Lecture Notes For Econometrics 203B[1]

Andres Santos
Department of Economics
U.C. Los Angeles
andres@econ.ucla.edu

First Draft: Winter, 2018
This Draft: Winter, 2022

[1]These notes rely heavily on the teaching material from several friends, including Brendan Beare, Ivan Canay, Graham Elliott, Jin Hahn, Pat Kline, Azeem Shaikh, Yixiao Sun, Alex Torgovitsky, and Frank Wolak.

Contents

Chapter 1

Basic Background

Throughout the course we will assume knowledge of the material in 203A, including a basic understanding of statistics and linear algebra. As part review, part extension of material you may have already seen, we first go over some basic concepts that we will employ repeatedly throughout the course. Please keep in mind that these do not exhaust the material from 203A that we will assume you to be familiar with.

1.1 The Normal Distribution

The normal distribution plays a crucial role in asymptotic analysis. As a result, we will repeatedly work with it and you should be very familiar with its basic properties. If you are interested in really getting into the weeds on normal random variables, I recommend the truly excellent book Bogachev (1998) (it is not for the faint of heart).

Starting with the scalar case, a random variable $X \in \mathbf{R}$ is said to follow a normal distribution with mean μ and variance $\sigma^2 > 0$ if for any set $A \subset \mathbf{R}$ we have

$$P(X \in A) = \int_A \frac{1}{\sigma}\phi(\frac{x-\mu}{\sigma})dx \qquad \phi(x) \equiv \frac{1}{\sqrt{2\pi}}\exp\{-\frac{x^2}{2}\}. \tag{1.1}$$

We employ the notation "$X \sim N(\mu, \sigma^2)$" to denote that X is normally distributed with mean μ and variance σ^2. A random variable Z satisfying $Z \sim N(0,1)$ is sometimes referred to as having a "standard normal" distribution. Finally, note that if $X \sim N(\mu, \sigma^2)$, then we obtain by the change of variables $z = (x-\mu)/\sigma$ that

$$P(\frac{X-\mu}{\sigma} \in A) = P(X \in \sigma A + \mu) = \int_{\sigma A+\mu} \frac{1}{\sigma}\phi(\frac{x-\mu}{\sigma})dx = \int_A \phi(z)dz, \tag{1.2}$$

where we used that $dz = dx/\sigma$ and $(\sigma A + \mu) \equiv \{b : \sigma a + \mu \text{ for some } a \in A\}$. Notice that by (1.1), the last equation in (1.2) corresponds to the probability that a standard

normal random variable belongs to a set A. Hence, these derivations imply the following

$$\text{If } X \sim N(\mu, \sigma^2) \text{ and } Z \equiv \frac{X - \mu}{\sigma} \text{ then } Z \sim N(0, 1). \tag{1.3}$$

Moving into more general settings, we next aim to define what it means for a vector $X \in \mathbf{R}^d$ to be normally distributed. An interesting and useful way to define the normal distribution in $\mathbf{R}^d$ (and even more abstract spaces) is as follows:

Definition 1.1.1. A random variable $X \in \mathbf{R}^d$ is normally distributed if and only if $(c'X)$ is normally distributed in $\mathbf{R}$ for all vectors $c \in \mathbf{R}^d$. ■

In other words, a random variable is normally distributed in $\mathbf{R}^d$ if all its linear transformations also follow a normal distribution in $\mathbf{R}$. This definition is in fact equivalent to one you may be more familiar with, which is that a random variable $X \in \mathbf{R}^d$ is normally distributed if it has probability density function (pdf) equal to

$$\frac{1}{(\det\{2\pi\Sigma\})^{1/2}} \exp\{-\frac{1}{2}(x - \mu)'\Sigma^{-1}(x - \mu)\}, \tag{1.4}$$

where $\mu = E[X]$ and $\Sigma = \text{Var}(X) = E[(X - \mu)(X - \mu)']$. Generalizing on the notation for the scalar case, we will write "$X \sim N(\mu, \Sigma)$" to denote that a random variable $X \in \mathbf{R}^d$ is normally distributed with $\mu = E[X]$ and $\Sigma = E[(X - \mu)(X - \mu)']$ – i.e. the pdf of the random variable $X \in \mathbf{R}^d$ equals (1.4). With some abuse of terminology, whenever $\mu = 0$ (in $\mathbf{R}^d$) and $\Sigma = I_d$ for I_d the $d \times d$ identity matrix, we also say that $Z \sim N(0, I_d)$ follows a standard normal distribution in $\mathbf{R}^d$. Using identical arguments to those in (1.2) it is moreover possible to easily relate any normal distribution to the standard normal distribution. Concretely, in parallel to (1.3), we have that

$$\text{If } X \sim N(\mu, \Sigma) \text{ and } Z \equiv \Sigma^{-1/2}(X - \mu) \text{ then } Z \sim N(0, I_d). \tag{1.5}$$

Here $\Sigma^{-1/2}$ is the matrix satisfying $(\Sigma^{-1/2})(\Sigma^{-1/2}) = \Sigma^{-1}$. Be careful and note $\Sigma^{-1/2}$ does not equal the matrix formed by taking square roots of each coordinate of Σ^{-1}.

Another interesting property follows from the formula of the pdf of X in equation (1.4). For simplicity, let us suppose that $X \in \mathbf{R}^2$ and write $X = (X_1, X_2)'$. Recall that two random variables with joint pdf f_{X_1,X_2} are independent if and only if

$$f_{X_1,X_2}(x_1, x_2) = f_{X_1}(x_1) f_{X_2}(x_2) \tag{1.6}$$

where f_{X_1} and f_{X_2} are the (marginal) pdfs of X_1 and X_2 respectively. In other words, X_1 and X_2 are independent if and only if their joint pdf can be factored into the product of two marginal pdfs. However, staring at equation (1.4) we can see that in order for the pdf to factor, the matrix Σ^{-1} must be diagonal (otherwise there will be a cross term

between x_1 and x_2). Moreover, remember that Σ is the variance matrix of X, so that

$$\Sigma = \begin{bmatrix} \text{Var}(X_1) & \text{Cov}(X_1, X_2) \\ \text{Cov}(X_1, X_2) & \text{Var}(X_2) \end{bmatrix}. \tag{1.7}$$

We conclude that Σ is diagonal if and only if $\text{Cov}(X_1, X_2) = 0$. Since in addition Σ is diagonal if and only if Σ^{-1} is diagonal, we obtain the following property

$$\text{If } (X_1, X_2) \sim N(\mu, \Sigma), \text{ then } X_1, X_2 \text{ are independent if and only if } \text{Cov}(X_1, X_2) = 0.$$

Finally, we get to a fundamental property of the normal distribution that we will employ heavily in this course. Suppose $X \in \mathbf{R}^d$ follows a normal distribution so that $X \sim N(\mu, \Sigma)$. For any $m \times d$ non-random matrix Ω, then note that $\Omega X \in \mathbf{R}^m$ is also a random variable. It is not hard to show employing Definition 1.1.1 that ΩX must be normally distributed. Through a bit more work, it is in fact possible to show that

$$\text{If } X \sim N(\mu, \Sigma), \text{ then } \Omega X \sim N(\Omega\mu, \Omega\Sigma\Omega'). \tag{1.8}$$

This is such a fundamental property of the normal distribution that it is definitional – indeed note the essential connection to our original Definition 1.1.1.

1.2 Conditional Expectations

In Econ 203A or other introductory courses to econometrics you may have seen a conditional expectation defined in terms of conditional pdfs. For our purposes, it will prove to be more useful to think of conditional expectations as projections.

In what follows we let $Y \in \mathbf{R}$ and $X \in \mathbf{R}^d$ be random variables. Consider the problem of approximating Y as well as possible employing only functions of X. More precisely suppose we aim to solve the following minimization problem

$$\inf_{f:\mathbf{R}^d \to \mathbf{R}} E[(Y - f(X))^2] \text{ s.t. } E[f^2(X)] < \infty. \tag{1.9}$$

It turns out that the infimum in (1.9) is in fact attained at a "unique" solution with "uniqueness" understood up to sets of measure zero.

Lemma 1.2.1. *Let $(Y, X) \in \mathbf{R} \times \mathbf{R}^d$ satisfy $E[Y^2] < \infty$. Then, there is $f^\star$ with*

$$E[(Y - f^\star(X))^2] \leq E[(Y - f(X))^2] \tag{1.10}$$

for all $f : \mathbf{R}^d \to \mathbf{R}$ satisfying $E[f^2(X)] < \infty$. Moreover, it follows that $E[(Y - f^\star(X))^2] = E[(Y - f(X))^2]$ if and only if $P(f(X) = f^\star(X)) = 1$ and $f^\star$ satisfies

(1.10) *if and only if for any function* $f : \mathbf{R}^d \to \mathbf{R}$ *with* $E[f^2(X)] < \infty$ *we have*

$$E[(Y - f^\star(X))f(X)] = 0. \tag{1.11}$$

PROOF: Omitted. If interested, you should look into the theory of projections in Hilbert spaces. Luenberger (1969) is a beautiful introduction into the subject. ■

Given Lemma 1.2.1 we may then define $E[Y|X]$ to be the "best" approximation to Y by functions of X. In other words we may set $E[Y|X]$ to equal

$$E[Y|X] \equiv \arg \min_{f:\mathbf{R}^d \to \mathbf{R}} E[(Y - f(X))^2] \text{ s.t. } E[f^2(X)] < \infty. \tag{1.12}$$

While this definition suffices for our purposes, it is worth noting that it is a bit more restrictive than required. Concretely, more generally we do not need Y to satisfy $E[Y^2] < \infty$ for $E[Y|X]$ to be well defined. In fact, in order for $E[Y|X]$ to be well defined we only need $E[|Y|] < \infty$. Nonetheless, the definition in (1.12) is convenient for us since we will be working with random variables that have second moments.

The definition of $E[Y|X]$ as a solution to an optimization problem (as in (1.12)) together with Lemma 1.2.1 enables us to establish various useful properties of the conditional expectation. We will repeatedly rely on the following properties:

1. If $Y = f(X)$ then $E[Y|X] = f(X)$.
2. If Y and X are independent, then $E[Y|X] = E[Y]$.
3. If $Y, Z \in \mathbf{R}$ have finite second moments, then $E[Y + Z|X] = E[Y|X] + E[Z|X]$.
4. If $E[f^2(X)] < \infty$, then $E[Yf(X)|X] = E[Y|X]f(X)$.
5. If $X = (X_1, X_2)$, then $E[E[Y|X_1, X_2]|X_1] = E[Y|X_1]$ (law of iterated expectations).

Example 1.2.1. A useful example to keep in mind involves the conditional expectations of bivariate normal random variables. Specifically, suppose that $(Y, X) \in \mathbf{R}^2$ follow a bivariate normal distribution and define $\beta \equiv \text{Cov}\{X, Y\}/\text{Var}\{X\}$. Then note that

$$\begin{aligned}
&\text{Cov}\{(Y - E[Y]) - (X - E[X])\beta, X\} \\
&\quad = \text{Cov}\{Y - X\beta, X\} = E[(Y - X\beta)X] - E[(Y - X\beta)]E[X] \\
&\quad = E[YX] - E[X^2]\beta - E[Y]E[X] + (E[X])^2\beta = \text{Cov}\{Y, X\} - \text{Var}\{X\}\beta = 0.
\end{aligned} \tag{1.13}$$

Recall, however, that: (i) linear functions of normal random variables are also normal, and hence $(Y - X\beta, X)$ are jointly normal; and (ii) normal random variables that have

zero covariance are also independent. Therefore, $(Y - E[Y]) - (X - E[X])\beta)$ and X are independent by the manipulations in (1.13), which implies that

$$E[(Y - E[Y]) - (X - E[X])\beta|X] = E[(Y - E[Y]) - (X - E[X])\beta] = 0.$$

In other words, we have just shown that $E[Y|X] = E[Y] + (X - E[X])\beta$. Crucially, for bivariate normal random variables, the conditional means are linear. ■

Remark 1.2.1. The conditional expectation is a special case of a "projection" operator. Projection operators play a fundamental role in linear models. Their role, however, is sometimes "hidden" in the background behind what appears to be arbitrary algebra. We'll highlight the roles of projections when studying ordinary least squares. ■

1.3 Asymptotic Analysis

Throughout Econ 203B we will rely heavily on asymptotic approximations rather than finite sample distributions whose computation may require unrealistic assumptions.

1.3.1 Preliminaries

For simplicity we will for the most part assume that the data is independent and identically distributed (abbreviated i.i.d.). Generalizing the material in our class to non i.i.d. settings is generally straightforward under appropriate conditions – i.e. provided we rule out that the dependence across observations is "too strong".

Formally, we say a sample $\{X_i\}_{i=1}^n$ is i.i.d. if the variables X_i are independent across i and each observation X_i follows exactly the same distribution. Let P denote the distribution of any X_i. Then, by independence, the *joint* distribution of the data $\{X_i\}_{i=1}^n$ is given by $P^n \equiv \bigotimes_{i=1}^n P$. This notation emphasizes that every parameter that is identifiable from the joint distribution $\{X_i\}_{i=1}^n$ must in fact be a function of P.

Before proceeding, we present a very useful inequality.

Lemma 1.3.1. *Let $X \in \mathbf{R}^d$ be a random variable and $\psi : \mathbf{R}^d \to \mathbf{R}_+$ be a positive function. Then for any $\epsilon > 0$ it follows that*

$$P(\psi(X) > \epsilon) \leq \frac{1}{\epsilon} E[\psi(X)].$$

Proof: This useful inequality is surprisingly easy to prove. Just note that we have

$$P(\psi(X) > \epsilon) = E[1\{\psi(X) > \epsilon\}] \leq E[\frac{\psi(X)}{\epsilon} 1\{\psi(X) > \epsilon\}] \leq \frac{1}{\epsilon} E[\psi(X)] \tag{1.14}$$

where the first equality follows by definition, the second is implied by $1 < \psi(X)/\epsilon$ whenever $\psi(X) > \epsilon$ and the third follows from $\psi(X) \geq 0$. ■

The inequality in Lemma 1.3.1 is often referred to as Markov's inequality. A special case of this inequality mysteriously has its own name: "Chebychev's inequality". In particular, Chebychev's inequality states that if $X \in \mathbf{R}$, then

$$P(|X| > \epsilon) = P(X^2 > \epsilon^2) \leq \frac{1}{\epsilon^2} E[X^2].$$

where the inequality simply follows from Lemma 1.3.1 with $\psi(x) = x^2$.

1.3.2 Convergence in Probability

The goal of asymptotic approximations is often to gain an understanding of the finite sample behavior of a statistic of interest. Heuristically, suppose there is a sequence of random variables Y_n whose distribution is unknown. However, suppose we can argue that as "n becomes large", Y_n gets "close" to another random variable X whose distribution we do know. Then, we could use the distribution of X to approximate that of Y_n at any n and hope that such approximation is accurate enough for our purposes.

But what does it mean for Y_n to get "close" to X? As it turns out there are multiple ways to define "getting close". An intuitive one is *convergence in probability.*

Definition 1.3.1. *Suppose $X \in \mathbf{R}^d$ and $Y_n \in \mathbf{R}^d$ is a sequence of random variables. Then we say Y_n converges in probability to X, and write $Y_n \xrightarrow{p} X$, if for any $\epsilon > 0$*

$$\lim_{n \to \infty} P(\|Y_n - X\| > \epsilon) = 0.$$

The rationale behind this definition is quite simple. If Y_n and X were non-random, then we would say that Y_n converges to X if $\|Y_n - X\| \to 0$. The only difference is that when (Y_n, X) are random, their distance $\|Y_n - X\|$ is itself a random variable. Thus, we instead demand that the probability that $\|Y_n - X\|$ be larger than ϵ converge to zero for any $\epsilon > 0$. Note that instead we could have required $\|Y_n - X\| \to 0$ with probability one – this concept, known as almost sure convergence, is stronger but we will not use it.

Our next result is often referred to as the law of large numbers

Lemma 1.3.2. *If $\{X_i\}_{i=1}^n$ is a scalar i.i.d. sequence satisfying $E[|X|] < \infty$, then*

$$\frac{1}{n} \sum_{i=1}^{n} X_i \xrightarrow{p} E[X].$$

PROOF: A full proof of this result is remarkably challenging. For simplicity, we will

therefore impose the stronger assumption that $\text{Var}\{X\} < \infty$. For any $\epsilon > 0$ then

$$P(|\frac{1}{n}\sum_{i=1}^{n} X_i - E[X]| > \epsilon) \leq \frac{1}{\epsilon^2} E[(\frac{1}{n}\sum_{i=1}^{n} X_i - E[X])^2]$$
$$= \frac{1}{\epsilon^2}\text{Var}\{\frac{1}{n}\sum_{i=1} X_i - E[X]\} = \frac{1}{\epsilon^2}\frac{1}{n^2}\sum_{i=1}^{n}\text{Var}\{X_i\} = \frac{1}{\epsilon^2}\frac{\text{Var}\{X_i\}}{n} \quad (1.15)$$

by Markov's inequality and the i.i.d. assumption on $\{X_i\}_{i=1}^n$. Taking limits, we get

$$\lim_{n\to\infty} P(|\frac{1}{n}\sum_{i=1}^{n} X_i - E[X]| > \epsilon) = 0,$$

which establishes the desired result. ■

1.3.3 Convergence in Distribution

Convergence in probability is undoubtedly a very useful concept. However, if our goal is to approximate the behavior of $Y_n \in \mathbf{R}^d$ by that of $X \in \mathbf{R}^d$, then we do not actually need Y_n to be "close" to X in $\mathbf{R}^d$. Instead, all we need is for Y_n to be "close" to X in the sense that the "behavior" of Y_n is close to the "behavior" of X. This observation leads to the concept of *convergence in distribution* for sequences of random variables.

For the next definition recall for any set $A \subseteq \mathbf{R}^d$, A^o denotes its interior, $\bar{A}$ its closure, and $\partial A = \bar{A} \setminus A^o$ is called the "boundary" of A.

Definition 1.3.2. *Let $X_n \in \mathbf{R}^d$ and $X \in \mathbf{R}^d$ be random variables. We then say X_n converges in distribution to X if for all sets A such that $P(X \in \partial A) = 0$ we have*

$$\lim_{n\to\infty} P(X_n \in A) = P(X \in A).$$

In other words, X_n converges in distribution to X if the probability it assigns to certain sets converges to the probability that X assigns to those sets. Notice that the qualifier that $P(X \in \partial A) = 0$ is important. In a previous definition you might have seen convergence in distribution in terms of convergence of the corresponding cdfs at continuity points. Specifically, such definition states that $X_n \xrightarrow{d} X$ if and only if

$$\lim_{n\to\infty} P(X_n \leq t) = P(X \leq t)$$

whenever t is a continuity point of the cdf. Notice this requirement maps into definition 1.3.2 by letting $A = (\infty, t]$ and noting that $\partial A = \{t\}$ and that the cdf of X is continuous at t precisely when $P(X = t) = 0$ (i.e. $P(X \in \partial A) = 0$).

Our next example illustrates the importance of demanding $P(X \in \partial A) = 0$

Example 1.3.1. Suppose X is the degenerate random variable satisfying $P(X = 0) = 1$ and let X_n be a sequence satisfying $X_n \sim N(0, \sigma_n^2)$ with $\sigma_n^2 > 0$ and $\sigma_n^2 \to 0$. It is straightforward to show that $X_n \xrightarrow{p} X$, which as we show below implies $X_n \xrightarrow{d} X$ (see Lemma 1.3.3). However, note that if we set $A = \{0\}$, then $X_n \sim N(0, \sigma_n^2)$ implies that

$$0 = \lim_{n\to\infty} P(X_n \in A) \leq P(X \in A) = 1. \tag{1.16}$$

Notice though that $A^o = \emptyset$, $\bar{A} = \{0\}$, and $\partial A = \{0\}$, therefore failure of convergence in (1.16) occurs precisely at a set A for which $P(X \in \partial A) > 0$. ■

The main tool for showing convergence in distribution is the central limit theorem.

Theorem 1.3.1. *If $\{X_i\}_{i=1}^n$ is an i.i.d. sample with $X_i \in \mathbf{R}^d$ and $E[\|X_i\|^2] < \infty$, then*

$$\sqrt{n}\{\frac{1}{n}\sum_{i=1}^{n} X_i - E[X]\} \xrightarrow{d} Z \tag{1.17}$$

where $Z \sim N(0, \Sigma)$ and $\Sigma \equiv E[(X - E[X])(X - E[X])']$.

PROOF: Omitted, though if you have not seen it before, then it is worth reading. ■

When a sequence X_n convergences in distribution to a limit X we only care about the distribution of X. Therefore, for notational simplicity we sometimes omit writing $X_n \xrightarrow{d} X$ and instead write the distribution of X in the "right hand side". For instance, instead of writing equation (1.17) we would use the expression

$$\sqrt{n}\{\frac{1}{n}\sum_{i=1}^{n} X_i - E[X]\} \xrightarrow{d} N(0, \Sigma).$$

An important issue to keep in mind when working with convergence in distribution is that *marginal* convergence of the sequences does not imply *joint* convergence. More concretely, suppose $X_n, Y_n \in \mathbf{R}^d$ and we have shown that $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} Y$. Such a conclusion *does not imply* that $(X_n, Y_n) \xrightarrow{d} (X, Y)$. As a counterexample let $X_n \sim N(0, 1)$ and $Y_n = (-1)^n X_n$. Then both X_n and Y_n have a standard normal distribution for all n, and hence they both converge *marginally*; i.e. $X_n \xrightarrow{d} N(0, 1)$ and $Y_n \xrightarrow{d} N(0, 1)$. However, (X_n, Y_n) do not converge jointly. For instance just note that

$$P(X_n = Y_n) = 1 - (-1)^n$$

fails to have a limit at all (since it alternates between one and zero).

We conclude this section with three useful lemmas that relate convergence in probability to converge in distribution. Their proof is a challenging but accessible problem.

Lemma 1.3.3. *If $X_n \xrightarrow{p} X$, then $X_n \xrightarrow{d} X$.*

Lemma 1.3.4. *If $X_n \xrightarrow{d} X$ and $P(X = c) = 1$ for some constant c (i.e. X is degenerate), then it follows that $X_n \xrightarrow{p} X$.*

Lemma 1.3.5. *If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$ for some constant c, then $(X_n, Y_n) \xrightarrow{d} (X, c)$.*

In summary, Lemma 1.3.3 shows that convergence in probability implies convergence in distribution – i.e. convergence in probability is a stronger requirement than convergence in distribution. In words, if X_n approaches X (i.e. $X_n \xrightarrow{p} X$), then it must also begin to "behave" like X (i.e. $X_n \xrightarrow{d} X$). In turn, Lemma 1.3.4 provides a partial converse, showing that if X_n converges in distribution to a degenerate random variable, then it must also converge in probability. In general, however, $X_n \xrightarrow{d} X$ *does not imply* $X_n \xrightarrow{p} X$. Finally, Lemma 1.3.5 establishes that marginal convergence can be combined to show joint convergence when one of the variables converges in probability.

1.3.4 Some Tools

The law of large numbers and the central limit theorem provide us with the basic building blocks for deriving asymptotic results. However, they concern only sample means and we will be interested in analysing more complex estimators. Fortunately, two key tools allow us to heavily leverage these simple building blocks.

The first such result is the continuous mapping theorem.

Theorem 1.3.2. *Suppose $X_n \in \mathbf{R}^d$ satisfies $X_n \xrightarrow{d} X$. If $f : \mathbf{R}^d \to \mathbf{R}^q$ is continuous on $\mathbb{D}_0$ and $P(X \in \mathbb{D}_0) = 1$, then it follows that $f(X_n) \xrightarrow{d} f(X)$ in $\mathbf{R}^q$.*

A common application of the continuous mapping theorem is sometimes called Slutsky's theorem. For illustrative purposes, suppose that $X_n \in \mathbf{R}^d$ and $Y_n \in \mathbf{R}$ satisfy $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$ for some constant $c \neq 0$. We are interested in the asymptotic distribution of the random variable X_n/Y_n. To this end, let us define $f : \mathbf{R}^d \times \mathbf{R}$ to be given by $f(x, y) = x/y$ for any $(x, y) \in \mathbf{R}^d \times \mathbf{R}$. Notice that f is continuous at any (x, y) with $y \neq 0$. Moreover, by Lemma 1.3.5 we know $(X_n, Y_n) \xrightarrow{d} (X, c)$. Therefore, applying the continuous mapping theorem (i.e. Theorem 1.3.2) we can conclude

$$\frac{X_n}{Y_n} = f(X_n, Y_n) \xrightarrow{d} f(X, c) = \frac{X}{c}.$$

These type of arguments are sometimes summarized as Slutsky's Theorem.

Theorem 1.3.3. *Suppose $X_n \in \mathbf{R}^d$ and $Y_n \in \mathbf{R}$ satisfy $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$ for some constant c. Then it follows that $X_n Y_n \xrightarrow{d} Xc$ and $X_n + Y_n \xrightarrow{d} X + c$.*

The second crucial tool for asymptotic analysis is the Delta method.

Theorem 1.3.4. *Suppose $X_n \in \mathbf{R}^d$, $\theta_0 \in \mathbf{R}^d$, and $f : \mathbf{R}^d \to \mathbf{R}^q$ is differentiable at θ_0 with $\nabla f(\theta_0)$ the $q \times d$ matrix of partial derivatives at θ_0. If $\sqrt{n}\{X_n - \theta_0\} \xrightarrow{d} Z$, then*

$$\sqrt{n}\{f(X_n) - f(\theta_0)\} \xrightarrow{d} \nabla f(\theta_0) Z. \tag{1.18}$$

In particular, note that if $\sqrt{n}\{X_n - \theta_0\} \xrightarrow{d} N(0, \Sigma)$, then we obtain that

$$\sqrt{n}\{f(X_n) - f(\theta_0)\} \xrightarrow{d} N(0, \nabla f(\theta_0) \Sigma \nabla f(\theta_0)'). \tag{1.19}$$

As an example of the delta method, suppose $\{Z_i, Y_i\}_{i=1}^n$ is an i.i.d. sample with (Z, Y) having finite second moments. For Σ the covariance matrix of (Z, Y) and $\bar{Z}_n$ and $\bar{Y}_n$ the sample means, the central limit theorem implies

$$\sqrt{n} \begin{pmatrix} \bar{Z}_n - E[Z] \\ \bar{Y}_n - E[Y] \end{pmatrix} \xrightarrow{d} N(0, \Sigma).$$

Suppose we are interested in the parameter $E[Z]E[Y]$ which we estimate by $\bar{Z}_n \bar{Y}_n$. The Delta method (i.e. Theorem 1.3.4) gives us a simple way to obtain the desired limiting distribution. Let $f : \mathbf{R}^2 \to \mathbf{R}$ be given by $f(z, y) = zy$ for any $(z, y) \in \mathbf{R}^2$ and set $\theta_0 = (E[Z], E[Y])'$. Then $\nabla f(\theta_0)$ is simply a row vector equal to $\nabla f(\theta_0) = (E[Y], E[Z])$. The first and second displays then state the equivalent conclusion that

$$\sqrt{n}\{\bar{Z}_n \bar{X}_n - E[Z]E[X]\} \xrightarrow{d} N(0, (E[Z], E[Y])\Sigma(E[Z], E[Y])').$$

While we are not ready to explore these concepts in more length in 203B, there are a couple of observations that are worth making in case you are interested in further studying the topic. For an in depth analysis I would recommend the (very) challenging but extraordinary book van der Vaart and Wellner (1996).

1. While we have worked with random variables in $\mathbf{R}^d$, we actually rarely used the properties of Euclidean spaces in our definitions and results.

2. Many of the definitions we stated only require a topology. This leads to a general study of random processes in topological spaces.

3. However, key complications arise in infinite dimensional spaces due to a possible lack of separability and compactness being a more demanding requirement.

1.3.5 Some Notation

In deriving asymptotic distributions we will employ tools like the Delta method and the continuous mapping theorem to leverage results such as law of large numbers and the

central limit theorem. For these manipulations, it will be convenient to rely on some very powerful notation known as stochastic order symbols.

First, for establishing asymptotic distributions, we will often ignore terms that converge in probability to zero. Our first piece of notation is precisely for such terms.

Definition 1.3.3. *For a sequence X_n, we write $X_n = o_p(1)$ to denote that $X_n \xrightarrow{p} 0$.*

For instance, if $\{X_i\}_{i=1}^n$ is an i.i.d. sample with $E[|X|] < \infty$, then the law of large numbers implies $\bar{X}_n \xrightarrow{p} E[X]$. We may write this result in the middle of a proof as

$$\bar{X}_n = E[X] + o_p(1), \tag{1.20}$$

where the term "$o_p(1)$" mathematically refers to the term $\bar{X}_n - E[X]$. The "$o_p(1)$" notation is particularly helpful together with the "$O_p(1)$" notation, which is defined as

Definition 1.3.4. *A sequence of random variables X_n satisfies $X_n = O_p(1)$ whenever*

$$\lim_{M \uparrow \infty} \lim_{n \to \infty} P(|X_n| > M) = 0.$$

Heuristically, "$X_n = O_p(1)$" means that the sequence of random variables is bounded in probability. As an example, suppose $\{Y_i\}_{i=1}^n$ is an i.i.d. sample with $E[Y^2] < \infty$ and let $\bar{Y}_n$ denote the sample mean. Then by the central limit theorem we obtain

$$\sqrt{n}\{\bar{Y}_n - E[Y]\} \xrightarrow{d} N(0, \sigma^2),$$

where $\sigma^2 \equiv \text{Var}\{Y\}$. Let $Z \sim N(0, \sigma^2)$ and then note that by the definition of convergence in distribution we obtain for any scalar positive M that

$$\lim_{n \to \infty} P(|\sqrt{n}\{\bar{Y}_n - E[X]\}| > M) = P(|Z| > M). \tag{1.21}$$

Therefore, taking limits for $M \uparrow \infty$ on both sides of equation (1.3.5) we can conclude

$$\lim_{M \uparrow \infty} \lim_{n \to \infty} P(|\sqrt{n}\{\bar{Y}_n - E[X]\}| > M) = \lim_{M \uparrow \infty} P(|Z| > M) = 0. \tag{1.22}$$

Thus, the sequence of random variables $\sqrt{n}\{\bar{Y}_n - E[Y]\}$ is bounded in probability and we can write $\sqrt{n}\{\bar{Y}_n - E[Y]\} = O_p(1)$.

1.4 Problems

1. Provide an example of a random vector $X = (X_1, X_2) \in \mathbf{R}^2$ such that $\text{Cov}(X_1, X_2) = 0$ but X_1 and X_2 are not independent.

2. Let $X \in \mathbf{R}^d$ be normally distributed, and Ω be a $m \times d$ matrix. Use Definition 1.1.1 to show ΩX is normally distributed as a random variable in $\mathbf{R}^m$.

3. Show that properties 1-5 of conditional expectations stated in Section 1.2 hold by using only Lemma 1.2.1 and the definition of $E[Y|X]$ as satisfying (1.12).

4. Let $(Y_n, X) \in \mathbf{R}^2$ be jointly normally distributed with $E[Y_n] = E[X] = 0$. Show that if $E[Y_n^2]E[X^2] - (E[Y_n X])^2 \to 0$ and $E[XY_n] \to E[X^2]$, then $Y_n \xrightarrow{p} X$ (Hint: Think of $E[Y_n|X]$).

5. Build an example with $Y_n \xrightarrow{d} X$ but Y_n not converging in probability to X.

6. Prove Lemmas 1.3.3, 1.3.4, and 1.3.5.

7. Let $\{X_i\}_{i=1}^n$ be an i.i.d. sample with $X \in \mathbf{R}$ satisfying $E[X^2] < \infty$. For $\bar{X}_n$ the sample mean, derive the asymptotic distribution of:

   (a) $\sqrt{n}\{\sqrt{|\bar{X}_n|} - \sqrt{|E[X]|}\}$ when $E[X] > 0$.

   (b) $n^{\frac{1}{4}}\{\sqrt{|\bar{X}_n|} - \sqrt{|E[X]|}\}$ when $E[X] = 0$.

8. Show that if $X_n = O_p(1)$ and $Y_n = o_p(1)$, then $X_n Y_n = o_p(1)$.

Chapter 2

Linear Regression

We next turn to the study of the linear regression model. Specifically, suppose we observe an i.i.d. sample $\{Y_i, X_i\}_{i=1}^n$ where $Y_i \in \mathbf{R}$ and $X_i \in \mathbf{R}^d$. The ordinary least squares estimator (OLS for short) is defined as the minimizer

$$\hat{\beta}_n = \arg\min_{b \in \mathbf{R}^d} \frac{1}{n} \sum_{i=1}^{n} (Y_i - X_i' b)^2. \tag{2.1}$$

In this chapter we will study the properties of this indispensable tool of applied work.

2.1 The Estimand

The obvious starting point of our analysis is to understand what exactly $\hat{\beta}_n$ is estimating. The analogy principle suggests that $\hat{\beta}_n$ must be an estimator for the population analogue to the optimization problem in (2.1). We will show such a statement to be essentially true, and that $\hat{\beta}_n$ indeed estimates the parameter β_0 solving

$$\beta_0 = \arg\min_{b \in \mathbf{R}^d} E[(Y - X'b)^2]. \tag{2.2}$$

Notice that the optimization problem in (2.2) is convex, and thus the solution is fully characterized by the first order conditions. By simple differentiation we obtain

$$\frac{\partial}{\partial b} E[(Y - X'b)^2] = -2E[(Y - X'b)X]. \tag{2.3}$$

Thus, if β_0 is the minimizer of (2.2) it is also, by convexity of the criterion function, characterized as the solution to the first order conditions – i.e. as the value of b that zeros (2.3). Plugging in we therefore can also characterize β_0 through the equation

$$E[(Y - X'\beta_0)X] = 0. \tag{2.4}$$

Equation (2.4) is a system of d linear equations with d unknowns. We know that a solution must exist but it may not be unique. In order for the solution to be unique we need $E[XX']$ to be full rank, which implies it is invertible. In fact under invertibility of the $d \times d$ matrix $E[XX']$ we can obtain a closed form solution for β_0, which equals

$$E[YX] - E[XX'\beta_0] = 0 \Rightarrow \beta_0 = \{E[XX']\}^{-1}E[YX]. \tag{2.5}$$

As a consequence of (2.4), β_0 is sometimes introduced through the linear "model":

$$Y = X'\beta_0 + \epsilon \qquad E[\epsilon X] = 0, \tag{2.6}$$

where $\epsilon = Y - X'\beta_0$ and therefore the condition that $E[\epsilon X] = 0$ is equivalent to condition (2.4). These specifications of β_0 are of course equivalent with each other. However, you should be wary that sometimes (2.6) can cause confusion as people try attribute some "mystical" meaning to ϵ, by giving it names such as "unobserved heterogeneity." In contrast, the characterizations in (2.3) and (2.4) emphasize β_0 is a simple function of the distribution of (Y, X). In summary, we now know what we are estimating, but not necessarily why we should care. The parameter β_0 may in fact have an important economic interpretation. Simply remember that such importance is not inherent in (2.6), which is in fact just *the definition* of the parameter β_0. Instead, we must carefully determine in each applications whether β_0 is (or is not) the parameter we should consider. Below, we examine different possible interpretations of β_0 to help inform our intuition.

2.1.1 Interpretation: Conditional Mean

Suppose $Y \in \mathbf{R}$ and $X \in \mathbf{R}^d$ and we assume that the conditional mean of Y given X is linear – i.e. for some $\gamma_0 \in \mathbf{R}^d$, the variables (Y, X) satisfy

$$E[Y|X] = X'\gamma_0. \tag{2.7}$$

Recall that by the law of iterated expectations $E[E[Y|X]X] = E[YX]$, and therefore

$$E[(Y - X'\gamma_0)X] = E[(Y - E[Y|X])X] = E[YX] - E[YX] = 0. \tag{2.8}$$

In other words, if the conditional mean is linear as in (2.7), then γ_0 is also a solution to the moment restrictions defining β_0. Equivalently, from (2.4) and (2.8) we obtain

$$0 = E[(Y - X'\gamma_0)X] = E[(Y - X'\beta_0)X] = E[XX'](\gamma_0 - \beta_0), \tag{2.9}$$

which implies that $\beta_0 = \gamma_0$ whenever the matrix $E[XX']$ is full rank. Therefore, if (2.7) holds and $E[XX']$ is invertible, then the OLS estimand β_0 in fact corresponds to the conditional expectation parameter – i.e. $E[Y|X] = X'\beta_0$.

We have already seen an example of linear conditional means: When (Y, X) are jointly normally distributed the conditional mean of Y given X is linear (as in (2.7)). While in general assuming the conditional mean is linear can be restrictive, there is one more important class of examples where it is satisfied.

Example 2.1.1. Suppose $Y \in \mathbf{R}$ and Z is a discrete random variable taking values in $\{z_1, \ldots, z_d\}$. Since Z can only take d values, $E[Y|Z]$ can only take d values as well, which we write as $(E[Y|Z = z_1], \ldots, E[Y|Z = z_d])'$. Therefore we can write

$$E[Y|Z] = \sum_{i=1}^{d} 1\{Z = z_i\} \times E[Y|Z = z_i],$$

where $1\{\cdot\}$ is the indicator function – i.e. $1\{Z = z_i\} = 1$ if $Z = z_i$ and equals zero if $Z \neq z_i$. Equivalently, we can define $X = (1\{Z = z_1\}, \ldots, 1\{Z = z_d\})'$ and $\beta_0 \equiv (E[Y|Z = z_1], \ldots, E[Y|Z = z_d])$ to arrive at the linear specification

$$E[Y|X] = X'\beta_0.$$

In general, whenever the conditioning variable is discrete, it is possible to write the conditional mean as a linear function of indicator variables for each possible value of the conditioning model. This specification is sometimes referred to as a *saturated* model. ■

It is tempting, and a common mistake, to interpret a linear conditional mean as implying that X is somehow causal. This misunderstanding arises from the relation

$$\frac{\partial}{\partial X} E[Y|X] = \frac{\partial}{\partial X} X'\beta_0 = \beta_0,$$

which suggests β_0 informs how changing X changes Y. This is nonsense. The conditional mean, just like the mean itself, is simply a summary description of the joint distribution of (Y, X) and has no intrinsic causal interpretation.

Example 2.1.2. A folk legend in some western cultures is that storks deliver babies. While ridiculous, we may let Y denote the birth rate in a country and X its stork population. The conditional mean $E[Y|X]$ then simply describes the average fertility rate in countries with a stork population equal to X. While I trust you do not find this relationship causal, you may be surprised to find that the OLS estimates of a regression of Y on X can be highly significant (Matthews, 2000). ■

2.1.2 Interpretation: (Good?) Approximation

We have seen that whenever the conditional mean of Y given X is linear, OLS gives us a consistent estimator for the conditional mean. However, aside from special cases, the

assumption that the conditional mean is linear is unrealistic. It is then natural to ask what exactly OLS is estimating in such cases. As the next Lemma shows, we can then think of the OLS estimand as the best "linear" approximation to the conditional mean.

Lemma 2.1.1. *If $Y \in \mathbf{R}$, $X \in \mathbf{R}^d$, $E[Y^2] < \infty$, and $E[XX']$ is full rank, then*

$$\beta_0 = \arg\min_{b \in \mathbf{R}^d} E[(Y - X'b)^2] = \arg\min_{b \in \mathbf{R}^d} E[(E[Y|X] - X'b)^2]. \tag{2.10}$$

PROOF: We first note that as previously discussed, $E[Y^2] < \infty$ and $E[XX']$ being full rank imply β_0 is indeed the unique minimizer of the criterion functions. Next observe

$$\begin{aligned} E[(Y - X'b)^2] &= E[(Y - E[Y|X] + E[Y|X] - X'b)^2] \\ &= E[(Y - E[Y|X])^2] + E[(E[Y|X] - X'b)^2] + 2E[(Y - E[Y|X])(E[Y|X] - X'b)] \end{aligned}$$

for any $b \in \mathbf{R}^d$. However, by the law of iterated expectations we may drop the last term (i.e. the "cross" product term). Therefore, we get that

$$\begin{aligned} \beta_0 &\equiv \arg\min_{b \in \mathbf{R}^d} E[(Y - X'b)^2] \\ &= \arg\min_{b \in \mathbf{R}^d} \{E[(Y - E[Y|X])^2] + E[(E[Y|X] - X'b)^2]\} \qquad (2.11) \\ &= \arg\min_{b \in \mathbf{R}^d} E[(E[Y|X] - X'b)^2] \qquad (2.12) \end{aligned}$$

where the final inequality follows from the fact that $E[(Y - E[Y|X])^2]$ does not actually depend on b; i.e. the *minimum* changes, but the *minimizer* does not. ■

Lemma 2.1.1 is extremely useful in conceptually understanding the estimand of the OLS coefficient $\hat{\beta}_n$. Because of Lemma 2.1.1, the OLS estimand is sometimes referred as the *best linear predictor* and abbreviated BLP – be cautious, however, specially if you go into IO, that BLP is sometimes also employed to refer to a widely used model of demand introduced in Berry et al. (1995).

Another point of caution is that the title *best linear predictor* is misconstrued as a justification that OLS is always interesting. To see whether this makes sense, let's dissect the definition more closely to see what each word means:

linear: Linear here means in the parameters β_0 and not in the regressors. For example, suppose $Y \in \mathbf{R}$, and $X \in \mathbf{R}$, and consider the following models

$$\begin{aligned} Y &= Z_1'\beta_1 + \epsilon_1 && \text{with } Z_1 = (1, X) \text{ and } E[\epsilon_1 Z_1] = 0 \\ Y &= Z_2'\beta_2 + \epsilon_2 && \text{with } Z_2 = (1, \sin(X), X^2) \text{ and } E[\epsilon_2 Z_2] = 0 \\ Y &= Z_3'\beta_3 + \epsilon_3 && \text{with } Z_3 = (X, \sin(X), \exp(X)) \text{ and } E[\epsilon_3 Z_3] = 0. \end{aligned}$$

All these models would be considered to be linear, because they are linear in the parameter being estimated (β_1, β_2, and β_3 respectively). ■

best: Best here means best *given the specified model*. For example, in the three specifications above, the OLS estimand would be considered to be the BLP in the sense that it is approximating $E[Y|X]$ as well as possible given the chosen specification (i.e. either $(1, X)$ or $(1, \sin(X), X^2)$, or $(X, \sin(X), \exp(X))$). In particular, note that as a result being the "best" given the chosen model provides no guarantees that the approximation is actually any good! ■

As a final point, it is important to note that Lemma 2.1.1 is a statement about $X'\beta_0$ being an approximation to the conditional mean $E[Y|X]$. Note that we cannot conclude from Lemma 2.1.1 that as a result β_0 is a "best" approximation to the derivative of $E[Y|X]$. In other words, β_0 is often not the solution to the optimization problem

$$\min_{b \in \mathbf{R}^d} E[\|\nabla E[Y|X] - b\|^2] \tag{2.13}$$

In fact, the solution to (2.13) is $E[\nabla E[Y|X]]$ which is often referred to as the "average derivative." Nonetheless, Yitzhaki (1996) shows that the OLS estimand β_0 can still be thought of as a weighted average of $\nabla E[Y|X]$. The weights themselves can be estimated from the data, which may help in interpreting β_0. However, we note that if you are interested in the average derivative $E[\nabla E[Y|X]]$, then there are numerous nonparametric methods for estimating it as well; see Powell et al. (1989).

2.1.3 Interpretation: Got a Model?

The discussion in Sections 2.1.1 and 2.1.2 hopefully highlight that OLS always has an interpretation as a *summary statistic*. In other words, it is indispensable as a data reduction tool that enables to uncover correlations present in the data.

In order to interpret an OLS coefficient as causal, however, it is important to have an explicit model in mind and show the parameter of interest indeed maps into the OLS coefficient. This exercise further enables us to think more precisely of what the errors in the regression are and what exactly the parameter β_0 means.

Example 2.1.3. Suppose we observe a sample of firms with output $Y_i \in \mathbf{R}_+$ and single input $X_i \in \mathbf{R}_+$. We are interested in estimating the production function, which equals

$$Y_i = \exp\{U_i\} X_i^{\beta_0} \tag{2.14}$$

where U_i is unobserved productivity factor of firm i. Notice that if we differentiate in

(2.14) with respect to X and plug into (2.14) we obtain the equality

$$\frac{dY/dX}{Y/X} = \frac{\exp\{U\}X^{\beta_0-1}\beta_0}{\exp\{U\}X^{\beta_0}/X} = \beta_0. \tag{2.15}$$

Hence, β_0 is equal to the output elasticity. Note, however, that β_0 does not tell us how a change in inputs X translates into a change in output Y for a particular firm since U remains unknown. To estimate β_0 we may take logs to arrive at the linear specification

$$\log(Y_i) = \beta_0 \log(X_i) + U_i. \tag{2.16}$$

It is at this point tempting to assume $E[U\log(X)] = 0$ and estimate β_0 by OLS. But is assuming $E[U\log(X)] = 0$ a credible assumption? It is here that having a model helps. Suppose the firm operates in a competitive environment so it takes the price P_Y at which it can sell its output and the prices P_X at which it can buy its inputs as given. A profit maximizing firm will therefore select inputs to maximize profits

$$\max_x \{P_Y \exp\{U_i\}x^{\beta_0} - xP_X\}. \tag{2.17}$$

However, working off the first order condition for maximizing (2.17) we arrive at

$$P_Y \exp\{U_i\}x^{\beta_0-1}\beta_0 - P_X = 0 \;\Rightarrow \log(\frac{\beta_0 P_Y}{P_X}) + U_i = (1-\beta_0)\log(x). \tag{2.18}$$

In particular, since the quantity X_i we observe was chosen by the firm to maximize its profits (i.e. to solve (2.18)) we find that X_i satisfies

$$\log(X_i) = \frac{1}{1-\beta_0}\{\log(\frac{\beta_0 P_Y}{P_X}) + U_i\},$$

which makes the assumption $E[U_i\log(X_i)] = 0$ implausible. Intuitively, more productive firms produce more leading to them requiring more inputs. Hence, OLS does not allow us to estimate β_0 consistently. To see procedure that address this problem in the estimation of production functions see, among others, Olley and Pakes (1996), Ackerberg et al. (2015), and Gandhi et al. (2011). ■

Example 2.1.4. Suppose we conduct a clinical trial in which participants are randomly selected into a treatment. We are concerned with a particular health outcome such as their blood pressure. For each individual i we define

$$\begin{aligned} Y_i(0) &= \text{ outcome without treatment} \\ Y_i(1) &= \text{ outcome with treatment} \end{aligned}$$

and set $D_i = 1$ if individual i was assigned to treatment and $D_i = 0$ otherwise. By construction, for each individual i we do not observe both $Y_i(0)$ and $Y_i(1)$ but rather

only the outcome corresponding to their treatment status – i.e. we observe

$$Y_i \equiv Y_i(1)D_i + (1 - D_i)Y_i(0).$$

Now suppose treatment is randomly assigned, so that $(Y_i(0), Y_i(1))$ are independent of D_i. We then obtain from the definition of Y_i that

$$\begin{aligned} E[Y_i|D_i] &= D_i E[Y_i(1)|D_i = 1] + (1 - D_i)E[Y_i(0)|D_i = 0] \\ &= E[Y_i(0)] + D_i E[Y_i(1) - Y_i(0)]; \end{aligned} \tag{2.19}$$

notice the similarity to the saturated model in Example 2.1.1. Since the conditional mean is linear, we can conclude from our discussion in Section 2.1.1 that a regression of Y_i on $X_i \equiv (1, D_i)$ estimates $E[Y_i(0)]$ (the coefficient in front of the constant) and $E[Y_i(1) - Y_i(0)]$ (the coefficient in front of D_i). In particular, note $E[Y_i(1) - Y_i(0)]$ is known as the *average treatment effect* (ATE for short). ■

2.2 The Estimator: The Basics

We next turn to study the basics of the OLS estimator itself, which recall was defined in (2.1). To this end, we start with basic notation and a geometric interpretation.

2.2.1 Some Notation

Throughout, we let $\{Y_i, X_i\}_{i=1}^n$ be an i.i.d. sample with $Y_i \in \mathbf{R}$ and $X_i \in \mathbf{R}^d$. Unless otherwise stated, X_i will be treated as a column vector; i.e. of dimension $d \times 1$. It will be immensely helpful to "stack" the observations into vectors and matrices so that we can rely on linear algebra for our analysis. To this end, we therefore define

$$\mathbb{Y}_n \equiv \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \qquad \mathbb{X}_n \equiv \begin{pmatrix} X_1' \\ \vdots \\ X_n' \end{pmatrix} \tag{2.20}$$

and note that $\mathbb{Y}_n$ is then a $n \times 1$ vector, while $\mathbb{X}_n$ is a $n \times d$ matrix.

For any column vector $(a_1, \ldots, a_q)' = a \in \mathbf{R}^q$ we denote its Euclidean norm by

$$\|a\| \equiv \{\sum_{i=1}^{q} a_i^2\}^{1/2} = \sqrt{a'a}. \tag{2.21}$$

Recall that the OLS estimator was defined in (2.1) as minimizing the sum of squared

residuals. Moreover, by some basic algebra we have for any $b \in \mathbf{R}^d$ that

$$\mathbb{X}_n b = \begin{pmatrix} X_1' b \\ \vdots \\ X_n' b \end{pmatrix}. \tag{2.22}$$

Therefore, employing our definitions in (2.20), the definition of $\hat{\beta}_n$, and (2.21) we obtain

$$\hat{\beta}_n \equiv \arg\min_{b \in \mathbf{R}^d} \frac{1}{n} \sum_{i=1}^{n} (Y_i - X_i' b)^2 = \arg\min_{b \in \mathbf{R}^d} (\mathbb{Y}_n - \mathbb{X}_n b)'(\mathbb{Y}_n - \mathbb{X}_n b) = \arg\min_{b \in \mathbf{R}^d} \|\mathbb{Y}_n - \mathbb{X}_n b\|^2, \tag{2.23}$$

where we dropped the scaling by $1/n$ since it does not affect the minimizing value.

Given the introduced notation we obtain a closed form solution for $\hat{\beta}_n$.

Lemma 2.2.1. *If the $d \times d$ matrix $\mathbb{X}_n' \mathbb{X}_n$ is invertible, then it follows that*

$$\hat{\beta}_n = (\mathbb{X}_n' \mathbb{X}_n)^{-1} \mathbb{X}_n' \mathbb{Y}_n = \left(\frac{1}{n} \sum_{i=1}^{n} X_i X_i' \right)^{-1} \frac{1}{n} \sum_{i=1}^{n} X_i Y_i.$$

PROOF: We start with the definition of $\hat{\beta}_n$ as a minimizer of the criterion

$$Q_n(b) \equiv \frac{1}{n} \sum_{i=1}^{n} (Y_i - X_i' b)^2. \tag{2.24}$$

Notice that Q_n is a convex function of b. Therefore, its minimizers are characterized as the set of zeros of the first order condition. Since we have that

$$\frac{\partial}{\partial b} Q_n(b) = -\frac{2}{n} \sum_{i=1}^{n} (Y_i - X_i' b) X_i \tag{2.25}$$

it follows that the set of minimizers must satisfy the d orthogonality conditions

$$\frac{1}{n} \sum_{i=1}^{n} (Y_i - X_i' b) X_i = 0. \tag{2.26}$$

Observe that (2.26) consists of d linear equations in a d dimensional unknown (note the similarity to (2.4)). Since $\hat{\beta}_n$ minimizes Q_n, it satisfies (2.26) and hence

$$\frac{1}{n} \sum_{i=1}^{n} X_i Y_i = \frac{1}{n} \sum_{i=1}^{n} X_i X_i' \hat{\beta}_n \quad \text{or equivalently} \quad \mathbb{X}_n' \mathbb{Y}_n = \mathbb{X}_n' \mathbb{X}_n \hat{\beta}_n, \tag{2.27}$$

where the equivalent characterization follows from simple algebra. Notice that so far we have not used the assumption that $\mathbb{X}_n' \mathbb{X}_n$ is invertible. It follows from (2.26) and (2.27) that invertibility of $\mathbb{X}_n' \mathbb{X}_n$ is needed to ensure that $\hat{\beta}_n$ is the *unique* solution to (2.26) –

i.e. Q_n has a *unique* minimizer. Under invertibility, result (2.27) implies that

$$\hat{\beta}_n = \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} X_i Y_i \quad \text{or equivalently} \quad \hat{\beta}_n = \left(\mathbb{X}_n'\mathbb{X}_n\right)^{-1}\mathbb{X}_n'\mathbb{Y}_n, \tag{2.28}$$

which establishes the claim of the Lemma. ■

As a last piece of notation it will also be helpful to write the true residuals as $U_i = Y_i - X_i'\beta_0$ (for β_0 the solution to (2.2)) and the fitted residuals as $\hat{U}_i \equiv Y_i - X_i'\hat{\beta}_n$. We stack both the true and fitted residuals into the $n \times 1$ column vectors

$$\mathbb{U}_n \equiv \begin{pmatrix} U_1 \\ \vdots \\ U_n \end{pmatrix} \qquad \hat{\mathbb{U}}_n \equiv \begin{pmatrix} \hat{U}_1 \\ \vdots \\ \hat{U}_n \end{pmatrix}. \tag{2.29}$$

2.2.2 Geometric Intuition

The characterization of $\hat{\beta}_n$ in Lemma 2.2.1 follows from basic calculus and linear algebra. In understanding the properties of OLS, however, it is helpful to have a "deeper" geometric interpretation in terms of projection.

We start with a small tangent on the properties of projections. Suppose $a \in \mathbf{R}^q$ is a vector and $\mathbf{V} \subset \mathbf{R}^q$ is a vector subspace of $\mathbf{R}^q$. Recall a vector subspace is a set that is closed under linear combinations of its elements; i.e. $\mathbf{V}$ is such that

$$\text{If } a_1, a_2 \in \mathbf{V} \text{ and } \lambda_1, \lambda_2 \in \mathbf{R}, \text{ then } a_1\lambda_1 + a_2\lambda_2 \in \mathbf{V}. \tag{2.30}$$

An important property of a linear subspace is that it can be represented in terms of an orthogonal basis. In particular, if $\mathbf{V} \subseteq \mathbf{R}^q$ is a linear subspace, then there exists a set $\{v_1, \ldots, v_m\}$ with (i) $0 \neq v_j \in \mathbf{R}^q$, (ii) $m \leq q$, (iii) $v_j'v_k = 0$ whenever $j \neq k$, and (iv)

$$\mathbf{V} = \{a \in \mathbf{R}^q : a = \sum_{j=1}^{m} a_j v_j \text{ for some } a_1, \ldots, a_m \in \mathbf{R}\}. \tag{2.31}$$

The number m is known as the dimension of the linear subspace $\mathbf{V}$. It is also helpful to introduce the *orthogonal complement* of $\mathbf{V}$ which we denote by $\mathbf{V}^\perp$ and define as

$$\mathbf{V}^\perp \equiv \{a \in \mathbf{R}^q : a'\tilde{a} = 0 \text{ for all } \tilde{a} \in \mathbf{V}\}. \tag{2.32}$$

Finally, for any $a \in \mathbf{R}^q$, we define its projection onto $\mathbf{V}$ as the vector $\Pi_{\mathbf{V}}a \in \mathbf{V}$ satisfying

$$\Pi_{\mathbf{V}}a \equiv \arg\min_{v \in \mathbf{V}} \|a - v\|^2. \tag{2.33}$$

The following Lemma summarizes some of the key properties of projections that we will use in our analysis. We leave its proof as an exercise.

Lemma 2.2.2. *Let $\mathbf{V} \subseteq \mathbf{R}^q$ be a vector subspace of $\mathbf{R}^q$. Then it follows*

(i) $\Pi_{\mathbf{V}} a$ *is the unique minimizer of* (2.33).

(ii) If $\{v_1, \ldots, v_m\}$ *is an orthogonal basis for* $\mathbf{V}$*, then it follows that*

$$\Pi_{\mathbf{V}} a = \sum_{j=1}^{m} \frac{v_j' a}{\|v_j\|^2} v_j.$$

(iii) $\mathbf{V}^{\perp}$ *is a vector subspace of* $\mathbf{R}^q$.

(iv) $(a - \Pi_{\mathbf{V}} a) \in \mathbf{V}^{\perp}$ *and in fact* $\Pi_{\mathbf{V}^{\perp}} a = a - \Pi_{\mathbf{V}} a$.

The following example gives a basic illustration of these concepts.

Example 2.2.1. We work with $\mathbf{R}^2$ so that we are able to draw a picture (see Figure 2.1). Suppose the linear subspace $\mathbf{V}$ is given by

$$\mathbf{V} = \{v \in \mathbf{R}^2 : v = \lambda(1,1)' \text{ for some } \lambda \in \mathbf{R}\},$$

and note that $\{(1,1)'\}$ is a basis for $\mathbf{V}$ – in fact, any vector of the form $(\lambda, \lambda)'$ with $0 \neq \lambda \in \mathbf{R}$ is a basis for $\mathbf{V}$. Let $a = (0,2)$, which note does not belong to $\mathbf{V}$. By Lemma 2.2.2(ii) we can use $(1,1)$ is a basis for $\mathbf{V}$ to find the projection of a onto $\mathbf{V}$ is equal to

$$\frac{(1,1)'(0,2)}{\|(1,1)\|^2} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

We then obtain that $a - \Pi_{\mathbf{V}} a = (-1,1)'$, which can be verified to be in $\mathbf{V}^{\perp}$ (in agreement with Lemma 2.2.2(iii)). In fact, since $\mathbf{V}^{\perp} = \{v \in \mathbf{R}^2 : v = \lambda(-1,1)' \text{ for some } \lambda \in \mathbf{R}\}$, we obtain $(-1,1) = \Pi_{\mathbf{V}^{\perp}} a$ as claimed by Lemma 2.2.2(iv). ■

What does this have to do with OLS? To appreciate the connection, define the set

$$\mathbf{V} \equiv \{v \in \mathbf{R}^n : v = \mathbb{X}_n b \text{ for some } b \in \mathbf{R}^d\}, \tag{2.34}$$

which note is a linear subspace of $\mathbf{R}^n$ – $\mathbf{V}$ as in (2.34) is often referred to as the *column space* of $\mathbb{X}_n$ because it corresponds to the set of vectors in $\mathbf{R}^n$ that can be written as linear combinations of the columns of $\mathbb{X}_n$. Employing our results in (2.23), we may then rewrite the minimization problem defining the OLS estimator as

$$\min_{b \in \mathbf{R}^d} \|\mathbb{Y}_n - \mathbb{X}_n b\|^2 = \min_{v \in \mathbf{V}} \|\mathbb{Y}_n - v\|^2 \tag{2.35}$$

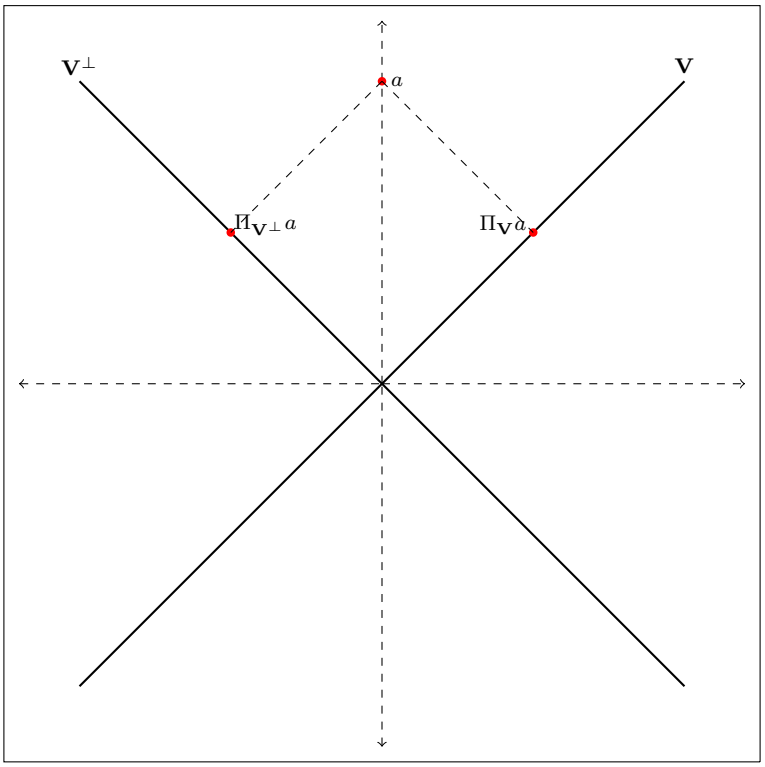

Figure 2.1: Simple Projection Example

By Lemma 2.2.2(ii), it is straightforward to find the projection of $\mathbb{Y}_n$ onto $\mathbf{V}$ if we have an orthogonal basis of $\mathbf{V}$. To this end, define the $n \times d$ matrix

$$[v_1, \ldots, v_d] \equiv \mathbb{X}_n (\mathbb{X}_n' \mathbb{X}_n)^{-1/2}, \tag{2.36}$$

where note $\mathbb{X}_n (\mathbb{X}_n' \mathbb{X}_n)^{-1/2}$ is a $n \times d$ matrix, and hence $v_j \in \mathbf{R}^n$. Finally, define

$$\mathbb{P}_n \equiv \mathbb{X}_n (\mathbb{X}_n' \mathbb{X}_n)^{-1} \mathbb{X}_n' \qquad \mathbb{M}_n \equiv I_n - \mathbb{P}_n, \tag{2.37}$$

where I_n is the $n \times n$ identity matrix – notice $\mathbb{P}_n$ and $\mathbb{M}_n$ are both $n \times n$.

Given this notation we obtain the following results.

Lemma 2.2.3. *Let $\mathbb{X}_n' \mathbb{X}_n$ be invertible, and $\mathbf{V}$, $\{v_1, \ldots, v_d\}$, and $\mathbb{P}_n$ be as defined in (2.34), (2.36), and (2.37) respectively. Then, it follows that*

(i) $\{v_1, \ldots, v_d\}$ is an orthogonal basis for $\mathbf{V}$.

(ii) For any $a \in \mathbf{R}^n$ we have $\Pi_{\mathbf{V}} a = \mathbb{P}_n a$ and $\Pi_{\mathbf{V}^\perp} a = \mathbb{M}_n a$

Proof: For part (i), suppose $v \in \mathbf{V}$, which by (2.34) means we can write $v = \mathbb{X}_n b$ for

some $b \in \mathbf{R}^d$. Letting $\tilde{b} \equiv (\mathbb{X}_n'\mathbb{X}_n)^{1/2}b$ we then obtain

$$[v_1, \ldots, v_d]\tilde{b} = \mathbb{X}_n(\mathbb{X}_n'\mathbb{X}_n)^{-1/2}(\mathbb{X}_n'\mathbb{X}_n)^{1/2}b = \mathbb{X}_n b = v, \tag{2.38}$$

thus verifying any $v \in \mathbf{V}$ can indeed be written as a linear combination of $\{v_1, \ldots, v_d\}$. To verify orthogonality of this basis, note that

$$[v_1, \ldots, v_d]'[v_1, \ldots, v_d] = (\mathbb{X}_n'\mathbb{X}_n)^{-1/2}\mathbb{X}_n'\mathbb{X}_n(\mathbb{X}_n'\mathbb{X}_n)^{-1/2} = I_d \tag{2.39}$$

for I_d the $d \times d$ identity matrix. In other words, $v_j'v_k = 0$ if $j \neq k$ and $v_j'v_j = 1$.

To verify part(ii), we use Lemma 2.2.2(ii). Notice that by (2.39) we have that $\|v_j\|^2 = 1$. Therefore, from Lemma 2.2.2(ii) and some algebra we obtain that

$$\begin{aligned}\Pi_{\mathbf{V}}a = \sum_{j=1}^{d}(v_j'a)v_j = [v_1, \ldots, v_d]&\begin{pmatrix} v_1'a \\ \vdots \\ v_d'a \end{pmatrix} \\ &= \{\mathbb{X}_n(\mathbb{X}_n'\mathbb{X}_n)^{-1/2}\}\{(\mathbb{X}_n'\mathbb{X}_n)^{-1/2}\mathbb{X}_n'a\} = \mathbb{P}_n a. \end{aligned} \tag{2.40}$$

Part(ii) of the Lemma therefore follows from (2.40) and Lemma 2.2.2(iv). ■

In particular, Lemma 2.2.3 tells us the solution to the projection problem in (2.35):

$$\mathbb{P}_n\mathbb{Y}_n = \arg\min_{v \in \mathbf{V}} \|\mathbb{Y}_n - v\|^2. \tag{2.41}$$

However, examining the matrix $\mathbb{P}_n$ more closely and employing Lemma 2.2.1 we obtain

$$\mathbb{P}_n\mathbb{Y}_n = \mathbb{X}_n(\mathbb{X}_n'\mathbb{X}_n)^{-1}\mathbb{X}_n'\mathbb{Y}_n = \mathbb{X}_n\hat{\beta}_n. \tag{2.42}$$

In other words: (i) For any $a \in \mathbf{R}^n$, $\mathbb{P}_n a$ equals the projection of a onto $\mathbf{V}$ (Lemma 2.2.3(ii)), (ii) When projecting $\mathbb{Y}_n$ onto $\mathbf{V}$ we obtain $\mathbb{P}_n\mathbb{Y}_n$ which is precisely the *predicted/fitted* values from OLS, and (iii) The residuals from this regression then satisfy

$$\hat{\mathbb{U}}_n = \mathbb{Y}_n - \mathbb{X}_n\hat{\beta}_n = \mathbb{Y}_n - \mathbb{P}_n\mathbb{Y}_n = \{I_n - \mathbb{P}_n\}\mathbb{Y}_n = \mathbb{M}_n\mathbb{Y}_n. \tag{2.43}$$

In other words, the residuals from the regression are the projection of $\mathbb{Y}_n$ onto $\mathbf{V}^\perp$ – loosely speaking, $\mathbf{V}^\perp$ is the component of $\mathbb{Y}_n$ that cannot be explained by $\mathbb{X}_n$.

We conclude this discussion with a simple set of properties of $\mathbb{P}_n$ and $\mathbb{M}_n$

Lemma 2.2.4. *If $\mathbb{X}_n'\mathbb{X}_n$ is invertible, then if follows that*

(i) $\mathbb{P}_n\mathbb{P}_n = \mathbb{P}_n$, $\mathbb{M}_n\mathbb{M}_n = \mathbb{M}_n$, *and* $\mathbb{P}_n\mathbb{X}_n = \mathbb{X}_n$.

(ii) $\mathbb{P}_n\mathbb{M}_n = 0$ *and* $\mathbb{M}_n\mathbb{X}_n = 0$.

The proof follows form simple algebra and is left as an exercise. The intuition of (2.2.4) is straightforward in terms of projections (recall Figure 2.1). Remember $\mathbb{P}_n$ projects any vector a onto the column space of $\mathbb{X}_n$ (i.e. $\mathbf{V}$ as in (2.34)). Therefore, $\mathbb{P}_n(\mathbb{P}_n a)$ equals the projection of $\mathbb{P}_n a$ onto the column space of $\mathbb{X}_n$. However, since $\mathbb{P}_n a$ is already in the column space of $\mathbb{X}_n$, it follows it equals its own projection, i.e. $\mathbb{P}_n\mathbb{P}_n a = \mathbb{P}_n a$. Since the argument holds for any a, we must have $\mathbb{P}_n\mathbb{P}_n = \mathbb{P}_n$. Similar arguments imply $\mathbb{P}_n\mathbb{X}_n = \mathbb{X}_n$ and $\mathbb{M}_n\mathbb{M}_n = \mathbb{M}_n$. In turn, part(ii) of Lemma 2.2.4 follows from $\mathbb{M}_n$ projecting onto the orthogonal complement of the column space of $\mathbb{X}_n$ (i.e. $\mathbf{V}^\perp$ as in (2.32) for $\mathbf{V}$ as defined in (2.34)), which implies $\mathbb{M}_n a = 0$ whenever $a \in \mathbf{R}^n$ can be written as a linear combination of the columns in $\mathbb{X}_n$.

2.2.2.1 Partitioned Regression

Employing the interpretation of OLS as a projection, it is straightforward to obtain a number of interesting results. One such result is the Frisch-Waugh-Lovell Theorem.

Suppose the regressor $X_i \in \mathbf{R}^d$ is partitioned into two components, so that

$$X_i = (X_{1i}', X_{2i}')'$$

with $X_{1i} \in \mathbf{R}^{d_1}$ and $X_{2i} \in \mathbf{R}^{d_2}$ (note $d_1 + d_2 = d$). We can then similarly decompose the OLS estimator $\hat{\beta}_n$ into the coefficients corresponding to X_{1i} (denoted $\hat{\beta}_{1n}$) and the coefficients corresponding to X_{2i} (denoted $\hat{\beta}_{2n}$), and we thus write

$$\begin{pmatrix} \hat{\beta}_{1n} \\ \hat{\beta}_{2n} \end{pmatrix} = \arg \min_{b_1 \in \mathbf{R}^{d_1}, b_2 \in \mathbf{R}^{d_2}} \frac{1}{n} \sum_{i=1}^{n} (Y_i - X_{1i}' b_1 - X_{2i}' b_2)^2. \tag{2.44}$$

As in Section 2.2.1 we can stack observations into matrices, and we therefore define

$$\mathbb{X}_{1n} \equiv \begin{pmatrix} X_{11}' \\ \vdots \\ X_{1n}' \end{pmatrix} \qquad \mathbb{X}_{2n} \equiv \begin{pmatrix} X_{21}' \\ \vdots \\ X_{2n}' \end{pmatrix}. \tag{2.45}$$

Notice that $\mathbb{X}_{1n}$ is an $n \times d_1$ matrix, $\mathbb{X}_{2n}$ is an $n \times d_2$ matrix, and $[\mathbb{X}_{1n}\ \mathbb{X}_{2n}] = \mathbb{X}_n$. Following Section 2.2.2 it is also helpful to define matrices that project onto the column space of $\mathbb{X}_{1n}$ and $\mathbb{X}_{2n}$ and their respective orthogonal complements. Formally, set

$$\mathbb{P}_{1n} \equiv \mathbb{X}_{1n}(\mathbb{X}_{1n}'\mathbb{X}_{1n})^{-1}\mathbb{X}_{1n}' \qquad \mathbb{M}_{1n} \equiv I_n - \mathbb{P}_{1n} \tag{2.46}$$

$$\mathbb{P}_{2n} \equiv \mathbb{X}_{2n}(\mathbb{X}_{2n}'\mathbb{X}_{2n})^{-1}\mathbb{X}_{2n}' \qquad \mathbb{M}_{2n} \equiv I_n - \mathbb{P}_{2n} \tag{2.47}$$

Notice that, as in Section 2.2.2, $\mathbb{P}_{1n}\mathbb{Y}_n$ and $\mathbb{P}_{2n}\mathbb{Y}_n$ give us the *fitted* values from regressing $\{Y_i\}_{i=1}^n$ on $\{X_{1i}\}_{i=1}^n$ and $\{X_{2i}\}_{i=1}^n$ respectively (though note, $\mathbb{P}_n\mathbb{Y}_n \neq \mathbb{P}_{1n}\mathbb{Y}_n + \mathbb{P}_{2n}\mathbb{Y}_n$).

The Frisch-Waugh-Lovell Theorem provides us with a different way to compute $\hat{\beta}_{2n}$.

Theorem 2.2.1. *If $\mathbb{X}_{1n}'\mathbb{X}_{1n}$ and $\mathbb{X}_{2n}'\mathbb{M}_{1n}\mathbb{X}_{2n}$ are invertible, then it follows that*

$$\hat{\beta}_{2n} = (\mathbb{X}_{2n}'\mathbb{M}_{1n}\mathbb{X}_{2n})^{-1}\mathbb{X}_{2n}'\mathbb{M}_{1n}\mathbb{Y}_n.$$

PROOF: Notice that we may write the estimators $\hat{\beta}_{1n}$ and $\hat{\beta}_{2n}$ (see (2.44)) as

$$\begin{pmatrix} \hat{\beta}_{1n} \\ \hat{\beta}_{2n} \end{pmatrix} = \arg \min_{b_1 \in \mathbf{R}^{d_1}, b_2 \in \mathbf{R}^{d_2}} \|\mathbb{Y}_n - \mathbb{X}_{1n}b_1 - \mathbb{X}_{2n}b_2\|^2. \tag{2.48}$$

Since $I_n = \mathbb{P}_{1n} + \mathbb{M}_{1n}$, $\mathbb{P}_{1n}\mathbb{M}_{1n} = 0$, and $\mathbb{P}_{1n}\mathbb{X}_{1n} = \mathbb{X}_{1n}$ by Lemma 2.2.4, it follows that

$$\begin{aligned} &\|\mathbb{Y}_n - \mathbb{X}_{1n}b_1 - \mathbb{X}_{2n}b_2\|^2 \\ &\quad = \|\mathbb{M}_{1n}\mathbb{Y}_n - \mathbb{M}_{1n}\mathbb{X}_{1n}b_1 - \mathbb{M}_{1n}\mathbb{X}_{2n}b_2\|^2 + \|\mathbb{P}_{1n}\mathbb{Y}_n - \mathbb{P}_{1n}\mathbb{X}_{1n}b_1 - \mathbb{P}_{1n}\mathbb{X}_{2n}b_2\|^2 \\ &\quad = \|\mathbb{M}_{1n}\mathbb{Y}_n - \mathbb{M}_{1n}\mathbb{X}_{2n}b_2\|^2 + \|\mathbb{P}_{1n}\mathbb{Y}_n - \mathbb{X}_{1n}b_1 - \mathbb{P}_{1n}\mathbb{X}_{2n}b_2\|^2 \end{aligned} \tag{2.49}$$

for any $b_1 \in \mathbf{R}^{d_1}$, $b_2 \in \mathbf{R}^{d_2}$. Since the first term in (2.49) does not depend on b_1 we get

$$\begin{aligned} \hat{\beta}_{2n} &= \arg \min_{b_2 \in \mathbf{R}^{d_2}} \{ \min_{b_1 \in \mathbf{R}^{d_1}} \|\mathbb{Y}_n - \mathbb{X}_{1n}b_1 - \mathbb{X}_{2n}b_2\|^2 \} \\ &= \arg \min_{b_2 \in \mathbf{R}^{d_2}} \{ \|\mathbb{M}_{1n}\mathbb{Y}_n - \mathbb{M}_{1n}\mathbb{X}_{2n}b_2\|^2 + \min_{b_1 \in \mathbf{R}^{d_1}} \|\mathbb{P}_{1n}\mathbb{Y}_n - \mathbb{X}_{1n}b_1 - \mathbb{P}_{1n}\mathbb{X}_{2n}b_2\|^2 \} \end{aligned} \tag{2.50}$$

Moreover, using the definition of $\mathbb{P}_{1n}$, we can rewrite the inner minimization problem as

$$\begin{aligned} &\mathbb{P}_{1n}\mathbb{Y}_n - \mathbb{X}_{1n}b_1 - \mathbb{P}_{1n}\mathbb{X}_{2n}b_2 \\ &\qquad = \mathbb{X}_{1n}(\mathbb{X}_{1n}'\mathbb{X}_{1n})^{-1}\mathbb{X}_{1n}'(\mathbb{Y}_n - \mathbb{X}_{2n}b_2) - \mathbb{X}_{1n}b_1 = \mathbb{X}_{1n}(\gamma - b_1) \end{aligned} \tag{2.51}$$

for $\gamma = (\mathbb{X}_{1n}'\mathbb{X}_{1n})^{-1}\mathbb{X}_{1n}'(\mathbb{Y}_n - \mathbb{X}_{2n}b_2)$. In words, result (2.51) is simply pointing out that since $\mathbb{P}_{1n}\mathbb{X}_{2n}$ and $\mathbb{P}_{1n}\mathbb{Y}_n$ can be expressed as a linear combination of the columns in $\mathbb{X}_{1n}$, it follows that $\mathbb{P}_{1n}(\mathbb{Y}_n - \mathbb{X}_{2n}b_2)$ can be expressed as a linear combination of the columns in $\mathbb{X}_{1n}$ (i.e. $\mathbb{X}_{1n}\gamma$). Putting together results (2.50) and (2.51) we arrive at

$$\min_{b_1 \in \mathbf{R}^{d_1}} \|\mathbb{P}_{1n}\mathbb{Y}_n - \mathbb{X}_{1n}b_1 - \mathbb{P}_{1n}\mathbb{X}_{2n}b_2\|^2 = \min_{b_1 \in \mathbf{R}^{d_1}} \|\mathbb{X}_{1n}(\gamma - b_1)\|^2 = 0; \tag{2.52}$$

i.e. it follows that the minimum of the inner minimization problem in (2.50) does not actually depend on b_2. From here we then conclude that $\hat{\beta}_{2n}$ must satisfy

$$\begin{aligned} \hat{\beta}_{2n} &= \arg \min_{b_2 \in \mathbf{R}^{d_2}} \|\mathbb{M}_{1n}\mathbb{Y}_n - \mathbb{M}_{1n}\mathbb{X}_{2n}b_2\|^2 \\ &\qquad = (\mathbb{X}_{2n}'\mathbb{M}_{1n}'\mathbb{M}_{1n}\mathbb{X}_{2n})^{-1}\mathbb{X}_{2n}'\mathbb{M}_{1n}'\mathbb{M}_{1n}\mathbb{Y}_n = (\mathbb{X}_{2n}'\mathbb{M}_{1n}\mathbb{X}_{2n})^{-1}\mathbb{X}_{2n}'\mathbb{M}_{1n}\mathbb{Y}_n, \end{aligned} \tag{2.53}$$

where the second equality follows from Lemma 2.2.1 and the final equality from $\mathbb{M}_{1n}' =$

$\mathbb{M}_{1n}$ and $\mathbb{M}_{1n}\mathbb{M}_{1n} = \mathbb{M}_{1n}$ by Lemma 2.2.4. ■

The formula for $\hat{\beta}_{2n}$ obtained by Theorem 2.2.1 can be understood as the outcome of a two stage regression. Specifically suppose that we do as follows:

STEP 1: Regress $\{Y_i\}$ and $\{X_{i2}\}_{i=1}^n$ on $\{X_{i1}\}_{i=1}^n$ only, and obtain the residuals from these regressions. Note from Section 2.2.2 we can write these residuals as $\mathbb{M}_{1n}\mathbb{Y}_n$ (from regressing $\mathbb{Y}_n$ on $\mathbb{X}_{1n}$) and $\mathbb{M}_{1n}\mathbb{X}_{2n}$ (from regression $\mathbb{X}_{2n}$ on $\mathbb{X}_{1n}$). ■

STEP 2: Regress the residuals $\mathbb{M}_{1n}\mathbb{Y}_n$ on $\mathbb{M}_{1n}\mathbb{X}_{2n}$, which leads to solving

$$\min_{b\in\mathbf{R}^{d_2}} \|\mathbb{M}_{1n}\mathbb{Y}_n - \mathbb{M}_{1n}\mathbb{X}_{2n}b_2\|^2. \tag{2.54}$$

The Frisch-Waugh-Lovell Theorem basically shows $\hat{\beta}_{2n}$ solves (2.54). ■

The interpretation of $\hat{\beta}_{2n}$ as a two step estimator also clarifies the intuition of Theorem 2.2.1. Essentially, Theorem 2.2.1 states that $\hat{\beta}_{2n}$ is the coefficient from projecting the "part" of $\mathbb{Y}_n$ that cannot be explained by $\mathbb{X}_{1n}$ (i.e. $\mathbb{M}_{1n}\mathbb{Y}_n$) onto the "part" of $\mathbb{X}_{2n}$ that cannot be explained by $\mathbb{X}_{1n}$ (i.e. $\mathbb{M}_{1n}\mathbb{X}_{2n}$). Finally, we also note that $\mathbb{M}_{1n}'\mathbb{M}_{1n} = \mathbb{M}_{1n}$ and Theorem 2.2.1 allow us to equivalently characterize $\hat{\beta}_{2n}$ as

$$\hat{\beta}_{2n} = \arg\min_{b\in\mathbf{R}^{d_2}} \|\mathbb{Y}_n - \mathbb{M}_{1n}\mathbb{X}_{2n}b_2\|^2;$$

i.e. we can also obtain $\hat{\beta}_{2n}$ by projecting $\mathbb{Y}_n$ onto the "part" of $\mathbb{X}_{2n}$ that cannot be explained by $\mathbb{X}_{1n}$. Heuristically, projecting $\mathbb{Y}_n = \mathbb{P}_{1n}\mathbb{Y}_n + \mathbb{M}_{1n}\mathbb{Y}_n$ onto $\mathbb{M}_{1n}\mathbb{X}_{2n}$ is the same as projecting $\mathbb{M}_{1n}\mathbb{Y}_n$ onto $\mathbb{M}_{1n}\mathbb{X}_{2n}$ because $\mathbb{P}_{1n}\mathbb{Y}_n$ and $\mathbb{M}_{1n}\mathbb{X}_{2n}$ are orthogonal.

We conclude this discussion with an example clarifying the algebra involved.

Example 2.2.2. Suppose $Y_i \in \mathbf{R}$, $Z_i \in \mathbf{R}^{d_z}$, and we aim to run the regression

$$\min_{a\in\mathbf{R},b\in\mathbf{R}^{d_z}} \frac{1}{n}\sum_{i=1}^n (Y_i - a - Z_i'b)^2.$$

We let $X_{1i} = 1$, $X_{2i} = Z_i$ and set $X_i = (1, Z_i')' \in \mathbf{R}^d$ where $d = 1 + d_z$; i.e. we add a vector of ones as a regressor in order to incorporate the constant. Then note

$$\mathbb{P}_{1n} = \mathbb{X}_{1n}(\mathbb{X}_{1n}'\mathbb{X}_{1n})^{-1}\mathbb{X}_{1n}' = \frac{1}{n}\mathbb{X}_{1n}\mathbb{X}_{1n}'$$

where we used $\mathbb{X}_{1n}'\mathbb{X}_{1n} = n$. Letting $\bar{Y}_n \equiv \sum_i Y_i/n$ and $\bar{Z}_n \equiv \sum_i Z_i/n$ be the sample means, then note that some algebra leads to the equalities

$$\begin{pmatrix} Y_1 - \bar{Y}_n \\ \vdots \\ Y_n - \bar{Y}_n \end{pmatrix} = \mathbb{M}_{1n}\mathbb{Y}_n \qquad \begin{pmatrix} (Z_1 - \bar{Z}_n)' \\ \vdots \\ (Z_n - \bar{Z}_n)' \end{pmatrix} = \mathbb{M}_{1n}\mathbb{X}_{2n}. \tag{2.55}$$

Hence, the matrix $\mathbb{M}_{1n}$ simply demeans vectors, which makes the result orthogonal to a vector of constants (since it has men zero). As a result, Theorem 2.2.1 here tells us we can compute the coefficient in front of Z_i by running the regression

$$\hat{\beta}_{2n} = \arg\min_{b \in \mathbf{R}^{d_z}} \frac{1}{n} \sum_{i=1}^{n} ((Y_i - \bar{Y}_n) - (Z_i - \bar{Z}_n)'b)^2;$$

i.e. by running a regression on the demeaned data. ■

2.2.2.2 Measures of Fit

It is common practice when running a regression to report a "measure of fit", such as the so called R^2 or *adjusted* R^2. These measures are meant to give a sense of how much of the variation in $\{Y_i\}_{i=1}^n$ can be explained by the regressors $\{X_i\}_{i=1}^n$.

We employ some of the derivations in Example 2.2.2. Suppose we have an i.i.d. sample $\{Y_i, X_i\}_{i=1}^n$ with $Y_i \in \mathbf{R}$ and $X_i \in \mathbf{R}^d$. We consider a regression in which we have factored out the constant (as in Example 2.2.2), so that

$$\hat{\beta}_n = \arg\min_{b \in \mathbf{R}^d} \frac{1}{n} \sum_{i=1}^{n} ((Y_i - \bar{Y}_n) - (X_i - \bar{X}_n)'b)^2. \tag{2.56}$$

With some abuse of notation, we stack the demeaned variables into vectors and matrices

$$\tilde{\mathbb{Y}}_n \equiv \begin{pmatrix} Y_1 - \bar{Y}_n \\ \vdots \\ Y_n - \bar{Y}_n \end{pmatrix} \qquad \tilde{\mathbb{X}}_n \equiv \begin{pmatrix} (X_1 - \bar{X}_n)' \\ \vdots \\ (X_n - \bar{X}_n)' \end{pmatrix}. \tag{2.57}$$

We similarly set the projection matrix onto the column space of $\tilde{\mathbb{X}}_n$ and the projection matrix onto the orthogonal complement of the column space of $\tilde{\mathbb{X}}_n$ by

$$\tilde{\mathbb{P}}_n \equiv \tilde{\mathbb{X}}_n (\tilde{\mathbb{X}}_n' \tilde{\mathbb{X}}_n)^{-1} \tilde{\mathbb{X}}_n' \qquad \tilde{\mathbb{M}}_n \equiv I_n - \tilde{\mathbb{P}}_n. \tag{2.58}$$

Given this notation, and employing that $\tilde{\mathbb{P}}_n \tilde{\mathbb{M}}_n = 0$ and $\tilde{\mathbb{P}}_n + \tilde{\mathbb{M}}_n = I_n$ we obtain

$$\|\tilde{\mathbb{Y}}_n\|^2 = \|\tilde{\mathbb{P}}_n \tilde{\mathbb{Y}}_n + \tilde{\mathbb{M}}_n \tilde{\mathbb{Y}}_n\|^2 = \|\tilde{\mathbb{P}}_n \tilde{\mathbb{Y}}_n\|^2 + \|\tilde{\mathbb{M}}_n \tilde{\mathbb{Y}}_n\|^2. \tag{2.59}$$

These terms in equation (2.59) have names that become self evident once we examine them more closely. In particular, note that by (2.57) we have

$$\|\tilde{\mathbb{Y}}_n\|^2 = \sum_{i=1}^{n} (Y_i - \bar{Y}_n)^2 \quad \text{(TSS)} \tag{2.60}$$

where TSS stands for *Total Sum of Squares*. Similarly, employing Lemma 2.2.1 and recalling that $\tilde{\mathbb{P}}_n\tilde{\mathbb{Y}}_n$ returns the fitted values from regressing $\tilde{\mathbb{Y}}_n$ on $\tilde{\mathbb{X}}_n$ we obtain that

$$\|\tilde{\mathbb{P}}_n\tilde{\mathbb{Y}}_n\|^2 = \|\tilde{\mathbb{X}}_n(\tilde{\mathbb{X}}_n'\tilde{\mathbb{X}}_n)^{-1}\tilde{\mathbb{X}}_n'\tilde{\mathbb{Y}}_n\|^2 = \|\tilde{\mathbb{X}}_n\hat{\beta}_n\|^2 = \sum_{i=1}^{n}(X_i\hat{\beta}_n - \bar{X}_n\hat{\beta}_n)^2 \quad \text{(ESS)} \tag{2.61}$$

where ESS stands for *Explained Sum of Squares*. Finally we have for the last term

$$\|\tilde{\mathbb{M}}_n\tilde{\mathbb{Y}}_n\|^2 = \|\tilde{\mathbb{Y}}_n - \tilde{\mathbb{P}}_n\tilde{\mathbb{Y}}_n\|^2 = \sum_{i=1}^{n}((Y_i - \bar{Y}_n) - (X_i - \bar{X}_n)\hat{\beta}_n)^2 \quad \text{(RSS)} \tag{2.62}$$

where RSS stands for *Residual Sum of Squares* (the residual being $(Y_i - \bar{Y}_n) - (X_i - \bar{X}_n)\hat{\beta}_n$). Given these definitions, we can then employ (2.59) to get

$$\text{TSS} = \text{ESS} + \text{RSS} \text{ or equivalently } 1 = \frac{\text{ESS}}{\text{TSS}} + \frac{\text{RSS}}{\text{TSS}}. \tag{2.63}$$

This is simply a decomposition of the proportion of the variance of $\{Y_i\}_{i=1}^n$ (i.e. TSS) into the proportion that can be "explained" by the regressors (i.e. ESS) and the proportion that remains "unexplained" (i.e. RSS). The R^2 of the regression is just the proportion of the variance that is "explained" by the regressors, i.e. it equals

$$R^2 \equiv \frac{\text{ESS}}{\text{TSS}} = \frac{\|\tilde{\mathbb{P}}_n\tilde{\mathbb{Y}}_n\|^2}{\|\tilde{\mathbb{Y}}_n\|^2}.$$

It follows from (2.63) that $0 \leq R^2 \leq 1$. Whether an R^2 is "high" or not, is very dependent on the dataset you are examining. For instance, in finance very small R^2 are common, while in wage regressions higher R^2 are common.

Remark 2.2.1. Be careful that, while standard, using the word "explain" here is deceiving. We have not really "explained" anything, the decomposition in (2.63) is just a consequence of projections, but has no economic or causal meaning attached to it. ■

Remark 2.2.2. Note that the decomposition in (2.63) alternatively allows us to write

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}} = 1 - \frac{\frac{1}{n}\sum_{i=1}^{n}((Y_i - \bar{Y}_n) - (X_i - \bar{X}_n)'\hat{\beta}_n)^2}{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2}.$$

Therefore, TSS/n can be seen as an estimator for $\text{Var}\{Y\}$ and RSS/n as an estimator for $\text{Var}\{U\}$ where U is the residual on the population regression. These estimators are biased, and sometimes they are replaced by unbiased counterparts, which leads to

$$1 - \frac{\text{RSS}/(n-d-1)}{\text{TSS}/(n-1)},$$

where recall n is the number of observations and d is the dimension of X. This measure is called the *adjusted* R^2. Remark 2.2.1 of course continues to apply. ■

2.3 The Estimator: Asymptotic Properties

We next proceed to establish the asymptotic properties of the OLS estimator. Through this section, we assume we posses an i.i.d. sample $\{Y_i, X_i\}_{i=1}^n$ with $Y_i \in \mathbf{R}$ and $X_i \in \mathbf{R}^d$. The estimand β_0 is understood to be the solution to the moment conditions

$$E[(Y - X'\beta_0)X] = 0 \tag{2.64}$$

as discussed in Section 2.1. We write the implied residual from (2.64) as $U_i = Y_i - X_i'\beta_0$. The estimator $\hat{\beta}_n$ is then understood to equal the solution to the least squares problem

$$\hat{\beta}_n = \arg\min_{b \in \mathbf{R}^d} \frac{1}{n} \sum_{i=1}^n (Y_i - X_i'b)^2, \tag{2.65}$$

and we write the corresponding residuals as $\hat{U}_i = (Y_i - X_i'\hat{\beta}_n)$. Throughout, we will rely heavily on the notation and manipulations of Section 2.2.

2.3.1 Consistency

A minimal requirement of the estimator $\hat{\beta}_n$ is that it be *consistent* for the estimand β_0. Recall consistency means that $\hat{\beta}_n$ converges to β_0 in probability as n tends to infinity.

We will establish consistency under the following Assumptions.

Assumption OLS-1. *(i)* $\{Y_i, X_i\}_{i=1}^n$ *is an i.i.d. sample; (ii)* $E[Y^2] < \infty$*; and (iii) The matrix* $E[XX']$ *is finite and invertible.*

We focus in the i.i.d. setting of Assumption OLS-1(i) for simplicity, though the requirement can be relaxed. The moment conditions imposed in Assumptions OLS-1(ii) and OLS-1(iii) are employed both to establish consistency and also to show that β_0 is indeed well defined as the solution to the moment conditions in (2.64).

Lemma 2.3.1. *If Assumption OLS-1 holds, then* β_0 *is the unique vector in* $\mathbf{R}^d$ *solving*

$$E[(Y - X'\beta_0)X] = 0.$$

Moreover, $\hat{\beta}_n$ *is a consistent estimator for* β_0*.*

Proof: We begin by establishing that β_0 is uniquely defined. First note that

$$E[|Y|\|X\|] \leq \{E[Y^2]\}^{1/2}\{E[\|X\|^2]\}^{1/2} < \infty \tag{2.66}$$

by the Cauchy-Schwarz inequaly and Assumption OLS-1(ii) and OLS-1(iii). Similarly,

$$E[|(X'b)|\|X\|] \leq \|b\|E[\|X\|^2] < \infty \tag{2.67}$$

for any $b \in \mathbf{R}^d$. In particular, results (2.66) and (2.67) imply that $E[(Y - X'b)X]$ is well defined (i.e. its absolute first moment is finite) for any $b \in \mathbf{R}^d$. For any $b \in \mathbf{R}^d$ then

$$E[(Y - X'b)X] = 0 \text{ iff } E[YX] - E[XX']b = 0 \text{ iff } b = \{E[XX']\}^{-1}E[XY]. \tag{2.68}$$

Thus, it follows that β_0 is well defined as the unique solution to the moment conditions in (2.64). Moreover, we obtain a closed form solution for β_0, which is

$$\beta_0 \equiv \{E[XX']\}^{-1}E[YX]. \tag{2.69}$$

In order to establish consistency, note that $E[XX']$ being finite implies that

$$\frac{1}{n}\mathbb{X}_n'\mathbb{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i X_i' \xrightarrow{p} E[XX'] \tag{2.70}$$

by the law of large numbers. Similarly, note that by the law of large number we have

$$\frac{1}{n}\mathbb{X}_n'\mathbb{Y}_n = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i \xrightarrow{p} E[XY]. \tag{2.71}$$

Next, let $\det\{A\}$ denote the determinant of a matrix A. Note that since the determinant is a continuous function, result (2.70) and the continuous mapping theorem implies

$$\det\{\frac{1}{n}\mathbb{X}_n'\mathbb{X}_n\} \xrightarrow{p} \det\{E[XX']\}. \tag{2.72}$$

However, since $E[XX']$ is invertible, we have $\det\{E[XX']\} > 0$, which together with (2.72) implies $\mathbb{X}_n'\mathbb{X}_n/n$ (and hence $\mathbb{X}_n'\mathbb{X}_n$) is invertible with probability tending to one. Therefore it follows from Lemma 2.2.1 and dividing/multiplying by n that

$$\hat{\beta}_n = (\mathbb{X}_n'\mathbb{X}_n)^{-1}\mathbb{X}_n'\mathbb{Y}_n + o_p(1) = (\frac{1}{n}\mathbb{X}_n'\mathbb{X}_n)^{-1}\frac{1}{n}\mathbb{X}_n'\mathbb{Y}_n + o_p(1). \tag{2.73}$$

Finally, combining results (2.70), (2.71), the continuous mapping theorem and (2.73)

$$\hat{\beta}_n = (\frac{1}{n}\mathbb{X}_n'\mathbb{X}_n)^{-1}\frac{1}{n}\mathbb{X}_n'\mathbb{Y}_n + o_p(1) = \{E[XX']\}^{-1}E[XY] + o_p(1) = \beta_0 + o_p(1); \tag{2.74}$$

i.e. we have shown that $\hat{\beta}_n \xrightarrow{p} \beta_0$. ■

2.3.2 Asymptotic Normality

Lemma 2.3.1 tells us $\hat{\beta}_n$ is consistent for β_0. This is reassuring, but not useful for hypothesis testing. For the latter, we need the asymptotic distribution of $\hat{\beta}_n$.

We obtain an asymptotic distribution under the following Assumption.

Assumption OLS-2. *(i)* $\{Y_i, X_i\}_{i=1}^n$ *is an i.i.d. sample; (ii)* $E[\|X\|^2 U^2] < \infty$*; and (iii) The matrix* $E[XX']$ *is finite and invertible.*

Notice that Assumption OLS-2 is almost identical to Assumption OLS-1 with the exception that we have changed the moment condition in Assumption OLS-2(ii). This is natural, since we will rely on central limit theorems instead of laws of large numbers.

Theorem 2.3.1. *If Assumption OLS-2 holds, then it follows that*

$$\sqrt{n}\{\hat{\beta}_n - \beta_0\} \xrightarrow{d} N(0, \{E[XX']\}^{-1} E[XX'U^2] \{E[XX']\}^{-1}).$$

Proof: The proof relies on some of the arguments employed in establishing Lemma 2.3.1. First recall that since $E[XX']$ is finite the law of large numbers implies that

$$\frac{1}{n}\mathbb{X}_n'\mathbb{X}_n = \frac{1}{n}\sum_{i=1}^n X_i X_i' \xrightarrow{p} E[XX']. \tag{2.75}$$

Since $E[XX']$ is invertible by Assumption OLS-2(iii), result (2.75) further implies

$$\lim_{n\to\infty} P(\frac{1}{n}\mathbb{X}_n'\mathbb{X}_n \text{ is invertible }) = 1. \tag{2.76}$$

Therefore, since Lemma 2.2.1 applies whenever $\mathbb{X}_n'\mathbb{X}_n$ is invertible, result (2.76) yields

$$\lim_{n\to\infty} P(\hat{\beta}_n = (\mathbb{X}_n'\mathbb{X}_n)^{-1}\mathbb{X}_n'\mathbb{Y}_n) = 1. \tag{2.77}$$

Next, note that since $U_i = (Y_i - X_i'\beta_0)$, we obtain by definition of $\mathbb{Y}_n$, $\mathbb{X}_n$, and $\mathbb{U}_n$ that

$$\mathbb{Y}_n = \mathbb{X}_n\beta_0 + \mathbb{U}_n. \tag{2.78}$$

Hence, combining results (2.77) and (2.78) with some algebra we arrive at the equality

$$\begin{aligned}\sqrt{n}\{\hat{\beta}_n - \beta_0\} &= \sqrt{n}\{(\mathbb{X}_n'\mathbb{X}_n)^{-1}\mathbb{X}_n'\mathbb{Y}_n - \beta_0\} + o_p(1) \\ &= \sqrt{n}\{(\mathbb{X}_n'\mathbb{X}_n)^{-1}\mathbb{X}_n'\{\mathbb{X}_n\beta_0 + \mathbb{U}_n\} - \beta_0\} + o_p(1) \\ &= (\frac{1}{n}\mathbb{X}_n'\mathbb{X}_n)^{-1}\frac{1}{\sqrt{n}}\mathbb{X}_n'\mathbb{U}_n + o_p(1).\end{aligned} \tag{2.79}$$

To conclude the proof, we need only study the last expression in (2.79). First note

$$\frac{1}{\sqrt{n}}\mathbb{X}_n'\mathbb{U}_n = \frac{1}{\sqrt{n}}\sum_{i=1}^n X_i U_i \xrightarrow{d} N(0, E[XX'U^2]) \tag{2.80}$$

by the central limit theorem and Assumption OLS-2(ii). On the other hand, by result (2.75) we have $\mathbb{X}_n'\mathbb{X}_n/n \xrightarrow{p} E[XX']$. Hence by Slutsky's theorem and the continuous

mapping theorem we can conclude from results (2.79) and (2.80) that

$$\sqrt{n}\{\hat{\beta}_n - \beta_0\} \xrightarrow{d} N(0, \{E[XX']\}^{-1}E[XX'U^2]\{E[XX']\}^{-1}), \tag{2.81}$$

which establishes the claim of the Theorem. ■

2.3.3 Variance Estimation

As a final results before proceeding to inference, we also need an estimator of the asymptotic variance of $\sqrt{n}\{\hat{\beta}_n - \beta_0\}$. The literature has made a distinction between two cases.

If $E[U^2|X] = \sigma^2$ with probability one (over X), then the residual U is called *homoskedastic*. Homoskedasticity is particularly helpful because it simplifies the asymptotic variance obtained in Theorem 2.3.1. Simply note that

$$\begin{aligned}\{E[XX']\}^{-1}&E[XX'U^2]\{E[XX']\}^{-1} \\ &= \{E[XX']\}^{-1}E[XX'\sigma^2]\{E[XX']\}^{-1} = \sigma^2\{E[XX']\}^{-1}\end{aligned} \tag{2.82}$$

where we employed the law of iterated expectations in the first equality.

On the other hand, if $E[U^2|X]$ is not constant in X, then the residual is called *heteroskedastic*. In that case, the asymptotic variance of Theorem 2.3.1 remains

$$\{E[XX']\}^{-1}E[XX'U^2]\{E[XX']\}^{-1} \tag{2.83}$$

2.3.3.1 Homoskedasticity

As argued in (2.82), homoskedasticity of the residuals and Theorem 2.3.1 imply that

$$\sqrt{n}\{\hat{\beta}_n - \beta_0\} \xrightarrow{d} N(0, \sigma^2\{E[XX']\}^{-1}).$$

Thus, in order to estimate the asymptotic variance it suffices to posses an estimator of the matrix $E[XX']$ and of the variance $\sigma^2 = E[U^2]$.

We have already seen in the proof of Lemma 2.3.1 and of Theorem 2.3.1 that $\mathbb{X}_n'\mathbb{X}_n/n$ is a consistent estimator for $E[XX']$ by the law of large numbers. We thus focus on estimating $\sigma^2 = E[U^2]$. If we knew the true residuals $\{U_i\}_{i=1}^n$, then

$$\frac{1}{n}\sum_{i=1}^{n} U_i^2,$$

would be a natural estimator and its consistency would follow by the law of large num-

bers. Since $\{U_i\}_{i=1}^n$ is unobserved, however, we instead employ

$$\hat{\sigma}_n^2 \equiv \frac{1}{n}\sum_{i=1}^n \hat{U}_i^2, \tag{2.84}$$

where recall $\hat{U}_i^2 = (Y_i - X_i'\hat{\beta}_n)^2$ is the regression residual for observation i.

Lemma 2.3.2. *If Assumption OLS-2 holds, $E[U^2|X] = \sigma^2$ with probability one, then*

$$\hat{\sigma}_n^2 \xrightarrow{p} \sigma^2 \text{ and } \{\mathbb{X}_n'\mathbb{X}_n/n\}^{-1}\hat{\sigma}_n^2 \xrightarrow{p} \{E[XX']\}^{-1}\sigma^2. \tag{2.85}$$

PROOF: We start by noting $\hat{U}_i = X_i'(\beta_0 - \hat{\beta}_n) + U_i$ and expanding the square to get

$$\frac{1}{n}\sum_{i=1}^n \hat{U}_i^2 = \frac{1}{n}\sum_{i=1}^n (X_i'(\hat{\beta}_n - \beta_0))^2 + \frac{2}{n}\sum_{1=1}^n U_i X_i'(\hat{\beta}_n - \beta_0) + \frac{1}{n}\sum_{i=1}^n U_i^2. \tag{2.86}$$

To address the first term in (2.86) we rely on the Cauchy-Schwarz inequality to obtain

$$\frac{1}{n}\sum_{i=1}^n (X_i'(\hat{\beta}_n - \beta_0))^2 \leq \frac{1}{n}\sum_{i=1}^n \|X_i\|^2\|\hat{\beta}_n - \beta_0\|^2 \xrightarrow{p} E[\|X\|^2] \times 0 = 0 \tag{2.87}$$

where the final result follows from $E[\|X\|^2] < \infty$, the law of large numbers, and $\hat{\beta}_n$ being consistent for β_0 by Lemma 2.3.1. By similar arguments, we also obtain

$$\{\frac{1}{n}\sum_{i=1}^n U_i X_i'\}(\hat{\beta}_n - \beta_0) \xrightarrow{p} E[UX'] \times 0 = 0 \times 0 = 0 \tag{2.88}$$

Finally note that Assumption OLS-2(ii) and $E[U^2|X] = \sigma^2$ implies that σ^2 is finite. Thus, combining results (2.86), (2.87) and (2.88) and the law of large numbers yields

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n U_i^2 + o_p(1) \xrightarrow{p} \sigma^2. \tag{2.89}$$

In turn, since $\mathbb{X}_n'\mathbb{X}_n/n \xrightarrow{p} E[XX']$ by the law of large numbers, and $E[XX']$ is invertible, the continuous mapping theorem yields $\{\mathbb{X}_n'\mathbb{X}_n/n\}^{-1}\hat{\sigma}_n^2 \xrightarrow{p} \{E[XX']\}^{-1}\sigma^2$. ■

Remark 2.3.1. Recall that $\hat{\mathbb{U}}_n = \mathbb{M}_n\mathbb{Y}_n$ – i.e. the matrix $\mathbb{M}_n$ returns residuals from projecting onto the column space of $\mathbb{X}_n$. Moreover, $\mathbb{M}_n\mathbb{X}_n = 0$ by Lemma 2.2.4 so

$$\sum_{i=1}^n \hat{U}_i^2 = \|\mathbb{M}_n\mathbb{Y}_n\|^2 = \|\mathbb{M}_n\mathbb{U}_n\|^2 = (\mathbb{U}_n'\mathbb{M}_n'\mathbb{M}_n\mathbb{U}_n) = \mathbb{U}_n'\mathbb{M}_n\mathbb{U}_n, \tag{2.90}$$

where we also employed $\mathbb{M}_n'\mathbb{M}_n = \mathbb{M}_n$. Recall for any two matrices A and B, trace$\{AB\}$ =

trace$\{BA\}$, and therefore from (2.90) and $E[U_i^2|X_i] = \sigma^2$ we obtain that

$$E[\sum_{i=1}^{n} \hat{U}_i^2|\{X_i\}_{i=1}^n] = E[\mathbb{U}_n'\mathbb{M}_n\mathbb{U}_n|\{X_i\}_{i=1}^n] = E[\text{trace}\{\mathbb{U}_n'\mathbb{M}_n\mathbb{U}_n\}|\{X_i\}_{i=1}^n]$$
$$= E[\text{trace}\{\mathbb{M}_n\mathbb{U}_n\mathbb{U}_n'\}|\{X_i\}_{i=1}^n] = E[\text{trace}\{\mathbb{M}_n\sigma^2 I_n\}|\{X_i\}_{i=1}^n] = \sigma^2\text{trace}\{\mathbb{M}_n\}$$

where in the last result we exploited that $\mathbb{M}_n$ is a function of $\{X_i\}_{i=1}^n$ only. Provided $\mathbb{X}_n'\mathbb{X}_n$ is invertible, we can then conclude from the definition of $\mathbb{M}_n$ that

$$\text{trace}\{\mathbb{M}_n\} = \text{trace}\{I_n - \mathbb{X}_n(\mathbb{X}_n'\mathbb{X}_n)^{-1}\mathbb{X}_n'\}$$
$$= \text{trace}\{I_n\} - \text{trace}\{(\mathbb{X}_n'\mathbb{X}_n)^{-1}\mathbb{X}_n'\mathbb{X}_n\} = \text{trace}\{I_n\} - \text{trace}\{I_d\} = n - d \quad (2.91)$$

where recall d is the dimension of X. These manipulations allow us to conclude that

$$E[\frac{1}{n-d}\sum_{i=1}^{n} \hat{U}_i^2|\{X_i\}_{i=1}^n] = \sigma^2$$

whenever $\mathbb{X}_n'\mathbb{X}_n$ is invertible; i.e. we obtain an unbiased (conditional on $\{X_i\}_{i=1}^n$) estimator of σ^2 by dividing by $n-d$ in place of n. ■

2.3.3.2 Heteroskedasticity

One of the original arguments cautioning against the homoskedasticity assumption was due to White (1982), who also proposed an estimator of the asymptotic variance of the OLS estimator that remained consistent in the presence of heteroskedasticity.

White (1982) was particularly concerned with a the conditional mean of Y given X not being linear; i.e. of $E[Y|X] \neq X'\beta_0$. In such a case, White (1982) note that

$$Y_i = X_i'\beta_0 + \{E[Y_i|X_i] - X_i'\beta_0\} + \{Y_i - E[Y_i|X_i]\} \equiv X_i'\beta_0 + U_i; \quad (2.92)$$

i.e. the regression error contains two components: (i) One arising from "misspecification" $(E[Y_i|X_i] - X_i'\beta_0)$ and (ii) One arising from the conditional mean $(Y_i - E[Y_i|X_i])$. Hence, by direct calculation and the law of iterated expectations we can obtain

$$E[U_i^2|X_i] = (E[Y_i|X_i] - X_i'\beta_0)^2 + E[(Y_i - E[Y_i|X_i])^2|X_i]$$
$$= (E[Y_i|X_i] - X_i'\beta_0)^2 + \text{Var}\{Y_i|X_i\}. \quad (2.93)$$

Therefore, Since under misspecification the first term in (2.93) necessarily is non-constant in X_i, misspecification naturally gives rise to heteroskedastic errors.

The key challenge in obtaining a consistent estimator for the asymptotic variance in Theorem 2.3.1 under heteroskedasticity is to obtain a consistent estimator for the

matrix $E[XX'U^2]$. To this end, we can employ the estimator

$$\frac{1}{n}\sum_{i=1}^{n} X_i X_i' \hat{U}_i^2.$$

Lemma 2.3.3. *If Assumption OLS-2 holds and $E[\|X\|^3|U|]$, $E[\|X\|^4]$ are finite, then*

$$\frac{1}{n}\sum_{i=1}^{n} X_i X_i' \hat{U}_i^2 \xrightarrow{p} E[XX'U^2]$$

$$\{\mathbb{X}_n'\mathbb{X}_n/n\}^{-1}\frac{1}{n}\sum_{i=1}^{n} X_i X_i' \hat{U}_i^2 \{\mathbb{X}_n'\mathbb{X}_n/n\}^{-1} \xrightarrow{p} \{E[XX']\}^{-1}E[XX'U^2]\{E[XX']\}^{-1}$$

Proof: The proof is similar to that of Lemma 2.3.2, though the arguments are more tedious as we need to rely on matrix operations. If you are uncomfortable with some of these arguments, just work through the proof assuming X_i is a scalar. In what follows, when we apply a norm $\|\cdot\|$ to a matrix it is understood to be the Euclidean norm in the components of the matrix. First, note that by expanding $\hat{U}_i = U_i + X_i'(\beta_0 - \hat{\beta}_n)$ and employing the triangle and Cauchy-Schwarz inequalities we can conclude.

$$\begin{aligned}\|\frac{1}{n}\sum_{i=1}^{n} X_i X_i' \hat{U}_i^2 - \frac{1}{n}\sum_{i=1}^{n} X_i X_i' U_i^2\| &= \|\frac{1}{n}\sum_{i=1}^{n} X_i X_i' \{2U_i X_i'(\hat{\beta}_n - \beta_0) + (X_i'(\hat{\beta}_n - \beta_0))^2\}\| \\ &\leq \frac{2}{n}\sum_{i=1}^{n} \|X_i X_i'\|\|X_i U_i\| \times \|\hat{\beta}_n - \beta_0\| + \frac{1}{n}\sum_{i=1}^{n} \|X_i X_i'\|\|X_i\|^2 \times \|\hat{\beta}_n - \beta_0\|^2. \quad (2.94)\end{aligned}$$

However, since we have assumed that $E[\|X\|^3|U|] < \infty$ and $E[\|X\|^4] < \infty$, the law of large numbers and Lemma 2.3.1 allow us to conclude that

$$\frac{1}{n}\sum_{i=1}^{n} \|X_i X_i'\|\|X_i U_i\| \times \|\hat{\beta}_n - \beta_0\| \xrightarrow{p} E[\|XX'\|\|X\||U|] \times 0 = 0 \quad (2.95)$$

$$\frac{1}{n}\sum_{i=1}^{n} \|X_i X_i'\|\|X_i\|^2 \times \|\hat{\beta}_n - \beta_0\| \xrightarrow{p} E[\|XX'\|\|X\|^2] \times 0 = 0 \quad (2.96)$$

Therefore, employing result (2.94), $E[XX'U^2] < \infty$, and the law of large numbers yields

$$\frac{1}{n}\sum_{i=1}^{n} X_i X_i' \hat{U}_i^2 = \frac{1}{n}\sum_{i=1}^{n} X_i X_i' U_i^2 + o_p(1) \xrightarrow{p} E[XX'U^2]. \quad (2.97)$$

The Lemma then follows from (2.97), $\mathbb{X}_n'\mathbb{X}_n/n \xrightarrow{p} E[XX']$ by the law of large numbers, $E[XX']$ being invertible, and the continuous mapping theorem. ■

Lema 2.3.3 establishes the consistency of the proposed variance estimator under heteroskedasticity. For this reason, these standard errors are often referred to as *robust*

standard errros. However, the consistency shown in Lemma 2.3.3 does not necessarily translate into reliable finite sample inference on $\hat{\beta}_n$. In fact, simulation evidence suggests that the standard errors implied by the variance estimator

$$\{\mathbb{X}_n'\mathbb{X}_n/n\}^{-1}\frac{1}{n}\sum_{i=1}^{n}X_iX_i'\hat{U}_i^2\{\mathbb{X}_n\mathbb{X}_n'/n\}^{-1} \tag{2.98}$$

are often "too small". As a result, in analogy to Remark 2.3.1, sometimes the estimator

$$\{\mathbb{X}_n'\mathbb{X}_n/n\}^{-1}\frac{1}{n-d}\sum_{i=1}^{n}X_iX_i'\hat{U}_i^2\{\mathbb{X}_n\mathbb{X}_n'/n\}^{-1} \tag{2.99}$$

are employed instead. The estimator in (2.99) is often referred to as HC1 standard errors. They are mechanically larger than the robust standard errors in (2.98) and automatically consistent by Lemma 2.3.3 (since $n/(n-d)\to 1$).

A number of alternatives have been proposed that are aimed to improve on finite sample performance. An influential option, known as HC2, was proposed by MacKinnon and White (1985). In order to introduce HC2, we first need some additional notation. Recall $\mathbb{P}_n$ is an $n\times n$ matrix, and let P_{ii} denote its i^{th} diagonal entry. We also let $\mathbf{e}_i\in\mathbf{R}^n$ denote the $n\times 1$ vector whose entries are zero with the exception of the i^{th} entry, which we set to equal one, and assume that errors are homoskedastic so that $E[U_i^2|X_i]=\sigma^2$. Since $\hat{\mathbb{U}}_n=\mathbb{M}_n\mathbb{Y}_n$, $\mathbb{M}_n\mathbb{M}_n'=\mathbb{M}_n$ and $\mathbb{M}_n\mathbb{X}_n=0$ by Lemma 2.2.4, we then obtain

$$\begin{aligned}E[\hat{U}_i^2|\{X_i\}_{i=1}^n] &= E[(\mathbf{e}_i'\mathbb{M}_n\mathbb{Y}_n)^2|\{X_i\}_{i=1}^n]\\ &= E[(\mathbf{e}_i'\mathbb{M}_n\mathbb{U}_n\mathbb{U}_n'\mathbb{M}_n'\mathbf{e}_i)|\{X_i\}_{i=1}^n]=\sigma^2\mathbf{e}_i'\{I_n-\mathbb{P}_n\}\mathbf{e}_i=\sigma^2(1-P_{ii}).\end{aligned} \tag{2.100}$$

Crucially, result (2.100) implies that the fitted residuals $\hat{U}_i$ are *heteroskedastic* even if the true residuals U_i are homoskedastic. However, equation (2.100) also implies that to recover homoskedasticity we need only divide by $(1-P_{ii})$; i.e. $E[\hat{U}_i^2/(1-P_{ii})|\{X_i\}]=\sigma^2$. Motivated by this observation, the HC2 standard errors are then given by

$$\{\mathbb{X}_n'\mathbb{X}_n/n\}^{-1}\frac{1}{n}\sum_{i=1}^{n}X_iX_i'\frac{\hat{U}_i^2}{1-P_{ii}}\{\mathbb{X}_n\mathbb{X}_n'/n\}^{-1}. \tag{2.101}$$

It is important to note that because $\mathbb{P}_n$ is idempotent, its diagonal entries P_{ii} satisfy $0\le P_{ii}\le 1$, and that since $P_{ii}\le 1$, the HC2 standard errors are larger than those in (2.98). Finally, we note that despite HC2 being motivated by a homoskedasticity assumption, they remain valid under heteroskedasticity, and in finite sample appear to perform better than HC1.

2.4 Inference

Having established the asymptotic distribution of the OLS estimator, we next proceed to study the properties of basic tests. For an in depth study of hypothesis testing, you should read Lehmann and Romano (2005).

2.4.1 Basic Background

Suppose we posses a sample $\{W_i\}_{i=1}^n$ of i.i.d. random variables, with W_i distributed according to P. The i.i.d. assumption implies the distribution of the sample $\{W_i\}_{i=1}^n$ is $\bigotimes_{i=1}^n P$, and hence P allows us to recover the entire distribution of the data.

Typically, in our analysis we will maintain a set of assumptions on P, which we may understand as restricting P to lie in a set of probability measures $\mathbf{P}$. When conducting a hypothesis test, we are in essence aiming to distinguish from the data whether P belongs to either of two disjoint subsets of $\mathbf{P}$. Concretely, for some set $\mathbf{P}_0 \subset \mathbf{P}$ and $\mathbf{P}_1 = \mathbf{P} \setminus \mathbf{P}_0$, we can write a null and alternative hypotheses as

$$H_0 : P \in \mathbf{P}_0 \qquad H_1 : P \in \mathbf{P}_1. \tag{2.102}$$

In a hypothesis testing problem, we examine the data $\{W_i\}_{i=1}^n$ and aim to decide whether the data provides enough evidence to reject the null hypothesis or not. Formally, we may therefore think of a *test* as a function mapping the data into $\{0, 1\}$ with a value of one corresponding to a decisions to reject the null hypothesis and a value of zero corresponding to a failure to reject; i.e. a test is a map $\phi_n : \{W_i\}_{i=1}^n \to \{0, 1\}$ such that

$$\phi_n(\{W_i\}_{i=1}^n) = \begin{cases} 0 & \text{if we fail to reject } H_0 \\ 1 & \text{if we reject } H_0 \end{cases}. \tag{2.103}$$

With some abuse of notation, we often write ϕ_n in place of $\phi_n(\{W_i\}_{i=1}^n)$. Notice that since ϕ_n is a function of the random sample $\{W_i\}_{i=1}^n$ our decision of whether to reject or fail to reject is also random. We may write the probability of rejection as

$$P(\phi_n = 1) = E_P[\phi_n] \tag{2.104}$$

where E_P means the expectations is taken over $\{W_i\}_{i=1}$ distributed according to $\bigotimes_{i=1}^n P$. In much of our analysis, we will skip the P subscript, but here we emphasize it since it is important whether $P \in \mathbf{P}_0$ or $P \in \mathbf{P}_1$.

When conducting a test, there are four different possible outcomes, which are:

	H_0 true ($P \in \mathbf{P}_0$)	H_1 true ($P \in \mathbf{P}_1$)
Fail to Reject ($\phi_n = 0$)	Correct	Type II error
Reject ($\phi_n = 1$)	Type I error	Correct

Ideally, we would like to minimize both the probability of making a Type I error and of making a Type II error. However, there is a tension between these two errors. If we want to minimize the probability of a Type I error, then we will demand additional "evidence" from the data in order to reject the null hypothesis. But the more "evidence" we demand from the data, the less likely we are to reject even if the null hypothesis is in fact false; i.e. the probability of a Type II error increases. For this reason, it is standard to first select the probability of a Type I error that we are "comfortable" with, and then examine the probability of a Type II error. The probability of a Type I error that we are "comfortable" with is known as the *level* of the test.

Definition 2.4.1. A test ϕ_n has level $\alpha \in [0,1]$ if $\sup_{P \in \mathbf{P}_0} E_P[\phi_n] \leq \alpha$. ■

In turn, instead of examining the probability of a Type II error, it is common to work with one minus the probability of a Type II error, known as the power of a test.

Definition 2.4.2. The power of a test ϕ_n against an alternative $P \in \mathbf{P}_1$ is $E_P[\phi_n]$. ■

We illustrate these concepts with a simple example.

Example 2.4.1. Suppose that $n = 1$ and W is normally distributed with unknown mean μ and variance equal to one. Under this maintained assumption, $\mathbf{P}$ then becomes

$$\mathbf{P} = \{N(\mu, 1) : \mu \in \mathbf{R}\}.$$

As a null hypothesis we consider the problem of testing whether P is such that $E_P[W] \leq 0$, against the alternative hypothesis that P is such that $E_P[W] > 0$. Following the notation in (2.102), this corresponds to setting $\mathbf{P}_0$ and $\mathbf{P}_1$ to equal

$$\mathbf{P}_0 = \{P \in \mathbf{P} : E_P[W] \leq 0\} \qquad \mathbf{P}_1 = \{P \in \mathbf{P} : E_P[W] > 0\}.$$

With some abuse of notation, it is common to write the hypothesis testing problem as

$$H_0 : \mu \leq 0 \qquad H_1 : \mu > 0.$$

A natural test is to decide to reject the null hypothesis when we observe a value of W that is "too large". Concretely, we may set $\phi(W) = 1\{W > c\}$ for some *critical value* c. For $Z \sim N(0,1)$, the level of the resulting test is then equal to

$$\sup_{P \in \mathbf{P}_0} E_P[\phi] = \sup_{\mu \leq 0} P(W > c) = \sup_{\mu \leq 0} P(Z > c - \mu) = P(Z > c).$$

Therefore, if we aim to have a test with level α, we must set c to be at least as large as the $1-\alpha$ quantile of Z, which is the number $c_{1-\alpha}$ solving the equation

$$P(Z > c_{1-\alpha}) = \alpha. \tag{2.105}$$

In turn the power of the test can be calculated for any $\mu > 0$ (i.e. $P \in \mathbf{P}_1$) to equal

$$E_P[\phi] = P(W > c) = P(Z > c - \mu).$$

Thus, in order to maximize the power of the test – or equivalently minimize the the probability of a Type II error – we must set c as small as possible. Since we want the level of the test to equal α, this means setting $c = c_{1-\alpha}$. ■

A key feature that makes Example 2.4.1 tractable, is that it is possible to compute the finite sample distribution of the test statistic due to the restrictive nature of $\mathbf{P}$. In most empirical applications, however, the finite sample properties are unknown and we need to rely on asymptotic approximations instead. Introducing asymptotic approximations (i.e. limits) into the definitions, however, gives raise to a number of important subtleties.

Consider first an appropriate extension of the concept of the level of a test (as in Definition 2.4.1) to an asymptotic framework. A common used notion, and the one we will employ in this course, is to say ϕ_n has (asymptotic) level α if it satisfies

$$\lim_{n\to\infty} E_P[\phi_n] \leq \alpha \text{ for all } P \in \mathbf{P}_0. \tag{2.106}$$

In particular, note we may rewrite (2.106) to more closely resemble Definition 2.4.1 as

$$\sup_{P\in\mathbf{P}_0} \lim_{n\to\infty} E_P[\phi_n] \leq \alpha. \tag{2.107}$$

While suitable for many problems, including most of the applications we study in this course, the notion of asymptotic level of a test in (2.107) can be inadequate. Intuitively, it is important to always remember that the end goal of asymptotic analysis is to provide an approximation to finite sample settings. Thus, whether (2.106) is a suitable notion depends on whether it is informative about the finite sample level of a test (as in Definition 2.4.1). To this end, it is therefore more appropriate to control $\sup_{P\in\mathbf{P}_0} E_P[\phi_n]$ directly, which leads to stating a test has asymptotic level α if it satisfies

$$\lim_{n\to\infty} \sup_{P\in\mathbf{P}_0} E_P[\phi_n] \leq \alpha; \tag{2.108}$$

i.e. notice that we have reversed the order of limits and supremum in (2.107) and (2.108). This reversal of supremum and limits can be shown to be inoccuous in many "standard" (sometimes referred to as "regular" problems) under suitable assumptions on $\mathbf{P}$. However, in many "non-standandard" (sometimes referred to as "irregular" problems), tests

satisfying (2.106) can dramatically fail (2.108), implying the finite sample level of a test can be quite far from α. We will see such an example when discussing "weak" instruments; see Staiger and Stock (1997) and the ensuing literature.

Turning next to power, a direct extension of Definition 2.4.2 would define the asymptotic power of a test ϕ_n against an alternative $P \in \mathbf{P}_1$ as the following limit

$$\lim_{n\to\infty} E_P[\phi_n]. \tag{2.109}$$

The problem with (2.109) is that as we get more observations (i.e. $n \uparrow \infty$) we eventually learn the true distribution of the data. As a result, in standard problems most "reasonable" tests will eventually reject any fixed $P \in \mathbf{P}_1$ with probability tending to one. A test that satisfies this weak requirement is referred to as being *consistent*.

Definition 2.4.3. A test ϕ_n is consistent if for any $P \in \mathbf{P}_1$ it follows that

$$\lim_{n\to\infty} E_P[\phi_n] = 1.$$

Because most tests are consistent, Definition 2.4.3 provides a weak criterion for distinguishing whether a test is "good" or "bad". For this reason, when comparing tests asymptotically it is often more informative to examine the *local power* of a test. Loosely speaking, a local power calculation involves examining a rejection probability along a sequence $\{P_n\}_{n=1}^{\infty} \subseteq \mathbf{P}_1$ instead of at a fixed $P \in \mathbf{P}_1$. If the sequence $\{P_n\}_{n=1}^{\infty}$ is chosen so that it approaches $\mathbf{P}_0$ at the right rate, then the limiting rejection probability will not be one nor zero. Here, the right rate means that the "distance" to $\mathbf{P}_0$ is proportional to the amount of sampling uncertainty. This concept is best illustrated with an example.

Example 2.4.2. We build on Example 2.4.1 by relaxing the normality requirement. Suppose we observe an i.i.d. sample $\{W_i\}_{i=1}^n$ with $W_i \in \mathbf{R}$ distributed according to P. For simplicity, we still assume that W has variance one, and thus we set

$$\mathbf{P} = \{P : E_P[(W - E_P[W])^2] = 1\}.$$

Our interest remains to test the null hypothesis that $E_P[W] \leq 0$ against the alternative hypothesis that $E_P[W] > 0$. As in Example 2.4.1 we therefore set

$$\mathbf{P}_0 = \{P \in \mathbf{P} : E_P[W] \leq 0\} \qquad \mathbf{P}_1 = \{P \in \mathbf{P} : E_P[W] > 0\}.$$

We base a test off the sample mean $\bar{W}_n \equiv \sum_{i=1}^n W_i/n$, which note for any $P \in \mathbf{P}$ satisfies

$$\sqrt{n}\{\bar{W}_n - E_P[W]\} \xrightarrow{d} N(0,1)$$

by the central limit theorem since $\mathrm{Var}_P(W) = 1$ for all $P \in \mathbf{P}$. Letting $c_{1-\alpha}$ be the $1-\alpha$ quantile of a standard normal distribution (as in (2.105)), we then set as our test

$\phi_n = 1\{\sqrt{n}\bar{W}_n > c_{1-\alpha}\}$. By exploiting that $E_P[W] \leq 0$ for all $P \in \mathbf{P}_0$, we can conclude

$$\sup_{P\in\mathbf{P}_0} \lim_{n\to\infty} P(\sqrt{n}\bar{W}_n > c_{1-\alpha}) \\ \leq \sup_{P\in\mathbf{P}_0} \lim_{n\to\infty} P(\sqrt{n}\{\bar{W}_n - E_P[W]\} > c_{1-\alpha}) = P(Z > c_{1-\alpha}) = \alpha. \quad (2.110)$$

To illustrate power considerations, we introduce an additional "naive" test $\phi_n^{(\mathrm{n})} = 1\{\sqrt{n}\bar{W}_n > 10c_{1-\alpha}\}$, which note satisfies $\phi_n^{(\mathrm{n})} \leq \phi_n$ with probability one. Hence, (2.107) holds with $\phi_n^{(\mathrm{n})}$ in place of ϕ_n by (2.110). Moreover, both ϕ_n and $\phi_n^{(\mathrm{n})}$ are consistent (as in Definition 2.4.3) since for any $c > 0$ and P with $E_P[W] > 0$ (i.e. $P \in \mathbf{P}_1$) we have

$$\lim_{n\to\infty} P(\sqrt{n}\bar{W}_n > c) = \lim_{n\to\infty} P(\sqrt{n}\{\bar{W}_n - E_P[W]\} > c - \sqrt{n}E_P[W]) \\ \geq \lim_{M\uparrow\infty} \lim_{n\to\infty} P(\sqrt{n}\{\bar{W}_n - E_P[W]\} > -M) = \lim_{M\uparrow\infty} P(Z > -M) = 1. \quad (2.111)$$

However, it is evident that ϕ_n is a a more powerful test than $\phi_n^{(\mathrm{n})}$. To get the asymptotic analysis to reflect this, we examine the *local power*. Let $\{P_n\}_{n=1}^n \subset \mathbf{P}_1$ satisfy

$$E_{P_n}[W] = \frac{\lambda}{\sqrt{n}}$$

and note that P_n approaches $\mathbf{P}_0$ in the sense that $E_{P_n}[W] \to 0$; i.e. P_n approaches the boundary between $\mathbf{P}_0$ and $\mathbf{P}_1$. If a triangular array central limit theorem applies, then

$$\lim_{n\to\infty} E_{P_n}[\phi_n] = \lim_{n\to\infty} P_n(\sqrt{n}\{\bar{W}_n - E_{P_n}[W]\} > c_{1-\alpha} - \lambda) = P(Z > c_{1-\alpha} - \lambda) \\ \lim_{n\to\infty} E_{P_n}[\phi_n^{(\mathrm{n})}] = \lim_{n\to\infty} P_n(\sqrt{n}\{\bar{W}_n - E_{P_n}[W]\} > 10c_{1-\alpha} - \lambda) = P(Z > 10c_{1-\alpha} - \lambda),$$

which shows the local power of ϕ_n is larger than that of $\phi_n^{(\mathrm{n})}$. ■

2.4.2 Wald Tests

The most common tests in linear regression models are Wald tests. These are tests in which a null hypothesis concerning an estimand are tested employing the corresponding estimator. For brevity we focus on Wald tests due to the close connection to the analysis in Section 2.3. However, you should not attributed Wald tests undue prominence as they can perform more poorly in finite samples than alternatives such as Score and Likelihood ratio tests; see Rothenberg (1984) and Newey and McFadden (1994).

2.4.2.1 Single Linear Restriction

We return to the linear regression setup and assume we posses an i.i.d. sample $\{Y_i, X_i\}_{i=1}^n$ with $Y \in \mathbf{R}$ and $X \in \mathbf{R}^d$. The estimand β_0 is understood to be as defined in (2.64) and the estimator $\hat{\beta}_n$ as in (2.65). Note β_0 depends on the distribution P of the data but, in contrast to Section 2.4.1, we suppress the dependence from the notation. We also let

$$\Sigma_0 \equiv \{E[XX']\}^{-1} E[XX'U^2] \{E[XX']\}^{-1}, \tag{2.112}$$

which by Theorem 2.3.1 is the asymptotic variance of the OLS estimator.

We first consider the problem of testing a single linear restriction. In particular, suppose that for some $r \in \mathbf{R}^d$ and $b \in \mathbf{R}$ we are interested in testing

$$H_0 : r'\beta_0 = b \qquad H_1 : r'\beta_0 \neq b. \tag{2.113}$$

A special case of this hypothesis is testing whether a specific coordinate of β_0 is equal to a conjectured value – e.g., for testing if the first coordinate equals zero, set $b = 0$ and $r = (1, 0, \ldots, 0)'$. A Wald test employs the estimator as a test statistic, so we use

$$|\sqrt{n}\{r'\hat{\beta}_n - b\}|. \tag{2.114}$$

In order to test the null hypothesis in (2.113) we will reject for "large" values of (2.114). However, to understand what "large" means, we must first characterize the asymptotic distribution of (2.114) under the null hypothesis.

Lemma 2.4.1. *Let Assumption OLS-2 hold and $\sigma_0^2 \equiv r'\Sigma_0 r$. If $r'\beta_0 = b$, then it follows*

$$|\sqrt{n}\{r'\hat{\beta}_n - b\}| \xrightarrow{d} |\sigma_0 Z|,$$

where Z follows a standard normal distribution.

Proof: First recall that if Assumption OLS-2 holds, then Theorem 2.3.1 implies that

$$\sqrt{n}\{\hat{\beta}_n - \beta_0\} \xrightarrow{d} N(0, \Sigma_0).$$

Hence, if $r'\beta_0 = b$, then the continuous mapping theorem and the properties of the normal distribution (see results (1.3) and (1.8) in Section 1.1) allow us to conclude that

$$|\sqrt{n}\{r'\hat{\beta}_n - b\}| = |r'\{\sqrt{n}(\hat{\beta}_n - \beta_0)\}| \xrightarrow{d} |N(0, r'\Sigma_0 r)| = |\sigma_0 Z|,$$

which establishes the claim of the Lemma. ■

Lemma 2.4.1 characterizes the asymptotic distribution of $|\sqrt{n}\{r'\hat{\beta}_n - b\}|$ under the null hypothesis. However, to employ Lemma 2.4.1 to conduct a test, we also need an

estimator for σ_0. Fortunately, given any consistent estimator $\hat{\Sigma}_n$ for Σ_0, the continuous mapping theorem implies $r'\hat{\Sigma}_n r$ will be consistent for σ_0^2. Since the specific choice of $\hat{\Sigma}_n$ may be context specific, we abstract from its specific formulation, though note the results in Section 2.3.3 apply under suitable conditions. Given such asymptotic variance estimator, we may then construct a Wald test, which in this context is defined as

$$\phi_n \equiv 1\{\frac{1}{\sqrt{r'\hat{\Sigma}_n r}}|\sqrt{n}\{r'\hat{\beta}_n - b\}| > c_{1-\alpha/2}\},$$

Our next Corollary implies the Wald test indeed satisfies (2.107).

Corollary 2.4.1. *If Assumption OLS-2 holds, $\hat{\Sigma}_n \xrightarrow{p} \Sigma_0$, $\sigma_0 > 0$, and $r'\beta_0 = b$, then*

$$\lim_{n\to\infty} P(|\frac{\sqrt{n}}{\sqrt{r'\hat{\Sigma}_n r}}\{r'\hat{\beta}_n - b\}| > c_{1-\alpha/2}) = \alpha,$$

where $c_{1-\alpha}$ denotes the $1-\alpha$ quantile of a standard normal random variable.

PROOF: By Lemma 2.4.1, $\hat{\Sigma}_n \xrightarrow{p} \Sigma_0$ and the continuous mapping theorem we obtain

$$\frac{1}{\sqrt{r'\hat{\Sigma}_n r}} \times |\sqrt{n}\{r'\hat{\beta}_n - b\}| \xrightarrow{d} \frac{1}{\sigma_0}|\sigma_0 Z| = |Z|. \tag{2.115}$$

Therefore, since $c_{\alpha/2} = -c_{1-\alpha/2}$ by symmetry of the cdf of a standard normal distribution around zero, we can conclude from result (2.115) that

$$\begin{aligned}\lim_{n\to\infty} P(|\frac{\sqrt{n}}{\sqrt{r'\hat{\Sigma}_n r}}\{r'\hat{\beta}_n - b\}| > c_{1-\alpha/2}) &= P(|Z| > c_{1-\alpha/2}) \\ = P(Z > c_{1-\alpha/2}) + P(Z < -c_{1-\alpha/2}) &= P(Z > c_{1-\alpha/2}) + P(Z < c_{\alpha/2}) = \alpha,\end{aligned} \tag{2.116}$$

which establishes the claim of the Corollary. ■

It is also useful to note that confidence intervals for $r'\beta_0$ are readily available as a consequence of Corollary 2.4.1. In particular, note that if we define

$$C_n \equiv [r'\hat{\beta}_n - c_{1-\alpha/2}\frac{\sqrt{r'\hat{\Sigma}_n r}}{\sqrt{n}}, r'\hat{\beta}_n + c_{1-\alpha/2}\frac{\sqrt{r'\hat{\Sigma}_n r}}{\sqrt{n}}], \tag{2.117}$$

then it follows $r'\beta_0$ belongs to C_n with asymptotic probability $1-\alpha$. This follows from

$$\begin{aligned}\lim_{n\to\infty} P(r'\beta_0 \in C_n) &= \lim_{n\to\infty} P(r'\hat{\beta}_n - c_{1-\alpha/2}\frac{\sqrt{r'\hat{\Sigma}_n r}}{\sqrt{n}} \leq r'\beta_0 \leq r'\hat{\beta}_n + c_{1-\alpha/2}\frac{\sqrt{r'\hat{\Sigma}_n r}}{\sqrt{n}}) \\ &= \lim_{n\to\infty} P(|\frac{\sqrt{n}}{\sqrt{r'\hat{\Sigma}_n r}}|\{r'\hat{\beta}_n - r'\beta_0\}| \leq c_{1-\alpha/2}) = 1-\alpha.\end{aligned} \tag{2.118}$$

As noted when we started this Section, a special case of our analysis concerns hypothesis

testing or confidence interval construction for a single coordinate of the vector β_0. By Corollary 2.4.1 and the analysis in (2.117) and (2.118), we only need the point estimate $\hat{\beta}_n$ and an estimator $\hat{\sigma}_n^2$ of the asymptotic variance. You will find in applied work both $\hat{\beta}_n$ and $\hat{\sigma}_n$ are often reported so that the reader may independently decide the confidence level α they deem appropriate and easily conduct a hypothesis test or build a corresponding confidence interval.

2.4.2.2 Multiple Linear Restrictions

As a generalization of the analysis in Section 2.4.2.1, we next examine the problem of testing a null hypothesis concerning multiple linear restrictions. In particular, suppose that R is a $p \times d$ matrix, $B \in \mathbf{R}^p$ is some vector, and we are interested in testing

$$H_0 : R\beta_0 = B \qquad H_1 : R\beta_0 \neq B. \tag{2.119}$$

Notice that an important special case of (2.119) consists of testing whether the entire vector β_0 equals zero, which corresponds to setting $R = I_d$ (for I_d the $d \times d$ identity matrix) and $B = 0$ (the zero vector in $\mathbf{R}^d$).

We again base a test off the estimator $R\hat{\beta}_n$ for $R\beta_0$. By Theorem 2.3.1, the continuous mapping theorem, and the properties of normal distributions (recall (1.8)), we obtain

$$R\sqrt{n}\{\hat{\beta}_n - \beta_0\} \xrightarrow{d} N(0, R\Sigma_0 R'). \tag{2.120}$$

The limiting distribution in (2.120) is known up to the covariance matrix $R\Sigma_0 R'$. Fortunately, given an estimator $\hat{\Sigma}_n$ for Σ_0 we can estimate $R\Sigma_0 R'$ by $R\hat{\Sigma}_n R'$. Thus, it is straightforward to obtain an asymptotic distribution for the Wald test statistic.

Lemma 2.4.2. *Let Assumption OLS-2 hold, $R\Sigma_0 R'$ be invertible, $\hat{\Sigma}_n \xrightarrow{p} \Sigma_0$, and $\{Z_j\}_{j=1}^p$ be i.i.d. standard normal random variables. If $R\beta_0 = B$, then it follows that*

$$\|(R\hat{\Sigma}_n R')^{-1/2}\sqrt{n}\{R\hat{\beta}_n - B\}\|^2 \xrightarrow{d} \sum_{j=1}^{p} Z_j^2.$$

Proof: We first employ that, as previously noted, $R\hat{\Sigma}_n R' \xrightarrow{p} R\Sigma_0 R'$ by the continuous mapping theorem. Therefore, employing Slutsky Theorem, $R\beta_0 = B$, Theorem 2.3.1, and $(R\Sigma_0 R')^{-1/2} R\Sigma_0 R' (R\Sigma_0 R')^{-1/2} = I_p$ we are able to conclude that

$$\begin{aligned}
&(R\hat{\Sigma}_n R')^{-1/2}\sqrt{n}\{R\hat{\beta}_n - B\} \\
&\quad = (R\hat{\Sigma}_n R')^{-1/2}\sqrt{n}R\{\hat{\beta}_n - \beta_0\} \xrightarrow{d} (R\Sigma_0 R')^{-1/2} \times N(0, R\Sigma_0 R') = N(0, I_p).
\end{aligned} \tag{2.121}$$

Since normal random variables that have zero covariance are also independent, the

vector $(Z_1, \ldots, Z_p)$ with each coordinate independent and following a standard normal distribution has law $N(0, I_d)$. The continuous mapping theorem and (2.121) then imply

$$\|(R\hat{\Sigma}_n R')^{-1/2}\sqrt{n}\{R\hat{\beta}_n - B\}\|^2 \xrightarrow{d} \sum_{j=1}^{p} Z_j^2,$$

which establishes the claim of the Lemma. ■

The asymptotic distribution derived in Lemma 2.4.2 is known as *chi squared distribution with p degrees of freedom*. We denote such a distribution by χ_p^2, whose quantiles are readily available in all commonly used statistical packages. Thus, Lemma 2.4.2 suggests employing the $1-\alpha$ quantile of a χ_p^2 random variable to construct a Wald test. Concretely, letting $c_{1-\alpha}$ solve $P(\chi_p^2 > c_{1-\alpha}) = \alpha$, the Wald test of (2.119) is defined as

$$\begin{aligned}\phi_n &= 1\{\|(R\hat{\Sigma}_n R')^{-1/2}\sqrt{n}\{R\hat{\beta}_n - B\}\|^2 > c_{1-\alpha}\} \\ &= 1\{n(R\hat{\beta}_n - B)'(R\hat{\Sigma}_n R')^{-1}(R\hat{\beta}_n - B) > c_{1-\alpha}\},\end{aligned} \tag{2.122}$$

where the equality follows by simple algebra. The ability of (2.122) to satisfy requirement (2.107) is then an immediate consequence of Lemma 2.4.2.

Corollary 2.4.2. *Let Assumption OLS-2 hold, $\hat{\Sigma}_n \xrightarrow{p} \Sigma_0$, $R\Sigma_0 R'$ be invertible, and $c_{1-\alpha}$ denote the $1-\alpha$ quantile of χ_p^2. If $R\beta_0 = B$, then it follows that*

$$\lim_{n\to\infty} P(\|(R\hat{\Sigma}_n R')^{-1/2}\sqrt{n}\{R\hat{\beta}_n - B\}\|^2 > c_{1-\alpha}) = \alpha.$$

PROOF: This is an immediate consequence of Lemma 2.4.2. ■

2.4.2.3 Non-Linear Restrictions

As an immediate extension of the analysis in Section 2.4.2.2, we consider testing

$$H_0 : f(\beta_0) = 0 \qquad H_1 : f(\beta_0) \neq 0, \tag{2.123}$$

where $f : \mathbf{R}^d \to \mathbf{R}^p$ is a possibly nonlinear transformation of β_0. Notice that since f is nonlinear, it is without loss of generality to employ zero as the value of the null hypothesis (vs. say b and B in (2.113) and (2.119) respectively).

Let $\nabla f(\beta_0)$ denote the $p \times d$ matrix of partial derivatives of f evaluated at the point β_0, and recall that the Delta method (see Theorem 1.3.4) implies that

$$\sqrt{n}\{f(\hat{\beta}_n) - f(\beta_0)\} \xrightarrow{d} N(0, \nabla f(\beta_0)\Sigma_0 \nabla f(\beta_0)').$$

Moreover, note that since $\hat{\beta}_n$ is consistent for β_0, a natural estimator for the asymptotic

variance of $f(\hat{\beta}_n)$ is just $\nabla f(\hat{\beta}_n)\hat{\Sigma}_n \nabla f(\hat{\beta}_n)'$. These observations lead us to an immediate extension of Lemma 2.4.2 and Corollary 2.4.2, whose proof is left as an exercise.

Lemma 2.4.3. *Let Assumption OLS-2 hold, $\hat{\Sigma}_n \xrightarrow{p} \Sigma_0$, ∇f be continuous, $\nabla f(\beta_0)\Sigma_0 \nabla f(\beta_0)'$ be invertible, and $c_{1-\alpha}$ denote the $1-\alpha$ quantile of χ^2_p. If $f(\beta_0) = 0$, then it follows that*

$$\lim_{n\to\infty} P(\|(\nabla f(\hat{\beta}_n)\hat{\Sigma}_n \nabla f(\hat{\beta}_n)')^{-1/2}\sqrt{n} f(\hat{\beta}_n)\|^2 > c_{1-\alpha}) = \alpha.$$

2.5 Problems

1. Suppose $Y \in \mathbf{R}$, $X \in \mathbf{R}$, $0 < \text{Var}\{X\} < \infty$, and we consider the linear regression

$$\hat{\beta}_n \equiv \arg\min_{b\in\mathbf{R}^2} \frac{1}{n}\sum_{i=1}^{n}(Y_i - (1, X_i)b)^2.$$

   Find the population parameter β_0 being estimated by $\hat{\beta}_n$ under the assumption that $E[Y|X] = X^2$ and X is uniformly distributed on $[0, 1]$.

2. Suppose $Y \in \mathbf{R}$ and $X \in \mathbf{R}^d$ satisfy $E[Y^2] < \infty$ and $E[\|\nabla E[Y|X]\|^2] < \infty$ Show that $b_0 \equiv E[\nabla E[Y|X]]$ is the solution to the optimization problem

$$\min_{b\in\mathbf{R}^d} E[\|\nabla E[Y|X] - b\|^2].$$

   Build an example in which β_0 (as in Lemma 2.1.1) is not equal to b_0 (Note: an "example" here would consist of a distribution of (Y, X) such that the implied β_0 and b_0 satisfy the stated requirements).

3. Following on Example 2.1.4, let D_i equal one if individual i attended college, and $D_i = 0$ otherwise. Further let $Y_i(1)$ denote income if individual i attended college, and $Y_i(0)$ denote the income if she did not attend college. Suppose we observe income $Y_i = D_i Y_i(1) + (1 - D_i)Y_i(0)$ and college attendance D_i.

   (a) For $\alpha_0 = E[Y_i(0)|D_i = 0]$ and $\beta_0 = E[Y_i(1)|D_i = 1] - E[Y_i(0)|D_i = 0]$, show

$$Y_i = \alpha_0 + D_i\beta_0 + \eta \qquad E[\eta|D_i] = 0.$$

   (b) Unlike Example in 2.1.4, do not assume $(Y_i(1), Y_i(0))$ are independent of D_i. Show that β_0 from part (a) satisfies the relation

$$\beta_0 = E[Y_i(1) - Y_i(0)|D_i = 1] + E[Y_i(0)|D_i = 1] - E[Y_i(0)|D_i = 0]$$

   (c) The quantity $E[Y_i(1) - Y_i(0)|D_i = 1]$ is known as the *average treatment effect on the treated* (ATEU). If individual choose to attend college, what do you expect the sign of ATEU to be? Explain your answer.

(d) The quantity $E[Y_i(0)|D_i = 1] - E[Y_i(0)|D_i = 0]$ can be interpreted as a *selection bias*. If more capable individuals choose to attend college, then can you sign the selection bias? Explain your answer.

(e) Is OLS consistent for the average treatment effect if there is no heterogeneity (i.e. $Y_i(1) - Y_i(0)$ is the same for all individuals)? What is there is heterogeneity (i.e. $Y_i(1) - Y_i(0)$ differs across individuals)?

4. Prove Lemma 2.2.2 (Hint: The representation in (2.31) and the proof of Lemma 2.2.1 should be particularly helpful.)

5. Prove Lemma 2.2.4.

6. Suppose $\{Y_i, X_i\}_{i=1}^n$ is an i.i.d. sample with $Y_i \in \mathbf{R}$ and $X_i \in \mathbf{R}^d$. Show $R^2 = 1$ if and only if there exists a constant $a_0 \in \mathbf{R}$ and vector $b_0 \in \mathbf{R}^d$ such that $Y_i = a_0 + X_i' b_0$ for all i.

7. Let $\{Y_i, X_i\}_{i=1}^n$ be an i.i.d. sample, $\hat{\beta}_n$ denote the OLS estimator of β_0, and suppose we are interested in estimating $E[X]'\beta_0$. Propose an estimator of $E[X]'\beta_0$ and derive its asymptotic distribution.

8. The following problem is based on Duflo et al. (2008), you should read it before starting this problem. The data is available for replication purposes at the *American Economic Review* website. I have downloaded and pre-processed it for you. The description of variables are in the paper and in the file `DDKDataGuide.m`.

   The following questions concern estimation. We'll return to this dataset for inference in another problem, so keep the code to your answers below.

   (a) Build a sample consisting only of observations for which none of the following variables are missing: `girl, stdmark, totalscore, tracking`. For the rest of the problem we will work with this subsample.

   (b) Compute the following summary statistics: (i) number of boys, (ii) number of students assigned to tracking schools, (iii) the average original score, (iv) number of unique schools in the study.

   (c) Estimate the ATE of being assigned to a tracking school on your score using a subsample consisting of only girls.

   (d) Repeated part (c) using only boys.

   (e) Estimate the ATE of girls and the ATE of boys using a single regression specification on the whole sample. What specification did you employ? The answers answers should match what you found in parts (c) and (d). Show mathematically why this is the case. (Hint: Build the right dummy variables)

(f) Simply based on point estimates, what group seems to benefit more from being assigned to tracking schools: students in the bottom half of the distribution or students in the upper half?

9. Suppose $W \in \mathbf{R}$ is known to be normally distributed with variance one and mean $\mu \in \{0, 2\}$. We aim to test the null hypothesis

$$H_0 : \mu = 0 \qquad H_1 : \mu = 2,$$

and we employ the test $\phi(W) = 1\{W > c\}$ for some value c. What should c be if we want to minimize the sum of the probabilities of a Type I and a Type II errors?

10. Suppose $\{W_i\}_{i=1}^n$ is an i.i.d. sample with W_i distributed according to P. The mean of W_i is unknown, but we know $\text{Var}_P(W) = 4$. We are interested in testing

$$H_0 : E_P[W] = 0 \qquad H_1 : E_P[W] \neq 0$$

(a) Consider the test $\phi_n = 1\{\sqrt{n}\bar{W}_n/2 > c_{1-\alpha}\}$. Show this test satisfies (2.107).

(b) Is the test in part (a) consistent?

(c) Show the test $\tilde{\phi}_n = 1\{|\sqrt{n}\bar{W}_n/2| > c_{1-\alpha/2}\}$ satisfies (2.107).

(d) Consider a sequence of local alternatives $P_n \subseteq \mathbf{P}_1$ such that $E_{P_n}[W] = \lambda/\sqrt{n}$, and a triangular array central limit theorem applies. For what values of λ is the local power of $\tilde{\phi}_n$ greater than the local power of ϕ_n?

11. Show the Wald test studied in Corollary 2.4.1 is consistent.

12. Prove Lemma 2.4.3.

13. Consider the following Table from Fehr and Goette (2007) reporting descriptive statistics from an individual-level randomized experiment:[1]

(a) Use the Delta method to construct a standard error estimate for the quantity

$$\hat{\eta}_n = \frac{\bar{Y}_n^A - \bar{Y}_n^B}{\bar{Y}_n^B}$$

where $\bar{Y}_n^A = 4131.33$ is the mean revenue of Group A and $\bar{Y}_n^B = 3005.75$ is the mean revenue of Group B. Assume observations are i.i.d.

(b) Let η denote the probability limit of $\hat{\eta}_n$. Then, we may approximate the revenue/wage elasticity as $\eta/0.25$. Use your answer to part (a) to construct a confidence interval for $\eta/0.25$.

[1] This problem was originally written by Pat Kline.

TABLE 1—DESCRIPTIVE STATISTICS

		Participating messengers		Difference groups A and B	Nonparticipating messengers, Veloblitz	Messengers, Flash
		Group A	Group B			
Four-week period prior to experiment	Mean revenues	3,500.67 (2,703.25)	3,269.94 (2,330.41)	241.67 [563.19]	1461.70 (1,231.95)	1637.49 (1,838.61)
	Mean shifts	12.14 (8.06)	10.95 (7.58)	1.20 [1.75]	5.19 (4.45)	6.76 (6.11)
	N	21	19		21	59
Treatment period 1	Mean revenues	4,131.33 (2,669.21)	3,005.75 (2,054.20)	1,125.59 [519.72]	844.21 (1,189.53)	1,408.23 (1,664.39)
	Mean shifts	14.00 (7.25)	9.85 (6.76)	4.15 [1.53]	3.14 (4.63)	6.32 (6.21)
	N	22	20		21	65
Treatment period 2	Mean revenues	2,734.03 (2,571.58)	3,675.57 (2,109.19)	−941.53 [513.2]	851.23 (1,150.31)	921.58 (1,076.47)
	Mean shifts	8.73 (7.61)	12.55 (7.49)	−3.82 [1.65]	3.29 (4.15)	4.46 (4.74)
	N	22	20		24	72

Notes: Standard deviations in parentheses, standard error of differences in brackets. Group A received the high commission rate in experimental period 1, group B in experimental period 2.
Source: Own calculations.

14. The following problem continues on Problem 8 and relies on the data in Duflo et al. (2008). Recall the variable descriptions are available in the file `DDKDataGuide.m`.

    (a) Compute the standard error for the ATE of girls under a homoskedasticity assumption.

    (b) Compute HC1 and HC2 standard errors for the ATE of girls under a heteroskedasticity assumption.

    (c) Report a confidence region for the ATE of girls employing the different standard errors computed in parts (a) and (b).

    (d) Suppose that we are concerned that the regression residuals of students in the same school may be correlated with each other. Are either of the standard errors computed in part (a) and (b) consistent? Justify your answer.

    (e) Revisiting problem 8(f), test the null hypothesis that the ATE of being assigned to a tracking school is higher for students in the bottom half of the distribution than for students in the upper half of the distribution. Use HC1 standard errors for this test (Hint: Note that under the i.i.d. assumption the ATE estimates for each subgroup are independent since they are based on different subsamples)

15. Suppose there are potential outcomes (Y_0, Y_1) and covariates $X \in \mathbf{R}^d$ satisfying

$$Y_0 = X'\beta_0 + \epsilon_0 \tag{2.124}$$

$$Y_1 = X'\beta_1 + \epsilon_1 \tag{2.125}$$

with $E[\epsilon_0|X] = E[\epsilon_1|X] = 0$. We conduct an experiment, where $D \in \{0, 1\}$ equals one whenever treatment is assigned, D is independent of (Y_0, Y_1, X), and

$$Y = Y_0 + D(Y_1 - Y_0).$$

As usual, assume that we have an i.i.d. sample $\{Y_i, D_i, X_i\}_{i=1}^n$.

(a) Find a closed form expression for $E[Y|X, D]$.

(b) Suppose that $E[XX']$ is full rank and finite and we conduct the regression

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg \min_{b_0, b_1 \in \mathbf{R}^d} \frac{1}{n} \sum_{i=1}^{n} (Y_i - (1 - D_i)X_i'b_0 - D_i X_i'b_1)^2.$$

Show that $\hat{\beta}_0 \xrightarrow{p} \beta_0$ and $\hat{\beta}_1 \xrightarrow{p} \beta_1$

(c) Find the joint asymptotic distribution of $\sqrt{n}\{\hat{\beta}_0 - \beta_0\}$ and $\sqrt{n}\{\hat{\beta}_1 - \beta_1\}$.

(d) Define the average treatment effect for an individual with covariates X as $E[Y_1 - Y_0|X]$. Using parts (a), (b), and (c) propose an estimator for $E[Y_1 - Y_0|X = x_0]$ and derive its asymptotic distribution (here $x_0 \in \mathbf{R}^d$ is some known fixed value for the covariates).

16. Let $(Y_i(1), Y_i(0), X_i, D_i), i = 1, \ldots, n$ be i.i.d. where $Y_i(1) \in \mathbf{R}$ and $Y_i(0) \in \mathbf{R}$ are potential outcomes under treatment and control, respectively, $X_i \in \mathbf{R}^k$ is a vector of observed, baseline covariates, and D_i is an indicator for receipt of treatment. As usual, define the observed outcome to be $Y_i = Y_i(1)D_i + Y_i(0)(1 - D_i)$. The parameter of interest is the average treatment effect,

$$\tau = E[Y_i(1) - Y_i(0)].$$

(a) A natural estimator of τ in this setting is the difference in means

$$\hat{\tau}_n^{\text{diff}} = \frac{1}{n_1} \sum_{1 \le i \le n : D_i = 1} Y_i - \frac{1}{n_0} \sum_{1 \le i \le n : D_i = 0} Y_i \ ,$$

where, for $d = 0, 1$, $n_d = |\{1 \le i \le n : D_i = d\}|$. Show that $\hat{\tau}_n^{\text{diff}}$ satisfies

$$\sqrt{n}(\hat{\tau}_n^{\text{diff}} - \tau) \xrightarrow{d} N(0, \sigma_{\text{diff}}^2)$$

with

$$\sigma_{\text{diff}}^2 = \frac{\text{Var}[Y_i(1)]}{P\{D_i = 1\}} + \frac{\text{Var}[Y_i(0)]}{P\{D_i = 0\}} \ .$$

Clearly state any additional assumptions needed to justify your answer.

(b) Empirical researchers often try to exploit X_i by defining an estimator $\hat{\tau}_n^{\text{reg}}$ as the ordinary least squares estimate of the coefficient on D_i in a regression

of Y_i on a constant, D_i and X_i. While $\hat{\tau}_n^{\text{reg}}$ and $\hat{\tau}_n^{\text{diff}}$ are both consistent for τ, the former estimator need not be more precise than $\hat{\tau}_n^{\text{diff}}$. Explain briefly why $\hat{\tau}_n^{\text{reg}}$ is consistent for τ.

(c) For this reason, it is useful to consider the following estimator:

$$\hat{\tau}_n^{\text{adj}} = \frac{1}{n_1} \sum_{1 \leq i \leq n : D_i = 1} \left(Y_i - (X_i - \bar{X}_n)' \hat{\gamma}_{1,n} \right) - \frac{1}{n_0} \sum_{1 \leq i \leq n : D_i = 0} \left(Y_i - (X_i - \bar{X}_n)' \hat{\gamma}_{0,n} \right),$$

where $\bar{X}_n = \frac{1}{n} \sum_{1 \leq i \leq n} X_i$ and, for $d = 0, 1$, $\hat{\gamma}_{n,d}$ is obtained as the ordinary least squares estimate of the coefficient on X_i in a regression of Y_i on a constant and X_i using *only* observations with $D_i = d$. This estimator is provably more precise that $\hat{\tau}_n^{\text{diff}}$. To see this, complete the following:

i. Show that

$$\sqrt{n}(\hat{\tau}_n^{\text{adj}} - \tau) \xrightarrow{d} N(0, \sigma^2_{\text{adj}})$$

with

$$\sigma^2_{\text{adj}} = \frac{\text{Var}[Y_i(1) - X_i' \gamma_1]}{P\{D_i = 1\}} + \frac{\text{Var}[Y_i(0) - X_i' \gamma_0]}{P\{D_i = 0\}} + (\gamma_1 - \gamma_0)' \text{Var}[X_i] (\gamma_1 - \gamma_0),$$

where, for $d = 0, 1$, $\gamma_d = \text{Var}[X_i]^{-1} \text{Cov}[Y_i(d), X_i]$. Clearly state any additional assumptions needed to justify your answer.

ii. Show that $\sigma^2_{\text{diff}} - \sigma^2_{\text{adj}} = \Delta' \text{Var}[X_i] \Delta \geq 0$, where

$$\Delta = \sqrt{\frac{P\{D_i = 0\}}{P\{D_i = 1\}}} \gamma_1 + \sqrt{\frac{P\{D_i = 1\}}{P\{D_i = 0\}}} \gamma_0.$$

Chapter 3

Instrumental Variables

We next turn to a class of linear models in which the parameter of interest is not identified as the estimand from a linear regression. Instead, in these models the parameter of interest β_0 is the solution to a system of equations

$$E[(Y - X'\beta_0)Z] = 0, \tag{3.1}$$

where $Y \in \mathbf{R}$, $X \in \mathbf{R}^{d_x}$, and $Z \in \mathbf{R}^{d_z}$. We note often Z and X may have some (but not all) components in common. If $Z = X$, then (3.1) in fact reduces to a linear regression.

3.1 Motivating Examples

As in Chapter 2, the principal challenge of employing model (3.1) is linking the parameter β_0 solving (3.1) to an economically interesting parameter of interest. In what follows, we revisit some of the canonical examples that have given raise to (3.1).

3.1.1 Measurement Error

Suppose $Y \in \mathbf{R}$, $X \in \mathbf{R}^d$, and we are interested in estimating the population linear regression coefficient β_0, defined as the solution to the moment equations

$$E[(Y - X'\beta_0)X] = 0. \tag{3.2}$$

Unfortunately, we do not observe X_i, but instead see a "noisy" observation $\tilde{X}_i$ with

$$\tilde{X}_i = X_i + \eta_i, \tag{3.3}$$

where $\eta_i \in \mathbf{R}^d$ is the measurement error. Notice this formulation allows certain coordinates of the vector of covariates X_i to be measured without error – such a setting simply

corresponds to the appropriate coordinate of η_i equaling zero with probability one.

It is not hard to see that simply regressing Y_i on $\tilde{X}_i$ does not yield a consistent estimator for β_0. Suppose for instance that $\eta_i \in \mathbf{R}^d$ is independent of other variables and satisfies $E[\eta] = 0$, which is sometimes referred to as "*classical measurement error*". By the standard arguments in Chapter 2, we obtain that the probability limit of the OLS estimator obtained from regressing Y_i on the mismeasured regressors $\tilde{X}_i$ equals

$$\{\frac{1}{n}\sum_{i=1}^{n}\tilde{X}_i\tilde{X}_i'\}^{-1}\frac{1}{n}\sum_{i=1}^{n}\tilde{X}_iY_i \xrightarrow{p} \{E[\tilde{X}\tilde{X}']\}^{-1}E[\tilde{X}Y]. \tag{3.4}$$

However, since η is independent of (Y, X) and $E[\eta] = 0$, we can simplify (3.4) to obtain

$$\begin{aligned}\{E[\tilde{X}\tilde{X}']\}^{-1}E[\tilde{X}Y] &= \{E[XX' + \eta X' + X\eta' + \eta\eta']\}^{-1}E[(X+\eta)Y] \\ &= \{E[XX'] + E[\eta\eta']\}^{-1}E[XX']\beta_0. \end{aligned} \tag{3.5}$$

On the other hand, suppose that we have available a random variable $Z \in \mathbf{R}^d$ such that $E[Z\eta'] = 0$ and $E[UZ] = 0$ for $U = (Y - X'\beta_0)$ – i.e. Z is uncorrelated with the measurement error and the regression residual. It then follows that

$$E[(Y - \tilde{X}'\beta_0)Z] = E[(Y - (X+\eta)'\beta_0)Z] = E[(Y - X'\beta_0)Z] = 0 \tag{3.6}$$

where the second and third equality follows from Z being uncorrelated with η and U respectively. Moreover, if $E[XZ']$ is full rank, then it follows that β_0 is the unique solution of the moment conditions in (3.6). When feasible, a common empirical approach for selecting Z is to employ an additional measurement of X. Concretely, whenever

$$Z_i = X_i + \epsilon_i$$

it follows that the stated requirements on Z are satisfied if the measurement error ϵ is uncorrelated with the measurement error η. Thus, here Z and $\tilde{X}$ may be seen as independent measurements of X that allow us to recover β_0 despite both being measured with error. For a recent empirical implementation of this approach see Chalfin and McCrary (2013), who employ multiple measurements of police force growth rates to address measurement error in a regression of crime growth on police force growth.

Remark 3.1.1. Suppose that only one variable is measured with error, which without loss of generality we assume to be the first coordinate. In such a case, it follows that

$$E[\eta\eta'] = \sigma^2\mathbf{e}_1\mathbf{e}_1' \tag{3.7}$$

where $\mathbf{e}_1 = (1, 0, \ldots, 0)'$. Moreover, by direct calculation it is possible to verify that

$$\{E[XX'] + \sigma^2 \mathbf{e}_1 \mathbf{e}_1'\}^{-1} = \{E[XX']\}^{-1} - \frac{\sigma^2}{1 + \sigma^2 \mathbf{e}_1' \{E[XX']\}^{-1} \mathbf{e}_1} \{E[XX']\}^{-1} \mathbf{e}_1 \mathbf{e}_1' \{E[XX']\}^{-1}. \quad (3.8)$$

Using (3.8) and our formula in (3.5) we find that the first coordinate of the OLS regression coefficient (i.e. the coordinate with the mismeasured variable) converges to

$$\mathbf{e}_1' \beta_0 \times \frac{1}{1 + \sigma^2 \mathbf{e}_1' \{E[XX']\}^{-1} \mathbf{e}_1}. \quad (3.9)$$

This result is often referred to as *attenuation bias* because it implies measurement error (subject to the discussed structure) shrinks the coefficient corresponding to the mismeasured variable towards zero. Notice that formula (3.8) also implies that measurement error in one variable can translate into inconsistency for the coefficients corresponding to other variables as well (even is said variables are measured without error). ■

3.1.2 Omitted Variables

Let $Y \in \mathbf{R}$, and $X = (X_1', X_2')' \in \mathbf{R}^d$ with $X_1 \in \mathbf{R}^{d_1}$ and $X_2 \in \mathbf{R}^{d_2}$. Similarly decompose $\beta_0 \in \mathbf{R}^d$ into $\beta_0 = (\beta_{10}', \beta_{20}')'$, which we assume solves

$$E[(Y - X'\beta_0)X] = 0. \quad (3.10)$$

Suppose the parameter of interest is β_{10}, but regrettably we do not observe X_2 (i.e. we only observe X_1). It is still possible to regress Y_i on X_{1i} only, but the resulting regression estimator is likely to be inconsistent for β_{10}. To see this, simply note that

$$\begin{aligned} \{\frac{1}{n}\sum_{i=1}^n X_{1i} X_{1i}'\}^{-1} \frac{1}{n}\sum_{i=1}^n X_{1i} Y_i &= \{\frac{1}{n}\sum_{i=1}^n X_{1i} X_{1i}'\}^{-1} \frac{1}{n}\sum_{i=1}^n X_{1i}(X_{1i}'\beta_{10} + X_{2i}'\beta_{20} + U_i) \\ &\xrightarrow{p} \beta_{10} + \{E[X_1 X_1']\}^{-1} E[X_1 X_2'] \beta_{20}, \end{aligned} \quad (3.11)$$

where we employed that $E[UX_1] = 0$ by condition (3.10). The "extra" term in the limit in (3.11) (i.e. $\{E[X_1 X_1]\}^{-1} E[X_1 X_2'] \beta_{20}$) is known as the *omitted variable bias*.

An estimation strategy that circumvents this problem becomes available if we find a variable $Z \in \mathbf{R}^d$ such that $E[ZX_2'] = 0$ and $E[ZU] = 0$. In such a case, we obtain

$$E[(Y - X_1'\beta_{10})Z] = E[(X_2'\beta_{20} + U)Z] = 0. \quad (3.12)$$

Moreover, β_{10} is the unique solution to (3.12) provided that $E[X_1' Z]$ has rank bigger than or equal to d_1 (recall $X_1 \in \mathbf{R}^{d_1}$). Heuristically, the omitted variable bias in (3.11) arises

from the regressor X_1 being correlated with the omitted variable X_2. In contrast, (3.12) avoids this problem by employing Z in the moment conditions, which is uncorrelated with X_2 but still with X_1 (as needed to identify β_{10}).

The concern for omitted variable bias is salient in the study of returns to education. Consider, for example, a simplified version of the model in Card (2001), in which

$$Y_i = a_0 + S_i b_i + \epsilon_i, \tag{3.13}$$

where Y_i is log earnings for person i, S_i is the education level, b_i is a person specific return to education, and ϵ_i is a person specific shock independent of S_i. If we are interested in estimating $E[b_i]$ (i.e. the average return to education) we may rewrite (3.13) as

$$Y_i = a_0 + S_i E[b] + \{S_i(b_i - E[b]) + \epsilon_i\}. \tag{3.14}$$

Thus, if we simply regress Y_i on a constant and S_i we will not consistently estimate $E[b_i]$ due to the omitted variable $S_i(b_i - E[b_i])$. Suppose, however, that we find a variable Z_i which affects the cost, but not the benefit of education – a commonly used such variable, for instance, is the distance of household to a four year college (Card, 1993). In order for such variable Z_i to solve the omitted variable problem it must then satisfy

$$E[S(b - E[b])Z] = 0. \tag{3.15}$$

As Card (2001) notes, while it is plausible to assume $E[(b_i - E[b_i])Z_i] = 0$, the requirement in (3.15) is more demanding. If we let π_0 solve $E[(S - Z\pi_0)Z] = 0$, and set $\xi = S - Z\pi_0$ (i.e. π_0 is the coefficient from the population regression of S on Z), then

$$E[S(b - E[b])Z] = \pi_0 E[Z^2(b - E[b])] + E[\xi(b - E[b])Z]. \tag{3.16}$$

Thus, a sufficient condition for (3.15) to hold is that $E[Z^2(b - E[b])] = 0$ and the more problematic $E[\xi(b - E[b])Z] = 0$. See Card (2001) for a discussion of the plausibility of this assumption and alternative estimating strategies.

3.1.3 Simultaneity

A canonical example of parameters being identified through moment restrictions as in (3.1) arises from demand estimation. Here, we base our discussion on Berry (1994).

Suppose there are J products, which compete in n different markets indexed by $1 \leq i \leq n$. In each market, consumers decide whether to purchase one of the J goods or none at all – when none of the J goods are chosen, we sometimes say the *outside option* was chosen instead. Under certain conditions on how individuals choose among the J goods, Berry (1994) shows that for S_{ij} and S_{i0} the market shares in market i of good j

and the outside good respectively we have that

$$\log(S_{ij}) - \log(S_{i0}) = W_{ij}'\gamma_0 + \alpha_0 P_{ij} + \xi_{ij}, \tag{3.17}$$

where W_{ij} are observable characteristics of good j in market i, P_{ij} is the price of good j in market i, and ξ_{ij} is a characteristic of good j in market i that is unobserved to the econometrician but observed to the consumers.

Estimating γ_0 and α_0 in (3.17) through linear regression would require us to assume

$$E[\xi_{ij} W_{ij}] = 0 \qquad E[\xi_{ij} P_{ij}] = 0. \tag{3.18}$$

The assumption that $E[\xi_{ij} P_{ij}] = 0$ is particularly problematic. In a model of oligopolistic competition, the price charged by the firm should be positively correlated with any unobservable characteristic that makes the good more appealing to consumers – intuitively, an appealing unobserved characteristic increases demand and allows the firm to charge a higher markup. A way to solve this problem is to employ a variable M_{ij} that is correlated with P_{ij} but not with ξ_{ij}. We may then obtain the system of equations

$$E\left[((\log(S_{ij}) - \log(S_{i0})) - W_{ij}'\gamma_0 - \alpha_0 P_{ij}) \begin{pmatrix} M_{ij} \\ W_{ij} \end{pmatrix}\right] = 0, \tag{3.19}$$

which maps into (3.1) with $Y_i = \log(S_{ij}) - \log(S_{i0})$, $X_i = (W_{ij}', P_{ij})'$ and $Z_i = (W_{ij}', M_{ij})'$. In the context of demand estimation, a common choice for M_{ij} is the price of an input used to produce good j (i.e. a cost shifter). Alternatively, in a setting of oligopolistic competition, M_{ij} can be any variable that is correlated with the firm's markup (and hence P_{ij}) but is not valued by consumers as an attribute of good j (and hence $E[M_{ij}\xi_{ij}] = 0$). Berry et al. (1995), for instance, employ observable characteristics of rival product in estimating automobile demands.

3.2 The Estimator

We next turn to discussing the estimation of the parameter β_0, which we assume satisfies

$$E[(Y - X'\beta_0)Z] = 0 \tag{3.20}$$

for $Y \in \mathbf{R}$, $X \in \mathbf{R}^{d_x}$, and $Z \in \mathbf{R}^{d_z}$. Notice that in our preceding examples we often arrived at the specification in (3.20) starting from a setting in which

$$Y_i = X_i'\beta_0 + U_i \tag{3.21}$$

but U_i failed to satisfy $E[UX] = 0$, and hence β_0 could not be estimated by ordinary least squares. While one might arrive at (3.20) through different arguments (see, e.g., Hansen and Singleton (1982)), it is always important that one have a model in mind in order to interpret β_0 as a parameter of interest. Indeed, notice when Z and X are of equal dimension and $E[ZX']$ is full rank, model (3.20) always defines a unique β_0 as a function of the distribution P of the data. In this regard, (3.20) does not confer any more meaning onto β_0 that the OLS moment restrictions $E[(Y - X'\beta_0)X] = 0$ do.

As a bit of terminology, we note that Z is referred to as an *instrument*. Components of X that are not also part of Z are often referred to as *endogenous*, while components of X that are also in Z are referred to as *exogenous*. You may find in the literature a concern for whether Z is *endogenous*. This terminology often arises when starting for a model as in (3.21), in which case *endogeneity* of Z means $E[UZ] \neq 0$ – i.e. Z does not yield the desired moment restrictions in (3.20). Finally, we note that Z should not only generate the moment restrictions in (3.20), but it should also be such that the equation (3.20) has a unique solution. Uniqueness is guaranteed by the rank of $E[ZX']$ being at least as large as the dimension of β_0, which means that

$$\text{rank}\{E[XZ']\} \geq d_x. \tag{3.22}$$

The requirement in (3.22) is known as *instrument relevance* or *the order/rank condition*.

3.2.1 Some Notation

We will assume availability of an i.i.d. sample $\{Y_i, X_i, Z_i\}_{i=1}^n$, with $Y_i \in \mathbf{R}$, $X_i \in \mathbf{R}^{d_x}$, and $Z_i \in \mathbf{R}^{d_z}$. Throughout, the parameter of interest β_0 is assumed to solve

$$E[(Y - X'\beta_0)Z] = 0. \tag{3.23}$$

Notice that potentially $d_z > d_x$, in which case (3.23) is a system of linear equations with more equations than unknowns. In such a case β_0 is deemed *overidentified*.

As in ordinary least squares, a simple estimation strategy is to employ as an estimator the solution to a sample analogue to (3.23) – i.e. to attempt to find a $b \in \mathbf{R}^{d_x}$ such that

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i'b)Z_i = 0. \tag{3.24}$$

However, when $d_z > d_x$ there is unlikely to exist a solution to the system of equations in (3.24). For this reason, we instead use as an estimator the value of b that makes (3.24)

as close to zero as possible. Concretely, for some $d_z \times d_z$ matrix $\hat{\Omega}_n$, we set

$$\hat{\beta}_n = \arg\min_{b \in \mathbf{R}^{d_x}} (\frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i'b)Z_i)'\hat{\Omega}_n(\frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i'b)Z_i). \tag{3.25}$$

Here, the matrix $\hat{\Omega}_n$ allows us to weight moment conditions differently, which can affect the asymptotic variance of $\hat{\beta}_n$. As a special case, we may set $\hat{\Omega}_n = I_{d_z}$, which yields

$$\hat{\beta}_n = \arg\min_{b \in \mathbf{R}^{d_x}} \|\frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i'b)Z_i\|^2.$$

As in Chapter 2, introducing appropriate notation can help us rely on linear algebra to greatly simplify our notation. To this end, we therefore define

$$\mathbb{Y}_n \equiv \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \qquad \mathbb{X}_n \equiv \begin{pmatrix} X_1' \\ \vdots \\ X_n' \end{pmatrix} \qquad \mathbb{Z}_n \equiv \begin{pmatrix} Z_1' \\ \vdots \\ Z_n' \end{pmatrix}, \tag{3.26}$$

and note that $\mathbb{Y}_n$ is a $n \times 1$ vector, $\mathbb{X}_n$ is a $n \times d_x$ matrix, and $\mathbb{Z}_n$ is a $n \times d_z$ matrix. We further denote the residual by $U_i = Y_i - X_i'\beta_0$, and set the corresponding $n \times 1$ vector as

$$\mathbb{U}_n = \mathbb{Y}_n - \mathbb{X}_n\beta_0.$$

Employing the introduced notation, we obtain a closed form solution for $\hat{\beta}_n$.

Lemma 3.2.1. *If $\hat{\Omega}_n$ is positive definite and $\mathbb{X}_n'\mathbb{Z}_n\hat{\Omega}_n\mathbb{Z}_n'\mathbb{X}_n$ is invertible, then*

$$\hat{\beta}_n = \arg\min_{b \in \mathbf{R}^{d_x}} \{\mathbb{Z}_n'(\mathbb{Y}_n - \mathbb{X}_n b)\}'\hat{\Omega}_n\{\mathbb{Z}_n'(\mathbb{Y}_n - \mathbb{X}_n b)\} \tag{3.27}$$

$$= \{\mathbb{X}_n'\mathbb{Z}_n\hat{\Omega}_n\mathbb{Z}_n'\mathbb{X}_n\}^{-1}\mathbb{X}_n'\mathbb{Z}_n\hat{\Omega}_n\mathbb{Z}_n'\mathbb{Y}_n \tag{3.28}$$

PROOF: The first characterization in (3.27) follows from the definition in (3.26) and direct calculation. With regards to (3.28), we first note that the condition that $\hat{\Omega}_n$ be positive ensures that the optimization problem defining $\hat{\beta}_n$ (as in (3.25)) is convex. As a result, $\hat{\beta}_n$ is characterized as a zero of the first order conditions. Here, that means

$$\mathbb{X}_n'\mathbb{Z}_n\hat{\Omega}_n\mathbb{Z}_n'(\mathbb{Y}_n - \mathbb{X}_n\hat{\beta}_n) = 0. \tag{3.29}$$

Whenever the $d_x \times d_x$ matrix $\mathbb{X}_n'\mathbb{Z}_n\hat{\Omega}_n\mathbb{Z}_n'\mathbb{X}_n$ is invertible, it follows that $\hat{\beta}_n$ is in fact the unique solution to (3.29). Thus, through matrix inversion we obtain

$$\hat{\beta}_n = \{\mathbb{X}_n'\mathbb{Z}_n\hat{\Omega}_n\mathbb{Z}_n'\mathbb{X}_n\}^{-1}\mathbb{X}_n'\mathbb{Z}_n\hat{\Omega}_n\mathbb{Z}_n'\mathbb{Y}_n,$$

which establishes the claim of the Lemma. ■

3.2.2 Consistency

We next turn to establishing the consistency of $\hat{\beta}_n$ for the parameter β_0 solving

$$E[(Y - X'\beta_0)Z] = 0. \tag{3.30}$$

Towards this end, we impose the following requirements.

Assumption IV-1. *$\{Y_i, X_i, Z_i\}_{i=1}^n$ is an i.i.d. sample with $Y \in \mathbf{R}$, $X \in \mathbf{R}^{d_x}$, $Z \in \mathbf{R}^{d_z}$, and such that (3.23) holds for some $\beta_0 \in \mathbf{R}^{d_z}$.*

Assumption IV-2. *(i) $\hat{\Omega}_n \xrightarrow{p} \Omega$ for some symmetric and positive matrix Ω; (ii) The rank of the $d_x \times d_z$ matrix $E[XZ']$ is equal to d_x.*

Assumption IV-3. *The moment $E[\|XZ'\|]$ is finite.*

Assumption IV-1 simply imposes the i.i.d. requirement and that the distribution of the data indeed satisfy (3.30) for some $\beta_0 \in \mathbf{R}^{d_x}$. When the dimension of Z is equal (or smaller) than that of X, there are at least as many unknowns as equations in (3.30) and a solution is guaranteed to exist. However, if the dimension of Z exceeds that of X, then a solution to (3.30) may fail to exist and thus Assumption IV-1 becomes necessary. Assumption IV-2(i) allows the weighting matrix $\hat{\Omega}_n$ to be random, but requires it to converge to a positive definite matrix ($\Omega > 0$). In turn, Assumption IV-2(ii) ensures identification of β_0 by guaranteeing that if a solution to (3.30) exists, then it must be unique. Finally, Assumption IV-3 introduces the appropriate moment conditions.

Lemma 3.2.2. *If Assumptions IV-1, IV-2, IV-3 hold, then $\hat{\beta}_n$ is consistent for β_0.*

Proof: First note that since Ω is a positive definite matrix by Assumption IV-2(i) and $E[XZ']$ has rank d_x by Assumption IV-2(ii), it follows that the $d_x \times d_x$ matrix

$$E[XZ']\Omega E[ZX'] \tag{3.31}$$

is invertible. Moreover, notice that since $E[\|XZ\|] < \infty$, we obtain by the law of large numbers that $\mathbb{X}_n'\mathbb{Z}_n/n \xrightarrow{p} E[XZ']$. Hence, since $\hat{\Omega}_n \xrightarrow{p} \Omega$ by Assumption IV-2, the continuous mapping theorem allows us to conclude that

$$(\frac{1}{n}\mathbb{X}_n'\mathbb{Z}_n)\hat{\Omega}_n(\frac{1}{n}\mathbb{Z}_n'\mathbb{X}_n) \xrightarrow{p} E[XZ']\Omega E[ZX']. \tag{3.32}$$

Thus, $\mathbb{X}_n'\mathbb{Z}_n\hat{\Omega}_n\mathbb{Z}_n'\mathbb{X}_n$ is invertible with probability tending to one, and by Lemma 3.2.1 we obtain that with probability tending to one

$$\begin{aligned}\hat{\beta}_n &= \{\mathbb{X}_n'\mathbb{Z}_n\hat{\Omega}_n\mathbb{Z}_n'\mathbb{X}_n\}^{-1}\mathbb{X}_n'\mathbb{Z}_n\hat{\Omega}_n\mathbb{Z}_n'\mathbb{Y}_n \\ &= \{\mathbb{X}_n'\mathbb{Z}_n\hat{\Omega}_n\mathbb{Z}_n'\mathbb{X}_n\}^{-1}\mathbb{X}_n'\mathbb{Z}_n\hat{\Omega}_n\mathbb{Z}_n'(\mathbb{X}_n\beta_0 + \mathbb{U}_n) \\ &= \beta_0 + \{\mathbb{X}_n'\mathbb{Z}_n\hat{\Omega}_n\mathbb{Z}_n'\mathbb{X}_n\}^{-1}\mathbb{X}_n'\mathbb{Z}_n\hat{\Omega}_n\mathbb{Z}_n'\mathbb{U}_n \end{aligned} \tag{3.33}$$

Finally, we note that since $E[ZU] = 0$ by Assumption IV-1, the law of large numbers, Assumption IV-2, and the continuous mapping theorem allow us to conclude that

$$\begin{aligned}&\{(\frac{1}{n}\mathbb{X}_n'\mathbb{Z}_n)\hat{\Omega}_n(\frac{1}{n}\mathbb{Z}_n'\mathbb{X}_n)\}^{-1}(\frac{1}{n}\mathbb{X}_n'\mathbb{Z}_n)\hat{\Omega}_n(\frac{1}{n}\mathbb{Z}_n'\mathbb{U}_n) \\ &\qquad\xrightarrow{p} (E[XZ']\Omega E[ZX'])^{-1}E[XZ']\Omega E[ZU] = 0. \end{aligned} \tag{3.34}$$

The claim of the Lemma therefore follows from (3.33) and (3.34). ■

3.2.3 Asymptotic Normality

The IV estimator $\hat{\beta}_n$ is not only consistent but also asymptotically normally distributed. In order to establish asymptotic normality we strengthen our moment conditions – as expected as we will rely on the central limit theorem instead of the law of large numbers.

Assumption IV-4. *The moment $E[ZZ'U^2]$ is finite*

The following Theorem establishes the asymptotic normality of $\hat{\beta}_n$.

Theorem 3.2.1. *If Assumptions IV-1, IV-2, IV-3, and IV-4 hold, then it follows that*

$$\begin{aligned}&\sqrt{n}\{\hat{\beta}_n - \beta_0\} \\ &\quad\xrightarrow{d} N(0, (E[XZ']\Omega E[ZX'])^{-1}E[XZ']\Omega E[ZZ'U^2]\Omega E[ZX'](E[XZ']\Omega E[ZX'])^{-1}).\end{aligned}$$

Proof: The proof relies on some of the same arguments we employed in establishing Lemma 3.2.2. First note $E[XZ']$ having rank d_x and Ω being positive definite implies

$$E[XZ']\Omega E[ZX'] \tag{3.35}$$

is invertible. Therefore, since $\{\mathbb{X}_n'\mathbb{Z}_n/n\} \xrightarrow{p} E[XZ']$ by the law of large numbers, and $\hat{\Omega}_n \xrightarrow{p} \Omega$, it follows from the continuous mapping theorem that

$$(\frac{1}{n}\mathbb{X}_n'\mathbb{Z}_n)\hat{\Omega}_n(\frac{1}{n}\mathbb{Z}_n'\mathbb{X}_n) \xrightarrow{p} E[XZ']\Omega E[ZX']. \tag{3.36}$$

We can therefore conclude $\mathbb{X}_n'\mathbb{Z}_n\hat{\Omega}_n\mathbb{Z}_n'\mathbb{X}_n$ is invertible with probability tending to one.

However, whenever $\mathbb{X}_n'\mathbb{Z}_n\hat{\Omega}_n\mathbb{Z}_n'\mathbb{X}_n$ is invertible, we may rely on Lemma 3.2.1 to obtain

$$\begin{aligned}\hat{\beta}_n &= \{\mathbb{X}_n'\mathbb{Z}_n\hat{\Omega}_n\mathbb{Z}_n'\mathbb{X}_n\}^{-1}\mathbb{X}_n'\mathbb{Z}_n\hat{\Omega}_n\mathbb{Z}_n'\mathbb{Y}_n \\ &= \{\mathbb{X}_n'\mathbb{Z}_n\hat{\Omega}_n\mathbb{Z}_n'\mathbb{X}_n\}^{-1}\mathbb{X}_n'\mathbb{Z}_n\hat{\Omega}_n\mathbb{Z}_n'(\mathbb{X}_n\beta_0 + \mathbb{U}_n) \\ &= \beta_0 + \{\mathbb{X}_n'\mathbb{Z}_n\hat{\Omega}_n\mathbb{Z}_n'\mathbb{X}_n\}^{-1}\mathbb{X}_n'\mathbb{Z}_n\hat{\Omega}_n\mathbb{Z}_n'\mathbb{U}_n \end{aligned} \tag{3.37}$$

with probability tending to one. In particular, rearranging terms we arrive at

$$\sqrt{n}\{\hat{\beta}_n - \beta_0\} = \{(\frac{1}{n}\mathbb{X}_n'\mathbb{Z}_n)\hat{\Omega}_n(\frac{1}{n}\mathbb{Z}_n'\mathbb{X}_n)\}^{-1}(\frac{1}{n}\mathbb{X}_n'\mathbb{Z}_n)\hat{\Omega}_n(\frac{1}{\sqrt{n}}\mathbb{Z}_n'\mathbb{U}_n) + o_p(1). \tag{3.38}$$

Next, observe that $E[ZU] = 0$ by Assumption IV-1, and $E[ZZ'U^2] < \infty$ by Assumption IV-4. Therefore, applying the central limit theorem yields

$$\frac{1}{\sqrt{n}}\mathbb{Z}_n'\mathbb{U}_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_iU_i \xrightarrow{d} N(0, E[ZZ'U^2]). \tag{3.39}$$

Hence, results (3.38) and (3.39), together with $\hat{\Omega}_n \xrightarrow{p} \Omega$, $\mathbb{Z}_n'\mathbb{X}_n/n \xrightarrow{p} E[ZX']$ by the law of large numbers, and the continuous mapping theorem imply that

$$\begin{aligned}&\sqrt{n}\{\hat{\beta}_n - \beta_0\} \\ &\quad\xrightarrow{d} N(0, (E[XZ']\Omega E[ZX'])^{-1}E[XZ']\Omega E[ZZ'U^2]\Omega E[ZX'](E[XZ']\Omega E[ZX'])^{-1}),\end{aligned}$$

which establishes the claim of the Theorem. ■

In order to employ Theorem 3.2.1 for inference, we still require a consistent estimator of the asymptotic variance of $\hat{\beta}_n$. To this end, we note that by the law of large numbers $\mathbb{X}_n'\mathbb{Z}_n/n$ is consistent for $E[XZ']$, while $\hat{\Omega}_n$ is consistent for Ω by assumption. Thus, given the characterization of the asymptotic variance of $\hat{\beta}_n$ in Theorem 3.2.1, we only require a consistent estimator for $E[ZZ'U^2]$. Such a consistent estimator is given by

$$\frac{1}{n}\sum_{i=1}^{n} Z_iZ_i'(Y_i - X_i'\hat{\beta}_n)^2. \tag{3.40}$$

Hence, we estimate the asymptotic variance of $\hat{\beta}_n$ obtained in Theorem 3.2.1 using

$$\{\frac{\mathbb{X}_n'\mathbb{Z}_n}{n}\hat{\Omega}_n\frac{\mathbb{Z}_n'\mathbb{X}_n}{n}\}^{-1}\frac{\mathbb{Z}_n'\mathbb{X}_n}{n}\hat{\Omega}_n(\frac{1}{n}\sum_{i=1}^{n} Z_iZ_i'(Y_i - X_i'\hat{\beta}_n)^2)\hat{\Omega}_n\frac{\mathbb{X}_n'\mathbb{Z}_n}{n}\{\frac{\mathbb{X}_n'\mathbb{Z}_n}{n}\hat{\Omega}_n\frac{\mathbb{Z}_n'\mathbb{X}_n}{n}\}^{-1}.$$

3.2.4 Weighting Matrix

In Lemma 3.2.2 and Theorem 3.2.1, we have in fact characterized the asymptotic behavior of a *family* of estimators indexed by the different possible choices of weighting

matrices $\hat{\Omega}_n$. We next discuss the two most prevalent choices of $\hat{\Omega}_n$.

3.2.4.1 Two Stage Least Squares

The Two Stage Least Squares (TSLS) estimator corresponds to setting $\hat{\Omega}_n$ to equal

$$\hat{\Omega}_n = (\frac{1}{n}\sum_{i=1}^{n} Z_i Z_i')^{-1} = (\frac{1}{n}\mathbb{Z}_n'\mathbb{Z}_n)^{-1}. \tag{3.41}$$

To appreciate where the name "Two Stage Least Squares" comes from, we plug in (3.41) into the formula for the IV estimator derived in Lemma 3.2.1 to obtain

$$\begin{aligned}\hat{\beta}_n &= \{\mathbb{X}_n'\mathbb{Z}_n\hat{\Omega}_n\mathbb{Z}_n'\mathbb{X}_n\}^{-1}\mathbb{X}_n'\mathbb{Z}_n\hat{\Omega}_n\mathbb{Z}_n'\mathbb{Y}_n \\ &= \{\mathbb{X}_n'\mathbb{Z}_n(\frac{1}{n}\mathbb{Z}_n'\mathbb{Z}_n)^{-1}\mathbb{Z}_n'\mathbb{X}_n\}^{-1}\mathbb{X}_n'\mathbb{Z}_n(\frac{1}{n}\mathbb{Z}_n'\mathbb{Z}_n)^{-1}\mathbb{Z}_n'\mathbb{Y}_n \\ &= \{\mathbb{X}_n'\mathbb{Z}_n(\mathbb{Z}_n'\mathbb{Z}_n)^{-1}\mathbb{Z}_n'\mathbb{X}_n\}^{-1}\mathbb{X}_n'\mathbb{Z}_n(\mathbb{Z}_n'\mathbb{Z}_n)^{-1}\mathbb{Z}_n'\mathbb{Y}_n. \end{aligned} \tag{3.42}$$

Recall that in Chapter 2, we had introduced the matrix $\mathbb{P}_n = \mathbb{X}_n(\mathbb{X}_n'\mathbb{X}_n)^{-1}\mathbb{X}_n'$, which had the property that $\mathbb{P}_n\mathbb{Y}_n$ returned the fitted values of regressing $\mathbb{Y}_n$ on $\mathbb{X}_n$ (see (2.37)). Mathematically, the matrix $\mathbb{P}_n$ simply returned the projection of any vector in $\mathbf{R}^n$ into the column space of $\mathbb{X}_n$. Analogously, we may define the matrix

$$\mathbb{P}_n^Z \equiv \mathbb{Z}_n(\mathbb{Z}_n'\mathbb{Z}_n)^{-1}\mathbb{Z}_n', \tag{3.43}$$

which projects any vector in $\mathbf{R}^n$ into the column space of $\mathbb{Z}_n$ – i.e. for any vector $v \in \mathbf{R}^n$ it returns the fitted values from regressing v on $\mathbb{Z}_n$.

Next, let $\hat{\mathbb{X}}_n = \mathbb{P}_n^Z\mathbb{X}_n$, which corresponds to the fitted values obtained from regressing $\mathbb{X}_n$ on $\mathbb{Z}_n$. Since $\mathbb{P}_n^Z = (\mathbb{P}_n^Z)'$ and $\mathbb{P}_n^Z\mathbb{P}_n^Z = \mathbb{P}_n^Z$, we then obtain from (3.42) that

$$\begin{aligned}\hat{\beta}_n = \{\mathbb{X}_n'\mathbb{Z}_n(\mathbb{Z}_n'\mathbb{Z}_n)^{-1}\mathbb{Z}_n'\mathbb{X}_n\}^{-1}\mathbb{X}_n'\mathbb{Z}_n(\mathbb{Z}_n'\mathbb{Z}_n)^{-1}\mathbb{Z}_n'\mathbb{Y}_n \\ = \{(\mathbb{P}_n^Z\mathbb{X}_n)'(\mathbb{P}_n^Z\mathbb{X}_n)\}^{-1}(\mathbb{P}_n^Z\mathbb{X}_n)'\mathbb{Y}_n = \{\hat{\mathbb{X}}_n'\hat{\mathbb{X}}_n\}^{-1}\hat{\mathbb{X}}_n'\mathbb{Y}_n.\end{aligned} \tag{3.44}$$

However, the right hand side of (3.44) is simply the regression coefficient obtained from regressing $\mathbb{Y}_n$ on $\hat{\mathbb{X}}_n$ (recall Lemma 2.2.1 for the formula). In words, we have shown that the two stage least squares estimator, defined as solving (3.25) with $\hat{\Omega}_n = \mathbb{Z}_n'\mathbb{Z}_n/n$, is numerically equivalent to the estimator obtained by following

STEP 1: Regress $\mathbb{X}_n$ on the set of instruments $\mathbb{Z}_n$ and obtain the fitted values $\hat{\mathbb{X}}_n$. ■

STEP 2: Regress $\mathbb{Y}_n$ on $\hat{\mathbb{X}}_n$ to obtain the estimator for β_0. ■

Notice, however, that for obtaining standard errors it is incorrect to simply employ the OLS formula of Theorem 2.3.1 with $\hat{\mathbb{X}}_n$ in place of $\mathbb{X}_n$. The reason is that such

a formula ignores that $\hat{\mathbb{X}}_n$ has been estimated in a preliminary step. In contrast, the asymptotic distribution of Theorem 3.2.1 properly accounts for both STEP 1 and STEP 2 and yields the correct standard errors for the two stage least squares estimator.

3.2.4.2 Efficient Estimation

An alternative approach for selecting the weighting matrix $\hat{\Omega}_n$ is to aim to make the asymptotic variance of the corresponding estimator as "small" as possible. In order to characterize such an "optimal" choice we first need to understand what "small" means.

Suppose we possess two estimators $\hat{\theta}_1 \in \mathbf{R}^d$ and $\hat{\theta}_2 \in \mathbf{R}^d$ for a common parameter $\theta_0 \in \mathbf{R}^d$. For simplicity, we let $\hat{\theta}_1$ and $\hat{\theta}_2$ be normally distributed with mean θ_0 so

$$(\hat{\theta}_1 - \theta_0) \sim N(0, \Sigma_1) \qquad (\hat{\theta}_2 - \theta_0) \sim N(0, \Sigma_2), \tag{3.45}$$

where we note that the covariance matrices of $\hat{\theta}_1$ and $\hat{\theta}_2$ are different. Because $\hat{\theta}_1$ and $\hat{\theta}_2$ are vectors, it may not be straightforward to "rank" these estimators. For instance, $\hat{\theta}_1$ may be a better estimator for the first coordinate of θ_0 (i.e. the variance of the first coordinate of $\hat{\theta}_1$ is smaller), while $\hat{\theta}_2$ may be a better estimator for the last coordinate of θ_0 (i.e. the variance of the d^{th} coordinate of $\hat{\theta}_2$ is smaller). Unambiguously stating that $\hat{\theta}_1$ has smaller mean squared error than $\hat{\theta}_2$ would require that

$$\mathrm{Var}\{c'(\hat{\theta}_1 - \theta_0)\} \leq \mathrm{Var}\{c'(\hat{\theta}_2 - \theta_0)\} \tag{3.46}$$

for all $c \in \mathbf{R}^d$ – as a special case (3.46) would imply all coordinates of $\hat{\theta}_1$ have a smaller variance than the corresponding coordinate of $\hat{\theta}_2$. However, since $\mathrm{Var}\{c'(\hat{\theta}_j - \theta_0)\} = c'\Sigma_j c$ (for $j \in \{1, 2\}$), the requirement in (3.46) is equivalent to demanding

$$\begin{aligned}\mathrm{Var}\{c'(\hat{\theta}_2 - \theta_0)\} - \mathrm{Var}\{c'(\hat{\theta}_1 - \theta_0)\} \geq 0 \text{ for all } c \in \mathbf{R}^d \\ \text{if and only if } c'(\Sigma_2 - \Sigma_1)c \geq 0 \text{ for all } c \in \mathbf{R}^d,\end{aligned} \tag{3.47}$$

which in turn is equivalent to the matrix $\Sigma_2 - \Sigma_1$ being positive semidefinite. We write $\Sigma_2 - \Sigma_1$ being positive semidefinite as "$\Sigma_2 - \Sigma_1 \geq 0$".

Returning to the choice of $\hat{\Omega}_n$, it follows from our discussion that we should aim to find a weighting matrix $\hat{\Omega}_n$ such that the asymptotic variance of Theorem 3.2.1 is as "small" as possible (in the sense of positive semi-definiteness). It is unclear whether an optimal $\hat{\Omega}_n$ even exists, but fortunately the following Lemma gives us a positive answer.

Lemma 3.2.3. *If A is a $d_z \times d_z$ positive definite matrix, and B is a $d_z \times d_x$ matrix with rank d_x, then $(B'B)^{-1}B'AB(B'B)^{-1} - (B'A^{-1}B)^{-1}$ is a positive definite matrix.*

We apply this Lemma with $B = \Omega^{1/2}E[ZX']$ and $A = \Omega^{1/2}E[ZZ'U^2]\Omega^{1/2}$. Under

theses choices, Lemma 3.2.3 enables us to conclude that

$$\{E[XZ']\Omega E[ZX']\}^{-1}E[XZ']\Omega E[ZZ^2U^2]\Omega E[ZX']\{E[XZ']\Omega E[ZX']\}^{-1}$$
$$= (B'B)^{-1}B'AB(B'B)^{-1} \geq (B'A^{-1}B)^{-1} = (E[XZ']\{E[ZZ'U^2]\}^{-1}E[ZX'])^{-1}.$$

In other words, regardless of the choice of $\hat{\Omega}_n$, the asymptotic variance of $\hat{\beta}_n$ has a lower bound (again in terms of positive semi-definiteness) that is equal to

$$(E[XZ']\{E[ZZ'U^2]\}^{-1}E[ZX'])^{-1}. \tag{3.48}$$

It follows that if we can find a choice of $\hat{\Omega}_n$ that delivers the asymptotic variance in (3.48), then we will have found an optimal estimator. In particular, note that if

$$\Omega = \{E[ZZ'U^2]\}^{-1},$$

then plugging into the asymtpotic variance in Theorem 3.2.1 we recover precisely (3.48)

$$(E[XZ']\Omega E[ZX'])^{-1}E[XZ']\Omega E[ZZ'U^2]\Omega E[ZX'](E[XZ']\Omega E[ZX'])^{-1}$$
$$= (E[XZ'](E[ZZ'U^2])^{-1}E[ZX'])^{-1}. \tag{3.49}$$

The "optimal" choice of $\Omega = (E[ZZ'U^2])^{-1}$ is unknown since it depends on the distribution of the data. Fortunately, Theorem 3.2.1 implies we will still attain the desired asymptotic variance if we employ a consistent estimator $\hat{\Omega}_n$ instead. The resulting procedure is sometimes referred to as three stage least squares.

STEP 1: Obtain an estimator $\tilde{\beta}_n$ that is consistent for β_0 – for instance by solving (3.25) with $\hat{\Omega}_n = I_{d_z}$, or by employing the two stage least squares estimator. ■

STEP 2: Employing $\tilde{\beta}_n$ create residuals $\tilde{U}_i = (Y_i - X_i'\tilde{\beta}_n)$ and set $\hat{\Omega}_n$ to equal

$$\hat{\Omega}_n = (\frac{1}{n}\sum_{i=1}^{n} Z_i Z_i'\tilde{U}_i^2)^{-1}.$$

STEP 3: Compute $\hat{\beta}_n$ by solving (3.25) employing $\hat{\Omega}_n$ from STEP 2. ■

In retrospect, the fact that the optimal choice of Ω is to set $\Omega = (E[ZZ'U^2])^{-1}$ is unsurprising. If we employ the definition of $\hat{\beta}_n$ in (3.25) we see that

$$\hat{\beta}_n = \arg\min_{b\in\mathbf{R}^d}(\frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i'b)Z_i)'\{E[(Y - X'\beta_0)^2ZZ']\}^{-1}(\frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i'b)Z_i)'.$$

Intuitively, the optimal estimator thus simply weights each of the moment conditions by how informative (i.e. how "precise") they are.

3.3 Inference

We next turn to inference. Once again, we focus on Wald tests for brevity, but emphasize other tests are available in this context; see in particular Newey and McFadden (1994). In addition, we discuss so called "overidentification tests".

3.3.1 Non-Linear Restrictions

Due to the similarities with the material in Section 2.4, we consider only the problem of testing nonlinear restrictions on β_0 by employing a Wald Test. Specifically, suppose that for some $f : \mathbf{R}^{d_x} \to \mathbf{R}^p$ we are interested in testing the null hypothesis

$$H_0 : f(\beta_0) = 0 \qquad H_1 : f(\beta_0) \neq 0 \tag{3.50}$$

We let Σ_0 denote the asymptotic variance matrix of Theorem 3.2.1, which recall is

$$\Sigma_0 \equiv (E[XZ']\Omega E[ZX'])^{-1} E[ZX']\Omega E[ZZ'U^2]\Omega E[XZ'](E[XZ']\Omega E[ZX'])^{-1}.$$

Setting $\nabla f(\beta_0)$ to denote the $p \times d_x$ matrix of partial derivatives evaluated at β_0, the Delta Method (recall Theorem 1.3.4) and Theorem 3.2.1 imply that

$$\sqrt{n}\{f(\hat{\beta}_n) - f(\beta_0)\} \xrightarrow{d} N(0, \nabla f(\beta_0)\Sigma_0 \nabla f(\beta_0)'). \tag{3.51}$$

In addition, since $\hat{\beta}_n$ is a consistent estimator for β_0, the continuous mapping theorem implies that $\nabla f(\hat{\beta}_n)$ is consistent for $\nabla f(\beta_0)$ provided ∇f is continuous at β_0. For any consistent estimator $\hat{\Sigma}_n$ for Σ_0, we may therefore employ $\nabla f(\hat{\beta}_n)\hat{\Sigma}_n \nabla f(\hat{\beta}_n)'$ as a consistent estimator for the asymptotic variance in (3.51).

The preceding discussion leads to the immediate validity of a Wald Test.

Lemma 3.3.1. *Let Assumptions IV-1, IV-2, IV-3, IV-4 hold, $\hat{\Sigma}_n \xrightarrow{p} \Sigma_0$, ∇f be continuous at β_0, $\nabla f(\beta_0)\Sigma_0 \nabla f(\beta_0)'$ be invertible, and $c_{1-\alpha}$ denote the $1-\alpha$ quantile of a χ^2_p distribution. If $f(\beta_0) = 0$, then it follows that*

$$\lim_{n\to\infty} P(\|(\nabla f(\hat{\beta}_n)\hat{\Sigma}_n \nabla f(\hat{\beta}_n)')^{-1/2}\sqrt{n} f(\hat{\beta}_n)\|^2 > c_{1-\alpha}) = \alpha.$$

PROOF: Identical to the proof of Lemma 2.4.3. ■

3.3.2 Overidentification Tests

An interesting distinction with the material in Chapter 2 occurs when the number of moments (i.e. d_z) is larger than the number of parameters (i.e. d_x). In such a setting,

there is no guarantee that there indeed exists a β_0 solving the restrictions

$$E[(Y - X'\beta_0)Z] = 0. \tag{3.52}$$

However, the null hypothesis that a solution to (3.52) exists can be tested by employing so called "Sargan tests", "J-tests", or more generally, "overidentification tests".

3.3.2.1 The J-Test

Whenever d_z is larger than d_x, the implied residuals from estimation are unlikely to zero the sample moment conditions exactly. Formally, we will find that

$$\begin{aligned}\min_{b\in\mathbf{R}^{d_x}} (\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(Y_i - X_i'b)Z_i)'\hat{\Omega}_n(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(Y_i - X_i'b)Z_i) \\ = (\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(Y_i - X_i'\hat{\beta}_n)Z_i)'\hat{\Omega}_n(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(Y_i - X_i'\hat{\beta}_n)Z_i) \neq 0\end{aligned} \tag{3.53}$$

Nonetheless, note that if we plugged in β_0 in place of $\hat{\beta}_n$ and employed the weighting matrix $\hat{\Omega}_n = \{E[ZZ'U^2]\}^{-1}$, then we would obtain that

$$\begin{aligned}(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(Y_i - X_i'\beta_0)Z_i)'\{E[ZZ'U^2]\}^{-1}(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(Y_i - X_i'\beta_0)Z_i) \\ = \|\{E[ZZ'U^2]\}^{-1/2}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}Z_iU_i\|^2 \\ \xrightarrow{d} \|N(0, \{E[ZZ'U^2]\}^{-1/2}E[ZZ'U^2]\{E[ZZ'U^2]\}^{-1/2})\|^2\end{aligned} \tag{3.54}$$

by the continuous mapping and central limit theorems. In other words, the residuals do not "explode" but converge in distribution (after scaling by $\sqrt{n}$).

Based on this observation the so called J test, or Sargan-Hansen test due to Sargan (1958) and Hansen (1982), employs the same principle but with $\hat{\beta}_n$ in place of β_0 and $\hat{\Omega}_n$ a consistent estimator for $\{E[ZZ'U^2]\}^{-1}$. Formally, the J statistic equals

$$J_n = \|\hat{\Omega}_n^{-1/2}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(Y_i - X_i'\hat{\beta}_n)Z_i\|^2.$$

Our next result obtains the asymptotic distribution of the J-statistic.

Theorem 3.3.1. *If Assumptions IV-1-IV-4 hold, and* $\hat{\Omega}_n \xrightarrow{p} \{E[ZZ'U^2]\}^{-1}$, *then*

$$J_n \xrightarrow{d} \chi^2_{d_z - d_x}$$

PROOF: We rely on some of the arguments employed in the proof of Theorem 3.2.1, and thus we skip some steps to avoid redundancies. We first rewrite J_n as

$$J_n = \|\hat{\Omega}_n^{1/2}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(Y_i - X_i\hat{\beta}_n)Z_i\|^2 = \|\hat{\Omega}_n^{1/2}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{U_i - X_i'(\hat{\beta}_n - \beta_0)\}Z_i\|^2, \quad (3.55)$$

and proceed to analyze the terms on the right hand side of (3.55). First note the central limit theorem implies $n^{-1/2}\sum_{i=1}^{n} U_i Z_i$ is asymptotically normally distributed and hence

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_i U_i = O_p(1). \quad (3.56)$$

Therefore, since $\hat{\Omega}_n \xrightarrow{p} \Omega$ by Assumption IV-2(i) and $O_p(1) \times o_p(1) = o_p(1)$ we obtain

$$\begin{aligned}\hat{\Omega}_n^{1/2}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_i U_i &= \Omega^{1/2}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_i U_i + \{\hat{\Omega}_n^{1/2} - \Omega^{1/2}\}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_i U_i \\ &= \Omega^{1/2}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_i U_i + o_p(1). \end{aligned} \quad (3.57)$$

By similar arguments, note that $\sqrt{n}\{\hat{\beta}_n - \beta_0\} = O_p(1)$ by Theorem 3.2.1 and hence $\hat{\Omega}_n \xrightarrow{p} \Omega$, $\mathbb{Z}_n'\mathbb{X}_n/n \xrightarrow{p} E[ZX']$ and $O_p(1) \times o_p(1) = o_p(1)$ allow us to conclude that

$$\begin{aligned}&\hat{\Omega}_n^{1/2}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_i X_i'(\hat{\beta}_n - \beta_0) \\ &\quad = \Omega^{1/2}E[ZX']\sqrt{n}\{\hat{\beta}_n - \beta_0\} + \{\hat{\Omega}_n^{1/2}\frac{1}{n}\sum_{i=1}^{n} Z_i X_i' - \Omega^{1/2}E[ZX']\}\sqrt{n}\{\hat{\beta}_n - \beta_0\} \\ &\quad = \Omega^{1/2}E[ZX']\sqrt{n}\{\hat{\beta}_n - \beta_0\} + o_p(1). \end{aligned} \quad (3.58)$$

Next, note that by Lemma 3.2.1 (see (3.38) for details) we further obtain that

$$\begin{aligned}\sqrt{n}\{\hat{\beta}_n - \beta_0\} &= \{(\frac{1}{n}\mathbb{X}_n'\mathbb{Z}_n)\hat{\Omega}_n(\frac{1}{n}\mathbb{Z}_n'\mathbb{X}_n)\}^{-1}(\frac{1}{n}\mathbb{X}_n'\mathbb{Z}_n)\hat{\Omega}_n(\frac{1}{\sqrt{n}}\mathbb{Z}_n'\mathbb{U}_n) + o_p(1) \\ &= \{E[XZ']\Omega E[ZX']\}^{-1}E[XZ']\Omega\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_i U_i + o_p(1), \end{aligned} \quad (3.59)$$

where in the second equality we again employed (3.56), $\mathbb{Z}_n'\mathbb{X}_n/n \xrightarrow{p} E[ZX']$, $\mathbb{X}_n'\mathbb{Z}_n/n \xrightarrow{p} E[XZ']$ and $O_p(1) \times o_p(1) = o_p(1)$. Further defining the $d_z \times d_z$ matrix

$$\mathbb{M} \equiv \{I - \Omega^{1/2}E[ZX']\{E[XZ']\Omega E[ZX']\}^{-1}E[XZ']\Omega^{1/2}\} \quad (3.60)$$

we can then combine results (3.57) and (3.59) and some algebra to conclude that

$$\begin{aligned}\hat{\Omega}_n^{1/2}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{U_i - X_i'(\hat{\beta}_n - \beta_0)\}Z_i \\ &= \{\Omega^{1/2} - \Omega^{1/2}E[ZX']\{E[XZ']\Omega E[ZX']\}^{-1}E[XZ']\Omega\}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_iU_i + o_p(1) \\ &= \mathbb{M}\{\Omega^{1/2}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_iU_i\} + o_p(1). \qquad (3.61)\end{aligned}$$

Finally, observe that since $\Omega = (E[ZZ'U^2])^{-1}$, the continuous mapping theorems implies

$$\{\Omega^{1/2}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_iU_i\} \xrightarrow{d} N(0, (E[ZZ'U^2])^{-1/2}E[ZZ'U^2](E[ZZ'U^2])^{-1/2}) = N(0, I_{d_z}) \qquad (3.62)$$

Thus, combining (3.55), (3.61) and (3.62) with the continuous mapping theorem yields

$$J_n \xrightarrow{d} \|\mathbb{M} \times N(0, I_{d_z})\|^2. \qquad (3.63)$$

The theorem follows from (3.63) because $\mathbb{M}$ is idempotent of rank $d_z - d_x$ and $\|\mathbb{A} \times N(0, I_d)\|^2 \sim \chi^2_a$ distribution whenever $\mathbb{A}$ is a $d \times d$ idempotent matrix of rank a. ■

Notice that if we had plugged in the true parameter β_0 into the moments (as in the derivations in (3.54)), then we would obtain a $\chi^2_{d_z}$ limiting distribution for our test statistic. In contrast, the asymptotic distribution of J_n, which employs $\hat{\beta}_n$ instead, equals $\chi^2_{d_z - d_x}$. The smaller degrees of freedom $d_z - d_x$ reflects that implicitly d_x moments are being employed to estimate $\beta_0 \in \mathbf{R}^{d_x}$. As a result, it is not possible to test the validity of those moments and we are left with only $d_z - d_x$ moments to test.

3.3.2.2 Incremental Sargan Tests

While time does not permit us to examine the local power properties of the J test in detail, you should think of it as considering violation of any particular moment as equally likely. (Very) Loosely speaking, a test has a "fixed" amount of power, and the J test "spends" it by assigning an equal amount on detecting a violation of every moment.

However, in empirical work it is not uncommon for researchers to "trust" some moment conditions more than others. In such a setting, the J test may not be the best choice of test as it spends all its power equally among the moment conditions (including the ones we trust). Instead, it may be preferable to employ a test that directs its power at detecting violations of the moment conditions we are unsure about while spending no power in detecting violations of the moment conditions that we trust. This strategy was, for example, employed by Eichenbaum et al. (1988) in estimating an

aggregate consumption and leisure model that relied on moment conditions linked to both intratemporal and intertemporal Euler equations. In their analysis, they devised a test specifically designed to detect violations of the intratemporal Euler equations. The resulting test is sometimes referred to as *Incremental Sargan Tests* (Arellano, 2003).

To describe the incremental Sargan test, suppose that $Z \in \mathbf{R}^{d_z}$ can be decomposed into $Z = (Z_1', Z_2')'$ with $Z_1 \in \mathbf{R}^{d_{z_1}}$ and $Z_2 \in \mathbf{R}^{d_{z_2}}$, and where $d_{z_1} + d_{z_2} = d_z$. Given this notation, we may decompose the moment restriction into the two sets

$$E[(Y - X'\beta_0)Z_1] = 0 \qquad E[(Y - X'\beta_0)Z_2] = 0. \tag{3.64}$$

Without loss of generality, we assume that we "trust" the moment conditions corresponding to Z_1, but are unsure of the validity of the moment conditions corresponding to Z_2. We will also assume that $d_{z_1} \geq d_x$, which means that the moment conditions that we trust suffice for identifying β_0. In addition, we let $\hat{\Omega}_{1n}$ and $\hat{\Omega}_n$ satisfy

$$\hat{\Omega}_{1n} \xrightarrow{p} \{E[Z_1 Z_1' U^2]\}^{-1} \qquad \hat{\Omega}_n \xrightarrow{p} \{E[ZZ'U^2]\}^{-1}, \tag{3.65}$$

where we note that $\hat{\Omega}_{1n}$ and $\hat{\Omega}_n$ are $d_{z_1} \times d_{z_1}$ and $d_z \times d_z$ matrices. Finally, we set

$$\begin{aligned} J_n &\equiv \inf_{b \in \mathbf{R}^{d_x}} \|\hat{\Omega}_n^{1/2} \frac{1}{\sqrt{n}} \sum_{i=1}^n (Y_i - X_i'b) Z_i\|^2 \\ J_{1n} &\equiv \inf_{b \in \mathbf{R}^{d_x}} \|\hat{\Omega}_{1n}^{1/2} \frac{1}{\sqrt{n}} \sum_{i=1}^n (Y_i - X_i'b) Z_{1i}\|^2; \end{aligned} \tag{3.66}$$

i.e. J_n is the J-statistic based on the full vector of moments (as in Section 3.3.2.1), while J_{1n} is the J-statistic based only the subset of moment conditions that we trust.

Intuitively, if we trust the moment conditions corresponding to Z_1, then we should not interpret violations of those moments in the sample as evidence against the model – i.e. violations would be attributed to sampling uncertainty instead. Motivated by this logic, the incremental Sargan test statistic is then defined as

$$J_n - J_{1n}; \tag{3.67}$$

i.e. the incremental Sargan test statistic subtracts from the usual J statistic the J statistic computed using only the "trusted" moments (the moments corresponding to Z_1). Under the null hypothesis that all moment restrictions are satisfied, it follows

$$J_n - J_{1n} \xrightarrow{d} \chi^2_{d_z - d_{z_1}}. \tag{3.68}$$

In particular, note that since $d_{z_1} \geq d_x$, the increment Sargan test statistic employs smaller critical values – this is intuitive since the test statistic itself is smaller when

$d_{z_1} > d_x$. We omit a proof, but refer to Arellano (2003) for a textbook treatment.

3.4 Extensions and Challenges

We conclude our discussion of instrumental variable methods with a brief description of the weak instrument problem and a discussion of treatment effect heterogeneity.

3.4.1 Weak Instruments

In the pursue of an exogenous instrument, it is not uncommon to find instruments that are weakly correlated with the endogenous variable. However, as argued by Bound et al. (1995), instruments that do not exhibit a "strong" enough correlation with the endogenous variable can render the asymptotic approximations we derived highly inaccurate.

The problem of weak instruments is most easily illustrated with scalar variables, so we suppose $Y, X, Z \in \mathbf{R}$. In this setting, the two stage least squares estimator equals

$$\hat{\beta}_n = (\mathbb{Z}_n'\mathbb{X}_n)^{-1}\mathbb{Z}_n'\mathbb{Y}_n = \frac{\sum_{i=1}^n Z_i Y_i}{\sum_{i=1}^n Z_i X_i} = \beta_0 + \frac{\sum_{i=1}^n Z_i U_i}{\sum_{i=1}^n Z_i X_i}, \tag{3.69}$$

where recall $U_i = (Y_i - X_i\beta_0)$; i.e. $\hat{\beta}_n$ is simply the ratio of two sample means. The exogeneity assumption on the instrument allows us to analyze the numerator in (3.69) by implying that $E[ZU] = 0$. In turn, the rank condition on the instrument affects the denominator in (3.69) by requiring that $E[XZ] \neq 0$. Therefore, we obtain

$$(\frac{1}{\sqrt{n}}\sum_{i=1}^n Z_i U_i, \frac{1}{n}\sum_{i=1}^n Z_i X_i) \xrightarrow{d} (N(0, E[Z^2U^2]), E[ZX]), \tag{3.70}$$

and hence if $E[ZX] \neq 0$ we can apply the continuous mapping theorem to conclude

$$\sqrt{n}\{\hat{\beta}_n - \beta_0\} \xrightarrow{d} \frac{1}{E[ZX]} N(0, E[Z^2U^2]) = N(0, \frac{E[Z^2U^2]}{(E[ZX])^2}). \tag{3.71}$$

Notice that, as expected, result (3.71) is a special case of Theorem 3.2.1.

The asymptotic approximation in (3.71) is implicitly modelling finite sample settings in which the instrument is "strong" in the sense that $E[ZX]$ is sufficiently different from zero. To appreciate this point, we decompose $\sqrt{n}\{\hat{\beta}_n - \beta_0\}$ into the following sum

$$\sqrt{n}\{\hat{\beta}_n - \beta_0\} = \frac{1}{\sqrt{n}}\sum_{i=1}^n Z_i U_i \times \frac{1}{E[ZX]} + \frac{1}{\sqrt{n}}\sum_{i=1}^n Z_i U_i \times \Big\{\frac{1}{\frac{1}{n}\sum_{i=1}^n Z_i X_i} - \frac{1}{E[ZX]}\Big\}. \tag{3.72}$$

Notice that the first term in the right hand side of (3.72) is the one that delivers the asymptotic distribution in (3.71) – sometimes in the literature, this is referred to as the

first order term. Therefore, the second term in the right hand side of (3.72) is effectively ignored by the asymptotic approximation in (3.71) and hence also by Theorem 3.2.1 – sometimes in the literature, this is referred to as the *second order term*. Of course, in finite samples, the second order term does not disappear and remains to influence the distribution of $\sqrt{n}\{\hat{\beta}_n - \beta_0\}$. Intuitively, the asymptotic approximation of Theorem 3.2.1 should therefore prove accurate in applications in which the effect of the *second order term* is indeed small relative to the effect of the *first order term*.

Intuitively, an approximation that ignores the second order term is accurate when the first order term is "much larger" in magnitude. From (3.72) this means

$$|\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_i U_i \times \frac{1}{E[ZX]} \times \Big(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_i U_i \times \Big\{\frac{1}{\frac{1}{n}\sum_{i=1}^{n} Z_i X_i} - \frac{1}{E[ZX]}\Big\}\Big)^{-1}| \gg 1, \quad (3.73)$$

where we are using "$\gg$" to, loosely speaking, signify "much larger than". Alternatively, with some simple algebra we may rewrite (3.73) as requiring that

$$\begin{aligned} & |\frac{1}{E[ZX]} \times \Big(\frac{E[ZX] - \frac{1}{n}\sum_{i=1}^{n} Z_i X_i}{E[ZX] \times \frac{1}{n}\sum_{i=1}^{n} Z_i X_i}\Big)^{-1}| \gg 1 \\ \text{if and only if} \quad & |\frac{\frac{1}{n}\sum_{i=1}^{n} Z_i X_i}{E[ZX] - \frac{1}{n}\sum_{i=1}^{n} Z_i X_i}| \gg 1. \end{aligned} \quad (3.74)$$

The left hand side of (3.74) consists of the ratio of two random variables. One way to compare them is to examine the ratio of the second moments. To this end note that

$$\begin{aligned} E[(\frac{1}{n}\sum_{i=1}^{n} Z_i X_i)^2] &= (E[ZX])^2 + \frac{1}{n}\text{Var}\{ZX\} \\ E[(E[ZX] - \frac{1}{n}\sum_{i=1}^{n} Z_i X_i)^2] &= \frac{1}{n}\text{Var}\{ZX\} \end{aligned} \quad (3.75)$$

Therefore, combining (3.74) and (3.75), we conclude that in employing an approximation in which (3.73) holds, Theorem 3.2.1 effectively models finite sample setting in which

$$\begin{aligned} & \frac{(E[ZX])^2 + \frac{1}{n}\text{Var}\{ZX\}}{\frac{1}{n}\text{Var}\{ZX\}} \gg 1 \\ \text{if and only if} \quad & \frac{(E[ZX])^2}{\frac{1}{n}\text{Var}\{ZX\}} \gg 1 \quad (3.76) \\ \text{if and only if} \quad & |E[ZX]| \gg \frac{1}{\sqrt{n}}(\text{Var}\{ZX\})^{\frac{1}{2}}. \quad (3.77) \end{aligned}$$

In summary, we have argued that in order for the asymptotic distribution of Theorem 3.2.1 to be accurate, we must be in an application in which the instrument Z is "strong". Here, strong not only means that the rank condition (i.e. $E[ZX] \neq 0$) is satisfied, but moreover that it is satisfied "strongly enough" (i.e. (3.76) holds).

In response to this concern, Staiger and Stock (1997) proposed an alternative asymptotic approximation. Their analysis departed from the observation that any approximation in which the distribution of (Z, X) is fixed and n is allowed to diverge to infinity will mechanically satisfy (3.76). Therefore, an approximation that models a finite sample setting in which instruments are not sufficiently strong should let the distribution of (Z, X) change with n. To this end, we will suppose that at sample size n, (Z, X) satisfy

$$E_{P_n}[ZX] = \frac{\pi}{\sqrt{n}}, \tag{3.78}$$

where we have employed the subscript P_n on the expectation to emphasize that the distribution of (X, Z) changes with the sample size. Revisiting our arguments in deriving the asymptotic distribution of $\sqrt{n}\{\hat{\beta}_n - \beta_0\}$, now note that in place of (3.70) we have

$$\begin{aligned}
&(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_i U_i, \frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_i X_i) \\
&\quad = (\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_i U_i, \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{Z_i X_i - E_{P_n}[ZX]\} + \pi) \xrightarrow{d} \begin{pmatrix} N_1 \\ N_2 \end{pmatrix} + \begin{pmatrix} 0 \\ \pi \end{pmatrix},
\end{aligned} \tag{3.79}$$

where (N_1, N_2) follow a bivariate normal distribution with $E[N_1] = E[N_2] = 0$ and are possibly correlated. Hence, result (3.69) and the continuous mapping theorem yield

$$\{\hat{\beta}_n - \beta_0\} = \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_i U_i}{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_i X_i} \xrightarrow{d} \frac{N_1}{N_2 + \pi}, \tag{3.80}$$

which is a substantially different approximation than that derived in Theorem 3.2.1.

The weak instruments literature is large and still evolving. For a nice, but slightly outdated, review see Stock et al. (2002). The main takeaways from the literature are:

1. The weak instruments problem may manifest itself regardless of sample size. The reason is that regardless of n, there is always a distribution of (Z, X) such that $E[ZX]$ is "too small" (i.e. (3.76) fails). Indeed, one of the first applications in which practitioners noticed the weak instruments problem is Angrist and Krueger (1991), which employed census data with hundreds of thousands of observations.

2. There exist inference procedures that are robust to weak instruments, in that they control size regardless of the "strength" of the instrument – for some influential papers, see Moreira (2003) and Kleibergen (2005). These procedures usually obtain confidence intervals through test inversion rather than employing an estimator plus/minus standard error as in Section 3.3.1.

3. In a sense, the weak instruments problem is diagnosable in the data. Within the context of our discussion, for example, we could examine in the data whether

$E[ZX]$ is "too small" by examining a confidence interval based on $\sum_{i=1}^{n} Z_i X_i$. For a discussion of what it means to be "too small" see Stock and Yogo (2002).

3.4.2 Heterogeneity

We next discuss an influential view of instrumental variables put forth by Imbens and Angrist (1994), who link instrumental variables to a *potential outcomes* framework.

3.4.2.1 LATE Theorem

The Imbens and Angrist (1994) setup is most easily understood in a setting in which $Y \in \mathbf{R}$ is the outcome of interest, $D \in \{0,1\}$ is an indicator for receipt of "treatment" and $Z \in \{0,1\}$ is an instrument. For each individual, we let $Y(1)$ and $Y(0)$ denote the outcome when they receive or fail to receive treatment respectively. As econometricians, we only observed the actual outcome corresponding to treatment status, which equals

$$\begin{aligned} Y &= DY(1) + (1-D)Y(0) \\ &= Y(0) + D(Y(1) - Y(0)). \end{aligned} \tag{3.81}$$

The instrument Z is assumed to affect the treatment decision, and we thus let $D(1)$ and $D(0)$ respectively denote the treatment decision when $Z = 1$ and $Z = 0$ respectively. As in (3.81), we only observe the actual treatment status of an individual

$$D = D(0) + Z(D(1) - D(0)). \tag{3.82}$$

Note that $D(1)$ and $D(0)$ are both random variables, so that changing the value of the instrument Z may affect individuals differently.

Given this framework, we will impose the following restrictions.

Assumption LATE-1. *(i)* $(Y(1), Y(0), D(1), D(0)) \perp Z$ *and (ii)* $P(D(1) \neq D(0)) > 0$.

Assumption LATE-1(i) may be understood as requiring that the instrument be exogenous (i.e. analagous to $E[ZU] = 0$), while Assumption LATE-1(ii) may be understood as requiring the rank condition (i.e. analogous to $E[ZX] \neq 0$). Within this setup, suppose we let $X = (1, D)'$, $\beta_0 = (\beta_{00}, \beta_{01})' \in \mathbf{R}^2$ solve the moment restrictions

$$E\Big[(Y - \beta_{00} - D\beta_{01})\begin{pmatrix} 1 \\ Z \end{pmatrix}\Big] = 0, \tag{3.83}$$

which we estimate by two stage least squares – notice here a solution is guaranteed to exist since (3.83) consists of two equations and two unknowns. The connection between

Assumption LATE-1 and (3.83) is unclear, and thus a natural question to ask is what exactly is the parameter β_{01} defined in (3.83). To answer this question, note

$$\beta_{01} = \frac{\text{Cov}\{Y, Z\}}{\text{Cov}\{D, Z\}}, \tag{3.84}$$

which can be shown by solving (3.83). Also note that for any random variable $W \in \mathbf{R}$

$$\begin{aligned} &E[WZ] - E[W]E[Z] \\ &\quad = E[W|Z=1]P(Z=1) - \{E[W|Z=1]P(Z=1) + E[W|Z=0]P(Z=0)\}P(Z=1) \\ &\quad = \{E[W|Z=1](1 - P(Z=1)) - E[W|Z=0](1 - P(Z=1))\}P(Z=1). \end{aligned} \tag{3.85}$$

Therefore, plugging (3.85) with W equal to Y and equal to D we obtain the equality

$$\beta_{01} = \frac{E[Y|Z=1] - E[Y|Z=0]}{E[D|Z=1] - E[D|Z=0]}. \tag{3.86}$$

Notice that so far we have not employed Assumption LATE-1(i). By employing equations (3.81) and (3.82) together with Assumption LATE-1(i) we can conclude that

$$\begin{aligned} &E[Y|Z=1] - E[Y|Z=0] \\ &\qquad = E[Y(0) + D(1)(Y(1) - Y(0))] - E[Y(0) + D(0)(Y(1) - Y(0))] \\ &\qquad = E[(D(1) - D(0))(Y(1) - Y(0))]. \end{aligned} \tag{3.87}$$

Moreover, by identical manipulations applied to the denominator in (3.86) we can obtain

$$E[D|Z=1] - E[D|Z=0] = E[D(1) - D(0)]. \tag{3.88}$$

Therefore, combining results (3.86), (3.87), and (3.88) we can then derive the expression

$$\beta_{01} = \frac{E[(Y(1) - Y(0))(D(1) - D(0))]}{E[D(1) - D(0)]}. \tag{3.89}$$

Unfortunately, β_{01} remains a hard to interpret parameter, and one that might not be of interest. Thus, to obtain interpretability of β_{01} Imbens and Angrist (1994) impose

Assumption LATE-2. *$D(1) \geq D(0)$ almost surely (or $D(0) \geq D(1)$ almost surely).*

Assumption LATE-2 is often referred to as a *monotonicity* assumption. For simplicity, let us suppose that we have $D(1) \geq D(0)$ almost surely. The term *monotonicity* then refers to the fact that Assumption LATE-2 demands that the instrument affect the treatment decision of *all individuals* in the same direction. Intuitively, we may group

individuals into three subgroups depending on the effect of the instrument:

stayers	$D(1) = D(0)$ individuals whose treatment decision is unaffected by instrument
compliers	$D(1) > D(0)$ individuals who are induced into treatment by instrument
defiers	$D(1) < D(0)$ individuals who are induced out of treatment by instrument

In words, Assumption LATE-2 therefore imposes that there be no defiers. Under this additional requirement, we can then simplify (3.89) to obtain the following key result.

Lemma 3.4.1. *If Assumptions LATE-1 and LATE-2 hold, then it follows that*

$$\beta_{01} = E[Y(1) - Y(0)|D(1) > D(0)]. \tag{3.90}$$

PROOF: First recall that our derivations leading to result (3.89) already established

$$\beta_{01} = \frac{E[(Y(1) - Y(0))(D(1) - D(0))]}{E[D(1) - D(0)]}. \tag{3.91}$$

Next observe that by definition of $(D(0), D(1))$ and Assumption LATE-2, it follows that $D(1) - D(0) \in \{0, 1\}$ – i.e. either $D(1) = D(0)$ or $D(1) > D(0)$ (which can only happen when $D(0) = 0$ and $D(1) = 1$). We thus obtain by direct calculation that

$$E[D(1) - D(0)] = P(D(1) - D(0) = 1) = P(D(1) > D(0)). \tag{3.92}$$

Moreover, by similar arguments and the law of iterated expectations we obtain that

$$\begin{aligned} &E[(Y(1)-Y(0))(D(1) - D(0))] \\ &\quad = E[Y(1) - Y(0)|D(1) - D(0) = 1] \times P(D(1) - D(0) = 1) \\ &\quad = E[Y(1) - Y(0)|D(1) > D(0)] \times P(D(1) > D(0)). \end{aligned} \tag{3.93}$$

The claim of the Lemma therefore follows form (3.91), (3.92), and (3.93). ■

Since $Y(1) - Y(0)$ is the treatment effect for an individual, Lemma 3.4.1 states β_{01} equals the average treatment effect for the compliers. This parameter is known as the *local average treatment effect* (LATE). The discussed derivations are to some extent generalizable to the inclusion of covariates and non-binary instruments. In these settings, under appropriate conditions, the probability limit of the two stage least squares estimator corresponds to a weighted average of LATEs for different subpopulations (defined by the covariates and the instruments). It is worth emphasizing that the LATE is not necessarily an interesting parameter (though it certainly can be). While the representation in (3.90) is fairly uncontroversial, you should be aware that there is a fair amount of debate on whether it is a useful conceptual framework – for an alternative conceptual framework see Heckman and Vytlacil (2005). We do not have time to cover

these arguments in detail, but if interested you should read Deaton (2009) and Heckman and Urzua (2010) for criticisms of LATE, and Imbens (2010) for a defense. This debate is part of a more broad contrast between structural and quasi-experimental work. The *Journal of Economic Perspectives* put together a special issue on the topic, which is quite an entertaining read. If you are under the sad misconception that econometrics is dry and boring, I encourage you to read some of the articles contained therein, including Angrist and Pischke (2010), Keane (2010), Nevo and Whinston (2010), and Sims (2010).

3.4.2.2 Model Implications

We return to the basic setup outlined by Assumptions LATE-1 and LATE-2 in which $Y \in \mathbf{R}$, $D \in \{0,1\}$, and $Z \in \{0,1\}$. As in Section 3.4.2.1, we let β_{00} and β_{01} satisfy

$$E\Big[(Y - \beta_{00} - D\beta_{01})\begin{pmatrix} 1 \\ Z \end{pmatrix}\Big] = 0. \tag{3.94}$$

Viewed from the lenses of our discussion on overidentification tests in Section 3.3.2, the restriction that there exist *some* solution to the moment restrictions in (3.94) is not testable – i.e. since we have two equations in two unknowns, any distribution of (Y, X, Z) satisfying a rank condition should also be such that (3.94) holds.

Assumptions LATE-1 and LATE-2, however, impose more restrictions than just the existence of a solution to (3.94). In fact, as noted by Balke and Pearl (1994) and Angrist and Imbens (1995), Assumptions LATE-1 and LATE-2 have important implications that go beyond the identification of LATE as established in Lemma 3.4.1. Many of these important implications are a consequence of the following result.

Lemma 3.4.2. *If Assumptions LATE-1, LATE-2 hold, then for any $A \subset \mathbf{R}$ we have*

$$P(Y \in A, D = 1|Z = 1) - P(Y \in A, D = 1|Z = 0) = P(Y(1) \in A, D(1) > D(0))$$
$$P(Y \in A, D = 0|Z = 0) - P(Y \in A, D = 0|Z = 1) = P(Y(0) \in A, D(1) > D(0)).$$

Proof: We establish only the first equality since the second one follows by identical arguments. To this end, first note that by Assumption LATE-1 we have

$$\begin{aligned} &P(Y \in A, D = 1|Z = 1) - P(Y \in A, D = 1|Z = 0) \\ &\qquad = P(Y(1) \in A, D(1) = 1) - P(Y(1) \in A, D(0) = 1). \end{aligned} \tag{3.95}$$

On the other hand, since $D(0), D(1) \in \{0,1\}$, the even $D(1) = 1$ can be divided into

$D(1) = D(0) = 1$ and $D(1) > D(0)$, so that we also have that

$$\begin{aligned} &P(Y(1) \in A, D(1) = 1) \\ &\qquad = P(Y(1) \in A, D(1) = 1, D(0) = 1) + P(Y(1) \in A, D(1) > D(0)). \end{aligned} \tag{3.96}$$

In contrast, by the monotonicity condition imposed in Assumption LATE-2, the event $D(0) = 1$ is equivalent to the event $D(0) = D(1) = 1$ – i.e. the even $D(0) = 1$ and $D(1) = 0$ has zero probability by assumption. Therefore we have

$$P(Y(1) \in A, D(0) = 1) = P(Y(1) \in A, D(1) = D(0) = 1). \tag{3.97}$$

The first equality in the Lemma thus follows by combining (3.95), (3.96), and (3.96). ■

One key implication of Lemma 3.4.2 is that we may identify many other aspects of the distribution of compliers besides the LATE. In fact, by evaluating the first equality in Lemma 3.4.2 at the set $A = \mathbf{R}$ we obtain that

$$P(D = 1|Z = 1) - P(D = 1|Z = 0) = P(D(1) > D(0)). \tag{3.98}$$

Hence, combining result (3.98) with the equalities in Lemma (3.4.2) then yields that

$$\frac{P(Y \in A, D = 1|Z = 1) - P(Y \in A, D = 1|Z = 0)}{P(D = 1|Z = 1) - P(D = 1|Z = 0)} = P(Y(1) \in A|D(1) > D(0))$$

and

$$\frac{P(Y \in A, D = 0|Z = 0) - P(Y \in A, D = 0|Z = 1)}{P(D = 0|Z = 0) - P(D = 0|Z = 1)} = P(Y(0) \in A, D(1) > D(0)).$$

In particular, these manipulations show that the entire marginal distributions of $(Y(0), Y(1))$ conditional on being a complier (i.e. $D(1) > D(0)$) is identified.

Another important implication of Lemma 3.4.2 is that Assumptions LATE-1 and LATE-2 in fact impose testable restrictions on the distribution of the (Y, D, Z). In particular, since probabilities must obviously be positive, Lemma 3.4.2 implies

$$\begin{aligned} P(Y \in A, D = 1|Z = 1) &\geq P(Y \in A, D = 1|Z = 0) \\ P(Y \in A, D = 0|Z = 0) &\geq P(Y \in A, D = 0|Z = 1) \end{aligned} \tag{3.99}$$

for any set $A \in \mathbf{R}$. Thus, we may test whether Assumptions LATE-1 and LATE-2 by examining the sample analogues to the inequalities in (3.99) and assessing whether any violations are statistically significant. Testing whether a potentially infinite system of inequalities (as indexed by $A \in \mathbf{R}$) holds is a statistically challenging problem that is beyond the scope of this course. Recent work, however, has indeed devised tests for this purpose that should read if interested; see, e.g., Kitagawa (2015).

3.5 Problems

1. Suppose Z and X are of equal dimension and $\mathbb{X}_n'\mathbb{Z}_n$ is invertible. Show all estimators with the structure in (3.25) with a positive definite matrix $\hat{\Omega}_n$ are then numerically equivalent (regardless of the choice of $\hat{\Omega}_n$).

2. Show that if $E[U^2|Z] = \sigma^2$ with probability one (i.e. U is homoskedastic), then two stage least squares is the efficient estimator in the sense that it has the smallest asymptotic variance among estimators with the structure in (3.25).

3. Prove Lemma 3.2.3 (Hint: Use that if Σ_1 and Σ_2 are positive definite matrices, then $\Sigma_1 \geq \Sigma_2$ if and only if $\Sigma_1^{-1} \leq \Sigma_2^{-1}$).

4. Show that if $d_x = d_z$, then the J-statistic is equal to zero with probability one. Do not assume that $\mathbb{X}_n'\mathbb{Z}_n$ is invertible.

5. The following problem examines the use of distance to a college as an instrument when estimating returns to education; see, e.g., Card (1993). The data is a subsample of the 1979 National Longitudinal Survey – a description is available in `SchoolingData.pdf`. In what follows, we assume the model

$$Y = E\beta_{0e} + W_i'\beta_{0w} + U \tag{3.100}$$

where Y is log wages, E is years of education, $W \in \mathbf{R}^{d_w}$ is a vector of covariates including a constant, β_{0e} are "returns" to education, and $\beta_{0w} \in \mathbf{R}^{d_w}$.

   (a) Estimate β_{0e} by ordinary least squares with the covariates W including a constant, `black`, `age`, and `age squared`.

   (b) Estimate β_{0e} using two stage least squares, the same specification as in part (a), instrumenting for E with `nearc4` and using all other variables as instruments for themselves.

   (c) Is the estimator for part (b) efficient (in the sense of selecting $\hat{\Omega}_n$ optimally)?

   (d) Estimate β_{0e} using two stage least squares, but this time instrument for E using `nearc4` and `nearc2` as instruments.

   (e) Consider your answer to parts (c) and (d). Is your estimator in part (b) efficient in a broader sense? (Hint: Related to part (d), consider whether you have used all information in part (b))

6. An alternative way to conduct an incremental Sargan tests is sometimes employed empirically; see Christiano and Eichenbaum (1992) for an example. Here we explore some of its properties, but keep in mind the version we discuss here can have poorer power properties than the test in Section 3.3.2.2 when $d_{z_1} > d_x$.

As in Section 3.3.2.2 partition $Z = (Z_1', Z_2')'$ with $Z_1 \in \mathbf{R}^{d_{z_1}}$ and $Z_2 \in \mathbf{R}^{d_{z_2}}$. Assume $d_{z_1} = d_x$, and let $\hat{\beta}_{1n}$ denote an IV estimator obtained employing only the moments corresponding to Z_1 – i.e. formally we let $\hat{\beta}_{1n}$ equal

$$\hat{\beta}_{1n} = \arg \min_{b \in \mathbf{R}^{d_x}} \| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (Y_i - X_i' b) Z_{1i} \|^2.$$

In order to test whether the moment restrictions hold, we could plug in $\hat{\beta}_{1n}$ into the moments corresponding to Z_2 to obtain the statistic

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (Y_i - X_i' \hat{\beta}_{1n}) Z_{2i} \tag{3.101}$$

(a) Derive the asymptotic distribution of $\sqrt{n}\{\hat{\beta}_{1n} - \beta_0\}$. Be precise about what assumptions you need to impose to obtain this result.

(b) Obtain the asymptotic distribution of (3.101). (Hint: Remember the waning in Section 1.3.3 that *marginal* convergence does not imply *joint* convergence. You should link (3.101) to an expression that depends on $n^{-1/2} \sum_i U_i Z_i$)

(c) Propose an estimator for the asymptotic variance of (3.101), but do not worry about showing its consistency formally.

(d) Based on your answers to parts (a)-(c) propose a test for the null hypothesis that all moment restrictions are satisfied. Your test should consist of a test statistic and an appropriate critical value.

7. This problem is based on Bound et al. (1995), which pointed out the dangers of weak instruments and used Angrist and Krueger (1991) as an illustration. You should look at both these articles before starting. The data employed in Angrist and Krueger (1991) is available at Josh Angrist's website. I have downloaded it and pre-processed it for you by restriction data to the 1930-1939 cohort. The description of the variables is available in `AKDataGuide.m`

(a) Why does quarter of birth satisfy the rank condition according to Angrist and Krueger (1991)?

(b) Create a dummy variable for each quarter of birth and for each year of birth.

(c) Compute the coefficient on education from regressing `lwklywage` on a constant, `educ`, `age, age squared`, and dummies for `married, race, smsa` and all geographical regions. Your answer should match column 1 in Table 1 of Bound et al. (1995).

(d) Compute the coefficient (and standard error) for the two stage least squares estimator of the coefficient corresponding to education by instrumenting education with dummies for quarter of birth and quarter of birth interacted with

year of birth. Keep the same specification as in part (c) and instrument all other variables with themselves.

(e) For each individual generate a random value of `quarter of birth` and re-estimate part (d). Repeat this exercise one hundred times and report the average (across simulations) of your estimate for returns to education.

8. Suppose $Y, X, Z \in \mathbf{R}$ and we are interested in estimating β_0 satisfying (3.1) using the two stage least squares estimator, which we denote by $\hat{\beta}_n$. We aim to understand the asymptotic properties of $\hat{\beta}_n$ under weak instruments, so we assume

$$E_{P_n}[ZX] = \frac{\pi}{\sqrt{n}}.$$

that at sample size n. Throughout, also assume a central limit theorem applies so

$$(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_i U_i, \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{Z_i X_i - E_{P_n}[ZX]\}) \xrightarrow{d} (N_1, N_2)$$

where $(N_1, N_2) \in \mathbf{R}^2$ follow a normal distribution with $E[N_1] = E[N_2] = 0$.

(a) Is $\hat{\beta}_n$ consistent for β_0? To answer this question find, for any $\epsilon > 0$,

$$\lim_{n\to\infty} P(|\hat{\beta}_n - \beta_0| > \epsilon).$$

(b) Using your analysis in part (a) show that for any $M > 0$ and $\epsilon > 0$ we have

$$\lim_{n\to\infty} P(|\hat{\beta}_n - \beta_0| > \epsilon) \leq P(|N_1| > \epsilon M) + P(|N_2 + \pi| \leq M).$$

(c) Assume for simplicity that $\pi > 0$ and $\text{Var}\{ZX\} = \text{Var}\{ZU\} = 1$. Show that

$$\lim_{n\to\infty} P(|\hat{\beta}_n - \beta_0| > \epsilon) \leq 3\Phi(-\frac{\epsilon\pi}{1+\epsilon})$$

where Φ is the c.d.f. of a standard normal random variable. (Hint: Use part (b) and set $M = \pi/(1+\epsilon)$.)

(d) Suppose we want to make sure that the limiting probability that $|\hat{\beta}_n - \beta_0|$ exceeds 0.1 is at most 0.1. How large would π need to be according to the bound in part (c)? Propose a test of whether π satisfies this requirement.

9. Consider the potential outcomes framework of Section 3.4.2 and suppose Assumption LATE-1 is satisfied. In addition suppose $D(0) = 0$ almost surely (i.e. nobody with $Z = 0$ takes the treatment). Show that two stage least squares estimator

$$\hat{\beta}_{01n} \equiv \frac{\sum_{i=1}^{n}(Y_i - \bar{Y}_n)(Z_i - \bar{Z}_n)}{\sum_{i=1}^{n}(D_i - \bar{D}_n)(Z_i - \bar{Z}_n)}$$

is consistent for the parameter $E[Y(1) - Y(0)|D(1) = 1]$, which is known as *the average treatment effect on the treated*.

10. In this problem we study the Hausman test. Let $Y \in \mathbf{R}$, $X \in \mathbf{R}^d$, $Z \in \mathbf{R}^d$, and

$$Y_i = X_i'\beta_0 + \epsilon_i$$

with $E[\epsilon Z] = 0$. As usual, assume we have an i.i.d. sample, the matrices $E[XZ']$ and $E[XX']$ are both finite and full rank, and $E[\epsilon^2 ZZ']$ and $E[\epsilon^2 XX']$ are finite. We also denote the OLS and IV estimators as $\hat{\beta}_n^{\text{OLS}}$ and $\hat{\beta}_n^{\text{IV}}$ respectively.

(a) Suppose that in addition $E[\epsilon X] = 0$. Show that it then follows that

$$\begin{pmatrix} \sqrt{n}\{\hat{\beta}_n^{\text{OLS}} - \beta_0\} \\ \sqrt{n}\{\hat{\beta}_n^{\text{IV}} - \beta_0\} \end{pmatrix} = \begin{bmatrix} (E[XX'])^{-1} & 0 \\ 0 & (E[ZX'])^{-1} \end{bmatrix} \begin{pmatrix} \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i \epsilon_i \\ \frac{1}{\sqrt{n}} \sum_{i=1}^n Z_i \epsilon_i \end{pmatrix} + o_p(1). \quad (3.102)$$

(b) Still assuming that $E[\epsilon X] = 0$, show that for some matrix Ω it follows that

$$T_n \equiv n(\hat{\beta}_n^{\text{OLS}} - \hat{\beta}_n^{\text{IV}})'\Omega^{-1}(\hat{\beta}_n^{\text{OLS}} - \hat{\beta}_n^{\text{IV}}) \xrightarrow{d} \chi_d^2$$

where χ_d^2 is a Chi-Squared random variable with d degrees of freedom. Provide a closed form expression for Ω.

(c) Suppose that we want to test the null hypothesis that $E[\epsilon X] = 0$. Use parts (a) and (b) to propose a level α test of this null hypothesis based on T_n.

(d) Suppose that $E[X\epsilon] \neq 0$. Show that the test you proposed on part (c) then rejects with probability tending to one.

11. Suppose $\{Y_i, X_i\}_{i=1}^n$ is an i.i.d. sample with $Y_i \in \mathbf{R}$ and $X_i \in \mathbf{R}$ satisfying

$$Y_i = \alpha_0 + X_i\beta_0 + U_i$$

for unknown $\alpha_0, \beta_0 \in \mathbf{R}$. We further suppose X_i takes three values $X_i \in \{0, 1, 2\}$ and that $E[U_i|X_i] = 0$ (which note is stronger than $E[U_i(1, X_i)'] = 0$).

(a) Show that the standard OLS "exogeneity" assumption $E[U_i X_i] = 0$ is equivalent to the following moment restrictions for identifying α_0 and β_0:

$$E[(Y_i - \alpha_0 - X_i\beta_0)] = 0$$
$$E[(Y_i - \alpha_0 - \beta_0)1\{X_i = 1\}] + 2E[(Y_i - \alpha_0 - 2\beta_0)1\{X_i = 2\}] = 0.$$

(b) Notice, however, that we have instead assumed the stronger requirement $E[U_i|X_i] = 0$. Show this condition implies the moment restrictions:

$$\begin{aligned} E[(Y_i - \alpha_0)1\{X_i = 0\}] &= 0 \\ E[(Y_i - \alpha_0 - \beta_0)1\{X_i = 1\}] &= 0 \\ E[(Y_i - \alpha_0 - 2\beta_0)1\{X_i = 2\}] &= 0. \end{aligned}$$

Further show that these moment conditions imply the restrictions in part (a).

(c) Suppose we estimate (α_0, β_0) by minimizing the following criterion function

$$\Big\{\frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix}(Y_i - \alpha - X_i\beta)1\{X_i = 0\}\\(Y_i - \alpha - X_i\beta)1\{X_i = 1\}\end{pmatrix}\Big\}'\hat{\Omega}_n\Big\{\frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix}(Y_i - \alpha - X_i\beta)1\{X_i = 0\}\\(Y_i - \alpha - X_i\beta)1\{X_i = 1\}\end{pmatrix}\Big\}$$

for $\hat{\Omega}_n$ a possibly data dependent matrix satisfying $\hat{\Omega}_n \xrightarrow{p} \Omega$ for some positive definite matrix Ω. Derive the asymptotic distribution of the resulting estimator and clearly state any regularity conditions you need.

(d) Using your results from part (c), derive a closed form expression for the efficient weighting matrix that reflects the structure of the problem. Propose an efficient estimator for (α_0, β_0).

(e) When is OLS efficient in this problem? Justify your answer.

12. Assume the standard potential outcomes model, in which there are two unobserved potential outcomes $(Y_i(0), Y_i(1))$ and for a binary treatment variable $D_i \in \{0, 1\}$ we observe the realized outcome

$$Y_i = Y_i(0) + D_i(Y_i(1) - Y_i(0)).$$

Also suppose we have an instrument Z_i which takes three values $Z_i \in \{0, 1, 2\}$, there similarly are three possible treatment status $(D_i(0), D_i(1), D_i(2))$ satisfying $D_i(2) \geq D_i(1) \geq D_i(0)$ almost surely, and we observe

$$D_i = \sum_{z=0}^{2} D_i(z)1\{Z_i = z\}.$$

Finally, assume that Z_i is independent of $(Y_i(0), Y_i(1), D_i(0), D_i(1), D_i(2))$.

(a) Show that for any $j \in \{0, 1\}$ it follows that

$$E[Y|Z = j + 1] - E[Y|Z = j] = E[(Y(1) - Y(0))(D(j + 1) - D(j))].$$

Is the monotonicity assumption $D(j + 1) \geq D(j)$ needed for this result?

(b) Show that for any $j \in \{0,1\}$ it follows that

$$E[D|Z=j+1] - E[D|Z=j] = E[D(j+1) - D(j)].$$

Is the monotonicity assumption $D(j+1) \geq D(j)$ needed for this result?

(c) Combine your answers to (a) and (b) to establish the following parameters

$$\text{LATE}(j, j+1) \equiv E[Y_i(1) - Y_i(0)|D_i(j+1) > D_i(j)]$$

are identified for any $j \in \{0,1\}$. Is the monotonicity assumption $D(j+1) \geq D(j)$ needed to establish this result?

(d) Propose estimators for LATE$(0,1)$ and LATE$(1,2)$ and establish their consistency. Clearly state any regularity conditions you need to impose.

13. Let $\{Y_i, X_i, Z_i\}_{i=1}^n$ be an i.i.d. sample and β_0 satisfy

$$E[(Y - X'\beta_0)Z] = 0,$$

where $Y \in \mathbf{R}$, $X \in \mathbf{R}^{d_x}$, and $Z \in \mathbf{R}^{d_z}$.

(a) Show that if $b \neq \beta_0$ and $\text{rank}\{E[ZX']\} \geq d_x$, then $E[(Y - X'b)Z] \neq 0$.

(b) Let $\hat{\Sigma}_n(\beta_0) \equiv (\sum_{i=1}^n (Y_i - X_i'\beta_0)^2 Z_i Z_i'/n)$ and define the following test statistic

$$T_n(\beta_0) \equiv (\frac{1}{\sqrt{n}} \sum_{i=1}^n (Y_i - X_i'\beta_0)Z_i)' \hat{\Sigma}_n^{-1}(\beta_0) (\frac{1}{\sqrt{n}} \sum_{i=1}^n (Y_i - X_i'\beta_0)Z_i).$$

whenever $\hat{\Sigma}_n(\beta_0)$ is invertible. Formally derive the asymptotic distribution of $T_n(\beta_0)$. Clearly state any assumpionts you need to establish this result.

(c) Based on part (a) propose a test of the null hypothesis that $\beta_0 = b$ for some specified vector b – i.e. propose a test of size α for

$$H_0 : \beta_0 = b \qquad H_1 : \beta_0 \neq b$$

Formally show that the limiting rejection probability of your test under the null hypothesis is indeed α.

(d) Show that if $\beta_0 \neq b$, then the test you proposed in part (c) is consistent – i.e. probability of rejection tends to one.

(e) Is part (c) is valid even in the presence of weak instruments? Explain your answer (it is OK to not have formal arguments for this part).

14. Let $\{Y_i, X_i, Z_i\}_{i=1}^n$ be an i.i.d. sample with $Y \in \mathbf{R}$, $\mathbf{X} \in \mathbf{R}^d$, and $Z \in \mathbf{R}^d$. Suppose

that the parameter of interest β_0 satisfies

$$E[(Y - X'\beta_0)Z] = 0$$

For a $d \times d$ positive definite and invertible matrix $\hat{\Omega}$ further define the estimator

$$\hat{\beta}_n = \arg\min_{b \in \mathbf{R}^d} (\frac{1}{n}\sum_{i=1}^n (Y_i - X_i'b)Z_i)'\hat{\Omega}_n(\frac{1}{n}\sum_{i=1}^n (Y_i - X_i'b)Z_i).$$

(a) Show that if the $d \times d$ matrix $\sum_{i=1}^n Z_i X_i'$ is invertible, then the value of $\hat{\beta}_n$ does not depend on the choice of $\hat{\Omega}_n$.

(b) Let $U_i = (Y_i - X_i'\beta_0)$. Formally establish the following asymptotic expansion

$$\sqrt{n}\{\hat{\beta}_n - \beta_0\} = (E[ZX'])^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n Z_i U_i + o_p(1).$$

Clearly state any assumptions you need to establish the result.

(c) Suppose you have another potential instrument $W \in \mathbf{R}$ satisfying the moment restriction $E[UW] = 0$. Formally establish the asymptotic distribution of

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n (Y_i - X_i'\hat{\beta}_n)W_i.$$

(**Hint:** To do this carefully you'll need to use the result from part (b)).

(d) Suppose that you trust the instrument Z, but you are unsure of the validity of W – i.e., you are not sure if $E[(Y - X'\beta_0)W] = 0$. Using your results from part (c) propose a level α test for the null hypothesis that $E[(Y - X'\beta_0)W] = 0$. You do not need to formally establish its properties, but I should be able to understand how to code your proposed test from your answer.

15. Consider the standard LATE model where we have potential outcomes $(Y(0), Y(1))$, potential binary treatment decisions $(D(0), D(1))$, a binary instrument Z, covariates $X \in \mathbf{R}^d$, and we observe (Y, D, X, Z) where

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

Throughout this problem, assume $D(1) \geq D(0)$ (almost surely) and that the instrument Z is independent of $(Y(0), Y(1), D(0), D(1), X)$.

(a) Recall a complier is an individual for whom $D(1) > D(0)$. Formally show

$$P(D(1) > D(0)) = P(D = 1|Z = 1) - P(D = 1|Z = 0)$$

Clearly state what assumptions you employed in showing the result.

(b) Show that for any set $A \subseteq \mathbf{R}^d$ for which $P(X \in A) > 0$ we have that

$$P(X \in A, D(1) > D(0)) = P(X \in A, D = 1|Z = 1) - P(X \in A, D = 1|Z = 0).$$

(c) Let α_0 and β_0 be the solutions to the following set of moment conditions

$$E\left[(1\{X \in A, D = 1\} - \alpha_0 - \beta_0 D)\begin{pmatrix} 1 \\ Z \end{pmatrix}\right] = 0.$$

Formally show that $\beta_0 = P(X \in A|D(1) > D(0))$.

(d) Using part (c) propose an estimator for $P(X \in A|D(1) > D(0))$ and obtain its asymptotic distribution. Clearly state any assumptions you need.

(e) Suppose that you are interested in examining whether the distribution of observable characteristics (i.e. X) for compliers is different than the distribution of observable characteristics for the overall population. Explain how results (a)-(d) would be helpful in this regard. No formal results are required.

Chapter 4

Panel Data

We have so far focused on the standard i.i.d. model in which each individual is observed exactly once. In this chapter we study settings in which we observe an individual for multiple time periods – a scenario that both provides for additional identifying power and raises concerns that the standard i.i.d. assumption is violated. Throughout, we focus on linear models and emphasize that, from a mathematical perspective, everything in this chapter is an application of the concepts introduced in Chapters 2 and 3.

4.1 Basic Model

Below, we introduce the basic notation and terminology of panel data models. In addition, we briefly discuss so called "clustered" standard errors.

4.1.1 Definitions and Notation

For individuals $1 \leq i \leq n$ and time periods $1 \leq t \leq T$ we consider the linear model

$$Y_{it} = X_{it}'\beta_0 + U_{it} \tag{4.1}$$

where Y_{it} is an outcome variable of interest, $X_{it} \in \mathbf{R}^d$ are regressors, and U_{it} is unobservable. It will be convenient to group the observations by individual, so we set

$$X_i \equiv \begin{pmatrix} X_{i1}' \\ \vdots \\ X_{iT}' \end{pmatrix} \qquad Y_i \equiv \begin{pmatrix} Y_{i1} \\ \vdots \\ Y_{iT} \end{pmatrix} \qquad U_i \equiv \begin{pmatrix} U_{i1} \\ \vdots \\ U_{iT} \end{pmatrix}. \tag{4.2}$$

Note that X_i is a $T \times d$ matrix, Y_i is a $T \times 1$ vector, and U_i is a $T \times 1$ vector as well. We can now also stack the individual level observations into

$$\mathbb{X}_n \equiv \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} \qquad \mathbb{Y}_n \equiv \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \qquad \mathbb{U}_n \equiv \begin{pmatrix} U_1 \\ \vdots \\ U_n \end{pmatrix} \tag{4.3}$$

where now note $\mathbb{X}_n$ is an $nT \times d$ matrix, $\mathbb{Y}_n$ is an $nT \times 1$ vector, and $\mathbb{U}_n$ is an $nT \times 1$ vector as well. Intuitively, $\mathbb{X}_n$, $\mathbb{Y}_n$, and $\mathbb{U}_n$ simply "stack" all observations corresponding to an individual first, and then "stacks" across individuals to obtain the whole sample.

Throughout this chapter we will work with variants of the basic assumption:

Assumption P-1. *$\{Y_i, X_i\}_{i=1}^n$ is an i.i.d. sample.*

In essence, Assumption P-1 imposes that the data be i.i.d. across individuals, but leaves the dependence between observations corresponding to the same individual unrestricted. Assumption P-1 is appropriate for so called *short panels*, which are applications in which there is a large number of individuals (n is large) but not many observations per individual (T is small). Theoretically, we derive distributional approximations in such applications by thinking of T as fixed, letting n diverge to infinity, and employing law of large numbers and central limit theorems designed for i.i.d. data (here an "observation" is (Y_i, X_i), which contains multiple time observations per individual). We also note that assuming each individual is observed for the same amount of time periods T is not an innocuous requirement – these settings are sometimes called *balanced* panels. In applications, it is often the case that individuals are observed for varying amounts of time, for example, due to individuals leaving the sample – these settings are sometimes called *unbalanced* panels. To the extent that the length of time for which an individual is observed is independent of other variables, these applications can be analyzed by treating $\{Y_i, X_i\}_{i=1}^n$ as an independent but no identically distributed sample.

While modeling a panel data set treating T fixed and letting n diverge to infinity is often appropriate, we note that there are multiple settings in which it leads to poor approximations. For instance, in so called *long panels* the time dimension T is large while the number of individuals n is small. In such a case, it is natural to treat n as fixed and let T diverge to infinity instead. The asymptotic study of *long panels* is as a result substantially different from what we have seen so far in that it requires invoking law of large numbers and central limit theorems that allow for dependence across time. Finally, we note that there is a third category of asymptotic approximations in which both n and T are allowed to diverge to infinity at the same time. This third category is often needed to study nonlinear panel data models because their identification can require a large number of observations per individual.

4.1.2 Clustered Data

Throughout this chapter we will often refer to the t index as "time" and the i index as "individual". However, it is worth emphasizing that these are just "labels" and that, more generally we can think of the data as being divided into $1 \leq g \leq G$ "groups" where observations across groups are assumed i.i.d. but the dependence between observations within the same group is left unrestricted. To map this setup into our previous discussion, simply think of an individual as a group (i.e. $g = i$ and $n = G$) and of the observations for that individual as belonging to the same group.

The following two examples illustrate the presence of clustered data outside the traditional panel data context. For a broader overview, see Cameron and Miller (2015).

Example 4.1.1. Hersch (1998) studied the magnitude of compensating differentials due to risk in different industries. In this study, the author considers the linear regression

$$Y_{i,g} = R_g\alpha_0 + Z'_{i,g}\gamma_0 + U_{i,g},$$

where $Y_{i,g}$ is the log-wage of individual i in industry g, R_g is a measure of risk of industry g, and $Z_{i,g}$ are individual level covariates such as education an experience. Here, a cluster is an industry and by clustering at the industry level we allow for arbitrary dependence among the residuals $U_{i,g}$ across individuals in the same industry. Observations of individuals in different industries are assumed independent of each other. ■

Example 4.1.2. In experimental development economics, it is common for randomization to occur at a unit level that contains mutliple individuals. For instance, Muralidharan and Sundararaman (2011) examine the impact of implementing a teacher performance pay program in different schools in India. The unit of randomization is a school, meaning every teacher in a given school is assigned to either the treatment or control group. In order to estimate an average treatment effect, the authors estimate

$$Y_{i,g} = D_g\alpha_0 + Z'_{i,g}\gamma_0 + U_{i,g},$$

where $Y_{i,g}$ is a test score for student i in school g, D_g is an indicator for whether that school is in the treatment group, and $Z_{i,g}$ are a set of additional covariates. In this setting, there is a natural concern that the residuals $U_{i,g}$ are correlated across students in the same school, and for this reason the authors cluster at the school level. ■

The presence of dependence within clusters requires us to appropriately adjust our inference procedures, for example, by adjusting our standard errors. This point was forcefully made in the context of linear regression by Moulton (1986), but of course applies to other estimation procedures. Below, we focus on linear regression for simplicity and return to our introduced notation for panel data. We additionally require:

Assumption C-1. *(i)* $\{Y_i, X_i\}_{i=1}^n$ *is an i.i.d. sample with* Y_i *and* X_i *satisfying* (4.2)*; (ii)* $E[\sum_{t=1}^T X_{it}U_{it}] = 0$*,* $E[X_i'X_i]$ *is finite and invertible, and* $\sum_{t=1}^T E[\|X_{it}\|^2 U_{it}^2] < \infty$*.*

Beyond the i.i.d. requirement and the *balanced* panel strucutre, Assumption C-1 is fairly weak simply demanding invertibility of the matrix $E[X_i'X_i]$ and standard second moment conditions. These requirements are sufficient for studying the asymptotic properties of the ordinary least squares estimator $\hat{\beta}_n$ defined as

$$\hat{\beta}_n = \arg\min_{b \in \mathbf{R}^d} \frac{1}{n} \sum_{i=1}^n \sum_{t=1}^T (Y_{it} - X_{it}'b)^2. \tag{4.4}$$

Our next Lemma establishes the asymptotic normality of $\hat{\beta}_n$.

Lemma 4.1.1. *If Assumption C-1 holds, then it follows that*

$$\sqrt{n}\{\hat{\beta}_n - \beta_0\} \xrightarrow{d} N(0, \{E[X_i'X_i]\}^{-1} E[(\sum_{t=1}^T X_{it}U_{it})(\sum_{t=1}^T X_{it}U_{it})'] \{E[X_i'X_i]\}^{-1}).$$

Proof: This lemma is really just Theorem 2.3.1 under different notation (you should convince yourself of that). We therefore keep the proof succinct and focus on algebraic manipulations that will become common throughout the chapter. First note that

$$\frac{1}{n} \sum_{i=1}^n \sum_{t=1}^T (Y_{it} - X_{it}'\hat{\beta}_n) X_{it} = 0$$

due to the first order condition defining $\hat{\beta}_n$. Plugging in that $Y_{it} = X_{it}'\beta_0 + U_{it}$ and employing the definitions of X_i, Y_i, and U_i given in (4.2) we then obtain

$$\begin{aligned} \sqrt{n}\{\hat{\beta}_n - \beta_0\} &= \{\frac{1}{n} \sum_{i=1}^n \sum_{t=1}^T X_{it}X_{it}'\}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \sum_{t=1}^T X_{it}U_{it} \\ &= \{\frac{1}{n} \sum_{i=1}^n X_i'X_i\}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i'U_i. \end{aligned} \tag{4.5}$$

The characterization in (4.5) highlights that we may proceed by applying standard law of large numbers and central limit theorems for i.i.d. data by treating the cluster as the unit of observation. In particular, by the law of large numbers and the continuous mapping theorem we obtain that as the number of clusters n tends to infinity

$$\{\frac{1}{n} \sum_{i=1}^n X_i'X_i\}^{-1} \xrightarrow{p} \{E[X_i'X_i]\}^{-1} \tag{4.6}$$

Similarly, applying a standard central limit theorem for i.i.d. data we also obtain that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i'U_i \xrightarrow{d} N(0, E[(X_i'U_i)(X_i'U_i)']), \tag{4.7}$$

where note that $E[(X_i'U_i)(X_i'U_i)']$ is finite due to T being finite and $E[\|X_{it}\|^2 U_{it}^2]$ being finite by Assumption C-1. Moreover, since $X_i'U_i = \sum_{t=1}^{T} X_{it}U_{it}$ we also have

$$E[(X_i'U_i)(X_i'U_i)'] = E[(\sum_{t=1}^{T} X_{it}U_{it})(\sum_{t=1}^{T} X_{it}U_{it})']. \tag{4.8}$$

The claim of the Lemma therefore follows from Slutzky and results (4.6) and (4.7). ■

We can estimate the asymptotic variance of the OLS estimator under clustered data by proceeding in analogy to our results in Chapter 2. Specifically, let

$$\hat{U}_{it} \equiv Y_{it} - X_{it}'\hat{\beta}_n,$$

for $\hat{\beta}_n$ as defined in (4.4). The asymptotic variance of $\hat{\beta}_n$ can then be estimated by

$$\{\frac{1}{n}\sum_{i=1}^{n} X_i'X_i\}^{-1}\{\frac{1}{n}\sum_{i=1}^{n}(\sum_{t=1}^{T} X_{it}\hat{U}_{it})(\sum_{t=1}^{T} X_{it}\hat{U}_{it})'\}\{\frac{1}{n}\sum_{i=1}^{n} X_i'X_i\}^{-1}, \tag{4.9}$$

which we note is simply the sample analogue to the limiting asymptotic variance obtained in Lemma 4.1.1. Its consistency can be obtained by similar arguments to those employed in Lemma 2.3.3 and we therefore leave the proof of its consistency as an exercise. We do note, however, that the middle term in (4.9) equals

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}(\sum_{t=1}^{T} X_{it}\hat{U}_{it})(\sum_{t=1}^{T} X_{it}\hat{U}_{it})' \\ = \underbrace{\frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{T} X_{it}X_{it}'\hat{U}_{it}^2}_{\text{standard term}} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{T}\sum_{\tilde{t}\neq t} X_{it}X_{i\tilde{t}}'\hat{U}_{it}\hat{U}_{i\tilde{t}}}_{\text{correlation within cluster}}.\end{aligned} \tag{4.10}$$

Hence, the cluster robust variance estimator is simply the sum of a standard variance estimator term that treats all observations as independent across i and t (the "standard term" term) and an adjustment that reflects the presence of possible correlation between observations that are within the same cluster (the "correlation within cluster" term). As in Chapter 2, we note that in analogy to HC1 and HC2 there are multiple variants of (4.9) that aim to improve the finite sample performance of the asymptotic variance estimators; see Cameron and Miller (2015) for additional discussion.

Remark 4.1.1. As is hopefully clear from the proof of Lemma 4.1.1, the standard

asymptotic analysis of $\hat{\beta}_n$ relies crucially on the number of clusters tending to infinity. In particular, $\hat{\beta}_n$ can fail to be approximately normal when we have only a few clusters *even if we have a large number of total observations* (but dependence within cluster is too strong). Recent work has shown normal approximations may also be poor when there is considerably heterogeneity across clusters, e.g. the number of observations per clusters varies considerably across clusters; see Carter et al. (2017). As an alternative, a number of statistical procedures have been developed that allow for inference when the number of clusters is small provided that each cluster has a large number of observations; see Ibragimov and Müller (2010, 2016), Bester et al. (2011), Canay et al. (2017), and Canay et al. (2018) among others. These procedures, however, often require the regressors to vary at the cluster level, which can be challenging in some applications; see, for instance, Example 4.1.2. ■

Remark 4.1.2. In a highly influential paper, Cameron et al. (2011) propose a generalization of clustering to so-called multiway clustering. For illustrative purposes, consider the type of employee-employer matched data sets popularized by Abowd et al. (1999). In this application we observe multiple employees $1 \leq i \leq n$ through time $1 \leq t \leq T$ and the firm at which individual i was employed at time t, which we denote by $\mathbf{J}(i,t)$. When estimating a regression of a dependent variable Y_{it} on covariates X_{it} with the form

$$Y_{it} = X_{it}'\beta_0 + U_{it},$$

it is natural to be concerned with two types of dependence across the U_{it}. First, we may be concerned that the residuals for the same individual may be correlated ($E[U_{it}U_{it'}] \neq 0$). Second, we may be concerned that the residuals for individuals in the same firm may be correlated ($E[U_{it}U_{kt}] \neq 0$ if $\mathbf{J}(i,t) = \mathbf{J}(k,t)$). Multi-way clustered designs in essence allow arbitrary dependence among these dimensions (firms and time) and imposes independence otherwise. While intuitive, the theoretical generalizations of standard clustering techniques to this setting are very challenging. Menzel (2017) shows, for example, that sample means may even fail to be normally distributed in the presence of multi-way clustering. ■

4.2 Random Effects

A canonical model for the residual U_{it} is to assume that it contains an individual specific component A_i that is time invariant, and an individual and time specific component V_{it}. This additional assumption allows us to specialize our equation in (4.1) to obtain that

$$Y_{it} = X_{it}'\beta_0 + \underbrace{A_i + V_{it}}_{U_{it}}. \tag{4.11}$$

In its most basic version, the *random effects* model assumes that X_{it} is exogenous in the sense that it is uncorrelated with both A_i nd V_{it} and additionally imposes that A_i and V_{it} are i.i.d. – this type of assumption contrasts with the *fixed effects* model which is concerned with the potential correlation between A_i and X_{it}.

Our next assumption formalizes the setup of the random effects model. In its statement, we let I_T denote the $T \times T$ identity matrix, set $V_i = (V_{i1}, \cdots, V_{iT})'$, and employ the notation that we introduced in display (4.2).

Assumption RE-1. *(i)* $\{Y_i, X_i\}_{i=1}^n$ *is an i.i.d. sample satisfying equation* (4.11)*; (ii)* $E[A_i|X_i] = 0$ *and* $E[V_i|X_i, A_i] = 0$*; (iii)* $E[V_iV_i'|X_i, A_i] = \sigma_V^2 I_T$ *and* $E[A_i^2|X_i] = \sigma_A^2$.

Assumption RE-1 imposes the mean independence of A_i given X_i and of V_i given (X_i, A_i). While this requirement can be slightly relaxed, we impose it here for simplicity. In turn, Assumption RE-1 imposes a homoskedasticity assumption on the A_i and V_i. Crucially, however, we note that Assumption RE-1 in fact implies that the covariance matrix of U_i is *not* diagonal. To see this, simply note that $U_{it} = A_i + V_{it}$ implies

$$E[U_iU_i'|X_i] = \begin{pmatrix} \sigma_A^2 + \sigma_V^2 & \cdots & \sigma_A^2 \\ \vdots & \ddots & \vdots \\ \sigma_A^2 & \cdots & \sigma_A^2 + \sigma_V^2 \end{pmatrix} \tag{4.12}$$

due to Assumption RE-1(iii). We also note Assumption RE-1(ii) implies that the parameter of interest β_0 must satisfy all the following moment conditions:

$$E[(Y_{it} - X_{it}'\beta_0)X_{i\tilde{t}}] = 0 \text{ for all } 1 \leq t \leq T \text{ and } 1 \leq \tilde{t} \leq T. \tag{4.13}$$

As in Chapter 3, this excess of moment conditions implies that we may potentially employ different linear combinations of the moment restrictions to obtain estimators with different asymptotic variances. In particular, we note that for any $T \times T$ positive definite matrix Ω and (Y_i, X_i) as in (4.2), the parameter β_0 satisfies

$$E[X_i'\Omega(Y_i - X_i\beta_0)] = 0 \tag{4.14}$$

(you should convince yourself (4.13) implies (4.14)).

For $\hat{\Omega}_n$ some $T \times T$ estimator of Ω, equation (4.14) suggests estimating β_0 by setting

$$\frac{1}{n}\sum_{i=1}^n X_i'\hat{\Omega}_n(Y_i - X_i\hat{\beta}_n^{\text{re}}) = 0; \tag{4.15}$$

where note a $\hat{\beta}_n^{\text{re}}$ solving the above equation exists provided the $d \times d$ matrix $\sum_i X_i'\hat{\Omega}_n X_i$

is invertible. In fact, under such invertibility condition, $\hat{\beta}_n^{\text{re}}$ is simply given by

$$\hat{\beta}_n^{\text{re}} = \{\frac{1}{n}\sum_{i=1}^{n} X_i'\hat{\Omega}_n X_i\}^{-1}\{\frac{1}{n}\sum_{i=1}^{n} X_i'\hat{\Omega}_n Y_i\}. \tag{4.16}$$

We will study the asymptotic distribution of $\hat{\beta}_n^{\text{re}}$ under the following assumption:

Assumption RE-2. *(i)* $\hat{\Omega}_n \xrightarrow{p} \Omega$ *and* $E[X_i'\Omega X_i]$ *is invertible; (ii)* $E[\|X_i\|^2] < \infty$

Besides requiring that the estimated matrix $\hat{\Omega}_n$ be consistent for some limit Ω, Assumption RE-2 imposes standard full rank requirements and moment conditions. Together with Assumption RE-1, Assumption RE-2 allow us to characterize the asymptotic distribution of the estimator $\hat{\beta}_n^{\text{re}}$ as a function of Ω.

Theorem 4.2.1. *Let* $\Sigma \equiv E[U_iU_i'|X_i]$ *and Assumptions RE-1 and RE-2 hold. Then:*

$$\sqrt{n}\{\hat{\beta}_n^{\text{re}} - \beta_0\} \xrightarrow{d} N(0, (E[X_i'\Omega X_i])^{-1}E[X_i'\Omega\Sigma\Omega X_i](E[X_i'\Omega X_i])^{-1}).$$

Proof: As in the proof of Lemma 2.3.3, when we apply a norm $\|\cdot\|$ to a matrix it is understood to be the Euclidean norm in the components of the matrix. Then note that the triangle inequality, $E[\|X_i\|^2] < \infty$ and the law of large numbers implies

$$\begin{aligned}\|\frac{1}{n}\sum_{i=1}^{n} X_i'\hat{\Omega}_n X_i - \frac{1}{n}\sum_{i=1}^{n} X_i'\Omega X_i\| &= \|\frac{1}{n}\sum_{i=1}^{n} X_i'\{\hat{\Omega}_n - \Omega\}X_i\| \\ &\leq \frac{1}{n}\sum_{i=1}^{n}\|X_i\|^2 \times \|\hat{\Omega}_n - \Omega\| \xrightarrow{p} E[\|X_i\|^2] \times 0 = 0, \end{aligned} \tag{4.17}$$

where the final result follows from the continuous mapping theorem and Assumption RE-2. Hence, since $E[X_i'\Omega X_i]$ is invertible, it follows from (4.17) that $\sum_i X_i'\hat{\Omega}_n X_i/n$ is invertible with probability tending to one and thus $\hat{\beta}_n^{\text{re}}$ (as in (4.15)) is well defined. Employing the characterization in (4.16) for $\hat{\beta}_n^{\text{re}}$ we then obtain

$$\begin{aligned}\sqrt{n}\{\hat{\beta}_n^{\text{re}} - \beta_0\} &= \sqrt{n}\{(\frac{1}{n}\sum_{i=1}^{n} X_i'\hat{\Omega}_n X_i)^{-1}(\frac{1}{n}\sum_{i=1}^{n} X_i'\hat{\Omega}_n Y_i) - \beta_0\} \\ &= \sqrt{n}\{(\frac{1}{n}\sum_{i=1}^{n} X_i'\hat{\Omega}_n X_i)^{-1}(\frac{1}{n}\sum_{i=1}^{n} X_i'\hat{\Omega}_n\{X_i\beta_0 + U_i\}) - \beta_0\} \\ &= (\frac{1}{n}\sum_{i=1}^{n} X_i'\hat{\Omega}_n X_i)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i'\hat{\Omega}_n U_i \end{aligned} \tag{4.18}$$

Let $\hat{\omega}_t'$ be the t^{th} row of the $T \times T$ matrix $\hat{\Omega}_n - \Omega$ and note we can then write $\hat{\Omega}_n - \Omega$ as

$$\hat{\Omega}_n - \Omega = \begin{pmatrix} \hat{\omega}_1' \\ \vdots \\ \hat{\omega}_T' \end{pmatrix}.$$

(each $\hat{\omega}_t$ is a $T \times 1$ vector). The definition of X_i in (4.2) and some algebra then yields

$$\begin{aligned} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i' \hat{\Omega}_n U_i - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i' \Omega U_i &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i' \{\hat{\Omega}_n - \Omega\} U_i \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sum_{t=1}^{T} X_{it} (\hat{\omega}_t' U_i) = \sum_{t=1}^{T} \{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_{it} U_i'\} \hat{\omega}_t. \quad (4.19) \end{aligned}$$

However, also note that $E[X_{it} U_i'] = 0$ for every $1 \leq t \leq T$ by Assumption RE-1(ii) and $E[\|X_{it} U_i'\|^2] < \infty$ by Assumptions RE-1(ii) and RE-2(ii). Hence, we obtain that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_{it} U_i' = O_p(1)$$

for all $1 \leq t \leq T$ by the central limit theorem. Moreover, since $\hat{\Omega}_n \xrightarrow{p} \Omega$ by Assumption RE-2(i) it follows that $\hat{\omega}_t = o_p(1)$ for all $1 \leq t \leq T$. Thus, we can conclude

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i' \{\hat{\Omega}_n - \Omega\} U_i = \sum_{t=1}^{T} \{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_{it} U_i'\} \hat{\omega}_t = \sum_{t=1}^{T} O_p(1) \times o_p(1) = o_p(1) \quad (4.20)$$

due to result (4.19). Hence, since $E[X_i' \Omega U_i] = 0$ by Assumption RE-1(ii) and $E[\|X_i' \Omega U_i\|^2] < \infty$ by Assumptions RE-1(ii) and RE-2(ii), we obtain by the central limit theorem, $E[U_i U_i' | X_i] = \Sigma$ and the law of iterated expectations that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i' \hat{\Omega}_n U_i = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i' \Omega U_i + o_p(1) \xrightarrow{d} N(0, E[X_i' \Omega \Sigma \Omega X_i]). \quad (4.21)$$

Since $\sum_i X_i' \Omega X_i / n \xrightarrow{p} E[X_i' \Omega X_i]$ by the law of large numbers, we can conclude from results (4.17), (4.18), and (4.21) together with the continuous mapping theorem that

$$\begin{aligned} \sqrt{n}\{\hat{\beta}_n^{\text{re}} - \beta_0\} &= (\frac{1}{n} \sum_{i=1}^{n} X_i' \hat{\Omega}_n X_i)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i' \hat{\Omega}_n U_i \\ &\xrightarrow{d} N(0, (E[X_i' \Omega X_i])^{-1} E[X_i' \Omega \Sigma \Omega X_i] (E[X_i' \Omega X_i])^{-1}), \quad (4.22) \end{aligned}$$

which established the claim of the theorem. ■

Perhaps unsurprisingly, the asymptotic variance of the estimator $\hat{\beta}_n^{\text{re}}$ depends on the

choice of weighting matrix Ω. This dependence implies there may potentially be an "optimal" weighting matrix in the sense discussed in Chapter 3. In fact, using Lemma 3.2.3 it is possible to show that the optimal choice of Ω is to set

$$\Omega = \Sigma^{-1} \equiv (E[U_i U_i' | X_i])^{-1}.$$

As in Chapter 3, the fact that the optimal Ω is to set $\Omega = \Sigma^{-1}$ is unsurprising, as it corresponds to weighting the moment conditions according to how informative they are (i.e. how "precise" they are). In practice, Ω is of course unknown, which leads to the following procedure often referred to as the *random effects* estimator based on (4.13):

STEP 1: Obtain an estimator $\tilde{\beta}_n$ that is consistent for β_0 – for instance by solving (4.15) with $\hat{\Omega}_n$ a fixed matrix such as $\hat{\Omega}_n = I_T$. ■

STEP 2: Employing $\tilde{\beta}_n$ create residuals $\tilde{U}_{it} = (Y_{it} - X_{it}'\tilde{\beta}_n)$ and motivated by (4.13) let

$$\hat{\sigma}_A^2 \equiv \frac{1}{nT(T-1)/2} \sum_{i=1}^{n} \sum_{t=1}^{T-1} \sum_{\tilde{t}=t+1}^{T} \tilde{U}_{it}\tilde{U}_{i\tilde{t}}$$

$$\hat{\sigma}_V^2 \equiv \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} (\tilde{U}_{it})^2 - \hat{\sigma}_A^2.$$

STEP 3: Employing $\hat{\sigma}_A^2$ and $\hat{\sigma}_V^2$, compute $\hat{\beta}_n^{\text{re}}$ by solving (4.15) with $\hat{\Omega}_n$ set to equal:

$$\hat{\Omega}_n \equiv \begin{pmatrix} \hat{\sigma}_A^2 + \hat{\sigma}_V^2 & \cdots & \hat{\sigma}_A^2 \\ \vdots & \ddots & \vdots \\ \hat{\sigma}_A^2 & \cdots & \hat{\sigma}_A^2 + \hat{\sigma}_V^2 \end{pmatrix}^{-1}.$$

Note there is no guarantee that $\hat{\sigma}_A^2 > 0$, which may indicate a violation of one of our assumptions (such as $E[V_i V_i' | X_i] = \sigma_V^2 I_T$). ■

In practice, in linear models the *random effects* estimator is not as popular as the *fixed effects* estimator, which we will see in the next section. One important reason for random effect estimators being less popular that they require X_i to be exogenous in the sense that it be uncorrelated with A_i – in contrast, the next section we will see how the fixed effects estimator dispenses with this assumption. Nonetheless, studying the random effects estimator is important for a number of reasons:

1. Random effects are more often employed in nonlinear models due to the difficulties that fixed effect approaches face in such settings (see the next section).

2. The random effects assumptions allow for X_{it} to contain covariates that do not vary at the individual level (e.g. race). The coefficient on such covariates will not

be identified by fixed effect approaches and hence some version of a random effects model may be required.

3. In panel data models, we can often think of different assumptions in terms of the moment restrictions that they imply (we will see more of this when studying dynamic panel data models). As in Chapter 3, these moment restrictions can be employed to improve on the efficiency of estimators. The random effects estimator is simply an illustration of this more general idea.

4. It is good practice to think how the assumptions one imposes impact the joint distribution of the observations for an individual; e.g. how Assumption RE-1(i) has strong implications for the correlation of U_{it} across time (as in (4.13).

Remark 4.2.1. The proof of Theorem 4.2.1 did not actually rely on the specific structure of $\Sigma \equiv E[U_i U_i' | X_i]$ derived in equation (4.13). In theory, it is therefore possible to dispense of this requirement and estimate Σ under more general conditions. However, notice that the matrix Σ contains $T(T-1)/2$ unknown parameters (since it is symmetric). As a result, estimating Σ in practice often requires some structure, such as that in (4.13), which reduces the dimensionality of the estimation problem. ■

4.3 Fixed Effects

As in the random effects model, we maintain that the residual U_{it} has the structure

$$U_{it} = A_i + V_{it},$$

where A_i is again an individual specific component that is time invariant, and U_{it} is a time and individual specific shock. While the random effects estimator focuses on employing the panel structure to improve on efficiency, the fixed effects estimator aims to utilize the panel structure to relax the requirement that X_{it} and A_i be uncorrelated.

Our next example is one of the earlier applications of fixed effects.

Example 4.3.1. We return to Example 2.1.3 from Chapter 2. Suppose now we observe a panel of firms $1 \leq i \leq n$ through $1 \leq t \leq T$ time periods. In each time period, a firm produces output Y_{it} from an input X_{it} according to the production function

$$Y_{it} = \exp\{V_{it} + A_i\} X_{it}^{\beta_0},$$

where A_i and V_{it} are unobserved to the econometrician. Recall β_0 is the output elasticity, which we aim to estimate by taking logs to the production function to obtain that

$$\log(Y_{it}) = \beta_0 \log(X_{it}) + A_i + V_{it}.$$

Now suppose that the firm knows its own productivity factor A_i but that it must choose X_{it} before the i.i.d. shock V_{it} is realized. The firm then maximizes the expected profit

$$\max_x \{P_Y E[\exp\{V_{it}\}] \exp\{A_i\} x^{\beta_0} - x P_X\}$$

where P_X and P_Y are the prices for X_{it} and Y_{it}. The first order condition then yields

$$\log(X_{it}) = \frac{1}{1-\beta_0}\{\log(\frac{\beta_0 P_Y E[\exp\{V_{it}\}]}{P_X}) + A_i\}; \tag{4.23}$$

see Example 2.1.3 for details. Intuitively, equation (4.23) states more productive firms will use more inputs and hence it is not reasonable to assume that X_{it} and A_i are uncorrelated (as required by a random effects approach). Fixed effects were proposed as a solution to this challenge as early as Hoch (1962) and Mundlak and Hoch (1965); see, e.g., Ackerberg et al. (2015) for approaches that allow A_i to vary through time. ■

In the fixed effects model, we maintain the linear specification so we require that

$$Y_{it} = X_{it}'\beta_0 + A_i + V_{it}. \tag{4.24}$$

Letting $V_i \equiv (V_{i1}, \ldots, V_{iT})'$ and employing the same notation as in (4.2) we then impose the following assumption which contains the main structure of the fixed effects model.

Assumption FE-1. *(i)* $\{Y_i, X_i\}_{i=1}^n$ *is i.i.d. and satisfies* (4.24)*; (ii)* $E[V_i|X_i, A_i] = 0$.

Crucially, unlike in the random effects model, we no longer require that $E[A_i|X_i] = 0$ in order to accommodate applications such as Example 4.3.1. As we will see below, it is sometimes helpful to consider A_i as a person-specific constant in the linear model (4.24). The unknown term A_i is referred to as the "fixed effect".

Before proceeding to estimation, it is worth noticing that A_i being both unobserved and time invariant causes some complications for identification. Suppose $X_{it} = (X_{it}^{(1)\prime}, X_{it}^{(2)\prime})'$ where $X_{it}^{(1)} \in \mathbf{R}^{d_1}$ and $X_{it}^{(2)} \in \mathbf{R}^{d_2}$, and let $X_{it}^{(2)}$ not vary through time – i.e. $X_{it}^{(2)} = X_{i\tilde{t}}^{(2)}$ for all $1 \leq t \leq \tilde{t} \leq T$. Writing $X_i^{(2)} = X_{it}^{(2)}$ for simplicity and decomposing β_0 into conforming subvectors $\beta_0 = (\beta_0^{(1)\prime}, \beta_0^{(2)\prime})'$ we then obtain

$$Y_{it} = X_{it}'\beta_0 + A_i + V_{it} = X_{it}^{(1)\prime}\beta_0^{(1)} + X_i^{(2)\prime}\beta_0^{(2)} + A_i + V_{it},$$

where by Assumption FE-1 we have $E[V_{it}|X_i, A_i] = 0$. Now, for any $\tilde{\beta}_0^{(2)}$ not necessarily equal to $\beta_0^{(2)}$ we can rewrite equation (4.24) as given by

$$Y_{it} = X_{it}^{(1)\prime}\beta_0^{(1)} + X_i^{(2)\prime}\tilde{\beta}_0^{(2)} + \underbrace{A_i + X_i^{(2)\prime}(\beta_0^{(2)} - \tilde{\beta}_0^{(2)})}_{\tilde{A}_i} + V_{it}, \tag{4.25}$$

where note Assumption FE-1 continues to be satisfied with the new fixed effect $\tilde{A}_i$ since

$$E[V_i|X_i, \tilde{A}_i] = E[V_i|X_i, A_i] = 0 \tag{4.26}$$

due to $\tilde{A}_i$ being a function of X_i and A_i. Intuitively, the fixed effect A_i captures all the information that is specific to an individual and time invariant. As a result, it is not possible to separately identify A_i from the effect of a person specific regressor that is time invariant, such as race or gender.

Remark 4.3.1. Notice that, in contrast, time invariant regressors are not an issue for the random effects model. The reason is that the stronger assumption $E[A_i|X_i] = 0$ prevents us from doing the manipulations in (4.25). In particular, while we still have $E[V_i|X_i, \tilde{A}_i] = 0$ (as in (4.26)), the new fixed effect $\tilde{A}_i$ now satisfies

$$E[\tilde{A}_i|X_i] = E[A_i + X_i^{(2)\prime}(\beta_0^{(2)} - \tilde{\beta}_0^{(2)})|X_i] = X_i^{(2)\prime}(\beta_0^{(2)} - \tilde{\beta}_0^{(2)}).$$

Thus, $\tilde{A}_i$ violates the condition $E[\tilde{A}_i|X_i] = 0$ imposed by the random effects model. ■

4.3.1 FE as Demeaning

The basic idea of the fixed effects estimator is to employ the variation across time for each individual in order to identify the parameter β_0. To this end, we denote the sample averages of the variables corresponding to individual i by

$$\bar{Y}_i \equiv \frac{1}{T}\sum_{t=1}^{T} Y_{it} \qquad \bar{X}_i \equiv \frac{1}{T}\sum_{t=1}^{T} X_{it} \qquad \bar{V}_i \equiv \frac{1}{T}\sum_{t=1}^{T} V_{it},$$

where note $\bar{Y}_i \in \mathbf{R}$, $\bar{X}_i \in \mathbf{R}^d$, and $\bar{V}_i \in \mathbf{R}$. It is also helpful to define

$$\dot{Y}_{it} \equiv Y_{it} - \bar{Y}_i \qquad \dot{X}_{it} \equiv X_{it} - \bar{X}_i \qquad \dot{V}_{it} \equiv V_{it} - \bar{V}_i,$$

which are simply the variables for individual i demeaned by the individual specific means – this transformation is sometimes referred to as the *within transformation* in the literature (i.e. the transformation is *within* individual i).

The linear structure of the model now lets us obtain for each individual i that

$$\begin{aligned}\dot{Y}_{it} &= Y_{it} - \bar{Y}_i \\ &= (X_{it}'\beta_0 + A_i + V_{it}) - (\bar{X}_i'\beta_0 + A_i + \bar{V}_i) \\ &= \dot{X}_{it}'\beta_0 + \dot{V}_{it}.\end{aligned} \tag{4.27}$$

Thus, by demeaning at the individual level we are able to get rid of the fixed effect A_i.

Moreover, Assumption FE-1(ii) and the law of iterated expectations imply

$$E[\dot{V}_{it}\dot{X}_{it}] = E[(V_{it} - \bar{V}_i)(X_{it} - \bar{X}_i)] = E[E[(V_{it} - \bar{V}_i)|X_i](X_{it} - \bar{X}_i)] = 0, \tag{4.28}$$

which together with (4.27) implies we may estimate β_0 by simply regressing $\dot{Y}_{it}$ on $\dot{X}_{it}$. The resulting estimator is precisely the fixed effects estimator, which we define by

$$\hat{\beta}_n^{\text{fe}} \equiv \arg\min_{b \in \mathbf{R}^d} \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} (\dot{Y}_{it} - \dot{X}_{it}'b)^2.$$

Since the fixed effects estimator is simply the OLS estimator applied to an appropriate transformation of the data, the assumptions required for its asymptotic analysis are the standard OLS assumptions adapted to this specific application. We thus impose:

Assumption FE-2. *(i) The matrix $\sum_{t=1}^{T} E[\dot{X}_{it}\dot{X}_{it}']$ is invertible; (ii) The variance matrix $E[(\sum_{t=1}^{T} \dot{X}_{it}\dot{V}_{it})(\sum_{t=1}^{T} \dot{X}_{it}\dot{V}_{it})']$ is finite.*

Assumption FE-2 is the standard full rank assumption imposed in OLS estimation. In the context of the fixed effects estimator, however, the full rank assumption in FE-2(i) rules out that there be a covariate in X_{it} that is time invariant (you should make sure you understand why). This restriction does simply reflect our discussion in the previous section pointing out that coefficients for time invariant regressors are not identified in the fixed effects model (see (4.25)). Finally, Assumption FE-2(ii) is just a second moment condition that will be used to invoke the central limit theorem.

Our next result derives the asymptotic distribution of the fixed effects estimator.

Theorem 4.3.1. *If Assumptions FE-1 and FE-2 hold, then it follows that:*

$$\sqrt{n}\{\hat{\beta}_n^{\text{fe}} - \beta_0\} \xrightarrow{d} N(0, (\sum_{t=1}^{T} E[\dot{X}_{it}\dot{X}_{it}'])^{-1} E[(\sum_{t=1}^{T} \dot{X}_{it}\dot{V}_{it})(\sum_{t=1}^{T} \dot{X}_{it}\dot{V}_{it})'](\sum_{t=1}^{T} E[\dot{X}_{it}\dot{X}_{it}'])^{-1})$$

PROOF: The proof is essentially re-doing the proof for consistency of OLS. We begin by noting that by definition $\hat{\beta}_n^{\text{fe}}$ must satisfy the first order condition

$$\frac{1}{n} \sum_{i=1}^{n} \sum_{t=1}^{T} (\dot{Y}_{it} - \dot{X}_{it}'\hat{\beta}_n^{\text{fe}})\dot{X}_{it} = 0. \tag{4.29}$$

Then observe that given Assumption FE-1(i) we can treat $\{\{\dot{Y}_{it}, \dot{X}_{it}\}_{t=1}^{T}\}_{i=1}^{n}$ as i.i.d. observations (over i). Hence, by a standard law of large numbers we obtain

$$\frac{1}{n} \sum_{i=1}^{n} \{\sum_{t=1}^{T} \dot{X}_{it}\dot{X}_{it}'\} \xrightarrow{p} E[\sum_{t=1}^{T} \dot{X}_{it}\dot{X}_{it}']. \tag{4.30}$$

Since $\sum_{t=1}^{T} E[\dot{X}_{it}\dot{X}_{it}']$ is invertible by Assumption FE-2(i), $\sum_{i=1}^{n}\{\sum_{t=1}^{T} \dot{X}_{it}\dot{X}_{it}'\}/n$ is invertible with probability approaching one. The first order condition in (4.29) together with the characterization of $\dot{Y}_{it}$ in (4.27) therefore yield that

$$\begin{aligned}\hat{\beta}_n^{\text{fe}} &= \{\frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{T}\dot{X}_{it}\dot{X}_{it}'\}^{-1}\frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{T}\dot{X}_{it}\dot{Y}_{it} \\ &= \beta_0 + \{\frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{T}\dot{X}_{it}\dot{X}_{it}'\}^{-1}\frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{T}\dot{X}_{it}\dot{V}_{it}. \end{aligned} \tag{4.31}$$

Moreover, $E[\sum_{t=1}^{T}\dot{V}_{it}\dot{X}_{it}] = 0$ (see (4.28)) and thus Assumption FE-2(ii) implies

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{\sum_{t=1}^{T}\dot{X}_{it}\dot{V}_{it}\} \xrightarrow{d} N(0, E[(\sum_{t=1}^{T}\dot{X}_{it}\dot{V}_{it})(\sum_{t=1}^{T}\dot{X}_{it}\dot{V}_{it})']) \tag{4.32}$$

by a standard central limit theorem (applied as $n \to \infty$ and where an "observation" is $\sum_{t=1}^{T}\dot{X}_{it}\dot{V}_{it}$). Hence, re-arranging terms in (4.31) and combining results (4.30) and (4.32) together with the continuous mapping theorem implies

$$\begin{aligned}\sqrt{n}\{\hat{\beta}_n^{\text{fe}} - \beta_0\} &= \{\frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{T}\dot{X}_{it}\dot{X}_{it}'\}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{t=1}^{T}\dot{X}_{it}\dot{V}_{it} \\ &\xrightarrow{d} N(0, (\sum_{t=1}^{T}E[\dot{X}_{it}\dot{X}_{it}'])^{-1}E[(\sum_{t=1}^{T}\dot{X}_{it}\dot{V}_{it})(\sum_{t=1}^{T}\dot{X}_{it}\dot{V}_{it})'](\sum_{t=1}^{T}E[\dot{X}_{it}\dot{X}_{it}'])^{-1}), \end{aligned} \tag{4.33}$$

which establishes the claim of the theorem. ■

Notice that the asymptotic variance derived in Theorem 4.3.1 is clustered at the individual level, meaning that the dependence of observations corresponding to the same individual is left unrestricted. Estimation of the asymptotic variance is straightforward and as usual we can proceed by employing sample analogues. In particular, denoting the fitted residuals by $\widehat{\dot{V}}_{it} \equiv \dot{Y}_{it} - \dot{X}_{it}'\hat{\beta}_n^{\text{fe}}$, we may estimate the asymptotic variance by

$$\{\frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{T}\dot{X}_{it}\dot{X}_{it}'\}^{-1}\{\frac{1}{n}\sum_{i=1}^{n}(\sum_{t=1}^{T}\dot{X}_{it}\widehat{\dot{V}}_{it})(\sum_{t=1}^{T}\dot{X}_{it}\widehat{\dot{V}}_{it})'\}\{\frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{T}\dot{X}_{it}\dot{X}_{it}'\}^{-1}.$$

Remark 4.3.2. Suppose instead we wish to impose a homoskedasticity assumption by requiring $E[V_iV_i'|X_i] = \sigma_V^2 I_T$ for I_T the $T \times T$ identity matrix. Since $\sum_{t=1}^{T}\dot{X}_{it}\dot{V}_{it} = \sum_{t=1}^{T}\dot{X}_{it}V_{it}$, we then obtain the middle term in the asymptotic variance equals

$$E[(\sum_{t=1}^{T}\dot{X}_{it}\dot{V}_{it})(\sum_{t=1}^{T}\dot{X}_{it}\dot{V}_{it})'] = E[(\sum_{t=1}^{T}\dot{X}_{it}V_{it})(\sum_{t=1}^{T}\dot{X}_{it}V_{it})'] = \sigma_V^2 E[\sum_{t=1}^{T}\dot{X}_{it}\dot{X}_{it}'].$$

Hence, under homoskedasticity, the asymptotic variance for the fixed effects estimator

obtained in Theorem 4.3.1 becomes $\sigma_V^2(\sum_{t=1}^T E[\dot{X}_{it}\dot{X}_{it}'])^{-1}$. Notice, however, that σ_V^2 is the variance of V_{it} *not* of $\dot{V}_{it}$. In fact, we have the relation

$$E[(V_{it} - \frac{1}{T}\sum_{t=1}^T V_{it})(V_{it} - \frac{1}{T}\sum_{t=1}^T V_{it})|X_i] \\ = E[V_{it}^2 - \frac{2}{T}V_{it}^2 + \frac{1}{T^2}\sum_{t=1}^T V_{it}^2|X_i] = \frac{T-1}{T}\sigma_V^2. \quad (4.34)$$

Hence, we have $E[(\dot{V}_{it})^2] = \sigma_V^2(T-1)/T$, and we can estimate σ_V^2 by employing

$$\hat{\sigma}_V^2 \equiv \frac{T}{T-1} \times \frac{1}{nT}\sum_{i=1}^n\sum_{t=1}^T(\hat{\dot{V}}_{it})^2.$$

Overall, the main point to keep in mind is that it is important to keep in mind how assumptions imposed on the original errors V_{it} affect the properties of $\dot{V}_{it}$. ■

4.3.2 FE as Dummy Variables

An alternative interpretation of the fixed effects estimator can be obtained by employing dummy variables. In particular, let us reconsider the linear model

$$Y_{it} = X_{it}'\beta_0 + A_i + V_{it} \quad (4.35)$$

and think of each A_i as an unknown parameter instead of an unknown person specific shock. In order to run the regression in (4.35) treating A_i as an unknown parameter we need to introduce dummy variables for each individual, which we denote by

$$D_i^{(j)} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases},$$

i.e. $D_i^{(j)}$ is the dummy variable for individual j and hence it is "on" only when $i = j$. In order to collect these dummy variables into a vector of regressors we define

$$D_{it} = \begin{pmatrix} D_i^{(1)} \\ \vdots \\ D_i^{(n)} \end{pmatrix}, \quad (4.36)$$

which note is of dimension $n \times 1$ (in contrast recall that $X_{it} \in \mathbf{R}^d$). Letting $\alpha_0 \equiv (A_1, \ldots, A_n)' \in \mathbf{R}^n$, we may rewrite equation (4.35) as equal to

$$Y_{it} = X_{it}'\beta_0 + D_{it}'\alpha_0 + V_{it}. \quad (4.37)$$

The corresponding ordinary least squares estimators for β_0 and α_0 are then given by

$$(\hat{\beta}_n^{\mathrm{d}}, \hat{\alpha}_n^{\mathrm{d}}) \equiv \arg \min_{b \in \mathbf{R}^d, a \in \mathbf{R}^n} \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} (Y_{it} - X_{it}' b - D_{it}' a)^2. \tag{4.38}$$

While we have rewritten the fixed effects model as a linear regression, it is important to note that it is not standard in that the dimension of the parameter (β_0, α_0) is $(d+n)$ and therefore grows with the sample size nT. However, the parameter of interest β_0 remains of dimension d only and it is only the *incidental* parameter α_0 that grows in dimension with the sample size. Neyman and Scott (1948) examined this challenge, called *the incidental parameter problem*, and showed that conducting inference on β_0 through linear regression remains straightforward even though α_0 is present. To examine their insight, we follow the notation in (4.2) and (4.3) and set

$$D_i \equiv \begin{pmatrix} D_{i1}' \\ \vdots \\ D_{iT}' \end{pmatrix} \qquad \mathbb{D}_n \equiv \begin{pmatrix} D_1 \\ \vdots \\ D_n \end{pmatrix}$$

where the dimensions of D_i and $\mathbb{D}_n$ are $T \times n$ and $(Tn) \times n$. Further recall the definitions

$$\mathbb{X}_n \equiv \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} \qquad \mathbb{Y}_n \equiv \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \qquad \mathbb{V}_n \equiv \begin{pmatrix} V_1 \\ \vdots \\ V_n \end{pmatrix}$$

where $\mathbb{X}_n$, $\mathbb{Y}_n$, and $\mathbb{V}_n$ are of dimension $(Tn) \times d$, $(Tn) \times 1$, and $(Tn) \times 1$. We can therefore re-write expression (4.37) in matrix notation as being equal to

$$\mathbb{Y}_n = \mathbb{X}_n \beta_0 + \mathbb{D}_n \alpha_0 + \mathbb{V}_n.$$

Employing the introduced notation, it is now straightforward to show that the fixed effects estimator $\hat{\beta}_n^{\mathrm{fe}}$ is in fact numerically equivalent to the estimator $\hat{\beta}_n^{\mathrm{d}}$.

Lemma 4.3.1. *If the matrix* $\{\sum_{i=1}^{n} \sum_{t=1}^{T} \dot{X}_{it} \dot{X}_{it}'\}$ *is invertible, then* $\hat{\beta}_n^{\mathrm{fe}} = \hat{\beta}_n^{\mathrm{d}}$.

PROOF: The proof simply applies the partitioned regression formula from Theorem 2.2.1. To this end, first note that by definition of $\mathbb{D}_n$ and direct calculation

$$\mathbb{D}_n' \mathbb{D}_n = T I_n.$$

Also let J_T denote a $T \times T$ matrix whose all entries are equal to one, which we write as

$$J_T \equiv \begin{pmatrix} 1 & \dots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \dots & 1 \end{pmatrix}.$$

Letting $B \otimes C$ denote the kronecker product between matrices B and C we then obtain

$$\mathbb{D}_n(\mathbb{D}_n'\mathbb{D}_n)^{-1}\mathbb{D}_n' = \mathbb{D}_n(\frac{1}{T}I_n)\mathbb{D}_n' = \frac{1}{T}(J_T \otimes I_n).$$

Note that pre-multiplying any $(nT) \times 1$ vector by the matrix $\mathbb{D}_n(\mathbb{D}_n'\mathbb{D}_n)^{-1}\mathbb{D}_n'$ simply returns an $nT \times 1$ vector consisting of individual specific averages (you should check this calculation). Hence, using the partitioned regression formula from Theorem 2.2.1 and noting that $\sum_{t=1}^T \dot{X}_{it} = \sum_{t=1}^T \dot{Y}_{it} = 0$ for any $1 \leq i \leq n$ we obtain that

$$\begin{aligned}\hat{\beta}_n^{\mathrm{d}} &= (\mathbb{X}_n'(I_{nT} - \frac{1}{T}(J_T \otimes I_n))\mathbb{X}_n)^{-1}\mathbb{X}_n'(I_{nT} - \frac{1}{T}(J_T \otimes I_n))\mathbb{Y}_n \\ &= (\sum_{i=1}^n \sum_{t=1}^T X_{it}\dot{X}_{it}')^{-1} \sum_{i=1}^n \sum_{t=1}^T X_{it}\dot{Y}_{it} = (\sum_{i=1}^n \sum_{t=1}^T \dot{X}_{it}\dot{X}_{it}')^{-1} \sum_{i=1}^n \sum_{t=1}^T \dot{X}_{it}\dot{Y}_{it} = \hat{\beta}_n^{\mathrm{fe}},\end{aligned}$$

which establishes the claim of the Lemma. ■

The intuition behind Lemma 4.3.1 is quite straightforward. Recall that the partitioned regression result of Theorem 2.2.1 allowed us to characterize $\hat{\beta}_n^{\mathrm{d}}$ as the coefficient obtained from regressing the residuals from regressing $\mathbb{Y}_n$ on $\mathbb{D}_n$ on the residuals from regressing $\mathbb{X}_n$ on $\mathbb{D}_n$. However, $\mathbb{D}_n$ is simply a matrix of dummy variables for each individual, and hence the fitted values of regressing a vector on $\mathbb{D}_n$ is simply the individual specific averages for that vector. As a result, the residuals of regressing a vector on $\mathbb{D}_n$ are just the values of that vector demeaned at the individual level, which connects us back to the original fixed effects estimator $\hat{\beta}_n^{\mathrm{fe}}$.

The representation of fixed effects as dummy variables easily generalizes to multiple categories. Perhaps the simplest generalization is one in which we write

$$Y_{it} = X_{it}'\beta_0 + A_i + B_t + V_{it},$$

and estimate β_0 by running a regression similar to (4.38) but including both dummy variables for individual (referred to as "person fixed effects") and dummy variables for time period (referred to as "time fixed effects"). More generally, whenever a discrete variable "`xxx`" is present (e.g. village, industry), practitioners refer to a regression that includes a dummy variable for each category of "`xxx`" as including "`xxx`" fixed effects.

Remark 4.3.3. It is worth emphasizing that the analysis in this section relies crucially on the linearity of the model in (4.35). The study of nonlinear panel data models is

substantially more challenging, with the presence of fixed effects often requiring both n and T diverging to infinity. For a recent review of this literature and the challenges involved in nonlinear panel data models, see Arellano and Bonhomme (2011). ■

Remark 4.3.4. Since $\hat{\beta}_n^{\mathrm{d}} = \hat{\beta}_n^{\mathrm{fe}}$, it follows from Theorem 4.3.1 that $\hat{\beta}_n^{\mathrm{d}}$ is asymptotically normally distributed. It is worth emphasizing that the fact that $\hat{\beta}_n^{\mathrm{d}}$ is asymptotically normally distributed even though the number of regressors is proportional to the sample size is in large part due to the special structure of the fixed effects regressor matrix $\mathbb{D}_n$. More generally, a large literature has strived to obtain the asymptotic distribution of OLS coefficients when the number of regressors is proportional to the sample size. Such results are challenging and require suitable generalizations of the structure present in the fixed effects regressors; see, e.g., Cattaneo et al. (2018) and references therein. ■

4.4 Dynamic Panel Models

We refer to a panel data model as dynamic, if the present value of the dependent variable is a function of past values of the dependent variable; e.g. we have

$$Y_{it} = \delta_0 Y_{i(t-1)} + \tilde{X}_{it}'\beta_0 + A_i + V_{it}, \tag{4.39}$$

where again we let $1 \leq t \leq T$ but additionally assume the observation Y_{i0} is available. In this setting, the regressors become $X_{it} = (\tilde{X}_{it}, Y_{i(t-1)})$ and we immediately see that the assumptions we imposed for the fixed effects estimator are violated. Specifically, notice that Assumption FE-1(ii) fails since, for example, we have from (4.39) that

$$E[V_{i2}|X_i, A_i] = E[V_{i2}|Y_{i0}, \ldots Y_{i(T-1)}, \tilde{X}_{i1}, \ldots, \tilde{X}_{iT}, A_i] = V_{i2}. \tag{4.40}$$

Estimation of dynamic panel data models therefore requires us to weaken assumption FE-1(ii) in order to allow the regressors X_{it} to possibly contain lagged values of the dependent variable. In what follows we continue to assume the linear structure

$$Y_{it} = X_{it}'\beta_0 + A_i + V_{it}, \tag{4.41}$$

but now replace Assumption FE-1 with a so-called *sequential exogeneity* assumption.

Assumption DYN-1. *(i)* $\{Y_i, X_i\}_{i=1}^n$ *is an i.i.d. sample that satisfies* (4.41)*; (ii) The moment restriction* $E[V_{it}|X_{it}, \ldots, X_{i1}, A_i] = 0$ *holds for all* $1 \leq t \leq T$*.*

It is straightforward to see that Assumption FE-1 implies Assumption DYN-1. Furthermore, we also note that Assumption DYN-1 potentially allows for the inclusion of lagged dependent variables. For instance, reconsidering model (4.39) and re-examining

the calculation in (4.40) note that Assumption DYN-1 now requires that

$$0 = E[V_{i2}|Y_{i0}, Y_{i1}, \tilde{X}_{i1}, \tilde{X}_{i2}, A_i] = E[V_{i2}|Y_{i0}, \tilde{X}_{i1}, \tilde{X}_{i2}, A_i, V_{i1}],$$

where the second equality holds due to V_{i1} being a function of Y_{i1}, Y_{i0}, and $\tilde{X}_{i1}$. In particular, in the context of model (4.39), Assumption DYN-1(ii) implies

$$E[V_{it}|V_{i(t-1)}, \ldots, V_{i1}] = 0;$$

which allows for a fairly flexible specification of the error process, but may be violated, for example, if V_{it} follows an autoregressive process.

While Assumption DYN-1 allows us to accommodate the inclusion of lagged dependent variables as regressors, it does not guarantee the consistency of the fixed effects estimator. As the next Lemma shows, the fixed effects estimator can be inconsistent under Assumption DYN-1 and a different estimator is required.

Lemma 4.4.1. *If Assumption FE-2 and DYN-1 holds, then it follows that*

$$\hat{\beta}_n^{\text{fe}} \xrightarrow{p} \beta_0 - \{\frac{1}{T}\sum_{t=1}^{T} E[\dot{X}_{it}\dot{X}_{it}']\}^{-1} E[\bar{X}_i\bar{V}_i].$$

Proof: The manipulations are similar to those in the proof of Theorem 4.3.1 and we therefore skip some of the details. First note that by a standard law of large numbers

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{T} \dot{X}_{it}\dot{X}_{it}' = \sum_{t=1}^{T} E[\dot{X}_{it}\dot{X}_{it}'] + o_p(1).$$

Since the matric $\sum_{t=1}^{T} E[\dot{X}_{it}\dot{X}_{it}']$ is invertible, we therefore obtain that its sample analogue is invertible with probability tending to one. We thus obtain

$$\begin{aligned}\hat{\beta}_n^{\text{fe}} &= \beta_0 + \{\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T} \dot{X}_{it}\dot{X}_{it}'\}^{-1}\frac{1}{n}\sum_{i=1}^{n}\frac{1}{T}\sum_{t=1}^{T}\dot{X}_{it}\dot{V}_{it} + o_p(1) \\ &= \beta_0 + \{\frac{1}{T}\sum_{t=1}^{T} E[\dot{X}_{it}\dot{X}_{it}']\}^{-1}\frac{1}{T}\sum_{t=1}^{T} E[\dot{X}_{it}\dot{V}_{it}] + o_p(1), \qquad (4.42)\end{aligned}$$

where the final result follows from the law of large numbers and the continuous mapping theorem. However, by direct algebra and Assumption DYN-1 we now obtain

$$\begin{aligned}E[\frac{1}{T}\sum_{t=1}^{T}\dot{X}_{it}\dot{V}_{it}] &= E[\frac{1}{T}\sum_{t=1}^{T}(X_{it} - \bar{X}_i)(V_{it} - \bar{V}_i)] \\ &= E[\frac{1}{T}\sum_{t=1}^{T} X_{it}(V_{it} - \bar{V}_i)] = -E[\frac{1}{T}\sum_{t=1}^{T} X_{it}\bar{V}_i] = -E[\bar{X}_i\bar{V}_i],\end{aligned}$$

which establishes the claim of the lemma. ■

In particular, we note that the term $E[\bar{X}_i\bar{V}_i]$ is not necessarily equal to zero under Assumption DYN-1(ii) as it contains the product of X_{it} with $V_{i(t+s)}$ for $s \geq 1$; i.e.

$$E[\bar{X}_i\bar{V}_i] = \frac{1}{T^2}E[\sum_{t=1}^{T} X_{it} \sum_{s=1}^{T} V_{is}] = \frac{1}{T^2}\sum_{t=2}^{T}\sum_{s=1}^{t-1} E[X_{it}V_{is}],$$

where the second equality follows by Assumption DYN-1(ii). It is interesting to note that in usual models, $\bar{V}_i \xrightarrow{p} 0$ as $T \to \infty$. As a result, even though $\hat{\beta}_n^{\text{fe}}$ is inconsistent, it follows its asymptotic "bias" decreases with the "length" (i.e. T) of the panel.

4.4.1 The GMM View

One way to interpret the differences between the random effects estimator $\hat{\beta}_n^{\text{re}}$ and the fixed effects estimator $\hat{\beta}_n^{\text{fe}}$ estimator is that they rely on different moment conditions for identification. For instance, under Assumption RE-1(ii) we know β_0 satisfies

$$E[(Y_{it} - X_{it}'\beta_0)X_{is}] = 0 \text{ for all } 1 \leq t, s \leq T, \tag{4.43}$$

which motivated defining the random effects estimator $\hat{\beta}_n^{\text{re}}$ by zeroing linear combinations of the sample analogues to the moment conditions in (4.43). In contrast, the fixed effects estimator relied on Assumption FE-1(ii) to derive the moment condition

$$E[(\dot{Y}_{it} - \dot{X}_{it}'\beta_0)\dot{X}_{it}] = 0 \text{ for all } 1 \leq t \leq T, \tag{4.44}$$

which motivated defining the fixed effects estimator $\hat{\beta}_n^{\text{fe}}$ as the solution to the sample analogue to the moment condition in (4.44).

More generally, it is often fruitful to think of the moment restrictions that a panel data model implies and simply base estimation based on said moment restrictions. In what follows, we apply this approach to estimate dynamic panel data models under the sequential exogeneity assumption (Assumption DYN-1), though we note that the principle is more widely applicable to other assumptions; see Arellano (2003).

The first step towards deriving moment restrictions implied by Assumption DYN-1 is to get rid of the fixed effects A_i. To this end, we note that as reflected by the inconsistency of the fixed effects estimator, simply demeaning at the individual level is not a good way to proceed. Instead, we rely on first differences, and to this let

$$\Delta X_{it} \equiv X_{it} - X_{i(t-1)} \qquad \Delta Y_{it} \equiv Y_{it} - Y_{i(t-1)} \qquad \Delta V_{it} \equiv V_{it} - V_{i(t-1)}. \tag{4.45}$$

Taking first differences to the equation in (4.41) we then get rid of the fixed effect since

$$\Delta Y_{it} = \Delta X_{it}'\beta_0 + \Delta V_{it} \text{ for } 2 \leq t \leq T. \tag{4.46}$$

Notice, however, that conducting linear regression on (4.46) does not deliver a consistent estimator as the regressor ΔX_{it} is not necessarily exogenous under Assumption DYN-1. Indeed, employing Assumption DYN-1 and a simple calculation implies that

$$E[\Delta X_{it}\Delta V_{it}] = E[(X_{it} - X_{i(t-1)})(V_{it} - V_{i(t-1)})] = -E[X_{it}V_{i(t-1)}].$$

Hence, estimation of β_0 in (4.46) requires us to find a suitable instrument for ΔX_{it}. A number of different approaches have been proposed that employ the structure of the panel data to obtain such instruments. Two prominent examples are:

Example 4.4.1. Arellano and Bond (1991) noted that the sequential exogeneity requirement of Assumption DYN-1(ii) implies that X_{is} is a potential instrument for ΔX_{it} in equation (4.46) for any $s \leq t-1$. In particular note that

$$E[X_{is}\Delta V_{it}] = E[X_{is}(V_{it} - V_{i(t-1)})] = 0 \text{ if } s \leq t-1,$$

which implies X_{is} satisfies the exclusion restriction. The rank condition for X_{is} then requires X_{is} to be suitably correlated with ΔX_{it} – when $s = t-1$ this is plausible as $\Delta X_{it} = X_{it} - X_{i(t-1)}$. Note that by the same arguments ΔX_{is} is also a potential instrument for ΔX_{it} for any $s \leq t-1$. ■

Example 4.4.2. Blundell and Bond (1998) observed that if the errors $\{V_{it}\}_{t=1}^T$ are serially uncorrelated, then additional instruments are available. Concretely note that if $E[V_{it}V_{is}] = 0$ for all $s \leq t-1$, then for any $s \leq t-2$ Assumption DYN-1(ii) yields

$$E[Y_{is}\Delta V_{it}] = E[(X_{is}'\beta_0 + A_i + V_{is})\Delta V_{it}] = E[V_{is}(V_{it} - V_{i(t-1)})] = 0.$$

By the same arguments, it is also possible to show that $E[\Delta Y_{is}\Delta V_{it}] = 0$ whenever $s \leq t-2$. As in Example 4.4.1, however, keep in mind the rank condition imposes additional requirements on these potential instruments. ■

The main takeaway from Examples 4.4.1 and 4.4.2 is that the panel structure (and suitable assumptions) allow us to obtain instruments Z_{it} for ΔX_{it}. Formally, we have

$$E[(\Delta Y_{it} - \Delta X_{it}'\beta_0)Z_{it}] = 0 \tag{4.47}$$

for appropriate instruments Z_{it}. For instance, following the Arellano and Bond (1991)

approach discussed in Example 4.4.1, we may set Z_{it} to equal the vector

$$Z_{it} = \begin{pmatrix} X_{i1} \\ \vdots \\ X_{i(t-1)} \end{pmatrix}$$

for any $t \geq 2$. Notice that in this instance, the dimension of Z_{it} may actually change with t (unlike, say, X_{it}). It is then notationally convenient to define

$$Z_i \equiv \begin{pmatrix} Z_{i2}' & 0 & \dots & \dots & 0 \\ 0 & Z_{i3}' & 0 & \dots & 0 \\ 0 & 0 & Z_{i4}' & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & Z_{iT}' \end{pmatrix},$$

which note is a matrix with $T-1$ rows and as many columns as there are total number of instruments. We additionally let ΔX_i, ΔY_i, and ΔV_i be given by

$$\Delta X_i \equiv \begin{pmatrix} \Delta X_{i2}' \\ \vdots \\ \Delta X_{iT}' \end{pmatrix} \qquad \Delta Y_i \equiv \begin{pmatrix} \Delta Y_{i2} \\ \vdots \\ \Delta Y_{iT} \end{pmatrix} \qquad \Delta V_i \equiv \begin{pmatrix} \Delta V_{i2} \\ \vdots \\ \Delta V_{iT} \end{pmatrix},$$

and note that ΔX_i is a $(T-1) \times d$ matrix, while ΔY_i and ΔV_i are $(T-1) \times 1$ vectors. Given these definitions, we may write the moment restrictions in (4.48) as equal to

$$E[Z_i'(\Delta Y_i - \Delta X_i \beta_0)] = E[Z_i' \Delta V_i] = 0. \tag{4.48}$$

Given the expression for the moment conditions obtained in (4.48), we can now see that the problem reduces to our analysis in Chapter 3. Specifically, letting d_z denote the number of columns in Z_i, and $\hat{\Omega}_n$ be a $d_z \times d_z$ positive definite matrix, we may estimate β_0 by minimizing a quadratic form in the sample analogues to (4.48) – i.e. we set

$$\hat{\beta}_n \equiv \arg\min_{b \in \mathbf{R}^d} (\frac{1}{n} \sum_{i=1}^n Z_i'(\Delta Y_i - \Delta X_i b))' \hat{\Omega}_n (\frac{1}{n} \sum_{i=1}^n Z_i'(\Delta Y_i - \Delta X_i b)). \tag{4.49}$$

We establish the asymptotic normality of $\hat{\beta}_n$ under the following conditions:

Assumption DYN-2. *(i)* $E[Z_i' \Delta V_i] = 0$*; (ii)* $\hat{\Omega}_n \xrightarrow{p} \Omega$ *for some positive definite matrix* Ω *such that* $E[\Delta X_i' Z_i] \Omega E[Z_i' \Delta X_i]$ *is invertible; (iii)* $E[Z_i' \Delta V_i \Delta V_i' Z_i]$ *is finite.*

Since we have not taken a stand on the exact manner in which Z_i is constructed, Assumption DYN-2 directly imposes that it satisfy the exclusion restriction. In turn,

Assumptions DYN-2(ii) and DYN-2(iii) impose regularity conditions that are analogous to those employed in establishing Theorem 3.2.1 (but adapted to the present setting). Together with Assumption DYN-1, these conditions suffice for deriving the asymptotic distribution of the estimator $\hat{\beta}_n$ as defined in (4.49).

Theorem 4.4.1. *Let Assumptions DYN-1 and DYN-2 hold, and define* $C \equiv E[Z_i \Delta X_i']$ *and* $\Sigma \equiv E[Z_i' \Delta V_i \Delta V_i' Z_i]$. *Then, it follows that*

$$\sqrt{n}\{\hat{\beta}_n - \beta_0\} \xrightarrow{d} N(0, \{C'\Omega C\}^{-1} C'\Omega\Sigma\Omega C\{C'\Omega C\}^{-1}).$$

PROOF: The proof is essentially the same as that of Theorem 3.2.1 and we therefore skip some of the details (if in doubt at a step, revisit the proof of Theorem 3.2.1). First observe that by definition of $\hat{\beta}_n$ it must satisfy the first order condition

$$\{\frac{1}{n}\sum_{i=1}^{n} \Delta X_i' Z_i\} \hat{\Omega}_n \{\frac{1}{n}\sum_{i=1}^{n} Z_i'(\Delta Y_i - \Delta X_i \hat{\beta}_n)\} = 0. \tag{4.50}$$

Moreover, by the law of large numbers and Assumption DYN-2(ii) we also have that

$$\{\frac{1}{n}\sum_{i=1}^{n} \Delta X_i' Z_i\} \hat{\Omega}_n \{\frac{1}{n}\sum_{i=1}^{n} Z_i' \Delta X_i\} \xrightarrow{p} E[\Delta X_i' Z_i] \Omega E[Z_i' \Delta X_i]. \tag{4.51}$$

by the continuous mapping theorem. Hence, Assumption DYN-2(ii) and result (4.51) imply that the matrix $\{\sum_{i=1}^n \Delta X_i' Z_i/n\} \hat{\Omega}_n \{\sum_{i=1}^n Z_i' \Delta X_i/n\}$ is invertible with probability tending to one. From result (4.50) we therefore obtain

$$\begin{aligned}\hat{\beta}_n &= \{(\frac{1}{n}\sum_{i=1}^{n} \Delta X_i' Z_i) \hat{\Omega}_n (\frac{1}{n}\sum_{i=1}^{n} Z_i' \Delta X_i)\}^{-1} \{\frac{1}{n}\sum_{i=1}^{n} \Delta X_i' Z_i\} \hat{\Omega}_n \{\frac{1}{n}\sum_{i=1}^{n} Z_i' \Delta Y_i\} + o_p(1) \\ &= \beta_0 + \{(\frac{1}{n}\sum_{i=1}^{n} \Delta X_i' Z_i) \hat{\Omega}_n (\frac{1}{n}\sum_{i=1}^{n} Z_i' \Delta X_i)\}^{-1} \{\frac{1}{n}\sum_{i=1}^{n} \Delta X_i' Z_i\} \hat{\Omega}_n \{\frac{1}{n}\sum_{i=1}^{n} Z_i' \Delta V_i\} + o_p(1)\end{aligned} \tag{4.52}$$

where the equality holds from (4.46). Next note that from a conventional central limit theorem and Assumptions DYN-2(i) and DYN-2(iii) we can conclude that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_i' \Delta V_i \xrightarrow{d} N(0, E[Z_i' \Delta V_i \Delta V_i' Z_i]). \tag{4.53}$$

Finally, combining results (4.51) and (4.53) with the expression (4.52), basic algebra,

the definitions of C and Σ, and the continuous mapping theorem implies

$$\begin{aligned}
&\sqrt{n}\{\hat{\beta}_n - \beta_0\} \\
&\quad = \{(\frac{1}{n}\sum_{i=1}^{n} \Delta X_i' Z_i)\hat{\Omega}_n(\frac{1}{n}\sum_{i=1}^{n} Z_i'\Delta X_i)\}^{-1}\{\frac{1}{n}\sum_{i=1}^{n} \Delta X_i' Z_i\}\hat{\Omega}_n\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_i'\Delta V_i\} + o_p(1) \\
&\quad \xrightarrow{d} N(0, \{C'\Omega C\}^{-1}C'\Omega\Sigma\Omega C\{C'\Omega C\}^{-1}),
\end{aligned}$$

which establishes the claim of the theorem. ■

As in Theorems 3.2.1 and 4.2.1, the asymptotic distribution of $\hat{\beta}_n$ depends on the choice of weighting matrix $\hat{\Omega}_n$ and its probability limit Ω. Unsurprisingly, given our previous results, the choice of Ω that minimizes the asymptotic variance is

$$\Omega = \{E[Z_i \Delta V_i \Delta V_i' Z_i']\}^{-1}.$$

Finally, as a word of warning, we note that in applications the instruments described in Examples 4.4.1 and 4.4.2 may be weak, and so you should keep in mind the discussion in Chapter 3.4.1. In addition, note that in principle the panel structure can lead to a large number of possible instruments, meaning the dimension of Z_i is "large". In practice, employing "many" instruments can lead to the asymptotic approximation of Theorem 4.4.1 to be inaccurate and therefore to inference problems such as poor size control; see, Bekker (1994) and the literature that followed.

4.5 Differences in Differences

A common strategy for employing the panel structure to identify causal treatment effects is the so called Difference in Differences (a.k.a. "diff in diff" or DiD) approach. Below we begin by examining the simplest model from which most intuition is often drawn. We then discuss common extensions encountered in applied work and emphasize that problems and complications can arise *very quickly*.

4.5.1 The Basic Model

Consider a panel data setting in which we observe $1 \leq i \leq n$ for two time periods. In each time period we observe an outcome variable Y_{it} and an indicator for treatment status D_{it}. We further adopt potential outcomes notation by letting $Y_{it}(d)$ denote outcome for individual i at time t and suppose the observed outcome Y_{it} is given by

$$Y_{it} = Y_{it}(0) + D_{it}(Y_{it}(1) - Y_{it}(0)). \tag{4.54}$$

To fix ideas, we consider the following canonical example of DiD.

Example 4.5.1. In 1992 New Jersey (NJ) raised state minimum wage from \$4.25 to \$5.05 while the minimum wage in Pennsylvania (PA) remained at \$4.25. To evaluate the impact of the change in the minimum wage, Card and Krueger (1994) interviewed fast food restaurants on both sides of the NJ/PA border and collected data on employment and wages before and after the change in minimum wage. Focusing on employment they find the following averages (across fast food establishments) in NJ and PA

	PA	NJ	Diff: NJ - PA
Avg. Employment Before	23.33	20.44	−2.89
Avg. Employment After	21.17	21.03	−0.14
Change in Avg. Employment	−2.16	0.59	2.76

To map into our notation, we let Y_{it} denote the number of employees working at fast food establishment i at time $t \in \{1, 2\}$ and D_{it} be an indicator for whether the minimum wage to which the fast food establishment is subject to changed. Within the potential out comes models, a parameter measuring the impact of the minimum wage change is

$$E[Y_{it}(1) - Y_{it}(0)|D_{it} = 1],$$

which is known as the average treatment effect on the treated (ATT). ■

The preceding example exhibits the key features of a DiD design: No individuals are treated in the first time period and some (but not all) individuals are treated in the second time period. We formalize this setup with the following assumption.

Assumption DiD-1. *(i)* $(Y_{i1}, D_{i1}, Y_{i2}, D_{i2})$ *are generated according to equation* (4.54)*; (ii)* (D_{i1}, D_{i2}) *satisfy* $P(D_{i1} = 0) = 1$ *and* $P(D_{i2} = 1) \in (0, 1)$.

Importantly, notice that there is no requirement status be independent of potential outcomes. Instead, we will aim to identify the ATT by employing the panel structure. In particular, we proceed by decomposing the ATT into the following two terms

$$\underbrace{E[Y_{i2}(1) - Y_{i2}(0)|D_{i2} = 1]}_{\text{ATT}} = \underbrace{E[Y_{i2}(1) - Y_{i1}(0)|D_{i2} = 1]}_{\text{identified}} - \underbrace{E[Y_{i2}(0) - Y_{i1}(0)|D_{i2} = 1]}_{\text{time trend}},$$

i.e. the first term contains both the effect of a change in status and a time trend, while the subtracted second term removes the time trend to obtain the ATT. The key assumption of the DiD framework is to require that the time trend for the treated units be the same as the time trend for the untreated units. This requirement, which we formalize below, is often referred to as the *parallel trends* assumption.

Assumption DiD-2. *(i)* $E[Y_{i2}(0) - Y_{i1}(0)|D_{i2} = 1] = E[Y_{i2}(0) - Y_{i1}(0)|D_{i2} = 0]$ *– i.e. the treated and control groups follow parallel time trends.*

Given Assumptions DiD-1 and DiD-2 it is then immediate that we may identify the ATT as the difference of two difference (hence differences in differences):

$$\text{ATT} = E[Y_{i2}(1) - Y_{i1}(0)|D_{i2} = 1] - E[Y_{i2}(0) - Y_{i1}(0)|D_{i2} = 0]$$

In practice, the ATT is often estimated by running a linear regression of Y_{it} on an appropriate set of time and treatment dummies. The following simple lemma derives a basic representation (though note it is not unique).

Lemma 4.5.1. *Let Assumptions DiD-1 and DiD-2 hold and consider the regression*

$$Y_{it} = \alpha + \gamma 1\{D_{i2} = 1\} + \lambda 1\{t = 2\} + \delta 1\{D_{i2} = 1, t = 2\} + \varepsilon_{it}.$$

Then it follows that $\delta = \text{ATT}$.

PROOF: Note that by Assumption DiD-1(ii), the variable $1\{D_{i2} = 1\}$ is an indicator for whether individual i is in the treatment group. Since the regression is fully saturated model in the possible values for time and treatment status, Example 2.1.1 implies

$$\begin{aligned}
E[Y_{it}|t = 1, D_{i2} = 0] &= \alpha \\
E[Y_{it}|t = 1, D_{i2} = 1] &= \alpha + \gamma \\
E[Y_{it}|t = 2, D_{i2} = 0] &= \alpha + \lambda \\
E[Y_{it}|t = 2, D_{i2} = 1] &= \alpha + \gamma + \lambda + \delta.
\end{aligned}$$

Rearranging terms, and employing the potential outcomes model we thus arrive at

$$\begin{aligned}
\delta &= E[Y_{i2}(1)|D_{i2} = 1] - (\alpha + \gamma) - \lambda \\
&= E[Y_{i2}(1) - Y_{i1}(0)|D_{i2} = 1] - E[Y_{i2}(0) - Y_{i1}(0)|D_{i2} = 0] = \text{ATT},
\end{aligned}$$

where the final equality holds by Assumption DiD-2. ■

The main convenience of writing the ATT as a coefficient in a linear regression lies in that we may readily apply our results from the preceding sections to conduct inference. Of particular concern in this literature is the possibility of clustered data (in the sense of Section 4.1.2). Due to Bertrand et al. (2004), it has become standard to at the very least cluster to account for possible serial correlation. In applications with multiple groups (e.g., states) it is also somewhat common to cluster at the group level – though recall that by our discussion in Section 4.1.2, such analysis often relies on approximations that require having a "large" number of groups.

4.5.1.1 Including Covariates

An alternative to the common trends assumptions holding at the group level is to require that it hold after we condition on a set of covariates. Specifically, suppose that in addition to observing $(Y_{i1}, D_{i1}, Y_{i2}, D_{i2})$ we observe a variable X_i that does not vary over time. We may then employ what is known as a *conditional* common trends assumptions.

Assumption DiD-3. $E[Y_{i2}(0) - Y_{i1}(0)|X_i, D_{i2} = 1] = E[Y_{i2}(0) - Y_{i1}(0)|X_i, D_{i2} = 0]$ – *i.e. the treated and control groups follow parallel time trends after conditioning on* X_i.

While Assumption DiD-3 may be deemed as more likely to hold than Assumption DiD-2, it is important to emphasize that mathematically neither assumption implies each other – i.e., formally neither assumptions is stronger (or weaker) than the other. To appreciate why, we need only consider the case of a binary covariate $X_i \in \{0, 1\}$ such as gender. In particular note that by the law of iterated expectations

$$\begin{aligned}
E[Y_{i2}(0) - Y_{i1}(0)|D_{i2} = 1] =& E[Y_{i2}(0) - Y_{i1}(0)|X_i = 1, D_{i2} = 1]P(X_i = 1|D_{i2} = 1) \\
&+ E[Y_{i2}(0) - Y_{i1}(0)|X_i = 0, D_{i2} = 1]P(X_i = 0|D_{i2} = 1) \\
E[Y_{i2}(0) - Y_{i1}(0)|D_{i2} = 0] =& E[Y_{i2}(0) - Y_{i1}(0)|X_i = 1, D_{i2} = 0]P(X_i = 1|D_{i2} = 0) \\
&+ E[Y_{i2}(0) - Y_{i1}(0)|X_i = 0, D_{i2} = 1]P(X_i = 0|D_{i2} = 0),
\end{aligned}$$

and therefore taking differences we obtain that under Assumption DiD-3 we have

$$\begin{aligned}
&E[Y_{i2}(0) - Y_{i1}(0)|D_{i2} = 1] - E[Y_{i2}(0) - Y_{i1}(0)|D_{i2} = 0] \\
&\quad = E[Y_{i2}(0) - Y_{i1}(0)|X_i = 1, D_{i2} = 1]\{P(X_i = 1|D_{i2} = 1) - P(X_i = 1|D_{i2} = 0)\} \\
&\quad + E[Y_{i2}(0) - Y_{i1}(0)|X_i = 0, D_{i2} = 1]\{P(X_i = 0|D_{i2} = 1) - P(X_i = 0|D_{i2} = 0)\}.
\end{aligned}$$

Crucially, note that unless the distribution of the covariates X_i is the same among the treated and the untreated groups, then it follows that the conditional common trends assumption (Assumption DiD-3) *does not* imply the common trends assumption (Assumption DiD-2). The fact that Assumption DiD-2 *does not* imply Assumption DiD-3 either is perhaps more intuitive and is left as an exercise.

By replicating the arguments in Section 4.5.1, it is straightforward to see that Assumption DiD-3 allows us to identify the ATT for subgroups defined by X_i – i.e. identify

$$E[Y_{i2}(1) - Y_{i1}(0)|X_i, D_{i2} = 1],$$

which may be useful for assessing the heterogenous impact of the treatment. The next Lemma, originally due to Abadie (2005), establishes that the conditional trends assumption also suffices for identifying the ATT.

Lemma 4.5.2. *If Assumptions DiD-1, DiD-3 hold, and* $0 < P(D_{i2} = 1|X_i) < 1$*, then*

$$\text{ATT} = E[Y_{i2} - Y_{i1}|D_{i2} = 1] - E[(Y_{i2} - Y_{i1})\frac{P(D_{i2} = 1|X_i)}{P(D_{i2} = 1)}\frac{P(D_{i2} = 0)}{P(D_{i2} = 0|X_i)}|D_{i2} = 0]$$

Proof: First note that Assumption DiD-1 and the law of iterated expectations imply

$$\begin{aligned} &E[(Y_{i2} - Y_{i1})\frac{P(D_{i2} = 1|X_i)}{P(D_{i2} = 1)}\frac{P(D_{i2} = 0)}{P(D_{i2} = 0|X_i)}|D_{i2} = 0] \\ &= E[E[(Y_{i2}(0) - Y_{i1}(0))|X_i, D_{i2} = 0]\frac{P(D_{i2} = 1|X_i)}{P(D_{i2} = 1)}\frac{P(D_{i2} = 0)}{P(D_{i2} = 0|X_i)}|D_{i2} = 0]. \quad (4.55) \end{aligned}$$

Next let F_X denote the distribution of X, $F_{X|D_{i2}=0}$ denote the conditional distribution of X given that $D_{i2} = 0$, and note the right hand side of (4.55) satisfies

$$\begin{aligned} &E[E[(Y_{i2}(0) - Y_{i1}(0))|X_i, D_{i2} = 0]\frac{P(D_{i2} = 1|X_i)}{P(D_{i2} = 1)}\frac{P(D_{i2} = 0)}{P(D_{i2} = 0|X_i)}|D_{i2} = 0] \\ &= \int E[(Y_{i2}(0) - Y_{i1}(0))|X_i, D_{i2} = 0]\frac{P(D_{i2} = 1|X_i)}{P(D_{i2} = 1)}\frac{P(D_{i2} = 0)}{P(D_{i2} = 0|X_i)}dF_{X|D_{i2}=0} \quad (4.56) \end{aligned}$$

However, by Bayes's rule it also follows that for any set A we must have that

$$\begin{aligned} P(X_i \in A|D_{i2} = 0)\frac{P(D_{i2} = 1|X_i \in A)}{P(D_{i2} = 1)}\frac{P(D_{i2} = 0)}{P(D_{i2} = 0|X_i \in A)} & \\ = \frac{P(X_i \in A, D_{i2} = 0)}{P(D_{i2} = 0)}\frac{P(D_{i2} = 1, X_i \in A)}{P(X_i \in A)P(D_{i2} = 1)}\frac{P(D_{i2} = 0)P(X_i \in A)}{P(D_{i2} = 0, X_i \in A)} & \\ = P(X_i \in A|D_{i2} = 1) \quad (4.57) & \end{aligned}$$

Therefore, combining results (4.55), (4.56), and (4.57) we finally conclude that

$$E[(Y_{i2} - Y_{i1})\frac{P(D_{i2} = 1|X_i)}{P(D_{i2} = 1)}\frac{P(D_{i2} = 0)}{P(D_{i2} = 0|X_i)}|D_{i2} = 0] = E[(Y_{i2}(0) - Y_{i1}(0))|D_{i2} = 1],$$

which immediately implies the claim of the Lemma. ■

There are a number of estimators for the ATT that follow the identification strategy of Lemma 4.5.2. However, many of these estimators can be challenging to implement as they may rely on nonparametric estimators for $P(D_{i2} = 1|X_i)$ when X_i is continuously distributed. As a result, it is not common to see in applied work that practitioners instead simply augment the regression of Lemma 4.5.1 and estimate

$$Y_{it} = \alpha + \gamma 1\{D_{i2} = 1\} + \lambda 1\{t = 2\} + \delta 1\{D_{i2} = 1, t = 2\} + X_i'\theta + \varepsilon_{it}.$$

This regression estimates a causal effect provided it is properly specified in the sense

$$E[\varepsilon_{it}|D_{i2}, X_i] = 0. \quad (4.58)$$

However, by direct calculation it also follows that if condition (4.58) holds, then we have

$$\begin{aligned} &E[Y_{i2}(0) - Y_{i1}(0)|D_{i2} = 0, X_i] \\ &\quad = E[Y_{i2}|D_{i2} = 0, X_i] - E[Y_{i1}|D_{i2} = 0, X_i] = (\alpha + \lambda + X_i'\theta) - (\alpha + X_i'\theta) = \lambda. \end{aligned}$$

In other words, under (4.58), we are ruling out that the effect of time on the outcomes depend on X_i, which was the entire motivation of the conditional trends assumption to begin with! In general, when including covariates in a DiD regression make sure to think through why you are doing it and how they are entering the regression.

4.5.2 Extensions and Complications

Below we discuss a number of extensions to the basic model and challenges/opportunities that often arise in application concerning differences in differences.

4.5.2.1 Multiple Time Periods

In applications it is common to observe individuals for multiple time periods both before treatment and after treatment. Building on the basic model of Section 4.5.1, we now assume we observe $1 \leq i \leq n$ for $1 \leq t \leq T$ time periods with treatment turning on for some individuals at time period $t^\star$. Formally we impose the following structure

Assumption DiDt-1. *(i) $\{Y_{it}, D_{it}\}_{t=1}^T$ are generated according to equation (4.54); (ii) There is a $t^\star$ such that the sequence $\{D_{it}\}_{t=1}^T$ satisfies $P(D_{it} = 0) = 1$ for all $1 \leq t < t^\star$, $P(D_{it^\star} = 1) \in (0, 1)$, and $P(D_{it^\star} = D_{is}) = 1$ for all $t^\star \leq s \leq T$.*

In particular note that, besides requiring that some individuals be treated at time $t^\star$, Assumption DiDt-1 further requires that no individuals change their treatment status in the subsequent time periods. The availability of multiple time periods implies there are now potentially additional parameters of interest to identify. For instance, returning to Example 4.5.1, it is natural to consider the dynamic effects of a raise in the minimum wage increase. Concretely, for any $t \geq t^\star$, we may now attempt to identify

$$\text{ATT}_t \equiv E[Y_{it}(1) - Y_{it}(0)|D_{it^\star} = 1].$$

As in the basic model, the standard approach is to rely on a parallel trends assumption. Following Section 4.5.1 we may decompose the ATT_t into

$$\text{ATT}_t = E[Y_{it}(1) - Y_{is}(0)|D_{it^\star} = 1] - E[Y_{it}(0) - Y_{is}(0)|D_{it^\star} = 1]$$

for any $1 \leq s < t^\star$ and then obtain identification by requiring the parallel trends

$$E[Y_{it}(0) - Y_{is}(0)|D_{it^\star} = 1] = E[Y_{it}(0) - Y_{is}(0)|D_{it^\star} = 0].$$

Unlike in Section 4.5.1, however, there may be multiple pre-treatment periods on which the parallel trends are imposed. Since it is often not credible to require parallel rends to hold for, e.g., one preceding period but not two preceding periods, the parallel trends requirement in the presence of multiple time periods often takes the form

Assumption DiDt-2. *(i)* $E[Y_{it}(0) - Y_{is}(0)|D_{it^\star} = 1] = E[Y_{it}(0) - Y_{is}(0)|D_{it^\star} = 0]$ *for all* $1 \leq s < t^\star$ *and* $t^\star \leq t \leq T$.

If there are multiple pre-treatment periods, then Assumption DiDt-2 in fact provides us with multiple potential ways to estimate ATT_t. In particular, we obtain

$$\begin{aligned} \text{ATT}_t &= E[Y_{it}(1) - Y_{is}(0)|D_{it^\star} = 1] - E[Y_{it}(0) - Y_{is}(0)|D_{it^\star} = 0] && (4.59) \\ &= E[Y_{it}(1) - Y_{i\tilde{s}}(0)|D_{it^\star} = 1] - E[Y_{it}(0) - Y_{i\tilde{s}}(0)|D_{it^\star} = 0] && (4.60) \end{aligned}$$

for any $1 \leq s < \tilde{s} < t^\star$. Hence, in the language of Section 3.3.2, the model is overidentified in the sense that it may potentially be rejected by the data – e.g. we may obtain estimates of ATT_t following both (4.59) and (4.60) and test whether they are statistically different from each other. A common approach to assess whether Assumption DiDt-2 is credible is to note that for (4.59) to equal (4.60) we must have

$$E[Y_{i\tilde{s}}(0) - Y_{is}(0)|D_{it^\star} = 1] = E[Y_{i\tilde{s}}(0) - Y_{is}(0)|D_{it^\star} = 0] \quad (4.61)$$

for all $1 \leq s < \tilde{s} < t^\star$ – i.e. the treated and un-treated individuals must follow *parallel trends* in the time periods preceding treatment. Given the importance of this implication it is common in DiD applications to see plots of pre-trends (i.e. (4.61)) and authors argue that they are similar. It is fairly unusual, however, to see actual empirical tests of (4.61). Tests of (4.61) are straightforward to devise and implement though not always employed in applied work (you may wonder why!).

As in the basic model, inference on ATT_t is often done by employing regression based estimators. The following lemma highlights this point

Lemma 4.5.3. *Let Assumptions DiDt-1 and DiDt-2 hold and consider the regression*

$$Y_{it} = \alpha + \gamma 1\{D_{it^\star} = 1\} + \sum_{r=2}^{T} \lambda_r 1\{t = r\} + \sum_{r=t^\star}^{T} \delta_r 1\{D_{it^\star} = 1, r = t\} + \varepsilon_{it}.$$

Then it follows that $\delta_t = \text{ATT}_t$.

Proof: Follows by the same arguments as in Lemma 4.5.1. ■

4.5.2.2 Multiple Time Periods and Multiple Groups

In many applications, we do not only observe multiple time periods, but multiple groups receiving treatment in different time periods – e.g. when studying minimum wage increases in the United States we may see the minimum wage change for different states at different times. Below we discuss recent work on identification and estimation in this context, which includes Callaway and Sant'Anna (2019), Abraham and Sun (2018), and De Chaisemartin and d'Haultfoeuille (2020).

First, note that in the classical differences in differences framework of Section 4.5.1 there are two groups: Those that are treated in period two (so $D_{i2} = 1$) and those that are untreated in period two (so $D_{i2} = 0$). We next generalize this observation by allowing each individual to belong to a group $1 \leq g \leq G$ with G possible larger than two. As in Section 4.5.1, however, we continue to assume that all members of a common group share the same treatment status at time t and we let $D_{g,t}$ be the corresponding dummy variable indicating treatment – this is sometimes referred to as a "sharp" design. For $Y_{i,g,t}$ the observed outcome for person i in group g at time t, we suppose

$$Y_{i,g,t} = Y_{i,g,t}(0) + D_{g,t}(Y_{i,g,t}(1) - Y_{i,g,t}(0)) \tag{4.62}$$

where $(Y_{i,g,t}(0), Y_{i,g,t}(1))$ are the potential outcomes of person i in group g at periods t.

We begin by imposing an assumption that formalizes the introduced framework.

Assumption DiDg-1. *(i) $\{Y_{i,g,t}, D_{g,t}\}_{i,g,t}$ are generated according to equation (4.62); (ii) Each group g has n_g members at each time period t.*

In Assumption DiDg-1(ii) we impose that the the number of observations in group g be constant through time for simplicity, but we note this requirement can be relaxed. We also note that Assumption DiDg-1 does not require a "real" panel structure in the sense that we observe the same individual for multiple time periods. In fact, Assumption DiDg-1 can be satisfied if we observe repeated cross sections – i.e., individual i in group g at time periods t and $t+1$ need not be the same person.

Given Assumption DiDg-1, we next introduce the following useful notation

$$\bar{Y}_{g,t} = \frac{1}{n_g}\sum_{i=1}^{n_g} Y_{i,g,t} \qquad \bar{Y}_{g,t}(0) = \frac{1}{n_g}\sum_{i=1}^{n_g} Y_{i,g,t}(0) \qquad \bar{Y}_{g,t}(1) = \frac{1}{n_g}\sum_{i=1}^{n_g} Y_{i,g,t}(1).$$

For $\mathbb{D} = \{D_{g,t}\}_{g,t}$, we next introduce a suitable version of the parallel trends assumption.

Assumption DiDg-2. *$E[\bar{Y}_{g,t}(0) - \bar{Y}_{g,t-1}(0)|\mathbb{D}] = E[\bar{Y}_{g',t}(0) - \bar{Y}_{g',t-1}(0)|\mathbb{D}]$ for any $1 \leq g, g' \leq G$ and $2 \leq t \leq T$ – i.e. groups have parallel trends in the absence of treatment.*

Following the logic of Sections 4.5.1 and 4.5.2.1, it is not hard to see that we may identify the average treatment effect on a treated group by using the time trend of an untreated group as a control – i.e. if we have a group g for which $D_{g,t} = 1$ and $D_{g,t-1} = 0$ and another group g' for which $D_{g',t} = D_{g',t-1} = 0$, then Assumption DiDg-2 yields

$$\begin{aligned} E[\bar{Y}_{g,t} - \bar{Y}_{g,t-1}|\mathbb{D}] - E[\bar{Y}_{g',t} - \bar{Y}_{g',t-1}|\mathbb{D}] & \\ &= E[\bar{Y}_{g,t}(1) - \bar{Y}_{g,t-1}(0)|\mathbb{D}] - E[\bar{Y}_{g',t}(0) - \bar{Y}_{g',t-1}(0)|\mathbb{D}] \\ &= E[\bar{Y}_{g,t}(1) - \bar{Y}_{g,t-1}(0)|\mathbb{D}] - E[\bar{Y}_{g,t}(0) - \bar{Y}_{g,t-1}(0)|\mathbb{D}] \\ &= E[\bar{Y}_{g,t}(1) - \bar{Y}_{g,t}(0)|\mathbb{D}]. \end{aligned} \tag{4.63}$$

De Chaisemartin and d'Haultfoeuille (2020) note, however, that many empirical papers do not obtain estimates by following the identification strategy in (4.63). Instead, it is not uncommon for researchers to simply augment the regression of Lemma 4.5.1 by including time and group fixed effects – i.e. they estimate the regression

$$Y_{i,g,t} = \alpha + \sum_{s=2}^{T} \lambda_s 1\{s = t\} + \sum_{s=2}^{G} \gamma_s 1\{s = g\} + \delta D_{g,t} + \varepsilon_{i,g,t}.$$

The resulting regression estimator for δ, which we denote by $\hat{\delta}_n^{\text{fe}}$, is then interpreted as being consistent for a causal parameter. But is $\hat{\delta}_n^{\text{fe}}$ actually consistent for a causal parameter? De Chaisemartin and d'Haultfoeuille (2020) show that in fact the answer is not necessarily (and in fact, often no!). To explain their result, define

$$\Delta_{g,t} = \frac{1}{n_g} \sum_{i=1}^{n_g} (Y_{i,g,t}(1) - Y_{i,g,t}(0)) \qquad w_{g,t} = \frac{D_{g,t} r_{g,t}}{\sum_{g,t} D_{g,t} r_{g,t}}$$

where $r_{g,t}$ are the residuals from regressing $\{D_{g,t}\}_{g,t}$ on time and group fixed effects. Intuitively, $\Delta_{g,t}$ represents an unbiased estimator of the treatment effect for group g and time t, while $w_{g,t}$ are weights that sum up to one *but are not necessarily positive.*

The next lemma characterizes the expectation of $\hat{\delta}_n^{\text{fe}}$ under our assumptions.

Lemma 4.5.4. *If Assumptions DiDg-1 and DiDg-2 holds, then it follows that*

$$E[\hat{\delta}_n^{\text{fe}}|\mathbb{D}] = \sum_{g,t} w_{g,t} E[\Delta_{g,t}|\mathbb{D}]$$

PROOF: The arguments are straightforward, but require a bit of algebra. First, note that by definition of $Y_{i,g,t}$, $\Delta_{g,t}$, $\bar{Y}_{g,t}$, and $\bar{Y}_{g,t}(0)$ we have

$$\bar{Y}_{g,t} = \frac{1}{n_g} \sum_{i=1}^{n_g} Y_{i,g,t} = \frac{1}{n_g} \sum_{i=1}^{n_g} (Y_{i,g,t}(0) + D_{g,t}(Y_{i,g,t}(1) - Y_{i,g,t}(0))) = \bar{Y}_{g,t}(0) + D_{g,t}\Delta_{g,t}$$

Hence, since $D_{g,t}$ is part of $\mathbb{D}$, we obtain for any time periods t, t' and groups g, g' that

$$E[\bar{Y}_{g,t} - \bar{Y}_{g,t'}|\mathbb{D}] - E[\bar{Y}_{g',t} - \bar{Y}_{g',t'}|\mathbb{D}] = E[\bar{Y}_{g,t}(0) - \bar{Y}_{g,t'}(0)|\mathbb{D}] - E[\bar{Y}_{g',t}(0) - \bar{Y}_{g',t'}(0)|\mathbb{D}] \\ + (D_{g,t}E[\Delta_{g,t}|\mathbb{D}] - D_{g,t'}E[\Delta_{g,t'}|\mathbb{D}]) - (D_{g',t}E[\Delta_{g',t}|\mathbb{D}] - D_{g',t'}E[\Delta_{g',t'}|\mathbb{D}]) \quad (4.64)$$

Thus, by applying the common trends assumption (Assumption DiDg-2), (4.64) yields

$$E[\bar{Y}_{g,t} - \bar{Y}_{g,t'}|\mathbb{D}] - E[\bar{Y}_{g',t} - \bar{Y}_{g',t'}|\mathbb{D}] \\ = (D_{g,t}E[\Delta_{g,t}|\mathbb{D}] - D_{g,t'}E[\Delta_{g,t'}|\mathbb{D}]) - (D_{g',t}E[\Delta_{g',t}|\mathbb{D}] - D_{g',t'}E[\Delta_{g',t'}|\mathbb{D}]) \quad (4.65)$$

We next proceed to study $\hat{\delta}_n^{\text{fe}}$. First, observe that by the Frisch-Waugh-Lovell Theorem (see Section 2.2.2.1) it follows that $\hat{\delta}_n^{\text{fe}}$ can be computed from the regression

$$\hat{\delta}_n^{\text{fe}} = \arg\min_{\delta \in \mathbf{R}} \sum_{i=1}^{n} (Y_{i,g,t} - r_{g,t}\delta)^2$$

Also note that since $r_{g,t}$ are by definition the residuals from regressing $D_{g,t}$ on group and time fixed effects we have the orthogonality relations

$$\sum_{g,t} r_{g,t}(D_{g,t} - r_{g,t}) = 0 \qquad \sum_{g=1}^{G} r_{g,t} = 0 \qquad \sum_{t=1}^{T} r_{g,t} = 0 \quad (4.66)$$

Obtaining a closed form solution for $\hat{\delta}_n^{\text{fe}}$ and using the first equality in (4.66) then implies

$$\hat{\delta}_n^{\text{fe}} = \frac{\sum_{i,g,t} Y_{i,g,t} r_{g,t}}{\sum_{i,g,t} r_{g,t}^2} = \frac{\sum_{i,g,t} Y_{i,g,t} r_{g,t}}{\sum_{g,t} n_g r_{g,t}^2} = \frac{\sum_{g,t} \bar{Y}_{g,t} r_{g,t}}{\sum_{g,t} r_{g,t} D_{g,t}} \quad (4.67)$$

We next study the numerator in (4.67), and to this end we note (4.66) yields

$$\sum_{g,t} \bar{Y}_{g,t} r_{g,t} = \sum_{g,t} \bar{Y}_{g,t} r_{g,t} - \sum_{g,t} \bar{Y}_{g,1} r_{g,t} - \sum_{g,t} \bar{Y}_{1,t} r_{g,t} + \sum_{g,t} \bar{Y}_{1,1} r_{g,t} \quad (4.68)$$

Note, however, that $E[r_{g,t}|\mathbb{D}] = r_{g,t}$ since $r_{g,t}$ is a function of $D_{g,t}$ and the fixed effects. Hence, combining result (4.68) with the equality established in (4.65) we finally obtain

$$\begin{aligned} &E[\sum_{g,t} \bar{Y}_{g,t} r_{g,t}|\mathbb{D}] \\ &\quad = \sum_{g,t} r_{g,t}((E[\bar{Y}_{g,t}|\mathbb{D}] - E[\bar{Y}_{g,1}|\mathbb{D}]) - (E[\bar{Y}_{1,t}|\mathbb{D}] - E[\bar{Y}_{1,1}|\mathbb{D}])) \\ &\quad = \sum_{g,t} r_{g,t}(D_{g,t}E[\Delta_{g,t}|\mathbb{D}] - D_{g,1}E[\Delta_{g,1}|\mathbb{D}]) - (D_{1,t}E[\Delta_{1,t}|\mathbb{D}] - D_{1,1}E[\Delta_{1,1}|\mathbb{D}]) \\ &\quad = \sum_{g,t} r_{g,t} D_{g,t} E[\Delta_{g,t}|\mathbb{D}], \end{aligned} \quad (4.69)$$

where the final equality follows from the orthogonality conditions in (4.66). Finally, combining results (4.67) and (4.69) we can conclude that

$$E[\hat{\delta}_n^{\text{fe}}|\mathbb{D}] = \frac{\sum_{g,t} r_{g,t} D_{g,t} E[\Delta_{g,t}|\mathbb{D}]}{\sum_{g,t} r_{g,t} D_{g,t}} = \sum_{g,t} w_{g,t} E[\Delta_{g,t}|\mathbb{D}],$$

which establishes the claim of the lemma. ■

The key implication of Lemma 4.5.4 is that because the weights $w_{g,t}$ may be negative, the expectation of $\hat{\delta}_n^{\text{fe}}$ may not have a clear causal interpretation. For example, it is not hard to construct examples in which average treatment effects are positive (i.e. $E[\Delta_{g,t}|\mathbb{D}] > 0$) but $E[\hat{\delta}_n^{\text{fe}}|\mathbb{D}]$ is negative. As we have emphasized in the previous sections, remember to think carefully through your regression specifications when applying differences in differences outside the basic setup of Section 4.5.1.

4.5.2.3 Changes in Changes

Athey and Imbens (2006) propose a generalization of the basic Differences in Differences framework of Section 4.5.1 sometimes referred to as the changes in changes model.

We maintain the basic setup in which we observe individuals $1 \leq i \leq n$ for two time periods with no individuals treated in the first time period and some (but not all) treated in the second time period. Employing the potential outcomes notation and letting D_i indicate whether individual i is treated in the second period we thus obtain

$$Y_{i1} = Y_{i1}(0) \qquad Y_{i2} = Y_{i2}(0) + D_i(Y_{i2}(1) - Y_{i2}(0))$$

In what follows we suppress i from the notation. The following assumption contains the key requirements of the changes in changes (CiC) model.

Assumption CiC-1. *(i)* (Y_1, Y_2, D) *with* $Y_1 = Y_1(0)$ *and* $Y_2 = Y_2(0) + D(Y_2(1) - Y_2(0))$*; (ii)* $Y_t(0) = g_t(U)$ *for strictly increasing functions* g_t *and continuously distributed* U*.*

Assumption CiC-1(i) formalizes the setting in which no individuals are treated in the first time period. The main requirement of the CiC model are in Assumption CiC-1, which is commonly known as a *rank invariance* assumption. In particular, note that since g_t is strictly increasing and U is common across time periods, Assumption CiC-1(ii) implies the "rank" of an individual when individuals are ordered according to $Y_1(0)$ is the same as when they are ordered according to $Y_2(0)$ – e.g., the individual with the largest value for $Y_1(0)$ also has the largest value for $Y_2(0)$. The requirement that U be continuously distributed is not essential and imposed here for simplicity; see Athey and Imbens (2006) for a more general treatment.

The rank invariance assumption has a lot of identifying power. In the next lemma, $F_{V|D=d}$ denotes the the c.d.f. of a random variable V conditional on $D = d$.

Lemma 4.5.5. *If Assumption CiC-1 holds, then it follows for any $y \in \mathbf{R}$ we have that*

$$P(Y_2(0) \leq y|D = 1) = F_{Y_1(0)|D_2=1}(F^{-1}_{Y_1(0)|D=0}(F_{Y_2(0)|D=0}(y)))$$

PROOF: First note that for any $y \in \mathbf{R}$, the strict monotonicity of g_t implies that

$$\begin{aligned} F_{Y_2(0)|D=1}(y) \equiv P(Y_2(0) \leq y|D = 1) &= P(U \leq g_2^{-1}(y)|D = 1) \\ &= P(Y_1(0) \leq g_1(g_2^{-1}(y))|D = 1) \equiv F_{Y_1(0)|D=1}(g_1(g_2^{-1}(y))). \quad (4.70) \end{aligned}$$

Thus, since $F_{Y_1(0)|D=1}$ can be identified from the data due to $Y_1(0) = Y_1$ by Assumption CiC-1(i), it follows that if we can identify $g_1(g_2^{-1}(y))$, then we can identify the distribution of $Y_2(0)$ conditional on $D = 1$. However, identical manipulations yield

$$F_{Y_2(0)|D=0}(y) = F_{Y_1(0)|D=0}(g_1(g_2^{-1}(y))).$$

Hence, for $F^{-1}_{Y_1(0)|D=0}(\tau)$ the τ-quantile of $Y_1(0)$ given $D = 0$, we can conclude that

$$g_1(g_2^{-1}(y)) = F^{-1}_{Y_1(0)|D=0}(F_{Y_2(0)|D=0}(y)),$$

which together with (4.70) establishes the claim of the lemma. ■

Since Lemma 4.5.5 implies $F_{Y_2(0)|D=1}$ is identified and $F_{Y_2(1)|D=1}$ is identified directly from the data, it follows that we can identify any functional of these two distributions. These include the ATT, but also other interesting parameters such as

$$F^{-1}_{Y_2(1)|D=1}(\tau) - F^{-1}_{Y_2(0)|D=1}(\tau)$$

which is the quantile treatment effect on the treated.

4.6 Problems

1. Consider a panel data setting with $1 \leq i \leq n$ and $1 \leq t \leq T$. Suppose a dependent variable Y_{it} is related to a regressor X_i according to

$$Y_{it} = X_i'\beta_0 + U_{it}$$

with $E[X_i U_{it}] = 0$ – i.e. in this setting the regressor of interest only varies at the individual level (think race or gender). Consider the following two estimators:

$$\hat{\beta}_n^{\mathrm{f}} \equiv \arg\min_{\beta \in \mathbf{R}^d} \frac{1}{n} \sum_{i=1}^{n} \sum_{t=1}^{T} (Y_{it} - X_i'\beta)^2 \tag{4.71}$$

$$\hat{\beta}_n^{\mathrm{m}} \equiv \arg\min_{\beta \in \mathbf{R}^d} \frac{1}{n} \sum_{i=1}^{n} (\frac{1}{T} \sum_{t=1}^{T} Y_{it} - X_i'\beta)^2. \tag{4.72}$$

(a) Show that $\hat{\beta}_n^{\mathrm{f}}$ and $\hat{\beta}_n^{\mathrm{m}}$ are numerically equivalent.

(b) Derive the asymptotic distributions of $\hat{\beta}_n^{\mathrm{f}}$ and $\hat{\beta}_n^{\mathrm{m}}$ with T fixed and n diverging to infinity. Impose whatever assumptions you need to accomplish this.

(c) Consider the following two variance estimates for $\hat{\beta}_n^{\mathrm{f}}$ and $\hat{\beta}_n^{\mathrm{m}}$:

$$\hat{\Sigma}_n^{\mathrm{f}} \equiv \{\frac{1}{n} \sum_{i=1}^{n} \sum_{t=1}^{T} X_i X_i'\}^{-1} \{\frac{1}{n} \sum_{i=1}^{n} (\sum_{t=1}^{T} X_i (Y_i - X_i'\hat{\beta}_n^{\mathrm{f}}))(\sum_{t=1}^{T} X_i (Y_i - X_i'\hat{\beta}_n^{\mathrm{f}}))'\} \{\frac{1}{n} \sum_{i=1}^{n} \sum_{t=1}^{T} X_i X_i'\}^{-1}$$

$$\hat{\Sigma}_n^{\mathrm{m}} \equiv \{\frac{1}{n} \sum_{i=1}^{n} X_i X_i'\}^{-1} \{\frac{1}{n} \sum_{i=1}^{n} X_i X_i' (\frac{1}{T} \sum_{t=1}^{T} Y_{it} - X_i'\hat{\beta}_n^{\mathrm{m}})^2\} \{\frac{1}{n} \sum_{i=1}^{n} X_i X_i'\}^{-1};$$

Show $\hat{\Sigma}_n^{\mathrm{f}}$ and $\hat{\Sigma}_n^{\mathrm{m}}$ are numerically equivalent.

(d) Based on parts (a)-(c), do you think it is better to (I) Estimate β_0 using the full sample and compute cluster robust standard errors, or (II) Estimate β_0 by regressing $\sum_{t=1}^{T} Y_{it}/T$ on X_i and computing standard errors in the usual way (as in Chapter 2). Justify your answer.

2. Suppose we observe a sample $\{Y_{it}\}$ with $1 \leq i \leq n$, $1 \leq t \leq T$, $n = T$, and

$$Y_{it} = \mu + \alpha_i + \gamma_t + \epsilon_{it}$$

where $\mu \in \mathbf{R}$ is a constant, $\{\gamma_t\}_{t=1}^{T}$ are i.i.d., $\{\alpha_i\}_{i=1}^{n}$ are i.i.d., and $\{\epsilon_{it}\}$ are i.i.d. Further assume that the variables $E[\alpha_i] = E[\gamma_t] = E[\epsilon_{it}] = 0$, and that α_i, γ_t, and ϵ_{it} are mutually independent (across $1 \leq i \leq n$ and $1 \leq t \leq T$).

(a) Characterize $E[Y_{it} Y_{i't'}]$ as a function of whether $i = i'$ (or not), $t = t'$ (or not), $\mathrm{Var}\{\alpha_i\}$, $\mathrm{Var}\{\gamma_t\}$, and $\mathrm{Var}\{\epsilon_{it}\}$. Note that if $\mathrm{Var}\{\alpha_i\}$ and $\mathrm{Var}\{\gamma_t\}$ are greater than zero, then this is a setting of multi-way clustering.

(b) Show that if $\mathrm{Var}\{\gamma_t\} = \mathrm{Var}\{\alpha_i\} = 0$, then it follows that (as $n = T \to \infty$):

$$\frac{1}{\sqrt{nT}} \sum_{i=1}^{n} \sum_{t=1}^{T} \{Y_{it} - E[Y_{it}]\} \xrightarrow{d} N(0, \mathrm{Var}(\epsilon_{it})).$$

(c) Show that if either $\text{Var}\{\gamma_t\} > 0$ or $\text{Var}\{\alpha_i\} > 0$, then (as $n = T \to \infty$):

$$\frac{1}{T\sqrt{n}} \sum_{i=1}^{n} \sum_{t=1}^{T} \{Y_{it} - E[Y_{it}]\} \xrightarrow{d} N(0, \text{Var}\{\alpha_i\} + \text{Var}\{\gamma_t\}).$$

Comparing to part (b), note that the distribution of the sample mean (even its rate of convergence) can depend on unknown features of the data distribution.

3. Throughout this problem let Assumption C-1 hold and additionally assume that $\sum_{t=1}^{T} E[\|X_{it}\|^4]$ and $\sum_{t=1}^{T} E[\|X_{it}\|^3 |U_{it}|]$ are finite.

(a) Show that, as n tends to infinity, the variance estimator in (4.9) is consistent for the asymptotic variance obtained in Lemma 4.1.1.

(b) Suppose the data is not clustered, meaning $E[X_{it} X_{i\tilde{t}}' U_{it} U_{i\tilde{t}}] = 0$ whenever $\tilde{t} \neq t$. Show that the variance estimator in (4.9) converges in probability to

$$\{\frac{1}{n} \sum_{i=1}^{n} X_i' X_i\}^{-1} \{\frac{1}{n} \sum_{i=1}^{n} \sum_{t=1}^{T} X_{it} X_{it}' \hat{U}_{it}^2\} \{\frac{1}{n} \sum_{i=1}^{n} X_i' X_i\}^{-1},$$

which is the asymptotic variance estimator one would use if we did not think clustering was of concern.

4. Use Lemma 3.2.3 to show that the optimal choice of Ω in Theorem 4.2.1 is to set $\Omega = (E[X_i' \Sigma X_i])^{-1}$ where recall $\Sigma = E[U_i U_i' | X_i]$.

5. This problem studies the so called correlated random effects model. Suppose that

$$Y_{it} = X_{it}' \beta_0 + A_i + V_{it} \tag{4.73}$$

where A_i an individual fixed effect, $X_{it} \in \mathbf{R}^d$ is possibly correlated with A_i and V_{it} is a person time specific shock independent of all other variables. Assume that the data is balanced, so that we have T observations for each individual.

(a) Suppose that, in addition to the stated conditions, we impose that A_i satisfies

$$A_i = \alpha_0 + \alpha_1 \bar{X}_i + \eta_i, \tag{4.74}$$

where η_i is independent of X_i and $\bar{X}_i = \sum_t X_{it}/T$. Employing equations (4.73) and (4.74), note that we can then write the model as

$$Y_{it} = \alpha_0 + \alpha_1 \bar{X}_i + X_{it}' \beta + \eta_i + V_{it}. \tag{4.75}$$

Show that the OLS estimator for (4.75) is consistent – this estimator is sometimes called the correlated random effects estimator.

(b) Suppose you instead estimate β_0 in (4.73) by running an OLS regression that includes X_{it} and a dummy variable for each individual. How does this estimator relate to the estimator from part (a)?

(c) What do you think of this statement "*Correlated Random Effects Estimators rely upon stronger assumptions than fixed effects estimators.*"

6. Let $Y_i = (Y_{i1}, \ldots, Y_{iT})'$ and $\{Y_i\}_{i=1}^n$ be an i.i.d. sample. Further suppose that

$$Y_{it} = A_i + B_t + V_{it}$$

with $E[V_{it}] = 0$ and where we treat A_i and B_t as parameters to be estimated. Let D_i be as in (4.36) and consider estimating $(A_1, \ldots, A_n)$ and $(B_1, \ldots, B_T)$ by

$$(\hat{\alpha}_n, \{\hat{\beta}_t\}_{t=1}^T) \equiv \arg \min_{a \in \mathbf{R}^n, \{b_t\}_{t=1}^T \in \mathbf{R}^T} \frac{1}{n} \sum_{i=1}^n \sum_{t=1}^T (Y_{it} - D_i' a - \sum_{\tilde{t}=1}^T 1\{t = \tilde{t}\} b_t)^2.$$

(a) Derive a closed form expression for $\hat{\alpha}_n$ by using Theorem 2.2.1.

(b) Writing $\hat{\alpha}_n = (\hat{A}_1, \ldots, \hat{A}_n)'$, show that for any j, $\hat{A}_j$ *is not consistent* for A_j.

7. In this problem we study the first differences estimator, which is an alternative to the fixed effects estimator. Throughout, let Assumption FE-1 hold, and define

$$\Delta X_{it} = X_{it} - X_{i(t-1)} \qquad \Delta Y_{it} = Y_{it} - Y_{i(t-1)} \qquad \Delta V_{it} = V_{it} - V_{i(t-1)}.$$

The first differences estimator, denoted $\hat{\beta}_n^{\text{fd}}$, is then defined as the minimizer to

$$\hat{\beta}_n^{\text{fd}} \equiv \arg \min_{b \in \mathbf{R}^d} \frac{1}{nT} \sum_{i=1}^n \sum_{t=2}^T (\Delta Y_{it} - \Delta X_{it}' b)^2.$$

(a) Impose regularity conditions (e.g. moment restrictions, rank restrictions on matrices) that allow you to show that $\hat{\beta}_n^{\text{fd}}$ is consistent for β_0. Establish the consistency formally.

(b) Show that if $T = 2$ then the fixed effects estimator and the first differences estimator are numerically equivalent.

(c) Impose regularity conditions that allow you to show that $\sqrt{n}\{\hat{\beta}_n^{\text{fd}} - \beta_0\}$ is asymptotically normally distributed. Establish the asymptotic normality result formally and characterize the asymptotic variance.

(d) Propose an estimator for the asymptotic variance you found in part (b). A formal proof of consistency is not needed.

8. Consider a panel data model, where each individual $1 \leq i \leq n$ is observed for $1 \leq t \leq T$ periods. Let $Y_i = (Y_{i1}, \ldots, Y_{iT})'$ with $Y_{it} \in \mathbf{R}$ denote wages, $X_i =$

$(X_{i1}, \ldots, X_{iT})'$ with $X_{it} \in \mathbf{R}^{d_x}$ equal a time varying regressor such as experience, and $Z_i \in \mathbf{R}^{d_z}$ be an observable individual specific characteristic that does not change through time such as gender. Further suppose that

$$Y_{it} = X_{it}'\beta_0 + Z_i'\delta_0 + \alpha_i + U_{it} \tag{4.76}$$

for unknown $\beta_0 \in \mathbf{R}^{d_x}$ and $\delta_0 \in \mathbf{R}^{d_z}$ and unobservable variables α_i and $U_i = (U_{i1}, \ldots, U_{iT})'$. Throughout the problem assume $\{Y_i, X_i, Z_i\}_{i=1}^n$ are i.i.d., $E[U_i] = 0$ with U_i is independent of (X_i, Z_i), and that $E[\alpha_i|Z_i] = 0$. However, we are still concerned that α_i is potentially correlated with X_i.

(a) Provide conditions under which β_0 is identified by taking first differences (i.e. by using the variables $\Delta Y_{it} \equiv Y_{it} - Y_{it-1}$ and $\Delta X_{it} \equiv X_{it} - X_{it-1}$). Can δ_0 be identified from the first difference equation? Justify your answer.

(b) Derive an asymptotic distribution for the first differences estimator for β_0.

(c) Suppose β_0 is known. Provide conditions under which δ_0 is identified.

(d) Employing the insights from parts (a)-(b) propose an estimator for δ_0 that uses the fixed effects estimator $\hat{\beta}_n^{\text{fe}}$ in its construction.

(e) Show the estimator you proposed in part (c) is consistent. Clearly state any additional regularity conditions you need to impose.

(f) Rigorously derive the asymptotic distribution for the estimator of δ_0 you proposed in part (c). Clearly state any additional conditions you need.

9. Consider a differences in differences setup in which we observe individuals for $1 \leq t \leq 3$ time periods. All individuals are untreated at $t \in \{1, 2\}$ and some (but not) all individuals are treated at $t = 3$. Let D_i be an indicator for whether individual i is treated at $t = 3$. Throughout assume the usual potential outcomes model, and note that we observe

$$Y_{it} = \begin{cases} Y_{it}(0) & \text{if } 1 \leq t \leq 2 \\ Y_{it}(0) + D_i(Y_{it}(1) - Y_{it}(0)) & \text{if } t = 3 \end{cases}.$$

Throughout the problem assume that $\{Y_{i1}, Y_{i2}, Y_{i3}, D_i\}_{i=1}^n$ is an i.i.d. sample.

(a) Suppose that we are willing to impose the parallel trend assumtion

$$E[Y_{i3}(0) - Y_{it}(0)|D_i = 1] = E[Y_{i3}(0) - Y_{it}(0)|D_i = 0] \text{ for } 1 \leq t \leq 2.$$

Show that the ATT $\equiv E[Y_{i3}(1) - Y_{i3}(0)|D_i = 1]$ is identified through

$$\text{ATT} = E[Y_{i3} - Y_{i2}|D_i = 1] - E[Y_{i3} - Y_{i2}|D_i = 0] \tag{4.77}$$

$$= E[Y_{i3} - Y_{i1}|D_i = 1] - E[Y_{i3} - Y_{i1}|D_i = 0] \tag{4.78}$$

(b) Propose two different estimators for ATT, one based on (4.77) (denote $\widehat{\text{ATT}_2}$) and a second one based on (4.78) (denote $\widehat{\text{ATT}_3}$).

(c) Derive the joint asymptotic distribution of $(\sqrt{n}\{\widehat{\text{ATT}_3} - \text{ATT}\}, \sqrt{n}\{\widehat{\text{ATT}_2} - \text{ATT}\})$. Formally state any assumptions you need to establish this result.

(d) By employing the result from part (c), propose a test for the null hypothesis that the parallel trends assumption indeed holds by comparing $\widehat{\text{ATT}_2}$ to $\widehat{\text{ATT}_3}$. Describe what the test would be if we want the size to be α.

10. Consider a standard panel data model in which we observe $1 \leq i \leq n$ individuals through $1 \leq t \leq T$ time periods. Also assume

$$Y_{it} = X_{it}'\beta_0 + A_i + V_{it}$$

where A_i and V_{it} satisfy $E[A_i|X_{i1}, \ldots, X_{iT}] = 0$ and $E[V_{it}|A_i, X_{i1}, \ldots, X_{iT}] = 0$. Throughout, also let $\dot{X}_{it} = X_{it} - \bar{X}_i$ where $\bar{X}_i = \sum_{t=1}^T X_{it}/T$.

(a) Suppose that the variables $\{V_{it}\}_{t=1}^T$ are homoskedastic in that they satisfy

$$E[V_{it}V_{is}|X_{i1}, \ldots, X_{iT}] = \begin{cases} \sigma_V^2 & \text{if } t = s \\ 0 & \text{if } t \neq s \end{cases} \tag{4.79}$$

A researcher decides to estimate β_0 by employing a fixed effects estimator, which we denote by $\hat{\beta}_n^{\text{fe}}$. Establish the asymptotic distribution of $\hat{\beta}_n^{\text{fe}}$. Clearly state what assumptions are needed, and employ condition (4.79) to simplify the expression for the asymptotic variance as much as possible.

(b) The researcher proposes to estimate the asymptotic variance of $\hat{\beta}_n^{\text{fe}}$ using

$$\left(\frac{1}{n}\sum_{i=1}^n\sum_{t=1}^T \dot{X}_{it}\dot{X}_{it}'\right)^{-1} \frac{T}{n(T-1)}\sum_{i=1}^n\sum_{t=1}^T \dot{X}_{it}\dot{X}_{it}'(\widehat{\dot{V}_{it}})^2 \left(\frac{1}{n}\sum_{i=1}^n\sum_{t=1}^T \dot{X}_{it}\dot{X}_{it}'\right)^{-1} \tag{4.80}$$

where $\widehat{\dot{V}_{it}} = \dot{Y}_{it} - \dot{X}_{it}\hat{\beta}_n^{\text{fe}}$. Derive the probability limit of the variance estimator in (4.80) under assumption (4.79) and the high level condition that

$$\frac{1}{n}\sum_{i=1}^n\sum_{t=1}^T \dot{X}_{it}\dot{X}_{it}'(\widehat{\dot{V}_{it}})^2 = \frac{1}{n}\sum_{i=1}^n\sum_{t=1}^T \dot{X}_{it}\dot{X}_{it}'(\dot{V}_{it})^2 + o_p(1).$$

Clearly state any other assumptions you need to establish your result.

(c) The researcher claims that since he has used fixed effects, he does not need to worry about correlation across time and does not need to cluster at the individual level. Is he correct? Justify your answer using parts (a) and (b).

(d) Would your answer to part (c) change if we relaxed assumption (4.79)? If yes, then provide an example. If no, then explain why. No formal results are

needed, an intuitive explanation suffices.

Chapter 5

Extremum Estimation

Both the Ordinary Least Squares model of Chapter 2 and the Instrumental Variables model of Chapter 3 are linear in the parameter of interest. Our next goal is to study maximum likelihood estimation, which often concerns models that are nonlinear in the parameter of interest. Understanding the asymptotic properties of these estimators, however, requires more sophisticated arguments and we therefore first present a general framework for studying so called extremum estimators.

5.1 Basic Setup

The estimands of both OLS and IV could be described as the solution to a minimization problem that was unknown due to its dependence on the distribution of the data. Focusing on OLS for concreteness, the parameter of interest β_0 was characterized as

$$\beta_0 = \arg\min_{b \in \mathbf{R}^d} E[(Y - X'b)^2]. \tag{5.1}$$

Trivially, β_0 is unknown simply because we do not know the distribution of (Y, X). As a result, the OLS estimator simply mimics unknown population quantities by employing sample analogues – this approach is sometimes referred to as the *plug-in* or *analogy* principle. Specifically, with regards to OLS, we obtain an estimator by minimizing

$$\hat{\beta}_n = \arg\min_{b \in \mathbf{R}^d} \frac{1}{n} \sum_{i=1}^{n} (Y_i - X_i'b)^2. \tag{5.2}$$

These simple observations are immediately generalizable to other models. In particular, suppose that the parameter of interest $\theta_0 \in \Theta \subseteq \mathbf{R}^d$ is characterized as

$$\theta_0 = \arg\min_{\theta \in \Theta} Q(\theta) \tag{5.3}$$

for some function $Q : \mathbf{R}^d \to \mathbf{R}$. The function Q is unknown, however, due to its dependence on the unknown distribution of the data – e.g., (5.1) corresponds to

$$Q(\theta) = E[(Y - X'\theta)^2]. \tag{5.4}$$

Since Q is unknown, we can employ the analogy principle, by simply employing as an estimator the minimizer of a sample analogue to Q. Formally, we therefore assume that there is some known function $Q_n : \mathbf{R}^d \to \mathbf{R}$ that depends on the data, and define

$$\hat{\theta}_n = \arg\min_{\theta \in \Theta} Q_n(\theta). \tag{5.5}$$

This approach follows verbatim the rationale behind OLS, which corresponds to setting

$$Q_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i'\theta)^2. \tag{5.6}$$

Estimators that fit the general framework in (5.3) and (5.5) are referred to as extremum estimator. Despite the generality of this framework, (5.3) and (5.5) possess enough structure to enable us to study the consistency and asymptotic distribution of extremum estimators in a unified manner. Before proceeding, however, we first introduce the maximum likelihood estimator to which we will return to illustrate our analysis.

Example 5.1.1. Suppose we have an i.i.d. sample $\{W_i\}_{i=1}^n$. The distribution of W_i is unknown, but we posit that its density is equal to $f(\cdot, \theta_0)$ for some $\theta_0 \in \Theta$. The maximum likelihood estimator (MLE) is then defined as the minimizer of

$$\hat{\theta}_n = \arg\min_{\theta \in \Theta} -\frac{1}{n}\sum_{i=1}^{n}\log(f(W_i, \theta)); \tag{5.7}$$

or, equivalently, note that $\hat{\theta}_n$ equals the maximizer of the log likelihood. Hence, MLE maps into our framework by setting $Q_n : \Theta \to \mathbf{R}$ to equal the function

$$Q_n(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\log(f(W_i, \theta)). \tag{5.8}$$

By the analogy principle, it therefore follows that $\hat{\theta}_n$ must be estimating the parameter

$$\theta_0 = \arg\min_{\theta \in \Theta} Q(\theta) \qquad Q(\theta) \equiv -E[f(W, \theta)]. \tag{5.9}$$

Crucially, notice that we can think of $\hat{\theta}_n$ as an estimator of θ_0 (for θ_0 as defined by (5.9)), even if we in fact were wrong in assuming that W is distributed according to $f(\cdot, \theta_0)$ for some $\theta_0 \in \Theta$. In other words, the extremum estimation framework allows us to naturally think of the MLE estimator under *misspecification*. ■

5.2 Consistency

The intuition behind why we would expect $\hat{\theta}_n$ to be consistent for θ_0 is straightforward: If θ_0 minimizes Q and Q_n begins to "look like" Q, then the minimizer of Q_n (i.e. $\hat{\theta}_n$) should converge to the minimizer of Q (i.e. θ_0). The key challenge in this story is simply correctly defining what it means for Q_n to "begin to look like" Q.

The following assumptions will suffice for this end.

Assumption EE-1. *(i)* θ_0 *is the unique minimizer of* $Q : \Theta \to \mathbf{R}$*; (ii)* $\hat{\theta}_n$ *minimizes* $Q_n : \Theta \to \mathbf{R}$*; (iii)* Q_n *satisfies* $\sup_{\theta \in \Theta} |Q_n(\theta) - Q(\theta)| \xrightarrow{p} 0$.

Assumption EE-1(i) and EE-1(ii) formalize our framework by imposing that θ_0 and $\hat{\theta}_n$ indeed satisfy (5.4) and (5.5) respectively. Assumption EE-1(i) is sometimes referred to as an *identification* assumption, in that it crucially requires that θ_0 be the *unique* minimizer to Q – in other words, if Q had multiple minimizers, then the parameter θ_0 would not be identified by the restriction that it minimized the function Q. In turn, Assumption EE-1(iii) is the key requiring that formally imposes that Q_n begin to "look like" Q. Assumption EE-1(iii) is in fact stronger than required, but it is often simple to verify – for weaker conditions see, for example, Knight (1999).

In order to understand the content of Assumption EE-1 it is helpful to ignore randomness temporarily and work with deterministic functions instead. For a sequence of functions $\{f_n\}_{n=1}^{\infty}$ we may then consider two modes of convergence to a limit f.

Definition 5.2.1. Suppose $f_n : T \to \mathbf{R}$. Then the sequence $\{f_n\}_{n=1}^{\infty}$ convergence *pointwise* to a function $f : T \to \mathbf{R}$ if for any $t \in T$, $|f_n(t) - f(t)| \to 0$ as $n \uparrow \infty$. ■

Definition 5.2.2. Suppose $f_n : T \to \mathbf{R}$. Then the sequence $\{f_n\}_{n=1}^{\infty}$ convergence *uniformly* to a function $f : T \to \mathbf{R}$ if $\sup_{t \in T} |f_n(t) - f(t)| \to 0$ as $n \uparrow \infty$. ■

It is clear that a sequence $\{f_n\}_{n=1}^{\infty}$ converging *uniformly* to a limit f must also converge *pointwise* to the same limit. As the next Example shows, however, the converse is not true: A sequence $\{f_n\}_{n=1}^{\infty}$ may converge *pointwise* to a limit f but not uniformly.

Example 5.2.1. Let $f_n : [0,1] \to \mathbf{R}$ be given by $f_n(t) = t^n$ and $f : [0,1] \to \mathbf{R}$ equal

$$f(t) = \begin{cases} 1 & \text{if } t = 1 \\ 0 & \text{if } 0 \leq t < 1 \end{cases}; \tag{5.10}$$

see Figure 5.1. The sequence $\{f_n\}_{n=1}^{\infty}$ then converges pointwise to f. This follows because if $t = 1$, then $f_n(t) = 1^n = 1 = f(1)$ for all n, while if $0 \leq t < 1$, then

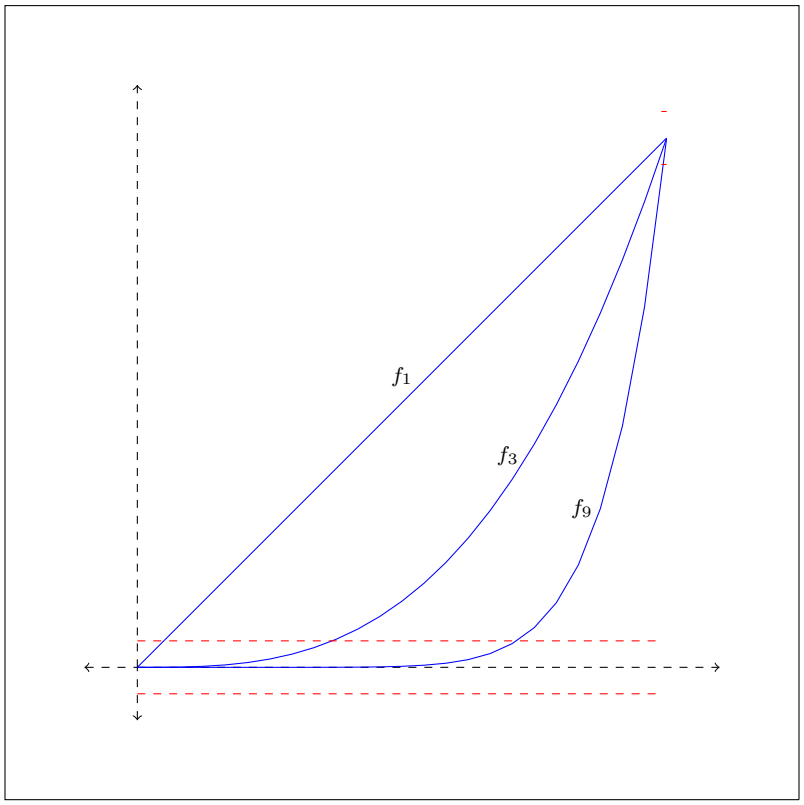

Figure 5.1: Example 5.2.1 Pointwise but not Uniform

$f_n(t) = t^n \to 0 = f(t)$. On the other hand, f_n does not converge uniformly to f. To see this, note uniform convergence requires that for any $\epsilon > 0$, there be an $n^\star$ such that

$$\sup_{0 \le t \le 1} |f_n(t) - f(t)| < \epsilon \tag{5.11}$$

for all $n > n^\star$. Intuitively, this simply means that we can draw an ϵ band around f and f_n will be within the band for all n sufficiently large; see Figure 5.1. However, for any $t \in [0, 1)$ we have from the definitions of f_n and f that

$$|f_n(t) - f(t)| = |t^n - 0| < \epsilon \Leftrightarrow n \log(t) < \log(\epsilon) \Leftrightarrow n > \frac{\log(\epsilon)}{\log(t)}. \tag{5.12}$$

Since $\log(\epsilon)/\log(t) \uparrow \infty$ as $t \uparrow 1$, result (5.12) implies that for any n we can find a t near one for which f_n is farther away from f than ϵ. This is seen in Figure 5.1, where f_n gets "far" from f as it raises from zero to one. ■

As illustrated by Example 5.2.1, the concept of uniform convergence requires that f_n be within an ϵ band of f for n sufficiently large. Returning to Assumption EE-1, we may then interpret Assumption EE-1(iii) as an "in probability" version of this same requirement. Formally, $\sup_{\theta \in \Theta} |Q_n(\theta) - Q(\theta)| \xrightarrow{p} 0$ if and only if for any $\epsilon > 0$ and $\eta > 0$

there exists an $n^\star$ such that for n larger than $n^\star$ we find that

$$P(\sup_{\theta\in\Theta}|Q_n(\theta)-Q(\theta)|<\epsilon)\geq 1-\delta;$$

i.e. Q_n is random, but it is within ϵ bands of Q with arbitrarily high probability.

Assumption EE-1(iii) is a somewhat technical condition but, as we next show, it enables us to establish the consistency of extremum estiamtors.

Theorem 5.2.1. *If Assumption EE-1 holds, Θ is compact, and $Q:\Theta\to\mathbf{R}$ is continuous, then it follows that $\hat{\theta}_n \xrightarrow{p} \theta_0$.*

PROOF: Fix $\epsilon>0$, and note that to establish the Theorem we aim to show that

$$\lim_{n\to\infty} P(\|\hat{\theta}_n-\theta_0\|<\epsilon)=1. \tag{5.13}$$

To this end, we define $N_\epsilon \equiv \{\theta\in\Theta : \|\theta-\theta_0\|<\epsilon\}$, which is simply an ϵ open neighborhood of θ_0. Setting the complement to be $N_\epsilon^c \equiv \{\theta\in\Theta : \|\theta-\theta_0\|\geq\epsilon\}$, notice N_ϵ^c is compact since it is a closed subset of a compact set Θ. Hence, we obtain

$$\delta \equiv \min_{\theta\in N_\epsilon^c} Q(\theta)-Q(\theta_0)>0, \tag{5.14}$$

because Q must attain a minimum on N_ϵ^c and the minimum must be larger than $Q(\theta_0)$ by Assumption EE-1(i). This is one of the main instances where compactness of Θ comes in, and the requirement can be dispensed with provided we have other ways of ensuring (5.14). One such condition, for example, is that Q being convex as in OLS.

Next, we define an event E_n to equal $E_n \equiv \{\sup_{\theta\in\Theta}|Q_n(\theta)-Q(\theta)|<\delta/2\}$ – i.e. E_n holds when the random function Q_n is within a $\delta/2$ band of the non-random function Q. Whenever the event E_n is true, then we are able to conclude that

$$|Q_n(\hat{\theta}_n)-Q(\hat{\theta}_n)|<\frac{\delta}{2}\Rightarrow Q(\hat{\theta})-Q_n(\hat{\theta})<\frac{\delta}{2}\Rightarrow Q_n(\hat{\theta}_n)>Q(\hat{\theta}_n)-\frac{\delta}{2}. \tag{5.15}$$

By similar arguments, it also follows that whenever E_n is true, it must be the case that

$$\begin{aligned}|Q_n(\theta_0)-Q(\theta_0)|<\frac{\delta}{2}\Rightarrow Q_n(\theta_0)-Q(\theta_0)<\frac{\delta}{2}\\ \Rightarrow Q_n(\theta_0)-\frac{\delta}{2}<Q(\theta_0)\Rightarrow Q_n(\hat{\theta}_n)-\frac{\delta}{2}<Q(\theta_0)\end{aligned} \tag{5.16}$$

where in the final inequality we used that $Q_n(\hat{\theta}_n)\leq Q(\theta_0)$ by definition of $\hat{\theta}_n$. Adding both sides of the final inequalities in (5.15) and (5.16) we obtain that E_n implies

$$Q(\theta_0)+Q_n(\hat{\theta}_n)>Q(\hat{\theta}_n)+Q_n(\hat{\theta}_n)-\delta\Rightarrow Q(\hat{\theta}_n)-Q(\theta_0)<\delta. \tag{5.17}$$

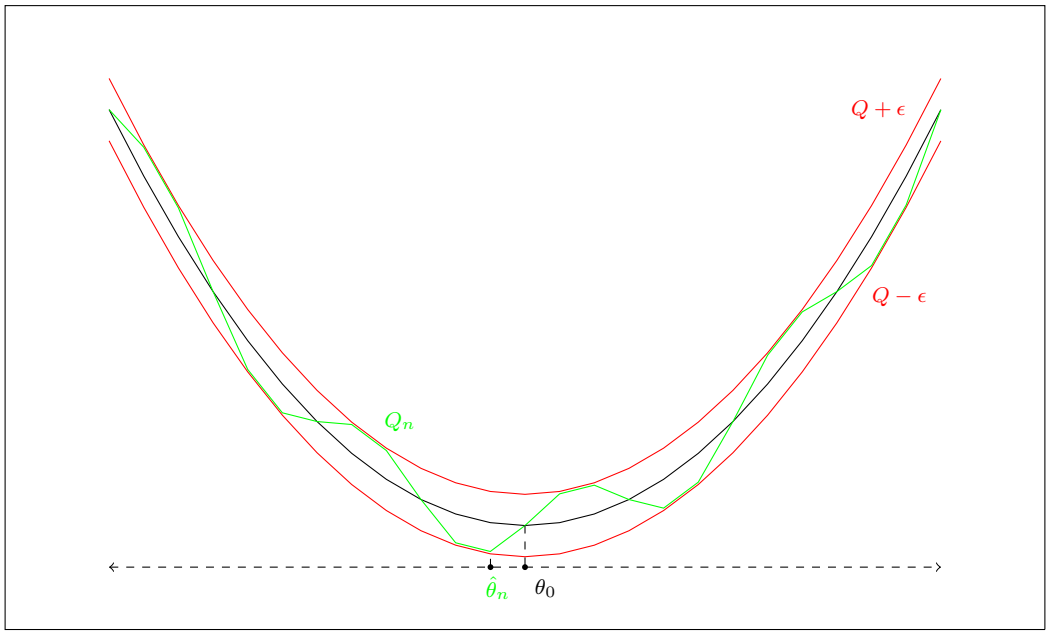

Figure 5.2: Theorem 5.2.1 Proof of Consistency

Hence, by definition of δ in (5.14) we can conclude that E_n, which implies (5.17), implies that $\hat{\theta}_n \in N_\epsilon$. Therefore, we can conclude that

$$\lim_{n\to\infty} P(\|\hat{\theta}_n - \theta_0\| < \epsilon) \geq \lim_{n\to\infty} P(E_n) = \lim_{n\to\infty} P(\sup_{\theta\in\Theta} |Q_n(\theta) - Q(\theta)| < \frac{\delta}{2}) = 1, \quad (5.18)$$

where the first equality follows by definition of E_n and the final equality is implied by Assumption EE-1(iii). Therefore, the claim of the Theorem follows from (5.18). ■

The proof of Theorem 5.2.1 is easy to understand graphically; see Figure 5.2. Consider drawing ϵ bands around Q. By Assumption EE-1(iii), the random function Q_n will be within those ϵ bands with probability tending to one. However, if Q_n is within those bands, it means that it must "follow" Q as it increases away from its minimizer θ_0. As a result the minimizer $\hat{\theta}_n$ of Q_n will not be far away from the minimizer θ_0 of Q provided Q_n is within the ϵ bands. In Figure 5.2 we draw this logic for one such realization of Q_n. More broadly, it is impossible to draw a Q_n that stays within the ϵ bands and whose minimum is far away from the minimum of Q (you should convince yourself of this!).

While Assumption EE-1(iii) is somewhat technical, there are multiple results in the literature that enable us to verify it. A key result in the *uniform law of large numbers*. Specifically, in problems such as MLE, the functions Q_n and Q have the structure

$$Q_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} m(W_i, \theta) \qquad Q(\theta) = E[m(W, \theta)] \quad (5.19)$$

where $\{W_i\}_{i=1}^n$ is an i.i.d. sample with $W_i \in \mathbf{R}^{d_w}$, $\theta \in \Theta$ and $m : \mathbf{R}^{d_w} \times \Theta \to \mathbf{R}$ is a known function. Notice that the law of large numbers then implies that

$$Q_n(\theta) \xrightarrow{p} Q(\theta). \tag{5.20}$$

for any $\theta \in \Theta$. However, as we discussed, Assumption EE-1(iii) is a stronger requirement than (5.20) in that it demands that the convergence be suitably uniform in θ; i.e.

$$\sup_{\theta \in \Theta} |Q_n(\theta) - Q(\theta)| = \sup_{\theta \in \Theta} |\frac{1}{n} \sum_{i=1}^n m(W_i, \theta) - E[m(W, \theta)]| \xrightarrow{p} 0 \tag{5.21}$$

The right hand side of (5.21) is known as the *uniform law of large numbers* in that the law of large number is required to hold *uniformly* in $\theta \in \Theta$.

Deriving conditions under which a uniform law of large numbers holds can be a challenging problem and the subject of an extensive literature; see van der Vaart and Wellner (1996). If you are planning to study Econometrics as a field, you will need to learn this literature in detail. However, for the purposes of this class, we will only present sufficient conditions for Assumption EE-1(iii) without proof in order to clarify Assumption EE-1(iii) can be easily verified in many applications of interest.

Lemma 5.2.1. *Let Q_n and Q be as in (5.19), $\{W_i\}_{i=1}^n$ be i.i.d., and suppose $m(w, \cdot) : \Theta \to \mathbf{R}$ is continuous at all θ, Θ is compact, and $E[\sup_{\theta \in \Theta} |m(W, \theta)|] < \infty$. Then*

$$\sup_{\theta \in \Theta} |Q_n(\theta) - Q(\theta)| \xrightarrow{p} 0.$$

PROOF: See Example 19.8 (and more generally Chapter 19) in van der Vaart (1998). ■

We conclude the discussion of consistency by returning to MLE.

Example 1.1.1 (cont). Recall that in this example we have an i.i.d. sample $\{W_i\}_{i=1}^n$ and we had posited W_i has density $f(\cdot, \theta_0)$ for some $\theta_0 \in \Theta$. In addition, we had

$$Q_n(\theta) \equiv -\frac{1}{n} \sum_{i=1}^n f(W_i, \theta) \qquad Q(\theta) = -E[f(W, \theta)]. \tag{5.22}$$

Notice Assumption EE-1(iii) can then be verified using Lemma 5.2.1 provided the density $f(w, \theta)$ is continuous in θ at all w, the density $f(w, \theta)$ is bounded uniformly in w and θ, and the parameter space Θ is compact. Assumption EE-1(ii) can be viewed as the definition of $\hat{\theta}_n$, and thus we need only verify Assumption EE-1(i). To this end, we let f_0 be the true density of W and consider two separate cases.

Case I: Suppose $f_0(W) = f(W, \theta_0)$ for some $\theta_0 \in \Theta$. In that case, we say the model is

properly specified. By Jensen's inequality, and using that $f(W,\theta_0) = f_0$ we then obtain

$$Q(\theta_0) - Q(\theta) = -E[\log(f(W,\theta_0))] + E[\log(f(W,\theta))] = E[\log(\frac{f(W,\theta)}{f(W,\theta_0)})]$$
$$\leq \log(E[\frac{f(W,\theta)}{f(W,\theta_0)}]) = \log(\int \frac{f(w,\theta)}{f(w,\theta_0)} f(w,\theta_0)dw) = \log(\int f(w,\theta)dw) = \log(1) = 0,$$

where we employed that $f(w,\theta)$ is a density even if $\theta \neq \theta_0$. If for all $\theta \in \Theta$ we have

$$P(f(W,\theta) \neq f(W,\theta_0)) > 0, \tag{5.23}$$

then Jensen's inequality holds strictly and therefore for all $\theta \neq \theta_0$ we obtain that $Q(\theta_0) - Q(\theta) < 0$. Hence, (5.23) holding for all $\theta \neq \theta_0$ suffices for verifying Assumption EE-1(i) when the model is properly specified.

Case II: Suppose that the model is *misspecified*, in that there is no $\theta \in \Theta$ such that $P(f(W,\theta) = f_0(W)) = 1$. To understand how to interpret θ_0 in that case, it is helpful to introduce the concept of *Kullback-Liebler Information Criterion* (KLIC) between two distributions f and $\tilde{f}$, which we denote by $\text{KLIC}(f,\tilde{f})$ and is given by

$$\text{KLIC}(f,\tilde{f}) = \int \log(\frac{f(w)}{\tilde{f}(w)}) f(w)dw. \tag{5.24}$$

Employing definition (5.24) and that $E[\log(f_0(W))]$ does not depend on θ, we obtain

$$\begin{aligned}\theta_0 &= \arg\min_{\theta\in\Theta} -E[\log(f(W,\theta))] \\ &= \arg\min_{\theta\in\Theta}\{-E[\log(f(W,\theta))] + E[\log(f_0(W))] - E[\log(f_0(W))]\} \\ &= \arg\min_{\theta\in\Theta} \text{KLIC}(f_0, f(\cdot,\theta)).\end{aligned} \tag{5.25}$$

In other words, under *misspecification*, θ_0 can be understood as the parameter whose corresponding density $f(\cdot,\theta_0)$ minimizes the distance (as measured by KLIC) to the true density f_0. Notice that unlike Case I, however, it is harder to provide conditions that ensure θ_0 is the unique minimizer of the KLIC (i.e. Assumption EE-1(i) holds). ■

5.3 Asymptotic Normality

Having established consistency, we next proceed to derive the asymptotic distribution of the extremum estimator $\hat{\theta}_n$. We will do so under the following conditions.

Assumption EE-2. *(i) θ_0 is an interior point of $\Theta \subseteq \mathbf{R}^d$; (ii) $Q_n : \Theta \to \mathbf{R}$ is twice continuously differentiable in θ; (iii) $\sqrt{n}\nabla_\theta Q_n(\theta_0) \xrightarrow{d} N(0,\Omega)$; (iv) $\sup_{\theta\in\Theta} \|\nabla^2_{\theta\theta'} Q_n(\theta) - B(\theta)\| \xrightarrow{p} 0$ for $B(\theta)$ a $d \times d$ continuous at θ_0 and $B(\theta_0)$ positive definite and invertible.*

Assumption EE-3. *(i)* $\hat{\theta}_n$ *is consistent for* θ_0*; (ii)* $\nabla_\theta Q_n(\hat{\theta}) = o_p(n^{-1/2})$.

Assumption EE-2 collects our assumptions on Q_n and Q, while Assumption EE-3 imposes conditions on $\hat{\theta}_n$. Some aspects of these Assumptions are worth discussing:

1. In Assumption EE-3(i) we directly impose that $\hat{\theta}_n$ is consistent for θ_0. We can of course rely on Theorem 5.2.1 to obtain sufficient conditions.

2. Superficially, Assumptions EE-2 and EE-3 do not appear to be employing the fact that $\hat{\theta}_n$ and θ_0 are defined as minimizers, but this is not the case. Consider, for example, Assumption EE-3. If $\hat{\theta}_n$ belongs to the interior of Θ, then the fact that it minimizes $Q_n : \Theta \to \mathbf{R}$ implies the first order condition

$$\nabla_\theta Q_n(\hat{\theta}_n) = 0,$$

which trivially verifies Assumption EE-3(ii). Relatedly, Assumption EE-2(iii) is often satisfied due to θ_0 being the minimizer of Q. Consider, for example, a setting such as MLE in which $Q : \Theta \to \mathbf{R}$ and $Q_n : \Theta \to \mathbf{R}$ have the structure

$$Q_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} m(W_i, \theta) \qquad Q(\theta) = E[m(W, \theta)] \tag{5.26}$$

Since θ_0 minimizes Q over Θ and θ_0 is an interior point of Θ by Assumption EE-2(i), θ_0 must satisfy the first order condition $E[\nabla_\theta m(W, \theta_0)] = 0$. By the central limit theorem we then obtain that for $\Omega = E[\nabla_\theta m(W, \theta_0)(\nabla_\theta m(W, \theta_0))']$ we have

$$\sqrt{n}\nabla_\theta Q_n(\theta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} m(W_i, \theta_0) \xrightarrow{d} N(0, \Omega).$$

3. The previous observation concerned the first order condition for θ_0. The second order condition is hidden in Assumption EE-2(iv). In particular, the requirement $B(\theta_0)$ be positive definite is reflecting that θ_0 is minimizer.

Given Assumptions EE-2 and EE-3 it is straightforward to obtain the asymptotic distribution of the extremum estimator $\hat{\theta}_n$. Surprisingly, the arguments are simpler than those employed to show consistency of extremum estimators in Theorem 5.2.1.

Theorem 5.3.1. *If Assumptions EE-2 and EE-3 hold, then it follows that*

$$\sqrt{n}\{\hat{\theta}_n - \theta_0\} \xrightarrow{d} N(0, B(\theta_0)^{-1}\Omega B(\theta_0)^{-1}).$$

PROOF: Somewhat unimaginatively we proceed by employing a Taylor expansion. In

particular applying the mean value theorem we obtain for some $\tilde{\theta}_n$ that

$$\nabla_\theta Q_n(\hat{\theta}_n) - \nabla_\theta Q_n(\theta_0) = \nabla^2_{\theta\theta'} Q_n(\tilde{\theta}_n)\{\hat{\theta}_n - \theta_0\}. \tag{5.27}$$

Notice that Assumptions EE-3(ii) controls $\nabla_\theta Q_n(\hat{\theta}_n)$ and Assumption EE-2(iii) controls $\nabla_\theta Q_n(\theta_0)$. Therefore, the only challenge is in studying the matrix $\nabla^2_{\theta\theta'} Q_n(\tilde{\theta}_n)$. To this end, we can apply the triangle inequality and Assumption EE-2(iv) to obtain

$$\begin{aligned} &|\nabla^2_{\theta\theta'} Q_n(\tilde{\theta}_n) - B(\theta_0)| \\ &\quad \leq \sup_{\theta\in\Theta} |\nabla^2_{\theta\theta'} Q_n(\theta) - B(\theta)| + |B(\tilde{\theta}_n) - B(\theta_0)| = |B(\tilde{\theta}_n) - B(\theta_0)| + o_p(1) \end{aligned} \tag{5.28}$$

Moreover, by the mean value theorem, $\tilde{\theta}_n$ is "between" $\hat{\theta}_n$ and θ_0. Therefore, $\hat{\theta}_n \xrightarrow{p} \theta_0$ by Assumption EE-3(i) implies that $\tilde{\theta}_n \xrightarrow{p} \theta_0$. By applying the continuous mapping theorem (since B is continuous at θ_0 by Assumption EE-2(iv)) we can then conclude $B(\tilde{\theta}_n) \xrightarrow{p} B(\theta_0)$. Hence, by employing result (5.28) we can conclude that

$$\nabla^2_{\theta\theta'} Q_n(\tilde{\theta}_n) \xrightarrow{p} B(\theta_0). \tag{5.29}$$

Since $B(\theta_0)$ is invertible by Assumption EE-2(iv), result (5.29) and the continuous mapping theorem implies that $\nabla^2_{\theta\theta'} Q_n(\tilde{\theta}_n)$ is invertible with probability tending to one and in addition $\{\nabla^2_{\theta\theta'} Q_n(\tilde{\theta}_n)\}^{-1} \xrightarrow{p} B(\theta_0)^{-1}$. Thus, returning to (5.27) we obtain

$$\begin{aligned} &\sqrt{n}\{\hat{\theta}_n - \theta_0\} \\ &\quad = \sqrt{n}\{\nabla_{\theta\theta'} Q_n(\tilde{\theta}_n)\}^{-1} \nabla_\theta Q_n(\hat{\theta}_n) - \sqrt{n}\{\nabla_{\theta\theta'} Q_n(\tilde{\theta}_n)\}^{-1} + o_p(1) && (5.30) \\ &\quad = \{B(\theta_0)^{-1} + o_p(1)\} \times o_p(1) - \{\nabla_{\theta\theta'} Q_n(\tilde{\theta}_n)\}^{-1} \sqrt{n} \nabla_\theta Q_n(\theta_0) + o_p(1) && (5.31) \\ &\quad \xrightarrow{d} N(0, B(\theta_0)^{-1} \Omega B(\theta_0)^{-1}), && (5.32) \end{aligned}$$

where the $o_p(1)$ term in (5.30) reflects $\nabla^2_{\theta\theta'} Q_n(\tilde{\theta}_n)$ is invertible with probability tending to one; Equality (5.31) follows from (5.29) and Assumption EE-2(iii); and result (5.32) follows from (5.29), Assumption EE-2(iii), and the continuous mapping theorem. The claim of the Theorem is therefore established. ■

Before returning to the discussion of MLE, we note that the theory of extremum estimation is substantially more general than it has been presented here. If you are interested, you should read Newey and McFadden (1994) for a general treatment of finite dimensional models. The analysis has also been extended to nonparametric and semiparametric models, in which θ is possible infinite dimensional (e.g. an unknown function); see, for example, van der Geer (2000) and Chen (2007).

Example 1.1.1 (cont). We return to maximum likelihood estimation, where recall

$$Q_n(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\log(f(W_i,\theta)) \qquad Q(\theta) = -E[\log(f(W,\theta))]. \tag{5.33}$$

In order to clarify the proof of Theorem 5.3.1, we revisit some of the arguments in the context of MLE. First, we apply the mean value theorem to obtain the expansion

$$\underbrace{-\frac{1}{n}\sum_{i=1}^{n}\nabla_\theta \log(f(W_i,\hat\theta_n))}_{\nabla_\theta Q_n(\hat\theta_n)} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}\log(f(W_i,\theta_0))}_{-\nabla_\theta Q_n(\theta_0)} = \underbrace{\{-\frac{1}{n}\sum_{i=1}^{n}\nabla_{\theta\theta'}\log(f(W_i,\tilde\theta_n))\}}_{\nabla^2_{\theta\theta'}Q_n(\tilde\theta_n)}\{\hat\theta_n-\theta_0\}$$

for some intermediate value $\tilde\theta_n$. Next note that since θ_0 is the minimizer of Q we obtain, under suitable regularity conditions, the first order condition

$$0 = \nabla_\theta Q(\theta_0) = \nabla_\theta E[\log(f(W,\theta_0))] = E[\nabla_\theta \log(f(W,\theta_0))]. \tag{5.34}$$

Therefore, provided $\Omega \equiv E[(\nabla_\theta \log(f(W,\theta_0)))(\nabla_\theta \log(f(W,\theta_0)))'] < \infty$, we can conclude

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\nabla_\theta \log(f(W_i,\theta_0)) \xrightarrow{d} N(0,\Omega), \tag{5.35}$$

which verifies Assumption EE-2(iii). Moreover, also note that Assumption EE-2(iv) can be verified by appealing to Lemma 5.2.1, in which case we obtain that

$$-\frac{1}{n}\sum_{i=1}^{n}\nabla^2_{\theta\theta'}\log(f(W_i,\tilde\theta_n)) \xrightarrow{p} -E[\nabla^2_{\theta\theta'}\log(f(W,\theta_0))] \equiv B(\theta_0). \tag{5.36}$$

Returning to result (5.3), we can combine results (5.35) and (5.36) to conclude that

$$\begin{aligned}\sqrt{n}\{\hat\theta_n-\theta_0\} &= \{-\frac{1}{n}\sum_{i=1}^{n}\nabla^2_{\theta\theta'}\log(f(W_i,\tilde\theta_n))\}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\nabla_\theta\log(f(W_i,\theta_0)) + o_p(1)\\ &= B(\theta_0)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\nabla_\theta\log(f(W_i,\theta_0)) + o_p(1) \xrightarrow{d} N(0, B(\theta_0)^{-1}\Omega B(\theta_0)^{-1})\end{aligned} \tag{5.37}$$

Notice that at no point in these derivations have we assumed that the model is properly specified (i.e. that the distribution of W_i indeed equals $f(\cdot,\theta_0)$). If we do have proper specification, however, then we can appeal to the information matrix equality:

$$B(\theta_0) \equiv -E[\nabla^2_{\theta\theta'}\log(f(W,\theta_0))] = E[\nabla_\theta\log(f(W,\theta_0))(\nabla_\theta\log(f(W,\theta_0)))'] = \Omega. \tag{5.38}$$

Hence, if the model is indeed correct, then the information matrix equality allows us to simplify the asymptotic variance to equal $B(\theta_0)^{-1}\Omega B(\theta_0)^{-1} = B(\theta_0)^{-1}$. In general,

however, we note that the information matrix equality in (5.38) can fail if the model is misspecified. The asymptotic variance $B(\theta_0)^{-1}\Omega B(\theta_0)^{-1}$ is in contrast robust to the model being misspecified. Its estimation is straightforward through sample analogues

$$\hat{\Omega} = \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta \log(f(W_i, \hat{\theta}_n))(\nabla_\theta \log(f(W_i, \hat{\theta}_n)))' \tag{5.39}$$

$$\hat{B}(\theta_0) = -\frac{1}{n}\sum_{i=1}^{n} \nabla^2_{\theta\theta'} \log(f(W_i, \hat{\theta}_n)) \tag{5.40}$$

and employing $\hat{B}(\theta_0)^{-1}\hat{\Omega}\hat{B}(\theta_0)^{-1}$ as an estimator of the asymptotic variance of the maximum likelihood estimator. ■

5.4 Problems

1. Let $X_i \in \mathbf{R}$ be continuously distributed with strictly increasing cdf F_X.

   (a) Show that if $\{X_i\}_{i=1}^n$ is an i.i.d. sample, then it follows that for any $c \in \mathbf{R}$

   $$\frac{1}{n}\sum_{i=1}^{n} 1\{X_i \leq c\} \xrightarrow{p} P(X_i \leq c).$$

   (b) Show for any $\eta > 0$ there are $c_0 \leq c_1 \leq \ldots \leq c_{K(\eta)}$ with $K(\eta) < \infty$, $c_0 = -\infty$ and $c_{K(\eta)} = +\infty$ such that for any $c \in \mathbf{R}$, there is $c_i \leq c \leq c_{i+1}$ satisfying

   $$|E[1\{X_i \leq c_i\} - 1\{X_i \leq c_{i+1}\}]| \leq \eta \ .$$

   (c) Using (a), (b), and that $1\{X \leq c\}$ is increasing in c show for any $\epsilon > 0$

   $$\lim_{n\to\infty} P\Big(\sup_{c\in\mathbf{R}} |\frac{1}{n}\sum_{i=1}^{n} 1\{X_i \leq c\} - P(X_i \leq c)| > \epsilon\Big) = 0.$$

2. For $Y \in \mathbf{R}$ and $X \in \mathbf{R}^{d_x}$, nonlinear least squares is given by the parametric model

   $$Y = g(X, \theta_0) + \epsilon \qquad E[\epsilon|X] = 0,$$

   where $(x, \theta) \mapsto g(x, \theta)$ is a parametric function defined up to an $\theta_0 \in \Theta \subset \mathbf{R}^k$.

   (a) For a known weight function $w(X) > 0$, define the criterion function Q^w by

   $$Q^w(\theta) = \frac{1}{2}E[(Y - g(X, \theta))^2 w(X)].$$

   Derive a condition under which θ_0 is the unique minimizer of Q_w on Θ.

(b) Let $Q_n^w : \Theta \to \mathbf{R}$ be the sample analogue to the criterion Q_n^w, given by

$$Q_n^w(\theta) = \frac{1}{2n}\sum_{i=1}^{n}(Y_i - g(X_i, \theta))^2 w(X_i).$$

Impose conditions on the function g and the parameter space Θ that enable you to show $\hat{\theta}_n = \arg\min_\Theta Q_n^w(\theta)$ is consistent for θ_0. Provide a proof.

(c) Provide additional assumptions on g, if you need them, to establish

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{L} N(0, \Sigma)$$

Provide a proof of this result and derive an explicit expression for Σ.

(d) Using your answer to part c) derive the optimal choice of $w(x)$. (i.e. what $w(x)$ minimizes the asymptotic variance). (Hint: If Ω is positive definite, then $(A'A)^{-1}A'\Omega A(A'A)^{-1}$ is positive semidefinite)

3. Suppose you have an i.i.d. sample $\{X_i\}_{i=1}^n$ with $X_i \sim N(\mu, \sigma^2)$ where the variance σ^2 is known but the mean μ is unknown. Recall the pdf of X_i is then $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$. Suppose the researcher knows that in fact $\mu \geq 0$.

(a) Write down the log-likelihood for this problem.

(b) What is the maximum likelihood estimator? Don't forget the researcher knows $\mu \geq 0$.

(c) For $\hat{\mu}$ the MLE estimator, derive the asymptotic distribution of $\sqrt{n}(\hat{\mu} - \mu)$ when $\mu > 0$.

(d) Derive the asymptotic distribution of $\sqrt{n}(\hat{\mu} - \mu)$ when in fact $\mu = 0$.

(e) In contrast to standard extremum estimation theory, in (d) $\sqrt{n}(\hat{\mu} - \mu)$ is not asymptotically normally distributed. What key assumption of extremum estimation is being violated?

4. Let $\{W_i\}_{i=1}^n$ be an i.i.d. sample, and suppose we posit that W has density $f(\cdot, \theta_0)$ for some $\theta_0 \in \Theta$. As argued, the asymptotic variance of the maximum likelihood estimator is given by $B(\theta_0)^{-1}\Omega B(\theta_0)^{-1}$, where $B(\theta_0)$ and Ω are given by

$$\begin{aligned} B(\theta_0) &= -E[\nabla^2_{\theta\theta'} \log(f(W, \theta_0))] \\ \Omega &= E[\nabla_\theta \log(f(W, \theta_0))(\nabla_\theta \log(f(W, \theta_0)))']. \end{aligned}$$

Using Lemma 5.2.1 provide conditions under which $\hat{B}(\theta_0)$ and $\hat{\Omega}$ as in (5.40) are consistent for $B(\theta_0)$ and Ω. Use such conditions to show that $\hat{B}(\theta_0)^{-1}\hat{\Omega}\hat{B}(\theta_0)^{-1}$ is a consistent estimator for the MLE asymptotic variance $B(\theta_0)^{-1}\Omega B(\theta_0)^{-1}$.

5. Suppose $\{W_i\}_{i=1}^n$ is an i.i.d. sample with $W \in \mathbf{R}^{d_w}$ and suppose that for some known function $g : \mathbf{R}^{d_w} \times \Theta \to \mathbf{R}^{d_g}$, the parameter of interest θ_0 is identified as the unique solution to

$$E[g(W, \theta_0)] = 0. \tag{5.41}$$

Throughout, assume θ_0 is indeed the unique $\theta \in \Theta$ for which (5.41) holds. The generalized method of moments estimator (GMM) of Hansen (1982) is then

$$\hat{\theta}_n = \arg\min_{\theta \in \Theta} (\frac{1}{n} \sum_{i=1}^n g(W_i, \theta))' \Omega (\frac{1}{n} \sum_{i=1}^n g(W_i, \theta))$$

where Ω is a $d_g \times d_g$ positive definite weighting matrix. In what follows we'll derive the asymptotic properties of the GMM estimator.

   (a) Provide conditions on g that let you verify Assumption EE-1(iii). Using these conditions, show $\hat{\theta}_n \xrightarrow{p} \theta_0$.

   (b) Provide conditions on g that let you verify Assumptions EE-2(iii) and EE-2(iv). Using these conditions, derive the asymptotic distribution of $\hat{\theta}_n$.

   (c) Propose a consistent estimator for the asymptotic variance of the GMM estimator (you do not need to formally show its consistency though).

6. Throughout this problem, let $Y \in \{0, 1\}$ be binary, $X \in \mathbf{R}^d$, and assume that

$$P(Y = 1|X) = F(X'\beta_0) \tag{5.42}$$

where $F : \mathbf{R} \to [0, 1]$ is the c.d.f. of a standard normal random variable and we know $\beta_0 \in \Theta \subset \mathbf{R}^d$ for some parameter space Θ.

   (a) Suppose that $E[XX']$ is full rank. Show that β_0 is the unique minimizer to

$$Q(\beta) \equiv E[(Y - F(X'\beta))^2].$$

   (b) Given an i.i.d. sample $\{Y_i, X_i\}_{i=1}^n$ define the function $Q_n : \mathbf{R}^d \to \mathbf{R}$ by

$$Q_n(\beta) \equiv \frac{1}{n} \sum_{i=1}^n (Y_i - F(X_i'\beta))^2$$

   and let $\hat{\beta}_n$ be the minimizer of Q_n over Θ. Using part (a) show that if $E[XX']$ is full rank and Θ is compact then $\hat{\beta}_n$ is consistent for β_0.

   (c) Let f denote the derivative of F. Verify that Assumption EE-2(iii) holds. What is Ω in Assumption EE-2(iii) in this example?

   (d) Assuming all the other conditions of Theorem 4.3.1 are satisfied, it follows that $\hat{\beta}_n$ is asymptotically normally distributed. Find an explicit formula for

the asymptotic variance of $\sqrt{n}\{\hat{\beta}_n - \beta_0\}$.

References

Abadie, A. (2005). Semiparametric difference-in-differences estimators. *The Review of Economic Studies*, **72** 1–19.

Abowd, J. M., Kramarz, F. and Margolis, D. N. (1999). High wage workers and high wage firms. *Econometrica*, **67** 251–333.

Abraham, S. and Sun, L. (2018). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Available at SSRN 3158747*.

Ackerberg, D. A., Caves, K. and Frazer, G. (2015). Identification properties of recent production function estimators. *Econometrica*, **83** 2411–2451.

Angrist, J. D. and Imbens, G. W. (1995). Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American statistical Association*, **90** 431–442.

Angrist, J. D. and Krueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics*, **106** 979–1014.

Angrist, J. D. and Pischke, J.-S. (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *The Journal of economic perspectives*, **24** 3–30.

Arellano, M. (2003). *Panel data econometrics*. Oxford university press.

Arellano, M. and Bond, S. (1991). Some tests of specification for panel data: Monte carlo evidence and an application to employment equations. *The review of economic studies*, **58** 277–297.

Arellano, M. and Bonhomme, S. (2011). Nonlinear panel data analysis.

Athey, S. and Imbens, G. W. (2006). Identification and inference in nonlinear difference-in-differences models. *Econometrica*, **74** 431–497.

Balke, A. and Pearl, J. (1994). Counterfactual probabilities: Computational methods, bounds and applications. In *Proceedings of the Tenth international conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 46–54.

Bekker, P. A. (1994). Alternative approximations to the distributions of instrumental variable estimators. *Econometrica: Journal of the Econometric Society* 657–681.

Berry, S., Levinsohn, J. and Pakes, A. (1995). Automobile prices in market equilibrium. *Econometrica: Journal of the Econometric Society* 841–890.

Berry, S. T. (1994). Estimating discrete-choice models of product differentiation. *The RAND Journal of Economics* 242–262.

Bertrand, M., Duflo, E. and Mullainathan, S. (2004). How much should we trust differences-in-differences estimates? *The Quarterly journal of economics*, **119** 249–275.

Bester, C. A., Conley, T. G. and Hansen, C. B. (2011). Inference with dependent data using cluster covariance estimators. *Journal of Econometrics*, **165** 137–151.

Blundell, R. and Bond, S. (1998). Initial conditions and moment restrictions in dynamic panel data models. *Journal of econometrics*, **87** 115–143.

Bogachev, V. I. (1998). *Gaussian Measures*. American Mathematical Society, Providence.

Bound, J., Jaeger, D. A. and Baker, R. M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American statistical association*, **90** 443–450.

Callaway, B. and Sant'Anna, P. H. (2019). Difference-in-differences with multiple time periods. *Available at SSRN 3148250*.

Cameron, A. C., Gelbach, J. B. and Miller, D. L. (2011). Robust inference with multiway clustering. *Journal of Business & Economic Statistics*, **29** 238–249.

Cameron, A. C. and Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. *Journal of Human Resources*, **50** 317–372.

Canay, I. A., Romano, J. P. and Shaikh, A. M. (2017). Randomization tests under an approximate symmetry assumption. *Econometrica*, **85** 1013–1030.

Canay, I. A., Santos, A. and Shaikh, A. M. (2018). The wild bootstrap with a" small" number of" large" clusters. Tech. rep., Centre for Microdata Methods and Practice, Institute for Fiscal Studies.

Card, D. (1993). Using geographic variation in college proximity to estimate the return to schooling. Tech. rep., National Bureau of Economic Research.

Card, D. (2001). Estimating the return to schooling: Progress on some persistent econometric problems. *Econometrica*, **69** 1127–1160.

Card, D. and Krueger, A. B. (1994). Minimum wages and employment: A case study of the fast-food industry in new jersey and pennsylvania. *The American Economic Review*, **84** 772–793.

Carter, A. V., Schnepel, K. T. and Steigerwald, D. G. (2017). Asymptotic behavior of at-test robust to cluster heterogeneity. *Review of Economics and Statistics*, **99** 698–709.

Cattaneo, M. D., Jansson, M. and Newey, W. K. (2018). Inference in linear regression models with many covariates and heteroscedasticity. *Journal of the American Statistical Association*, **113** 1350–1361.

Chalfin, A. and McCrary, J. (2013). Are us cities under-policed? theory and evidence. *NBER Working Paper*, **18815**.

Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. *Handbook of econometrics*, **6** 5549–5632.

Christiano, L. J. and Eichenbaum, M. (1992). Current real-business-cycle theories and aggregate labor-market fluctuations. *The American Economic Review* 430–450.

De Chaisemartin, C. and d'Haultfoeuille, X. (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review*, **110** 2964–96.

Deaton, A. S. (2009). Instruments of development: Randomization in the tropics, and the search for the elusive keys to economic development. Tech. rep., National Bureau of Economic Research.

Duflo, E., Dupas, P. and Kremer, M. (2008). Peer effects and the impact of tracking: Evidence from a randomized evaluation in kenya.

Eichenbaum, M. S., Hansen, L. P. and Singleton, K. J. (1988). A time series analysis of representative agent models of consumption and leisure choice under uncertainty. *The Quarterly Journal of Economics*, **103** 51–78.

Fehr, E. and Goette, L. (2007). Do workers work more if wages are high? evidence from a randomized field experiment. *The American Economic Review*, **97** 298–317.

Gandhi, A., Navarro, S. and Rivers, D. A. (2011). On the identification of production functions: How heterogeneous is productivity?

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society* 1029–1054.

Hansen, L. P. and Singleton, K. J. (1982). Generalized instrumental variables estimation of nonlinear rational expectations models. *Econometrica: Journal of the Econometric Society* 1269–1286.

Heckman, J. J. and Urzua, S. (2010). Comparing iv with structural models: What simple iv can and cannot identify. *Journal of Econometrics*, **156** 27–37.

Heckman, J. J. and Vytlacil, E. (2005). Structural equations, treatment effects, and econometric policy evaluation. *Econometrica*, **73** 669–738.

Hersch, J. (1998). Compensating differentials for gender-specific job injury risks. *The American Economic Review*, **88** 598–607.

Hoch, I. (1962). Estimation of production function parameters combining time-series and cross-section data. *Econometrica: journal of the Econometric Society* 34–53.

Ibragimov, R. and Müller, U. K. (2010). t-statistic based correlation and heterogeneity robust inference. *Journal of Business & Economic Statistics*, **28** 453–468.

Ibragimov, R. and Müller, U. K. (2016). Inference with few heterogeneous clusters. *Review of Economics and Statistics*, **98** 83–96.

Imbens, G. W. (2010). Better late than nothing: Some comments on deaton (2009) and heckman and urzua (2009). *Journal of Economic literature*, **48** 399–423.

Imbens, G. W. and Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, **62** 467–475.

Keane, M. P. (2010). A structural perspective on the experimentalist school. *The Journal of Economic Perspectives*, **24** 47–58.

Kitagawa, T. (2015). A test for instrument validity. *Econometrica*, **83** 2043–2063.

Kleibergen, F. (2005). Testing parameters in gmm without assuming that they are identified. *Econometrica*, **73** 1103–1123.

Knight, K. (1999). Epi-convergence in distribution and stochastic equi-semicontinuity. *Unpublished manuscript*, **37** 28–29.

Lehmann, E. and Romano, J. (2005). *Testing Statistical Hypotheses*. Springer Verlag.

Luenberger, D. (1969). *Optimization by Vector Space Methods*. John Wiley & Sons.

MacKinnon, J. G. and White, H. (1985). Some heteroskedasticity consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics* 305–325.

Matthews, R. (2000). Storks deliver babies (p= 0.008). *Teaching Statistics*, **22** 36–38.

Menzel, K. (2017). Bootstrap with clustering in two or more dimensions. *arXiv preprint arXiv:1703.03043*.

Moreira, M. J. (2003). A conditional likelihood ratio test for structural models. *Econometrica*, **71** 1027–1048.

Moulton, B. R. (1986). Random group effects and the precision of regression estimates. *Journal of econometrics*, **32** 385–397.

Mundlak, Y. and Hoch, I. (1965). Consequences of alternative specifications in estimation of cobb-douglas production functions. *Econometrica: Journal of the Econometric Society* 814–828.

Muralidharan, K. and Sundararaman, V. (2011). Teacher performance pay: Experimental evidence from india. *Journal of political Economy*, **119** 39–77.

Nevo, A. and Whinston, M. D. (2010). Taking the dogma out of econometrics: Structural modeling and credible inference. *The Journal of Economic Perspectives*, **24** 69–81.

Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of econometrics*, **4** 2111–2245.

Neyman, J. and Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica: Journal of the Econometric Society* 1–32.

Olley, G. S. and Pakes, A. (1996). The dynamics of productivity in the telecommunications equipment industry. *Econometrica*, **64** 1263–1297.

Powell, J. L., Stock, J. H. and Stoker, T. M. (1989). Semiparametric estimation of index coefficients. *Econometrica: Journal of the Econometric Society* 1403–1430.

Rothenberg, T. J. (1984). Approximating the distributions of econometric estimators and test statistics. *Handbook of econometrics*, **2** 881–935.

Sargan, J. D. (1958). The estimation of economic relationships using instrumental variables. *Econometrica: Journal of the Econometric Society* 393–415.

Sims, C. A. (2010). But economics is not an experimental science. *The Journal of Economic Perspectives*, **24** 59–68.

Staiger, D. and Stock, J. H. (1997). Instrumental variables regression with weak instruments. *Econometrica*, **65** 557–586.

Stock, J. H., Wright, J. H. and Yogo, M. (2002). A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business & Economic Statistics*, **20** 518–529.

Stock, J. H. and Yogo, M. (2002). Testing for weak instruments in linear iv regression.

van der Geer, S. A. (2000). *Empirical Processes in M-estimation*, vol. 6. Cambridge university press.

van der Vaart, A. (1998). *Asymptotic Statistics*. Cambridge University Press.

van der Vaart, A. and Wellner, J. (1996). *Weak Convergence and Empirical Processes*. Springer Verlag.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the Econometric Society* 1–25.

Yitzhaki, S. (1996). On using linear regressions in welfare economics. *Journal of Business & Economic Statisttics* 478–486.