ldots, X_n\}$ are independent and identically distributed (IID), the condition can be easily confirmed by the fact that $\mathrm{Var}(\overline{X}_n) = \mathrm{Var}(X)/n \to 0$ as $n \to \infty$ as long as $\mathrm{Var}(X)$ is finite.[4]

An estimator which converges in probability to its population value is called a **consistent estimator**. That is, letting $\widehat{\theta}_n$ be an estimator of $\theta_0$ obtained from a sample of size $n$, if $\widehat{\theta}_n \xrightarrow{p} \theta_0$, we say that $\widehat{\theta}_n$ is consistent for $\theta_0$. WLLN states that the sample mean is a consistent estimator of the population mean under IID sampling.

**Exercise 1.2.3** Suppose that $\{X_1, \ldots, X_n\}$ are independent, and

$$X_n = \begin{cases} 0 & \text{with probability } 1 - 1/n \\ n & \text{with probability } 1/n. \end{cases}$$

Prove that $X_n$ converges to zero in probability.

[3] Since there is a "weak" LLN, there is also a "strong" version of LLN (SLLN), which states that $\Pr(\lim_{n\to\infty} \overline{X}_n = \mathbb{E}\overline{X}_n) = 1$. The proof of SLLN is much more complicated than that of WLLN. SLLN implies WLLN.

[4] That said, perhaps somewhat surprisingly, $\mathrm{Var}(X) < \infty$ is not a necessary condition for WLLN in general.

**Exercise 1.2.4** Suppose that we have $n$ IID observations $\{X_1, \ldots, X_n\}$ drawn from Uniform$[0, 1]$. Define

$$a_n \equiv \max\{X_1, \ldots, X_n\}.$$

Prove that $a_n$ converges to one in probability.

**Exercise 1.2.5** (**Continuous mapping theorem**) Prove the following claim: Suppose that $g(\cdot)$ is a continuous function. Then, if $Z_n \xrightarrow{p} \mu$, we have $g(Z_n) \xrightarrow{p} g(\mu)$.

## 1.3 Laudau's notation: big $O$ and little $o$

**Landau's notation** is a convenient tool for representing asymptotic behavior of a sequence, also known as "Big-$O$, little-$o$ notation" or "asymptotic notation". Let $a_n$ be a sequence of real numbers, and $p_n$ be a sequence of positive numbers. We write

$$a_n = O(p_n)$$

if there exists a constant $C$ such that

$$|a_n| \le C p_n \text{ for all sufficiently large } n.$$

When $a_n = O(p_n)$ holds, we say that $a_n$ is of order $p_n$. Intuitively, if $p_n \to \infty$, this means that $a_n$ grows not faster than $p_n$. Similarly, if $p_n \to 0$, $a_n = O(p_n)$ means that $a_n$ converges to zero at the same rate or faster than $p_n$. For example, $a_n = O(n^{-1})$ implies that $a_n$ converges to zero at a rate at least $n^{-1}$ because $a_n$ does not diverge even if it is multiplied by $n$ (i.e., $n|a_n| \le C$). When $a_n$ is bounded by some finite constant, then we can write $a_n = O(1)$.

By the definition of $O$ symbol, the following properties can be easily verified:

**Lemma 1.3.1**

- $a_n = O(p_n)$ & $p_n = O(q_n) \Longrightarrow a_n = O(q_n)$
  - **Proof.** By assumption, $|a_n| \le C_1 p_n$ and $p_n \le C_2 q_n$. Then, $|a_n| \le C_1 C_2 q_n$.
- $a_n = O(p_n)$ & $b_n = O(q_n) \Longrightarrow a_n b_n = O(p_n q_n)$
  - **Proof.** By assumption, $|a_n| \le C_1 p_n$ and $|b_n| \le C_2 q_n$. Then, $|a_n b_n| \le C_1 C_2 p_n q_n$.
- $a_n = O(p_n)$ & $b_n = O(q_n) \Longrightarrow a_n + b_n = O(\max\{p_n, q_n\})$
  - **Proof.** By assumption, $|a_n| \le C_1 p_n$ and $|b_n| \le C_2 q_n$. Then, the triangle inequality gives $|a_n + b_n| \le |a_n| + |b_n| \le C_1 p_n + C_2 q_n \le \max\{C_1, C_2\}(p_n + q_n) \le 2\max\{C_1, C_2\}\max\{p_n, q_n\}$.

For illustrations, the orders of the sequences in Exercise 1.1.2 are as follows:

1. $a_n = n^2 - n = O(n^2) + O(n) = O(n^2)$
2. $a_n = (-1)^n/\sqrt{n} = O(n^{-1/2})$

3. $a_n = \sqrt{n+1} - \sqrt{n} = O(1)$

Little-$o$ notation $a_n = o(p_n)$ is a special case of Big-$O$ notation $a_n = O(p_n)$. If

$$\frac{|a_n|}{p_n} \to 0 \text{ as } n \to \infty,$$

we write

$$a_n = o(p_n).$$

Obviously, if $a_n = o(p_n)$ is true, so is $a_n = O(p_n)$. If a sequence is $o(1)$, it means that the sequence is just a convergent to zero. For example, if we have $a_n = o(n^{-1})$, $a_n$ goes to zero so quickly that it can converge to zero even when multiplied by $n$.

By the definitions of $O$ and $o$, the following properties are evident:

**Lemma 1.3.2**

- $a_n = o(p_n)$ & $p_n = o(q_n) \Longrightarrow a_n = o(q_n)$
  - **Proof.** By assumption, $|a_n|/p_n \to 0$ and $p_n/q_n \to 0$. Then, $|a_n|/q_n = (|a_n|/p_n)(p_n/q_n) \to 0$.
- $a_n = o(p_n)$ & $b_n = o(q_n) \Longrightarrow a_n b_n = o(p_n q_n)$
  - **Proof.** By assumption, $|a_n|/p_n \to 0$ and $|b_n|/q_n \to 0$. Then, $|a_n b_n|/(p_n q_n) \to 0$.
- $a_n = o(p_n)$ & $b_n = o(q_n) \Longrightarrow a_n + b_n = o(\max\{p_n, q_n\})$
  - **Proof.** By assumption, $|a_n|/p_n \to 0$ and $|b_n|/q_n \to 0$. It is clear to see that $|a_n|/\max\{p_n, q_n\} \to 0$ and $|b_n|/\max\{p_n, q_n\} \to 0$. Then, the triangle inequality gives $|a_n + b_n|/\max\{p_n, q_n\} \le |a_n|/\max\{p_n, q_n\} + |b_n|/\max\{p_n, q_n\} \to 0$.
- ♠ $a_n = o(p_n)$ & $b_n = O(q_n) \Longrightarrow a_n b_n = o(p_n q_n)$
- ♣ $a_n = o(p_n)$ & $b_n = O(q_n) \Longrightarrow a_n + b_n = O(\max\{p_n, q_n\})$

**Exercise 1.3.3** Prove ♠ and ♣.

So far, we have considered asymptotic behavior of real sequences. If $a_n$ is a sequence of random variables, we write

$$a_n = O_P(p_n)$$

if for any given number $\epsilon > 0$ there exists a constant $C$ such that

$$\Pr(|a_n| > C p_n) < \epsilon \text{ for all sufficiently large } n.$$

In words, $a_n = O_P(p_n)$ means that the probability that $|a_n|/p_n$ is unbounded is arbitrarily small for large $n$. In particular, a random variable that is bounded in probability is $O_P(1)$. It is very useful to remember that when

we would like to show $a_n = O_P(1)$, we only need to verify $\mathbb{E}|a_n| \leq C < \infty$. Indeed, by Markov's inequality, for any $\kappa > 0$

$$\Pr(|a_n| > \kappa) \leq \frac{C}{\kappa}.$$

Since $\kappa$ is arbitrary, the right-hand side term can be made arbitrarily small.

We can verify that the same results as those in Lemma 1.3.1 hold for $O_P$ as well (proofs are omitted).

Now, letting $\widehat{\theta}_n$ be an estimator of $\theta_0$, when $\widehat{\theta}_n - \theta_0 = O_P(n^{-1/2})$ is satisfied (i.e., $\sqrt{n}(\widehat{\theta}_n - \theta_0) = O_P(1)$), we say that $\widehat{\theta}_n$ is **root-n-consistent** for $\theta_0$. The root-n rate is often also referred to as the "parametric" rate. Most of the estimators introduced in this textbook, except those in Chapter 8, are root-n-consistent.

The probabilistic version of little $o$ is $o_P$. We denote

$$a_n = o_P(p_n)$$

if

$$\frac{|a_n|}{p_n} \xrightarrow{p} 0 \text{ as } n \to \infty.$$

Thus, if $\widehat{\theta}_n$ is a consistent estimator for $\theta_0$, we may write $\widehat{\theta}_n - \theta_0 = o_P(1)$. All the results in Lemma 1.3.2 hold with $O$ and $o$ replaced with $O_P$ and $o_P$, respectively (proofs are omitted).

## 1.4 Convergence in distribution

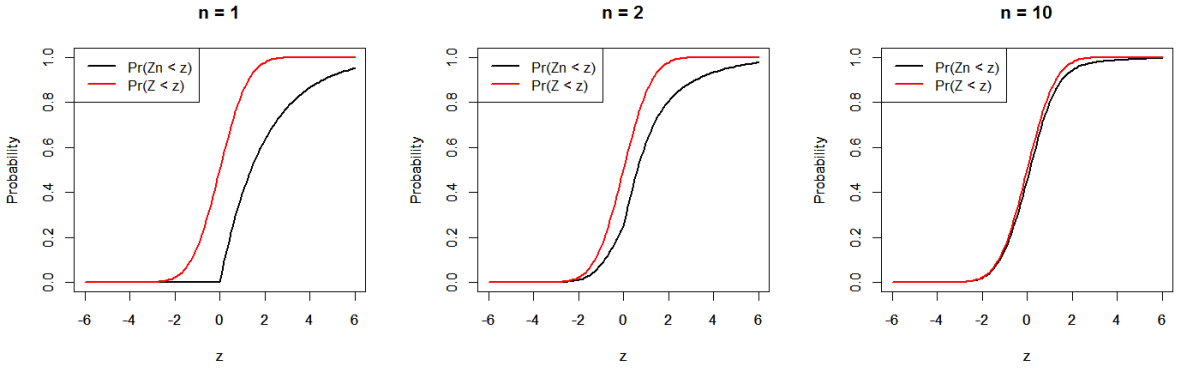
Let $Z_n \in \mathbb{R}$ be a sequence of random variables, and $Z \in \mathbb{R}$ be a random variable with distribution function $F_Z(z) \equiv \Pr(Z \leq z)$.

**Definition 1.4.1** *If*

$$\Pr(Z_n \leq z) \to F_Z(z) \text{ for all } z \in \mathbb{R}$$

*holds, we say that* $Z_n$ ***converges in distribution*** *to* $Z$ *as* $n \to \infty$*, and write* $Z_n \xrightarrow{d} Z$.

When $Z_n \xrightarrow{d} Z$, $F_Z(\cdot)$ is referred to as $Z_n$'s **asymptotic distribution** or **limiting distribution**. Throughout the rest of this section, $Z$ is assumed to be continuous for simplicity of discussion.




Note that convergence in probability is a stronger concept than convergence in distribution. That is, $Z_n \xrightarrow{p} Z$ implies $Z_n \xrightarrow{d} Z$, but the converse is not true. We formally prove the former result below. The proof of the latter is easy by constructing a counterexample.

**Lemma 1.4.2** $Z_n \xrightarrow{p} Z \Longrightarrow Z_n \xrightarrow{d} Z$.

**Proof.** For any positive constant $\kappa > 0$,

$$\begin{aligned}
\Pr(Z_n \le z) &= \Pr(Z_n \le z, \underbrace{|Z_n - Z| \le \kappa}_{Z_n - \kappa \le Z \le Z_n + \kappa}) + \Pr(Z_n \le z, |Z_n - Z| > \kappa) \\
&\le \Pr(Z_n + \kappa \le z + \kappa, Z \le Z_n + \kappa) + \Pr(|Z_n - Z| > \kappa) \\
&= \Pr(Z \le Z_n + \kappa \le z + \kappa) + \Pr(|Z_n - Z| > \kappa) \\
&\le F_Z(z + \kappa) + \Pr(|Z_n - Z| > \kappa).
\end{aligned}$$

Similarly,

$$\begin{aligned}
F_Z(z - \kappa) &= \Pr(Z \le z - \kappa, \underbrace{|Z_n - Z| \le \kappa}_{Z_n - \kappa \le Z \le Z_n + \kappa}) + \Pr(Z \le z - \kappa, |Z_n - Z| > \kappa) \\
&\le \Pr(Z \le z - \kappa, Z_n - \kappa \le Z) + \Pr(|Z_n - Z| > \kappa) \\
&= \Pr(Z_n - \kappa \le Z \le z - \kappa) + \Pr(|Z_n - Z| > \kappa) \\
&\le \Pr(Z_n \le z) + \Pr(|Z_n - Z| > \kappa).
\end{aligned}$$

Combining these inequalities,

$$F_Z(z - \kappa) - \Pr(|Z_n - Z| > \kappa) \le \Pr(Z_n \le z) \le F_Z(z + \kappa) + \Pr(|Z_n - Z| > \kappa). \tag{1.4.1}$$

Since $F_Z(\cdot)$ is assumed to be continuous, by choosing sufficiently small $\kappa > 0$, it holds that

$$\begin{aligned}
F_Z(z + \kappa) &\le F_Z(z) + \epsilon \\
F_Z(z - \kappa) &\ge F_Z(z) - \epsilon
\end{aligned}$$

for a small $\epsilon > 0$. Then, taking the limit $n \to \infty$ of each term in (1.4.1), we have

$$F_Z(z) - \epsilon \le \Pr(Z_n \le z) \le F_Z(z) + \epsilon$$

with probability approaching one as $n \to \infty$. Since $\epsilon$ can be made arbitrarily small, this completes the proof. ■

Let $\{X_1, \ldots, X_n\}$ be an IID sample of $X$ of size $n$ with $\mathbb{E}X = \mu$ and $\mathrm{Var}(X) = \sigma^2 < \infty$. Then, as $n \to \infty$,

$$\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2),$$

where $N(0, \sigma^2)$ is a normal random variable with mean zero and variance $\sigma^2$; in other words,

$$\lim_{n \to \infty} \Pr\left(\sqrt{n}(\overline{X}_n - \mu) \le z\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{z} \exp\left(-\frac{x^2}{2\sigma^2}\right) \mathrm{d}x \quad \text{for all } z \in \mathbb{R}.$$

This result is known as the (Lindeberg-Lévy) **central limit theorem** (CLT). The CLT states that if $n$ is sufficiently large, the distribution of $\sqrt{n}(\overline{X}_n - \mu)$ can be approximated by that of $N(0, \sigma^2)$. The proof of CLT is rather complicated, and thus is omitted here.[5]

[5] In a special case where $X$ has a finite moment of any order, the proof can be greatly simplified by using the properties of the **moment generating function**.

Let $\widehat{\theta}_n$ be an estimator of $\theta_0$ obtained from a sample of size $n$. If $\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{d} \mathbf{N}(\mathbf{0}, \Omega)$, where $\mathbf{N}(\mathbf{0}, \Omega)$ is a multivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix $\Omega$, we say that $\widehat{\theta}_n$ is **asymptotically normally distributed** at root-n rate. $\Omega$ is called the asymptotic covariance matrix of $\sqrt{n}(\widehat{\theta}_n - \theta_0)$.

**Exercise 1.4.3** Prove that $Z_n \xrightarrow{d} Z$ does not imply $Z_n \xrightarrow{p} Z$ by presenting a counterexample.

## 1.5 Asymptotic properties of OLS estimator

Consider a linear regression model:

$$Y = X^\top \beta_0 + \varepsilon, \tag{1.5.1}$$

where $Y$ is a dependent variable, $X$ is a vector of explanatory variables, and $\varepsilon$ is an unobserved random variable (i.e., error term) satisfying $\mathbb{E}[\varepsilon \mid X] = 0$. $\beta_0$ is the true coefficient vector to be estimated. When the data $\{(Y_i, X_i) : 1 \le i \le n\}$ are available, the **ordinary least squares** (OLS) estimator $\widehat{\beta}_n^{ols}$ of $\beta_0$ is given by

$$\begin{aligned} \widehat{\beta}_n^{ols} &= \operatorname*{argmin}_\beta \frac{1}{n}\sum_{i=1}^n (Y_i - X_i^\top \beta)^2 \\ &= \left(\frac{1}{n}\sum_{i=1}^n X_i X_i^\top\right)^{-1} \frac{1}{n}\sum_{i=1}^n X_i Y_i, \end{aligned} \tag{1.5.2}$$

provided that $\frac{1}{n}\sum_{i=1}^n X_i X_i^\top$ is a nonsingular matrix.

Here, note that the conditional expectation function $\mathbb{E}[Y \mid X]$ is a minimizer of the **mean squared error** (MSE) among all possible regression functions:

$$\mathbb{E}[Y \mid X] = \operatorname*{argmin}_{g \in \mathcal{G}} \mathbb{E}[(Y - g(X))^2], \tag{1.5.3}$$

where $\mathcal{G}$ is the set of all measurable functions of $X$. Since we have assumed that $\mathbb{E}[\varepsilon \mid X] = 0$ (i.e., $\mathbb{E}[Y \mid X] = X^\top \beta_0$), we can restrict the above MSE minimization with respect to the class of linear functions of $X$:

$$\begin{aligned} \beta_0 &= \operatorname*{argmin}_{\beta \in \mathbb{R}^{\dim(X)}} \mathbb{E}[(Y - X^\top \beta)^2] \\ &= \mathbb{E}[XX^\top]^{-1}\mathbb{E}[XY]. \end{aligned}$$

Thus, we can see that the OLS objective function in (1.5.2) is the sample analog of the population MSE, and that the estimated regression function, say $\widehat{g}_n(X) \equiv X^\top \widehat{\beta}_n^{ols}$, is an estimator of $\mathbb{E}[Y \mid X]$.[6] By (1.5.1), we can

[6] Another interpretation of the OLS estimator is that it is a sample solution to the moment equation $\mathbb{E}[Y \mid X] = X^\top \beta$. Premultiplying $X$ on both sides of this equation and taking the expectation yield

$$\begin{aligned} \mathbb{E}[X\mathbb{E}[Y \mid X]] = \mathbb{E}[XX^\top]\beta_0 &\iff \mathbb{E}[XY] = \mathbb{E}[XX^\top]\beta_0 \\ &\iff \beta_0 = \mathbb{E}[XX^\top]^{-1}\mathbb{E}[XY]. \end{aligned}$$

Note that, in the above characterization of $\beta_0$, the multiplier does not need to be $X$. As easily confirmed, for any general $h(X)$, we can obtain $\beta_0 = \mathbb{E}[h(X)X^\top]^{-1}\mathbb{E}[h(X)Y]$ as long as $\mathbb{E}[h(X)X^\top]$ exists and is nonsingular.

rewrite the right-hand side of (1.5.2) as

$$\begin{aligned}\widehat{\beta}_n^{ols} &= \left(\frac{1}{n}\sum_{i=1}^n X_i X_i^\top\right)^{-1} \frac{1}{n}\sum_{i=1}^n X_i(X_i^\top \beta_0 + \varepsilon_i) \\ &= \beta_0 + \left(\frac{1}{n}\sum_{i=1}^n X_i X_i^\top\right)^{-1} \frac{1}{n}\sum_{i=1}^n X_i\varepsilon_i.\end{aligned}$$

**Proof of consistency** To prove the consistency of $\widehat{\beta}_n^{ols}$, it is sufficient to show that

$$\left(\frac{1}{n}\sum_{i=1}^n X_i X_i^\top\right)^{-1} \frac{1}{n}\sum_{i=1}^n X_i\varepsilon_i \xrightarrow{p} \mathbf{0}. \tag{1.5.4}$$

Assume that the data are IID and that $\mathbb{E}[XX^\top]$ exists and is nonsingular. Then, by WLLN, $\frac{1}{n}\sum_{i=1}^n X_i X_i^\top \xrightarrow{p} \mathbb{E}[XX^\top]$ element-wisely. Further, by the continuity of matrix inversion, we can show that $\left(\frac{1}{n}\sum_{i=1}^n X_i X_i^\top\right)^{-1} \xrightarrow{p} \mathbb{E}[XX^\top]^{-1}$. In addition, it holds that $\frac{1}{n}\sum_{i=1}^n X_i\varepsilon_i \xrightarrow{p} \mathbf{0}$ since

$$\mathbb{E}[X_i\varepsilon_i] = \mathbb{E}[X_i\mathbb{E}[\varepsilon_i \mid X_i]] = \mathbf{0}$$

by the law of iterated expectations (LIE) and $\mathbb{E}[\varepsilon_i \mid X_i] = 0$. Finally, we can prove (1.5.4) by Lemma B.1.1.

**Proof of root-n-consistency** Suppose that $\mathbb{E}[\varepsilon^2 \mid X] = \sigma^2 < \infty$ (i.e., **homoskedasticity**) for simplicity. Write

$$\begin{aligned}\sqrt{n}(\widehat{\beta}_n^{ols} - \beta_0) &= \left(\frac{1}{n}\sum_{i=1}^n X_i X_i^\top\right)^{-1} \frac{1}{\sqrt{n}}\sum_{i=1}^n X_i\varepsilon_i \\ &= \left(\mathbf{X}_n^\top\mathbf{X}_n/n\right)^{-1}\mathbf{X}_n^\top\mathcal{E}_n/\sqrt{n},\end{aligned}$$

where $\mathbf{X}_n = (X_1, \ldots, X_n)^\top$, and $\mathcal{E}_n = (\varepsilon_1, \ldots, \varepsilon_n)^\top$. The independence and homoskedasticity assumptions imply that $\mathbb{E}[\mathcal{E}_n\mathcal{E}_n^\top \mid \{X_i\}_{i=1}^n] = \sigma^2 I_n$. Hence,

$$\begin{aligned}\mathbb{E}\left[||\sqrt{n}(\widehat{\beta}_n^{ols} - \beta_0)||^2 \mid \{X_i\}_{i=1}^n\right] &= \text{trace}\left\{\left(\mathbf{X}_n^\top\mathbf{X}_n/n\right)^{-1}\mathbf{X}_n^\top\mathbb{E}[\mathcal{E}_n\mathcal{E}_n^\top \mid \{X_i\}_{i=1}^n]\mathbf{X}_n\left(\mathbf{X}_n^\top\mathbf{X}_n/n\right)^{-1}\right\}/n \\ &= \sigma^2\text{trace}\left\{\left(\mathbf{X}_n^\top\mathbf{X}_n/n\right)^{-1}\right\} = O(\dim(X)).\end{aligned}$$

Then, by Markov's inequality, $||\sqrt{n}(\widehat{\beta}_n^{ols} - \beta_0)|| = O_P(1)$.

**Proof of asymptotic normality** By $\left(\frac{1}{n}\sum_{i=1}^n X_i X_i^\top\right)^{-1} \xrightarrow{p} \mathbb{E}[XX^\top]^{-1}$ and Slutsky's theorem B.1.2, the asymptotic distribution of $\sqrt{n}(\widehat{\beta}_n^{ols} - \beta_0)$ is equal to that of $\mathbb{E}[XX^\top]^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n X_i\varepsilon_i$. The mean of $\frac{1}{\sqrt{n}}\sum_{i=1}^n X_i\varepsilon_i$ is zero, and the variance of it is

$$\begin{aligned}\text{Var}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n X_i\varepsilon_i\right) = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n X_i X_j^\top\varepsilon_i\varepsilon_j\right] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n X_i X_i^\top\varepsilon_i^2\right] \\ &= \sigma^2\frac{1}{n}\sum_{i=1}^n\mathbb{E}[X_i X_i^\top] = \sigma^2\mathbb{E}[XX^\top],\end{aligned}$$

where the second equality follows by the independence assumption, and the third equality is by LIE and the homoskedasticity assumption. Hence, by CLT, we have

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n X_i\varepsilon_i \xrightarrow{d} \mathbf{N}(\mathbf{0}, \sigma^2\mathbb{E}[XX^\top]),$$

implying that

$$\sqrt{n}(\widehat{\beta}_n^{ols} - \beta_0) \xrightarrow{d} \mathbf{N}(\mathbf{0}, \sigma^2 \mathbb{E}[XX^\top]^{-1}).$$

**Exercise 1.5.1** Prove (1.5.3).

**Exercise 1.5.2** Suppose now that the error term is heteroskedastic in the following sense: $\mathbb{E}[\varepsilon^2 \mid X] = \sigma^2(X)$. Derive the asymptotic distribution of $\sqrt{n}(\widehat{\beta}_n^{ols} - \beta_0)$.

**Exercise 1.5.3** Consider a linear regression model:

$$Y = X^\top \beta_0 + \varepsilon,$$

where $\mathbb{E}[\varepsilon \mid X] = 0$. Define the **ridge regression** estimator:

$$\widehat{\beta}_n^{ridge} \equiv \left( \frac{1}{n} \sum_{i=1}^n X_i X_i^\top + \lambda I_{\dim(X)} \right)^{-1} \frac{1}{n} \sum_{i=1}^n X_i Y_i,$$

where $\lambda$ is a given positive number. Prove that (1) $\widehat{\beta}_n^{ridge}$ is a biased estimator of $\beta_0$, and that (2) if $\lambda \to 0$, $\widehat{\beta}_n^{ridge}$ is consistent for $\beta_0$.

# Chapter 2

# Two-Stage Least Squares

## 2.1 Endogeneity and exogeneity

Consider the following simple regression model:

$$Y = \beta_0 + X\beta_1 + \varepsilon, \tag{2.1.1}$$

where $Y$ is a dependent variable, $X$ is a scalar explanatory variable, and $\varepsilon$ is an unobserved error term satisfying $\mathbb{E}\varepsilon = 0$. $\beta_0$ and $\beta_1$ are unknown parameters to be estimated. Taking the expectation of both sides of (2.1.1), we obtain $\beta_0 = \mathbb{E}Y - \mathbb{E}[X]\beta_1$. Then, the model (2.1.1) can be rewritten as

$$Y - \mathbb{E}Y = (X - \mathbb{E}X)\beta_1 + \varepsilon. \tag{2.1.2}$$

Further, premultiplying both sides of (2.1.2) by $X$ and taking the expectation of them yield

$$\underbrace{\mathbb{E}[X(Y - \mathbb{E}Y)]}_{\mathrm{Cov}(X,Y)} = \underbrace{\mathbb{E}[X(X - \mathbb{E}X)]}_{\mathrm{Var}(X)}\beta_1 + \underbrace{\mathbb{E}[X\varepsilon]}_{\mathrm{Cov}(X,\varepsilon)}.$$

Assuming that $\mathrm{Var}(X) > 0$, divide the above equation by $\mathrm{Var}(X)$ to obtain

$$\frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)} = \beta_1 + \frac{\mathrm{Cov}(X,\varepsilon)}{\mathrm{Var}(X)}. \tag{2.1.3}$$

Notice that the sample analog of the left-hand side of (2.1.3) is exactly the OLS estimator of the slope parameter $\beta_1$, and under standard conditions we have

$$\widehat{\beta}_{n1}^{ols} \xrightarrow{p} \beta_1 + \frac{\mathrm{Cov}(X,\varepsilon)}{\mathrm{Var}(X)}.$$

This implies that the OLS estimator $\widehat{\beta}_{n1}^{ols}$ cannot correctly (i.e., consistently) estimate $\beta_1$ if $\mathrm{Cov}(X,\varepsilon) \neq 0$. When $\mathrm{Cov}(X,\varepsilon) = 0$ holds true, we say that $X$ is **exogenous**. If it is not satisfied, we say that $X$ is **endogenous**.

Exogeneity: $X$ is uncorrelated with the error term $\varepsilon$.
Endogeneity: $X$ is correlated with the error term $\varepsilon$.

Under exogeneity, the parameter $\beta_1$ can be characterized simply by

$$\beta_1 = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}.$$

On the other hand, if $X$ is endogenous, $\mathrm{Cov}(X,Y)/\mathrm{Var}(X)$ entails a bias, so-called **endogeneity bias**, that amounts to $\mathrm{Cov}(X,\varepsilon)/\mathrm{Var}(X)$. In general, the magnitude and even the sign of the endogeneity bias are uncertain because $\varepsilon$ is unobservable. Hence, in the presence of endogeneity, the OLS estimator is inconsistent, and thus the OLS estimate is uninformative about the true impact of $X$ on $Y$.

There are many potential sources of endogeneity. The two most important sources are from **omitted variables** and from **simultaneity**.

**Omitted variables** Consider the following multiple regression model with two explanatory variables $X_1$ and $X_2$:

$$Y = \beta_0 + X_1\beta_1 + X_2\beta_2 + \varepsilon. \tag{2.1.4}$$

Here, we assume that both $X_1$ and $X_2$ are exogenous: $\mathbb{E}[X_1\varepsilon] = \mathbb{E}[X_2\varepsilon] = 0$. Then, when both $X_1$ and $X_2$ are observable, the slope parameters $\beta_1$ and $\beta_2$ can be consistently estimated simply by regressing $Y$ on $(X_1, X_2)$.

Now consider a case in which $X_2$ is unobservable for some reason. Excluding unobservable $X_2$ from the model (2.1.4), we regress $Y$ only on $X_1$ based on the following simple regression model:

$$Y = \alpha + X_1\beta_1 + \eta, \tag{2.1.5}$$

where $\alpha = \beta_0 + \mathbb{E}[X_2]\beta_2$, and $\eta$ is a new error term defined by

$$\eta = (X_2 - \mathbb{E}X_2)\beta_2 + \varepsilon$$

such that $\mathbb{E}\eta = 0$. In this simple regression model, the explanatory variable $X_2$, which needs to be included in the model if $\beta_2 \neq 0$, is omitted from the analysis. In order to correctly estimate $\beta_1$ based on the model (2.1.5), $X_1$ needs to be an exogenous variable in the sense that $\mathbb{E}[X_1\eta] = 0$. However, by the definition of $\eta$,

$$\begin{aligned}\mathbb{E}[X_1\eta] &= \mathbb{E}\left[X_1\left(X_2 - \mathbb{E}X_2\right)\beta_2 + X_1\varepsilon\right] \\ &= \mathrm{Cov}(X_1, X_2)\beta_2 + \mathrm{Cov}(X_1, \varepsilon) \\ &= \mathrm{Cov}(X_1, X_2)\beta_2.\end{aligned}$$

Hence, unless either $\mathrm{Cov}(X_1, X_2) = 0$ or $\beta_2 = 0$ (or both), $X_1$ is an endogenous variable in the model (2.1.5): $\mathbb{E}[X_1\eta] \neq 0$.

**Simultaneity** As an example, consider estimating the effect of police on crime. Let $X_i$ be the number of police officers in a district $i$ and $Y_i$ be the number of crime incidents in this district. It is expected that there exists a "simultaneity" between police and crime; that is, if crime rate increases, a larger police force will be needed, at the same time, increasing police force reduces crime. This relationship can be expressed as:

$$\begin{aligned}Y_i &= \beta_0 + X_i\beta_1 + \varepsilon_i \quad (\beta_1\text{: \# of police officers} \to \text{\# of crimes}) \\ X_i &= \gamma_0 + Y_i\gamma_1 + u_i \quad (\gamma_1\text{: \# of crimes} \to \text{\# of police officers}),\end{aligned}$$

where $(\gamma_0, \gamma_1)$ is another set of unknown parameters, and $u$ is an unobserved error term. For simplicity of discussion, we do not consider other factors that may influence on $Y$ and $X$. If $X_i$ is an endogenous variable such that $\mathbb{E}[X_i\varepsilon_i] \neq 0$, simply regressing $Y_i$ on $X_i$ does not give us a consistent estimator of $\beta_1$.

According to the second model,

$$\begin{aligned}\mathbb{E}[X_i\varepsilon_i] &= \mathbb{E}[(\gamma_0 + Y_i\gamma_1 + u_i)\varepsilon_i] \\ &= \mathbb{E}[Y_i\varepsilon_i]\gamma_1 + \mathbb{E}[u_i\varepsilon_i].\end{aligned}$$

For the second term on the right-hand side, if there are unobservable regional factors that can affect both $X$ and $Y$, we would have $\mathbb{E}[u_i\varepsilon_i] \neq 0$. Further, the first term is necessarily non-zero as long as $\gamma_1 \neq 0$ (i.e., simultaneity exists) because

$$\begin{aligned}\mathbb{E}[Y_i\varepsilon_i] &= \mathbb{E}[(\beta_0 + X_i\beta_1 + \varepsilon_i)\varepsilon_i] \\ &= \mathbb{E}[X_i\varepsilon_i]\beta_1 + \underbrace{\mathbb{E}[\varepsilon_i^2]}_{>0}.\end{aligned}$$

Thus, endogeneity problem arises when simultaneity exists.

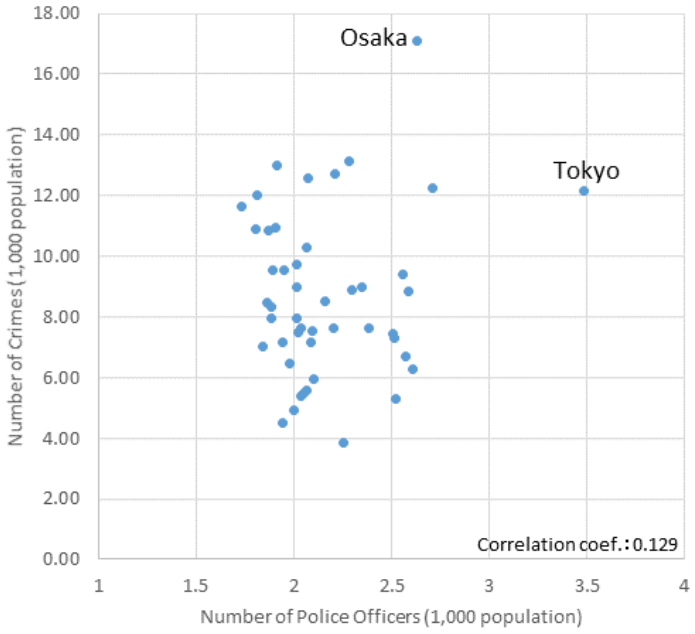



Figure 2.1: Simultaneity between police and crime.

**Exercise 2.1.1** Give specific examples of endogeneity bias caused by (1) omitted variables and (2) simultaneity (other than those mentioned above).

## 2.2 Instrumental variables

A common approach to account for the endogeneity problem is to use **instrumental variables** (IV). Again, consider the following simple regression model:

$$Y = \beta_0 + X\beta_1 + \varepsilon.$$

Here, we assume that $X$ is endogenous: $\mathbb{E}[X\varepsilon] \neq 0$. An instrumental variable for $X$, which we denote by $Z$, is a random variable that satisfies the following conditions:

$$\begin{aligned} &\mathbb{E}[Z\varepsilon] = \mathrm{Cov}(Z,\varepsilon) = 0 && \textbf{Exclusion restriction} \\ &\mathrm{Cov}(Z,X) \neq 0 && \textbf{Relevance condition} \end{aligned}$$

The exclusion restriction says that the instrumental variable $Z$ is uncorrelated with the error term $\varepsilon$. This implies that $Z$ is not a direct determinant of $Y$; if $Z$ directly affects $Y$, the information about $Z$ must be contained in $\varepsilon$, leading to a correlation of $Z$ and $\varepsilon$ (this is why the condition is called "exclusion" restriction). The relevance condition requires that the instrumental variable $Z$ is a determinant of the endogenous variable $X$.

**Example 2.2.1** As a typical econometric application, suppose that $Y$ is an individual's wage and $X$ is his education in years. It would be expected that the higher the individual's ability, the higher his wage. However, since one's ability is hard to measure correctly, the effect of ability on $Y$ will be partly included in the error term $\varepsilon$. Also, there must exist a positive correlation between one's ability and education. Consequently, a positive correlation between $X$ and $\varepsilon$ arises naturally; that is, $X$ should be treated as an endogenous variable. Here, let $Z$ be the individual's parents' education level. It would be legitimate to suppose that the parents' education does not "directly" affect his own wage $Y$ (but influences on $Y$ only through $X$), and it is surely correlated with his education level. Thus, the parents' education level can be a valid instrument for $X$.

**Example 2.2.2** Consider once again the simultaneous relationship of police and crime. Here, let $Y_{it}$ and $X_{it}$ be the number of crime incidents and the size of police force in city $i$ year $t$, respectively. As an instrumental variable for $X_{it}$, [Levitt, 1997] employed an indicator variable $Z_{it}$ which takes one if $t$ is the mayoral election year in city $i$. In years when there is an election, the incumbent mayors would have an incentive to increase the police force to reduce crime (which was indeed confirmed as statistically significant in [Levitt, 1997]). Also, the presence of elections should not have "direct" impacts on the number of crimes. Thus, $Z_{it}$ is considered to be a valid instrument for $X_{it}$.

Now, suppose that we have a valid instrumental variable $Z$ satisfying both the exclusion restriction and relevance condition. Then, premultiplying both sides of (2.1.2) by $Z$ and taking the expectation of them yield

$$\underbrace{\mathbb{E}[Z(Y-\mathbb{E}Y)]}_{\mathrm{Cov}(Z,Y)} = \underbrace{\mathbb{E}[Z(X-\mathbb{E}X)]}_{\mathrm{Cov}(Z,X)\neq 0}\beta_1 + \underbrace{\mathbb{E}[Z\varepsilon]}_{\mathrm{Cov}(Z,\varepsilon)=0}.$$

Thus, we obtain

$$\beta_1 = \frac{\mathrm{Cov}(Z,Y)}{\mathrm{Cov}(Z,X)}.$$

Here, suppose that the sample $\{(Y_i, X_i, Z_i) : 1 \leq i \leq n\}$ is available, and let $\widehat{\mathrm{Cov}}_n(Z,Y)$ and $\widehat{\mathrm{Cov}}_n(Z,X)$ be the sample covariances of $(Z,Y)$ and $(Z,X)$, respectively. Then, the above result implies that $\beta_1$ can be estimated

by

$$\widehat{\beta}_{n1} \equiv \frac{\widehat{\mathrm{Cov}}_n(Z,Y)}{\widehat{\mathrm{Cov}}_n(Z,X)}.$$

Further, note that

$$\begin{aligned}
\widehat{\mathrm{Cov}}_n(Z,Y) &= \frac{1}{n}\sum_{i=1}^{n} Z_i(Y_i - \overline{Y}_n) \\
&= \frac{1}{n}\sum_{i=1}^{n} Z_i((\beta_0 + X_i\beta_1 + \varepsilon_i) - (\beta_0 + \overline{X}_n\beta_1 + \overline{\varepsilon}_n)) \\
&= \frac{1}{n}\sum_{i=1}^{n} Z_i((X_i - \overline{X}_n)\beta_1 + (\varepsilon_i - \overline{\varepsilon}_n)) \\
&= \widehat{\mathrm{Cov}}_n(Z,X)\beta_1 + \widehat{\mathrm{Cov}}_n(Z,\varepsilon),
\end{aligned}$$

where $\overline{Y}_n$, $\overline{X}_n$, and $\overline{\varepsilon}_n$ are the sample averages of $Y$, $X$, and $\varepsilon$, respectively. Hence,

$$\widehat{\beta}_{n1} = \beta_1 + \frac{\widehat{\mathrm{Cov}}_n(Z,\varepsilon)}{\widehat{\mathrm{Cov}}_n(Z,X)}. \tag{2.2.1}$$

The second term on the right-hand side of (2.2.1) is the finite sample estimation bias.[1] If $\mathrm{Cov}(Z,X)$ is close to zero, i.e., if $Z$ is almost irrelevant to $X$, $\widehat{\mathrm{Cov}}_n(Z,X)$ is also close to zero. Thus, in this case, the estimation bias of $\widehat{\beta}_{n1}$ can be very large. This problem is known as the **weak instruments problem**. There are several formal tests for weak instruments (see, e.g., [Stock and Yogo, 2005]); however, the details are omitted here.

**Exercise 2.2.3** Suppose that we would like to estimate the causal impact of aircraft noise on land prices by a simple regression analysis. In this regression model, we should suspect that the aircraft noise variable is endogenous. Explain why, and give an example of a valid instrumental variable in this analysis.

## 2.3 2SLS procedure

In the following, we consider a more general case with multiple regressors and multiple instrumental variables:

$$Y = D \cdot \alpha_0 + X^\top \beta_0 + \varepsilon,$$

where $X = (X_1, \ldots, X_{d_x})^\top$ is a vector of exogenous explanatory variables (including a constant term), and $D$ is an endogenous explanatory variable. Suppose that we have a vector of instrumental variables $Z_1 = (Z_{11}, \ldots, Z_{1d_z})^\top$ for $D$.

We modify the above estimation procedure for a simple regression model as follows. First, let $Z = (Z_1^\top, X^\top)^\top$ be a vector of all exogenous variables, and assume that the inverse matrix $\mathbb{E}[ZZ^\top]^{-1}$ exists. Further, let $\mathcal{P}_Z(A)$ be the orthogonal projection of $A$ onto $Z$; that is, $\mathcal{P}_Z(A) \equiv Z^\top \mathbb{E}[ZZ^\top]^{-1}\mathbb{E}[ZA]$. When $A$ is a vector $A = (A_1, \ldots A_k)^\top$, $\mathcal{P}_Z(A)$ is defined as

$$\mathcal{P}_Z(A) = (Z^\top \mathbb{E}[ZZ^\top]^{-1}\mathbb{E}[ZA_1], \ldots, Z^\top \mathbb{E}[ZZ^\top]^{-1}\mathbb{E}[ZA_k])^\top.$$

[1] Thus, this IV-based estimator is not an unbiased estimator.

Rewrite the model as follows:

$$\begin{aligned} Y &= \mathcal{P}_Z(D)\cdot\alpha_0 + X^\top\beta_0 + u \\ &= \mathcal{P}_Z(H)^\top\theta_0 + u, \end{aligned} \tag{2.3.1}$$

where $H = (D, X^\top)^\top$, $\theta_0 = (\alpha_0, \beta_0^\top)^\top$, and $u = (D - \mathcal{P}_Z(D))\cdot\alpha_0 + \varepsilon$. The second equality in (2.3.1) is due to the fact that $\mathcal{P}_Z(X) = X$ (see the supplementary material in this chapter). By the exclusion restriction and (2.7.1) below, we can find that

$$\begin{aligned} \mathbb{E}[\mathcal{P}_Z(H)u] &= \mathbb{E}\begin{bmatrix} \mathcal{P}_Z(D)u \\ Xu \end{bmatrix} \\ &= \mathbb{E}\begin{bmatrix} \mathcal{P}_Z(D)(D - \mathcal{P}_Z(D)) \\ X(D - \mathcal{P}_Z(D)) \end{bmatrix}\cdot\alpha_0 = \underset{(1+d_x)\times 1}{\mathbf{0}}. \end{aligned}$$

This result implies that $\theta_0$ can be estimated by the least squares regression of $Y$ on $(\mathcal{P}_Z(D), X)$, or equivalently, on $\mathcal{P}_Z(H)$. Figure 2.2 provides a simple geometry of the above discussion.

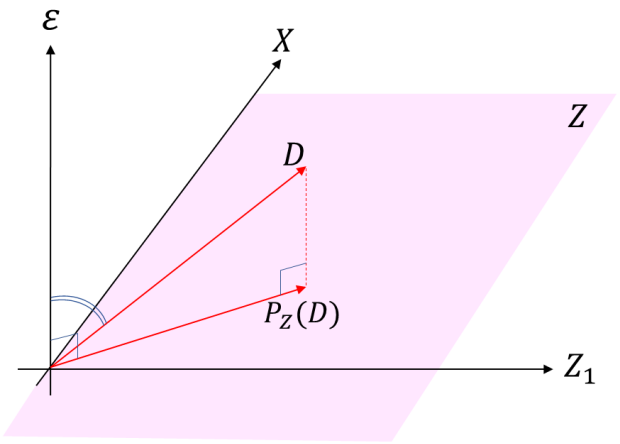



Figure 2.2: Orthogonal projection of $D$ onto $Z = (Z_1, X)$.

Now, suppose that we have data of $n$ observations $\{(Y_i, D_i, X_i, Z_{1i}) : 1 \le i \le n\}$. For a sample analog to the above procedure, the parameter $\theta_0$ can be estimated in the following two steps:

**Step (1)**: perform a least squares regression of $D$ on $Z = (Z_1^\top, X^\top)^\top$, and compute the predicted value of $D$ by $\widehat{D} = Z^\top\widehat{\gamma}_n$, where

$$\widehat{\gamma}_n \equiv \left(\frac{1}{n}\sum_{i=1}^n Z_i Z_i^\top\right)^{-1}\frac{1}{n}\sum_{i=1}^n Z_i D_i.$$

(At this step, it is important to check the significance of $\widehat{\gamma}_n$. If all elements of $\widehat{\gamma}_n$ are close to zero, the weak instruments problem can occur.) Let $\widehat{H} = (\widehat{D}, X^\top)^\top$.

**Step (2)**: perform a least squares regression of $Y$ on $\widehat{H}$ to estimate $\theta_0$:

$$\widehat{\theta}_n^{2sls} \equiv \left(\frac{1}{n}\sum_{i=1}^n \widehat{H}_i\widehat{H}_i^\top\right)^{-1}\frac{1}{n}\sum_{i=1}^n \widehat{H}_i Y_i.$$

The above estimation procedure is called the **two-stage least squares** (2SLS) estimator.

Noting the second equality in (2.3.1), the 2SLS estimator can be, in fact, implemented in one step. To do so, we introduce the following matrix notations: $\mathbf{Y}_n = (Y_1, \ldots, Y_n)^\top$, $\mathbf{D}_n = (D_1, \ldots, D_n)^\top$, $\mathbf{X}_n = (X_1, \ldots, X_n)^\top$, $\mathbf{Z}_n = (Z_1, \ldots, Z_n)^\top$, and $\mathbf{H}_n = (H_1, \ldots, H_n)^\top$. Further, let $\mathcal{P}_{\mathbf{Z}_n}$ be an empirical projection matrix, which is defined by

$$\mathcal{P}_{\mathbf{Z}_n} \equiv \mathbf{Z}_n(\mathbf{Z}_n^\top \mathbf{Z}_n)^{-1}\mathbf{Z}_n^\top.$$

Then, $\widehat{\mathbf{D}}_n = (\widehat{D}_1, \ldots, \widehat{D}_n)^\top$ is obtained by $\widehat{\mathbf{D}}_n = \mathcal{P}_{\mathbf{Z}_n}\mathbf{D}_n$. In addition, by the same argument as the fact $\mathcal{P}_Z(X) = X$, we can easily see that $\mathbf{X}_n = \mathcal{P}_{\mathbf{Z}_n}\mathbf{X}_n$. Therefore, $\widehat{\mathbf{H}}_n \equiv [\widehat{\mathbf{D}}_n \,;\, \mathbf{X}_n] = \mathcal{P}_{\mathbf{Z}_n}\mathbf{H}_n$. Consequently, noting the fact that $\mathcal{P}_{\mathbf{Z}_n}\mathcal{P}_{\mathbf{Z}_n} = \mathcal{P}_{\mathbf{Z}_n}$,[2] the 2SLS estimator $\widehat{\theta}_n^{2sls}$ can be numerically equivalently expressed as

$$\begin{aligned}
\widehat{\theta}_n^{2sls} &= \left(\widehat{\mathbf{H}}_n^\top \widehat{\mathbf{H}}_n\right)^{-1} \widehat{\mathbf{H}}_n^\top \mathbf{Y}_n \\
&= \left(\mathbf{H}_n^\top \mathcal{P}_{\mathbf{Z}_n}\mathcal{P}_{\mathbf{Z}_n}\mathbf{H}_n\right)^{-1} \mathbf{H}_n^\top \mathcal{P}_{\mathbf{Z}_n}\mathbf{Y}_n \\
&= \left(\mathbf{H}_n^\top \mathbf{Z}_n(\mathbf{Z}_n^\top \mathbf{Z}_n)^{-1}\mathbf{Z}_n^\top \mathbf{H}_n\right)^{-1} \mathbf{H}_n^\top \mathbf{Z}_n(\mathbf{Z}_n^\top \mathbf{Z}_n)^{-1}\mathbf{Z}_n^\top \mathbf{Y}_n.
\end{aligned} \tag{2.3.2}$$

## 2.4 Asymptotic properties

The asymptotic properties of the 2SLS estimator (2.3.2) can be relatively easily derived. First, noting that $\mathbf{Y}_n = \mathbf{H}_n^\top \theta_0 + \mathcal{E}_n$, where $\mathcal{E}_n = (\varepsilon_1, \ldots, \varepsilon_n)^\top$, we can obtain

$$\widehat{\theta}_n^{2sls} = \theta_0 + \left[\mathbf{H}_n^\top \mathbf{Z}_n(\mathbf{Z}_n^\top \mathbf{Z}_n)^{-1}\mathbf{Z}_n^\top \mathbf{H}_n\right]^{-1} \mathbf{H}_n^\top \mathbf{Z}_n(\mathbf{Z}_n^\top \mathbf{Z}_n)^{-1}\mathbf{Z}_n^\top \mathcal{E}_n.$$

by WLLN, under standard conditions, we have $\mathbf{H}_n^\top \mathbf{Z}_n/n \xrightarrow{p} \mathbb{E}[HZ^\top]$, $\mathbf{Z}_n^\top \mathbf{Z}_n/n \xrightarrow{p} \mathbb{E}[ZZ^\top]$, and $\mathbf{Z}_n^\top \mathcal{E}_n/n \xrightarrow{p} \mathbb{E}[Z\varepsilon] = \underset{(d_z+d_x)\times 1}{\mathbf{0}}$ as the sample size increases to infinity. Thus,

$$\begin{aligned}
\widehat{\theta}_n^{2sls} - \theta_0 &= \left[(\mathbf{H}_n^\top \mathbf{Z}_n/n)(\mathbf{Z}_n^\top \mathbf{Z}_n/n)^{-1}(\mathbf{Z}_n^\top \mathbf{H}_n/n)\right]^{-1} (\mathbf{H}_n^\top \mathbf{Z}_n/n)(\mathbf{Z}_n^\top \mathbf{Z}_n/n)^{-1}(\mathbf{Z}_n^\top \mathcal{E}_n/n) \\
&\xrightarrow{p} \left[\mathbb{E}[HZ^\top]\mathbb{E}[ZZ^\top]^{-1}\mathbb{E}[ZH^\top]\right]^{-1} \mathbb{E}[HZ^\top]\mathbb{E}[ZZ^\top]^{-1} \underbrace{\mathbb{E}[Z\varepsilon]}_{=\mathbf{0}} = \mathbf{0}
\end{aligned}$$

by Lemma B.1.1 and the continuous mapping theorem, implying that the 2SLS estimator $\widehat{\theta}_n^{2sls}$ is consistent for $\theta_0$.

Next, suppose for simplicity that the error terms $\varepsilon_i$'s are IID with variance $\mathbb{E}[\varepsilon^2] = \sigma^2$. Further, assume that $\varepsilon$ is independent of $Z$ (note that this is a stronger condition than the uncorrelation between $\varepsilon$ and $Z$). Then, by CLT, it would hold that $\mathbf{Z}_n^\top \mathcal{E}_n/\sqrt{n} \xrightarrow{d} \mathbf{N}(\mathbf{0}, \mathbb{E}[ZZ^\top]\sigma^2)$. Hence, by Slutsky's theorem B.1.2, we obtain the asymptotic normality of $\sqrt{n}(\widehat{\theta}_n^{2sls} - \theta_0)$ as follows:

$$\begin{aligned}
\sqrt{n}(\widehat{\theta}_n^{2sls} - \theta_0) &= \left[(\mathbf{H}_n^\top \mathbf{Z}_n/n)(\mathbf{Z}_n^\top \mathbf{Z}_n/n)^{-1}(\mathbf{Z}_n^\top \mathbf{H}_n/n)\right]^{-1} (\mathbf{H}_n^\top \mathbf{Z}_n/n)(\mathbf{Z}_n^\top \mathbf{Z}_n/n)^{-1}\mathbf{Z}_n^\top \mathcal{E}_n/\sqrt{n} \\
&\xrightarrow{d} \mathbf{N}\left(\mathbf{0}, \sigma^2\left[\mathbb{E}[HZ^\top]\mathbb{E}[ZZ^\top]^{-1}\mathbb{E}[ZH^\top]\right]^{-1}\right).
\end{aligned}$$

**Exercise 2.4.1** Explain what happens if we create $\widehat{D}$ using only $Z_1$ without $X$ and perform a 2SLS estimation. Is such an estimator consistent? If yes, prove that, and if no, explain under what additional condition(s) the estimator becomes consistent.

[2] This property is called *idempotency*.

## 2.5 Durbin-Wu-Hausman test for endogeneity

Note that the 2SLS estimator is consistent and asymptotically normal even when the variable $D$ is in fact exogenous. Of course, if $D$ is an exogenous variable, we can use the OLS estimator, and, in general, the OLS estimator is more efficient than the 2SLS estimator when both estimators are available. Thus, statistically testing the appropriateness of using OLS or the necessity of using 2SLS is useful.

Here, we consider the following null hypothesis

$$\mathbb{H}_0 : \mathrm{Cov}(D, \varepsilon) = 0,$$

and the alternative hypothesis is its negation

$$\mathbb{H}_1 : \mathrm{Cov}(D, \varepsilon) \neq 0.$$

Under the null hypothesis $\mathbb{H}_0$, both OLS and 2SLS estimators of $\alpha_0$, say $\widehat{\alpha}_n^{ols}$ and $\widehat{\alpha}_n^{2sls}$, respectively, are consistent and asymptotically normal. However under $\mathbb{H}_1$, the OLS estimator $\widehat{\alpha}_n^{ols}$ converges to $\alpha_0 +$ endogeneity bias, while the 2SLS estimator is still consistent for $\alpha_0$. Thus the difference between the OLS and 2SLS estimators is a valid test statistic for endogeneity.

Suppose that $\mathbb{H}_0$ is true. Then, since both $\widehat{\alpha}_n^{ols}$ and $\widehat{\alpha}_n^{2sls}$ are asymptotically distributed as normal, we can show that the difference of them $\widehat{\alpha}_n^{ols} - \widehat{\alpha}_n^{2sls}$ is also asymptotically normal. Thus, letting

$$T_n \equiv \frac{\widehat{\alpha}_n^{ols} - \widehat{\alpha}_n^{2sls}}{\sqrt{\mathrm{Var}(\widehat{\alpha}_n^{ols} - \widehat{\alpha}_n^{2sls})}},$$

we have $T_n \xrightarrow{d} N(0,1)$ and also $T_n^2 \xrightarrow{d} \chi^2(1)$ under $\mathbb{H}_0$, where $\chi^2(1)$ denotes the chi-square distribution with one degree of freedom. Thus, if $T_n^2 > 3.841\ldots$ for instance, we can reject $\mathbb{H}_0$ at the 5% significance level.

In a more general setting where $q$ endogenous regressors $D = (D_1, \ldots, D_q)^\top$ exist, the statistic $T_n$ can be defined as $T_n \equiv \left[\mathrm{Cov}(\widehat{\alpha}_n^{ols} - \widehat{\alpha}_n^{2sls})\right]^{-1/2} (\widehat{\alpha}_n^{ols} - \widehat{\alpha}_n^{2sls})$, and it holds that $T_n^\top T_n \xrightarrow{d} \chi^2(q)$. The above testing procedure for endogeneity is called the **Durbin-Wu-Hausman test**.

## 2.6 A causal interpretation of the 2SLS estimate

In this section, for simplicity, suppose that we have a dummy endogenous variable $D \in \{0,1\}$ and a dummy instrumental variable $Z \in \{0,1\}$ only, with no other covariates. Here, we are interested in estimating the "causal" impact of $D$ on $Y$, rather than a mere correlation between them.

For each $d \in \{0,1\}$, define the **potential outcome** $Y(d)$ as the outcome that will be obtained when $D = d$. Note that we never observe outcomes for the same individual under different treatment status at the same time. That is, when $D = 1$ (resp. $D = 0$) is realized, we can observe only $Y(1)$ (resp. $Y(0)$), but $Y(0)$ (resp. $Y(1)$) is an unobservable "counterfactual" outcome. Thus, the "observed" outcome $Y$ can be written as

$$Y = D \cdot Y(1) + (1 - D) \cdot Y(0).$$

In this framework, the causal effect of $D$ on $Y$ is defined as

$$Y(1) - Y(0).$$

This quantity is referred to as the **treatment effect**, and $D$ is called the **treatment variable** in this context. In this way, we can define the causal effect of the treatment $D$ on the outcome $Y$ as the difference of the potential outcomes $Y(0)$ and $Y(1)$. This approach of defining causal relationship between variables is called the **Rubin's Causal Model**. Since we cannot observe both potential outcomes $Y_i(0)$ and $Y_i(1)$ at the same time (unless there is someone completely identical to $i$), it is generally impossible to estimate the individual-specific treatment effects. Therefore, studies on treatment effects typically aim at estimating treatment effects averaged over some subpopulation; for example,

$$\text{ATE (the average treatment effect)} : \mathbb{E}[Y(1) - Y(0)]$$
$$\text{CATE (the conditional average treatment effect)} : \mathbb{E}[Y(1) - Y(0) \mid X = x]$$
$$\text{ATET (the average treatment effect on the treated)} : \mathbb{E}[Y(1) - Y(0) \mid D = 1].$$

In a seminal paper, [Imbens and Angrist, 1994] studied conditions under which the 2SLS estimate can be interpreted as the **LATE** (local average treatment effect) – the average treatment effect for those whose treatment status is altered by the instrument. Note that when both $D$ and $Z$ are dummy variables, we can classify individuals into the following four latent types:

$$\begin{aligned}
&\textbf{Complier:} && Z = 0 \Rightarrow D = 0 \ \ \& \ \ Z = 1 \Rightarrow D = 1 \\
&\textbf{Defier:} && Z = 0 \Rightarrow D = 1 \ \ \& \ \ Z = 1 \Rightarrow D = 0 \\
&\textbf{Always-taker:} && Z = 0 \Rightarrow D = 1 \ \ \& \ \ Z = 1 \Rightarrow D = 1 \\
&\textbf{Never-taker:} && Z = 0 \Rightarrow D = 0 \ \ \& \ \ Z = 1 \Rightarrow D = 0
\end{aligned}$$

For example, consider estimating the causal effect of college degree ($D = 1$: college graduate or higher, $D = 0$: otherwise) on annual income ($Y$). As an instrumental variable for $D$, we consider father's education level ($Z = 1$: the individual's father has a college degree or higher, $Z = 0$: otherwise). Then, the compliers are those who will (not) go on to a college if their fathers (do not) have a college degree, the always-takers are those who will go on to a college irrespective of their fathers' education level, and so forth. If we can assume that there are no defiers, we can distinguish the latent types from the observed data in the following sense:

| | | $Z$ | |
|---|---|---|---|
| | | 0 | 1 |
| $D$ | 0 | Complier or Never-Taker | Never-taker |
| | 1 | Always-taker | Complier or Always-taker |

The assumption of no defiers is called **monotonicity assumption** because, when no defiers exist, increasing $Z$ from 0 to 1 never decreases $D$. In addition, assume that $Z$ is independent of the potential outcomes (i.e., the exclusion restriction) and the individuals' types such that

$$\Pr(\text{Always-taker} \mid Z = 1) = \Pr(\text{Always-taker} \mid Z = 0) = \pi_A$$
$$\Pr(\text{Never-taker} \mid Z = 1) = \Pr(\text{Never-taker} \mid Z = 0) = \pi_N$$
$$\Pr(\text{Complier} \mid Z = 1) = \Pr(\text{Complier} \mid Z = 0) = \pi_C.$$

Under these conditions, one can show that

$$\begin{aligned}\underset{\textbf{LATE}}{\mathbb{E}[Y(1)-Y(0)\mid \text{Complier}]} &= \frac{\mathbb{E}[Y\mid Z=1]-\mathbb{E}[Y\mid Z=0]}{\mathbb{E}[D\mid Z=1]-\mathbb{E}[D\mid Z=0]}\\ &= \frac{\text{Cov}(Z,Y)}{\text{Cov}(Z,D)}.\end{aligned} \tag{2.6.1}$$

The term on the left-hand side of (2.6.1) is called **LATE** (the local average treatment effect). The first equality of (2.6.1) can be shown as follows. Observe that

$$\begin{aligned}\mathbb{E}[Y\mid Z=1] &= \mathbb{E}[Y\mid Z=1,\text{AT}]\cdot\Pr(\text{AT}\mid Z=1)+\mathbb{E}[Y\mid Z=1,\text{NT}]\cdot\Pr(\text{NT}\mid Z=1)+\mathbb{E}[Y\mid Z=1,\text{C}]\cdot\Pr(\text{C}\mid Z=1)\\ &= \mathbb{E}[Y(1)\mid Z=1,\text{AT}]\cdot\pi_A+\mathbb{E}[Y(0)\mid Z=1,\text{NT}]\cdot\pi_N+E[Y(1)\mid Z=1,\text{C}]\cdot\pi_C\\ &= \mathbb{E}[Y(1)\mid \text{AT}]\cdot\pi_A+E[Y(0)\mid \text{NT}]\cdot\pi_N+E[Y(1)\mid \text{C}]\cdot\pi_C,\end{aligned}$$

where the last equality follows from the exclusion restriction. Similarly, we can show that

$$\mathbb{E}[Y\mid Z=0]=\mathbb{E}[Y(1)\mid \text{AT}]\cdot\pi_A+\mathbb{E}[Y(0)\mid \text{NT}]\cdot\pi_N+\mathbb{E}[Y(0)\mid \text{C}]\cdot\pi_C$$

Thus,

$$\begin{aligned}\mathbb{E}[Y\mid Z=1]-E[Y\mid Z=0] &= \{\mathbb{E}[Y(1)\mid \text{Complier}]-\mathbb{E}[Y(0)\mid \text{Complier}]\}\cdot\pi_C\\ &= \textbf{LATE}\cdot\pi_C\end{aligned}$$

Here, since we have assumed that there are no defiers, we must have

$$\begin{aligned}\mathbb{E}[D\mid Z=1] &= \pi_C+\pi_A\\ \mathbb{E}[D\mid Z=0] &= \pi_A.\end{aligned}$$

Therefore, it holds that

$$\mathbb{E}[D\mid Z=1]-\mathbb{E}[D\mid Z=0]=\pi_C.$$

Finally, combining these results gives the desired equality, provided that $\pi_C>0$.

The second equality of (2.6.1) is left for an exercise. Recall that the slope coefficient $\alpha_0$ in the regression model $Y=\beta_0+D\cdot\alpha_0+\varepsilon$ can be expressed as $\alpha_0=\text{Cov}(Z,Y)/\text{Cov}(Z,D)$. This means that the 2SLS estimator for $\alpha_0$ can be seen as an estimator of the LATE parameter.

**Exercise 2.6.1** Prove the following equality: $\dfrac{\text{Cov}(Z,Y)}{\text{Cov}(Z,D)}=\dfrac{\mathbb{E}[Y\mid Z=1]-\mathbb{E}[Y\mid Z=0]}{\mathbb{E}[D\mid Z=1]-\mathbb{E}[D\mid Z=0]}$.

## 2.7 Numerical simulations with R

In this section, we see how the endogeneity problem is serious and can be solved by the method of instrumental variables. We first generate data of sample size 500 as follows.

```
N   <- 500      # sample size
X   <- rnorm(N) # exogenous regressor
```

```
err <- rnorm(N) # error term

Z1 <- runif(N, -1, 1)  # instrumental variable
D  <- 1 + Z1 + 0.5*err # endogenous regressor
Y  <- 1 + D + X + err  # outcome variable
```

We create the variable $D$ so that it is strongly correlated with the error term; i.e., $D$ is endogenous. The variable $Z_1$ is an instrument for $D$.

```
>  cov(D, err)  # endogeneity
[1] 0.5744865
>  cov(Z1, err) # exclusion restriction
[1] 0.03393532
>  cov(Z1, D)   # relevance condition
[1] 0.3617284
```

First, we run an OLS regression without accounting for the endogeneity of $D$. The result is as follows.

```
> # OLS estimation #
>
>  lm(Y ~ D + X)

Call:
lm(formula = Y ~ D + X)

Coefficients:
(Intercept)            D            X
0.1229       1.8853       0.9780
```

Recalling that the true coefficient of $D$ is one, we can see that the OLS coefficient estimate for $D$ is severely biased. Next, we implement the 2SLS estimation. Following (2.3.2), we can run the 2SLS regression in the following manner.

```
# 2SLS estimation #

H <- cbind(1,D, X)
Z <- cbind(1,Z1,X)

Proj  <- Z%*%solve(t(Z)%*%Z)%*%t(Z)
theta <- solve(t(H)%*%Proj%*%H)%*%t(H)%*%Proj%*%Y

>  theta
[,1]
0.9056225
D 1.0944848
X 0.9809257
```

Then, we can observe that the 2SLS estimator performs accurately. However, as mentioned above, the performance of the 2SLS estimator is crucially dependent on the "strength" of the instruments. To see this, we recreate $D$ so that the correlation between $D$ and $Z_1$ is much weaker than the current case.

```
# Weak instruments problem #
```

```
D  <- 1 + 0.2*Z1 + 0.5*err # endogenous regressor
Y  <- 1 + D + X + err      # outcome variable

>  cov(Z1, D) # relevance condition
[1] 0.08591981
```

```
# 2SLS estimation #

H <- cbind(1,D, X)
Z <- cbind(1,Z1,X)

Proj  <- Z%*%solve(t(Z)%*%Z)%*%t(Z)
theta <- solve(t(H)%*%Proj%*%H)%*%t(H)%*%Proj%*%Y

>  theta
[,1]
0.6027511
D 1.3973392
X 0.9839573
```

As can be seen from this result, the estimation performance of the 2SLS is clearly deteriorated when the instrument is weak.

## Appendix: A property of the orthogonal projection

Using the formula in (B.4.1), denoting $\mathbb{E}_{AB} \equiv \mathbb{E}[AB^\top]$, we have

$$
\begin{aligned}
\mathbb{E}[ZZ^\top]^{-1}\mathbb{E}[ZX^\top] &= \begin{pmatrix} \mathbb{E}_{Z_1Z_1} & \mathbb{E}_{Z_1X} \\ \mathbb{E}_{XZ_1} & \mathbb{E}_{XX} \end{pmatrix}^{-1} \begin{pmatrix} \mathbb{E}_{Z_1X} \\ \mathbb{E}_{XX} \end{pmatrix} \\
&= \begin{pmatrix} (\mathbb{E}_{Z_1Z_1} - \mathbb{E}_{Z_1X}\mathbb{E}_{XX}^{-1}\mathbb{E}_{XZ_1})^{-1} & -(\mathbb{E}_{Z_1Z_1} - \mathbb{E}_{Z_1X}\mathbb{E}_{XX}^{-1}\mathbb{E}_{XZ_1})^{-1}\mathbb{E}_{Z_1X}\mathbb{E}_{XX}^{-1} \\ -(\mathbb{E}_{XX} - \mathbb{E}_{XZ_1}\mathbb{E}_{Z_1Z_1}^{-1}\mathbb{E}_{Z_1X})^{-1}\mathbb{E}_{XZ_1}\mathbb{E}_{Z_1Z_1}^{-1} & (\mathbb{E}_{XX} - \mathbb{E}_{XZ_1}\mathbb{E}_{Z_1Z_1}^{-1}\mathbb{E}_{Z_1X})^{-1} \end{pmatrix} \begin{pmatrix} \mathbb{E}_{Z_1X} \\ \mathbb{E}_{XX} \end{pmatrix} \\
&= \begin{pmatrix} \mathbf{0}_{d_z \times d_x} \\ -(\mathbb{E}_{XX} - \mathbb{E}_{XZ_1}\mathbb{E}_{Z_1Z_1}^{-1}\mathbb{E}_{Z_1X})^{-1}\mathbb{E}_{XZ_1}\mathbb{E}_{Z_1Z_1}^{-1}\mathbb{E}_{Z_1X} + (\mathbb{E}_{XX} - \mathbb{E}_{XZ_1}\mathbb{E}_{Z_1Z_1}^{-1}\mathbb{E}_{Z_1X})^{-1}\mathbb{E}_{XX} \end{pmatrix} \\
&= \begin{pmatrix} \mathbf{0}_{d_z \times d_x} \\ (\mathbb{E}_{XX} - \mathbb{E}_{XZ_1}\mathbb{E}_{Z_1Z_1}^{-1}\mathbb{E}_{Z_1X})^{-1}(\mathbb{E}_{XX} - \mathbb{E}_{XZ_1}\mathbb{E}_{Z_1Z_1}^{-1}\mathbb{E}_{Z_1X}) \end{pmatrix} = \begin{pmatrix} \mathbf{0}_{d_z \times d_x} \\ I_{d_x} \end{pmatrix}.
\end{aligned}
\tag{2.7.1}
$$

Hence,

$$
\mathcal{P}_Z(X) = \left(Z^\top \mathbb{E}[ZZ^\top]^{-1}\mathbb{E}[ZX^\top]\right)^\top = \left(\begin{bmatrix} Z_1^\top, & X^\top \end{bmatrix} \begin{pmatrix} \mathbf{0}_{d_z \times d_x} \\ I_{d_x} \end{pmatrix}\right)^\top = X.
$$

That is, the orthogonal projection of $X$ onto $Z = (Z_1, X)$ is $X$ itself.

# Chapter 3

# Maximum Likelihood Estimation

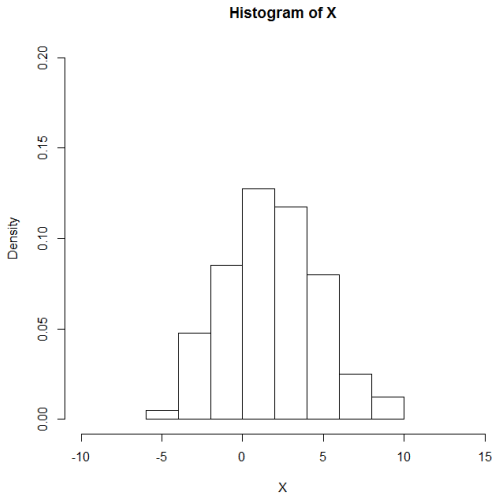
## 3.1 Maximum likelihood principle

The maximum likelihood method is a basic statistical technique for estimating parameters by finding the "most likely" data-generating process to explain the observed data. To what extent it is "likely" is measured by the joint probability/density.

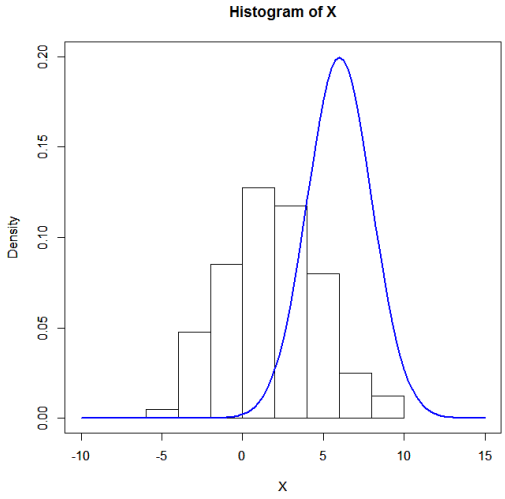
Least Squares = estimate parameters by minimizing the sum of squared errors

Maximum Likelihood = estimate parameters by maximizing the probability of observing the data

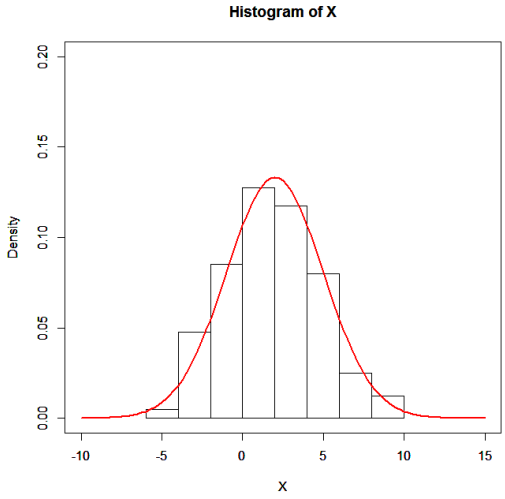
For example, suppose that a random variable $X$ is distributed as normal; but we do not know its mean $\mathbb{E}X = 2$ and variance $\mathrm{Var}(X) = 9$. We have a set of observations $\{X_1, \ldots, X_n\}$ of sample size $n$ independently drawn from the same distribution as $X$. The relative frequency histogram of this data is given below ($n = 200$):




As a candidate distribution, for example, consider $N(6, 2)$, which is shown by a blue curve in the next figure:




If $N(6,2)$ is the true distribution of $X$, the values in the neighborhood of 6 would be most likely observable. In other words, it is less likely that our data are generated from $N(6,2)$. The probability distribution that generates our data with highest likelihood is the one that best fits the histogram:




The red curve in the above figure is the normal distribution with its mean and variance being equal to the sample average (1.85) and sample variance (8.55) of $X$, respectively. Indeed, as stated below, the sample average and sample variance are the **maximum likelihood estimators** (MLE) of $\mathbb{E}X$ and $\mathrm{Var}(X)$, respectively.

**Example 3.1.1 (Normal random variable)** Suppose that we have a set of observations $\{X_1, \ldots, X_n\}$ of size $n$, which are known to be IID normal $N(\mu_0, \sigma_0^2)$, but $\mu_0$ and $\sigma_0^2$ are unknown. The joint density of the data can be written as

$$\text{Joint density of } \{X_1, \ldots, X_n\} = \prod_{i=1}^{n} f_X(X_i)$$

by the IID assumption, where $f_X(x) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{(x-\mu_0)^2}{2\sigma_0^2}\right)$. Since $\mu_0$ and $\sigma_0^2$ are unknown, we consider their candidate values $\mu$ and $\sigma^2$, respectively, and let

$$L_n(\theta) \equiv \prod_{i=1}^{n} f_X(X_i; \mu, \sigma^2),$$

where $\theta = (\mu, \sigma^2)$, and $f_X(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$. The function $L_n(\theta)$ is the joint density of the data under $\theta$, which is called the **likelihood function**. The likelihood function is a function of the parameter $\theta$ such that it coincides with the true joint density when evaluated at $\theta = \theta_0$. The MLE for $\theta_0$ is defined as the maximizer of the "log" of the likelihood, the so-called **log-likelihood function**; that is,

$$\begin{aligned}\widehat{\theta}_n^{mle} &\equiv \operatorname*{argmax}_{\theta} \log L_n(\theta) \\ &= \operatorname*{argmax}_{(\mu,\sigma^2)} \sum_{i=1}^n \log f_X(X_i; \mu, \sigma^2) \\ &= \operatorname*{argmax}_{(\mu,\sigma^2)} \sum_{i=1}^n \left[ -\frac{1}{2}\log\sigma^2 - \frac{(X_i-\mu)^2}{2\sigma^2} \right] = \operatorname*{argmin}_{(\mu,\sigma^2)} \left[ \frac{n}{2}\log\sigma^2 + \frac{\sum_{i=1}^n (X_i-\mu)^2}{2\sigma^2} \right].\end{aligned} \tag{3.1.1}$$

A theoretical reason for maximizing the log-likelihood function, rather than maximizing the likelihood function itself, will be described later.[1]

From the last line of (3.1.1), we can see that the MLE $\widehat{\mu}_n^{mle}$ of $\mu_0$ can be obtained independently by solving the problem $\min_\mu \sum_{i=1}^n (X_i - \mu)^2$, whose solution is clearly $\widehat{\mu}_n^{mle} = \overline{X}_n$, where $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Hence, the sample average of $X$ is the MLE of $\mu_0$. For the MLE $\widehat{\sigma}_n^{2,mle}$ of $\sigma_0^2$, by the first-order condition it satisfies that

$$0 = \frac{n}{\widehat{\sigma}_n^{2,mle}} - \frac{\sum_{i=1}^n (X_i - \overline{X}_n)^2}{\left[\widehat{\sigma}_n^{2,mle}\right]^2},$$

which further implies that

$$\frac{\sum_{i=1}^n (X_i - \overline{X}_n)^2}{\widehat{\sigma}_n^{2,mle}} = n \quad \Rightarrow \quad \widehat{\sigma}_n^{2,mle} = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2.$$

Thus, the MLE of $\sigma_0^2$ is the sample variance of $X$. It should be noted that the MLE variance estimator is not an unbiased variance estimator.[2]

**Example 3.1.2 (Linear regression)** We consider the following linear regression model with normally distributed error term:

$$\begin{aligned} Y_i &= X_i^\top \beta_0 + \varepsilon_i, \quad i = 1, \ldots, n \\ \varepsilon_i &\sim N(0, \sigma_0^2) \end{aligned}$$

Note that the second line of the above is equivalent to $Y_i - X_i^\top \beta_0 \sim N(0, \sigma_0^2)$. Thus, the conditional density of $Y_i$ given $X_i$ can be characterized by

$$Y_i | X_i \sim N(X_i^\top \beta_0, \sigma_0^2).$$

Denote the marginal density of $X_i$ as $f_X(\cdot)$, and let $\phi(\cdot; \mu, \sigma^2)$ be the normal density function with mean $\mu$ and variance $\sigma^2$. Then, the joint density of $(Y_i, X_i)$ can be written as

$$f_{Y,X}(Y_i, X_i) = \phi(Y_i; X_i^\top \beta_0, \sigma_0^2) \cdot f_X(X_i).$$

If the data are IID, the joint density of our data is

$$\text{Joint density of } \{(Y_i, X_i) : 1 \le i \le n\} = \prod_{i=1}^n \phi(Y_i; X_i^\top \beta_0, \sigma_0^2) \cdot f_X(X_i).$$

[1] An intuitive reason for this is that, without taking the log, the product of the likelihood (probability) tends to zero as $n$ increases.

[2] As is well known, an unbiased variance estimator is obtained by $\widehat{\mathrm{Var}}_n(X) = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$.

Thus, letting $L_n(\theta) \equiv \prod_{i=1}^n \phi(Y_i; X_i^\top \beta, \sigma^2) \cdot f_X(X_i)$ be the likelihood function with $\theta = (\beta, \sigma^2)$, the log-likelihood function is given by

$$\log L_n(\theta) = \sum_{i=1}^n \log \phi(Y_i; X_i^\top \beta, \sigma^2) + \sum_{i=1}^n \log f_X(X_i).$$

Since the second term on the right-hand side is irrelevant to the unknown parameters, the MLE for $\theta_0 = (\beta_0, \sigma_0^2)$ can be obtained by maximizing only the first term:

$$\begin{aligned} \widehat{\theta}_n^{mle} &\equiv \operatorname*{argmax}_\theta \log L_n(\theta) \\ &= \operatorname*{argmax}_{(\beta,\sigma^2)} \left[ \sum_{i=1}^n \log \phi(Y_i; X_i^\top \beta, \sigma^2) \right] = \operatorname*{argmin}_{(\beta,\sigma^2)} \left[ \frac{n}{2} \log \sigma^2 + \frac{\sum_{i=1}^n (Y_i - X_i^\top \beta)^2}{2\sigma^2} \right], \end{aligned}$$

where the last equality follows from the same argument as in (3.1.1). Further, by the similar arguments in Example 3.1.1, it is easy to see that the MLE $\widehat{\beta}_n^{mle}$ coincides with the OLS estimator and that the MLE $\widehat{\sigma}_n^{2,mle}$ is obtained as the sample variance of the OLS residuals.

**Example 3.1.3 (Flipping an uneven coin)** Suppose that we have a possibly uneven coin such that the probability of heads is unknown and may not be equal to 0.5. Let $X$ be a dummy variable defined by

$$X = 1 \text{ for "head"},$$
$$X = 0 \text{ for "tail"}.$$

We use $p_0$ to denote the coin's true probability of heads: $\mathbb{E}X = p_0$. Suppose that we have a set of coin-flipping data $\{X_1, \ldots, X_n\}$ of $n$ independent trials. A natural estimator of $p_0$ is the sample average, i.e., the sample proportion of heads: $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. In fact, the sample average $\overline{X}_n$ is the least-squares estimator for $p_0$. To see this, let

$$\widehat{p}_n^{ls} \equiv \operatorname*{argmin}_p \frac{1}{n} \sum_{i=1}^n (X_i - p)^2.$$

The first-order condition of the minimization problem gives

$$\begin{aligned} -\frac{2}{n} \sum_{i=1}^n (X_i - \widehat{p}_n^{ls}) = 0 \quad &\Rightarrow \quad \frac{1}{n} \sum_{i=1}^n X_i - \widehat{p}_n^{ls} = 0 \\ &\Rightarrow \quad \widehat{p}_n^{ls} = \overline{X}_n. \end{aligned}$$

As shown below, $\overline{X}_n$ is also an MLE for $p_0$. The probability distribution of $X$ is

$$\Pr(X = 1) = p_0, \quad \Pr(X = 0) = 1 - p_0$$

This type of probability distribution is called the **Bernoulli distribution**. By the independence, the joint probability of $\{X_1, \ldots, X_n\}$ is equal to

$$\text{Joint probability of } \{X_1, \ldots, X_n\} = \prod_{i=1}^n p_0^{X_i} (1 - p_0)^{1 - X_i}.$$

Then, the likelihood function is given by $L_n(p) \equiv \prod_{i=1}^n p^{X_i} (1-p)^{1-X_i}$, and the MLE is defined as

$$\widehat{p}_n^{mle} \equiv \operatorname*{argmax}_{p \in (0,1)} \log L_n(p)$$

$$= \underset{p\in(0,1)}{\operatorname{argmax}} \sum_{i=1}^{n} \{X_i \log p + (1 - X_i) \log(1 - p)\}.$$

The first-order condition for the above maximization implies that

$$\begin{aligned} 0 &= \sum_{i=1}^{n} \left\{ \frac{X_i}{\widehat{p}_n^{mle}} - \frac{1 - X_i}{1 - \widehat{p}_n^{mle}} \right\} \\ &= \sum_{i=1}^{n} \left\{ \frac{X_i - \widehat{p}_n^{mle}}{\widehat{p}_n^{mle} \cdot (1 - \widehat{p}_n^{mle})} \right\}. \end{aligned}$$

Therefore, $\widehat{p}_n^{mle}$ must satisfy

$$\sum_{i=1}^{n} (X_i - \widehat{p}_n^{mle}) = 0 \Rightarrow \widehat{p}_n^{mle} = \overline{X}_n.$$

We generalize the above discussion. Suppose we have a set of IID observations $\{X_1, \ldots, X_n\}$ of sample size $n$. Here, $X$ may be a scalar or a vector. For continuous (resp. discrete) $X$, let $p(x; \theta_0)$ be the population density (resp. probability) function of $X$, where $\theta_0$ is the vector of unknown parameters to be estimated. The functional form of $p(x; \theta_0)$ is known up to $\theta_0$; that is, $\theta_0$ is the only unknown component.[3] Then, the **likelihood function** is defined as

$$L_n(\theta) \equiv \prod_{i=1}^{n} p(X_i; \theta),$$

and the **log-likelihood function** is

$$\ell_n(\theta) \equiv \log L_n(\theta) = \sum_{i=1}^{n} \log p(X_i; \theta).$$

The **maximum likelihood estimator** (MLE) of $\theta_0$ is defined as the maximizer of the log-likelihood function $\ell_n(\theta)$:

$$\widehat{\theta}_n^{mle} \equiv \underset{\theta}{\operatorname{argmax}}\, \ell_n(\theta),$$

or numerically equivalently,

$$\begin{aligned} \widehat{\theta}_n^{mle} &\equiv \underset{\theta}{\operatorname{argmax}}\, \frac{1}{n} \ell_n(\theta) \\ &= \underset{\theta}{\operatorname{argmax}}\, \frac{1}{n} \sum_{i=1}^{n} \log p(X_i; \theta). \end{aligned}$$

Dividing the objective function by $n$ is for technical convenience in deriving the asymptotic properties of MLE.[4]

**Exercise 3.1.4** Suppose that we have $n$ IID observations $\{X_1, \ldots, X_n\}$ drawn from a Poisson distribution with parameter $\lambda_0$; i.e.,

$$\Pr(X = k) = \frac{\lambda_0^k \exp(-\lambda_0)}{k!}.$$

Find the MLE of $\lambda_0$.

[3] If the model specification of $p$ is incorrect, then the resulting MLE is not consistent for $\theta_0$. Note that, for example in the case of linear regression models, we may use a least-squares method to consistently estimate the parameters even when $p$ is partly unknown. In this sense, in general, maximum likelihood methods require more restrictive assumptions than the other types of estimators.

[4] As $n$ increases to infinity, $\ell_n(\theta)$ may also diverge to infinity since it is a summation of $n$ non-zero terms. However, by dividing it by $n$, under standard conditions, we can have $\frac{1}{n}\ell_n(\theta) \xrightarrow{p} \mathbb{E}[\log p(X_i; \theta)]$ for all $\theta$ by the "uniform" law of large numbers.

**Exercise 3.1.5** Suppose that we have $n$ IID observations $\{X_1, \ldots, X_n\}$ drawn from Uniform$[0, \alpha_0]$. Show that the MLE of $\alpha_0$ is given by

$$\widehat{\alpha}_n^{mle} = \max\{X_1, \ldots, X_n\}.$$

### 3.1.1 Kullback-Leibler divergence

Let $p(x;\theta_1)$ and $p(x;\theta_2)$ be two density (or probability) functions such that $p(x;\theta_1) = p(x;\theta_2)$ for all $x$ if and only if $\theta_1 = \theta_2$. The following quantity is called the **Kullback-Leibler divergence** (KL divergence) between $p(x;\theta_1)$ and $p(x;\theta_2)$:[5]

$$\begin{aligned} K\left(p(X;\theta_1)\|p(X;\theta_2)\right) &\equiv \mathbb{E}_{\theta_1}\left[\log \frac{p(X;\theta_1)}{p(X;\theta_2)}\right] \\ &= \begin{cases} \int \left[\log \frac{p(x;\theta_1)}{p(x;\theta_2)}\right] p(x;\theta_1)\mathrm{d}x & \text{if } X \text{ is continuous} \\ \sum_{x\in\{x_1,\ldots,x_k\}} \left[\log \frac{p(x;\theta_1)}{p(x;\theta_2)}\right] p(x;\theta_1) & \text{if } X \text{ is discrete with its support } \{x_1,\ldots,x_k\} \end{cases} \end{aligned}$$

where $\mathbb{E}_\theta[\,\cdot\,]$ denotes the expectation under $p(x;\theta)$ (note that $\mathbb{E}[\,\cdot\,] \equiv \mathbb{E}_{\theta_0}[\,\cdot\,]$, where $\theta_0$ is the true value of $\theta$). It is straightforward to see that

$$K\left(p(X;\theta_1)\|p(X;\theta_2)\right) = 0$$

if $\theta_1 = \theta_2$ since $\log 1 = 0$. Moreover, it holds that for any $\theta_1 \neq \theta_2$,

$$K\left(p(X;\theta_1)\|p(X;\theta_2)\right) > 0.$$

Formally, we have the following result.

**Theorem 3.1.6** *For any density (or probability) functions $p(x;\theta_1)$ and $p(x;\theta_2)$ such that $p(x;\theta_1) = p(x;\theta_2)$ for all $x$ if and only if $\theta_1 = \theta_2$, it holds that $K(p(x;\theta_1)\|p(x;\theta_2)) \geq 0$. The equality holds if and only if $\theta_1 = \theta_2$.*

**Proof.** We only prove the case when $p(x;\theta)$ is a density function. First, note that

$$\begin{aligned} K(p(x;\theta_1)\|p(x;\theta_2)) &= \mathbb{E}_{\theta_1}\left[\log \frac{p(X;\theta_1)}{p(X;\theta_2)}\right] \\ &= \mathbb{E}_{\theta_1}\left[-\log \frac{p(X;\theta_2)}{p(X;\theta_1)}\right]. \end{aligned}$$

Then, since $-\log(\cdot)$ is a convex function, by Jensen's inequality B.1.3,

$$\begin{aligned} K(p(x;\theta_1)\|p(x;\theta_2)) &\geq -\log \mathbb{E}_{\theta_1}\left[\frac{p(X;\theta_2)}{p(X;\theta_1)}\right] \\ &= -\log \int \frac{p(x;\theta_2)}{p(x;\theta_1)} p(x;\theta_1)\mathrm{d}x \\ &= -\log \int p(x;\theta_2)\mathrm{d}x \\ &= -\log 1 = 0. \end{aligned}$$

[5] Note that the KL divergence is not symmetric: $K\left(p(X;\theta_1)\|p(X;\theta_2)\right) \neq K\left(p(X;\theta_2)\|p(X;\theta_1)\right)$, in general. In addition, it does not satisfy the triangle inequality. Thus, KL divergence is not a distance function.

This proves the first argument. For the second argument, The "if"-part is trivial. To prove the "only if"-part, note that, if a function $\varphi(\cdot)$ is strictly convex, $\mathbb{E}[\varphi(X)] = \varphi(\mathbb{E}[X])$ occurs only when $X = \mathbb{E}[X]$ (i.e., when $X$ is a constant). Since $-\log(\cdot)$ is a strictly convex function, $K(p(x;\theta_1)\|p(x;\theta_2)) = 0$ holds only when

$$\begin{aligned}\frac{p(X;\theta_2)}{p(X;\theta_1)} &= \mathbb{E}_{\theta_1}\left[\frac{p(X;\theta_2)}{p(X;\theta_1)}\right] \\ &= \int \frac{p(x;\theta_2)}{p(x;\theta_1)} p(x;\theta_1)\mathrm{d}x \\ &= \int p(x;\theta_2)\mathrm{d}x = 1.\end{aligned}$$

This implies the desired result. ■

### 3.1.2 MLE minimizes the KL divergence

From Theorem 3.1.6, for the true parameter $\theta_0$, we can obtain the following trivial equality:

$$\theta_0 = \operatorname*{argmin}_{\theta} K(p(X;\theta_0)\|p(X;\theta)).$$

Note that

$$\begin{aligned}K(p(X;\theta_0)\|p(X;\theta)) &= \mathbb{E}\left[\log\frac{p(X;\theta_0)}{p(X;\theta)}\right] \\ &= \mathbb{E}[\log p(X;\theta_0)] - \mathbb{E}[\log p(X;\theta)].\end{aligned}$$

Since the first term on the right-hand side is irrelevant to $\theta$, it further holds that

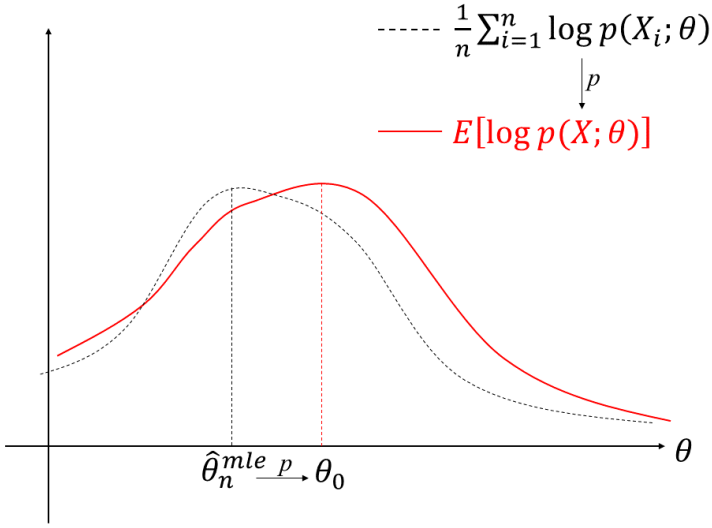
$$\theta_0 = \operatorname*{argmax}_{\theta} \mathbb{E}[\log p(X;\theta)], \tag{3.1.2}$$

implying that the true parameter $\theta_0$ is the maximizer of the population log-likelihood function. Thus, maximizing the population log-likelihood is equivalent to minimize the KL divergence.

Here, by the law of large numbers, we can expect that

$$\frac{1}{n}\sum_{i=1}^{n} \log p(X_i;\theta) \xrightarrow{p} \mathbb{E}[\log p(X;\theta)],$$

for all $\theta$ (see also Footnote 4). Since $\widehat{\theta}_n^{mle} = \operatorname{argmax}_\theta \frac{1}{n}\sum_{i=1}^{n} \log p(X_i;\theta)$, in view of (3.1.2), it holds that $\widehat{\theta}_n^{mle} \xrightarrow{p} \theta_0$. For more precise discussion, see, for example, [Newey and McFadden, 1994].



## 3.2 Cramer-Rao lower bound

Now, we introduce two important concepts. The **score function** (or simply score) is the partial derivative of the log-likelihood $\log p(x;\theta)$ with respect to $\theta$, which we denote by $s(x;\theta)$. The variance covariance matrix of the score under $p(x;\theta)$ is called the **Fisher information matrix**, which we denote by $I(\theta)$. Observe that the expectation of the score function (under $p(x;\theta)$) is always zero:

$$\begin{aligned}\mathbb{E}_\theta[s(X;\theta)] &= \int \left\{\frac{\partial}{\partial\theta}\log p(x;\theta)\right\} p(x;\theta)\mathrm{d}x \\ &= \int \frac{\frac{\partial}{\partial\theta}p(x;\theta)}{p(x;\theta)} p(x;\theta)\mathrm{d}x \\ &= \int \frac{\partial}{\partial\theta}p(x;\theta)\mathrm{d}x \\ &= \frac{\partial}{\partial\theta}\underbrace{\int p(x;\theta)\mathrm{d}x}_{=1} = \mathbf{0},\end{aligned} \tag{3.2.1}$$

and, therefore

$$\begin{aligned}I(\theta) &\equiv \mathrm{Cov}_\theta\left[s(X;\theta)\right] \\ &= \mathbb{E}_\theta\left[(s(X;\theta) - \mathbb{E}_\theta s(X;\theta))(s(X;\theta) - \mathbb{E}_\theta s(X;\theta))^\top\right] \\ &= \mathbb{E}_\theta\left[s(X;\theta)s(X;\theta)^\top\right].\end{aligned}$$

When estimating $\theta_0$, we can consider many alternative estimators $\widehat{\theta}_n$'s, including OLS, MLE, etc. An important criterion for the choice of estimator is the **efficiency**; the more efficient estimator is the one that has the smaller variance. Surprisingly or not, under certain regularity conditions – see the appendix of this chapter, the inverse of the Fisher information matrix $[I(\theta_0)]^{-1}$ gives the "lower bound", known as **Cramer-Rao lower bound**, of the covariance matrix for all possible estimators of $\theta_0$; that is, for any "regular" estimator $\widehat{\theta}_n$ of $\theta_0$,

$$\mathrm{Cov}\left(\sqrt{n}(\widehat{\theta}_n - \theta_0)\right) \geq [I(\theta_0)]^{-1}. \tag{3.2.2}$$

Here, for any square matrices $A$ and $B$ of the same size, $A \geq B$ means that $A - B$ is positive-semidefinite. As described below, the MLE $\widehat{\theta}_n^{mle}$ generally attains this lower bound; that is, (3.2.2) holds with equality. In other words, whenever the MLE is available, it is the most efficient estimator we can use in the class of regular estimators. An example of "irregular" estimator will be provided in the appendix of this chapter.

We can relatively easily check (3.2.2) if we restrict our attention to the class of unbiased estimators. Suppose that the data $\{X_1,\ldots,X_n\}$ are IID with density function $p(x;\theta_0)$, where $\theta_0$ is a scalar true parameter for simplicity, and that we have an **unbiased estimator** $\widehat{\theta}_n$ such that

$$\mathbb{E}_\theta[\widehat{\theta}_n] - \theta = 0 \tag{3.2.3}$$

for any $\theta$. Note that the above expectation is taken with respect to the data $\{X_1,\ldots,X_n\}$ since any estimator is a function of the data sample; thus, we may write $\widehat{\theta}_n = \theta_n(X_1,\ldots,X_n)$.

**Proof of (3.2.2) for unbiased estimators.** Since (3.2.3) holds for all $\theta$, differentiating both sides of (3.2.3) with respect to $\theta$, we obtain

$$\begin{aligned}
0 = \frac{\partial}{\partial\theta}\mathbb{E}_\theta[\widehat{\theta}_n - \theta] &= \frac{\partial}{\partial\theta}\int[\theta_n(x_1,\ldots,x_n)-\theta]\prod_{i=1}^n p(x_i;\theta)\mathrm{d}x_1\cdots\mathrm{d}x_n \\
&= \int[\theta_n(x_1,\ldots,x_n)-\theta]\frac{\partial}{\partial\theta}\prod_{i=1}^n p(x_i;\theta)\mathrm{d}x_1\cdots\mathrm{d}x_n - 1 \\
&= \int[\theta_n(x_1,\ldots,x_n)-\theta]\left\{\frac{\partial}{\partial\theta}\log\left(\prod_{i=1}^n p(x_i;\theta)\right)\right\}\prod_{i=1}^n p(x_i;\theta)\mathrm{d}x_1\cdots\mathrm{d}x_n - 1 \\
&= \int[\theta_n(x_1,\ldots,x_n)-\theta]\left\{\frac{\partial}{\partial\theta}\sum_{i=1}^n \log p(x_i;\theta)\right\}\prod_{i=1}^n p(x_i;\theta)\mathrm{d}x_1\cdots\mathrm{d}x_n - 1 \\
&= \int[\theta_n(x_1,\ldots,x_n)-\theta]\sum_{i=1}^n s(x_i;\theta)\prod_{i=1}^n p(x_i;\theta)\mathrm{d}x_1\cdots\mathrm{d}x_n - 1 \\
&= \mathbb{E}_\theta\left[(\widehat{\theta}_n-\theta)\sum_{i=1}^n s(X_i;\theta)\right] - 1.
\end{aligned}$$

Recall that $\mathbb{E}_\theta[\widehat{\theta}_n - \theta] = 0$ and $\mathbb{E}_\theta[s(X_i;\theta)] = 0$ by (3.2.1). Then, by Cauchy-Schwarz inequality B.1.4 and the IID assumption, we have

$$\begin{aligned}
1 &\le \left|\mathbb{E}_\theta\left[\sqrt{n}(\widehat{\theta}_n-\theta)\cdot\frac{1}{\sqrt{n}}\sum_{i=1}^n s(X_i;\theta)\right]\right| \\
&\le \mathrm{Var}_\theta\left(\sqrt{n}(\widehat{\theta}_n-\theta)\right)\mathrm{Var}_\theta\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n s(X_i;\theta)\right) \\
&= \mathrm{Var}_\theta\left(\sqrt{n}(\widehat{\theta}_n-\theta)\right) I(\theta),
\end{aligned}$$

and, thus,

$$\mathrm{Var}_\theta\left(\sqrt{n}(\widehat{\theta}_n-\theta)\right) \ge \frac{1}{I(\theta)}.$$

■

## 3.3 Asymptotic properties

Suppose that we have a consistent MLE $\widehat{\theta}_n^{mle}$ of $\theta_0$. By the first-order condition, the MLE $\widehat{\theta}_n^{mle}$ satisfies that

$$\frac{1}{n}\sum_{i=1}^n s(X_i;\widehat{\theta}_n^{mle}) = \mathbf{0}.$$

Applying the mean-value expansion to $s(X_i;\widehat{\theta}_n^{mle})$ around $\theta_0$, we have

$$\mathbf{0} = \frac{1}{n}\sum_{i=1}^n s(X_i;\theta_0) + \frac{1}{n}\sum_{i=1}^n H(X_i;\overline{\theta}_n)(\widehat{\theta}_n^{mle}-\theta_0),$$

where $H(X_i;\theta) \equiv \partial s(X_i;\theta)/(\partial\theta^\top)$, or equivalently, $H(X_i;\theta) \equiv \partial^2 \log p(X_i;\theta)/(\partial\theta\partial\theta^\top)$, i.e., the Hessian matrix of $\log p(X_i;\theta)$, and $\overline{\theta}_n \in [\widehat{\theta}_n^{mle},\theta_0]$. Then, if $\frac{1}{n}\sum_{i=1}^n H(X_i;\overline{\theta}_n)$ is nonsingular, we obtain

$$\sqrt{n}(\widehat{\theta}_n^{mle}-\theta_0) = -\left[\frac{1}{n}\sum_{i=1}^n H(X_i;\overline{\theta}_n)\right]^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n s(X_i;\theta_0). \tag{3.3.1}$$

As shown in Section 3.2, we have $\mathbb{E}[s(X;\theta_0)] = 0$ and $\mathrm{Cov}[s(X;\theta_0)] = I(\theta_0)$. Since $X_i$'s are IID, so are $s(X_i;\theta_0)$'s. Then, we can apply CLT to $\frac{1}{\sqrt{n}}\sum_{i=1}^n s(X_i;\theta_0)$ to obtain

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n s(X_i;\theta_0) \xrightarrow{d} \mathbf{N}(\mathbf{0}, I(\theta_0)).$$

Further, note that $\widehat{\theta}_n^{mle}$ is a consistent estimator of $\theta_0$ and so is $\overline{\theta}_n$ clearly by its definition. Then, we have $\frac{1}{n}\sum_{i=1}^n H(X_i;\overline{\theta}_n) \xrightarrow{p} \mathbb{E}[H(X;\theta_0)]$ by WLLN, assuming that $H(x;\theta)$ is continuous at $\theta_0$. Finally, combining these results and applying Slutsky's theorem B.1.2 give the following result:

$$\sqrt{n}(\widehat{\theta}_n^{mle} - \theta_0) \xrightarrow{d} \mathbf{N}\left(\mathbf{0}, \mathbb{E}[H(X;\theta_0)]^{-1} I(\theta_0) \mathbb{E}[H(X;\theta_0)]^{-1}\right).$$

From this result, one might think that the MLE is not necessarily the most efficient estimation method. However, that is not the case; that is, the MLE is always the most efficient because of the following fact:

$$I(\theta_0) = -\mathbb{E}[H(X;\theta_0)]. \tag{3.3.2}$$

This equality is known as the **information matrix equality**.

**Proof of (3.3.2).** Since (3.2.1) holds for any $\theta$, differentiating both sides of (3.2.1) with respect to $\theta$, we have

$$\begin{aligned}
\mathbf{0} = \frac{\partial}{\partial \theta^\top}\mathbb{E}_\theta[s(X;\theta)] &= \frac{\partial}{\partial \theta^\top}\int s(x;\theta)p(x;\theta)\mathrm{d}x \\
&= \int H(x;\theta)p(x;\theta)\mathrm{d}x + \int s(x;\theta)\frac{\frac{\partial}{\partial\theta^\top}p(x;\theta)}{p(x;\theta)}p(x;\theta)\mathrm{d}x \\
&= \int H(x;\theta)p(x;\theta)\mathrm{d}x + \int s(x;\theta)s(x;\theta)^\top p(x;\theta)\mathrm{d}x.
\end{aligned}$$

The last line implies that $-\mathbb{E}_\theta[H(X;\theta)] = \mathbb{E}_\theta[s(X;\theta)s(X;\theta)^\top]$. Finally, the result follows from the definition of $I(\theta)$. ■

Therefore, thanks to the information matrix equality, the asymptotic distribution of the MLE can be simply written in two ways as

$$\begin{aligned}
\sqrt{n}(\widehat{\theta}_n^{mle} - \theta_0) &\xrightarrow{d} \mathbf{N}\left(\mathbf{0}, -\mathbb{E}[H(X;\theta_0)]^{-1}\right) \\
\sqrt{n}(\widehat{\theta}_n^{mle} - \theta_0) &\xrightarrow{d} \mathbf{N}\left(\mathbf{0}, [I(\theta_0)]^{-1}\right).
\end{aligned}$$

## 3.4 An application: Binary response models

As an important application of the maximum likelihood method, we consider estimating binary response models. Let $D$ be a binary outcome variable defined by

$$D = \mathbf{1}\{X^\top\beta_0 \geq \varepsilon\},$$

where $X$ is a vector of explanatory variables, $\varepsilon$ is an unobserved error term, and $\beta_0$ is the vector of parameters to be estimated.

First of all, note that for any constant $c > 0$, it holds that $D = \mathbf{1}\{X^\top(c\beta_0) \geq \eta\}$, where $\eta$ is a new error term defined as $\eta = c\varepsilon$. Thus, when the distribution of $\varepsilon$ is fully unrestricted, $\beta_0$ cannot be "identified" without some

scale normalization (more formal discussion will be provided in Chapter 5). To overcome this issue, we typically impose a scale restriction on $\varepsilon$ by assuming that $\varepsilon$ is distributed as either the standard normal or standard logistic with location 0 and scale 1. The resulting model based on the former assumption is called the **probit model** and the one based on the latter is called the **logit model**.

Then, under appropriate scale normalization, we can estimate $\beta_0$ by the maximum likelihood method in the following manner. Observe that the conditional probability of $D = 1$ given $X$ is equal to

$$\Pr(D = 1 \mid X) = \Pr(\varepsilon \leq X^\top \beta_0 \mid X) = F(X^\top \beta_0),$$

where $F(\cdot)$ is the distribution function of $\varepsilon$, which is a known function. Suppose that we have the data of $n$ IID observations $\{(D_i, X_i) : 1 \leq i \leq n\}$. Then, similarly as in Example 3.1.3, the (conditional) likelihood function for the data is given by

$$L_n(\beta) \equiv \prod_{i=1}^{n} F(X_i^\top \beta)^{D_i} (1 - F(X_i^\top \beta))^{1-D_i},$$

and thus the MLE of $\beta_0$ is defined as

$$\begin{aligned} \widehat{\beta}_n^{mle} &\equiv \operatorname*{argmax}_\beta \log L_n(\beta) \\ &= \operatorname*{argmax}_\beta \sum_{i=1}^{n} \left\{ D_i \cdot \log F(X_i^\top \beta) + (1 - D_i) \cdot \log(1 - F(X_i^\top \beta)) \right\}. \end{aligned}$$

Next, we derive the Fisher information matrix. To this end, we need to calculate the score function, which is the derivative of $D \log F(X^\top \beta) + (1 - D) \log(1 - F(X^\top \beta)$ with respect to $\beta$:

$$\begin{aligned} s(X; \beta) &= X \frac{D \cdot f(X^\top \beta)}{F(X^\top \beta)} - X \frac{(1 - D) \cdot f(X^\top \beta)}{1 - F(X^\top \beta)} \\ &= X \frac{(D - F(X^\top \beta)) f(X^\top \beta)}{F(X^\top \beta) \cdot (1 - F(X^\top \beta))}, \end{aligned}$$

where $f(\cdot)$ is the density function of $\varepsilon$. Thus, noting that

$$\mathbb{E}[(D - F(X^\top \beta_0))^2 | X] = F(X^\top \beta_0)(1 - F(X^\top \beta_0)),$$

the information matrix $I(\beta_0)$ is obtained by

$$\begin{aligned} I(\beta_0) &= \mathbb{E}\left[ s(X; \beta_0) s(X; \beta_0)^\top \right] \\ &= \mathbb{E}\left[ XX^\top \frac{(D - F(X^\top \beta_0))^2 [f(X^\top \beta_0)]^2}{[F(X^\top \beta_0) \cdot (1 - F(X^\top \beta_0))]^2} \right] \\ &= \mathbb{E}\left[ XX^\top \frac{[f(X^\top \beta_0)]^2}{F(X^\top \beta_0) \cdot (1 - F(X^\top \beta_0))} \right] \end{aligned}$$

by LIE. Hence, we have

$$\sqrt{n}(\widehat{\beta}_n^{mle} - \beta_0) \xrightarrow{d} \mathbf{N}\left( \mathbf{0}, \mathbb{E}\left[ XX^\top \frac{[f(X^\top \beta_0)]^2}{F(X^\top \beta_0) \cdot (1 - F(X^\top \beta_0))} \right]^{-1} \right).$$

It is also straightforward to confirm the information matrix equality. By the chain rule, the Hessian $H(X; \beta_0)$ can be written as

$$H(X; \beta_0) = \partial s(X; \beta_0) / \partial \beta^\top$$

$$= -XX^\top \frac{[f(X^\top\beta_0)]^2}{F(X^\top\beta_0)\cdot(1-F(X^\top\beta_0))} + X(D - F(X^\top\beta_0))\cdot\frac{\partial}{\partial\beta^\top}\left[\frac{f(X^\top\beta_0)}{F(X^\top\beta_0)\cdot(1-F(X^\top\beta_0))}\right].$$

By LIE, we can see that the second term on the right-hand side will be zero after taking the expectation. Thus, we have $-\mathbb{E}[H(X;\beta_0)] = I(\beta_0)$, as desired.

**Exercise 3.4.1** Consider an ordered choice model with three ordered outcomes:

$$\begin{aligned} Y^* = X^\top\beta_0 + \varepsilon, \qquad D = 1 &\iff Y^* \le c_1 \\ D = 2 &\iff c_1 < Y^* \le c_2 \\ D = 3 &\iff c_2 < Y^*, \end{aligned}$$

where $Y^*$ is an unobservable latent variable, and $\theta_0 = (\beta_0, c_1, c_2)$ is the set of unknown parameters to be estimated. Suppose that we have an IID sample $\{(D_i, X_i) : 1 \le i \le n\}$. The error term $\varepsilon$ is assumed to be independent of $X$ and has a known distribution function $F$. Derive the log-likelihood function for this model and define the MLE $\widehat{\theta}_n^{mle}$ of $\theta_0$.

## 3.5 Likelihood ratio test

Partition the estimation parameters as $\theta_0 = (\theta_{01}, \theta_{02})$, where $\theta_{01}$ is a $q_1 \times 1$ vector, and $\theta_{02}$ is a $q_2 \times 1$ vector. Suppose that we would like to test the following null hypothesis:

$$\mathbb{H}_0 : \theta_{02} = \theta_{02}^*.$$

The alternative hypothesis is that at least one element of $\theta_{02}$ is not equal to the corresponding element of $\theta_{02}^*$, which is written as

$$\mathbb{H}_1 : \theta_{02} \ne \theta_{02}^*.$$

By the maximum likelihood principle, it holds that

$$\ell_n(\widehat{\theta}_{n1}^{mle}, \widehat{\theta}_{n2}^{mle}) \ge \ell_n(\widetilde{\theta}_{n1}^{mle}, \theta_{02}^*),$$

where $(\widehat{\theta}_{n1}^{mle}, \widehat{\theta}_{n2}^{mle})$ is the MLE of $(\theta_{01}, \theta_{02})$ under $\mathbb{H}_1$, and $\widetilde{\theta}_{n1}^{mle}$ is the "restricted" MLE of $\theta_{01}$ under $\mathbb{H}_0$; namely, $\widetilde{\theta}_{n1}^{mle} = \operatorname{argmax}_{\theta_1} \ell_n(\theta_1, \theta_{02}^*)$. If $\mathbb{H}_1$ is true, then we can expect that the former maximized log-likelihood will be significantly larger than the latter. Thus, we can use the the difference of these log-likelihoods as a valid test statistic for $\mathbb{H}_0$. Let

$$\begin{aligned} T_n &\equiv 2\left(\ell_n(\widehat{\theta}_{n1}, \widehat{\theta}_{n2}) - \ell_n(\widetilde{\theta}_{n1}, \theta_{02}^*)\right) \\ &= 2\log\frac{L_n(\widehat{\theta}_{n1}, \widehat{\theta}_{n2})}{L_n(\widetilde{\theta}_{n1}, \theta_{02}^*)}, \end{aligned}$$

where the superscript "*mle*" is omitted for simplicity. This test statistic is called the **likelihood ratio statistic**. Then, under certain regularity conditions, we can show that

$$T_n \xrightarrow{d} \chi^2(q_2) \tag{3.5.1}$$

under $\mathbb{H}_0$. Thus, if the computed $T_n$ is sufficiently large compared with the critical value of $\chi^2(q_2)$, then we can reject $\mathbb{H}_0$. This testing procedure is called the **likelihood ratio test**.

For a better understanding, it would be useful to provide a sketch of the proof of (3.5.1) for a simpler case, where $\theta_0$ is a scaler and the null hypothesis is $\mathbb{H}_0 : \theta_0 = \theta_0^*$. In this case, the test statistic is simplified as follows: $T_n \equiv 2\left(\ell_n(\widehat{\theta}_n) - \ell_n(\theta_0^*)\right)$. By the second order Taylor expansion, we have

$$\begin{aligned}\ell_n(\theta_0^*) - \ell_n(\widehat{\theta}_n) &= (\theta_0^* - \widehat{\theta}_n)\cdot \ell_n'(\widehat{\theta}_n) + \frac{1}{2}(\theta_0^* - \widehat{\theta}_n)^2 \cdot \ell_n''(\overline{\theta}_n) \\ &= \frac{1}{2}(\theta_0^* - \widehat{\theta}_n)^2 \cdot \ell_n''(\overline{\theta}_n)\end{aligned}$$

where $\ell_n'(\widehat{\theta}_n)$ and $\ell_n''(\widehat{\theta}_n)$ are the first and second derivatives of $\ell_n$ evaluated at $\widehat{\theta}_n$, respectively, and $\overline{\theta}_n \in [\theta_0^*, \widehat{\theta}_n]$. The second equality holds from the first order condition of the maximum likelihood estimation. Hence, we can write

$$\begin{aligned}T_n &= -(\theta_0^* - \widehat{\theta}_n)^2 \cdot \ell_n''(\overline{\theta}_n) \\ &= (\sqrt{n}(\widehat{\theta}_n - \theta_0^*))^2 \cdot \left(-\frac{1}{n}\ell_n''(\overline{\theta}_n)\right).\end{aligned}$$

Observe that under $\mathbb{H}_0 : \theta_0 = \theta_0^*$, we can have $\sqrt{n}(\widehat{\theta}_n - \theta_0^*) = \sqrt{n}(\widehat{\theta}_n - \theta_0)$, and

$$-\frac{1}{n}\ell_n''(\overline{\theta}_n) = -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \log p(X_i, \overline{\theta}_n)}{\partial\theta\partial\theta} \xrightarrow{p} -\mathbb{E}[H(X;\theta_0^*)] = I(\theta_0).$$

Since $\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{d} N(0, [I(\theta_0)]^{-1})$,

$$T_n \approx \Bigg[ \underbrace{\frac{\sqrt{n}(\widehat{\theta}_n - \theta_0)}{\sqrt{I(\theta_0)}}}_{\xrightarrow{d} N(0,1)} \Bigg]^2 \xrightarrow{d} \chi^2(1).$$

**A numerical simulation with R** The following code gives the simulated distributions of $T_n$ for the coin-flipping data in Example 3.1.3 with $n = 500$ and $\mathbb{H}_0 : p = p^*$ for $p^* \in \{0.5, 0.55, 0.6\}$. Here, the true value of $p$ is set to 0.5.

We first define the log-likelihood function:

```
LL <- function(p){

       LogP <- X*log(p) + (1 - X)*log(1 - p)
       ell  <- sum(LogP)
       return(ell)

       }
```

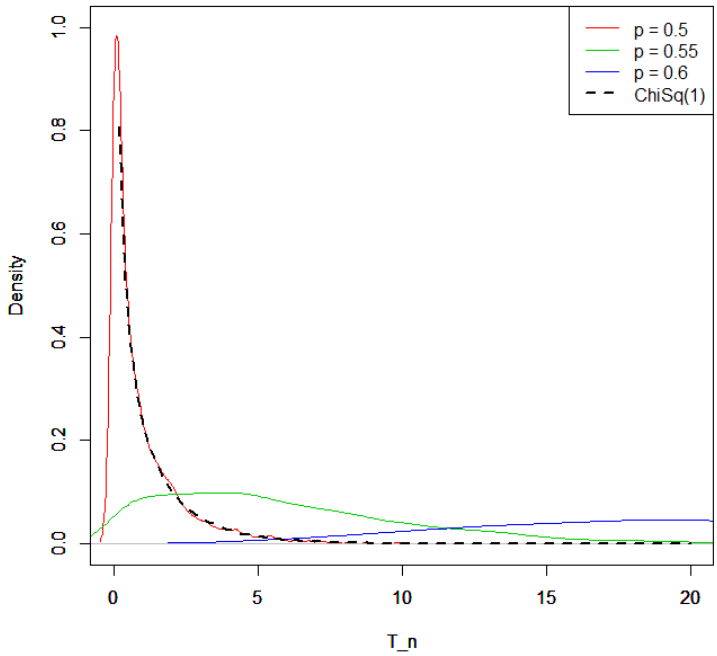
Next, we calculate the test statistic $T_n$ for each $\mathbb{H}_0$ for 5000 replicated datasets:

```
nrep <- 5000
T0   <- numeric(nrep)
T1   <- numeric(nrep)
T2   <- numeric(nrep)
```

```
for(i in 1:nrep){

        X   <- rbinom(500, 1, 0.5)  # flipping an even coin 500 times
        mle <- mean(X)              # MLE = sample average

        T0[i] <- 2*(LL(mle) - LL(0.5 ))
        T1[i] <- 2*(LL(mle) - LL(0.55))
        T2[i] <- 2*(LL(mle) - LL(0.6 ))

        }
```

Finally, plot the densities of the calculated $T_n$'s:

```
xlm <- c(0,20)
ylm <- c(0,1)
xlb <- "T_n"
ylb <- "Density"

plot(density(T0), xlim = xlm, ylim = ylm, col = 2, main = "", xlab = xlb, ylab = ylb)
par(new = T)
plot(density(T1), xlim = xlm, ylim = ylm, col = 3, main = "", xlab = xlb, ylab = ylb)
par(new = T)
plot(density(T2), xlim = xlm, ylim = ylm, col = 4, main = "", xlab = xlb, ylab = ylb)
par(new = T)
curve(dchisq(x,1), xlim = xlm, ylim = ylm, lwd = 2, lty = 2, main = "", xlab = xlb, ylab = ylb)
legend(
       "topright",
       c("p = 0.5", "p = 0.55", "p = 0.6", "ChiSq(1)"),
       lwd = c(1,1,1,2),
       lty = c(1,1,1,2),
       col = c(2,3,4,1)
       )
```



## 3.6 Akaike's Information Criterion

When the number of regressors is large and only some of them are meaningful, including all of them in the model may cause an **overfitting** problem. That is, using too many explanatory variables results in a model that adapts to the data too well, and shows poor performance in out-of-sample predictions. In an extreme case, when the number of regressors is equal to the sample size $n$, it is possible to achieve a "perfect" fit to the data with zero residuals, but such a model has no predictive power.

In order to avoid the overfitting problem, we need to select the regressors that should be included in the model using some criterion. Note that the log-likelihood is not a valid variable selection criterion since it monotonically increases in the number of included regressors; model selection based on the log-likelihood value always results in the largest model. A "good" model should be defined as a model whose likelihood function well approximates the "true" data distribution, rather than the empirical distribution of the observed data.

Here, let $f_X(x)$ be the true density function of $X$. The parameterized density $p(x;\theta_0)$ may or may not be equal to $f_X(x)$ (if they are not same, $\theta_0$ is interpreted as a "pseudo" true parameter). Let $\widehat{\theta}_n = \theta_n(X_1,\ldots,X_n)$ be the MLE from our model and $p(x;\widehat{\theta}_n)$ be the estimated likelihood function. Then, taking the sampling error involved in the estimate of $\widehat{\theta}_n$ into account, we may conclude that our modelling is optimal if the following statistic is sufficiently small:

$$\begin{aligned}\mathbb{E}_{data}K(f_X(X)\|p(X;\widehat{\theta}_n)) &= \mathbb{E}_{data}\mathbb{E}_X\left[\log\frac{f_X(X)}{p(X;\widehat{\theta}_n)}\right] \quad \left(= \int\int \log\frac{f(x)}{p(x;\theta_n(x_1,\ldots,x_n))} f_X(x)\mathrm{d}x\prod_{i=1}^n f(x_i)\mathrm{d}x_i\right)\\ &= \mathbb{E}_X[\log f_X(X)] - \mathbb{E}_{data}\mathbb{E}_X[\log p(X;\widehat{\theta}_n)],\end{aligned}$$

which is the expected KL divergence between the true data distribution and the estimated distribution. Since the first term is independent of the model selection, minimizing the KL divergence over the whole model space is equal to maximizing $\mathbb{E}_{data}\mathbb{E}_X[\log p(X;\widehat{\theta}_n)]$.

By using the second order Taylor expansion around $\theta_0$, we have

$$\log p(X;\widehat{\theta}_n) \approx \log p(X;\theta_0) + s(X;\theta_0)^\top(\widehat{\theta}_n - \theta_0) + R_n(X;\widehat{\theta}_n),$$

where $R_n(X;\widehat{\theta}_n) \equiv \frac{1}{2}(\widehat{\theta}_n - \theta_0)^\top H(X;\theta_0)(\widehat{\theta}_n - \theta_0)$. Further, by (3.2.1),

$$\mathbb{E}_X[\log p(X;\widehat{\theta}_n)] - \mathbb{E}_X[\log p(X;\theta_0)] \approx \mathbb{E}_X[R_n(X;\widehat{\theta}_n)]. \tag{3.6.1}$$

Also, letting $\overline{s}_n(\theta_0) \equiv \frac{1}{n}\sum_{i=1}^n s(X_i;\theta_0)$, observe that by (3.3.1) and the information matrix equality (3.3.2),

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \log p(X_i;\widehat{\theta}_n) - \mathbb{E}_X[\log p(X;\theta_0)] &\approx \frac{1}{n}\sum_{i=1}^n\left(\log p(X_i;\widehat{\theta}_n) - \log p(X_i;\theta_0)\right)\\ &\approx \overline{s}_n(\theta_0)^\top(\widehat{\theta}_n - \theta_0) + \frac{1}{n}\sum_{i=1}^n R_n(X_i;\widehat{\theta}_n)\\ &\approx -\overline{s}_n(\theta_0)^\top\underbrace{\left[\frac{1}{n}\sum_{i=1}^n H(X_i;\overline{\theta}_n)\right]^{-1}}_{\xrightarrow{p} -[I(\theta_0)]^{-1}}\overline{s}_n(\theta_0) + \mathbb{E}_X[R_n(X;\widehat{\theta}_n)].\end{aligned} \tag{3.6.2}$$

Combining (3.6.1) and (3.6.2) yields

$$\frac{1}{n}\sum_{i=1}^n \log p(X_i;\widehat{\theta}_n) - \mathbb{E}_X[\log p(X;\widehat{\theta}_n)] \approx \overline{s}_n(\theta_0)^\top[I(\theta_0)]^{-1}\overline{s}_n(\theta_0).$$

Then, under the IID assumption,

$$\begin{aligned}\mathbb{E}_{data}\left[\frac{1}{n}\sum_{i=1}^{n}\log p(X_i;\widehat{\theta}_n) - \mathbb{E}_X[\log p(X;\widehat{\theta}_n)]\right] &\approx \mathbb{E}_{data}\left[\overline{s}_n(\theta_0)^\top [I(\theta_0)]^{-1}\overline{s}_n(\theta_0)\right] \\ &= \text{trace}\left\{[I(\theta_0)]^{-1}\mathbb{E}_{data}\left[\overline{s}_n(\theta_0)\overline{s}_n(\theta_0)^\top\right]\right\} \\ &= \text{trace}\left\{[I(\theta_0)]^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\underbrace{\mathbb{E}_{data}\left[s(X_i;\theta_0)s(X_i;\theta_0)^\top\right]}_{=I(\theta_0)}\right)\right\}/n \\ &= \frac{\text{trace}\left\{I_{\dim(\theta_0)}\right\}}{n} = \frac{\dim(\theta_0)}{n}.\end{aligned}$$

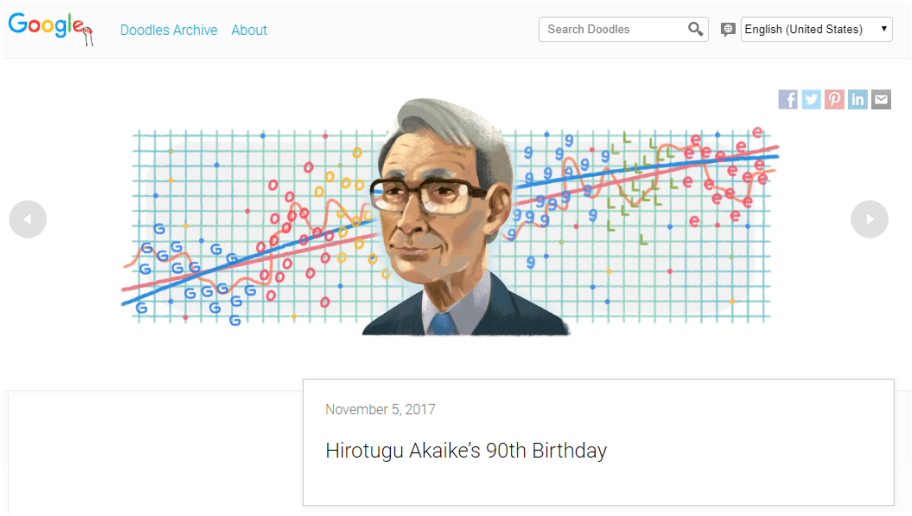
Finally, this implies that we can approximate $\mathbb{E}_{data}\mathbb{E}_X[\log p(X;\widehat{\theta}_n)]$ by

$$\begin{aligned}\mathbb{E}_{data}\mathbb{E}_X[\log p(X;\widehat{\theta}_n)] &\approx \frac{1}{n}\sum_{i=1}^{n}\log p(X_i;\widehat{\theta}_n) - \frac{\dim(\theta_0)}{n} \\ &= \frac{1}{n}(\ell_n(\widehat{\theta}_n) - \dim(\theta_0)) = -\frac{1}{2n}\text{AIC},\end{aligned}$$

where AIC is defined as

$$\text{AIC} \equiv -2\ell_n(\widehat{\theta}_n) + 2\dim(\theta_0).$$

Hence, minimizing the AIC is approximately equivalent to choosing a model which has minimum KL divergence to the true data distribution. This way of model selection was first proposed in [Akaike, 1973], which is known as **Akaike's Information Criterion**.




It should be noted that the derivation of the AIC statistic involves a lot of large sample approximations, and it is known that the AIC tends to select a less parsimonious model under finite sample size. Then, to improve the finite sample model selection performance, a number of modified versions of AIC has been proposed in the literature.

## Appendix: Irregular estimators

Let $\{X_1, \ldots, X_n\}$ be an IID sample drawn from $N(\mu_0, 1)$. Then, the MLE for $\mu_0$ is the sample average $\overline{X}_n$, and we have the asymptotic normality result: $\sqrt{n}(\overline{X}_n - \mu_0) \xrightarrow{d} N(0,1)$. Note that this argument holds for any given $\mu_0$.

Now, we consider the following estimator:

$$\widehat{\mu}_n = \begin{cases} \overline{X}_n & \text{if} \quad |\overline{X}_n| \geq n^{-1/4} \\ 0 & \text{if} \quad |\overline{X}_n| < n^{-1/4} \end{cases}$$

This estimator is known as **Hodge's estimator**. Namely, Hodge's estimator returns the MLE when $|\overline{X}_n| \geq n^{-1/4}$ and zero when $|\overline{X}_n|$ does not exceed $n^{-1/4}$.

To derive the limiting distribution of Hodge's estimator, we first consider the case where $\mu_0 \neq 0$. Then, since $|\mu_0| - n^{-1/4} > 0$ for sufficiently large $n$, we have

$$\begin{aligned} \Pr(\widehat{\mu}_n \neq \overline{X}_n) &= \Pr(|\mu_0 - (\mu_0 - \overline{X}_n)| < n^{-1/4}) \\ &\leq \Pr(|\mu_0| - |\mu_0 - \overline{X}_n| < n^{-1/4}) \\ &= \Pr(|\mu_0| - n^{-1/4} < |\mu_0 - \overline{X}_n|) \to 0 \end{aligned}$$

as $n \to \infty$. This implies that $\widehat{\mu}_n$ has the same limiting distribution as the MLE; that is, $\sqrt{n}(\widehat{\mu}_n - \mu_0) \xrightarrow{d} N(0,1)$.

Next, consider the case where $\mu_0 = 0$. In this case, for any $\kappa > 0$,

$$\begin{aligned} \Pr\left(\left|\sqrt{n}\widehat{\mu}_n\right| > \kappa\right) &= \Pr\left(\left|\sqrt{n}\widehat{\mu}_n\right| > \kappa, \widehat{\mu}_n = 0\right) + \Pr\left(\left|\sqrt{n}\widehat{\mu}_n\right| > \kappa, \widehat{\mu}_n \neq 0\right) \\ &= \Pr(\left|\sqrt{n}\widehat{\mu}_n\right| > \kappa, \widehat{\mu}_n \neq 0) \leq \Pr(\widehat{\mu}_n \neq 0). \end{aligned}$$

Note that

$$\begin{aligned} \Pr(\widehat{\mu}_n \neq 0) &= \Pr(|\overline{X}_n| \geq n^{-1/4}) \\ &= \Pr(\sqrt{n}|\overline{X}_n - \mu_0| \geq n^{1/4}) \to 0 \end{aligned}$$

as $n \to \infty$ since $\sqrt{n}|\overline{X}_n - \mu_0|$ is asymptotically equivalent to $|N(0,1)|$. Hence, $\sqrt{n}\widehat{\mu}_n \xrightarrow{p} 0$ and, in view of Lemma 1.4.2, $\sqrt{n}\widehat{\mu}_n \xrightarrow{d} 0$. To summarize, the limiting distribution of Hodge's estimator is

$$\sqrt{n}(\widehat{\mu}_n - \mu_0) \xrightarrow{d} \begin{cases} N(0,1) & \text{if} \quad \mu_0 \neq 0 \\ 0 & \text{if} \quad \mu_0 = 0. \end{cases}$$

Although it is tempting to interpret this result as indicating that Hodge's estimator is better than the MLE, Hodge's estimator has a serious drawback. When $\mu_0$ is away from zero, $\widehat{\mu}_n$ behaves almost like the MLE. However, as $\mu_0$ gets closer to zero, it differs from the MLE quite often. To see this, we investigate the performance of $\widehat{\mu}_n$ when $\mu_0 = h/n^{1/4}$ for some $0 < h < 1$. The data $\{X_1, \ldots, X_n\}$ are drawn from $N(h/n^{1/4}, 1)$ for each $n$. The asymptotic distribution of the MLE does not change under $\mu_0 = h/n^{1/4}$: $\sqrt{n}(\overline{X}_n - h/n^{1/4}) \xrightarrow{d} N(0,1)$. On the other hand,

$$\begin{aligned} \Pr(\widehat{\mu}_n = 0) &= \Pr(|\overline{X}_n| < n^{-1/4}) \\ &= \Pr(-n^{-1/4} < \overline{X}_n < n^{-1/4}) \\ &= \Pr\left(-\sqrt{n}(n^{-1/4} + hn^{-1/4}) < \sqrt{n}(\overline{X}_n - hn^{-1/4}) < \sqrt{n}(n^{-1/4} - hn^{-1/4})\right) \\ &= \Pr\left(-n^{1/4}(1+h) < \sqrt{n}(\overline{X}_n - hn^{-1/4}) < n^{1/4}(1-h)\right) \to 1. \end{aligned}$$

Hence,

$$\sqrt{n}(\widehat{\mu}_n - h/n^{1/4}) = hn^{1/4} + o_P(1) \to \infty.$$

Thus, the MSE of Hodge's estimator $\mathbb{E}[\{\sqrt{n}(\widehat{\mu}_n - \mu_0)\}^2]$ explodes to infinity as $n$ increases when $\mu_0$ is close to zero, whereas $\mathbb{E}[\{\sqrt{n}(\overline{X}_n - \mu_0)\}^2] = 1$ holds for any $\mu_0$. Figure 3.1 shows the graphs of the MSE of Hodge's estimator for three different $n$'s. The R code to create this figure is attached below. Hodge's estimator "buys" its extremely better asymptotic behavior exactly at $\mu_0 = 0$ at the cost of erratic behavior close to zero.

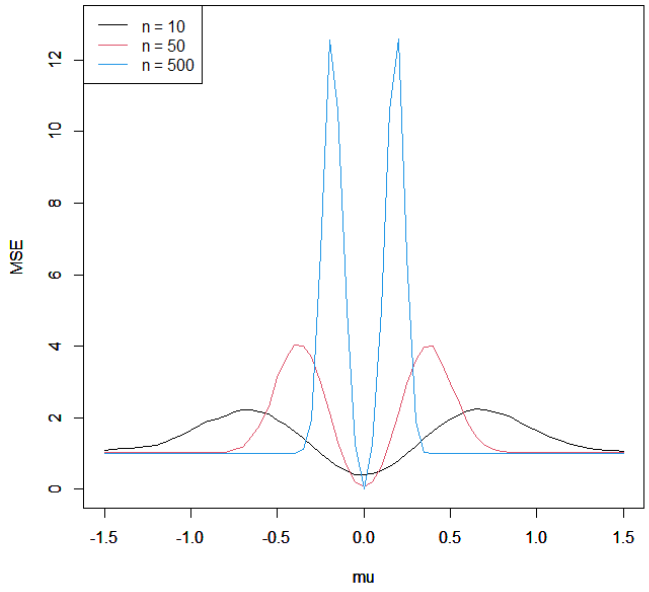



Figure 3.1: MSE of Hodge's estimator

The above discussion shows that in order to have a meaningful discussion on the efficiency of estimators, we need to rule out pathological estimators such as Hodge's estimator and focus on "regular" estimators. An estimator $\widehat{\theta}_n$ is called **regular** if a small change in the true parameter $\theta_0$ changes the distribution of the estimator only slightly. Hodge's estimator is a typical example of "irregular" estimators. Within the class of regular estimators, we can show the optimality of MLE. For more precise discussion, see, for example, Section 8.5 of [Van der Vaart, 2000].

**R code to create Figure 3.1**

```
ns  <- c(10, 50, 500)
mus <- seq(-1.5, 1.5, 0.05)

mse <- function(n, mu){

        se <- function(seed){
           set.seed(seed)
           X <- rnorm(n, mean = mu, sd = 1)
           Xbar <- mean(X)
           H <- 0*(abs(Xbar) < n^{-1/4}) + Xbar*(abs(Xbar) >= n^{-1/4})
           se <- (sqrt(n)*(H - mu))^2
           return(se)
```

```
            }

   R <- numeric(2000)
   for(r in 1:2000) R[r] <- se(r)
   return(mean(R))

   }

result <- matrix(0, length(ns), length(mus))
for(i in 1:length(ns)) for(j in 1:length(mus)) result[i,j] <- mse(ns[i],mus[j])

plot(mus, result[1,], ylim = c(0,13), xlab = "mu", ylab = "MSE", type = "l", col = 1)
par(new = T)
plot(mus, result[2,], ylim = c(0,13), xlab = "mu", ylab = "MSE", type = "l", col = 2)
par(new = T)
plot(mus, result[3,], ylim = c(0,13), xlab = "mu", ylab = "MSE", type = "l", col = 4)
legend("topleft", legend = c("n = 10", "n = 50", "n = 500"), lty = c(1, 1, 1), col = c(1, 2, 4))
```

# Chapter 4

# Generalized Method of Moments

## 4.1 Moment conditions

Consider the following multiple regression model:

$$Y = X_1^\top \theta_0 + \varepsilon.$$

We assume that the regressor $\underset{d_{x1}\times 1}{X_1}$ is exogenous: $\mathbb{E}[X_1\varepsilon] = \mathbf{0}$, and that $\mathbb{E}[X_1 X_1^\top]$ is nonsingular. Then, one can estimate $\theta_0$ by the OLS estimator based on the following moment equations:

$$\underset{d_{x1}\times 1}{\mathbf{0}} = \mathbb{E}[X_1\varepsilon] = \mathbb{E}[X_1(Y - X_1^\top\theta_0)] \iff \theta_0 = \mathbb{E}[X_1 X_1^\top]^{-1}\mathbb{E}[X_1 Y].$$

Note here that the number of moment equations $\mathbb{E}[X_1(Y - \underset{d_{x1}\times 1}{X_1^\top\theta_0})] = \mathbf{0}$ available is exactly the same as the number of elements in $\underset{d_{x1}\times 1}{\theta_0}$. In this case we say that $\theta_0$ is **just-identified**.[1]

Now suppose that we know that not only $X_1$ but also $\underset{d_{x2}\times 1}{X_2}$ satisfies the same moment equality $\mathbb{E}[X_2\varepsilon] = \mathbf{0}$.[2] For simplicity suppose that $d_{x1} = d_{x2}$. These additional $d_{x2}$ moment conditions $\mathbb{E}[X_2(Y - \underset{d_{x2}\times 1}{X_1^\top\theta_0})] = \mathbf{0}$ can be utilized in the estimation of $\theta_0$. That is, if $\mathbb{E}[X_2 X_1^\top]^{-1}$ exists, we can characterize $\theta_0$ alternatively by

$$\theta_0 = \mathbb{E}[X_2 X_1^\top]^{-1}\mathbb{E}[X_2 Y].$$

This situation is called **over-identification**; that is, the number of available moment equations is larger than the number of unknowns. In other words, we can write down $\theta_0$ as a function of the moments in several different ways.

The above example is a special case of a more general class of moment restriction models. Let $g(W;\theta)$ be an $\mathbb{R}^J$-valued function, where $\theta$ is a $K \times 1$ vector of parameters, and $W$ is a vector of observable random variables. We assume that $J \geq K$. Suppose that the following moment conditions hold:

$$\mathbb{E}[g(W;\theta_0)] = \underset{J\times 1}{\mathbf{0}}, \tag{4.1.1}$$

[1] It is well-known in linear algebra that when the number of independent linear equations equals the number of unknowns, the system can be uniquely solved.

[2] $X_2$ could be a function of $X_1$; see also footnote 6 in Chapter 1.

where $\theta_0$ is the true value of $\theta$. In the above multiple regression model, $g(W;\theta_0)$ corresponds to $(X_1^\top, X_2^\top)^\top(Y - X_1^\top\theta_0)$, and $W$ corresponds to $(Y, X_1, X_2)$; thus, in this case, $J = d_{x1} + d_{x2}$ and $K = d_{x1}$ (note that any two linearly dependent moment conditions are counted as the same moment condition). If $J = K$, we say that $\theta_0$ is just-identified, and if $J > K$, it is over-identified.

**Example 4.1.1 (Linear regression)** Consider the following linear regression model with exogenous regressors:

$$Y = X^\top\theta_0 + \varepsilon, \quad \mathbb{E}[X\varepsilon] = \mathbf{0}.$$

Then, setting $g(W;\theta) = X(Y - X^\top\theta)$ gives (4.1.1) with $W = (Y, X)$.

**Example 4.1.2 (Instrumental variables regression)** Consider the following linear regression model with both endogenous and exogenous regressors:

$$Y = D^\top\alpha_0 + X^\top\beta_0 + \varepsilon, \quad \mathbb{E}[D\varepsilon] \neq \mathbf{0}, \quad \mathbb{E}[X\varepsilon] = \mathbf{0}.$$

Suppose that we have a set of instrumental variables $Z_1$ for $D$, and let $Z = (Z_1^\top, X^\top)^\top$. Then, we have $\mathbb{E}[Z\varepsilon] = \mathbf{0}$, which gives (4.1.1) by setting $g(W;\theta) = Z(Y - D^\top\alpha - X^\top\beta)$ with $\theta_0 = (\alpha_0^\top, \beta_0^\top)^\top$ and $W = (Y, D, X, Z_1)$.

**Example 4.1.3 (Probit model 1)** Consider the following probit model:

$$D = \mathbf{1}\{X^\top\theta_0 \geq \varepsilon\}, \quad \varepsilon \sim N(0,1).$$

Assuming that $\varepsilon$ is independent of $X$, the conditional expectation of $D$ given $X$ is $\mathbb{E}[D \mid X] = \Phi(X^\top\theta_0)$, where $\Phi$ is the standard normal distribution function. Trivially, this implies that

$$\mathbb{E}[D - \Phi(X^\top\theta_0) \mid X] = 0.$$

This "conditional" moment restriction implies the "unconditional" moment restrictions $\mathbb{E}[X(D - \Phi(X^\top\theta_0))] = \mathbf{0}$ by LIE:

$$\begin{aligned}\mathbb{E}[X(D - \Phi(X^\top\theta_0))] &= \mathbb{E}\left\{\mathbb{E}[X(D - \Phi(X^\top\theta_0)) \mid X]\right\} \\ &= \mathbb{E}\left\{X\mathbb{E}[D - \Phi(X^\top\theta_0) \mid X]\right\} = \mathbb{E}\left(X \times 0\right) = \mathbf{0}.\end{aligned}$$

Hence, setting $g(W;\theta) = X(D - \Phi(X^\top\theta))$ with $W = (D, X)$, the probit model also fits in the framework of (4.1.1).

**Example 4.1.4 (Probit model 2)** There are other moment conditions that can be used to estimate the probit model. First, note that the true parameter $\theta_0$ can be characterized as the maximizer of the population log-likelihood function (see Section 3.1.2):

$$\theta_0 = \operatorname*{argmax}_\theta \mathbb{E}\left[D \cdot \log\Phi(X^\top\theta) + (1 - D) \cdot \log(1 - \Phi(X^\top\theta))\right].$$

The first order condition for this maximization problem is

$$\mathbb{E}\left[X\eta(\theta_0)\right] = \mathbf{0},$$

where

$$\eta(\theta_0) \equiv \frac{(D - \Phi(X^\top\theta_0)) \cdot \phi(X^\top\theta_0)}{\Phi(X^\top\theta_0) \cdot (1 - \Phi(X^\top\theta_0))},$$

and $\phi$ is the standard normal density function. Thus, the score function $X\eta(\theta)$ serves as $g(W;\theta)$

Here, the term $\eta(\theta_0)$ is referred to as the "generalized residual" for the probit model; this is the conditional expectation of $-\varepsilon$ given $(D, X)$. Namely, using the formula for the expectation of a truncated random variable (see Appendix B.3) and the fact that $-e\phi(e) = \phi'(e)$, we have

$$\begin{aligned}
\mathbb{E}[-\varepsilon \mid X, D=1] = \mathbb{E}[-\varepsilon \mid X, X^\top\theta_0 \geq \varepsilon] &= -\frac{\int_{-\infty}^{X^\top\theta_0} e\phi(e)\mathrm{d}e}{\Phi(X^\top\theta_0)} \\
&= \frac{\int_{-\infty}^{X^\top\theta_0} \phi'(e)\mathrm{d}e}{\Phi(X^\top\theta_0)} = \underbrace{\frac{\phi(X^\top\theta_0)}{\Phi(X^\top\theta_0)}}_{=\eta(\theta_0)|_{D=1}} \\
\mathbb{E}[-\varepsilon \mid X, D=0] = \mathbb{E}[-\varepsilon \mid X, X^\top\theta_0 < \varepsilon] &= -\frac{\int_{X^\top\theta_0}^{\infty} e\phi(e)\mathrm{d}e}{1-\Phi(X^\top\theta_0)} \\
&= \frac{\int_{X^\top\theta_0}^{\infty} \phi'(e)\mathrm{d}e}{1-\Phi(X^\top\theta_0)} = \underbrace{-\frac{\phi(X^\top\theta_0)}{1-\Phi(X^\top\theta_0)}}_{=\eta(\theta_0)|_{D=0}}.
\end{aligned}$$

## 4.2 GMM procedure

When the data of sample size $n$ are available, the sample analogue of $\mathbb{E}[g(W;\theta)]$ can be obtained by

$$\overline{g}_n(\theta) \equiv \frac{1}{n}\sum_{i=1}^{n} g(W_i;\theta).$$

Then, by the law of large numbers, we can expect that $\overline{g}_n(\theta_0) \approx \mathbf{0}$ for sufficiently large $n$. This implies that one can consider estimating $\theta_0$ by solving

$$\overline{g}_n(\theta) = \mathbf{0} \tag{4.2.1}$$

with respect to $\theta$. This type of estimator is called the **method of moments** estimator. The method of moments estimator has a unique solution when $J = K$ (i.e., just-identification) under appropriate rank conditions. However, in general $J > K$, the parameter value that exactly satisfies (4.2.1) may not exist.

The idea of the **generalized method of moments** (GMM) is to obtain an estimator of $\theta_0$ by finding a $\theta$ that makes $\overline{g}_n(\theta)$ as close to zero as possible. To do this, let $\Omega_n$ be a $J \times J$ positive definite symmetric "weight" matrix. Then, the GMM estimator of $\theta_0$ is defined by

$$\widehat{\theta}_n^{gmm} \equiv \operatorname*{argmin}_{\theta} \overline{g}_n(\theta)^\top \Omega_n \overline{g}_n(\theta). \tag{4.2.2}$$

Since the objective function in (4.2.2) has a quadratic form and $\Omega_n$ is positive definite, it holds that $\overline{g}_n(\theta)^\top\Omega_n\overline{g}_n(\theta) \geq 0$ for any $\theta$. One can simply choose an identity matrix $I_J$ of dimension $J$ as $\Omega_n$. In this case, $\widehat{\theta}_n^{gmm}$ is defined as the minimizer of $||\overline{g}_n(\theta)||^2$. However, as discussed later, the choice of the weight matrix affects the efficiency of the estimator.

**Example 4.2.1 (Linear regression (cont.))** Letting $g(W_i;\theta) = X_i(Y_i - X_i^\top\theta)$, we have

$$\overline{g}_n(\theta) = \mathbf{X}_n^\top(\mathbf{Y}_n - \mathbf{X}_n\theta)/n,$$

where $\mathbf{X}_n = (X_1, \ldots, X_n)^\top$, and $\mathbf{Y}_n = (Y_1, \ldots, Y_n)^\top$. Thus, the GMM estimator $\widehat{\theta}_n^{gmm}$ is obtained by

$$\widehat{\theta}_n^{gmm} = \underset{\theta}{\operatorname{argmin}} (\mathbf{Y}_n - \mathbf{X}_n\theta)^\top \mathbf{X}_n \Omega_n \mathbf{X}_n^\top (\mathbf{Y}_n - \mathbf{X}_n\theta).$$

Solving the first order condition gives

$$\begin{aligned} -2\mathbf{X}_n^\top \mathbf{X}_n \Omega_n \mathbf{X}_n^\top (\mathbf{Y}_n - \mathbf{X}_n \widehat{\theta}_n^{gmm}) = \mathbf{0} &\iff \mathbf{X}_n^\top \mathbf{X}_n \Omega_n \mathbf{X}_n^\top (\mathbf{Y}_n - \mathbf{X}_n \widehat{\theta}_n^{gmm}) = \mathbf{0} \\ &\iff \widehat{\theta}_n^{gmm} = \left[\mathbf{X}_n^\top \mathbf{X}_n \Omega_n \mathbf{X}_n^\top \mathbf{X}_n\right]^{-1} \mathbf{X}_n^\top \mathbf{X}_n \Omega_n \mathbf{X}_n^\top \mathbf{Y}_n. \end{aligned}$$

This implies that the OLS estimator is a special case of the GMM estimator in which the weight matrix $\Omega_n$ is chosen as $\Omega_n = (\mathbf{X}_n^\top \mathbf{X}_n)^{-1}$.

**Example 4.2.2 (Instrumental variables regression (cont.))** Let $H = (D^\top, X^\top)^\top$ and $g(W_i; \theta) = Z_i(Y_i - H_i^\top \theta)$. Then, we have

$$\overline{g}_n(\theta) = \mathbf{Z}_n^\top (\mathbf{Y}_n - \mathbf{H}_n\theta)/n,$$

where $\mathbf{Z}_n = (Z_1, \ldots, Z_n)^\top$, and $\mathbf{H}_n = (H_1, \ldots, H_n)^\top$. Similarly as above, the GMM estimator $\widehat{\theta}_n^{gmm}$ can be obtained by

$$\widehat{\theta}_n^{gmm} = \underset{\theta}{\operatorname{argmin}} (\mathbf{Y}_n - \mathbf{H}_n\theta)^\top \mathbf{Z}_n \Omega_n \mathbf{Z}_n^\top (\mathbf{Y}_n - \mathbf{H}_n\theta).$$

By easy calculations,

$$\widehat{\theta}_n^{gmm} = \left[\mathbf{H}_n^\top \mathbf{Z}_n \Omega_n \mathbf{Z}_n^\top \mathbf{H}_n\right]^{-1} \mathbf{H}_n^\top \mathbf{Z}_n \Omega_n \mathbf{Z}_n^\top \mathbf{Y}_n.$$

Thus, when setting the weight matrix $\Omega_n$ as $\Omega_n = (\mathbf{Z}_n^\top \mathbf{Z}_n)^{-1}$, we can see that the GMM estimator coincides with the 2SLS estimator (see (2.3.2)).

## 4.3 Asymptotic properties

Let $\Omega \equiv \operatorname{plim}_{n\to\infty} \Omega_n$. Suppose that the true parameter $\theta_0$ can be characterized as

$$\theta_0 = \underset{\theta}{\operatorname{argmin}}\, \mathbb{E}[g(W;\theta)]^\top \Omega \mathbb{E}[g(W;\theta)].$$

Then, under some regularity conditions, one can show that the GMM estimator $\widehat{\theta}_n^{gmm}$ is consistent for $\theta_0$ (see, e.g., [Newey and McFadden, 1994]).

The asymptotic distribution of $\widehat{\theta}_n^{gmm}$ can be derived as follows. Hereinafter, we suppress the superscript "*gmm*" for notational simplicity. By the first order condition for the minimization in (4.2.2), it holds that

$$\overline{g}_n'(\widehat{\theta}_n)^\top \Omega_n \overline{g}_n(\widehat{\theta}_n) = \underset{K\times 1}{\mathbf{0}},$$

where $\overline{g}_n'(\theta) = \underset{J\times K}{\partial \overline{g}_n(\theta)/\partial\theta^\top}$. Applying the mean-value expansion to $\overline{g}_n(\widehat{\theta}_n)$ around $\theta_0$, we have

$$\mathbf{0} = \overline{g}_n'(\widehat{\theta}_n)^\top \Omega_n \left[\overline{g}_n(\theta_0) + \overline{g}_n'(\overline{\theta}_n)(\widehat{\theta}_n - \theta_0)\right],$$

where $\overline{\theta}_n \in [\widehat{\theta}_n, \theta_0]$. This implies that

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) = -\left[\overline{g}_n'(\widehat{\theta}_n)^\top \Omega_n \overline{g}_n'(\overline{\theta}_n)\right]^{-1} \overline{g}_n'(\widehat{\theta}_n)^\top \Omega_n (\sqrt{n}\overline{g}_n(\theta_0))$$

$$= -\left[\overline{g}_n'(\widehat{\theta}_n)^\top \Omega_n \overline{g}_n'(\overline{\theta}_n)\right]^{-1} \overline{g}_n'(\widehat{\theta}_n)^\top \Omega_n \left[\frac{1}{\sqrt{n}} \sum_{i=1}^n g(W_i;\theta_0)\right].$$

Letting $V \equiv \mathbb{E}[g(W;\theta_0)g(W;\theta_0)^\top]$ and assuming that the data are IID, by CLT, we have

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n g(W_i;\theta_0) \xrightarrow{d} \mathbf{N}(\mathbf{0}, V).$$

In addition, by WLLN, $\overline{g}_n'(\theta_0) \xrightarrow{p} M \equiv \mathbb{E}[\partial g(W;\theta_0)/\partial\theta^\top]$. Note that if $\widehat{\theta}_n$ is a consistent estimator of $\theta_0$, so is $\overline{\theta}_n$ by its definition. Thus, by the continuous mapping theorem, we obtain both $\overline{g}_n'(\widehat{\theta}_n) \xrightarrow{p} M$ and $\overline{g}_n'(\overline{\theta}_n) \xrightarrow{p} M$. Finally, by Slutsky's theorem B.1.2, we have the asymptotic normality of $\sqrt{n}(\widehat{\theta}_n - \theta_0)$ as follows:

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{d} \mathbf{N}\left(\mathbf{0}, (M^\top \Omega M)^{-1} M^\top \Omega V \Omega M (M^\top \Omega M)^{-1}\right). \tag{4.3.1}$$

**Exercise 4.3.1** Consider a linear regression model:

$$Y = X^\top \beta_0 + \varepsilon,$$

where the conditional median of $\varepsilon$ given $X$ is zero: $\text{Med}[\varepsilon \mid X] = 0$. Describe how to construct a GMM estimator of $\beta_0$.

**Exercise 4.3.2** Consider an endogenous probit model:

$$Y = \mathbf{1}\{\beta_0 + D\alpha_0 \geq \varepsilon\}, \quad \varepsilon \sim N(0,1),$$

where $D$ is potentially dependent on $\varepsilon$. Suppose that there is an IV $Z_1$ for $D$ such that it is a determinant of $D$ and is independent of $\varepsilon$. An analyst tries to estimate $\theta_0 = (\beta_0, \alpha_0)^\top$ using GMM with $g(W;\theta) = Z(Y - \Phi(\beta + D\alpha))$ by analogy from Example 4.1.3, where $Z = (1, Z_1)^\top$, and $W = (Y, D, Z_1)$. However, such an estimator is generally inconsistent. Explain why.

## 4.4 Two-step optimal GMM

It can be seen from (4.3.1) that the variance of the GMM estimator depends on the choice of the weight matrix $\Omega_n$. An optimal GMM estimator that achieves the smallest possible asymptotic variance can be obtained by setting $\Omega_n = V_n^{-1}$ such that $\text{plim}_{n\to\infty} \Omega_n = V^{-1}$. With this weight matrix, the asymptotic covariance matrix in (4.3.1) can be simplified to

$$(M^\top V^{-1} M)^{-1} M^\top V^{-1} V V^{-1} M (M^\top V^{-1} M)^{-1} = (M^\top V^{-1} M)^{-1}.$$

The optimality of this weight matrix can be easily shown as follows. For any $\Omega$, observe that

$$\begin{aligned}
&(M^\top \Omega M)^{-1} M^\top \Omega V \Omega M (M^\top \Omega M)^{-1} - (M^\top V^{-1} M)^{-1} \\
&= (M^\top \Omega M)^{-1} M^\top \Omega V^{1/2} \left\{I_J - V^{-1/2} M (M^\top V^{-1} M)^{-1} M^\top V^{-1/2}\right\} V^{1/2} \Omega M (M^\top \Omega M)^{-1} \\
&= (M^\top \Omega M)^{-1} M^\top \Omega V^{1/2} \left\{I_J - P(P^\top P)^{-1} P^\top\right\} V^{1/2} \Omega M (M^\top \Omega M)^{-1},
\end{aligned} \tag{4.4.1}$$

where $I_J$ is an identity matrix of dimension $J$, and $P \equiv V^{-1/2}M$. Note that the matrix $I_J - P(P^\top P)^{-1}P^\top$ is idempotent, and thus its eigenvalues are either 0 or 1.[3] Therefore, the left-hand side in (4.4.1) is positive semidefinite, implying the optimality of $\Omega_n = V_n^{-1}$.

Recall that $V = \text{Cov}[g(W;\theta_0)]$. Thus, its sample analog $V_n$ can be constructed by

$$V_n = \frac{1}{n}\sum_{i=1}^{n} g(W_i;\theta_0)g(W_i;\theta_0)^\top.$$

However, since the true parameter $\theta_0$ is unknown, $V_n$ is also unknown, and the optimal GMM is not feasible. To estimate $V_n$, we first need to obtain a consistent GMM estimator $\widehat{\theta}_n^{gmm}$ of $\theta_0$ with an arbitrary weighting matrix, such as $\Omega_n = I_J$. This estimator is consistent but not fully efficient, in general. Then, we can consistently estimate $V_n$ by

$$\widehat{V}_n = \frac{1}{n}\sum_{i=1}^{n} g(W_i;\widehat{\theta}_n^{gmm})g(W_i;\widehat{\theta}_n^{gmm})^\top,$$

and re-estimate $\theta_0$ by

$$\widehat{\theta}_n^{opt} \equiv \underset{\theta}{\text{argmin}}\, \overline{g}_n(\theta)^\top \widehat{V}_n^{-1}\overline{g}_n(\theta).$$

The estimator $\widehat{\theta}_n^{opt}$ is called the **two-step optimal GMM estimator**. The two-step estimator has the same asymptotic distribution as the infeasible optimal GMM:

$$\sqrt{n}(\widehat{\theta}_n^{opt} - \theta_0) \xrightarrow{d} \mathbf{N}\left(\mathbf{0}, (M^\top V^{-1} M)^{-1}\right).$$

**Example 4.4.1 (Probit model 2 (cont.))** Let $g(W;\theta) = X\eta(\theta)$, where the $\eta(\theta)$ is as defined in Example 4.1.4. Since $g(W;\theta)$ is the score function of the log-likelihood, the matrix $V$ is equal to the Fisher information matrix $I(\theta)$. Furthermore, recalling the definition $M \equiv \mathbb{E}[\partial g(W;\theta_0)/\partial\theta^\top]$, we can see that $M$ is the expected Hessian matrix of the log-likelihood: $M = \mathbb{E}[H(W;\theta_0)]$. Finally, by the information matrix equality (3.3.2), we have

$$\sqrt{n}(\widehat{\theta}_n^{opt} - \theta_0) \xrightarrow{d} \mathbf{N}\left(\mathbf{0}, [I(\theta_0)]^{-1}\right).$$

Thus, the optimal GMM estimator is asymptotically equivalent to the MLE in this case.

**Exercise 4.4.2** 1. Consider the multiple regression models in Example 4.1.1. Prove that the OLS estimator is equivalent to the optimal GMM estimator under homoskedasticity.

2. Consider the IV regression models in Example 4.1.2. Prove that the 2SLS estimator is equivalent to the optimal GMM estimator under homoskedasticity.

**Exercise 4.4.3** Consider the following model: $Y = \theta_0 + \varepsilon$ with $\mathbb{E}[\varepsilon] = 0$ and, for a $J \times 1$ vector of random variables $X$, $\mathbb{E}[X\varepsilon] = \mathbf{0}$. Assuming that IID data $\{(Y_i, X_i) : 1 \le i \le n\}$ are available, derive the optimal GMM estimator of $\theta_0$.

[3] This can be easily proved. Let $A$ be an idempotent matrix such that $AA = A$, and $\lambda$ and $x$ be an eigenvalue of $A$ and its corresponding eigenvector, respectively. Then, we have

$$\lambda x = Ax = AAx = A(\lambda x) = \lambda Ax = \lambda^2 x.$$

Since the eigenvector $x$ is a nonzero vector by definition, this is possible only when $\lambda = \lambda^2$, implying that $\lambda \in \{0, 1\}$.

## 4.5 Over-identification test

Suppose that we would like to know the validity of our moment conditions by testing the following hypotheses:

$$\mathbb{H}_0 : \mathbb{E}[g(W;\theta_0)] = \mathbf{0}$$
$$\mathbb{H}_1 : \mathbb{E}[g(W;\theta)] \neq \mathbf{0} \text{ for all } \theta.$$

In the case of just-identification $J = K$, it is possible to find a GMM estimator that exactly solves $\overline{g}_n(\widehat{\theta}_n^{gmm}) = \mathbf{0}$. Note that such $\widehat{\theta}_n^{gmm}$ may exist even when $\mathbb{H}_1$ is true.[4] Therefore, it is generally impossible to statistically test the validity of the moment conditions in a just-identified model.

Suppose now that $J > K$. Then, by solving $K$ moment equations, we can set them equal to zero. If all $J$ moment conditions are actually valid, then the remaining $J - K$ moment conditions should also be close to zero, otherwise they are away from zero. This implies that in an over-identified model, we can test the validity of $J - K$ over-identifying moment conditions. Under $\mathbb{H}_0$, by CLT, we have

$$\sqrt{n}V^{-1/2}\overline{g}_n(\theta_0) \xrightarrow{d} \mathbf{N}(\mathbf{0}, I_J).$$

Hence, $n\overline{g}_n(\theta_0)V^{-1}\overline{g}_n(\theta_0)$ has a chi-square limit distribution, and we can test $\mathbb{H}_0$ based on this fact. However note that $\theta_0$ and $V$ are unknown. Thus, replacing them by their consistent estimators, we can show under regularity conditions that

$$n\overline{g}_n(\widehat{\theta}_n^{gmm})^\top \widehat{V}_n^{-1}\overline{g}_n(\widehat{\theta}_n^{gmm}) \xrightarrow{d} \chi^2(J-K)$$

under $\mathbb{H}_0$. Thus, if the value of $n\overline{g}_n(\widehat{\theta}_n^{gmm})^\top \widehat{V}_n^{-1}\overline{g}_n(\widehat{\theta}_n^{gmm})$ is sufficiently large compared with the critical value of $\chi^2(J-K)$, then we can reject $\mathbb{H}_0$. This testing approach is called **Hansen's over-identification test** (also referred to as Hansen's $J$ test). Note that the test statistic of the over-identification test is $n$ times the objective function for the optimal GMM. Thus, the over-identification test can be easily performed as a by-product of the optimal GMM.

[4] For example, consider a linear regression model $Y = X^\top\theta_0 + \varepsilon$, where $X$ is endogenous: $\mathbb{E}[X\varepsilon] \neq \mathbf{0}$. Set $g(W;\theta) = X(Y - X^\top\theta)$ based on the wrong model assumption of $\mathbb{E}[X\varepsilon] = \mathbf{0}$. Then, we can find a $\widehat{\theta}_n^{gmm}$ that exactly solves $\overline{g}_n(\widehat{\theta}_n^{gmm}) = \mathbf{0}$ (i.e., the OLS estimator), and this estimator is clearly inconsistent.

# Chapter 5

# Identification

A minimal requirement for an estimator is consistency, that is, the estimator converges in probability to its true value as the sample size tends to infinity. A necessary condition for the existence of consistent estimators for a parameter is that the parameter can be "identified". In econometrics, **identification** means that the parameter of interest can be "uniquely" determined from the observable population data (not from the sample observations). More intuitively, identification means that if we could obtain infinite observations, the true parameter is knowable as a unique value. For example, for maximum likelihood estimation of a parameter $\theta_0$, we can say "$\theta_0$ is identified" if $\theta_0$ uniquely solves the maximization problem in (3.1.2): i.e., $\{\theta : \operatorname{argmax}_\theta \mathbb{E}[\log p(X;\theta)]\}$ is a singleton. If there is another parameter value, say $\theta_1$, that also maximizes $\mathbb{E}[\log p(X;\theta)]$, we cannot distinguish which of the two is the true value. In this case, $\theta_0$ is not identified, and it is impossible to construct a consistent estimator for $\theta_0$.

In the following, we discuss the identification problem in econometrics more formally with some examples. Although there are a number of similar identification concepts in econometrics, all of them eventually ask if the model parameters $\theta_0$ can be recovered from the population observable data. In this chapter, we mainly follow the definitions and terminology in [Hurwicz, 1950] and [Matzkin, 2013].

> The many terms for identification that appear in the econometrics literature include (in alphabetical order): Bayesian identification, causal identification, essential identification, eventual identification, exact identification, first order identification, frequentist identification, generic identification, global identification, identification arrangement, identification at infinity, identification by construction, identification of bounds, ill-posed identification, irregular identification, local identification, nearly-weak identification, nonparametric identification, non-robust identification, nonstandard weak identification, overidentification, parametric identification, partial identification, point identification, sampling identification, semiparametric identification, semi-strong identification, set identification, strong identification, structural identification, thin-set identification, underidentification, and weak identification.
>
> A quote from [Lewbel, 2019]

## 5.1 A numerical illustration: a heteroskedastic probit model

Note that the identification problem is indeed crucial in empirical studies. Even when there seem to be no problems with the data and statistical programming, if identification is not achieved, the resulting estimates are not reliable, or you simply cannot obtain any estimates.

Suppose that a researcher considers estimating the following two heteroskedastic probit models:

$$\text{Model 1 } \ Y = \mathbf{1}\{\beta_0 + X\beta_1 \geq \varepsilon\}, \ \text{ where } \ \varepsilon \sim N(0, \sigma_1^2(X)), \ \ \sigma_1(X) = |X\beta_2|$$
$$\text{Model 2 } \ Y = \mathbf{1}\{\beta_0 + X\beta_1 \geq \varepsilon\}, \ \text{ where } \ \varepsilon \sim N(0, \sigma_2^2(X)), \ \ \sigma_2(X) = |(1 + \sqrt{X})\beta_2|.$$

Here, we assume that $X$ is one-dimensional and non-negative. At first glance, both models are quite similar and seem to be valid. However, as shown numerically below, Model 1 is a faulty model while Model 2 is not.

We generate 1000 observations for each model, where the true parameters $(\beta_0, \beta_1, \beta_2)$ are fixed at $(0, 1, 1)$ for both models:

```
# sample size #
 n <- 1000

# true parameters #
 b0 <- 0
 b1 <- 1
 b2 <- 1

# standard deviation functions #
 sd1 <- function(a) abs(X*a)
 sd2 <- function(a) abs((1 + sqrt(X))*a)

# data generation #
 X <- runif(n,0,2) # X ~ Uniform[0,2]
 e1 <- rnorm(n, mean = 0, sd = sd1(b2))
 e2 <- rnorm(n, mean = 0, sd = sd2(b2))
 Y1 <- ifelse(b0 + X*b1 > e1, 1, 0)
 Y2 <- ifelse(b0 + X*b1 > e2, 1, 0)
```

The log-likelihood functions are as follows:

```
 LL1 <- function(p){
        P1 <- pnorm(p[1] + X*p[2], mean = 0, sd = sd1(p[3]))
        P0 <- 1 - P1
        LL <- Y1*log(P1) + (1 - Y1)*log(P0)
        return(-sum(LL))
        }

 LL2 <- function(p){
        P1 <- pnorm(p[1] + X*p[2], mean = 0, sd = sd2(p[3]))
        P0 <- 1 - P1
        LL <- Y2*log(P1) + (1 - Y2)*log(P0)
        return(-sum(LL))
        }
```

The obtained MLEs are

```
MLE1 <- optim(c(0,1,1), LL1, method = "BFGS")
>  MLE1
$par
[1] -0.01150047  2.94351522  2.85940987
$convergence
[1] 0

MLE2 <- optim(c(0,1,1), LL2, method = "BFGS")
>  MLE2
$par
[1] -0.0326636  0.9913823  0.9722578
$convergence
[1] 0
```

Here, `convergence = 0` means that the optimization has been successfully completed. However, the estimated parameters for Model 1 are clearly inconsistent with the true values, whereas we can correctly estimate the true parameters in Model 2.

Where does this difference come from? The only difference between Model 1 and Model 2 is the specification of the standrd deviation function. In fact, the one used in Model 1 has a serious problem that makes the model unidentifiable, whereas the specification in Model 2 does not have such a problem.

In this example, since we know the values of the true parameters, we can detect that there is something wrong with Model 1. However, in real situations where we have no prior knowledge about the true model, it is generally a difficult task to determine whether the estimation results are legitimate or not. One thing we can say for sure is that if the model is not identifiable, the resulting parameter estimates are not consistent.

## 5.2 Definition of identification

Suppose that a dependent variable $Y$ is uniquely determined by a **structural equation**

$$h(Y, X, \varepsilon) = 0,$$

where $X$ is a vector of observable covariates and $\varepsilon$ is an unobserved error. We can treat the marginal distribution of $X$ as known. [Hurwicz, 1950] defines a **structure** as a system of the structural equation and the distribution of $(X, \varepsilon)$. More specifically, the pair $S \equiv (h, F_{\varepsilon,X})$ forms a structure, where $F_{\varepsilon,X}$ denotes the joint distribution of $(X, \varepsilon)$. Then, each $S$ determines the distribution of observable variables $F_{Y,X}$.

Let $F^1_{Y,X}$ and $F^2_{Y,X}$ be the distributions of $(Y, X)$ generated by $S^1 \equiv (h^1, F^1_{\varepsilon,X})$ and $S^2 \equiv (h^2, F^2_{\varepsilon,X})$, respectively.

**Definition 5.2.1** *The structures $S^1$ and $S^2$ are said to be **observationally equivalent** if*

$$F^1_{Y,X} = F^2_{Y,X}.$$

In words, when $S^1$ and $S^2$ are observationally equivalent, we cannot distinguish them from observable information alone because they generate the same distribution of observable variables.

A **model** is defined as a set of such structures with some restrictions on $h$ and $F_{\varepsilon,X}$. Let us denote the true structure as $S^* \equiv (h^*, F^*_{\varepsilon,X})$. Denote the set of functions $h$ that satisfy the restrictions that $h^*$ is assumed to

satisfy by $H$, and denote the set of distributions $F_{\varepsilon,X}$ that satisfy the restrictions that $F^*_{\varepsilon,X}$ is assumed to satisfy is denoted by $\Gamma$. Then, we can define a set of "admissible" structures as $\mathcal{S}_{H,\Gamma} \equiv H \times \Gamma$.

**Definition 5.2.2** *Let $F_{Y,X}$ be the distribution generated by a structure $S \in \mathcal{S}_{H,\Gamma}$. Then, we say that $S^*$ is identified in $\mathcal{S}_{H,\Gamma}$ if*

$$F_{Y,X} \neq F^*_{Y,X}$$

*for all $S \neq S^*$.*

Next, consider a **feature** (or simply a parameter) of a structure, $\theta \equiv \theta(S)$.

**Definition 5.2.3 (Point-identification)** *The feature $\theta^* \equiv \theta(S^*)$ is point-identified (or simply "identified") if for any $S \in \mathcal{S}_{H,\Gamma}$ that is observationally equivalent to $S^*$,*

$$\theta(S) = \theta(S^*).$$

The above definition says that when the observable variables contain sufficiently rich information to uniquely determine the value of a feature $\theta$ under $\mathcal{S}_{H,\Gamma}$, it is identifiable.

**Example 5.2.4 (Linear regression models)** Consider the following linear regression model:

$$Y = X^\top \beta + \varepsilon,$$

where $Y \in \mathbb{R}$, $X \in \mathbb{R}^k$, and $\varepsilon \in \mathbb{R}$. The structural equation is given by

$$h(Y, X, \varepsilon) = Y - X^\top \beta - \varepsilon.$$

The set of admissible $h(Y, X, \varepsilon)$'s can be formed simply by

$$H \equiv \{h : h(Y, X, \varepsilon) = Y - X^\top \beta - \varepsilon, \beta \in \mathbb{R}^k\}.$$

For an admissible set of distributions $F_{\varepsilon,X}$, for example, consider

$$\Gamma_1 \equiv \{F_{\varepsilon,X} : \mathbb{E}[\varepsilon \mid X] = 0\}.$$

Then, in the class of models defined by $H \times \Gamma_1$, we can have

$$\mathbb{E}[XY] = \mathbb{E}[XX^\top]\beta$$

by LIE. The above relationship is not sufficient to point-identify $\beta$ because the matrix $\mathbb{E}[XX^\top]$ may not be of full rank. That is, there are possibly infinite number of different $\beta$'s that satisfy the above equality.

Here, we add one more restriction on $\Gamma_1$. Namely, let

$$\Gamma_2 \equiv \{F_{\varepsilon,X} : \mathbb{E}[\varepsilon \mid X] = 0, \mathbb{E}[XX^\top] \text{ is nonsingular.}\}.$$

Then, for any $S_1, S^* \in H \times \Gamma_2$ that are observationally equivalent, we have

$$\beta(S_1) = \mathbb{E}_1[XX^\top]^{-1}\mathbb{E}_1[XY]$$
$$\beta(S^*) = \mathbb{E}^*[XX^\top]^{-1}\mathbb{E}^*[XY],$$

where $\mathbb{E}_1$ denotes the expectation under $F^1_{Y,X}$ generated by $S_1$, and $\mathbb{E}^*$ denotes the expectation under $F^*_{Y,X}$ generated by $S^*$. Since $S_1$ and $S^*$ are observationally equivalent, we have $F^1_{Y,X} = F^*_{Y,X}$, implying that $\mathbb{E}_1[XX^\top]^{-1} = \mathbb{E}^*[XX^\top]^{-1}$ and $\mathbb{E}_1[XY] = \mathbb{E}^*[XY]$ hold. Thus, $\beta(S_1) = \beta(S^*)$ for any $S_1$ that is observationally equivalent to $S^*$, i.e., $\beta^* \equiv \beta(S^*)$ is identified.

As shown in the above example, in general, a parameter of interest $\theta^*$ can be identified if we can find a mapping, say $\psi$, such that $\theta^* = \psi(\text{the moments of observed random variables})$ holds. In this case, we say that the identification is "constructive". Since any observationally equivalent distributions have the same moments of all order, the requirement in Definition 5.2.3 is met automatically. However, it is usually difficult to find such closed-form expressions for $\theta^*$ (of course, the lack of closed-form solution does not imply the failure of identification). There are many different, but closely related, notions of identification in the econometrics literature. For a recent comprehensive survey, see [Lewbel, 2019].

**Example 5.2.5 (Binary response models)** Consider the following binary response model:

$$Y = \mathbf{1}\{X^\top \beta \geq \varepsilon\},$$

where $X \in \mathbb{R}^k$, and $\varepsilon \in \mathbb{R}$. The structural equation is

$$h(Y, X, \varepsilon) = Y - \mathbf{1}\{X^\top \beta \geq \varepsilon\}.$$

For simplicity, the following assumption is maintained throughout this part: $\Pr(X^\top \beta = X^\top \beta^*) = 1$ if and only if $\beta = \beta^*$ (which is the full-rankness condition in this context). Further, let

$$\begin{aligned} H_1 &\equiv \{h : h(Y, X, \varepsilon) = Y - \mathbf{1}\{X^\top \beta \geq \varepsilon\}, \beta \in \mathbb{R}^k\} \\ \Gamma_1 &\equiv \{F_{\varepsilon,X} : \varepsilon \text{ is independent of } X,\ \varepsilon \sim N(0,1)\} \\ \Gamma_2 &\equiv \{F_{\varepsilon,X} : \varepsilon \text{ is independent of } X,\ \varepsilon \sim N(0,\sigma^2)\}. \end{aligned}$$

Note that clearly $\Gamma_1 \subset \Gamma_2$ holds by their definitions. As it turns out, $\beta$ can be identified under $H_1 \times \Gamma_1$ but not identified under $H_1 \times \Gamma_2$. To see this, first focus on the models in $H_1 \times \Gamma_2$. Then, for any observationally equivalent $S_1, S^* \in H_1 \times \Gamma_2$, we have

$$\begin{aligned} \mathbb{E}_1[Y \mid X] &= \Phi\left(\frac{X^\top \beta_1}{\sigma_1}\right) \\ \mathbb{E}^*[Y \mid X] &= \Phi\left(\frac{X^\top \beta^*}{\sigma^*}\right). \end{aligned}$$

Since $S$ and $S^*$ are observationally equivalent ($F^1_{Y,X} = F^*_{Y,X}$), we have $\mathbb{E}_1[Y \mid X] = \mathbb{E}^*[Y \mid X]$ almost surely (a.s.), which implies

$$\Phi\left(\frac{X^\top \beta_1}{\sigma_1}\right) = \Phi\left(\frac{X^\top \beta^*}{\sigma^*}\right) \text{ a.s.} \Longrightarrow \Pr\left(\frac{X^\top \beta_1}{\sigma_1} = \frac{X^\top \beta^*}{\sigma^*}\right) = 1 \Longrightarrow \frac{\beta_1}{\sigma_1} = \frac{\beta^*}{\sigma^*}, \tag{5.2.1}$$

where the first $\Longrightarrow$ is by the fact that $\Phi(\cdot)$ is strictly increasing and continuous. This tells us that the exact value of $\beta^*$ is not identified in $H_1 \times \Gamma_2$; in order to achieve point identification of $\beta^*$, we need to impose an additional restriction on the model which ensures $\sigma_1 = \sigma^*$. What we can identify within $H_1 \times \Gamma_2$ is only the relative scale of the elements of $\beta^*$. For example, letting $\beta^*_k$ be the $k$-th element of $\beta^*$, it holds by (5.2.1) that

$$(\beta_{1k}/\sigma_1)/(\beta_{1k'}/\sigma_1) = (\beta^*_k/\sigma^*)/(\beta^*_{k'}/\sigma^*) \iff \beta_{1k}/\beta_{1k'} = \beta^*_k/\beta^*_{k'}$$

for all observationally equivalent $S_1$. In this situation, we say that $\beta^*$ is identified "up to scale". The above result is intuitively understandable from the fact that for any constant $c > 0$ the following two models

$$Y = \mathbf{1}\{X^\top \beta \geq \varepsilon\} \text{ and } Y = \mathbf{1}\{X^\top (c\beta) \geq c\varepsilon\}$$

can generate exactly the same distributions for the observable data. Thus, for identification of $\beta^*$, we need to introduce some scale normalization restriction. A convenient way to do so is to assume that $\varepsilon$ is distributed as the standard normal (i.e., probit model). Then, for models in $H_1 \times \Gamma_1$, since both $\sigma_1$ and $\sigma^*$ are fixed at one, we have $\beta_1 = \beta^*$ for any $S_1$ observationally equivalent to $S^*$ from (5.2.1), implying that $\beta^*$ is identifiable.

Other possible identification restrictions are, for example, as follows:

$$H_2 \equiv \{h : h(Y, X, \varepsilon) = Y - \mathbf{1}\{X^\top \beta \geq \varepsilon\}, \beta_1 = 1, \beta_{-1} \in \mathbb{R}^{k-1}\}, \text{ where } \beta = (\beta_1, \beta_{-1}^\top)^\top,$$

(assuming that the first element of $\beta$ is known to be positive)

$$H_3 \equiv \{h : h(Y, X, \varepsilon) = Y - \mathbf{1}\{X^\top \beta \geq \varepsilon\}, \beta \in \mathbb{R}^k, ||\beta|| = 1\}.$$

Then, we can show that $(\beta_{-1}, \sigma^2)$ and $(\beta, \sigma^2)$ are identifiable in $H_2 \times \Gamma_2$ and $H_3 \times \Gamma_2$, respectively.

It would be worth noting that the normality assumption on $\varepsilon$ is in fact not necessary for identification. With some scale normalization, $\beta^*$ can be identified even when the distribution function of $\varepsilon$ is completely unknown (see, e.g., [Manski, 1975]).

**Exercise 5.2.6** Suppose that the true system $S^*$ belongs to $\mathcal{S}_{H,\Gamma} \equiv H \times \Gamma$, where

$$H \equiv \{h : h(Y, X, \varepsilon) = Y - \phi(X^\top \beta) - \varepsilon,\ \beta \in \mathbb{R}^{\dim(X)},\ \phi \text{ is a known and strictly increasing function}\}$$
$$\Gamma \equiv \{F_{\varepsilon,X} : \mathbb{E}[\varepsilon \mid X] = 0,\ \mathbb{E}[XX^\top] \text{ is nonsingular}\}$$

Prove that $\beta(S^*)$ is point-identified.

## 5.3 Partial identification

In the above discussion, we have considered under what conditions a parameter of interest $\theta^* \equiv \theta(S^*)$ can be identified as a unique value. However, in practice, it is often the case that such identification conditions are quite restrictive and not testable. Without these conditions, although the exact value of $\theta^*$ may not be identified, if we can still identify $\theta_L^*$ and $\theta_U^*$ such that

$$\theta_L^* \leq \theta^* \leq \theta_U^*,$$

we can infer the value of $\theta^*$ (for example, if $\theta_L^*$ is positive, we can say at least that $\theta^*$ is positive). In this situation, we say that $\theta^*$ is "partially" identified (or set-identified), and $[\theta_L^*, \theta_U^*]$ is called the identified interval, or more generally, the identified set. More formal definition is as follows:

**Definition 5.3.1 (Partial-identification)** *The feature $\theta^* \equiv \theta(S^*)$ is **partially identified** if for any $S \in \mathcal{S}_{H,\Gamma}$ that is observationally equivalent to $S^*$,*

$$\theta(S), \theta(S^*) \in \Theta^*.$$

*The set* $\Theta^*$ *is called the* ***identified set****.*

The identified set $\Theta^*$ is said to be "informative" if $\Theta^*$ is a bounded set, and is "uninformative" if not. Note that the identified set is not unique; for an extreme example, $\mathbb{R}^{\dim(\theta)}$ is always an (uninformative) identified set for $\theta^*$. In the literature on partial identification, we are usually interested in finding the smallest (i.e., the most informative) identified set.

**Example 5.3.2 (Missing data)** Let $Y \in \{0,1\}$ be a dummy response variable of interest. Suppose that $Y$ is not observable for some individuals for some reason, and let $D \in \{0,1\}$ be a dummy variable indicating the observability of $Y$:

$$\begin{cases} Y \text{ is observed} & \text{if } D = 1 \\ Y \text{ is unobserved} & \text{if } D = 0. \end{cases}$$

This situation is quite common when the data are collected from a questionnaire survey. Suppose that we would like to know the population ratio of $Y = 1$, $\mathbb{E}[Y]$. Then, a common practice is to estimate it by estimating instead $\mathbb{E}[Y \mid D = 1]$ using only observable data subset. However, in order for the equality $\mathbb{E}[Y] = \mathbb{E}[Y \mid D = 1]$ to hold, we need a strong condition, such as the independence between $Y$ and $D$, the so-called **missing at random** assumption.

Even when $Y$ and $D$ are dependent, in general, we can construct an informative identified interval for $\mathbb{E}[Y]$. First, observe that

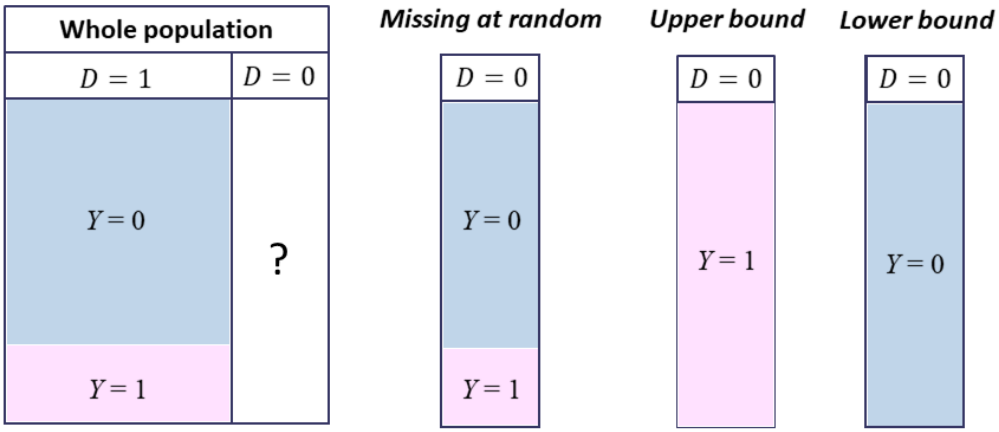
$$\mathbb{E}[Y] = \mathbb{E}[Y \mid D = 1]\Pr(D = 1) + \mathbb{E}[Y \mid D = 0]\Pr(D = 0).$$

For the terms on the right-hand side, only $\mathbb{E}[Y \mid D = 0]$ is an unknown component. Note that we must have $0 \le \mathbb{E}[Y \mid D = 0] \le 1$ since $Y$ is a dummy variable. Thus, the following inequalities must hold:

$$\mathbb{E}[Y \mid D = 1]\Pr(D = 1) \le \mathbb{E}[Y] \le \mathbb{E}[Y \mid D = 1]\Pr(D = 1) + \Pr(D = 0).$$

Hence, the identified interval for $\mathbb{E}[Y]$ is given by

$$\Theta'_{\mathbb{E}[Y]} \equiv \Big[\mathbb{E}[Y \mid D = 1]\Pr(D = 1),\ \mathbb{E}[Y \mid D = 1]\Pr(D = 1) + \Pr(D = 0)\Big].$$



In order to obtain a more informative identified interval than $\Theta'_{\mathbb{E}[Y]}$, we need to add more assumptions on $\mathbb{E}[Y \mid D = 0]$. For example, suppose that we know that $\mathbb{E}[Y \mid D = 1] \leq \mathbb{E}[Y \mid D = 0]$.[1] Then, the resulting identified interval for $\mathbb{E}[Y]$ is

$$\Theta''_{\mathbb{E}[Y]} \equiv \Big[\mathbb{E}[Y \mid D = 1],\ \mathbb{E}[Y \mid D = 1]\Pr(D = 1) + \Pr(D = 0)\Big].$$

Clearly, $\Theta''_{\mathbb{E}[Y]} \subseteq \Theta'_{\mathbb{E}[Y]}$.

**Example 5.3.3 (Regression with interval-valued dependent variables)** It is fairly common that the exact value of variables such as wage or household wealth is not directly asked by surveys; instead, they usually ask the same question in the form of intervals. For example, let $Y$ be a wage variable, whose exact level is unknown, but we know that

$$Y \in [Y_L, Y_L + \Delta],$$

where both the value of $Y_L$ and that of $\Delta$ are known, $-\infty < Y_L < \infty$, and $0 < \Delta < \infty$. Suppose that we would like to estimate the impact of an explanatory variable $X$ on $Y$ using the following simple regression model:

$$Y = \beta_0 + \beta_1 X + \varepsilon, \quad \mathbb{E}[\varepsilon \mid X] = 0.$$

For simplicity, we assume that $\mathbb{E}[X] = 0$. Then, if $Y$ were observed, we can identify $\beta_1$ by

$$\beta_1 = \frac{\mathbb{E}[XY]}{\mathbb{E}[X^2]},$$

assuming that $\mathbb{E}[X^2] > 0$. When $Y$ is unknown, researchers often try to estimate $\beta_1$ by the sample analog of $\mathbb{E}[X(Y_L + \Delta/2)]/\mathbb{E}[X^2]$. However, this approach does not have a theoretical justification.

A more conservative and safer approach is to use partial identification. Note that there exists a latent random variable $t \in [0, 1]$ such that

$$Y = Y_L + t\Delta.$$

Thus, we can write

$$\beta_1 = \frac{\mathbb{E}[XY_L]}{\mathbb{E}[X^2]} + \frac{\mathbb{E}[X \cdot t\Delta]}{\mathbb{E}[X^2]}.$$

Observe that

$$\begin{aligned}
\mathbb{E}[X \cdot t\Delta] &= \mathbb{E}[X \cdot t\Delta \mathbf{1}\{X > 0\}] + \mathbb{E}[X \cdot t\Delta \mathbf{1}\{X < 0\}] \\
&\leq \mathbb{E}[X \cdot 1\Delta \mathbf{1}\{X > 0\}] + \mathbb{E}[X \cdot 0\Delta \mathbf{1}\{X < 0\}] \\
&= \mathbb{E}[X \cdot \Delta \mathbf{1}\{X > 0\}].
\end{aligned}$$

This implies that the upper bound of $\beta_1$ is obtained by

$$\beta_1 \leq \frac{\mathbb{E}[XY_L]}{\mathbb{E}[X^2]} + \frac{\mathbb{E}[X \cdot \Delta \mathbf{1}\{X > 0\}]}{\mathbb{E}[X^2]}.$$

Similarly, we can show that $\mathbb{E}[X \cdot t\Delta] \geq \mathbb{E}[X \cdot \Delta \mathbf{1}\{X < 0\}]$, and thus the lower bound of $\beta_1$ is

$$\beta_1 \geq \frac{\mathbb{E}[XY_L]}{\mathbb{E}[X^2]} + \frac{\mathbb{E}[X \cdot \Delta \mathbf{1}\{X < 0\}]}{\mathbb{E}[X^2]}.$$

[1] For example, let $Y$ be the yes/no answer to a question about the experience of illegal drug use. Then, it would be somewhat likely that the respondents who did not answer this question have more experience of drug use ($Y$ = yes) than those who did answer the question. Thus, it should be reasonable to assume that $\Pr(Y = \text{yes} \mid D = 1) \leq \Pr(Y = \text{yes} \mid D = 0)$ in this example.

Consequently, we can obtain the identified interval for $\beta_1$ as

$$\Theta_{\beta_1} \equiv \left[ \frac{\mathbb{E}[XY_L]}{\mathbb{E}[X^2]} + \frac{\mathbb{E}[X \cdot \Delta \mathbf{1}\{X < 0\}]}{\mathbb{E}[X^2]}, \ \frac{\mathbb{E}[XY_L]}{\mathbb{E}[X^2]} + \frac{\mathbb{E}[X \cdot \Delta \mathbf{1}\{X > 0\}]}{\mathbb{E}[X^2]} \right].$$

From this, we can observe that the length of the identified interval is proportional to $\Delta$. For more general discussion on this topic, see [Bontemps et al., 2012]

**Exercise 5.3.4** Consider a simple regression model:

$$Y = \beta_0 + \beta_1 X + \varepsilon, \ \ \mathbb{E}\varepsilon = 0,$$

where the regressor $X$ is suspected to be endogenous: $\mathrm{Cov}(X, \varepsilon) \neq 0$. Suppose that we have an "imperfect" instrument $Z \in \mathbb{R}$ for $X$ such that $\mathrm{Cov}(X, Z) > 0$ and $\underline{c} \leq \mathrm{Cov}(Z, \varepsilon) \leq \overline{c}$, where $\underline{c}$ and $\overline{c}$ are known constants. Derive an informative identified interval for $\beta_1$.

**Exercise 5.3.5** Imagine you are a seller at an online store like Amazon. You have sold out your products to 300 people at the site. Among these 300 buyers, 100 of them posted the product's review with a star rating. The distribution of the star-ratings is as given in Table 5.1. Suppose that you would like to know the value of $\mu \equiv \mathbb{E}[\text{the number of } \star\text{'s}|\text{buyers}]$.

1. If you assume that the missing-at-random (MAR) holds, what is the estimate of $\mu$?
2. Do you think the MAR assumption is credible in this context? Why?
3. When imposing no assumptions on the non-respondents, we can only partially identify $\mu$. (a) Derive the identified set, and (b) report the computed identified set.
4. Suppose you would like to obtain a smaller identified set than the one obtained in 3. Explain specifically (a) what kind of additional assumption(s) you would introduce and (b) to what extent the resulting identified set is smaller than the one in 3.

Table 5.1: Distribution of star-ratings

| | |
|---|---|
| ★ ★ ★ ★ ★ | 20 |
| ★ ★ ★★ | 32 |
| ★ ★ ★ | 24 |
| ★★ | 9 |
| ★ | 15 |
| Total | 100 |

# Chapter 6

# Structural Estimation

One of the most important tasks in econometrics is to empirically simulate the impact of new policies. However, there is a huge difficulty to quantify the policy impact if the policy is purely hypothetical or there are no similar policies implemented in the past. In such situations, pure causal inference methods are infeasible simply because we cannot observe data "after" the policy.

One possible approach to circumvent this problem is to use a **structural estimation** technique. Structural estimation is a method to construct econometric models explicitly based on economic theory and estimate the parameters of functions (such as utility function, supply function, etc.) that govern the behavior of individuals and firms. Once these functions are recovered, even when we do not have any data for the policy, we can predict how individuals and firms behave in the hypothetical situation using some economic theory as a guideline.[1]

Note that the structural estimation does not refer to a particular estimation method or a type of econometric models. Rather, it is a way of thinking about econometric models. As shown below, even a simple OLS regression can be viewed as a structural estimation approach under certain microeconomic assumptions. Then, the validity of a structural econometric model and its prediction is reduced to the validity of the economic assumptions on which the model is based.

## 6.1 OLS estimation of production functions

Consider a problem of estimating the agricultural production function of farmers. For simplicity, suppose that each farmer has only two inputs, labor $L$ (the number of workers) and capital $K$ (the size of farm land). The amount of agricultural output is denoted by $P$. Then, simply running a linear regression of $P$ on $(L, K)$ is not recognized as a "structural" estimation, unless there is a particular reason to believe that the production function is a linear function. Thus, the obtained regression coefficients cannot be viewed as the estimates of the deep

[1] The idea of structural estimation is closely related to the well-known **Lucas critique** of macroeconomic forecasting:

> *Given that the structure of an econometric model consists of optimal decision-rules of economic agents, and that optimal decision-rules vary systematically with changes in the structure of series relevant to the decision maker, it follows that any change in policy will systematically alter the structure of econometric models.* a quote from [Lucas, 1976].

His statement can be interpreted as that (atheoretical) "reduced-form" econometric models should not be used for the purpose of predicting the effect of a new policy. Rather, we should model the "deep" (i.e., policy invariant) parameters that characterize individual behavior based on microeconomic foundations.

parameters.

A more reasonable (i.e., more justifiable in terms of economic theory) functional form of the production function would be a Cobb–Douglas production function:

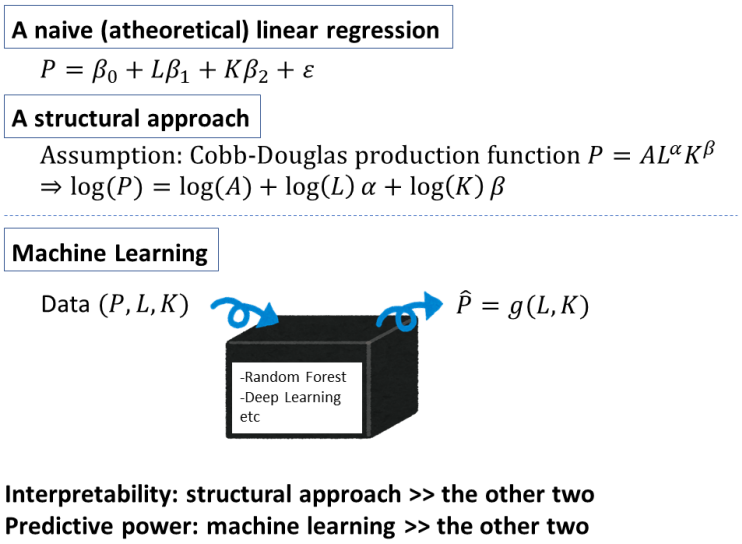
$$P = AL^{\alpha}K^{\beta},$$

where $A$ denotes an idiosyncratic productivity disturbance. The parameters $\alpha$ and $\beta$ are the output elasticities of labor and capital, respectively. Taking the log of both sides yields the following linear regression model:

$$\log(P) = c + \log(L)\alpha + \log(K)\beta + \varepsilon,$$

where $c = \mathbb{E}[\log(A)]$, and $\varepsilon = \log(A) - c$. Hence, by simply running an OLS regression of $\log(P)$ on $(\log(L), \log(K))$, we can estimate the elasticity parameters.

It should be important to note that, although atheoretical (non-structural) estimation approaches cannot be used to simulate counterfactual policy changes, they can be used for the purpose of predicting the other farmers' output level $P$ under the same market condition. If one's research interest is of the latter type, it is not necessary to consider a structural estimation; rather, it has been shown that using modern **machine learning** techniques is superior to the classical econometric methods for this purpose (see, e.g., [Bajari et al., 2015]).




## 6.2 Discrete choice models: the random utility framework

Discrete choice models, such as the logit model, can be viewed as structural econometric models based on the revealed preference approach. Suppose that there are two goods, say Good 1 and Good 2. Let $D_i$ be a dummy outcome variable that takes one if consumer $i$ chooses Good 1 and zero if Good 2 is chosen. Following the revealed preference approach, we can assume that $D_i = 1$ if and only if $U_{i1} > U_{i2}$, where $U_{i1}$ and $U_{i2}$ are the utilities obtained from Good 1 and Good 2, respectively.

For each individual $i$, let $X_i$ be a vector of "observable" individual characteristics, and $\varepsilon_{ji}$ be an "unobservable" random variable that may be specific to Good $j$.[2] Then, without loss of generality, we can write $i$'s utility of

[2]Note that $\varepsilon_{ji}$ is unobservable only to researchers, but of course $i$ knows the value of her own $\epsilon$.

choosing Good $j$ as $U_{ij} = U_{ij}(X_i, \varepsilon_{ji})$.

Here, unlike the traditional economic model of consumer demand, it is allowed that we cannot fully observe the variables that determine the utility. Namely, the utility can depend on an unobservable random factor $\varepsilon$. This framework is called **random utility model** (RUM).[3] To facilitate the discussion, suppose that the utility function takes the following form:

$$U_{ij}(X_i, \varepsilon_{ji}) \equiv V_j(X_i) + \varepsilon_{ji}, \quad \text{for } j = 1, 2.$$

Then, we have

$$\begin{aligned} D_i = 1 &\iff U_{i1}(X_i, \varepsilon_{1i}) > U_{i2}(X_i, \varepsilon_{2i}) \\ &\iff V_1(X_i) + \varepsilon_{1i} > V_2(X_i) + \varepsilon_{2i} \\ &\iff V_1(X_i) - V_2(X_i) > \varepsilon_{2i} - \varepsilon_{1i}. \end{aligned}$$

Now, we assume that the error terms $\varepsilon_{ji}$'s are independent of $X_i$ and are IID as **Type-1 Extreme Value** (also known as **Gumbel** distribution). It is known that the difference of two independent Type-1 Extreme Value variables has the standard logistic distribution. Therefore, we obtain

$$\Pr(D_i = 1 \mid X_i) = \frac{\exp[V_1(X_i) - V_2(X_i)]}{1 + \exp[V_1(X_i) - V_2(X_i)]}.$$

In applications, we often assume a linear functional form for $V_1(X_i) - V_2(X_i)$ such that

$$V_1(X_i) - V_2(X_i) = X_i^\top \beta_0.$$

for some $\beta_0$. From this perspective, we can see that the regression coefficient $\beta_0$ in the binary response models corresponds to the difference of the marginal utilities; that is,

$$\begin{aligned} \beta_0 &= \frac{\partial V_1(X_i)}{\partial X} - \frac{\partial V_2(X_i)}{\partial X} \\ &= \frac{\partial U_{i1}(X_i, \varepsilon_{1i})}{\partial X} - \frac{\partial U_{i2}(X_i, \varepsilon_{2i})}{\partial X} \end{aligned}$$

under the assumptions made here.

## 6.3 Estimation of entry games

In the last few decades, there has been a growing interest in the econometric analysis of game theoretic models (see, e.g., [De Paula, 2013] for a survey on this topic). An important early example is the estimation of **entry games** (e.g., [Bresnahan and Reiss, 1990, Berry, 1992]). Suppose that there are two players competing in an entry game across $n$ different markets. The players are labeled by $j = 1, 2$. In applications, these labels refer to the actual name of the firms (e.g., JAL and ANA; Walmart and Kmart; IKEA and Nitori). A player's entry decision depends on its profit, which in turn depends on whether the rival also enters the market or not.

Let $D_{ij}$ denote whether player $j$ enters the market $i$ ($D_{ij} = 1$) or not ($D_{ij} = 0$), where $i = 1, \ldots, n$. For each player $j$, we assume that the payoff from entering the market $i$ is given by

$$u_{ij}(d_{-j}) \equiv \rho d_{-j} + X_{ij}^\top \beta - \varepsilon_{ij},$$

[3] Based on the random utility framework, Daniel McFadden developed a set of econometric methods for analyzing discrete choice behavior. He was awarded the Nobel Prize in economics for this work.

where $d_{-j} \in \{0,1\}$ denotes the action chosen by $j$'s opponent, $X_{ij}$ and $\varepsilon_{ij}$ denote observable and unobservable (to econometricians) payoff covariates, and $\rho$ and $\beta$ are the parameters of interest. We assume that $(\varepsilon_{i1}, \varepsilon_{i2})$ are independent of $(X_{i1}, X_{i2})$ for all $i$. The parameter $\rho$ captures the **strategic interaction effect**, in which if it is negative (resp. positive) the model exhibits strategic substitutes (resp. complements). Note that the parameters $(\rho, \beta)$ can be heterogeneous among the players in general, but we assume that they are homogeneous to simplify the discussion. Then, the payoff matrix of the game in market $i$ is given by

| | $D_{i2} = 0$ | $D_{i2} = 1$ |
|---|---|---|
| $D_{i1} = 0$ | $(0, 0)$ | $(0, X_{i2}^\top \beta - \varepsilon_{i2})$ |
| $D_{i1} = 1$ | $(X_{i1}^\top \beta - \varepsilon_{i1}, 0)$ | $(\rho + X_{i1}^\top \beta - \varepsilon_{i1}, \rho + X_{i2}^\top \beta - \varepsilon_{i2})$ |

Hereinafter, we omit the market index $i$ for simplicity when there is no confusion. Now, assume that the realizations of $(X_1, X_2)$ and $(\varepsilon_1, \varepsilon_2)$ are known to both players in each market, i.e., we assume a **complete information game**. Further, we adopt the pure strategy Nash equilibrium as the solution concept of this game. In addition, suppose that the game is a simultaneous-move game. Then, we can see that the players' entry decisions follow the following simultaneous binary choice model:

$$\begin{aligned} D_1 &= \mathbf{1}\{\rho D_2 + X_1^\top \beta \geq \varepsilon_1\} \\ D_2 &= \mathbf{1}\{\rho D_1 + X_2^\top \beta \geq \varepsilon_2\}. \end{aligned} \tag{6.3.1}$$

Then, we have the following relationship between the realized outcomes $(D_1, D_2)$ and the error terms $(\varepsilon_1, \varepsilon_2)$:

$$\begin{aligned} (D_1, D_2) = (1,1) &\implies \varepsilon_1 \leq \rho + X_1^\top \beta,\ \varepsilon_2 \leq \rho + X_2^\top \beta, \\ (D_1, D_2) = (1,0) &\implies \varepsilon_1 \leq X_1^\top \beta,\ \quad \varepsilon_2 > \rho + X_2^\top \beta, \\ (D_1, D_2) = (0,1) &\implies \varepsilon_1 > \rho + X_1^\top \beta,\ \varepsilon_2 \leq X_2^\top \beta, \\ (D_1, D_2) = (0,0) &\implies \varepsilon_1 > X_1^\top \beta,\ \quad \varepsilon_2 > X_2^\top \beta. \end{aligned} \tag{6.3.2}$$

As conventionally assumed in the literature on entry games, we assume strategic substitutes, i.e., $\rho < 0$. Then, the relationship in (6.3.2) can be visually summarized in Figure 6.1. As shown in the figure, the space of $(\varepsilon_1, \varepsilon_2)$ cannot be partitioned into non-overlapping regions associated with the four alternative realizations of $(D_1, D_2)$. Both $(D_1, D_2) = (0,1)$ and $(D_1, D_2) = (1,0)$ can occur when $(\varepsilon_1, \varepsilon_2)$ fall into the shaded area:

$$\begin{aligned} \rho + X_1^\top \beta < \varepsilon_1 \leq X_1^\top \beta \\ \rho + X_2^\top \beta < \varepsilon_2 \leq X_2^\top \beta \end{aligned} \implies (D_1, D_2) = (0,1) \text{ or } (1,0)$$

That is, **multiple Nash equilibria** exist in this region.

This non-uniqueness of model-consistent decisions is called **incompleteness** and has been extensively studied in the literature on simultaneous equations models for discrete outcomes (see, e.g., [Tamer, 2003]). When the model is incomplete, the standard maximum likelihood estimation is infeasible because the likelihood function is not well-defined.[4] In what follows, we discuss two approaches to overcome this problem.

### 6.3.1 Rewriting the model in terms of the number of entrants

Notice that even though there are multiple equilibria in the entry decisions, "the number of entrants" can be uniquely determined in our model under $\rho < 0$. Namely, let $\mathbf{N} \in \{0,1,2\}$ be the number of players choosing 1

[4] If we simply construct the likelihood function based on the relationship (6.3.2), then it sums up to a value larger than one.

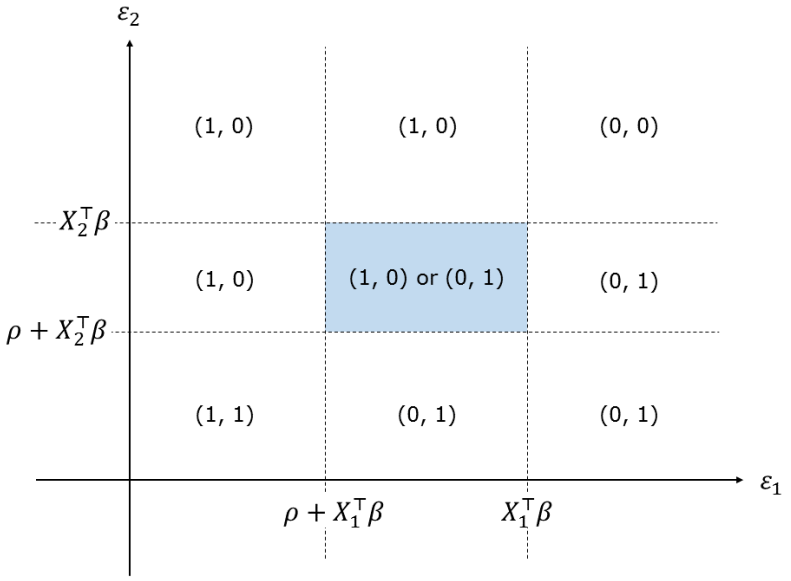



Figure 6.1: Nash equilibrium

and $F_j(\cdot)$ and $F(\cdot,\cdot)$ be the marginal and joint distribution function of $\varepsilon_j$'s, respectively. Then, we have the conditional likelihood function for each action as follows

$$\begin{aligned}
\text{No entrants} \quad & \Pr(\mathbf{N}=0 \mid X_1, X_2) = 1 - F_1(X_1^\top\beta) - F_2(X_2^\top\beta) + F(X_1^\top\beta, X_2^\top\beta) \\
\text{Duopoly} \quad & \Pr(\mathbf{N}=2 \mid X_1, X_2) = F(\rho + X_1^\top\beta, \rho + X_2^\top\beta) \\
\text{Monopoly} \quad & \Pr(\mathbf{N}=1 \mid X_1, X_2) = 1 - \Pr(\mathbf{N}=0 \mid X_1, X_2) - \Pr(\mathbf{N}=2 \mid X_1, X_2).
\end{aligned}$$

Thus, the model can be estimated by solving the following maximum likelihood problem:

$$\max_{(\rho,\beta,\ldots)} \sum_{i=1}^{n} \sum_{k\in\{0,1,2\}} \mathbf{1}\{\mathbf{N}_i = k\} \log \Pr(\mathbf{N}_i = k \mid X_{i1}, X_{i2}).$$

Here, the set of estimation parameters may include not only $\rho$ and $\beta$ but also some additional parameters, such as a correlation parameter between $\varepsilon_1$ and $\varepsilon_2$. This approach is pursued in [Bresnahan and Reiss, 1990] and [Berry, 1992], which highlights that the presence of multiple equilibria does not necessarily imply the lack of point identification.

### 6.3.2 Stochastic equilibrium selection rule

An alternative approach is to explicitly introduce a stochastic (or possibly deterministic) equilibrium selection mechanism. Unlike the above approach, this approach allows us to identify the entry behavior in the region of multiplicity. The simplest form of this approach is to assume a pure random equilibrium selection. More specifically, suppose that the players would choose $(D_1, D_2) = (1, 0)$ with probability $\lambda$ when in the multiple equilibria region. Then, the likelihood function for each realization can be well-defined as follows:

$$\Pr(0,0 \mid X_1, X_2) = 1 - F_1(X_1^\top\beta) - F_2(X_2^\top\beta) + F(X_1^\top\beta, X_2^\top\beta)$$

$$\begin{aligned}
\Pr(1,1 \mid X_1, X_2) &= F(\rho + X_1^\top \beta, \rho + X_2^\top \beta) \\
\Pr(0,1 \mid X_1, X_2) &= F_2(X_2^\top \beta) - F(\rho + X_1^\top \beta, X_2^\top \beta) - \lambda P_{mul}(X_1, X_2) \\
\Pr(1,0 \mid X_1, X_2) &= 1 - \sum_{(d_1,d_2)\in\{(0,0),(1,1),(0,1)\}} \Pr(d_1, d_2 \mid X_1, X_2),
\end{aligned}$$

where

$$P_{mul}(X_1, X_2) \equiv F(X_1^\top \beta, X_2^\top \beta) - F(\rho + X_1^\top \beta, X_2^\top \beta) - F(X_1^\top \beta, \rho + X_2^\top \beta) + F(\rho + X_1^\top \beta, \rho + X_2^\top \beta),$$

which is the probability that $(\varepsilon_1, \varepsilon_2)$ reside in the multiple equilibria region for given $(X_1, X_2)$. Then, we can estimate the model by solving the following maximum likelihood problem:

$$\max_{(\rho,\beta,\lambda,\ldots)} \sum_{i=1}^{n} \sum_{(d_1,d_2)\in\{0,1\}^2} \mathbf{1}\{(D_{i1}, D_{i2}) = (d_1, d_2)\} \log \Pr(d_1, d_2 | X_{i1}, X_{i2}).$$

### 6.3.3 When the sign of $\rho$ is unknown

So far, we have assumed that the sign of the strategic interaction effect is known a priori. Although such assumption may be justified by economic theory in particular situations, in general, we do not have prior knowledge about $\rho$. When the the sign of $\rho$ is unknown, it is difficult to obtain point identification of the model.

To tackle this problem, [Ciliberto and Tamer, 2009] proposed a partial identification approach. For each realization $(D_1, D_2) = (d_1, d_2)$, we can generally write

$$\begin{aligned}
\Pr(d_1, d_2 \mid X_1, X_2) &= \int \Pr(d_1, d_2 \mid X_1, X_2, \varepsilon_1, \varepsilon_2) \mathrm{d}F \\
&= \int_{\mathcal{R}_{uni}^{(d_1,d_2)}(X_1,X_2)} \Pr(d_1, d_2 \mid X_1, X_2, \varepsilon_1, \varepsilon_2) \mathrm{d}F + \int_{\mathcal{R}_{mul}^{(d_1,d_2)}(X_1,X_2)} \Pr(d_1, d_2 \mid X_1, X_2, \varepsilon_1, \varepsilon_2) \mathrm{d}F \\
&= \int_{\mathcal{R}_{uni}^{(d_1,d_2)}(X_1,X_2)} \mathrm{d}F + \int_{\mathcal{R}_{mul}^{(d_1,d_2)}(X_1,X_2)} \Pr(d_1, d_2 \mid X_1, X_2, \varepsilon_1, \varepsilon_2) \mathrm{d}F.
\end{aligned}$$

Here, $\mathcal{R}_{uni}^{(d_1,d_2)}(X_1, X_2)$ denotes the region of $(\varepsilon_1, \varepsilon_2)$ where $(D_1, D_2) = (d_1, d_2)$ uniquely occurs for given $(X_1, X_2)$, and $\mathcal{R}_{mul}^{(d_1,d_2)}(X_1, X_2)$ is the multiple equilibria region. The probability function $\Pr(d_1, d_2 \mid X_1, X_2, \varepsilon_1, \varepsilon_2)$ plays the role of equilibrium selection function for $(d_1, d_2)$ in the multiplicity region.[5] Whatever the functional form of $\Pr(d_1, d_2 \mid X_1, X_2, \varepsilon_1, \varepsilon_2)$ is, since $0 \leq \Pr(d_1, d_2 \mid X_1, X_2, \varepsilon_1, \varepsilon_2) \leq 1$ must hold, we have

$$\int_{\mathcal{R}_{uni}^{(d_1,d_2)}(X_1,X_2)} \mathrm{d}F \leq \Pr(d_1, d_2 \mid X_1, X_2) \leq \int_{\mathcal{R}_{uni}^{(d_1,d_2)}(X_1,X_2)} \mathrm{d}F + \int_{\mathcal{R}_{mul}^{(d_1,d_2)}(X_1,X_2)} \mathrm{d}F. \tag{6.3.3}$$

Thus, the identified set of the model parameters can be defined by the set of parameters that satisfy (6.3.3). For practical implementation of this approach, see [Ciliberto and Tamer, 2009].

## 6.4 Estimation of first-price auction models

To be added in a future update.

## 6.5 BLP Demand Estimation

To be added in a future update.

[5] If we can impose some parametric functional form assumption on $\Pr(d_1, d_2 \mid X_1, X_2, \varepsilon_1, \varepsilon_2)$ for $(\varepsilon_1, \varepsilon_2) \in \mathcal{R}_{mul}^{(d_1,d_2)}(X_1, X_2)$, then we may point identify the model. Such an approach was taken in [Bajari et al., 2010].

# Chapter 7

# Bootstrap

The bootstrap is a method to estimate the distribution of an estimator or that of a test statistic by repeatedly resampling the original data. The bootstrap was introduced by [Efron, 1979] and has continued to be one of the central interests of statistical research in all areas. The bootstrap has several advantages over the classical inference methods that are based on the asymptotic distributions. One is its algorithmic simplicity. It is often the case that in complicated econometric models, the statistic of interest has a very complicated form of asymptotic variance, which makes its estimation cumbersome. If we use a bootstrap method instead, we may circumvent such complicated computation. A more important advantage of using bootstrap is that under some regularity conditions, the approximated distribution obtained by the bootstrap is at least as accurate as and often more accurate than the approximation obtained in the first-order asymptotic theory. That is, the bootstrap is not just a convenient alternative to the conventional asymptotic method, but it sometimes exhibits theoretically desirable properties compared to the conventional method.

The contents of this chapter are mostly based on [Horowitz, 2001] and [Horowitz, 2019]. Detailed mathematical proofs will be omitted.[1]

## 7.1 The basic idea of the bootstrap method

Suppose that we have an IID sample $\{X_1, \ldots, X_n\}$ of size $n$ obtained from an unknown distribution $F_0$. Let $T_n \equiv T_n(X_1, \ldots, X_n)$ be the statistic of which we would like to know the distribution. A conventional approach is to derive its asymptotic distribution as $n \to \infty$. For example, let $T_n$ be the $t$-statistic for testing the null hypothesis $\mathbb{H}_0 : \mathbb{E}X = 0$; namely,

$$T_n = \frac{\overline{X}_n}{s_n},$$

where $\overline{X}_n$ amd $s_n$ are the sample mean and sample standard deviation, respectively. Then, by CLT, we know that $T_n$ is distributed as the standard normal $N(0,1)$ as $n \to \infty$. This approximation is actually very accurate if the sample size is not small. However, for more involved test statistics in complicated models, the asymptotic approximation can be inaccurate for moderate sample sizes.

[1] For those who are interested in the theory of bootstrap, I suggest take a look at [Hall, 1992] first, among many others.

Now, consider a general statistic $T_n$ and let $G_n(\cdot \mid F)$ be the distribution function of $T_n$ when the data are drawn from the distribution function $F$:

$$\begin{aligned} G_n(t \mid F) &\equiv P_F(T_n(X_1, \ldots, X_n) \le t) \\ &= \int \mathbf{1}\{T_n(x_1, \ldots, x_n) \le t\} \mathrm{d}F^n. \end{aligned}$$

The true distribution of $T_n$ under $F_0$ can be written as $G_n(\cdot \mid F_0)$. Thus, if we happen to know the exact form of $G_n(\cdot \mid F_0)$, we can conduct any statistical inference on $T_n$; however, this is a rare case since $F_0$ is usually unknown.

For data $\{X_1, \ldots, X_n\}$ that are drawn from $F_0$, we have the empirical distribution function $F_n$ defined as

$$F_n(t) \equiv \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{X_i \le t\}.$$

Using this empirical distribution, we can compute $G_n(\cdot \mid F_n)$, instead of $G_n(\cdot \mid F_0)$, in the following manner:

**Step 1.** Generate a bootstrap sample $\{X_1^*, \ldots, X_n^*\}$ from $F_n$.

**Step 2.** Compute the test statistic $T_n^* = T_n(X_1^*, \ldots, X_n^*)$.

**Step 3.** Repeat Steps 1 and 2 many times, and compute the empirical distribution of $T_n^*$.

Step 1 can be simply performed by a simple random sampling from $\{X_1, \ldots, X_n\}$ with replacement. By increasing the number of repetitions in Step 3, we can approximate $G_n(\cdot \mid F_n)$ arbitrarily precisely.[2] The above procedure is called the **nonparametric bootstrap**.

By a uniform LLN, we have $|F_n(z) - F_0(z)| \to 0$ over all $z$ with probability one. (Formally, this result is known as the **Glivenko-Cantelli theorem**.) From this result, for sufficiently large $n$, we can anticipate that $G_n(\cdot \mid F_n) \approx G_n(\cdot \mid F_0)$ in some sense if $G_n$ is continuous in $F$. This is the consistency property of bootstrap. A more formal definition is as follows.

Taking the limit $n \to \infty$ of $G_n(\cdot \mid F_0)$, we define the asymptotic true distribution of $T_n$ as $G_\infty(\cdot \mid F_0)$.

**Definition 7.1.1 (Bootstrap consistency)** *The bootstrap distribution estimator $G_n(\cdot \mid F_n)$ is said to be **consistent** if, for any $\kappa > 0$,*

$$\Pr\left( |G_n(t \mid F_n) - G_\infty(t \mid F_0)| > \kappa \text{ for all } t \in \mathbb{R} \right) \to 0.$$

Note that in the above definition, the bootstrap distribution $G_n(t \mid F_n) = P_{F_n}(T_n(X_1^*, \ldots, X_n^*) \le t)$ is a random CDF because $F_n$ depends on the original sampling $\{X_1, \ldots, X_n\}$. Thus, the probability is taken over this randomness. $G_\infty(\cdot \mid F_0)$ is not a stochastic function. Once a consistent bootstrap distribution $G_n(\cdot \mid F_n)$ is obtained, we can perform statistical inference based on the bootstrap critical values and the confidence intervals directly obtained from $G_n(\cdot \mid F_n)$.

It is difficult to provide general and easy-to-check conditions for a bootstrap estimator to be consistent. In many practical situations, it is often the case that the test statistic of interest can be expressed as $T_n = (\overline{g}_n - t_n)/s_n$

[2] Note that, theoretically, the exact form of $G_n(t \mid F_n)$ can be obtained by calculating the proportion of all possible bootstrap samples that have $\{T_n^* \le t\}$. However, this would require $n^n$ calculations, which is not possible in practice except when $n$ is very small. Thus, in Step 3, we basically generate "as many bootstrap samples as possible".

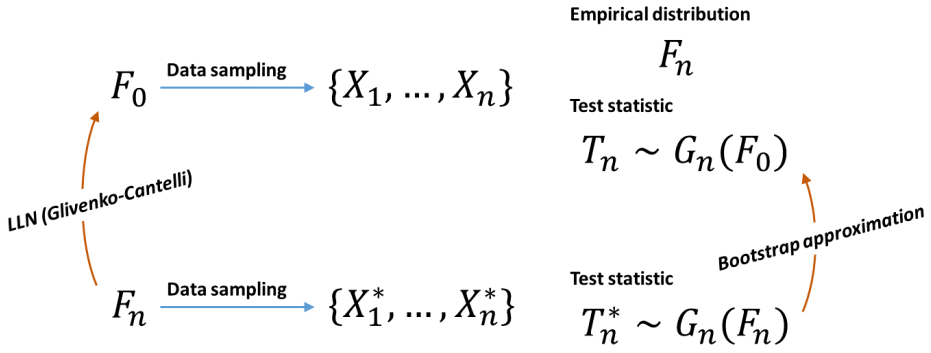



Figure 7.1: The idea of bootstrap

for some sequences $t_n$ and $s_n$, where $\overline{g}_n = n^{-1}\sum_{i=1}^n g(X_i)$ with an IID sample $\{X_1, \ldots, X_n\}$. For this special case, [Mammen, 1992] shows that the nonparametric bootstrap distribution of $T_n^* = (\overline{g}_n^* - \overline{g}_n)/s_n$, where $\overline{g}_n^* = n^{-1}\sum_{i=1}^n g(X_i^*)$, is consistent for the distribution of $T_n$ if and only if $T_n \xrightarrow{d} N(0,1)$. Many of commonly used estimators and test statistics, such as least squares estimators, ML estimators, $t$-statistic, etc, have this form and are asymptotically normal. Thus, the nonparametric bootstrap is consistent for most applications. Nonetheless, it is important to know when the bootstrap may fail.

One such example is the estimation of the distribution of the maximum or minimum value of a sample. Let $\{X_1, \ldots, X_n\}$ be an IID sample drawn from Uniform$[0, \alpha_0]$. As we have seen in Exercise 3.1.5, the ML estimator of $\alpha_0$ is given by

$$\widehat{\alpha}_n^{mle} = \max\{X_1, \ldots, X_n\}.$$

Define $T_n \equiv n(\alpha_0 - \widehat{\alpha}_n^{mle})/\alpha_0$. One can show that the limiting distribution of $T_n$ is the standard exponential. In particular, we have $\Pr(T_n = 0) = 0$ for all $n$.

Now, let $\{X_1^*, \ldots, X_n^*\}$ be a bootstrap sample that is obtained by sampling the original data $\{X_1, \ldots, X_n\}$ randomly with replacement. Then, the bootstrap analog of $T_n$ is $T_n^* \equiv n(\widehat{\alpha}_n^{mle} - \widehat{\alpha}_n^*)/\widehat{\alpha}_n^{mle}$, where $\widehat{\alpha}_n^* = \max\{X_1^*, \ldots, X_n^*\}$ (note that the "population" value for the upper bound of the bootstrap samples is $\widehat{\alpha}_n^{mle}$). Hence,

$$\begin{aligned} P_{F_n}(T_n^* = 0) &= \int \mathbf{1}\{\widehat{\alpha}_n^* = \widehat{\alpha}_n^{mle}\} \mathrm{d}F_n^n \\ &= \int (1 - \mathbf{1}\{\widehat{\alpha}_n^{mle} \notin \text{bootstrap sample}\}) \mathrm{d}F_n^n \\ &= 1 - (1 - (1/n))^n \to 1 - e^{-1}. \end{aligned}$$

Thus, the bootstrap distribution of $T_n^*$ is not consistent for the distribution of $T_n$. This result would be intuitively understandable noting the fact that the bootstrap maximum never exceeds the maximum of the original sample.

As in the above example, when the target parameter is on a boundary, simple nonparametric bootstrap methods often fail. For other examples, see [Horowitz, 2001] and [Horowitz, 2019]

## 7.2 Asymptotic refinements

Let $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ be an IID sample from the distribution of the random vector $\mathbf{X} \sim F_0$. Further, let $\mathbf{Z}(\cdot)$ be a vector-valued function of $\mathbf{X}$, and define $\zeta \equiv \mathbb{E}[\mathbf{Z}(\mathbf{X})]$, $\mathbf{Z}_i \equiv \mathbf{Z}(\mathbf{X}_i)$, and $\overline{\mathbf{Z}} \equiv n^{-1}\sum_{i=1}^n \mathbf{Z}_i$. The parameter of interest is $H(\zeta)$ and we estimate it by $H(\overline{\mathbf{Z}})$.

For example, for a simple linear regression with a standardized regressor, we have the slope parameter $\beta = H(\zeta)$ and its estimate $\widehat{\beta} = H(\overline{\mathbf{Z}})$, where $\mathbf{X} = (X, Y)$, $\mathbf{Z}(\mathbf{X}) = (XY, X^2)$,

$$\zeta = \Big(\underbrace{\mathbb{E}[XY]}_{\zeta_1}, \underbrace{\mathbb{E}[X^2]}_{\zeta_2}\Big), \quad \overline{\mathbf{Z}} = \Big(\underbrace{\frac{1}{n}\sum_{i=1}^n X_i Y_i}_{\overline{\mathbf{Z}}_1}, \underbrace{\frac{1}{n}\sum_{i=1}^n X_i^2}_{\overline{\mathbf{Z}}_2}\Big), \quad H(\zeta) = \frac{\zeta_1}{\zeta_2}, \quad H(\overline{\mathbf{Z}}) = \frac{\overline{\mathbf{Z}}_1}{\overline{\mathbf{Z}}_2}.$$

In the following, we focus on estimating the distribution of the statistic

$$T_n = \frac{\sqrt{n}(H(\overline{\mathbf{Z}}) - H(\zeta))}{s_n},$$

where $s_n^2$ is a consistent estimator of the variance of $\sqrt{n}(H(\overline{\mathbf{Z}}) - H(\zeta))$. We assume that $H$ is sufficiently smooth such that it has sufficiently many derivatives. In addition, we assume the following condition:

**Cramér Condition** Let $\tau$ be a vector of constants with the same dimension as $\mathbf{Z}$. Let $i \equiv \sqrt{-1}$. $\mathbf{Z}$ satisfies the Cramér condition:

$$\limsup_{||\tau|| \to \infty} |\mathbb{E}\exp(i\tau'\mathbf{Z})| < 1.$$

That is, the Cramér condition requires that the absolute value of the characteristic function $\phi_{\mathbf{Z}}(\tau) \equiv \mathbb{E}\exp(i\tau'\mathbf{Z})$ is strictly bounded by one uniformly in $\tau$. This condition is satisfied if $\mathbf{Z}$ is continuous and has a proper density function.[3]

Let $G_n(\cdot \mid F_0)$ and $G_\infty(\cdot \mid F_0)$ denote the distribution and the asymptotic distribution of $T_n$, respectively. Similarly, $G_n(\cdot \mid F_n)$ and $G_\infty(\cdot \mid F_n)$ are defined as the bootstrap counterparts of $G_n(\cdot \mid F_0)$ and $G_\infty(\cdot \mid F_0)$, respectively. Then, under the Cramér condition, we have the following result:

$$\begin{aligned} G_n(t \mid F_0) &= G_\infty(t \mid F_0) + \frac{1}{n^{1/2}} g_1(t, F_0) + \frac{1}{n} g_2(t, F_0) + \frac{1}{n^{3/2}} g_3(t, F_0) + O(n^{-2}) \\ G_n(t \mid F_n) &= G_\infty(t \mid F_n) + \frac{1}{n^{1/2}} g_1(t, F_n) + \frac{1}{n} g_2(t, F_n) + \frac{1}{n^{3/2}} g_3(t, F_n) + O(n^{-2}) \end{aligned} \tag{7.2.1}$$

uniformly over $t$. Here, $g_1$ and $g_3$ are even functions and $g_2$ is an odd function with respect to the first argument. The series expansion of a distribution function as given in (7.2.1) is called an **Edgeworth expansion**. From this result, we can see that the conventional asymptotic method that approximates $G_n(t \mid F_0)$ by its asymptotic analog $G_\infty(t \mid F_0)$ has an approximation error of order $n^{-1/2}$.

Note that in our setup, both $T_n \xrightarrow{d} N(0,1)$ and $T_n^* \xrightarrow{d} N(0,1)$ hold. That is, the first terms on the right-hand side of (7.2.1) are both $\Phi(t)$, the standard normal CDF. When the asymptotic distribution of a statistic does

[3] Note that $|\phi_{\mathbf{Z}}(\tau)| \leq 1$ always holds for all $\tau$:

$$\begin{aligned} |\phi_{\mathbf{Z}}(\tau)| &\leq \mathbb{E}|\exp(i\tau'\mathbf{Z})| \\ &= \mathbb{E}|\cos(\tau'\mathbf{Z}) + i\sin(\tau'\mathbf{Z})| \quad \text{(Euler's formula)} \\ &= \mathbb{E}\Big[(\cos^2(\tau'\mathbf{Z}) + \sin^2(\tau'\mathbf{Z}))^{1/2}\Big] = 1. \end{aligned}$$

For a formal derivation of the Cramér condition, see, e.g., Section 2.4 of [Hall, 1992].

not depend on the DGP, such as $G_\infty(t \mid F) = N(0,1)$ for any $F$ as in the present case, we say that the statistic is **asymptotically pivotal**. Most test statistics are asymptotically pivotal, but most estimators are not.

Then, it follows from (7.2.1) that the gap between the true distribution and the bootstrap distribution is

$$G_n(t \mid F_0) - G_n(t \mid F_n) = \frac{1}{n^{1/2}}[g_1(t, F_0) - g_1(t, F_n)] + \frac{1}{n}[g_2(t, F_0) - g_2(t, F_n)] + O(n^{-3/2}).$$

The leading term on the right-hand side is $n^{-1/2}[g_1(t, F_0) - g_1(t, F_n)]$. Since $F_n$ converges to $F_0$, this term converges to zero faster than $n^{-1/2}$. Indeed, one can show that $[g_1(t, F_0) - g_1(t, F_n)] = O(n^{-1/2})$, and therefore the approximation error by the bootstrap is of order $n^{-1}$, i.e., an **asymptotic refinement**. Thus, the bootstrap is more accurate than the conventional asymptotic approximation.

The bootstrap is even more accurate when approximating the symmetrical distribution $\Pr(|T_n| \le t)$. Noting that $g_1$ and $g_3$ are even functions and $g_2$ is an odd function (i.e., $g_1(t, F) = g_1(-t, F)$, $g_2(t, F) = -g_2(-t, F)$, and $g_3(t, F) = g_3(-t, F)$) and that $\Phi(-t) = 1 - \Phi(t)$,

$$\begin{aligned}
P_{F_0}(|T_n| \le t) &= P_{F_0}(T_n \le t) - P_{F_0}(T_n \le -t) \\
&= G_n(t \mid F_0) - G_n(-t \mid F_0) \\
&= \Phi(t) - \Phi(-t) + \frac{1}{n^{1/2}}[g_1(t, F_0) - g_1(-t, F_0)] + \frac{1}{n}[g_2(t, F_0) - g_2(-t, F_0)] \\
&\quad + \frac{1}{n^{3/2}}[g_3(t, F_0) - g_3(-t, F_0)] + O(n^{-2}) \\
&= 2\Phi(t) - 1 + \frac{2}{n} g_2(t, F_0) + O(n^{-2}).
\end{aligned}$$

Similarly,

$$P_{F_n}(|T_n^*| \le t) = 2\Phi(t) - 1 + \frac{2}{n} g_2(t, F_n) + O(n^{-2}).$$

Further, similarly as above, one can show that $[g_2(t, F_0) - g_2(t, F_n)] = O(n^{-1/2})$. Thus, the order of the error in the bootstrap approximation to the symmetrical distribution is $n^{-3/2}$.

Note that if the statistic $T_n$ is not asymptotically pivotal (for example, not standardized $T_n = \sqrt{n}(H(\overline{\mathbf{Z}}) - H(\zeta))$), the leading term of the difference $G_n(t \mid F_0) - G_n(t \mid F_n)$ will be $G_\infty(t \mid F_0) - G_\infty(t \mid F_n)$, which is typically of order $n^{-1/2}$. Then, the error of the bootstrap approximation of the distribution of a statistic that is not asymptotically pivotal converges to zero at the same rate as that of the conventional asymptotic approximation (i.e., no asymptotic refinements); there is no merit of using bootstrap other than computational simplicity.

Table 7.1: Approximation errors for estimating $G_n(\cdot \mid F_0)$

| | Asymptotic distribution | Bootstrap | | |
|---|---|---|---|---|
| | | Pivotal | Pivotal + Symmetricity | Not pivotal |
| The speed of approximation | $n^{-1/2}$ | $n^{-1}$ | $n^{-3/2}$ | $n^{-1/2}$ |

## 7.3 Other bootstrap resampling schemes

### 7.3.1 Parametric bootstrap

The consistency property of the nonparametric bootstrap method is essentially a consequence of the consistency of $F_n$ for $F_0$. Moreover, the size of the error in the bootstrap approximation is determined by the size of $|F_n - F_0|$. Thus, to improve the performance of bootstrap, $F_n$ should be obtained by the most efficient available estimator. Clearly, if we know the form of $F_0$ perfectly, we should use $F_0$ to generate bootstrap samples, rather than $F_n$. When it is known that $F_0$ belongs to a parametric family with some unknown parameters $\theta$ such that $F_0(\cdot) = F(\cdot; \theta_0)$, we can use $F(\cdot; \theta_n)$ as a bootstrap sample generator with $\theta_n$ being a consistent estimator of $\theta_0$ such as an ML estimator. This approach is called the **parametric bootstrap**.

### 7.3.2 Residual bootstrap

Suppose that we are interested in estimating the regression model

$$Y_i = \mathbf{X}_i^\top \beta_0 + \varepsilon_i, \;\; \mathbb{E}(\varepsilon_i \mid \mathbf{X}_i) = 0.$$

We assume that the data $\{(\mathbf{X}_i, Y_i) : 1 \le i \le n\}$ are IID such that $\mathbb{E}(\varepsilon_i^2) = \sigma^2$ (i.e., homoskedasticity). $\mathbf{X}_i$ includes a constant term. For this setup, we can generate bootstrap samples in the following manner:

**Step 1.** Obtain the OLS estimate $\widehat{\beta}_n$ of $\beta_0$.

**Step 2.** Compute the residuals $\widehat{\varepsilon}_i = Y_i - \mathbf{X}_i^\top \widehat{\beta}_n$ for all $i = 1, \ldots, n$.

**Step 3.** Create a bootstrap sample $\{(Y_i^*, \mathbf{X}_i) : 1 \le i \le n\}$, where $Y_i^* = \mathbf{X}_i^\top \widehat{\beta}_n + \widehat{\varepsilon}_i^*$, and $\varepsilon_i^*$'s are obtained by a random sampling from $\{\widehat{\varepsilon}_1, \ldots, \widehat{\varepsilon}_n\}$ with replacement.

In the above procedure, only the residuals are resampled repeatedly to create bootstrap samples, and the regressors $\mathbf{X}_i$'s are the same as the original sample and are not resampled. This approach is called the **residual bootstrap**. Note that in Step 3, we implicitly assume that the errors are identically distributed.

If we know the exact form of the distribution function of the error terms, as in the parametric bootstrap, we can draw new error terms from it, rather than sampling from the residuals. Such a situation is common in parametric limited dependent variable models such as probit and logit.

### 7.3.3 Wild bootstrap

We continue to consider the same regression model as above. However, we now assume that the error terms are heteroskedastic: $\mathbb{E}(\varepsilon_i^2) = \sigma_i^2$. In this case, using the residual bootstrap is not appropriate.

Letting $\widehat{\beta}_n$ denote the OLS estimate of $\beta_0$ and $\widehat{\varepsilon}_i = Y_i - \mathbf{X}_i^\top \widehat{\beta}_n$, the **wild bootstrap** generates new $Y_i^*$'s as $Y_i^* = \mathbf{X}_i^\top \widehat{\beta}_n + u_i^*$, where

$$u_i^* \equiv W_i \widehat{\varepsilon}_i, \;\; i = 1, \ldots, n$$

and $\{W_i : 1 \le i \le n\}$ is a randomly generated IID sequence that is independent of the whole data and satisfies $\mathbb{E}W_i = 0$ and $\mathbb{E}W_i^2 = 1$. Common choices of distributions for $W_i$ include the standard normal and the two-point

distribution that takes $(1-\sqrt{5})/2$ with probability $(\sqrt{5}+1)/(2\sqrt{5})$ and $(\sqrt{5}+1)/2$ with probability $(\sqrt{5}-1)/(2\sqrt{5})$. It can be easily confirmed that

$$\mathbb{E}[u_i^* \mid \{(\mathbf{X}_j, Y_j)\}_{j=1}^n] = \mathbb{E}[W_i \mid \{(\mathbf{X}_j, Y_j)\}_{j=1}^n]\widehat{\varepsilon}_i = 0$$
$$\mathbb{E}[(u_i^*)^2 \mid \{(\mathbf{X}_j, Y_j)\}_{j=1}^n] = \mathbb{E}[W_i^2 \mid \{(\mathbf{X}_j, Y_j)\}_{j=1}^n]\widehat{\varepsilon}_i^2 = \widehat{\varepsilon}_i^2.$$

Hence, the heteroskedasticity in the original sample is retained in the bootstrap sample.

# Chapter 8

# Nonparametric Regression

Consider a general regression model: $Y = g(X) + \varepsilon$. As shown in (1.5.3), in terms of minimization of MSE, the best regression function is the conditional expectation function: $g(X) = \mathbb{E}[Y \mid X]$. A linear regression model $Y = X^\top\beta + \varepsilon$ is a special case when we impose a funcitonal form assumption $g(X) = X^\top\beta$, and thus its optimality hinges on whether the assumption $\mathbb{E}[Y \mid X] = X^\top\beta$ is satisfied or not. In reality, however, it would be extremely rare that the data follow such a linear relationship exactly, and if the functional form assumption does not hold, what the linear regression produces is only a "linear approximation" of $\mathbb{E}[Y \mid X]$. As shown in the figure below, linear approximation is not always informative and sometimes even misleading.

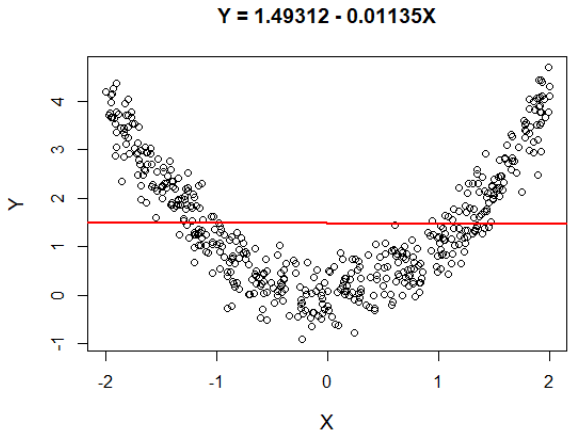



Figure 8.1: Linear approximation of a quadratic function

Thus, ideally, we would like to estimate $\mathbb{E}[Y \mid X]$ without imposing any a priori functional form assumptions. Such a regression approach is called **nonparametric**. In contrast, if one assumes a specific functional form $g(X, \theta)$ with a finite dimensional parameter vector $\theta$, this approach is **parametric**. The most typical parametric regression models are linear regression models. For another example, if the range of a dependent variable is restricted to $(0, 1)$, a logistic regression model $g(X, \theta) = \exp(X^\top\theta)/[1 + \exp(X^\top\theta)]$ is popularly used. The intermediate case between nonparametric and parametric models is called **semiparamtric**. As explained below, estimating a fully nonparametric regression model is almost infeasible in practice due to the curse of

dimensionality unless the dimension of $X$ is less than three or at most four. On the other hand, assuming a full parametric specification has a risk of model misspecification biases. Then, the main idea of semiparametric regression models is to introduce mild functional form assumptions to facilitate the estimation while retaining the flexibility of nonparametric models for certain variables. For example, if one assumes that the impact of a subset $X_1$ of $X$ on $Y$, say $\varphi(\cdot)$, is nonlinear and its functional form is left unspecified, and that the remaining subset $X_2$ is linearly related to $Y$, then the resulting regression model would be

$$Y = \varphi(X_1) + X_2^\top \beta + \varepsilon.$$

This type of semiparametric model is called a partially linear regression model, which will be discussed in detail later.

Note that if $X$ is a discrete random variable, the estimation of $\mathbb{E}[Y \mid X]$ is straightforward; split the data into subsamples according to the realized value of $X$, and compute the average of $Y$ for each subsample. Thus, hereinafter, we assume that $X$ is a continuous random variable.

## 8.1 Nonparametric regression

Suppose we would like to estimate the following regression model:

$$Y = g_0(X) + \varepsilon, \tag{8.1.1}$$

where $\mathbb{E}[\varepsilon \mid X] = 0$ (i.e., $g_0(X) = \mathbb{E}[Y \mid X]$). If we additionally assume a parametric regression model $g(X, \theta_0) = g_0(X)$, once a consistent estimator $\widehat{\theta}_n$ for $\theta_0$ is obtained, we can estimate $g(x, \widehat{\theta}_n) \approx g_0(x)$ for any $x$. However, the correct parametric specification is usually unknown to us. Then, we consider estimating $g_0(\cdot)$ with its functional form fully unrestricted, except for some smoothness constraints.

Let $\mathcal{X} \subseteq \mathbb{R}^d$ be the support of $X$. For any given $x \in \mathcal{X}$, we would like to develop a consistent estimator of $g_0(x)$. There are two major approaches used in nonparametric regression: local regression and global regression. The local approach is to estimate $g_0(x)$ by fitting a simple model using only the observations in the neighborhood of $x$ (the evaluation point $x \in \mathcal{X}$ does not have to be an observaed data point). Since any local estimator only gives an estimate of $g_0(\cdot)$ at a specific point $x$, if one wants to recover the whole shape of $g_0(\cdot)$, it is necessary to repeat the estimation at different points many times. In contrast, the global approach considers a basis expansion of the function $g_0(x) \approx p(x)^\top \theta_0$ and estimates the whole functional form of $g_0(\cdot)$ at once. Here, $p(\cdot)$ is a vector of known basis functions of increasing dimension, and $\theta_0$ is the corresponding coefficient vector. This type of nonparametric regression is called **series regression** estimator or **sieve method**.

### 8.1.1 $k$-nearest-neighbor regression and kernel regression

We first describe the local approach. For simplicity, we temporarily assume that $X$ is one-dimensional. One simplest local regression method is the **$k$-nearest-neighbor ($k$NN) regression**. The $k$NN regression estimator simply computes the average of $Y$ over the $k$-closest observations to $x$ on $\mathcal{X}$. Suppose we have $n$ IID observations $\{(Y_i, X_i) : 1 \le i \le n\}$. Re-ordering the observations in the increasing order of $|X_i - x|$, so that $|X^{(1)} - x| \le |X^{(2)} - x| \le \cdots \le |X^{(n)} - x|$, we estimate $g_0(x)$ by

$$\widehat{g}_n^{knn}(x) \equiv \frac{1}{k} \sum_{i=1}^{k} Y^{(i)}.$$

Clearly, to establish the consistency of the $k$NN estimator, $|X^{(k)} - x|$ must get closer to zero as $n$ increases. On the other hand, $k$ should increase to infinity to apply the law of large numbers to $\widehat{g}_n^{knn}(x)$. That is, there is a bias-variance trade-off in terms of the choice of $k$. Therefore, to meet these conditions simultaneously, we need to deliberately choose $k$ so that it grows to infinity but not too fast.

Another popularly used local regression method is the **kernel regression**. The idea of kernel regression is similar to that of $k$NN regression, but differs in that it introduces a **bandwidth** parameter $h$ such that only the observations whose $X$ value falling into the interval $[x - h,\ x + h]$ are used to compute the estimator. For example, a **local constant regression** estimator is given by

$$\widehat{g}_n^{const}(x) \equiv \frac{\sum_{i=1}^n \mathbf{1}\{|X_i - x| \le h\} Y_i}{\sum_{i=1}^n \mathbf{1}\{|X_i - x| \le h\}}.$$

The above estimator is slightly inefficient in practice because, once the criterion $|X_i - x| \le h$ is met, the observations adjacent to $x$ and those not that close to $x$ have the same contributions to the estimator. Thus, it is possible to improve the estimator by assigning larger weights for the observations closer to the evaluation point $x \in \mathcal{X}$.

To this end, we introduce a general weighting function $K(\cdot)$, which is referred to as the **kernel function**. The kernel function is a continuous density function (i.e., $\int K(u)du = 1$), and is assumed to be symmetric around zero. Some typical choices for the kernel function are: uniform kernel $K(u) = \frac{1}{2}\mathbf{1}\{|u| \le 1\}$, Epanechnikov kernel $K(u) = \frac{3}{4}(1 - u^2)\mathbf{1}\{|u| \le 1\}$, and Gaussian kernel $K(u) = \frac{1}{\sqrt{2\pi}} \exp(-u^2/2)$; see Figure 8.2.

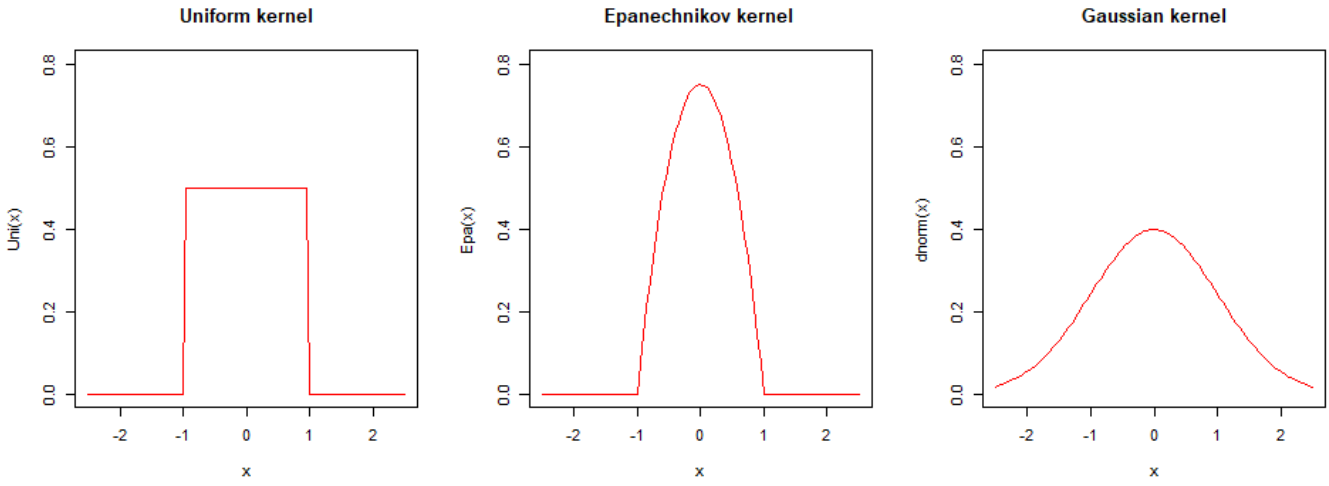



Figure 8.2: Kernel density functions

One can easily see that the local constant estimator given above is numerically equivalent to

$$\widehat{g}_n^{const}(x) = \frac{(nh)^{-1} \sum_{i=1}^n K\left(\frac{X_i - x}{h}\right) Y_i}{(nh)^{-1} \sum_{i=1}^n K\left(\frac{X_i - x}{h}\right)}$$

with $K$ being the uniform kernel. Dividing both the numerator and denominator by $nh$ is for technical convenience. From a straightforward calculation, one can find that this $\widehat{g}_n^{const}(x)$ is equivalent to the solution of the

following weighted least squares problem:

$$\widehat{g}_n^{const}(x) = \underset{a}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - a)^2 K\left(\frac{X_i - x}{h}\right).$$

Hence, the local constant estimator can be viewed as fitting a constant-only model to the (weighted) local observations in the neighborhood of $x$ (this is why the estimator is called local "constant").

As an extension of the local constant approach, we can consider fitting a linear regression model to the local observations instead of simply taking the average or fitting a constant-only model. Such an approach is called the **local linear regression**. That is, the local linear kernel regression estimator is defined as $\widehat{g}_n^{ll}(x) \equiv \widehat{a}_n$, where

$$(\widehat{a}_n, \widehat{b}_n) = \underset{(a,b)}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - a - b(X_i - x))^2 K\left(\frac{X_i - x}{h}\right).$$

It is known that by including an additional regressor $(X_i - x)$, the local linear regression estimator achieves a nice bias-correction property that the local constant estimator does not have (for more details, see, e.g., [Li and Racine, 2007]). As a further generalization, one may consider including higher order polynomials $(X_i - x)^2$, $(X_i - x)^3$, ... as additional regressors. These estimators are called the local polynomial estimators.

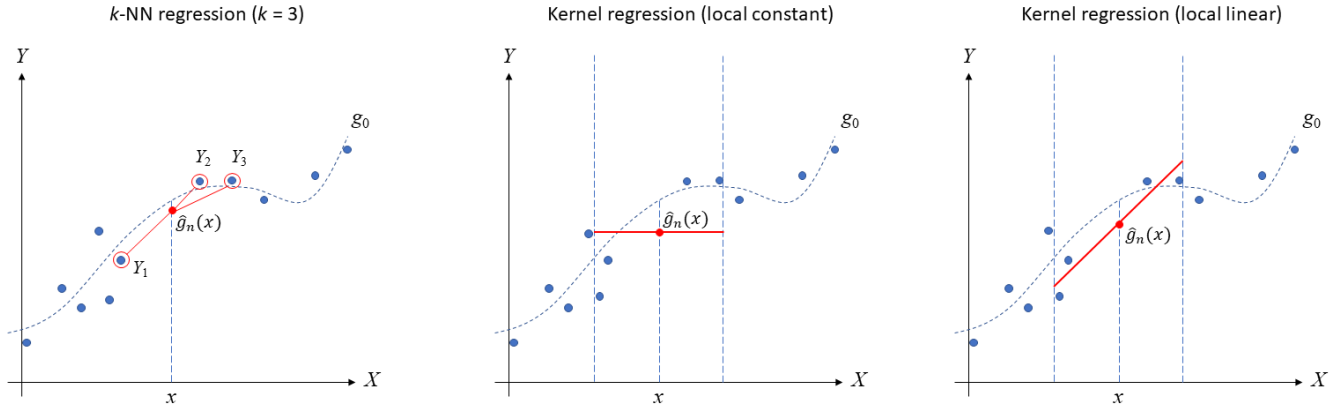



Figure 8.3: Loacal nonparametric regression

Here, let us provide a sketch for the proof of the consistency of $\widehat{g}_n^{const}(x)$. Let $f_X$ be the density function of $X$. We first show that the numerator term of $\widehat{g}_n^{const}(x)$ converges to $\mathbb{E}[Y \mid X = x] f_X(x)$ in probability. Observe that

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) Y_i\right] &= \frac{1}{h}\mathbb{E}\left[K\left(\frac{X - x}{h}\right) Y\right] \quad \text{(IID assumption)} \\
&= \frac{1}{h}\int K\left(\frac{w - x}{h}\right) \mathbb{E}[Y \mid X = w] f_X(w) \mathrm{d}w \quad \text{(LIE)} \\
&= \int K(u) \mathbb{E}[Y \mid X = x + uh] f_X(x + uh) \mathrm{d}u \quad \text{(change of variables: } u = (w - x)/h\text{)}.
\end{aligned}$$

Recalling that $\int K(u)\mathrm{d}u = 1$ and assuming that $\mathbb{E}[Y \mid X = \cdot]$ and $f_X(\cdot)$ are bounded continuous functions, we have

$$\lim_{h \to 0} \mathbb{E}\left[\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) Y_i\right] = \mathbb{E}[Y \mid X = x] f_X(x)$$

by the bounded convergence theorem. By similar but slightly more tedious calculations, we can show that the variance of $(nh)^{-1}\sum_{i=1}^n K\left(\frac{X_i-x}{h}\right)Y_i$ is of order $O((nh)^{-1})$. Thus, if $h \to 0$ and $nh \to \infty$ as $n \to \infty$, by Chebyshev's inequality, we get $(nh)^{-1}\sum_{i=1}^n K\left(\frac{X_i-x}{h}\right)Y_i \xrightarrow{p} \mathbb{E}[Y \mid X = x]f_X(x)$.

By the same argument as above, we can prove that the denominator of $\widehat{g}_n^{const}(x)$, $(nh)^{-1}\sum_{i=1}^n K\left(\frac{X_i-x}{h}\right)$, converges to $f_X(x)$ in probability (this is the so-called **kernel density estimator**) if $h \to 0$ and $nh \to \infty$ as $n \to \infty$. Combining these results together, the consistency of $\widehat{g}_n^{const}(x)$ is proved.

Note that, similar to the $k$NN regression estimator, there is a bias-variance trade-off for the choice of the bandwidth $h$. As we have discussed, as $n$ increases, $h$ must converge to zero sufficiently slowly so that $nh$ can increase to infinity. Intuitively, the effective sample size for a kernel regression is of order $O(nh)$ for each evaluation point, and thus to apply the law of large numbers, $nh$ must grow to infinity.

So far, we have focused on the case where the dimension of $X$ is just one. When $X = (X_1, \ldots, X_d)$ is a general $d$-dimensional variable, the aforementioned estimators require some modifications. Specifically, we introduce a "multivariate" kernel weighting function, say $\mathbf{K}(u)$, where $u = (u_1, \ldots, u_d) \in \mathbb{R}^d$, and we can estimate $g_0(x)$ in a similar manner using $\mathbf{K}(u)$ in the place of $K(u)$. One simple way to construct $\mathbf{K}(u)$ is to take the product of one-dimensional kernel functions: $\mathbf{K}(u) = \prod_{j=1}^d K(u_j)$. For example, if one employs the same bandwith $h$ for all $X_j$'s, the product uniform kernel function is

$$\mathbf{K}\left(\frac{X_i - x}{h}\right) = \frac{1}{2^d}\mathbf{1}\{|X_{i,1} - x_1| \le h\}\mathbf{1}\{|X_{i,2} - x_2| \le h\} \times \cdots \times \mathbf{1}\{|X_{i,d} - x_d| \le h\}.$$

This means that only the observations contained in the $d$-dimensional cube centered at $x$ with each side length of $2h$ can be used for estimating $g_0(x)$. If the data are uniformly distributed on $\mathcal{X}$, the size of the subsample in each cube is of order $O(nh^d)$. Therefore, since $h$ is assumed to tend to zero, the size of the effective sample decreases "exponentially" (not proportionally) with respect to $d$. Thus, nonparametric estimation of $g_0(x)$ becomes less and less feasible quickly and inefficient as $d$ grows. This issue is known as the **curse of dimensionality**. As far as I can see in the empirical literature, full nonparametric regression has rarely been considered when $d$ is larger than three, except when the sample size is very large (e.g., $10^5$ or more).

### 8.1.2 Series regression

Again, we first focus on the estimation of the model in (8.1.1) with a one-dimensional $X$, and extend the discussion to more general cases later. The idea of approximating a general continuous function by a linear combination of basis functions has a long history. Perhaps one of the most familiar to readers would be the Maclaurin expansion (i.e., the Taylor expansion at 0): assuming that $g_0(\cdot)$ is sufficiently many times differentiable, at each $x \in \mathcal{X}$

$$g_0(x) = g_0(0) + g_0^{(1)}(0)x + \frac{g_0^{(2)}(0)}{2!}x^2 + \cdots + \frac{g_0^{(k)}(0)}{k!}x^k + \cdots,$$

where $g_0^{(k)}$ denotes the $k$-th order derivative of $g_0$. This suggests that if $g_0(\cdot)$ is sufficiently smooth, we can approximate $g_0(x)$ by $g_0(x) \approx p(x)^\top\theta_0$ with $p(x) = (1, x, x^2, \ldots, x^k)^\top$, the **power series** of order $k$, and $\theta_0 = (g_0(0), g_0^{(1)}(0), \ldots, g_0^{(k)}(0)/k!)^\top$.

A similar well-known result is the Weierstrass approximation theorem. The theorem states that for any continuous function on a closed interval $\mathcal{X}$, the same power-series approximation holds in the sense that the approximation error $g_0(x) - p(x)^\top\theta_0$ can be made arbitrarily small for any $x \in \mathcal{X}$ by choosing a sufficiently large polynomial order and an appropriate coefficient vector $\theta_0$. Remarkably, the differentiability is not required here.

Historically, the most important series expansion result for humankind would be the **Fourier series** expansion. The French mathematician Joseph Fourier (1768-1830) claimed that any periodic function can be decomposed into a linear combination of sine and cosine curves. Although its formal proof was not given by Fourier, it was later shown that his claim was after all true for a wide class of functions. Then, named after his discovery, the following series expansion is called the Fourier series expansion: for a continuous function $g_0(\cdot)$ on a closed interval $\mathcal{X}$,

$$g_0(x) = a_0 + a_1 \sin(\omega x) + b_1 \cos(\omega x) + a_2 \sin(\omega 2x) + b_2 \cos(\omega 2x) + \cdots + a_k \sin(\omega k x) + b_k \cos(\omega k x) + \cdots,$$

where $\omega$ is a constant determined by the size of $\mathcal{X}$.

In Figure 8.4, we provide numerical results for the power series approximation and the Fourier series approximation. In the figure, the true functional form of $g_0(\cdot)$ is presented as the dotted curve. We can clearly observe that in both series expansions, more and more precise approximation can be achieved by increasing the order of the basis functions.

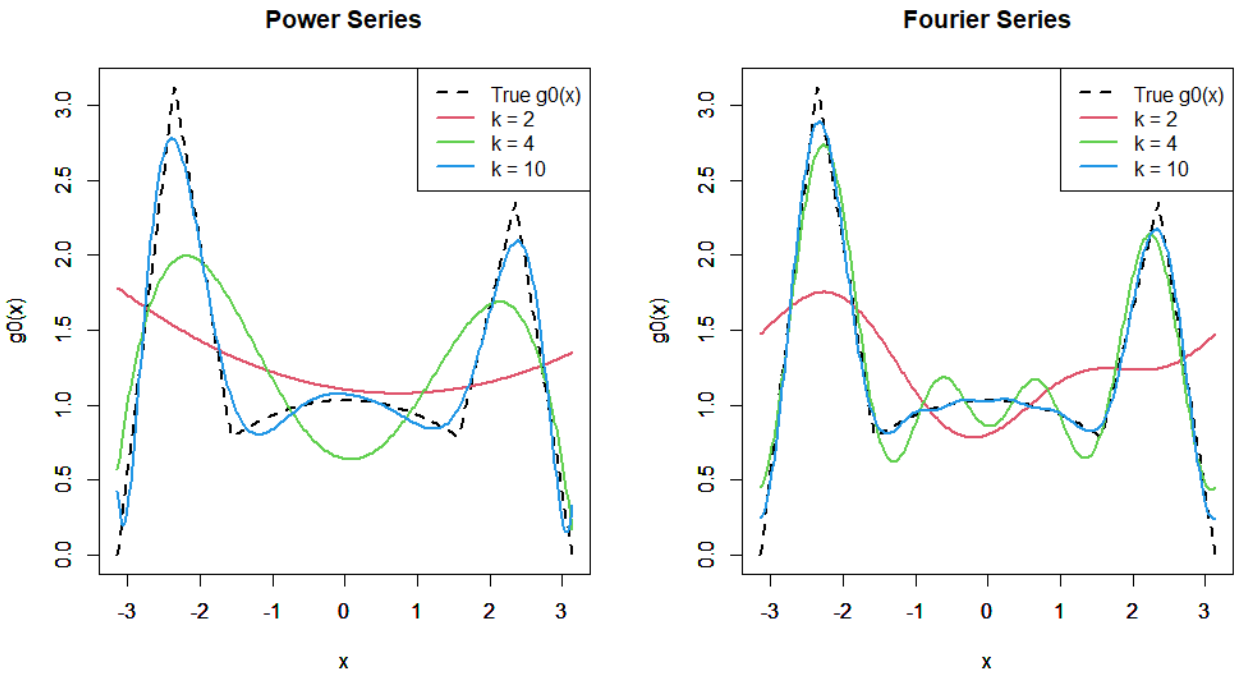



Figure 8.4: Series approximation

As such, with suitable conitnuity conditions, a fairly large class of functions can be approximated by some series expansion. In addition to those introduced above, basis functions that are often adopted in applications include splines, wavelets, artificial neural networks, and so forth (see [Chen, 2007] for their definitions and other examples). Now, let $\{p_1(x), p_2(x), \ldots\}$ be a sequence of $\mathbb{R}$-valued basis functions, and $p(x) = (p_1(x), \ldots, p_k(x))^\top$. We assume that for a sufficiently large $k$, there exists a $k \times 1$ vector $\theta_0$ such that the target function $g_0(\cdot)$ can be well-approximated by $p(\cdot)^\top \theta_0$ uniformly over $\mathcal{X}$ in the following sense:

$$|g_0(x) - p(x)^\top \theta_0| \leq O(k^{-\alpha}) \text{ for any } x \in \mathcal{X},$$

where $\alpha > 0$ is a constant that depends on the choice of the basis function, the smoothness of $g_0$, and the dimension of $X$ (for details, see [Chen, 2007]). The larger $\alpha$, the smoother the function $g_0$ is (the larger number of derivatives $g_0$ has).[1] Using this series approximation, we can rewrite (8.1.1) as follows:

$$\begin{aligned} Y &= g_0(X) + \varepsilon \\ &= p(X)^\top \theta_0 + \eta \end{aligned}$$

where $\eta \equiv g_0(X) - p(X)^\top \theta_0 + \varepsilon$. Thus, the nonparametric regression model can be transformed into a "linear regression" model with increasing dimension. Indeed, we can estimate $\theta_0$ simply by running an OLS regression of $Y$ on $p(X)$:[2]

$$\begin{aligned} \widehat{\theta}_n &\equiv \operatorname*{argmin}_\theta \frac{1}{n}\sum_{i=1}^n (Y_i - p(X_i)^\top \theta)^2 \\ &= \left(\frac{1}{n}\sum_{i=1}^n p(X_i)p(X_i)^\top\right)^{-1} \frac{1}{n}\sum_{i=1}^n p(X_i)Y_i \end{aligned}$$

Once we obtain $\widehat{\theta}_n$, for any given $x \in \mathcal{X}$, we can estimate $g_0(x)$ by $\widehat{g}_n(x) \equiv p(x)^\top \widehat{\theta}_n$. Thus, compared to the kernel regression approach, series regression is much easier to implement.

To prove the consistency of $\widehat{g}_n(x)$, by the triangle inequality, observe that

$$\begin{aligned} |\widehat{g}_n(x) - g_0(x)| &\le |p(x)^\top(\widehat{\theta}_n - \theta_0)| + |g_0(x) - p(x)^\top \theta_0| \\ &\le |p(x)^\top(\widehat{\theta}_n - \theta_0)| + O(k^{-\alpha}) \end{aligned}$$

Hence, it suffices to show that $|p(x)^\top(\widehat{\theta}_n - \theta_0)| = o_P(1)$. Here, assuming that the basis functions are bounded, we have $||p(x)|| = (p_1^2(x) + \cdots + p_k^2(x))^{1/2} = O(\sqrt{k})$. Further, under some regularity conditions, we can show that

$$||\widehat{\theta}_n - \theta_0|| = O_P\left(\sqrt{\frac{k}{n}} + k^{-\alpha}\right). \tag{8.1.2}$$

A sketch of the proof of this result can be found in the appendix of this chapter. Then, combining these results with Cauchy-Schwarz inequality gives that $|p(x)^\top(\widehat{\theta}_n - \theta_0)| \le ||p(x)|| \cdot ||\widehat{\theta}_n - \theta_0|| = O_P(k/\sqrt{n} + k^{1/2-\alpha})$.[3] Thus, if $\alpha > 1/2$ (i.e., $g_0$ is sufficiently smooth) and $k$ grows to infinity sufficiently slowly so that $k/\sqrt{n} \to 0$ is maintained, we obtain $\widehat{g}_n(x) - g_0(x) \xrightarrow{p} 0$ for any $x \in \mathcal{X}$, as desired.

Now we turn to the cases where $X = (X_1, \ldots, X_d)$ is a general $d$-dimensional variable. As can be inferred from a multivariate Taylor expansion, if $X$ is $d$-dimensional, the number of basis functions required to approximate the target function increases in the order $O(k^d)$. In general, we can construct a multivariate basis functions by taking the Kronecker product of univariate basis functions:

$$p(x) = p(x_1) \otimes p(x_2) \otimes \cdots \otimes p(x_d).$$

[1] Among the basis functions listed above, while the power series is quite popular, it is known that the power series expansion has relatively a "slower" convergence rate than the others.

[2] Here, we implicitly assume for simplicity that $\frac{1}{n}\sum_{i=1}^n p(X_i)p(X_i)^\top$ is a nonsingular matrix. However, noting that the dimension of this matrix grows to infinity along with $n$, it is generally possible that the matrix is not invertible for finite $n$. In that case, the inverse should be replaced by a generalized inverse.

[3] This evaluation is actually a rough bound. For more sophisticated results, see, for example, [Belloni et al., 2015] and [Chen and Christensen, 2015].

Thus, if $d$ is not small, the number of coefficient parameters to be estimated easily reaches several hundreds, resulting in a serious loss of efficiency. This is the series-regression version of the curse of dimensionality.

### 8.1.3 An empirical illustration: estimation of Engel curve

As an empirical illustration, we nonparametrically estimate the Engel curve for food consumption using the local constant kernel regression and the series regression based on B-splines; a B-spline function is a piecewise polynomial defined by a set of "cut points", called **knots**.

Install **np** and **splines** packages, and load them. The data used here is `Engel95`, a random sample taken from the U.K. Family Expenditure Survey 1995, which can be imported from the **np** package.

```
library(np) # run nonparametric kernel regression
library(splines) # compute b-spline functions

data("Engel95") # sample size n = 1655
food <- Engel95$food # share of food consumption
exp  <- Engel95$logexp # log of total expenditure
```

We first run the kernel regression using the **npreg** function. This function automatically computes an optimal bandwidth parameter $h$, so that we do not need to specify it.

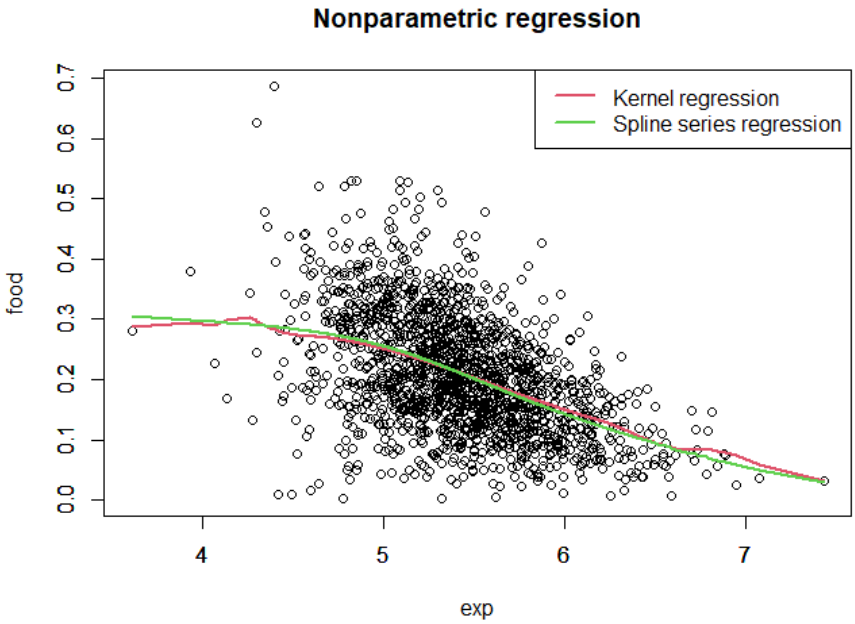
```
# Kernel regression #
ghat_kernel <- npreg(food ~ exp)
```

Next, we perform the series regression. The B-spline basis function can be computed using the **bs** function. Here, we set the 25%, 50%, and 75% empirical quantiles of `exp` as the knot values and create B-splines with $k = 6$. After the B-splines are created, we simply run an OLS regression.

```
# Spline series regression #
knots <- quantile(exp, (1:3)/4)
p <- bs(exp, knots = knots) # compute the B-splines with k = 6
ghat_spline <- predict(lm(food ~ p))
```

To summarize the estimation results of nonparametric regression, we present them visually. Running the code below, we can observe that both the kernel and series regression perform very similarly.

```
plot(exp, food, xlim = range(exp), ylim = range(food),
       main = "Nonparametric regression", xlab = "exp", ylab = "food")
par(new = T)
plot(exp[order(exp)], ghat_kernel$mean[order(exp)], xlab = "", ylab = "",
       xlim = range(exp), ylim = range(food), type = "l", lwd = 2, col = 2)
par(new = T)
plot(exp[order(exp)], ghat_spline[order(exp)], xlab = "", ylab = "",
       xlim = range(exp), ylim = range(food), type = "l", lwd = 2, col = 3)
legend("topright", c("Kernel regression", "Spline series regression"),
       lty = c(1,1), lwd = c(2,2), col = c(2,3))
```




## 8.2 Semiparamtric regression

### 8.2.1 Partially linear models

### 8.2.2 Generalized additive models

### 8.2.3 Functional coefficient models

## Appendix: Proof of (8.1.2)

Let $P = (p(X_1), \ldots, p(X_n))^\top$, $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$, $\mathcal{E} = (\varepsilon_1, \ldots, \varepsilon_n)^\top$, $v_i = g_0(X_i) - p(X_i)^\top \theta_0$, and $\mathbf{V} = (v_1, \ldots, v_n)^\top$. Using these notations, we can write $\mathbf{Y}/n = P\theta_0 + \mathbf{V} + \mathcal{E}$. Then,

$$\begin{aligned}
\widehat{\theta}_n - \theta_0 &= \left[P^\top P/n\right]^{-1} P^\top \mathbf{Y}/n - \theta_0 \\
&= \left[P^\top P/n\right]^{-1} P^\top (P\theta_0 + \mathbf{V} + \mathcal{E})/n - \theta_0 \\
&= \left[P^\top P/n\right]^{-1} P^\top \mathbf{V}/n + \left[P^\top P/n\right]^{-1} P^\top \mathcal{E}/n
\end{aligned}$$

Assume that $\mathbb{E}[P^\top P/n]$ is positive definite for all $k$. Applying a matrix law of large numbers to $P^\top P/n$, we have $||P^\top P/n - \mathbb{E}[P^\top P/n]|| = o_P(1)$.[4] Then, with sufficeitnly large $n$, the minimum eigenvalue of $\left[P^\top P/n\right]^{-1}$ is bounded by some postive constant with a high probability. Hence, noting that the maximum eigenvalue of an idempotent matrix is at most one (see Footnote 3, Chapter 4),

$$\begin{aligned}
||\left[P^\top P/n\right]^{-1} P^\top \mathbf{V}/n||^2 &= \mathbf{V}^\top P \left[P^\top P/n\right]^{-2} P^\top \mathbf{V}/n^2 \\
&\leq O_P(1) \cdot \mathbf{V}^\top P \left[P^\top P/n\right]^{-1} P^\top \mathbf{V}/n^2
\end{aligned}$$

[4] A rigorous proof of this result is not that easy since $P^\top P/n$ is a matrix whose dimension increases to infinity along with $n$. For interested readers, see, for example, Lemma 4 of [Horowitz and Mammen, 2004].

$$\begin{aligned} &\leq O_P(1) \cdot \mathbf{V}^\top \mathbf{V}/n \\ &= O_P(1) \cdot \frac{1}{n} \sum_{i=1}^{n} |v_i|^2 \leq O_P(k^{-2\alpha}), \end{aligned}$$

where in the last inequality, we have used $\max_i |v_i| = O(k^{-\alpha})$. This implies that $||\left[P^\top P/n\right]^{-1} P^\top \mathbf{V}/n|| = O(k^{-\alpha})$.

Next, observe that $\mathbb{E}[\mathcal{E}\mathcal{E}^\top \mid \{X_i\}_{i=1}^n]$ is a diagonal matrix whose typical element is $\mathbb{E}[\varepsilon_i^2 \mid \{X_i\}_{i=1}^n]$, which we assume to be bounded by some large constant $C$, uniformly in $i$. Then, we have

$$\begin{aligned} \mathbb{E}\left[||\left[P^\top P/n\right]^{-1} P^\top \mathcal{E}/n||^2 \mid \{X_i\}_{i=1}^n\right] &= \text{trace}\left\{\left[P^\top P/n\right]^{-1} P^\top \mathbb{E}[\mathcal{E}\mathcal{E}^\top \mid \{X_i\}_{i=1}^n] P \left[P^\top P/n\right]^{-1}\right\}/n^2 \\ &\leq C\text{trace}\left\{\left[P^\top P/n\right]^{-1}\right\}/n = O(k/n). \end{aligned}$$

Thus, by Markov's inequality, $||\left[P^\top P/n\right]^{-1} P^\top \mathcal{E}/n||^2 = O_P(k/n)$, implying that $||\left[P^\top P/n\right]^{-1} P^\top \mathcal{E}/n|| = O_P(\sqrt{k/n})$. Finally, (8.1.2) holds by the triangle inequality.

# Part II

# Econometric Analysis of Cross-Sectional Dependence Models

# Chapter 9

# Cross-Sectional Dependence

*Humans are social animals* (Aristotle).

## 9.1 Social interaction

We cannot live alone away from other people. Our behavior is inevitably affected by those around us and the social groups we belong to, and vice versa. This interaction with others is called **social interaction**. Thus, in order to precisely understand the nature of human behavior, such feature, i.e. social interaction, should not (or often cannot) be overlooked.

In the literature of econometrics, there is a long history in modeling the dependence structure of time series data – the interactions between the past and present. Estimation of econometric models with social interactions (i.e., "cross-sectional" interactions) is a relatively young research theme (as compared to time series literature), and now growing attention has been devoted to this field.

In an early seminal paper, [Manski, 1993] considered a linear social interaction model, which is called a **linear-in-means model**, and distinguished three types of effects among the social interaction effects: an individual's outcome can be affected by the average outcome in the group to which he/she belongs (**endogenous effects**), by the average individual characteristics in the group (**contextual effects**), and by the common environment of the group (**correlated effects**).

For example, consider a student's academic achievement as the dependent variable of interest, say, $Y$. Let $X$ be a determinant of academic achievement, such as whether or not belonging to an academic club. Further, denote $e$ as the quality of class teacher. There is an endogenous effect if individual achievement $Y$ tends to vary with the mean achievement $\overline{Y}$ of the students in the same school, class room, or other reference groups ($\overline{Y} \to Y$). Similarly, there is a contextual effect if achievement $Y$ tends to vary with $\overline{X}$, where $\overline{X}$ is, for example, the ratio of students belonging to academic clubs ($\overline{X} \to Y$). There are correlated effects if the students in the same class tend to achieve similarly because they are taught by the same teacher ($e \to Y$).

Importantly, the three hypotheses have different policy implications. Consider, for example, an educational intervention providing a tutoring program to some of the students but not to the others. If the endogenous effect exists, then an effective tutoring program not only directly helps the tutored students but, as their achievement rises, indirectly helps other students with a feedback to further achievement gains by the tutored students. This

is the so-called **spillover effect** or **social multiplier effect**. Exogenous effects and correlated effects do not possess such multiplier mechanism. Thus, identification of the three social interaction effects separately is of great importance to policy implementation and evaluation. However, as described in Chapter 10, this is not an easy task because of the so-called **reflection problem**.

## 9.2 Social network

**Social network** is a type of platform that facilitates social interactions between individuals. Here, the term "social network" is used to refer not only to the onilne SNS (social networking service), such as Facebook or Twitter, but to the connection of individuals in a broader sense (friendship, classmates, colleagues, international trade alliance, author-coauthor relationship, etc). Mathematically, a social network can be represented by a directed or undirected **graph**, where individuals are represented as nodes/vertices and connections between them are represented as edges of the graph. For example, Figure 9.1 shows the international trade network in 1954 ([Arpino et al., 2017]). The black edges indicate that both countries that are connected by the edge are members of the GATT (The General Agreement on Tariffs and Trade); gray edges indicate that at least one of the two countries is not a GATT member.

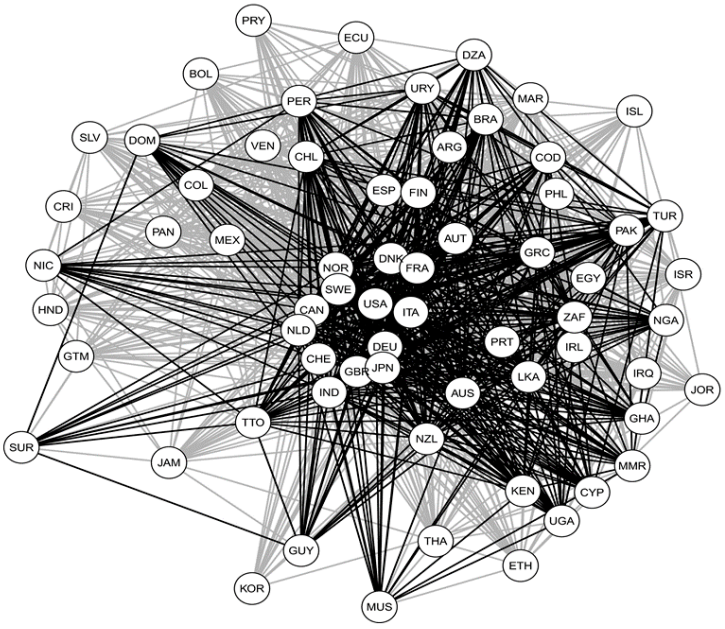



Figure 9.1: Trade partners in 1954 and GATT membership, [Arpino et al., 2017].

For another example, Figure 9.2 shows the co-authorship network for 1,767 published papers in the field of statistics ([Said et al., 2010]).

Identification of linear-in-means social interaction models with network structure has been investigated by [Bramoullé et al., 2009]. They show that, when social interactions are structured through a social network, endogenous and contextual effects can be identified (i.e., the reflection problem does not occur) on "most" networks. This result will be described in detail in Chapter 11.

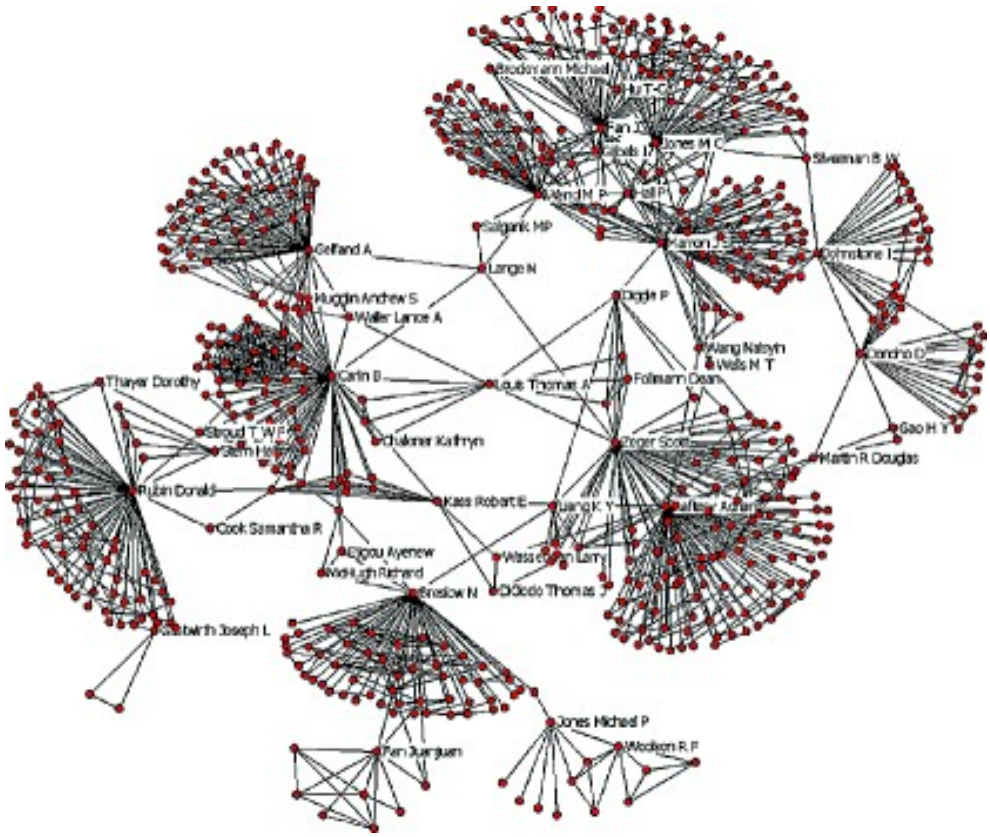



Figure 9.2: Co-authorship social network of statisticians, [Said et al., 2010].

## 9.3 Spatial dependence

Data of observations where each observation is identified by its geographical location is called spatial data. **Spatial dependence** (or, almost synonymously, **spatial autocorrelation**) – the tendency for nearby observations to be more highly correlated than those that are further apart – distinguishes the analysis of spatial data from that of non-spatial data. Spatial dependence is a common phenomenon observed in data from diverse areas such as climatology, geology, epidemiology, real estate market, crime, spatial price competition, etc.

Figure 9.3 shows the spatial distribution of tuberculosis cases from 2002 to 2009 for each municipality in Brazil ([Harling and Castro, 2014]). In this figure, the higher rates of tuberculosis are found in the northwestern municipalities and in the coastal municipalities, indicating the presence of spatial dependence.

Figure 9.4 presents a map of Tokyo 23 wards and the value of the rate of household burglary per 1,000 households for each district in 2011 ([Hoshino, 2018]). The map indicates particularly low and high rates of household burglaries in the central and the northeast zones, respectively. The **Moran's I** statistic, which is shown in the lower right of the figure (0.3968), is a type of correlation coefficient measuring the strength of spatial dependence, ranging from -1 (negative spatial dependence) to 1 (positive spatial dependence). In Chapter 12, we introduce several methods to statistically measure the degree of spatial correlation in data.

The field of statistics focusing on models that explicitly account for such spatial dependence is called **spatial statistics**, and particularly in the context of econometrics, it is called **spatial econometrics**. Since the pioneering work by [Anselin, 1988], spatial econometric methods have been widely applied in a number of empirical

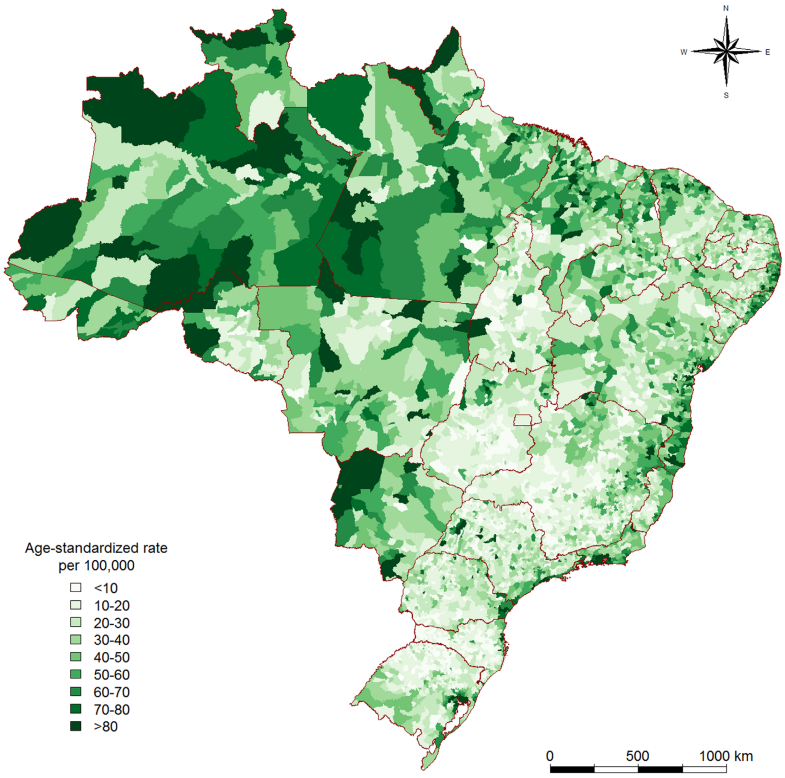



Figure 9.3: Municipal tuberculosis notification rates per 100,000 in Brazil 2002-2009, [Harling and Castro, 2014].

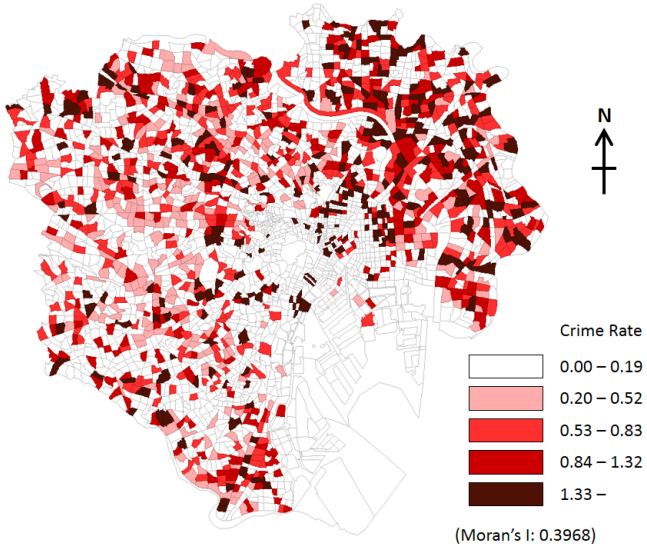



Figure 9.4: Distribution of household burglary rates in Tokyo 23-wards area 2011, [Hoshino, 2018].

researches in economics. An overview and introductory discussion of spatial econometric models are given in Chapters 13 and 14.

One may have noticed that the spatial dependence can be seen as a special form of social interaction where the reference social group is formed based on geographical proximity. Or, one can see social interaction as a special

case of spatial dependence where the "space" is defined by social and economic distance between individuals. Indeed, spatial econometric models and social interaction models have similarities in many aspects.

# Chapter 10

# The Reflection Problem

## 10.1 Linear-in-means model

[Manski, 1993] considered three hypotheses in order to explain the observation that individuals belonging to the same social group tend to behave similarly:

**Endogenous effects** individuals' outcome can be affected by the mean outcome in the group.

**Contextual effects** individuals' outcome can be affected by the exogenous characteristics of the group (also referred to as **exogenous effects**).

**Correlated effects** individuals in the same group tend to behave similarly because they face similar institutional environments.

Let $Y$ be an outcome variable of interest (e.g., student's academic achievement), and $Z$ be a $K \times 1$ vector of individual characteristics. Also, let $X$ denote a $J \times 1$ vector of attributes characterizing an individual's reference group. Then, to account for the three distinct social effects, [Manski, 1993] considered the following model:

$$Y = \alpha + \beta \mathbb{E}[Y \mid X] + \mathbb{E}[Z \mid X]^\top \gamma + Z^\top \eta + U, \quad \mathbb{E}[U \mid X, Z] = X^\top \delta \tag{10.1.1}$$

where $(\alpha, \beta, \gamma, \eta, \delta)$ are unknown parameters. The model in (10.1.1) is called the **linear-in-means model**. It follows that the conditional expectation of $Y$ given $(X, Z)$ has the linear form

$$\mathbb{E}[Y \mid X, Z] = \alpha + \underbrace{\beta \mathbb{E}[Y \mid X]}_{\text{endogenous effect}} + \underbrace{\mathbb{E}[Z \mid X]^\top \gamma}_{\text{contextual effect}} + Z^\top \eta + \underbrace{X^\top \delta}_{\text{correlated effect}} \tag{10.1.2}$$

If $\beta \neq 0$, $Y$ varies with $\mathbb{E}[Y \mid X]$, the mean of $Y$ among those individuals in the reference group characterized by $X$ (endogenous effect); if $\gamma \neq 0$, $Y$ varies with $\mathbb{E}[Z \mid X]$, the mean of $Z$ among those individuals in the reference group (contextual effect); and if $\delta \neq 0$, individuals in the reference group $X$ tend to behave similarly (correlated effect).

An important special case is when $X$ is a vector of indicator variables identifying which social group individuals belong to (i.e., $J$ groups in total). In this case, $\mathbb{E}[Y \mid X]$ and $\mathbb{E}[Z \mid X]$ simply denote the group means of $Y$ and $Z$, respectively, and $\delta$ represents the group-specific fixed effects.

## 10.2 The reflection problem

Taking the expectation of (10.1.2) with respect to $Z$ conditional on $X$ gives that $\mathbb{E}[Y \mid X]$ solves the equation

$$\mathbb{E}[Y \mid X] = \alpha + \beta \mathbb{E}[Y \mid X] + \mathbb{E}[Z \mid X]^\top (\gamma + \eta) + X^\top \delta \tag{10.2.1}$$

that is, $\mathbb{E}[Y \mid X]$ can be characterized as a fixed point of

$$H(p) \equiv \alpha + \beta p + \mathbb{E}[Z \mid X]^\top (\gamma + \eta) + X^\top \delta.$$

If $\beta \neq 1$, we can solve (10.2.1) with respect to $\mathbb{E}[Y \mid X]$ as

$$\mathbb{E}[Y \mid X] = \frac{\alpha}{1-\beta} + \mathbb{E}[Z \mid X]^\top \frac{(\gamma + \eta)}{1-\beta} + X^\top \frac{\delta}{1-\beta}. \tag{10.2.2}$$

Thus, $\mathbb{E}[Y \mid X]$ is a linear function of $(1, \mathbb{E}[Z \mid X], X)$, which implies that the endogenous effects can be expressed as the linear combination of the contextual effects and the correlated effects. Hence, in the linear-in-means model (10.1.1), the endogenous effects cannot be distinguished from the other two effects.

Specifically, inserting the right-hand side of (10.2.2) into (10.1.2) yields

$$\begin{aligned}
\mathbb{E}[Y \mid Z, X] &= \alpha + \frac{\alpha\beta}{1-\beta} + \mathbb{E}[Z \mid X]^\top \left( \gamma + \frac{\beta(\gamma + \eta)}{1-\beta} \right) + Z^\top \eta + X^\top \left( \delta + \frac{\delta\beta}{1-\beta} \right) \\
&= \frac{\alpha}{1-\beta} + \mathbb{E}[Z \mid X]^\top \frac{\gamma + \beta\eta}{1-\beta} + Z^\top \eta + X^\top \frac{\delta}{1-\beta}.
\end{aligned}$$

**Theorem 10.2.1 ([Manski, 1993])** *Suppose that $\beta \neq 1$ and $(1, \mathbb{E}[Z \mid X], Z, X)$ are linearly independent. Then, the composite parameters $(\alpha/(1-\beta), (\gamma + \beta\eta)/(1-\beta), \eta, \delta/(1-\beta))$ are identified.*

The interpretation of the theorem is as follows. Suppose that a dataset $\{(Y_i, Z_i, X_i) : 1 \leq i \leq n\}$ is available. The theorem states that what we can estimate from the data are the four composite parameters: $\alpha/(1-\beta)$, $(\gamma + \beta\eta)/(1-\beta)$, $\eta$, and $\delta/(1-\beta)$, no matter how large $n$ is. On the other hand, the number of unknown parameters $(\alpha, \beta, \gamma, \eta, \delta)$ is five. Therefore, we cannot identify these parameters uniquely (except for $\eta$). In particular, the three social effects $(\beta, \gamma, \delta)$ cannot be distinguished. This is called the **reflection problem**.[1]

If one has additional information on some parameter values, the identification result can be improved. For example, suppose that contextual effects and correlated effects do not exist: $\gamma = \delta = 0$. Then, in this case, we have

$$\mathbb{E}[Y \mid Z, X] = \frac{\alpha}{1-\beta} + \mathbb{E}[Z \mid X]^\top \frac{\beta\eta}{1-\beta} + Z^\top \eta.$$

**Proposition 10.2.2 ([Manski, 1993])** *Suppose that $\beta \neq 1$, $\gamma = \delta = 0$, and $(1, \mathbb{E}[Z \mid X], Z)$ are linearly independent. Then, the composite parameters $(\alpha/(1-\beta), \beta\eta/(1-\beta), \eta)$ are identified.*

The proposition implies that, since $\eta$ is identified, the endogenous effect $\beta$ can be also identified from $\beta\eta/(1-\beta)$ if $\eta \neq 0$.

[1] Even in the presence of the reflection problem, if the value of $(\gamma + \beta\eta)/(1-\beta)$ is non-zero, then either $\gamma$ or $\beta\eta$ (or both) is non-zero. Hence, one can still determine whether "some" social effects exist or not.

## 10.3 Numerical simulation with R

A direct consequence obtained from Theorem 10.2.1 and Proposition 10.2.2 is that when $\gamma \neq 0$ the parameters of interest in model (10.1.1), except for $\eta$, cannot be estimated because $(1, \mathbb{E}[Y \mid X], \mathbb{E}[Z \mid X], Z)$ is linearly dependent; however, if both $\gamma$ and $\delta$ are zero, the parameters can be estimated.

We can check this result with a numerical simulation in **R**. The data-generating process used in the simulation is as follows.

```
N <- 300                                # sample size
X <- c(t(matrix(rnorm(10), 10, 30))) # 10 groups, 30 individuals for each.
Z <- X + X^2 + rnorm(N)                 # E[Z | X] = X + X^2
U <- rnorm(N)                           # no correlated effect

alpha <- 1
beta  <- 0.5
gamma <- 1
eta   <- 1

EZ_X <- X + X^2
EY_X <- (alpha + EZ_X*(gamma + eta)) / (1 - beta)
Y    <- alpha + beta*EY_X + gamma*EZ_X + eta*Z + U
```

The data are comprised of ten social groups with 30 members for each group, and thus, the total sample size is 300. Here, we assume that there is no correlated effect, but the contextual effect is non-zero ($\gamma = 1$). The conditional expectation of $Z$ given $X$ is given by $\mathbb{E}[Z \mid X] = X + X^2$, and that of $Y$ can be computed following equation (10.2.2). Now, with this simulated dataset, if we run a linear regression of $Y$ on $(1, \mathbb{E}[Y \mid X], \mathbb{E}[Z \mid X], Z)$, ...

```
>  lm(Y ~ EY_X + EZ_X + Z)

Call:
lm(formula = Y ~ EY_X + EZ_X + Z)

Coefficients:
(Intercept)         EY_X         EZ_X            Z
         0.5158       0.7568           NA       1.0693
```

we cannot obtain the estimate of $\gamma$ (`NA` stands for "not available"). If we reverse the order of $\mathbb{E}[Y \mid X]$ and $\mathbb{E}[Z \mid X], Z)$,

```
>  lm(Y ~ EZ_X + EY_X + Z)

Call:
lm(formula = Y ~ EZ_X + EY_X + Z)

Coefficients:
(Intercept)         EZ_X         EY_X            Z
          2.029        3.027           NA        1.069
```

now the value of $\beta$ becomes `NA`, implying that the social effects $\beta$ and $\gamma$ cannot be separately identified, i.e., the reflection problem. However, note that the estimate of $\eta$ is unaffected by the order of regressors and is very

close to its true value, that is, $\eta$ can be identified even in the presence of the reflection problem. This result is consistent with Theorem 10.2.1.

Next, using the same dataset, we restrict the value of $\gamma$ to zero, as required in Proposition 10.2.2.

```
>  EY_X <- (alpha + EZ_X*eta) / (1 - beta)
>  Y    <- alpha + beta*EY_X + eta*Z + U
>  lm(Y ~ EY_X + Z)

Call:
lm(formula = Y ~ EY_X + Z)

Coefficients:
(Intercept)          EY_X             Z
         1.0022        0.5137       1.0693
```

Then, we can estimate all parameters in the model correctly.

## 10.4 A game theoretic interpretation

We can interpret the linear-in-means model as a game theoretic model of incomplete information. For simplicity, in this section we assume that contextual effects and correlated effects do not exist: $\gamma = \delta = 0$.

Let $q \in \mathbb{R}$ be an action, and the actual action made by individual $i$ is denoted by $Y_i$. Now, assume that the payoff of individual $i$ choosing $q$ can be written as the following quadratic function:

$$u_i(q, \{Y_j : X_j = X_i\}) = \left[\alpha + Z_i^\top \eta + U_i\right] q - \frac{q^2}{2} + T_i(q, \{Y_j : X_j = X_i\}),$$

where $T_i(q, \{Y_j : X_j = X_i\})$ is a term representing the endogenous interaction between $i$ and the other members of the group to which $i$ belongs. For example, consider

$$T_i(q, \{Y_j : X_j = X_i\}) \equiv \frac{\beta}{|\{j : X_j = X_i\}|} \sum_{j:X_j=X_i} Y_j q, \quad \beta > 0,$$

where $|A|$ denotes the cardinality of the set $A$. This specification presumes that the source of social interactions is "complementarity": the average of group members' outcomes positively affects the individual's own marginal payoff.

Suppose that individual characteristics and error term, $Z_i$ and $U_i$, are unobservable to the other individuals including those in the same group. This implicitly postulates that each social group is sufficiently large such that the members of the group are not identifiable with each other. Then, it is impossible to predict precisely the actual actions $\{Y_j : X_j = X_i\}$ from the point of view of $i$. In this situation, it is conventional to assume that each individual's behavior is characterized by a Bayesian–Nash equilibrium. That is, a rational individual would choose an action to maximize her expected payoff:

$$\begin{aligned}
Y_i &= \operatorname*{argmax}_q \mathbb{E}_i\left[u_i(q, \{Y_j : X_j = X_i\})\right] \\
&= \operatorname*{argmax}_q \left\{\left[\alpha + Z_i^\top \eta + U_i\right] q - \frac{q^2}{2} + \frac{\beta}{|\{j : X_j = X_i\}|} \sum_{j:X_j=X_i} \mathbb{E}_i[Y_j] q\right\} \\
&= \alpha + \frac{\beta}{|\{j : X_j = X_i\}|} \sum_{j:X_j=X_i} \mathbb{E}_i[Y_j] + Z_i^\top \eta + U_i,
\end{aligned}$$

where $\mathbb{E}_i[Y_j]$ is the conditional expectation of $Y_j$ given the information set of $i$. Notice that, because we have assumed that $Z_j$ and $U_j$ are unknown to $i$, they are not included in $i$'s information set. Further, if $(Z, U)$'s are independent among individuals, the only available information to compute $\mathbb{E}_i[Y_j]$ is that $j$ is in the same group as $i$. Consequently, we may have $\mathbb{E}_i[Y_j] = \mathbb{E}[Y|X = X_i]$; this results in the linear-in-means model:

$$Y_i = \alpha + \beta\mathbb{E}[Y|X = X_i] + Z_i^\top \eta + U_i.$$

The linear-in-means model assumes that an individual is affected equally by all the other members of the same group, and that she forms beliefs about their behavior utilizing only the group-level covariates. As suggested in [Lee et al., 2014], such framework is not appropriate for, for example, friendship networks, where the size of each social group is relatively small. When the size of group is small, the values of the individuals characteristics $Z$ of those in the same group would be common knowledge within the members; that is $\mathbb{E}_i[Y_j] = \mathbb{E}[Y \mid Z = Z_j, X = X_i]$. [Lee et al., 2014] called this paradigm of belief formation "heterogeneous rational expectations". Heterogeneous rational expectations imply that the predicted value of the outcome of different members with different individual characteristics will be different. For more details, see their paper.

# Chapter 11

# Social Interactions through Social Networks

## 11.1 Common types of social network graphs

**Social network** is a type of platform that facilitates social interactions between individuals. In social network analysis, networks are typically represented as **graphs**. The followings are examples of social networks that we often observe in real data. *Facebook*: an undirected graph (a mutual consent is required to become Friends), *Twitter*: a directed graph (a mutual consent is not required to be a Follower), *married couples*: a disconnected undirected graphs, *leader and followers*: a directed star graph, *classroom membership*: a complete graph, and so forth.

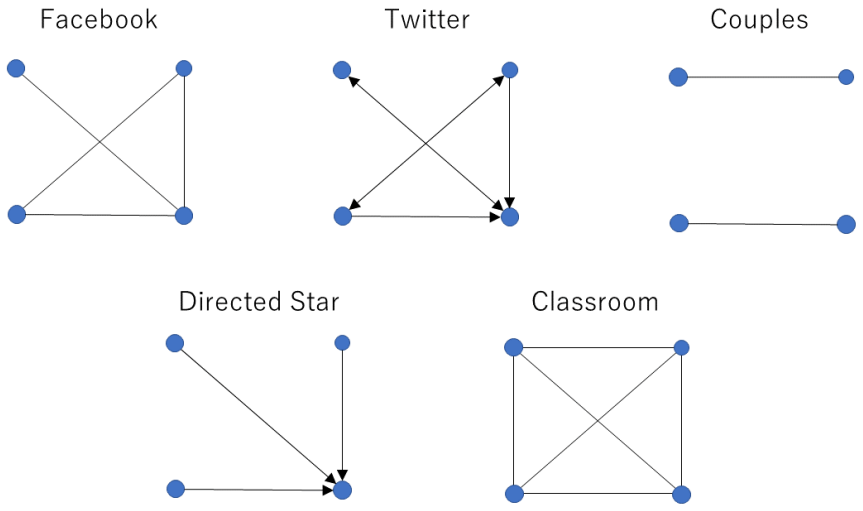



Figure 11.1: Social network graphs

## 11.2 Linear-in-means social network model

When social interactions occur on social networks, such as friendship networks, the reference group for each individual is formed by the agents who are connected to her. Thus, since the members of the reference group are distinguishable in this situation (unless the reference group is extremely large), Manski's linear-in-means model would not be appropriate.

Suppose there is a sample of $n$ individuals that form social networks. Each individual $i$ belongs to one of these groups, and $i$ does not necessarily interact with all the members of the group to which $i$ belongs, but may have a specific reference group (close friends) $P_i$ of size $n(i)$. Individual $i$ is excluded from her own reference group $P_i$, that is, $i \notin P_i$. The reference is either directed or undirected. For both cases, if $j$ affects $i$, we obtain $j \in P_i$. When the friendship is directed, $j \in P_i$ does not imply the converse $i \in P_j$. We say that an individual $i$ is **isolated** if $P_i$ is empty. Note that while an isolated individual is not affected by the other individuals, she can still affect the others if the reference is asymmetric. Throughout this section, we assume that not all individuals are isolated.

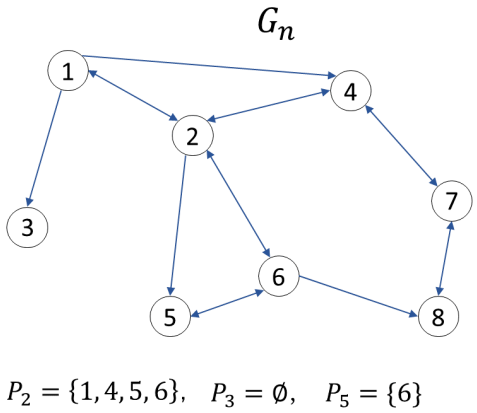



Figure 11.2: Reference groups. The arrows represent friendship nominations.

Let $Y$ be an outcome variable of interest, and $X$ be the vector of individual characteristics. Then, we consider the following linear-in-means social network model ([Bramoullé et al., 2009]):

$$Y_i = \alpha + \beta \frac{\sum_{j \in P_i} Y_j}{n(i)} + X_i^\top \gamma + \frac{\sum_{j \in P_i} X_j^\top}{n(i)} \delta + \epsilon_i, \tag{11.2.1}$$

where $\beta$ captures the endogenous effect and $\delta$ the contextual effect. For simplicity of discussion, we do not consider the correlated effects here.[1] The conditional expectation of $\epsilon$ given $\mathbf{X}_n = (X_1, \ldots, X_n)^\top$ is assumed to be zero: $\mathbb{E}[\epsilon \mid \mathbf{X}_n] = 0$ (i.e., $X$ is exogenous).

Let $\underset{n \times n}{G_n}$ be the weighted social interaction matrix (this is a directed adjacency matrix weighted by $n(i)^{-1}$).[2]

[1] In the linear-in-means social network models as in (11.2.1), one can account for the correlated effects by including *network fixed effects* in the model. For more details, see, for example, [Lee, 2007] and [Bramoullé et al., 2009].

[2] Throughout this chapter, we assume that $G_n$ is non-stochastic. However, this assumption is indeed debatable. Identification and estimation of social interaction models with stochastic and endogenous $G_n$ is an important research topic in the recent literature.

Namely, the $(i,j)$-th element of $G_n$ is given by

$$(G_n)_{i,j} = \begin{cases} 1/n(i) & \text{if } j \in P_i \\ 0 & \text{otherwise} \end{cases}$$

Note that the sum of each $i$'s row of $G_n$ is one, unless $i$ is isolated. We can write model (11.2.1) in matrix form as

$$\mathbf{Y}_n = \alpha \mathbf{1}_n + \beta G_n \mathbf{Y}_n + \mathbf{X}_n \gamma + G_n \mathbf{X}_n \delta + \mathcal{E}_n, \tag{11.2.2}$$

where $\mathbf{1}_n$ is an $n \times 1$ vector of ones, $\mathbf{Y}_n = (Y_1, \ldots, Y_n)^\top$, and $\mathcal{E}_n = (\epsilon_1, \ldots, \epsilon_n)^\top$. Throughout the rest, we assume that $(\mathbf{1}_n, \mathbf{X}_n)$ is linearly independent.

**A game theoretic interpretation** As opposed to Manski's linear-in-means model, the model in (11.2.1) can be seen as a game theoretic model of "complete information". Assume that the payoff of individual $i$ choosing an action $q \in \mathbb{R}$ can be written as the following quadratic function:

$$u_i(q, \{Y_j\}_{j \in P_i}) = \left[ \alpha + X_i^\top \gamma + \frac{\sum_{j \in P_i} X_j^\top}{n(i)} \delta + \epsilon_i \right] q - \frac{q^2}{2} + \beta \frac{\sum_{j \in P_i} Y_j}{n(i)} q, \quad \beta > 0.$$

Here, $Y_j$'s, $j \in P_i$, are observable to $i$ under complete information. This specification presumes that the source of social interactions is **complementarity**: the average of friends' outcomes positively affects the individual's own marginal payoff. Then, based on the payoff-maximization principle, the set of realized outcomes $\{Y_1, \ldots, Y_n\}$ can be characterized by a pure Nash equilibrium. That is,

$$\begin{aligned} Y_i &= \operatorname*{argmax}_{q \in \mathbb{R}} u_i(q, \{Y_j\}_{j \in P_i}) \\ &= \alpha + \beta \frac{\sum_{j \in P_i} Y_j}{n(i)} + X_i^\top \gamma + \frac{\sum_{j \in P_i} X_j^\top}{n(i)} \delta + \epsilon_i \end{aligned}$$

for all $i = 1, \ldots, n$.

For another example, one can consider

$$u_i(q, \{Y_j\}_{j \in P_i}) = \left[ \alpha + X_i^\top \gamma + \frac{\sum_{j \in P_i} X_j^\top}{n(i)} \delta + \epsilon_i \right] q - \frac{q^2}{2} - \frac{\beta}{2} \left( q - \frac{\sum_{j \in P_i} Y_j}{n(i)} \right)^2, \quad \beta > 0.$$

In this case, the source of social interactions is described by **conformity**: $i$'s payoff is positively affected by the degree to which she conforms with her close friends' actions. Solving the payoff-maximization problem, we can find that we obtain the same model as in (11.2.1). This implies that it is generally impossible to identify the source of social interactions from model (11.2.1).

## 11.3 Using network structure to identify social interactions

In the following, assume that $|\beta| < 1$. Then, we can show that the inverse matrix $(I_n - \beta G_n)^{-1}$ exists. For simplicity of presentation, assume that the dimension of $X$ is equal to one: $\dim(X) = 1$. Then, the "reduced-form" of model (11.2.2) can be written as

$$\mathbf{Y}_n = \alpha (I_n - \beta G_n)^{-1} \mathbf{1}_n + \gamma (I_n - \beta G_n)^{-1} \mathbf{X}_n + \delta (I_n - \beta G_n)^{-1} G_n \mathbf{X}_n + (I_n - \beta G_n)^{-1} \mathcal{E}_n$$

$$= \alpha(I_n - \beta G_n)^{-1}\mathbf{1}_n + (I_n - \beta G_n)^{-1}(\gamma I_n + \delta G_n)\mathbf{X}_n + (I_n - \beta G_n)^{-1}\mathcal{E}_n.$$

Note that it follows from the Neumann series expansion (Lemma B.2.2) that

$$\begin{aligned}(I_n - \beta G_n)^{-1} &= I_n + \beta G_n + \beta^2 G_n G_n + \beta^3 G_n G_n G_n + \cdots \\ &= \sum_{t=0}^{\infty} \beta^t G_n^t,\end{aligned}$$

where $G_n^0 = I_n$. Thus, we have

$$(I_n - \beta G_n)^{-1}\mathbf{X}_n = \mathbf{X}_n + \beta G_n \mathbf{X}_n + \beta^2 G_n^2 \mathbf{X}_n + \cdots,$$

where the first term on the right-hand side is the individual's own characteristic variables, the second term represents the average characteristics of her direct friends multiplied by $\beta$, the third term is the average characteristics of friends' friends multiplied by $\beta^2$, and so forth (see also Appendix A.5). This implies that, if $\beta \neq 0$, the characteristics of other individuals in the same network can affect own outcome through the social interactions, even when they are not her direct friends, and, conversely, her own characteristics affect the outcomes of the others in the same network. In this sense, the matrix $(I_n - \beta G_n)^{-1}$ is often called the **social multiplier** matrix.

Here, assume that there are no isolated individuals. In this case, all row-sums of $G_n^t$ for all $t = 0, 1, \ldots$ are equal to one. Therefore, it holds that

$$\alpha(I_n - \beta G_n)^{-1}\mathbf{1}_n = \alpha \sum_{t=0}^{\infty} \beta^t G_n^t \mathbf{1}_n = \alpha \sum_{t=0}^{\infty} \beta^t \mathbf{1}_n = \frac{\alpha}{1-\beta}\mathbf{1}_n. \tag{11.3.1}$$

Now, let $\theta = (\alpha, \beta, \gamma, \delta)$ be the vector of unknown parameters to be estimated. We say that the matrices $I_n$, $G_n$, and $G_n^2$ are linearly independent if and only if

$$\mu_0 I_n + \mu_1 G_n + \mu_2 G_n^2 = 0$$

implies that $\mu_0 = \mu_1 = \mu_2 = 0$.

**Proposition 11.3.1** ([Bramoullé et al., 2009]) *Suppose that $\gamma\beta + \delta \neq 0$.*

*(i) If the matrices $I_n$, $G_n$, and $G_n^2$ are linearly independent, $\theta$ is identified.*

*(ii) If the matrices $I_n$, $G_n$, and $G_n^2$ are linearly dependent and there are no isolated individuals, $\theta$ is not identified.*

**Proof.** (i) Consider two parameter vectors $\theta = (\alpha, \beta, \gamma, \delta)$ and $\theta' = (\alpha', \beta', \gamma', \delta')$. We show that

$$\alpha(I_n - \beta G_n)^{-1}\mathbf{1}_n + (I_n - \beta G_n)^{-1}(\gamma I_n + \delta G_n)\mathbf{X}_n = \alpha'(I_n - \beta' G_n)^{-1}\mathbf{1}_n + (I_n - \beta' G_n)^{-1}(\gamma' I_n + \delta' G_n)\mathbf{X}_n \tag{11.3.2}$$

implies $\theta = \theta'$. Note that the above equality implies that

$$\begin{aligned}&\alpha(I_n - \beta G_n)^{-1}\mathbf{1}_n = \alpha'(I_n - \beta' G_n)^{-1}\mathbf{1}_n \\ &(I_n - \beta G_n)^{-1}(\gamma I_n + \delta G_n)\mathbf{X}_n = (I_n - \beta' G_n)^{-1}(\gamma' I_n + \delta' G_n)\mathbf{X}_n\end{aligned}$$

by the linear independence of $(\mathbf{1}_n, \mathbf{X}_n)$. Multiplying the both sides of the second equality by $I_n - \beta G_n$ yields

$$\begin{aligned}\gamma I_n + \delta G_n &= (I_n - \beta G_n)(I_n - \beta' G_n)^{-1}(\gamma' I_n + \delta' G_n) \\ &= (I_n - \beta' G_n)^{-1}(\gamma' I_n + \delta' G_n) - \beta G_n (I_n - \beta' G_n)^{-1}(\gamma' I_n + \delta' G_n).\end{aligned}$$

Noting that $G_n(I_n - \beta' G_n)^{-1} = (I_n - \beta' G_n)^{-1} G_n$, further multiplying the both sides by $I_n - \beta' G_n$,

$$\begin{aligned}&\gamma I_n + \delta G_n - \gamma\beta' G_n - \delta\beta' G_n^2 = \gamma' I_n + \delta' G_n - \gamma'\beta G_n - \delta'\beta G_n^2 \\ &\Longrightarrow (\gamma - \gamma')I_n + (\delta - \delta' + \gamma'\beta - \gamma\beta')G_n + (\delta'\beta - \delta\beta')G_n^2 = 0. \qquad (11.3.3)\end{aligned}$$

Hence, we obtain $\gamma = \gamma'$, $\delta + \gamma'\beta = \delta' + \gamma\beta'$ and $\delta'\beta = \delta\beta'$. Let $\lambda = \beta'/\beta$ such that $\beta' = \lambda\beta$ and $\delta' = \lambda\delta$. Then,

$$\begin{aligned}\gamma = \gamma',\ \ \delta + \gamma'\beta = \delta' + \gamma\beta' \quad &\Longrightarrow \quad \delta + \gamma\beta = \delta' + \gamma\beta' \\ &\Longrightarrow \quad \delta + \gamma\beta = \lambda(\delta + \gamma\beta) \\ &\Longrightarrow \quad \lambda = 1 \quad (\because \gamma\beta + \delta \neq 0) \quad \Longrightarrow \quad \beta = \beta',\ \ \delta = \delta'.\end{aligned}$$

Finally, $\alpha = \alpha'$ follows from $\alpha(I_n - \beta G_n)^{-1}\mathbf{1}_n = \alpha'(I_n - \beta G_n)^{-1}\mathbf{1}_n$.

(ii) When the matrices $I_n$, $G_n$, and $G_n^2$ are linearly dependent, we can find $\theta' = (\alpha', \beta', \gamma', \delta')$ that satisfy (11.3.2) and (11.3.3) but at least one of the following inequalities holds:

$$\gamma - \gamma' \neq 0,\ \ \delta - \delta' + \gamma'\beta - \gamma\beta' \neq 0,\ \ \delta'\beta - \delta\beta' \neq 0.$$

This implies that at least $\beta$ and $\delta$ cannot be separately identified. For the identification of $\alpha$, because no individual is isolated,

$$\alpha(I_n - \beta G_n)^{-1}\mathbf{1}_n = \alpha'(I_n - \beta' G_n)^{-1}\mathbf{1}_n \iff \alpha/(1-\beta) = \alpha'/(1-\beta')$$

by (11.3.1). Since $\beta$ is not identified, $\alpha$ is also not identified. ■

In Proposition 11.3.1, we have assumed that $\gamma\beta + \delta \neq 0$. The role of this assumption should be clear from the following equality. Noting that $(I_n - \beta G_n)^{-1} = I_n + \beta G_n (I_n - \beta G_n)^{-1}$, if $\gamma\beta + \delta = 0$, we have

$$\begin{aligned}\mathbb{E}[\mathbf{Y}_n \mid \mathbf{X}_n] &= \alpha(I_n - \beta G_n)^{-1}\mathbf{1}_n + \gamma(I_n - \beta G_n)^{-1}\mathbf{X}_n + \delta(I_n - \beta G_n)^{-1} G_n \mathbf{X}_n \\ &= \alpha(I_n - \beta G_n)^{-1}\mathbf{1}_n + \mathbf{X}_n\gamma + (\gamma\beta + \delta)(I_n - \beta G_n)^{-1} G_n \mathbf{X}_n \\ &= \alpha(I_n - \beta G_n)^{-1}\mathbf{1}_n + \mathbf{X}_n\gamma.\end{aligned}$$

Thus, the contextual effect vanishes when $\gamma\beta + \delta = 0$.

The next result shows that if $\mathbb{E}[G_n\mathbf{Y}_n \mid \mathbf{X}_n]$ is a linear function of $(\mathbf{1}_n, \mathbf{X}_n, G_n\mathbf{X}_n)$, the matrices $I_n$, $G_n$, and $G_n^2$ become linearly dependent, and thus $\theta$ is not identified. This result corresponds to the reflection problem of Manski's model (recall equation (10.2.2) and Theorem 10.2.1).

**Proposition 11.3.2** ([Bramoullé et al., 2009]) *Suppose that $\gamma\beta + \delta \neq 0$ and that no individuals are isolated. If there are $(\lambda_0, \lambda_1, \lambda_2)$ satisfying $\mathbb{E}[G_n\mathbf{Y}_n \mid \mathbf{X}_n] = \lambda_0\mathbf{1}_n + \lambda_1\mathbf{X}_n + \lambda_2 G_n\mathbf{X}_n$ for any $\mathbf{X}_n$, then $I_n$, $G_n$ and $G_n^2$ are linearly dependent.*

**Proof.** First, by assumption

$$\begin{aligned}\mathbb{E}[G_n\mathbf{Y}_n \mid \mathbf{X}_n] &= \alpha/(1-\beta)\mathbf{1}_n + (I_n - \beta G_n)^{-1}(\gamma G_n + \delta G_n^2)\mathbf{X}_n \\ &= \lambda_0\mathbf{1}_n + \lambda_1\mathbf{X}_n + \lambda_2 G_n\mathbf{X}_n\end{aligned}$$

for any $\mathbf{X}_n$. This implies that $\lambda_0 = \alpha/(1-\beta)$ and $(I_n - \beta G_n)^{-1}(\gamma G_n + \delta G_n^2) = \lambda_1 I_n + \lambda_2 G_n$. For the second equality, by multiplying the both sides by $I_n - \beta G_n$ yields

$$\begin{aligned}&\gamma G_n + \delta G_n^2 = \lambda_1 I_n + \lambda_2 G_n - \beta\lambda_1 G_n - \beta\lambda_2 G_n^2 \\ \Longrightarrow\ &\lambda_1 I_n + (\lambda_2 - \beta\lambda_1 - \gamma)G_n - (\beta\lambda_2 + \delta)G_n^2 = 0.\end{aligned}$$

Here assume that $I_n$, $G_n$, and $G_n^2$ are linearly independent. Then, it must hold that $\lambda_1 = 0$ and $\lambda_2 = \gamma$, and hence $\gamma\beta + \delta = 0$. This is a contradiction with the assumption that $\gamma\beta + \delta \neq 0$. Therefore, $I_n$, $G_n$, and $G_n^2$ are linearly dependent. ■

**Example 11.3.3 (Classroom interaction: single classroom)** Consider a school classroom of size $n$, where each student is interacted with all the other $n-1$ students in the same classroom. In this case, the social interaction matrix $G_n$ is characterized by a weighted complete graph with the weights being all equal to $1/(n-1)$. A typical element of $G_n^2$ is given by

$$(G_n^2)_{i,j} = \begin{cases} \sum_{k=1}^n (G_n)_{i,k}(G_n)_{k,j} = (n-2)/(n-1)^2 & \text{if } i \neq j \\ \sum_{k=1}^n (G_n)_{i,k}(G_n)_{k,j} = 1/(n-1) & \text{if } i = j \end{cases}$$

Therefore, we have

$$\frac{1}{n-1}I_n + \frac{n-2}{n-1}G_n = G_n^2,$$

and thus $\theta$ cannot be identified.

**Example 11.3.4 (Classroom interaction: multiple classrooms)** Now assume that (without loss of generality) there are two classrooms of sizes $n_1$ and $n_2$. In this case, the social interaction matrix $G_n$ is a block-diagonal matrix

$$G_n = \begin{pmatrix} G_{n1} & 0 \\ 0 & G_{n2} \end{pmatrix},$$

where each $G_{nj}$ is a weighted complete graph with weights equal to $1/(n_j-1)$, for $j = 1, 2$. By the same argument as above, we can see that

$$G_n^2 = \begin{pmatrix} \frac{1}{n_1-1}I_{n_1} + \frac{n_1-2}{n_1-1}G_{n1} & 0 \\ 0 & \frac{1}{n_2-1}I_{n_2} + \frac{n_2-2}{n_2-1}G_{n2} \end{pmatrix}.$$

Therefore, if $n_1 \neq n_2$, we can identify $\theta$. In words, the variations in classroom sizes have identification power; see [Lee, 2007].

**Example 11.3.5 (Intransitive network)** Suppose now that individuals interact through a network. In addition, suppose that we can find an **intransitive triad** (i.e., a friend of my friend is not necessarily my friend) in the network. For example, suppose that there is a set of three individuals $(i, j, k)$ satisfying $(G_n)_{i,j} > 0$, $(G_n)_{j,k} > 0$, but $(G_n)_{i,k} = 0$, such that the shortest path between $i$ and $k$ is of length 2. In this case, the $(i, k)$-th element of $G_n^2$ is larger than zero, while that of $G_n$ is zero (and of course that of $I_n$ is also zero). Therefore, the presence of the intransitive triad guarantees the linear independence.

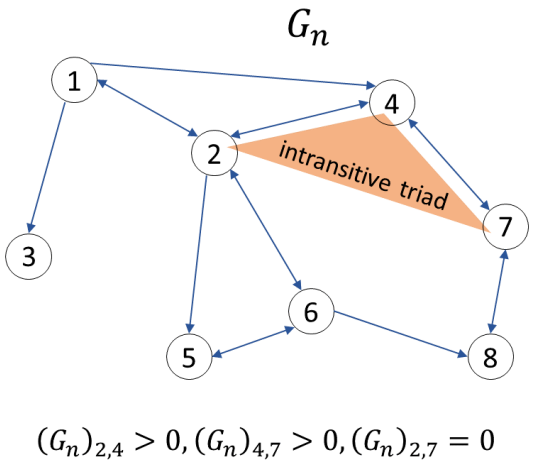



Figure 11.3: Intransitive triad

---

**Exercise 11.3.6** Prove that if $|\beta| < 1$ the matrix $I_n - \beta G_n$ is nonsingular. (Hint: use Lemma B.2.2)

**Exercise 11.3.7** Prove that all row-sums of $G_n^t$ for all $t = 0, 1, \ldots$ are one when there are no isolated individuals.

## 11.4 Estimation

Let us recall that the model to be estimated is (11.2.2):

$$\mathbf{Y}_n = \alpha_0 \mathbf{1}_n + \beta_0 G_n \mathbf{Y}_n + \mathbf{X}_n \gamma_0 + G_n \mathbf{X}_n \delta_0 + \mathcal{E}_n,$$

where $\theta_0$ is the true parameter value of $\theta = (\alpha, \beta, \gamma, \delta)$. Given the identification results as above, we here provide an approach to estimate $\theta_0$ In this section, we do not restrict the dimension of $X$ to be one.

First of all, it should be clear that the OLS estimator that simply regresses $\mathbf{Y}_n$ on $(\mathbf{1}_n, G_n \mathbf{Y}_n, \mathbf{X}_n, G_n \mathbf{X}_n)$ is inconsistent because of the endogeneity of $G_n \mathbf{Y}_n$; if $i$ and $j$ are friends such that $i \in P_j$,

$$\begin{aligned}
\mathbb{E}[Y_j \epsilon_i] &= \mathbb{E}\left[\left(\alpha + \frac{\beta_0}{n(j)} \sum_{k \in P_j} Y_k + X_j^\top \gamma_0 + \frac{1}{n(j)} \sum_{k \in P_j} X_k^\top \delta_0 + \epsilon_j\right) \epsilon_i\right] \\
&= \frac{\beta_0}{n(j)} \underbrace{\mathbb{E}\left[\sum_{k \in P_j} Y_k \epsilon_i\right]}_{\neq 0} + \mathbb{E}[\epsilon_j \epsilon_i].
\end{aligned}$$

Thus, even when $\epsilon_i$ and $\epsilon_j$ are uncorrelated, unless $\beta_0 = 0$, $\mathbb{E}[Y_j \epsilon_i] \neq 0$ in general.

To account for this endogeneity problem, we can use a 2SLS estimator with instrumental variables. Fortunately, a set of reasonable instrumental variables for $G_n \mathbf{Y}_n$ can be easily found. Observe that

$$G_n \mathbf{Y}_n = \alpha_0 G_n \mathbf{1}_n + \beta_0 G_n^2 \mathbf{Y}_n + G_n \mathbf{X}_n \gamma_0 + G_n^2 \mathbf{X}_n \delta_0 + G_n \mathcal{E}_n.$$

This expression implies that $G_n^2\mathbf{X}_n$ is a valid instrument for $G_n\mathbf{Y}_n$, which is not included in the outcome equation (11.2.2), and is correlated with $G_n\mathbf{Y}_n$ (assuming that $\delta_0 \neq 0$) but not with $\mathcal{E}_n$.[3]

The 2SLS estimation procedure is as follows. First, run a least squares regression of $G_n\mathbf{Y}_n$ on $\mathbf{Z}_n = (\mathbf{1}_n, \mathbf{X}_n, G_n\mathbf{X}_n, G_n^2\mathbf{X}_n)$, and compute the predicted value of $G_n\mathbf{Y}_n$, say $\widehat{G_n\mathbf{Y}_n}$, by

$$\widehat{G_n\mathbf{Y}_n} = \mathbf{Z}_n(\mathbf{Z}_n^\top\mathbf{Z}_n)^{-1}\mathbf{Z}_n^\top G_n\mathbf{Y}_n.$$

In the second stage, run a least squares regression of $\mathbf{Y}_n$ on $(\mathbf{1}_n, \widehat{G_n\mathbf{Y}_n}, \mathbf{X}_n, G_n\mathbf{X}_n)$ to obtain the estimator $\widehat{\theta}_n$ of $\theta$. Note that since $(\mathbf{1}_n, \mathbf{X}_n, G_n\mathbf{X}_n)$ is a subset of $\mathbf{Z}_n$, it holds that $(\mathbf{1}_n, \mathbf{X}_n, G_n\mathbf{X}_n) = \mathbf{Z}_n(\mathbf{Z}_n^\top\mathbf{Z}_n)^{-1}\mathbf{Z}_n^\top(\mathbf{1}_n, \mathbf{X}_n, G_n\mathbf{X}_n)$, and the second-stage regression is numerically equivalent to regressing $\mathbf{Y}_n$ on $\mathbf{Z}_n(\mathbf{Z}_n^\top\mathbf{Z}_n)^{-1}\mathbf{Z}_n^\top\mathbf{H}_n$, where $\mathbf{H}_n = (\mathbf{1}_n, G_n\mathbf{Y}_n, \mathbf{X}_n, G_n\mathbf{X}_n)$. Hence, the 2SLS estimator $\widehat{\theta}_n$ can be obtained actually in one step by

$$\widehat{\theta}_n = \left[\mathbf{H}_n^\top\mathbf{Z}_n(\mathbf{Z}_n^\top\mathbf{Z}_n)^{-1}\mathbf{Z}_n^\top\mathbf{H}_n\right]^{-1}\mathbf{H}_n^\top\mathbf{Z}_n(\mathbf{Z}_n^\top\mathbf{Z}_n)^{-1}\mathbf{Z}_n^\top\mathbf{Y}_n. \tag{11.4.1}$$

Under some regularity conditions, the 2SLS estimator $\widehat{\theta}_n$ is consistent for $\theta_0$ and asymptotically normally distributed at $\sqrt{n}$ rate; see Section 2.4.

[3]Note that if the matrices $I_n$, $G_n$, and $G_n^2$ are linearly dependent, $G_n^2\mathbf{X}_n$ is perfectly collinear with $(\mathbf{X}_n, G_n\mathbf{X}_n)$, and thus $G_n^2\mathbf{X}_n$ cannot be used as an identifying instrument. From this, one can view the identification condition in Proposition 11.3.1 that $I_n$, $G_n$, and $G_n^2$ are linearly independent as a condition ensuring that $G_n^2\mathbf{X}_n$ becomes a valid instrument for $G_n\mathbf{Y}_n$.

# Chapter 12

# Spatial Data

## 12.1 Spatial data

**Spatial data** are data that are related to geographic information as part of that data. Geographic information refers to all kind of data that identify the spatial location of each data unit, which is stored not only in the form of point data (i.e., longitude and latitude), but also in the form of spatial polygon data (e.g., regions, districts, and municipalities) and mesh data (e.g., the distribution of air pollutants, satellite images, and brain images).

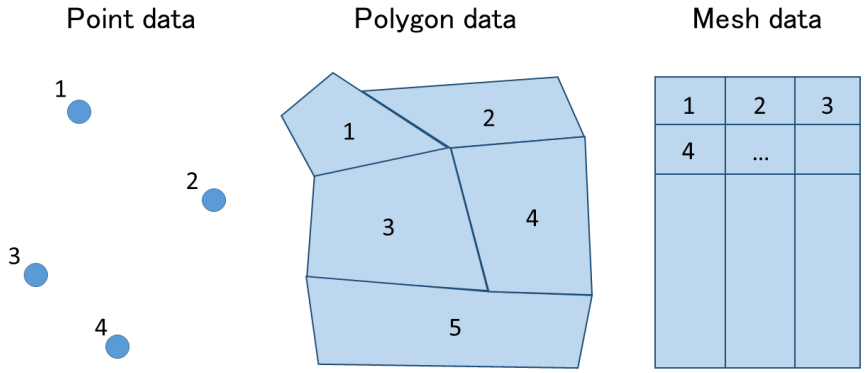



Figure 12.1: Spatial data

An integrated computational system, which is designed to store, manipulate, and visualize spatial data, is called **GIS** (geographic information system). The world's most popular GIS software is ArcGIS (ESRI). In these days, free and open source GIS softwares, such as QGIS and Grass GIS, have been becoming more popular. The statistical software **R** is also able to process and analyze spatial information with additional packages (such as **sf**, **spdep**, etc).


Figure 12.2: Open-source free GIS

## 12.2 Moran's I and Geary's C

### 12.2.1 Spatial weight matrix

The first step in any spatial data analysis is to plot the data on a map to find out if there is a specific correlational pattern in the data. Figure 9.3 and 9.4 are examples of plotted spatial data (polygon data). In these examples, we can find a tendency that districts of similar values are spatially clustered. This phenomenon is called **spatial autocorrelation**, more precisely speaking, a "positive" spatial autocorrelation. Negative spatial autocorrelation means that nearby districts have dissimilar values. Although negative spatial autocorrelation is rarely observed in real data, a spatial competition under strategic substitutes potentially may generate a negative correlation (e.g., where to open convenience stores).

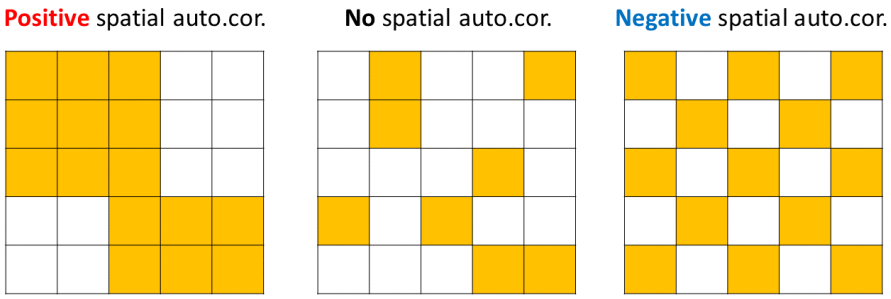



Figure 12.3: Spatial autocorrelation

The main objectives of spatial data analysis are to quantify and to analyze the mechanism of spatial autocorrelation in the data. In measuring the degree of spatial autocorrelation, one thing that needs clarification is what "nearby" means. How to measure the proximity between two spatial units depends on the form of spatial data. For point data, the proximity can be defined simply by the Euclidean distance (or, possibly other distance measures, such as road distance and travel time). For polygon data, the proximity is often determined by whether the units share a common border or not. For this type of proximity measure, there are two common approaches: **Rook contiguity** and **Queen contiguity** (based on the movement of chess pieces). The rook contiguity defines neighbors by the existence of a common edge between two spatial units, and the queen contiguity is based on the existence of either a common edge or vertex between them (see Figure 12.4). Note that, if one specifies a

representative point of each polygon (e.g., location of administrative office, center of gravity), the distance-based proximity can be used for polygon data as well.

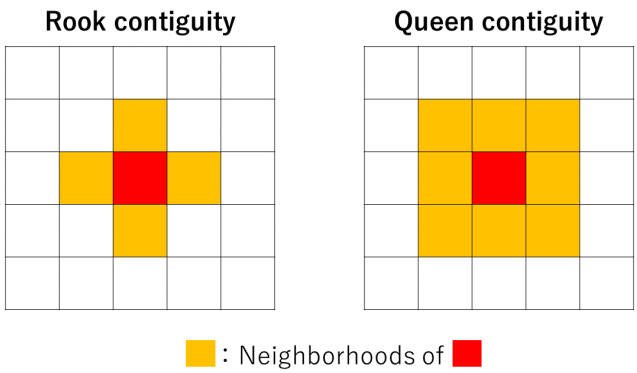



Figure 12.4: Rook contiguity and Queen contiguity

Now, consider a variable $X$, and suppose that we have sample data $\{X_1, \ldots, X_n\}$ of size $n$, where the subscript denotes each spatial unit. Further, let $w_{ij}$ $(1 \le i, j \le n)$ be positive constants satisfying

(i) $w_{ii} = 0$ for all $i$

(ii) $w_{ij} \ge 0$ if $i$ and $j$ are close

(iii) $w_{ij} = 0$ if $i$ and $j$ are distant enough

(iv) $\sum_{j=1}^{n} w_{ij} = 1$ for all $i$.

Such $w_{ij}$ is called the **spatial weight** between $i$ and $j$. Using the spatial weight, we can define the weighted "neighborhood average" of $X$ around unit $i$ by

$$X_i^* \equiv \sum_{j=1}^{n} w_{ij} X_j.$$

Note that $i$ is not included in its own neighborhood ($w_{ii} = 0$).

## Examples of spatial weights

**Polygon data** In the case of polygon data, a commonly used spatial weight is as follows:

$$w_{ij} = \frac{1}{\#\text{ of adjacent districts to } i} \quad \text{if } j \text{ is adjacent to } i$$

$$w_{ij} = 0 \quad \text{if } j \text{ is not adjacent to } i.$$

With this spatial weight, $X_i^*$ can be interpreted as the average of $X$ over the districts that are adjacent to $i$. Note that the requirements $w_{ii} = 0$ and $\sum_{j=1}^{n} w_{ij} = 1$ for all $i$ means that there are no isolated districts, such as islands. The latter condition is introduced for expositional simplicity, and can be relaxed in practice.

**Point data** In the case of point data, we first need to calculate the distance between every pair of spatial units, say $d(i,j)$. Define,

$$v_{ij} = \frac{1}{d(i,j)^r} \quad \text{if } d(i,j) \le q \text{ and } j \neq i$$

$$v_{ij} = 0 \quad \text{otherwise}$$

where $r$ is some positive number, which is typically 1 or 2, and $q$ is a pre-specified threshold value such that the objects with distance larger than $q$ are assumed to be independent. Note that the $v_{ij}$'s defined above do not generally satisfy $\sum_{j=1}^n v_{ij} = 1$. Thus, we normalize them in the following way:

$$w_{ij} = \frac{v_{ij}}{\sum_{j'=1}^n v_{ij'}}.$$

This type of spatial weight is often called the distance-based spatial weight.

An $n \times n$ matrix $W_n$ whose $(i,j)$-th entry is given by $w_{ij}$ is called the **spatial weight matrix**: namely,

$$W_n = \begin{pmatrix} 0 & w_{12} & \cdots & w_{1n} \\ w_{21} & 0 & \cdots & w_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n1} & w_{n2} & \cdots & 0 \end{pmatrix}$$

One can view that the spatial weight matrix $W_n$ is a spatial version of the weighted social interaction matrix $G_n$. Note that $W_n$ is not necessarily a symmetric matrix. Using the spatial weight matrix, we can simply write $(X_1^*, \ldots, X_n^*)^\top = W_n \mathbf{X}_n$, where $\mathbf{X}_n = (X_1, \ldots, X_n)^\top$.

Now, we are ready to more formally describe the definition of spatial autocorrelation. We say that there exists a positive (resp. negative) spatial autocorrelation in the data if $\mathbf{X}_n$ and $W_n\mathbf{X}_n$ are positively (resp. negatively) correlated. As one may notice, the determination of the existence of spatial autocorrelation is dependent and sensitive to the (somewhat arbitrarily) chosen spatial weights matrix $W_n$.

### 12.2.2 Moran's I and Geary's C

There are two major statistics to measure the magnitude of spatial autocorrelation. The one is **Moran's I** statistic, and the other is **Geary's C** statistic, which are defined by

$$I \equiv \frac{\sum_{i=1}^n \sum_{j=1}^n w_{ij}(X_i - \overline{X}_n)(X_j - \overline{X}_n)}{\sum_{i=1}^n (X_i - \overline{X}_n)^2}$$

and

$$C \equiv \frac{n-1}{2n} \frac{\sum_{i=1}^n \sum_{j=1}^n w_{ij}(X_i - X_j)^2}{\sum_{i=1}^n (X_i - \overline{X}_n)^2}$$

respectively, where $\overline{X}_n = n^{-1} \sum_{i=1}^n X_i$. Since Moran's I is a spatial extension of the standard (i.e., Pearson's) correlation coefficient between $\mathbf{X}_n$ and $W_n\mathbf{X}_n$, $I$ takes the value on the range $[-1, 1]$:

| | |
|---|---|
| $0 < I \le 1$ | Positive spatial autocorrelation |
| $I = 0$ | No spatial autocorrelation |
| $-1 \le I < 0$ | Negative spatial autocorrelation. |

On the other hand, Geary's C ranges in value from 0 to 2, and the degree of spatial autocorrelation decreases as $C$ increases:

$$\begin{aligned} 1 < C \le 2 &\quad \text{Negative spatial autocorrelation} \\ C = 1 &\quad \text{No spatial autocorrelation} \\ 0 \le C < 1 &\quad \text{Positive spatial autocorrelation.} \end{aligned}$$

Figure 12.5 provides numerical examples of Moran's I and Geary's C statistic based on the rook contiguity-based spatial weight.

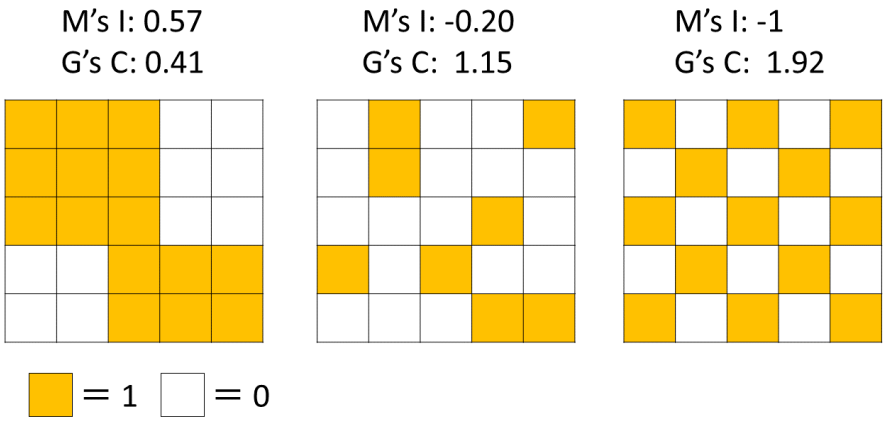



Figure 12.5: Numerical examples: Moran's I and Geary's C

## 12.3 Spatial random variables

### 12.3.1 Spatial stochastic process

In applications, it is mostly the case that $X$ is a random variable rather than a non-stochastic variable. Suppose that each realized value of $X$ is uniquely characterized by its location $s$ in $\mathbb{R}^2$, and we may write $X \equiv X(s)$. The collection of random variables $\{X(s) : s \in \mathcal{S}_n\}$ is called as a **spatial stochastic process** (random field) with the sampling region $\mathcal{S}_n \subseteq \mathbb{R}^2$. The sampling region $\mathcal{S}_n$ can be dependent on the sample size $n$, and it is usually required that $\mathcal{S}_n$ expands as $n$ increases in each direction in $\mathbb{R}^2$, as stated below.

If $X(s)$ is independent of $X(s')$ for all $s \neq s'$, a random sample $\{X_1, \ldots, X_n\}$ of $X$ from $n$ distinct locations $\{s_i \in \mathcal{S}_n : 1 \le i \le n\}$, where $X_i \equiv X(s_i)$, can be virtually treated as the standard non-spatial data. In this case, letting $\sigma^2(s) = \mathrm{Var}(X(s)) < \infty$, we have

$$\mathrm{Var}\left(\overline{X}_n\right) = \frac{1}{n}\overline{\sigma}_n^2 \to 0,$$

as $n \to \infty$, where $\overline{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^n \sigma^2(s_i)$. Thus, by Chebyshev's inequality (1.2.2), for any positive constant $\kappa > 0$,

$$\Pr\left(\left|\overline{X}_n - \mathbb{E}\overline{X}_n\right| \ge \kappa\right) \le \frac{\mathrm{Var}(\overline{X}_n)}{\kappa^2} \to 0,$$

implying that the WLLN holds: $\overline{X}_n \xrightarrow{p} \mathbb{E}\overline{X}_n$. Similarly, it can be straightforwardly verified that the CLT also holds (under some additional regularity conditions).

However, the above argument is not realistic for spatial data in that $\{X_1, \ldots, X_n\}$ are independent. It is clear that if the dependence between the variables is strong such that $\mathrm{Var}\left(\overline{X}_n\right)$ does not converge to zero, standard large sample theory would not be applicable. In contrast, even when they are dependent, it is possible to show that the law of large numbers and the central limit theorem hold only if the degree of dependence is sufficiently weak.

Here, for simplicity, assume that the data points are located on a regular lattice in $\mathbb{R}^2$, and that $\mathrm{Cov}(X(s), X(s')) \equiv C(s, s') = \rho$, where $0 < \rho < \infty$, if $(s, s')$ are neighboring grid-points, and $C(s, s') = 0$ otherwise.

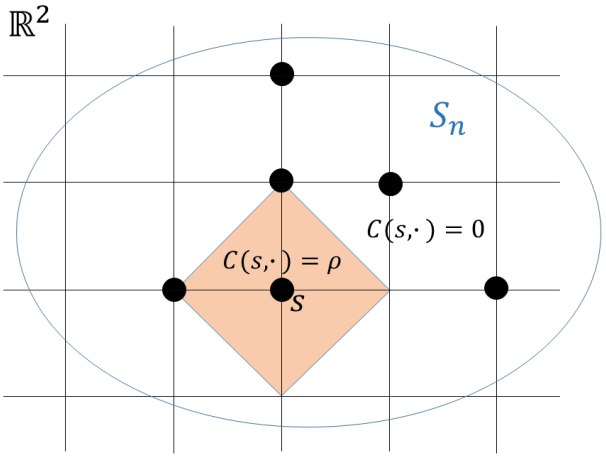



Figure 12.6: A type of weak spatial dependence

The assumption that the covariance between the variables at two different locations depends only on the distance between them is called **isotropy** (i.e., directions do not matter).

Then, noting that there are at most four neighboring data points for each $s$, we have

$$\mathrm{Var}(\overline{X}_n) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2(s_i) + \frac{1}{n^2}\sum_{i=1}^{n}\underbrace{\sum_{j\neq i}^{n} C(s_i, s_j)}_{\leq 4\rho}$$

$$\leq \frac{1}{n}\overline{\sigma}_n^2 + \frac{1}{n}4\rho \to 0$$

as $n \to \infty$. Therefore, the WLLN holds, and so does CLT with some weak additional conditions. Although the spatial dependence structure introduced in this example is too simplistic for real data, more realistic dependence structures are proposed in the literature, for example, *spatial mixing* and *spatial near epoch dependence*, which are both natural spatial extensions of the dependence concept used in time series analysis.

### 12.3.2 Spatial sampling

As mentioned above, if the degree of spatial dependence is sufficiently weak, one can rely on standard large sample theory. It is important to note that such assumption implicitly requires that the sampling region must be

(weakly) expanding: $\mathcal{S}_n \subseteq \mathcal{S}_{n+1} \subseteq \cdots \subseteq \mathbb{R}^2$, with the minimum distance between observations being fixed away from zero. Large sample theory based on this type of data sampling framework is called **increasing domain asymptotics**.

On the other hand, suppose that the sampling region is fixed and bounded: $\mathcal{S}_n = \mathcal{S}_{n+1} = \cdots \subset \mathbb{R}^2$. In this case, as the sample size $n$ increases, the observations get denser and denser. Consequently, for a given data point $s$, the number of observations that are spatially dependent of $s$ can increase to infinity as $n$ increases. This case is called **infill asymptotics** (or fixed domain asymptotics). As can be easily imagined, asymptotic theory under the infill asymptotics is much more involved than the one under the increasing domain asymptotics. The infill asymptotic theory has been an important research field particularly in the literature on high frequency time series.

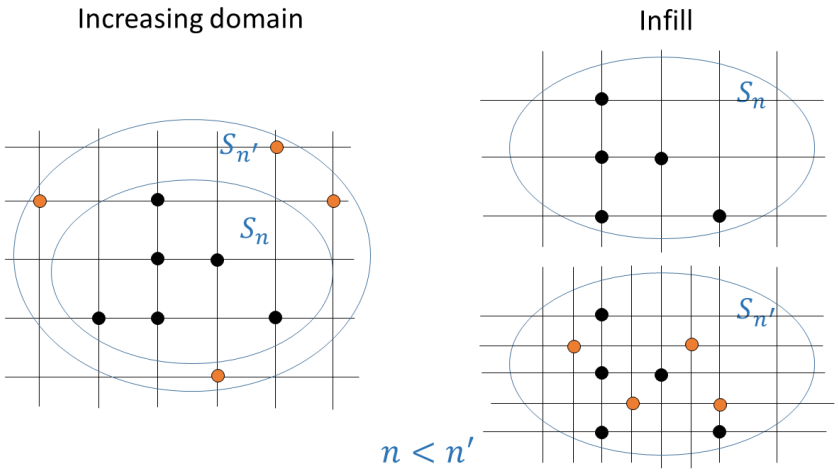



Figure 12.7: Increasing domain asymptotics and infill asymptotics

# Chapter 13

# Spatial Econometrics

Spatial econometrics is a subfield of econometrics that deals with spatial interaction (spatial autocorrelation) across spatial units. Spatial econometric tools have become popular as empirical data analysis techniques for a range of diverse topics, including local economic growth, property markets, local competition between municipalities, local crime, and so forth. There are two most commonly used spatial econometric models: the **spatial lag model** (or spatial autoregressive model) and the **spatial error model**. In terms of the three types of social interaction effects classified by [Manski, 1993], the spatial lag model is a model that accounts for possible endogeneity of spatial interaction, and the spatial error model accounts for the spatially correlated effects.

## 13.1 Spatial lag model

Let $Y$ be a dependent variable of interest, and $X$ be the vector of explanatory variables including a constant term. Suppose that a sample $\{(Y_i, X_i) : 1 \leq i \leq n\}$ of size $n$ is observed. Further, let $w_{ij}$ be a spatial weight term between $i$ and $j$. Recall that $w_{ii} = 0$ for normalization. Then, the following model is called the **spatial lag model** (SLM):

$$Y_i = \rho_0 \sum_{j=1}^{n} w_{ij} Y_j + X_i^\top \beta_0 + \varepsilon_i, \quad i = 1, \ldots, n. \tag{13.1.1}$$

Here, $\rho_0$ is a parameter that captures the spatial endogenous effect, $\beta_0$ is a vector of unknown coefficients, and $\varepsilon$ is an unobserved random variable such that $\mathbb{E}\varepsilon = 0$. The SLM assumes that the average outcome of the neighboring districts influences on own outcome.

For example, imagine that there are two districts, which are almost identical in terms of sociodemographic and economic characteristics. Suppose that one of them is surrounded by areas with higher crime rates, and the other is located in a low crime rate area. Then, if the crime rate of the former district is higher than that of the latter, this suggests the existence of spatial endogenous effect, and the spatial parameter $\rho_0$ will be estimated as positive.

Let $W_n = (w_{ij})_{i,j=1}^n$ be the spatial weight matrix, and write $\mathbf{Y}_n = (Y_1, \ldots, Y_n)^\top$, $\mathbf{X}_n = (X_1, \ldots, X_n)^\top$, and $\mathcal{E}_n = (\varepsilon_1, ...., \varepsilon_n)^\top$. Then, we can rewrite model (13.1.1) in matrix form as

$$\mathbf{Y}_n = \rho_0 W_n \mathbf{Y}_n + \mathbf{X}_n \beta_0 + \mathcal{E}_n.$$

As one can see, the form of the SLM is quite similar to that of the social interaction model in (11.2.2). Indeed, the SLM can be interpreted in the same manner as in the social interaction model. That is, the SLM can be viewed as a game model of complete information with the realized outcomes $\{Y_1, \ldots, Y_n\}$ being a result of Nash equilibrium behavior.[1] In addition, if $|\rho_0| < 1$ and the spatial weight matrix is normalized such that each row sums to unity, the matrix $I_n - \rho_0 W_n$ is nonsingular, and thus the inverse matrix $(I_n - \rho_0 W_n)^{-1}$ exists. Thus, the reduced-form of the SLM can be obtained by

$$\mathbf{Y}_n = (I_n - \rho_0 W_n)^{-1} \mathbf{X}_n \beta_0 + (I_n - \rho_0 W_n)^{-1} \mathcal{E}_n.$$

By Neumann series expansion (Appendix B.2), we have

$$(I_n - \rho_0 W_n)^{-1} \mathbf{X}_n \beta_0 = \mathbf{X}_n \beta_0 + \rho_0 W_n \mathbf{X}_n \beta_0 + \rho_0^2 W_n^2 \mathbf{X}_n \beta_0 + \cdots.$$

Similarly to the social interaction model, the second term on the right-hand side represents the effect of the neighbors' characteristics, and the third term is the effect from the neighbors' neighbors' characteristics, and so forth. This implies that a marginal increase in $X$ affects not only own outcome $Y$ but the outcomes of the others, and vice versa. In the context of spatial econometrics, the matrix $(I_n - \rho_0 W_n)^{-1}$ is called the **spatial multiplier** matrix.

### 13.1.1 Maximum likelihood estimation

For the same reason mentioned in Section 11.4, a simple OLS estimator that regresses $\mathbf{Y}_n$ on $(W_n \mathbf{Y}_n, \mathbf{X}_n)$ is not consistent because of the endogeneity of $W_n \mathbf{Y}_n$. To deal with this problem, one may use a maximum likelihood approach. Specifically, assume that the error terms $(\varepsilon_1, \ldots, \varepsilon_n)$ are IID as normal $N(0, \sigma_0^2)$. Then, the conditional joint distribution of $\mathbf{Y}_n$ given $\mathbf{X}_n$ is characterized by

$$\mathbf{Y}_n \mid \mathbf{X}_n \sim \mathbf{N}\Big((I_n - \rho_0 W_n)^{-1} \mathbf{X}_n \beta_0,\ (I_n - \rho_0 W_n)^{-1} (I_n - \rho_0 W_n^\top)^{-1} \sigma_0^2\Big),$$

and one can show that the log-likelihood function of the SLM is given by

$$\begin{aligned} \ell_n^{SLM}(\rho, \beta, \sigma^2) = &-\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 + \log|I_n - \rho W_n| \\ &- \frac{1}{2\sigma^2}((I_n - \rho W_n)\mathbf{Y}_n - \mathbf{X}_n\beta)^\top((I_n - \rho W_n)\mathbf{Y}_n - \mathbf{X}_n\beta), \end{aligned}$$

where $|A_n|$ denotes the determinant of the matrix $A_n$. Then, we can obtain a consistent estimator of $(\rho_0, \beta_0, \sigma_0^2)$ as the maximizer of the above log-likelihood function; see [Lee, 2004] for details on the theoretical properties of this estimator.

However, this maximum likelihood estimation has two serious drawbacks. First, the assumption that $(\varepsilon_1, \ldots, \varepsilon_n)$ are IID normal is very restrictive for spatial data, since spatial units are naturally correlated and heterogeneous in several characteristics, such as area size and population. Second, the computation of the objective function $\ell_n^{SLM}(\rho, \beta, \sigma^2)$ can be quite expensive when the sample size is not small; calculating the determinant of a large matrix is very time-consuming and often computationally intractable. Fortunately, as described below, one can obtain a consistent estimator of the model parameters by the 2SLS method of [Kelejian and Prucha, 1998], without relying on the IID normality assumption.

[1] In this case, the source of spatial autocorrelation is the presence of local strategic interaction. However, in many applications, what mechanisms generate the spatial autocorrelation is less concerned, and researchers often introduce a model like (13.1.1) a priori.

### 13.1.2 2SLS estimation

To simplify the exposition, in the following we assume that the regressor $X$ is non-stochastic and bounded in absolute value.[2] Then, since $\mathbb{E}\varepsilon = 0$, this automatically implies that the regressor $X$ is exogenous: $\mathbb{E}[X\varepsilon] = X\mathbb{E}[\varepsilon] = 0$.

Suppose that the matrix $I_n - \rho_0 W_n$ is nonsingular with $|\rho_0| < 1$. Then, we have

$$\begin{aligned} W_n\mathbf{Y}_n &= W_n(I_n - \rho_0 W_n)^{-1}\mathbf{X}_n\beta_0 + W_n(I_n - \rho_0 W_n)^{-1}\mathcal{E}_n \\ &= W_n\mathbf{X}_n\beta_0 + \rho_0 W_n^2\mathbf{X}_n\beta_0 + \rho_0^2 W_n^3\mathbf{X}_n\beta_0 + \cdots + W_n(I_n - \rho_0 W_n)^{-1}\mathcal{E}_n. \end{aligned}$$

The first line implies that the ideal instrumental variable for the endogenous regressor $W_n\mathbf{Y}_n$ is $W_n(I_n - \rho W_n)^{-1}\mathbf{X}_n\beta$. However, we cannot use this since $\rho_0$ and $\beta_0$ are unknown in practice. The second line shows that the ideal instrument $W_n(I_n - \rho_0 W_n)^{-1}\mathbf{X}_n\beta_0$ is expressed as a linear combination of $\{W_n\mathbf{X}_n, W_n^2\mathbf{X}_n, \ldots\}$. Thus, a subset of columns of $\{W_n\mathbf{X}_n, W_n^2\mathbf{X}_n, \ldots\}$ can be used as a feasible and reasonable instrument for $W_n\mathbf{Y}_n$. Specifically, let $\mathbf{Z}_{n1}$ be a matrix of instrumental variables composed of a subset of linearly independent columns of $\{W_n\mathbf{X}_n, W_n^2\mathbf{X}_n, \ldots\}$, and define $\mathbf{Z}_n = (\mathbf{X}_n, \mathbf{Z}_{n1})$. Then, a consistent estimator of $\theta_0 = (\rho_0, \beta_0^\top)^\top$ can be obtained the 2SLS method as follows. In the first step, we regress the endogenous variable $W_n\mathbf{Y}_n$ on $\mathbf{Z}_n$, and obtain the predicted value of $W_n\mathbf{Y}_n$ by

$$\widehat{W_n\mathbf{Y}_n} = \mathbf{Z}_n(\mathbf{Z}_n^\top\mathbf{Z}_n)^{-1}\mathbf{Z}_n^\top W_n\mathbf{Y}_n.$$

In the second step, $\theta$ is estimated by a linear regression of $\mathbf{Y}_n$ on $(\widehat{W_n\mathbf{Y}_n}, \mathbf{X}_n)$. Let $\widehat{\theta}_n$ be the resulting estimator of $\theta_0$. In the same manner as in (11.4.1), we can derive the closed-form expression for the 2SLS estimator $\widehat{\theta}_n$ as

$$\widehat{\theta}_n = \left[\mathbf{H}_n^\top\mathbf{Z}_n(\mathbf{Z}_n^\top\mathbf{Z}_n)^{-1}\mathbf{Z}_n^\top\mathbf{H}_n\right]^{-1}\mathbf{H}_n^\top\mathbf{Z}_n(\mathbf{Z}_n^\top\mathbf{Z}_n)^{-1}\mathbf{Z}_n^\top\mathbf{Y}_n, \tag{13.1.2}$$

where $\mathbf{H}_n = (W_n\mathbf{Y}_n, \mathbf{X}_n)$. Asymptotic properties of the 2SLS estimator (13.1.2) can be derived in the same way as in Section 2.4.

## 13.2 Spatial error model

It is highly possible that the unobserved determinant of $Y$, $\varepsilon$, involves some spatial factors. If they are spatially correlated, the model has spatial correlated effects. To account for the spatial correlated effects (i.e., spatial autocorrelation of $\varepsilon$), we can consider the following model:

$$\begin{aligned} Y_i &= X_i^\top\beta_0 + \varepsilon_i \\ \varepsilon_i &= \lambda_0\sum_{j=1}^n w_{ij}\varepsilon_j + u_i, \quad i = 1, \ldots, n \end{aligned} \tag{13.2.1}$$

where $u_i$ is an idiosyncratic (non-spatial) error term such that $\mathbb{E}u = 0$, and $\lambda_0$ is the parameter that captures the spatial correlated effect. The above model is called the **spatial error model** (SEM). In contrast to the SLM, the SEM does not admit the existence of spatial multiplier effects; that is, in SEM, an increase in own $X$ solely influences on own $Y$ and not on the others' outcomes. Thus, these two models have quite different policy implications.

[2] The assumption of non-stochastic regressors is often adopted in the literature of spatial econometrics, partly for the purpose of avoiding tedious technical discussions on the underlying spatial random process. In the non-stochastic regressors framework, we should interpret the analysis as being conditional on the realization of these variables.

### 13.2.1 Maximum likelihood estimation

Again, we retain the assumption that $X$ is non-stochastic. When the matrix $I_n - \lambda_0 W_n$ is nonsingular with $|\lambda_0| < 1$, the SEM can be re-written in matrix form as

$$\mathbf{Y}_n = \mathbf{X}_n\beta_0 + (I_n - \lambda_0 W_n)^{-1}U_n,$$

where $U_n = (u_1, \ldots, u_n)^\top$. Then, since each $\varepsilon$ can be expressed as a linear combination of $(u_1, \ldots, u_n)$, we have $\mathbb{E}\varepsilon = 0$, and thus $\mathbb{E}[X\varepsilon] = 0$ holds; that is, $X$ is exogenous. Therefore, for the estimation of SEM, a simple OLS regression yields a consistent estimate of $\beta_0$, although not efficient and uninformative for the spatial parameter $\lambda_0$.

In order to estimate $\lambda_0$, we can consider a maximum likelihood estimation. Assuming now that $(u_1, \ldots, u_n)$ are IID as normal $N(0, \sigma_0^2)$, the conditional joint distribution of $\mathbf{Y}_n$ given $\mathbf{X}_n$ is characterized by

$$\mathbf{Y}_n \mid \mathbf{X}_n \sim \mathbf{N}\left(\mathbf{X}_n\beta_0,\ (I_n - \lambda_0 W_n)^{-1}(I_n - \lambda_0 W_n^\top)^{-1}\sigma_0^2\right).$$

After some calculations, the log-likelihood function of the SEM can be obtained as follows:

$$\begin{aligned}\ell_n^{SE}(\lambda, \beta, \sigma^2) = &-\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 + \log|I_n - \lambda W_n| \\ &- \frac{1}{2\sigma^2}(\mathbf{Y}_n - \mathbf{X}_n\beta)^\top(I_n - \lambda W_n)^\top(I_n - \lambda W_n)(\mathbf{Y}_n - \mathbf{X}_n\beta).\end{aligned}$$

By maximizing the above log-likelihood function, we can obtain a consistent estimator of $(\lambda_0, \beta_0, \sigma_0^2)$. However, the IID normality assumption is restrictive as mentioned above. Also, again since the log-likelihood function involves the determinant of the matrix $I_n - \lambda W_n$, there can be significant difficulties in the practical computation of this estimator. In order to overcome this issue, [Kelejian and Prucha, 1999] developed a method of moments (MM) estimator that is not dependent on the IID normality assumption and is computationally easy irrespective of the size of the sample.

### 13.2.2 Method of moments estimation

The estimation procedure of the MM estimator is implemented in two steps. The first step estimates the coefficients $\beta_0$ by the OLS regression of $\mathbf{Y}_n$ on $\mathbf{X}_n$, and let $\widehat{\beta}_n$ be the resulting estimator. Here, we assume that $(u_1, \ldots, u_n)$ are homoskedastic with variance $\mathbb{E}[u_i^2] = \sigma_0^2$ for all $i$.[3] In the second step, we estimate $(\lambda_0, \sigma_0^2)$ by the MM approach. The estimator is based on the following three moment equalities:

$$\mathbb{E}\left[U_n^\top U_n/n\right] = \sigma_0^2,\ \ \mathbb{E}\left[\overline{U}_n^\top \overline{U}_n/n\right] = \sigma_0^2 \mathrm{trace}\left\{W_n^\top W_n/n\right\},\ \ \mathbb{E}\left[\overline{U}_n^\top U_n/n\right] = 0,$$

where $\overline{U}_n = W_n U_n$. Note that for any $n \times 1$ vectors $A_n$ and $B_n$, $A_n^\top B_n = \mathrm{trace}\{A_n B_n^\top\}$. In the third equality, we have used the fact that the diagonal elements of $W_n$ are zero. Let $\overline{\mathcal{E}}_n = W_n\mathcal{E}_n$ and $\overline{\overline{\mathcal{E}}}_n = W_n W_n \mathcal{E}_n$. Since we have

$$U_n = \mathcal{E}_n - \lambda_0\overline{\mathcal{E}}_n \ \text{ and } \ \overline{U}_n = \overline{\mathcal{E}}_n - \lambda_0\overline{\overline{\mathcal{E}}}_n,$$

[3] The homoskedasticity assumption is introduced for simplicity. A heteroskedasticity-robust version of the estimation procedure is explained in [Kelejian and Prucha, 2010].

the above three equalities can be rewritten in terms of $\mathcal{E}_n$ as

$$\begin{aligned}
&\mathbb{E}\left[\mathcal{E}_n^\top \mathcal{E}_n/n\right] &&-2\lambda_0\mathbb{E}\left[\mathcal{E}_n^\top \overline{\mathcal{E}}_n/n\right] &&+\lambda_0^2\mathbb{E}\left[\overline{\mathcal{E}}_n^\top \overline{\mathcal{E}}_n/n\right] &&-\sigma_0^2 &&= 0,\\
&\mathbb{E}\left[\overline{\mathcal{E}}_n^\top \overline{\mathcal{E}}_n/n\right] &&-2\lambda_0\mathbb{E}\left[\overline{\mathcal{E}}_n^\top \overline{\overline{\mathcal{E}}}_n/n\right] &&+\lambda_0^2\mathbb{E}\left[\overline{\overline{\mathcal{E}}}_n^\top \overline{\overline{\mathcal{E}}}_n/n\right] &&-\sigma_0^2\mathrm{trace}\left\{W_n^\top W_n/n\right\} &&= 0,\\
&\mathbb{E}\left[\overline{\mathcal{E}}_n^\top \mathcal{E}_n/n\right] &&-\lambda_0\mathbb{E}\left[\overline{\mathcal{E}}_n^\top \overline{\mathcal{E}}_n/n + \overline{\overline{\mathcal{E}}}_n^\top \mathcal{E}_n/n\right] &&+\lambda_0^2\mathbb{E}\left[\overline{\mathcal{E}}_n^\top \overline{\overline{\mathcal{E}}}_n/n\right] &&&&= 0.
\end{aligned}$$

Equivalently,

$$\Gamma_n(\lambda_0, \lambda_0^2, \sigma_0^2)^\top - \gamma_n = \mathbf{0},$$

where

$$\Gamma_n \equiv \begin{pmatrix}
2\mathbb{E}\left[\mathcal{E}_n^\top \overline{\mathcal{E}}_n/n\right] & -\mathbb{E}\left[\overline{\mathcal{E}}_n^\top \overline{\mathcal{E}}_n/n\right] & 1\\
2\mathbb{E}\left[\overline{\mathcal{E}}_n^\top \overline{\overline{\mathcal{E}}}_n/n\right] & -\mathbb{E}\left[\overline{\overline{\mathcal{E}}}_n^\top \overline{\overline{\mathcal{E}}}_n/n\right] & \mathrm{trace}\left\{W_n^\top W_n/n\right\}\\
\mathbb{E}\left[\overline{\mathcal{E}}_n^\top \overline{\mathcal{E}}_n/n + \overline{\overline{\mathcal{E}}}_n^\top \mathcal{E}_n/n\right] & -\mathbb{E}\left[\overline{\mathcal{E}}_n^\top \overline{\overline{\mathcal{E}}}_n/n\right] & 0
\end{pmatrix} \quad \gamma_n \equiv \begin{pmatrix}
\mathbb{E}\left[\mathcal{E}_n^\top \mathcal{E}_n/n\right]\\
\mathbb{E}\left[\overline{\mathcal{E}}_n^\top \overline{\mathcal{E}}_n/n\right]\\
\mathbb{E}\left[\overline{\mathcal{E}}_n^\top \mathcal{E}_n/n\right]
\end{pmatrix}.$$

Thus, the true value of $(\lambda, \sigma^2)$ can be characterized as the minimizer of the objective function $||\Gamma_n(\lambda, \lambda^2, \sigma^2)^\top - \gamma_n||^2$, where $||\cdot||$ denotes the Euclidean norm. This implies that we can estimate $(\lambda_0, \sigma_0^2)$ by minimizing the sample analog of $||\Gamma_n(\lambda, \lambda^2, \sigma^2)^\top - \gamma_n||^2$, where the expectations in the definition of $\Gamma_n$ and $\gamma_n$ are dropped, and $\mathcal{E}_n$ is replaced by $\widehat{\mathcal{E}}_n = \mathbf{Y}_n - \mathbf{X}_n\widehat{\beta}_n$. Let the resulting MM estimator of $(\lambda_0, \sigma_0^2)$ be $(\widehat{\lambda}_n, \widehat{\sigma}_n^2)$. [Kelejian and Prucha, 1999] proved that the MM estimator $(\widehat{\lambda}_n, \widehat{\sigma}_n^2)$ is consistent under mild regularity conditions. The limiting distribution of $(\widehat{\lambda}_n, \widehat{\sigma}_n^2)$ is complicated and is omitted here (see, e.g., [Kelejian and Prucha, 2010]).

Once a consistent estimate $(\widehat{\lambda}_n, \widehat{\sigma}_n^2)$ is obtained, we can use a generalized least squares (GLS) estimator of $\beta_0$ to obtain a more efficient estimate. Specifically, let

$$\widehat{\Omega}_n \equiv (I_n - \widehat{\lambda}_n W_n)^{-1}(I_n - \widehat{\lambda}_n W_n^\top)^{-1}.$$

Then, the GLS estimator of $\beta_0$ is defined by

$$\widehat{\beta}_n^{GLS} = \left[\mathbf{X}_n^\top \widehat{\Omega}_n^{-1}\mathbf{X}_n\right]^{-1}\mathbf{X}_n^\top \widehat{\Omega}_n^{-1}\mathbf{Y}_n. \tag{13.2.2}$$

The GLS estimator $\widehat{\beta}_n^{GLS}$ is asymptotically efficient, with its covariance matrix being $(\sigma^2/n)\left[\mathbf{X}_n^\top \Omega_n^{-1}\mathbf{X}_n/n\right]^{-1}$, where $\Omega_n \equiv (I_n - \lambda_0 W_n)^{-1}(I_n - \lambda_0 W_n^\top)^{-1}$.

## 13.3 Empirical analysis with R: Household burglary in Tokyo

It is commonly observed that different locations sharing the same social and economic conditions do not necessarily experience the same crime intensity. Instead, spatial clusters of crimes (i.e., "hot spots") exist. Accordingly, it is important to incorporate spatial information in quantitative criminology research. In this empirical analysis, we investigate the determinants of household burglary risk in Tokyo.

The dataset used here is the same as the one in [Hoshino, 2018].[4] The variables used in the analysis are created from several data sources. For crime data, we use the number of household burglaries recorded in 2011 by the Tokyo Metropolitan Police Department for each district in Tokyo. The socio-demographic and economic information for each district are taken from the Census 2010 and the Commercial Statistics Survey 2009. The definitions of the variables are summarized in the following table. All variables are constructed at the district level.

Variables and their definitions

| | Variable | Definition |
|---|---|---|
| Dependent variable | *burglary* | $1,000 \times \frac{\#\text{ of recorded household burglaries}}{\#\text{ of households}}$ |
| Explanatory variables | *nhmem* | Average number of household members. |
| | *dnsty* | Residential density (# of residences / Area of district). |
| | *owner* | Portion of owner-occupied houses. |
| | *elder* | Portion of households headed by the elderly (over 65 years old). |
| | *high* | Portion of high-rise residential buildings. |
| | *manag* | # of workers in managerial positions / Labor force population. |
| | *retail* | Total area of retail stores in district / Area of district. |
| | Ward dummies | Dummies for the 23 special wards of Tokyo.† |

(†: Here, 22 dummy variables are introduced. Nerima ward is omitted as a benchmark.)

After deleting the observations with missing values, a sample of $n = 3025$ observations is used in this analysis. With this dataset, we estimate SLM (13.1.1) and SEM (13.2.1) by the 2SLS method and the MM method, respectively. The spatial weight matrix is constructed based on Queen-contiguity and is row-normalized.

Set the **R** working directory appropriately, and read the csv files.

```
# Data #

 data <- as.data.frame(read.csv("TokyoCrime.csv"))
 n    <- nrow(data) # sample size (n = 3025)

# Spatial weight matrix #

 W <- as.matrix(read.csv("TokyoCrime_W.csv", header = FALSE))
```

The variables used are defined as follows.

```
# Variables #

 Y  <- data$burglary
 X1 <- with(data, cbind(nhmem, dnsty, owner, elder, high, manag, retail))

 ward23 <- as.numeric(data$ward)
 wdum   <- matrix(0, n, 22)
 for(i in 1:22) wdum[,i] <- ifelse(ward23 == i, 1, 0)
```

[4] The data files are available on request.

```
X <- cbind(1, X1, wdum)
```

We first estimate the SLM. In accordance with the notation in Section 13.1.2, define the variables as follows.

```
H  <- cbind(W%*%Y, X)
Z1 <- cbind(W%*%X1, W%*%W%*%X1, W%*%W%*%W%*%X1)
Z  <- cbind(X, Z1)
```

As the instruments for $W_n\mathbf{Y}_n$, we employ the first, second, and third-order spatial lags of the explanatory variables (excluding the constant and the ward dummies). Then, following (13.1.2), we obtain the 2SLS estimate in the following manner.

```
# 2SLS estimation #

P     <- Z%*%solve(t(Z)%*%Z)%*%t(Z)
theta <- solve(t(H)%*%P%*%H)%*%t(H)%*%P%*%Y
```

We compute the covariance matrix of the 2SLS estimator under homoskedasticity. Once the covariance matrix is obtained, we can then calculate the standard error and t-value of the estimated parameter.

```
# Covariance matrix under homoskedasticity #

sigma2 <- mean((Y - H%*%theta)^2)
Q_ZZ   <- t(Z)%*%Z/n
Q_ZH   <- t(Z)%*%H/n
Cov    <- (sigma2/n)*solve(t(Q_ZH)%*%solve(Q_ZZ)%*%Q_ZH)

sd   <- diag(Cov)^0.5 # Standard error
tval <- theta/sd      # t-value
```

The summary of the results is as follows.

```
>  cbind(theta, sd, tval)
                       sd
        0.4257077087 0.18952982  2.246125184
        0.6074076919 0.23175842  2.620865663
nhmem  -0.1652634554 0.09324095 -1.772434221
dnsty   5.9766729090 3.18795297  1.874768216
owner   0.0733078500 0.16210661  0.452220004
elder   0.0625461529 0.35319744  0.177085524
high   -0.5562633330 0.12004293 -4.633870187
manag  -0.3134209803 0.92201410 -0.339930790
retail -0.4174281178 0.41219978 -1.012683981

== The results for the ward dummies are omitted to save space. ==
```

The results show that the spatial autocorrelation in the household burglary rate is positive and significant. It indicates that a one-point increase in the average burglary rate of neighboring districts increases own district's burglary rate by about 0.4. Among the other explanatory variables, we find that the *high* variable significantly negatively affect the burglary rate. This result would be a reflection of the fact that high-rise buildings are likely equipped with high-security system.

Next, we estimate the SEM. The fist step of the estimation procedure is to estimate $\beta_0$ by OLS.

```
# OLS estimation #

 beta <- lm(Y ~ X - 1)$coef
```

The second step is to implement the MM estimation. In line with Section 13.2.2, we define the following quantities.

```
 E  <- Y - X%*%beta
 E1 <- W%*%E
 E2 <- W%*%W%*%E

 Gam1 <- c(2*t(E)%*%E1/n,              -t(E1)%*%E1/n, 1)
 Gam2 <- c(2*t(E1)%*%E2/n,             -t(E2)%*%E2/n, sum(diag(t(W)%*%W))/n)
 Gam3 <- c(t(E1)%*%E1/n + t(E2)%*%E/n, -t(E1)%*%E2/n, 0)
 Gam  <- rbind(Gam1, Gam2, Gam3)
 gam  <- c(t(E)%*%E/n, t(E1)%*%E1/n, t(E1)%*%E/n)
```

Then, the objective function to be minimized is defined as follows.

```
# Objective function in the MM estimation #

 ObjF <- function(param){

        g <- Gam%*%c(param[1], param[1]^2, param[2]^2) - gam
        sum(g^2)

        }
```

Here, the vector `param` is a 2-dimensional vector with its first element corresponding to the spatial parameter $\lambda$ and the second element to the standard error $\sigma$. The minimizer of `ObjF` can be found by using the command `optim`. To do this, we need to specify initial candidate values for $\lambda$ and $\sigma$, which are set to 0.5 and the standard deviation of the residuals (`sd(E)`), respectively.

```
 NLS    <- optim(c(0.5, sd(E)), ObjF)
 lambda <- NLS$par[1]
 sigma2 <- NLS$par[2]^2

>  c(lambda, sigma2)
[1] 0.1597388 0.9959481
```

Using these estimated values, we re-estimate $\beta_0$ by the GLS method. Following (13.2.2), the GLS estimate $\widehat{\beta}_n^{GLS}$ is computed as follows.

```
# GLS estimation #

 Omega    <- solve(diag(n) - lambda*W)%*%solve(diag(n) - lambda*t(W))
 beta_GLS <- solve(t(X)%*%solve(Omega)%*%X)%*%t(X)%*%solve(Omega)%*%Y

>  cbind(beta, beta_GLS)
          beta
X         0.8835338084  0.88768536
Xnhmem   -0.1741576539 -0.17705278
```

```
Xdnsty   7.5336515722  7.19534826
Xowner   0.0612327015  0.05990058
Xelder   0.0172327270  0.01519167
Xhigh   -0.6863309538 -0.66143487
Xmanag  -0.4887735638 -0.40733820
Xretail -0.4189122834 -0.49152027

== The results for the ward dummies are omitted to save space. ==
```

Since the impact of the spatially correlated effects is not strong, the OLS estimate and the one obtained by GLS are only slightly different. Finally, we compute the standard errors and t-values of the GLS coefficients.

```
# Covariance matrix #

 Cov <- (sigma2/n)*solve(t(X)%*%solve(Omega)%*%X/n)

 sd   <- diag(Cov)^0.5 # Standard error
 tval <- beta_GLS/sd   # t-value
```

The summary of the results is as follows.

```
>  cbind(beta_GLS, sd, tval)
                    sd
        0.88768536 0.20913531  4.24455045
nhmem  -0.17705278 0.09752507 -1.81545921
dnsty   7.19534826 3.15978566  2.27716340
owner   0.05990058 0.16498658  0.36306333
elder   0.01519167 0.36085584  0.04209899
high   -0.66143487 0.10933201 -6.04978255
manag  -0.40733820 0.95136201 -0.42816319
retail -0.49152027 0.42438051 -1.15820651

== The results for the ward dummies are omitted to save space. ==
```

When comparing the parameter values estimated from the SLM and those from the SEM, the signs and relative significances are similar overall. In the SEM, the *dnsty* variable shows a significantly positive impact on the burglary rate: it should be natural to expect fewer household burglaries in areas with fewer residences.

# Chapter 14

# Binary Response Models with Spatial Interactions

In this chapter, we consider binary response models with spatial interactions. The outcome variable of interest is now a dummy variable, say $D$. For example, suppose that $D_i = 1$ if assault and violent crimes occurred once or more in district $i$ in a given year, and $D_i = 0$ otherwise. As we have seen in Section 13.3, crime data often exhibit spatial correlation for several reasons. Thus, in order to precisely examine the determinants of $D$, it is required to incorporate the spatial dependence in the model. In Section 14.1, we introduce the spatial-lag probit model and the spatial-error probit model, which are natural extensions of the linear spatial lag model (in Section 13.1) and the linear spatial error model (in Section 13.2), respectively. The estimation of the spatial probit models is much more involved as compared to the linear models, and the standard maximum likelihood approach is not feasible due to computational complexity. Then in Section 14.2, for model estimation, we develop GMM procedures that can be implemented relatively easily.[1]

## 14.1 Spatial probit models

### 14.1.1 Two spatial-lag probit models

Let $D$ be a binary response variable of interest, and $X$ be the vector of explanatory variables. Suppose that a sample $\{(D_i, X_i) : 1 \le i \le n\}$ of size $n$ is observed. Further, let $w_{ij}$ be a spatial weight term between $i$ and $j$ with $w_{ii} = 0$ for normalization.

First, we consider the following model, which can be seen as a direct extension of the linear spatial lag model in Section 13.1 to binary outcome:

$$\begin{aligned} D_i &= \mathbf{1}\{Y_i^* \ge 0\} \\ Y_i^* &= \rho_0 \sum_{j=1}^{n} w_{ij} D_j + X_i^\top \beta_0 - \varepsilon_i, \quad i = 1, \ldots, n. \end{aligned} \tag{14.1.1}$$

Here, $\rho_0$ is the parameter that represents the magnitude of the spatial endogenous effect for the binary outcome $D$, and $Y^*$ is an unobservable "latent" dependent variable (propensity variable). Hereinafter, we assume that the

[1] Alternatively, some researchers have suggested the use of Bayesian approach (see, e.g., [LeSage and Pace, 2009]).

error terms $\varepsilon_i$'s are IID as the standard normal $N(0,1)$. Note that the standard probit model cannot be used to estimate this model because the regressor $\sum_{j=1}^n w_{ij}D_j$ is not independent of $\varepsilon_i$ through the spatial interactions, which causes the endogeneity issue.

Not only the endogeneity problem, the model (14.1.1) has a more serious issue that hinders the estimation. If we interpret $Y^*$ as the "marginal payoff" of taking an action $D = 1$, observing that one's payoff is dependent on the others' actions, then the realized outcomes $\{D_1, \ldots, D_n\}$ based on (14.1.1) can be viewed as a pure strategy Nash equilibrium under complete information. In general, there is potentially a large number of multiple solutions in this model. As discussed in Section 6.3, even in the case of two individuals only, the estimation of the model (14.1.1) involves a certain difficulty. As such, the model (14.1.1) has been rarely studied in the literature of spatial econometrics.[2]

Now, suppose instead that the propensity variable $Y^*$ is spatially autocorrelated. Such a model is more tractable than the one in (14.1.1). Namely, consider the following model:

$$\begin{aligned} D_i &= \mathbf{1}\{Y_i^* \geq 0\} \\ Y_i^* &= \rho_0 \sum_{j=1}^n w_{ij} Y_j^* + X_i^\top \beta_0 - \varepsilon_i, \quad i = 1, \ldots, n. \end{aligned} \tag{14.1.2}$$

Let $W_n = (w_{ij})_{i,j=1}^n$ be the spatial weight matrix, and write $\mathbf{Y}_n^* = (Y_1^*, \ldots, Y_n^*)^\top$, $\mathbf{X}_n = (X_1, \ldots, X_n)^\top$, and $\mathcal{E}_n = (\varepsilon_1, \ldots, \varepsilon_n)^\top$. Then, we can rewrite the second equation in (14.1.2) in matrix form as

$$\begin{aligned} \mathbf{Y}_n^* &= \rho_0 W_n \mathbf{Y}_n^* + \mathbf{X}_n \beta_0 - \mathcal{E}_n \\ &= (I_n - \rho_0 W_n)^{-1} \mathbf{X}_n \beta_0 - (I_n - \rho_0 W_n)^{-1} \mathcal{E}_n. \end{aligned}$$

Let the $i$-th row of $(I_n - \rho_0 W_n)^{-1}\mathbf{X}_n$ be $\overline{x}_i(\rho_0)$ and the $i$-th element of $(I_n - \rho_0 W_n)^{-1}\mathcal{E}_n$ be $\overline{\varepsilon}_i(\rho_0)$. Then, we can write

$$D_i = \mathbf{1}\{\overline{x}_i(\rho_0)^\top \beta_0 \geq \overline{\varepsilon}_i(\rho_0)\}.$$

Note that, in order to estimate the parameters by a (full) maximum likelihood procedure, we need to compute the joint probability of $\{D_1, \ldots, D_n\}$ given $\mathbf{X}_n$. However, by assumption,

$$\begin{pmatrix} \overline{\varepsilon}_1(\rho_0) \\ \vdots \\ \overline{\varepsilon}_n(\rho_0) \end{pmatrix} = (I_n - \rho_0 W_n)^{-1} \mathcal{E}_n \sim \mathbf{N}\Big( \underset{n\times 1}{\mathbf{0}}, \underset{n\times n}{(I_n - \rho_0 W_n)^{-1}(I_n - \rho_0 W_n^\top)^{-1}} \Big),$$

implying that the reduced form error terms $\overline{\varepsilon}_i(\rho_0)$'s are not independent with non-zero covariances; thus, the likelihood function involves the evaluation of an $n$-dimensional integral

$$\Pr(D_1, \ldots, D_n \mid \mathbf{X}_n) = \int_{\mathcal{A}_n} \int_{\mathcal{A}_{n-1}} \cdots \int_{\mathcal{A}_1} \phi(e_1, \ldots, e_n; \rho_0) \mathrm{d}e_1 \cdots \mathrm{d}e_{n-1} \mathrm{d}e_n,$$

where $\phi(\cdot, \ldots, \cdot; \rho_0)$ is the joint density function of $n$-variate normal distribution with mean $\mathbf{0}$ and covariance matrix $(I_n - \rho_0 W_n)^{-1}(I_n - \rho_0 W_n^\top)^{-1}$, and

$$\mathcal{A}_i \equiv \begin{cases} \big(-\infty, \overline{x}_i(\rho_0)^\top \beta_0\big] & \text{if } D_i = 1 \\ \big(\overline{x}_i(\rho_0)^\top \beta_0, \infty\big) & \text{if } D_i = 0. \end{cases}$$

[2] One way to deal with this problem is to explicitly assume an equilibrium selection rule when in the multiple equilibria region (see [Soetevent and Kooreman, 2007]), as in Subsection 6.3.2.

This high dimensional integration makes the standard maximum likelihood estimation infeasible for large sample sizes. To circumvent this problem, researchers have proposed several alternatives to the maximum likelihood method that can be implemented in practice (at the cost of some efficiency loss). Among them, we will describe a GMM estimator for the model (14.1.2) in Section 14.2.

### 14.1.2 Spatial-error probit model

Similarly to Section 13.2, we can also consider to incorporate spatially autocorrelated error terms in the binary response model:

$$\begin{aligned} D_i &= \mathbf{1}\{Y_i^* \geq 0\} \\ Y_i^* &= X_i^\top \beta_0 - \varepsilon_i, \\ \varepsilon_i &= \lambda_0 \sum_{j=1}^n w_{ij}\varepsilon_j + u_i, \quad i = 1, \ldots, n. \end{aligned} \tag{14.1.3}$$

where $u_i$'s are idiosyncratic (non-spatial) error terms that are IID as the standard normal $N(0,1)$, and $\lambda_0$ is the parameter that captures the spatial correlated effect. Then, we can call this model the spatial-error probit.

Letting $U_n = (u_1, \ldots, u_n)^\top$, we can rewrite the second equation in (14.1.3) in matrix form as

$$\mathbf{Y}_n^* = \mathbf{X}_n\beta_0 - (I_n - \lambda_0 W_n)^{-1}U_n.$$

If we denote the $i$-th element of $(I_n - \lambda_0 W_n)^{-1}U_n$ as $\overline{u}_i(\lambda)$, then we can write

$$D_i = \mathbf{1}\{X_i^\top \beta_0 \geq \overline{u}_i(\lambda_0)\}.$$

Analogously to the above discussion, we observe that

$$\begin{pmatrix} \overline{u}_1(\lambda_0) \\ \vdots \\ \overline{u}_n(\lambda_0) \end{pmatrix} = (I_n - \lambda_0 W_n)^{-1}U_n \sim \mathbf{N}\Big(\underset{n\times 1}{\mathbf{0}},\ \underset{n\times n}{(I_n - \lambda_0 W_n)^{-1}(I_n - \lambda_0 W_n^\top)^{-1}}\Big).$$

Again, this implies that estimating the parameters by a (full) maximum likelihood procedure involves a high-dimensional integration, which is intractable when the sample size is large.

## 14.2 GMM estimation of spatial probit models

In this section, we mainly discuss GMM estimation of the spatial-lag probit model in (14.1.2). GMM estimation of the spatial-error probit model (14.1.3) can be implemented in a similar manner, and is omitted here.

Here, let $\ell_{ij}(\rho_0)$ be the $(i,j)$-th element of the matrix $(I_n - \rho_0 W_n)^{-1}$. Then, the reduced form error term $\overline{\varepsilon}_i(\rho_0)$ can be written as

$$\overline{\varepsilon}_i(\rho_0) = \sum_{j=1}^n \ell_{ij}(\rho_0)\varepsilon_j.$$

Recall that we have assumed that $\varepsilon_j$'s are IID standard normal. Thus, it holds that

$$\ell_{ij}(\rho_0)\varepsilon_j \sim N(0, \ell_{ij}^2(\rho_0)).$$

Moreover, by the reproductive property of normal distribution, the distribution of $\overline{\varepsilon}_i(\rho_0)$ is characterized as

$$\overline{\varepsilon}_i(\rho_0) \sim N(0, V_i(\rho_0)), \text{ where } V_i(\rho_0) \equiv \sum_{j=1}^{n} \ell_{ij}^2(\rho_0).$$

Hence, the distribution of $\overline{\varepsilon}_i(\rho_0)/\sqrt{V_i(\rho_0)}$ is standard normal. Then, since

$$\begin{aligned} D_i &= \mathbf{1}\{\overline{x}_i(\rho_0)^\top \beta_0 \geq \overline{\varepsilon}_i(\rho_0)\} \\ &= \mathbf{1}\{\xi_i(\rho_0)^\top \beta_0 \geq r_i\}, \text{ where } \xi_i(\rho_0) \equiv \frac{\overline{x}_i(\rho_0)}{\sqrt{V_i(\rho_0)}} \text{ and } r_i \equiv \frac{\overline{\varepsilon}_i(\rho_0)}{\sqrt{V_i(\rho_0)}} \sim N(0,1), \end{aligned}$$

we can compute the generalized residual for this model, the conditional expectation of $r_i$ given $(D_i, X_1, \ldots, X_n)$, by

$$\eta_i(\theta_0) \equiv \frac{(D_i - \Phi(\xi_i(\rho_0)^\top \beta_0)) \cdot \phi(\xi_i(\rho_0)^\top \beta_0)}{\Phi(\xi_i(\rho_0)^\top \beta_0) \cdot (1 - \Phi(\xi_i(\rho_0)^\top \beta_0))}$$

as in Example 4.1.4, where $\theta_0 = (\rho_0, \beta_0^\top)^\top$.

Now, let $Z_i$ be the vector of instrumental variables. Note that since the number of unknown parameters is larger than the dimension of $X$, we need to add (at least one) more exogenous variables to estimate $\rho_0$. For example, letting $X_{1i}$ be a sub-vector of $X_i$, one can use $Z_i = (X_i^\top, \sum_{j=1}^n w_{i,j} X_{1i}^\top)^\top$. Then, the GMM estimator of the true parameter $\theta_0$ is defined by

$$\widehat{\theta}_n = \operatorname*{argmin}_\theta \overline{g}_n(\theta)^\top \Omega_n \overline{g}_n(\theta), \tag{14.2.1}$$

where

$$\overline{g}_n(\theta) \equiv \frac{1}{n} \sum_{i=1}^{n} Z_i \eta_i(\theta).$$

Note that the above estimation procedure requires calculating the inverse matrix $(I_n - \rho W_n)^{-1}$ many times, which is computationally extremely burdensome when $n$ is not small. For computational tractability, in practice, one can use the Neumann series approximation $(I_n - \rho W_n)^{-1} \approx \sum_{t=0}^{T} \rho^t W_n^t$ with $T$ being three or four. Asymptotic properties of the GMM estimator (14.2.1) can be derived in the same way as in Section 4.3. See also [Pinkse and Slade, 1998] for more details.

# Appendix A

# Introductory Graph Theory

In history, the paper written by Leonhard Euler on the *Seven Bridges of Königsberg* published in 1736 is regarded as the first paper in graph theory. In this paper, Euler solved the famous Königsberg bridges problem which asks "whether you can cross each of the seven bridges in Figure A.1 exactly once and return to the starting point".

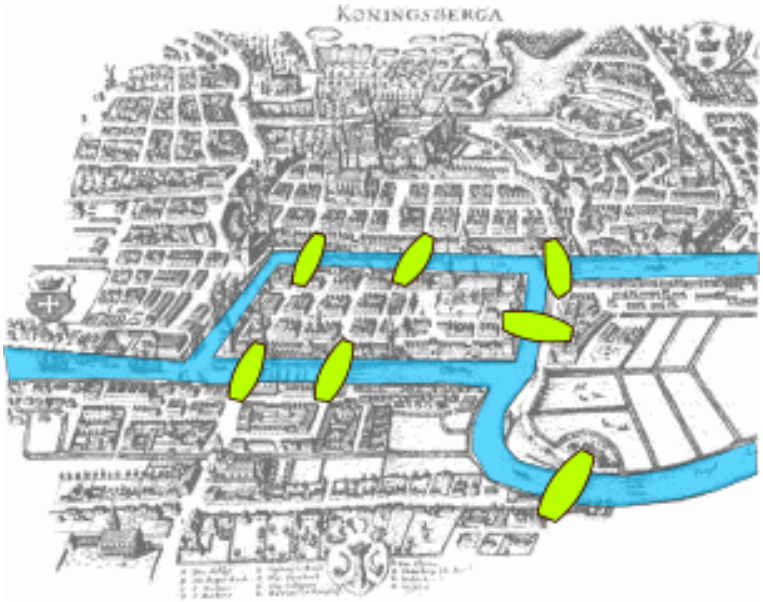



Figure A.1: Seven Bridges of Königsberg

(The figure is taken from Wikipedia.)

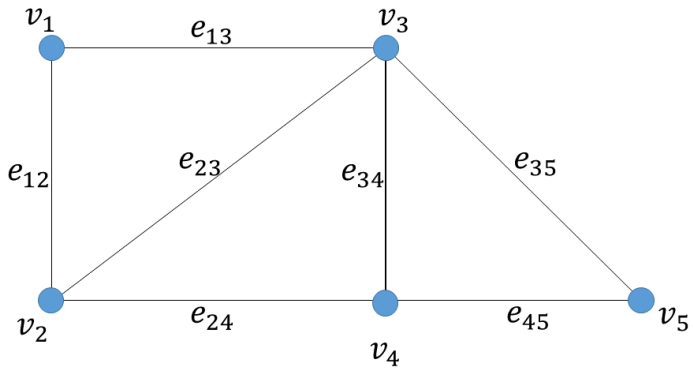
The answer to the Königsberg bridges problem is: NO.

## A.1 Basic terminology

Mathematically, a **graph** $G$ is a pair of sets $G = (V, E)$, where $V = \{v_i\}$ is a set of **vertices** (also referred to as nodes, agents, individuals, etc, depending on the context), and $E = \{e_{ij}\}$ is a set of **edges** (links between nodes). If $i = j$, this edge $e_{ii}$ is called **loop** or self-loop (self-link). We say that a graph has **multiple edges** if there are multiple distinct edges connecting the same pair of vertices (i.e., $E$ is a *multiset* of edges rather than an ordinary

set). A **simple graph** is a graph that has no loops and no multiple edges. In the following, we mainly focus on simple graphs.

A **directed graph** is a graph whose edges have a *direction* (directed graphs distinguish between $e_{ij}$ and $e_{ji}$). A graph is an **undirected graph** if it is not directed (no distinction between $e_{ij}$ and $e_{ji}$). Figure A.2 is an example of simple undirected graph. In the following, except when explicitly stated, we will restrict our attention to undirected graphs.




$$V = \{v_1, v_2, v_3, v_4, v_5\}$$
$$E = \{e_{12}, e_{13}, e_{23}, e_{24}, e_{34}, e_{35}, e_{45}\}$$

Figure A.2: A graph $G = (V, E)$

A **complete graph** is a simple undirected graph in which every pair of vertices is connected by an edge. The complete graph with $n$ vertices is often denoted by $K_n$ (Figure A.3). It is clear that a complete graph $K_n$ has $n(n-1)/2$ edges.

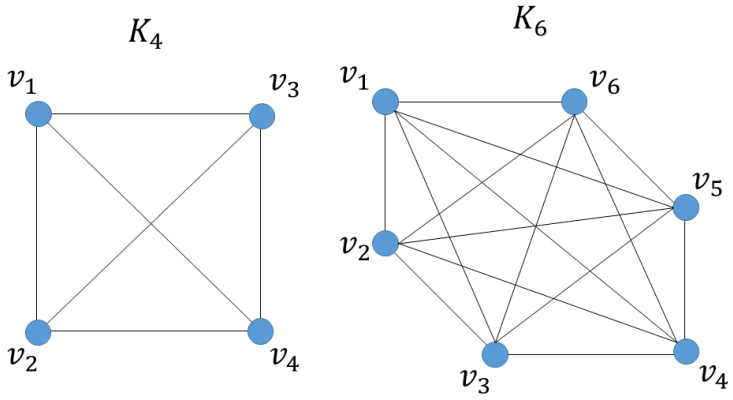



Figure A.3: Complete graphs

## A.2 Paths, Cycles and Connectivity

For a graph $G = (V, E)$, a **walk** from $v_i$ to $v_t$ is a finite sequence of edges of the form $e_{ij}, e_{jk}, \ldots, e_{ls}, e_{st}$. If the starting vertex and the ending vertex are same, i.e., $v_i = v_t$, we say that the walk is **closed**. The number of edges included in a walk is called its **length**. A walk $e_{ij}, e_{jk}, \ldots, e_{ls}, e_{st}$ in which all the edges $\{e_{ij}, e_{jk}, \ldots, e_{ls}, e_{st}\}$ are distinct is called a **trail**. In addition, a walk in which all the vertices $\{v_i, v_j, v_k, \ldots v_l, v_s, v_t\}$ are distinct (except, possibly, $v_i = v_t$) is a **path**. A closed path is called a **cycle**. For vertices $v_i$ and $v_j$, the **distance** from $v_i$ to $v_j$ is the length of the shortest path from $v_i$ to $v_j$. When a graph is directed, there may exist one-way edges in the graph. Therefore, in this case the distance from $v_i$ to $v_j$ does not generally coincide with that from $v_j$ to $v_i$.

We say that a graph is **connected** if there is a path between each pair of vertices, otherwise it is said to be disconnected (Figure A.4). When a graph is disconnected, it is possible that there is no path between a pair of vertices. In this case, we set the distance between them to infinity.

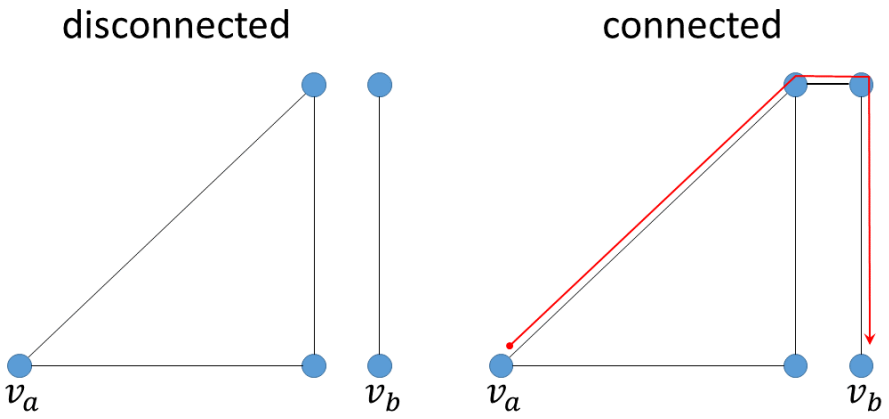



Figure A.4: Connected/disconnected graph

The next result provides an easy-to-check sufficient condition for a graph to be connected.

**Proposition A.2.1** *Let a graph $G = (V, E)$ be a simple graph with $n \geq 2$ vertices. If the number of edges in $G$ is larger than $(n-1)(n-2)/2$, then the graph is connected.*

**Proof.** We prove the result by contradiction. Suppose that $G$ is disconnected and is comprised of $k > 1$ connected graph components. It is clear that the maximum number of edges that $G$ can contain decreases as $k$ increases. Thus, it is sufficient to consider the case $k = 2$. Let $G_1$ and $G_2$ be connected graphs with $n_1$ and $n_2$ vertices, respectively, such that $G = G_1 \cup G_2$ $(n_1 + n_2 = n)$. Here, note that, in order to attain the maximum number of edges, (without loss of generality) $G_1$ must be a single vertex without edges and $G_2$ must be a complete graph $K_{n-1}$. Then, since the number of edges in $K_{n-1}$ is equal to $(n-1)(n-2)/2$, the number of edges in $G$ is at most $(n-1)(n-2)/2$, which is a contradiction with the assumption. Thus, $G$ is connected. ■

## A.3 Degree

For a graph $G = (V, E)$, if there is an edge $e_{ij} \in E$ between $v_i$ and $v_j$, we say that $v_i$ and $v_j$ are **adjacent**. In a complete graph, all pairs of vertices are adjacent. When $v_i$ and $v_j$ are adjacent, we say that the vertices $v_i$ and $v_j$ are **incident** with the edge $e_{ij}$.

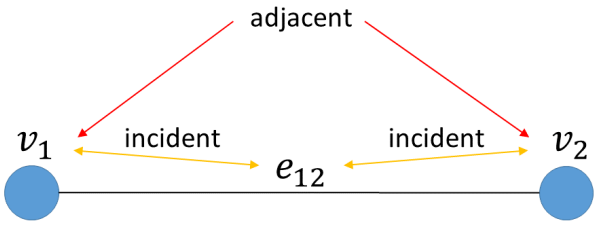



Figure A.5: Adjacency and incidence

The **degree** of a vertex $v$ of a graph $G$ is the number of edges incident with $v$, which we denote as $d_G(v)$. A loop is counted twice in $d_G(v)$. A vertex of degree $d_G(v) = 0$ and $d_G(v) = 1$ are said to be **isolated vertex** and **end vertex** of $G$, respectively. When the graph $G = V(V, E)$ is simple with no loops and multiple edges, the degree of a vertex $v_i \in V$ is equal to the number of vertices adjacent to $v_i$, namely,

$$d_G(v_i) = |\{v_j \in V : e_{ij} \in E\}|,$$

where $|A|$ denotes the cardinality of set $A$. For a complete graph $K_n$, the degrees of all vertices are equal to $n - 1$.

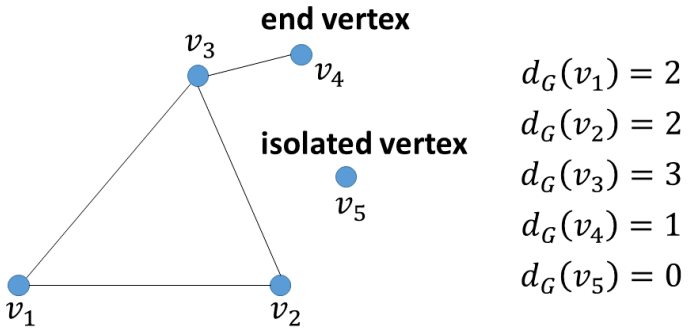



Figure A.6: Degree

The next result is often referred to as the "handshaking" lemma:

**Proposition A.3.1 (Handshaking)** *For any graph $G = (V, E)$, $\sum_{v \in V} d_G(v) = 2|E|$. (The total number of hands shaken is twice the number of handshakes.)*

**Proof.** Let $\mathbf{1}(v \sim e)$ be an indicator that takes 1 if $v$ is incident with $e$ and 0 otherwise. Then, we can write $d_G(v) = \sum_{e \in E} \mathbf{1}(v \sim e)$. Note that each edge contributes 2 to the sum of degrees: $\sum_{v \in V} \mathbf{1}(v \sim e) = 2$ holds for each $e \in E$. Thus,

$$\begin{aligned} \sum_{v \in V} d_G(v) &= \sum_{v \in V} \sum_{e \in E} \mathbf{1}(v \sim e) \\ &= \sum_{e \in E} \underbrace{\sum_{v \in V} \mathbf{1}(v \sim e)}_{=2} = 2|E|. \end{aligned}$$

■

As a corollary of the above result, we can easily show that the number of vertices of odd degree is even.

## A.4 Eulerian graph

A connected graph $G$ is an **Eulerian graph** if there exists a closed trail containing every edge of $G$. Such a trail is called an **Eulerian trail**. In other words, an Eulerian graph is a graph that can be drawn unicursally without lifting one's pencil from the paper and without repeating any lines.

Then, the problem of Königsberg bridges is equivalent to asking whether the lower-right graph in Figure A.7 has an Eulerian trail. In 1736, Euler proved that if the degree of each vertex of $G$ is an even number, then $G$ has an Eulerian trail. From this result, we can see that the upper graphs in Figure A.7 are Eulerian, and the lower graphs are not.

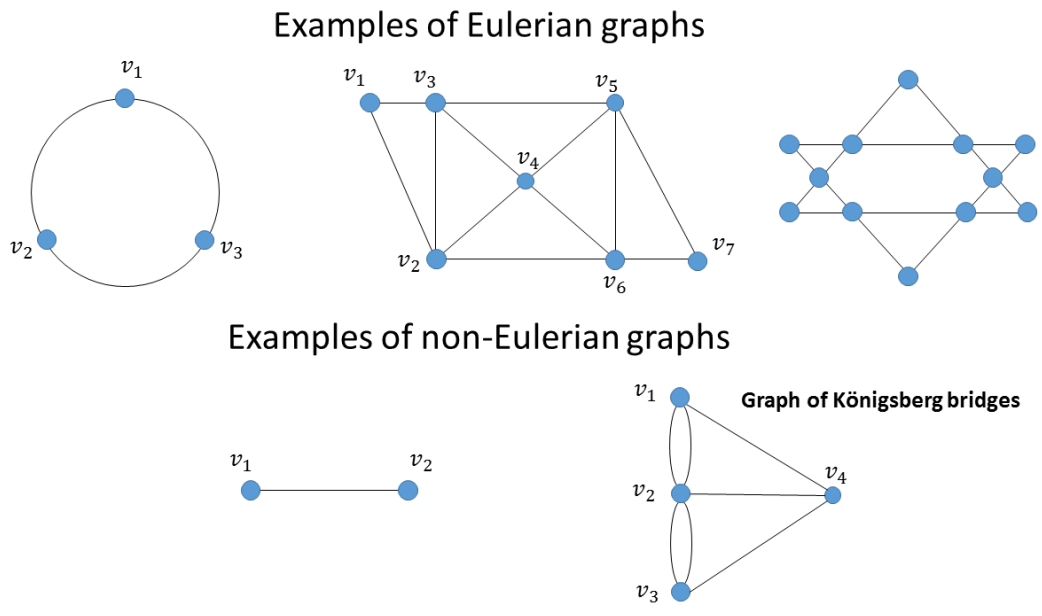



Figure A.7: Eulerian graphs and non-Eulerian graphs

Before providing the proof of this result, we prove the following useful lemma.

**Lemma A.4.1** *If $G$ is a graph in which the degree of each vertex is at least 2, then $G$ contains a cycle (recall: cycle = closed path).*

**Proof.** If $G$ has loops or multiple edges, the result trivially holds. Then, suppose that $G$ is a simple graph. Let $v$ be any vertex of $G$, and construct a walk starting from $v$, say $v \to v_1 \to v_2 \to \cdots$, by choosing $v_{i+1}$ such that $v_{i+1}$ is adjacent to $v_i$ and $v_{i+1} \neq v_{i-1}$. The existence of such walk is guaranteed by the simplicity of $G$ and the assumption made. Since the number of vertices if finite, we must eventually choose a vertex that has been chosen before. Letting $v_k$ be the first such vertex, the walk contains a cycle starting from $v_k$. ■

Now, we state Euler's (1736) result formally as follows:

**Theorem A.4.2** *A connected graph $G = (V, E)$ is Eulerian if and only if the degree of each vertex of $G$ is even.*

**Proof.** (Eulerian $\Rightarrow$ the degrees of the vertices are even.) Suppose that $P$ is an Eulerian trail of $G$. Whenever $P$ passes through a vertex, there is a contribution of 2 to the degree of that vertex. Since each edge appears exactly once in $P$, the degree of each vertex must be even.

(The degrees of the vertices are even $\Rightarrow$ Eulerian.) We prove the result by induction with respect to $q = |E|$. If $q \leq 2$, the result trivially holds. Then, suppose that $q > 2$ and the result holds for any connected graph $G' = (V', E')$ with $|E'| \leq q - 1$. Since the degrees of the vertices of $G$ are at least 2, by Lemma A.4.1, we can construct a cycle $C : v_1 \to v_2 \to \cdots \to v_1$ in $G$. Let $G_1 = (V_1, E_1)$ be a graph obtained by eliminating the edges in $C$ from $G$. Since $|E_1| \leq q - 1$, $G_1$ is Eulerian by assumption. Here, we can choose a vertex $v^* \in V_1$ which is included in the cycle $C$. Noting that an Eulerian trail can be started at any vertex and it will end at the same vertex, we can construct an Eulerian trail of $G_1$ starting and ending at $v^*$, say $P^*$. Then, the cycle $v_1 \to v_2 \to \cdots \to \underbrace{v^* \to \cdots \to v^*}_{P^*} \to \cdots \to v_1$ is an Eulerian trail of $G$. ■

## A.5 Adjacency matrix

For a graph $G = (V, E)$ with $n$ vertices $\{v_1, \ldots, v_n\}$, its **adjacency matrix** $A_n = (a_{i,j})$ is the $n \times n$ matrix whose $(i, j)$-th element is the number of edges joining $v_i$ and $v_j$. If $G$ has no loops, the diagonal elements of $A_n$ are zero. Further, if $G$ has no multiple edges, each $(i, j)$-th element of $A_n$ is an indicator whether $v_i$ is adjacent to $v_j$ or not. Adjacency matrix of an undirected graph is symmetric, while that of an directed graph can be asymmetric (Figure A.8). The adjacency matrix $A_n$ contains all the information about the graph $G$.

### Some useful properties of adjacency matrix

1. Using the adjacency matrix, the degree of vertex $v_i$ can be calculated as

$$d_G(v_i) = \sum_{j=1}^{n} a_{i,j}$$

which is the $i$-th row sum of $A_n$. In the case of directed graph, the $i$-th row sum of $A_n$ corresponds to the "out-degree" of $v_i$. By Proposition A.3.1, it follows that the total number of edges is equal to one-half of the sum of all entries of $A_n$.

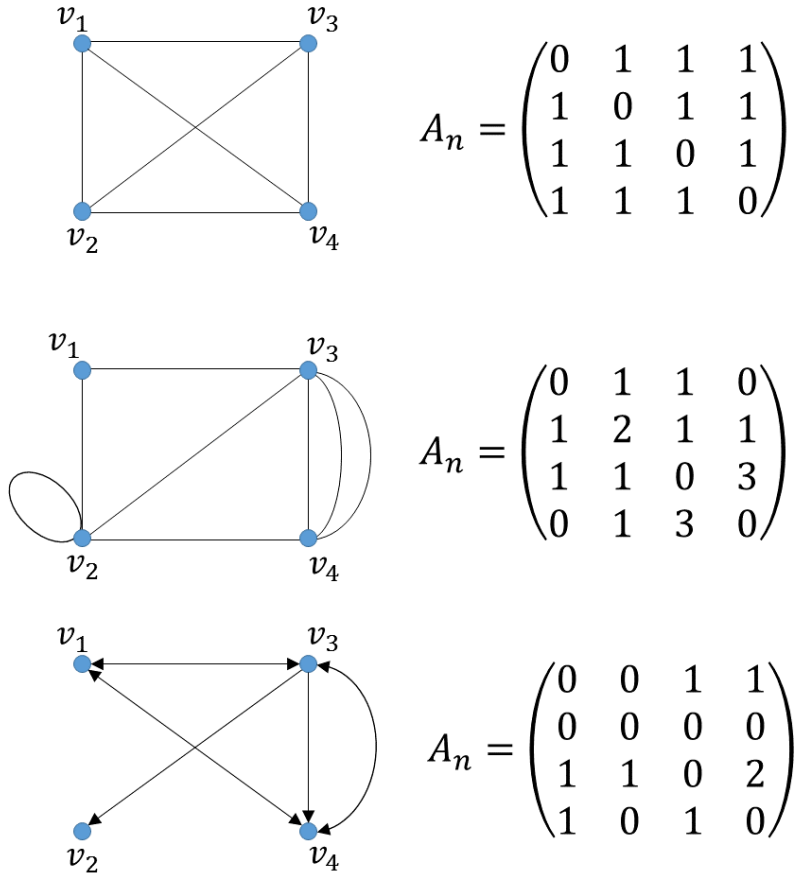



Figure A.8: Adjacency matrix

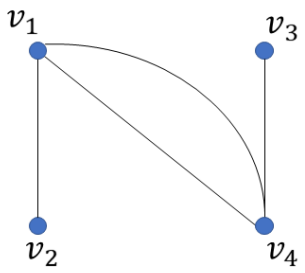
2. For a non-negative integer $\ell$, the $(i,j)$-th element of $A_n^\ell$ equals to the number of walks of length $\ell$ from $v_i$ to $v_j$. (We define $A_n^0$ to be the identity matrix $I_n$.) Recall that $a_{ik}$ tells us the number of edges from $v_i$ to $v_k$. Hence, the number of walks of length 2 from $v_i$ to $v_j$ passing through $v_k$ is obtained by $a_{ik}a_{kj}$, and the total number of such walks is $\sum_{k=1}^n a_{ik}a_{kj} = (A_n^2)_{i,j}$.

   By using this property, one can easily find the shortest path length between a given pair of vertices. For example, suppose that $(A_n)_{i,j} = (A_n^2)_{i,j} = \cdots = (A_n^{\ell-1})_{i,j} = 0$ and $(A_n^\ell)_{i,j} > 0$. Then, the shortest path length between $v_i$ and $v_j$ is $\ell$.

## A.6 Centrality in networks

In empirical network analysis, we often want to identify which agents (vertices/nodes) are most "central". The definition of "centrality" varies by contexts and purposes.

The most basic centrality measure is the **degree centrality** of a vertex $v$, which is simply defined as the degree of $v$, $d_G(v)$. The degree centrality is based on an idea that a person who has many connections is the most important person. For example, see the graph in Figure A.10. For this graph, the most central vertices in terms of the degree centrality are $v_3$ and $v_5$. However, in this example, one might consider that $v_4$ is more central than $v_3$ and $v_5$.




$$A_n = \begin{pmatrix} 0 & 1 & 0 & 2 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 2 & 0 & 1 & 0 \end{pmatrix}$$

$$A_n^2 = \begin{pmatrix} 5 & 0 & 2 & 0 \\ 0 & 1 & 0 & 2 \\ 2 & 0 & 1 & 0 \\ 0 & 2 & 0 & 5 \end{pmatrix}$$

Figure A.9: Number of walks

Here, let us denote the distance (i.e., the length of the shortest path) between the vertices $v_i$ and $v_j$ by $\Delta(v_i, v_j)$. The **eccentricity** $e(v)$ of a node $v$ in a connected graph $G = (V, E)$ is the maximum distance between $v$ and $u \in V$, that is

$$e(v) = \max_{u \in V} \Delta(v, u).$$

When $G$ is not connected, we define $e(v) = \infty$ for all $v$. Then, the reciprocal of the eccentricity value can be also used as a measure of centrality. Based on this centrality measure, the most central vertex in the graph in Figure A.10 is $v_4$.

More sophisticated centrality measure can be constructed based on the accessibility to other vertices, including not only adjacent vertices but also those that are not adjacent in the graph. Here, recall that the $(i, j)$-th element of the $\ell$-th power of the adjacency matrix $A_n^\ell$ represents the number of walks of length $\ell$ from $v_i$ to $v_j$, where $n$ is the number of vertices in the graph. Thus, letting $\mathbf{1}_n$ be the vector of ones of length $n$, the $i$-th element of $A_n^\ell \mathbf{1}_n$ is equal to:

$$(A_n^\ell \mathbf{1}_n)_i = \text{the number of vertices reachable from } v_i \text{ by length-}\ell \text{ walk.}$$

Let $\beta \in (0, 1)$ be an "propagation factor" for an increase of the length. (For example, one may interpret $\beta$ as the magnitude of peer effects; or, for another example, the rate of infection of a disease to neighboring residents.) Then, we can consider the $i$-th element of $C_n(\beta)$, where

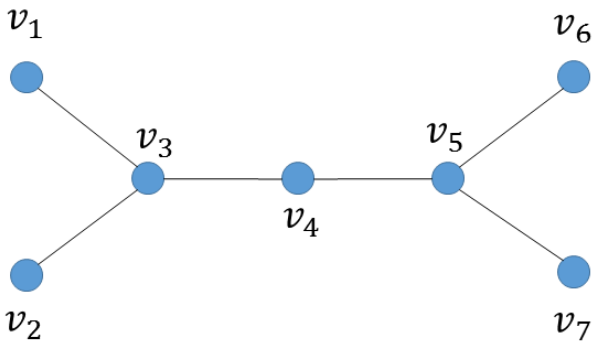
$$C_n(\beta) = \left( \sum_{\ell=0}^{\infty} \beta^\ell A_n^\ell \right) \mathbf{1}_n,$$

as the $i$-th centrality value, which is called the **Bonacich centrality**.

In the third and the fourth row of the table in Figure A.10, we report the Bonacich centrality of each vertex with $\beta = 1/2.5$ and $\beta = 1/5$, respectively. As we can see, in terms of Bonacich centrality, the centrality of the vertices 3, 4 and 5 becomes more (less) prominent if the propagation factor $\beta$ is large (small).

It is important to note that a higher-order matrix polynomial $\beta^\ell A_n^\ell$ can quickly diverge to infinity for some choice of $\beta$. When $\beta$ is appropriately chosen so that $C_n(\beta)$ exists, using the Neumann series formula, it can be exactly obtained by

$$C_n(\beta) = (I_n - \beta A_n)^{-1} \mathbf{1}_n,$$




| | Nodes: 1, 2, 6, 7 | Nodes: 3, 5 | Node: 4 |
|---|---|---|---|
| Degree centrality | 1 | 3 | 2 |
| 1/eccentricity | 1/4 | 1/3 | 1/2 |
| BC ($\beta = 1/2.5$) | 3.4 | 6.1 | 5.8 |
| BC ($\beta = 1/5.0$) | 1.4 | 1.9 | 1.8 |

Figure A.10: Centrality measures

where $I_n$ is the $n \times n$ identity matrix (see Appendix B.2 for details). An easy-to-check sufficient condition for the existence of $C_n(\beta)$ is that $\beta \cdot \max_{v \in V} d_G(v) < 1$.

In the literature on network games, the Bonacich centrality has an important role in identifying the *key player*, the player who, once removed from the network, causes the most significant impact on the aggregate outcome such as total welfare of the network (see, e.g., [Ballester et al., 2006]).

# Appendix B

# Supplementary Mathematical Notes

## B.1 Some supplementary results in probability theory.

**Lemma B.1.1** *Let $X_n \in \mathbb{R}$ and $Y_n \in \mathbb{R}$ be sequences of random variables such that $X_n \xrightarrow{p} \overline{X}$ and $Y_n \xrightarrow{p} \overline{Y}$, respectively, where both $\overline{X}$ and $\overline{Y}$ are finite. Then, it holds that $X_n Y_n \xrightarrow{p} \overline{XY}$.*

**Proof.** By the triangle inequality,

$$\begin{aligned} |X_n Y_n - \overline{XY}| &= |X_n Y_n - \overline{X} Y_n + \overline{X} Y_n - \overline{XY}| \\ &\leq |X_n Y_n - \overline{X} Y_n| + |\overline{X} Y_n - \overline{XY}| \\ &= |Y_n| \cdot |X_n - \overline{X}| + |\overline{X}| \cdot |Y_n - \overline{Y}|. \end{aligned}$$

The second term on the right-hand side clearly converges to zero in probability. For the first term, note that $|Y_n|$ is not necessarily finite for some $n$. Let $\kappa > 0$ and $\eta > 0$ be positive numbers.

$$\begin{aligned} \Pr(|Y_n| \cdot |X_n - \overline{X}| > \kappa) &= \Pr(|Y_n| \cdot |X_n - \overline{X}| > \kappa, |Y_n| \leq 1/\eta) + \Pr(|Y_n| \cdot |X_n - \overline{X}| > \kappa, |Y_n| > 1/\eta) \\ &\leq \Pr(|X_n - \overline{X}| > \kappa\eta) + \Pr(|Y_n| > 1/\eta) \\ &\leq \Pr(|X_n - \overline{X}| > \kappa\eta) + \Pr(|\overline{Y}| > 1/(2\eta)) + \Pr(|Y_n - \overline{Y}| > 1/(2\eta)). \end{aligned}$$

By choosing sufficiently small $\eta$, the terms on the right-hand side all converge to zero in probability by assumption. Thus, we have

$$\lim_{n \to \infty} \Pr(|Y_n| \cdot |X_n - \overline{X}| > \kappa) = 0.$$

Since ths choice of $\kappa$ is arbitrary, this gives the desired result. ■

**Lemma B.1.2 (Slutsky's theorem)** *Let $X_n \in \mathbb{R}$ and $Y_n \in \mathbb{R}$ be sequences of random variables such that $X_n \xrightarrow{d} \overline{X}$ and $Y_n \xrightarrow{p} c$, respectively. Then,*

$$\text{(i)} \quad X_n + Y_n \xrightarrow{d} \overline{X} + c,$$

$$\text{(ii)} \quad X_n Y_n \xrightarrow{d} c\overline{X}.$$

*More generally, $g(X_n, Y_n) \xrightarrow{d} g(\overline{X}, c)$ holds for any continuous function $g(\cdot, \cdot)$.*

The proof is omitted (see, e.g., Lemma 2.8 in [Van der Vaart, 2000]).

**Lemma B.1.3 (Jensen's inequality)** *Let $X \in \mathbb{R}$ be a random variable and $\varphi(\cdot)$ be a convex function. Then,*

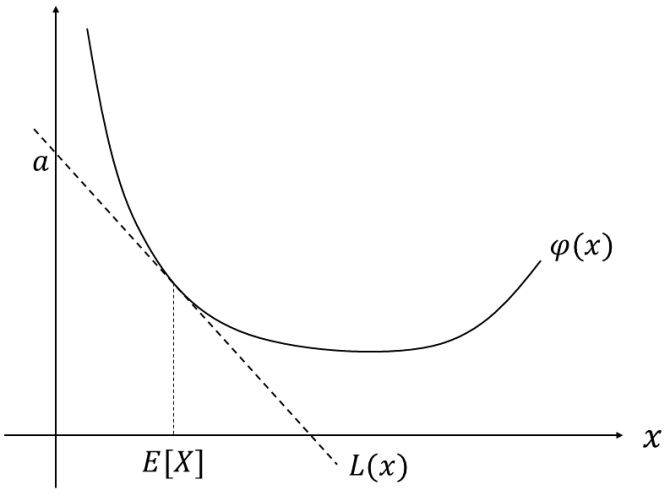
$$\mathbb{E}[\varphi(X)] \geq \varphi(\mathbb{E}[X]).$$

*If $\varphi(\cdot)$ is concave, then*

$$\mathbb{E}[\varphi(X)] \leq \varphi(\mathbb{E}[X]).$$

**Proof.** Suppose that $\varphi(\cdot)$ is a convex function. Let $L(x) = a + bx$ be a straight line which is tangent to $\varphi(x)$ at $x = \mathbb{E}[X]$. Since $\varphi(\cdot)$ is convex, for all $x$, $\varphi(x) \geq L(x)$. Thus,

$$\mathbb{E}[\varphi(X)] \geq \mathbb{E}[L(X)] = \mathbb{E}[a + bX] = a + b\mathbb{E}[X] = L(\mathbb{E}[X]) = \varphi(\mathbb{E}[X]).$$




■

**Lemma B.1.4 (Cauchy-Schwarz inequality)** *For any random variables $X \in \mathbb{R}$ and $Y \in \mathbb{R}$, we have*

$$|\mathbb{E}[XY]| \leq \sqrt{\mathbb{E}[X^2]}\sqrt{\mathbb{E}[Y^2]}.$$

**Proof.** When either $\mathbb{E}[X^2] = 0$ or $\mathbb{E}[Y^2] = 0$ is true, the result follows trivially with equality. Thus, we may consider the case where $\mathbb{E}[X^2], \mathbb{E}[Y^2] > 0$. Let us define $W = (X - cY)^2$ for some constant $c$. Since $W$ is a non-negative random variable, we must have

$$\begin{aligned} 0 &\leq \mathbb{E}[W] \\ &= \mathbb{E}[X^2] - 2c\mathbb{E}[XY] + c^2\mathbb{E}[Y^2]. \end{aligned}$$

Further, if we choose $c = \frac{\mathbb{E}[XY]}{\mathbb{E}[Y^2]}$, we have

$$\mathbb{E}[X^2] - 2c\mathbb{E}[XY] + c^2\mathbb{E}[Y^2] = \mathbb{E}[X^2] - \frac{\mathbb{E}[XY]^2}{\mathbb{E}[Y^2]}.$$

Combining these results yields the desired result. ■

## B.2 Neumann series expansion

Let $A = (a_{i,j})$ be a matrix of finite-dimension $n \times m$.[1] We use $||A||_\infty$ to denote its infinity-norm: $||A||_\infty \equiv \max_{1\le i\le n} \sum_{j=1}^{m} |a_{i,j}|$.

**Lemma B.2.1** *The infinity norm is sub-multiplicative. That is, for matrices* $\underset{n\times m}{A} = (a_{i,j})$ *and* $\underset{m\times p}{B} = (b_{j,k})$, $||AB||_\infty \le ||A||_\infty ||B||_\infty$ *holds.*

**Proof.**

$$\begin{aligned}
||AB||_\infty &= \max_{1\le i\le n} \sum_{k=1}^{p} \left| \sum_{j=1}^{m} a_{i,j} b_{j,k} \right| \\
&\le \max_{1\le i\le n} \sum_{k=1}^{p} \sum_{j=1}^{m} |a_{i,j}| \cdot |b_{j,k}| \\
&= \max_{1\le i\le n} \sum_{j=1}^{m} |a_{i,j}| \sum_{k=1}^{p} |b_{j,k}| \\
&\le \max_{1\le i\le n} \sum_{j=1}^{m} |a_{i,j}| \max_{1\le j\le m} \sum_{k=1}^{p} |b_{j,k}| = ||A||_\infty ||B||_\infty .
\end{aligned}$$

■

The series of matrices $\sum_{t=0}^{\infty} A^t$ is called the **Neumann series**. The Neumann series has the following important property:

**Lemma B.2.2 (Neumann series expansion)** *Suppose that* $A_n$ *is an* $n\times n$ *symmetric matrix such that* $||A_n||_\infty < 1$. *Then,* $I_n - A_n$ *is nonsingular, and it holds that* $(I_n - A_n)^{-1} = \sum_{t=0}^{\infty} A_n^t$.

**Proof.** The first part is an elementary result in matrix algebra. For the second part, observe that, for some $t$,

$$\begin{aligned}
(I_n - A_n)(I_n + A_n + A_n^2 + \cdots + A_n^t) &= (I_n + A_n + A_n^2 + \cdots + A_n^t) - \underbrace{A_n(I_n + A_n + A_n^2 + \cdots + A_n^t)}_{A_n + A_n^2 + \cdots + A_n^{t+1}} \\
&= I_n - A_n^{t+1}.
\end{aligned}$$

Multiplying both sides by $(I_n - A_n)^{-1}$ yields

$$\begin{aligned}
&(I_n + A_n + A_n^2 + \cdots + A_n^t) = (I_n - A_n)^{-1} - (I_n - A_n)^{-1} A_n^{t+1}, \\
&\text{and thus } (I_n - A_n)^{-1} - (I_n + A_n + A_n^2 + \cdots + A_n^t) = (I_n - A_n)^{-1} A_n^{t+1}.
\end{aligned}$$

Since the infinity norm is sub-multiplicative, noting that $||A_n||_\infty < 1$ by assumption,

$$\begin{aligned}
||(I_n - A_n)^{-1} - (I_n + A_n + A_n^2 + \cdots + A_n^t)||_\infty &= ||(I_n - A_n)^{-1} A_n^{t+1}||_\infty \\
&\le ||(I_n - A_n)^{-1}||_\infty ||A_n^{t+1}||_\infty \\
&\le ||(I_n - A_n)^{-1}||_\infty ||A_n||_\infty^{t+1} \to 0
\end{aligned}$$

as $t \to \infty$. ■

[1] This is just for simplicity. In general, we can consider a linear operator $A$ on an infinite dimensional Banach space.

## B.3 Expectation of a truncated random variable

Let $X$ be a continuous random variable with its probability distribution function $F(\cdot)$ and density function $f(\cdot)$. Further, let $a$ be a constant. Then, the conditional distribution of $X$ given that $X \leq a$ is obtained by

$$\begin{aligned} F(x \mid X \leq a) &= \Pr(X \leq x \mid X \leq a) \\ &= \frac{\Pr(X \leq \min\{x, a\})}{\Pr(X \leq a)} = \begin{cases} 1 & \text{if } x \geq a \\ \frac{F(x)}{F(a)} & \text{if } x < a \end{cases}. \end{aligned}$$

By differentiating the both sides with respect to $x$, the conditional density of $X$ given that $X \leq a$ is given by

$$f(x \mid X \leq a) = \begin{cases} 0 & \text{if } x \geq a \\ \frac{f(x)}{F(a)} & \text{if } x < a \end{cases}.$$

Thus, the expectation of $X$ conditional on $X \leq a$ is

$$\begin{aligned} \mathbb{E}[X \mid X \leq a] &= \int_{-\infty}^{\infty} x f(x \mid X \leq a) \mathrm{d}x \\ &= \frac{\int_{-\infty}^{a} x f(x) \mathrm{d}x}{F(a)}. \end{aligned}$$

By analogous arguments, we can show that the expectation of $X$ conditional on $X > a$ is obtained by

$$\begin{aligned} \mathbb{E}[X \mid X > a] &= \int_{-\infty}^{\infty} x f(x \mid X > a) \mathrm{d}x \\ &= \frac{\int_{a}^{\infty} x f(x) \mathrm{d}x}{1 - F(a)}. \end{aligned}$$

## B.4 The inverse of a partitioned matrix

Consider a partitioned matrix

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}.$$

Then, the inverse of $A$ is given by

$$A^{-1} = \begin{pmatrix} (A_{11} - A_{12}A_{22}^{-1}A_{21})^{-1} & -(A_{11} - A_{12}A_{22}^{-1}A_{21})^{-1}A_{12}A_{22}^{-1} \\ -(A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1}A_{21}A_{11}^{-1} & (A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1} \end{pmatrix}. \tag{B.4.1}$$

# Bibliography


[Akaike, 1973] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory*, pages 267–281. Akadémiai Kiadó Location Budapest, Hungary.

[Anselin, 1988] Anselin, L. (1988). *Spatial Econometrics: Methods and Models*. Springer.

[Arpino et al., 2017] Arpino, B., Benedictis, L. D., and Mattei, A. (2017). Implementing propensity score matching with network data: the effect of the general agreement on tariffs and trade on bilateral trade. *Journal of the Royal Statistical Society: Series C*, (3):537–554.

[Bajari et al., 2010] Bajari, P., Hong, H., and Ryan, S. P. (2010). Identification and estimation of a discrete game of complete information. *Econometrica*, 78(5):1529–1568.

[Bajari et al., 2015] Bajari, P., Nekipelov, D., Ryan, S. P., and Yang, M. (2015). Machine learning methods for demand estimation. *American Economic Review*, 105(5):481–85.

[Ballester et al., 2006] Ballester, C., Calvó-Armengol, A., and Zenou, Y. (2006). Who's who in networks. wanted: The key player. *Econometrica*, 74(5):1403–1417.

[Belloni et al., 2015] Belloni, A., Chernozhukov, V., Chetverikov, D., and Kato, K. (2015). Some new asymptotic theory for least squares series: Pointwise and uniform results. *Journal of Econometrics*, 186(2):345–366.

[Berry, 1992] Berry, S. T. (1992). Estimation of a model of entry in the airline industry. *Econometrica*, pages 889–917.

[Bontemps et al., 2012] Bontemps, C., Magnac, T., and Maurin, E. (2012). Set identified linear models. *Econometrica*, 80(3):1129–1155.

[Bramoullé et al., 2009] Bramoullé, Y., Djebbari, H., and Fortin, B. (2009). Identification of peer effects through social networks. *Journal of Econometrics*, 150(1):41–55.

[Bresnahan and Reiss, 1990] Bresnahan, T. F. and Reiss, P. C. (1990). Entry in monopoly market. *The Review of Economic Studies*, 57(4):531–553.

[Chen, 2007] Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. volume 6 of *Handbook of Econometrics*, pages 5549–5632. Elsevier.

[Chen and Christensen, 2015] Chen, X. and Christensen, T. M. (2015). Optimal uniform convergence rates and asymptotic normality for series estimators under weak dependence and weak conditions. *Journal of Econometrics*, 188(2):447–465.

[Ciliberto and Tamer, 2009] Ciliberto, F. and Tamer, E. (2009). Market structure and multiple equilibria in airline markets. *Econometrica*, 77(6):1791–1828.

[De Paula, 2013] De Paula, A. (2013). Econometric analysis of games with multiple equilibria. *Annual Review of Econnomics*, 5(1):107–131.

[Efron, 1979] Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26.

[Hall, 1992] Hall, P. (1992). *The bootstrap and Edgeworth expansion*. Springer.

[Harling and Castro, 2014] Harling, G. and Castro, M. C. (2014). A spatial analysis of social and economic determinants of tuberculosis in brazil. *Health & Place*, 25:56–67.

[Horowitz, 2001] Horowitz, J. L. (2001). The bootstrap. In Heckman, J. J. and Leamer, E., editors, *Handbook of Econometrics*, volume 5, chapter 52, pages 3159–3228. Elsevier.

[Horowitz, 2019] Horowitz, J. L. (2019). Bootstrap methods in econometrics. *Annual Review of Economics*, 11:193–224.

[Horowitz and Mammen, 2004] Horowitz, J. L. and Mammen, E. (2004). Nonparametric estimation of an additive model with a link function. *The Annals of Statistics*, 32(6):2412–2443.

[Hoshino, 2018] Hoshino, T. (2018). Semiparametric spatial autoregressive models with endogenous regressors: With an application to crime data. *Journal of Business & Economic Statistics*, 36(1):160–172.

[Hurwicz, 1950] Hurwicz, L. (1950). Generalization of the concept of identification. *Statistical Inference in Dynamic Economic Models*, 10:245–57.

[Imbens and Angrist, 1994] Imbens, G. W. and Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475.

[Kelejian and Prucha, 1998] Kelejian, H. H. and Prucha, I. R. (1998). A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances. *The Journal of Real Estate Finance and Economics*, 17(1):99–121.

[Kelejian and Prucha, 1999] Kelejian, H. H. and Prucha, I. R. (1999). A generalized moments estimator for the autoregressive parameter in a spatial model. *International Economic Review*, 40(2):509–533.

[Kelejian and Prucha, 2010] Kelejian, H. H. and Prucha, I. R. (2010). Specification and estimation of spatial autoregressive models with autoregressive and heteroskedastic disturbances. *Journal of Econometrics*, 157(1):53–67.

[Lee, 2004] Lee, L.-F. (2004). Asymptotic distributions of quasi-maximum likelihood estimators for spatial autoregressive models. *Econometrica*, 72(6):1899–1925.

[Lee, 2007] Lee, L.-F. (2007). Identification and estimation of econometric models with group interactions, contextual factors and fixed effects. *Journal of Econometrics*, 140(2):333–374.

[Lee et al., 2014] Lee, L.-f., Li, J., and Lin, X. (2014). Binary choice models with social network under heterogeneous rational expectations. *Review of Economics and Statistics*, 96(3):402–417.

[LeSage and Pace, 2009] LeSage, J. and Pace, R. K. (2009). *Introduction to spatial econometrics*. Chapman and Hall/CRC.

[Levitt, 1997] Levitt, S. D. (1997). Using electoral cycles in police hiring to estimate the effect of police on crime. *The American Economic Review*, pages 270–290.

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

[Li and Racine, 2007] Li, Q. and Racine, J. S. (2007). *Nonparametric econometrics: theory and practice*. Princeton University Press.

[Lucas, 1976] Lucas, R. E. (1976). Econometric policy evaluation: A critique. *Carnegie-Rochester Conference Series on Public Policy*, 1:19–46.

[Mammen, 1992] Mammen, E. (1992). Bootstrap, wild bootstrap, and asymptotic normality. *Probability Theory and Related Fields*, 93(4):439–455.

[Manski, 1975] Manski, C. F. (1975). Maximum score estimation of the stochastic utility model of choice. *Journal of Econometrics*, 3(3):205–228.

[Manski, 1993] Manski, C. F. (1993). Identification of endogenous social effects: The reflection problem. *The Review of Economic Studies*, 60(3):531–542.

[Matzkin, 2013] Matzkin, R. L. (2013). Nonparametric identification in structural economic models. *Annual Review of Econonomics*, 5(1):457–486.

[Newey and McFadden, 1994] Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. In *Handbook of econometrics*, volume 4, pages 2111–2245. Elsevier.

[Pinkse and Slade, 1998] Pinkse, J. and Slade, M. E. (1998). Contracting in space: An application of spatial statistics to discrete-choice models. *Journal of Econometrics*, 85(1):125–154.

[Said et al., 2010] Said, Y. H., Wegman, E. J., and Sharabati, W. K. (2010). Author–coauthor social networks and emerging scientific subfields. In *Data Analysis and Classification*, pages 257–268. Springer.

[Soetevent and Kooreman, 2007] Soetevent, A. R. and Kooreman, P. (2007). A discrete-choice model with social interactions: with an application to high school teen behavior. *Journal of Applied Econometrics*, 22(3):599–624.

[Stock and Yogo, 2005] Stock, J. and Yogo, M. (2005). Testing for weak instruments in linear IV regression. In *Identification and inference for econometric models: Essays in honor of Thomas Rothenberg*.

[Tamer, 2003] Tamer, E. (2003). Incomplete simultaneous discrete response model with multiple equilibria. *The Review of Economic Studies*, 70(1):147–165.

[Van der Vaart, 2000] Van der Vaart, A. W. (2000). *Asymptotic statistics*. Cambridge university press, 3 edition.

# Index