Empirical Analysis I
Fall 2017

(Azeem Shaikh)

Notes by Takuma Habu

University of Chicago

Contents

1 Large sample theory

In econometrics, there are three typical problems we wish to address.

Suppose $X_1, X_2, \ldots, X_n$ are independently and identically distributed according to the cumulative distribution function P. Then, our goal is to "learn" some "features" of P from the data.

(i) Estimate $\theta(P)$. Construct a function called an *estimator*, $\hat{\theta}_n = \hat{\theta}_n(X_1, X_2, \ldots, X_n)$, that provides a "best guess" for $\theta(P)$. For example, $\mu(P) = \mathbb{E}[X_i]$, $\sigma^2(P) = \mathrm{Var}_P[X_i]$, coefficients from the regressions ($\beta(P)$).

(ii) Test a hypothesis about $\theta(P)$. For example, a test could be of the form $\theta(P) = \theta_0$, and we want to construct a function, $\phi_n = \phi_n(X_1, X_2, \ldots, X_n) \in [0,1]$, that determines the probability with which you reject the hypothesis.

(iii) Construct a confidence region for $\theta(P)$. Construct a random set, $C_n = C_n(X_1, X_2, \ldots, X_n)$, that contains $\theta(P)$ with some pre-specified probability.

Studying finite sample properties of (i) and (iii) is typically difficult (unless we make some strong assumptions about P). This is why we wish to study large-sample properties.

1.1 Existence of moments

Definition 1.1. We say that the mean of a random vector X, denoted $\mathbb{E}[X]$, exists if $\mathbb{E}[|X|] < \infty$.

If X is a random variable on $\mathbb{R}$, then we can split $\mathbb{E}[X]$ into positive and negative parts:

$$\mathbb{E}[X] = \mathbb{E}\left[X^+\right] - \mathbb{E}\left[X^-\right],$$

where

$$X^+ \coloneqq \max\{X, 0\}, \; X^- \coloneqq \max\{-X, 0\}.$$

Then, notice that

$$\mathbb{E}[|X|] = \mathbb{E}\left[X^+\right] + \mathbb{E}\left[X^-\right].$$

Thus, the requirement that $\mathbb{E}[|X_i|] < \infty$ is equivalent to requiring that $\mathbb{E}\left[X_i^+\right], \mathbb{E}\left[X_i^-\right] < \infty$ (i.e. they both exist) since $\mathbb{E}\left[X_i^+\right], \mathbb{E}\left[X_i^-\right] \geq 0$ by construction.

1.1.1 Jensen's inequality

The Jensen's inequality is another useful inequality.

Lemma 1.1. *(Jensen's Inequality) Let $I \subseteq \mathbb{R}$ be a convex set (i.e. an interval on $\mathbb{R}$) and $f : I \to \mathbb{R}$ a convex function. Then, for any random variable X such that $\mathbb{P}(x \in I) = 1$, $\mathbb{E}[|X|] < \infty$ and $\mathbb{E}[|f(X)|] < \infty$, then*

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)].$$

If f is a concave function (i.e. $-f$ is convex), then the inequality above is reversed.

Proof. Recall that f is (weakly) convex if, for any $x, y \in I$ and any $\alpha \in [0,1]$,

$$f(\alpha x + (1-\alpha) y) \leq \alpha f(x) + (1-\alpha) f(y).$$

Let $c \coloneqq \mathbb{E}[X]$. If $c \notin \mathrm{int}(I)$, then it must be that all of the probability mass of X is at c—otherwise, the expected value would not be at the boundary. In other words,

$$c \notin \mathrm{int}(I) \Rightarrow \mathbb{P}(X = c) = 1.$$

In this case, the expectations operator essentially does nothing so that

$$\mathbb{E}\left[f\left(x\right)\right] = f\left(c\right) \geq f\left(c\right) = f\left(\mathbb{E}\left[x\right]\right);$$

i.e. the inequality holds trivially.

Suppose instead that $c \in \text{int}\left(I\right)$. Define

$$\Delta_{+,h(c)} \coloneqq \frac{f\left(c+h\right) - f\left(c\right)}{h},\ \Delta_{-,h(c)} \coloneqq \frac{f\left(c\right) - f\left(c-h\right)}{h}.$$

We claim that $\Delta_{+,h(c)}$ is decreasing as $h \downarrow 0$ and that $\Delta_{-,h(c)}$ is increasing as $h \downarrow 0$. To see this, first note that, given $z, x \in I$ with $z > x$, we can express any point $y \in (x, z)$ as a convex combination of x and z:

$$y = \lambda x + \left(1-\lambda\right) z,\ \lambda = \frac{z-y}{z-x} \in \left(0,1\right),$$

where we note that

$$\lambda + \left(1-\lambda\right) = \frac{z-y}{z-x} + \left(1 - \frac{z-y}{z-x}\right) = \frac{z-y}{z-x} + \frac{y-x}{z-x} = 1.$$

Since f is convex and $\lambda = \left(z-y\right)/\left(z-x\right) \in (0,1)$,

$$\begin{aligned} f\left(y\right) &\leq \frac{z-y}{z-x} f\left(x\right) + \left(1 - \frac{z-y}{z-x}\right) f\left(z\right) \\ &= \frac{z-y}{z-x}\left(f\left(x\right) - f\left(z\right)\right) + f\left(z\right) \\ \Leftrightarrow \frac{f\left(y\right) - f\left(z\right)}{z-y} &\leq \frac{f\left(x\right) - f\left(z\right)}{z-x} \\ \Leftrightarrow \frac{f\left(z\right) - f\left(x\right)}{z-x} &\leq \frac{f\left(z\right) - f\left(y\right)}{z-y} \end{aligned} \tag{1.1}$$

Similarly, since $\left(z-y\right)/\left(z-x\right) = 1 - \left(y-x\right)/\left(z-x\right)$, we can also obtain

$$\begin{aligned} f\left(y\right) &\leq \left(1 - \frac{y-x}{z-x}\right) f\left(x\right) + \frac{y-x}{z-x} f\left(z\right) \\ &= \frac{y-x}{z-x}\left(f\left(z\right) - f\left(x\right)\right) + f\left(x\right) \\ \Leftrightarrow \frac{f\left(y\right) - f\left(x\right)}{y-x} &\leq \frac{f\left(z\right) - f\left(x\right)}{z-x}. \end{aligned} \tag{1.2}$$

Combining (1.1) and (1.2), for any $x, y, z \in I$ with $x \leq y \leq z$, we obtain the *Chordal Slope Lemma*:

$$\frac{f\left(y\right) - f\left(x\right)}{y-x} \leq \frac{f\left(z\right) - f\left(x\right)}{z-x} \leq \frac{f\left(z\right) - f\left(y\right)}{z-y}. \tag{1.3}$$

Let $z = c + 2h$, $y = c + h$ and $x = c$, with $h > 0$ such that $x, y, z \in I$, then, from the first inequality of (1.3), we have

$$\frac{f\left(c+h\right) - f\left(c\right)}{h} \leq \frac{f\left(c+2h\right) - f\left(c\right)}{2h}.$$

Hence, $\Delta_{+,h(c)}$ is increasing in h so that $\Delta_{+,h(c)}$ is decreasing as $h \downarrow 0$. Similarly, setting $z = c$, $y = c - h$ and $x = c - 2h$, and using the second inequality of (1.3), we have

$$\frac{f\left(c\right) - f\left(c-2h\right)}{2h} \leq \frac{f\left(c\right) - f\left(c-h\right)}{h}$$

so that $\Delta_{-,h(c)}$ is increasing as $h \downarrow 0$. These imply that $\Delta_{-,h(c)}$ and $\Delta_{+,h(c)}$ are monotone functions of h. Hence, their (one-sided) limits exist. Define

Don't we also need Δ's to be bounded?

$$D_+(c) \coloneqq \lim_{h \downarrow 0} \Delta_{+,h(c)},\ D_-(c) \coloneqq \lim_{h \downarrow 0} \Delta_{-,h(c)}.$$

Letting $z = c + h$, $y = c$ and $x = c - h$ in (1.3) gives

$$\Delta_{-,h(c)} = \frac{f(c) - f(c-h)}{h} \leq \frac{f(c+h) - f(c-h)}{2h} \leq \frac{f(c+h) - f(c)}{h} = \Delta_{+,h(c)}.$$

Above implies that $D_+(c) > -\infty$ (since it is bounded below by $\Delta_{-,h(c)}$) and that $D_-(c) < \infty$ (since it is bounded above by $\Delta_{+,h(c)}$).

Choose $m \in [D_-(c), D_+(c)]$ and define

$$L(x) \coloneqq f(c) + m(x - c).$$

We now claim that $L(x) \leq f(x)$ for all $x \in I$. We consider three cases: $x = c$, $x > c$ and $x < c$.

(i) Suppose $x = c$, then the the inequality holds trivially.

(ii) Suppose now that $x = c + h > c$, then

$$\begin{aligned} m \leq D_+(c) &\leq \Delta_{+,h(c)} \\ &= \frac{f(c+h) - f(c)}{h} = \frac{f(x) - f(c)}{x - c} \\ \Leftrightarrow m(x - c) &\leq f(x) - f(c) \\ \Leftrightarrow f(x) &\geq f(c) + m(x - c) = L(x). \end{aligned}$$

(iii) Suppose instead that $x = c - h < c$, then

$$\begin{aligned} m \geq D_-(c) &\geq \Delta_{-,h(c)} \\ &= \frac{f(c) - f(c-h)}{h} = \frac{f(c) - f(x)}{c - x} \\ \Leftrightarrow f(x) &\geq f(c) + m(x - c) = L(x). \end{aligned}$$

Thus, in all cases, we have that $L(x) \leq f(x)$.

Taking expectations of both sides, we obtain

$$\mathbb{E}[L(X)] \leq \mathbb{E}[f(X)].$$

Since L is an affine function by construction and $\mathbb{E}$ operator is linear—i.e. $\mathbb{E}[L(X)] = L(\mathbb{E}(X))$—we realise that

$$\begin{aligned} L(\mathbb{E}(X)) = \mathbb{E}[L(X)] &\leq \mathbb{E}[f(X)] \\ \Leftrightarrow f(c) + m(\mathbb{E}(X) - c) &\leq \mathbb{E}[f(X)]. \end{aligned}$$

Since $c = \mathbb{E}[x]$,

$$\begin{aligned} f(\mathbb{E}[X]) + m(\mathbb{E}(X) - \mathbb{E}[X]) &\leq \mathbb{E}[f(X)] \\ \Leftrightarrow f(\mathbb{E}[X]) &\leq \mathbb{E}[f(X)]. \end{aligned}$$

To show that the inequality is reversed when f is concave, we use the fact that when f is concave then $-f$ is convex. Thus,

$$\begin{aligned} -f(\mathbb{E}[X]) &\leq \mathbb{E}[-f(X)] \\ \Leftrightarrow f(\mathbb{E}[X]) &\geq \mathbb{E}[f(X)]. \end{aligned}$$

■

Remark. The proof is long but we can break it down into the following steps:

(i) Set $\mathbb{E}[X] = c$.

(ii) Prove the Chordal Slope Lemma—easy to see why the inequalities hold if you draw (see Figure 1.1); if $x \leq y \leq z$,

$$\frac{f(y) - f(x)}{y - x} \leq \frac{f(z) - f(x)}{z - x} \leq \frac{f(z) - f(y)}{z - y}.$$

The trick here is to remember that

$$f(y) = \frac{z - y}{z - x} f(x) + \frac{y - x}{z - x} f(z).$$

(iii) Show that $\Delta_{+,h(c)}$ and $\Delta_{-,h(c)}$ are monotone (use $x = c$, $y = c + h$, $z = c + 2h$ for $\Delta_{+,h(c)}$ and $x = c - 2h$, $y = c - h$, $z = c$ for $\Delta_{-,h(c)}$). This implies that their limits, $D_+(c)$ and $D_-(c)$, as $h \downarrow 0$ exist. Show, using Chordal Slope Lemma, that $\Delta_{-,h(c)} \leq \Delta_{+,h(c)}$ ($x = c - h$, $y = c$, $z = c + h$). This implies that the limits are bounded above and below respectively.

(iv) Choose $m \in [D_-(c), D_+(c)]$, and define $L(x) \coloneqq f(c) + m(x - c)$. Show that $f(x) \geq L(x)$ for all $x \in I$.

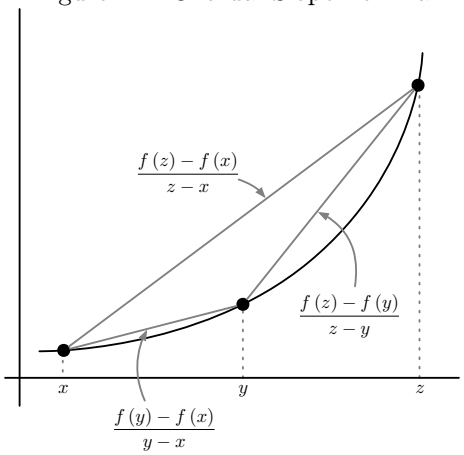
(a) Case 1: $x = c$. Holds with equality.

(b) Case 2: $x > c$. Set $x = c + h$ and then consider the left-most inequality of the Chordal Slope Lemma while using the fact that $\Delta_{+,h(c)} \geq D_+(c) \geq m$.

(c) Case 3: $x < c$. Set $x = c - h$ and then consider the right-most inequality of the Chordal Slope Lemma while using the fact that $\Delta_{-,h(c)} \leq D_-(c) \leq m$.

(v) Take expectation of $f(x) \geq L(x)$ and use the fact that $\mathbb{E}[L(x)] = L(\mathbb{E}[x])$ (since L is linear) and that $\mathbb{E}[x] = c$.

Figure 1.1: Chordal Slope Lemma



Remark. The proof shows that if I is open, then $f(x) = \sup_{L \in \mathcal{L}} L(x)$, where

$$L \coloneqq \{L \in \mathcal{L} : L(x) \leq f(x) \text{ and } L(x) \text{ is affine}\}.$$

In fact, it suffices to take

$$\mathcal{L} \coloneqq \{f(c) + D_+(c)(x - c), \forall x \in I\}$$

Note that the open interval assumption is essential as, without it, we can construct a counterexample with a function with a large jump at the boundary. In addition, if $D_+(c)$ is sufficiently continuous, we could replace I with countable dense subset of I.

1.1.2 Existence of higher moments imply existence of lower moments

We now use the Jensen's Inequality to prove the fact that existence of higher moments imply existence of lower moments. We start with some definitions.

Definition 1.2. *(Raw and centred moments)* Let X be a random variable:

- ▷ the kth *raw moment* of X is given by $\mathbb{E}[X^k]$;
- ▷ the kth *centred moment* of X is given by $\mathbb{E}[(X - \mathbb{E}[X])^k]$.

Proposition 1.1. *(Existence of higher moments imply existence of lower moments) Let X be a random vector. Then,*

$$\mathbb{E}\left[|X|^k\right] < \infty \Rightarrow \mathbb{E}\left[|X|^j\right] < \infty, \ \forall k \geq j \geq 1.$$

Proof. Let $k \geq j \geq 1$ and define

$$\begin{aligned} f(x) &\coloneqq x^{\frac{k}{j}}, \\ Y &\coloneqq |X|^j, \\ Y_n &\coloneqq \min\left\{|X|^j, n\right\}, \end{aligned}$$

where $n \in \mathbb{N}$. We want to use Jensen's inequality on Y_n and $f(Y_n)$. We first need to show that: $\mathbb{E}[|Y_n|] < \infty$ and $\mathbb{E}[|f(Y_n)|] < \infty$.

- ▷ $\mathbb{E}[|Y_n|] < \infty$: This follows from the fact that Y_n is bounded between 0 and n.
- ▷ $\mathbb{E}[|f(Y_n)|] < \infty$: Since f is an increasing function and $f(Y_n) \geq 0$ so that $f(Y_n) = |f(Y_n)|$:

$$\begin{aligned} & Y_n \leq Y \\ \Rightarrow\ & f(Y_n) \leq f(Y) \\ & |f(Y_n)| \leq f(Y) \\ \Rightarrow\ & \mathbb{E}[|f(Y_n)|] \leq \mathbb{E}[f(Y)] \\ & = \mathbb{E}\left[\left(|X|^j\right)^{\frac{k}{j}}\right] = \mathbb{E}\left[|X|^k\right] < \infty, \end{aligned} \tag{1.4}$$

where the last inequality is given by assumption.

Given $k \geq j \geq 1$, $f(x)$ is convex:

$$f'(x) = \frac{k}{j} x^{\frac{k}{j}-1} \geq 0, \ f''(x) = \left(\frac{k}{j} - 1\right) \frac{k}{j} x^{\frac{k}{j}-2}.$$

We can use Jensen's Inequality to obtain that:

$$f(\mathbb{E}[Y_n]) \leq \mathbb{E}[f(Y_n)].$$

But

$$\begin{aligned} f(\mathbb{E}[Y_n]) &\leq \mathbb{E}[f(Y_n)] = \mathbb{E}[|f(Y_n)|] \\ &\leq \mathbb{E}\left[|X|^k\right] < \infty \end{aligned}$$

where the second line follows from (1.4). Then,

$$\begin{aligned} f\left(\mathbb{E}\left[Y_n\right]\right) &\leq \mathbb{E}\left[|X|^k\right] \\ \Rightarrow \left(\mathbb{E}\left[Y_n\right]\right)^{\frac{k}{j}} &\leq \mathbb{E}\left[|X|^k\right] \\ \Rightarrow \mathbb{E}\left[Y_n\right] &\leq \left(\mathbb{E}\left[|X|^k\right]\right)^{\frac{j}{k}}. \end{aligned}$$

Since $\mathbb{E}[|X|^k] < \infty$, and $j/k \leq 1$, $(\mathbb{E}[|X|^k])^{\frac{j}{k}} \leq \mathbb{E}[|X|^k] < \infty$ so that $\mathbb{E}[Y_n]$ is bounded above. Together with the fact that $Y_{n+1} \geq Y_n \geq 0$ and $Y_n \to Y$, by the Monotone Convergence Theorem,[1]

$$\infty > \lim_{n\to\infty}\left(\mathbb{E}\left[|X|^k\right]\right)^{\frac{j}{k}} = \left(\mathbb{E}\left[|X|^k\right]\right)^{\frac{j}{k}} \geq \lim_{n\to\infty}\mathbb{E}\left[Y_n\right] = \mathbb{E}\left[\lim_{n\to\infty} Y_n\right] = \mathbb{E}\left[Y\right] = \mathbb{E}\left[|X|^j\right]. \quad \blacksquare$$

1.2 Convergence in probability

Definition 1.3. *(Convergence in Probability).* Let $\{X_n : n \geq 1\}$ and X be random vectors on $\mathbb{R}^k$. X_n is said to *converge in probability* to X, denoted $X_n \xrightarrow{\text{p}} X$, if, as $n \to \infty$,

$$\mathbb{P}\left(|X_n - X| > \varepsilon\right) \to 0, \ \forall \varepsilon > 0,$$

where $|\cdot|$ is the usual Euclidean norm. Equivalently,

- ▷ $\lim_{n\to\infty}\mathbb{P}\left(|X_n - X| > \varepsilon\right) = 0$ for any $\varepsilon > 0$;
- ▷ for any $\delta, \varepsilon > 0$, there exists $N \in \mathbb{N}$ such that, $\mathbb{P}\left(|X_n - X| > \varepsilon\right) \leq \delta$ for all $n \geq N$.

In the special case when X_n are random variables (i.e. $k = 1$), we say that a sequence of random variables $\{X_n : n \geq 1\}$ converges in probability to $+\infty$ if, for any $\varepsilon > 0$, as $n \to \infty$, $\mathbb{P}\left(X_n > \varepsilon\right) \to 1$. In such a case, we write $X_n \xrightarrow{\text{p}} +\infty$. Similarly, we say that $\{X_n : n \geq 1\}$ converges in probability to $-\infty$ if, for any $\varepsilon > 0$, as $n \to \infty$, $\mathbb{P}\left(X_n > \varepsilon\right) \to 0$, and write $X_n \xrightarrow{\text{p}} -\infty$.

1.2.1 Markov's inequality

Lemma 1.2. *(Markov's inequality). For any random variable X,*

$$\mathbb{P}\left(|X| > \varepsilon\right) \leq \frac{\mathbb{E}\left[|X|^q\right]}{\varepsilon^q}, \ \forall q, \varepsilon > 0,$$

where $|\cdot|$ is the usual Euclidean norm.

Proof. We want to use the fact that if $g\left(x\right) \leq f\left(x\right) \Rightarrow \mathbb{E}\left[g\left(x\right)\right] \leq \mathbb{E}\left(f\left(x\right)\right)$.[2] Notice that we can express the left-hand side using the indicator function:

$$\mathbb{P}\left(|X| > \varepsilon\right) = \mathbb{E}\left[\mathbf{1}_{\{|X|>\varepsilon\}}\right].$$

[1]

Theorem 1.1. (Monotone Convergence Theorem) *For any sequence of nonnegative measurable functions $\{f_n\}$ such that $0 \leq f_n \leq f_{n+1}$ for all $n \in \mathbb{N}$, and that $\{f_n\}$ converges point-wisely (Lebesgue) almost everywhere to some integrable function $f : X \to \mathbb{R}$, f is integrable and*

$$\lim_{n\to\infty}\int_X f_n d\lambda = \int_X f d\lambda.$$

[2] This property follows from the monotonicity property of the integral.

Thus, it is sufficient for us to show that the following holds:

$$\mathbf{1}_{\{|X|>\varepsilon\}} \leq \frac{|X|^q}{\varepsilon^q}. \tag{1.5}$$

By construction, the indicator function takes on values 0 or 1. If its value is zero (which is the case when $|X| \leq \varepsilon$), since $|X| \geq 0$ and $\varepsilon > 0$, the right-hand side is (weakly) larger than zero so that the inequality holds. If, instead, the value of the indicator function is 1, which is the case when $|X| > \varepsilon$, then

$$\frac{|X|}{\varepsilon} > 1 \Rightarrow \frac{|X|^q}{\varepsilon^q} > 1, \forall q > 0.$$

Thus, equation (1.5) holds in general. Taking expectation of both sides of equation (1.5) gives

$$\mathbb{E}\left[\mathbf{1}_{\{|X|>\varepsilon\}}\right] = \mathbb{P}\left(|X| > \varepsilon\right) \leq \mathbb{E}\left[\frac{|X|^q}{\varepsilon^q}\right] = \frac{\mathbb{E}\left[|X|^q\right]}{\varepsilon^q}. \qquad \blacksquare$$

Markov's inequality is useful in many proofs involving convergence in probability. We can also use it to construct confidence intervals.

Example 1.1. *(Constructing confidence sets with Markov's inequality).* Suppose that $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} X$, where X has the CDF $P = \text{Bernoulli}\,(q)$ with $q \in (0,1)$. The mean and the variance of a Bernoulli distribution are given by $\mu(P) = q$ and $\sigma^2(P) = q(1-q)$ respectively.[3] Let $\alpha \in (0,1)$. We would like to construct a (random) set $C_n \coloneqq C_n(X_1, X_2, \ldots, X_n)$ such that the probability that the C_n contains the true mean is greater than $1-\alpha$;[4] i.e.

$$\mathbb{P}\left(\mu(P) \in C_n\right) \geq 1 - \alpha.$$

C_n would then be the *confidence set* of level $1-\alpha$ for $\mu(P)$. For example, if $\alpha = 5\%$, then C_n gives us the interval in which the the probability that C_n contains the true mean is 95%.

Define the sample mean as

$$\bar{X}_n \coloneqq \frac{1}{n}\left(\sum_{i=1}^{n} X_i\right).$$

We can use the Markov's inequality to construct this set (later, we will use the Central Limit Theorem). Letting $q = 2$, we obtain:

$$\begin{aligned}
\mathbb{P}\left(\left|\bar{X}_n - \mu(P)\right| > \varepsilon\right) &\leq \frac{\mathbb{E}\left[\left(\bar{X}_n - \mu(P)\right)^2\right]}{\varepsilon^2} \\
&= \frac{\operatorname{Var}\left[\bar{X}_n\right]}{\varepsilon^2} = \frac{\frac{1}{n^2}\operatorname{Var}\left[\sum_{i=1}^n X\right]}{\varepsilon^2} = \frac{\frac{1}{n^2}\left(\sum_{i=1}^n \operatorname{Var}[X] + 2\sum_{i\neq j}\operatorname{Cov}\left[X_i, X_j\right]\right)}{\varepsilon^2} \\
&= \frac{1}{n^2}\frac{n\operatorname{Var}(X)}{\varepsilon^2} = \frac{q(1-q)}{n\varepsilon^2},
\end{aligned}$$

where we used the fact that X_i's are identically distributed (i.e. $\operatorname{Var}[X_i] = \operatorname{Var}[X]$ for all i) and independently distributed (i.e. $\operatorname{Cov}[X_i, X_j] = 0$ for all $i \neq j$).

[3] Let Y be a random variable on $\mathbb{R}$ with some distribution. A Bernoulli distribution is then constructed by setting $X = \mathbf{1}_{\{Y_i \leq x\}}$.Then,

$$\begin{aligned}
\mathbb{E}[X] &= \mathbb{E}\left[\mathbf{1}_{\{Y_i\leq x\}}\right] = \mathbb{P}\left(Y_i \leq x\right) = q, \\
\operatorname{Var}[X] &= \mathbb{E}\left[\mathbf{1}^2_{\{Y_i\leq x\}}\right] - \mathbb{E}\left[\mathbf{1}_{\{Y_i\leq x\}}\right]^2 \\
&= \mathbb{E}\left[\mathbf{1}_{\{Y_i\leq x\}}\right] - \mathbb{E}\left[\mathbf{1}_{\{Y_i\leq x\}}\right]^2 = q(1-q),
\end{aligned}$$

where we used the fact that $\mathbf{1}^2_{\{Y_i\leq x\}} = \mathbf{1}_{\{Y_i\leq x\}}$.

[4] Of course if set set $C_n = (0,1)$, then this would be true but we will ignore such a trivial set.

Note that $q(1-q)$ is maximised at $1/4$ when $q = 1/2$. Thus,

$$\mathbb{P}\left(\left|\bar{X}_n - \mu(P)\right| > \varepsilon\right) \leq \frac{q(1-q)}{n\varepsilon^2} \leq \frac{1}{4n\varepsilon^2}, \ \forall q.$$

Since $\mathbb{P}(X > x) = 1 - \mathbb{P}(X \leq x)$, we can write

$$\mathbb{P}\left(\left|\bar{X}_n - \mu(P)\right| \leq \varepsilon\right) \geq 1 - \frac{1}{4n\varepsilon^2}.$$

We can then choose $\bar{\varepsilon}$ such that

$$1 - \frac{1}{4n\bar{\varepsilon}^2} = 1 - \alpha \Leftrightarrow \bar{\varepsilon} = \frac{1}{\sqrt{4\alpha n}}.$$

Then, by construction,

$$\mathbb{P}\left(\left|\bar{X}_n - \mu(P)\right| \leq \bar{\varepsilon}\right) \geq 1 - \alpha.$$

We can then define the confidence set as

$$C_n \coloneqq \left[\bar{X}_n - \bar{\varepsilon}, \bar{X}_n + \bar{\varepsilon}\right].$$

Equivalently, we can write the confidence set as

$$C_n \coloneqq \left\{x \in \mathbb{R} : \left|\bar{X}_n - x\right| \leq \bar{\varepsilon}\right\}.$$

Note that

$$\begin{aligned}\mathbb{P}(\mu(P) \in C_n) &= \mathbb{P}\left(\bar{X}_n - \bar{\varepsilon} \leq \mu(P) \leq \bar{X}_n + \bar{\varepsilon}\right) \\ &= \mathbb{P}\left(\left|\bar{X}_n - \mu(P)\right| \leq \bar{\varepsilon}\right) \geq 1 - \alpha.\end{aligned}$$

1.2.2 Weak Law of Large Numbers (WLLN)

We now introduce a useful result called the Weak Law of Large Numbers which tells us that with iid random variables, the sample mean converges in probability to the population mean.

Definition 1.4. *(iid).* Let $\{X_n : n \geq 1\}$ be random variables on $\mathbb{R}$ with CDF P. The sequence of random variables $\{X_n : n \geq 1\}$ is said to be independently distributed if

$$\mathbb{P}(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n) = \prod_{i=1}^{n} \mathbb{P}(X_i \leq x_i).$$

Moreover, we say that $\{X_n : n \geq 1\}$ is identically distributed if

$$\mathbb{P}(X_i \leq x) = \mathbb{P}(X_j \leq x), \ \forall i, j \in \{1, 2, \ldots n\}.$$

If $\{X_n : n \geq 1\}$ satisfies both of these properties, the sequence is said to be *independently and identically distributed* (iid).

Theorem 1.2. *(Weak Law of Large Numbers)*[a] *Let $\{X_n : n \geq 1\}$ be a sequence of iid random variables on $\mathbb{R}$ with CDF P. Suppose that $\mu(P)$ exists, then*

$$\bar{X}_n \xrightarrow{\text{p}} \mu(P).$$

[a] Strong law of large numbers is also called *almost sure* convergence; i.e. $\mathbb{P}\left(\lim_{n\to\infty} \bar{X}_n = \mu\right) = 1$, written $\bar{X}_n \xrightarrow{\text{a.s.}} \mu(P)$. The proof is more complex than that for the weak law of large numbers.

Proof. To simplify the proof, we assume that $\sigma^2(P) < \infty$. Given the definition of convergence in probability, we wish to show that, as $n \to \infty$,

$$\mathbb{P}\left(\left|\bar{X}_n - \mu(P)\right| > \varepsilon\right) \to 0, \ \forall \varepsilon > 0.$$

By Markov's inequality, we know that

$$\mathbb{P}\left(\left|\bar{X}_n - \mu(P)\right| > \varepsilon\right) \leq \frac{\mathbb{E}\left[\left|\bar{X}_n - \mu(P)\right|^q\right]}{\varepsilon^q}. \tag{1.6}$$

Let $q = 2$. Consider the numerator on the right-hand side:

$$\begin{aligned}
\mathbb{E}\left[\left|\bar{X}_n - \mu(P)\right|^2\right] &= \mathbb{E}\left[\left(\bar{X}_n - \mathbb{E}[X]\right)^2\right] \\
\text{(definition)} &= \operatorname{Var}\left(\bar{X}_n\right) \\
\text{(independence)} &= \frac{1}{n^2}\operatorname{Var}\left(\sum_{i=1}^{n} X_i\right) \\
\text{(identically distributed)} &= \frac{1}{n}\operatorname{Var}(X) = \frac{\sigma^2(P)}{n}.
\end{aligned}$$

Substituting the expression into (1.6) gives

$$\mathbb{P}\left(\left|\bar{X}_n - \mu(P)\right| > \varepsilon\right) \leq \frac{1}{n}\left(\frac{\sigma^2(P)}{\varepsilon^2}\right).$$

As $n \to \infty$, the right-hand side tends to zero so that the left-hand side must also tend to zero. ■

1.2.3 Convergence in marginal and joint probabilities

The following proposition says that convergences in marginal probabilities imply convergence in joint probability. The proposition generalises the Weak Law of Large Numbers from one-dimensional random variables to k-dimensional random vectors.

Specifically, suppose $\{X_n : n \geq 1\}$, $\{Y_n : n \geq 1\}$, X and Y are random variables and that $X_n \xrightarrow{\text{p}} X$ and $Y_n \xrightarrow{\text{p}} Y$. Then, the proposition tells us that

$$(X_n, Y_n) \xrightarrow{\text{p}} (X, Y).$$

This is a very useful property of convergence in probability especially when combined with the Continuous Mapping Theorem.

Before proceeding, we can write a random vector X_i on $\mathbb{R}^k$ as

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

where $X_{i,j}$ denotes the jth element of the random vector X_i.

The following Lemma are useful for the proof. The following is an extension of the idea that if $x + y > 1$, then either x and/or y must be greater than $1/2$.

Lemma 1.3. *Let $X = (X_1, X_2, \ldots, X_k)$ be random vector on $\mathbb{R}^k$, then*

$$\mathbb{P}\left(\sum_{j=1}^{k} |X_j| > \varepsilon\right) \leq \mathbb{P}\left(\bigcup_{j=1}^{k}\left\{|X_j| > \frac{\varepsilon}{k}\right\}\right).$$

Proof. Define

$$\begin{aligned}
A &= \left\{X \in \mathbb{R}^k : \sum_{j=1}^{k} |X_j| > \varepsilon\right\}, \\
B_j &= \left\{X \in \mathbb{R}^k : |X_j| > \frac{\varepsilon}{k}\right\}, \\
B &= \bigcup_{j=1}^{k} B_j.
\end{aligned}$$

We want to show that $A \subseteq B$ which would imply the desired inequality. Suppose, by way of contradiction, that $X \in A$ but $X \notin B$. Then, it must be that

$$|X_j| \leq \frac{\varepsilon}{k}, \forall j = 1, 2, \ldots, k \Rightarrow \sum_{j=1}^{k} |X_j| \leq \sum_{j=1}^{k} \frac{\varepsilon}{k} = \varepsilon.$$

This contradicts the assumption that $X \in A$. Hence, $A \subseteq B$. ■

The following is a formalisation of the idea that if we have two events, then the probability of either of the two events happening (i.e. the union of the two events) must be less than the sum of probability of individual events (because the events may be overlapping) (draw a Venn diagram!).

Lemma. *(Boole's Inequality). For a countable set of events $B_1, B_2, \ldots, B_k$,*

$$\mathbb{P}\left(\bigcup_{j=1}^{k} B_j\right) \leq \sum_{j=1}^{k} \mathbb{P}\left(B_j\right).$$

Proof. Admitted. ■

Proposition 1.2. *(Convergences in marginal probabilities imply convergence in joint probability). Let $\{X_n : n \geq 1\}$ and X be random vectors on $\mathbb{R}^k$. Let $X_{n,j}$ denote the jth element of sequence X_n. Then,*

$$X_{n,j} \xrightarrow{\text{p}} X_j, \ \forall j = 1, 2, \ldots, k \Rightarrow X_n \xrightarrow{\text{p}} X.$$

Proof. Fix $\varepsilon > 0$. We want to show that $\mathbb{P}\left(|X_n - X| > \varepsilon\right) = 0$. Since $|\cdot|$ is the Euclidean norm,

$$\begin{aligned}
\mathbb{P}\left(|X_n - X| > \varepsilon\right) &= \mathbb{P}\left(\left[\sum_{j=1}^{k} \left(X_{n,j} - X_j\right)^2\right]^{\frac{1}{2}} > \varepsilon\right) \\
&= \mathbb{P}\left(\sum_{j=1}^{k} \left(X_{n,j} - X_j\right)^2 > \varepsilon^2\right) \\
\text{(lemma above)} &\leq \mathbb{P}\left(\bigcup_{j=1}^{k} \left\{\left(X_{n,j} - X_j\right)^2 > \frac{\varepsilon^2}{k}\right\}\right) \\
\text{(Boole's Inequality)} &\leq \sum_{j=1}^{k} \mathbb{P}\left(\left(X_{n,j} - X_j\right)^2 > \frac{\varepsilon^2}{k}\right) \\
&= \sum_{j=1}^{k} \mathbb{P}\left(|X_{n,j} - X_j| > \frac{\varepsilon}{\sqrt{k}}\right).
\end{aligned}$$

Since $X_{n,j} \xrightarrow{\text{p}} X_j$, then, for any $\varepsilon > 0$,

$$\mathbb{P}\left(|X_{n,j} - X_j| > \frac{\varepsilon}{\sqrt{k}}\right) \to 0, \ \forall j = 1, 2, \ldots, k$$

so that $\mathbb{P}\left(|X_n - X| > \varepsilon\right) \to 0$ as $n \to \infty$. ■

1.2.4 Continuous Mapping Theorem for convergence in probability

Recall that if $\{X_n : n \geq 1\}$, $\{Y_n : n \geq 1\}$, X and Y are random variables and that $X_n \xrightarrow{\text{p}} X$ and $Y_n \xrightarrow{\text{p}} Y$, then

$$(X_n, Y_n) \xrightarrow{\text{p}} (X, Y).$$

The Continuous Mapping Theorem then allows us to define a function $g(X_n, Y_n)$ so that

$$g(X_n, Y_n) \xrightarrow{\text{p}} g(X, Y).$$

For example, g could be $X_n + Y_n$, $X_n Y_n$, X_n / Y_n (if $Y \neq 0$) etc. We will use this property many times in proving consistency.

Theorem 1.3. *(Continuous Mapping Theorem for Convergence in Probability) Let $\{X_n : n \geq 1\}$ and X be random vectors on $\mathbb{R}^k$. Suppose that $g : \mathbb{R}^k \to \mathbb{R}^d$ is continuous at each point in the set $C \subseteq \mathbb{R}^k$ such that $\mathbb{P}(X \in C) = 1$. Then,*

$$X_n \xrightarrow{\text{p}} X \Rightarrow g(X_n) \xrightarrow{\text{p}} g(X).$$

Proof. Recall that g is said to be continuous at a point $x \in C$ if, for any $\varepsilon > 0$, there exists $\delta > 0$ such that $|x - y| < \delta \Rightarrow |g(x) - g(y)| \leq \varepsilon$. Fix $\varepsilon > 0$. We want to show that $\mathbb{P}(|g(X_n) - g(X)| > \varepsilon) \to 0$ as $n \to \infty$. The circumstances in which the convergence fails is if X_n and X are "close" but $g(X_n)$ and $g(X)$ are "far". Define such points into a "problematic set" denoted B_δ; i.e. for $\delta > 0$,

$$B_\delta \coloneqq \left\{x \in \mathbb{R}^k : \exists y \in \mathbb{R}^k \text{ } s.t. \text{ } |x - y| < \delta \text{ and } |g(x) - g(y)| > \varepsilon\right\}.$$

Then, we can rewrite the expression we want to show as

$$\begin{aligned}\mathbb{P}(|g(X_n) - g(X)| > \varepsilon) = \mathbb{P}(\{|g(X_n) - g(X)| > \varepsilon\} \cap \{X \notin B_\delta\}) \\ + \mathbb{P}(\{|g(X_n) - g(X)| > \varepsilon\} \cap \{X \in B_\delta\}).\end{aligned} \tag{1.7}$$

Let us consider the first term of the right-hand side of (1.7). If $x \notin B_\delta$, then, for all $y \in \mathbb{R}^k$, either $|x - y| \geq \delta$ or $|g(x) - g(y)| \leq \varepsilon$. Setting $x = X$ and $y = X_n$, if $X \notin B_\delta$, then either $|X - X_n| = |X_n - X| \geq \delta$ or $|g(X) - g(X_n)| = |g(X_n) - g(X)| \leq \varepsilon$. Hence,

$$\begin{aligned}\{|g(X_n) - g(X)| > \varepsilon\} \cap \{X \notin B_\delta\} &= \{|g(X_n) - g(X)| > \varepsilon\} \\ &\quad \cap \{\{|X_n - X| \geq \delta\} \cup \{|g(X_n) - g(X)| \leq \varepsilon\}\} \\ &= \{|g(X_n) - g(X)| > \varepsilon\} \cap \{|X_n - X| \geq \delta\} \\ &\quad \cup \underbrace{\{|g(X_n) - g(X)| > \varepsilon\} \cap \{|g(X_n) - g(X)| \leq \varepsilon\}}_{=\emptyset} \\ &= \{|g(X_n) - g(X)| > \varepsilon\} \cap \{|X_n - X| \geq \delta\}\end{aligned}$$

Then,

$$\begin{aligned}\mathbb{P}(\{|g(X_n) - g(X)| > \varepsilon\} \cap \{X \notin B_\delta\}) &= \mathbb{P}(\{|g(X_n) - g(X)| > \varepsilon\} \cap \{|X_n - X| \geq \delta\}) \\ &\leq \mathbb{P}(|X_n - X| \geq \delta) \leq \mathbb{P}\left(|X_n - X| > \frac{\delta}{2}\right).\end{aligned}$$

Now consider the second term of the right-hand side of (1.7). By construction of B_δ,

$$\begin{aligned}\mathbb{P}(\{|g(X_n) - g(X)| > \varepsilon\} \cap \{X \in B_\delta\}) &= \mathbb{P}(X \in B_\delta) \\ &= \mathbb{P}(X \in \{B_\delta \cap C\}) \because \mathbb{P}(X \in C) = 1.\end{aligned}$$

Together, we have

$$\mathbb{P}\left(\left|g\left(X_n\right)-g\left(X\right)\right|>\varepsilon\right)\leq\mathbb{P}\left(\left|X_n-X\right|>\frac{\delta}{2}\right)+\mathbb{P}\left(X\in\left\{B_\delta\cap C\right\}\right).$$

Note that $\mathbb{P}\left(X\in\left\{B_\delta\cap C\right\}\right)\to 0$ as $\delta\downarrow 0$ since $X\in C$ so that we may choose δ sufficiently small so that $X\notin B_\delta$ (by continuity).[5] Thus, if $X_n\xrightarrow{\mathrm{p}}X$, then $\mathbb{P}\left(\left|X_n-X\right|\geq\delta/2\right)\to 0$ as $n\to 0$ so that the right-hand side tends to zero, which implies that the left-hand side must also tend to zero. That is, $g\left(X_n\right)\xrightarrow{\mathrm{p}}g\left(X\right)$. ■

Remark 1.1. Suppose the function g_n point-wisely converges to g. This does not imply that $g_n\left(X_n\right)\xrightarrow{\mathrm{p}}g\left(X\right)$. To see this, let $X_n\coloneqq n^{-1}\xrightarrow{\mathrm{p}}0=:X$ so that $X_n\xrightarrow{\mathrm{p}}X$. Let

$$g_n\left(x\right)=g\left(x\right)=\begin{cases}1 & \text{if } x>0\\ 0 & \text{if } x\leq 0\end{cases}.$$

Then, $g_n\left(x\right)\to g\left(x\right)$ for all $x\in\mathbb{R}$ although neither functions are continuous everywhere (there is a "jump" at $x=0$). Notice that since $X_n>1$ for any finite n,

$$g\left(X_n\right)=1.$$

However, in the limit, $g\left(X\right)=g\left(0\right)=0$ so that $g_n\left(X_n\right)\xnrightarrow{\mathrm{p}}g\left(X\right)$.

1.2.5 Consistency of estimators

Let $X_1,X_2,\ldots X_n\overset{\text{iid}}{\sim}P$ on $\mathbb{R}$ and suppose that $\mu\left(P\right)$ exists. We wish to construct an estimator of $\mu\left(P\right)$. A natural estimator for $\mu\left(P\right)$ is $\bar{X}_n$ since, by WLLN,

$$\bar{X}_n\xrightarrow{\mathrm{p}}\mu\left(P\right);$$

i.e. $\bar{X}_n$ is *consistent* for $\mu\left(P\right)$. Note that:

- ▷ this is an application of the *analog principle* to construct estimators, where we construct an estimand for the unknown P first, and use this to construct an estimator of desired property;
- ▷ consistency requires $\bar{X}_n$ to converge to the true/population value of the distribution, $\mu\left(P\right)$. Thus, although $\bar{X}_n+1$ would converge in probability to a constant, it is not consistent for $\mu\left(P\right)$.
- ▷ consistency is related, but not equivalent to unbiasedness. Unbiasedness requires that

  $$\mathbb{E}\left[\bar{X}_n\right]=\mu\left(P\right);$$

  i.e. it is not an asymptotic property. Note that unbiasedness does not imply consistency, nor does consistency imply unbiasedness.

Example 1.2. *(Unbiasedness and consistency)* Let $X_1,X_2,\ldots,X_n$ be iid random variables, and $\mathbb{E}\left[\left|X_i\right|\right]<\infty$. Then, recall that, by WLLN, $\bar{X}_n\xrightarrow{\mathrm{p}}\mathbb{E}\left[X_i\right]$. Moreover, $\bar{X}_n$ is unbiased since

$$\mathbb{E}\left[\bar{X}_n\right]=\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}X_i\right]=\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[X_i\right]=\mathbb{E}\left[X_i\right].$$

[5] Suppose that $g\left(x\right)$ jumps up at some point $\bar{x}$ by amount ε. Then, $\left|g\left(\bar{x}\right)-g\left(y\right)\right|\geq\varepsilon$ however we close we choose y to $\bar{x}$ since "gap" does not shrink. In other words, there is $\delta>0$ such that $\left|x-y\right|<\delta\Rightarrow\left|g\left(x\right)-g\left(y\right)\right|\leq\varepsilon$.

Notice that $\bar{X}_n + 1/n$ is a consistent but biased estimator since

$$\begin{aligned}\mathbb{E}\left[\bar{X}_n + \frac{1}{n}\right] &= \mathbb{E}\left[X_i\right] + \frac{1}{n} \neq \mathbb{E}\left[X_i\right] \\ \left(\bar{X}_n, \frac{1}{n}\right) &\xrightarrow{\text{p}} \left(\mathbb{E}\left[X_i\right], 0\right), \\ \Rightarrow \bar{X}_n + \frac{1}{n} &\xrightarrow{\text{p}} \mathbb{E}\left[X_i\right] + 0 = \mathbb{E}\left[X_i\right].\end{aligned}$$

Notice also that X_1 is an unbiased but inconsistent estimator since $\mathbb{E}\left[X_1\right] = \mathbb{E}\left[X_i\right]$ but X_1 does not converge in probability to $\mathbb{E}\left[X_i\right]$ as $n \to \infty$ (since X_1 is not affected by n).

Example 1.3. *(Empirical CDF)* In an iid setting, a natural estimator of P is the distribution that puts equal mass on $X_1, X_2, \ldots, X_n$. We denote such an estimator $\hat{P}_n$ and refer to it as the *empirical distribution* of P. The CDF corresponding to the empirical distribution is:

$$\hat{F}_n\left(x\right) \coloneqq \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{X_i \leq x\}}.$$

Since X_i's are iid (so that $\mathbf{1}_{\{X_i \leq x\}}$ are also iid) and $\mathbb{E}[\mathbf{1}_{\{X_i \leq x\}}] = \mathbb{P}(X_i \leq x) = F\left(x\right) \in [0,1] < \infty$. Then, by WLLN, for any $x \in \mathbb{R}$,

$$\hat{F}_n\left(x\right) \xrightarrow{\text{p}} F\left(x\right).$$

That is, the CDF of the empirical distribution converges point-wisely in probability to the CDF of the actual distribution so that $\hat{F}_n\left(x\right)$ is consistent for $F\left(x\right)$. In fact, it can further be shown that $\hat{F}_n\left(x\right)$ converges uniformly to $F\left(x\right)$ (Glivenko–Cantelli Theorem);[6] i.e.

$$\sup_{x \in \mathbb{R}} \left|\hat{F}_n\left(x\right) - F\left(x\right)\right| \xrightarrow{\text{p}} 0.$$

In other words, the CDF of the empirical distribution is a good estimator of the CDF of the actual distribution.

Example 1.4. *(Constructing an estimator of $\sigma^2\left(P\right)$)* Let $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} P$ on $\mathbb{R}$. Suppose that $\mathbb{E}[X_i^2] < \infty$. We wish to construct an estimator of $\sigma^2\left(P\right) \coloneqq \text{Var}\left(X_i\right)$. A natural estimator is

$$s_n^2 \coloneqq \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2.$$

Let's check if this is a consistent estimator of $\sigma^2\left(P\right)$. We can write

$$\begin{aligned}s_n^2 &= \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i^2 - 2X_i\bar{X}_n + \bar{X}_n^2\right) \\ &= \frac{1}{n-1}\left[\left(\sum_{i=1}^{n} X_i^2\right) - 2\bar{X}_n\left(\sum_{i=1}^{n} X_i\right) + \left(\sum_{i=1}^{n} \bar{X}_n^2\right)\right] \\ &= \frac{1}{n-1}\left[\left(\sum_{i=1}^{n} X_i^2\right) - 2n\bar{X}_n^2 + n\bar{X}_n^2\right] \\ &= \frac{n}{n-1}\left[\frac{1}{n}\left(\sum_{i=1}^{n} X_i^2\right) - \bar{X}_n^2\right].\end{aligned}$$

Thus, we may write s_n^2 as a function g, parameterised as

$$s_n^2 = g\left(\frac{n}{n-1}, \frac{1}{n}\left(\sum_{i=1}^{n} X_i^2\right), \bar{X}_n\right),$$

[6] Proof is long!

where $g(a,b,c) = a\left(b - c^2\right)$.

As $n \to \infty$,

$$\begin{aligned} \frac{n}{n-1} &\to 1, \\ \frac{1}{n}\sum_{i=1}^{n} X_i^2 &\xrightarrow{\text{p}} \mathbb{E}\left[X_i^2\right], \\ \bar{X}_n &\xrightarrow{\text{p}} \mathbb{E}\left[X\right]. \end{aligned}$$

(The second follows by WLLN since X_i^2's are iid and are (assumed to be) finite.) Since we are given that $\mathbb{E}[X_i^2] < \infty$, this also implies that $\mathbb{E}\left[|X|\right] < \infty$ so that $\mathbb{E}\left[X\right]$ exists (Proposition 1.1). Hence, by Proposition 1.2,

$$\left(\frac{n}{n-1}, \frac{1}{n}\sum_{i=1}^{n} X_i^2, \bar{X}_n\right) \xrightarrow{\text{p}} \left(1, \mathbb{E}\left[X_i^2\right], \mathbb{E}\left[X\right]\right).$$

Since g is continuous (in particular, it is continuous at the limit), by the Continuous Mapping Theorem,

$$s_n^2 \xrightarrow{\text{p}} \mathbb{E}\left[X_i^2\right] - \mathbb{E}\left[X\right]^2 = \sigma^2(P).$$

This shows that s_n^2 is a consistent estimator of $\sigma^2(P)$.

Remark. To show that s_n^2 is an unbiased estimator of $\sigma^2(P)$; i.e. $\mathbb{E}\left[s_n^2\right] = \sigma^2(P)$.

$$\mathbb{E}\left[s_n^2\right] = \frac{1}{n-1}\left(n\mathbb{E}\left[X^2\right] - n\mathbb{E}\left[\bar{X}_n^2\right]\right).$$

Recall that, for any random variable Y, $\text{Var}(Y) = \mathbb{E}\left[Y^2\right] - \mathbb{E}\left[Y\right]^2 \Leftrightarrow \mathbb{E}\left[Y^2\right] = \text{Var}\left[Y\right] + \mathbb{E}\left[Y\right]^2$. Applying this to $\mathbb{E}[\bar{X}_n^2]$, we have

$$\begin{aligned} \mathbb{E}\left[\bar{X}_n^2\right] &= \text{Var}\left[\bar{X}_n\right] + \mathbb{E}\left[\bar{X}_n\right]^2 \\ &= \text{Var}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] + \mathbb{E}\left[\bar{X}_n^2\right] \\ &= \frac{1}{n}\sigma^2(P) + \mu(P)^2. \end{aligned}$$

Similarly,

$$\begin{aligned} \mathbb{E}\left[X^2\right] &= \text{Var}\left[X\right] + \mathbb{E}\left[X\right]^2 \\ &= \sigma^2(P) + \mu(P)^2. \end{aligned}$$

Therefore,

$$\begin{aligned} \mathbb{E}\left[s_n^2\right] &= \frac{1}{n-1}\left[n\left(\sigma^2(P) + \mu(P)^2\right) - n\left(\frac{1}{n}\sigma^2(P) + \mu(P)^2\right)\right] \\ &= \frac{1}{n-1}(n-1)\sigma^2(P) \\ &= \sigma^2(P). \end{aligned}$$

Note that

$$\bar{s}_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2$$

would be a consistent but biased estimator of $\sigma^2(P)$.

We can also show that a similar estimator for the variance in multivariate case is consistent.

Example 1.5. Let $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} P = X$ on $\mathbb{R}^k$. Suppose that $\mathbb{E}[XX'] < \infty$. We wish to construct an estimator of $\Sigma(P)_{k\times k} \coloneqq \text{Var}[X]$. A natural estimator is

$$\hat{\Sigma}_n = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)\left(X_i - \bar{X}_n\right)'.$$

We wish to show that $\hat{\Sigma}_n \xrightarrow{\text{p}} \Sigma(P)$. Write

$$\begin{aligned}
\hat{\Sigma}_n =& \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)\left(X_i - \bar{X}_n\right)' \\
=& \frac{1}{n-1}\sum_{i=1}^{n}\left[\left(X_i - \bar{X}_n\right)X_i' - \left(X_i - \bar{X}_n\right)\bar{X}_n'\right] \\
=& \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)X_i' - \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)\bar{X}_n' \\
=& \frac{1}{n-1}\left[\sum_{i=1}^{n}X_iX_i' - \sum_{i=1}^{n}\left(\bar{X}_nX_i'\right)\right] - \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i\bar{X}_n' - \bar{X}_n\bar{X}_n'\right) \\
=& \frac{1}{n-1}\left[\sum_{i=1}^{n}X_iX_i' - \bar{X}_n\left(\sum_{i=1}^{n}X_i\right)'\right] - \frac{1}{n-1}\left(\sum_{i=1}^{n}X_i\bar{X}_n' - \sum_{i=1}^{n}\bar{X}_n\bar{X}_n'\right) \\
=& \frac{n}{n-1}\left[\frac{1}{n}\left(\sum_{i=1}^{n}X_iX_i'\right) - \bar{X}_n\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)'\right] - \frac{n}{n-1}\left(\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)\bar{X}_n' - \bar{X}_n\bar{X}_n'\right) \\
=& \frac{n}{n-1}\left[\left(\frac{1}{n}\sum_{i=1}^{n}X_iX_i'\right) - \bar{X}_n\bar{X}_n'\right].
\end{aligned}$$

We can express this as a continuous function g,

$$\hat{\Sigma}_n = g\left(\frac{n}{n-1}, \frac{1}{n}\sum_{i=1}^{n}X_iX_i', \bar{X}_n\right),$$

where $g(a,b,c) = a(b-c)$. As $n \to \infty$,

$$\begin{aligned}
\frac{n}{n-1} &\to 1, \\
\frac{1}{n}\sum_{i=1}^{n}X_iX_i' &\xrightarrow{\text{p}} \mathbb{E}[XX'], \\
\bar{X}_n &\xrightarrow{\text{p}} \mathbb{E}[X],
\end{aligned}$$

where the last two convergence in probabilities follow via WLLN. Since we are given that $\Sigma(P) < \infty$, which also implies that $\mathbb{E}[|X|] < \infty$, we can use Proposition 1.2 to obtain that

$$\left(\frac{n}{n-1}, \frac{1}{n}\sum_{i=1}^{n}X_iX_i', \bar{X}_n\right) \xrightarrow{\text{p}} \left(1, \mathbb{E}[XX'], \mathbb{E}[X]\right).$$

Since g is continuous, by the Continuous Mapping Theorem,

$$\hat{\Sigma}_n \xrightarrow{\text{p}} \mathbb{E}[XX'] - \mathbb{E}[X]\mathbb{E}[X] = \Sigma(P).$$

1.3 Convergence in moments

Definition 1.5. *(Convergence in qth moment)* Let $\{X_n : n \geq 1\}$ and X be random vectors on $\mathbb{R}^k$. X_n is said to *converge in qth moment to X such that $q \geq 1$*, denoted $X_n \xrightarrow{\text{L}^q} X$, if, as $n \to \infty$,

$$\mathbb{E}\left[|X_n - X|^q\right] \to 0.$$

The following shows that convergence in moments implies convergence in probability.

Proposition 1.3. *(Convergence in moments implies convergence in probability) Let $\{X_n : n \geq 1\}$ and X be random vectors on $\mathbb{R}^k$. Then,*

$$X_n \xrightarrow{\text{L}^q} X \Rightarrow X_n \xrightarrow{\text{P}} X.$$

Proof. (Proposition 1.3) Immediate from Markov's inequality, which states that

$$\mathbb{P}\left(|X_n - X| > \varepsilon\right) \leq \frac{\mathbb{E}\left[|X_n - X|^q\right]}{\varepsilon^q}.$$

Thus, if $\mathbb{E}\left[|X_n - X|^q\right] \to 0$, then $X_n \xrightarrow{\text{P}} X$. ■

Example 1.6. *(Converse to Proposition 1.3 does not hold in general).* Suppose $q = 1$ and let

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

Then, $X_n \to 0$ as $n \to \infty$ so that,

$$X_n \xrightarrow{\text{P}} X.$$

However,

$$\mathbb{E}\left[|X_n - X|\right] = \mathbb{E}\left[X_n\right] = 1 \neq 0 = \mathbb{E}\left[X\right].$$

Thus, in general, convergence in probability does not imply convergence in moments.

Proposition 1.4. *Convergence in higher moments imply convergence in lower moments.*

Proof. Suppose $k \geq j \geq 1$ and that

$$\mathbb{E}\left[\left|X_n - X\right|^k\right] \to 0.$$

Define $Z \coloneqq \left|X_n - X\right|^k$, then for sufficiently large n, it must be that

$$\mathbb{E}\left[Z_n\right] = \mathbb{E}\left[\left|X_n - X\right|^k\right] < \infty.$$

Consider the concave function $f(z) = z^{j/k}$. Then, by Proposition 1.1,

$$\infty > \mathbb{E}\left[\left|X_n - X\right|^j\right] = \mathbb{E}\left[f\left(Z_n\right)\right].$$

We can therefore apply Jensen's Inequality to conclude that

$$\mathbb{E}\left(\left|X_n - X\right|^k\right)^{\frac{j}{k}} = f\left(\mathbb{E}\left[Z_n\right]\right) \geq \mathbb{E}\left[f\left(Z_n\right)\right] = \mathbb{E}\left[\left|X_n - X\right|^j\right] \geq 0.$$

Given that the left-hand side converges to zero as n goes to infinity, we can conclude that $\mathbb{E}[|X_n - X|^j] \to 0$ as $n \to \infty$. ■

1.4 Convergence in distribution

Convergence in distribution is sometimes referred to as convergence in law or weak convergence. If $\{X_n : n \geq 1\}$ converges to a "common" distribution, e.g. $N(0,1)$, we may write $X_n \xrightarrow{\text{d}} N(0,1)$.

Definition 1.6. *(Convergence in Distribution)* Let $\{X_n : n \geq 1\}$ and X be random vectors on $\mathbb{R}^k$. X_n is said to *converge in distribution* to X, denoted $X_n \xrightarrow{\text{d}} X$ if, for all x for which $\mathbb{P}(X \leq x)$ is continuous, as $n \to \infty$,

$$\mathbb{P}(X_n \leq x) \to \mathbb{P}(X \leq x).$$

The following gives equivalent, alternative definitions of convergence in distribution.

Theorem 1.4. *(Portmanteau Lemma) Let $\{X_n : n \geq 1\}$ and X be random variable on $\mathbb{R}^k$. Then, the following statements are equivalent.*

(i) $X_n \xrightarrow{\text{d}} X$.

(ii) $\mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)]$ *as* $n \to \infty$, *for any real-valued, continuous and bounded function* f.

(iii) $\mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)]$ *as* $n \to \infty$, *for any real-valued, Lipschitz-continuous and bounded function* f.[a]

(iv) $\liminf_{n\to\infty} \mathbb{E}[f(X_n)] \geq \mathbb{E}[f(X)]$ *for any real-valued, continuous and nonnegative function* f.

(v) $\liminf_{n\to\infty} \mathbb{P}(X_n \in G) \geq \mathbb{P}(X \in G)$ *for any open set* G.

(vi) $\limsup_{n\to\infty} \mathbb{P}(X_n \in F) \leq \mathbb{P}(X \in F)$ *for any closed in* F.

(vii) $\mathbb{P}(X_n \in B) \to \mathbb{P}(X \in B)$ *for every Borel set* $B \in \mathcal{B}$ *such that* $\mathbb{P}(x \in \{\text{cl}(B) \setminus \text{int}(B)\}) = 0$.

[a] Let $f : [a,b] \to \mathbb{R}$ be a real-valued function. f is said to be *Lipschitz-continuous* on $[a,b]$ if there exists $K \in \mathbb{R}_+$ such that, for any $x, y \in [a,b]$,

$$|f(x) - f(y)| \leq K|x - y|.$$

The example below illustrates why we need the condition that $\mathbb{P}(X \leq x)$ is continuous.

Example 1.7. Let $X_n = 1/n$ and $X = 0$—i.e. they are degenerate random variables. Clearly, $X_n \to X$ as $n \to \infty$ in the usual sense. Note that

$$\mathbb{P}(X_n \leq x) = \begin{cases} 1 & x \geq X_n = \frac{1}{n} \\ 0 & x < X_n = \frac{1}{n} \end{cases}, \quad \mathbb{P}(X \leq x) = \begin{cases} 1 & x \geq X = 0 \\ 0 & x < X = 0 \end{cases}.$$

If $x \neq 0$, then $\mathbb{P}(X_n \leq x) \to \mathbb{P}(X \leq x)$ as $n \to \infty$. However, if $x = 0$, then $\mathbb{P}(X_n \leq 0) = 0$ for any n while $\mathbb{P}(X \leq 0) = 1$. The definition of convergence in distribution implies that we can ignore such discontinuous points.

Remark 1.2. (Theorem 1.4) Note that $G = (0,1)$ is an open set so that property (v) can be written as

$$\liminf_{n\to\infty} \mathbb{P}(X_n \in (0,1)) = \liminf_{n\to\infty} \mathbb{P}(0 < X_n < 1) \geq \mathbb{P}(0 < X < 1).$$

Similarly, if we set $F = [0,1]$, we can write property (vi) as

$$\limsup_{n\to\infty} \mathbb{P}(X_n \in [0,1]) = \limsup_{n\to\infty} \mathbb{P}(0 \leq X_n \leq 1) \leq \mathbb{P}(0 \leq X \leq 1).$$

1.4.1 Convergence in probability implies convergence in distribution

We will show that convergence in distribution is a weaker notion than convergence in probability.

Lemma 1.4. *Let* $\{X_n : n \geq 1\}$, $\{Y_n : n \geq 1\}$ *and* X *be random vectors on* $\mathbb{R}^k$. *Then,*

$$X_n \xrightarrow{\text{d}} X \text{ and } Y_n - X_n \xrightarrow{\text{p}} 0 \Rightarrow Y_n \xrightarrow{\text{d}} X.$$

Proposition 1.5. *Let* $\{X_n : n \geq 1\}$ *and* X *be random vectors. Then,*

$$X_n \xrightarrow{\text{p}} X \Rightarrow X_n \xrightarrow{\text{d}} X.$$

Proof. (Lemma 1.4) Using Portmanteau Lemma (iii), we want to show that, for any real-valued, Lipschitz-continuous and bounded function f,

$$\begin{aligned} &\mathbb{E}\left[f\left(Y_n\right)\right] \to \mathbb{E}\left[f\left(X\right)\right] \\ \Leftrightarrow\ &\mathbb{E}\left[f\left(Y_N\right) - f\left(X\right)\right] \to 0. \end{aligned}$$

By assumption,

$$X_n \xrightarrow{\text{d}} X \Leftrightarrow \mathbb{E}\left[f\left(X_n\right)\right] \to \mathbb{E}\left[f\left(X\right)\right] \Leftrightarrow \mathbb{E}\left[f\left(X_n\right) - f\left(X\right)\right] \to 0,$$

where we used Portanteau Lemma again. Then, to prove the Lemma, it suffices to show that $\mathbb{E}[f(Y_n)] \to \mathbb{E}[f\left(X_n\right)]$ since

$$\underbrace{\mathbb{E}\left[f\left(X_n\right) - f\left(X\right)\right]}_{\to 0} + \underbrace{\mathbb{E}\left[f\left(Y_n\right) - f\left(X_n\right)\right]}_{\to 0} = \mathbb{E}\left[f\left(Y_n\right) - f\left(X\right)\right] \to 0.$$

Notice that

$$\left|\mathbb{E}\left[f(Y_n) - f\left(X_n\right)\right]\right| \to 0 \Rightarrow \mathbb{E}\left[f(Y_n) - f\left(X_n\right)\right] \to 0.$$

Since $|\cdot|$ is a convex function, by Jensen's inequality,[7]

$$\left|\mathbb{E}\left[f\left(Y_n\right) - f\left(X_n\right)\right]\right| \leq \mathbb{E}\left[\left|f\left(Y_n\right) - f\left(X_n\right)\right|\right]. \tag{1.8}$$

We want to use the fact that $Y_n - X_n \xrightarrow{\text{p}} 0$ so let us split $\left|f\left(Y_n\right) - f\left(X_n\right)\right|$ into two parts: fixing $\varepsilon > 0$, then

$$\begin{aligned} \left|f\left(Y_n\right) - f\left(X_n\right)\right| = &\left|f\left(Y_n\right) - f\left(X_n\right)\right| \mathbf{1}_{\{|Y_n - X_n| > \varepsilon\}} \\ &+ \left|f\left(Y_n\right) - f\left(X_n\right)\right| \mathbf{1}_{\{|Y_n - X_n| \leq \varepsilon\}}. \end{aligned} \tag{1.9}$$

Consider the second term on the right-hand side. Since f is Lipschitz continuous, there exists $K \geq 0$ such that

$$\left|f\left(Y_n\right) - f\left(X_n\right)\right| \leq K\left|Y_n - X_n\right|.$$

Moreover, when $|Y_n - X_n| \leq \varepsilon$, then

$$\left|f\left(Y_n\right) - f\left(X_n\right)\right| \leq K\left|Y_n - X_n\right| \leq K\varepsilon,$$

so that

$$\left|f\left(Y_n\right) - f\left(X_n\right)\right| \mathbf{1}_{\{|Y_n - X_n| \leq \varepsilon\}} \leq K\varepsilon.$$

Now the first term. We know that f is bounded. Denote the bound as B, then

$$\left|f\left(Y_n\right) - f\left(X_n\right)\right| \leq 2B$$

[7] This also follows as $\mathbb{E}[\cdot]$ is an integral and noting that integral of the absolute value of a function cannot be lower than the integral of the function itself.

so that

$$|f(Y_n) - f(X_n)| \mathbf{1}_{\{|Y_n - X_n| > \varepsilon\}} \leq 2B\mathbf{1}_{\{|Y_n - X_n| > \varepsilon\}}.$$

We can now write (1.9) as

$$\begin{aligned} |f(Y_n) - f(X_n)| &\leq 2B\mathbf{1}_{\{|Y_n - X_n| > \varepsilon\}} + K\varepsilon \\ \Rightarrow \mathbb{E}[|f(Y_n) - f(X_n)|] &\leq 2B\mathbb{E}\left[\mathbf{1}_{\{|Y_n - X_n| > \varepsilon\}}\right] + K\varepsilon \\ &= 2B\mathbb{P}(|Y_n - X_n| > \varepsilon) + K\varepsilon. \end{aligned} \tag{1.10}$$

Since $Y_n - X_n \xrightarrow{\text{p}} 0$, we know that, as $n \to \infty$,

$$\mathbb{P}(|Y_n - X_n| > \varepsilon) \to 0$$

so that the first term on the right-hand side of equation (1.10) converges to zero. Since we can set ε to be arbitrarily small, we can therefore make the right-hand side of equation (1.10) arbitrarily small as $n \to \infty$. Thus, the right-hand side of equation (1.8) can be made arbitrarily small so that $\mathbb{E}[f(Y_n)] - \mathbb{E}[f(X_n)] \to 0$ as $n \to \infty$. ■

Proof. (Proposition 1.5) By Lemma lemma 1.4, we have that

$$\tilde{X}_n \xrightarrow{\text{d}} \tilde{X} \text{ and } \tilde{Y}_n - \tilde{X}_n \xrightarrow{\text{p}} 0 \Rightarrow \tilde{Y}_n \xrightarrow{\text{d}} \tilde{X}.$$

Setting

$$\begin{aligned} \tilde{X}_n &= X, \\ \tilde{X} &= X, \\ \tilde{Y}_n &= X_n \end{aligned}$$

gives the result. Note that $X \xrightarrow{\text{d}} X$ holds trivially. ■

Example 1.8. *(Converse to Proposition 1.5 does not hold in general)* Let $X \sim N(0,1)$ and $X_n \overset{\text{d}}{=} -X$. Since $N(0,1)$ is symmetric, $X_n \xrightarrow{\text{d}} X$ (in fact, they have the same distribution for all $n \geq 1$). However, notice that $X_n \overset{\text{p}}{\not\to} X$ since $X_n - X \overset{\text{d}}{=} -X - X \overset{\text{d}}{=} -2X \overset{\text{d}}{=} 2X$.

This means that convergence in distribution is a weaker notion than convergence in probability. However, the converse holds for an important special case as shown below.

Lemma 1.5. *(Special case in which converse to Proposition* 1.5 *holds). Let* $\{X_n : n \geq 1\}$ *be random vectors and* $c \in \mathbb{R}^k$ *a constant. Then,*

$$X_n \xrightarrow{\text{d}} c \Rightarrow X_n \xrightarrow{\text{p}} c.$$

Proof. Fix $\varepsilon > 0$. We want to show that, as $n \to \infty$,

$$\mathbb{P}(|X_n - c| > \varepsilon) \to 0.$$

The idea is to use property (vi) of Portmanteau Lemma. However, we require a closed set for this. Let $B_\delta(c)$ denote the open ball around c with radius δ. By construction, B_δ is open and so its complement, denoted $B_\delta^c(c)$, is closed. Choose any $\delta \in (0, \varepsilon)$, then

$$\{|X_n - c| > \varepsilon\} \subseteq B_\delta^c(c).$$

Hence,

$$\mathbb{P}(|X_n - c| > \varepsilon) \leq \mathbb{P}(X_n \in B_\delta^c(c)).$$

By Portmanteau Lemma, then

$$\limsup_{n \to \infty} \mathbb{P}(X_n \in B_\delta^c(c)) \leq \mathbb{P}(c \in B_\delta^c(c)) = 0,$$

where the last equality comes from the fact that $c \notin B_\delta^c(c)$ by construction. Finally, because

$$\mathbb{P}\left(X_n \in B_\delta^c(c)\right) \leq \limsup_{n\to\infty} \mathbb{P}\left(X_n \in B_\delta^c(c)\right)$$

we must have that $\mathbb{P}\left(|X_n - c| > \varepsilon\right) \to 0$ as $n \to \infty$. ■

Example 1.9. *(Convergences in marginal distributions does not imply convergence in joint distribution).* Recall that, by Proposition 1.2, convergence in marginal probabilities implies convergence in joint probabilities. A similar result does not hold for convergence in distribution. To see this, consider

$$\begin{pmatrix} X_n \\ Y_n \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{matrix} 1 & (-1)^n \\ (-1)^n & 1 \end{matrix}\right).$$

Note that $X_n \xrightarrow{\mathrm{d}} N(0,1)$ and $Y_n \xrightarrow{\mathrm{d}} N(0,1)$. However, the joint density does not ever converge as it "flips" from being perfectly positive and negative correlated between X_n and Y_n.[8]

Once again, there is an important exception to this, which we explore below.

Lemma 1.6. *(Special case in which convergences in marginal distributions imply convergence in joint distribution). Let $\{X_n : n \geq 1\}$, $\{Y_n : n \geq 1\}$ and X be random vectors and $c \in \mathbb{R}^k$ a constant. Then,*

$$X_n \xrightarrow{\mathrm{d}} X \text{ and } Y_n \xrightarrow{\mathrm{d}} c \Rightarrow (X_n, Y_n) \xrightarrow{\mathrm{d}} (X, c).$$

Proof. First, note that, using the definition of $|\cdot|$ as the Euclidean norm:

$$\begin{aligned} |(X_n, Y_n) - (X_n, c)| &= \left[(X_n - X_n)^2 + (Y_n - c)^2\right]^{\frac{1}{2}} \\ &= |Y_n - c|. \end{aligned}$$

By Lemma 1.5, $Y_n \xrightarrow{\mathrm{d}} c$ implies $Y_n \xrightarrow{\mathrm{p}} c$ so that $|Y_n - c| \xrightarrow{\mathrm{p}} 0$. That is,[9]

$$\begin{aligned} Y_n \xrightarrow{\mathrm{d}} c &\Rightarrow |(X_n, Y_n) - (X_n, c)| \xrightarrow{\mathrm{p}} 0 \\ &\Leftrightarrow (X_n, Y_n) \xrightarrow{\mathrm{p}} (X_n, c). \end{aligned}$$

Then, by Lemma 1.4, if we can show that $(X_n, c) \xrightarrow{\mathrm{d}} (X, c)$, then we would immediately have

$$(X_n, c) \xrightarrow{\mathrm{d}} (X, c) \text{ and } (X_n, Y_n) \xrightarrow{\mathrm{p}} (X_n, c) \Rightarrow (X_n, Y_n) \xrightarrow{\mathrm{d}} (X, c).$$

To show that $(X_n, c) \xrightarrow{\mathrm{d}} (X, c)$, by Portmanteau Lemma, it is sufficient to show that, for any real-valued, continuous and bounded function f,

$$\mathbb{E}\left[f(X_n, c)\right] \to \mathbb{E}\left[f(X, c)\right].$$

Since $X_n \xrightarrow{\mathrm{d}} X$, then, by Portmanteau Lemma, for any real-valued, continuous and bounded function g, $\mathbb{E}\left[g(X_n)\right] \to \mathbb{E}\left[g(X)\right]$. Since c is a constant, we can define f so that

$$g(X_n) = f(X_n, c),$$

and f will still be real-valued, continuous and bounded. Hence, $\mathbb{E}\left[f(X_n, c)\right] \to \mathbb{E}\left[f(X, c)\right]$ and we are done. ■

[8] Note that every bounded sequence has a convergence subsequent. In this case, if we create a subsequent X_k consists of all n that are odd, then the joint distribution converges.

[9] Note

$$\begin{aligned} |(X_n, Y_n) - (X_n, c)| \xrightarrow{\mathrm{p}} 0 &\Leftrightarrow \mathbb{P}\left(||(X_n, Y_n) - (X_n, c)|| > \varepsilon\right) \to 0 \\ &\Leftrightarrow \mathbb{P}\left(|(X_n, Y_n) - (X_n, c)| > \varepsilon\right) \to 0 \\ &\Leftrightarrow (X_n, Y_n) - (X_n, c) \xrightarrow{\mathrm{p}} 0. \end{aligned}$$

1.4.2 Continuous Mapping Theorem for convergence in distribution

We have already seen that continuous functions preserve convergence in probability. Below shows that the same property holds for convergence in distribution.

Theorem 1.5. *(Continuous Mapping Theorem for convergence in distribution) Let* $\{X_n : n \geq 1\}$ *and* X *be random vectors on* $\mathbb{R}^k$*. Suppose that* $g : \mathbb{R}^k \to \mathbb{R}^d$ *is continuous at each point in the set* C *such that* $\mathbb{P}(X \in C) = 1$*. Then,*

$$X_n \xrightarrow{\mathrm{d}} X \Rightarrow g(X_n) \xrightarrow{\mathrm{d}} g(X).$$

Lemma 1.7. *(Slutsky's Lemma). Let* $\{X_n : n \geq 1\}$*,* $\{Y_n : n \geq 1\}$ *and* Y *be random vectors, and* $c \in \mathbb{R}^k$ *a constant. If* $X_n \xrightarrow{\mathrm{d}} X$ *and* $Y_n \xrightarrow{\mathrm{d}} c$*, then* $X_n' Y_n \xrightarrow{\mathrm{d}} X'c$ *and* $X_n + Y_n \xrightarrow{\mathrm{d}} X + c$*.*

Proof. (Continuous Mapping Theorem for convergence in distribution) By Portmanteau Lemma (vi), it is sufficient to show that for any closed set H,

$$\limsup_{n\to\infty} \mathbb{P}(g(X_n) \in H) \leq \mathbb{P}(g(X) \in H).$$

Define

$$g^{-1}(H) \coloneqq \left\{x \in \mathbb{R}^k : g(X) \in H\right\}.$$

Then,

$$\mathbf{1}_{\{g(X_n)\in H\}} = \mathbf{1}_{\{X_n \in g^{-1}(H)\}}. \tag{1.11}$$

However, we do not know whether $g^{-1}(H)$ is closed. By definition, $g^{-1}(H)$ is closed if g is continuous. Notice that

$$g^{-1}(H) \subseteq \mathrm{cl}\left(g^{-1}(H)\right).$$

We now claim that

$$\mathrm{cl}\left(g^{-1}(H)\right) \subseteq \left\{g^{-1}(H) \cup C^c\right\}.$$

To see why this holds, suppose first that $x \in \mathrm{cl}(g^{-1}(H))$. Then, since a closure is closed, there exists a sequence $x_n \in g^{-1}(H)$ with $x_n \to x \in \mathrm{cl}\left(g^{-1}(H)\right)$. However, we do not know if $x \in \left\{g^{-1}(H) \cup C^c\right\}$. Note that x is either in C or C^c. If $x \in C$, then g is continuous so that $g(x_n) \to g(x) \in H$. Then, by (1.11), $x \in g^{-1}(H)$. If instead $x \notin C$, then $x \in C^c$. Hence, $x \in \left\{g^{-1}(H) \cup C^c\right\}$ so that the claim holds.

Finally,

$$\begin{aligned}
\limsup_{n\to\infty} \mathbb{P}(g(X_n) \in H) &= \limsup_{n\to\infty} \mathbb{P}\left(X_n \in g^{-1}(H)\right) \\
&\leq \limsup_{n\to\infty} \mathbb{P}\left(X_n \in \mathrm{cl}\left(g^{-1}(H)\right)\right) \\
\text{(by Portmanteau Lemma)} &\leq \mathbb{P}\left(X \in \mathrm{cl}\left(g^{-1}(H)\right)\right) \\
&\leq \mathbb{P}\left(X \in g^{-1}(H) \cup C^c\right) \\
(\because \mathbb{P}(X \in C) = 1) &= \mathbb{P}\left(X \in g^{-1}(H)\right) \\
&= \mathbb{P}(g(X) \in H).
\end{aligned}$$

■

Proof. (Slutsky's Lemma) Note that summation and multiplications are continuous operations. Thus, we can apply the Continuous Mapping Theorem to the joint distribution (X_n, Y_n) to obtain the result. ■

1.5 Central Limit Theorem

Theorem 1.6. *(Univariate Central Limit Theorem) Let* $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} P$ *on* $\mathbb{R}$ *and* $\sigma^2(P) < \infty$*. Then,*

$$\sqrt{n}\left(\bar{X}_n - \mu(P)\right) \xrightarrow{\text{d}} N\left(0, \sigma^2(P)\right).$$

The following lemma allows us to generalise the univariate central limit theorem to multivariate cases.

Lemma 1.8. *(Cramér-Wold Device) Let* $\{X_n : n \geq 1\}$ *and* X *be random vectors. Then,*

$$X_n \xrightarrow{\text{d}} X \Leftrightarrow t'X_n \xrightarrow{\text{d}} t'X, \forall t' \in \mathbb{R}^k.$$

Theorem 1.7. *(Multivariate Central Limit Theorem) Let* $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} P$ *on* $\mathbb{R}^k$*, and* $\Sigma(P) < \infty$ *(i.e. every entry in the variance-covariance matrix is finite). Then,*

$$\sqrt{n}\left(\bar{X}_n - \mu(P)\right) \xrightarrow{\text{d}} N\left(0, \Sigma(P)\right).$$

We omit the proof of the Univariate Central Limit Theorem and the Cramér-Wold Device.

Proof. (Multivariate Central Limit Theorem) By the Cramér-Wold Device ($\Leftarrow$), it is sufficient to show that

$$t'\sqrt{n}\left(\bar{X}_n - \mu(P)\right) \xrightarrow{\text{d}} t'N\left(0, \Sigma(P)\right).$$

Note that

$$\begin{aligned} t'\sqrt{n}\left(\bar{X}_n - \mu(P)\right) &= t'\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu(P)\right) \\ &= \frac{\sqrt{n}}{n}\sum_{i=1}^{n} t'\left(X_i - \mu(P)\right). \end{aligned}$$

Since $X_i = (X_{i,1}, X_{i,2}, \ldots, X_{i,k})'$,

$$Y_i \coloneqq t'X_i = \sum_{j=1}^{k} t_j X_{i,j}$$

is a linear combination of all the elements in X_i, which are individually iid. Note that

$$\begin{aligned} \mathbb{E}[Y_i] &= \mathbb{E}[t'X_i] = t'\left(\mathbb{E}[X_i]\right), \\ \text{Var}[Y_i] &= \text{Var}[t'X_i] \\ &= t'\Sigma(P)\, t. \end{aligned}$$

By assumption, $\Sigma(P) < \infty$ so that $\text{Var}[Y_i] < \infty$ since the latter is a finite sum of finite numbers. We can then apply the univariate Central Limit Theorem to Y_i to obtain that

$$\begin{aligned} &\sqrt{n}\left(\bar{Y}_n - \mathbb{E}[Y_i]\right) \xrightarrow{\text{d}} N\left(0, \text{Var}[Y_i]\right) \\ \Rightarrow\ &\sqrt{n}\left(t'\bar{X}_n - t'\mu(P)\right) \xrightarrow{\text{d}} N\left(0, t'\Sigma(P)\, t\right) \\ \Rightarrow\ &t'\sqrt{n}\left(\bar{X}_n - \mu(P)\right) \xrightarrow{\text{d}} t'N\left(0, \Sigma(P)\right). \end{aligned}$$

Hence, by the Cramér-Wold Device, above implies (in fact, is equivalent to)

$$\sqrt{n}\left(\bar{X}_n - \mu(P)\right) \xrightarrow{\text{d}} N\left(0, \Sigma(P)\right). \qquad \blacksquare$$

1.6 Hypothesis testing

We wish to apply the concept we developed above to test hypotheses about $\mu(P)$. For example, we wish to test the *null hypothesis*,

$$H_0 : \mu(P) \leq 0,$$

against the *alternative hypothesis*,

$$H_1 : \mu(P) > 0.$$

There are two types of errors:

- ▷ Type I error, which is the probability of rejecting H_0 when it is true (i.e. false rejection/positive);
- ▷ Type II error, which is the probability of not rejecting H_0 when it is false (i.e. false negative).

Generally, it is not possible to minimise both type I and II errors (since type II error can be under any distribution).

1.6.1 Consistency in level

For a test $\phi_n = \phi_n(X_1, X_2, \ldots, X_n) \in [0,1]$, $\mathbb{E}_P[\phi_n]$ is the probability of rejecting H_0 when viewed as a function of P, called the power function of the test. For P satisfying H_0, $\mathbb{E}_P[\phi_n]$ is the probability of type I error, and for a P satisfying H_1, $1 - \mathbb{E}_P[\phi_n]$ is the probability of type II error.

Definition 1.7. *(Consistency in level)* We say that a test $\phi_n = \phi_n(X_1, X_2, \ldots, X_n) \in [0,1]$ is *consistent in level* if

$$\limsup_{n\to\infty} \mathbb{E}_P[\phi_n] \leq \alpha,$$

for P satisfying H_0 and where $\alpha \in (0,1)$ is the significance level of the test.

We restrict attention to tests of the form:

$$\phi_n = \mathbf{1}_{\{T_n > c_n\}},$$

where T_n is the *test statistic* which is a function of data such that "large" values provide evidence against H_0; and c_n is the *critical value* which gives the definition of "large". Therefore, the requirement of consistency in level is given by

$$\limsup_{n\to\infty} \mathbb{E}_P[\phi_n] = \limsup_{n\to\infty} \mathbb{E}_P\left[\mathbf{1}_{\{T_n > c_n\}}\right] = \limsup_{n\to\infty} \mathbb{P}(T_n > c_n) \leq \alpha$$

for P satisfying H_0.

Example 1.10. *(Test statistic for the sample mean)* Suppose that $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} P$ on $\mathbb{R}$ and that $\sigma^2(P) < \infty$. By the Central Limit Theorem,

$$\sqrt{n}\left(\bar{X}_n - \mu(P)\right) \xrightarrow{\text{d}} N\left(0, \sigma^2(P)\right).$$

In Example 1.4, we showed that $s_n^2 \xrightarrow{\text{p}} \sigma^2(P)$. Suppose that $\sigma^2(P) > 0$. Then, by the Continuous Mapping Theorem,

$$\frac{\sqrt{n}\left(\bar{X}_n - \mu(P)\right)}{s_n} \xrightarrow{\text{d}} N(0,1). \tag{1.12}$$

In particular, we use the fact that $f(x) = 1/\sqrt{x}$ is a continuous function when $x \neq 0$.

Consider the following null and alternative hypotheses,

$$H_0 : \mu(P) = 0,$$
$$H_1 : \mu(P) \neq 0.$$

Then, under the null,

$$\frac{\sqrt{n}\bar{X}_n}{s_n} \xrightarrow{\mathrm{d}} N(0,1).$$

So a natural candidate for a test statistic is

$$T_n \coloneqq \frac{\sqrt{n}\bar{X}_n}{s_n}.$$

To conduct the test at level α, the critical value is based on the standard normal distribution. But since the test is two sided, we would reject whenever we see that T_n is both too low or too high. So with significance level α,

$$c_n \coloneqq \Phi^{-1}\left(1-\frac{\alpha}{2}\right) \coloneqq z_{1-\frac{\alpha}{2}},$$

where Φ is the CDF of $N(0,1)$. Note that $\Phi(x) = \mathbb{P}(z \leq x)$ and we want to choose c_n such that the probability of z being greater than c_n is $1-\alpha/2$ (and the probability of z being less than c_n is $\alpha/2$); i.e. c_n is the $1-\alpha/2$th quantile of the standard normal distribution. The test is therefore

$$\phi_n = \mathbf{1}_{\{|T_n|>c_n\}}.$$

We want to show that this is consistent in level. That is, we wish to show that

$$\limsup_{n\to\infty} \mathbb{E}_P[\phi_n] \leq \alpha$$

under the null. Notice that

$$\mathbb{E}_p[\phi_n] = \mathbb{P}\left(\frac{\sqrt{n}\left|\bar{X}_n\right|}{s_n} > z_{1-\frac{\alpha}{2}}\right).$$

We also have that

$$\frac{\sqrt{n}\left|\bar{X}_n\right|}{s_n} \xrightarrow{\mathrm{d}} \left|N(0,1)\right|,$$

which follows from the Continuous Mapping Theorem noting that $|\cdot|$ is a continuous operation.

To show that the test is consistent, we appeal to Portmanteau Lemma (vi), which requires a closed set, but we do not have a closed set here. However, we can make it closed:

$$\mathbb{E}_p[\phi_n] \leq \mathbb{P}\left(\frac{\sqrt{n}\left|\bar{X}_n\right|}{s_n} > z_{1-\frac{\alpha}{2}}\right) \leq \mathbb{P}\left(\frac{\sqrt{n}\sqrt{n}\left|\bar{X}_n\right|}{s_n} \geq z_{1-\frac{\alpha}{2}}\right).$$

Taking lim sup of both sides, we obtain

$$\begin{aligned}\limsup_{n\to\infty} \mathbb{E}_p[\phi_n] &\leq \limsup_{n\to\infty} \mathbb{P}\left(\frac{\sqrt{n}\left|\bar{X}_n\right|}{s_n} \geq z_{1-\frac{\alpha}{2}}\right) \\ &\leq \mathbb{P}\left(|Z| \geq z_{1-\frac{\alpha}{2}}\right),\end{aligned}$$

where $Z \sim N(0,1)$. Convergence in distribution requires $\mathbb{P}(|Z| \geq z)$ to be continuous at $z = z_{z_{1-\frac{\alpha}{2}}}$, which is true in the case of normal distribution (it is continuous everywhere). Finally,

$$\begin{aligned}\limsup_{n\to\infty} \mathbb{E}_p[\phi_n] &\leq \mathbb{P}(|Z| \geq z_{1-\alpha}) \\ &= \mathbb{P}\left(Z < z_{\frac{\alpha}{2}}\right) + \mathbb{P}\left(Z > z_{1-\frac{\alpha}{2}}\right) \\ &= \frac{\alpha}{2} + \frac{\alpha}{2} = \alpha.\end{aligned}$$

That is, the test is consistent in level α.

Example 1.11. Suppose we wish to test the following hypotheses, at level α,

$$\begin{aligned} H_0 &: \mu(P) \leq 0, \\ H_1 &: \mu(P) > 0. \end{aligned}$$

The test is now one-sided and we would only wish to reject the null in one direction.

To conduct the test at level α, the critical value is based on the standard normal distribution:

$$c_n \coloneqq \Phi^{-1}(1-\alpha) \coloneqq z_{1-\alpha},$$

where Φ is the CDF of $N(0,1)$. Note that $\Phi(x) = \mathbb{P}(z \leq x)$ and we want to choose c_n such that the probability of z being less than c_n is $1-\alpha$ (or, equivalently, we want to choose c_n such that the probability of z being higher than c_n is α); i.e. c_n is the $1-\alpha$th quantile of the standard normal distribution. The test is then

$$\phi_n = \mathbf{1}_{\{T_n > c_n\}}.$$

We want to show that this is consistent in level. That is, we wish to show that

$$\limsup_{n\to\infty} \mathbb{E}_P[\phi_n] \leq \alpha$$

under the null; i.e. $\mu(P) \leq 0$. Following the same steps as before

$$\mathbb{E}_p[\phi_n] = \mathbb{P}(T_n > c_n) = \mathbb{P}\left(\frac{\sqrt{n}\bar{X}_n}{s_n} > z_{1-\alpha}\right).$$

Add and subtract $\mu(P)$ to $\bar{X}_n$ yields

$$\mathbb{E}_p[\phi_n] = \mathbb{P}\left(\frac{\sqrt{n}\left(\bar{X}_n - \mu(P)\right)}{s_n} + \frac{\sqrt{n}\mu(P)}{s_n} > z_{1-\alpha}\right), \tag{1.13}$$

where we notice that, under the null, $\mu(P) \leq 0$, and so

$$\mathbb{E}_p[\phi_n] \leq \mathbb{P}\left(\frac{\sqrt{n}\left(\bar{X}_n - \mu(P)\right)}{s_n} > z_{1-\alpha}\right). \tag{1.14}$$

To apply Portmanteau Lemma (vi), we again close the set:

$$\mathbb{E}_p[\phi_n] \leq \mathbb{P}\left(\frac{\sqrt{n}\left(\bar{X}_n - \mu(P)\right)}{s_n} > z_{1-\alpha}\right) \leq \mathbb{P}\left(\frac{\sqrt{n}\left(\bar{X}_n - \mu(P)\right)}{s_n} \geq z_{1-\alpha}\right).$$

Taking lim sup of both sides, we obtain

$$\begin{aligned} \limsup_{n\to\infty} \mathbb{E}_p[\phi_n] &\leq \limsup_{n\to\infty} \mathbb{P}\left(\frac{\sqrt{n}\left(\bar{X}_n - \mu(P)\right)}{s_n} \geq z_{1-\alpha}\right) \\ &\leq \mathbb{P}(Z \geq z_{1-\alpha}), \end{aligned}$$

where $Z \sim N(0,1)$. Finally,

$$\begin{aligned} \limsup_{n\to\infty} \mathbb{E}_p[\phi_n] &\leq \mathbb{P}(Z \geq z_{1-\alpha}) \\ &= 1 - \mathbb{P}(Z < z_{1-\alpha}) = 1 - \Phi(z_{1-\alpha}) \\ &= 1 - (1-\alpha) = \alpha. \end{aligned}$$

That is, the test is consistent in level α.

Remark 1.3. Suppose we alter the hypothesis test to

$$H_0 : \mu(P) \geq 0,$$
$$H_1 : \mu(P) < 0.$$

That is, we "flipped" the inequalities. How would we show that consistency in level of this test? Since normal distribution is symmetric, note that

$$-\frac{\sqrt{n}\left(\bar{X}_n - \mu(P)\right)}{s_n} \xrightarrow{\text{p}} N(0,1)$$

So we can use as the test statistic

$$T_n \coloneqq -\frac{\sqrt{n}\bar{X}_n}{s_n},\ c_n = z_{1-\alpha}.$$

Then, under the null,

$$\begin{aligned}
\mathbb{E}_p[\phi_n] = \mathbb{P}(T_n > c_n) &= \mathbb{P}\left(-\frac{\sqrt{n}\bar{X}_n}{s_n} > z_{1-\alpha}\right) \\
&= \mathbb{P}\left(-\frac{\sqrt{n}\left(\bar{X}_n - \mu(P)\right)}{s_n} - \underbrace{\frac{\sqrt{n}\mu(P)}{s_n}}_{\geq 0} > z_{1-\alpha}\right) \\
&\leq \mathbb{P}\left(-\frac{\sqrt{n}\left(\bar{X}_n - \mu(P)\right)}{s_n} > z_{1-\alpha}\right) \\
&\leq \mathbb{P}\left(-\frac{\sqrt{n}\left(\bar{X}_n - \mu(P)\right)}{s_n} \geq z_{1-\alpha}\right) \\
\Rightarrow \limsup_{n\to\infty} \mathbb{E}_p[\phi_n] &\leq \mathbb{P}(Z \geq z_{1-\alpha}) = \alpha.
\end{aligned}$$

The following proposition allows us to consider cases in which the test statistic itself is a converging sequence (e.g. the test statistic is distributed according to student-t distribution, which converges to normal).

Proposition 1.6. *Suppose $T_n \xrightarrow{\text{d}} T$ and $c_n \xrightarrow{\text{p}} c$ and that $\mathbb{P}(X \leq x)$ is continuous at c. Then, as $n \to \infty$,*

$$\mathbb{P}(T_n \leq c_n) \to \mathbb{P}(T \leq c).$$

Proof. We can write what we wish to show as

$$\mathbb{P}(T_n - c_n \leq 0) \to \mathbb{P}(T - c \leq 0).$$

By assumption, $\mathbb{P}(T \leq x)$ is continuous at c, which means that $\mathbb{P}(T - c \leq 0)$ is continuous. Thus, it suffices to show that $T_n - c_n \xrightarrow{\text{d}} T - c$. But this follows immediately from applying the Slutsky's Lemma given that, by assumption, $T_n \xrightarrow{\text{d}} T$ and $c_n \xrightarrow{\text{p}} c$. ■

1.6.2 p-value of a test

Definition 1.8. *(p-value)* p-value of a test is the smallest value of α for which we reject the null hypothesis

$$\hat{p}_n \coloneqq \inf\{\alpha \in (0,1) : T_n > c_n\},$$

where c_n depends on α.

Following on from the previous example, the p-value can be defined as:

$$\begin{aligned}\hat{p}_n &= \inf\left\{\alpha \in (0,1) : \frac{\sqrt{n}\bar{X}_n}{s_n} > z_{1-\alpha}\right\} \\ &= \inf\left\{\alpha \in (0,1) : \alpha > 1 - \Phi\left(\frac{\sqrt{n}\bar{X}_n}{s_n}\right)\right\} \\ &= 1 - \Phi\left(\frac{\sqrt{n}\bar{X}_n}{s_n}\right).\end{aligned}$$

For a two-sided test, the p-value is given by

$$\begin{aligned}\hat{p}_n &= \inf\left\{\alpha \in (0,1) : \frac{\sqrt{n}\left|\bar{X}_n\right|}{s_n} > z_{1-\frac{\alpha}{2}}\right\} \\ &= 2\left(1 - \Phi\left(\frac{\sqrt{n}\left|\bar{X}_n\right|}{s_n}\right)\right).\end{aligned}$$

1.6.3 Deriving the confidence region of a test

Definition 1.9. *(Confidence set/region)* A confidence set/region of level $1-\alpha$ for $\mu(P)$, denoted $C_n \coloneqq C_n(X_1, X_2, \ldots, X_n)$, is such that the probability that the true mean is contained in the set is greater than $1-\alpha$, $\alpha \in (0,1)$; i.e.

$$\mathbb{P}(\mu(P) \in C_n) \geq 1 - \alpha.$$

We have previously constructed a confidence set for Bernoulli distributed iid random variables using Markov's Inequality (see Example 1.1), which did not rely on asymptotics. Below, we construct a confidence set/region using Central Limit Theorem, which relies on asymptotic properties.

Example 1.12. *(Confidence region using CLT)* $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} P = \text{Bernoulli}(q)$, where $q \in (0,1)$. Let α be given. We wish to construct a confidence region for $\mu(P) = q$ at level $1-\alpha$. Recall that

$$\bar{X}_n \xrightarrow{\text{p}} \mu(P) = q.$$

Since $\sigma^2(P) = q(1-q)$, a natural candidate for $\sigma^2(P)$ is

$$s_n^2 = \bar{X}_n\left(1 - \bar{X}_n\right)$$

Thus, we can write s_n^2 as a function g, parameterised as $s_n^2 = g(\bar{X}_n)$, where $g(a) = a(1-a)$. Since g is continuous, by the Continuous Mapping Theorem,

$$s_n^2 \xrightarrow{\text{p}} \sigma^2(P).$$

Since $\sigma^2(P) > 0$, by Slutsky's Lemma,

$$\frac{\sqrt{n}\left(\bar{X}_n - \mu(P)\right)}{\sqrt{\bar{X}_n\left(1 - \bar{X}_n\right)}} \xrightarrow{\text{d}} N(0,1).$$

Define

$$c_n \coloneqq z_{1-\frac{\alpha}{2}} \frac{\sqrt{\bar{X}_n\left(1 - \bar{X}_n\right)}}{\sqrt{n}},$$

and

$$C_n \coloneqq \left[\bar{X}_n - c_n, \bar{X}_n + c_n\right]. \tag{1.15}$$

To show that $\mathbb{P}(\mu(P) \in C_n) \to 1-\alpha$:

$$\begin{aligned}
\mathbb{P}(\mu(P) \in C_n) &= \mathbb{P}\left(\bar{X}_n - c_n \leq \mu(P) \leq \bar{X}_n + c_n\right) \\
&= \mathbb{P}\left(\left|\bar{X}_n - \mu(P)\right| \leq c_n\right) \\
&= \mathbb{P}\left(\left|\bar{X}_n - \mu(P)\right| \leq z_{1-\frac{\alpha}{2}} \frac{\sqrt{\bar{X}_n\left(1-\bar{X}_n\right)}}{\sqrt{n}}\right) \\
&= \mathbb{P}\left(\frac{\sqrt{n}\left|\bar{X}_n - \mu(P)\right|}{\sqrt{\bar{X}_n\left(1-\bar{X}_n\right)}} \leq z_{1-\frac{\alpha}{2}}\right) \\
&\to \mathbb{P}\left(|z| \leq z_{1-\frac{\alpha}{2}}\right) \\
&= \mathbb{P}\left(z_{\frac{\alpha}{2}} \leq z \leq z_{1-\frac{\alpha}{2}}\right) \\
&= 1 - \left(\frac{\alpha}{2} + \frac{\alpha}{2}\right) = 1-\alpha.
\end{aligned}$$

Note that $\to$ is used in the same sense as in the definition of convergence in distribution; i.e. as $n \to \infty$.

We can write confident regions in the following equivalent way:

$$C_n \coloneqq \left\{x \in \mathbb{R} : \frac{\sqrt{n}\left|\bar{X}_n - x\right|}{\sqrt{\bar{X}_n\left(1-\bar{X}_n\right)}} \leq z_{1-\frac{\alpha}{2}}\right\}.$$

Remark 1.4. The probability $\mathbb{P}(\mu(P) \in C_n)$ above is called the *coverage probability*. The actual coverage probability based on data may be poor in finite samples, in particular, when q is close to 0 or 1. To see this, suppose that the true mean is given by $q = (1-\varepsilon)^{\frac{1}{n}} > 0$, and that we have n (independent) observations with $X_1, X_2, \ldots, X_n = 1$ so that $\bar{X}_n = 1$. Then, from (1.15), since sample variance is zero in this case, the confidence region is given simply by $C_n = 1$. Thus, we realise that the confidence region does not contain the true mean, q,

$$(1-\varepsilon)^{\frac{1}{n}} = q \notin C_n = 1.$$

The probability of observing all X_i's equal to one is given by

$$\mathbb{P}(X_1 = 1, X_2 = 1, \ldots, X_n = 1) = (1-\varepsilon)^{\frac{n}{n}} = (1-\varepsilon).$$

We know that $\mu(P) \notin C_n$ if we observe all X_i's equal to one, so that

$$\mathbb{P}(\mu(P) \notin C_n) \geq 1-\varepsilon.$$

This is an inequality since there may be other values of observed X_i's that mean that $\mu(P)$ would not lie within the confidence region. The coverage probability is given by

$$\begin{aligned}
\mathbb{P}(\mu(P) \in C_n) &= 1 - \mathbb{P}(\mu(P) \notin C_n) \\
&\leq 1 - (1-\varepsilon) = \varepsilon.
\end{aligned}$$

That is, $\mathbb{P}(\mu(P) \in C_n)$ can be made arbitrary small by fixing ε small. Notice that confidence region using Markov's inequality does not have this problem (although the region is wider).

1.6.4 Testing multidimensional hypothesis

Let $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} P$ on $\mathbb{R}^k$ with $\Sigma(P) < \infty$ being the $k \times k$ variance-covariance matrix. Assume that $\Sigma(P) < \infty$. Note that $\mu(P)$ is a $k \times 1$ vector in this context. The test is

$$\begin{aligned}
H_0 &: \mu(P) = \mathbf{0}, \\
H_1 &: \mu(P) \neq \mathbf{0}.
\end{aligned}$$

From the Central Limit Theorem, since X_i's are iid

$$\sqrt{n}\left(\bar{X}_n - \mu\left(P\right)\right) \xrightarrow{\mathrm{d}} z \sim N\left(0, \Sigma\left(P\right)\right).$$

A useful fact to remember is that if $\Sigma\left(P\right)$ is invertible, and z is a $k \times 1$ vector, then

$$z \sim N\left(0, \Sigma\left(P\right)\right) \Rightarrow z' \Sigma^{-1}\left(P\right) z \sim \chi^2_k.$$

In this case, we therefore have that

$$n\left(\bar{X}_n - \mu\left(P\right)\right)' \Sigma^{-1}\left(P\right)\left(\bar{X}_n - \mu\left(P\right)\right) \xrightarrow{\mathrm{d}} \chi^2_k.$$

Although appears that can may use this to construct a test statistic, notice that we do not in fact know $\Sigma\left(P\right)$. However, we can estimate $\Sigma\left(P\right)$ using:

$$\hat{\Sigma}_n = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)\left(X_i - \bar{X}_n\right)',$$

which we already showed is consistent; i.e. $\hat{\Sigma}_n \xrightarrow{\mathrm{p}} \Sigma\left(P\right)$. Then, since we have assumed that Σ^{-1} is invertible, by the Continuous Mapping Theorem,[10] we have that

$$\hat{\Sigma}_n^{-1} \xrightarrow{\mathrm{p}} \Sigma^{-1}\left(P\right).$$

Then, by the Continuous Mapping Theorem (and the useful fact), we have that

$$n\left(\bar{X}_n - \mu\left(P\right)\right)' \hat{\Sigma}_n^{-1}\left(\bar{X}_n - \mu\left(P\right)\right) \xrightarrow{\mathrm{d}} \chi^2_k. \tag{1.16}$$

Suppose we wish to test the following hypothesis:

$$\begin{aligned} H_0 &: \mu\left(P\right) = 0, \\ H_1 &: \mu\left(P\right) \neq 0. \end{aligned}$$

We can then let

$$\begin{aligned} T_n &\coloneqq n\bar{X}_n' \hat{\Sigma}_n^{-1} \bar{X}_n, \\ c_n &\coloneqq c_{k,1-\alpha}, \end{aligned}$$

where $c_{k,1-\alpha}$ is the $1-\alpha$th quantile of χ^2_k. Since $\mu\left(P\right) = 0$ under the null, we know that

$$T_n \coloneqq n\bar{X}_n' \hat{\Sigma}_n^{-1} \bar{X}_n \xrightarrow{\mathrm{d}} \chi^2_k.$$

The test is,

$$\phi_n = \mathbf{1}_{\{T_n > c_n\}}.$$

To show that the test is consistent in level,

$$\begin{aligned} \limsup_{n\to\infty} \mathbb{E}_P\left[\phi_n\right] &= \limsup_{n\to\infty} \mathbb{P}\left(n\bar{X}_n' \hat{\Sigma}_n^{-1} \bar{X}_n > c_{k,1-\alpha}\right) \\ &\leq \mathbb{P}\left(T \geq c_{k,1-\alpha}\right) = 1 - \left(1-\alpha\right) = \alpha, \end{aligned}$$

where $T \sim \chi^2_k$.

[10] The inverse is given by dividing the adjugate matrix by the determinant. Since determinants are polynomials, it is therefore continuous. That is, the inverse mapping is a continuous operation.

1.6.5 Delta method

Theorem 1.8. *(Delta method) Let $\{X_n : n \geq 1\}$ and X be random vectors, $c \in \mathbb{R}^k$ a constant, and τ_n a sequence of constants such that $\tau_n \to \infty$ and $\tau_n (X_n - c) \xrightarrow{d} X$. Let $g : \mathbb{R}^k \to \mathbb{R}^d$ be a function that is continuous and differentiable at c. Denote $Dg(c)$ as the $d \times k$ matrix of partials of g evaluated at c. Then,*

$$\tau_n (g(X_n) - g(c)) \xrightarrow{d} Dg(c) X.$$

In particular, if

$$X \sim N(0, \Sigma) \Rightarrow \tau_n (g(X_n) - g(c)) \xrightarrow{d} N\left(0, Dg(c) \Sigma Dg(c)'\right).$$

Proof. Because g is differentiable at c, this means that $g(c)$ can be well-approximated by a linear function at c. The linear approximation of $g(X_n)$ around c is then given by (Taylor expansion),

$$g(X_n) = g(c) + Dg(c)(X_n - c) + R(X_n - c), \tag{1.17}$$

where $R(X_n - c)$ is the error. Note that $R(0) = 0$ (when $x = c$), and also that $R(h) = o(|h|)$ as $|h| \to 0$.[11] That is,

$$\frac{R(h)}{|h|} \to 0. \tag{1.18}$$

Rearranging and then multiplying both sides by τ_n yields:

$$\begin{aligned} g(X_n) - g(c) &= Dg(c)(X_n - c) + R(X_n - c) \\ \Rightarrow \tau_n (g(X_n) - g(c)) &= Dg(c) \tau_n (X_n - c) + \tau_n R(X_n - c). \end{aligned}$$

If we can show that $\tau_n R(X_n - c) \xrightarrow{p} 0$, then we can use the Slutsky's Lemma to argue that $Dg(c) \tau_n (X_n - c) \xrightarrow{d} Dg(c) X$.

We first argue that $|X_n - c| \xrightarrow{p} 0$. Multiplying by τ_n / τ_n while noting that $\tau_n (X_n - c) \xrightarrow{d} X$ and $1/\tau_n \to 0$, we see that

$$\frac{1}{\tau_n} \tau_n |X_n - c| \xrightarrow{d/p} 0 \times |X| = 0.$$

by Slutsky's Lemma.

Now we can rewrite $\tau_n R(X_n - c)$ as:

$$\tau_n R(X_n - c) = \tau_n |X_n - c| \, b\left(\frac{R(X_n - c)}{|X_n - c|}\right),$$

where

$$b\left(\frac{R(X_n - c)}{|X_n - c|}\right) = \begin{cases} \frac{R(X_n - c)}{|X_n - c|} & X_n \neq c \\ 0 & X_n = c \end{cases}.$$

so that $b(\cdot)$ is continuous. We can set b this way since we would still have that $\tau_n R(0) = 0$. But, notice that (1.18) implies that $b \to 0$ as $|X_n - c| \to 0$ which we showed is the case as $n \to \infty$. Since we have $\tau_n (X_n - c) \xrightarrow{d} X$ and $b(\cdot) \xrightarrow{d} 0$, by Slutsky's Lemma,

$$\tau_n R(X_n - c) = \tau_n |X_n - c| \, b\left(\frac{R(X_n - c)}{|X_n - c|}\right) \xrightarrow{d} |X| \times 0 = 0. \qquad \blacksquare$$

Remark. Note that the theorem holds even if $Dg(c)$ is singular.

[11] This means that $R(h) \to 0$ faster than $|h| \to 0$; i.e. $\lim_{h \to 0} R(h) / |h| = 0$.

Example 1.13. Suppose $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} P = \text{Bernoulli}(q)$, with $q \in (0,1)$. Then, by the Central Limit Theorem,

$$\sqrt{n}\left(\bar{X}_n - q\right) \xrightarrow{\text{d}} N\left(0, g\left(q\right)\right),$$

where $g(q) = q(1-q)$. A natural estimator for $g(q)$ is $g\left(\bar{X}_n\right)$. Since, $Dg(q) = 1 - 2q$, the Delta Method implies that

$$\sqrt{n}\left(g\left(\bar{X}_n\right) - g\left(q\right)\right) \xrightarrow{\text{d}} Dg\left(q\right) N\left(0, g\left(q\right)\right) = N\left(0, \left(1-2q\right)^2 q\left(1-q\right)\right).$$

When $q = 1/2$, we have that

$$\sqrt{n}\left(g\left(\bar{X}_n\right) - g\left(q\right)\right) \xrightarrow{\text{d}} N\left(0, 0\right) = 0.$$

That is, $\sqrt{n}\left(g\left(\bar{X}_n\right) - g\left(q\right)\right) \xrightarrow{\text{p}} 0$. We get a degenerate limit distribution because the Delta Method considers Taylor expansion only up to the first order. To get a non-degenerate limiting distribution, we use second-order Taylor expansion, which gives us

$$g\left(\bar{X}_n\right) - g\left(q\right) = Dg\left(q\right)\left(\bar{X}_n - q\right) + \frac{D^2 g\left(q\right)}{2}\left(\bar{X}_n - q\right)^2 + R\left(\bar{X}_n - q\right),$$

where $R(0) = 0$ and $R(h) = o\left(h^2\right)$; i.e. as $h \to 0$,

$$\frac{R\left(h\right)}{h^2} \to 0. \tag{1.19}$$

Since $D^2 g(q) = -2$, when $q = 1/2$, above simplifies to

$$g\left(\bar{X}_n\right) - g\left(q\right) = -\left(\bar{X}_n - q\right)^2 + R\left(\bar{X}_n - q\right).$$

Multiplying both sides by n, we obtain

$$n\left(g\left(\bar{X}_n\right) - g\left(q\right)\right) = -n\left(\bar{X}_n - q\right)^2 + nR\left(\bar{X}_n - q\right).$$

Consider the first term on the right-hand side. Rearranging and using the Continuous Mapping Theorem gives that:

$$\begin{aligned} -n\left(\bar{X}_n - q\right)^2 &= -\left[\sqrt{n}\left(\bar{X}_n - q\right)\right]^2 \\ &\xrightarrow{\text{d}} -\left[N\left(0, \frac{1}{4}\right)\right]^2 \overset{\text{d}}{=} -\left[\frac{1}{2}N\left(0,1\right)\right]^2 \\ &\overset{\text{d}}{=} -\left[\frac{1}{2}N\left(0,1\right)\right]^2 \overset{\text{d}}{=} -\frac{1}{4}\chi_1^2. \end{aligned}$$

We can also write the second as

$$nR\left(\bar{X}_n - q\right) = n\left(\bar{X}_n - q\right)^2 b\left(\frac{R\left(\bar{X}_n - q\right)}{\left(\bar{X}_n - q\right)^2}\right),$$

where b is defined as before to ensure that it is continuous at $X_n = q$. By (1.19), and the fact that $\bar{X}_n \xrightarrow{\text{p}} q$, we know that

$$b\left(\frac{R\left(\bar{X}_n - q\right)}{\left(\bar{X}_n - q\right)^2}\right) \to 0 \Rightarrow b\left(\frac{R\left(\bar{X}_n - q\right)}{\left(\bar{X}_n - q\right)^2}\right) \xrightarrow{\text{d}} 0.$$

And, from before, we have already shown that

$$n\left(\bar{X}_n - q\right)^2 \xrightarrow{\text{d}} \frac{1}{4}\chi^2.$$

Then, by Slutsky's Lemma, we must have

$$nR\left(\bar{X}_n - q\right) = n\left(\bar{X}_n - q\right)^2 b\left(\frac{R\left(\bar{X}_n - q\right)}{\left(\bar{X}_n - q\right)^2}\right) \xrightarrow{\text{d}} \frac{1}{4}\chi_1^2 \times 0 = 0.$$

Together, we showed that, when $q = 1/2$,

$$n\left(g\left(\bar{X}_n\right) - g\left(q\right)\right) \xrightarrow{\text{d}} -\frac{1}{4}\chi_1^2.$$

1.6.6 Correlation

Suppose $\left(X_1, Y_1\right), \left(X_2, Y_2\right), \ldots, \left(X_n, Y_n\right) \overset{\text{iid}}{\sim} P$ on $\mathbb{R}^2$ and that $\mathbb{E}\left[X_i^2\right], \mathbb{E}\left[Y_i^2\right] < \infty$.

Definition 1.10. *(Correlation)* Suppose $\text{Var}\left[X_i\right], \text{Var}\left[Y_i\right] > 0$. Then, correlation between X_i and Y_i is defined as

$$\rho_{X,Y}\left(P\right) \coloneqq \frac{\text{Cov}\left[X_i, Y_i\right]}{\sqrt{\text{Var}\left[X_i\right]}\sqrt{\text{Var}\left[Y_i\right]}}.$$

$\rho_{X,Y}\left(P\right)$ measures the strength of linear relationship between X and Y.

Remark 1.5. *(Existence of covariance).* Recall the definition of covariance.

$$\begin{aligned}\text{Cov}\left[X_i, Y_i\right] &\coloneqq \mathbb{E}\left[\left(X_i - \mathbb{E}\left[X_i\right]\right)\left(Y_i - \mathbb{E}\left[Y_i\right]\right)\right] \\ &= \mathbb{E}\left[X_i Y_i - \mathbb{E}\left[X_i\right] Y_i - X_i \mathbb{E}\left[Y_i\right] + \mathbb{E}\left[X_i\right]\mathbb{E}\left[Y_i\right]\right] \\ &= \mathbb{E}\left[X_i Y_i\right] - \mathbb{E}\left[X_i\right]\mathbb{E}\left[Y_i\right].\end{aligned}$$

Recall that existence of $\mathbb{E}\left[X_i^2\right]$ and $\mathbb{E}\left[Y_i^2\right]$ imply existence of $\mathbb{E}\left[\left|X_i\right|\right]$ and $\mathbb{E}\left[\left|Y_i\right|\right]$. But do they also imply existence of $\mathbb{E}\left[\left|X_i Y_i\right|\right]$? That is, does existence of $\mathbb{E}\left[X_i^2\right]$ and $\mathbb{E}\left[Y_i^2\right]$ imply existence of $\text{Cov}\left[X_i, Y_i\right]$?

First note that, in general,

$$\begin{aligned}&\left(\left|u\right| - \left|v\right|\right)^2 \geq 0 \\ \Leftrightarrow &\left|u\right|^2 - 2\left|u\right|\left|v\right| + \left|v^2\right| \geq 0 \\ \Leftrightarrow &\frac{1}{2}\left|u^2\right| + \frac{1}{2}\left|v^2\right| \geq \left|u\right|\left|v\right| = \left|uv\right|.\end{aligned}$$

Letting $u = X_i$ and $v = Y_i$, above implies that

$$\begin{aligned}&\left|X_i Y_i\right| \leq \frac{1}{2}\left|X_i^2\right| + \frac{1}{2}\left|Y_i^2\right| \\ \Rightarrow &\mathbb{E}\left[\left|X_i Y_i\right|\right] \leq \frac{1}{2}\mathbb{E}\left[X_i^2\right] + \frac{1}{2}\mathbb{E}\left[Y_i^2\right] < \infty.\end{aligned}$$

That is, if $\mathbb{E}\left[X_i^2\right], \mathbb{E}\left[Y_i^2\right] < \infty$, then $\mathbb{E}\left[\left|X_i Y_i\right|\right] < \infty$ and so covariance exists.

We prove some properties of correlation using the following inequality. Note that Cauchy-Schwartz is useful in many other contexts.

Lemma 1.9. *(Cauchy-Schwartz Inequality) For any random variables u and v with $\mathbb{E}\left[u^2\right], \mathbb{E}\left[v^2\right] < \infty$,*

$$\mathbb{E}\left[uv\right]^2 \leq \mathbb{E}\left[u^2\right]\mathbb{E}\left[v^2\right]. \tag{1.20}$$

Furthermore, inequality is strict unless there exists α such that $\mathbb{P}\left(u = \alpha v\right) = 1$.

Proof. Suppose either $\mathbb{E}\left[u^2\right]$ or $\mathbb{E}\left[v^2\right]$ equals zero. Then the right-hand side of (1.20) is clearly zero. Notice also that

$$\mathbb{E}\left[u^2\right] = 0 \Rightarrow \mathbb{E}\left[u\right] = 0.$$

Hence, in this case, the (1.20) holds.

Now suppose that $\mathbb{E}\left[u^2\right] > 0$ and $\mathbb{E}\left[v\right]^2 > 0$. For any α, note that $\mathbb{E}\left[(u - \alpha v)^2\right] \geq 0$. Notice that

$$\mathbb{E}\left[(u - \alpha v)^2\right] = \mathbb{E}\left[u^2\right] - 2\alpha\mathbb{E}\left[uv\right] + \alpha^2\mathbb{E}\left[v^2\right]$$

is minimised when

$$\alpha = \frac{\mathbb{E}\left[uv\right]}{\mathbb{E}\left[v^2\right]}.$$

At this value of α, it must still be the case that $\mathbb{E}\left[(u - \alpha v)^2\right] \geq 0$; i.e.

$$\begin{aligned}
\mathbb{E}\left[u^2\right] - 2\frac{\mathbb{E}\left[uv\right]^2}{\mathbb{E}\left[v^2\right]} + \left(\frac{\mathbb{E}\left[uv\right]}{\mathbb{E}\left[v^2\right]}\right)^2 \mathbb{E}\left[v^2\right] &\geq 0 \\
\Leftrightarrow \mathbb{E}\left[u^2\right] - \frac{\mathbb{E}\left[uv\right]^2}{\mathbb{E}\left[v^2\right]} &\geq 0 \\
\Leftrightarrow \mathbb{E}\left[u^2\right]\mathbb{E}\left[v^2\right] &\geq \mathbb{E}\left[uv\right]^2.
\end{aligned}$$

■

Proposition 1.7. *Suppose $\rho_{X,Y}(P)$ exists. Then,*

$$\left|\rho_{X,Y}(P)\right| \leq 1.$$

Moreover, $\left|\rho_{X,Y}(P)\right| = 1$ if and only if there exists $a, b \in \mathbb{R}$ such that $\mathbb{P}(a + bX_i = Y_i) = 1$.

Proof. The Cauchy-Schwartz's inequality gives that

$$\frac{\mathbb{E}\left[uv\right]^2}{\mathbb{E}\left[u^2\right]\mathbb{E}\left[v^2\right]} \leq 1. \tag{1.21}$$

First, note that adding or subtracting constants do not alter variance or covariance. Define $\tilde{X}_i \coloneqq X_i - \mu_X$ and $\tilde{Y}_i \coloneqq Y_i - \mu_Y$ so that $\mathbb{E}\left[\tilde{X}_i\right] = \mathbb{E}\left[\tilde{Y}_i\right] = 0$, then

$$\begin{aligned}
\rho_{X,Y}(P) &= \frac{\mathrm{Cov}\left[X_i, Y_i\right]}{\sqrt{\mathrm{Var}\left[X_i\right]}\sqrt{\mathrm{Var}\left[Y_i\right]}} = \rho_{\tilde{X},\tilde{Y}}(P) \\
&= \frac{\mathrm{Cov}\left[\tilde{X}_i, \tilde{Y}_i\right]}{\sqrt{\mathrm{Var}\left[\tilde{X}_i\right]}\sqrt{\mathrm{Var}\left[\tilde{Y}_i\right]}} \\
&= \frac{\mathbb{E}\left[\tilde{X}_i\tilde{Y}_i\right]}{\sqrt{\mathbb{E}\left[\tilde{X}_i^2\right]}\sqrt{\mathbb{E}\left[\tilde{Y}_i^2\right]}} \\
\Rightarrow \rho_{\tilde{X},\tilde{Y}}(P)^2 &= \frac{\mathbb{E}\left[\tilde{X}_i\tilde{Y}_i\right]^2}{\mathbb{E}\left[\tilde{X}_i^2\right]\mathbb{E}\left[\tilde{Y}_i^2\right]}.
\end{aligned}$$

Using (1.21), we can now write

$$\rho_{\tilde{X},\tilde{Y}}(P)^2 \leq 1 \Rightarrow \left|\rho_{X,Y}(P)\right| \leq 1.$$

Recall also that, if and only if $\mathbb{P}(u = \alpha v) = 1$, the Cauchy-Schwart'z inequality holds with equality. Here, $u = \alpha v$ is equivalent to

$$\begin{aligned}\tilde{X}_i = \alpha \tilde{Y}_i \Leftrightarrow & X_i - \mu_X = \alpha (Y_i - \mu_Y) \\ \Leftrightarrow & \underbrace{\frac{\mu_X}{-\alpha}}_{=a} + \underbrace{\frac{\mu_Y}{\alpha}}_{=b} X_i = Y_i.\end{aligned}$$

Hence, $|\rho_{X,Y}(P)| = 1$ if and only if there exists $a, b \in \mathbb{R}$ such that $\mathbb{P}(a + bX_i = Y_i) = 1$. ■

A natural estimator of $\rho_{X,Y}(P)$ is

$$\hat{\rho}_{X,Y} = \frac{\hat{\sigma}_{X,Y,n}}{S_{X,n} S_{Y,n}},$$

where

$$\begin{aligned}\hat{\sigma}_{X,Y,n} &= \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{X}_n\right)\left(Y_i - \bar{Y}_n\right), \\ S_{X,n} &= \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{X}_n\right)^2}, \\ S_{Y,n} &= \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} \left(Y_i - \bar{Y}_n\right)^2}.\end{aligned}$$

We can then show, using the Delta Method, that if $\mathbb{E}\left[X_i^4\right] < \infty$, then

$$\sqrt{n}\left(\hat{\rho}_{X,Y,n} - \rho_{X,Y}(P)\right) \xrightarrow{\mathrm{d}} N(0, v),$$

where v depends on τ and fourth centred moment. Note that we can use the Delta Method because $\hat{\rho}_{X,Y,n}$ is a smooth function of averages of $\left(X_i, Y_i, X_i Y_i, X_i^2, Y_i^2\right)$.

1.6.7 Deriving a test statistic for the median

Suppose $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} P$ on $\mathbb{R}$ and the CDF is given by F. Then, we can define the median as

$$\theta \coloneqq \inf\left\{x \in \mathbb{R} : F(x) \geq \frac{1}{2}\right\}.$$

Define the estimator as

$$\hat{\theta}_n \coloneqq \inf\left\{x \in \mathbb{R} : \hat{F}_n(x) \geq 0.5\right\},$$

where $\hat{F}_n(x) \coloneqq \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{X_i \leq x\}}$ is the empirical CDF. Before providing the proof, we introduce the following theorem.

Theorem 1.9. *(Berry-Esseen Central Limit Theorem) Suppose $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} P$ on $\mathbb{R}$ with $0 < \sigma^2(P) < \infty$. Then,*

$$\sup_{x \in \mathbb{R}} \left| \mathbb{P}\left(\frac{\sqrt{n}\left(\bar{X}_n - \mu(P)\right)}{\sigma(P)} \leq x\right) - \Phi(x) \right| \leq \frac{c}{\sqrt{n}} \frac{\mathbb{E}\left[\left|X_i - \mu(P)\right|^3\right]}{\sigma^3(P)},$$

where c is a constant and does not depend on n or P.

Remark. Note that $\sup|\cdot|$ implies uniform convergence.

Equipped with the theorem, we can prove the following.

Proposition 1.8. *(Limiting distribution of the median) Suppose* $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} P$ *on* $\mathbb{R}$ *and the CDF is given by* F*, which is continuous and differentiable at* θ*. Then,*

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow{\text{d}} N\left(0, \frac{1}{4f^2(\theta)}\right).$$

Proof. First, notice that we cannot use the Delta Method here since θ is not smooth.

Suppose that n is even (the proof would be almost identical if it was odd), this means that $\hat{\theta}_n$ is given by the $n/2$th highest X_i's (since θ is defined using inf). Consider

$$\begin{aligned}\mathbb{P}\left(\sqrt{n}\left(\hat{\theta}_n - \theta\right) \leq x\right) &= \mathbb{P}\left(\hat{\theta}_n \leq \theta + \frac{x}{\sqrt{n}}\right) \\ &= \mathbb{P}\left(\sum_{i=1}^{n} Z_i \leq \frac{n}{2}\right)\end{aligned}$$

where $Z_i \coloneqq \mathbf{1}_{\left\{X_i > \theta + \frac{x}{\sqrt{n}}\right\}}$. The last equality follows because when n is even, $\hat{\theta}_n$ is the $n/2 - 1$th element so that there can be at most $n/2$ observations larger than $n/2$th element. Then, since F is continuous at θ, it must mean that $F(\theta) = 1/2$.[12]

Let

$$\begin{aligned}\mu_n &\coloneqq \mathbb{E}[Z_i] = \mathbb{E}\left[\mathbf{1}_{\left\{X_i > \theta + \frac{x}{\sqrt{n}}\right\}}\right] = 1 - F\left(\theta + \frac{x}{\sqrt{n}}\right) \to \frac{1}{2}, \\ \sigma_n^2 &\coloneqq \operatorname{Var}[Z_i] = \mu_n(1 - \mu_n) \to \frac{1}{4}.\end{aligned}$$

Then

$$\begin{aligned}\mathbb{P}\left(\sum_{i=1}^{n} Z_i \leq \frac{n}{2}\right) &= \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i \leq \frac{1}{2}\right) \\ &= \mathbb{P}\left(\bar{Z}_n \leq \frac{1}{2}\right) \\ &= \mathbb{P}\left(\frac{\sqrt{n}\left(\bar{Z}_n - \mu_n\right)}{\sigma_n} \leq \underbrace{\frac{\sqrt{n}\left(\frac{1}{2} - \mu_n\right)}{\sigma_n}}_{\coloneqq z_n}\right).\end{aligned}$$

We can write z_n as

$$z_n = \frac{\sqrt{n}\left(\frac{1}{2} - \mu_n\right)}{\sigma_n}.$$

[12] Suppose that $F(\theta) < 0.5$; i.e. $\theta \notin \{x \in \mathbb{R} : F(x) \geq 0.5\}$. Then, since $F(x)$ is continuous, in particular, it is right-continuous, there exists $\varepsilon > 0$ such that

$$F(\theta + \varepsilon) < 0.5.$$

but then this means that $\theta + \varepsilon$ is a greater lower bound than θ, contradicting the definition of θ as the greatest lower bound of the $\{x \in \mathbb{R} : F(x) \geq 0.5\}$. Hence, $F(\theta) \geq 0.5$.

Now suppose that $F(\theta) > 0.5$. Then, since $F(x)$ is continuous, in particular, it is left-continuous, there exists $\varepsilon > 0$ such that

$$F(\theta - \varepsilon) > 0.5.$$

But since $\theta > \theta - \varepsilon$, then θ is not a lower bound, which contradicts its definition as the infimum. Hence, $F(\theta) \leq 0.5$.

Together, it must be the case that $F(\theta) = 0.5$.

First, note that, since $\sigma_n \to 1/2$. Using the fact that $F(\theta) = 1/2$ and $\mu_n = 1 - F(\theta + x/\sqrt{n})$:

$$\begin{aligned}\sqrt{n}\left(\frac{1}{2} - \mu_n\right) &= \sqrt{n}\left(\frac{1}{2} - \mu_n + \frac{1}{2} - \frac{1}{2}\right) \\ &= \sqrt{n}\left(1 - \mu_n - \frac{1}{2}\right) \\ &= \sqrt{n}\left(1 - \left(1 - F\left(\theta + \frac{x}{\sqrt{n}}\right)\right) - F(\theta)\right) \\ &= \sqrt{n}\left(F\left(\theta + \frac{x}{\sqrt{n}}\right) - F(\theta)\right).\end{aligned}$$

Recall that Taylor expansion to the first-order is:

$$F(a) - F(c) = f(c)(x - c) + R(a - c),$$

where $R(h)/|h| \to 0$. Using this

$$\sqrt{n}\left(F\left(\theta + \frac{x}{\sqrt{n}}\right) - F(\theta)\right) \to f(\theta)\sqrt{n}\frac{x}{\sqrt{n}} = xf(\theta).$$

We therefore can conclude that, as $n \to \infty$,

$$z_n \to \frac{xf(\theta)}{1/2} = 2xf(\theta). \tag{1.22}$$

By the Berry-Esseen Central Limit Theorem

$$\begin{aligned}&\sup_{c\in\mathbb{R}}\left|\mathbb{P}\left(\frac{\sqrt{n}\left(\bar{Z}_n - \mu_n\right)}{\sigma_n} \le z_n\right) - \Phi(z_n)\right| \le \frac{c}{\sqrt{n}}\frac{\mathbb{E}\left[|Z_i - \mu_n|^3\right]}{\sigma^3(P)} \\ &\Rightarrow \left|\mathbb{P}\left(\frac{\sqrt{n}\left(\bar{Z}_n - \mu_n\right)}{\sigma_n} \le z_n\right) - \Phi(z_n)\right| \le \frac{c}{\sqrt{n}}\frac{\mathbb{E}\left[|Z_i - \mu_n|^3\right]}{\sigma^3(P)}.\end{aligned}$$

Note that, for any finite n, $|(Z_i - \mu_n)^3| \le 1$ so that the right-hand side must tend to zero as $n \to \infty$, which also implies that

$$\mathbb{P}\left(\frac{\sqrt{n}\left(\bar{Z}_n - \mu_n\right)}{\sigma_n} \le z_n\right) \to \Phi(z_n).$$

Since we know (1.22), we therefore have that

$$\mathbb{P}\left(\frac{\sqrt{n}\left(\bar{Z}_n - \mu_n\right)}{\sigma_n} \le z_n\right) = \mathbb{P}\left(\sqrt{n}\left(\hat{\theta}_n - \theta\right) \le x\right) \to \Phi(2xf(\theta)).$$

Since $\Phi(2xf(\theta))$ is the CDF of $N\left(0, 1/4f^2(\theta)\right)$, we have that

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow{\text{d}} N\left(0, \frac{1}{4f^2(\theta)}\right).$$

■

Remark. We cannot use the standard Central Limit Theorem that requires the same sample to be repeated Here, we have a Triangular array of of data, which means that each X_i needs to be sub-indexed:

$$\begin{array}{cccccc} n\backslash i & X_1 & X_2 & \cdots & X_n \\ n=1 & X_{1,1} & X_{2,1} & \cdots & X_{n,1} \\ n=2 & X_{1,2} & X_{2,2} & \cdots & X_{n,2} \\ n=3 & X_{1,3} & X_{2,3} & \cdots & X_{n,3} \\ \vdots & \vdots & \vdots & & \vdots \end{array}.$$

1.7 Tightness

A sequence of random variables may not converge in distribution but may satisfy a weaker property called *tightness* (or boundedness in probability). For example, suppose $X \sim N\left(\mu, \sigma^2\right)$. Then, for any $p \in (0,1)$, there exists finite $x > 0$ such that

$$\mathbb{P}\left(|X| \leq x\right) = \mathbb{P}\left(\{X \leq x\} \cup \{-X \geq x\}\right) = p.$$

Definition 1.11. *(Tightness) A sequence of random vectors* $\{X_n : n \geq 1\}$ on $\mathbb{R}^k$ is *tight* if, for any $\varepsilon > 0$, there exists a (finite) constant $B > 0$ such that

$$\inf_n \mathbb{P}\left(|X_n| \leq B\right) \geq 1 - \varepsilon,$$

where $|\cdot|$ is the Euclidean norm. Equivalently, $\{X_n : n \geq 1\}$ is tight if, for any $\varepsilon > 0$, there exists a finite constant $M_\varepsilon < \infty$ and $n_\varepsilon \in \mathbb{N}$ such that

$$\mathbb{P}\left(|X_n| > M_\varepsilon\right) < \varepsilon, \forall n \geq n_\varepsilon.$$

Proposition 1.9. *If* $X_n \xrightarrow{\text{d}} X$*, then* X_n *is tight.*

Proof. (Proposition (1.9)) Since taking the norm is a continuous operation, by the Continuous Mapping Theorem,

$$X_n \xrightarrow{\text{d}} X \Rightarrow |X_n| \xrightarrow{\text{d}} |X|.$$

Since *any finite sequence of random vectors are tight*, there exists $B > 0$ such that

$$\mathbb{P}\left(|X| < B\right) \geq 1 - \frac{\varepsilon}{2} \tag{1.23}$$

for any $\varepsilon > 0$. Fix ε and taking $B > 0$ that satisfies (1.23), then

$$\begin{aligned} \mathbb{P}\left(|X_n| \leq B\right) &\geq \mathbb{P}\left(|X_n| < B\right), \ \forall n \\ \Rightarrow \liminf_{n\to\infty} \mathbb{P}\left(|X_n| \leq B\right) &\geq \liminf_{n\to\infty} \mathbb{P}\left(|X_n| < B\right) \\ &\geq \mathbb{P}\left(|X| < B\right) \\ &\geq 1 - \frac{\varepsilon}{2} \end{aligned}$$

where the third line follows from Portmanteau's Lemma (v) and the last line uses (1.23).

Now, recall that $\liminf X_n \equiv \lim_{N\to\infty} \inf_{n>N} X_n$. So,

$$\liminf_{n\to\infty} \mathbb{P}\left(|X_n| \leq B\right) \geq 1 - \frac{\varepsilon}{2}$$

means that as $N \to \infty$, $\inf_{n>N} \mathbb{P}\left(|X_n| \leq B\right)$ is greater than (or equal to) $1 - \varepsilon/2$. This means that there exists $N \in \mathbb{N}$ such that

$$\mathbb{P}\left(|X_n| \leq B\right) \geq 1 - \varepsilon, \ \forall n \geq N.$$

For each $|X_n|$, $n < N$, they are individually tight so that we can find $B_n > 0$ such that

$$\mathbb{P}\left(|X_n| \leq B_n\right) \geq 1 - \varepsilon, \ \forall n < N.$$

Define $B^* = \max\left\{B_1, B_2, \ldots, B_{N-1}, B\right\} > 0$, then

$$\inf_n \mathbb{P}\left(|X_n| \leq B^*\right) \geq 1 - \varepsilon.$$

Alternatively, denote $F_{|X_n|}$ as the CDF of $|X_n|$, then

$$\mathbb{P}\left(|X_n| > M_\varepsilon\right) = 1 - F_{|X_n|}\left(M_\varepsilon\right).$$

Since $|X_n| \xrightarrow{\text{d}} |X|$, then

$$\mathbb{P}\left(|X_n| > M_\varepsilon\right) = 1 - F_{|X_n|}\left(M_\varepsilon\right) \to 1 - F_{|X|}\left(M_\varepsilon\right)$$

Now, choose M_ε sufficiently large such that $F_{|X|}\left(M_\varepsilon\right) \to 1$. That is, we can choose M_ε so that

$$\mathbb{P}\left(|X_n| > M_\varepsilon\right) < \varepsilon. \qquad \blacksquare$$

Proposition (1.9) establishes that convergence in distribution implies tightness. However, the converse is generally not true. In other words, tightness is a weaker notion than convergence in distribution.

Example 1.14. *(Tightness does not imply convergence in distribution) Suppose* $X_{2n} \sim \text{Uniform}\left[0,1\right]$ *and* $X_{2n+1} \sim \text{Uniform}\left[2,3\right]$ *so that even and odd elements in the sequence have different distributions. Clearly, the sequence does not converge in distribution. However, the sequence is tight. For example, for any* $\varepsilon > 0$*, we can take* $B = 3$ *so that*

$$\liminf_{n\to\infty} \mathbb{P}\left(|X_n| \leq 3\right) = 1 > 1 - \varepsilon.$$

The following theorem gives a partial converse of Proposition (1.9).

Theorem 1.10. *(Prokhorov's Theorem) If a sequence of random vectors* $\{X_n : n \geq 1\}$ *is tight, then there exists* $\{X_{n_j}\}_{j=1}^{\infty}$ *and* X *such that* $X_{n_j} \xrightarrow{\text{d}} X$*.*

1.7.1 τ_n-consistency

Recall that $\hat{\theta}_n$ is consistent for $\theta\left(P\right)$ if $\hat{\theta}_n \xrightarrow{\text{p}} \theta\left(p\right)$. An even stronger notion of convergence is that $\tau_n(\hat{\theta}_n - \theta\left(P\right))$ is tight for $\tau_n \to \infty$; i.e. even if you "blow up" the sequence, the sequence still is bounded in probability. The proposition below shows the sense in which τ_n-consistency is a stronger notion that consistency.

Definition 1.12. *(τ_n-consistency)* If the sequence $\tau_n(\hat{\theta}_n - \theta\left(P\right))$ is tight for $\tau_n \to \infty$, we say that $\hat{\theta}_n$ is τ_n-consistent for $\theta\left(P\right)$.

Proposition 1.10. *Let* $\hat{\theta}_n$ *be* τ_n*-consistent for* $\theta\left(P\right)$*, then* $\hat{\theta}_n$ *is consistent for* θ*.*

Proof. (Proposition 1.10) We want to show that if $\tau_n(\hat{\theta}_n - \theta\left(P\right))$ is tight for $\tau_n \to \infty$, then $\hat{\theta}_n \xrightarrow{\text{p}} \theta$. That is, as $n \to \infty$,

$$\mathbb{P}\left(\left|\hat{\theta}_n - \theta\left(P\right)\right| > \varepsilon\right) \to 0, \ \forall \varepsilon > 0$$

Equivalently, we want to show that for any $\varepsilon, \epsilon > 0$, there exists $N \in \mathbb{N}$ such that

$$\mathbb{P}\left(\left|\hat{\theta}_n - \theta\left(P\right)\right| > \varepsilon\right) \leq \epsilon, \forall n \geq N.$$

If $\tau_n(\hat{\theta}_n - \theta\left(P\right))$ is tight then, for any $\epsilon > 0$, there exists a constant $B > 0$ such that

$$\begin{aligned} 1 - \epsilon &\leq \inf_n \mathbb{P}\left(\tau_n \left|\left(\hat{\theta}_n - \theta\left(P\right)\right)\right| \leq B\right) \\ &= \inf_n \mathbb{P}\left(\left|\left(\hat{\theta}_n - \theta\left(P\right)\right)\right| \leq \frac{B}{\tau_n}\right). \end{aligned}$$

(We assume n is sufficiently large so that $\tau_n > 0$.) Since $\tau_n \to \infty$ as $n \to \infty$, there exists $N \in \mathbb{N}$ such that

$$\frac{B}{\tau_n} \leq \varepsilon, \ \forall n \geq N.$$

Then,

$$\begin{aligned} 1-\epsilon &\leq \inf_n \mathbb{P}\left(\left|\left(\hat{\theta}_n - \theta\left(P\right)\right)\right| \leq \frac{B}{\tau_n}\right) \\ &\leq \inf_{n\geq N} \mathbb{P}\left(\left|\left(\hat{\theta}_n - \theta\left(P\right)\right)\right| \leq \frac{B}{\tau_n}\right) \\ &\leq \inf_{n\geq N} \mathbb{P}\left(\left|\left(\hat{\theta}_n - \theta\left(P\right)\right)\right| \leq \varepsilon\right). \end{aligned}$$

Because

$$\inf_{n\geq N} \mathbb{P}\left(\left|\left(\hat{\theta}_n - \theta\left(P\right)\right)\right| \leq \varepsilon\right) \leq \mathbb{P}\left(\left|\left(\hat{\theta}_n - \theta\left(P\right)\right)\right| \leq \varepsilon\right),\ \forall n \geq N,$$

it follows that

$$\begin{aligned} & 1-\epsilon \leq \mathbb{P}\left(\left|\left(\hat{\theta}_n - \theta\left(P\right)\right)\right| \leq \varepsilon\right),\ \forall n \geq N \\ \Leftrightarrow\ & 1-\mathbb{P}\left(\left|\left(\hat{\theta}_n - \theta\left(P\right)\right)\right| \leq \varepsilon\right) \leq \epsilon,\ \forall n \geq N \\ \Leftrightarrow\ & \mathbb{P}\left(\left|\left(\hat{\theta}_n - \theta\left(P\right)\right)\right| > \varepsilon\right) \leq \epsilon,\ \forall n \geq N; \end{aligned}$$

i.e. $\hat{\theta}_n$ is consistent for θ. ■

Recall that, under the conditions of the Central Limit Theorem, $\sqrt{n}(\bar{x}_n - \mu\left(P\right))$ converges in distribution to a normal distribution; so that $\sqrt{n}(\bar{x}_n - \mu\left(P\right))$ is tight. Thus, we realise that $\bar{x}_n$ is a $\sqrt{n}$-consistent estimator of the population mean.

1.8 Stochastic order

Here, we introduce some useful notations called *stochastic order symbols*.

Notation 1. *(Stochastic order symbols)*

- ▷ If $X_n \xrightarrow{\text{p}} 0$, we write $X_n = o_P\left(1\right)$ (X_n is "little oh p"). More generally, we say that $X_n = o_p\left(R_n\right)$ if $X_n = Y_n R_n$ for some $Y_n = o_P\left(1\right)$.
- ▷ If X_n is tight, we write $X_n = O_P\left(1\right)$ (X_n is "big oh p"). More generally, we say that $X_n = O_P\left(R_n\right)$ if $X_n = Y_n R_n$ for some $Y_n = O_P\left(1\right)$.

These symbols expresses that the sequence X_n converges in probability to zero or is bounded in probability at "rate" R_n. For deterministic sequences, X_n and R_n, the stochastic order symbols refer to the derivatives (which are written as o and O). More generally,

- ▷ $\{X_n : n \geq 1\}$ is $o_P\left(n^\delta\right)$ for $\delta \in \mathbb{R}$ if $n^{-\delta} X_n \xrightarrow{\text{p}} 0$.
- ▷ $\{X_n : n \geq 1\}$ is $O_P\left(n^\delta\right)$ for $\delta \in \mathbb{R}$ if for any $\varepsilon > 0$, there exists a constant $B > 0$ such that

$$\inf_n \mathbb{P}\left(n^{-\delta}\left|X_n\right| \leq B\right) \geq 1-\varepsilon.$$

Example 1.15. Suppose $\{U_i\}_{i=1}^{\infty}$ are iid random vectors with $\mathbb{E}\left[U_i\right] = 0$ and $\text{Var}\left[U_i\right] = \sigma^2$. Then,

- ▷ By WLLN, $n^{-1}\sum U_i \xrightarrow{\text{p}} 0$ so $n^{-1}\sum U_i = o_P\left(1\right)$.
- ▷ By CLT, $n^{-1/2}\sum U_i \xrightarrow{\text{d}} N\left(0, \sigma^2\right)$ so that $\sum U_i = O_P\left(n^{1/2}\right)$ or $n^{-1/2}\sum U_i = O_P\left(1\right)$.

Proposition 1.11. $o_P\left(1\right) \Rightarrow O_P\left(1\right)$.

Proof. Suppose $X_n = o_P(1)$ so that $X_n \xrightarrow{p} 0$. This, in turn, implies that $X_n \xrightarrow{d} 0$. Since we know that convergence in distribution implies tightness, we obtain that $X_n = O_P(1)$; i.e. $o_P(1) \Rightarrow O_P(1)$.

We can show prove this more directly. $X_n = o_P(1)$ means that, for any $\varepsilon > 0$,

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

Let $M_\varepsilon = \varepsilon$ and n_ε sufficiently large such that, for $n > n_\varepsilon$,

$$\mathbb{P}(|X_n| > M_\varepsilon) < \varepsilon.$$

■

The converse is not generally true as can be seen from the following example.

Example 1.16. $(O_P(1) \not\Rightarrow o_P(1))$

▷ Consider a (degenerate) random variable $X_n = (-1)^n$. Then, for all $n \in N$, we have that $|X_n| \leq 1$ so that X_n is bounded, and so X_n is bounded in probability. However, this sequence does not converge to zero so that $X_n \neq o_P(1)$.

▷ Let $X_n, X \sim N(0,1)$ so that $X_n \xrightarrow{d} X$ trivially. Since convergence in distribution implies tightness, $X_n = O_P(1)$. However, we know that convergence in distribution does not imply convergence in probability; i.e. $X_n \neq o_P(1)$.

Proposition 1.12. *(Rules of "Calculus" for o_P and O_P)*

(i) $o_P(1) + o_P(1) = o_P(1)$.

(ii) $o_P(1) + O_P(1) = O_P(1)$.

(iii) $o_P(1)\, O_P(1) = o_P(1)$.

(iv) $(1 + o_P(1))^{-1} = O_P(1)$.

(v) $o_P(O_P(1)) = o_P(1)$.

Proof. We provide the proof in turn below.

(i) Let $X_n, Y_n = o_P(1)$, then $X_n \xrightarrow{p} 0$ and $Y_n \xrightarrow{p} 0$. Then, by Continuous Mapping Theorem (since convergence in marginal probabilities implies convergence in joint probabilities) gives that $X_n + Y_n \xrightarrow{p} 0$. That is, $o_P(1) + o_P(1) = o_P(1)$.

(ii) Let $X_n = o_P(1)$, then $X_n \xrightarrow{p} 0 \Rightarrow X_n \xrightarrow{d} 0$ so that X_n is tight (i.e. $o_P(1) \Rightarrow O_P(1)$); i.e. $X_n = O_P(1)$. Let $Y = O_P(1)$. Then, for any $\varepsilon > 0$, there exists $B_X, B_Y > 0$ such that

$$\inf_n \mathbb{P}(|X_n| \leq B_X) \geq 1 - \frac{\varepsilon}{2},$$
$$\inf_n \mathbb{P}(|Y_n| \leq B_Y) \geq 1 - \frac{\varepsilon}{2}.$$

We want to show that, for any $\varepsilon > 0$, there exists $B > 0$,

$$\inf_n \mathbb{P}(|X_n + Y_n| \leq B) \geq 1 - \varepsilon.$$

Let $B = B_X + B_Y > 0$.

$$\begin{aligned}
\inf_n \mathbb{P}\left(|X_n + Y_n| \leq B\right) &\geq \inf_n \mathbb{P}\left(|X_n| + |Y_n| \leq B_X + B_Y\right) \\
&\geq \inf_n \mathbb{P}\left(\{|X_n| \leq B_X\} \cap \{|Y_n| \leq B_Y\}\right) \\
&\geq \inf_n \left\{\mathbb{P}\left(|X_n| \leq B_X\right) + \mathbb{P}\left(|Y_n| \leq B_Y\right) - 1\right\} \\
&\geq \inf_n \mathbb{P}\left(|X_n| \leq B_X\right) + \inf_n \mathbb{P}\left(|Y_n| \leq B_Y\right) - 1 \\
&\geq \left(1 - \frac{\varepsilon}{2}\right) + \left(1 - \frac{\varepsilon}{2}\right) - 1 = 1 - \varepsilon.
\end{aligned}$$

The first inequality follow since the event $\{|X_n + Y_n| \leq B\}$ contains the event $|X_n| + |Y_n| \leq B$ (Triangle Inequality implies $|X_n + Y_n| \leq |X_n| + |Y_n|$. The second inequality follows since $\{|X_n| + |Y_n| \leq B\}$ contains the event $\{|X_n| \leq B_X\} \cap \{|Y_n| \leq B_Y\}$. The third inequality follows since, for any events A and B, $\mathbb{P}(A \cap B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cup B) \Rightarrow \mathbb{P}(A \cap B) \geq \mathbb{P}(A) + \mathbb{P}(B) - 1$. The penultimate inequality uses the fact that $\inf_n \{a_n + b_n\} \geq \inf_n \{a_n\} + \inf_n \{b_n\}$.

(iii) Let $X_n = o_P(1)$ and $Y_n = O_P(1)$. We want to show that $X_n Y_n = o_P(1)$. Suppose not, so that there exists is $\varepsilon > 0$ such that $\mathbb{P}(|X_n Y_n| > \varepsilon)$ does not converge to 0. This implies that there exists subsequences, indexed by n_j , and a constant $\delta > 0$ such that $\mathbb{P}(|X_{n_j}, Y_{n_j}| > \varepsilon) \to \delta$. By Prokhorov's Theorem, then there exists subsubsequences, indexed by n_{j_k} such that $Y_{n_{j_k}} \xrightarrow{\mathrm{d}} Y$. Since $X_{n_{j_k}} \xrightarrow{\mathrm{p}} 0$, by Slutsky's Lemma, $X_{n_{j_k}} Y_{n_{j_k}} \xrightarrow{\mathrm{p}} 0$, which contradicts the assumptions above.

(iv) Let $X_n = o_P(1)$, then $X_n \xrightarrow{\mathrm{p}} 0$. Let $f(x) = 1/(1+x)$ which is continuous everywhere except at $x = -1$. By the Continuous Mapping Theorem, $f(X_n) \xrightarrow{\mathrm{p}} g(0) = 1$, which, in turn, implies that $f(X_n) \xrightarrow{\mathrm{d}} 1$. Since convergence in distribution implies tightness, we have $f(X_n) = O_P(1)$.

(v) Recall that $X_n = o_p(R_n)$ if $X_n = Y_n R_n$ for some $Y_n = o_P(1)$. Hence, $o_P(O_P(1))$ means that $R_n = O_P(1)$, where $Y_n = o_P(1)$. We therefore want to show that $Y_n R_n = o_P(1)$. Note that $Y_n R_n = o_P(1) O_p(1)$. By property (iii), it follows that $Y_n R_n = o_P(1)$; i.e. $o_P(O_P(1)) = o_P(1)$.

■

Example 1.17. Consider $\{X_{n_j}\} = o_P(1)$ for $j = 1, 2, \ldots, J$. Let $Y_n = \sum_{j=1}^{J} X_n$. Since $o_P(1) + o_P(1) = o_P(1)$, for any finite J, $Y_n = o_P(1)$. What if $J \to \infty$? Then, Y_n may not be $o_P(1)$ or even $O_P(1)$. To see this, suppose $\{X_{n_j}\} = Z_n\left(\frac{j}{n}\right)$ where $Z_n \sim U[0,1]$. Then,

$$Y_n = \sum_{j=1}^{J}\left(X_{n_j}\right) = \sum_{j=1}^{J} Z_n\left(\frac{j}{n}\right) = \frac{Z_n}{n}\frac{J(J+1)}{2} \to \infty$$

as $J, n \to \infty$.

2 Conditional expectations

Let (Y, X) be random vectors such that $Y \in \mathbb{R}$, $\mathbf{X} \in \mathbb{R}^k$ and (for now) $\mathbb{E}\left[Y^2\right] < \infty$. Define

$$\mathbb{M} \coloneqq \left\{m\left(\mathbf{X}\right) : m : \mathbb{R}^k \to \mathbb{R}, \mathbb{E}\left[m^2\left(\mathbf{X}\right)\right] < \infty\right\}.$$

Thus, $m \in \mathbb{M}$ is a function that maps random vectors to a random variable with finite variance.

Definition 2.1. *(Conditional expectation)* The conditional expectation of Y given X, denoted $\mathbb{E}\left[Y|X\right]$ is given by

$$\mathbb{E}\left[Y|X\right] \in \inf_{m \in \mathbb{M}} \mathbb{E}\left[\left(Y - m\left(\mathbf{X}\right)\right)^2\right]. \tag{2.1}$$

Theorem 2.1. *(Orthogonality condition) Let $(Y, \mathbf{X})$ be random vectors such that $Y \in \mathbb{R}$, $\mathbf{X} \in \mathbb{R}^k$, $\mathbb{E}\left[Y^2\right] < \infty$. Then, $m^*\left(\mathbf{X}\right) \in \mathbb{M}$ solves (2.1) if and only if*

$$\mathbb{E}\left[\left(Y - m^*\left(\mathbf{X}\right)\right) m\left(\mathbf{X}\right)\right] = 0, \ \forall m\left(\mathbf{X}\right) \in \mathbb{M}. \tag{2.2}$$

Further, if $\tilde{m}\left(\mathbf{X}\right)$ is another solution to (2.1), then

$$\mathbb{P}\left(\tilde{m}\left(\mathbf{X}\right) = m^*\left(\mathbf{X}\right)\right) = 1.$$

We define $\mathbb{E}\left[Y|X\right]$ to be the function in $\mathbb{M}$ that minimises the squared distance between Y and $m\left(\mathbf{X}\right)$. This is the sense in which $\mathbb{E}\left[Y|\mathbf{X}\right]$ can be thought of as the "best" predictor of Y given $\mathbf{X}$ (given, of course, the square loss function assumed defined in (2.1)). Moreover, it is possible to show that there exists $m^* \in \mathbb{M}$ satisfying (2.1). Theorem 1.10 tells us that conditional expectation satisfies an orthogonality condition. It also gives us the sense in which the conditional expectation is unique.

Proof. (Theorem 1.10) ($\Rightarrow$) Suppose $m^*\left(\mathbf{X}\right) \in \mathbb{M}$ satisfies (2.2). We want to show that $m^*\left(\mathbf{X}\right)$ is a solution to (2.1). We first want to introduce $m^*\left(\mathbf{X}\right)$ into the expression inside inf in (2.1)—as usual, we add and subtract $m^*\left(\mathbf{X}\right)$ in (2.1):[13]

$$\begin{aligned}
\mathbb{E}\left[\left(Y - m\left(\mathbf{X}\right)\right)^2\right] &= \mathbb{E}\left[\left(Y - m^*\left(\mathbf{X}\right) + m^*\left(\mathbf{X}\right) - m\left(\mathbf{X}\right)\right)^2\right] \\
&= \mathbb{E}\left[\left(Y - m^*\left(\mathbf{X}\right)\right)^2\right] + \mathbb{E}\left[\left(m^*\left(\mathbf{X}\right) - m\left(\mathbf{X}\right)\right)^2\right] \\
&\quad + 2\mathbb{E}\left[\left(Y - m^*\left(\mathbf{X}\right)\right)\left(m^*\left(\mathbf{X}\right) - m\left(\mathbf{X}\right)\right)\right].
\end{aligned}$$

If $m^*\left(\mathbf{X}\right)$ satisfies (2.2), then $\mathbb{E}[(Y - m^*\left(\mathbf{X}\right))m\left(\mathbf{X}\right)] = 0$ for any $m\left(\mathbf{X}\right) \in \mathbb{M}$. In particular, since $m^*\left(\mathbf{X}\right) \in \mathbb{M}$, $\mathbb{E}[(Y - m^*\left(\mathbf{X}\right))m^*\left(X\right)] = 0$. Hence, the last term in the expression above must be zero; i.e.

$$\mathbb{E}\left[\left(Y - m\left(\mathbf{X}\right)\right)^2\right] = \mathbb{E}\left[\left(Y - m^*\left(\mathbf{X}\right)\right)^2\right] + \mathbb{E}\left[\left(m^*\left(\mathbf{X}\right) - m\left(\mathbf{X}\right)\right)^2\right]. \tag{2.3}$$

Since $\mathbb{E}[(m^*\left(\mathbf{X}\right) - m\left(\mathbf{X}\right))^2] \geq 0$, we obtain that

$$\mathbb{E}\left[\left(Y - m\left(\mathbf{X}\right)\right)^2\right] \geq \mathbb{E}\left[\left(Y - m^*\left(\mathbf{X}\right)\right)^2\right], \ \forall m\left(\mathbf{X}\right) \in \mathbb{M}.$$

That is, $m^*\left(\mathbf{X}\right)$ solves (2.1).

[13] To see this, first, note that $(x - y)^2 = x^2 - 2xy + y^2$. Define $x \coloneqq (\hat{x} - b)$ and $y \coloneqq (\hat{y} - b)$. Then,

$$\begin{aligned}
(\hat{x} - b)^2 - 2\left(\hat{x} - b\right)\left(\hat{y} - b\right) + (\hat{y} - b)^2 &= \left((\hat{x} - b) - (\hat{y} - b)\right)^2 \\
&= (\hat{x} - \hat{y})^2.
\end{aligned}$$

($\Leftarrow$) Now suppose that $m^*(\mathbf{X})$ solves (2.1). For any $\alpha \in \mathbb{R}$, it must be that

$$\mathbb{E}\left[\left(Y - m^*(X)\right)^2\right] \leq \mathbb{E}\left[\left(Y - m^*(X) - \alpha m(X)\right)^2\right].$$

(Similar trick to how we proved the Cauchy-Schwartz inequality.) Expanding the right-hand side,

$$\begin{aligned}&\mathbb{E}\left[\left(Y - m^*(\mathbf{X}) - \alpha m(\mathbf{X})\right)^2\right]\\ =&\mathbb{E}\left[\left(Y - m^*(\mathbf{X})\right)^2\right] + \alpha^2\mathbb{E}\left[\left(m(\mathbf{X})\right)^2\right] - 2\alpha\mathbb{E}\left[\left(Y - m^*(\mathbf{X})\right)m(\mathbf{X})\right]\end{aligned}$$

Thus,

$$\begin{aligned}&0 \leq \alpha^2\mathbb{E}\left[\left(m(\mathbf{X})\right)^2\right] - 2\alpha\mathbb{E}\left[\left(Y - m^*(\mathbf{X})\right)m(\mathbf{X})\right]\\ \Rightarrow\ &0 \geq 2\alpha\mathbb{E}\left[\left(Y - m^*(\mathbf{X})\right)m(\mathbf{X})\right] - \alpha^2\mathbb{E}\left[\left(m(\mathbf{X})\right)^2\right].\end{aligned}$$

The only possible way in which the expression above holds for any $\alpha \in \mathbb{R}$ is if

$$\mathbb{E}\left[\left(Y - m^*(\mathbf{X})\right)m(\mathbf{X})\right] = 0.$$

That is, if (2.2) holds.

Now suppose that $\tilde{m}(\mathbf{X})$ and $m^*(\mathbf{X})$ both solve (2.1). Then,

$$\mathbb{E}\left[\left(Y - \tilde{m}(\mathbf{X})\right)^2\right] = \mathbb{E}\left[\left(Y - m^*(\mathbf{X})\right)^2\right].$$

Recall (2.3) which has to hold for any $m(\mathbf{X})$, in particular, for $\tilde{m}(\mathbf{X})$ so that

$$\mathbb{E}\left[\left(Y - \tilde{m}(\mathbf{X})\right)^2\right] = \mathbb{E}\left[\left(Y - m^*(\mathbf{X})\right)^2\right] + \mathbb{E}\left[\left(m^*(\mathbf{X}) - \tilde{m}(\mathbf{X})\right)^2\right].$$

The two expression imply that

$$\mathbb{E}\left[\left(m^*(\mathbf{X}) - \tilde{m}(\mathbf{X})\right)^2\right] = 0,$$

which implies $\mathbb{P}\left(\tilde{m}(\mathbf{X}) = m^*(x)\right) = 1$. ■

So far, we assumed that $\mathbb{E}\left[Y^2\right] < \infty$. If we instead only assume $\mathbb{E}\left[|Y|\right] < \infty$, then we would define $\mathbb{E}\left[Y|\mathbf{X}\right]$ to be any $m^*(\mathbf{X})$ with $\mathbb{E}\left[|m^*(\mathbf{X})|\right] < \infty$ such that, for any Borel set $\mathbf{B}$ in $\mathcal{B} \in \mathbb{R}^k$,

$$\mathbb{E}\left[\left(Y - m^*(\mathbf{X})\right)\mathbf{1}_{\{\mathbf{X}\in\mathbf{B}\}}\right] = 0. \tag{2.4}$$

The following can be proved using (2.4).

Proposition 2.1. *(Properties of conditional expectations)*

(i) *If $Y = f(X)$, then $\mathbb{E}\left[Y|X\right] = f(X)$.*

(ii) $\mathbb{E}\left[Y + X|Z\right] = \mathbb{E}\left[Y|Z\right] + \mathbb{E}\left[X|Z\right]$.

(iii) $\mathbb{E}\left[f(X)\,Y|X\right] = f(X)\,\mathbb{E}\left[Y|X\right]$.

(iv) *If $\mathbb{P}\left(Y \geq 0\right) = 1$, then $\mathbb{P}\left(\mathbb{E}\left[Y|X\right] \geq 0\right) = 1$.*

(v) *(Law of Iterated Expectations) For any Borel set B in $\mathbb{R}^k$, $\mathbb{E}\left[Y - \mathbb{E}\left[Y|X\right]\right] = 0$, which, in turn, implies that $\mathbb{E}\left[Y\right] = \mathbb{E}\left[\mathbb{E}\left[Y|X\right]\right]$. More generally,*

$$\mathbb{E}\left[\mathbb{E}\left[Y|X_1, X_2\right]|X_1\right] = \mathbb{E}\left[Y|X_1\right].$$

(vi) *(Independence) If $X \perp\!\!\!\perp Y$, $\mathbb{E}\left[Y|X\right] = \mathbb{E}\left[Y\right]$.*

(vii) *(Mean independence) If $\mathbb{E}\left[Y|X\right] = c$ (i.e. y is mean independent of X), then, by Law of Iterated Expectations,*

$$\mathbb{E}\left[\mathbb{E}\left[Y|X\right]\right] = \mathbb{E}\left[Y\right] = \mathbb{E}\left[c\right] = c.$$

Proof. See below.

(i) Setting $m^*(X) = f(X)$,

$$\mathbb{E}\left[(Y - f(X))\,\mathbf{1}_{\{X\in B\}}\right] = \mathbb{E}\left[(Y - Y)\,\mathbf{1}_{\{X\in B\}}\right] = 0$$

for any B.

(ii) $\mathbb{E}[Y|Z] = m^*(Z)$ is such that

$$\mathbb{E}\left[(Y - m^*(Z))\,\mathbf{1}_{\{Z\in B\}}\right] = 0, \forall B \in \mathcal{B}\left(\mathbb{R}^k\right)$$

and $\mathbb{E}[X|Z] = \tilde{m}^*(Z)$ is such that

$$\mathbb{E}\left[(X - \tilde{m}^*(Z))\,\mathbf{1}_{\{Z\in B\}}\right] = 0, \forall B \in \mathcal{B}\left(\mathbb{R}^k\right).$$

Summing the two,

$$\begin{aligned}\mathbb{E}\left[(Y - m^*(Z))\,\mathbf{1}_{\{Z\in B\}}\right] + \mathbb{E}\left[(X - m^*(Z))\,\mathbf{1}_{\{Z\in B\}}\right] &= \mathbb{E}\left[(Y + X - m^*(Z) - \tilde{m}^*(Z))\,\mathbf{1}_{\{Z\in B\}}\right] \\ &= 0, \forall B \in \mathcal{B}\left(\mathbb{R}^k\right)\end{aligned}$$

Hence, $\mathbb{E}[Y + X|Z] = m^*(Z) + \tilde{m}^*(Z) = \mathbb{E}[Y|Z] + \mathbb{E}[X|Z]$.

(iii) Admitted.

(iv) [To do]

(v) From (2.4), by definition,

$$\mathbb{E}\left[(Y - \mathbb{E}[Y|X])\,\mathbf{1}_{\{\mathbb{E}[Y|X]\in B\}}\right] = 0.$$

Setting $B = \mathcal{B}$ gives

$$\mathbb{E}[(Y - \mathbb{E}[Y|X])] = 0 \Rightarrow \mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|X]].$$

Now consider, $\mathbb{E}[Y|X_1, X_2] = m^*(X_1, X_2)$ and $\mathbb{E}[Y|X_1] = n^*(X_1)$, which implies that

$$\mathbb{E}\left[(Y - m^*(X_1, X_2))\,\mathbf{1}_{\{(X_1,X_2)\in A\}}\right] = 0, \forall A \in \mathcal{B}\left(\mathbb{R}^2\right) \tag{2.5}$$

$$\mathbb{E}\left[(Y - n^*(X_1))\,\mathbf{1}_{\{X_1\in B\}}\right] = 0, \forall B \in \mathcal{B}(\mathbb{R}). \tag{2.6}$$

We want to show that $n^*(X_1) = \mathbb{E}[m^*(X_1, X_2)|X_1]$; i.e.

$$\mathbb{E}\left[(m^*(X_1, X_2) - n^*(X_1))\,\mathbf{1}_{\{X_1\in\mathbf{B}\}}\right] = 0, \forall B \in \mathcal{B}.$$

Notice that

$$\begin{aligned}\mathbb{E}\left[(m^*(X_1, X_2) - n^*(X_1))\,\mathbf{1}_{\{X_1\in B\}}\right] &= \mathbb{E}\left[(-[Y - m^*(X_1, X_2)] + [Y - n^*(X_1)])\,\mathbf{1}_{\{X_1\in B\}}\right] \\ &= \mathbb{E}\left[(Y - n^*(X_1))\,\mathbf{1}_{\{X_1\in B\}}\right] - \mathbb{E}\left[(Y - m^*(X_1, X_2))\,\mathbf{1}_{\{X_1\in B\}}\right],\end{aligned} \tag{2.7}$$

where we realise that the first term is equal to zero due to (2.6). For the second term, notice that in (2.5) holds for any $X_1 \in B$. In particular, choose $\mathbf{A}$ such that

$$\mathbf{1}_{\{(X_1,X_2)\in(B,\mathcal{B}(\mathbb{R}))\}} = \mathbf{1}_{\{X_1\in B\}}.$$

We therefore see that (2.5) implies that the second term in (2.7) is zero.

(vi) Setting $B = \mathcal{B}$ yields

$$\mathbb{E}\left[(Y - m^*(X))\right] = 0 \Rightarrow \mathbb{E}[Y] = \mathbb{E}\left[m^*(X)\right].$$

Consider $A = \{m^*(X) \neq \mathbb{E}[Y]\} \in \mathcal{B}$. Then,

$$\begin{aligned}
\mathbb{E}\left[(Y - m^*(X))\mathbf{1}_{\{X\in A\}}\right] &= 0 \\
\Rightarrow \mathbb{E}\left[m^*(X)\mathbf{1}_{\{X\in A\}}\right] &= \mathbb{E}\left[Y\mathbf{1}_{\{X\in A\}}\right] \\
&= \mathbb{E}[Y]\,\mathbb{E}\left[\mathbf{1}_{\{X\in A\}}\right] \because X \perp\!\!\!\perp Y \\
&= \mathbb{E}[Y]\,\mathbb{P}(X \in A).
\end{aligned}$$

Now, consider

$$\begin{aligned}
\mathbb{E}\left[m^*(X)\mathbf{1}_{\{X\in A\}}\right] &= \mathbb{E}\left[m^*(X)\mathbf{1}_{\{X\in\mathcal{B}\}}\right] - \mathbb{E}\left[m^*(X)\mathbf{1}_{\{X\in A^c\}}\right] \\
&= \mathbb{E}\left[m^*(X)\right] - \mathbb{E}[Y].
\end{aligned}$$

Hence,

$$\begin{aligned}
\mathbb{E}\left[m^*(X)\right] - \mathbb{E}[Y] &= \mathbb{E}[Y]\,\mathbb{P}(X \in A) \\
\mathbb{E}\left[m^*(X)\right] &= \mathbb{E}[Y]\left[1 + \mathbb{P}(X \in A)\right].
\end{aligned}$$

Since we know $\mathbb{E}[Y] = \mathbb{E}\left[m^*(X)\right]$, above implies that $\mathbb{P}(X \in A) = 0$; i.e. $m^*(X) = \mathbb{E}[Y|X] = \mathbb{E}[Y]$ with probability one.

■

Note that independence implies mean independence, which in turn, implies uncorrelatedness. The converse statements do not hold.

Example 2.1. *(Mean independence $\not\Rightarrow$ independence)* Suppose $Y|X \sim N\left(0, \sigma^2 X\right)$. Then $\mathbb{E}[Y|X] = 0$ so that Y is mean independent of X. However, Y and X and clearly correlated.

Example 2.2. *(Mean independence $\Rightarrow$ zero covariance)* Suppose $\mathbb{E}[Y|X] = \mathbb{E}[Y]$, then

$$\begin{aligned}
\mathrm{Cov}[X, Y] &= \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y] \\
\text{(LIE)} &= \mathbb{E}\left[\mathbb{E}[XY|X]\right] - \mathbb{E}[X]\,\mathbb{E}[Y] \\
&= \mathbb{E}\left[X\mathbb{E}[Y|X]\right] - \mathbb{E}[X]\,\mathbb{E}[Y] \\
&= \mathbb{E}\left[X\mathbb{E}[Y]\right] - \mathbb{E}[X]\,\mathbb{E}[Y] \\
&= \mathbb{E}[X]\,\mathbb{E}[Y] - \mathbb{E}[X]\,\mathbb{E}[Y] = 0.
\end{aligned}$$

Example 2.3. *(Uncorrelatedness $\not\Rightarrow$ mean independence)* Suppose $X \sim N(0, 1)$ and $Y \overset{\mathrm{d}}{=} X^2$. Then,

$$\mathrm{Cov}[X, Y] = \mathbb{E}\left[X^3\right] - \mathbb{E}[X]\,\mathbb{E}\left[X^2\right] = 0 - 0 = 0.$$

However, $\mathbb{E}[Y|X] = \mathbb{E}\left[X^2|X\right] = X^2$.

3 Linear Regressions

Let $(Y, \mathbf{X}, u)$ be random variables/vectors where $Y, u \in \mathbb{R}$ and $\mathbf{X} \in \mathbb{R}^{k+1}$, such that

$$\mathbf{X} = (X_0, X_1, , \ldots, X_k)', \; X_0 = 1,$$
$$\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)',$$

and

$$Y = \mathbf{X}'\boldsymbol{\beta} + u. \tag{3.1}$$

3.1 Three interpretations of linear regressions

3.1.1 Linear Conditional Expectation

Assume $\mathbb{E}[Y|\mathbf{X}] = \mathbf{X}'\boldsymbol{\beta}$. Define $u \coloneqq Y - \mathbb{E}[Y|\mathbf{X}]$. Then

$$Y = \mathbb{E}[Y|\mathbf{X}] + u = \mathbf{X}'\boldsymbol{\beta} + u$$

so that (3.1) can be interpreted as linear expectation of Yconditional on $\mathbf{X}$. Note the following.

- ▷ The parameter $\boldsymbol{\beta}$ has no causal interpretation—it is simply a convenient way of summarising a feature of distribution of Y and $\mathbf{X}$;
- ▷ By construction $\mathbb{E}[u|\mathbf{X}] = \mathbb{E}[(Y - \mathbb{E}[Y|\mathbf{X}])|\mathbf{X}] = \mathbb{E}[Y|\mathbf{X}] - \mathbb{E}[Y|\mathbf{X}] = 0$. Then, by Law of Iterated Expectations,
$$\mathbb{E}[u] = \mathbb{E}[\mathbb{E}[u|\mathbf{X}]] = \mathbb{E}[0] = 0.$$
- ▷ Since $\mathbb{E}[u|\mathbf{X}] = 0$, u is mean independent of $\mathbf{X}$. This also implies that $\mathbf{X}$ and u are orthogonal:
$$\mathbb{E}[\mathbf{X}u] = \mathbb{E}[\mathbb{E}[\mathbf{X}u|\mathbf{X}]] = \mathbb{E}[\mathbb{E}[u|\mathbf{X}]\mathbf{X}] = 0.$$

3.1.2 "Best" Linear Approximation to Conditional Expectation / "Best" Linear Predictor of $Y|\mathbf{X}$

Define $u \coloneqq Y - \mathbf{X}'\boldsymbol{\beta}$. For the first interpretation, consider

$$\min_{\mathbf{b}\in\mathbb{R}^{k+1}} \quad \mathbb{E}\left[\left(\mathbb{E}[Y|\mathbf{X}] - \mathbf{X}'\mathbf{b}\right)^2\right]. \tag{3.2}$$

For the second interpretation, consider

$$\min_{\mathbf{b}\in\mathbb{R}^{k+1}} \quad \mathbb{E}\left[\left(Y - \mathbf{X}'\mathbf{b}\right)^2\right]. \tag{3.3}$$

Proposition 3.1. *Any solution to (3.2) is a solution to (3.3) (and vice versa).*

Proof. We want to express (3.2) to appear like (3.3); since the former is missing the Y term, it is natural to add and subtract Y in (3.2):

$$\begin{aligned}
\mathbb{E}\left[\left(\mathbb{E}[Y|\mathbf{X}] - \mathbf{X}'\mathbf{b}\right)^2\right] &= \mathbb{E}\left[\left(\underbrace{\mathbb{E}[Y|\mathbf{X}] - Y}_{\coloneqq v} + Y - \mathbf{X}'\mathbf{b}\right)^2\right] \\
&= \mathbb{E}\left[v^2\right] + 2\mathbb{E}\left[v\left(Y - \mathbf{X}'\mathbf{b}\right)\right] \\
&\quad + \mathbb{E}\left[\left(Y - \mathbf{X}'\mathbf{b}\right)^2\right] \\
&= \mathbb{E}\left[v^2\right] + 2\mathbb{E}[vY] - 2\mathbb{E}[v\mathbf{X}'\mathbf{b}] + \mathbb{E}\left[\left(Y - \mathbf{X}'\mathbf{b}\right)^2\right] \\
&= \mathbb{E}\left[v^2\right] + 2\mathbb{E}[vY] + \mathbb{E}\left[\left(Y - \mathbf{X}'\mathbf{b}\right)^2\right],
\end{aligned} \tag{3.4}$$

where we used the fact that

$$\begin{aligned}\mathbb{E}\left[v\mathbf{X}'\mathbf{b}\right] &= \mathbb{E}\left[\left(\mathbb{E}\left[Y|\mathbf{X}\right]-Y\right)\mathbf{X}'\mathbf{b}\right]\\ &= \mathbb{E}\left[\mathbb{E}\left[Y|\mathbf{X}\right]\mathbf{X}'\mathbf{b}\right]-\mathbb{E}\left[Y\mathbf{X}'\mathbf{b}\right]\\ \text{(LIE)} &= \mathbb{E}\left[Y\mathbf{X}'\mathbf{b}\right]-\mathbb{E}\left[Y\mathbf{X}'\mathbf{b}\right]\\ &= 0.\end{aligned}$$

Then, notice that the first two terms of (3.4) are constants with respect to b. Hence, minimising (3.4) with respect to $\mathbf{b}$ must give the same solution as minimising (3.3) with respect to $\mathbf{b}$. ■

Let $\boldsymbol{\beta}$ be a solution to (3.2) or (3.3). Define $u \coloneqq Y - \mathbf{X}'\boldsymbol{\beta}$, then we can interpret (3.1) as either the best "Best" Linear Approximation to Conditional Expectation or "Best" Linear Predictor of $Y|\mathbf{X}$, which we denote as $\text{BLP}\left(Y|\mathbf{X}\right)$. As before, no causal interpretation can be drawn from $\boldsymbol{\beta}$.

To derive properties of u, consider the first-order condition of (3.3):

$$\begin{aligned}\mathbf{0} &= 2\mathbb{E}\left[\mathbf{X}\left(Y-\mathbf{X}'\boldsymbol{\beta}\right)\right]\\ &= \mathbb{E}\left[\mathbf{X}u\right].\end{aligned} \tag{3.5}$$

3.1.3 Causal model interpretation

Assume that $Y = g\left(\mathbf{X}, u\right)$, where $\mathbf{X}$ represents observed determinants of Y, and u represents the unobserved determinants of Y. In other words, the assumption says that if we are given $\mathbf{X}$ and u , then we can compute Y. Then, the effect of a unit change in X_i on Y, holding X_{-i} and u constant is $\partial g/\partial X_i$. Thus, if we assume that $g\left(\mathbf{X}, u\right) = \mathbf{X}'\boldsymbol{\beta} + u$, then

$$\frac{\partial g\left(\mathbf{X}, u\right)}{\partial X_i} = \beta_i.$$

In this set up, $\mathbb{E}\left[u\right]$ may not equal zero but we can normalise it so that it is zero by replacing:

$$\begin{aligned}\beta_0 &= \beta_0 + \mathbb{E}\left[u\right],\\ u &= u - \mathbb{E}\left[u\right].\end{aligned}$$

However, we do not know statements about the relationships between observed and unobserved determinants of Y: e.g. $\mathbb{E}\left[X_i u\right]$ or $\mathbb{E}\left[u|\mathbf{X}\right]$.

3.2 Linear Regression when $\mathbb{E}\left[\mathbf{X}u\right] = \mathbf{0}$

Define $(Y, \mathbf{X}, u)$ and $\boldsymbol{\beta}$ as above and assume that: $\mathbb{E}\left[Y^2\right] < \infty$, $\mathbb{E}\left[\mathbf{X}\mathbf{X}'\right]$ exists, $\mathbb{E}\left[\mathbf{X}u\right] = \mathbf{0}$, and there is no perfect collinearity in $\mathbf{X}$.

Definition 3.1. *(Perfect collinearity)* $\mathbf{X}$ is said to be *perfectly collinear* if there is $\mathbf{c} \neq \mathbf{0}$, $\mathbf{c} \in \mathbb{R}^{k+1}$ such that $\mathbb{P}\left(\mathbf{c}'\mathbf{X} = 0\right) = 1$.

Lemma 3.1. $\mathbb{E}\left[\mathbf{X}\mathbf{X}'\right]$ *is invertible if and only if there is no perfect collinearity in* $\mathbf{X}$.

Proof. We can state the Lemma equivalently as: $\mathbb{E}\left[\mathbf{X}\mathbf{X}'\right]$ is not invertible if and only if there is perfect collinearity in $\mathbf{X}$. We will prove this statement.

($\Rightarrow$) We first want to show that, if $\mathbb{E}\left[\mathbf{X}\mathbf{X}'\right]$ is not invertible, then there is perfect collinearity in $\mathbf{X}$. If $\mathbb{E}\left[\mathbf{X}\mathbf{X}'\right]$ is not invertible, then there exists $\mathbf{c} \neq 0$ such that $\mathbb{E}\left[\mathbf{X}\mathbf{X}'\right]\mathbf{c} = \mathbf{0}$. Then,

$$0 = \mathbf{c}'\mathbb{E}\left[\mathbf{X}\mathbf{X}'\right]\mathbf{c} = \mathbb{E}\left[\mathbf{c}'\mathbf{X}\mathbf{X}'\mathbf{c}\right] = \mathbb{E}\left[\left(\mathbf{c}'\mathbf{X}\right)^2\right].$$

For this to be true, then it must be that

$$\mathbb{P}\left(\mathbf{c}'\mathbf{X} = 0\right) = 1;$$

i.e. there is perfect collinearity in $\mathbf{X}$.

($\Leftarrow$) We will prove the equivalent statement that, if there is perfect collinearity, then $\mathbb{E}[\mathbf{X}\mathbf{X}']$ is not invertible. Assume that there is perfect collinearity in $\mathbf{X}$, then there exists $\mathbf{c} \neq \mathbf{0}$ such that $\mathbb{P}(\mathbf{c}'\mathbf{X} = 0) = 1$, then

$$\begin{aligned} 0 = \mathbb{E}[\mathbf{c}'\mathbf{X}] &= \mathbb{E}[\mathbf{X}'\mathbf{c}] \\ &= \mathbb{E}[\mathbb{E}[\mathbf{X}'\mathbf{c}|\mathbf{X}]] \\ \Rightarrow \mathbf{X}0 = \mathbf{0} &= \mathbb{E}[\mathbb{E}[\mathbf{X}\mathbf{X}'\mathbf{c}|\mathbf{X}]] \\ &= \mathbb{E}[\mathbf{X}\mathbf{X}'\mathbf{c}] \\ &= \mathbb{E}[\mathbf{X}\mathbf{X}']\,\mathbf{c}. \end{aligned}$$

That is, $\mathbb{E}[\mathbf{X}\mathbf{X}']$ is not invertible. ■

From the Lemma, given our assumption that there is no perfect collinearity in $\mathbf{X}$, we know that $\mathbb{E}[\mathbf{X}\mathbf{X}']$ is invertible. Using this fact, we can rewrite the expression $\mathbb{E}[\mathbf{X}u] = \mathbf{0}$ as

$$\begin{aligned} \mathbf{0} &= \mathbb{E}[\mathbf{X}u] \\ &= \mathbb{E}[\mathbf{X}(Y - \mathbf{X}'\boldsymbol{\beta})] \\ &= \mathbb{E}[\mathbf{X}Y] - \mathbb{E}[\mathbf{X}\mathbf{X}']\,\boldsymbol{\beta} \\ \Rightarrow \boldsymbol{\beta} &= \mathbb{E}[\mathbf{X}\mathbf{X}']^{-1}\,\mathbb{E}[\mathbf{X}Y]. \end{aligned} \tag{3.6}$$

What happens if $\mathbb{E}[\mathbf{X}\mathbf{X}']$ is not invertible? This means that there exists multiple solutions to (3.6). However, any solutions $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}$ satisfy:

$$\mathbb{P}\left(\mathbf{X}'\hat{\boldsymbol{\beta}} = \mathbf{X}'\tilde{\boldsymbol{\beta}}\right) = 1.$$

This follows from the fact that, by construction, $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}$ are solutions to the minimisation problem (3.2)/(3.3). Although having multiple solutions is not a problem under the first and second interpretations, it would matter for the last interpretation; i.e. multiple solutions are a problem in cases where we are interested in causal relationships.

3.2.1 Solving for subvectors of $\boldsymbol{\beta}$

Suppose we partition $\mathbf{X}$ and $\boldsymbol{\beta}$ into subvectors $(\mathbf{X}_1)_{k_1\times 1}$ and $(\mathbf{X}_2)_{k_2\times 1}$, and accordingly, to $(\boldsymbol{\beta}_1)_{k_1\times 1}$ and $(\boldsymbol{\beta}_2)_{k_2\times 1}$. Then,

$$\begin{aligned} Y &= \mathbf{X}'\boldsymbol{\beta} + u \\ &= \begin{pmatrix} \mathbf{X}_1' & \mathbf{X}_2' \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} + u \\ &= \mathbf{X}_1'\boldsymbol{\beta}_1 + \mathbf{X}_2'\boldsymbol{\beta}_2 + u. \end{aligned}$$

We also have that, by assumption,

$$\mathbb{E}[\mathbf{X}u] = \mathbf{0} \Rightarrow \mathbb{E}\left[\begin{pmatrix} \mathbf{X}_1 u \\ \mathbf{X}_2 u \end{pmatrix}\right] = \begin{pmatrix} \mathbf{0}_{k_1\times 1} \\ \mathbf{0}_{k_2\times 1} \end{pmatrix}. \tag{3.7}$$

One way to solve this is to use (3.6):

$$\begin{aligned} \boldsymbol{\beta} &= \mathbb{E}[\mathbf{X}\mathbf{X}']^{-1}\,\mathbb{E}[\mathbf{X}Y] \\ \Leftrightarrow \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} &= \mathbb{E}\left[\begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix} \begin{pmatrix} \mathbf{X}_1' & \mathbf{X}_2' \end{pmatrix}\right]^{-1} \mathbb{E}\left[\begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix} Y\right] \\ &= \begin{bmatrix} \mathbb{E}[\mathbf{X}_1\mathbf{X}_1'] & \mathbb{E}[\mathbf{X}_1\mathbf{X}_2'] \\ \mathbb{E}[\mathbf{X}_2\mathbf{X}_1'] & \mathbb{E}[\mathbf{X}_1\mathbf{X}_2'] \end{bmatrix}^{-1} \mathbb{E}\left[\begin{pmatrix} \mathbf{X}_1 Y \\ \mathbf{X}_2 Y \end{pmatrix}\right]. \end{aligned}$$

However, there is another way. As shown below, we can estimate $\boldsymbol{\beta}_1$ in three steps:

(i) "regress" Y on $\mathbf{X}_2$ (i.e. BLP$(Y|\mathbf{X})$) and obtain the "residuals" $\tilde{Y}$;[14]

[14] Strictly speaking BLP is not the same as regressing.

(ii) "regress" X_1 on $\mathbf{X}_2$ and obtain the "residuals" $\tilde{\mathbf{X}}_1$;

(iii) "regress" $\tilde{Y}$ on $\tilde{\mathbf{X}}_1$ and the coefficient on $\tilde{\mathbf{X}}_1$ is equal to $\boldsymbol{\beta}_1$.

This method gives meaning to the phrase: $\boldsymbol{\beta}_1$ is the effect of $\mathbf{X}_1$ on Y when controlling for the effects of $\mathbf{X}_2$.

First, define

$$\begin{aligned}\tilde{Y} &\coloneqq Y - \text{BLP}\left(Y|\mathbf{X}_2\right),\\ \tilde{\mathbf{X}}_1 &\coloneqq \mathbf{X}_1 - \text{BLP}\left(\mathbf{X}_1|\mathbf{X}_2\right),\end{aligned}$$

where

$$\underbrace{\begin{pmatrix}\tilde{X}_{11}\\ \tilde{X}_{12}\\ \vdots\\ \tilde{X}_{1k_1}\end{pmatrix}}_{=\tilde{\mathbf{X}}_1} = \underbrace{\begin{pmatrix}X_{11}\\ X_{12}\\ \vdots\\ X_{1k_1}\end{pmatrix}}_{=\mathbf{X}_1} - \underbrace{\begin{pmatrix}\text{BLP}\left(X_{11}|\mathbf{X}_2\right)\\ \text{BLP}\left(X_{12}|\mathbf{X}_2\right)\\ \vdots\\ \text{BLP}\left(X_{1k_1}|\mathbf{X}_2\right)\end{pmatrix}}_{=\text{BLP}(\mathbf{X}_1|\mathbf{X}_2)}.$$

That is, $\tilde{Y}$ can be thought of as the residual from regressing Y on $\mathbf{X}_2$, and $\tilde{\mathbf{X}}_1$ as the collection of residual from regressing $\tilde{\mathbf{X}}_1$ on $\mathbf{X}_2$ component wise.

$$\begin{aligned}\text{BLP}\left(Y|\mathbf{X}_2\right) &= \mathbf{X}_2'\boldsymbol{\gamma}_Y,\\ \text{BLP}\left(\mathbf{X}_1|\mathbf{X}_2\right) &= \mathbf{X}_2'\boldsymbol{\gamma}_{\mathbf{X}_1},\end{aligned}$$

By the first-order conditions, we realise that

$$\begin{aligned}\mathbb{E}\left[\mathbf{X}_2\tilde{Y}\right] &= 0,\\ \begin{pmatrix}\mathbb{E}[\mathbf{X}_2\tilde{X}_{11}]\\ \mathbb{E}[\mathbf{X}_2\tilde{X}_{12}]\\ \vdots\\ \mathbb{E}[\mathbf{X}_2\tilde{X}_{1k_1}]\end{pmatrix} = \begin{pmatrix}\mathbf{0}_{k_2\times 1}\\ \mathbf{0}_{k_2\times 1}\\ \vdots\\ \mathbf{0}_{k_2\times 1}\end{pmatrix} &\Leftrightarrow \mathbb{E}\left[\tilde{\mathbf{X}}_1\mathbf{X}_2'\right] = \mathbf{0}_{k_1\times k_2}.\end{aligned} \tag{3.8}$$

That last expression is a general result that the "residual" from BLP $(X_{1i}|\mathbf{X}_2)$, $\tilde{X}_{1i}$, is orthogonal to $\mathbf{X}_2$ (in expectation).

Proposition 3.2. *Consider the regression*

$$\tilde{Y} = \tilde{\mathbf{X}}_1'\tilde{\boldsymbol{\beta}} + \tilde{u},$$

where

$$\begin{aligned}\tilde{\boldsymbol{\beta}} &= \mathbb{E}\left[\tilde{\mathbf{X}}_1\tilde{\mathbf{X}}_1'\right]^{-1}\mathbb{E}\left[\tilde{\mathbf{X}}_1\tilde{Y}\right]\\ &= \text{Var}\left[\tilde{\mathbf{X}}_1\right]^{-1}\text{Cov}\left[\tilde{\mathbf{X}}_1,\tilde{Y}\right]\end{aligned} \tag{3.9}$$

and $\mathbb{E}\left[\tilde{\mathbf{X}}\tilde{u}\right] = \mathbf{0}$.[15] *Then,*

$$\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}_1.$$

[15] Note that we can write $\mathbb{E}\left[\tilde{\mathbf{X}}_1\tilde{\mathbf{X}}_1'\right] = \text{Var}\left[\tilde{\mathbf{X}}_1\right]$ whenever $\mathbf{X}_2$ contains a constant.

Proof. First note that since $\mathbb{E}[\mathbf{X}\mathbf{X}']$ is invertible by assumption, and $\tilde{\mathbf{X}}_1$ is linear combination of $\mathbf{X}$, $\mathbb{E}\left[\tilde{\mathbf{X}}_1\tilde{\mathbf{X}}_1'\right]$ is invertible.

$$\begin{aligned}\tilde{\boldsymbol{\beta}} &= \mathbb{E}\left[\tilde{\mathbf{X}}_1\tilde{\mathbf{X}}_1'\right]^{-1}\mathbb{E}\left[\tilde{\mathbf{X}}_1\tilde{Y}\right]\\ &= \mathbb{E}\left[\tilde{\mathbf{X}}_1\tilde{\mathbf{X}}_1'\right]^{-1}\mathbb{E}\left[\tilde{\mathbf{X}}_1\left(Y-\text{BLP}\left(Y|\mathbf{X}_2\right)\right)\right]\\ &= \mathbb{E}\left[\tilde{\mathbf{X}}_1\tilde{\mathbf{X}}_1'\right]^{-1}\left(\mathbb{E}\left[\tilde{\mathbf{X}}_1 Y\right]-\mathbb{E}\left[\tilde{\mathbf{X}}_1\text{BLP}\left(Y|\mathbf{X}_2\right)\right]\right).\end{aligned}$$

Since $\tilde{\mathbf{X}}_1$ is the residual from regressing $\mathbf{X}_1$ on $\mathbf{X}_2$, in particular, it is orthogonal to $\mathbf{X}_2$. To see this, from (3.8), we have

$$\mathbb{E}\left[\tilde{\mathbf{X}}_1\text{BLP}\left(Y|\mathbf{X}_2\right)\right] = \mathbb{E}\left[\tilde{\mathbf{X}}_1\left(\mathbf{X}_2'\boldsymbol{\gamma}_Y\right)\right] = \mathbb{E}\left[\tilde{\mathbf{X}}_1\mathbf{X}_2'\right]\boldsymbol{\gamma}_Y = \mathbf{0}_{k_1\times 1}.$$

Hence,

$$\begin{aligned}\tilde{\boldsymbol{\beta}} &= \mathbb{E}\left[\tilde{\mathbf{X}}_1\tilde{\mathbf{X}}_1'\right]^{-1}\mathbb{E}\left[\tilde{\mathbf{X}}_1 Y\right]\\ &= \mathbb{E}\left[\tilde{\mathbf{X}}_1\tilde{\mathbf{X}}_1'\right]^{-1}\mathbb{E}\left[\tilde{\mathbf{X}}_1\left(\mathbf{X}_1'\boldsymbol{\beta}_1+\mathbf{X}_2'\boldsymbol{\beta}_2+u\right)\right]\\ &= \mathbb{E}\left[\tilde{\mathbf{X}}_1\tilde{\mathbf{X}}_1'\right]^{-1}\left(\mathbb{E}\left[\tilde{\mathbf{X}}_1\mathbf{X}_1'\right]\boldsymbol{\beta}_1+\mathbb{E}\left[\tilde{\mathbf{X}}_1\mathbf{X}_2'\right]\boldsymbol{\beta}_2+\mathbb{E}\left[\tilde{\mathbf{X}}_1 u\right]\right).\end{aligned}$$

Using the fact that $\mathbb{E}\left[\tilde{\mathbf{X}}_1\mathbf{X}_2'\right]=\mathbf{0}$ again (from (3.8)) and

$$\begin{aligned}\mathbb{E}\left[\tilde{\mathbf{X}}_1 u\right] &= \mathbb{E}\left[\left(\mathbf{X}_1-\text{BLP}\left(\mathbf{X}_1|\mathbf{X}_2\right)\right)u\right] = \mathbb{E}\left[\mathbf{X}_1 u\right]-\mathbb{E}\left[\left(\mathbf{X}_2'\boldsymbol{\gamma}_{\mathbf{X}_1}\right)u\right]\\ &= \mathbb{E}\left[\mathbf{X}_1 u\right]-\left(\boldsymbol{\gamma}_{\mathbf{X}_1}\mathbb{E}\left[\mathbf{X}_2 u\right]\right)' = \mathbf{0}\end{aligned}$$

using (3.7). Thus,

$$\begin{aligned}\tilde{\boldsymbol{\beta}} &= \mathbb{E}\left[\tilde{\mathbf{X}}_1\tilde{\mathbf{X}}_1'\right]^{-1}\mathbb{E}\left[\tilde{\mathbf{X}}_1\mathbf{X}_1'\right]\boldsymbol{\beta}_1\\ &= \mathbb{E}\left[\tilde{\mathbf{X}}_1\tilde{\mathbf{X}}_1'\right]^{-1}\mathbb{E}\left[\tilde{\mathbf{X}}_1\left(\tilde{\mathbf{X}}_1'+\text{BLP}\left(\mathbf{X}_1|\mathbf{X}_2\right)'\right)\right]\boldsymbol{\beta}_1\\ &= \boldsymbol{\beta}_1+\mathbb{E}\left[\tilde{\mathbf{X}}_1\tilde{\mathbf{X}}_1'\right]^{-1}\mathbb{E}\left[\tilde{\mathbf{X}}_1\text{BLP}\left(\mathbf{X}_1|\mathbf{X}_2\right)'\right]\boldsymbol{\beta}_1.\end{aligned}$$

Note that

$$\begin{aligned}\mathbb{E}\left[\tilde{\mathbf{X}}_1\text{BLP}\left(\mathbf{X}_1|\mathbf{X}_2\right)'\right] &= \mathbb{E}\left[\tilde{\mathbf{X}}_1\left(\mathbf{X}_2'\boldsymbol{\gamma}_{\mathbf{X}_1}\right)'\right] = \mathbb{E}\left[\tilde{\mathbf{X}}_1\boldsymbol{\gamma}_{\mathbf{X}_1}'\mathbf{X}_2\right]\\ &= \mathbb{E}\left[\begin{pmatrix}\tilde{X}_{11}\\ \tilde{X}_{12}\\ \vdots\\ \tilde{X}_{1k_1}\end{pmatrix}\begin{pmatrix}\gamma_{\mathbf{X}_1 1}\\ \gamma_{\mathbf{X}_1 2}\\ \vdots\\ \gamma_{\mathbf{X}_1 k_2}\end{pmatrix}'\mathbf{X}_2\right]\\ &= \mathbb{E}\left[\begin{pmatrix}\gamma_{\mathbf{X}_1 1}\tilde{X}_{11} & \gamma_{\mathbf{X}_1 2}\tilde{X}_{11} & \cdots & \gamma_{\mathbf{X}_1 k_2}\tilde{X}_{11}\\ \gamma_{\mathbf{X}_1 1}\tilde{X}_{12} & \gamma_{\mathbf{X}_1 2}\tilde{X}_{12} & \cdots & \gamma_{\mathbf{X}_1 k_2}\tilde{X}_{12}\\ \vdots & \vdots & \ddots & \vdots\\ \gamma_{\mathbf{X}_1 1}\tilde{X}_{1k_1} & \gamma_{\mathbf{X}_1 2}\tilde{X}_{1k_1} & \cdots & \gamma_{\mathbf{X}_1 k_2}\tilde{X}_{1k_1}\end{pmatrix}_{k_1\times k_2}(\mathbf{X}_2)_{k_2\times 1}\right]\\ &= \mathbf{0}_{k_1\times 1},\end{aligned}$$

where the last equality holds since $\mathbb{E}\left[\tilde{X}_{1i}X_{2j}\right]=0$ for all $i=1,2,\ldots,k_1$ and $j=1,2,\ldots k_2$ (from (3.8)). We therefore obtain that

$$\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}_1.$$

■

Remark 3.1. Notice that we showed the following in the proof above.

- ▷ $\mathbb{E}\left[\tilde{\mathbf{X}}_1 \text{BLP}\left(Y|\mathbf{X}_2\right)\right] = \mathbf{0}_{k_1\times 1}$: This means that the "residual" from "regressing" $\mathbf{X}_1$ on $\mathbf{X}_2$ is orthogonal to $\text{BLP}\left(Y|\mathbf{X}_2\right)$ from "regressing" Y on $\mathbf{X}_2$.
- ▷ $\mathbb{E}\left[\tilde{\mathbf{X}}_1 u\right] = \mathbf{0}_{k_1\times 1}$: This comes from the assumption that $\mathbb{E}\left[\mathbf{X}u\right] = 0$, which holds by construction of BLP.
- ▷ $\mathbb{E}\left[\tilde{\mathbf{X}}_1 \text{BLP}\left(\mathbf{X}_1|\mathbf{X}_2\right)'\right] = \mathbf{0}_{k_1\times 1}$: This means that the "residual" from "regressing" $\mathbf{X_1}$ on $\mathbf{X}_2$ is orthogonal to $\text{BLP}\left(\mathbf{X}_1|\mathbf{X}_2\right)$ from "regressing" $\mathbf{X}_1$ on $\mathbf{X}_2$.
- ▷ $\mathbb{E}\left[\tilde{\mathbf{X}}_1 \mathbf{X}_2'\right] = \mathbf{0}_{k_1\times k_2}$: This means that the "residual" from regressing $\mathbf{X}_1$ on $\mathbf{X}_2$ is orthogonal to $\mathbf{X}_2$.

Corollary 3.1. *If X_2 is the constant term, then*

$$\begin{aligned}\tilde{Y} &= Y - \mathbb{E}\left[Y\right],\\ \tilde{\mathbf{X}}_1 &= \mathbf{X}_1 - \mathbb{E}\left[\mathbf{X}_1\right].\end{aligned}$$

This is because $\text{BLP}\left[Y|X_2\right]$ *would solve* $\min_{\beta_0}\mathbb{E}\left[\left(Y-\beta_0\right)^2\right]$*, which gives the first-order condition as* $\mathbb{E}\left[\left(Y-\beta_0\right)\right] = 0 \Rightarrow \mathbb{E}\left[Y\right] = \beta_0$*. Similarly for* $\text{BLP}\left[\mathbf{X}_1|\mathbf{X}_2\right]$*. Substituting the expressions above into (3.9),*

$$\begin{aligned}\tilde{\boldsymbol{\beta}} &= \mathbb{E}\left[\left(\mathbf{X}_1 - \mathbb{E}\left[\mathbf{X}_1\right]\right)\left(\mathbf{X}_1 - \mathbb{E}\left[\mathbf{X}_1\right]\right)'\right]^{-1}\mathbb{E}\left[\left(\mathbf{X}_1 - \mathbb{E}\left[\mathbf{X}_1\right]\right)\left(Y - \mathbb{E}\left[Y\right]\right)\right]\\ &= \text{Var}\left[\mathbf{X}_1\right]^{-1}\text{Cov}\left[\mathbf{X}_1, Y\right]\\ &= \boldsymbol{\beta}_1.\end{aligned}$$

3.2.2 Omitted variable bias

Suppose $k = 2$ so that

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u.$$

Consider

$$Y = \beta_0^* + \beta_1^* X_1 + u^*$$

with $\mathbb{E}\left[u^*\right] = 0$ and $\mathbb{E}\left[X_1 u^*\right] = 0$. Then, using the corollary above,

$$\begin{aligned}\beta_1^* &= \frac{\text{Cov}\left[X_1, Y\right]}{\text{Var}\left[X_1\right]} = \frac{\text{Cov}\left[X_1, \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u\right]}{\text{Var}\left[X_1\right]}\\ &= \frac{\text{Cov}\left[X_1, \beta_1 X_1 + \beta_2 X_2\right]}{\text{Var}\left[X_1\right]} \because \mathbb{E}\left[X_1 u\right] = 0\\ &= \beta_1 + \beta_2\frac{\text{Cov}\left[X_1, X_2\right]}{\text{Var}\left[X_1\right]}.\end{aligned}$$

Thus, in general, we see that $\beta_1^* \neq \beta_1$ due to the *omitted variable bias* term. Note that when X_1 and X_2 are scalars, the sign of the bias is given by the covariance between X_1 and X_2; however, the direction of bias is difficult to determine if X_1 and X_2 are vectors.

More generally, if

$$Y = \beta_0 + \mathbf{X}_1'\boldsymbol{\beta}_1 + \mathbf{X}_2'\boldsymbol{\beta}_2 + u,$$

and we estimate

$$Y = \beta_0^* + \mathbf{X}_1'\boldsymbol{\beta}_1^* + u^*.$$

Then,

$$\begin{aligned}\boldsymbol{\beta}_1^* &= \text{Var}\left[\mathbf{X}_1\right]^{-1}\text{Cov}\left[\mathbf{X}_1, Y\right]\\ &= \boldsymbol{\beta}_1 + \text{Var}\left[\mathbf{X}_1\right]^{-1}\text{Cov}\left[\mathbf{X}_1, \mathbf{X}_2\right]\boldsymbol{\beta}_2.\end{aligned}$$

3.2.3 Measurement error

Suppose we partition $\mathbf{X}$ into $X_0 = 1 \in \mathbb{R}$ and $\mathbf{X}_1 \in \mathbb{R}^k$, and partition $\boldsymbol{\beta}$ correspondingly. We then have

$$Y = \beta_0 + \mathbf{X}_1'\boldsymbol{\beta}_1 + u.$$

Let us assume that $\mathbf{X}_1$ is unobserved and that we instead observe, $\hat{\mathbf{X}}_1$, given by

$$\hat{\mathbf{X}}_1 = \mathbf{X}_1 + \mathbf{v},$$

with the *classical measurement error assumptions*: $\mathbb{E}[\mathbf{v}] = \mathbf{0}$, $\text{Cov}[\mathbf{X}_1, \mathbf{v}] = \mathbf{0}$ and $\text{Cov}[u, \mathbf{v}] = \mathbf{0}$.

Now consider the following regression

$$Y = \beta_0^* + \hat{\mathbf{X}}_1'\boldsymbol{\beta}_1^* + u^*,$$

with $\mathbb{E}[u^*] = 0$ and $\mathbb{E}[\hat{\mathbf{X}}_1 u^*] = \mathbf{0}$. Using Corollary 3.1,

$$\boldsymbol{\beta}_1^* = \text{Var}\left[\hat{\mathbf{X}}_1\right]^{-1} \text{Cov}\left[\hat{\mathbf{X}}_1, Y\right].$$

Expanding Y,

$$\boldsymbol{\beta}_1^* = \text{Var}\left[\hat{\mathbf{X}}_1\right]^{-1} \text{Cov}\left[\hat{\mathbf{X}}_1, Y\right] = \text{Var}\left[\hat{\mathbf{X}}_1\right]^{-1} \text{Cov}\left[\hat{\mathbf{X}}_1, \beta_0 + \mathbf{X}_1'\boldsymbol{\beta}_1 + u\right].$$

Since β_0 is a constant and $\text{Cov}[\mathbf{X}_1 + \mathbf{v}, u] = \text{Cov}[\mathbf{X}_1, u] + \text{Cov}[\mathbf{v}, u] = \mathbf{0}$, then

$$\begin{aligned}\boldsymbol{\beta}_1^* &= \text{Var}\left[\hat{\mathbf{X}}_1\right]^{-1} \text{Cov}\left[\hat{\mathbf{X}}_1, \mathbf{X}_1'\boldsymbol{\beta}_1\right] = \text{Var}\left[\hat{\mathbf{X}}_1\right]^{-1} \text{Cov}\left[\mathbf{X}_1 + \mathbf{v}, \mathbf{X}_1'\right]\boldsymbol{\beta}_1 \\ &= \text{Var}\left[\hat{\mathbf{X}}_1\right]^{-1} \text{Var}\left[\mathbf{X}_1\right]\boldsymbol{\beta}_1,\end{aligned}$$

where we used the fact that $\text{Cov}[\mathbf{X}_1, \mathbf{v}] = \mathbf{0}$. We may further write

$$\begin{aligned}\boldsymbol{\beta}_1^* &= \text{Var}\left[\mathbf{X}_1 + \mathbf{v}\right]^{-1} \text{Var}\left[\mathbf{X}_1\right]\boldsymbol{\beta}_1 \\ &= \left(\text{Var}\left[\mathbf{X}_1\right] + \text{Var}\left[\mathbf{v}\right]\right)^{-1} \text{Var}\left[\mathbf{X}_1\right]\boldsymbol{\beta}_1.\end{aligned}$$

In case $\mathbf{X}_1$ is a scalar, we obtain that

$$\beta_1^* = \frac{\text{Var}[X_1]}{\text{Var}[X_1] + \text{Var}[v]}\beta_1,$$

where we see that

$$0 < \frac{\text{Var}[X_1]}{\text{Var}[X_1] + \text{Var}[v]} < 1.$$

This term is known as *attenuation bias.*

Attenuation bias occurs more generally. Suppose we instead partition $\mathbf{X}$ into $X_0 = 1 \in \mathbb{R}$, $X_1 \in \mathbb{R}$, $\mathbf{X}_2 \in \mathbb{R}^{k_2}$, and partition $\boldsymbol{\beta}$ correspondingly. We then have

$$Y = \beta_0 + \beta_1 X_1 + \mathbf{X}_2'\boldsymbol{\beta}_2 + u.$$

We assume again that X_1 is unobserved and that we instead observe

$$\hat{X}_1 = X_1 + v,$$

with $\mathbb{E}[v] = 0$, $\text{Cov}[X_1, v] = 0$, $\text{Cov}[\mathbf{X}_2, v] = \mathbf{0}$, $\text{Cov}[u, v] = 0$. Now consider the following regression

$$Y = \beta_0^* + \beta_1^*\hat{X}_1 + \mathbf{X}_2'\boldsymbol{\beta}_2^* + u^*$$

with $\mathbb{E}[u^*] = 0$, $\text{Cov}\left[\hat{X}_1, u^*\right] = 0$ and $\text{Cov}[\mathbf{X}_2, u^*] = \mathbf{0}$. Define

$$\tilde{\hat{X}}_1 \coloneqq \hat{X}_1 - \text{BLP}\left(\hat{X}_1|1, \mathbf{X}_2\right),$$
$$\tilde{Y} \coloneqq Y - \text{BLP}\left(Y|1, \mathbf{X}_2\right).$$

We can obtain β_1^* from the regression, $\tilde{Y} = \tilde{\beta}\tilde{\hat{X}}_1 + \tilde{u}$ as

$$\beta_1^* = \tilde{\beta} = \frac{\text{Cov}\left[\tilde{\hat{X}}_1, \tilde{Y}\right]}{\text{Var}\left[\tilde{\hat{X}}_1\right]}.$$

Since $\mathbb{E}[\tilde{\hat{X}}_1 \text{BLP}\left(Y|1, \mathbf{X}_2\right)] = 0$ (see Remark 3.1) implies $\text{Cov}[\tilde{\hat{X}}_1, \text{BLP}\left(Y|1, \mathbf{X}_2\right)] = 0$,

$$\beta_1^* = \frac{\text{Cov}\left[\tilde{\hat{X}}_1, \tilde{Y}\right]}{\text{Var}\left[\tilde{\hat{X}}_1\right]} = \frac{\text{Cov}\left[\tilde{\hat{X}}_1, Y - \text{BLP}\left(Y|1, \mathbf{X}_2\right)\right]}{\text{Var}\left[\tilde{\hat{X}}_1\right]} = \frac{\text{Cov}\left[\tilde{\hat{X}}_1, Y\right]}{\text{Var}\left[\tilde{\hat{X}}_1\right]}.$$

Substituting for Y,

$$\beta_1^* = \frac{\text{Cov}\left[\tilde{\hat{X}}_1, \beta_0 + \beta_1 X_1 + \mathbf{X}_2'\boldsymbol{\beta}_2 + u\right]}{\text{Var}\left[\tilde{\hat{X}}_1\right]} = \frac{\text{Cov}\left[\tilde{\hat{X}}_1, \beta_1 X_1\right]}{\text{Var}\left[\tilde{\hat{X}}_1\right]},$$

where we note that β_0 is a constant, $\mathbb{E}[\tilde{\hat{X}}_1 \mathbf{X}_2'] = \mathbf{0}$ and $\mathbb{E}[\tilde{\hat{X}}_1 u] = 0$ (see Remark 3.1).

Now, notice that $\tilde{\hat{X}}_1 = \tilde{X}_1 + v$, where $\tilde{X}_1 \coloneqq X_1 - \text{BLP}\left(X_1|1, \mathbf{X}_2\right)$. To see this,

$$\begin{aligned}\tilde{\hat{X}}_1 &= \hat{X}_1 - \text{BLP}\left(\hat{X}_1|1, \mathbf{X}_2\right) = X_1 - \text{BLP}\left(X_1 + v|1, \mathbf{X}_2\right) + v \\ &= X_1 - \text{BLP}\left(X_1|1, \mathbf{X}_2\right) + v \because \mathbb{E}\left[\mathbf{X}_2 v\right] = 0 \\ &= \tilde{X}_1 + v.\end{aligned}$$

We can now write:

$$\begin{aligned}\beta_1^* &= \frac{\text{Cov}\left[\tilde{\hat{X}}_1, \beta_1 X_1\right]}{\text{Var}\left[\tilde{\hat{X}}_1\right]} = \frac{\text{Cov}\left[\tilde{X}_1 + v, X_1\right]}{\text{Var}\left[\tilde{X}_1 + v\right]}\beta_1 = \frac{\text{Cov}\left[\tilde{X}_1, X_1\right]}{\text{Var}\left[\tilde{X}_1\right] + \text{Var}\left[v\right]}\beta_1 \\ &= \frac{\text{Cov}\left[\tilde{X}_1, \tilde{X}_1 + \text{BLP}\left(X_1|1, \mathbf{X}_2\right)\right]}{\text{Var}\left[\tilde{X}_1 + v\right]}\beta_1.\end{aligned}$$

Since $\mathbb{E}\left[\tilde{X}_1 \text{BLP}\left(X_1|1, \mathbf{X}_2\right)\right] = 0$ (see Remark 3.1). We therefore obtain that

$$\beta_1^* = \frac{\text{Var}\left[\tilde{X}_1\right]}{\text{Var}\left[\tilde{X}_1\right] + \text{Var}\left[v\right]}\beta_1,$$

where we again see that

$$0 < \frac{\text{Var}\left[\tilde{X}_1\right]}{\text{Var}\left[\tilde{X}_1\right] + \text{Var}\left[v\right]} < 1.$$

3.2.4 Estimating $\boldsymbol{\beta}$

Let $(Y, \mathbf{X}, u)$ be random variables/vectors where $Y, u \in \mathbb{R}$ and $\mathbf{X} \in \mathbb{R}^{k+1}$, where

$$\mathbf{X} = (X_0, X_1, , \ldots, X_k)', \ X_0 = 1,$$
$$\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)',$$

be a parameter such that

$$Y = \mathbf{X}'\boldsymbol{\beta} + u. \tag{3.10}$$

Assume that:

- ▷ $\mathbb{E}[\mathbf{X}u] = 0$;
- ▷ $\mathbb{E}[\mathbf{X}\mathbf{X}'] < \infty$;
- ▷ there is no perfect collinearity in $\mathbf{X}$;
- ▷ $(Y, \mathbf{X}) \sim P$ and we have samples $\left(Y^1, \mathbf{X}^1\right), \left(Y^2, \mathbf{X}^2\right), \ldots, \left(Y^n, \mathbf{X}^n\right) \overset{\text{iid}}{\sim} P$.

We denote by superscripts the index for sample.

Recall that

$$\boldsymbol{\beta} = \mathbb{E}[\mathbf{X}\mathbf{X}']^{-1}\mathbb{E}[\mathbf{X}Y].$$

A natural estimator for $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}}_n = \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}^i\mathbf{X}^{i\prime}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}^iY^i\right).$$

This (analog) estimator is called the *ordinary least square* (OLS) estimator.

The name comes from the fact that $\hat{\boldsymbol{\beta}}_n$ solves

$$\hat{\boldsymbol{\beta}}_n = \underset{\mathbf{b}\in\mathbb{R}^{k+1}}{\arg\min} \quad \frac{1}{n}\sum_{i=1}^{n}\left(Y^i - \mathbf{X}^{i\prime}\mathbf{b}\right)^2.$$

To see this, notice that the first-order condition is given by

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}^i\left(Y^i - \mathbf{X}^{i\prime}\hat{\boldsymbol{\beta}}_n\right) &= \mathbf{0}_{(k+1)\times 1} \\ \Rightarrow \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{X}^i\mathbf{X}^{i\prime}\right)\hat{\boldsymbol{\beta}}_n &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}^iY^i\end{aligned} \tag{3.11}$$

Given the assumption that there is no perfect collinearity in $\mathbf{X}$, we know that $\mathbb{E}[\mathbf{X}\mathbf{X}']$ is invertible, so that $\det(\mathbb{E}[\mathbf{X}\mathbf{X}']) \neq 0$. Then, by the Continuous Mapping Theorem, while noting that taking the determinant is a continuous operation (polynomial which is continuous), we obtain that

$$\det\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}^i\mathbf{X}^{i\prime}\right) \xrightarrow{\text{P}} \det(\mathbb{E}[\mathbf{X}\mathbf{X}']).$$

This also means that $\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}^i\mathbf{X}^{i\prime}$ is invertible as n becomes large with probability 1 (since it converges to the population variance which is positive definite).

We define the *fitted values* to be

$$\hat{Y}^i = \mathbf{X}^i\hat{\boldsymbol{\beta}}_n$$

and the *fitted residuals* to be

$$\hat{u}^i \coloneqq Y^i - \hat{Y}^i = Y^i - \mathbf{X}^{i\prime}\hat{\boldsymbol{\beta}}_n.$$

With these definitions, we realise that (3.11) can be written as

$$\sum_{i=1}^{n} \mathbf{X}^i \hat{u}^i = \mathbf{0}_{(k+1)\times 1} \tag{3.12}$$
$$\Leftrightarrow \begin{pmatrix} \sum_{i=1}^n X_0^i \hat{u}^i \\ \sum_{i=1}^n X_1^i \hat{u}^i \\ \vdots \\ \sum_{i=1}^n X_{k+1}^i \hat{u}^i \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Note that since $X_0^i = 1$ for all i, the first-order condition implies that

$$\sum_{i=1}^{n} \hat{u}^i = 0, \tag{3.13}$$

when we have a constant in the regression.

3.2.5 Interpreting OLS as projection

Let us denote the collections of sample observations as

$$\begin{aligned}
\mathbb{Y}_{n\times 1} &\coloneqq \left(Y^1, Y^2, \ldots, Y^n\right)', \\
\mathbb{X}_{n\times(k+1)} &\coloneqq \left(\mathbf{X}^1, \mathbf{X}^2, \ldots, \mathbf{X}^n\right)' = \begin{pmatrix} 1 & X_1^1 & X_2^1 & \cdots & X_k^1 \\ 1 & X_1^2 & X_2^2 & \cdots & X_k^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_1^n & X_2^n & \cdots & X_k^n \end{pmatrix} \\
\mathbb{U}_{n\times 1} &\coloneqq \left(u^1, u^2, \ldots, u^n\right)', \\
\hat{\mathbb{Y}}_{n\times 1} &\coloneqq \left(\hat{Y}^1, \hat{Y}^2, \ldots, \hat{Y}^n\right)' = \mathbb{X}\hat{\boldsymbol{\beta}}_n \\
\hat{\mathbb{U}}_{n\times 1} &\coloneqq \left(\hat{u}^1, \hat{u}^2, \ldots, \hat{u}^n\right)', \\
&= \mathbb{Y} - \hat{\mathbb{Y}} = \mathbb{Y} - \mathbb{X}\hat{\boldsymbol{\beta}}_n.
\end{aligned}$$

Note that, whereas $\mathbf{X}$ has dimension $(k+1)\times 1$, $\mathbb{X}$ consists of observations of $\mathbf{X}$ transposed and then stacked across different observations.

Given these definitions, we can write the OLS estimator as

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_n &= \left(\mathbb{X}'\mathbb{X}\right)^{-1}\mathbb{X}'\mathbb{Y}, \\
\mathbb{X}'\hat{\mathbb{U}} &= \mathbf{0}_{(k+1)\times 1},
\end{aligned}$$

where $\hat{\boldsymbol{\beta}}_n$ solves the following problem:

$$\hat{\boldsymbol{\beta}}_n = \min_{\mathbf{b}\in\mathbb{R}^{k+1}} \quad \left|\mathbb{Y} - \mathbb{X}\mathbf{b}\right|^2,$$

where $|\cdot|$ is the Euclidean norm. Let us interpret this minimisation problem.

First, recall that a column space of a matrix A is the set of all possible linear combinations (i.e. span) of its column vectors. Then,

$$\left\{\mathbb{X}\mathbf{b} : \mathbf{b} \in \mathbb{R}^{k+1}\right\}$$

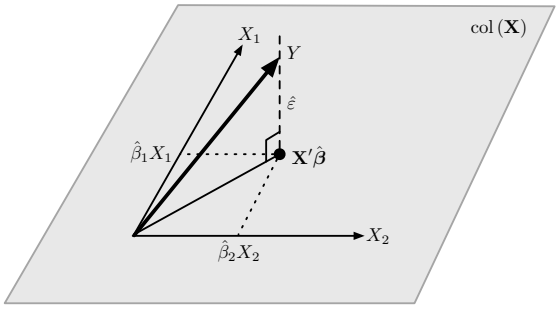
is the column space of $\mathbb{X}$, denoted $\operatorname{col}[\mathbb{X}]$ since

$$\mathbb{X}\mathbf{b} = \begin{pmatrix} 1 & X_1^1 & X_2^1 & \cdots & X_k^1 \\ 1 & X_1^2 & X_2^2 & \cdots & X_k^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_1^n & X_2^n & \cdots & X_k^n \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{pmatrix} = \begin{pmatrix} b_0 + b_1 X_1^1 + b_2 X_2^1 + \cdots + n X_k^1 \\ b_0 + b_1 X_1^2 + b_2 X_2^2 + \cdots + n X_k^2 \\ \vdots \\ b_0 + b_1 X_1^n + b_2 X_2^n + \cdots + n X_k^n \end{pmatrix}.$$

Now, define $\mathbb{P}$ as:

$$\begin{aligned}\mathbb{P} &\coloneqq \mathbb{X}\left(\mathbb{X}'\mathbb{X}\right)^{-1}\mathbb{X}' \\ \Rightarrow \mathbb{P}\mathbb{Y} &= \mathbb{X}\left(\mathbb{X}'\mathbb{X}\right)^{-1}\mathbb{X}'\mathbb{Y} \\ &= \mathbb{X}\hat{\boldsymbol{\beta}}_n.\end{aligned}$$

Since $\mathbb{X}\hat{\boldsymbol{\beta}}_n \in \operatorname{col}[\mathbb{X}]$, we realise that $\mathbb{P}\mathbb{Y}$ is a projection of $\mathbb{Y}$ onto $\operatorname{col}[\mathbb{X}]$. This is why $\mathbb{P}$ is called the *projection matrix*.



Let us also define

$$\mathbb{M} \coloneqq \mathbb{I} - \mathbb{P} = \mathbb{I} - \mathbb{X}\left(\mathbb{X}'\mathbb{X}\right)^{-1}\mathbb{X}'.$$

Note the following properties of $\mathbb{P}$ and $\mathbb{M}$:

▷ symmetric:

$$\begin{aligned}\mathbb{P}' &= \left(\mathbb{X}\left(\mathbb{X}'\mathbb{X}\right)^{-1}\mathbb{X}'\right)' = \mathbb{X}\left(\left(\mathbb{X}'\mathbb{X}\right)^{-1}\right)\mathbb{X}' \\ &= \mathbb{X}\left(\left(\mathbb{X}'\mathbb{X}\right)'\right)^{-1}\mathbb{X}' = \mathbb{X}\left(\mathbb{X}'\mathbb{X}\right)^{-1}\mathbb{X}' = \mathbb{P}, \\ \mathbb{M}' &= \left(\mathbb{I}-\mathbb{P}\right)' = \mathbb{I}-\mathbb{P} = \mathbb{M}.\end{aligned}$$

▷ idempotent:

$$\begin{aligned}\mathbb{P}^2 &= \left(\mathbb{X}\left(\mathbb{X}'\mathbb{X}\right)^{-1}\mathbb{X}'\right)\left(\mathbb{X}\left(\mathbb{X}'\mathbb{X}\right)^{-1}\mathbb{X}'\right) = \mathbb{P} \\ \mathbb{M}^2 &= \left(\mathbb{I}-\mathbb{P}\right)\left(\mathbb{I}-\mathbb{P}\right) = \mathbb{I}\left(\mathbb{I}-\mathbb{P}\right) - \mathbb{P}\left(\mathbb{I}-\mathbb{P}\right) \\ &= \left(\mathbb{I}-\mathbb{P}\right) - \mathbb{P} + \mathbb{P}^2 = \left(\mathbb{I}-\mathbb{P}\right) - \mathbb{P} + \mathbb{P} \\ &= \left(\mathbb{I}-\mathbb{P}\right) = \mathbb{M}.\end{aligned}$$

$\mathbb{M}$ is called the *residual maker matrix* since it projects $\mathbb{Y}$ onto the space orthogonal to $\operatorname{col}(\mathbb{X})$:

$$\begin{aligned}\mathbb{M}\mathbb{X} &= \left(\mathbb{I}-\mathbb{P}\right)\mathbb{X} = \mathbb{X} - \mathbb{X}\left(\mathbb{X}'\mathbb{X}\right)^{-1}\mathbb{X}'\mathbb{X} = 0, \\ \mathbb{M}\mathbb{Y} &= \left(\mathbb{I}-\mathbb{P}\right)\mathbb{Y} = \mathbb{Y} - \mathbb{P}\mathbb{Y} = \mathbb{Y} - \mathbb{X}\hat{\boldsymbol{\beta}}_n = \hat{\mathbb{U}}.\end{aligned}$$

These imply that

$$\mathbb{M}\mathbb{U} = \mathbb{M}\left(\mathbb{Y} - \mathbb{X}\boldsymbol{\beta}\right) = \hat{\mathbb{U}}.$$

3.2.6 Estimating subvectors of $\hat{\boldsymbol{\beta}}_n$

Recall that we can partition $\mathbf{X}$ and $\boldsymbol{\beta}$ and obtain subvectors of $\boldsymbol{\beta}$ by regressing residuals. Suppose:

$$Y = \mathbf{X}'\boldsymbol{\beta} = \mathbf{X}_1'\boldsymbol{\beta}_1 + \mathbf{X}_2'\boldsymbol{\beta}_2 + u$$

and we partition $\mathbb{X}$ correspondingly into $\mathbb{X}_1$ and $\mathbb{X}_2$ and define $\mathbb{P}_1$, $\mathbb{P}_2$, $\mathbb{M}_1$, $\mathbb{M}_2$ accordingly. Note:

$$(\mathbb{X}_i)_{n\times k_i} = \left(\mathbf{X}_{11}^1, \mathbf{X}_{12}^2, \ldots, \mathbf{X}_{1k_i}^n\right)' = \begin{pmatrix} X_{11}^1 & X_{12}^1 & \cdots & X_{1k_i}^1 \\ X_{11}^2 & X_{12}^2 & \cdots & X_{1k_i}^2 \\ \vdots & \vdots & \ddots & \vdots \\ X_{11}^n & X_{12}^n & \cdots & X_{1k_i}^n \end{pmatrix}, \forall i = 1, 2.$$

In this notation, we have

$$\mathbb{Y} = \mathbb{X}_1\hat{\boldsymbol{\beta}}_{1,n} + \mathbb{X}_2\hat{\boldsymbol{\beta}}_{2,n} + \hat{\mathbb{U}}.$$

We want to get rid of $\mathbb{X}_2$, we can use the fact that $\mathbb{M}_2\mathbb{X}_2 = 0$; i.e. multiplying both sides by $\mathbb{M}_2$:

$$\begin{aligned} \mathbb{M}_2\mathbb{Y} &= \mathbb{M}_2\mathbb{X}_1\hat{\boldsymbol{\beta}}_{1,n} + \mathbb{M}_2\mathbb{X}_2\hat{\boldsymbol{\beta}}_{2,n} + \mathbb{M}_2\hat{\mathbb{U}} \\ &= \mathbb{M}_2\mathbb{X}_1\hat{\boldsymbol{\beta}}_{1,n} + \mathbb{M}_2\hat{\mathbb{U}}. \end{aligned}$$

Note that

$$\mathbb{M}_2\hat{\mathbb{U}} = (\mathbb{I} - \mathbb{P}_2)\,\hat{\mathbb{U}} = \hat{\mathbb{U}} - \mathbb{P}_2\hat{\mathbb{U}} = \hat{\mathbb{U}}$$

where we use the fact that

$$\mathbb{P}\hat{\mathbb{U}} = \mathbb{P}\mathbb{Y} - \mathbb{P}\mathbb{X}\hat{\boldsymbol{\beta}}_n = \mathbb{X}\hat{\boldsymbol{\beta}}_n - \mathbb{X}\hat{\boldsymbol{\beta}}_n = \mathbf{0}$$

implies that $\mathbb{P}_2\hat{\mathbb{U}} = \mathbf{0}$. We therefore can write

$$\mathbb{M}_2\mathbb{Y} = \mathbb{M}_2\mathbb{X}_1\hat{\boldsymbol{\beta}}_{1,n} + \hat{\mathbb{U}}.$$

We wish to make a square matrix, so

$$\begin{aligned} (\mathbb{M}_2\mathbb{X}_1)'\,\mathbb{M}_2\mathbb{Y} &= (\mathbb{M}_2\mathbb{X}_1)'\,\mathbb{M}_2\mathbb{X}_1\hat{\boldsymbol{\beta}}_{1,n} + (\mathbb{M}_2\mathbb{X}_1)'\,\hat{\mathbb{U}}. \\ &= (\mathbb{M}_2\mathbb{X}_1)'\,\mathbb{M}_2\mathbb{X}_1\hat{\boldsymbol{\beta}}_{1,n} + \mathbb{X}_1'\mathbb{M}_2'\hat{\mathbb{U}} \\ [\mathbb{M}_2 = \mathbb{M}_2'] &= (\mathbb{M}_2\mathbb{X}_1)'\,\mathbb{M}_2\mathbb{X}_1\hat{\boldsymbol{\beta}}_{1,n} + \mathbb{X}_1'\mathbb{M}_2\hat{\mathbb{U}} \\ &= (\mathbb{M}_2\mathbb{X}_1)'\,\mathbb{M}_2\mathbb{X}_1\hat{\boldsymbol{\beta}}_{1,n} + \mathbb{X}_1'\hat{\mathbb{U}} \\ &= (\mathbb{M}_2\mathbb{X}_1)'\,\mathbb{M}_2\mathbb{X}_1\hat{\boldsymbol{\beta}}_{1,n}, \end{aligned}$$

where we use the fact that $\mathbb{X}'\hat{\mathbb{U}} = \mathbf{0} \Rightarrow \mathbb{X}_1'\hat{\mathbb{U}} = \mathbf{0}$. Finally, we obtain that

$$\hat{\boldsymbol{\beta}}_{1,n} = \left[(\mathbb{M}_2\mathbb{X}_1)'\,(\mathbb{M}_2\mathbb{X}_1)\right]^{-1}(\mathbb{M}_2\mathbb{X}_1)'\,(\mathbb{M}_2\mathbb{Y}).$$

Notice that this is the residual regression. The expression above is called the *Frisch-Waugh-Lowell decomposition*. Note that we could simplify the expression above to

$$\hat{\boldsymbol{\beta}}_{1,n} = \left[\mathbb{X}_1'\mathbb{M}_2\mathbb{X}_1\right]^{-1}\mathbb{X}_1'\mathbb{M}_2\mathbb{Y}.$$

Remark 3.2. Recall that when we partitioned $\mathbf{X}$ to $\mathbf{X}_1$ and $\mathbf{X}_2$ (and $\boldsymbol{\beta}$ correspondingly), we can obtain $\boldsymbol{\beta}_1$ by: (i) obtaining the residual, $\tilde{Y}$, by regression Y on $\mathbf{X}_2$; (ii) obtaining the residual, $\tilde{\mathbf{X}}_1$, by regressing $\mathbf{X}_1$ on $\mathbf{X}_2$; and (iii) regressing $\tilde{Y}$ on $\tilde{\mathbf{X}}_1$, and

$$\beta_1 = \mathbb{E}\left[\tilde{\mathbf{X}}_1\tilde{\mathbf{X}}_1'\right]^{-1}\mathbb{E}\left[\tilde{\mathbf{X}}_1\tilde{Y}\right].$$

Here, since $\mathbb{M}_2$ is the residual maker matrix:

$$\begin{aligned} \tilde{Y}' &= \mathbb{M}_2\mathbb{Y}, \\ \tilde{\mathbf{X}}_1' &= \mathbb{M}_2\mathbb{X}_1. \end{aligned}$$

Thus, the regression in step (iii) becomes $\mathbb{M}_2\mathbb{Y} = \mathbb{M}_2\mathbb{X}_1\tilde{\boldsymbol{\beta}} + \mathbb{M}_2\mathbb{U}$ so that

$$\tilde{\boldsymbol{\beta}} = \left((\mathbb{M}_2\mathbb{X}_1)'\,\mathbb{M}_2\mathbb{X}_1\right)^{-1}(\mathbb{M}_2\mathbb{X}_1)'\,(\mathbb{M}_2\mathbb{Y}),$$

which is the expression we obtained above.

3.2.7 Measures of fit

Define

$$R^2 \coloneqq \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS}, \tag{3.14}$$

where:

- ▷ $ESS = \sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}_n\right)^2$ is the *explained sum of squares;*
- ▷ $TSS = \sum_{i=1}^{n} \left(Y_i - \bar{Y}_n\right)^2$ is the *total sum of squares*;
- ▷ $SRR = \sum_{i=1}^{n} \left(\hat{u}^i\right)^2$ is the *sum of square residuals.*

To show the equality (3.14), it would be sufficient to show that $ESS + SSR = TSS$.

$$\begin{aligned} TSS &= \sum_{i=1}^{n} \left(Y^i - \bar{Y}_n\right)^2 \\ &= \sum_{i=1}^{n} \left(\underbrace{Y^i - \hat{Y}^i}_{=\hat{u}^i} + \hat{Y}^i - \bar{Y}_n \right)^2 \\ &= \underbrace{\sum_{i=1}^{n} \left(\hat{u}^i\right)^2}_{=SRR} + 2\sum_{i=1}^{n} \hat{u}^i \left(\hat{Y}^i - \bar{Y}_n\right) + \underbrace{\sum_{i=1}^{n} \left(\hat{Y}^i - \bar{Y}_n\right)^2}_{=ESS}. \end{aligned}$$

It now suffices to show the term in the middle is zero.

$$\begin{aligned} \sum_{i=1}^{n} \hat{u}^i \left(\hat{Y}^i - \bar{Y}_n\right) &= \sum_{i=1}^{n} \hat{u}^i \hat{Y}^i - \sum_{i=1}^{n} \hat{u}^i \bar{Y}_n \\ &= \sum_{i=1}^{n} \hat{u}^i \mathbf{X}^{i\prime} \hat{\boldsymbol{\beta}}_n - \bar{Y}_n \sum_{i=1}^{n} \hat{u}^i \\ &= 0, \end{aligned}$$

where we use the first-order condition, (3.12) and (3.13) to conclude that

$$\begin{aligned} \left(\sum_{i=1}^{n} \hat{u}^i \mathbf{X}^{i\prime}\right) \hat{\boldsymbol{\beta}}_n = 0, \\ \sum_{i=1}^{n} \hat{u}^i = 0. \end{aligned}$$

Therefore, we have that $ESS + SSR = TSS$. This also implies that

$$0 \leq R^2 \leq 1,$$

where:

- ▷ $R^2 = 1 \Leftrightarrow SSR = \sum_{i=1}^{n} \left(\hat{u}^i\right)^2 = 0 \Rightarrow \hat{u}^i = 0, \forall i \Rightarrow Y^i = \hat{Y}^i = \mathbf{X}^{i\prime}\hat{\boldsymbol{\beta}}_n$. So $R^2 = 1$ means that we can perfectly predict Y^i.
- ▷ $R^2 = 0 \Leftrightarrow ESS = \sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}_n\right)^2 = 0 \Rightarrow \hat{Y}_i = \bar{Y}_n, \forall i \Rightarrow \hat{\beta}_j = 0, \forall j = 1, 2, \ldots, k$. Hence, $R^2 = 0$ means that $\mathbf{X}^i$'s do not help us to predict Y^i.

Note that $R^2 = 0$ occurs when the model contains only an intercept. Thus, R^2 measures the "goodness of fit" of a model with additional explanatory variables relative to the fit of a model that includes only an intercept term. This also means that R^2 is not appropriate for models without constants.

What happens to R^2 as we add more variables; i.e. we increase k? A higher k would weakly increase the R^2 since adding an irrelevant variables leaves the fit unchanged while adding relevant would improve the fit.

Writing

$$R^2 = 1 - \frac{SSR}{TSS} = 1 - \frac{\sum_{i=1}^n \left(\hat{u}^i\right)^2}{\sum_{i=1}^n \left(Y_i - \bar{Y}_n\right)^2},$$

observe that R^2 can be viewed as an estimator of

$$1 - \frac{\text{Var}\left[u\right]}{\text{Var}\left[Y\right]}. \tag{3.15}$$

By replacing the estimate of $\text{Var}\left[u\right]$ and $\text{Var}\left[Y\right]$ with bias-adjusted versions (based on the degrees of freedom), we can obtain the *adjusted* R^2:

$$\bar{R}^2 = 1 - \frac{\frac{1}{n-(k+1)} \sum_{i=1}^n \left(\hat{u}^i\right)^2}{\frac{1}{n-1} \sum_{i=1}^n \left(Y_i - \bar{Y}_n\right)^2} = 1 - \frac{(n-1)}{n-k-1} \frac{\sum_{i=1}^n \left(\hat{u}^i\right)^2}{\sum_{i=1}^n \left(Y_i - \bar{Y}_n\right)^2}.$$

We see that $\bar{R}^2$ is not increasing in k. However, although $\bar{R}^2$ is bounded above at 1, it is no longer bounded below at zero; i.e. $\bar{R}^2$ can be negative. Note that there is no strong theoretical justification for $\bar{R}^2$.

Importantly, R^2 and $\bar{R}^2$ does not validate a causal interpretation whether it is high or low.[16]

3.2.8 Properties of the OLS estimator

Suppose we have $(Y, \mathbf{X}, u)$ such that

$$Y = \mathbf{X}'\boldsymbol{\beta} + u,$$

where $\mathbf{X} = (X_0, X_1, \ldots, X_k)$ with $X_0 = 1$. Assume that $\mathbb{E}\left[\mathbf{X}u\right] = \mathbf{0}$, $\mathbb{E}\left[\mathbf{X}\mathbf{X}'\right]$ exists and that there is no perfect collinearity in $\mathbf{X}$. Finally, assume that $\left(Y^1, \mathbf{X}^1\right), \cdots, (Y^n, \mathbf{X}^n) \overset{\text{iid}}{\sim} P$, where $(Y, \mathbf{X}) \sim P$. Let $\hat{\boldsymbol{\beta}}_n$ be the OLS estimator of $\boldsymbol{\beta}$.

The assumptions made in each of the subsections below are self-contained.

Unbiasedness

Proposition 3.3. *(The OLS estimator is unbiased). Suppose further that* $\mathbb{E}\left[u|\mathbf{X}\right] = 0$ *(i.e. u is mean independent of* $\mathbf{X}$*). Then,*

$$\mathbb{E}\left[\hat{\boldsymbol{\beta}}_n\right] = \boldsymbol{\beta}.$$

Proof. Recall that

$$\hat{\boldsymbol{\beta}}_n = \left(\mathbb{X}'\mathbb{X}\right)^{-1} \mathbb{X}'\mathbb{Y},$$
$$\mathbb{Y} = \mathbb{X}\boldsymbol{\beta} + \mathbb{U}.$$

Then,

$$\begin{aligned}\hat{\boldsymbol{\beta}}_n &= \left(\mathbb{X}'\mathbb{X}\right)^{-1} \mathbb{X}' \left(\mathbb{X}\boldsymbol{\beta} + \mathbb{U}\right) \\ &= \boldsymbol{\beta} + \left(\mathbb{X}'\mathbb{X}\right)^{-1} \mathbb{X}'\mathbb{U}.\end{aligned}$$

[16] For example, in a randomised drug test, and we test if the dose of drugs led to a difference between the control and the treatment groups. If we find $R^2 = 0$, we would then make a causal inference that the drug is ineffective.

Note that the first-order condition for OLS gives that $\mathbb{X}'\hat{\mathbb{U}} = 0$, not $\mathbb{X}'\mathbb{U} = \mathbf{0}$ so that we cannot immediately conclude that the second term is zero. However,

$$\begin{aligned}\mathbb{E}\left[\hat{\boldsymbol{\beta}}_n|\mathbf{X}^1,\mathbf{X}^2,\ldots,\mathbf{X}^n\right] &= \boldsymbol{\beta} + \left(\mathbb{X}'\mathbb{X}\right)^{-1}\mathbb{X}'\mathbb{E}\left[\mathbb{U}|\mathbf{X}^1,\mathbf{X}^2,\ldots,\mathbf{X}^n\right] \\ &= \boldsymbol{\beta}\end{aligned}$$

since $\mathbb{E}\left[u^i|\mathbf{X}^i\right] = 0, \forall i$. This follows from the fact that $\mathbb{E}\left[u|\mathbf{X}\right] = 0$ and $\left(Y^i, \mathbf{X}^i\right)$ are iid, which, in turn, implies that $\mathbb{E}\left[u^i|\mathbf{X}^1,\mathbf{X}^2,\ldots,\mathbf{X}^n\right] = \mathbb{E}\left[u^i|\mathbf{X}^i\right] = 0$.

By Law of Iterated Expectations,

$$\begin{aligned}\mathbb{E}\left[\mathbb{E}\left[\hat{\boldsymbol{\beta}}_n|\mathbf{X}^1,\mathbf{X}^2,\ldots,\mathbf{X}^n\right]\right] &= \mathbb{E}\left[\boldsymbol{\beta}\right] \\ \Leftrightarrow \mathbb{E}\left[\hat{\boldsymbol{\beta}}_n\right] &= \boldsymbol{\beta}.\end{aligned}$$

■

Gauss-Markov Theorem

Definition 3.2. We say that u is *homoscedastic* if $\text{Var}\left[u|\mathbf{X}\right]$ is constant and *heteroscedastic* otherwise.

Proposition 3.4. *Suppose further that* $\mathbb{E}\left[u|\mathbf{X}\right] = 0$, $\text{Var}\left[u|\mathbf{X}\right] = \sigma^2$ *(i.e. u is homoscedastic), then* $\hat{\boldsymbol{\beta}}_n$ *is the "best" estimator of* $\boldsymbol{\beta}$ *in the sense of having the "smallest"* $\text{Var}\left[\mathbb{A}'\mathbb{Y}|\mathbf{X}^1,\mathbf{X}^2,\ldots,\mathbf{X}^n\right]$ *among all estimators of form* $\mathbb{A}'\mathbb{Y}$ *for* $\mathbb{A} = \mathbb{A}\left(\mathbf{X}^1,\mathbf{X}^2,\ldots,\mathbf{X}^n\right)$ *such that* $\mathbb{E}\left[\mathbb{A}'\mathbb{Y}|\mathbf{X}^1,\mathbf{X}^2,\ldots,\mathbf{X}^n\right] = \boldsymbol{\beta}$ *(i.e unbiased).*

Proof. Since $\mathbb{E}\left[u|\mathbf{X}\right] = 0$, then

$$\begin{aligned}\mathbb{E}\left[\mathbb{A}'\mathbb{Y}|\mathbf{X},\mathbf{X}^2,\ldots,\mathbf{X}^n\right] &= \mathbb{E}\left[\mathbb{A}'\left(\mathbb{X}\boldsymbol{\beta}+\mathbb{U}\right)|\mathbf{X},\mathbf{X}^2,\ldots,\mathbf{X}^n\right] \\ &= \mathbb{A}'\mathbb{E}\left[\mathbb{X}|\mathbf{X},\mathbf{X}^2,\ldots,\mathbf{X}^n\right]\boldsymbol{\beta} + \mathbb{A}'\mathbb{E}\left[\mathbb{U}|\mathbf{X},\mathbf{X}^2,\ldots,\mathbf{X}^n\right] \\ &= \mathbb{A}'\mathbb{X}\boldsymbol{\beta}.\end{aligned}$$

Thus, $\mathbb{A}$ is unbiased if and only if $\mathbb{A}'\mathbb{X} = \mathbb{I}$.

Note that

$$\begin{aligned}\text{Var}\left[\mathbb{A}'\mathbb{Y}|\mathbf{X},\mathbf{X}^2,\ldots,\mathbf{X}^n\right] &= \mathbb{A}'\,\text{Var}\left[\mathbb{Y}|\mathbf{X},\mathbf{X}^2,\ldots,\mathbf{X}^n\right]\mathbb{A} \\ &= \mathbb{A}'\,\text{Var}\left[\mathbb{U}|\mathbf{X},\mathbf{X}^2,\ldots,\mathbf{X}^n\right]\mathbb{A} \\ &= \mathbb{A}'\mathbb{A}\sigma^2.\end{aligned}$$

The OLS estimator is given by $\mathbb{A}' = \left(\mathbb{X}'\mathbb{X}\right)^{-1}\mathbb{X}'$. In this case,

$$\begin{aligned}\text{Var}\left[\mathbb{A}'\mathbb{Y}|\mathbf{X},\mathbf{X}^2,\ldots,\mathbf{X}^n\right] &= \left(\mathbb{X}'\mathbb{X}\right)^{-1}\mathbb{X}'\mathbb{X}\left(\mathbb{X}'\mathbb{X}\right)^{-1}\sigma^2 \\ &= \left(\mathbb{X}'\mathbb{X}\right)^{-1}\sigma^2.\end{aligned}$$

Thus, to show that $\text{Var}\left[\mathbb{A}'\mathbb{Y}|\mathbf{X},\mathbf{X}^2,\ldots,\mathbf{X}^n\right] \geq \text{Var}\left[\mathbb{A}'\mathbb{Y}|\mathbf{X},\mathbf{X}^2,\ldots,\mathbf{X}^n\right]$ for any $\mathbb{A}$ such that $\mathbb{A}'\mathbb{X} = \mathbb{I}$ is equivalent to showing that, for any $\mathbb{A}$ such that $\mathbb{A}'\mathbb{X} = \mathbb{I}$,

$$\mathbb{A}'\mathbb{A}\sigma^2 \geq \left(\mathbb{X}'\mathbb{X}\right)^{-1}\sigma^2 \Rightarrow \mathbb{A}'\mathbb{A} - \left(\mathbb{X}'\mathbb{X}\right)^{-1} \geq \mathbf{0}.$$

Since we are dealing with matrices, the last inequality means that we need to show that $\mathbb{A}'\mathbb{A} - \left(\mathbb{X}'\mathbb{X}\right)^{-1}$ is positive semi-definite for all $\mathbb{A}$ such that $\mathbb{A}'\mathbb{X} = \mathbb{I}$. That is, we want to show that for any $\mathbf{c} \in \mathbb{R}^{k+1}$ that is nonzero, we have

$$\mathbf{c}'\left(\mathbb{A}'\mathbb{A} - \left(\mathbb{X}'\mathbb{X}\right)^{-1}\right)\mathbf{c} \geq 0.$$

Define $\mathbb{C}_{n\times(k+1)} \coloneqq \mathbb{A} - \mathbb{X}\left(\mathbb{X}'\mathbb{X}\right)^{-1}$. Then,

$$\begin{aligned}
\mathbb{X}'\mathbb{C} &= \mathbb{X}'\left(\mathbb{A} - \mathbb{X}\left(\mathbb{X}'\mathbb{X}\right)^{-1}\right) \\
&= \mathbb{X}'\mathbb{A} - \mathbb{X}'\mathbb{X}\left(\mathbb{X}'\mathbb{X}\right)^{-1} \\
&= \mathbb{X}'\mathbb{A} - \mathbb{I} \\
&= \mathbf{0}_{(k+1)\times(k+1)}, \\
\Leftrightarrow \mathbb{C}'\mathbb{X} &= \mathbf{0}_{(k+1)\times(k+1)},
\end{aligned}$$

where we use fact that $\mathbb{A}'\mathbb{X} = \mathbb{I} \Rightarrow \mathbb{X}'\mathbb{A} = \mathbb{I}' = \mathbb{I}$. This implies that

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

Hence,

$$\begin{aligned}
\mathbb{A}'\mathbb{A} - \left(\mathbb{X}'\mathbb{X}\right)^{-1} &= \left(\mathbb{C} + \mathbb{X}\left(\mathbb{X}'\mathbb{X}\right)^{-1}\right)'\left(\mathbb{C} + \mathbb{X}\left(\mathbb{X}'\mathbb{X}\right)^{-1}\right) - \left(\mathbb{X}'\mathbb{X}\right)^{-1} \\
&= \left(\mathbb{C}' + \left(\mathbb{X}'\mathbb{X}\right)^{-1}\mathbb{X}'\right)\left(\mathbb{C} + \mathbb{X}\left(\mathbb{X}'\mathbb{X}\right)^{-1}\right) - \left(\mathbb{X}'\mathbb{X}\right)^{-1} \\
&= \mathbb{C}'\mathbb{C} + \mathbb{C}'\mathbb{X}\left(\mathbb{X}'\mathbb{X}\right)^{-1} + \left(\mathbb{X}'\mathbb{X}\right)^{-1}\mathbb{X}'\mathbb{C} \\
&\quad + \left(\mathbb{X}'\mathbb{X}\right)^{-1}\mathbb{X}'\mathbb{X}\left(\mathbb{X}'\mathbb{X}\right)^{-1} - \left(\mathbb{X}'\mathbb{X}\right)^{-1} \\
&= \mathbb{C}'\mathbb{C} + \mathbf{0} + \mathbf{0} + \left(\mathbb{X}'\mathbb{X}\right)^{-1} - \left(\mathbb{X}'\mathbb{X}\right)^{-1} \\
&= \mathbb{C}'\mathbb{C}.
\end{aligned}$$

Let $\mathbf{c} \in \mathbb{R}^{k+1}$ be any nonzero vector and consider $\mathbb{C}$. Since $\mathbb{C}\mathbf{c}$ is a $n \times 1$ column vector, we can denote it by

$$\mathbb{C}\mathbf{c} = \left(\alpha_1, \alpha_2, \ldots, \alpha_n\right)'.$$

Then,

$$\mathbf{c}'\mathbb{C}'\mathbb{C}\mathbf{c} = \left(\mathbb{C}\mathbf{c}\right)'\mathbb{C}\mathbf{c} = \sum_{i=1}^{n}\alpha_i^2 \geq 0.$$

Therefore, $\mathbb{C}'\mathbb{C}$ is positive semi-definite and we are done. ■

Remark 3.3. Note that OLS estimator is the best estimator among *linear* and *unbiased* estimators with *homoscedastic* errors. If we were to relax these requirements, then there are other estimators which are "better".

Consistency Without requiring further assumptions, we can show that the OLS estimator is consistent.

Proposition 3.5. $\hat{\boldsymbol{\beta}}_n \xrightarrow{\text{p}} \boldsymbol{\beta}$.

Proof. Recall that

$$\begin{aligned}
\boldsymbol{\beta} &= \mathbb{E}\left[\mathbf{X}\mathbf{X}'\right]^{-1}\mathbb{E}\left[\mathbf{X}Y\right], \\
\hat{\boldsymbol{\beta}}_n &= \left(\frac{1}{n}\sum_{n=1}^{n}\mathbf{X}^i\mathbf{X}^{i\prime}\right)^{-1}\left(\frac{1}{n}\sum_{n=1}^{n}\mathbf{X}^iY^i\right).
\end{aligned}$$

First, since $\mathbb{E}\left[\mathbf{X}^i\mathbf{X}^{i\prime}\right] = \mathbb{E}\left[\mathbf{X}\mathbf{X}'\right]$ exists and $\mathbf{X}^i$'s are iid,

$$\frac{1}{n}\sum_{n=1}^{n}\mathbf{X}^i\mathbf{X}^{i\prime} \xrightarrow{\text{p}} \mathbb{E}\left[\mathbf{X}\mathbf{X}'\right].$$

Since we are given that there is no perfect collinearity in $\mathbf{X}$, then $\mathbb{E}[\mathbf{X}\mathbf{X}']$ is invertible. Then, by the Continuous Mapping Theorem,

$$\left(\frac{1}{n}\sum_{n=1}^{n}\mathbf{X}^i\mathbf{X}^{i\prime}\right)^{-1} \xrightarrow{\text{p}} \mathbb{E}[\mathbf{X}\mathbf{X}']^{-1}.$$

Similarly, since $\mathbb{E}\left[\mathbf{X}^iY^i\right] = \mathbb{E}[\mathbf{X}Y] = \mathbb{E}[\mathbf{X}(\mathbf{X}'\boldsymbol{\beta}+u)] = \mathbb{E}[\mathbf{X}\mathbf{X}']\boldsymbol{\beta}$ and $\mathbb{E}[\mathbf{X}\mathbf{X}']$ exists, and $\left(\mathbf{X}^i, Y^i\right)$'s are iid,

$$\frac{1}{n}\sum_{n=1}^{n}\mathbf{X}^iY^i \xrightarrow{\text{p}} \mathbb{E}\left[\mathbf{X}^iY^i\right].$$

Since convergence in marginal probabilities implies converge in joint probabilities:

$$\left(\left(\frac{1}{n}\sum_{n=1}^{n}\mathbf{X}^i\mathbf{X}^{i\prime}\right)^{-1}, \frac{1}{n}\sum_{n=1}^{n}\mathbf{X}^iY^i\right) \xrightarrow{\text{p}} \left(\mathbb{E}[\mathbf{X}\mathbf{X}']^{-1}, \mathbb{E}\left[\mathbf{X}^iY^i\right]\right).$$

Noting that multiplication is a continuous operation, and given above, by the Continuous Mapping Theorem,

$$\hat{\boldsymbol{\beta}}_n = \left(\frac{1}{n}\sum_{n=1}^{n}\mathbf{X}^i\mathbf{X}^{i\prime}\right)^{-1}\left(\frac{1}{n}\sum_{n=1}^{n}\mathbf{X}^iY^i\right) \xrightarrow{\text{p}} \mathbb{E}[\mathbf{X}\mathbf{X}']^{-1}\mathbb{E}[\mathbf{X}Y] = \boldsymbol{\beta}. \qquad \blacksquare$$

Limit distribution

Proposition 3.6. *Suppose further that* $\text{Var}[\mathbf{X}u]$ *exists, then*

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\right) \xrightarrow{\text{d}} N(0, \Omega),$$

where

$$\Omega = \mathbb{E}[\mathbf{X}\mathbf{X}']^{-1}\,\text{Var}[\mathbf{X}u]\,\mathbb{E}[\mathbf{X}\mathbf{X}']^{-1}.$$

Proof. Substituting $Y^i = \mathbf{X}^{i\prime}\boldsymbol{\beta} + u^i$ into $\hat{\boldsymbol{\beta}}_n$ yields

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_n &= \left(\frac{1}{n}\sum_{n=1}^{n}\mathbf{X}^i\mathbf{X}^{i\prime}\right)^{-1}\left(\frac{1}{n}\sum_{n=1}^{n}\mathbf{X}^i\left(\mathbf{X}^{i\prime}\boldsymbol{\beta}+u^i\right)\right) \\
&= \boldsymbol{\beta} + \left(\frac{1}{n}\sum_{n=1}^{n}\mathbf{X}^i\mathbf{X}^{i\prime}\right)^{-1}\left(\frac{1}{n}\sum_{n=1}^{n}\mathbf{X}^iu^i\right) \\
\Rightarrow \sqrt{n}\left(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\right) &= \left(\frac{1}{n}\sum_{n=1}^{n}\mathbf{X}^i\mathbf{X}^{i\prime}\right)^{-1}\left(\sqrt{n}\frac{1}{n}\sum_{n=1}^{n}\mathbf{X}^iu^i\right).
\end{aligned}$$

By assumption, $\text{Var}[\mathbf{X}u]$ exists and $\mathbf{X}^iu^i$'s are iid, then by the Central Limit Theorem:

$$\sqrt{n}\left(\frac{1}{n}\sum_{n=1}^{n}\mathbf{X}^iu^i\right) \xrightarrow{\text{d}} N(0, \text{Var}[\mathbf{X}u]),$$

where we used the fact that that $\mathbb{E}[\mathbf{X}u] = \mathbf{0}$ by assumption. We already showed previously that

$$\left(\frac{1}{n}\sum_{n=1}^{n}\mathbf{X}^i\mathbf{X}^{i\prime}\right)^{-1} \xrightarrow{\text{p}} \mathbb{E}[\mathbf{X}\mathbf{X}']^{-1}.$$

Hence, by Slutsky's Lemma,

$$\begin{aligned}\sqrt{n}\left(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\right) &= \left(\frac{1}{n}\sum_{n=1}^{n}\mathbf{X}^i\mathbf{X}^{i\prime}\right)^{-1}\left(\sqrt{n}\frac{1}{n}\sum_{n=1}^{n}\mathbf{X}^i u^i\right)\\ &\xrightarrow{\text{d}} \mathbb{E}\left[\mathbf{X}\mathbf{X}'\right]^{-1} N\left(0, \text{Var}\left[\mathbf{X}u\right]\right).\end{aligned}$$

Since, $\mathbf{X}\mathbf{X}'$ is symmetric, we obtain that

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\right) \xrightarrow{\text{d}} N\left(0, \mathbb{E}\left[\mathbf{X}\mathbf{X}'\right]^{-1}\text{Var}\left[\mathbf{X}u\right]\mathbb{E}\left[\mathbf{X}\mathbf{X}'\right]^{-1}\right). \qquad \blacksquare$$

Corollary 3.2. *Suppose, in addition, that* $\mathbb{E}\left[u|\mathbf{X}\right] = 0$ *and* $\text{Var}\left[u|\mathbf{X}\right] = \sigma^2$ *(i.e. u is homoscedastic), then*

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\right) \xrightarrow{\text{d}} N\left(0, \sigma^2\mathbb{E}\left[\mathbf{X}\mathbf{X}'\right]^{-1}\right)$$

Proof. Notice that

$$\begin{aligned}\text{Var}\left[\mathbf{X}u\right] &= \mathbb{E}\left[\left(\mathbf{X}u - \mathbb{E}\left[\mathbf{X}u\right]\right)\left(\mathbf{X}u - \mathbb{E}\left[\mathbf{X}u\right]\right)'\right]\\ \left[\mathbb{E}\left[\mathbf{X}u\right] = 0\right] &= \mathbb{E}\left[\mathbf{X}\mathbf{X}'u^2\right]\\ \left[\text{LIE}\right] &= \mathbb{E}\left[\mathbf{X}\mathbf{X}'\mathbb{E}\left[u^2|\mathbf{X}\right]\right].\end{aligned}$$

Since,

$$\begin{aligned}\text{Var}\left[u|\mathbf{X}\right] &= \mathbb{E}\left[\left(u - \mathbb{E}\left[u|\mathbf{X}\right]\right)^2|\mathbf{X}\right]\\ &= \mathbb{E}\left[u^2|\mathbf{X}\right] - \mathbb{E}\left[u|\mathbf{X}\right]^2\\ &= \mathbb{E}\left[u^2|\mathbf{X}\right] = \sigma^2.\end{aligned}$$

we can write

$$\text{Var}\left[\mathbf{X}u\right] = \mathbb{E}\left[\mathbf{X}\mathbf{X}'\right]\sigma^2.$$

Thus,

$$\begin{aligned}\Omega &= \mathbb{E}\left[\mathbf{X}\mathbf{X}'\right]^{-1}\text{Var}\left[\mathbf{X}u\right]\mathbb{E}\left[\mathbf{X}\mathbf{X}'\right]^{-1}\\ &= \mathbb{E}\left[\mathbf{X}\mathbf{X}'\right]^{-1}\mathbb{E}\left[\mathbf{X}\mathbf{X}'\right]\sigma^2\mathbb{E}\left[\mathbf{X}\mathbf{X}'\right]^{-1}\\ &= \sigma^2\mathbb{E}\left[\mathbf{X}\mathbf{X}'\right]^{-1}.\end{aligned} \qquad \blacksquare$$

3.2.9 Estimating Ω

Recall our assumptions: $\mathbb{E}\left[\mathbf{X}u\right] = \mathbf{0}$, $X_0 = 1$; $\mathbb{E}\left[\mathbf{X}\mathbf{X}'\right]$ exists; no perfect collinearity in $\mathbf{X}$; $\text{Var}\left[\mathbf{X}u\right]$ exists. Then, we showed that

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\right) \xrightarrow{\text{d}} N\left(0, \Omega\right),$$

where

$$\Omega = \mathbb{E}\left[\mathbf{X}\mathbf{X}'\right]^{-1}\text{Var}\left[\mathbf{X}u\right]\mathbb{E}\left[\mathbf{X}\mathbf{X}'\right]^{-1}.$$

Since we do not observe u, we do not know Ω. Here, we focus on deriving a consistent estimator of Ω.

Under homoscedasticity Suppose u is homoscedastic so that

$$\mathbb{E}\left[u|\mathbf{X}\right]=0,\mathrm{Var}\left[u|\mathbf{X}\right]=\sigma^2.$$

Recall that, in this case, Ω simplifies to

$$\Omega=\sigma^2\mathbb{E}\left[\mathbf{X}\mathbf{X}'\right]^{-1}.$$

So, a natural estimator for Ω is:[17]

$$\tilde{\Omega}_n\coloneqq\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}^i\mathbf{X}^{i\prime}\right)^{-1}\hat{\sigma}_n^2,$$
$$\hat{\sigma}_n^2\coloneqq\frac{1}{n}\sum_{i=1}^{n}\left(\hat{u}^i\right)^2.$$

We know from before that (see consistency of OLS)

$$\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}^i\mathbf{X}^{i\prime}\right)^{-1}\xrightarrow{\mathrm{p}}\mathbb{E}\left[\mathbf{X}^i\mathbf{X}^{i\prime}\right]^{-1}.$$

Thus, it remains to show that $\hat{\sigma}_n^2$ is consistent. Note that if we could observe u_i and define $\hat{\sigma}_n^2$ using u^i, then it would be easy to show that $\hat{\sigma}_n^2$ is consistent (use WLLN). But we cannot observe u^i. However, we can still introduce u^i to show that $\hat{\sigma}_n^2$ is consistent—to do that write out $\hat{u}^i$ while adding and subtracting $\mathbf{X}_i'\boldsymbol{\beta}$:

$$\hat{u}^i=Y^i-\mathbf{X}^{i\prime}\hat{\boldsymbol{\beta}}_n=Y^i-\mathbf{X}^{i\prime}\boldsymbol{\beta}+\mathbf{X}^{i\prime}\boldsymbol{\beta}-\mathbf{X}^{i\prime}\hat{\boldsymbol{\beta}}_n=u^i-\mathbf{X}^{i\prime}\left(\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}\right).$$

This implies that

$$\left(\hat{u}^i\right)^2=\left(u^i\right)^2-2u^i\mathbf{X}^{i\prime}\left(\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}\right)+\left[\mathbf{X}^{i\prime}\left(\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}\right)\right]^2,\tag{3.16}$$

which, in turn, gives that

$$\hat{\sigma}_n^2=\frac{1}{n}\sum_{i=1}^{n}\left(u^i\right)^2-2\frac{1}{n}\sum_{i=1}^{n}u^i\mathbf{X}^{i\prime}\left(\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}\right)+\frac{1}{n}\sum_{i=1}^{n}\left[\mathbf{X}^{i\prime}\left(\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}\right)\right]^2.$$

Let us consider each term in turn:

- ▷ $\frac{1}{n}\sum_{i=1}^{n}\left(u^i\right)^2$: Since u^i's are iid, and $\mathbb{E}\left[u^i\right]=0<\infty$, WLLN immediately gives us that $\frac{1}{n}\sum_{i=1}^{n}\left(u^i\right)^2\xrightarrow{\mathrm{p}}\mathrm{Var}\left[u\right]$.
- ▷ $\frac{1}{n}\sum_{i=1}^{n}u^i\mathbf{X}^{i\prime}(\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta})$. Notice that the betas can be taken outside of the summation since they do not depend on i. Then, we use the fact that $\hat{\boldsymbol{\beta}}_n\xrightarrow{\mathrm{p}}\boldsymbol{\beta}$ and $\frac{1}{n}\sum_{i=1}^{n}u^i\mathbf{X}^{i\prime}\xrightarrow{\mathrm{p}}\mathbb{E}\left[\mathbf{X}u\right]=\mathbf{0}$, and the Continuous Mapping Theorem (as well as the fact that convergence in marginal probabilities implies convergence in joint probabilities) to conclude that

$$\left(\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}\right)\frac{1}{n}\sum_{i=1}^{n}u^i\mathbf{X}^{i\prime}=o_P\left(1\right)o_P\left(1\right)=o_P\left(1\right).$$

- ▷ $\frac{1}{n}\sum_{i=1}^{n}\left[\mathbf{X}^{i\prime}\left(\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}\right)\right]^2$: Since the term is positive, it would be sufficient to show that its upper bound is zero. With $\mathbf{u}$ and $\mathbf{v}$ being $(k+1)\times 1$ vectors, Cauchy-Schwartz inequality tells us that

$$\left(\mathbf{u}\cdot\mathbf{v}\right)^2=\left(\sum_{j=1}^{k}u_jv_j\right)^2\leq\left(\sum_{j=1}^{k}u_j^2\right)\left(\sum_{j=1}^{k}v_j^2\right)=\left(\mathbf{u}\cdot\mathbf{u}\right)\left(\mathbf{v}\cdot\mathbf{v}\right)=\left|\mathbf{u}\right|^2\left|\mathbf{v}\right|^2.$$

[17] Note that, here, we are interested in consistency and not unbiasedness. We could divide by $1/\left(n-k-1\right)$ to obtain the same result using unbiased estimators.

So, for each i,

$$\left(\mathbf{X}^{i\prime}\left(\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}\right)\right)^2=\left(\sum_{j=1}^{k}X_j^{i\prime}\left(\hat{\beta}_{n,j}-\beta_j\right)\right)^2\leq\left(\sum_{j=1}^{k}\left(X_j^{i\prime}\right)^2\right)\left(\sum_{j=1}^{k}\left(\hat{\beta}_{n,j}-\beta_j\right)^2\right)=\left|\mathbf{X}^i\right|^2\left|\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}\right|^2, \tag{3.17}$$

where $|\cdot|$ is the Euclidean norm. Summing across i and dividing by n, we obtain that

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\left[\mathbf{X}^{i\prime}\left(\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}\right)\right]^2&\leq\frac{1}{n}\sum_{i=1}^{n}\left(\left|\mathbf{X}^i\right|^2\left|\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}\right|^2\right)\\&=\left(\frac{1}{n}\sum_{i=1}^{n}\left|\mathbf{X}^i\right|^2\right)\left|\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}\right|^2\end{aligned}$$

That $\mathbb{E}\left[\mathbf{X}\mathbf{X}'\right]<\infty$ implies that $\left(X_j^i\right)^2<\infty$ for all $j=0,1\ldots,k$:

$$\mathbb{E}\left[\mathbf{X}\mathbf{X}'\right]=\begin{pmatrix}X_0\\X_1\\\vdots\\X_k\end{pmatrix}\begin{pmatrix}X_0&X_1&\cdots&X_k\end{pmatrix}=\mathbb{E}\begin{pmatrix}X_0^2&X_0X_1&\cdots&X_0X_k\\X_0X_1&X_1^2&\cdots&X_1X_k\\\vdots&\vdots&\ddots&\vdots\\X_0X_k&X_1X_k&\cdots&X_k^2\end{pmatrix}<\infty. \tag{3.18}$$

This means that $\left(\frac{1}{n}\sum_{i=1}^{n}\left|\mathbf{X}^i\right|^2\right)=O_P\left(1\right)$. To see this, note that

$$\left(\frac{1}{n}\sum_{i=1}^{n}\left|\mathbf{X}^i\right|^2\right)=\left(\frac{1}{n}\sum_{i=1}^{n}\sum_{j=0}^{k}\left(X_j^i\right)^2\right)=\left(\sum_{j=0}^{k}\frac{1}{n}\sum_{i=1}^{n}\left(X_j^i\right)^2\right).$$

Since X_j^i's are iid, by WLLN, we know that $\frac{1}{n}\sum_{i=1}^{n}\left(X_j^i\right)^2\xrightarrow{\text{p}}\mathbb{E}\left[\left(X_j^i\right)^2\right]$. Then, by the fact that convergence in marginal probabilities implies convergence in joint probabilities, using the Continuous Mapping Theorem,

$$\sum_{j=0}^{k}\frac{1}{n}\sum_{i=1}^{n}\left(X_j^i\right)^2\xrightarrow{\text{p}}\sum_{j=0}^{k}\mathbb{E}\left[\left(X_j^i\right)^2\right].$$

Since convergence in probability implies convergence in distribution, which, in turn, implies tightness, it follows that $\left(\frac{1}{n}\sum_{i=1}^{n}\left|\mathbf{X}^{i\prime}\right|^2\right)=O_P\left(1\right)$. We also know that $\left(\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}\right)^2=o_P\left(1\right)o_P\left(1\right)=o_P\left(1\right)$. Therefore,

$$\frac{1}{n}\sum_{i=1}^{n}\left[\mathbf{X}^{i\prime}\left(\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}\right)\right]^2\leq\left(\frac{1}{n}\sum_{i=1}^{n}\left|\mathbf{X}^i\right|^2\right)\left|\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}\right|^2=O_P\left(1\right)o_p\left(1\right)=o_P\left(1\right).$$

Thus, we obtain the result that

$$\hat{\sigma}_n^2=\text{Var}\left[u\right].$$

Therefore, we have that (since convergence in marginal probabilities implies convergence in joint probabilities),

$$\left(\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}^i\mathbf{X}^{i\prime}\right)^{-1},\hat{\sigma}_n^2\right)\xrightarrow{\text{p}}\left(\mathbb{E}\left[\mathbf{X}^i\mathbf{X}^{i\prime}\right]^{-1},\text{Var}\left[u\right]\right),$$

and applying the Continuous Mapping Theorem (with $g\left(a,b\right)=ab$, which is continuous) yields that

$$\tilde{\Omega}_n=\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}^i\mathbf{X}^{i\prime}\right)^{-1}\hat{\sigma}_n^2\xrightarrow{\text{p}}\text{Var}\left[\mathbf{X}u\right].$$

Under heteroscedasticity A natural estimator here is

$$\hat{\Omega}_n = \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}^i\mathbf{X}^{i\prime}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}^i\mathbf{X}^{i\prime}\left(\hat{u}^i\right)^2\right)\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}^i\mathbf{X}^{i\prime}\right)^{-1}. \tag{3.19}$$

We already showed that $(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}^i\mathbf{X}^{i\prime})^{-1} \xrightarrow{\text{p}} \mathbb{E}[\mathbf{X}^i\mathbf{X}^{i\prime}]^{-1}$ so our focus here is the term in the middle, which we hope converges in probability to $\text{Var}[\mathbf{X}u]$. As before, we want to introduce u^i in the expression—so add and subtract to obtain

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}^i\mathbf{X}^{i\prime}\left(\hat{u}^i\right)^2 &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}^i\mathbf{X}^{i\prime}\left(\left(\hat{u}^i\right)^2-\left(u^i\right)^2+\left(u^i\right)^2\right)\\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}^i\mathbf{X}^{i\prime}\left(u^i\right)^2+\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}^i\mathbf{X}^{i\prime}\left(\left(\hat{u}^i\right)^2-\left(u^i\right)^2\right).\end{aligned} \tag{3.20}$$

Since $\mathbf{X}^i$'s and u^i's are both iid and, by assumption, $\text{Var}[\mathbf{X}u]$ exists, applying WLLN gives us that

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}^i\mathbf{X}^{i\prime}\left(u^i\right)^2 \xrightarrow{\text{p}} \mathbb{E}\left[\mathbf{X}^i\mathbf{X}^{i\prime}\left(u^i\right)^2\right].$$

Now, we consider the second term, which is a $(k+1)\times(k+1)$ matrix. Let us consider the (j,j') element of this matrix (which is a scalar):

$$\frac{1}{n}\sum_{i=1}^{n}X_j^iX_{j'}^i\left(\left(\hat{u}^i\right)^2-\left(u^i\right)^2\right).$$

Taking absolute value, and using the Triangle Inequality,

$$\begin{aligned}\left|\frac{1}{n}\sum_{i=1}^{n}X_j^iX_{j'}^i\left(\left(\hat{u}^i\right)^2-\left(u^i\right)^2\right)\right| &\le \frac{1}{n}\sum_{i=1}^{n}\left|X_j^iX_{j'}^i\left(\hat{u}^i\right)^2-\left(u^i\right)^2\right|\\ &= \frac{1}{n}\sum_{i=1}^{n}\left|X_j^iX_{j'}^i\right|\left|\left(\hat{u}^i\right)^2-\left(u^i\right)^2\right|.\end{aligned}$$

Notice that we cannot take $|\left(\hat{u}^i\right)^2-\left(u^i\right)^2|$ out of the summation as we could do with $(\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta})$ in the homoscedastic case. So we use a "trick" instead—take the maximum so that the term would no longer depend on i:

$$\frac{1}{n}\sum_{i=1}^{n}\left|X_j^iX_{j'}^i\right|\left|\left(\hat{u}^i\right)^2-\left(u^i\right)^2\right| \le \left(\max_{1\le i\le n}\left|\left(\hat{u}^i\right)^2-\left(u^i\right)^2\right|\right)\frac{1}{n}\sum_{i=1}^{n}\left|X_j^iX_{j'}^i\right|. \tag{3.21}$$

Notice that $X_j^iX_{j'}^i$ is finite (see (3.18)) given that $\mathbb{E}[\mathbf{X}\mathbf{X}'] < \infty$. Hence,

$$\frac{1}{n}\sum_{i=1}^{n}\left|X_j^iX_{j'}^i\right| \xrightarrow{\text{p}} \mathbb{E}\left[X_jX_{j'}\right]$$

so that $\frac{1}{n}\sum_{i=1}^{n}\left|X_j^iX_{j'}^i\right| = O_P(1)$. To work with the max term, we first prove the following Lemma.

Lemma 3.2. *Let $\mathbf{Z}_1,\mathbf{Z}_2,\ldots,\mathbf{Z}_n$ be iid random vectors with $\mathbb{E}[|\mathbf{Z}_i|^r] < \infty$. Then,*

$$\max_{1\le i\le n}|\mathbf{Z}_i| = o_P\left(n^{\frac{1}{r}}\right).$$

That is,

$$n^{-\frac{1}{r}}\max_{1\le i\le n}|\mathbf{Z}_i| \xrightarrow{\text{p}} 0.$$

Proof. Writing out the condition for convergence in probability:[18]

$$\begin{aligned}\mathbb{P}\left(n^{-\frac{1}{r}}\max_{1\leq i\leq n}|\mathbf{Z}_i|>\varepsilon\right)&=\mathbb{P}\left(\max_{1\leq i\leq n}|\mathbf{Z}_i|^r>\varepsilon^r n\right)\\&\leq\mathbb{P}\left(\bigcup_{i=1}^n\{|\mathbf{Z}_i|^r>\varepsilon^r n\}\right)\\&\leq\sum_{i=1}^n\mathbb{P}\left(|\mathbf{Z}_i|^r>\varepsilon^r n\right),\end{aligned}$$

where the last inequality uses the Boole's inequality. Now, at this stage, we might be tempted to use the Markov's Inequality directly—but this will not help since we would end up with a constant:

$$\sum_{i=1}^n\mathbb{P}\left(|\mathbf{Z}_i|^r>\varepsilon^r n\right)\leq\sum_{i=1}^n\frac{\mathbb{E}\left[|\mathbf{Z}_i|^r\right]}{\varepsilon^r n}=\frac{\mathbb{E}\left[|\mathbf{Z}_i|^r\right]}{\varepsilon^r}.$$

So, we use another trick. Notice that

$$\mathbb{P}\left(|\mathbf{Z}_i|^r>\varepsilon^r n\right)=\mathbb{P}\left(|\mathbf{Z}_i|^r\mathbf{1}_{\{|\mathbf{Z}_i|^r>\varepsilon^r n\}}>\varepsilon^r n\right).$$

Using the Markov Inequality on this expression, we get that

$$\begin{aligned}\sum_{i=1}^n\mathbb{P}\left(|\mathbf{Z}_i|^r>\varepsilon^r n\right)&\leq\sum_{i=1}^n\frac{\mathbb{E}\left[|\mathbf{Z}_i|^r\mathbf{1}_{\{|\mathbf{Z}_i|^r>\varepsilon^r n\}}\right]}{\varepsilon^r n}\\&=\frac{\mathbb{E}\left[|\mathbf{Z}_i|^r\mathbf{1}_{\{|\mathbf{Z}_i|^r>\varepsilon^r n\}}\right]}{\varepsilon^r}.\end{aligned}$$

Then, as $n\to\infty$, the indicator function becomes always zero so that

$$\mathbb{P}\left(n^{-\frac{1}{r}}\max_{1\leq i\leq n}|\mathbf{Z}_i|>\varepsilon\right)\to 0$$

as $n\to\infty$.[19] ■

Recall (3.16), which can be rearranged to

$$\left(\hat{u}^i\right)^2-\left(u^i\right)^2=-2u^i\mathbf{X}^{i\prime}\left(\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}\right)+\left[\mathbf{X}^{i\prime}\left(\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}\right)\right]^2.$$

Using the Triangle Inequality:

$$\begin{aligned}\left|\left(\hat{u}^i\right)^2-\left(u^i\right)^2\right|&=\left|-2u^i\mathbf{X}^{i\prime}\left(\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}\right)+\left[\mathbf{X}^{i\prime}\left(\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}\right)\right]^2\right|\\&\leq\left|2u^i\mathbf{X}^{i\prime}\left(\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}\right)\right|+\left|\left[\mathbf{X}^{i\prime}\left(\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}\right)\right]^2\right|\\&=2\left|u^i\mathbf{X}^{i\prime}\right|\left|\left(\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}\right)\right|+\left|\left[\mathbf{X}^{i\prime}\left(\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}\right)\right]^2\right|,\end{aligned}$$

where we note that $|\cdot|$ is the Euclidean norm. Using Cauchy-Schwartz inequality,

$$\left|\left(\hat{u}^i\right)^2-\left(u^i\right)^2\right|\leq 2\left|u^i\mathbf{X}^{i\prime}\right|\left|\left(\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}\right)\right|+\left|\mathbf{X}^i\right|^2\left|\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}\right|^2.$$

[18] To see the second inequality, consider the case when $n=2$. Then, $\mathbb{P}\left(\max\left\{|Z_1|^r,|Z_2|^r\right\}>\varepsilon^r n\right)$. So it's clear that this is the probability that at least one of the two $|Z_i|^r$ must be larger than $\varepsilon^r n$. This is the same as the probability that one (or both) individually are larger than $\varepsilon^r n$; i.e.

$$\mathbb{P}\left(\max\left\{|Z_1|^r,|Z_2|^r\right\}>\varepsilon^r n\right)=\mathbb{P}\left(\left\{|Z_1|^r>\varepsilon^r n\right\}\cup\left\{|Z_2|^r>\varepsilon^r n\right\}\right).$$

[19] Strictly speaking, we are using the Dominance Convergence Theorem based on the assumption that $\mathbb{E}\left[|Z_i|^r\right]<\infty$.

Taking max of both sides:

$$\begin{aligned}\max_{1\leq i\leq n}\left|\left(\hat{u}^i\right)^2-\left(u^i\right)^2\right| &\leq 2\max_{1\leq i\leq n}\left|u^i\mathbf{X}^{i\prime}\right|\left|\left(\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}\right)\right|+\max_{1\leq i\leq n}\left|\mathbf{X}^i\right|^2\left|\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}\right|^2\\ &=2\left|\left(\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}\right)\right|\max_{1\leq i\leq n}\left|u^i\mathbf{X}^{i\prime}\right|+\left|\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}\right|^2\max_{1\leq i\leq n}\left|\mathbf{X}^i\right|^2\end{aligned}$$

Now we are ready to use the Lemma. Write

$$\left|\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}\right|^2\max_{1\leq i\leq n}\left|\mathbf{X}^i\right|^2=n\left|\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}\right|^2\max_{1\leq i\leq n}\frac{\left|\mathbf{X}^i\right|^2}{n}.$$

Let $\mathbf{Z}_i=\left|\mathbf{X}_i\right|^2$ with $r=1$. We need to check that $\mathbb{E}\left[\left|\mathbf{X}^i\right|^2\right]$ is finite. Recall from (3.17) that $\left|\mathbf{X}^i\right|^2=\sum_{j=0}^k\left(X_j^i\right)^2$ so that it is a finite sum of the diagonals of $\mathbf{X}\mathbf{X}'$. By assumption $\mathbb{E}\left[\mathbf{X}\mathbf{X}'\right]<\infty$, then we conclude that $\mathbb{E}\left[\left|\mathbf{X}^i\right|^2\right]$ is the sum of the diagonals of $\mathbb{E}\left[\mathbf{X}\mathbf{X}'\right]$ so that it is finite. Thus, we can use the Lemma and obtain that

$$\max_{1\leq i\leq n}\frac{\left|\mathbf{X}^i\right|^2}{n}=o_P\left(1\right).$$

Note that since $\sqrt{n}\left(\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}\right)\xrightarrow{\mathrm{d}}N\left(0,\Omega\right)$. By the Continuous Mapping Theorem, letting $g\left(\mathbf{x}\right)=\left|\mathbf{x}\right|^2$, which is a continuous function, we know that

$$n\left|\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}\right|^2\xrightarrow{\mathrm{d}}g\left[N\left(0,\Omega\right)\right].$$

That is, it converges to some distribution so that

$$n\left|\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}\right|^2=O_P\left(1\right).$$

Together, we have

$$n\left|\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}\right|^2\max_{1\leq i\leq n}\frac{\left|\mathbf{X}^i\right|^2}{n}=O_P\left(1\right)o_P\left(1\right)=o_P\left(1\right).$$

Similarly,

$$\left|\left(\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}\right)\right|\max_{1\leq i\leq n}\left|u^i\mathbf{X}^{i\prime}\right|=\sqrt{n}\left|\left(\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}\right)\right|\max_{1\leq i\leq n}\frac{\left|u^i\mathbf{X}^{i\prime}\right|}{\sqrt{n}}.$$

We use the Lemma again: let $\mathbf{Z}_i=\left|u^i\mathbf{X}^{i\prime}\right|$ with $r=2$. Note that

$$\mathbb{E}\left[\left|u^i\mathbf{X}^{i\prime}\right|^2\right]=\mathbb{E}\left[\left|\begin{pmatrix}u^iX_0^i & u^iX_1^i & \cdots & u^iX_k^i\end{pmatrix}\right|^{1/2}\right]=\sum_{j=1}^k\mathbb{E}\left[\left(u^iX_j^i\right)^2\right].$$

Since $\mathrm{Var}\left[\mathbf{X}u\right]$ exists by assumption, we know that $\mathbb{E}\left[\mathbf{X}\mathbf{X}'u^2\right]<\infty$. Since $\mathbb{E}\left[\left|u^i\mathbf{X}^{i\prime}\right|^2\right]$ is a finite sum of the diagonals $\mathbb{E}\left[\mathbf{X}\mathbf{X}'u^2\right]$, it is finite. Hence, we can appeal to the Lemma to gives us that

$$\max_{1\leq i\leq n}\frac{\left|u^i\mathbf{X}^{i\prime}\right|}{\sqrt{n}}=o_P\left(1\right).$$

Since $\sqrt{n}\left(\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}\right)$ converges in distribution to a normal distribution,

$$\sqrt{n}\left|\left(\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}\right)\right|=O_P\left(1\right).$$

Hence,

$$2\left|\left(\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}\right)\right|\max_{1\leq i\leq n}\left|u^i\mathbf{X}^{i\prime}\right|=O_P\left(1\right)o_P\left(1\right)=o_P\left(1\right).$$

We therefore obtain that

$$\max_{1\leq i\leq n}\left|\left(\hat{u}^i\right)^2-\left(u^i\right)^2\right|=o_P\left(1\right).$$

This implies that, from (3.21),

$$\max_{1\leq i\leq n}\left|\left(\hat{u}^i\right)^2-\left(u^i\right)^2\right|\frac{1}{n}\sum_{i=1}^{n}\left|X_j^iX_{j'}^i\right|=o_P\left(1\right)O_P\left(1\right)=o_P\left(1\right),$$

which, in turn, implies that

$$\left|\frac{1}{n}\sum_{i=1}^{n}X_j^iX_{j'}^i\left(\left(\hat{u}^i\right)^2-\left(u^i\right)^2\right)\right|=o_P\left(1\right).$$

We therefore obtain using (3.20) that

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}^i\mathbf{X}^{i\prime}\left(\hat{u}^i\right)^2\xrightarrow{\text{p}}\mathbb{E}\left[\mathbf{X}^i\mathbf{X}^{i\prime}\left(u^i\right)^2\right].$$

Finally, noting that convergence in marginal probabilities implies convergence in joint probabilities:

$$\left(\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}^i\mathbf{X}^{i\prime}\right]^{-1},\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}^i\mathbf{X}^{i\prime}\left(\hat{u}^i\right)^2\right)\xrightarrow{\text{p}}\left(\mathbb{E}\left[\mathbf{X}^i\mathbf{X}^{i\prime}\right]^{-1},\mathbb{E}\left[\mathbf{X}^i\mathbf{X}^{i\prime}\left(u^i\right)^2\right]\right),$$

while setting $g\left(a,b\right)=aba$, which is a continuous function, and applying the Continuous Mapping Theorem yields that

$$\begin{aligned}\hat{\Omega}_n&=\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}^i\mathbf{X}^{i\prime}\right]^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}^i\mathbf{X}^{i\prime}\left(\hat{u}^i\right)^2\right)\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}^i\mathbf{X}^{i\prime}\right]^{-1}\\&\xrightarrow{\text{p}}\mathbb{E}\left[\mathbf{X}^i\mathbf{X}^{i\prime}\right]^{-1}\mathbb{E}\left[\mathbf{X}^i\mathbf{X}^{i\prime}\left(u^i\right)^2\right]\mathbb{E}\left[\mathbf{X}^i\mathbf{X}^{i\prime}\right]^{-1}\end{aligned}$$

as we wanted.

3.2.10 Inference

Recall our assumptions: $\mathbb{E}\left[\mathbf{X}u\right]=\mathbf{0}$, $X_0=1$; $\mathbb{E}\left[\mathbf{X}\mathbf{X}'\right]$ exists; no perfect collinearity in $\mathbf{X}$; $\text{Var}\left[\mathbf{X}u\right]$ exists (i.e. same assumption as when we showed $\hat{\Omega}_n\xrightarrow{\text{p}}\Omega$). Additionally, assume that $\text{Var}\left[\mathbf{X}u\right]$ is invertible, which implies that Ω is invertible.

Testing a single linear restriction Suppose we wish to test, at level α,

$$\begin{aligned}H_0&:\mathbf{r}'\boldsymbol{\beta}=c,\\H_1&:\mathbf{r}'\boldsymbol{\beta}\neq c,\end{aligned}$$

where $\mathbf{r}=\left(k+1\right)\times 1$ and c is a scalar. For example, if $\mathbf{r}=\left(1,0,\ldots\right)'$, then the test is $\beta_0=c$. Alternatively, if $\mathbf{r}=\left(1,-1,0,0,\ldots\right)'$, then the test is $\beta_0-\beta_1=c$.

Recall that

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}\right)\xrightarrow{\text{d}}N\left(0,\Omega\right).$$

Since linear operations are continuous, by the Continuous Mapping Theorem,

$$\sqrt{n}\left(\mathbf{r}'\hat{\boldsymbol{\beta}}_n-\mathbf{r}'\boldsymbol{\beta}\right)\xrightarrow{\text{d}}N\left(0,\mathbf{r}'\Omega\mathbf{r}\right).$$

Note that, if $\mathbf{r} \neq \mathbf{0}$, then $\mathbf{r}'\Omega\mathbf{r} > 0$ (i.e. Ω is positive definite) so that, by the Continuous Mapping Theorem,

$$\left(\mathbf{r}'\hat{\Omega}_n\mathbf{r}\right)^{-1} \xrightarrow{\text{p}} \left(\mathbf{r}'\Omega\mathbf{r}\right)^{-1}.$$

Then, by Slutsky's Lemma,

$$\frac{\sqrt{n}\left(\mathbf{r}'\hat{\boldsymbol{\beta}}_n - \mathbf{r}'\boldsymbol{\beta}\right)}{\sqrt{\mathbf{r}'\hat{\Omega}_n\mathbf{r}}} \xrightarrow{\text{d}} N(0,1).$$

Thus, we can consider the following test statistic (under the null)

$$T_n = \frac{\sqrt{n}\left(\mathbf{r}'\hat{\boldsymbol{\beta}}_n - c\right)}{\sqrt{\mathbf{r}'\hat{\Omega}_n\mathbf{r}}}$$

and the test is

$$\phi_n = \mathbf{1}_{\left\{|T_n| > z_{1-\frac{\alpha}{2}}\right\}}.$$

By construction, this test is consistent in level α.

Testing multiple linear restrictions (Wald Test) Suppose we wish to test, at level α,

$$H_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{c},$$
$$H_1 : \mathbf{R}\boldsymbol{\beta} \neq \mathbf{c},$$

where $\mathbf{R} = p \times (k+1)$ and $c = p \times 1$; i.e. there are p linear restrictions. For example, this allows us to test

$$\begin{pmatrix} \beta_0 \\ \beta_2 + \beta_3 \\ \beta_1 - \beta_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix},$$

by setting

$$\mathbf{R} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 1 & 1 & 0 & \cdots \\ 0 & 1 & -1 & 0 & 0 & \cdots \end{pmatrix}_{3\times(k+1)}.$$

Just as we required $\mathbf{r} \neq \mathbf{0}$, here, to rule out redundant restrictions (e.g. $2\beta_1 + 2\beta_2 = 2c$ and $\beta_1 + \beta_2 = c$), we require the rows of $\mathbf{R}$ to be linearly independent. This means that, given that Ω is invertible by assumption, $\mathbf{R}\Omega\mathbf{R}'$ is also invertible.[20] By the Continuous Mapping Theorem (linear operations are continuous),

$$\sqrt{n}\left(\mathbf{R}\hat{\boldsymbol{\beta}}_n - \mathbf{R}\boldsymbol{\beta}\right) \xrightarrow{\text{d}} N\left(0, \mathbf{R}\Omega\mathbf{R}'\right).$$

By the Continuous Mapping Theorem, we also have that

$$\left(\mathbf{R}\hat{\Omega}_n\mathbf{R}'\right)^{-1} \xrightarrow{\text{p}} \left(\mathbf{R}\Omega\mathbf{R}'\right)^{-1}.$$

[20] To see this suppose

$$\Omega = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

and that this is invertible.

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} a & b \\ c & d \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}.$$

However, if

$$\begin{pmatrix} 2 & 2 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} a & b \\ c & d \end{pmatrix}\begin{pmatrix} 2 & 1 \\ 2 & 1 \end{pmatrix} = \begin{pmatrix} 2a+2c & 2b+2d \\ a+c & b+d \end{pmatrix}\begin{pmatrix} 2 & 1 \\ 2 & 1 \end{pmatrix}$$
$$= \begin{pmatrix} 4a+4c+4b+4d & 2a+2c+2b+2d \\ 2a+2c+2b+2d & a+c+b+d \end{pmatrix}.$$

Notice that the first row and the second row are linearly dependent so that the matrix is not invertible.

Recall that, if

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\right) \xrightarrow{\mathrm{d}} z \sim N\left(0, \Omega\right),$$

then $z'\Omega^{-1}z \sim \chi^2_k$. Using this, we can obtain that

$$n\left(\mathbf{R}\hat{\boldsymbol{\beta}}_n - \mathbf{R}\boldsymbol{\beta}\right)\left(\mathbf{R}\hat{\Omega}_n\mathbf{R}'\right)^{-1}\left(\mathbf{R}\hat{\boldsymbol{\beta}}_n - \mathbf{R}\boldsymbol{\beta}\right)' \xrightarrow{\mathrm{d}} \chi^2_p.$$

The test statistic is then given by

$$T_n = n\left(\mathbf{R}\hat{\boldsymbol{\beta}}_n - \mathbf{c}\right)\left(\mathbf{R}\hat{\Omega}_n\mathbf{R}'\right)^{-1}\left(\mathbf{R}\hat{\boldsymbol{\beta}}_n - \mathbf{c}\right)'$$

and

$$\phi_n = \mathbf{1}_{\{T_n > c_{p,1-\alpha}\}},$$

where $c_{p,1-\alpha}$ is the $1-\alpha$th quantile of χ^2_p distribution. By construction, this is consistent in level α.

Remark 3.4. *(Construction confidence regions).* We can use the expression above to construct C_n such that

$$\mathbb{P}\left(\beta \in c_n\right) \to 1 - \alpha.$$

This is given by

$$\begin{aligned} C_n &= \left\{c \in \mathbb{R}^{k+1} : H_0 : \beta = c \text{ is not rejected}\right\} \\ &= \left\{c \in \mathbb{R}^{k+1} : n\left(\mathbf{R}\hat{\boldsymbol{\beta}}_n - c\right)\left(\mathbf{R}\hat{\Omega}_n\mathbf{R}'\right)^{-1}\left(\mathbf{R}\hat{\boldsymbol{\beta}}_n - c\right)' \leq c_{p,1-\alpha}\right\}. \end{aligned}$$

Tests of non-linear restrictions Suppose we wish to test, at level α,

$$\begin{aligned} H_0 &: f\left(\boldsymbol{\beta}\right) = \mathbf{c}, \\ H_1 &: f\left(\boldsymbol{\beta}\right) \neq \mathbf{c}, \end{aligned}$$

where $f : \mathbb{R}^{k+1} \to \mathbb{R}^p$ is continuously differentiable at $\boldsymbol{\beta}$. Denote as $D_{\boldsymbol{\beta}} f\left(\boldsymbol{\beta}\right)$ the $p \times (k+1)$ matrix of partials of f, evaluated at $\boldsymbol{\beta}$ with linearly independent rows. Then, by the Delta method,

$$\sqrt{n}\left(f\left(\hat{\boldsymbol{\beta}}_n\right) - f\left(\boldsymbol{\beta}\right)\right) \xrightarrow{\mathrm{d}} N\left(0, D_{\boldsymbol{\beta}} f\left(\boldsymbol{\beta}\right) \Omega D_{\boldsymbol{\beta}} f\left(\boldsymbol{\beta}\right)'\right),$$

where we note that $D_{\boldsymbol{\beta}} f\left(\boldsymbol{\beta}\right) \Omega D_{\boldsymbol{\beta}} f\left(\boldsymbol{\beta}\right)'$ is non-singular (for the same reason as before). By the Continuous Mapping Theorem, then

$$D_{\boldsymbol{\beta}} f\left(\hat{\boldsymbol{\beta}}_n\right) \hat{\Omega}_n D_{\boldsymbol{\beta}} f\left(\hat{\boldsymbol{\beta}}_n\right)' \xrightarrow{\mathrm{p}} D_{\boldsymbol{\beta}} f\left(\boldsymbol{\beta}\right) \Omega D_{\boldsymbol{\beta}} f\left(\boldsymbol{\beta}\right)'.$$

Then we have that

$$n\left(f\left(\hat{\boldsymbol{\beta}}_n\right) - f\left(\boldsymbol{\beta}\right)\right)\left(D_{\boldsymbol{\beta}} f\left(\hat{\boldsymbol{\beta}}_n\right) \hat{\Omega}_n D_{\boldsymbol{\beta}} f\left(\hat{\boldsymbol{\beta}}_n\right)'\right)^{-1}\left(f\left(\hat{\boldsymbol{\beta}}_n\right) - f\left(\boldsymbol{\beta}\right)\right) \xrightarrow{\mathrm{d}} \chi^2_p.$$

The test statistic is then given by

$$T_n = n\left(f\left(\hat{\boldsymbol{\beta}}_n\right) - \mathbf{c}\right)\left(D_{\boldsymbol{\beta}} f\left(\hat{\boldsymbol{\beta}}_n\right) \hat{\Omega}_n D_{\boldsymbol{\beta}} f\left(\hat{\boldsymbol{\beta}}_n\right)'\right)^{-1}\left(f\left(\hat{\boldsymbol{\beta}}_n\right) - \mathbf{c}\right)'$$

and

$$\phi_n = \mathbf{1}_{\{T_n > c_{p,1-\alpha}\}},$$

where $c_{p,1-\alpha}$ is the $1-\alpha$th quantile of χ^2_p distribution. By construction, this is consistent in level α.

Remark 3.5. *(When $p = 1$).* If $p = 1$, then we would have

$$\frac{\sqrt{n}\left|f\left(\hat{\boldsymbol{\beta}}_n\right) - f\left(\boldsymbol{\beta}\right)\right|}{\sqrt{D_{\boldsymbol{\beta}} f\left(\hat{\boldsymbol{\beta}}_n\right) \hat{\Omega}_n D_{\boldsymbol{\beta}} f\left(\hat{\boldsymbol{\beta}}_n\right)'}} \xrightarrow{\mathrm{d}} N\left(0,1\right).$$

The test statistic is then given by

$$T_n = \frac{\sqrt{n}\left|f\left(\hat{\boldsymbol{\beta}}_n\right) - f\left(\mathbf{c}\right)\right|}{\sqrt{D_{\boldsymbol{\beta}} f\left(\hat{\boldsymbol{\beta}}_n\right) \hat{\Omega}_n D_{\boldsymbol{\beta}} f\left(\hat{\boldsymbol{\beta}}_n\right)'}}.$$

and the test is

$$\phi_n = \mathbf{1}_{\left\{T_n > z_{1-\frac{\alpha}{2}}\right\}},$$

where z is the $1 - \alpha/2$th quantile of standard normal distribution.

3.3 Linear Regression when $\mathbb{E}\left[\mathbf{X}u\right] \neq \mathbf{0}$

Suppose that we have $(Y, \mathbf{X}, u)$ where $Y, u \in \mathbb{R}$ and $\mathbf{X} \in \mathbb{R}^{k+1}$, with $X_0 = 1$, and $\boldsymbol{\beta}$ such that

$$Y = \mathbf{X}'\boldsymbol{\beta} + u.$$

Note that we can always normalise β_0 such that $\mathbb{E}\left[u\right] = 0$.

Definition 3.3. *(Exogenous and endogenous variables).*

- ▷ Any X_j with $\mathbb{E}\left[X_j u\right] = 0$ is said to be *exogenous*.
- ▷ Any X_j with $\mathbb{E}\left[X_j u\right] \neq 0$ is said to be *endogenous*.

Since we do not assume that $\mathbb{E}\left[\mathbf{X}u\right] = \mathbf{0}$, we are using the causal model interpretation of linear regressions.

3.3.1 Motivating examples for when $\mathbb{E}\left[\mathbf{X}u\right] \neq \mathbf{0}$

The examples below show that even if the "true" model is such that $\mathbb{E}\left[\mathbf{X}u\right] = \mathbf{0}$, we can still have endogeneity problems in the estimated model if: (i) the estimated model omits relevant variables; (ii) there are measurement errors; and (iii) there are simultaneity problems.

Omitted Variables Suppose $k = 2$ and that the true model is given by

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u.$$

Assume that the model is causal but we still have that $\mathbb{E}\left[\mathbf{X}u\right] = 0$. However, suppose we cannot observe X_2 so that we estimate the model as

$$Y = \beta_0^* + \beta_1^* X_1 + u^*.$$

Using the true model, we can rewrite the estimated model as:

$$Y = \underbrace{\left(\beta_0 + \mathbb{E}\left[X_2\right]\beta_2\right)}_{=\beta_0^*} + \underbrace{\beta_1}_{=\beta_1^*} X_1 + \left(\underbrace{u + \left(X_2 - \mathbb{E}\left[X_2\right]\right)\beta_2}_{=u^*}\right).$$

Then, note that

$$\mathbb{E}\left[u^*\right] = \mathbb{E}\left[u\right] + \mathbb{E}\left[X_2 - \mathbb{E}\left[X_2\right]\right]\beta_2 = 0,$$

however,

$$\begin{aligned}\text{Cov}\left[X_1, u^*\right] &= \text{Cov}\left[X_1, u + \left(X_2 - \mathbb{E}\left[X_2\right]\right)\beta_2\right] \\ &= \text{Cov}\left[X_1, X_2\right]\beta_2.\end{aligned}$$

Thus, $\text{Cov}\left[X_1, u^*\right] \neq 0$ if $\text{Cov}\left[X_1, X_2\right] \neq 0$ and/or $\beta_2 \neq 0$. Hence, we realise that $\mathbb{E}\left[X_1 u^*\right] \neq 0$ in general so that X_1 is an endogenous variable in the estimated model.

Measurement error Suppose we partition $\mathbf{X}$ into $X_0 = 1$ and $\mathbf{X}_1 \in \mathbb{R}^k$, and partition $\boldsymbol{\beta}$ correspondingly. The true model is given by

$$Y = \beta_0 + \mathbf{X}_1'\boldsymbol{\beta}_1 + u.$$

We suppose that this is the causal model and that $\mathbb{E}\left[\mathbf{X}u\right] = \mathbf{0}$. However, assume that $\mathbf{X}_1$ is unobserved and that we instead observe

$$\hat{\mathbf{X}}_1 = \mathbf{X}_1 + \mathbf{v},$$

where $\mathbb{E}\left[\mathbf{v}\right] = \mathbf{0}$, $\text{Cov}\left[u, \mathbf{v}\right] = \mathbf{0}$ and $\text{Cov}\left[\mathbf{X}_1, \mathbf{v}\right] = \mathbf{0}$. The model we therefore estimate is

$$Y = \beta_0^* + \hat{\mathbf{X}}_1'\boldsymbol{\beta}_1^* + u^*.$$

Using the true model, we can rewrite the estimated model as

$$Y = \beta_0 + \hat{\mathbf{X}}_1'\boldsymbol{\beta}_1 + \underbrace{\left(u - \mathbf{v}'\boldsymbol{\beta}_1\right)}_{=u^*}.$$

Then, note that

$$\mathbb{E}\left[u^*\right] = \mathbb{E}\left[u\right] - \mathbb{E}\left[\mathbf{v}'\right]\boldsymbol{\beta}_1 = 0,$$

however,

$$\begin{aligned}\text{Cov}\left[\hat{\mathbf{X}}_1, u^*\right] &= \text{Cov}\left[\mathbf{X}_1 + \mathbf{v}, u - \mathbf{v}'\boldsymbol{\beta}_1\right] \\ &= -\text{Var}\left[\mathbf{v}\right]\boldsymbol{\beta}_1\end{aligned}$$

Thus, unless $\text{Var}\left[\mathbf{v}\right] = \mathbf{0}$ (i.e. measurement error is a constant) or $\boldsymbol{\beta}_1 = \mathbf{0}$ (i.e. $\mathbf{X}_1$ are unrelated to Y), then $\mathbb{E}\left[\hat{\mathbf{X}}_1 u^*\right] \neq \mathbf{0}$; i.e. $\hat{\mathbf{X}}_1$ is endogenous.

Simultaneity Let superscript d denote demand-side variables and superscript s denote supply-side variables. Consider the following demand and supply equations:

$$\begin{aligned}Q^d &= \beta_0^d + \beta_1^d\tilde{P} + u^d, \\ Q^s &= \beta_0^s + \beta_1^s\tilde{P} + u^s\end{aligned}$$

with $\mathbb{E}\left[u^d\right] = \mathbb{E}\left[u^s\right] = 0$ and that $\mathbb{E}\left[u^d u^s\right] = 0$. What we observe in the data is the equilibrium outcome determined by the market clearing condition, $Q^d = Q^s$; i.e.

$$\begin{aligned}\beta_0^d + \beta_1^d\tilde{P} + u^d &= \beta_0^s + \beta_1^s\tilde{P} + u^s \\ \Rightarrow \tilde{P} &= \frac{\left(\beta_0^s - \beta_0^d\right) + \left(u^s - u^d\right)}{\beta_1^d - \beta_1^s}.\end{aligned}$$

Thus, since $\tilde{P}$ contains u^s and u^d, $\mathbb{E}\left[\tilde{P}u^d\right], \mathbb{E}\left[\tilde{P}u^s\right] \neq 0$; i.e. $\tilde{P}$ is endogenous in both equations.

3.3.2 What happens to the OLS estimator if $\mathbb{E}[\mathbf{X}u] \neq \mathbf{0}$?

In general, when $\mathbb{E}[\mathbf{X}u] \neq 0$, the OLS estimator is both biased and inconsistent.

First, note that when $\mathbb{E}[\mathbf{X}u] \neq \mathbf{0}$, then $\mathbb{E}[u|\mathbf{X}] \neq \mathbf{0}$.[21] Then,

$$\mathbb{E}\left[\hat{\boldsymbol{\beta}}_n|\mathbf{X}^1, \mathbf{X}^2, \ldots, \mathbf{X}^n\right] = \boldsymbol{\beta} + (\mathbb{X}\mathbb{X}')^{-1}\mathbb{X}'\underbrace{\mathbb{E}\left[\mathbb{U}|\mathbf{X}^1, \mathbf{X}^2, \ldots, \mathbf{X}^n\right]}_{\neq \mathbf{0}} \neq \boldsymbol{\beta}$$

so that

$$\mathbb{E}\left[\mathbb{E}\left[\hat{\boldsymbol{\beta}}_n|\mathbf{X}^1, \mathbf{X}^2, \ldots, \mathbf{X}^n\right]\right] = \boldsymbol{\beta} + \mathbb{E}\left[(\mathbb{X}\mathbb{X}')^{-1}\mathbb{X}'\mathbb{E}\left[\mathbb{U}|\mathbf{X}^1, \mathbf{X}^2, \ldots, \mathbf{X}^n\right]\right] \neq \boldsymbol{\beta}.$$

That is, the OLS estimator is biased.

Similarly, when $\mathbb{E}[\mathbf{X}u] \neq \mathbf{0}$, notice that

$$\begin{aligned}\mathbb{E}[\mathbf{X}\mathbf{X}']^{-1}\mathbb{E}[\mathbf{X}Y] &= \mathbb{E}[\mathbf{X}\mathbf{X}']^{-1}\mathbb{E}[\mathbf{X}(\mathbf{X}'\boldsymbol{\beta} + u)] \\ &= \boldsymbol{\beta} + \underbrace{\mathbb{E}[\mathbf{X}\mathbf{X}']^{-1}\mathbb{E}[\mathbf{X}u]}_{\neq \mathbf{0}}.\end{aligned}$$

Hence, we although we would still have that $\hat{\boldsymbol{\beta}}_n \xrightarrow{\text{p}} \mathbb{E}[\mathbf{X}\mathbf{X}']^{-1}\mathbb{E}[\mathbf{X}Y]$, we no longer have that $\hat{\boldsymbol{\beta}}_n \xrightarrow{\text{p}} \boldsymbol{\beta}$. That is, the OLS estimator is inconsistent.

3.3.3 Estimating $\boldsymbol{\beta}$

We maintain the same assumptions as before, however, suppose further that there exists *instruments*, $\mathbf{Z} \in \mathbb{R}^{l+1}$, such that

- ▷ *(instrument exogeneity).* $\mathbb{E}[\mathbf{Z}u] = \mathbf{0}$.
- ▷ *(instrument relevance/rank condition).* $\mathbb{E}[\mathbf{Z}\mathbf{X}']$ has rank $k+1$.
- ▷ *(order condition).* $l + 1 \geq k + 1$ (a necessary condition for rank condition).
- ▷ $\mathbf{Z}$ includes all exogenous $\mathbf{X}_j$ (in particular, $Z_0 = X_0 = 1$).
- ▷ No perfect collinearity in $\mathbf{Z}$.
- ▷ $\mathbb{E}[\mathbf{Z}\mathbf{Z}']$ and $\mathbb{E}[\mathbf{Z}\mathbf{X}']$ exist.

Now we solve for $\boldsymbol{\beta}$. First, note that

$$\begin{aligned}\mathbb{E}[\mathbf{Z}u] = \mathbf{0} &\Rightarrow \mathbb{E}[\mathbf{Z}(\mathbf{Y} - \mathbf{X}'\boldsymbol{\beta})] = \mathbf{0} \\ &\Rightarrow \mathbb{E}[\mathbf{Z}\mathbf{Y}] = \mathbb{E}[\mathbf{Z}\mathbf{X}']\boldsymbol{\beta}. \end{aligned} \tag{3.22}$$

When $\boldsymbol{\beta}$ is exactly identified ($l + 1 = k + 1$) Suppose first that $l + 1 = k + 1$, then $\mathbb{E}[\mathbf{Z}\mathbf{X}']$ is an invertible square matrix so that, from (3.22), we can immediately obtain that

$$\boldsymbol{\beta} = \mathbb{E}[\mathbf{Z}\mathbf{X}']^{-1}\mathbb{E}[\mathbf{Z}\mathbf{Y}]. \tag{3.23}$$

[21] To show that $\mathbb{E}[\mathbf{X}u] \neq 0 \Rightarrow \mathbb{E}[u|\mathbf{X}] \neq 0$, we can show the contrapositive; i.e. $\mathbb{E}[u|\mathbf{X}] = 0 \Rightarrow \mathbb{E}[\mathbf{X}u] = 0$. Then,

$$\mathbf{X}\mathbb{E}[u|\mathbf{X}] = 0 \Leftrightarrow \mathbb{E}[\mathbf{X}u|\mathbf{X}] = 0 \Rightarrow \mathbb{E}[\mathbb{E}[\mathbf{X}u|\mathbf{X}]] = \mathbb{E}[\mathbf{X}u] = 0.$$

When $\boldsymbol{\beta}$ is over-identified ($l+1>k+1$) Now suppose that $l+1>k+1$, then $\mathbb{E}[\mathbf{ZX}']$ is not invertible. To proceed, we first prove the following Lemma.

Proposition 3.7. *(Rank Inequality). For any comformable matrices A and B,*

$$\operatorname{rank}(AB) \leq \min\{\operatorname{rank}(A), \operatorname{rank}(B)\}.$$

Lemma 3.3. *Suppose there is no perfect collinearity in $\mathbf{Z}$, then*

$$\operatorname{rank}(\mathbb{E}[\mathbf{ZX}']) = \operatorname{rank}(\boldsymbol{\Pi}),$$

where $\boldsymbol{\Pi}$ is such that $\operatorname{BLP}(\mathbf{X}|\mathbf{Z}) = \boldsymbol{\Pi}'\mathbf{Z}$. In particular, if $\operatorname{rank}[\mathbb{E}[\mathbf{ZX}']] = k+1$, then

$$\boldsymbol{\Pi}'\mathbb{E}[\mathbf{ZX}'] = \boldsymbol{\Pi}'\mathbb{E}\left[\mathbf{Z}\left(\boldsymbol{\Pi}'\mathbf{Z}\right)'\right] = \boldsymbol{\Pi}'\mathbb{E}[\mathbf{ZZ}']\boldsymbol{\Pi}$$

is invertible.

Proof. From the fact that $\operatorname{BLP}(\mathbf{X}|\mathbf{Z}) = \boldsymbol{\Pi}'\mathbf{Z}$,

$$\mathbf{X} = \boldsymbol{\Pi}'\mathbf{Z} + \mathbf{v},$$

with $\mathbb{E}[\mathbf{Zv}'] = \mathbf{0}$ (we use the fact that there is no collinearity for the existence of $\boldsymbol{\Pi}$ so that $\mathbb{E}[\mathbf{ZZ}']$ is invertible). Hence,

$$\mathbb{E}[\mathbf{ZX}'] = \mathbb{E}\left[\mathbf{Z}\left(\boldsymbol{\Pi}'\mathbf{Z}+\mathbf{v}\right)'\right] = \mathbb{E}[\mathbf{ZZ}'\boldsymbol{\Pi}] + \mathbb{E}[\mathbf{Zv}'] = \mathbb{E}[\mathbf{ZZ}'\boldsymbol{\Pi}].$$

Using Rank Inequality,

$$\begin{aligned}
\operatorname{rank}(\mathbb{E}[\mathbf{ZZ}']\boldsymbol{\Pi}) &\leq \min\{\operatorname{rank}(\mathbb{E}[\mathbf{ZZ}']), \operatorname{rank}(\boldsymbol{\Pi})\} \\
&\leq \operatorname{rank}(\boldsymbol{\Pi}) \\
&= \operatorname{rank}\left(\mathbb{E}[\mathbf{ZZ}']\mathbb{E}[\mathbf{ZZ}']^{-1}\boldsymbol{\Pi}\right) \\
&\leq \min\{\operatorname{rank}(\mathbb{E}[\mathbf{ZZ}']), \operatorname{rank}(\boldsymbol{\Pi})\} \\
&\leq \operatorname{rank}\left(\mathbb{E}[\mathbf{ZZ}']^{-1}\boldsymbol{\Pi}\right)
\end{aligned}$$

so that

$$\operatorname{rank}(\mathbb{E}[\mathbf{ZZ}']\boldsymbol{\Pi}) = \operatorname{rank}(\mathbb{E}[\mathbf{ZX}']) = \operatorname{rank}(\boldsymbol{\Pi}).$$

Finally, notice that $\boldsymbol{\Pi}'\mathbb{E}[\mathbf{ZX}']$ is a $(k+1)\times(k+1)$ square matrix with rank $k+1$; hence it is invertible. ■

Multiplying both sides of (3.22) by $\boldsymbol{\Pi}'$,

$$\begin{aligned}
\boldsymbol{\Pi}'\mathbb{E}[\mathbf{ZY}] &= \boldsymbol{\Pi}'\mathbb{E}[\mathbf{ZX}']\boldsymbol{\beta} \\
&= \boldsymbol{\Pi}'\mathbb{E}[\mathbf{ZZ}']\boldsymbol{\Pi}\boldsymbol{\beta}.
\end{aligned}$$

Since we showed that $\boldsymbol{\Pi}'\mathbb{E}[\mathbf{ZZ}']\boldsymbol{\Pi}$ is invertible, then

$$\boldsymbol{\beta} = \left(\boldsymbol{\Pi}'\mathbb{E}[\mathbf{ZZ}']\boldsymbol{\Pi}\right)^{-1}\boldsymbol{\Pi}'\mathbb{E}[\mathbf{ZY}]. \tag{3.24}$$

Note that $\mathbb{E}[\mathbf{ZX}'] = \mathbb{E}\left[\mathbf{Z}\left(\boldsymbol{\Pi}'\mathbf{Z}\right)'\right] = \mathbb{E}[\mathbf{ZZ}']\boldsymbol{\Pi}$. Since $\mathbb{E}[\mathbf{ZZ}']$ is invertible by assumption, for $\mathbb{E}[\mathbf{ZX}']$ to have rank $k+1$, we need $\boldsymbol{\Pi}$ to have full rank. When $\boldsymbol{\beta}$ is exactly identified (i.e. $l+1=k+1$) and there is single endogenous regressor, say X_k (i.e. $Z_j = X_j$ for all $j<k$), then

$$\mathbf{X} = \begin{pmatrix} X_0 \\ X_1 \\ \vdots \\ X_{k-1} \\ X_k \end{pmatrix}, \ \mathbf{Z} = \begin{pmatrix} Z_0 \\ Z_1 \\ \vdots \\ Z_{k-1} \\ Z_k \end{pmatrix} = \begin{pmatrix} X_0 \\ X_1 \\ \vdots \\ X_{k-1} \\ Z_k \end{pmatrix}.$$

Consider $\mathbf{\Pi}'\mathbf{Z} = \text{BLP}(\mathbf{X}|\mathbf{Z}) \Rightarrow \mathbf{X} = \mathbf{\Pi}'\mathbf{Z} + \mathbf{v}$:

$$\begin{pmatrix} X_0 \\ X_1 \\ \vdots \\ X_{k-1} \\ X_k \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \cdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \\ \Pi_{k,1} & \Pi_{k,2} & \cdots & \Pi_{k,k-1} & \Pi_{k,k} \end{pmatrix}}_{(k\times 1)+(l\times 1)} \begin{pmatrix} X_0 \\ X_1 \\ \vdots \\ X_{k-1} \\ Z_k \end{pmatrix} + \mathbf{v}.$$

Thus, in this case, instrument relevance is equivalent to requiring that $\Pi_{k,k} \neq 0$ since if $\Pi_{k,k} = 0$, then $\text{rank}(\mathbf{\Pi}) = k < k+1$. Therefore, we can interpret the instrument relevance condition as ensuring that the instruments are correlated with the endogenous variables (here X_k and Z_k must be correlated).

3.3.4 Solving for subvectors of $\boldsymbol{\beta}$

Splitting out the constant term Suppose $l+1 = k+1$, and we partition $\mathbf{X}$ to $X_0 = 1$ and $\mathbf{X}_1 \in \mathbb{R}^k$, $\boldsymbol{\beta}$ into β_0 and $\boldsymbol{\beta}_1$ correspondingly, and we also partition $\mathbf{Z}$ into $Z_0 = 1$ and $\mathbf{Z}_1 \in \mathbb{R}^{l=k}$. Assume that

$$Y = \beta_0 + \mathbf{X}_1'\boldsymbol{\beta}_1 + u.$$

Note that $\mathbb{E}[u] = 0$ implies that

$$\begin{aligned} 0 =& \mathbb{E}[u] = \mathbb{E}[Y - \beta_0 - \mathbf{X}_1'\boldsymbol{\beta}_1] \\ \Rightarrow& \mathbb{E}[Y] = \beta_0 + \mathbb{E}[\mathbf{X}_1']\boldsymbol{\beta}_1 \end{aligned}$$

Subtracting $\mathbb{E}[Y]$ from Y gives

$$Y - \mathbb{E}[Y] = (\mathbf{X}_1' - \mathbb{E}[\mathbf{X}_1'])\boldsymbol{\beta}_1 + u.$$

Multiplying both sides by $\mathbf{Z}_1$ and taking expectation yields

$$\begin{aligned} \mathbb{E}[\mathbf{Z}_1(Y - \mathbb{E}[Y])] &= \mathbb{E}[\mathbf{Z}_1(\mathbf{X}_1' - \mathbb{E}[\mathbf{X}_1'])\boldsymbol{\beta}_1] + \underbrace{\mathbb{E}[\mathbf{Z}_1 u]}_{=\mathbf{0}} \\ &= \mathbb{E}\left[\mathbf{Z}_1(\mathbf{X}_1 - \mathbb{E}[\mathbf{X}_1])'\right]\boldsymbol{\beta}_1 \\ \Rightarrow \mathbb{E}[(\mathbf{Z}_1 - \mathbb{E}[\mathbf{Z}_1])(Y - \mathbb{E}[Y])] &= \mathbb{E}\left[(\mathbf{Z}_1 - \mathbb{E}[\mathbf{Z}_1])(\mathbf{X}_1 - \mathbb{E}[\mathbf{X}_1])'\right]\boldsymbol{\beta}_1 \\ \Rightarrow \boldsymbol{\beta}_1 &= \text{Cov}[\mathbf{Z}_1, \mathbf{X}_1]^{-1}\text{Cov}[\mathbf{Z}_1, Y]. \end{aligned} \tag{3.25}$$

General case Suppose we partition $\mathbf{X}$ into $\mathbf{X}_1$ and $\mathbf{X}_2$ where $\mathbf{X}_1$ is endogenous and $\mathbf{X}_2$ is exogenous, and analogously partition $\boldsymbol{\beta}$ and $\mathbf{Z}$ (i.e. $\mathbf{Z}_2 = \mathbf{X}_2$), then

$$Y = \mathbf{X}_1'\boldsymbol{\beta}_1 + \mathbf{X}_2'\boldsymbol{\beta}_2 + u.$$

Instead of subtracting $\mathbb{E}[Y]$, we now want to subtract $\text{BLP}(Y|\mathbf{Z}_2)$ from Y. Consider first $\text{BLP}(Y|\mathbf{Z}_2)$:[22]

$$\begin{aligned}\text{BLP}(Y|\mathbf{Z}_2) &= \text{BLP}(\mathbf{X}_1'\boldsymbol{\beta}_1 + \mathbf{X}_2'\boldsymbol{\beta}_2 + u|\mathbf{Z}_2) \\ &= \text{BLP}(\mathbf{X}_1'\boldsymbol{\beta}_1|\mathbf{Z}_2) + \underbrace{\text{BLP}(\mathbf{X}_2'\boldsymbol{\beta}_2|\mathbf{Z}_2)}_{=\mathbf{X}_2'\boldsymbol{\beta}_2 \because \mathbf{Z}_2=\mathbf{X}_2} + \underbrace{\text{BLP}(u|\mathbf{Z}_2)}_{=0 \cdot \because \mathbb{E}[\mathbf{Z}u]=\mathbf{0}} \\ &= \text{BLP}(\mathbf{X}_1|\mathbf{Z}_2)'\boldsymbol{\beta}_1 + \mathbf{X}_2'\boldsymbol{\beta}_2.\end{aligned}$$

Then,

$$\begin{aligned}Y - \text{BLP}(Y|\mathbf{Z}_2) &= (\mathbf{X}_1'\boldsymbol{\beta}_1 + \mathbf{X}_2'\boldsymbol{\beta}_2 + u) - \left(\text{BLP}(\mathbf{X}_1|\mathbf{Z}_2)'\boldsymbol{\beta}_1 + \mathbf{X}_2'\boldsymbol{\beta}_2\right) \\ \Rightarrow Y^* &= \mathbf{X}_1^{*\prime}\boldsymbol{\beta}_1 + u, \qquad (3.26)\end{aligned}$$

where $Y^* \coloneqq Y - \text{BLP}(Y|\mathbf{Z}_2)$ and $\mathbf{X}_1^* \coloneqq X_1 - \text{BLP}(\mathbf{X}_1|\mathbf{Z}_2)$, which can be understood as deviations from the best linear predictor. Notice that Y^* is the "residual" from "regressing" Y on $\mathbf{Z}_2$, and $\mathbf{X}_1^*$ is the "residual" from "regressing" $\mathbf{X}_1$ on $\mathbf{Z}_2$.

In the exactly identified case (i.e. $l+1 = k+1$), we multiply (3.26) through by $\mathbf{Z}_1$ and take expectations to obtain that

$$\begin{aligned}\mathbb{E}[\mathbf{Z}_1 Y^*] &= \mathbb{E}[\mathbf{Z}_1\mathbf{X}_1^{*\prime}]\boldsymbol{\beta}_1 + \mathbb{E}[\mathbf{Z}_1 u] = \mathbb{E}[\mathbf{Z}_1\mathbf{X}_1^{*\prime}]\boldsymbol{\beta}_1 \\ \Rightarrow \boldsymbol{\beta}_1 &= \mathbb{E}[\mathbf{Z}_1\mathbf{X}_1^{*\prime}]^{-1}\mathbb{E}[\mathbf{Z}_1 Y^*] \\ &= \text{Cov}[\mathbf{Z}_1, \mathbf{X}_1^*]^{-1}\text{Cov}[\mathbf{Z}_1, Y^*].\end{aligned}$$

In the over-identified case (i.e. $l+1 > k+1$), we multiply (3.26) by $\hat{\mathbf{X}}_1^* = \text{BLP}(\mathbf{X}_1^*|\mathbf{Z}_1)$:

$$\begin{aligned}\boldsymbol{\beta}_1 &= \mathbb{E}\left[\hat{\mathbf{X}}_1^*\mathbf{X}_1^{*\prime}\right]^{-1}\mathbb{E}\left[\hat{\mathbf{X}}_1^* Y^*\right] \\ &= \mathbb{E}\left[\hat{\mathbf{X}}_1^*\left(\hat{\mathbf{X}}_1^* + \mathbf{v}\right)'\right]^{-1}\mathbb{E}\left[\hat{\mathbf{X}}_1^* Y^*\right] \\ &= \mathbb{E}\left[\hat{\mathbf{X}}_1^*\hat{\mathbf{X}}_1^{*\prime}\right]^{-1}\mathbb{E}\left[\hat{\mathbf{X}}_1^* Y^*\right],\end{aligned}$$

since, by construction, $\mathbb{E}\left[\hat{\mathbf{X}}_1^*\mathbf{v}\right] = \mathbf{0}$.

[22] To see this note that $\text{BLP}(Y|\mathbf{Z}_2) = \mathbf{Z}_2'\boldsymbol{\gamma}$ solves $\min_{\boldsymbol{\gamma}} \mathbb{E}\left[\left(Y - \mathbf{Z}_2'\boldsymbol{\gamma}\right)^2\right]$.The first-order condition gives that

$$\begin{aligned}&\mathbb{E}[\mathbf{Z}_2(Y - \mathbf{Z}_2'\boldsymbol{\gamma})] = 0 \\ \Rightarrow\ &\mathbb{E}[\mathbf{Z}_2(\mathbf{X}_1'\boldsymbol{\beta}_1 + \mathbf{X}_2'\boldsymbol{\beta}_2 + u)] = \mathbb{E}[\mathbf{Z}_2\mathbf{Z}_2']\boldsymbol{\gamma} \\ \Rightarrow\ &\mathbb{E}[\mathbf{Z}_2\mathbf{X}_1']\boldsymbol{\beta}_1 + \mathbb{E}[\mathbf{Z}_2\mathbf{X}_2']\boldsymbol{\beta}_2 + \mathbb{E}[\mathbf{Z}_2 u] = \mathbb{E}[\mathbf{Z}_2\mathbf{Z}_2']\boldsymbol{\gamma}.\end{aligned}$$

Now, consider $\text{BLP}\left(\mathbf{X}_1'\boldsymbol{\beta}_1|\mathbf{Z}_2\right) = \mathbf{Z}_2'\boldsymbol{\theta}$ which solves $\min_{\boldsymbol{\theta}} \mathbb{E}\left[\left(\mathbf{X}_1'\boldsymbol{\beta}_1 - \mathbf{Z}_2'\boldsymbol{\theta}\right)^2\right]$. The first-order condition is

$$\mathbb{E}[\mathbf{Z}_2\mathbf{X}_1']\boldsymbol{\beta}_1 = \mathbb{E}[\mathbf{Z}_2\mathbf{Z}_2']\boldsymbol{\theta}.$$

Similarly, $\text{BLP}\left(\mathbf{X}_2'\boldsymbol{\beta}_2|\mathbf{Z}_2\right) = \mathbf{Z}_2'\boldsymbol{\tau}$ implies

$$\mathbb{E}[\mathbf{Z}_2\mathbf{X}_2']\boldsymbol{\beta}_2 = \mathbb{E}[\mathbf{Z}_2\mathbf{Z}_2']\boldsymbol{\tau},$$

and $\text{BLP}(u|\mathbf{Z}_2) = \mathbf{Z}_2\boldsymbol{\phi}$ implies

$$\mathbb{E}[\mathbf{Z}_2 u] = \mathbb{E}[\mathbf{Z}_2\mathbf{Z}_2']\boldsymbol{\phi},$$

We can therefore write

$$\mathbb{E}[\mathbf{Z}_2\mathbf{Z}_2']\boldsymbol{\theta} + \mathbb{E}[\mathbf{Z}_2\mathbf{Z}_2']\boldsymbol{\tau} + \mathbb{E}[\mathbf{Z}_2\mathbf{Z}_2']\boldsymbol{\phi} = \mathbb{E}[\mathbf{Z}_2\mathbf{Z}_2']\boldsymbol{\gamma};$$

i.e.

$$\boldsymbol{\gamma} = \boldsymbol{\theta} + \boldsymbol{\tau} + \boldsymbol{\phi}.$$

We therefore have that

$$\text{BLP}(Y|\mathbf{Z}_2) = \text{BLP}\left(\mathbf{X}_1'\boldsymbol{\beta}_1|\mathbf{Z}_2\right) + \text{BLP}\left(\mathbf{X}_2'\boldsymbol{\beta}_2|\mathbf{Z}_2\right) + \text{BLP}(u|\mathbf{Z}_2).$$

Remark 3.6. We can summarise the steps in the following way:

(i) "regress" Y on $\mathbf{Z}_2$ to obtain the "residuals" Y^*;

(ii) "regress" $\mathbf{X}_1$ on $\mathbf{Z}_2$ to obtain the "residuals" $\mathbf{X}_1^*$;

(iii) Then, we have two cases:

(a) exactly identified case: "regress" $\mathbf{Z}_1 Y^*$ on $\mathbf{Z}_1 \mathbf{X}_1^*$ and then coefficient on $\mathbf{Z}_1 \mathbf{X}_1^*$ is equal to $\boldsymbol{\beta}_1$;

(b) over-identified case: first "predict" $\mathbf{X}_1^*$ using $\mathbf{Z}_1$ to obtain $\hat{\mathbf{X}}_1^*$. Then "regress" $\hat{\mathbf{X}}_1^* Y^*$ on $\hat{\mathbf{X}}_1^* \mathbf{X}_1^*$ and the coefficient on $\hat{\mathbf{X}}_1^* \mathbf{X}_1^*$ is equal to $\boldsymbol{\beta}_1$.

3.3.5 Motivating examples revisited

Omitted variables Recall that the we assumed that the true model is given by

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$$

with $\mathbb{E}[\mathbf{X}u] = \mathbf{0}$. However, we assumed that we cannot observe X_2 so that we estimate the model as:

$$Y = \beta_0^* + \beta_1^* X_1 + u^*,$$

where

$$\begin{aligned} \beta_0^* &= \beta_0 + \mathbb{E}[X_2]\,\beta_2, \\ \beta_1^* &= \beta_1, \\ u^* &= u + (X_2 - \mathbb{E}[X_2])\,\beta_2. \end{aligned}$$

We then found that $\mathbb{E}[X_1 u^*] \neq 0$ in general so that X_1 is an endogenous variable in the estimated model.

Suppose we now have $\mathbf{Z} = (1, Z_1)'$ such that $\mathbf{Z}$ satisfies

▷ instrument exogeneity: i.e. $\mathbb{E}[\mathbf{Z}u^*] = \mathbf{0}$. Then, we have $\mathbb{E}[u^*] = 0$ and $\mathbb{E}[Z_1 u^*] = 0$ so that $\mathrm{Cov}[Z_1, u^*] = 0$. In other words, $\mathbb{E}[Z_1 u] = 0$ and $\mathrm{Cov}[Z_1, X_2] = 0$.

▷ instrument relevance: i.e. $\mathbb{E}[\mathbf{Z}\mathbf{X}']$ has rank $k+1$. This is equivalent to requiring that $\Pi_1 \neq 0$ in $X_1 = \Pi_0 + \Pi_1 Z_1 = \mathrm{BLP}(X_1 | 1, Z_1)$. Since $\Pi_1 = \mathrm{Cov}[X_1, Z_1] / \mathrm{Var}[Z_1]$, $\Pi_1 \neq 0$ also means that $\mathrm{Cov}[X_1, Z_1] \neq 0$.

3.3.6 Measurement error

Recall that we partitioned $\mathbf{X}$ into $X_0 = 1$ and $\mathbf{X}_1 \in \mathbb{R}^k$, and $\boldsymbol{\beta}$ correspondingly. The true model is given by

$$Y = \beta_0 + \mathbf{X}_1' \boldsymbol{\beta}_1 + u$$

with $\mathbb{E}[\mathbf{X}u] = \mathbf{0}$. However, we supposed that $\mathbf{X}_1$ is unobserved and that we instead observe

$$\hat{\mathbf{X}}_1 = \mathbf{X}_1 + \mathbf{v}$$

where $\mathbb{E}[\mathbf{v}] = \mathbf{0}$, $\mathrm{Cov}[u, \mathbf{v}] = \mathbf{0}$ and $\mathrm{Cov}[\mathbf{X}_1, \mathbf{v}] = \mathbf{0}$. The model we therefore estimate is

$$Y = \beta_0^* + \hat{\mathbf{X}}_1' \boldsymbol{\beta}_1^* + u^*,$$

where

$$\begin{aligned} \beta_0^* &= \beta_0, \\ \boldsymbol{\beta}_1^* &= \boldsymbol{\beta}_1, \\ u^* &= u - \mathbf{v}' \boldsymbol{\beta}_1. \end{aligned}$$

We then found that $\mathbb{E}\left[\hat{\mathbf{X}}_1 u^*\right] \neq \mathbf{0}$ in general so that $\hat{\mathbf{X}}_1$ is an endogenous variable in the estimated model.

Suppose we now have $\mathbf{Z}_1$, which is another measurement of $\mathbf{X}_1$ so that

$$\mathbf{Z}_1 = \mathbf{X}_1 + \mathbf{w},$$

where $\mathbb{E}[\mathbf{w}] = \mathbf{0}$, $\text{Cov}[u, \mathbf{w}] = \mathbf{0}$ and $\text{Cov}[\mathbf{X}_1, w] = \mathbf{0}$. We also assume that $\mathbf{w}$ brings some new information beyond $\hat{\mathbf{X}}_1$ so that $\text{Cov}[\mathbf{v}, \mathbf{w}] = \mathbf{0}$. Then,

▷ $\mathbf{Z}_1$ satisfies strictly exogeneity—i.e. $\mathbb{E}[\mathbf{Z}_1 u^*] = \mathbf{0}$—by assumption since

$$\begin{aligned}
\mathbb{E}[\mathbf{Z}_1 u^*] &= \mathbb{E}[\mathbf{Z}_1 (u - \mathbf{v}'\boldsymbol{\beta}_1)] = \mathbb{E}[\mathbf{Z}_1 u] - \mathbb{E}[\mathbf{Z}_1 \mathbf{v}']\boldsymbol{\beta}_1 \\
&= \mathbb{E}[(\mathbf{X}_1 + \mathbf{w}) u] - \mathbb{E}[(\mathbf{X}_1 + \mathbf{w}) \mathbf{v}']\boldsymbol{\beta}_1 \\
&= \underbrace{\mathbb{E}[\mathbf{X}_1 u]}_{=\mathbf{0}} + \underbrace{\mathbb{E}[\mathbf{w}u]}_{=\mathbf{0}\because \text{Cov}[\mathbf{w},u]=\mathbf{0}} - \left(\underbrace{\mathbb{E}[\mathbf{X}_1 \mathbf{v}']}_{=\mathbf{0}} + \underbrace{\mathbb{E}[\mathbf{w}\mathbf{v}']}_{=\mathbf{0}\because \text{Cov}[\mathbf{v},\mathbf{w}]=\mathbf{0}} \right) \boldsymbol{\beta}_1 \\
&= \mathbf{0}.
\end{aligned}$$

▷ $\mathbf{Z}_1$ satisfies instrument relevance since

$$\begin{aligned}
\mathbb{E}\left[\mathbf{Z}\hat{\mathbf{X}}'\right] &= \mathbb{E}\left[\begin{pmatrix} 1 \\ \mathbf{Z}_1 \end{pmatrix} \begin{pmatrix} 1 & \hat{\mathbf{X}}_1' \end{pmatrix} \right] = \mathbb{E}\begin{bmatrix} 1 & \mathbb{E}\left[\hat{\mathbf{X}}_1'\right] \\ \mathbb{E}[\mathbf{Z}_1] & \mathbb{E}\left[\mathbf{Z}_1 \hat{\mathbf{X}}_1'\right] \end{bmatrix} \\
&= \mathbb{E}\begin{bmatrix} 1 & \mathbb{E}\left[\hat{\mathbf{X}}_1'\right] \\ \mathbb{E}[\mathbf{Z}_1] & \mathbb{E}\left[(\mathbf{X}_1 + \mathbf{w})(\mathbf{X}_1 + \mathbf{v})'\right] \end{bmatrix} = \mathbb{E}\begin{bmatrix} 1 & \mathbb{E}\left[\hat{\mathbf{X}}_1'\right] \\ \mathbb{E}[\mathbf{Z}_1] & \mathbb{E}[\mathbf{X}_1 \mathbf{X}_1'] \end{bmatrix} \\
&= \mathbb{E}[\mathbf{X}\mathbf{X}'].
\end{aligned}$$

And $\mathbb{E}[\mathbf{X}\mathbf{X}']$ has full rank (since $\mathbf{X}$ has no perfect collinearity by assumption).

Simultaneity Recall that we have the following demand and supply equations:

$$\begin{aligned}
Q^d &= \beta_0^d + \beta_1^d \tilde{P} + u^d, \\
Q^s &= \beta_0^s + \beta_1^s \tilde{P} + u^s
\end{aligned}$$

with $\mathbb{E}\left[u^d\right] = \mathbb{E}[u^s] = 0$ and that $\mathbb{E}\left[u^d u^s\right] = 0$, which implied that

$$\tilde{P} = \frac{\left(\beta_0^s - \beta_0^d\right) + \left(u^s - u^d\right)}{\beta_1^d - \beta_1^s}$$

so that $\mathbb{E}\left[\tilde{P} u^d\right], \mathbb{E}\left[\tilde{P} u^s\right] \neq 0$; i.e. $\tilde{P}$ is endogenous in both equations. Suppose we now can observe Z_1 that affects only the supply: i.e.

$$\begin{aligned}
Q^d &= \beta_0^d + \beta_1^d \tilde{P} + u^d, \\
Q^s &= \beta_0^s + \beta_1^s \tilde{P} + \beta_2^s Z_1 + u^s
\end{aligned}$$

where $\mathbb{E}\left[Z_1 u^d\right] = \mathbb{E}[Z_1 u^s] = 0$. Then,

$$P = \frac{\left(\beta_0^s - \beta_0^d\right) + \beta_2^s Z_1 + \left(u^s - u^d\right)}{\beta_1^d - \beta_1^s}.$$

Consider $Q^d = \beta_0^d + \beta_1^d P + u^d$ using Z_1 as an instrument, then notice that:

▷ Z_1 satisfies instrument exogeneity since $\mathbb{E}\left[Z_1 u^d\right] = 0$ by assumption;

▷ Z_1 satisfies instrument relevance if

$$\begin{aligned}\operatorname{Cov}[P, Z_1] &= \operatorname{Cov}\left[\frac{\left(\beta_0^s - \beta_0^d\right) + \beta_2^s Z_1 + \left(u^s - u^d\right)}{\beta_1^d - \beta_1^s}, Z_1\right] \\ &= \frac{\beta_2^s}{\beta_1^d - \beta_1^s} \operatorname{Var}[Z_1] \neq 0\end{aligned}$$

if $\beta_2^s \neq 0$.

3.3.7 Estimating $\boldsymbol{\beta}$

Suppose we have $(Y, \mathbf{X}, \mathbf{Z}, u)$ such that $\mathbf{X} \in \mathbb{R}^{k+1}$ with $X_0 = 1$ and $\mathbf{Z} \in \mathbb{R}^{l+1}$ with $Z_0 = 1$, and that $\mathbf{Z}_1$ includes any exogenous X_j's. Suppose also that $\mathbb{E}[\mathbf{Z}\mathbf{Z}']$ and $\mathbb{E}[\mathbf{Z}\mathbf{X}']$ exist, and that there is no perfect collinearity in $\mathbf{Z}$. We also have $\mathbb{E}[\mathbf{Z}u] = 0$ and $\operatorname{rank}(\mathbb{E}[\mathbf{Z}\mathbf{X}']) = k+1$. Finally, we suppose that that sample is iid:

$$\left(Y^1, \mathbf{X}^1, \mathbf{Z}^1\right), \left(Y^2, \mathbf{X}^2, \mathbf{Z}^2\right), \ldots, \left(Y^n, \mathbf{X}^n, \mathbf{Z}^n\right) \overset{\text{iid}}{\sim} (Y, \mathbf{X}, \mathbf{Z}).$$

We consider two cases:

▷ *instrument variables (IV) estimator:* in the exactly identified case $l + 1 = k + 1$;

▷ *two-stage least squares (TSLS) estimator*: in the over-identified case $l + 1 > k + 1$.

IV estimator Recall that, when $l + 1 = k + 1$, the population estimator is given by (3.23):

$$\boldsymbol{\beta} = \mathbb{E}[\mathbf{Z}\mathbf{X}']^{-1} \mathbb{E}[\mathbf{Z}Y].$$

So the natural estimator is

$$\hat{\boldsymbol{\beta}}_n = \left(\frac{1}{n}\sum_{i=1}^n \mathbf{Z}^i \mathbf{X}^{i\prime}\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^n \mathbf{Z}^i Y^i\right). \tag{3.27}$$

Note that this implies:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \mathbf{Z}^i \left(\mathbf{Y}^i - \mathbf{X}^{i\prime}\hat{\boldsymbol{\beta}}_n\right) &= 0 \\ \Rightarrow \frac{1}{n}\sum_{i=1}^n \mathbf{Z}^i \hat{u}^i &= 0,\end{aligned}$$

where $\hat{u}^i = \mathbf{Y}^i - \mathbf{X}^{i\prime}\hat{\boldsymbol{\beta}}_n$.

The IV estimator can also be written as

$$\hat{\boldsymbol{\beta}}_n = \left(\mathbb{Z}'\mathbb{X}\right)^{-1} \mathbb{Z}'\mathbb{Y},$$

where $\mathbb{Z} = \left(Z^1, Z^2, \ldots, Z^n\right)'$. In this case, defining $\hat{\mathbb{U}} = \left(\hat{u}^1, \hat{u}^2, \ldots, \hat{u}^n\right)' = \mathbb{Y} - \mathbb{X}\hat{\boldsymbol{\beta}}_n$, we realise that

$$\mathbb{Z}\hat{\mathbb{U}} = 0.$$

TSLS estimator Recall (3.24):

$$\begin{aligned}\boldsymbol{\beta} &= \left(\boldsymbol{\Pi}'\mathbb{E}[\mathbf{Z}\mathbf{X}']\right)^{-1} \boldsymbol{\Pi}'\mathbb{E}[\mathbf{Z}Y] \\ &= \left(\boldsymbol{\Pi}'\mathbb{E}[\mathbf{Z}\mathbf{Z}']\,\boldsymbol{\Pi}\right)^{-1} \boldsymbol{\Pi}'\mathbb{E}[\mathbf{Z}Y]\end{aligned}$$

where $\text{BLP}(\mathbf{X}|\mathbf{Z}) = \mathbf{\Pi}'\mathbf{Z}$ so that

$$\mathbf{\Pi} = \mathbb{E}[\mathbf{Z}\mathbf{Z}']^{-1}\mathbb{E}[\mathbf{Z}\mathbf{X}'].$$

So the natural estimator is

$$\begin{aligned}\hat{\boldsymbol{\beta}}_n^1 &= \left(\hat{\mathbf{\Pi}}_n'\left(\frac{1}{n}\sum_{i=1}^n \mathbf{Z}^i\mathbf{X}^{i\prime}\right)\right)^{-1}\hat{\mathbf{\Pi}}_n'\left(\frac{1}{n}\sum_{i=1}^n \mathbf{Z}^i Y^i\right)\\ &= \left(\frac{1}{n}\sum_{i=1}^n \hat{\mathbf{\Pi}}_n'\mathbf{Z}^i\mathbf{X}^{i\prime}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n \hat{\mathbf{\Pi}}_n'\mathbf{Z}^i Y^i\right)\end{aligned}$$

or

$$\hat{\boldsymbol{\beta}}_n^2 = \left(\hat{\mathbf{\Pi}}_n'\left(\frac{1}{n}\sum_{i=1}^n \mathbf{Z}^i\mathbf{Z}^{i\prime}\right)\hat{\mathbf{\Pi}}_n\right)^{-1}\hat{\mathbf{\Pi}}_n'\left(\frac{1}{n}\sum_{i=1}^n \mathbf{Z}^i Y^i\right).$$

where

$$\hat{\mathbf{\Pi}}_n = \left(\frac{1}{n}\sum_{i=1}^n \mathbf{Z}^i\mathbf{Z}^{i\prime}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n \mathbf{Z}^i\mathbf{X}^{i\prime}\right).$$

To see that $\hat{\boldsymbol{\beta}}_n^1 = \hat{\boldsymbol{\beta}}_n^2$, recall that $\mathbf{X}^i = \hat{\mathbf{\Pi}}_n'\mathbf{Z}^i + \hat{\mathbf{v}}^i$ so that

$$\begin{aligned}\hat{\boldsymbol{\beta}}_n^1 &= \left(\hat{\mathbf{\Pi}}_n'\left(\frac{1}{n}\sum_{i=1}^n \mathbf{Z}^i\left(\hat{\mathbf{\Pi}}_n'\mathbf{Z}^i + \hat{\mathbf{v}}^i\right)'\right)\right)^{-1}\hat{\mathbf{\Pi}}_n'\left(\frac{1}{n}\sum_{i=1}^n \mathbf{Z}^i Y^i\right)\\ &= \left(\hat{\mathbf{\Pi}}_n'\left(\frac{1}{n}\sum_{i=1}^n \mathbf{Z}^i\mathbf{Z}^{i\prime}\right)\hat{\mathbf{\Pi}}_n\right)^{-1}\hat{\mathbf{\Pi}}_n'\left(\frac{1}{n}\sum_{i=1}^n \mathbf{Z}^i Y^i\right) = \hat{\boldsymbol{\beta}}_n^2,\end{aligned}$$

where we use the fact that

$$\frac{1}{n}\sum_{i=1}^n \mathbf{Z}^i\hat{\mathbf{v}}^{i\prime} = 0$$

by construction.

Observe that $\hat{\boldsymbol{\beta}}_n$ satisfies

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^n \hat{\mathbf{\Pi}}_n'\mathbf{Z}^i\left(Y^i - \mathbf{X}^{i\prime}\hat{\boldsymbol{\beta}}_n\right) = 0\\ &\qquad\Rightarrow \frac{1}{n}\sum_{i=1}^n \hat{\mathbf{\Pi}}_n'\mathbf{Z}^i\hat{u}^i = 0.\end{aligned}$$

Recall that $\mathbf{Z}^i$ contains all the exogenous X_j^i and that the corresponding column in the matrix $\hat{\mathbf{\Pi}}_n$ has one in the jth row with zero everywhere else. Hence, the condition above requires that $\hat{u}^i$ is orthogonal to any exogenous regressors; however, $\hat{u}^i$ may not be orthogonal to other components of $\mathbf{Z}^i$.

Notice also that:

- ▷ $\hat{\boldsymbol{\beta}}_n^1$ has the interpretation as the IV estimator, (3.27), where we replace $\mathbf{Z}^i$ with $\hat{\mathbf{\Pi}}_n'\mathbf{Z}^i$.
- ▷ $\hat{\boldsymbol{\beta}}_n^2$ has the TSLS interpretation: (i) We first regress $\mathbf{X}^i$ on $\mathbf{Z}^i$ to obtain $\hat{\mathbf{\Pi}}_n'\mathbf{Z}^i$; (ii) We then regress Y^i on $\hat{\mathbf{\Pi}}_n'\mathbf{Z}^i$.
- ▷ if $\hat{\mathbf{\Pi}}_n'$ is invertible (i.e. $l+1 = k+1$), then the IV and TSLS estimators correspond exactly.

We can also write the TSLS estimator using matrices. Define

$$\begin{aligned}\hat{\mathbb{X}} &= \left(\hat{X}^1, \hat{X}^2, \ldots, \hat{X}^n\right) = \mathbb{P}_Z \mathbb{X}, \\ \mathbb{P}_Z &= \mathbb{Z}\left(\mathbb{Z}'\mathbb{Z}\right)^{-1}\mathbb{Z}'.\end{aligned}$$

Then,

$$\begin{aligned}\hat{\boldsymbol{\beta}}_n &= \left(\hat{\mathbb{X}}'\mathbb{X}\right)^{-1}\left(\hat{\mathbb{X}}'\mathbb{Y}\right) \\ &= \left(\mathbb{X}'\mathbb{P}_Z\mathbb{X}\right)^{-1}\left(\mathbb{X}'\mathbb{P}_Z\mathbb{Y}\right).\end{aligned}$$

3.3.8 Estimating subvectors of $\hat{\boldsymbol{\beta}}_n$

Splitting out the constant term Partition $\mathbf{X}$ into $X_0 = 1$ and $\mathbf{X}_1 \in \mathbb{R}^k$, $\boldsymbol{\beta}$ correspondingly, and $\mathbf{Z}$ into $Z_0 = 1$ and $\mathbf{Z}_1 \in \mathbb{R}^l$:

$$Y = \beta_0 + \mathbf{X}_1'\boldsymbol{\beta}_1 + u.$$

The first-order condition gives that

$$\begin{aligned}\mathbf{0} &= \frac{1}{n}\sum_{i=1}^n \mathbf{Z}^i\left(Y^i - \mathbf{X}^{i\prime}\hat{\boldsymbol{\beta}}_n\right) \\ &= \frac{1}{n}\sum_{i=1}^n \begin{pmatrix} 1 \\ Z_1^i \\ \vdots \\ Z_k^i \end{pmatrix}\left(Y^i - \begin{pmatrix} 1 & X_1^i & \cdots & X_k^i \end{pmatrix}\begin{pmatrix} \hat{\beta}_{n,0} \\ \hat{\beta}_{n,1} \\ \vdots \\ \hat{\beta}_{n,k} \end{pmatrix}\right) \\ &= \frac{1}{n}\sum_{i=1}^n \begin{pmatrix} 1 \\ Z_2^i \\ \vdots \\ Z_k^i \end{pmatrix}_{k\times 1}\left(Y^i - \left(\hat{\beta}_{n,0} + \hat{\beta}_{n,1}X_2^i + \cdots + \hat{\beta}_{n,k}X_k^i\right)\right)_{1\times 1}. \qquad (3.28)\end{aligned}$$

Taking the first row,

$$\begin{aligned}0 &= \frac{1}{n}\sum_{i=1}^n\left(Y^i - \left(\hat{\beta}_{n,0} + \hat{\beta}_{n,1}X_1^i + \cdots + \hat{\beta}_{n,k}X_k^i\right)\right) \\ \Rightarrow \hat{\beta}_{n,0} &= \bar{Y}_n - \hat{\beta}_{n,1}\left(\frac{1}{n}\sum_{i=1}^n X_1^i\right) - \cdots - \hat{\beta}_{n,k}\left(\frac{1}{n}\sum_{i=1}^n X_k^i\right) \\ &= \bar{Y}_n - \hat{\beta}_{n,1}\bar{X}_1 - \cdots - \hat{\beta}_{n,k}\bar{X}_k \\ &= \bar{Y}_n - \begin{pmatrix} \bar{X}_1 & \bar{X}_2 & \cdots & \bar{X}_k \end{pmatrix}\begin{pmatrix} \hat{\beta}_{n,1} \\ \hat{\beta}_{n,2} \\ \vdots \\ \hat{\beta}_{n,k} \end{pmatrix} = \bar{Y}_n - \bar{\mathbf{X}}_n'\hat{\boldsymbol{\beta}}_{1,n}.\end{aligned}$$

Taking the last k rows of (3.28),

$$\begin{aligned}
\mathbf{0} &= \frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} Z_1^i \\ Z_2^i \\ \vdots \\ Z_k^i \end{pmatrix}_{k\times 1}\left(Y^i - \left(\hat{\beta}_{n,0} + \hat{\beta}_{n,1}X_2^i + \cdots + \hat{\beta}_{n,k}X_k^i\right)\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbf{Z}_1^i\left(Y^i - \left(\underbrace{\left(\hat{\beta}_{n,0}\right)}_{=\bar{Y}_n - \bar{\mathbf{X}}_n'\hat{\boldsymbol{\beta}}_{1,n}} + \underbrace{\left(\hat{\beta}_{n,1}X_2^i + \cdots + \hat{\beta}_{n,k}X_k^i\right)}_{=\mathbf{X}_1^{i\prime}\hat{\boldsymbol{\beta}}_{1,n}}\right)\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbf{Z}_1^i\left(Y^i - \bar{Y}_n - \left(\left(\mathbf{X}_1^{i\prime} - \bar{\mathbf{X}}_n'\right)\hat{\boldsymbol{\beta}}_{1,n}\right)\right).
\end{aligned}$$

Hence,

$$\hat{\boldsymbol{\beta}}_{1,n} = \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{Z}_1^i\left(\mathbf{X}_1^i - \bar{\mathbf{X}}_n\right)'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{Z}_1^i\left(Y^i - \bar{Y}_n\right)\right).$$

General case We partition $\mathbf{X}$ into $\mathbf{X}_1$ and $\mathbf{X}_2$ where $\mathbf{X}_2$ are exogenous (and partition $\boldsymbol{\beta}$ correspondingly), and partition $\mathbf{Z}$ into $\mathbf{Z}_1$ and $\mathbf{Z}_2$ such that $\mathbf{Z}_2 = \mathbf{X}_2$. Define, for $i \in \{1, 2\}$,

$$\begin{aligned}
\mathbb{X}_i &= \left(\mathbf{X}_i^1, \mathbf{X}_i^2, \ldots, \mathbf{X}_i^n\right)', \\
\mathbb{Z}_i &= \left(\mathbf{Z}_i^1, \mathbf{Z}_i^2, \ldots, \mathbf{Z}_i^n\right)', \\
\mathbb{P}_i &= \mathbb{Z}_i\left(\mathbb{Z}_i'\mathbb{Z}_i\right)^{-1}\mathbb{Z}_i', \\
\mathbb{M}_i &= \mathbb{I} - \mathbb{P}_i.
\end{aligned}$$

We then have

$$\mathbb{Y} = \mathbb{X}_1\hat{\boldsymbol{\beta}}_{1,n} + \mathbb{X}_2\hat{\boldsymbol{\beta}}_{2,n} + \hat{\mathbb{U}}.$$

Multiplying both sides by $\mathbb{M}_2$ allows us to eliminate $\mathbb{X}_2$.

$$\begin{aligned}
\mathbb{M}_2\mathbb{Y} &= \mathbb{M}_2\mathbb{X}_1\hat{\boldsymbol{\beta}}_{1,n} + \mathbb{M}_2\mathbb{X}_2\hat{\boldsymbol{\beta}}_{2,n} + \mathbb{M}_2\hat{\mathbb{U}} \\
&= \mathbb{M}_2\mathbb{X}_1\hat{\boldsymbol{\beta}}_{1,n} + \hat{\mathbb{U}},
\end{aligned} \tag{3.29}$$

where we use the fact that $\mathbb{M}_2\mathbb{X}_2\hat{\boldsymbol{\beta}}_{2,n} = \mathbf{0}$ since $\mathbb{M}_2$ is orthogonal to the space of $\mathbb{X}_2$, and $\mathbb{M}_2\hat{\mathbb{U}} = \hat{\mathbb{U}}$ since $\hat{\mathbb{U}}$ is in the space of $\mathbb{M}_2$.

In the exactly identified case (i.e. $l+1 = k+1$), we can multiply (3.29) through by $\mathbb{Z}'$ to obtain that

$$\begin{aligned}
\mathbb{Z}'\mathbb{M}_2\mathbb{Y} &= \mathbb{Z}'\mathbb{M}_2\mathbb{X}_1\hat{\boldsymbol{\beta}}_{1,n} + \mathbb{Z}'\hat{\mathbb{U}} \\
\Rightarrow \hat{\boldsymbol{\beta}}_{1,n} &= \left(\mathbb{Z}'\mathbb{M}_2\mathbb{X}_1\right)^{-1}\mathbb{Z}'\mathbb{M}_2\mathbb{Y}.
\end{aligned}$$

This is analogous to the population case.

Recall in the over-identified case for the population estimator, we regressed $Y^* = Y - \text{BLP}\left(Y|\mathbf{Z}_2\right)$ on $\mathbf{X}_1^* = X_1 - \text{BLP}\left(\mathbf{X}_1|\mathbf{Z}_2\right)$. Here, Y^* is analogous to $\mathbb{M}_2\mathbb{Y}$ and $\mathbb{M}_2\mathbb{X}_1$ is analogous to $\mathbf{X}_1^*$. We then multiplied through by $\hat{\mathbf{X}}_1^* = \text{BLP}\left(\mathbf{X}_1^*|\mathbf{Z}_1\right)$, which is analogous to $\left(\mathbb{P}_1\mathbb{M}_2\mathbb{X}_1\right)'$. Multiplying through by this gives

$$\left(\mathbb{P}_1\mathbb{M}_2\mathbb{X}_1\right)'\mathbb{M}_2\mathbb{Y} = \left(\mathbb{P}_1\mathbb{M}_2\mathbb{X}_1\right)'\mathbb{M}_2\mathbb{X}_1\hat{\boldsymbol{\beta}}_{1,n} + \left(\mathbb{P}_1\mathbb{M}_2\mathbb{X}_1\right)'\hat{\mathbb{U}}.$$

It can be shown that $\left(\mathbb{P}_1\mathbb{M}_2\mathbb{X}_1\right)'\hat{\mathbb{U}} = 0$ so that

$$\hat{\boldsymbol{\beta}}_{1,n} = \left(\left(\mathbb{P}_1\mathbb{M}_2\mathbb{X}_1\right)'\mathbb{M}_2\mathbb{X}_1\right)^{-1}\left(\mathbb{P}_1\mathbb{M}_2\mathbb{X}_1\right)'\mathbb{M}_2\mathbb{Y}.$$

Once again, this is analogous to the population case.

3.3.9 Properties of the TSLS estimator

Let $(Y, \mathbf{X}, \mathbf{Z}, u)$ where $Y, u \in \mathbb{R}$, $\mathbf{X} \in \mathbb{R}^{k+1}$, $\mathbf{Z} \in \mathbb{R}^{l+1}$, $X_0 = Z_0 = 1$, and Z contains all exogenous X_j's such that

$$Y = \mathbf{X}'\boldsymbol{\beta} + u.$$

Assume that $\mathbb{E}[\mathbf{Z}\mathbf{Z}']$ and $\mathbb{E}[\mathbf{Z}\mathbf{X}]'$ exist, $\mathbb{E}[\mathbf{Z}u] = \mathbf{0}$, $\operatorname{rank}(\mathbb{E}[\mathbf{Z}\mathbf{X}']) = k+1$ and there is no perfect collinearity in $\mathbf{Z}$. We have n iid samples, where

$$\left(Y^1, \mathbf{X}^1, \mathbf{Z}^1\right), \left(Y^2, \mathbf{X}^2, \mathbf{Z}^2\right), \ldots, \left(Y^n, \mathbf{X}^n, \mathbf{Z}^n\right) \overset{\text{iid}}{\sim} (Y, \mathbf{X}, \mathbf{Z}).$$

Recall (3.24):

$$\begin{aligned}\boldsymbol{\beta} &= \mathbb{E}\left[\boldsymbol{\Pi}'\mathbf{Z}\mathbf{X}'\right]^{-1}\mathbb{E}\left[\boldsymbol{\Pi}'\mathbf{Z}Y\right] \\ &= \mathbb{E}\left[\boldsymbol{\Pi}'\mathbf{Z}\mathbf{Z}'\boldsymbol{\Pi}\right]^{-1}\mathbb{E}\left[\boldsymbol{\Pi}'\mathbf{Z}Y\right],\end{aligned}$$

and the TSLS estimator:

$$\begin{aligned}\hat{\boldsymbol{\beta}}_n &= \left(\frac{1}{n}\sum_{i=1}^n \hat{\boldsymbol{\Pi}}_n'\mathbf{Z}^i\mathbf{X}^{i\prime}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n \hat{\boldsymbol{\Pi}}_n'\mathbf{Z}^iY^i\right) \\ &= \left(\frac{1}{n}\sum_{i=1}^n \hat{\boldsymbol{\Pi}}_n'\mathbf{Z}^i\mathbf{Z}^{i\prime}\hat{\boldsymbol{\Pi}}_n\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n \hat{\boldsymbol{\Pi}}_n'\mathbf{Z}^iY^i\right),\end{aligned} \tag{3.30}$$

where

$$\hat{\boldsymbol{\Pi}}_n = \left(\frac{1}{n}\sum_{i=1}^n \mathbf{Z}^i\mathbf{Z}^{i\prime}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n \mathbf{Z}^i\mathbf{X}^{i\prime}\right).$$

Consistency Under the given conditions, $\hat{\boldsymbol{\beta}}_n \xrightarrow{\text{p}} \boldsymbol{\beta}$.

We focus on (3.30). Since $\mathbf{Z}^i$ and $\mathbf{X}^i$ are iid and $\mathbb{E}[\mathbf{Z}\mathbf{Z}']$ and $\mathbb{E}[\mathbf{Z}\mathbf{X}]'$ exist, by WLLN,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \mathbf{Z}^i\mathbf{X}^i &\xrightarrow{\text{p}} \mathbb{E}[\mathbf{Z}\mathbf{X}'], \\ \frac{1}{n}\sum_{i=1}^n \mathbf{Z}^i\mathbf{Z}^{i\prime} &\xrightarrow{\text{p}} \mathbb{E}[\mathbf{Z}\mathbf{Z}'].\end{aligned} \tag{3.31}$$

By the Continuous Mapping Theorem, since there is no perfect collinearity in $\mathbf{Z}$ (i.e. $\mathbb{E}[\mathbf{Z}\mathbf{Z}']$ is invertible),

$$\left(\frac{1}{n}\sum_{i=1}^n \mathbf{Z}^i\mathbf{Z}^{i\prime}\right)^{-1} \xrightarrow{\text{p}} \mathbb{E}[\mathbf{Z}\mathbf{Z}']^{-1}.$$

Since convergence in marginal probabilities implies convergence in joint probabilities, and by the Continuous Mapping Theorem again,

$$\hat{\boldsymbol{\Pi}}_n \xrightarrow{\text{p}} \boldsymbol{\Pi}.$$

Rewrite (3.30) as

$$\hat{\boldsymbol{\beta}}_n = \left(\hat{\boldsymbol{\Pi}}_n'\left(\frac{1}{n}\sum_{i=1}^n \mathbf{Z}^i\mathbf{Z}^{i\prime}\right)\hat{\boldsymbol{\Pi}}_n\right)^{-1}\left(\hat{\boldsymbol{\Pi}}_n'\left(\frac{1}{n}\sum_{i=1}^n \mathbf{Z}^iY^i\right)\right).$$

Since $\mathbb{E}\left[\mathbf{Z}^iY^i\right] = \mathbb{E}[\mathbf{Z}Y] = \mathbb{E}[\mathbf{Z}(\mathbf{X}'\boldsymbol{\beta} + u)] = \mathbb{E}[\mathbf{Z}\mathbf{X}']\boldsymbol{\beta}$ and $\mathbb{E}[\mathbf{Z}\mathbf{X}']$ exists, and $\left(\mathbf{Z}^i, Y^i\right)$'s are iid,

$$\frac{1}{n}\sum_{n=1}^n \mathbf{Z}^iY^i \xrightarrow{\text{p}} \mathbb{E}\left[\mathbf{Z}^iY^i\right].$$

Since convergence in marginal probabilities implies convergence in joint probabilities, and by the Continuous Mapping Theorem,

$$\hat{\mathbf{\Pi}}_n' \left(\frac{1}{n} \sum_{n=1}^{n} \mathbf{Z}^i Y^i \right) \xrightarrow{\text{p}} \mathbf{\Pi}' \mathbb{E} \left[\mathbf{Z}^i Y^i \right] = \mathbb{E} \left[\mathbf{\Pi}' \mathbf{Z}^i Y^i \right].$$

Using the same reasoning and (3.31),

$$\hat{\mathbf{\Pi}}_n' \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{Z}^i \mathbf{Z}^{i\prime} \right) \hat{\mathbf{\Pi}}_n \xrightarrow{\text{p}} \mathbf{\Pi}' \mathbb{E} \left[\mathbf{Z}^i \mathbf{Z}^{i\prime} \right] \mathbf{\Pi}.$$

Since $\text{rank}\left(\mathbb{E}\left[\mathbf{Z}\mathbf{X}'\right]\right) = \text{rank}\left(\mathbf{\Pi}\right) = k+1$ then $\mathbb{E}\left[\mathbf{\Pi}'\mathbf{Z}\mathbf{Z}'\mathbf{\Pi}\right]$ is invertible, so, by the Continuous Mapping Theorem,

$$\hat{\mathbf{\Pi}}_n' \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{Z}^i \mathbf{Z}^{i\prime} \right) \hat{\mathbf{\Pi}}_n \xrightarrow{\text{p}} \mathbb{E} \left[\mathbf{\Pi}' \mathbf{Z}\mathbf{Z}' \mathbf{\Pi} \right].$$

Applying the Continuous Mapping Theorem (together with the fact that convergence in marginal probabilities implies convergence in joint probabilities), we obtain that

$$\hat{\boldsymbol{\beta}}_n \xrightarrow{\text{p}} \boldsymbol{\beta} = \mathbb{E} \left[\mathbf{\Pi}' \mathbf{Z}\mathbf{Z}' \mathbf{\Pi} \right]^{-1} \mathbb{E} \left[\mathbf{\Pi}' \mathbf{Z} Y \right].$$

Limit distribution Assume further that $\text{Var}\left[\mathbf{Z}u\right]$ exist. Then,

$$\sqrt{n} \left(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta} \right) \xrightarrow{\text{d}} N\left(0, \Omega\right),$$

where

$$\Omega = \mathbb{E} \left[\mathbf{\Pi}' \mathbf{Z}\mathbf{Z}' \mathbf{\Pi} \right]^{-1} \mathbf{\Pi}' \, \text{Var}\left[\mathbf{Z}u\right] \mathbf{\Pi} \mathbb{E} \left[\mathbf{\Pi}' \mathbf{Z}\mathbf{Z}' \mathbf{\Pi} \right]^{-1}.$$

To see this, substituting the expression for Y^i gives

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_n &= \left(\frac{1}{n} \sum_{i=1}^{n} \hat{\mathbf{\Pi}}_n' \mathbf{Z}^i \mathbf{Z}^{i\prime} \hat{\mathbf{\Pi}}_n \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} \hat{\mathbf{\Pi}}_n' \mathbf{Z}^i \left(\mathbf{X}^{i\prime} \boldsymbol{\beta} + u^i \right) \right) \\
&= \left(\frac{1}{n} \sum_{i=1}^{n} \hat{\mathbf{\Pi}}_n' \mathbf{Z}^i \mathbf{X}^{i\prime} \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} \hat{\mathbf{\Pi}}_n' \mathbf{Z}^i \mathbf{X}^{i\prime} \right) \boldsymbol{\beta} + \left(\frac{1}{n} \sum_{i=1}^{n} \hat{\mathbf{\Pi}}_n' \mathbf{Z}^i \mathbf{X}^{i\prime} \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} \hat{\mathbf{\Pi}}_n' \mathbf{Z}^i u^i \right) \\
&= \boldsymbol{\beta} + \left(\frac{1}{n} \sum_{i=1}^{n} \hat{\mathbf{\Pi}}_n' \mathbf{Z}^i \mathbf{X}^{i\prime} \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} \hat{\mathbf{\Pi}}_n' \mathbf{Z}^i u^i \right).
\end{aligned}$$

Rearranging and multiplying by $\sqrt{n}$ yields

$$\sqrt{n} \left(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta} \right) = \left(\frac{1}{n} \sum_{i=1}^{n} \hat{\mathbf{\Pi}}_n' \mathbf{Z}^i \mathbf{X}^{i\prime} \right)^{-1} \hat{\mathbf{\Pi}}_n' \left(\sqrt{n} \frac{1}{n} \sum_{i=1}^{n} \mathbf{Z}^i u^i \right).$$

Recall that we already showed that

$$\left(\frac{1}{n} \sum_{i=1}^{n} \hat{\mathbf{\Pi}}_n' \mathbf{Z}^i \mathbf{X}^{i\prime} \right)^{-1} \hat{\mathbf{\Pi}}_n' = \left(\hat{\mathbf{\Pi}}_n' \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{Z}^i \mathbf{Z}^{i\prime} \right) \hat{\mathbf{\Pi}}_n \right)^{-1} \hat{\mathbf{\Pi}}_n' \xrightarrow{\text{p}} \left(\mathbf{\Pi}' \mathbb{E} \left[\mathbf{Z}^i \mathbf{Z}^{i\prime} \right] \mathbf{\Pi} \right)^{-1} \mathbf{\Pi}'. \quad (3.32)$$

Since $\text{Var}\left[\mathbf{Z}u\right]$ and $\mathbf{Z}^i$, $\mathbf{X}^i$, Y^i are iid, by the Central Limit Theorem,

$$\left(\sqrt{n} \frac{1}{n} \sum_{i=1}^{n} \mathbf{Z}^i u^i \right) \xrightarrow{\text{d}} N\left(0, \text{Var}\left[\mathbf{Z}u\right]\right).$$

By Slutsky's Lemma,

$$\left(\hat{\mathbf{\Pi}}_n'\left(\frac{1}{n}\sum_{i=1}^n \mathbf{Z}^i\mathbf{Z}^{i\prime}\right)\hat{\mathbf{\Pi}}_n\right)^{-1}\hat{\mathbf{\Pi}}_n'\left(\sqrt{n}\frac{1}{n}\sum_{i=1}^n \mathbf{Z}^i u^i\right) \xrightarrow{\text{d}} \left(\mathbf{\Pi}'\mathbb{E}\left[\mathbf{Z}^i\mathbf{Z}^{i\prime}\right]\mathbf{\Pi}\right)^{-1}\mathbf{\Pi}' N\left(0, \text{Var}\left[\mathbf{Z}u\right]\right).$$

That is,

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\right) \xrightarrow{\text{d}} N\left(0, \Omega\right),$$

where $\Omega = \mathbb{E}\left[\mathbf{\Pi}'\mathbf{Z}\mathbf{Z}'\mathbf{\Pi}\right]^{-1}\mathbf{\Pi}'\,\text{Var}\left[\mathbf{Z}u\right]\mathbf{\Pi}\mathbb{E}\left[\mathbf{\Pi}'\mathbf{Z}\mathbf{Z}'\mathbf{\Pi}\right]^{-1}$.

Efficiency Recall that we solve for $\boldsymbol{\beta}$ using $\mathbf{\Pi}'\mathbb{E}\left[\mathbf{Z}u\right] = 0$. Alternatively, we could have used any other matrix $\mathbf{\Gamma}$ with some dimension such that $\text{rank}\left(\mathbf{\Gamma}\mathbb{E}\left[\mathbf{Z}\mathbf{X}'\right]\right) = k+1$, and use it to solve for $\boldsymbol{\beta}$:

$$\tilde{\boldsymbol{\beta}}_n = \left(\frac{1}{n}\sum_{i=1}^n \mathbf{\Gamma}'\mathbf{Z}^i\mathbf{X}^{i\prime}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n \mathbf{\Gamma}'\mathbf{Z}^i Y^i\right).$$

As before, we will have that

$$\sqrt{n}\left(\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\right) \xrightarrow{\text{d}} N\left(0, \tilde{\Omega}\right),$$

where $\tilde{\Omega} = \mathbb{E}\left[\mathbf{\Gamma}'\mathbf{Z}\mathbf{X}'\right]^{-1}\mathbf{\Gamma}'\,\text{Var}\left[\mathbf{Z}u\right]\mathbf{\Gamma}\left(\mathbb{E}\left[\mathbf{\Gamma}'\mathbf{Z}\mathbf{X}'\right]^{-1}\right)'$ (the transpose in the last term is there as the term might not be symmetric).

Proposition 3.8. *If* $\mathbb{E}\left[u|\mathbf{Z}\right] = 0$ *and* $\text{Var}\left[u|\mathbf{Z}\right] = \sigma^2$ *, then* $\mathbf{\Gamma} = \mathbf{\Pi}$ *is the "best" in the sense that* $\Omega \leq \tilde{\Omega}$ *for any* $\mathbf{\Gamma}$ *that satisfies* $\text{rank}\left(\mathbf{\Gamma}\mathbb{E}\left[\mathbf{Z}\mathbf{X}'\right]\right) = k+1$.

Proof. [To do: hint is to show the equivalent statement that $\Omega^{-1} \geq \tilde{\Omega}^{-1}$.] ■

Of course, if these conditions are not imposed, then there are other better estimators.

3.3.10 Estimating Ω

The natural estimator here is

$$\hat{\Omega}_n = A\left(\frac{1}{n}\sum_{i=1}^n \mathbf{Z}^i\mathbf{Z}^{i\prime}\left(\hat{u}^i\right)^2\right)A',$$

where $A \coloneqq \left(\hat{\mathbf{\Pi}}_n'\left(\frac{1}{n}\sum_{i=1}^n \mathbf{Z}^i\mathbf{Z}^{i\prime}\right)\hat{\mathbf{\Pi}}_n\right)^{-1}\hat{\mathbf{\Pi}}_n'$. Note that we already showed that A converges in probability (see (3.32)). To show that $\left(\frac{1}{n}\sum \mathbf{Z}^i\mathbf{Z}^{i\prime}\left(\hat{u}^i\right)^2\right) \xrightarrow{\text{p}} \text{Var}\left[\mathbf{Z}u\right]$ is isomorphic when we showed that $\left(\frac{1}{n}\sum \mathbf{X}^i\mathbf{X}^{i\prime}\left(\hat{u}^i\right)^2\right) \xrightarrow{\text{p}} \text{Var}\left[\mathbf{X}u\right]$ in the case of OLS estimator.

Crucially, $\hat{u}^i$ is the predicted residual from the regression of Y^i on $\mathbf{X}^i$—and does not involve $\mathbf{Z}^i$. That is, $\hat{u}^i \neq Y^i - \hat{\mathbf{X}}^{i\prime}\boldsymbol{\beta}$ and so we *do not* use the residual from regressing Y^i on $\hat{\mathbf{X}}^{i\prime}$.

3.3.11 Weak instruments

When the rank condition—i.e. $\text{rank}\left(\mathbb{E}\left[\mathbf{Z}\mathbf{X}'\right]\right) = k+1$—is "close" to being less than $k+1$, then the approximation of distribution of $\sqrt{n}\left(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\right)$ may be poor in finite samples.

Example 3.1. *(Weak instruments).* Let X^i be a random variable and Z^i deterministic. Assume

$$Y^i = \beta X^i + u^i,$$
$$X^i = \pi Z^i + v^i,$$

and

$$\begin{pmatrix} u^i \\ v^i \end{pmatrix}_{i=1,2,\ldots,n} \overset{\text{iid}}{\sim} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_2^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}\right).$$

In this setup, the relevance condition requires that $\pi \neq 0$. The IV estimator in this case is

$$\begin{aligned} \hat{\beta}_n &= \frac{\frac{1}{n}\sum_{i=1}^n Z^i Y^i}{\frac{1}{n}\sum_{i=1}^n Z^i X^i} \\ \Rightarrow \sqrt{n}\left(\hat{\beta}_n - \beta\right) &= \sqrt{n}\frac{\frac{1}{n}\sum_{i=1}^n Z^i u^i}{\frac{1}{n}\sum_{i=1}^n Z^i X^i} \\ &= \frac{\sqrt{n}\frac{1}{n}\sum_{i=1}^n Z^i u^i}{\frac{1}{n}\sum_{i=1}^n \left(Z^i\right)^2 \pi + \frac{1}{n}\sum_{i=1}^n Z^i v^i} \coloneqq \frac{A}{B}. \end{aligned}$$

where we substituted that $X^i = \pi Z^i + v^i$. Since u_i and v_i are jointly normal, we realise that

$$\begin{pmatrix} A \\ B \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ \overline{Z_n^2}\pi \end{pmatrix}, \begin{pmatrix} \overline{Z_n^2}\sigma_1^2 & \frac{1}{\sqrt{n}}\overline{Z_n^2}\sigma_{12} \\ \frac{1}{\sqrt{n}}\overline{Z_n^2}\sigma_{12} & \frac{1}{n}\overline{Z_n^2}\sigma_2^2 \end{pmatrix}\right),$$

where $\overline{Z_n^2} = \frac{1}{n}\sum_{i=1}^n \left(Z^i\right)^2$.[23]

If we use asymptotic theories, since

$$\begin{aligned} \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n Z^i u^i\right) &\xrightarrow{\text{d}} N\left(0, \operatorname{Var}\left[Z^i u^i\right]\right), \\ \frac{1}{n}\sum_{i=1}^n Z^i v^i &\xrightarrow{\text{p}} \mathbb{E}\left[Z^i v^i\right] = Z^i \mathbb{E}\left[v^i\right] = 0, \\ \frac{1}{n}\sum_{i=1}^n \left(Z^i\right)^2 \pi &\xrightarrow{\text{p}} \mathbb{E}\left[\left(Z^i\right)^2\right]\pi \equiv \overline{Z^2}\pi, \end{aligned}$$

by Slutsky's Lemma, we have

$$\begin{aligned} \sqrt{n}\left(\hat{\beta}_n - \beta\right) &\xrightarrow{\text{d}} \frac{1}{\overline{Z^2}\pi} N\left(0, \overline{Z^2}\sigma_1^2\right) \\ &\overset{\text{d}}{=} N\left(0, \frac{\sigma_1^2}{\overline{Z^2}\pi^2}\right), \end{aligned}$$

where $\overline{Z^2}$ is such that $\overline{Z_n^2} \xrightarrow{\text{p}} \overline{Z^2}$. Thus, for the approximation to be "good", we would like to effectively treat the denominator (B) as a constant. Alternatively, we would like to have the mean

[23]

$$\begin{aligned} \mathbb{E}[A] &= \mathbb{E}\left[\sqrt{n}\frac{1}{n}\sum_{i=1}^n Z^i u^i\right] = \left(\sqrt{n}\frac{1}{n}\sum_{i=1}^n Z^i\right)\mathbb{E}\left[u^i\right] = 0, \\ \mathbb{E}[B] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \left(Z^i\right)^2 \pi + \frac{1}{n}\sum_{i=1}^n Z^i v^i\right] = \frac{1}{n}\sum_{i=1}^n \left(Z^i\right)^2 \pi + \left(\frac{1}{n}\sum_{i=1}^n Z^i\right)\mathbb{E}\left[v^i\right] = \overline{Z_n^2}\pi, \\ \operatorname{Var}[A] &= \operatorname{Var}\left[\sqrt{n}\frac{1}{n}\sum_{i=1}^n Z^i u^i\right] = \frac{1}{n}\sum_{i=1}^n \left(Z^i\right)^2 \operatorname{Var}\left[u^i\right] = \overline{Z_n^2}\sigma_1^2, \\ \operatorname{Var}[B] &= \operatorname{Var}\left[\frac{1}{n}\sum_{i=1}^n \left(Z^i\right)^2 \pi + \frac{1}{n}\sum_{i=1}^n Z^i v^i\right] = \frac{1}{n^2}\sum_{i=1}^n \left(Z^i\right)^2 \operatorname{Var}\left[v^i\right] = \frac{1}{n}\overline{Z_n^2}\sigma_2^2, \\ \operatorname{Cov}[A, B] &= \operatorname{Cov}\left[\sqrt{n}\frac{1}{n}\sum_{i=1}^n Z^i u^i, \frac{1}{n}\sum_{i=1}^n \left(Z^i\right)^2 \pi + \frac{1}{n}\sum_{i=1}^n Z^i v^i\right] \\ &= \operatorname{Cov}\left[\sum_{i=1}^n Z^i u^i, \sum_{i=1}^n Z^i v^i\right]\frac{1}{\sqrt{n}n} = \frac{1}{\sqrt{n}n}\sum_{i=1}^n \left(Z^i\right)^2 \operatorname{Cov}\left[u^i, v^i\right] = \frac{1}{\sqrt{n}}\overline{Z_n^2}\sigma_2^2. \end{aligned}$$

$\overline{Z_i^2}\pi$ to be much larger than the standard deviation of the denominator B so denominator is almost like a constant; i.e.

$$\overline{Z^2}\pi \gg \sqrt{\frac{1}{n}\overline{Z_n^2}\sigma_2^2},$$

which, in turn, requires that

$$\pi \gg \frac{1}{\sqrt{n}}.$$

Thus, if π is small (i.e. we have weak instruments), then the approximation can be poor. ???

Note that we cannot directly test if the rank condition holds or if $\pi = 0$ since such tests rely on asymptotics which can be poor for the reasons above.

Below is an example of a hypothesis that does not suffer from weak instrument problems: Anderson-Rubin Test, which involves regressing u^i under the null—i.e. $Y_i - \mathbf{X}^{i\prime}\mathbf{c}$—on $\mathbf{Z}^i$ and examines whether all the coefficients on $\mathbf{Z}^i$ equal zero. It can be shown that this test does not suffer from weak instruments problems.

Example 3.2. *(Anderson-Rubin Test).* Consider the test:

$$H_0 : \boldsymbol{\beta} = \mathbf{c},$$
$$H_1 : \boldsymbol{\beta} \neq \mathbf{c}$$

at level α, where

$$Y^i = \mathbf{X}^{i\prime}\boldsymbol{\beta} + u^i.$$

Under the null, the error term is given by $u^i(\mathbf{c}) \coloneqq Y^i - \mathbf{X}^{i\prime}\mathbf{c}$. To test the hypothesis, we can test the orthogonality condition that holds under the null; i.e.

$$\mathbb{E}\left[\mathbf{Z}^i\left(Y^i - \mathbf{X}^{i\prime}\mathbf{c}\right)\right] = \mathbf{0}.$$

The test statistic is give by

$$T_n = n\bar{w}_n(\mathbf{c})\,\hat{\Sigma}_n\bar{w}_n(\mathbf{c}),$$
$$c_n = \chi^2_{l+1,1-\alpha},$$

where

$$\bar{w}_n(c) \coloneqq \frac{1}{n}\sum_{i=1}^n \mathbf{Z}^i\left(Y^i - \mathbf{X}^{i\prime}\mathbf{c}\right),$$
$$\hat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^n \left(u_i(\mathbf{c}) - \bar{w}_n(\mathbf{c})\right)\left(u_i(\mathbf{c}) - \bar{w}_n(\mathbf{c})\right)'.$$

The test is given by

$$\phi_n = \mathbf{1}_{\{T_n > c_n\}}.$$

3.3.12 Interpretation for TSLS under heterogeneous effects

The model $Y^i = \mathbf{X}^{i\prime}\boldsymbol{\beta} + u^i$ with constant $\boldsymbol{\beta}$ imposes that the effect of a change in $\mathbf{X}$ on Y is the same for every observation; i.e. the model assumes homogeneous effects. We can relax this assumption and allow for heterogeneity by allowing $\boldsymbol{\beta}$ to be random. In this case, we may write the model as

$$Y^i = \mathbf{X}^{i\prime}\boldsymbol{\beta}^i,$$

where we absorbed u_i into β_0^i. This is called the *random coefficient model*.

Let us focus on the case when $k = 1$ and $X^i = D^i \in \{0, 1\}$ so that

$$Y^i = \beta_0^i + \beta_1^i D^i. \tag{3.33}$$

We call Y^i as the *outcome* and D^i as *treatment*. Notice that the outcome (Y^i) equals β_0 when untreated ($D^i = 0$) and equals $\beta_0^i + \beta_1^i$ when treated ($D^i = 1$).

We define *potential outcomes* (also referred to as the counterfactual or latent outcomes), Y_0^i and Y_1^i, as

$$\begin{aligned} Y_0^i &:= \beta_0^i, \\ Y_1^i &:= \beta_0^i + \beta_1^i. \end{aligned}$$

We refer to these as potential or counterfactual outcomes because we can only observe one of the two potential outcomes for each individual i in the data.

We call the difference

$$Y_1^i - Y_0^i = \beta_1^i$$

as the *treatment effect*—i.e. the difference in outcome between when the individual is treated and when he is not. We refer to

$$\mathbb{E}\left[Y_1^i - Y_0^i\right] = \mathbb{E}\left[\beta_1^i\right]$$

as the *average treatment effect* among the population.

We can write Y^i in terms of the potential outcomes:

$$\begin{aligned} Y^i &= \underbrace{Y_0^i}_{\beta_0^i} + \underbrace{\left(Y_1^i - Y_0^i\right)}_{=\beta_1^i} D^i. && (3.34) \\ &= Y_1^i D^i + Y_0^i \left(1 - D^i\right). && (3.35) \end{aligned}$$

Now, let us consider the estimate of the slope coefficient D^i from regressing Y^i on $\left(1, D^i\right)$.

Case 1: When $\left(Y_0^i, Y_1^i\right)$ are independent of D^i If $\left(Y_0^i, Y_1^i\right) \perp\!\!\!\perp D^i$, then

$$\begin{aligned} \mathbb{E}\left[Y_1^i | D^i = 1\right] &= \mathbb{E}\left[Y_1^i | D^i = 0\right] = \mathbb{E}\left[Y_1^i\right], \\ \mathbb{E}\left[Y_0^i | D^i = 0\right] &= \mathbb{E}\left[Y_0^i | D^i = 1\right] = \mathbb{E}\left[Y_0^i\right]. \end{aligned}$$

That is, the potential outcomes are the same between those in the treatment and those in the control groups. The independence condition can be assumed, for example, when the treatment is randomised so that the selection (for treatment) process is independent of the individuals' (unobservable) characteristics.

To understand what the slope coefficient measures, the model we have is

$$Y^i = \beta_0^i + \beta_1^i D^i + u^i.$$

Running OLS of Y^i on $\left(1, D^i\right)$ then gives:[24]

$$\hat{\beta}_1 = \left(\frac{1}{N}\sum_{i=1}^{N}\left(D^i - \bar{D}\right)^2\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\left(D^i - \bar{D}\right)\left(Y^i - \bar{Y}\right)\right).$$

[24] Note that

$$\begin{aligned} \mathbb{E}\left[u^i D^i\right] &= \mathbb{E}\left[\left(Y^i - \beta_0^i - \beta_1^i D^i\right) D^i\right] \\ &= \mathbb{E}\left[\left(Y_1^i D^i + Y_0^i\left(1 - D^i\right)\right) D^i\right] - \mathbb{E}\left[\beta_0^i D^i\right] - \mathbb{E}\left[\beta_1^i \left(D^i\right)^2\right]. \end{aligned}$$

Since $D^i \in \{0, 1\}$, $D^i = \left(D^i\right)^2$ and $\mathbb{P}\left(\left(1 - D^i\right) D^i = 0\right) = 1$. We also have that $\beta_0^i = Y_0^i$ and $\beta_1^i = Y_1^i - Y_0^i$. Hence,

$$\begin{aligned} \mathbb{E}\left[u^i D^i\right] &= \mathbb{E}\left[Y_1^i D^i\right] - \mathbb{E}\left[Y_0^i D^i\right] - \mathbb{E}\left[\left(Y_1^i - Y_0^i\right) D^i\right]. \\ &= \mathbb{E}\left[Y_1^i\right]\mathbb{E}\left[D^i\right] - \mathbb{E}\left[Y_0^i\right]\mathbb{E}\left[D^i\right] - \mathbb{E}\left[Y_1^i D^i\right] - \mathbb{E}\left[Y_0^i D^i\right] \\ &= 0, \end{aligned}$$

where we used the assumption that $\left(Y_0^i, Y_1^i\right) \perp\!\!\!\perp D^i$.

Since[25]

$$\frac{1}{N}\sum_{i=1}^{N}\left(D^i-\bar{D}\right)^2 \xrightarrow{\text{p}} \text{Var}\left[D^i\right],$$
$$\frac{1}{N}\sum_{i=1}^{N}\left(D^i-\bar{D}\right)\left(Y^i-\bar{Y}\right) \xrightarrow{\text{p}} \text{Cov}\left[Y^i,D^i\right],$$

by the Continuous Mapping Theorem,

$$\hat{\beta}_1 \xrightarrow{\text{p}} \beta_1 \coloneqq \frac{\text{Cov}\left[Y^i,D^i\right]}{\text{Var}\left[D^i\right]}.$$

To understand what β_1—i.e. the probability limit of the OLS estimator—measures in this case, note that

[25]These take a little more effort than you think to prove (we can't simply take limits since $\bar{D}$ and $\bar{Y}$ are not iid).

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}\left(D^i-\bar{D}\right)^2 &= \frac{1}{N}\sum_{i=1}^{N}\left(\left(D^i\right)^2-2D^i\bar{D}+\left(\bar{D}\right)^2\right)\\ &= \frac{1}{N}\sum_{i=1}^{N}\left(D^i\right)^2-\left(\bar{D}\right)^2.\end{aligned}$$

Then, since D^i's are iid, by WLLN,

$$\frac{1}{N}\sum_{i=1}^{N}\left(D^i\right)^2 \xrightarrow{\text{p}} \mathbb{E}\left[\left(D^i\right)^2\right].$$

Of course,

$$\bar{D}=\frac{1}{N}\sum_{i=1}^{N}D^i \xrightarrow{\text{p}} \mathbb{E}\left[D^i\right].$$

Since taking a square is a continuous operation, by the continuous mapping theorem,

$$\left(\bar{D}^i\right)^2 \xrightarrow{\text{p}} \mathbb{E}\left[\bar{D}^i\right]^2.$$

Combining everything gives

$$\frac{1}{N}\sum_{i=1}^{N}\left(D^i-\bar{D}\right)^2 \xrightarrow{\text{p}} \mathbb{E}\left[\left(D^i\right)^2\right]-\mathbb{E}\left[\bar{D}^i\right]^2=\text{Var}\left[D^i\right].$$

Similarly,

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}\left(D^i-\bar{D}^i\right)\left(Y^i-\bar{Y}\right) &= \frac{1}{N}\sum_{i=1}^{N}\left(D^i-\bar{D}\right)Y^i+\frac{1}{N}\sum_{i=1}^{N}\left(-\bar{Y}D^i+\bar{D}\bar{Y}\right)\\ &= \frac{1}{N}\sum_{i=1}^{N}\left(D^i-\bar{D}\right)Y^i+\left(-\bar{Y}\bar{D}+\bar{D}\bar{Y}\right)\\ &= \frac{1}{N}\sum_{i=1}^{N}D^iY^i-\bar{D}\bar{Y}.\end{aligned}$$

By WLLN,

$$\frac{1}{N}\sum_{i=1}^{N}D^iY^i \xrightarrow{\text{p}} \mathbb{E}\left[D^iY^i\right],$$
$$\bar{Y}=\frac{1}{N}\sum_{i=1}^{N}Y^i \xrightarrow{\text{p}} \mathbb{E}\left[Y^i\right],$$

we also have from before that $\bar{D} \xrightarrow{\text{p}} \mathbb{E}\left[D^i\right]$. Combining everything gives us

$$\frac{1}{N}\sum_{i=1}^{N}\left(D^i-\bar{D}^i\right)\left(Y^i-\bar{Y}\right) \xrightarrow{\text{p}} \mathbb{E}\left[D^iY^i\right]-\mathbb{E}\left[Y^i\right]\mathbb{E}\left[D^i\right]=\text{Cov}\left[D^i,Y^i\right].$$

$$\begin{aligned}\beta_1 &= \frac{\text{Cov}\left[Y^i, D^i\right]}{\text{Var}\left[D^i\right]} = \frac{\mathbb{E}\left[Y^i D^i\right] - \mathbb{E}\left[Y^i\right]\mathbb{E}\left[D^i\right]}{\mathbb{E}\left[\left(D^i\right)^2\right] - \mathbb{E}\left[D^i\right]\mathbb{E}\left[D^i\right]} \\ &= \frac{\mathbb{E}\left[Y^i D^i\right] - \mathbb{E}\left[Y^i\right]\mathbb{E}\left[D^i\right]}{\mathbb{E}\left[D^i\right]\left(1 - \mathbb{E}\left[D^i\right]\right)}, \end{aligned} \tag{3.36}$$

where we again used the fact that $\left(D^i\right)^2 = D^i$. Consider now the numerator:

$$\begin{aligned}\mathbb{E}\left[Y^i D^i\right] &= \mathbb{E}\left[\left(Y_1^i D^i + Y_0^i\left(1 - D^i\right)\right) D^i\right] \\ &= \mathbb{E}\left[Y_1^i \underbrace{\left(D^i\right)^2}_{=D^i} + Y_0^i \underbrace{\left(1 - D^i\right) D^i}_{=0}\right] \\ &= \mathbb{E}\left[Y_1^i D^i\right] = \mathbb{E}\left[Y_1^i\right]\mathbb{E}\left[D^i\right], \\ \mathbb{E}\left[Y^i\right]\mathbb{E}\left[D^i\right] &= \mathbb{E}\left[Y_1^i D^i + Y_0^i\left(1 - D^i\right)\right]\mathbb{E}\left[D^i\right] \\ &= \mathbb{E}\left[Y_1^i\right]\mathbb{E}\left[D^i\right]^2 + \mathbb{E}\left[Y_0^i\right]\mathbb{E}\left[D^i\right]\underbrace{\mathbb{E}\left[1 - D^i\right]}_{=1-\mathbb{E}[D^i]}, \\ \Rightarrow \text{Cov}\left[Y^i, D^i\right] &= \mathbb{E}\left[Y_1^i\right]\mathbb{E}\left[D^i\right] - \left(\mathbb{E}\left[Y_1^i\right]\mathbb{E}\left[D^i\right]^2 + \mathbb{E}\left[Y_0^i\right]\mathbb{E}\left[D^i\right]\left(1 - \mathbb{E}\left[D^i\right]\right)\right) \\ &= \mathbb{E}\left[Y_1^i\right]\mathbb{E}\left[D^i\right]\left(1 - \mathbb{E}\left[D^i\right]\right) - \mathbb{E}\left[Y_0^i\right]\mathbb{E}\left[D^i\right]\left(1 - \mathbb{E}\left[D^i\right]\right) \\ &= \mathbb{E}\left[Y_1^i - Y_0^i\right]\mathbb{E}\left[D^i\right]\left(1 - \mathbb{E}\left[D^i\right]\right). \end{aligned}$$

So,

$$\begin{aligned}\beta_1 &= \frac{\mathbb{E}\left[Y_1^i - Y_0^i\right]\mathbb{E}\left[D^i\right]\left(1 - \mathbb{E}\left[D^i\right]\right)}{\mathbb{E}\left[D^i\right]\left(1 - \mathbb{E}\left[D^i\right]\right)} \\ &= \mathbb{E}\left[Y_1^i - Y_0^i\right]. \end{aligned}$$

That is, the OLS estimator of the coefficient on D^i gives the *average treatment effect* in the case $\left(Y_0^i, Y_1^i\right) \perp\!\!\!\perp D^i$.

Case 2: When $\left(Y_0^i, Y_1^i\right)$ is not independent of D^i We would not have $\left(Y_0^i, Y_1^i\right) \perp\!\!\!\perp D^i$ if there are unobservable factors that affect whether an individual is treated as well as the outcome. What do we know about β_1 when $\left(Y_0^i, Y_1^i\right)$ is not independent of D^i?

Recall that

$$\beta_1 = \frac{\mathbb{E}\left[Y^i D^i\right] - \mathbb{E}\left[Y^i\right]\mathbb{E}\left[D^i\right]}{\mathbb{E}\left[D^i\right]\left(1 - \mathbb{E}\left[D^i\right]\right)}$$

Then,

$$\begin{aligned}
\mathbb{E}\left[Y^i D^i\right] &= \mathbb{E}\left[Y_1^i D^i\right] \\
&= \mathbb{E}\left[Y_1^i \cdot 1 | D^i = 1\right] \mathbb{P}\left(D^i = 1\right) + \mathbb{E}\left[Y_1^i \cdot 0 | D^i = 0\right] \mathbb{P}\left(D^i = 0\right) \\
&= \mathbb{E}\left[Y_1^i | D^i = 1\right] \mathbb{E}\left[D^i\right] \\
\mathbb{E}\left[Y^i\right] \mathbb{E}\left[D^i\right] &= \mathbb{E}\left[Y_1^i D^i + Y_0^i \left(1 - D^i\right)\right] \mathbb{E}\left[D^i\right] \\
&= \left(\mathbb{E}\left[Y_1^i D^i\right] + \mathbb{E}\left[Y_0^i \left(1 - D^i\right)\right]\right) \mathbb{E}\left[D^i\right] \\
&= \mathbb{E}\left[Y_1^i | D^i = 1\right] \mathbb{E}\left[D^i\right]^2 + \mathbb{E}\left[D^i\right] \mathbb{E}\left[Y_0^i \cdot 0 | D^i = 1\right] \mathbb{P}\left(D^i = 1\right) \\
&\quad + \mathbb{E}\left[D^i\right] \mathbb{E}\left[Y_0^i \cdot 1 | D^i = 0\right] \mathbb{P}\left(D^i = 0\right) \\
&= \mathbb{E}\left[Y_1^i | D^i = 1\right] \mathbb{E}\left[D^i\right]^2 + \mathbb{E}\left[Y_0^i | D^i = 0\right] \mathbb{E}\left[D^i\right] \left(1 - \mathbb{E}\left[D^i\right]\right) \\
\mathbb{E}\left[Y^i D^i\right] - \mathbb{E}\left[Y^i\right] \mathbb{E}\left[D^i\right] &= \mathbb{E}\left[Y_1^i | D^i = 1\right] \mathbb{E}\left[D^i\right] \\
&\quad - \left(\mathbb{E}\left[Y_1^i | D^i = 1\right] \mathbb{E}\left[D^i\right]^2 + \mathbb{E}\left[Y_0^i | D^i = 0\right] \mathbb{E}\left[D^i\right] \left(1 - \mathbb{E}\left[D^i\right]\right)\right) \\
&= \mathbb{E}\left[Y_1^i | D^i = 1\right] \mathbb{E}\left[D^i\right] \left(1 - \mathbb{E}\left[D^i\right]\right) - \mathbb{E}\left[Y_0^i | D^i = 0\right] \mathbb{E}\left[D^i\right] \left(1 - \mathbb{E}\left[D^i\right]\right) \\
&= \left(\mathbb{E}\left[Y_1^i | D^i = 1\right] - \mathbb{E}\left[Y_0^i | D^i = 0\right]\right) \mathbb{E}\left[D^i\right] \left(1 - \mathbb{E}\left[D^i\right]\right).
\end{aligned}$$

Hence,

$$\begin{aligned}
\beta_1 &= \frac{\left(\mathbb{E}\left[Y_1^i | D^i = 1\right] - \mathbb{E}\left[Y_0^i | D^i = 0\right]\right) \mathbb{E}\left[D^i\right] \left(1 - \mathbb{E}\left[D^i\right]\right)}{\mathbb{E}\left[D^i\right] \left(1 - \mathbb{E}\left[D^i\right]\right)} \\
&= \mathbb{E}\left[Y_1^i | D^i = 1\right] - \mathbb{E}\left[Y_0^i | D^i = 0\right]. \qquad (3.37)
\end{aligned}$$

Thus, β_1 is now the difference in the potential outcome for the treated (Y_1^i) among those were treated and the potential outcome for the untreated for those were not treated (Y_0^i). However, without independence, we cannot rule out the possibility that the differences are caused only by the treatment; i.e. there may be other unobserved factors that affect the differences in the potential outcome, and the effect of such factors are included in β_1. This means, for example, that we may observe $\mathbb{E}\left[Y_1^i - Y_0^i\right] \neq 0$ even when the treatment has no effect (we may simply be capturing the environmental effect). What can we do about this? We can use instruments.

Suppose that we can find an instrument $Z^i \in \{0, 1\}$ for D^i. Now, we first regress the model (again, the error term is absorbed in π_0)

$$D^i = \pi_0^i + \pi_1^i Z^i.$$

As before, we can define potential treatments as

$$\begin{aligned}
D_0^i &\coloneqq \pi_0^i, \\
D_1^i &\coloneqq \pi_0^i + \pi_1^i,
\end{aligned}$$

and write outcome D^i as a function of potential outcomes and the instrument:

$$\begin{aligned}
D^i &= D_0^i + \left(D_1^i - D_0^i\right) Z^i \\
&= D_1^i Z^i + D_0^i \left(1 - Z^i\right). \qquad (3.38)
\end{aligned}$$

To understand what (3.38) means, recall that, previously, we had two groups of potential outcomes for Y^i: one for when $D^i = 0$ (Y_0^i) and the other when $D^i = 1$ (Y_1^i). With the instrument, we split each group into two further (sub)groups. The idea is that such a decomposition allows us to identify part of the treatment effect due solely to the treatment (from the effect of other factors).

There are four cases:

▷ instrument treatment has no effect on D^i:

- ▷ an individual may choose not to be treated independent of whether they receive instrumental treatment—these are called *never takers* since they never take on treatment. In this case $D_1^i = D_0^i = 0$ so that $D^i = 0$ whether $Z^i = 0$ or $Z^i = 1$.
- ▷ an individual may instead choose to be treated independent of whether they receive instrumental treatment—these are called *always takers* since they always take on treatment. In this case, $D_1^i = D_0^i = 1$ so that $D^i = 1$ whether $Z^i = 0$ or $Z^i = 1$.

▷ instrument treatment has an effect on D^i:

- ▷ an individual may choose to switch as a result of the instrument treatment, from not being treated to being treated—these are called *compliers.* In this case, $D_1^i > D_0^i$. Since $D^i \in \{0, 1\}$, this implies that $1 = D_1^i > D_0^i = 0$—i.e. when treated in the first stage ($Z^i = 1$), then this individual is treated in the second stage ($D_1^i = 1$); if untreated in the first stage ($Z^i = 0$), then the individual does not switch and is not treated in the second stage ($D_0^i = 0$).
- ▷ an individual may also choose to switch from being treated to not being treated—these are called *defiers*. This implies that $D_1^i < D_0^i$ so that $0 = D_1^i < D_0^i = 1$—i.e. when treated in the first stage ($Z^i = 1$), then this individual is untreated in the second stage ($D_1^i = 0$); if untreated in the first stage ($Z^i = 0$), then the individual switches and is treated in the second stage ($D_0^i = 1$).

Above can be summarised in the table below:

	$D_0^i = 0$	$D_0^i = 1$
$D_1^i = 0$	Never takers	Defiers
$D_1^i = 1$	Compliers	Always takers

To proceed, we assume that the following conditions hold:

▷ instrument exogeneity: $\left(Y_1^i, Y_0^i, D_1^i, D_0^i\right) \perp\!\!\!\perp Z^i$.

This ensures that the expectations of potential outcomes are no different among those for whom $Z^i = 1$ and those for whom $Z^i = 0$.

▷ instrument relevance: $\mathbb{P}\left(D_1^i \neq D_0^i\right) > 0$.

If the instrument is completely irrelevant to the determination of whether an individual is in the treatment or the control group, then we would expect $\pi_1^i = 0$ in all cases (recall that the coefficients can differ across observations here); i.e. $\mathbb{P}\left(D_0^i = D_1^i\right) = 1$. This condition therefore says that instruments do have an impact on D.

▷ (uniform) monotonicity: $\mathbb{P}\left(D_1^i \geq D_0^i\right) = 1$.

Thus, this rules out the case $D_1^i < D_0^i$; i.e. monotonicity rules out existence of defiers.

Recall the results for solving for subvectors of $\boldsymbol{\beta}$ when we split out the constant term, (3.25). The model in this case is given by

$$\begin{aligned} Y^i &= \beta_0^i + \beta_1^i D^i + u^i, \\ D^i &= \delta_0^i + \delta_1^i Z^i + v^i. \end{aligned}$$

Since we are in the just-identified case, we can use the IV estimator:[26]

$$\hat{\beta}_{IV} = \left(\frac{1}{N}\sum_{i=1}^{N}\left(Z^i - \bar{Z}\right)\left(D^i - \bar{D}\right)\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\left(Z^i - \bar{Z}\right)\left(Y^i - \bar{Y}\right)\right).$$

[26] Note that

$$\begin{aligned} \mathbb{E}\left[Z^i u^i\right] &= \mathbb{E}\left[Z^i\left(Y^i - \beta_0^i - \beta_1^i D^i\right)\right] \\ &= \mathbb{E}\left[Z^i Y^i\right] - \mathbb{E}\left[Z^i \beta_0^i\right] - \mathbb{E}\left[Z^i \beta_1^i D^i\right], \end{aligned}$$

Define β_1 as its probability limit:

$$\frac{1}{N}\sum_{i=1}^{N}\left(Z^i-\bar{Z}\right)\left(D^i-\bar{D}\right)\xrightarrow{\text{p}}\text{Cov}\left[Z^i,D^i\right],$$
$$\frac{1}{N}\sum_{i=1}^{N}\left(Z^i-\bar{Z}\right)\left(Y^i-\bar{Y}\right)\xrightarrow{\text{p}}\text{Cov}\left[Z^i,Y^i\right],$$
$$\Rightarrow\hat{\beta}_{IV}\xrightarrow{\text{p}}\beta_1\coloneqq\frac{\text{Cov}\left[Z^i,Y^i\right]}{\text{Cov}\left[Z^i,D^i\right]}.$$

How do we interpret this coefficient?

First, write

$$\beta_1=\frac{\text{Cov}\left[Z^i,Y^i\right]/\text{Var}\left[Z^i\right]}{\text{Cov}\left[Z^i,D^i\right]/\text{Var}\left[Z^i\right]}.$$

Using the same steps as when we derived (3.37),

$$\frac{\text{Cov}\left[Y^i,Z^i\right]}{\text{Var}\left[Z^i\right]}=\mathbb{E}\left[Y^i|Z^i=1\right]-\mathbb{E}\left[Y^i|Z^i=0\right],\tag{3.39}$$
$$\frac{\text{Cov}\left[D^i,Z^i\right]}{\text{Var}\left[Z^i\right]}=\mathbb{E}\left[D^i|Z^i=1\right]-\mathbb{E}\left[D^i|Z^i=0\right].\tag{3.40}$$

Substituting the expression for D^i in the expression for the denominator, (3.40), yields

$$\begin{aligned}\frac{\text{Cov}\left[D^i,Z^i\right]}{\text{Var}\left[Z^i\right]}&=\mathbb{E}\left[D_1^iZ^i+D_0^i\left(1-Z^i\right)|Z^i=1\right]-\mathbb{E}\left[D_1^iZ^i+D_0^i\left(1-Z^i\right)|Z^i=0\right]\\&=\mathbb{E}\left[D_1^i|Z^i=1\right]-\mathbb{E}\left[D_0^i|Z^i=0\right]\\[\text{exogeneity}]&=\mathbb{E}\left[D_1^i-D_0^i\right]\\&=\mathbb{E}\left[D_1^i-D_0^i|D_1^i>D_0^i\right]\mathbb{P}\left(D_1^i>D_0^i\right)+\mathbb{E}\left[D_1^i-D_0^i|D_1^i<D_0^i\right]\mathbb{P}\left(D_1^i<D_0^i\right)\\&\quad+\mathbb{E}\left[D_1^i-D_0^i|D_1^i=D_0^i\right]\mathbb{P}\left(D_1^i=D_0^i\right)\\[\text{monotonicity}]&=\mathbb{E}\left[D_1^i-D_0^i|D_1^i>D_0^i\right]\mathbb{P}\left(D_1^i>D_0^i\right)\\[\text{binary}]&=\mathbb{P}\left(D_1^i>D_0^i\right).\end{aligned}$$

where

$$\begin{aligned}\mathbb{E}\left[Z^iY^i\right]&=\mathbb{E}\left[Z^i\left(Y_1^iD^i+Y_0^i\left(1-D^i\right)\right)\right]\\&=\mathbb{E}\left[Z^i\left(Y_1^i\left(D_1^iZ^i+D_0^i\left(1-Z^i\right)\right)+Y_0^i\left(1-\left(D_1^iZ^i+D_0^i\left(1-Z^i\right)\right)\right)\right)\right]\\&=\mathbb{E}\left[Y_1^iD_1^i\left(Z^i\right)^2+Y_1^iD_0^i\left(1-Z^i\right)Z^i+Y_0^i\left(Z^i-D_1^i\left(Z^i\right)^2-D_0^i\left(1-Z^i\right)Z^i\right)\right]\\&=\mathbb{E}\left[Y_1^iD_1^iZ^i+Y_0^i\left(1-D_1^i\right)Z^i\right]\\&=\mathbb{E}\left[Y_1^iD_1^i+Y_0^i\left(1-D_1^i\right)\right]\mathbb{E}\left[Z^i\right]\\\mathbb{E}\left[Z^i\beta_0^i\right]&=\mathbb{E}\left[Z^iY_0^i\right]=\mathbb{E}\left[Z^i\right]\mathbb{E}\left[Y_0^i\right],\\\mathbb{E}\left[Z^i\beta_1^iD^i\right]&=\mathbb{E}\left[Z^i\beta_1^i\left(D_1^iZ^i+D_0^i\left(1-Z^i\right)\right)\right]\\&=\mathbb{E}\left[\beta_1^i\left(D_1^i\left(Z^i\right)^2+D_0^i\left(1-Z^i\right)Z^i\right)\right]\\&=\mathbb{E}\left[\beta_1^iD_1^iZ^i\right]=\mathbb{E}\left[\left(Y_1^i-Y_0^i\right)D_1^iZ^i\right]\\&=\mathbb{E}\left[Y_1^iD_1^iZ^i\right]-\mathbb{E}\left[Y_0^iD_1^iZ^i\right]\\&=\left(\mathbb{E}\left[Y_1^iD_1^i\right]-\mathbb{E}\left[Y_0^iD_1^i\right]\right)\mathbb{E}\left[Z^i\right].\end{aligned}$$

So

$$\begin{aligned}\mathbb{E}\left[Z^iu^i\right]&=\left(\mathbb{E}\left[Y_1^iD_1^i+Y_0^i\left(1-D_1^i\right)\right]-\mathbb{E}\left[Y_0^i\right]-\mathbb{E}\left[Y_1^iD_1^i\right]+\mathbb{E}\left[Y_0^iD_1^i\right]\right)\mathbb{E}\left[Z^i\right]\\&=\left(\mathbb{E}\left[Y_1^iD_1^i\right]+\mathbb{E}\left[Y_0^i\right]-\mathbb{E}\left[Y_0^iD_1^i\right]-\mathbb{E}\left[Y_0^i\right]-\mathbb{E}\left[Y_1^iD_1^i\right]+\mathbb{E}\left[Y_0^iD_1^i\right]\right)\mathbb{E}\left[Z^i\right]\\&=0.\end{aligned}$$

(monotonicity implies that $\mathbb{P}\left(D_1^i < D_0^i\right) = 0$ and that, if $D_1^i > D_0^i$, then $\mathbb{E}\left[D_1^i - D_0^i\right] = 1$).

For the numerator, (3.39), first note that

$$\begin{aligned}
\mathbb{E}\left[Y^i | Z^i = 1\right] &= \mathbb{E}\left[Y_1^i D^i + Y_0^i \left(1 - D^i\right) | Z^i = 1\right] \\
&= \mathbb{E}\left[Y_1^i \left(D_1^i Z^i + D_0^i \left(1 - Z^i\right)\right) + Y_0 \left(1 - \left(D_1^i Z^i + D_0^i \left(1 - Z^i\right)\right)\right) | Z^i = 1\right] \\
&= \mathbb{E}\left[Y_1^i D_1^i + Y_0^i \left(1 - D_1^i\right) | Z^i = 1\right] \\
\text{[exogeneity]} &= \mathbb{E}\left[Y_1^i D_1^i + Y_0^i \left(1 - D_1^i\right)\right], \\
\mathbb{E}\left[Y | Z^i = 0\right] &= \mathbb{E}\left[Y_1^i \left(D_1^i Z^i + D_0^i \left(1 - Z^i\right)\right) + Y_0 \left(1 - \left(D_1^i Z^i + D_0^i \left(1 - Z^i\right)\right)\right) | Z^i = 0\right] \\
&= \mathbb{E}\left[Y_1^i D_0^i + Y_0^i \left(1 - D_0^i\right) | Z^i = 0\right] \\
\text{[exogeneity]} &= \mathbb{E}\left[Y_1^i D_0^i + Y_0^i \left(1 - D_0^i\right)\right].
\end{aligned}$$

Then,

$$\begin{aligned}
\frac{\text{Cov}\left[Y^i, Z^i\right]}{\text{Var}\left[Z^i\right]} &= \mathbb{E}\left[Y_1^i D_1^i + Y_0^i \left(1 - D_1^i\right)\right] - \mathbb{E}\left[Y_1^i D_0^i + Y_0^i \left(1 - D_0^i\right)\right] \\
&= \mathbb{E}\left[Y_1^i D_1^i - Y_1^i D_0^i - Y_0^i D_1^i + Y_0^i D_0^i\right] \\
&= \mathbb{E}\left[\left(Y_1^i - Y_0^i\right)\left(D_1^i - D_0^i\right)\right] \\
&= \mathbb{E}\left[\left(Y_1^i - Y_0^i\right)\left(D_1^i - D_0^i\right) | D_1^i > D_0^i\right] \mathbb{P}\left(D_1^i > D_0^i\right) \\
&\quad + \mathbb{E}\left[\left(Y_1^i - Y_0^i\right)\left(D_1^i - D_0^i\right) | D_1^i < D_0^i\right] \mathbb{P}\left(D_1^i < D_0^i\right) \\
&\quad + \mathbb{E}\left[\left(Y_1^i - Y_0^i\right)\left(D_1^i - D_0^i\right) | D_1^i = D_0^i\right] \mathbb{P}\left(D_1^i = D_0^i\right) \\
\text{[monotonicity]} &= \mathbb{E}\left[\left(Y_1^i - Y_0^i\right)\left(D_1^i - D_0^i\right) | D_1^i > D_0^i\right] \mathbb{P}\left(D_1^i > D_0^i\right) \\
&= \mathbb{E}\left[Y_1^i - Y_0^i | D_1^i > D_0^i\right] \mathbb{P}\left(D_1^i > D_0^i\right)
\end{aligned}$$

Together, we realise that

$$\begin{aligned}
\beta_1 &= \frac{\text{Cov}\left[Y^i, Z^i\right] / \text{Var}\left[Z^i\right]}{\text{Cov}\left[D^i, Z^i\right] / \text{Var}\left[Z^i\right]} = \frac{\mathbb{E}\left[Y_1^i - Y_0^i | D_1^i > D_0^i\right] \mathbb{P}\left(D_1^i > D_0^i\right)}{\mathbb{P}\left(D_1^i > D_0^i\right)} \\
&= \mathbb{E}\left[Y_1^i - Y_0^i | D_1^i > D_0^i\right].
\end{aligned}$$

The condition that $D_1^i > D_0^i$ is equivalent to $D_1^i = 1$ and $D_0^i = 0$. Thus, we realise that β_1 measures the the *average effect of the treatment among the compliers* as shown in the table above. This estimate is called the *local average treatment effect* (LATE).

Remark 3.7. Things to keep in mind when estimating LATE.

(i) Different choice of instrumental variable Z^i will lead to different LATE

(ii) Monotonicity condition is important (and often difficult to justify).

(iii) Does knowing LATE answer the question? Sometimes we are interested in how the treatment affects other groups of individuals.

(iv) No covariates.

If we do not have monotonicity, then

$$\begin{aligned}
\beta_1 &= \frac{\mathbb{E}\left[\left(Y_1^i - Y_0^i\right)\left(D_1^i - D_0^i\right)\right]}{\mathbb{E}\left[D_1^i - D_0^i\right]} \\
&= \mathbb{E}\left[\frac{D_1^i - D_0^i}{\mathbb{E}\left[D_1^i - D_0^i\right]}\left(Y_1^i - Y_0^i\right)\right].
\end{aligned}$$

Thus, we realise that the monotonicity allows us to ensure that the weight put on $Y_1^i - Y_0^i$ all have the same sign.

4 Maximum Likelihood (ML) Estimators

4.1 Unconditional ML estimator

Suppose we have $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} P = P_{\theta_0}$, where $\theta_0 \in \Theta$. Suppose that each P_θ has a density p_θ with respect to a common measure μ.

Definition 4.1. *(Unconditional likelihood)* The *likelihood*, denoted $\ell_n(\theta)$, is given by the joint density of $X_1, X_2, \ldots, X_n$ under θ evaluated at the realised values $x_1, x_2, \ldots, x_n$:

$$\ell_n(\theta) \coloneqq \prod_{i=1}^{n} p_\theta(x_i),$$

where the iid assumption allows us to write the joint density as product of marginal densities. Sometimes, it will be more convenient to work with *log-likelihood* functions:

$$L_n(\theta) \coloneqq \frac{1}{n} \log \ell_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \log p_\theta(x_i).$$

Since log is a monotone function, in particular, it preserves the maximum.

We express joint density as product of $p_\theta(x_i)$ as we assume independence (and identically distributed means that every x_i has the same p_θ).

Definition 4.2. *(ML estimator)* The maximum likelihood estimator is given by

$$\hat{\theta}_n \in \arg\max_{\theta \in \Theta} \ell_n(\theta).$$

Note that this is equivalent to

$$\hat{\theta}_n \in \arg\max_{\theta \in \Theta} L_n(\theta).$$

To understand the intuition behind the ML estimator, suppose that we observe a random sample, $x_1, x_2, \ldots, x_n$ that came from a discrete distribution. If an estimate of θ must be selected, we would not consider any value of θ for which it would have been impossible to obtain the observed data, $x_1, x_2, \ldots, x_n$. Furthermore, suppose that the probability $p(x_1, x_2, \ldots, x_n|\theta)$ of obtaining the actual observed data is very high when θ has a particular value, say $\theta = \hat{\theta}$, and is very small for every other value of θ. Then, we would naturally estimate the value of θ to be $\hat{\theta}$. When the sample comes from a continuous distribution, it would again be natural to find a value of θ for which the probability density $p(x_1, x_2, \ldots, x_n|\theta)$ is large, and to use this value as an estimate of θ. For any given observed data, we are led by this reasoning to consider a value of θ for which the likelihood function is a maximum and use this value as an estimate of θ_0.

Thus, maximum likelihood means that we choose the parameter that makes the likelihood of having the observed data maximum. With discrete distributions, the likelihood is the same as the probability. We choose the parameter for the density that maximises the probability of the data coming from it.

Theoretically, if we had no observed data, maximising the likelihood function will give us a function of n random variables $X_1, X_2, \ldots, X_n$. This is what we call the ML estimator.

Note that $\hat{\theta}_n$ may not be unique. In that case, we can simply take any maximiser. It may also be that $\hat{\theta}_n$ does not exist. In that case, we can take a "near" maximiser.

Example 4.1. *(Uniformly distributed sample)* Suppose $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} P = P_{\theta_0} = \text{unif}[0, \theta_0]$ with $\theta \in (0, \infty)$. The density function for each observation is given by

$$p_\theta(x) = \frac{1}{\theta} \mathbf{1}_{\{0 \le x \le \theta\}} = \begin{cases} 1/\theta & \text{if } 0 \le x \le \theta \\ 0 & \text{otherwise} \end{cases}.$$

The likelihood function is then

$$\begin{aligned}\ell_n(\theta) &= \prod_{i=1}^{n} p_\theta(X_i) = \prod_{i=1}^{n} \frac{1}{\theta}\mathbf{1}_{\{0\le X_i\le\theta\}} \\ &= \left(\frac{1}{\theta}\right)^n \mathbf{1}_{\{0\le X_1,X_2,\ldots,X_n\le\theta\}} \\ &= \begin{cases} 1/\theta^n & \text{if } 0\le X_i\le\theta, \forall i \\ 0 & \text{otherwise}\end{cases}.\end{aligned}$$

We can see that the ML estimator of θ must be a value of θ for which $\theta \ge X_i$ for all $i = 1, 2, \ldots, n$, and that maximises $1/\theta^n$ among all such values. Since $1/\theta^n$ is a decreasing function of θ, the estimate will be the smallest possible value of θ such that $\theta \ge X_i$ for $i = 1, 2, \ldots, n$. That is,

$$\hat{\theta}_n = \max_{1\le i\le n} X_i.$$

Remark 4.1. Notice that $\max_{1\le i\le n} X_i < \theta$ with probability one so that $\hat{\theta}_n$ surely underestimates the value of θ. This suggests that $\hat{\theta}_n$ may not be a suitable estimator of θ.

Example 4.2. *(Uniformly distributed sample in which ML estimator does not exist)* Suppose we alter the support of P in the previous example to the interval $(0, \theta)$ with $\theta \in (0, \infty)$. The density function is now:

$$p_\theta(x) = \frac{1}{\theta}\mathbf{1}_{\{0<x<\theta\}} = \begin{cases} 1/\theta & \text{if } 0<x<\theta \\ 0 & \text{otherwise}\end{cases}.$$

The ML estimator is the value of θ for which $\theta > X_i$ for all i that maximises $1/\theta^n$ among all such values. Note that the possible values of θ no longer include the value $\theta = \max_{1\le i\le n} X_i$ since θ must be strictly greater than the observed values. Since we can choose θ to be arbitrarily close to the value of $\max_{1\le i\le n} X_i$ but cannot be chosen to equal this value, it follows that the MLE of θ does not exist in this case.

Example 4.3. *(Non-unique ML estimator)* Suppose we alter P so that the density function is given by

$$p_\theta(x) = \mathbf{1}_{\{\theta\le x\le\theta+1\}} = \begin{cases} 1 & \text{if } \theta\le x\le\theta+1 \\ 0 & \text{otherwise}\end{cases},$$

where $\theta \in \mathbb{R}$. In this example, the likelihood function is

$$\ell_n(\theta) = \begin{cases} 1 & \text{if } \theta\le X_i\le\theta+1, \forall i \\ 0 & \text{otherwise}\end{cases}.$$

The condition that $\theta \le X_i$ for all i is equivalent to the condition that $\theta \le \min_{1\le i\le n} X_i$, and the condition that $\theta + 1 \ge X_i$ for all i is equivalent to the condition that $\theta \ge \max_{1\le i\le n} X_i - 1$. Therefore, we can rewrite the likelihood function as

$$\ell_n(\theta) = \begin{cases} 1 & \text{if } \max_{1\le i\le n} X_i - 1 \le \theta \le \min_{1\le i\le n} X_i \\ 0 & \text{otherwise}\end{cases}.$$

That is, we can select any value in the interval $[\max_{1\le i\le n} X_i - 1, \min_{1\le i\le n} X_i]$ as the ML estimator for θ; i.e. ML estimator is not unique in this example.

Example 4.4. Suppose $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} P = P_{\theta_0} = \text{Bernoulli}(\theta_0)$ with $\theta_0 \in (0, 1)$. Then

$$\begin{aligned}p_\theta(x) &= \theta\mathbf{1}_{\{x=1\}} + (1-\theta)\mathbf{1}_{\{x=0\}} = \theta^x(1-\theta)^{1-x} \\ &= \begin{cases}\theta & \text{if } x=1 \\ 1-\theta & \text{if } x=0\end{cases}.\end{aligned}$$

The likelihood function is then

$$\begin{aligned}\ell_n(\theta) &= \prod_{i=1}^{n} p_\theta(X_i) = \prod_{i=1}^{n} \theta^{X_i}(1-\theta)^{1-X_i} \\ &= \theta^{\sum_{i=1}^n X_i}(1-\theta)^{\sum_{i=1}^n (1-X_i)} \\ &= \theta^{n\bar{X}_n}(1-\theta)^{n\left(1-\bar{X}_n\right)} = \left[\theta^{\bar{X}_n}(1-\theta)^{1-\bar{X}_n}\right]^n.\end{aligned}$$

Consider the log-likelihood function,

$$\begin{aligned}L_n(\theta) &= \frac{1}{n}\sum_{i=1}^{n} \log p_\theta(X_i) = \frac{1}{n}\sum_{i=1}^{n} \log\left(\theta^{X_i}(1-\theta)^{1-X_i}\right) \\ &= \log(\theta)\frac{1}{n}\sum_{i=1}^{n} X_i + \log(1-\theta)\frac{1}{n}\sum_{i=1}^{n}(1-X_i) \\ &= \log(\theta)\bar{X}_n + \log(1-\theta)\left(1-\bar{X}_n\right).\end{aligned}$$

The ML estimator is given by

$$\hat{\theta}_n = \arg\max_{\theta\in\Theta} L_n(\theta) = \max_{\theta\in(0,1)} \log(\theta)\bar{X}_n + \log(1-\theta)\left(1-\bar{X}_n\right)$$

The first-order condition is

$$\begin{aligned}0 &= \frac{\bar{X}_n}{\hat{\theta}_n} - \frac{1-\bar{X}_n}{1-\hat{\theta}_n} \\ &= \frac{\left(1-\hat{\theta}_n\right)\bar{X}_n - \hat{\theta}_n\left(1-\bar{X}_n\right)}{\hat{\theta}_n\left(1-\hat{\theta}_n\right)} \\ &= \frac{\bar{X}_n - \hat{\theta}_n}{\hat{\theta}_n\left(1-\hat{\theta}_n\right)} \\ \Rightarrow \hat{\theta}_n &= \bar{X}_n.\end{aligned}$$

We confirm that we have found the maximum by checking the second-order condition:

$$-\frac{\bar{X}_n}{\theta^2} - \frac{1-\bar{X}_n}{(1-\theta)^2} < 0.$$

Example 4.5. Suppose $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} P = P_{\theta_0}$ where P_θ is the distribution of $X = \max\{Z-\theta, 0\}$ with $Z \sim N(0,1)$; i.e. a *censored* distribution. The probability of observing $x = 0$ is $\mathbb{P}(z \leq \theta)$. Since Z is standard normal,

$$p_{\theta_0}(0) = \mathbb{P}(z \leq \theta) = \Phi(\theta).$$

The probability of observing $x > 0$ is just the same as probability of observing z whenever $z > \theta$, so

$$p_{\theta_0}(x) = \phi(z) = \phi(x+\theta).$$

Combining the two, we can write

$$\begin{aligned}p_\theta(x) &= \Phi(\theta)\mathbf{1}_{\{x=0\}} + \phi(x+\theta)\mathbf{1}_{\{x>0\}} \\ &= \begin{cases}\Phi(\theta) & \text{if } x = 0 \\ \phi(x+\theta) & \text{if } x > 0\end{cases}.\end{aligned}$$

The likelihood function is then given by

$$\begin{aligned}\ell_n(\theta) &= \prod_{i=1}^{n} p_\theta(X_i) \\ &= \left(\prod_{i=1}^{n} \Phi(\theta)\mathbf{1}_{\{X_i=0\}}\right)\left(\prod_{i=1}^{n} \phi(X_i+\theta)\mathbf{1}_{\{X_i>0\}}\right).\end{aligned}$$

As in the example above, we may not always be able to obtain an explicit characterisation of $\hat{\theta}_n$. However, we can still study its properties through implicit characterisation.

4.2 Conditional ML estimators

Suppose we have $(Y_1, X_1), (Y_2, X_2), \ldots, (Y_n, X_n) \overset{\text{iid}}{\sim} P$. Assume that the conditional distribution of Y_i given X_i, $Y_i|X_i$, is P_{θ_0} with $\theta_0 \in \Theta$. Denote by P_X the distribution of X_i itself. Suppose also that each P_θ has a density $p_\theta(y|x)$ with respect to a common measure μ.

Definition 4.3. *(Conditional Likelihood)* The conditional likelihood is the joint density of $Y_1, Y_2, \ldots, Y_n$ given $X_1, X_2, \ldots, X_n$ under θ evaluated at $(y_1, x_1), (y_2, x_2), \ldots, (y_n, x_n)$:

$$\ell_n(\theta) \coloneqq \prod_{i=1}^{n} p_\theta(y_i|x_i).$$

The conditional ML estimator, $\hat{\theta}_n$, is given by

$$\hat{\theta}_n \in \arg\max_{\theta\in\Theta} \ell_n(\theta).$$

The conditional log-likelihood is defined analogously as

$$L_n(\theta) \coloneqq \frac{1}{n}\sum_{i=1}^{n} \log p_\theta(y_i|x_i).$$

Notice that assuming X_i's to be constant gives the unconditional likelihood.

Example 4.6. *(Probit)* Suppose we have $(Y_1, X_1), (Y_2, X_2), \ldots, (Y_n, X_n) \overset{\text{iid}}{\sim} P$ on $\{0,1\} \times \mathbb{R}^{k+1}$ and assume that the first element of X_i equals one. Suppose that

$$\begin{aligned}p_\theta(y|x) &= \Phi(x'\theta)\mathbf{1}_{\{y=1\}} + [1-\Phi(x'\theta)]\mathbf{1}_{\{y=0\}} \\ &= \begin{cases}\Phi(x'\theta) & \text{if } y=1 \\ 1-\Phi(x'\theta) & \text{if } y=0\end{cases} \\ &= \Phi(x'\theta)^y [1-\Phi(x'\theta)]^{1-y}.\end{aligned}$$

The likelihood function is then

$$\begin{aligned}\ell_n(\theta) &= \prod_{i=1}^{n} \Phi(X_i'\theta)^{Y_i}[1-\Phi(X_i'\theta)]^{1-Y_i} \\ \Rightarrow L_n(\theta) &= \frac{1}{n}\sum_{i=1}^{n}\left(Y_i \log(X_i'\theta) + (1-Y_i)\log(1-\Phi(X_i'\theta))\right).\end{aligned}$$

This gives the Probit model with $Y = \mathbf{1}_{\{X'\theta \geq \varepsilon\}}$ where $\varepsilon \sim N(0,1)$ and independent of X. If we replace Φ with the logistic CDF (i.e. replace ε with the Logit distribution), then we will get the Logit model. Note that CDF of Logit is given by

$$G(X'\theta) = \frac{\exp(X'\theta)}{1+\exp(X'\theta)}.$$

4.3 Properties of ML estimators

4.3.1 Consistency

By WLLN, we know that

$$L_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log p_\theta(Y_i|X_i) \xrightarrow{\text{p}} \mathbb{E}[\log p_\theta(Y|X)] =: L(\theta).$$

So maybe $\hat{\theta}_n$ would converge in probability to the maximum of $L(\theta)$, and hopefully to θ_0 (note that θ_0 is "buried" in $\mathbb{E}[\log p_\theta(Y_i|X_i)]$).

For consistency to hold, we require that the (log) likelihood is uniquely maximised at θ_0 (if not, how do we know which maximiser $\hat{\theta}_n$ "should" converge to?). The following Lemmas provide sufficient conditions.

Lemma 4.1. *(Uniqueness of ML estimators) Suppose that for every $\theta \neq \theta_0$,*

$$\mathbb{P}(p_\theta(Y|X) \neq p_{\theta_0}(Y|X)) > 0,$$

then, $L(\theta) \coloneqq \mathbb{E}[\log p_\theta(Y|X)]$ is uniquely maximised at $\theta = \theta_0$.

Proof. Define M to be the log-likelihood ratio between θ model and θ_0 model:

$$\begin{aligned} M(\theta) &\coloneqq L(\theta) - L(\theta_0) \\ &= \mathbb{E}_{\theta_0}[\log p_\theta(Y|X)] - \mathbb{E}[\log p_{\theta_0}(Y|X)] \\ &= \mathbb{E}_{\theta_0}\left[\log \frac{p_\theta(Y|X)}{p_{\theta_0}(Y|X)}\right], \end{aligned}$$

where the expectation is taken under the true density/distribution parameterised by θ_0. Since $M(\theta_0) = 0$, we need to show that, for any $\theta \neq \theta_0$, $M(\theta) < 0$. We can use Jensen's inequality to obtain an upper bound on M. By Jensen's Inequality,

$$M(\theta) = \mathbb{E}_{\theta_0}\left[\log \frac{p_\theta(Y|X)}{p_{\theta_0}(Y|X)}\right] \leq \log \mathbb{E}_{\theta_0}\left[\frac{p_\theta(Y|X)}{p_{\theta_0}(Y|X)}\right].$$

Writing out the expectation in integral, we can see that

$$\begin{aligned} \mathbb{E}_{\theta_0}\left[\frac{p_\theta(Y|X)}{p_{\theta_0}(Y|X)}\right] &= \int\int \frac{p_\theta(y|x)}{p_{\theta_0}(y|x)} \underbrace{p_{\theta_0}(y|x)\, d\mu(y)}_{=dP_{\theta_0}(y|x)} dP_X(x) \\ &= \int \underbrace{\int p_\theta(y|x)\, d\mu(y)}_{=dP_\theta(y|x)} dP_X(x) \\ &= \int \underbrace{\int 1 dP_\theta(y|x)}_{=1} dP_X(x) \\ &= 1. \end{aligned} \tag{4.1}$$

Thus, we have

$$M(\theta) \leq \log(1) = 0. \tag{4.2}$$

Since log is a strictly concave function, above inequality holds with equality if and only if $p_\theta(Y|X)/p_{\theta_0}(Y|X)$ is a constant.[27] In other words, the inequality above is strict unless

$$\mathbb{P}\left(\frac{p_\theta(Y|X)}{p_{\theta_0}(Y|X)} = c\right) = 1$$

[27] To see this, note that $\mathbb{E}[\log(c)] = \log(\mathbb{E}[c])$ if and only if c is a constant.

for some constant c. We want to use our assumption. Note that

$$p_\theta(Y|X) \neq p_{\theta_0}(Y|X) \Rightarrow \frac{p_\theta(Y|X)}{p_{\theta_0}(Y|X)} \neq 1$$

so that the assumption rules out the case when $c = 1$. Now suppose that $c > 1$, then it must be that

$$\mathbb{E}_{\theta_0}\left[\frac{p_\theta(Y|X)}{p_{\theta_0}(Y|X)}\right] > 1,$$

which contradicts (4.1). Finally, suppose $c < 1$, then it must be that

$$\begin{aligned}\mathbb{E}_{\theta_0}\left[\frac{p_\theta(Y|X)}{p_{\theta_0}(Y|X)}\right] < 1 \Rightarrow \log \mathbb{E}_{\theta_0}\left[\frac{p_\theta(Y|X)}{p_{\theta_0}(Y|X)}\right] < 0 \\ \Rightarrow M(\theta) < 0\end{aligned}$$

and so we are fine. Hence, we conclude that (4.2) must hold with strict inequality; i.e. θ_0 is the unique maximiser. ■

Note that, if there is no X (i.e. unconditional ML case), the condition becomes:

$$\mathbb{P}(p_\theta(Y) \neq p_{\theta_0}(Y)) > 0, \ \forall \theta \neq \theta_0.$$

Lemma 4.2. *(Consistency of ML estimators) Let $\hat{\theta}_n$ be such that:*

(i) (near maximiser) $L_n\left(\hat{\theta}_n\right) \geq L_n(\theta_0) - o_P(1)$*;*

(ii) (uniform convergence) $\sup_{\theta\in\Theta} |L_n(\theta) - L(\theta)| \xrightarrow{\text{p}} 0$*;*

(iii) (well-separatedness condition) $\sup_{\theta\in\Theta\setminus B_\delta(\theta_0)} L(\theta) < L(\theta_0)$ *for any* $\delta > 0$.

Then, $\hat{\theta}_n \xrightarrow{\text{p}} \theta_0$.

Proof. We want to show that, for any $\varepsilon > 0$, $\mathbb{P}(|\hat{\theta}_n - \theta_0| > \varepsilon) \to 0$. We do this in two steps. First, we show that

$$\left|L(\theta_0) - L\left(\hat{\theta}_n\right)\right| \xrightarrow{\text{p}} 0.$$

Then, we show that if $|\hat{\theta}_n - \theta_0| > \varepsilon$, then it implies that $L(\theta_0) - L(\hat{\theta}_n) > \eta/2 > 0$;[28] i.e. $\mathbb{P}(|\hat{\theta}_n - \theta_0| > \varepsilon) \leq \mathbb{P}(L(\theta_0) - L(\hat{\theta}_n) > \eta/2)$. Since we know the right-hand side converges in probability to zero, we will then have the desired result.

Step 1: From (i), we have $L_n(\hat{\theta}_n) \geq L_n(\theta_0) - o_P(1)$. Rearranging gives

$$L_n(\theta_0) \leq L_n\left(\hat{\theta}_n\right) + o_P(1).$$

From (ii), we know that $L_n(\theta_0)$ converges uniformly in probability to $L(\theta_0)$ so that, for sufficiently large n,

$$L(\theta_0) \leq L_n\left(\hat{\theta}_n\right) + o_P(1).$$

Given the expression we want to show, we need to introduce $L(\hat{\theta}_0)$. So subtract this term from both sides:

$$L(\theta_0) - L\left(\hat{\theta}_n\right) \leq L_n\left(\hat{\theta}_n\right) - L\left(\hat{\theta}_n\right) + o_P(1).$$

[28] Notice that $L(\theta_0)$ comes first (which ensures that the term is nonnegative since $L(\theta_0)$ is the maximum likelihood).

Since we know that $L(\theta_0)$ is the maximiser, the left-hand side is nonnegative. Hence, we can take the absolute value of both sides:

$$\begin{aligned}\left|L(\theta_0) - L\left(\hat{\theta}_n\right)\right| &\leq \left|L_n\left(\hat{\theta}_n\right) - L\left(\hat{\theta}_n\right) + o_P(1)\right| \\ &\leq \left|L_n\left(\hat{\theta}_n\right) - L\left(\hat{\theta}_n\right)\right| + |o_P(1)|,\end{aligned}$$

where the second inequality follows from the Triangle Inequality. Since $\hat{\theta}_n \in \Theta$, by definition,

$$\left|L_n\left(\hat{\theta}_n\right) - L\left(\hat{\theta}_n\right)\right| \leq \sup_{\theta\in\Theta} |L_n(\theta) - L(\theta)|.$$

Hence, we can write,

$$\left|L(\theta_0) - L\left(\hat{\theta}_n\right)\right| \leq \sup_{\theta\in\Theta} |L_n(\theta) - L(\theta)| + |o_P(1)|.$$

From (ii), we know that $\sup_{\theta\in\Theta} |L_n(\theta) - L(\theta)| \xrightarrow{\text{p}} 0$, we therefore obtain that

$$\left|L(\theta_0) - L\left(\hat{\theta}_n\right)\right| \xrightarrow{\text{p}} 0. \tag{4.3}$$

Step 2: We want now to show that if $|\hat{\theta}_n - \theta_0| > \varepsilon$, then it implies that $L(\theta_0) - L(\hat{\theta}_n) > \eta > 0$. So suppose that $|\hat{\theta}_n - \theta_0| > \varepsilon$, then

$$\hat{\theta}_n \in \Theta \backslash B_\varepsilon(\theta_0),$$

and by the definition of supremum,

$$\Rightarrow L\left(\hat{\theta}_n\right) \leq \sup_{\theta\in\Theta\backslash B_\varepsilon(\theta_0)} L(\theta) \tag{4.4}$$

We now want to introduce $L(\theta_0)$, so subtract from both sides, and rearrange to obtain

$$L(\theta_0) - L\left(\hat{\theta}_n\right) \geq L(\theta_0) - \sup_{\theta\in\Theta\backslash B_\varepsilon(\theta_0)} L(\theta).$$

From (iii), we realise that the right-hand side is strictly positive. Define

$$\eta \coloneqq L(\theta_0) - \sup_{\theta\in\Theta\backslash B_\varepsilon(\theta_0)} L(\theta) > 0.$$

Thus, we have

$$L(\theta_0) - L\left(\hat{\theta}_n\right) > \frac{\eta}{2} > 0.$$

(division by two is to ensure that the inequality is strict). We have therefore shown that $|\hat{\theta}_n - \theta_0| > \varepsilon$ implies $L(\theta_0) - L(\hat{\theta}_n) > \eta/2$. In other words, the event $|\hat{\theta}_n - \theta_0| > \varepsilon$ is contained in the event $L(\theta_0) - L(\hat{\theta}_n) > \eta/2$. It follows then that

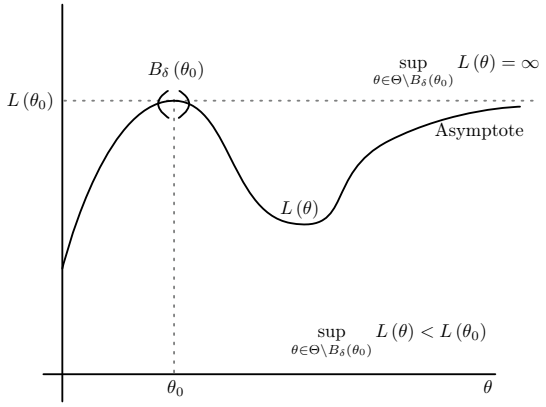
$$\begin{aligned}\mathbb{P}\left(\left|\hat{\theta}_n - \theta_0\right| > \varepsilon\right) &\leq \mathbb{P}\left(L(\theta_0) - L\left(\hat{\theta}_n\right) > \frac{\eta}{2}\right) \\ &= \mathbb{P}\left(\left|L(\theta_0) - L\left(\hat{\theta}_n\right)\right| > \frac{\eta}{2}\right),\end{aligned}$$

where we used the fact that $L(\theta_0) - L(\hat{\theta}_n)$ is nonnegative. But (4.3) implies that the right-hand sides converges to zero; i.e.

$$\mathbb{P}\left(\left|\hat{\theta}_n - \theta_0\right| > \varepsilon\right) \to 0,$$

and we are done. ■

What does well-separatedness mean? Consider $L(\theta)$ as drawn in the figure below, where $L(\theta)$ asymptotes to $L(\theta_0)$ as $\theta \to \infty$. The condition says that, if we were to "carve out" a neighbourhood around θ_0, $B_\delta(\theta_0)$, from Θ, and take the supremum of $L(\theta)$ in that set, then the likelihood at that supremum must be lower than $L(\theta_0)$. Notice also that uniqueness is guaranteed by the well-separatedness condition.



We now wish to provide some sufficient conditions for the assumptions in Lemma 4.2.

Lemma 4.3. *(A sufficient condition for near maximiser):* $L_n\left(\hat{\theta}_n\right) \geq \sup_{\theta\in\Theta} L_n(\theta) - o_P(1)$.

Lemma 4.4. *(Sufficient conditions for well-separatedness). Suppose $L(\theta)$ is continuous, Θ is compact, and $L(\theta)$ is uniquely maximised over Θ at $\theta = \theta_0$, then, for all $\delta > 0$,*

$$\sup_{\theta\in\Theta\setminus B_\delta(\theta_0)} L(\theta) < L(\theta_0).$$

Lemma 4.5. *(Sufficient conditions for uniform convergence). Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} P$ on $\mathbb{R}^d$. Let S be the support of P, $\Theta \subseteq \mathbb{R}^k$ be compact, $f : S \times \Theta \to \mathbb{R}$ such that $f(x,\theta)$ is continuous in θ for each x. Suppose there exists a dominating function $F : S \to \mathbb{R}$ such that $|f(x,\theta)| \leq F(x)$ for all x and θ, and $\mathbb{E}[F(X_i)] < \infty$. Then,*

$$\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n} f(x,\theta) - \mathbb{E}[f(x,\theta)]\right| \xrightarrow{\text{p}} 0.$$

Proof. (Lemma 4.3) Immediate from the fact that $\sup_{\theta\in\Theta} L_n(\theta) \geq L_n(\theta_0)$. ■

Proof. (Lemma 4.4) Compactness and continuity of $L(\theta)$ guarantees the existence of $\theta^* = \sup_{\theta\in\Theta\setminus B_\delta(\theta_0)} L(\theta) \in \Theta\setminus B_\delta(\theta_0)$. By construction, $\theta^* \neq \theta_0$. Since $L(\theta)$ is unique maximised at θ_0 by assumption, it follows that $L(\theta^*) < L(\theta_0)$. ■

Proof. (Lemma 4.5) Admitted. ■

Example 4.7. *(Probit).* Recall that the Probit model is given by $Y_i \in \{0,1\}$ and $X_i \in \mathbb{R}^k$, where

$$p_\theta(Y|X) = \begin{cases} \Phi(X'\theta) & \text{if } Y = 1 \\ 1 - \Phi(X'\theta) & \text{if } Y = 0 \end{cases}.$$

Equivalently,

$$\log p_\theta(Y|X) = Y \log \Phi(X'\theta) + (1-Y)\log(1 - \Phi(X'\theta)).$$

Suppose that Θ is compact, there is no perfect collinearity in X_i and the support of X_i is bounded. To check if $L(\theta)$ is uniquely maximised at θ_0, we can check that

$$\mathbb{P}\left(p_\theta\left(Y_i|X_i\right) \neq p_{\theta_0}\left(Y_i|X_i\right)\right) > 0, \ \forall \theta \neq \theta_0.$$

To see that this holds, note that

$$\begin{aligned}\mathbb{P}\left(p_\theta\left(Y_i|X_i\right) \neq p_{\theta_0}\left(Y_i|X_i\right)\right) &= \mathbb{P}\left(\Phi\left(X_i'\theta\right) \neq \Phi\left(X_i'\theta_0\right)\right) \\ [\Phi \text{ is strictly increasing}] &= \mathbb{P}\left(X_i'\theta \neq X_i'\theta_0\right) \\ [\text{no perfect collinearity in } X_i] &= \mathbb{P}\left(X_i'\left(\theta - \theta_0\right) \neq 0\right) > 0.\end{aligned}$$

(Recall Definition 3.1: X_i is perfectly collinear if there is $c \neq 0$, $c \in \mathbb{R}^k$, such that $\mathbb{P}\left(c'X_i = 0\right) = 1$. Hence, there is no perfect collinearity in X_i, then $\mathbb{P}\left(c'X_i = 0\right) < 1$, which, in turn, implies that $\mathbb{P}\left(c'X_i \neq 0\right) > 0$). In addition, since $\log p_\theta$ is continuous for every θ and x, $L(\theta)$ is continuous. Finally, since Θ is compact, we realise that the sufficient conditions for well-separateness are satisfied.

To verify uniform convergence, we use Lemma 4.5 while setting $f\left(\theta, Y, X\right) = \log p_\theta\left(Y|X\right)$. Note that f is continuous in θ for each y, x and Θ is compact. Consider

$$\left|\log p_\theta\left(Y|X\right)\right| = \left|Y \log \Phi\left(X'\theta\right) + \left(1 - Y\right) \log\left(1 - \Phi\left(X'\theta\right)\right)\right|.$$

We need to find the bound so taking supremum of both sides yields

$$\sup_{y,x,\theta} \left|\log p_\theta\left(Y|X\right)\right| = \sup_{y,x,\theta} \left|Y \log \Phi\left(X'\theta\right) + \left(1 - Y\right) \log\left(1 - \Phi\left(X'\theta\right)\right)\right|.$$

Since Y takes values 0 or 1, we can write

$$\sup_{Y,X,\theta} \left|\log p_\theta\left(Y|X\right)\right| = \max\left\{\sup_{X,\theta} \left|\log \Phi\left(X'\theta\right)\right|, \sup_{X,\theta} \left|\left(1 - \Phi\left(X'\theta\right)\right)\right|\right\}.$$

Since X_i's support is bounded and $\Theta \in \theta$ is a compact set, there exists a bound; i.e.

$$\sup_{Y,X,\theta} \left|\log p_\theta\left(Y|X\right)\right| \leq M < \infty.$$

Setting the constant M as the dominating function F, we satisfy the requirement of Lemma 4.5.

Finally, we obtain that

$$L_n\left(\theta\right) = \frac{1}{n}\sum_{i=1}^{n} \log p_\theta\left(Y_i|X_i\right) \overset{\text{uniformly p}}{\to} L\left(\theta\right) = \mathbb{E}\left[\log p_\theta\left(Y_i|X_i\right)\right].$$

Example 4.8. (Example in which well-separatedness is satisfied yet Lemmas 4.4 and 4.5 do not apply). Recall that when $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\to} P = \text{Unif}\left(0, \theta_0\right)$, $\theta_0 \in [0, \infty)$, the ML estimator is given by

$$\hat{\theta}_n = \max_{1 \leq i \leq n} X_i.$$

Notice that we cannot verify using Lemmas 4.4 and 4.5 since $L(\theta)$ is not continuous and Θ is not compact. However, we can nevertheless verify consistency directly. For any $\varepsilon > 0$,

$$\begin{aligned}\mathbb{P}\left(\left|\hat{\theta}_n - \theta_0\right| > \varepsilon\right) &= \mathbb{P}\left(\theta_0 - \max\left\{X_1, X_2, \ldots, X_n\right\} > \varepsilon\right) \\ &= \mathbb{P}\left(\max\left\{X_1, X_2, \ldots, X_n\right\} < \theta_0 - \varepsilon\right) \\ &= \mathbb{P}\left(X_i < \theta_0 - \varepsilon\right)^n \\ &= \left(\frac{\theta_0 - \varepsilon}{\theta_0}\right)^n \to 0.\end{aligned}$$

Hence, $\hat{\theta}_n$ is consistent.

4.3.2 Misspecification

What if $Y|X$ is not equal to P_θ for any θ? We can still expect $\hat{\theta}_n$ to converge in probability to the maximiser of $L(\theta) = \mathbb{E}\left[\log p_\theta\left(Y|X\right)\right]$, but what is that?

Let $f\left(y|x\right)$ be the true density of $Y|X$. Note that

$$\begin{aligned}
\underset{\theta\in\Theta}{\arg\max}\ L\left(\theta\right) &\equiv \underset{\theta\in\Theta}{\arg\max}\ \mathbb{E}_{\theta_0}\left[\log p_\theta\left(Y|X\right)\right] \\
&\equiv -\underset{\theta\in\Theta}{\arg\min}\ \mathbb{E}_{\theta_0}\left[\log p_\theta\left(Y|X\right)\right] \\
&\equiv \underset{\theta\in\Theta}{\arg\min}\ -\mathbb{E}_{\theta_0}\left[\log p_\theta\left(Y|X\right)\right] + \mathbb{E}_{\theta_0}\left[\log f\left(Y|X\right)\right] \\
&\equiv \underset{\theta\in\Theta}{\arg\min}\ \mathbb{E}_{\theta_0}\left[\log f\left(Y|X\right) - \log p_\theta\left(Y|X\right)\right] \\
&\equiv \underset{\theta\in\Theta}{\arg\min}\ \mathbb{E}_{\theta_0}\left[\log \frac{f\left(Y|X\right)}{p_\theta\left(Y|X\right)}\right].
\end{aligned}$$

where we adding, $\mathbb{E}_{\theta_0}[\log f(Y|X)]$, which is a constant with respect to θ does not alter the minimiser. The transformed objective function is called the *Kullback-Leibler divergence* between f and p_θ. Note that the divergence is zero if $f = p_\theta$. We realise that, when we misspecify P_θ, the ML estimator minimises the divergence between f and p_θ.

Consider the negative of the divergence,

$$-\mathbb{E}_{\theta_0}\left[\log \frac{f\left(Y|X\right)}{p_\theta\left(Y|X\right)}\right] = \mathbb{E}_{\theta_0}\left[\log \frac{p_\theta\left(Y|X\right)}{f\left(Y|X\right)}\right].$$

By Jensen's inequality,

$$\mathbb{E}_{\theta_0}\left[\log \frac{p_\theta\left(Y|X\right)}{f\left(Y|X\right)}\right] \leq \log \mathbb{E}_{\theta_0}\left[\frac{p_\theta\left(Y|X\right)}{f\left(Y|X\right)}\right].$$

Consider

$$\begin{aligned}
\mathbb{E}_{\theta_0}\left[\frac{p_\theta\left(Y|X\right)}{f\left(Y|X\right)}\right] &= \int\int \frac{p_\theta\left(y|x\right)}{f\left(y|x\right)} \underbrace{f\left(y|x\right) d\mu\left(y\right)}_{=dF(y|x)} dP_X\left(x\right) \\
&= \int \underbrace{\int p_\theta\left(y|x\right) d\mu\left(y\right)}_{=dP_\theta(y|x)} dP_X\left(x\right) \\
&= \int \underbrace{\int 1 dP_\theta\left(y|x\right)}_{=1} dP_X\left(x\right) \\
&= 1,
\end{aligned}$$

where $F\left(y|x\right)$ denotes CDF of $f\left(y|x\right)$. Hence,

$$\mathbb{E}_{\theta_0}\left[\log \frac{p_\theta\left(Y|X\right)}{f\left(Y|X\right)}\right] \leq 0 \Rightarrow -\mathbb{E}_{\theta_0}\left[\log \frac{p_\theta\left(Y|X\right)}{f\left(Y|X\right)}\right] \geq 0;$$

i.e. Kullback-Leibler divergence is always nonnegative. Thus, we can think of it like a "distance" (though it isn't technically because the it is not symmetric).

4.3.3 Limiting distribution of $\hat{\theta}_n$

We want to provide a sufficient condition under which

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{\mathrm{d}} N\left(0, \Omega\right).$$

Proposition 4.1. *Suppose* $\hat{\theta}_n \in \arg\max_{\theta\in\Theta} L_n(\theta)$, $\theta_0 \in \arg\max_{\theta\in\Theta} L(\theta)$, *and* Θ *is compact. Assume that* $\hat{\theta}_n \xrightarrow{\text{p}} \theta_0 \in \text{int}(\Theta)$, *and that* $\log p_\theta(y|x)$ *is twice continuously differentiable in* θ *for all* (y,x). *Suppose that*

$$A \coloneqq \mathbb{E}\left[D_\theta \log p_{\theta_0}(Y_i|X_i)\, D_{\theta'} \log p_{\theta_0}(Y_i|X_i)\right],$$
$$B \coloneqq \mathbb{E}\left[D^2_{\theta,\theta'} \log p_{\theta_0}(Y_i|X_i)\right]$$

exist and that B *is invertible. Assume that there exists two dominating functions* $M_1(y,x)$ *and* $M_2(y,x)$ *with* $\mathbb{E}[M_1(Y_i,X_i)] < \infty$ *and* $\mathbb{E}[M_2(Y_i,X_i)] < \infty$ *such that:*

(i) for some $\delta > 0$, $\left|D_{\theta_j} \log p_\theta(Y_i|X_i)\right| \leq M_1(y,x)$ *for all* $1 \leq j \leq k$, $\theta \in B_\delta(\theta_0)$;

(ii) for some $\delta > 0$, $\left|D_{\theta_j,\theta_l} \log p_\theta(Y_i|X_i)\right| \leq M_2(y,x)$ *for all* $1 \leq j,l \leq k$, $\theta \in B_\delta(\theta_0)$.

Then,

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{\text{d}} N(0,\Omega),$$

where $\Omega = B^{-1}AB^{-1}$.

Proposition 4.2. *Given the condition as given in Proposition 4.1,*

$$B = -A.$$

Hence,

$$\begin{aligned}\Omega &= B^{-1}AB^{-1} = -B^{-1}BB^{-1} \\ &= (-B)^{-1} = A^{-1}.\end{aligned}$$

Definition 4.4. *(Information matrix)* The term

$$-B = -\mathbb{E}_{\theta_0}\left[D^2_{\theta,\theta'} \log p_\theta(y|x)\right]$$

is called the *Fisher information* matrix or simply the *information matrix.*

Proof. (Proposition 4.1) Since $\theta_0 \in \text{int}(\Theta)$, and θ_0 maximises $L(\theta)$, we know that the first-order condition holds such that

$$0 = D_\theta L(\theta_0) = D_\theta \mathbb{E}\left[\log p_{\theta_0}(Y_i|X_i)\right].$$

Assumption (i) allows us to interchange D_θ and $\mathbb{E}$ so that

$$0 = D_\theta L(\theta_0) = \mathbb{E}\left[D_\theta \log p_{\theta_0}(Y_i|X_i)\right]. \tag{4.5}$$

Since $\hat{\theta}_n \xrightarrow{\text{p}} \theta_0 \in \text{int}(\Theta)$, $\hat{\theta}_n \in \text{int}(\Theta)$ with probability approaching one. When that happens, we know that

$$0 = D_\theta L_n\left(\hat{\theta}_n\right).$$

Note that

$$D_\theta L_n\left(\hat{\theta}_n\right) = \frac{1}{n}\sum_{i=1}^n D_\theta \log p_{\hat{\theta}_n}(Y_i|X_i) = \begin{bmatrix} \frac{\partial \log p_{\hat{\theta}_n}(Y_i|X_i)}{\partial \theta^1} \\ \frac{\partial \log p_{\hat{\theta}_n}(Y_i|X_i)}{\partial \theta^2} \\ \vdots \\ \frac{\partial \log p_{\hat{\theta}_n}(Y_i|X_i)}{\partial \theta^k} \end{bmatrix}_{k\times 1}.$$

By the Mean-Value Theorem,[29] we have

$$0 = D_\theta L_n\left(\hat{\theta}_n\right) = D_\theta L_n\left(\theta_0\right) + H_n\left(\hat{\theta}_n - \theta_0\right)$$

where H_n is a $k \times k$ matrix whose the jth row is given by the jth row of $D^2_{\theta,\theta'} L_n(\theta)$ evaluated at $\theta = \tilde{\theta}_{n,j}$ where $\tilde{\theta}_{n,j}$ lies on the line segment between $\hat{\theta}_n$ and θ_0; i.e.

$$H_n = \begin{bmatrix} \frac{\partial^2 L_n\left(\tilde{\theta}_{n,1}\right)}{\partial \theta_1^2} & \frac{\partial^2 L_n\left(\tilde{\theta}_{n,1}\right)}{\partial \theta_1 \partial \theta_2} & \cdots & \frac{\partial^2 L_n\left(\tilde{\theta}_{n,1}\right)}{\partial \theta_1 \partial \theta_k} \\ \frac{\partial^2 L_n\left(\tilde{\theta}_{n,2}\right)}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 L_n\left(\tilde{\theta}_{n,2}\right)}{\partial \theta_2^2} & \cdots & \frac{\partial^2 L_n\left(\tilde{\theta}_{n,2}\right)}{\partial \theta_2 \partial \theta_k} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 L_n\left(\tilde{\theta}_{n,k}\right)}{\partial \theta_k \partial \theta_1} & \frac{\partial^2 L_n\left(\tilde{\theta}_{n,k}\right)}{\partial \theta_k \partial \theta_2} & \cdots & \frac{\partial^2 L_n\left(\tilde{\theta}_{n,k}\right)}{\partial \theta_k^2} \end{bmatrix}_{k \times k}$$

where $\tilde{\theta}_{n,j} = \hat{\theta}_n + t\theta_0$ for some $t \in (0,1)$ for all j. Note, in particular, that $\tilde{\theta}_{n,j}$ does not mean the jth element of $\tilde{\theta}_n$.

Rearranging $0 = D_\theta L_n(\theta_0) + H_n(\hat{\theta}_n - \theta_0)$ and multiplying both sides by $\sqrt{n}$,

$$\begin{aligned} -H_n\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) &= \sqrt{n} D_\theta L_n\left(\theta_0\right) \\ &= \sqrt{n}\frac{1}{n}\sum_{i=1}^{n} D_\theta \log p_{\theta_0}\left(Y_i|X_i\right). \end{aligned}$$

By the Central Limit Theorem.

$$-H_n\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) = \sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n} D_\theta \log p_{\theta_0}\left(Y_i|X_i\right) - \underbrace{\mathbb{E}\left[D_\theta \log p_{\theta_0}\left(Y|X\right)\right]}_{=0 \because \text{FOC}}\right] \xrightarrow{\text{d}} N\left(0, A\right), \quad (4.6)$$

where

$$A \coloneqq \mathbb{E}\left[\left(D_\theta \log p_{\theta_0}\left(Y_i|X_i\right)\right)\left(D_\theta \log p_{\theta_0}\left(Y_i|X_i\right)\right)'\right].$$

To show that $H_n \xrightarrow{\text{p}} B$, it suffices to show that $D^2_{\theta,\theta'} L_n(\tilde{\theta}_{n,j}) \xrightarrow{\text{p}} B$ for every j. If $\tilde{\theta}_{n,j} \in B_\delta(\theta_0)$, we have

$$\left|D^2_{\theta,\theta'} L_n\left(\tilde{\theta}_{n,j}\right) - \mathbb{E}\left[D^2_{\theta,\theta'} \log p_{\tilde{\theta}_{n,j}}\left(Y_i|X_i\right)\right]\right| \leq \sup_{\theta \in B_\delta(\theta_0)} \left|D^2_{\theta,\theta'} L_n\left(\theta\right) - \mathbb{E}\left[D^2_{\theta,\theta'} \log p_\theta\left(Y_i|X_i\right)\right]\right|, \ \forall j.$$

Hence, for any $\varepsilon > 0$,

Why do we add $\mathbb{P}\left(\tilde{\theta}_{n,j} \notin B_\delta\left(\theta_0\right)\right)$?

$$\begin{aligned} &\mathbb{P}\left(\left|D^2_{\theta,\theta'} L_n\left(\tilde{\theta}_{n,j}\right) - \mathbb{E}\left[D^2_{\theta,\theta'} \log p_{\tilde{\theta}_{n,j}}\left(Y_i|X_i\right)\right]\right| > \varepsilon\right) \\ &\leq \mathbb{P}\left(\sup_{\theta \in B_\delta(\theta_0)} \left|D^2_{\theta,\theta'} L_n\left(\theta\right) - \mathbb{E}\left[D^2_{\theta,\theta'} \log p_\theta\left(Y_i|X_i\right)\right]\right| > \varepsilon\right), \ \forall j \\ &\leq \mathbb{P}\left(\sup_{\theta \in B_\delta(\theta_0)} \left|D^2_{\theta,\theta'} L_n\left(\theta\right) - \mathbb{E}\left[D^2_{\theta,\theta'} \log p_\theta\left(Y_i|X_i\right)\right]\right| > \varepsilon\right) + \mathbb{P}\left(\tilde{\theta}_{n,j} \notin B_\delta\left(\theta_0\right)\right), \ \forall j \end{aligned}$$

By assumption (ii), we know that there exists a dominating function $M_2(y,x)$ such that $|D_{\theta_j,\theta_l} \log p_\theta(y|x)| \leq M_2(y,x)$ for all $1 \leq j,l \leq k$, $\theta \in B_\delta(\theta_0)$, and $\mathbb{E}[M_2(y,x)] < \infty$. We also have, by assumption,

[29]

Theorem 4.1. *(Mean Value Theorem) Let $f[a,b] \to \mathbb{R}$ be a continuous function on the closed interval $[a,b]$ and differentiable in the open interval (a,b). Then, there exists $c \in (a,b)$ such that*

$$f(b) = f(a) + f'(c)(b-a).$$

The Mean Value Theorem applies to real-valued functions—this is why we apply the theorem component wise here.

$\log p_\theta(y,x)$ is twice continuously differentiable in θ for every (y,x) so that $D^2_{\theta,\theta'}\log p_\theta(y|x)$ is continuous in θ for every (y,x), and Θ is compact. We therefore satisfy the sufficient conditions for uniform convergence (Lemma 4.5) so that:[30]

$$\mathbb{P}\left(\sup_{\theta\in B_\delta(\theta_0)}\left|D^2_{\theta,\theta'}L_n(\theta)-\mathbb{E}\left[D^2_{\theta,\theta'}\log p_\theta(Y_i|X_i)\right]\right|>\varepsilon\right)\to 0,\ \forall j.$$

Since $\tilde{\theta}_{n,j}$ lies on the line segment between $\hat{\theta}_n$ and θ_0, $\hat{\theta}_n \xrightarrow{\mathrm{p}} \theta_0$ implies that $\tilde{\theta}_{n,j} \xrightarrow{\mathrm{p}} \theta_0$ for all j. Thus, $\mathbb{P}(\tilde{\theta}_{n,j}\notin B_\delta(\theta_0))\to 0$, for all j. Together, we obtain that

$$\mathbb{P}\left(\left|D^2_{\theta,\theta'}L_n\left(\tilde{\theta}_{n,j}\right)-\mathbb{E}\left[D^2_{\theta,\theta'}\log p_{\tilde{\theta}_{n,j}}(Y_i|X_i)\right]\right|>\varepsilon\right)\to 0,\forall j;$$

i.e.

$$D^2_{\theta,\theta'}L_n\left(\tilde{\theta}_n\right)\xrightarrow{\mathrm{p}}\mathbb{E}\left[D^2_{\theta,\theta'}\log p_{\tilde{\theta}_{n,j}}(Y_i|X_i)\right]\xrightarrow{\mathrm{p}}\mathbb{E}\left[D^2_{\theta,\theta'}\log p_{\theta_0}(Y_i|X_i)\right]=:B,$$

where we use again the fact that $\tilde{\theta}_{n,j}\xrightarrow{\mathrm{p}}\theta_0$ for all j.

Thus, we have that, by Slutsky's Lemma,

$$\begin{aligned}-H_n\sqrt{n}\left(\hat{\theta}_n-\theta_0\right)&\xrightarrow{\mathrm{d}}-BN(0,\Omega)\\&\overset{\mathrm{d}}{=}N(0,B\Omega B')\end{aligned}$$

where$\sqrt{n}\left(\hat{\theta}_n-\theta_0\right)\xrightarrow{\mathrm{d}}N(0,\Omega)$. However, we now from (4.6) that $-H_n\sqrt{n}(\hat{\theta}_n-\theta_0)\overset{\mathrm{d}}{\sim}N(0,A)$, hence,

$$B\Omega B'=A.$$

Since B is symmetric,

$$\begin{aligned}B\Omega B'&=B\Omega B\\\Rightarrow B^{-1}B\Omega BB^{-1}&=B^{-1}AB^{-1}\\\Rightarrow\Omega&=B^{-1}AB^{-1},\end{aligned}$$

where we used the fact that B is invertible by assumption. That is,

$$\sqrt{n}\left(\hat{\theta}_n-\theta_0\right)\xrightarrow{\mathrm{d}}N\left(0,B^{-1}AB^{-1}\right).$$ ■

Proof. (Proposition 4.2) Recall (4.5):

$$0=D_\theta L(\theta_0)=\mathbb{E}_{\theta_0}\left[D_\theta\log p_{\theta_0}(Y_i|X_i)\right].$$

Writing the last expression in terms of integrals,

$$\begin{aligned}0&=\mathbb{E}_{\theta_0}\left[D_\theta\log p_{\theta_0}(Y_i|X_i)\right]\\&=\int\int D_\theta\log p_{\theta_0}(y|x)\,p_{\theta_0}(y|x)\,d\mu(y)\,dP_X(x).\end{aligned}$$

[30]Note that $D^2_{\theta,\theta'}L_n(\theta)=\frac{1}{n}\sum_{i=1}^n D^2_{\theta,\theta'}\log p_\theta(Y_i|X_i)$ so that $D^2_{\theta,\theta'}\log p_\theta(Y_i|X_i)$ corresponds to $f(x,\theta)$ in the Lemma. $M_2(y,x)$ corresponds $F(x)$.

Suppose we evaluate the derivative at some $\theta \in \Theta$, $\theta \neq \theta_0$, then

$$\begin{aligned}
&\int\int D_\theta \log p_\theta(y|x)\, p_\theta(y|x)\, d\mu(y)\, dP_X(x) \\
&= \int\int D_\theta p_\theta(y|x) \frac{1}{p_\theta(y|x)} p_\theta(y|x)\, d\mu(y)\, dP_X(x) \\
&= \int\int D_\theta p_\theta(y|x)\, d\mu(y)\, dP_X(x) \\
&= \int D_\theta \underbrace{\int p_\theta(y|x)\, d\mu(y)}_{=1} dP_X(x) \\
&= 0.
\end{aligned}$$

That is, the first-order condition is satisfied for any $\theta \in \Theta$. This allow to drop the subscript 0:

$$0 = \int\int D_\theta \log p_\theta(y|x)\, p_\theta(y|x)\, d\mu(y)\, dP_X(x).$$

Since this is an identity, we can differentiate the expression with respect to θ to obtain (using chain rule):

$$\begin{aligned}
0 = &\int\int D^2_{\theta,\theta'} \log p_\theta(y|x)\, p_\theta(y|x)\, d\mu(y)\, dP_X(x) \\
&+ \int\int D_\theta \log p_\theta(y|x)\, D_{\theta'} p_\theta(y|x)\, d\mu(y)\, dP_X(x).
\end{aligned}$$

Consider the first term,

$$\begin{aligned}
\int\int D^2_{\theta,\theta'} \log p_\theta(y|x)\, p_\theta(y|x)\, d\mu(y)\, dP_X(x) &= \int\int D^2_{\theta,\theta'} \log p_\theta(y|x)\, dP_\theta(y|x)\, dP_X(x) \\
&= \int\int D^2_{\theta,\theta'} \log p_{\theta_0}(y|x)\, dP_{\theta_0}(y|x)\, dP_X(x) \\
&= \mathbb{E}_{\theta_0}\left[D^2_{\theta,\theta'} \log p_\theta(y|x)\right] \\
&= B.
\end{aligned}$$

Now, consider the second term. Notice that

$$D_{\theta'} \log p_\theta(y|x) = \frac{D_{\theta'} p_\theta(y|x)}{p_\theta(y|x)}.$$

Thus, multiplying by $p_\theta(y|x)/p_\theta(y|x)$ inside the integral gives,

$$\begin{aligned}
&\int\int D_\theta \log p_\theta(y|x) \frac{D_{\theta'} p_\theta(y|x)}{p_\theta(y|x)} p_\theta(y|x)\, d\mu(y)\, dP_X(x) \\
=&\int\int D_\theta \log p_\theta(y|x)\, D_{\theta'} \log p_\theta(y|x)\, p_\theta(y|x)\, d\mu(y)\, dP_X(x) \\
=&\int\int D_\theta \log p_\theta(y|x)\, D_{\theta'} \log p_\theta(y|x)\, dP_\theta(y|x)\, dP_X(x) \\
=&\int\int D_\theta \log p_{\theta_0}(y|x)\, D_{\theta'} \log p_{\theta_0}(y|x)\, dP_{\theta_0}(y|x)\, dP_X(x) \\
=&\mathbb{E}\left[D_\theta \log p_{\theta_0}(Y_i|X_i)\, D_{\theta'} \log p_{\theta_0}(Y_i|X_i)\right] \\
=&A.
\end{aligned}$$

Hence, we now have that

$$0 = B + A \Rightarrow -B = A.$$

■

Note that $\hat{\theta}_n$ may not always have a normal limit distribution. This is the case when the support depends "heavily" on the parameter.

Example 4.9. Let $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} P = \text{Unif}[0, \theta_0]$ with $\theta_0 \in [0, \infty)$. In this case, the ML estimator is given by

$$\hat{\theta}_n = \max_{1 \leq i \leq n} X_i.$$

We can show that $-n\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{\text{d}} \text{Exp}(\theta_0)$, where the CDF is given by

$$F(t) = \begin{cases} 0 & \text{if } t < 0 \\ 1 - \exp\left(-\frac{t}{\theta_0}\right) & \text{if } t \geq 0 \end{cases}.$$

To see this, first, note that

$$\mathbb{P}\left(-n\left(\hat{\theta}_n - \theta_0\right) \leq t\right) = 0 \text{ if } t < 0.$$

If $t \geq 0$, then

$$\begin{aligned} \mathbb{P}\left(-n\left(\hat{\theta}_n - \theta_0\right) \leq t\right) &= \mathbb{P}\left(\hat{\theta}_n \geq \theta_0 - \frac{t}{n}\right) \\ &= 1 - \mathbb{P}\left(\hat{\theta}_n < \theta_0 - \frac{t}{n}\right) \\ &= 1 - \mathbb{P}\left(\max_{1 \leq i \leq n} X_i < \theta_0 - \frac{t}{n}\right). \end{aligned}$$

Note that the probability on the right-hand side is the same as the probability that every X_i is smaller than $\theta_0 - \frac{t}{n}$. Since X_i's are iid,

$$\mathbb{P}\left(\max_{1 \leq i \leq n} X_i < \theta_0 - \frac{t}{n}\right) = \mathbb{P}\left(X_i < \theta_0 - \frac{t}{n}\right)^n.$$

Since X_i's are uniformly distributed, we can express the right-hand side using the CDF:

$$\mathbb{P}\left(X_i < \theta_0 - \frac{t}{n}\right)^n = \left(\frac{\theta_0 - \frac{t}{n}}{\theta_0}\right)^n = \left(1 + \frac{1}{n}\left(-\frac{t}{\theta_0}\right)\right)^n.$$

Recall the definition of exp:

$$\exp(x) = \lim_{n \to \infty} \left(1 + \frac{x}{n}\right)^n.$$

Hence, as $n \to \infty$,

$$\begin{aligned} &\mathbb{P}\left(X_i < \theta_0 - \frac{t}{n}\right)^n \to \exp\left(-\frac{t}{\theta_0}\right) \\ \Rightarrow\ &\mathbb{P}\left(-n\left(\hat{\theta}_n - \theta_0\right) \leq t\right) \to 1 - \exp\left(-\frac{t}{\theta_0}\right). \end{aligned}$$

Therefore, as $n \to \infty$,

$$-n\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{\text{d}} \text{Exp}(\theta_0).$$

4.4 Inference

Let $f : \mathbb{R}^k \to \mathbb{R}^p$ be a continuously differentiable function at θ_0. Assume also that $D_\theta(f(\theta_0))$ has linearly independent rows. We wish to test

$$\begin{aligned} H_0 &: f(\theta_0) = 0, \\ H_1 &: f(\theta_0) \neq 0. \end{aligned}$$

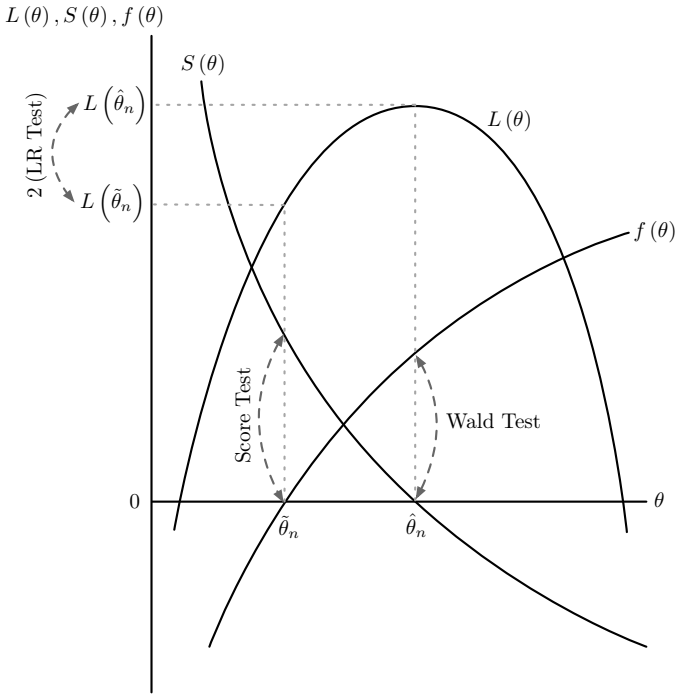
Hence, there are p number of restrictions on the parameters of the model.

Let $\tilde{\theta}_n$ be the maximiser of $L_n(\theta)$ in $\theta \in \Theta$ such that $f(\theta) = 0$. Then, $\tilde{\theta}_n$ is called the *restricted* ML estimator, which contrasts with $\hat{\theta}_n$—the *unrestricted* ML estimator. We denote the value of $L_n(\theta)$ evaluated at $\hat{\theta}_n$ and $\tilde{\theta}_n$ as $L_n(\hat{\theta}_n)$ and $L_n(\tilde{\theta}_n)$ respectively. $S(\theta) \coloneqq D_\theta \log p_\theta(Y_i|X_i)$ is called the *score* and $S_i(\theta) \coloneqq D_{\theta_i} \log p_\theta(Y_i|X_i)$ is called the *ith score*.

We introduce three tests:

- ▷ Wald Test: Compare $f(\hat{\theta}_n)$ with zero (since $f(\tilde{\theta}_n) = 0$);
- ▷ Score Test: Compare $S(\tilde{\theta}_n)$ with zero (since $S(\hat{\theta}_n) = 0$);
- ▷ Likelihood Ratio Test: Compare $L(\hat{\theta}_n)$ with $L(\tilde{\theta})$.

Figure below shows the graphical example in which θ is a scalar and $f(\cdot)$ is increasing. Note that $S(\theta)$ is decreasing in θ since $L(\theta)$ is concave.



In all these tests, we assume that

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{\mathrm{d}} N(0, \Omega),$$

where

$$\Omega = (-B)^{-1} = \left(-\mathbb{E}\left[D^2_{\theta,\theta'} \log p_{\theta_0}(Y_i|X_i)\right]\right)^{-1}$$
$$\text{or } A^{-1} = \mathbb{E}\left[D_\theta \log p_{\theta_0}(Y_i|X_i)\, D_{\theta'} \log p_{\theta_0}(Y_i|X_i)\right]^{-1}.$$

4.4.1 Wald Test

Using the Delta Method,

$$\sqrt{n}\left(f\left(\hat{\theta}_n\right) - f(\theta_0)\right) \xrightarrow{\mathrm{d}} D_\theta f(\theta_0)\, N(0, \Omega)$$
$$\stackrel{\mathrm{d}}{=} N\left(0, D_\theta f(\theta_0)\, \Omega D_\theta f(\theta_0)'\right).$$

Since we know $\Omega = (-B)^{-1} = A^{-1}$, we have a few options to estimate Ω using the analog principle:

$$\hat{\Omega} = \hat{B} = \frac{1}{n}\sum_{i=1}^{n} D^2_{\theta,\theta'} \log p_{\hat{\theta}_n}(Y_i|X_i),$$
$$\text{or } \hat{A} = \frac{1}{n}\sum_{i=1}^{n} D_\theta \log p_{\hat{\theta}_n}(Y_i|X_i)\, D_{\theta'} \log p_{\hat{\theta}_n}(Y_i|X_i),$$

so that

$$\hat{\Omega}_n = \left(-\frac{1}{n}\sum_{i=1}^{n} D^2_{\theta,\theta'} \log p_{\hat{\theta}_n}(Y_i|X_i)\right)^{-1}$$
$$\text{or } \left(\frac{1}{n}\sum_{i=1}^{n} D_\theta \log p_{\hat{\theta}_n}(Y_i|X_i)\, D_{\theta'} \log p_{\hat{\theta}_n}(Y_i|X_i)\right)^{-1}.$$

We also estimate $D_\theta f(\theta_0)$ as $D_\theta f(\hat{\theta}_n)$. Then, we have that

$$n\left(f\left(\hat{\theta}_n\right) - f(\theta_0)\right)\left(D_\theta f\left(\hat{\theta}_n\right)\hat{\Omega}_n D_\theta f\left(\hat{\theta}_n\right)'\right)^{-1}\left(f\left(\hat{\theta}_n\right) - f(\theta_0)\right)' \xrightarrow{\text{d}} \chi^2_p.$$

Since $f(\theta_0) = 0$ under the null, the test statistic is given by

$$T_n = n f\left(\hat{\theta}_n\right)\left(D_\theta f\left(\hat{\theta}_n\right)\hat{\Omega}_n D_\theta f\left(\hat{\theta}_n\right)'\right)^{-1} f\left(\hat{\theta}_n\right)'$$

and the critical value is given by the $1-\alpha$th quantile of χ^2_p distribution.

Note that the Wald Test is not invariant to how restrictions are formulated. For example, given a scalar q, $\beta/(1-\alpha) = q$ (a nonlinear restriction) and $\beta - (1-\alpha)q = 0$ (linear restriction) are equivalent restrictions but may lead to different values of the test statistic.

4.4.2 Score Test (or Lagrange Multiplier Test)

Idea: Let $\tilde{\theta}_n$ be the maximiser of $L_n(\theta)$ in $\theta \in \Theta$ such that $f(\theta) = 0$. Then, maybe, if H_0 is true, then $D_\theta L_n(\tilde{\theta}_n) \approx 0$ since $D_\theta L_n(\hat{\theta}_n) = 0$.

As in the derivation of the limit distribution of $\hat{\theta}_n$ (Proposition 4.1), we can apply the Mean Value Theorem component wise to obtain

$$D_\theta L_n\left(\tilde{\theta}_n\right) = D_\theta L_n(\theta_0) + H_n\left(\tilde{\theta}_n - \theta_0\right), \tag{4.7}$$

where H_n has jth row equal to the jth row of $D^2_{\theta,\theta'} L_n(\theta)$ evaluated at $\theta = \tilde{\theta}_{n,j}$ where $\tilde{\theta}_{n,j}$ lies in the line segment between $\tilde{\theta}_n$ and θ_0. In an analogous way to the proof in Proposition 4.1, we can show that, under the null,

$$H_n \xrightarrow{\text{p}} B = \mathbb{E}\left[\left(D^2_{\theta,\theta'} \log p_\theta(Y_i|X_i)\right)\right].$$

Now, applying the Mean Value Theorem to $f\left(\tilde{\theta}_n\right)$, we obtain

$$f\left(\tilde{\theta}_n\right) = f(\theta_0) + F_n\left(\tilde{\theta}_n - \theta_0\right), \tag{4.8}$$

where the jth row of F_n is given the jth row of $D_\theta f(\theta)$ evaluated at $\theta = \theta^*_{n,j}$ where $\theta^*_{n,j}$ lies in the line segment between $\tilde{\theta}_n$ and θ_0. As before, we can show that $F_n \xrightarrow{\text{p}} \mathbb{E}[D_\theta f(\theta_0)] = D_\theta f(\theta_0)$ (the

expectation term is vacuous here since we know f and θ_0 under the null). Moreover, under the null, $f(\theta_0) = 0$ and, by construction, $f\left(\tilde{\theta}_n\right) = 0$, so, under the null, (4.8) simplifies to

$$\begin{aligned} F_n\left(\tilde{\theta}_n - \theta_0\right) &= 0 \\ \Rightarrow F_n\sqrt{n}\left(\tilde{\theta}_n - \theta_0\right) &= 0. \end{aligned} \tag{4.9}$$

Then, multiplying (4.7) through by $H_n^{-1}\sqrt{n}$, we obtain

$$\begin{aligned} H_n^{-1}\sqrt{n}D_\theta L_n\left(\tilde{\theta}_n\right) &= H_n^{-1}\sqrt{n}D_\theta L_n(\theta_0) + H_n^{-1}H_n\sqrt{n}\left(\tilde{\theta}_n - \theta_0\right) \\ &= H_n^{-1}\sqrt{n}D_\theta L_n(\theta_0) + \sqrt{n}\left(\tilde{\theta}_n - \theta_0\right). \end{aligned}$$

Multiplying through by F_n,

$$F_nH_n^{-1}\sqrt{n}D_\theta L_n\left(\tilde{\theta}_n\right) = F_nH_n^{-1}\sqrt{n}D_\theta L_n(\theta_0) + F_n\sqrt{n}\left(\tilde{\theta}_n - \theta_0\right).$$

Using (4.9), under the null, we can rewrite above as

$$F_nH_n^{-1}\sqrt{n}D_\theta L_n\left(\tilde{\theta}_n\right) = F_nH_n^{-1}\sqrt{n}D_\theta L_n(\theta_0)$$

and notice/recall that

$$\begin{aligned} H_n^{-1} &\xrightarrow{\text{p}} B^{-1} = \mathbb{E}\left[\left(D^2_{\theta,\theta'}\log p_\theta(Y_i|X_i)\right)\right]^{-1}, \\ F_n &\xrightarrow{\text{p}} D_\theta f(\theta_0), \\ \sqrt{n}D_\theta L_n(\theta_0) &\xrightarrow{\text{d}} N(0, A), \end{aligned}$$

where the last result comes from the proof of Proposition 4.1. Thus, by Slutsky's Lemma,

$$\begin{aligned} F_nH_n^{-1}\sqrt{n}D_\theta L_n\left(\tilde{\theta}_n\right) &\xrightarrow{\text{d}} D_\theta f(\theta_0) B^{-1}N(0, A) \\ &\overset{\text{d}}{=} N\left(0, D_\theta f(\theta_0) B^{-1}AB^{-1}D_\theta f(\theta_0)'\right) \\ &\overset{\text{d}}{=} N\left(0, D_\theta f(\theta_0)\Omega D_\theta f(\theta_0)'\right). \end{aligned}$$

Since we do not in fact know F_n or H_n, we replace them with $D_\theta f\left(\tilde{\theta}_n\right)$ and $D^2_{\theta,\theta'}L_n\left(\tilde{\theta}_n\right)$ respectively, and the result above will still hold. Therefore, we can write

$$\sqrt{n}D_\theta f\left(\tilde{\theta}_n\right)\left(D^2_{\theta,\theta'}L_n\left(\tilde{\theta}_n\right)\right)^{-1}D_\theta L_n\left(\tilde{\theta}_n\right) \xrightarrow{\text{d}} N\left(0, D_\theta f(\theta_0)\Omega D_\theta f(\theta_0)'\right).$$

Define $S \coloneqq D_\theta f\left(\tilde{\theta}_n\right)\left(D^2_{\theta,\theta'}L_n\left(\tilde{\theta}_n\right)\right)^{-1}D_\theta L_n\left(\tilde{\theta}_n\right)$. We also estimate $D_\theta f(\theta_0)$ with $D_\theta f\left(\tilde{\theta}_n\right)$. Then, we can obtain that

$$\begin{aligned} & nS'\left(D_\theta f\left(\tilde{\theta}_n\right)\tilde{\Omega}_n D_\theta f\left(\tilde{\theta}_n\right)'\right)^{-1}S \\ &= nS'\left(D_\theta f\left(\tilde{\theta}_n\right)'\right)^{-1}\tilde{\Omega}_n^{-1}\left(D_\theta f\left(\tilde{\theta}_n\right)\right)^{-1}S \\ &= nS'\left(D_\theta f\left(\tilde{\theta}_n\right)'\right)^{-1}\tilde{\Omega}_n^{-1}\left(D_\theta f\left(\tilde{\theta}_n\right)\right)^{-1}D_\theta f\left(\tilde{\theta}_n\right)\left(D^2_{\theta,\theta'}L_n\left(\tilde{\theta}_n\right)\right)^{-1}D_\theta L_n\left(\tilde{\theta}_n\right) \\ &= n\left(D_\theta L_n\left(\tilde{\theta}_n\right)\right)'\left(D^2_{\theta,\theta'}L_n\left(\tilde{\theta}_n\right)\right)^{-1}\tilde{\Omega}_n^{-1}\left(D^2_{\theta,\theta'}L_n\left(\tilde{\theta}_n\right)\right)^{-1}D_\theta L_n\left(\tilde{\theta}_n\right) \\ &= n\left(D_\theta L_n\left(\tilde{\theta}_n\right)\right)'\left(D^2_{\theta,\theta'}L_n\left(\tilde{\theta}_n\right)\right)^{-1}\tilde{\Omega}_n^{-1}\left(D^2_{\theta,\theta'}L_n\left(\tilde{\theta}_n\right)\right)^{-1}D_\theta L_n\left(\tilde{\theta}_n\right) \\ &= n\left(D_\theta L_n\left(\tilde{\theta}_n\right)\right)'\tilde{\Omega}_n^{-1}D_\theta L_n\left(\tilde{\theta}_n\right) \\ &\xrightarrow{\text{d}} \chi^2_p, \end{aligned}$$

where

$$\tilde{\Omega}_n = D^2_{\theta,\theta'} L_n\left(\tilde{\theta}_n\right) = \left(-\frac{1}{n}\sum_{i=1}^{n} D^2_{\theta,\theta'} \log p_{\tilde{\theta}_n}\left(Y_i|X_i\right)\right)^{-1}.$$

The test statistic is then given by

$$T_n = n\left(D_\theta L_n\left(\tilde{\theta}_n\right)\right)' \tilde{\Omega}_n^{-1} D_\theta L_n\left(\tilde{\theta}_n\right)$$

and the critical value is given by the $1-\alpha$th quantile of χ^2_p distribution.

4.4.3 Likelihood Ratio Test

Idea: If H_0 is true, then $\ell_n(\tilde{\theta}_n) = \ell_n(\hat{\theta}_n)$; i.e. the likelihood with the constrained maximiser is equal to the likelihood with the unconstrained maximiser.

This suggests that we look at the ratio

$$\frac{\ell_n\left(\hat{\theta}_n\right)}{\ell_n\left(\tilde{\theta}_n\right)}$$

or, equivalently,

$$L_n\left(\hat{\theta}_n\right) - L_n\left(\tilde{\theta}_n\right)$$

and expect this to be "small". It can be shown that

$$2\left[L_n\left(\hat{\theta}_n\right) - L_n\left(\tilde{\theta}_n\right)\right] \xrightarrow{\mathrm{d}} \chi^2_p.$$

It can also be shown that for simple hypothesis tests (e.g. $H_0 : \theta_0 = \bar{\theta}$ vs $H_1 : \theta_0 = \check{\theta}$), then the Likelihood Ratio test is *uniformly most powerful* (Neyman-Pearson Lemma).