

ECON21030 (S25): Econometrics - Honors

Lecturer: Joseph Hardwick

Notes by: Aden Chen

Wednesday 28th May, 2025

Contents

1	Probability	4
1.1	Expectation	4
1.2	Mean Independence	5
1.3	Moments	5
1.4	Probability Inequalities	5
1.5	Random Vectors	7
1.6	The Binning Estimator	7
1.7	Conditional Expectation	7
2	Linear Regression	8
3	Estimation and Large Sample Theory	10
3.1	Convergence	10
3.2	Weak Law of Large Numbers	10
3.3	Continuous Mapping Theorem and Slutsky's Theorem	12
3.4	Central Limit Theorem	13
3.5	Delta Method	15
3.6	Estimators	16
4	Ordinary Least Squares Estimation	17
4.1	Matrix Notation	18
4.2	Projection	19
4.3	R squared	20
5	Finite Sample Properties of OLS	21
6	Large Sample Properties of OLS	23
6.1	Estimation of Σ : Homoskedasticity	24
6.2	Estimation of Σ : Heteroskedasticity	25
7	Partitioned Regression	27
7.1	Components of the Variance	27
7.2	Omitted Variable Bias	28
7.3	The Population Case	30
8	Inference	32
8.1	Finite Sample, Homoskedasticity	32
8.2	Heteroskedasticity	33
8.3	Linear Restrictions	34
9	Regression Specification	37
9.1	Weighted Regression	37
9.2	Log Specifications	37
9.3	Functional Forms	38
9.4	Dummy Variables	39

10 Regression Cheat Sheet	40
11 The Language of Causality	42
11.1 Random Assignment	44
12 Selection on Observables	46
12.1 Example Specification 1	46
12.2 Example Specification 2	47
13 Difference-in-Differences	48
13.1 Regression Implementation	48
13.2 Conditional Common Trends	49
13.3 Triple Difference-in-Differences	50
14 Instrumental Variables: The Simple Case	51
14.1 Standard Errors	52
14.2 Weak Instruments	53
14.3 Causes of Endogeneity	53
15 Instrumental Variables	55
15.1 Identification Setup	57
15.2 Exact Identification: The IV Estimator	57
15.3 Overidentification, Generalized Method of Moments	59
15.4 Consistency and Asymptotic Normality of GMM Estimators	61
15.5 Optimal GMM Estimation	62
15.6 GMM Under Conditional Homoskedasticity, Two Stage Least Squares	63
15.7 GMM Under Heteroskedasticity	64
16 Appendix A: A List of Theorems	66
17 Appendix B: Common Distributions	72
Index	73

1 Probability

Definition 1.1. A random variable X is **absolutely continuous** if there exists a density function f_X such that

$$F_X(x) = \int_{-\infty}^x f_X(t) dt.$$

Remark 1.2. Absolutely continuous distributions assign probability 0 to any finite set of points. ☞

1.1 Expectation

Proposition 1.3.

- \mathbb{E} is linear.
- If $X \leq Y$ with probability 1, then $\mathbb{E}X \leq \mathbb{E}Y$.

Theorem 1.4 (Jensen's Inequality). If X is such that $\mathbb{E}X$ and $\mathbb{E}g(X)$ exist and g is convex, then

$$g(\mathbb{E}X) \leq \mathbb{E}g(X)$$

where the inequality is strict if g is strictly convex and X is not constant.

Proof. From the convexity of g we know $g(x) \geq g(y) + g'(y)(x - y)$ for any x and y . Setting $y = \mu =: \mathbb{E}X$ gives

$$g(X) \geq g(\mu) + g'(\mu)(X - \mu), \quad \forall x, y.$$

Taking expectation on both sides gives the desired result. \square

Example 1.5. Wages are often modeled using a log-normal distribution: $\log w \sim \mathcal{N}(\mu, \sigma^2)$. Then, $\mathbb{E} \log w = \mu$, but $\mathbb{E}w = \mathbb{E}(\exp \log w) \geq e^\mu$ (the inequality is strict when $\sigma^2 > 0$). It turns out that $\mathbb{E}w = \exp(\mu + \sigma^2/2)$. ☞

Proposition 1.6 (Properties of Conditional Expectation).

- *Linearity.*
- *(Law of iterated expectation)* $\mathbb{E}(Y) = \mathbb{E}(\mathbb{E}(Y|X))$.
- *(Taking out what is known)* $\mathbb{E}(f(X) + g(X)Y|X) = f(X) + g(X)\mathbb{E}(Y|X)$.

Proposition 1.7 (Law of total variance).

$$\mathbb{V}(Y) = \mathbb{E}(\mathbb{V}(Y|X)) + \mathbb{V}(\mathbb{E}(Y|X)),$$

where

$$\mathbb{V}(Y|X) := \mathbb{E} \{ [Y - \mathbb{E}(Y|X)]^2 | X \} = \mathbb{E}(Y^2|X) - \mathbb{E}(Y|X)^2.$$

1.2 Mean Independence

Definition 1.8. Y is **mean independent** of X if $\mathbb{E}(Y|X) = \mathbb{E}(Y)$.

Proposition 1.9.

$$\begin{aligned} X \text{ is independent of } Y &\implies X \text{ is mean independent of } Y \\ &\implies \mathbb{C}(X, Y) = 0. \end{aligned}$$

The second implication follows from the law of iterated expectations:

$$\mathbb{C}(X, Y) = \mathbb{E}(X\mathbb{E}(Y|X)) - \mathbb{E}(X)\mathbb{E}(Y) = \mathbb{E}(X\mathbb{E}(Y)) - \mathbb{E}(X)\mathbb{E}(Y) = 0.$$

The converse of the last implication is not true in general, but true for jointly normal random variables.

1.3 Moments

Definition 1.10. If $\mathbb{E}(X^k)$ exists, then

- $\mathbb{E}(X^k)$ is the **k -th moment of X** .
- $\mathbb{E}[(X - \mathbb{E}X)^k]$ is the **k -th central moment of X** . The case $k = 2$ gives the variance of X .

Proposition 1.11 (Existence of Moments). *Suppose $\mathbb{E}(|X|^k) < \infty$ for some $k > 0$. Then for $0 < r < k$, $\mathbb{E}(|X|^r) < \infty$.*

Proof. First note

$$|X|^r \leq \mathbb{1}_{|X| < 1} + |X|^k \mathbb{1}_{|X| \geq 1}.$$

Taking expectation on both sides gives

$$\begin{aligned} \mathbb{E}|X|^r &\leq \mathbb{P}(|X| < 1) + \mathbb{E}\left(|X|^k \mathbb{1}_{|X| \geq 1}\right) \\ &\leq \mathbb{P}(|X| < 1) + \mathbb{E}\left(|X|^k\right) < \infty. \end{aligned}$$

□

Remark 1.12. Using the binomial theorem, we can then show that the k -th moment exists if and only if the r -th central moment exists. ☕

1.4 Probability Inequalities

Theorem 1.13 (Chebychev's Inequality). *Suppose X^r is a non-negative integrable random variable for some $r > 0$. Then for any $\delta > 0$, we have*

$$\mathbb{P}(X \geq \delta) \leq \frac{\mathbb{E}(X^r)}{\delta^r}.$$

Proof. Note that $X^r \geq \delta^r \mathbb{1}_{X \geq \delta}$ and take expectations on both sides. □

Remark 1.14. We can bound the probability that X is large using its moments. When $r = 1$, this is called Markov's Inequality. ☕

Lemma 1.15. $Y = 0$ almost surely if and only if $\mathbb{E}Y^2 = 0$.

Proof. If $\mathbb{E}Y^2 = 0$, then $Y^2 = 0$ as. Otherwise, suppose $\mathbb{P}(Y^2 > 0) = \epsilon$ for some $\epsilon > 0$. Write $\{Y^2 > 0\} = \bigcup_n \{Y^2 > n^{-1}\}$. We have

$$0 < \epsilon = \mathbb{P}(Y^2 > 0) \leq \sum_n \mathbb{P}\left(Y^2 > \frac{1}{n}\right),$$

where we used Boole's inequality.¹ There thus exists N such that $\mathbb{P}(Y^2 > N^{-1}) > 0$. We have

$$Y^2 \geq \frac{1}{N} \mathbb{1}\left(Y^2 > \frac{1}{N}\right)$$

and so $\mathbb{E}(Y^2) \geq N^{-1} \mathbb{P}(Y^2 > N^{-1}) > 0$. □

Theorem 1.16 (Cauchy-Schwarz Inequality). *If $\mathbb{E}(X^2)$ and $\mathbb{E}(Y^2)$ exist, then*

$$|\mathbb{E}(XY)|^2 \leq \mathbb{E}(X^2)\mathbb{E}(Y^2)$$

with equality if and only if $X = aY$ almost surely for some constant a .

Proof. If $Y = 0$ as the inequality is trivial. If not, $\mathbb{E}(Y^2) > 0$ and we can write

$$\begin{aligned} 0 &\leq \frac{\mathbb{E}\{[X\mathbb{E}(Y^2) - Y\mathbb{E}(XY)]^2\}}{\mathbb{E}(Y^2)} \\ &\leq \frac{\mathbb{E}(X^2)\mathbb{E}(Y^2) - 2\mathbb{E}(XY)^2\mathbb{E}(Y^2) + \mathbb{E}(Y^2)\mathbb{E}(XY)^2}{\mathbb{E}(Y^2)} \\ &= \mathbb{E}(X^2)\mathbb{E}(Y^2) - \mathbb{E}(XY)^2. \end{aligned}$$

We have equality if and only if $X\mathbb{E}(Y^2) - Y\mathbb{E}(XY) = 0$ as, which holds if and only if $X = aY$ as for some constant a . □

Corollary 1.17. *The correlation is bounded between -1 and 1 , with equality if and only if $X - \mathbb{E}X = b(Y - \mathbb{E}Y)$ for some constant b , which holds if and only if $X = a + bY$ for some constants a, b .*

Theorem 1.18 (Holder's Inequality). *If X and Y are random variables, then*

$$\mathbb{E}(|XY|) \leq \mathbb{E}(|X|^p)^{1/p} \mathbb{E}(|Y|^q)^{1/q}$$

for any $p, q > 0$ such that $1/p + 1/q = 1$.

¹ $\mathbb{P}(\bigcup A_i) \leq \sum \mathbb{P}(A_i)$.

1.5 Random Vectors

Definition 1.19. If X and Y are random vectors, then

$$C(X, Y) := \mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)'].$$

Proposition 1.20. Let X be a random vector such that $\mathbb{V}(X)$ exists. If A is a constant matrix and b a constant vector, then

$$\mathbb{V}(AX + b) = A \mathbb{V}(X) A'.$$

1.6 The Binning Estimator

Consider sample $\{Y_i, X_i\}_{i=1}^n$ with X discrete. The **binning estimator** of $\mathbb{E}(Y|X \in B)$ is

$$\hat{\mu}(B) = \frac{\sum Y_i \mathbb{1}(X_i \in B)}{\sum \mathbb{1}(X_i \in B)}.$$

With continuous X we may use a moving bin of the form $x \pm h$. For large sample we can use smaller h .

1.7 Conditional Expectation

Suppose $\mathbb{E}X^2 < \infty$ and $\mathbb{E}Y_i < \infty$ for each i . Consider the problem of minimizing

$$\mathbb{E}(Y - g(X))^2.$$

The solution is the **best predictor of Y under square loss**. That is,

$$g^* \in \arg \min_{g \in L^2(X)} \mathbb{E}(Y - g(X))^2.$$

Then, $g^*(X) = \mathbb{E}(Y|X)$.

Proof.

$$\begin{aligned} \mathbb{E}(Y - g(X))^2 &= \mathbb{E}[Y - \mathbb{E}(Y|X)]^2 + \mathbb{E}[\mathbb{E}(Y|X) - g(X)]^2 \\ &\quad + 2\mathbb{E}[(Y - \mathbb{E}(Y|X))(\mathbb{E}(Y|X) - g(X))], \end{aligned}$$

where the last term is 0 by the law of iterated expectation. □

2 Linear Regression

Proposition 2.1. Note that $\mathbb{E}(XX')$ is always positive semidefinite. Moreover, if it is invertible, then it is positive definite.

Definition 2.2. There is **perfect collinearity** in X if there exists a constant vector $a \neq 0$ such that $a'X = 0$ almost surely.

Example 2.3 (Perfect Collinearity).

- Suppose $X = (X_1, X_2)'$ and $X_1 = 3X_2$. If we take $c = (3, -1)'$, then $c'X = 3X_1 - X_2 = 0$.
- $X = (X_1, X_1^2)'$ is perfect collinear if X_1 is Bernoulli.



Proposition 2.4. Suppose X is a $(k \times 1)$ random vector and $\mathbb{E}(XX')$ exists. Then $\mathbb{E}(XX')$ is invertible if and only if there is no perfect collinearity in X .

Proof. If $X'a = 0$ a.s., then

$$\mathbb{E}(XX')a = \mathbb{E}(X(X'a)) = \mathbb{E}(X \cdot 0) = 0.$$

So $\mathbb{E}(XX')$ is not full rank and not invertible. If for any $c \in \mathbb{R}^k \setminus \{0\}$ we have $c'X \neq 0$ with positive probability, then

$$c'\mathbb{E}(XX')c = \mathbb{E}[(X'c)^2] > 0.$$

Thus $\mathbb{E}(XX')$ is positive definite and in particular invertible. □

We may restrict $L^2(X)$ to a smaller subset

$$H(X) = \{f : f(X) = X'a \text{ for some } a \in \mathbb{R}^k\}.$$

Then, the **best linear predictor** of Y given X is found by solving

$$\min_{b \in \mathbb{R}^k} \mathbb{E}(Y - X'b)^2.$$

Differentiation gives the FOC $2\mathbb{E}(XX')b^* - 2\mathbb{E}(XY) = 0$. Provided X_j are not perfectly collinear random variables, $\mathbb{E}(XX')$ is full rank, and so

$$b^* = \mathbb{E}(XX')^{-1}\mathbb{E}(XY).$$

Define the **prediction error** or **residual** to be $U = Y - X'b^*$. We have $\mathbb{E}(XU) = 0$.

Proof. $\mathbb{E}(XU) = \mathbb{E}(XY) - \mathbb{E}(XX')b^* = 0$ by the FOC. □

Consider next the problem

$$\min_{b \in \mathbb{R}^k} \mathbb{E}(\mathbb{E}(Y|X) - X'b)^2.$$

The solution is the best linear approximation to $\mathbb{E}(Y|X)$ under square loss. Write

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

By the law of iterated expectation, we have

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

Thus,

$$\mathbb{E}[(\mathbb{E}(Y|X) - X'b)^2] = \mathbb{E}(Y - X'b)^2 + \text{constant}.$$

Remark 2.5. Thus, regression can be interpreted as

- the best linear predictor of Y given X , and
- the best linear approximation to $\mathbb{E}(Y|X)$ under square loss.

Intuition: Write $Y = \mathbb{E}[Y|X] + U$, where $\mathbb{E}[U|X] = 0$. Then the projection of U onto X is 0. ☕

3 Estimation and Large Sample Theory

3.1 Convergence

Definition 3.1. A sequence of random variables X_n **converges in probability** to X ($X_n \xrightarrow{p} X$) if for each $\epsilon > 0$ we have

$$\mathbb{P}(|X_n - X| > \epsilon) \longrightarrow 0.$$

Example 3.2. Let $X_n = 2^n$ with probability $1/n$ and 0 otherwise. For arbitrary $\epsilon > 0$, we have

$$\mathbb{P}(|X_n - 0| > \epsilon) \leq \mathbb{P}(X_n = 2^n) = \frac{1}{n} \longrightarrow 0.$$



Definition 3.3. We say X_n **converges in r -th mean** to X for some $r > 0$ if

$$\mathbb{E}(|X_n - X|^r) \longrightarrow 0.$$


Note that Chebyshev gives $\mathbb{P}(|X_n - X| > \epsilon) \leq \mathbb{E}(|X_n - X|^r)/\epsilon^r \rightarrow 0$ and so:

Proposition 3.4. If X_n converges in r -th mean to X , then $X_n \xrightarrow{p} X$.

Example 3.5. The converse is not true. Consider $X_n = 2^n$ with probability $1/n^2$ and 0 otherwise. $X_n \xrightarrow{p} 0$ but any moments of X_n explodes. The converse can be established under additional assumptions, e.g.:

Suppose $|X_n| \leq K$ wp 1 for some constant K and choose $\epsilon > 0$. We have

$$|X_n|^r \leq K^r \mathbb{1}(|X_n| > \epsilon) + \epsilon^r \mathbb{1}(|X_n| \leq \epsilon).$$

So $\mathbb{E}(|X_n|^r) \leq K^r \mathbb{P}(|X_n| > \epsilon) + \epsilon^r$. Thus if $X_n \xrightarrow{p} 0$, we have $\mathbb{E}(|X_n|^r) \rightarrow 0$ also. 

3.2 Weak Law of Large Numbers

Theorem 3.6 (Weak Law of Large Numbers). Suppose $\{X_i\}_{i \geq 1}$ is an iid sequence of random variables with $\mathbb{E}(X_i) = \mu$. Then, $\bar{X}_n \xrightarrow{p} \mu$.

We prove this statement with the additional assumption that the second moment exists. This is not required but makes the proof easier.

Proof. We have

$$\mathbb{P}(|\bar{X}_n - \mu| > \epsilon) \leq \frac{\mathbb{E}[|\bar{X}_n - \mu|^2]}{\epsilon^2} = \frac{\mathbb{V}(\bar{X}_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \longrightarrow 0.$$


□

Theorem 3.7 (WLLN for Moments). If $\mathbb{E}(X_i^k) < \infty$, then

$$\frac{1}{n} \sum_i X_i^k \xrightarrow{p} \mathbb{E}(X^k).$$

Example 3.8. Let $\{X_i\}_{i \geq 1}$ be an iid sample drawn from F . Define the **empirical distribution** of F by

$$\hat{F}(x) := \frac{1}{n} \sum_i \mathbb{1}(X_i < x).$$

The weak law of large numbers gives $\hat{F}_n(x) \xrightarrow{P} \mathbb{E}[\mathbb{1}(X < x)] = F(x)$. 

Proposition 3.9. Let X_n be a sequence of $(k \times 1)$ random vectors. Then,

- $X_n \xrightarrow{P} X$ if and only if $X_{n,i} \xrightarrow{P} X_i$ for $i = 1, \dots, k$.
- $X_n \rightarrow X$ in r -th mean if and only if $X_{n,i} \rightarrow X_i$ in r -th mean for $i = 1, \dots, k$.

Proof. We prove the first result. Note that

$$|X_{n,i} - X_i| \leq \|X_n - X\| \leq \sqrt{K} \max_{i \leq k} |X_{n,i} - X_i|.$$

Then, if $X_{n,i} \xrightarrow{P} X_i$ for all i , then

$$\begin{aligned} \mathbb{P}(\|X_n - X\| > \epsilon) &\leq \mathbb{P}\left(\max_{i \leq k} |X_{n,i} - X_i| > \epsilon/\sqrt{K}\right) \\ &= \mathbb{P}\left(\bigcup_{i=1}^k |X_{n,i} - X_i| > \epsilon/\sqrt{K}\right) \\ &\leq \sum_{i=1}^k \mathbb{P}\left(|X_{n,i} - X_i| > \epsilon/\sqrt{K}\right) \rightarrow 0. \end{aligned}$$

On the other hand, if $X_n \xrightarrow{P} X$, then $\mathbb{E}(|X_{n,i} - X_i| > \epsilon) \leq \mathbb{P}(\|X_n - X\| > \epsilon) \rightarrow 0$. \square

Corollary 3.10. The weak law of large numbers for random vectors.

Definition 3.11. A sequence of random variables X_n **converges in distribution** to X ($X_n \xrightarrow{d} X$) if

$$F_{X_n}(x) \rightarrow F_X(x)$$

for all x at which F_X is continuous.

Remark 3.12.

- To see why we require convergence only at continuity points of F_X , note that F_X is only right-continuous for discrete X . Then, consider $X_n := 1/n$ and $X := 0$. We have $X_n \rightarrow X$ a.s., but $F_{X_n}(0) = 0 \neq 1 = F_X(0)$.
- This is the weakest notion of convergence. Convergence in probability implies convergence in distribution, but the converse is not true: Let $Y_n := (-1)^n X$, where $X \sim \mathcal{N}(0, 1)$. We have $Y_n \xrightarrow{d} \mathcal{N}(0, 1)$, but Y_n does not converge in probability.



3.3 Continuous Mapping Theorem and Slutsky's Theorem

Theorem 3.13 (Continuous Mapping Theorem). *Let $g : \mathbb{R}^k \rightarrow \mathbb{R}^n$ be continuous on $S \subset \mathbb{R}^k$ with $\mathbb{P}(X \in S) = 1$. Then the following hold:*

(i) *if $X_n \xrightarrow{p} X$, then $g(X_n) \xrightarrow{p} g(X)$.*

(ii) *If $X_n \xrightarrow{d} X$, then $g(X_n) \xrightarrow{d} g(X)$.*

Remark 3.14. The theorem does not hold for $X_n \rightarrow X$ in r -th moment. Consider $X_n = n$ with probability n^{-2} and 0 otherwise. We have $\mathbb{E}|X_n - 0| \rightarrow 0$ but

$$\mathbb{E}(|X_n^2 - 0^2|) = 1.$$



Theorem 3.15 (Slutsky's Theorem). *Suppose $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$ for some constant c . Then,*

$$X_n + Y_n \xrightarrow{d} X + c, \quad X_n Y_n \xrightarrow{d} Xc, \quad X_n / Y_n \xrightarrow{d} X/c \text{ provided } c \neq 0.$$

Remark 3.16. It turns out that under our assumption,

$$\begin{pmatrix} X_n \\ Y_n \end{pmatrix} \xrightarrow{d} \begin{pmatrix} X \\ c \end{pmatrix}.$$

We may then apply the continuous mapping theorem.



Example 3.17. Let $\{(X_i - \bar{X}, Y_i - \bar{Y})\}_{i \geq 1}$ be a sequence of (2×1) iid random vectors with $\mathbb{E}(X_i^2) < \infty$, $\mathbb{E}(Y_i^2) < \infty$. Define

$$\begin{aligned} \hat{\rho} &:= \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2} \sqrt{\sum (Y_i - \bar{Y})^2}} \\ &= \frac{\frac{1}{n} \sum X_i Y_i - \bar{X} \bar{Y}}{\sqrt{\frac{1}{n} \sum X_i^2 - \bar{X}^2} \sqrt{\frac{1}{n} \sum Y_i^2 - \bar{Y}^2}}. \end{aligned}$$

The weak law of large numbers gives

$$\begin{aligned} \frac{1}{n} \sum X_i Y_i &\xrightarrow{p} \mathbb{E}(XY), \\ \frac{1}{n} \sum X_i^2 &\xrightarrow{p} \mathbb{E}(X^2), \\ \frac{1}{n} \sum Y_i^2 &\xrightarrow{p} \mathbb{E}(Y^2). \end{aligned}$$

Thus,

$$\left(\bar{X}, \bar{Y}, \frac{1}{n} \sum X_i Y_i, \frac{1}{n} \sum X_i^2, \frac{1}{n} \sum Y_i^2 \right) \xrightarrow{p} (\mathbb{E}(X), \mathbb{E}(Y), \mathbb{E}(XY), \mathbb{E}(X^2), \mathbb{E}(Y^2)).$$

Now let

$$g(x, y, s, t, w) := \frac{s - xy}{\sqrt{t - x^2}\sqrt{w - y^2}}.$$

Note that g is continuous at all points except where $t = x^2$ and $w = y^2$. Provided neither X nor Y are constant, we have

$$\mathbb{E}(X^2) > \mathbb{E}(X)^2, \quad \mathbb{E}(Y^2) > \mathbb{E}(Y)^2.$$

By the continuous mapping theorem we then have

$$\hat{\rho} \xrightarrow{p} \frac{\mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)}{\sqrt{\mathbb{E}(X^2) - \mathbb{E}(X)^2}\sqrt{\mathbb{E}(Y^2) - \mathbb{E}(Y)^2}} = \frac{\mathbb{C}(X, Y)}{\sqrt{\mathbb{V}(X)\mathbb{V}(Y)}}.$$



Proposition 3.18. *Let $A_n \in \mathbb{R}^{P \times K}$ be a sequence of matrices converging in probability to a constant matrix A . Let B_n be a sequence of $(K \times 1)$ random vectors such that $B_n \xrightarrow{d} \mathcal{N}(\mu, \Sigma)$. Then,*

$$A_n B_n \xrightarrow{d} A \mathcal{N}(\mu, \Sigma) \sim \mathcal{N}(A\mu, A\Sigma A').$$

Proof. Since the columns of A_n , denoted $\text{vec}(A_n)$ converges in probability to $\text{vec}(A)$ (that is, we use the Frobenius norm for matrices), a constant vector, we have

$$\begin{pmatrix} B_n \\ \text{vec}(A_n) \end{pmatrix} \xrightarrow{d} \begin{pmatrix} \mathcal{N}(\mu, \Sigma) \\ \text{vec}(A) \end{pmatrix}.$$

The continuous mapping theorem then gives

$$A_n B_n \xrightarrow{d} A \mathcal{N}(\mu, \Sigma).$$

Since linear transformations of multivariate normal are also multivariate normal, we have

$$A_n B_n \xrightarrow{d} \mathcal{N}(A\mu, A\Sigma A').$$

□

3.4 Central Limit Theorem

Theorem 3.19 (Central Limit Theorem). *Let $\{X_i\}_{i \geq 1}$ be an iid sequence of $(K \times 1)$ random vectors with mean μ and finite variance matrix Σ . Then*

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

Example 3.20. Let X_i be iid with $\mathbb{E}(X_i) = \mu$, $\mathbb{V}(X_i) = \sigma^2$, and skewness $\kappa = \mathbb{E}\left(\left(\frac{X_i - \mu}{\sigma}\right)^3\right)$. It turns out that the skewness of the sample mean is

$$\mathbb{E}\left(\left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}\right)^3\right) = \frac{\kappa}{\sqrt{n}}.$$


If $X_i \sim \text{Gamma}(k, \theta)$, where k and θ are shape and scale parameters. Then, $\sum X_i \sim \text{Gamma}(kn, \theta)$ and $\bar{X}_n \sim \text{Gamma}(kn, \theta/n)$ and we have $\mathbb{E}(X_i) = k\theta$ and $\mathbb{V}(X_i) = k\theta^2$. It turns out that the skewness of gamma is


$$\kappa = \mathbb{E} \left(\left(\frac{X_i - k\theta}{\sqrt{k\theta^2}} \right)^3 \right) = \frac{2}{\sqrt{k}}.$$

For the sample mean, we have

$$\mathbb{E} \left(\left(\frac{\bar{X}_n - k\theta}{\sqrt{(k\theta^2)/n}} \right)^3 \right) = \frac{\kappa}{\sqrt{n}} = \frac{2}{\sqrt{nk}}.$$



Example 3.21. If X_i is iid Cauchy, then \bar{X}_n are Cauchy. This occurs because the Cauchy distribution does not have mean or variance. 

Remark 3.22. The Berry-Esseen theorem gives a finite sample bound on the inaccuracy of the normal approximation using the third moment. 

Example 3.23. Let $\{X_i\}_{i \geq 1}$ be iid with mean μ and variance $\sigma^2 > 0$. We wish to test $H_0 : \mu = \mu_0$ at the significance level α .

We have

$$S_n^2 := \frac{1}{n-1} \sum (X_i - \bar{X}_n)^2 = \frac{n}{n-1} \left(\frac{1}{n} \sum X_i^2 - \bar{X}_n^2 \right).$$

The weak law of large numbers gives

$$\frac{1}{n} \sum X_i^2 \xrightarrow{p} \mathbb{E}(X_i^2), \quad \bar{X}_n \xrightarrow{p} \mathbb{E}(X_i).$$

It follows that

$$\left(\frac{n}{n-1}, \frac{1}{n} \sum X_i^2, \bar{X}_n \right) \xrightarrow{p} (1, \mathbb{E}(X_i)^2, \mathbb{E}(X_i)).$$

The continuous mapping theorem with

$$g(x, y, z) = \frac{1}{\sqrt{x(y - z^2)}}$$

gives (since $\mathbb{E}(X_i^2) > \mathbb{E}(X_i)^2$ from $\sigma^2 > 0$)

$$\frac{1}{\sqrt{S_n^2}} \xrightarrow{p} \frac{1}{\sqrt{\sigma^2}} = \frac{1}{\sigma}.$$

By Slutsky's theorem, we have then that


$$\sqrt{n} \frac{\bar{X}_n - \mu}{S_n} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sqrt{S_n^2}} \xrightarrow{d} \frac{\mathcal{N}(0, \sigma^2)}{\sigma} = \mathcal{N}(0, 1).$$

Note on terminology:

- $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ is a **pivot** since its distribution does not depend on the unknown parameters. It enables use to construct confidence intervals by “inverting the pivot.” This is called “test inversion.” (Note however that the pivot is not a test statistic.)
- Under H_0 , $\sqrt{n}(\bar{X}_n - \mu_0)/\sigma$ is a **test statistic**.
- The first term in the display above is not a test statistic since we do not know μ . It is though called a **asymptotic pivot** since its asymptotic distribution does not depend on the unknown parameters.

Under H_0 , then,

$$\mathbb{P} \left(\left| \frac{\sqrt{n}(\bar{X}_n - \mu_0)}{S_n} \right| > z_{1-\frac{\alpha}{2}} \right) \longrightarrow \alpha,$$

with which we can construct a *asymptotically correct* confidence interval for μ . 

3.5 Delta Method

Theorem 3.24 (Delta Method). *Let $\{X_n\}_{n \geq 1}$ be a sequence of $(K \times 1)$ random vectors and suppose*

$$n^r(X_n - c) \xrightarrow{d} X$$

for some $r > 0$ and constant vector c . Let $g : \mathbb{R}^k \rightarrow \mathbb{R}^d$ be differentiable at the point c . Then,

$$n^r(g(X_n) - g(c)) \xrightarrow{d} Dg(c)X.$$

In particular, if $X \sim \mathcal{N}(0, \Sigma)$, then


$$n^r(g(X_n) - g(c)) \xrightarrow{d} \mathcal{N}(0, Dg(c)\Sigma Dg(c)').$$

Remark 3.25. Intuition: we use the approximation

$$\sqrt{n}(g(X_n) - g(\mu)) \approx g'(\mu)\sqrt{n}(X_n - \mu)$$

and apply the WLLN. More precisely, we can use

$$\sqrt{n}(g(X_n) - g(\mu)) = g'(\mu)\sqrt{n}(X_n - \mu) + h(X_n)\sqrt{n}(X_n - \mu),$$


where $h(\cdot)$ satisfies $\lim_{X \rightarrow \mu} h(X) = h(\mu) = 0$. Thus $h(X_n) \xrightarrow{p} h(\mu) = 0$ if $X_n \xrightarrow{p} \mu$. 

Example 3.26. Suppose we wish to estimate $\theta \in \mathbb{R}^d$ that solves the equation

$$g(\theta) = \mathbb{E}(h(X)),$$

where $g : \mathbb{R}^d \rightarrow \mathbb{R}^k$ and h are known functions. Given sample, we solve

$$g(\hat{\theta}) = \frac{1}{n} \sum h(X_i).$$


A solution to this equation is a **method of moments estimator**. 


3.6 Estimators


Definition 3.27. We say $\hat{\theta}$ is biased downward if $\mathbb{E}(\hat{\theta}) < \theta$.

Example 3.28. Let $\{X_i\}$ be iid from G with mean μ and variance σ^2 . Define $S_n^2 = n^{-1} \sum (X_i - \bar{X}_n)^2$. This is biased downward. To see this, note that $n^{-1} \sum (X_i - \mu)^2$ is unbiased (if we know μ). The estimator S_n^2 is biased downward because

$$S_n^2 = \min_{a \in \mathbb{R}} \frac{1}{n} \sum (X_i - a)^2 \leq \frac{1}{n} \sum (X_i - \bar{X}_n)^2.$$

However, S_n^2 is consistent: we have $S_n^2 = n^{-1} \sum x_i^2 - (\bar{X}_n)^2$, where the first term converges to probability to $\mathbb{E}(X^2)$ and the second to $\mathbb{E}(X)^2$. 

Example 3.29. Let $\{X_i\}$ be iid from G with mean μ and variance σ^2 . Consider linear estimators of the form $\hat{\mu} = \sum a_i X_i$. If the estimator is unbiased, then $\sum a_i = 1$. To find the best linear estimator, we optimize subject to $\sum a_i = 1$ to obtain $a_i = n^{-1}$. 

Remark 3.30. Trade off between bias and variance: ridge regression, shrinkage estimator. 

We think of the asymptotic distribution of an estimator as a non-degenerate distribution:

Definition 3.31. Suppose $\hat{\theta}_n$ is a consistent estimator of θ . If for some sequence $\tau_n \rightarrow \infty$ and non-degenerate random variable X ,

$$\tau_n(\hat{\theta}_n - \theta) \xrightarrow{d} X,$$

we say that F_X is the **asymptotic distribution** of $\hat{\theta}_n$, where F_X is the distribution function of X .

4 Ordinary Least Squares Estimation

Theorem 4.1. Suppose $\mathbb{E}(y^2) < \infty$ and $\mathbb{E}(x_j^2) < \infty$ for each $j = 1, \dots, k$. The function $g(x) := \mathbb{E}(y|x)$ is the best predictor of y given x under square loss. That is,

$$\mathbb{E}(y|x) \in \arg \min_g \mathbb{E}[(y - g(x))^2].$$

We may interpret the linear model as a best linear approximation to the conditional mean function. We choose b to solve

$$\min_{b \in \mathbb{R}^k} \mathbb{E}(\mathbb{E}(y|x) - x'b)^2.$$

If $\mathbb{E}(xx')$ is invertible, we have

$$\beta = \mathbb{E}(xx')^{-1} \mathbb{E}(xy).$$

Example 4.2. Consider the simple linear model $y = \beta_0 + \beta_1 x_1 + u$. We have

$$\beta = \frac{1}{\mathbb{V}(x_1)} \begin{bmatrix} \mathbb{E}[x_1^2] & -\mathbb{E}[x_1] \\ -\mathbb{E}[x_1] & 1 \end{bmatrix} \begin{bmatrix} \mathbb{E}(y) \\ \mathbb{E}(x_1 y) \end{bmatrix}.$$

Thus,

$$\beta_1 = \frac{\mathbb{C}(y, x_1)}{\mathbb{V}(x_1)} = \sqrt{\frac{\mathbb{V}(y)}{\mathbb{V}(x_1)}} \cdot \text{Corr}(y, x_1)$$

and

$$\beta_0 = \mathbb{E}(y) - \beta_1 \mathbb{E}(x_1).$$



Remark 4.3. By writing

$$\tilde{W} = \alpha_0 + \alpha_1 S + U; \quad \mathbb{E}[SU] = \mathbb{E}[U] = 0$$

in a prediction context, $\mathbb{E}[SU] = \mathbb{E}[U] = 0$ is true *by construction*. In a causal content, they are *assumptions*; we are saying that “correlation is causation.” ☕

Given a sample $\{y_i, x_i\}_{i=1}^n$, we can solve the analogous sample problem

$$\min_{b \in \mathbb{R}^k} \frac{1}{n} \sum_i (y_i - x_i' b)^2.$$

The FOC is

$$\sum_i x_{ij} (y_i - x_i' \hat{\beta}) = 0, \quad 1 \leq j \leq k.$$

Or, equivalently,

$$\sum_i x_i (y_i - x_i' \hat{\beta}) = 0.$$

Thus we have:

$$\left(\frac{1}{n} \sum x_i x_i' \right) \hat{\beta} = \frac{1}{n} \sum x_i y_i,$$

and if $\sum x_i x_i'$ is invertible, then

$$\hat{\beta} = \left(\frac{1}{n} \sum x_i x_i' \right)^{-1} \frac{1}{n} \sum x_i y_i, \quad \sum x_i \hat{u}_i = 0.$$

4.1 Matrix Notation

The model $y_i = x_i'\beta + u_i$ for each i can equivalently be written as

$$Y = X\beta + U,$$

where

$$X = \begin{bmatrix} -x_1' \\ \vdots \\ -x_n' \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix}.$$

is called the **design matrix**. The least squares minimization problem is

$$\min_{b \in \mathbb{R}^k} \sum (y_i - x_i'b)^2 \equiv \min_{b \in \mathbb{R}^k} (Y - Xb)'(Y - Xb).$$

Note that by definition, \hat{Y} is the projection of Y onto the column space of X . By the projection theorem below, \hat{Y} is unique.

We may rewrite the FOC using matrix notation as:



$$X'X\hat{\beta} = X'Y, \quad X'\hat{U} = 0.$$

Remark 4.4. Recall that

$$X = \begin{bmatrix} -x_1' \\ \vdots \\ -x_n' \end{bmatrix}.$$

Thus

$$\begin{aligned} X'X &= \begin{pmatrix} \sum x_{i1}^2 & \sum x_{i1}x_{i2} & \cdots & \sum x_{i1}x_{ik} \\ \vdots & \vdots & \ddots & \vdots \\ \sum x_{ik}x_{i1} & \sum x_{ik}x_{i2} & \cdots & \sum x_{ik}^2 \end{pmatrix} \\ &= \sum_i \begin{pmatrix} x_{i1}^2 & x_{i1}x_{i2} & \cdots & x_{i1}x_{ik} \\ \vdots & \vdots & \ddots & \vdots \\ x_{ik}x_{i1} & x_{ik}x_{i2} & \cdots & x_{ik}^2 \end{pmatrix} = \sum_i x_i x_i'. \end{aligned}$$

Remark 4.5. The FOC can be equivalently written as $X'\hat{U} = 0$, since $\hat{U} = Y - X\hat{\beta}$. Thus $\sum_i x_i \hat{u}_i = 0$; each column of X is orthogonal to \hat{U} . The FOC, in other words, states that the residual is orthogonal to the column space of X .  

Thus:

Proposition 4.6. *If X is full column rank, then the OLS estimator is given by*

$$\hat{\beta} = (X'X)^{-1}X'Y.$$

Proof. We show that X is full column rank if and only if it is invertible. If X is not full column rank, there exists $c \neq 0$ such that $Xc = 0$. Then $X'Xc = 0$. Conversely, if $X'X$ is not invertible, there exists $a \neq 0$ such that $X'Xa = 0$, and so $a'X'Xa = 0$. This happens precisely when $Xa = 0$. \square

Remark 4.7. $X'X$ is invertible with probability approaching 1 as $n \rightarrow \infty$ if $\mathbb{E}[X'X]$ is invertible in the population. ☕

4.2 Projection

Theorem 4.8 (Projection Theorem). *Let $y \in \mathbb{R}^n$ and let S be any nonempty subspace of \mathbb{R}^n . There exists a unique point \hat{y} such that $\|y - \hat{y}\|$ is minimized over S . A necessary and sufficient condition for \hat{y} is that $y - \hat{y}$ is orthogonal to every vector in S .*

In the case of ordinary least squares, we have the necessary and sufficient condition for \hat{Y} is

$$X'(Y - \hat{Y}) = 0$$

for all $x \in S(X)$, the column space of X . Equivalently, we have

$$X'(Y - \hat{Y}) = 0.$$

Since $\hat{Y} \in S(X)$, we may write $\hat{Y} = Xb$ for some $b \in \mathbb{R}^k$. Then, from $X'Y = X'Xb$ we have

$$b = (X'X)^{-1}X'Y$$

and

$$\hat{Y} = X(X'X)^{-1}X'Y.$$

Definition 4.9.

- $P_X := X(X'X)^{-1}X'$ is called the **projection matrix**.
- $M_X := I_n - X(X'X)^{-1}X'$ is called the **residual maker**.

Proposition 4.10. M_X projects a vector on to the $n - k$ dimensional vector space orthogonal to the column space of X .

Proof. The residual $P_X Y$ is orthogonal to the orthogonal complement of the column space of X . By the projection theorem we have the desired result. \square

Proposition 4.11 (Projection Matrices, Elementary Properties).

- P_X is symmetric and idempotent.
- $P_X M_X = M_X P_X = 0$, since $(P_X Y)' M_X Y = Y' P_X M_X Y = 0$.
- $P_X X = X$, $M_X X = 0$.
- For every Y , $Y = P_X Y + M_X Y = \hat{Y} + \hat{U}$. Thus $\|Y\|^2 = \|P_X Y\|^2 + \|M_X Y\|^2$.

Remark 4.12. $(A^{-1})' = (A')^{-1}$. ☕

4.3 R squared

Definition 4.13. The **coefficient of determination**, R^2 , is defined by

$$\begin{aligned} R^2 &:= \frac{\text{ESS}}{\text{TSS}} = \frac{\text{SSR}}{\text{TSS}} \\ &= \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}. \end{aligned}$$

Remark 4.14.

- If we include a constant, then $0 \leq R^2 \leq 1$ and the equalities above hold. In the univariate case, $R^2 = [\text{Corr}(y_i, x_i)]^2$; in general, $R^2 = [\text{Corr}(y_i, \hat{y}_i)]^2$.
- It turns out that $\text{SSR} = \|M_X M_C Y\|^2$, $\text{ESS} = \|P_X M_C Y\|^2$, and $\text{TSS} = \|M_C Y\|^2$.
- R^2 measures how well the model fits the data relative to a model with only a constant and no other regressors. OLS maximizes R^2 over the class of linear estimators. Adding new regressors weakly increases R^2 .
- R^2 is an estimator of the **population R^2**

$$R_{\text{pop}}^2 := 1 - \frac{\mathbb{V}(u)}{\mathbb{V}(y)},$$

since SSR/n is an estimator of $\mathbb{V}(u)$ and TSS/n an estimator of $\mathbb{V}(y)$.

- Intuitively, the model is soaking up the variance of y :

$$\mathbb{V}(Y) = \mathbb{E}(\mathbb{V}(Y|X)) + \mathbb{V}(\mathbb{E}(Y|X)).$$

- SSR weakly decreases (and so R^2 weakly increases) when a regressor is added to the model. The **adjusted R^2** penalizes the additional regressors:

$$\bar{R}^2 := 1 - \frac{n-1}{n-k-1} \frac{\text{SSR}}{\text{TSS}} = 1 - \frac{n-1}{n-k-1} (1 - R^2) \leq R^2,$$

where k is the total number of explanatory variables in the model (excluding the intercept).



5 Finite Sample Properties of OLS

Assumption 5.1 (Gauss Markov).

MLR.1: The model under consideration is

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u \equiv x' \beta + u.$$

MLR.2: We observe an iid sample of $\{y_i, x_i\}_{i=1}^n$.

MLR.3: There is no perfect collinearity in the sample (so a unique $\hat{\beta}$ exists).

MLR.4: $\mathbb{E}(u|x) = 0$.

MLR.5: (**Homoskedasticity**) $\mathbb{V}(y|x) = \sigma^2$.

Remark 5.2. If MLR.5 does not hold but we know the precise form of the heteroskedasticity, say $\sigma(x_i)^2$, we can transform the model to restore homoskedasticity. This is called **generalized least squares** (GLS). The finite sample estimate, however, is quite noisy because we also need to estimate $\sigma(x_i)^2$. Thus, people now usually just use OLS and correct the standard errors. ☕

Proposition 5.3. *Assuming MLR.1–4, we have:*

- (i) *The OLS estimator is unbiased.*
- (ii) *Let $\Omega := \mathbb{V}(U|X) = \mathbb{E}(U'U|X)$. Then,*

$$\mathbb{V}(\hat{\beta}|X) = (X'X)^{-1} X' \Omega X (X'X)^{-1}.$$

Thus, assuming also MLR.5, we have $\Omega = \sigma^2 I_n$ and so

$$\mathbb{V}(\hat{\beta}|X) = \sigma^2 (X'X)^{-1}.$$

Proof.

- (i) From $\mathbb{E}(U|X) = 0$, we have $\mathbb{E}(Y|X) = X\beta$ and so

$$\mathbb{E}[\hat{\beta}|X] = \mathbb{E}\left((X'X)^{-1} X' Y | X\right) = (X'X)^{-1} X' X \beta = \beta.$$

The law of iterated expectations gives $\mathbb{E}(\hat{\beta}) = \beta$.

- (ii) We have

$$\mathbb{V}(\hat{\beta}|X) = \mathbb{V}(\beta + (X'X)^{-1} X' U) = (X'X)^{-1} \mathbb{V}(U|X) (X'X)^{-1}.$$

Since $\mathbb{E}[U|X] = 0$ by MLR.4, we have $\mathbb{V}(U|X) = \mathbb{E}(U'U|X)$. Note that when $i \neq j$,

$$\mathbb{E}(u_i u_j | x_i, x_j) = \mathbb{E}(u_i \mathbb{E}[u_j | u_i, x_i, x_j] | x_i, x_j) = 0$$

and on the diagonal, we have

$$\mathbb{E}(u_i^2 | x_i) = \mathbb{V}(u_i | x_i) = \sigma^2.$$

□

Remark 5.4. The same argument shows that under heteroskedasticity, the variance of U is of the form

$$\mathbb{V}(U|X) = \begin{pmatrix} f(x_1) & 0 & \cdots & 0 \\ 0 & f(x_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & f(x_n) \end{pmatrix}.$$

☕

Theorem 5.5 (Gauss-Markov). *Under MLR.1–5, the OLS estimator is the **best linear unbiased estimator**. That is, it achieves the smallest variance in the class of linear estimators² that are also unbiased conditional on X . More precisely, let $\hat{\beta}$ be the OLS estimator and let $\tilde{\beta} = A(X)Y$ satisfy $\mathbb{E}(\tilde{\beta}|X) = \beta$. We have then that $\mathbb{V}(\tilde{\beta}|X) - \mathbb{V}(\hat{\beta}|X)$ is positive semidefinite.*

Proof. Write $\tilde{\beta} = AY$, where A is a $k \times N$ matrix that depends only on X . Note that we have

$$\mathbb{E}(\tilde{\beta}|X) = \mathbb{E}(AY|X) = AX\beta = \beta.$$

This implies that $(AX - I)\beta = 0$ for any β . Thus $AX = I_k$.

Next, note that

$$\mathbb{V}(AY|X) = A \mathbb{V}(Y|X) A' = A \mathbb{V}(U|X) A' = \sigma^2 AA'.$$

The OLS estimator corresponds to the case $A = (X'X)^{-1}X'$. It remains to show that $\sigma^2 AA' - \sigma^2 (X'X)^{-1}$ is positive semidefinite for any A such that $AX = I_k$. Now note that

$$\begin{aligned} & [A - (X'X)^{-1}X'] [A - (X'X)^{-1}X']' \\ &= AA' - (X'X)^{-1}X'A' - AX(X'X)^{-1} + (X'X)^{-1}X'X(X'X)^{-1} \\ &= AA' - (X'X)^{-1}, \end{aligned}$$

where the last equality follows from recalling $AX = I_k$. It remains to recall that CC' is positive semidefinite for any matrix C . □

Corollary 5.6. *Let r be an arbitrary $k \times 1$ vector. The Gauss-Markov theorem implies that $r'\hat{\beta}$ is the best linear unbiased estimator of $r'\beta$, since*

$$\mathbb{V}(r'\tilde{\beta}|X) - \mathbb{V}(r'\hat{\beta}|X) = r' [\mathbb{V}(\tilde{\beta}|X) - \mathbb{V}(\hat{\beta}|X)] r \geq 0.$$

In particular, taking $r = e_j$, we have $\mathbb{V}(\hat{\beta}_j|X) \leq \mathbb{V}(\tilde{\beta}_j|X)$.

²A linear estimator of β_j is a linear combinations of the $\{y_i\}$, with coefficients depending on $\{x_i\}$.

6 Large Sample Properties of OLS

We drop MLR.5, and weaken MLR.3 and MLR.4 in the next proposition:

Proposition 6.1. *Consider the model*

$$y = x'\beta + u.$$

And assume $\{y_i, x_i\}_{i=1}^n$ is iid.

- (i) Suppose $\mathbb{E}(ux) = 0$ and $\mathbb{E}(xx')$ is invertible (so that $n^{-1} \sum x_i x_i'$ is invertible with probability approaching 1). Then, the OLS estimator is consistent; $\hat{\beta} \xrightarrow{p} \beta$.

Note that if we are doing prediction, $\mathbb{E}(ux) = 0$ is true by construction. When dealing with causality, however, this is an assumption we need to argue.

- (ii) If $\mathbb{V}(xu)$ also exists, then $\hat{\beta}$ is asymptotically normal. Specifically, we have $\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(0, \Sigma)$, where $\Sigma := \mathbb{E}(xx')^{-1} \mathbb{V}(xu) \mathbb{E}(xx')^{-1}$.

Proof.

- (i) First note that

$$\left(\frac{1}{n} \sum x_i x_i', \frac{1}{n} \sum x_i y_i \right) \xrightarrow{p} (\mathbb{E}(xx'), \mathbb{E}(xy)),$$

since the components each convergences in probability by the law of large numbers.

Let $g(x, y) = x^{-1}y$ (where x and y are matrices). The function g is continuous if x is invertible. By the continuous mapping theorem, we have

$$\left(\frac{1}{n} \sum x_i x_i' \right)^{-1} \frac{1}{n} \sum x_i y_i \xrightarrow{p} \mathbb{E}(xx')^{-1} \mathbb{E}(xy).$$

Plugging in $y = x'\beta + u$ and using $\mathbb{E}(xu) = 0$, the right hand side can be rewritten as

$$\beta + \mathbb{E}(xx')^{-1} \mathbb{E}(xu) = \beta.$$

This proves that $\hat{\beta}$ is consistent.

- (ii) Next, suppose that $\mathbb{V}(xu)$ exists. Applying the central limit theorem to the iid sequence $\{x_i(y_i - x_i'\beta)\}_{i \geq 1}$ gives

$$\frac{1}{\sqrt{n}} \sum x_i u_i \xrightarrow{d} \mathcal{N}(0, \mathbb{V}(xu)) = \mathcal{N}(0, \mathbb{E}(u^2 x x')).$$

Slutsky's theorem then gives

$$\sqrt{n}(\hat{\beta}_n - \beta) = \left(\frac{1}{n} \sum x_i x_i' \right)^{-1} \frac{1}{\sqrt{n}} \sum x_i u_i \xrightarrow{d} \mathcal{N}(0, \Sigma).$$

□

6.1 Estimation of Σ : Homoskedasticity

Throughout this subsection, we maintain MLR.4 and MLR.5: $\mathbb{E}(u|x) = 0$, $\mathbb{V}(u|x) = \sigma^2$. Note first that

$$\mathbb{V}(xu) = \mathbb{E}(u^2xx') = \mathbb{E}(\mathbb{E}(u^2|x)xx') = \sigma^2\mathbb{E}(xx')$$

and thus

Proposition 6.2 (Asymptotic Covariance Under Homoskedasticity).

$$\Sigma = \sigma^2\mathbb{E}(xx')^{-1}.$$

Proposition 6.3 (Consistent Estimators).

(i) *The estimator*

$$\hat{\sigma}^2 := \frac{1}{n-k} \sum \hat{u}_i^2 = \frac{\text{SSR}}{n-k}.$$

is consistent and unbiased.

(ii) *The estimator*

$$\hat{\Sigma} := \hat{\sigma}^2 \left(\frac{1}{n} \sum x_i x_i' \right)^{-1}$$

is consistent.

Proof.

(i) Note first that $\hat{U} = M_X Y = M_X(X\beta + U) = M_X U$. We have thus that

$$\text{SSR} = \|M_X U\|^2 = U'U - U'P_X U.$$

Thus,

$$\begin{aligned} \hat{\sigma}^2 &= \frac{n}{n-k} [U'U - U'X(X'X)^{-1}X'U] \\ &= \frac{n}{n-k} \left[\frac{1}{n} \sum u_i^2 - \left(\frac{1}{n} \sum u_i x_i' \right) \left(\frac{1}{n} \sum x_i x_i' \right)^{-1} \left(\frac{1}{n} \sum x_i u_i \right) \right] \\ &\xrightarrow{p} \sigma^2 - 0 \cdot \mathbb{E}(xx')^{-1} \cdot 0 = \sigma^2 \end{aligned}$$

by the continuous mapping theorem. The second equality can be derived using the law of large numbers; the constant can be dealt with using the continuous mapping theorem.

Next, we show that $\hat{\sigma}^2$ is unbiased. Note that

$$\begin{aligned} \mathbb{E}(\text{SSR} | X) &= \mathbb{E}(U'U | X) - \mathbb{E}(U'P_X U | X) \\ &= \sum \mathbb{E}(u_i^2 | X) - \mathbb{E}(U'P_X U | X) \\ &= n\sigma^2 - \mathbb{E}(U'P_X U | X). \end{aligned}$$

Now, since $U'P_XU$ is a scalar,

$$\begin{aligned}\mathbb{E}(U'P_XU|X) &= \mathbb{E}(\text{tr}(U'P_XU)|X) = \mathbb{E}(\text{tr}(P_XUU'|X)) \\ &= \text{tr}(P_X\mathbb{E}(UU'|X)) = \text{tr}(P_X\sigma^2 I_n) \\ &= \sigma^2 \text{tr}(X(X'X)^{-1}X') = \sigma^2 \text{tr}((X'X)^{-1}X'X) \\ &= k\sigma^2.\end{aligned}$$

We used the properties that tr can be exchanged with \mathbb{E} (since it is a finite summation), that tr is linear, and that tr is invariant under cyclic permutations. Combining the above gives $\mathbb{E}(\text{SSR}|X) = (n-k)\sigma^2$.

- (ii) The weak law of large numbers gives $n^{-1} \sum x_i x_i' \xrightarrow{p} \mathbb{E}(xx')$. It remains to apply the continuous mapping theorem.

□

6.2 Estimation of Σ : Heteroskedasticity

If we can observe u_i , we may just use the estimator

$$\hat{\Sigma}^{\text{ideal}} := \left(\frac{1}{n} \sum x_i x_i' \right)^{-1} \left(\frac{1}{n} \sum u_i^2 x_i x_i' \right) \left(\frac{1}{n} \sum x_i x_i \right)^{-1} \xrightarrow{p} \Sigma.$$

But we cannot observe u_i in practice.

Lemma 6.4. *Let $\{Z_i\}_{i \geq 1}$ be a sequence of identically distributed random vectors such that $\mathbb{E}(\|Z_i\|^r) < \infty$. Then,*

$$\frac{\max_{1 \leq i \leq n} \|Z_i\|}{n^{1/r}} \xrightarrow{p} 0.$$

Proof. Fix $\epsilon > 0$ and note that

$$\begin{aligned}\mathbb{P} \left(\max_{1 \leq i \leq n} \|Z_i\| > \epsilon n^{1/r} \right) &= \mathbb{P} \left(\bigcup_{i=1}^n \{\|Z_i\|^r > \epsilon^r n\} \right) \\ &\leq \sum_{i=1}^n \mathbb{P}(\|Z_i\|^r > \epsilon^r n) \\ &= \sum_{i=1}^n \mathbb{P}(\|Z_i\|^r \mathbf{1}(\|Z_i\|^r > \epsilon^r n) > \epsilon^r n).\end{aligned}$$

Using Markov's inequality, the right side of the preceding display is bounded above by

$$\frac{1}{n\epsilon^r} \sum_{i=1}^n \mathbb{E} [\|Z_i\|^r \mathbf{1}(\|Z_i\|^r > \epsilon^r n)] = \frac{1}{\epsilon^r} \mathbb{E} [\|Z_i\|^r \mathbf{1}(\|Z_i\|^r > \epsilon^r n)] \longrightarrow 0.$$

Convergence comes from the dominated convergence theorem, since $\mathbb{E}(\|Z_i\|^r) < \infty$.

□

Proposition 6.5. *The estimator*

$$\hat{\Sigma} := \left(\frac{1}{n} \sum x_i x_i' \right)^{-1} \left(\frac{1}{n} \sum \hat{u}_i^2 x_i x_i' \right) \left(\frac{1}{n} \sum x_i x_i \right)^{-1}$$

is consistent. Note that the estimator can equivalently be written as

$$\hat{\Sigma} \equiv n(X'X)^{-1} X' \hat{\Omega} X (X'X)^{-1},$$

where

$$\hat{\Omega} := \begin{pmatrix} \hat{u}_1^2 & & 0 \\ & \hat{u}_2^2 & \\ & & \ddots \\ 0 & & & \hat{u}_n^2 \end{pmatrix}.$$

Proof. In light of the continuous mapping theorem, we need only show

$$\frac{1}{n} \sum \hat{u}_i^2 x_i x_i' \longrightarrow \mathbb{E}(u^2 x x').$$

Since $n^{-1} \sum u_i^2 x_i x_i' \xrightarrow{p} \mathbb{E}(u^2 x x')$, we need only show

$$\frac{1}{n} \sum (\hat{u}_i^2 - u_i^2) x_i x_i' \xrightarrow{p} 0.^3$$

We show this by proving each element of the matrix converges to 0 in probability. Fix j, k and observe that the (j, k) element of the matrix

$$\left| \frac{1}{n} \sum (\hat{u}_i^2 - u_i^2) x_{ij} x_{ik} \right| \leq \max_{1 \leq i \leq n} |\hat{u}_i^2 - u_i^2| \cdot \frac{1}{n} \sum |x_{ij} x_{ik}|.$$

Since $n^{-1} \sum |x_{ij} x_{ik}| \xrightarrow{p} \mathbb{E}(|x_{ij} x_{ik}|)$, it is sufficient to prove

$$\max_{1 \leq i \leq n} |\hat{u}_i^2 - u_i^2| \xrightarrow{p} 0.$$

To that end, note that since $\hat{u}_i = y_i - x_i' \hat{\beta}_n = x_i'(\beta - \hat{\beta}_n) + u_i$, we have

$$\begin{aligned} |\hat{u}_i^2 - u_i^2| &= |x_i'(\beta - \hat{\beta}_n)(\hat{u}_i + u_i)| \\ &= |x_i'(\beta - \hat{\beta}_n)(x_i'(\beta - \hat{\beta}_n) + 2u_i)| \\ &\leq (x_i'(\beta - \hat{\beta}_n))^2 + 2|u_i x_i'(\beta - \hat{\beta}_n)| \end{aligned}$$

Using the Cauchy-Schwarz inequality,

$$\begin{aligned} \max |\hat{u}_i^2 - u_i^2| &\leq \|\beta - \hat{\beta}_n\|^2 \max \|x_i\|^2 + 2\|\beta - \hat{\beta}_n\| \max \|x_i u_i\| \\ &= \|\sqrt{n}(\beta - \hat{\beta}_n)\|^2 \frac{\max \|x_i\|^2}{n} + 2\|\sqrt{n}(\beta - \hat{\beta}_n)\| \frac{\max \|x_i u_i\|}{\sqrt{n}}. \end{aligned}$$

Since $\sqrt{n}(\beta - \hat{\beta}_n)$ converges in distribution, we need only show the other two terms converges in probability to 0. This is given by the previous lemma. \square

³We sometimes write $X_n = o_p(1)$ to mean $X_n \xrightarrow{p} 0$.

7 Partitioned Regression

Proposition 7.1. Consider the model $Y = X_1\beta_1 + X_2\beta_2 + U$. Let $\tilde{X}_2 := M_{X_1}X_2$ be the vector of residuals from a regression of (each column of) X_2 on X_1 . Denote $\tilde{Y} := M_{X_1}Y = \hat{U}$ similarly. Then, the OLS estimator $\hat{\beta}_2$ equals:

- The OLS estimator of Y on \tilde{X}_2 .
- The OLS estimator of \tilde{Y} on \tilde{X}_2 .

In particular, we have the formula

$$\hat{\beta}_2 = (X_2'M_{X_1}X_2)^{-1}X_2'M_{X_1}Y.$$

Remark 7.2. This is sometimes called “controlling for” or “partialling out” X_1 . ☕

Proof. Note that

$$X_2'M_{X_1}Y = X_2'M_{X_1}X_2\beta_2 + X_2'M_{X_1}\hat{U}.$$

The last term is 0 since

$$X_2'M_{X_1}\hat{U} = X_2'(\hat{U} - P_{X_1}\hat{U}) = X_2'\hat{U} = 0.$$

(We pre-multiply by M_{X_1} to get rid of the X_1 term, and pre-multiply by X_2' to get rid of the \hat{U} term.) Thus,

$$\begin{aligned}\hat{\beta}_2 &= (X_2'M_{X_1}X_2)^{-1}X_2'M_{X_1}Y \\ &= (\tilde{X}_2'\tilde{X}_2)^{-1}\tilde{X}_2'\tilde{Y} = (\tilde{X}_2'\tilde{X}_2)^{-1}\tilde{X}_2'M_{X_1}Y.\end{aligned}$$

□

7.1 Components of the Variance

Thus, if MLR.5 holds, then

$$\mathbb{V}(\hat{\beta}_2|X) = \sigma^2(\tilde{X}_2'\tilde{X}_2)^{-1}.$$

In particular, if X_2 contains a single regressor x_j with estimated coefficient $\hat{\beta}_j$, then

$$\mathbb{V}(\hat{\beta}_j|X) = \frac{\sigma^2}{\text{SSR}_j} = \frac{\sigma^2}{\text{SST}_j(1 - R_j^2)}$$

where SSR_j denotes the SSR of a regression of x_j on all the other regressors.

Remark 7.3. We have smaller variance if sample size is large, if R_j^2 is small, and if x_j is spread out.

- When SST_j is large, $\mathbb{V}(\hat{\beta}_j|X)$ is small. SST_j increases when sample size increases or when x_j become more spread out. When x_j is clustered, small variations in y can drastically change the slope.

- When R_j^2 is large, $\mathbb{V}(\hat{\beta}_j|X)$ is large. The more variability in x_j can be accounted for by the other regressors, the harder it is to estimate β_j . When there is high correlation between X_1 and X_2 —when there is near perfect multi-collinearity—it is hard to partial out the effect of one from the other, since there is less variations in x_j left after partialling out the other covariates.
- When adding more regressors, R_j^2 increases, and $\mathbb{V}(\hat{\beta}_j|X)$ increases.



7.2 Omitted Variable Bias

Proposition 7.4 (Long and Short Regression). *Suppose x_1 and x_2 has dimension d_1 and d_2 . Consider the short regression*

$$y = x_1' \beta_{1s} + u_s, \quad \mathbb{E}[xu_s] = 0$$

and the long regression

$$y = x_1' \beta_{1l} + x_2' \beta_{2l} + u_l, \quad \mathbb{E}[xu_l] = 0.$$

Then,

$$\beta_{1a} = \beta_{1l} + \alpha \beta_{2l},$$

where α is a $d_1 \times d_2$ dimensional matrix, the j th column of which is the population regression coefficient vector from regressing the j th component of x_2 onto x_1 .

Proof. Note that

$$\begin{aligned} \beta_{1s} &= \mathbb{E}[x_1 x_1']^{-1} \mathbb{E}[x_1 (x_1' \beta_{1l} + x_2' \beta_{2l} + u_l)] \\ &= \beta_{1l} + \underbrace{\mathbb{E}[x_1 x_1'] \mathbb{E}[x_1 x_2']}_{\alpha} \beta_{2l}. \end{aligned}$$

□

Example 7.5. Consider

$$y = \beta_0 + \beta_1 x_1 + u, \quad \mathbb{E}[xu] = 0.$$

Partialling out the constant, we have

$$\tilde{y} = y - \bar{y}, \quad \tilde{x}_1 = x_1 - \bar{x}_1.$$

Then,

$$\beta_1 = \frac{\mathbb{E}[\tilde{x}_1 \tilde{y}]}{\mathbb{E}[\tilde{x}_1^2]} = \frac{\mathbb{C}(y, x_1)}{\mathbb{V}(x_1)}.$$



Example 7.6 (Omitted Variable Bias). Suppose x_1 and x_2 are scalar random variables and

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u,$$

where $(\beta_0, \beta_1, \beta_2)$ represents the best linear predictor of y given x_1 and x_2 . Suppose we omit x_2 and instead specify the regression

$$y = b_0 + b_1 x_1 + v,$$

where (b_0, b_1) represents the best linear predictor of y given x_1 . We have then that

$$\begin{aligned} b_1 &= \frac{\mathbb{C}(y, x_1)}{\mathbb{V}(x_1)} \\ &= \frac{\beta_1 \mathbb{V}(x_1) + \beta_2 \mathbb{C}(x_1, x_2)}{\mathbb{V}(x_1)} = \beta_1 + \beta_2 \frac{\mathbb{C}(x_1, x_2)}{\mathbb{V}(x_1)}, \end{aligned}$$

where the last term is sometimes called the **omitted variable bias**. The sign of this bias is determined by the signs of $\mathbb{C}(x_1, x_2)$ and β_2 . If both are positive, then the bias is positive.

But note that in so saying, we are assuming that the long regression is the true model and interpreting the regression causally. We do not need to include covariates that:

- do not affect the dependent variable y , or
- are not correlated with the independent variable x_1 .



Example 7.7. Consider the models

$$y = b_0 + b_1 x_1 + v,$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u.$$

If we interpret them as causal models, we may write $v = \beta_2 x_2 + \beta_3 x_3 + u$. In a prediction context, however, this is incorrect since b_1 is general is not equal to β_1 . We have

$$\hat{b}_1 = \hat{\beta}_1 + \hat{\beta}_2 \hat{\delta}_1 + \hat{\beta}_3 \hat{\gamma}_1.$$

This shows that \hat{b}_1 is both biased and inconsistent *as an estimator of β_1* .



Example 7.8. Consider $X_3 = [1 \text{ sqft}_i]$ and $X_4 = [\text{bdrm}_i]$. In the regression $Y_i = X_3 \delta_0 + X_4 \delta_1$ we have

$$\hat{\delta}_1 = (X_4' M_{X_3}' M_{X_3} X_4)^{-1} X_4' M_{X_3}' Y = (\hat{r}' \hat{r})^{-1} \hat{r}' Y.$$

where $\hat{r} := M_{X_3} X_4$ is the vector of residuals in the regression

$$A : \quad \text{bdrm}_s = \alpha_0 + \alpha_1 \text{sqft}_i + r_i.$$

Note that $\hat{\delta}_1$ is proportional to the partial correlation between price and the number of bedrooms. In particular,

$$\hat{\delta}_1 = \frac{\hat{r}' Y}{\hat{r}' \hat{r}} = \frac{\hat{r}' Y}{\text{SSR}_A}.$$

Thus, noting that \hat{r} only depends on X , we have under MLR.5 that

$$\mathbb{V}(\hat{\delta}_1|X) = \frac{\hat{r}' \mathbb{V}(Y|X) \hat{r}}{\text{SSR}_A^2} = \frac{\sigma^2}{\text{SSR}_A} = \frac{\sigma^2}{\text{SST}_A(1 - R_A^2)}.$$

Now consider the regression

$$C : Y_i = \gamma_0 + \text{bdrm}_i \gamma_1 + u_i.$$

We have

$$\hat{\gamma}_1 = (X_4' M_c X_4)^{-1} X_4' M_c [c \hat{\beta}_0 + X_4 \hat{\beta}_1 + X_s \hat{\beta}_2 + \hat{\epsilon}],$$

where $X_s = [\text{sqft}_i]$. Then,

$$\hat{\gamma}_1 = \hat{\beta}_1 + (X_4' M_c X_4)^{-1} X_4' M_c X_s \hat{\beta}_2 + (X_4' M_c X_4)^{-1} X_4' M_c \hat{\epsilon},$$

where the last term is 0 by the normal equations. So,


$$\hat{\gamma}_1 = \hat{\beta}_1 + \hat{\beta}_2 \frac{\widehat{C}(\text{bdrm}_i, \text{sqft}_i)}{\widehat{V}(\text{bdrm}_i)} = \hat{\beta}_1 + \hat{\beta}_2 \hat{\alpha}_1.$$

We have *no* omitted variable bias if:

- $\hat{\alpha}_1 = 0$: bdrm does not vary linearly with sqft, so we can't falsely attribute the effect of sqft on the price of bdrm.
- $\hat{\beta}_2 = 0$: sqft does not affect price anyway.

Slightly different comparison: $\mathbb{E}(\hat{\gamma}_1|X)$ vs β_1 . Suppose MLR.1-4 hold in $Y = X\beta + \epsilon$. Then

$$\mathbb{E}(\hat{\gamma}_1|X) = \mathbb{E}(\hat{\beta}_1 + \hat{\beta}_2 \hat{\alpha}_1|X) = \beta_1 + \beta_2 \hat{\alpha}_1.$$

We have *no* omitted variable bias unless $\beta_2 = 0$ or $\hat{\alpha}_1 = 0$. 

7.3 The Population Case

Suppose we partition x into (x_1, x_2) and β into $\beta = (\beta_1, \beta_2)$, each with dimensions (k_1, k_2) and write

$$y = x_1' \beta_1 + x_2' \beta_2 + u, \quad \mathbb{E}(xu) = 0.$$

Let $\tilde{\beta}_1$ represent the best linear predictor of y given x_1 and write

$$y = x_1' \tilde{\beta}_1 + \tilde{y}, \quad \mathbb{E}(x_1 \tilde{y}) = 0.$$

For each j write

$$x_{2j} = \tilde{\gamma}_j' x_1 + \tilde{x}_{2j}, \quad \mathbb{E}(x_1 \tilde{x}_{2j}) = 0.$$

Write

$$\tilde{\gamma} = \begin{bmatrix} -\tilde{\gamma}_1' - \\ \vdots \\ -\tilde{\gamma}_{k_2}' - \end{bmatrix}.$$

Then

$$x_2 = \tilde{\gamma}x_1 + \tilde{x}_2.$$

Finally, consider

$$\tilde{y} = \tilde{x}_2'\bar{\beta}_2 + v, \quad \mathbb{E}(\tilde{x}_2v) = 0.$$

We have that:

Theorem 7.9 (Yule-Frisch-Waugh-Lovell). β_2 is both the best linear predictor of \tilde{y} given \tilde{x}_2 and the best linear predictor of y given \tilde{x}_2 .

Proof. Since $\mathbb{E}(\tilde{x}_2x_1') = 0$ (as \tilde{x}_2 is the residual of a linear projection of x_2 onto x_1), we have

$$\bar{\beta}_2 := \mathbb{E}(\tilde{x}_2\tilde{x}_2')^{-1}\mathbb{E}(\tilde{x}_2\tilde{y}) = \mathbb{E}(\tilde{x}_2\tilde{x}_2')^{-1}\mathbb{E}(\tilde{x}_2y).$$

Thus $\bar{\beta}_2$ represents both the best linear predictor of y given \tilde{x}_2 and the best linear predictor of \tilde{y} given \tilde{x}_2 . We show next that $\bar{\beta}_2 = \beta_2$:

$$\begin{aligned} \bar{\beta}_2 &= \mathbb{E}(\tilde{x}_2\tilde{x}_2')^{-1}\mathbb{E}(\tilde{x}_2[x_1'\beta_1 + x_2'\beta_2 + u]) \\ &= \mathbb{E}(\tilde{x}_2\tilde{x}_2')^{-1}\mathbb{E}(\tilde{x}_2x_2')\beta_2 + \mathbb{E}(\tilde{x}_2\tilde{x}_2')^{-1}\mathbb{E}(\tilde{x}_2u) \\ &= \mathbb{E}(\tilde{x}_2\tilde{x}_2')^{-1}\mathbb{E}(\tilde{x}_2\tilde{x}_2')\beta_2 + \mathbb{E}(\tilde{x}_2\tilde{x}_2')^{-1}\mathbb{E}([x_2 - \tilde{\gamma}x_1]u) \\ &= \beta_2, \end{aligned}$$

where the third equality follows from

$$\mathbb{E}(\tilde{x}_2\tilde{x}_2') = \mathbb{E}(\tilde{x}_2[x_2 - \tilde{\gamma}x_1]') = \mathbb{E}(\tilde{x}_2x_2'),$$

and the fourth because

$$\mathbb{E}(xu) = \mathbb{E}\left(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} u\right) = 0.$$

□

Remark 7.10.

- We can think of $x_2'\beta_2$ as the best linear predictor of y given x_2 after “controlling for” x_1 .
- $x_2'\beta_2$ is generally not the best linear predictor of y given x_2 :



8 Inference

8.1 Finite Sample, Homoskedasticity

Suppose $Y|X \sim \mathcal{N}(X\beta, \sigma^2 I_n)$, so that

$$Y = X\beta + U, \quad \mathbb{E}(U|X) = 0, \quad \mathbb{V}(U|X) = \sigma^2 I_n.$$

This set of assumptions can be generated by the following:

- (y_i, x_i) are multivariate normal. (We have $(y_1, \dots, y_n, x'_1, \dots, x'_n)$ is multivariate normal, and hence $Y|x_1, \dots, x_n$ is multivariate normal.)
- MLR.1–5 with normality: MLR.6: $y_i|x_i \sim \mathcal{N}(x'_i\beta, \sigma^2)$.

Under these assumptions, we have

$$(X'X)^{-1}X'Y|X \sim \mathcal{N}(\beta, \sigma^2(X'X)^{-1}).$$

Thus

$$\hat{\beta}_j|X \sim \mathcal{N}\left(\beta_j, [\sigma^2(X'X)^{-1}]_{j,j}\right)$$

and we may write

$$\frac{\hat{\beta}_j - \beta_j}{\sigma \sqrt{e'_j(X'X)^{-1}e_j}}|X \sim \mathcal{N}(0, 1),$$

where e_j counts the $(j+1)$ -th column of the identity matrix (so e_0 is the first column).

However, we usually need to estimate σ^2 : Recall that we have an unbiased and consistent estimator of σ^2 :

$$\hat{\sigma}^2 := \frac{\text{SSR}}{n-k}.$$

Note we also have the following result:

Proposition 8.1. *Under homoskedasticity, the estimator*

$$\hat{\sigma}^2 := \frac{\text{SSR}}{n-k} \sim \frac{\sigma^2}{n-k} \cdot \chi^2_{n-k}$$

is unbiased, consistent, and independent of $\hat{\beta}$.

Proof (*Sketch, Intuition*). Since

$$(n-k)\hat{\sigma}^2 = \text{SSR} = (M_X Y)'(M_X Y) = U' M_X U.$$

It turns out that

$$\underbrace{U'U}_{\sigma^2 \chi^2_n} = \underbrace{U'P_X U}_{\sigma^2 \chi^2_k} + \underbrace{U'M_X U}_{\sigma^2 \chi^2_{n-k}}.$$

□

8.1.1 Hypothesis on a Single Coefficient

Consequently, we have

$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma} \sqrt{e'_j (X'X)^{-1} e_j}} \bigg| X \sim t_{n-k}.$$

The test

$$\phi_n := \mathbb{1} \left(\left| \frac{\hat{\beta}_j - \beta_j^0}{\hat{\sigma} \sqrt{e'_j (X'X)^{-1} e_j}} \right| > t_{n-k, 1-\alpha/2} \right)$$

has rejection probability α under the null hypothesis $H_0 : \beta_j = \beta_j^0$. The rejection probability is strictly greater than α when H_0 is false. The power curve achieves its minimum α at $\beta_j = \beta_j^0$ and asymptotes at 1.

Definition 8.2. $se \hat{\beta}_j := \hat{\sigma} \sqrt{e'_j (X'X)^{-1} e_j}$ is called the **standard error** of $\hat{\beta}_j$.

Conditional on the data, the p -value is given by

$$\begin{aligned} \hat{p}_n &:= \inf_{\alpha \in (0,1)} \{ |T_n| > t_{n-k, 1-\frac{\alpha}{2}} \} \\ &= \inf_{\alpha \in (0,1)} \left\{ |T_n| > F^{-1} \left(1 - \frac{\alpha}{2} \right) \right\}. \end{aligned}$$

In this case, since F^{-1} is strictly increasing and continuous, we get

$$\alpha = 2[1 - F(|T_n|)] = 2F(-|T_n|).$$

Definition 8.3. The p -value is the smallest significance level at which we would reject the null hypothesis.

Remark 8.4. The introductory definition of p being the probability of getting more “extreme” test-statistics only works if the null is a singleton. ☕

8.2 Heteroskedasticity

We assume MLR.1–4 hold, but not MLR.5. Recall that

$$\begin{aligned} \mathbb{V}(\hat{\beta}|X) &= \mathbb{V}(\beta + (X'X)^{-1} X'U|X) \\ &= (X'X)^{-1} X' \Omega X (X'X)^{-1}, \end{aligned}$$

where $\Omega := \mathbb{V}(U|X)$.

When $y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik} + u_i$, we have by Yule-Frisch-Waugh-Lovell,

$$\hat{\beta}_j = \frac{\sum_i \hat{r}_{ij} y_i}{\sum_i \hat{r}_{ij}^2},$$

where \hat{r}_{ij} is the i -th residual of a regression of x_j on all other regressors. Thus, denoting $\sigma_i := \mathbb{V}(y_i|x_i)$, we have

$$\mathbb{V}(\hat{\beta}_j|X) = \frac{\sum_i \hat{r}_{ij}^2 \sigma_i^2}{\left(\sum_i \hat{r}_{ij}^2\right)^2}.$$

Estimating σ_i^2 with \hat{u}_i^2 , we have the **heteroskedasticity robust standard error**:

$$\text{se } \hat{\beta}_j := \sqrt{\frac{\sum_i \hat{r}_{ij}^2 \hat{u}_i^2}{(\sum_i \hat{r}_{ij}^2)^2}}.$$

Remark 8.5. In the simple regression $y_i = \beta_0 + \beta_1 x_i + u_i$, we have

$$\text{se } \hat{\beta}_1 = \sqrt{\frac{\sum \hat{u}_i^2 (x_i - \bar{x})^2}{(\sum (x_i - \bar{x})^2)^2}}.$$

This equals the homoskedasticity standard error when \hat{u}_i^2 is constant. ☕

Remark 8.6. The heteroskedasticity robust standard error is not necessarily larger. If, for example, u has small variance for small and large x , then $\hat{\beta}$ tend to have smaller variance than the homoskedasticity standard error. If u has large variance for small and large x , then $\hat{\beta}$ has large variance. ☕

Remark 8.7. People often use OLS to estimate $\hat{\beta}$ and then use the heteroskedasticity robust standard errors to estimate $\text{se } \hat{\beta}$. Note that the OLS estimator is no longer efficient when the errors are heteroskedastic.

The only advantage of computing homoskedasticity standard errors seems to be that it gives a exact distribution (instead of an approximation by the central limit theorem, though for large n , $t_{n-k} \sim \mathcal{N}$, and this advantage is moot), at the cost of assuming the *very strong* assumptions of MLR.5 and MLR.6. ☕

8.3 Linear Restrictions

8.3.1 A Single Linear Restriction

Suppose

$$\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{d} \mathcal{N}(0, V),$$

where $V = \mathbb{E}(xx')^{-1} \mathbb{E}(u^2 xx') \mathbb{E}(xx')^{-1}$. Suppose V is non-singular and $\hat{V}_n \xrightarrow{p} V$ is a consistent estimator of V .

Consider

$$H_0 : r' \beta = c \quad \text{vs} \quad H_1 : r' \beta \neq c.$$

The continuous mapping theorem gives

$$\sqrt{n}(r' \hat{\beta} - r' \beta) \xrightarrow{d} \mathcal{N}(0, r' V r),$$

and so by Slutsky's theorem we have the asymptotic pivot


$$T_n := \frac{\sqrt{n}(r' \hat{\beta} - r' \beta)}{\sqrt{r' \hat{V}_n r}} \xrightarrow{d} \mathcal{N}(0, 1).$$

- The test $\phi_n := \mathbb{1}(|T_n| > z_{1-\alpha/2})$ is of asymptotic size α .
- We can use $\mathbb{1}(T_n > z_{1-\alpha})$ to test $H_0 : r'\beta \leq c$.
- An asymptotic $1 - \alpha$ confidence interval for $r'\beta$ is

$$\left[r'\hat{\beta} - z_{1-\frac{\alpha}{2}} \sqrt{\frac{r'\hat{V}_n r}{n}}, r'\hat{\beta} + z_{1-\frac{\alpha}{2}} \sqrt{\frac{r'\hat{V}_n r}{n}} \right]$$

Example 8.8. We may want to test the hypothesis that firms exhibit constant returns to scale. With the production function $Q = AK^{\beta_1}L^{\beta_2}U$. We may write

$$\log Q = \beta_0 + \beta_1 \log A + \beta_2 \log K + \beta_3 \log L + u$$

and test $H_0 : \beta_1 + \beta_2 = 1$. 

8.3.2 Multiple Linear Restrictions

Proposition 8.9. A positive definite and symmetric matrix A has a square root $A^{1/2}$ with inverse $A^{-1/2} = (A^{-1})^{1/2}$.

Consider

$$H_0 : R\beta = c \quad \text{vs} \quad H_1 : R\beta \neq c,$$

where R is a $p \times k$ -dimensional matrix of full row rank (so none of the restrictions are redundant) and c is a $p \times 1$ vector. The continuous mapping theorem gives

$$\sqrt{n}(R\hat{\beta} - R\beta) \longrightarrow \mathcal{N}(0, RVR').$$

Note that RVR' is full rank (because R and V are) and hence positive definite (because V is). Slutsky's Theorem then gives

$$(R\hat{V}_n R')^{-1/2} \sqrt{n}(R\hat{\beta}_n - R\beta) \longrightarrow \mathcal{N}(0, I_p),$$

since

$$(RVR')^{-1/2} RVR' (RVR')^{-1/2} = I_p.$$

We may use the contours of the multivariate normal distribution to test the hypothesis.

Alternatively, note that the continuous mapping theorem gives

Proposition 8.10 (Testing Multiple Linear Restrictions). *If \hat{V} is a consistent estimator of the asymptotic variance of $\hat{\beta}$, then*

$$\begin{aligned} n \cdot (R\hat{\beta}_n - R\beta)' (R\hat{V}_n R')^{-1/2} (R\hat{V}_n R')^{-1/2} (R\hat{\beta}_n - R\beta) \\ = n \cdot (R\hat{\beta}_n - R\beta)' (R\hat{V}_n R')^{-1} (R\hat{\beta}_n - R\beta) \xrightarrow{d} \chi_p^2. \end{aligned}$$

We may use this asymptotic pivot to derive a confidence set for $R\beta$ and conduct hypothesis tests.

- An asymptotic $1 - \alpha$ confidence set for $R\beta$ is given by

$$C_n := \left\{ c \in \mathbb{R}^p : n \cdot (R\hat{\beta} - c)'(R\hat{V}_n R')^{-1}(R\hat{\beta} - c) \leq \chi_{p,1-\alpha}^2 \right\}.$$

Since $(RV R')^{-1}$ is positive definite, this is an ellipsoid in \mathbb{R}^p centered at $R\hat{\beta}$.

- Under H_0 , we can reject if

$$T_n := n \cdot (R\hat{\beta}_n - c)'(R\hat{V}_n R')^{-1}(R\hat{\beta}_n - c) > \chi_{p,1-\alpha}^2.$$

Remark 8.11. If R_1 and R_2 are related by a invertible linear transformation, then the tests conducted by R_1 and R_2 are equivalent. The way in which we write the restrictions does not matter. ☕

Remark 8.12. The finite sample homoskedasticity version of this is called a F -test. To see this, note that if $U_2 \sim \chi_{d_2}^2$, then as $d_2 \rightarrow \infty$, $U_2/d_2 \xrightarrow{p} 1$. Thus, $F = (U_1/d_1)/(U_2/d_2) \approx U_1/d_1 = \chi_{d_1}^2/d_1$. ☕

9 Regression Specification

9.1 Weighted Regression

If we know the error variance σ_i^2 , we could transform the model

$$\frac{y_i}{\sigma_i} = \left(\frac{1}{\sigma_i}, \frac{x_{i1}}{\sigma_i}, \dots, \frac{x_{ik}}{\sigma_i} \right) \beta + \frac{u_i}{\sigma_i}.$$

This restores MLR.5, since $\mathbb{E}((u_i/\sigma_i)^2|x_i) = \mathbb{E}(u_i^2|x_i)/\sigma_i^2 = 1$. Gauss-Markov then hold, since MLR.1–4 still hold under the transformation.

Example 9.1. Suppose the model for observation i is

$$y_i = \beta_0 + \beta_1 x_i + u_i,$$

where MLR.1–5 hold. Let $\mathbb{V}(u|x) = \sigma^2$. For groups $k = 1, \dots, K$, we observe

$$\bar{y}_k = \frac{1}{n_k} \sum_{i \in \text{group } k} y_i, \quad \bar{x}_k = \frac{1}{n_k} \sum_{i \in \text{group } k} x_i.$$

The model for averages is

$$\bar{y}_k = \beta_0 + \beta_1 \bar{x}_k + \bar{u}_k, \quad \mathbb{V}(\bar{u}_k|X) = \frac{\sigma^2}{n_k}.$$

We transform the model to restore MLR.5:

$$\sqrt{n_k} \cdot \bar{y}_k = \beta_0 \sqrt{n_k} + \beta_1 \sqrt{n_k} \cdot \bar{x}_k + \sqrt{n_k} \cdot \bar{u}_k, \quad \mathbb{V}(\sqrt{n_k} \cdot \bar{u}_k|X) = \sigma^2.$$



9.2 Log Specifications

Recall the approximation

$$\log x' - \log x = \log \left(\frac{x'}{x} \right) \approx \frac{x' - x}{x}.$$

9.2.1 Level-Log

We have $\mathbb{E}(y|x = t) \approx \alpha + \beta \log t$.

$$\mathbb{E}(y|x = t + \Delta t) - \mathbb{E}(y|x = t) \approx \beta \frac{\Delta t}{t}.$$

So $\beta/100$ is approximately the change in the conditional mean of y when t increases by 1%.

9.2.2 Log-Level

We have $\mathbb{E}(\log y|x = t) \approx \alpha + \beta t$.

$$\begin{aligned}\beta &\approx \mathbb{E}(\log y|x = t + 1) - \mathbb{E}(\log y|x = t) \\ &\approx \log \mathbb{E}(y|x = t + 1) - \log \mathbb{E}(y|x = t) \approx \frac{\mathbb{E}(y|x = t + 1) - \mathbb{E}(y|x = t)}{\mathbb{E}(y|x = t)}.\end{aligned}$$

So 100β is approximately the percentage change in the mean of y resulting from a unit increase in x .

Remark 9.2. We are committing a sin by interchanging \mathbb{E} with \log . If this is a causal model (note that we need a lot more assumptions to say this), then we may differentiate the actual conditional mean function: Differentiating $\log y = \alpha + \beta t$ gives

$$\frac{1}{y} \frac{\partial y}{\partial t} = \beta_1,$$

and so

$$\frac{100\Delta y}{y} \approx 100t \cdot \Delta t.$$



A less bad interpretation is as follows:

$$\begin{aligned}\exp \beta - 1 &\approx \exp(\mathbb{E}(\log y|x = t + 1) - \mathbb{E}(\log y|x = t)) - 1 \\ &= \frac{\exp(\mathbb{E}(\log y|x = t + 1)) - \exp(\mathbb{E}(\log y|x = t))}{\exp(\mathbb{E}(\log y|x = t))}.\end{aligned}$$

Thus when $\beta \approx 0$ we have $\beta \approx \exp(\beta) - 1$ is the proportional change in the conditional geometric mean of y given a unit increase in x .

9.2.3 Log-Log

We have $\mathbb{E}(\log(y)|x = t) \approx \alpha + \beta \log t$. Then,

$$\mathbb{E}(\log(y)|x = t + \Delta t) - \mathbb{E}(\log(y)|x = t) \approx \beta[\log(t + \Delta t) - \log t],$$

which gives

$$\frac{\mathbb{E}(y|x = t + \Delta t) - \mathbb{E}(y|x = t)}{\mathbb{E}(y|x = t)} \approx \beta \cdot \frac{\Delta t}{t}.$$

So a 1% increase in x is associated with a $\beta\%$ change in the conditional mean of y (also known as an “elasticity”).

9.3 Functional Forms

Example 9.3 (Polynomials). Consider the approximation

$$\mathbb{E}[\text{wage}|\text{exper}] \approx \beta_0 + \beta_1 \text{exper} + \beta_2 \text{exper}^2.$$

- We may estimate the turning point to be $-\hat{\beta}_1/2\hat{\beta}_2$.
- The effect of experience diminishes if $\beta_2 < 0$.
- The average partial effect can be estimated using

$$\frac{1}{n} \sum (\hat{\beta}_1 + 2\hat{\beta}_2 \text{exper}_i) = \hat{\beta}_1 + 2\hat{\beta}_2 \overline{\text{exper}}_n.$$



Example 9.4 (Kinks and Discontinuities).

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + d_i(\gamma + \delta x_i) + \epsilon_i \\ &= \beta_0 + \beta_1 x_i + \gamma d_i + \delta d_i x_i + \epsilon_i. \end{aligned}$$

If $d_i = \mathbb{1}(x_i \geq \bar{x})$, then the jump is $\gamma + \delta \cdot \bar{x}$.



9.4 Dummy Variables

Example 9.5 (Dummy Variables). For a continuous random variable, we transform it into a categorical variable described by mutually exclusive ($d_i d_j = 0$ when $i \neq j$) and exhaustive ($\sum d_j = 1$) dummies d_j . We may for instance estimate the group averages using

$$\mathbb{E}(y_i | d_1, \dots, d_k) = \sum \beta_j d_j,$$

where we omit the intercept to avoid perfect collinearity. We may also specify

$$\mathbb{E}(y_i | d_1, \dots, d_k) = \alpha_0 + \sum_{j=1}^{k-1} \alpha_j d_j$$

where α_j is the difference in mean between d_j and d_k (the base group).



Definition 9.6. We say a regression is **saturated** if it has one parameter for each possible value of the regressors.

10 Regression Cheat Sheet

Let x be a scalar. $y = \beta x + u$, $\mathbb{E}[xu] = 0$.

(i)

$$\beta = \frac{\mathbb{E}[xy]}{\mathbb{E}[x^2]}, \quad \hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2}.$$

(ii) If $\mathbb{E}[u|x] = 0$ and $\mathbb{E}[u^2|x] = \mathbb{E}[u^2]$, then

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}\left(0, \frac{\mathbb{E}[u^2]}{\mathbb{E}[x^2]}\right).$$

(iii) $\text{se } \hat{\beta} = \hat{\sigma} / \sqrt{\sum x_i^2}$.

Let x be a scalar. $y = \beta_0 + \beta_1 x + u$, $\mathbb{E}[xu] = 0$.

(i)

$$\beta_1 = \frac{\mathbb{C}(y, x)}{\mathbb{V}(x)}, \quad \hat{\beta}_1 = \frac{\sum (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x})^2}.$$

(ii) If $\mathbb{E}[u|x] = 0$ and $\mathbb{E}[u^2|x] = \mathbb{E}[u^2]$, then

$$\sqrt{n}(\hat{\beta}_1 - \beta_1) \xrightarrow{d} \mathcal{N}\left(0, \frac{\mathbb{E}[u^2]}{\mathbb{V}(x)}\right)$$

(iii) $R^2 = (\text{Corr}(y_i, x_i))^2$.

(iv) $\text{se } \hat{\beta}_1 = \hat{\sigma} / \sqrt{\sum x_i^2 - (\sum x_i)^2} = \hat{\sigma} / \sqrt{\sum (x_i - \bar{x})^2}$.

Let x be a $(k \times 1)$ vector. $y = x' \beta + u$, $\mathbb{E}[xu] = 0$.

(i)

$$\beta = \mathbb{E}[xx']^{-1} \mathbb{E}[xy'], \quad \hat{\beta} = (X'X)^{-1} X'Y.$$

(ii)

$$\mathbb{V}(\hat{\beta}|X) = (X'X)^{-1} X \Omega X (X'X)^{-1}, \quad \Omega := \mathbb{V}(u|X).$$

(iii) If $\mathbb{V}(xu)$ exists, then

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(0, \Sigma),$$

where

$$\begin{aligned} \Sigma &= \mathbb{E}[xx']^{-1} \mathbb{V}(xu) \mathbb{E}[xx']^{-1} \\ &= \mathbb{E}[xx']^{-1} \mathbb{E}[u^2 xx'] \mathbb{E}[xx']^{-1}. \end{aligned}$$

- (iv) $R^2 = (\text{Corr}(y_i, \hat{y}_i))^2$.
- (v) Under MLR.4 or if we include an intercept, $\mathbb{E}[U|X] = 0$. Consequently, $\Omega = \mathbb{E}[U'U|X]$.
- (vi) Under MLR.4 and MLR.5, $\Omega = \sigma^2 I_n$ and $\Sigma = \sigma^2 \mathbb{E}[xx']^{-1}$.
- (vii) Under MLR.5, we have the consistent and unbiased estimator

$$\hat{\sigma}^2 := \frac{\sum \hat{u}_i^2}{n-k} = \frac{\text{SSR}}{n-k} \sim \frac{\sigma^2}{n-k} \cdot \chi_{n-k}^2.$$

Consequently, we have the consistent estimator

$$\hat{\Sigma} := \sigma^2 \left(\frac{1}{n} \sum x_i x_i' \right)^{-1}.$$

- (viii) Without MLR.5, we have the consistent estimator

$$\hat{\Sigma} := n(X'X)^{-1} X' \hat{\Omega} X (X'X)^{-1} = \left(\frac{1}{n} \sum x_i x_i' \right)^{-1} \left(\frac{1}{n} \sum \hat{u}_i^2 x_i x_i' \right) \left(\frac{1}{n} \sum x_i x_i' \right)^{-1}.$$

- (ix) Write $x = (w', x_j)'$, where x_j is a scalar. Then, $\hat{\beta}_j = (\hat{V}'\hat{V})^{-1}\hat{V}'Y$, where \hat{V} is the vector of residuals of a regression of x_j on w . Thus, if MLR.5 holds, then

$$\mathbb{V}(\hat{\beta}_j|X) = \sigma^2(\hat{V}'\hat{V})^{-1} = \frac{\sigma^2}{\text{SSR}_j} = \frac{\sigma^2}{\text{SST}_j(1-R_j^2)}.$$

- (x) The standard error is given by

$$\text{se } \hat{\beta}_j := \hat{\sigma} \sqrt{e_j'(X'X)^{-1}e_j}.$$

The heteroskedasticity robust standard error is given by

$$\text{se } \hat{\beta}_j = \sqrt{\frac{\sum_i \hat{r}_{ij}^2 \hat{u}_i^2}{\left(\sum \hat{r}_{ij}^2 \right)^2}},$$

where \hat{r}_{ij} denotes the i th residual of a regression of x_j on all other regressors.

11 The Language of Causality

Example 11.1 (A Primer, CPS). Let D_i be the indicator of an individual being sampled. Let Y_i denote the wage of individual i . Let X_i be the indicator of an individual being of Type 1, who loves to respond to surveys. All other individuals are of Type 0, who hate to respond to surveys. We observe $D_i Y_i$ instead of Y_i . Suppose that

$$\begin{aligned} \mathbb{E}[Y|X = 1] &= 100, & \mathbb{E}[Y|X = 0] &= 50, \\ \mathbb{P}(D = 1|X = 1) &= x_1 = \frac{1}{2}, & \mathbb{P}(D = 1|X = 0) &= x_0 = \frac{1}{18}, \\ \mathbb{P}(X = 1) &= \frac{1}{10}, & \mathbb{P}(X = 0) &= \frac{9}{10}. \end{aligned}$$

Every 1000 individuals surveyed, we expect to get around 50 responses from Type 1 individuals and 50 responses from Type 0 individuals. We have

$$\mathbb{P}(X = 1|D = 1) = \frac{1}{2}, \quad \mathbb{P}(X = 0|D = 1) = \frac{1}{2}$$

and so

$$\frac{1}{n} \sum Y_i D_i \xrightarrow{p} \mathbb{E}[YD] = \mathbb{E}[Y|D = 1]\mathbb{P}(D = 1).$$

Thus,

$$\frac{1}{n} \sum \frac{Y_i D_i}{\mathbb{P}(D = 1)} \xrightarrow{p} \mathbb{E}[Y|D = 1].$$

If Y were independent of D (responses missing at random), then we have $\mathbb{E}[Y] = \mathbb{E}[Y|D = 1]$. Thus if we know $\mathbb{P}(D = 1)$, an unbiased and consistent estimator for $\mathbb{E}[Y]$ would be $n^{-1} \sum Y_i D_i / \mathbb{P}(D = 1)$. This is not, however, true of our stylized world (assuming $Y \perp\!\!\!\perp D|X$):

$$\begin{aligned} \mathbb{E}[Y|D = 1] &= \mathbb{E}[Y|D = 1, X = 1]\mathbb{P}(X = 1|D = 1) + \mathbb{E}[Y|D = 1, X = 0]\mathbb{P}(X = 0|D = 1) \\ &= 100 \cdot \frac{1}{2} + 50 \cdot \frac{1}{2} = 75 \neq \mathbb{E}[Y|D = 0]. \end{aligned}$$

In particular $Y \not\perp\!\!\!\perp D$, since people who are sampled are Type 1 more often than their relative frequency in the population and Type 1 people earn more.

Now suppose we are willing to assume non-response is “at random” conditional on Type: $Y \perp\!\!\!\perp D|X$. Then

$$\begin{aligned} \mathbb{E}[Y] &= \mathbb{E}[Y|X = 1]\mathbb{P}(X = 1) + \mathbb{E}[Y|X = 0]\mathbb{P}(X = 0) \\ &= \mathbb{E}[Y|D = 1, X = 1]\mathbb{P}(X = 1) + \mathbb{E}[Y|D = 1, X = 0]\mathbb{P}(X = 0) \\ &= \frac{\mathbb{E}[Y\mathbb{1}(D = 1, X = 1)]}{\mathbb{E}[\mathbb{1}(D = 1, X = 1)]}\mathbb{P}(X = 1) + \frac{\mathbb{E}[Y\mathbb{1}(D = 1, X = 0)]}{\mathbb{E}[\mathbb{1}(D = 1, X = 0)]}\mathbb{P}(X = 0). \end{aligned}$$

We may use the sample analogue principle:

$$\begin{aligned}
 \hat{\mathbb{E}}(Y|D = 1, X = 1) &= \overbrace{\frac{\frac{1}{n} \sum Y_i \mathbb{1}(D_i = 1, X_i = 1)}{\mathbb{P}(D = 1, X = 1)}}^{\text{Average of Type 1, obtained through reweighing each response by response rate}} \cdot \underbrace{\mathbb{P}(D = 1)}_{\text{Reweigh by the proportion of Type 1 in the population}} \\
 &= \frac{\frac{1}{n} \sum Y_i \mathbb{1}(D_i = 1, X_i = 1)}{\mathbb{P}(D = 1|X = 1)} = \frac{1}{n} \sum \frac{Y_i D_i}{\mathbb{P}(X_i)},
 \end{aligned}$$

and similarly for the other term.⁴ Then,

$$\hat{\mathbb{E}}[Y] = \frac{1}{n} \sum \frac{Y_i D_i}{\mathbb{P}(X_i)}$$

This is called inverse propensity score weighting. The population analogue of this is

$$\mathbb{E}[Y] = \mathbb{E}\left(\frac{YD}{\mathbb{P}(X)}\right)$$

This is the weights used in the CPS (base weights).



Definition 11.2.

- The **average treatment effect** (ATE) is defined as

$$\mathbb{E}[Y(1) - Y(0)]$$

- The **average treatment effect on the treated** (ATT) is defined as

$$\mathbb{E}[Y(1) - Y(0)|D = 1]$$

- The **average treatment effect on the untreated** (ATU) is defined as

$$\mathbb{E}[Y(1) - Y(0)|D = 0]$$

Proposition 11.3. *We have the decomposition*

$$\text{ATE} = \text{ATT} \cdot \mathbb{P}(D = 1) + \text{ATU} \cdot \mathbb{P}(D = 0).$$

With a binary treatment D , we may always without loss of generality write the conditional mean as

$$\mathbb{E}(Y|D) = \beta_0 + \beta_1 D.$$

However, the difference in means β_1 is not necessarily the treatment effect:

$$\begin{aligned}
 \beta_1 &= \mathbb{E}[y(1)|D = 1] - \mathbb{E}[y(0)|D = 0] \\
 &= \underbrace{\mathbb{E}[y(1)|D = 1] - \mathbb{E}[y(0)|D = 1]}_{\text{ATT}} + \underbrace{\mathbb{E}[y(0)|D = 1] - \mathbb{E}[y(0)|D = 0]}_{\text{Selection Bias}}.
 \end{aligned}$$

⁴It turns out that using the estimated propensity scores is more efficient, see Hirano, Imbens, and Ridder (2003) Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Scores.

We may similarly decompose β_1 as the ATU and the selection bias in $y(1)$. Combining the two decompositions, we can write β_1 as a weighted average of the ATE and the selection biases:

$$\begin{aligned} & \mathbb{E}(y(1)|D = 1) - \mathbb{E}(y(0)|D = 0) \\ &= \text{ATE} \\ &+ \{\mathbb{E}[y(1)|D = 1] - \mathbb{E}[y(1)|D = 0]\} \mathbb{P}(D = 0) \\ &+ \{\mathbb{E}[y(0)|D = 1] - \mathbb{E}[y(0)|D = 0]\} \mathbb{P}(D = 1). \end{aligned}$$

Example 11.4 (Selection Biases).

- Selection bias on $y(0)$: manager upgrades worse-performing stores.
- Selection bias on $y(1)$: manager upgrades stores based on their forecasted performance.

When there is no selection bias in $y(0)$, $\text{ATE} = \text{ATT}$. When there is no selection bias in $y(1)$, $\text{ATE} = \text{ATU}$. 


Example 11.5. In a regression $Y_i = \beta_0 + \beta_1 D_i + v_i$, we have

$$\hat{\beta}_1 = \frac{\frac{1}{n} \sum Y_i D_i}{\frac{1}{n} \sum D_i} \xrightarrow{p} \frac{\mathbb{E}[YD]}{\mathbb{E}[D]} = \mathbb{E}[Y|D = 1].$$



Example 11.6. If there are a control group and k possible treatments groups. Suppose participants are randomly assigned to either control or one of the treatments and let x_{ij} denote the indicator for individual i taking treatment j . We may specify

$$\mathbb{E}[Y_i|x_i] = \beta_0 + \sum_{j=1}^k \beta_j x_{ij} + \epsilon_i,$$

where $\beta_0 = \mathbb{E}[y(0)]$ and $\beta_j = \mathbb{E}(y(j) - y(0))$. 

11.1 Random Assignment

Example 11.7 (Identification Under Random Assignment). Under random assignment, or

$$D_i \perp\!\!\!\perp (y_i(0), y_i(1)),$$

we have

$$\begin{aligned} \beta_1 &= \mathbb{E}[y(1)|D = 1] - \mathbb{E}[y(0)|D = 0] \\ &= \mathbb{E}(y(1)) - \mathbb{E}(y(0)) = \text{ATE}. \end{aligned}$$

Thus, under randomized treatment assignment, estimating a linear regression produces an unbiased estimate of the ATE.

With random assignment to k different outcomes, we have $k + 1$ potential outcomes

$$Y_i = y_i(0) + \sum_{j=1}^k (y_j(i) - y_i(0)).$$

Using the model

$$\mathbb{E}[Y_i|x_i] = \beta_0 + \sum_{j=1}^k \beta_j x_{ij},$$

we can estimate

$$\beta_0 = \mathbb{E}(y(0)), \quad \beta_j = \mathbb{E}(y(j) - y(0)).$$



11.1.1 Covariate Balance Test

A statistical justification for random assignment is a so-called “balance test” where we regress treatment dummy D_i :

$$D_i = \gamma_0 + x_i' \gamma_1 + \epsilon_i.$$

We can the conduct a F -test of $H_0 : \gamma_1 = 0_{d_x}$ vs. $H_1 : \gamma_1 \neq 0_{d_x}$.

Remark 11.8.

- A rejection of the null hypothesis does not imply that the treatment assignment is not random. At the significance level α , we have a probability of α of rejecting the null hypothesis when it is true.
- Perhaps more importantly, D_i could be correlated with some unobserved variables. But always note that what we really need is

$$D_i \perp\!\!\!\perp (y_i(0), y_i(1))$$

instead of $D_i \perp\!\!\!\perp X$.



12 Selection on Observables

Definition 12.1. We say treatment assignment is **unconfounded** if, conditional on observed covariates x_i , treatment is randomly assigned:

$$D_i \perp\!\!\!\perp (y_i(0), y_i(1)) | x_i.$$

Remark 12.2. Conditional independence does not imply independence. The converse is false also. In particular, conditional independence may fail to hold when we conditional on too much variables. See the next example. ☕

Example 12.3. Let W denote wage and D_i indicate individual i is employed. If we assume $W \perp\!\!\!\perp D$, then $F_W(\cdot | D = 1) = F_W(\cdot | D = 0)$. Now suppose $D = 1$ if and only if $W \geq R$, where R is a reservation wage. “Missing at random” would require that those with high market wages also tend to have high reservation wages. Note that missing at random does not imply that missing at random conditional on reservation wages:

$$\begin{aligned} \mathbb{E}[W | D = 1, R = r] &= \mathbb{E}[W | W \geq R, R = r] \geq r, \\ \mathbb{E}[W | D = 0, R = r] &\leq r. \end{aligned}$$



Under unconfoundedness, we have

$$\begin{aligned} \mathbb{E}(y_1 | D, x) &= \mathbb{E}(y_1 | x) \\ \mathbb{E}(y_0 | D, x) &= \mathbb{E}(y_0 | x). \end{aligned}$$

Thus, to estimate the ATE, we can write

$$\begin{aligned} \text{ATE} &= \mathbb{E}\{\mathbb{E}[y_1 - y_0 | x]\} \\ &= \mathbb{E}\{\mathbb{E}(y_1 | D = 1, x) - \mathbb{E}(y_0 | D = 0, x)\} \\ &= \mathbb{E}\{\mathbb{E}(y | D = 1, x) - \mathbb{E}(y | D = 0, x)\}. \end{aligned}$$

Each term in the last equality can be estimated using observed data.

Since in general $\mathbb{P}(X = \cdot; D = 1) \neq \mathbb{P}(X = \cdot)$, we cannot simplify the last equality to $\mathbb{E}[y | D = 1] - \mathbb{E}[y | D = 0]$. However, if $\mathbb{P}(X = \cdot; D = 1) = \mathbb{P}(X = \cdot)$, then

$$\text{ATE} = \mathbb{E}[y | D = 1] - \mathbb{E}[y | D = 0].$$

We thus have the following:

Proposition 12.4. *If, in addition to unconfoundedness, we have $D_i \perp\!\!\!\perp x_i$, then the naive comparison is the ATE.*

12.1 Example Specification 1

Suppose

$$\begin{aligned} \mathbb{E}(y(0) | x) &= \alpha_0 + x' \beta_2; \\ \mathbb{E}(y(1) | x) &= \alpha_1 + x' \beta_2. \end{aligned}$$

In particular, this is true if we are imposing homogeneous treatment effects.

We may approximate conditional means with a linear model, since

$$\begin{aligned}\mathbb{E}[Y|D, x] &= \mathbb{E}(y_0 + D(y_1 - y_0)|D, x) \\ &= \underbrace{\alpha_0}_{\beta_0} + x'\beta_2 + D(\underbrace{\alpha_1 - \alpha_0}_{\beta_1}) \\ &= \beta_0 + \beta_1 D_i + x'_i \beta_2.\end{aligned}$$

We now have

$$\text{ATE} = \mathbb{E}(Y|D = 1, x) - \mathbb{E}(Y|D = 0, x) = \alpha_1 - \alpha_0 = \beta_1$$

and so we may estimate the ATE using $\hat{\beta}_1$.

12.2 Example Specification 2

We may think that potential outcomes should have different β :

$$\begin{aligned}\mathbb{E}[y(0)|x] &= \alpha_0 + x'\beta^0; \\ \mathbb{E}[y(1)|x] &= \alpha_1 + x'\beta^1.\end{aligned}$$

The ATE under this model is

$$\mathbb{E}[y(1) - y(0)] = \mathbb{E}[(\alpha_1 - \alpha_0) + x'(\beta_1 - \beta_0)].$$

A natural estimator of the ATE is then

$$\hat{\tau} = \frac{1}{n} \sum [y_i(1) - y_i(0)] = \hat{\alpha}_1 - \hat{\alpha}_0 + \bar{x}'_n(\hat{\beta}^1 - \hat{\beta}^0).$$

Note that

$$\begin{aligned}\mathbb{E}[Y|D, x] &= \mathbb{E}[y(0) + D(y(1) - y(0))|D, x] \\ &= \alpha_0 + (\alpha_1 - \alpha_0)D + x'\beta^0 + Dx'(\beta^1 - \beta^0),\end{aligned}$$

Estimating this model is mathematically equivalent to estimating the two regressions separately (since we may split the SSR into the SSR's of $D = 0$ and $D = 1$).

It is convenient to center x about 0:

$$\begin{aligned}\mathbb{E}[y(0)|x] &= \delta_0 + (x - \mu_x)'\beta^0; \\ \mathbb{E}[y(1)|x] &= \delta_1 + (x - \mu_x)'\beta^1,\end{aligned}$$

since in this model, the ATE is just the difference in intercepts:

$$\mathbb{E}[y(1) - y(0)] = \delta_1 - \delta_0.$$

We don't know μ_x , but may replace it with \bar{x}_n to obtain $\hat{\tau}$:

$$Y_i = \delta_0 + \tau D_i + [x_i - \bar{x}_n]'\beta_0 + D_i[x_i - \bar{x}_n]'\rho + \epsilon_i.$$

But be cautious! We calculating the standard error, statistical packages will not automatically treat \bar{x}_n as a random variable, although this error is small when n is large.

13 Difference-in-Differences

Example 13.1. Two time periods. Treated units has dummy $G_i = 1$. We want the ATT:

$$ATT = \mathbb{E}[y_1 - y_0 | G = 1, T = 1].$$



Assumption 13.2 (Common Trends).

$$\begin{aligned} & \mathbb{E}[y_0 | G = 1, T = 1] - \mathbb{E}[y_0 | G = 1, T = 0] \\ &= \mathbb{E}[y_0 | G = 0, T = 1] - \mathbb{E}[y_0 | G = 0, T = 0]. \end{aligned}$$

In other words, the common trends assumption states that the selection bias in y_0 is constant over time.

Remark 13.3. It might be more plausible to assume parallel trends in $f(y)$, but be careful about the interpretation of the *ATT*: in general, $\mathbb{E}(f(y_1) - f(y_0)) \neq f(\mathbb{E}[y_1 - y_0])$. ☕

In the two period setting, we have

$$\begin{aligned} \tau^{\text{DiD}} &= \{ \mathbb{E}[y | G = 1, T = 1] - \mathbb{E}[y | G = 1, T = 0] \} \\ &\quad - \{ \mathbb{E}[y | G = 0, T = 1] - \mathbb{E}[y | G = 0, T = 0] \} \\ &= \{ \mathbb{E}[y | G = 1, T = 1] - \mathbb{E}[y | G = 1, T = 0] \} \\ &\quad - \{ \mathbb{E}[y_0 | G = 0, T = 1] - \mathbb{E}[y_0 | G = 0, T = 0] \} \\ &= \{ \mathbb{E}[y | G = 1, T = 1] - \mathbb{E}[y | G = 1, T = 0] \} \\ &\quad - \{ \mathbb{E}[y_0 | G = 1, T = 1] - \mathbb{E}[y_0 | G = 1, T = 0] \} \\ &= ATT. \end{aligned}$$

13.1 Regression Implementation

We may write the model for y_0 as

$$\mathbb{E}[y_0 | G, T] = \beta_0 + \beta_1 G + \beta_2 T + \beta_3 GT.$$

The common trends assumption holds if and only if $\beta_3 = 0$. Since y_1 is observed if and only if $GT = 1$, we have

$$\mathbb{E}[Y | G, T] = \mathbb{E}[y_0 | G, T] + \mathbb{E}[y_1 - y_0 | G = 1, T = 1] GT.$$

We thus have

$$\mathbb{E}[Y | G, T] = \beta_0 + \beta_1 G + \beta_2 T + [\beta_3 + ATT] GT.$$

Under common trends, the coefficient on GT is the ATT. In general, the coefficient on GT is the ATT plus the difference in untreated (average) potential outcome trends between the two groups.

Remark 13.4. Since we would assume $Y | G, T = 1$ to be correlated with $Y | G, T = 0$, we often use **clustered standard errors**, with the cluster being the group. ☕

13.2 Conditional Common Trends

Remark 13.5. Including controls can also reduce the standard errors on the interaction term. ☕

Suppose now we believe only the following conditional common trends assumption:

Assumption 13.6 (Conditional Common Trends).

$$\begin{aligned} & \mathbb{E}[y_0|G = 1, T = 1, X = x] - \mathbb{E}[y_0|G = 1, T = 0, X = x] \\ &= \mathbb{E}[y_0|G = 0, T = 1, X = x] - \mathbb{E}[y_0|G = 0, T = 0, X = x]. \end{aligned}$$

We may, for example, model y_0 as

$$\mathbb{E}[y_0|G, T, X] = \beta_0 + \beta_1 G + \beta_2 T + \beta_3 GT + X' \delta.$$

Remark 13.7.

- Under Assumption 13.6, we have $\beta_3 = 0$.
- Note that

$$\begin{aligned} & \mathbb{E}[y_0|G = 1, T = 1] - \mathbb{E}[y_0|G = 1, T = 0] \\ &= \beta_2 + \{\mathbb{E}[X|G = 1, T = 1] - \mathbb{E}[X|G = 1, T = 0]\}' \delta. \end{aligned}$$

The same calculation for $G = 0$ shows that common trends (without conditioning) holds if the composition of covariates across time changes in the same way in both groups, or if $\delta = 0$. ☕

Using the model above for y_0 , we have

$$\begin{aligned} \mathbb{E}[Y|G, T, X] &= \mathbb{E}[y_0|G, T, X] + \mathbb{E}[y_1 - y_0|G = 1, T = 1, X] GT \\ &= \beta_0 + \beta_1 G + \beta_2 T + X' \delta + \text{ATT}(X) GT \end{aligned}$$


Under conditional common trends and assuming a homogeneous effect across covariate values, the coefficient on GT is the ATT. In this case the model reduces to

$$\mathbb{E}[Y|G, T, X] = \beta_0 + \beta_1 G + \beta_2 T + \text{ATT} GT + X' \delta.$$

Formally (symbolically), this is the same specification as the naive difference in differences, with the term $X' \delta$ appended to the formula.

We may also assume instead (for example) that $\text{ATT}(X) = X' \theta$ and estimate θ .

13.3 Triple Difference-in-Differences

Example 13.8. The manager chooses to upgrade stores in city A, but all the upgraded stores were in urban areas. We may no longer believe that changes in sales in the absence of upgrades would be the same. However, we may use the difference in trends between urban and suburban stores in the city where no upgrades occurred to see what the difference in trends would have been in city A had there been no upgrade. 

Assumption 13.9 (Triple Difference-in-Differences). “In the absence of treatment, the difference in the trend in sales across urban and suburban stores is the same in city B as it would have been in city A.”

Let $C = 1$ denote city A, where treatment occurred.

$$\begin{aligned} E[Y|G, T, C] = & \beta_0 + \beta_1 G + \beta_2 T + \beta_3 GT \\ & + C(\gamma_0 + \gamma_1 + \gamma_2 T + [\gamma_3 + \text{ATT}]GT). \end{aligned}$$

Under Assumption 13.9, $\gamma_3 = 0$, and the coefficient on CGT is the ATT.

14 Instrumental Variables: The Simple Case

Consider the simple linear regression model

$$y = \beta_0 + \beta_1 x + v.$$

If we interpret the coefficients as those in the best linear predictor of y given x , then $\mathbb{E}[xu] = 0$ is true by contraction. However, this may not be true if we interpret the coefficients as the causal effect of x on y .

Definition 14.1. We say x_1 is **exogenous** if $\mathbb{E}[vx_1] = 0$. Otherwise, we say x_1 is **endogenous**.

We aim to use instrumental variables to address endogeneity in the model.

Assumption 14.2 (Instrumental Variables). In the model $y = \beta_0 + \beta_1 x + v$, instrumental variables satisfy the following:

- (i) Relevance: $\mathbb{C}(x, z) \neq 0$.
- (ii) Validity: $\mathbb{C}(z, v) = 0$.
- (iii) Exclusion: z is not itself a determinant of y . In particular, $z \perp\!\!\!\perp v$ and $z \perp\!\!\!\perp y|x$.

Remark 14.3. The relevance assumption is empirically testable, we may just run the regression

$$x = \pi_0 + \pi_1 z + \epsilon.$$




Under Assumption 14.2, we have

$$\mathbb{C}(z, y) = \mathbb{C}(z, \beta_0 + \beta_1 x + v) = \beta_1 \mathbb{C}(z, x) + \mathbb{C}(z, v),$$

where the last equality comes from the validity assumption. Thus,

$$\beta_1 = \frac{\mathbb{C}(z, y)}{\mathbb{C}(z, x)}.$$

Remark 14.4. We may then think of the OLS estimator as a special case of the IV estimator, where we use x as its own instrument. 

Proposition 14.5. Suppose the model is $y = \beta_0 + \beta_1 x + v$, where x is a scalar. Suppose z satisfies Assumption 14.2. Then, the IV estimator for β_1

$$\hat{\beta}_1^{\text{IV}} = \frac{\sum (z_i - \bar{z}) y_i}{\sum (z_i - \bar{z}) x_i}$$

is consistent and asymptotically normal. If $\mathbb{E}[v|z] = 0$ and $\mathbb{V}(v|z) = \sigma^2$, then

$$\sqrt{n}(\hat{\beta}_1^{\text{IV}} - \beta_1) \xrightarrow{d} \mathcal{N}\left(0, \frac{\sigma^2}{\mathbb{V}(x) \text{Corr}(x, z)^2}\right)$$

Remark 14.6. From this we see that the IV instrument has a larger asymptotic variance than the OLS estimator, $\sigma^2/\mathbb{V}(x)$. ☕

Proof. Consistency comes from the continuous mapping theorem, noting that sample variance is a consistent estimator of the population variance. To see asymptotic normality, write

$$\sqrt{n}(\hat{\beta}_1^{\text{IV}} - \beta_1) = \sqrt{n} \frac{\sum (z_i - \bar{z})(y_i - x_i \beta_1)}{\sum (z_i - \bar{z})x_i} = \frac{\frac{1}{\sqrt{n}} \sum (z_i - \bar{z})v}{\frac{1}{n} \sum (z_i - \bar{z})x_i}.$$

By the central limit theorem,

$$\frac{1}{\sqrt{n}} \sum (z_i - \bar{z})v \xrightarrow{d} \mathcal{N}(0, \sigma^2 \mathbb{V}(z)),$$

since

$$\mathbb{V}((z - \mu_z)v) = \mathbb{E}((z - \mu_z)^2 \mathbb{V}(v|z)) = \sigma^2 \mathbb{V}(z)$$

by the law of total variance. By the weak law of large numbers,

$$\frac{1}{n} \sum (z_i - \bar{z})x_i \xrightarrow{p} \mathbb{C}(x, z).$$

□

14.1 Standard Errors

Remark 14.7. We may estimate $\mathbb{V}(x)$ consistently using the sample variance. We may estimate $\text{Corr}(x, z)^2$ consistently using the R^2 of the regression $x = \pi_0 + \pi_1 z + v$. We may estimate σ^2 consistently using

$$\hat{\sigma}^2 = \frac{\sum (y_i - \hat{\beta}_0^{\text{IV}} - \hat{\beta}_1^{\text{IV}} x_i)^2}{n - 2}.$$

☕

In large samples, the variance of $\hat{\beta}_1^{\text{IV}}$ is approximately

$$\frac{\sigma^2/n}{\mathbb{V}(x) \text{Corr}(x, z)^2}.$$

The standard error computed by regression packages is

$$\sqrt{\frac{\hat{\sigma}^2}{\text{SST}_x \cdot R_{x,z}^2}}$$

This is typically larger than the OLS standard error $\hat{\sigma}/\sqrt{\text{SST}_x}$, since we “lose information” by first regressing x on z .

14.2 Weak Instruments

Suppose $\mathbb{C}(z, v)$ is small but non-zero. We have

$$\hat{\beta}_1^{\text{IV}} \xrightarrow{p} \frac{\mathbb{C}(z, y)}{\mathbb{C}(z, x)} = \beta_1 + \frac{\mathbb{C}(z, v)}{\mathbb{C}(z, x)}.$$

Even if $\mathbb{C}(z, v)$ is small, small $\mathbb{C}(z, x)$ may mean the inconsistency in the IV estimator is worse than OLS.

We run essentially a t -test of π_1 “is near 0” to test for weak instruments.

14.3 Causes of Endogeneity

14.3.1 Omitted Variables

Suppose the true model is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u, \quad \mathbb{E}[u|x] = 0$$

but we specified the model as

$$y = \beta_0 + \beta_1 x_1 + v.$$

Then,

$$\mathbb{C}(x_1, v) = \mathbb{C}(x_1, \beta_2 x_2 + u) = \mathbb{C}(x_1, x_2) \beta_2.$$

The OLS estimator in the short model is

$$\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2 \frac{\widehat{\mathbb{C}}(x_1, x_2)}{\widehat{\mathbb{V}}(x_1)} \xrightarrow{p} \beta_1 + \beta_2 \frac{\mathbb{C}(x_1, x_2)}{\mathbb{V}(x_1)}.$$

Remark 14.8. The sign of this bias is determined by the signs of $\mathbb{C}(x_1, x_2)$ and β_2 . If both are positive, then the bias is positive. ☕

14.3.2 Measurement Error

The key intuition in this case is that when we spread x out, the slope of the regression line will be smaller.

Suppose x_1^* is a scalar and

$$y = \beta_0 + \beta_1 x_1^* + u, \quad \mathbb{E}[u] = 0, \quad \mathbb{C}(x_1^*, u) = 0.$$

Suppose we observe a noisy signal of x_1^* , given by $x_1 = x_1^* + v$, with

$$\mathbb{E}(v) = 0, \quad \mathbb{C}(x_1^*, v) = 0, \quad \mathbb{C}(u, v) = 0.$$

These are the assumptions of **classical measurement error**.

The true model can then be rewritten as

$$y = \beta_0 + \beta_1 x_1 + \epsilon,$$

where

$$\mathbb{C}(x_1, \epsilon) = \mathbb{C}(x_1, u - \beta_1 v) = \mathbb{C}(x_1^* + v, u - \beta_1 v) = -\beta_1 \mathbb{V}(v).$$

The OLS estimator is not consistent since

$$\begin{aligned}\hat{\beta}_1^{\text{OLS}} &\xrightarrow{p} \frac{\mathbb{C}(x_1, y)}{\mathbb{V}(x_1)} \\ &= \beta_1 + \frac{\mathbb{C}(x_1, \epsilon)}{\mathbb{V}(x_1)} = \beta_1 \left(1 - \frac{\mathbb{V}(v)}{\mathbb{V}(x_1)} \right) \\ &= \beta_1 \left(1 - \frac{\mathbb{V}(v)}{\mathbb{V}(x_1^*) + \mathbb{V}(v)} \right).\end{aligned}$$

It follows that $\hat{\beta}_1^{\text{OLS}}$ will tend be smaller in magnitude than the true causal effect β_1 .

An instrument for x_1 in this case can be another noisy measurement of x_1^* .⁵

$$z_1 = x_1^* + w,$$

where

$$\mathbb{E}(w) = 0, \quad \mathbb{C}(x_1^*, w) = 0, \quad \mathbb{C}(u, w) = 0$$

and $\mathbb{C}(v, w) = 0$. Validity holds since

$$\mathbb{C}(z_1, \epsilon) = \mathbb{C}(x_1^* + w, u - \beta_1 v) = 0.$$

For relevance, note that

$$\mathbb{C}(z_1, x_1) = \mathbb{C}(x_1^* + w, x_1^* + w) = \mathbb{V}(x_1^*),$$

which is nonzero provided x_1^* is not constant.

⁵An example of this method being used can be found in Ashenfelter and Krueger (1994) Estimates of the Economic Return to Schooling from a New Sample of Twins.

15 Instrumental Variables

Assumption 15.1 (Instrumental Variables). Suppose we observe outcomes y , covariates w , treatment status d , and instrument z .

- (i) **Relevance:** Instrument is correlated with the treatment (conditional on covariates).
- (ii) **Exclusion:** Instrument has no direct effect on potential outcomes.

$$Y = y(D, z) = y(D).$$

The first equality writes the observed outcome in the potential outcomes framework; the second is our assumption.

- (iii) **Exogeneity:** Instrument is independent of potential outcomes (conditional on covariates).

$$y(d) \perp\!\!\!\perp z|w.$$

Remark 15.2. Under Assumption 15.1, the covariance between instrument and observed outcome occurs only through changes in the treatment, so we can measure the causal effect by adjusting for how much the treatment moves with the instrument. ☕

Assumption 15.3 (Constant, linear treatment effects). For each d ,

$$y(d) = x(w, d)' \beta + u, \quad \mathbb{E}[u|w] = 0.$$

In particular, u depends on the individual and not d .

What grounds do we have to assume that u does not depend on d ? The following proposition gives an alternative way of viewing Assumption 15.3 that can be easier to argue for.

Proposition 15.4. *Assumption 15.3 is equivalent to assuming linear conditional means and constant treatment effects conditional on w .*

Remark 15.5. Note, however, that unconditionally, the treatment effect need not be constant, since we allow for interactions. ☕

Proof. First assume linear conditional means and constant treatment effects conditional on w . We have

$$y(d) = x(w, d)' \beta + u_d, \quad \mathbb{E}(u_d|w) = 0,$$

where u_d can depend *a priori* on d . Assuming constant conditional treatment effects, we have

$$\begin{aligned} y(d) &= \mathbb{E}(y(d) - y(d^*)|w) + y(d^*) \\ &= [x(w, d) - x(w, d^*)]' \beta + y(d^*). \end{aligned}$$

Since $u_d = y(d) - x(w, d)' \beta$, it follows that $u = u_{d^*} \equiv u$ for all d .

The converse is straightforward. □

Proposition 15.6. Assumption 15.1 and Assumption 15.3 imply $\mathbb{E}(u|z, w) = 0$. In particular, we have *instrumental validity*, $\mathbb{E}(zu) = 0$.

Proof. Note that

$$\begin{aligned}\mathbb{E}[u|z, w] &= \mathbb{E}[y(d) - x(w, d)' \beta | z, w] \\ &= \mathbb{E}[y(d)|w] - x(w, d)' \beta = 0,\end{aligned}$$

where the second line follows from the exogeneity assumption. \square

Remark 15.7.

- $\mathbb{E}(zu) = 0$ is the starting point of our estimation, and what we assumed at the outset in the previous section. However, stating from Assumption 15.1 and Assumption 15.3 is informative, for it lays bare the precise and tangible assumptions that can be argued for.
- We have similarly that $\mathbb{E}(u|w) = 0 = \mathbb{E}(uw)$. The set of assumptions that give rise to this result is however drastically different. Here, $\mathbb{E}(u|w) = 0$ results from the assumption that we correctly specified the conditional mean of $y(d)|w$. It is also worthy or note that the terms in w need not be determinants of y , some of them can be control variables. Their only purpose is to specify the conditional mean.




Now, let (y, x, u) be a random vector such that y and u are scalar random variables and $x \in \mathbb{R}^{k \times 1}$ with the first component being 1. Let $\beta \in \mathbb{R}^{k \times 1}$ be a constant vector of unknown parameters such that $y = x' \beta + u$.

Definition 15.8.

- If $\mathbb{E}[ux_j] = 0$ for some j , x_j is **exogenous**.
- If $\mathbb{E}[ux_j] \neq 0$ for some j , x_j is **endogenous**.

Note that x_0 can always be made exogenous by shifting β_0 such that $\mathbb{E}[x_0 u] = \mathbb{E}[u] = 0$.

Example 15.9. Since $\mathbb{E}(wu) = 0$, control variables are exogenous. Endogenous variables can include combinations of d and w . 

Pre-multiplying by x and taking expectations, we have $\mathbb{E}(yx) = \mathbb{E}(xx')\beta + \mathbb{E}(xu)$. It follows that $\mathbb{E}(xx')^{-1} \mathbb{E}[xy] = \beta + \mathbb{E}(xx')^{-1} \mathbb{E}(xu)$ and thus

$$\hat{\beta}^{\text{OLS}} = (X'X)^{-1} X'Y \xrightarrow{P} \beta + \mathbb{E}[xx']^{-1} \mathbb{E}(xu).$$

Under endogeneity, the OLS estimator of β is inconsistent.

15.1 Identification Setup

The goal is to use a random vector $z \in \mathbb{R}^{l+1}$ such that the **instrumental validity** assumption $\mathbb{E}(zu) = 0$ is satisfied. The components of z are called **instrumental variables**. The vector z contains any exogenous components of x (recall that this includes the intercept and components of w) and variables that are (partially) correlated with endogenous variables but uncorrelated with u that are outside the main specification. We write $z_0 = 1$ and

$$z = (z_0, z_1, \dots, z_l)' \in \mathbb{R}^{l+1}.$$

Assumption 15.10.

- $\mathbb{E}(zz') < \infty$ and there is no perfect collinearity in z .
- **Instrument relevance / rank condition:** $\mathbb{E}(zx')$ has full rank $k + 1$. Note that a necessary condition for the rank condition is the **order condition**, $l \geq k$.

Remark 15.11. If we set $\mathbb{E}[u] = 0$ and $z_0 = 1$, then $\mathbb{E}[zu] = 0$ is equivalent to

$$\mathbb{C}(z_j, u) = 0, \quad \forall j.$$



We now pre-multiply the model by z and take expectations to get

$$\mathbb{E}(zy) = \beta \mathbb{E}(zx') + \mathbb{E}(zu) = \beta \mathbb{E}(zx').$$

This will be the starting point of our estimation.

15.2 Exact Identification: The IV Estimator

Assuming $l = k$ (there are exactly as many instruments as regressors), $\mathbb{E}[zx']$ is a square. We may thus write

$$\begin{aligned} \beta &= [\mathbb{E}(zx')]^{-1} \mathbb{E}(zy) \\ &= \underbrace{[\mathbb{E}(zz')^{-1} \mathbb{E}(zx')]^{-1}}_{\text{First Stage Coefficients}} \underbrace{\mathbb{E}(zz')^{-1} \mathbb{E}(zy)}_{\text{Reduced Form Coefficients}}, \end{aligned}$$

where the first stage and reduced form refers respectively to the regressions

$$\begin{aligned} x' &= z' \pi + v', & \mathbb{E}[zv'] &= 0; \\ y &= z' \delta + \epsilon, & \mathbb{E}[z\epsilon] &= 0. \end{aligned}$$

Proposition 15.12. *In the case of exact identification, the IV estimator*

$$\hat{\beta}^{\text{IV}} := \left(\frac{1}{n} \sum z_i x_i' \right)^{-1} \left(\frac{1}{n} \sum z_i y_i \right) = (Z'X)^{-1} Z'Y$$

coincides with the Two Stage Least Squares (2SLS) estimator

$$\hat{\beta}_{2\text{SLS}} := (X'P_Z X)^{-1} X'P_Z Y.$$

Moreover, both estimators are consistent and asymptotically normal.

Proof. The first statement follows from expanding the projection matrix $P_Z = Z(Z'Z)^{-1}Z'$. Consistency and asymptotic normality comes from the law of larger numbers and the central limit theorem. \square

Example 15.13 (Simple IV Estimation). If $x = (1, x_1)'$ and $z = (1, z_1)'$. Then

$$\beta = \begin{pmatrix} 1 & \pi_0 \\ 0 & \pi_1 \end{pmatrix}^{-1} \begin{pmatrix} \delta_0 \\ \delta_1 \end{pmatrix} = \frac{1}{\pi_1} \begin{pmatrix} \pi_1 & -\pi_0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \delta_0 \\ \delta_1 \end{pmatrix}.$$

In particular,

$$\beta_1 = \frac{\delta_1}{\pi_1} = \frac{\mathbb{C}(y, z_1)/\mathbb{V}(z_1)}{\mathbb{C}(x_1, z_1)/\mathbb{V}(z_1)}.$$



The notation $[\mathbb{E}(zz')^{-1}\mathbb{E}(zx')]^{-1}$ is justified by the following result:

Proposition 15.14. *Suppose there is no perfect collinearity in z and $\mathbb{E}[zz'] < \infty$. Then, $\mathbb{E}[zx']$ is full rank if and only if $\mathbb{E}[zz']^{-1}\mathbb{E}[zx']$ is full rank.*

Proof. Since the rank of $\mathbb{E}[zx']$ must be weakly larger than that of $\mathbb{E}[zz']^{-1}\mathbb{E}[zx']$, we have if $\mathbb{E}[zz']^{-1}\mathbb{E}[zx']$ is full rank, then $\mathbb{E}[zx']$ is full rank also.

If $\mathbb{E}[zx']$ is full rank, then we have $\mathbb{E}[zz']^{-1}\mathbb{E}[zx']$ is full rank, because $\mathbb{E}[zz']$ is full rank also by there being no perfect collinearity in z . \square

The matrix $\mathbb{E}[zz']\mathbb{E}[zx']$ is the matrix of coefficients of the best predictors of each x_j given z . Thus if we let $x_j = z'\gamma_j + v_j$, $\mathbb{E}[zv_j] = 0$, then

$$\mathbb{E}[zz']^{-1}\mathbb{E}[zx'] = \begin{pmatrix} | & | & & | \\ \gamma_0 & \gamma_1 & \cdots & \gamma_k \\ | & | & & | \end{pmatrix}.$$

If there is a single endogenous regressor x_k and $k = I$, then

$$\mathbb{E}[zz']^{-1}\mathbb{E}[zx'] = \begin{pmatrix} 1 & 0 & \cdots & 0 & \gamma_{k,0} \\ 0 & 1 & \cdots & 0 & \gamma_{k,1} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & \gamma_{k,I-1} \\ 0 & 0 & \cdots & 0 & \gamma_{k,I} \end{pmatrix}$$

is full rank if and only if $\gamma_{k,I} \neq 0$. Thus, with a single endogenous regressor and an exactly identified system, the rank condition holds if and only if a regression of x_k on other x 's and the excluded instrument z_I produces a non-zero coefficient on z_I . That is, x_k must be correlated with z_I “after controlling for x_0, \dots, x_{k-1} .”

Note that the top left submatrix is the identity matrix, thus the rank condition holds if and only if the bottom right submatrix is full rank. This submatrix contains the coefficients on the excluded instruments.

15.2.1 Partitioned Regression

We can use partitioned regression to derive the slopes for the endogenous variables. Partition x and z as follows:

$$x = (w', d')', \quad z = (w', r')',$$

where $w \in \mathbb{R}^{k_1}$ contains a constant and any exogenous variables, $d \in \mathbb{R}^{k_2}$ contains the endogenous variables, and $r \in \mathbb{R}^{k_2}$ contains the excluded instruments. The first stage regression is

$$(w', d') = w'\pi_1 + r'\pi_2 + v', \quad \mathbb{E}[zv'] = 0.$$

This is partitioning the matrix π into two horizontal blocks, π_1 and π_2 . We can write a subset of these projections (of d onto z) as

$$d' = w'\pi_{1d} + r'\pi_{2d} + v'_d, \quad \mathbb{E}[zv'_d] = 0,$$

where π_{1d} consists of the k_2 right-most columns of π_1 and π_{2d} is defined analogously. The projection of w onto z is just w . Now write

$$y = x'\beta + u = w'\beta_1 + d'\beta_2 + u.$$

Inserting the first stage regression into $y = x'\beta + u$ gives

$$y = w'[\beta_1 + \pi_{1d}\beta_2] + r'\pi_{2d}\beta_2 + (v'_d\beta_2 + u),$$

which we can write as $y = z'\delta + \epsilon$. It is easy to verify that $\mathbb{E}[z\epsilon] = 0$. Writing $\tilde{r} := r - \text{BLP}(r|w)$, we have from partitioned regression that

$$\begin{aligned} \pi_{2d}\beta_2 &= \mathbb{E}[\tilde{r}\tilde{r}']^{-1}\mathbb{E}[\tilde{r}y], \\ \pi_{2d} &= \mathbb{E}[\tilde{r}\tilde{r}']^{-1}\mathbb{E}[\tilde{r}d']. \end{aligned}$$

It thus follows that

$$\begin{aligned} \beta_2 &= \{\mathbb{E}[\tilde{r}\tilde{r}']^{-1}\mathbb{E}[\tilde{r}d']\}^{-1}\mathbb{E}[\tilde{r}\tilde{r}']^{-1}\mathbb{E}[\tilde{r}y] \\ &= \mathbb{E}[\tilde{r}d']^{-1}\mathbb{E}[\tilde{r}y]. \end{aligned}$$

15.3 Overidentification, Generalized Method of Moments

If $I > k$, the moment condition

$$\mathbb{E}[zu] = \mathbb{E}[z(y - x'\beta)] = 0$$

has a solution by the model specification and the IV validity assumption, but its sample analog may not have a solution, since this would require

$$\sum z_i y_i = Z'Y = Z'X\hat{\beta} = \sum z_i x'_i \hat{\beta}.$$

We have a $\mathbb{R}^{(I+1)}$ vector on the left hand side but only $k + 1 < I + 1$ columns on the right.

One obvious solution is to discard extra instruments to reduce the case to that of exact identification. This is a waste of data. A more optimal way approach is to solve instead the equation

$$CZ'Y = CZ'X\hat{\beta},$$

where C is a full rank $(k + 1) \times (I + 1)$ matrix. This gives rise to a **GMM estimator**

$$\hat{\beta} = (CZ'X)^{-1}(CZ'Y).$$

It turns out that the optimal C can be consistently estimated.

Suppose $C \in \mathbb{R}^{(k+1) \times (I+1)}$ and $C\mathbb{E}[zx']$ is full rank (this is the **relevance** condition). Then $Cz \in \mathbb{R}^{k+1}$ “selects” $k + 1$ linear combinations of z . Since $Czy = Cz x' \beta + Cz u$ and $C\mathbb{E}[zx']$ is square, we may write

$$\begin{aligned} \beta &= \mathbb{E}(Czx')^{-1} \mathbb{E}(Czy) \\ &= \underbrace{\left[\mathbb{E}(Cz[Cz]')^{-1} \mathbb{E}(Czx') \right]^{-1}}_{\text{First Stage Coefficients}} \underbrace{\mathbb{E}(Cz[Cz]')^{-1} \mathbb{E}(Czy)}_{\text{Reduced Form Coefficients}}, \end{aligned}$$

where the first stage and reduced form refer respectively to the regressions

$$\begin{aligned} x' &= [cz]' \pi + v', & \mathbb{E}[czv'] &= 0; \\ y &= [cz]' \delta + \epsilon, & \mathbb{E}[cz\epsilon] &= 0. \end{aligned}$$

15.3.1 Two Stage Least Squares

The **two stage least squares** estimator is obtained by setting C to π in the population regression $x' = z'C + u$. That is,

$$\hat{\beta}_{2SLS} := (X'P_ZX)^{-1}X'P_ZY.$$

We may interpret this an IV estimator where we use $\pi'z$, the projection of x onto z , as instruments. Here, π is the matrix of coefficients in the population regression

$$x' = z'\pi + v', \quad \mathbb{E}[zv'] = 0.$$

It is easy to verify that instrument validity holds.

15.3.2 Simultaneous Equations

Example 15.15 (Simultaneous Equations). Consider the following supply and demand systems:

$$\begin{aligned} q_D &= \beta_0 + \beta_1 p + u, & \mathbb{E}[u] &= 0; \\ q_S &= \gamma_0 + \gamma_1 p + v, & \mathbb{E}[v] &= 0. \end{aligned}$$

Suppose that $\mathbb{E}(uv) = 0$. We only observe supply and demand in equilibrium $q_D = q_S$, at which we have

$$p = \frac{\gamma_0 - \beta_0 + v - u}{\beta_1 - \gamma_1}.$$

Thus p is endogenous in the equations

$$\begin{aligned} q &= \beta_0 + \beta_1 p + u, \\ q &= \gamma_0 + \gamma_1 p + u, \end{aligned}$$

since, for example,

$$\mathbb{C}(p, u) = \mathbb{C}\left(\frac{\gamma_0 - \beta_0 + v - u}{\beta_1 - \gamma_1}, u\right) = -\frac{\mathbb{V}(u)}{\beta_1 - \gamma_1}.$$

Now suppose the model is given by

$$\begin{aligned} q_D &= \beta_0 + \beta_1 p + u, \quad \mathbb{E}[u] = 0; \\ q_S &= \gamma_0 + \gamma_1 p + \gamma_2 z + u, \quad \mathbb{E}[v] = \mathbb{E}[vz] = 0, \end{aligned}$$

where z is an exogenous “supply shifter.” Thus $\mathbb{E}(zu) = 0$ also. Solving for the equilibrium price now gives

$$p = \frac{\gamma_0 - \beta_0 + \gamma_2 z + v - u}{\beta_1 - \gamma_1}.$$

The parameters of the demand equation can now be estimated consistently, because z is a valid instrument for p . Relevance holds if $\gamma_2 \neq 0$, since

$$\mathbb{C}(p, z) = \frac{\gamma_2 \mathbb{V}(z)}{\beta_1 - \gamma_1}.$$



15.4 Consistency and Asymptotic Normality of GMM Estimators

Proposition 15.16 (Consistency and Asymptotic Normality of GMM estimators). *Let $C \in \mathbb{R}^{(k+1) \times (l+1)}$ and suppose $\hat{C} \xrightarrow{P} C$. The estimator based on \hat{C} is consistent and asymptotically normal.*

Proof. For consistency,

$$\begin{aligned} \hat{\beta} &:= (\hat{C}Z'X)^{-1} \hat{C}Z'Y \\ &= \beta + \left(\hat{C} \cdot \frac{Z'X}{n} \right)^{-1} \hat{C} \cdot \frac{Z'U}{n} \\ &\xrightarrow{P} \beta + (CE[zx']^{-1})CE[zu] = \beta. \end{aligned}$$

For normality, write

$$\begin{aligned} \sqrt{n}(\hat{\beta} - \beta) &= \left(\hat{C} \cdot \frac{Z'X}{n} \right)^{-1} \hat{C} \cdot \frac{Z'U}{\sqrt{n}} \\ &= \left(\hat{C} \sum z_i x_i' \right)^{-1} \frac{\hat{C}}{\sqrt{n}} \sum z_i u_i \\ &\xrightarrow{d} (CE[z_i x_i'])^{-1} C \cdot \mathcal{N}(0, \mathbb{E}[u_i^2 z_i z_i']) = \mathcal{N}(0, V), \end{aligned}$$

where

$$V = (C\mathbb{E}[z_i x_i'])^{-1} C C' (\mathbb{E}[z_i x_i'] C')^{-1}$$

$$\Omega := \mathbb{E}[u_i^2 z_i z_i'].$$

□

15.5 Optimal GMM Estimation

We may now pick the optimal C :

Proposition 15.17 (Optimal Choice of C). *Assume $\Omega := \mathbb{E}(u^2 z z')$ is invertible and let $Q := \mathbb{E}(z x')$. Then,*

$$C_{\text{OGMM}} := Q' \Omega^{-1} = \mathbb{E}(x z') [\mathbb{E}(u^2 z z')]^{-1}$$

minimizes the asymptotic variance of the GMM estimator.

Proof. With C_{OGMM} , the asymptotic variance is

$$\begin{aligned} V_{\text{OGMM}} &= (C_{\text{OGMM}} Q)^{-1} C_{\text{OGMM}} \Omega C_{\text{OGMM}}' (C_{\text{OGMM}} Q)^{-1} \\ &= (Q' \Omega^{-1} Q)^{-1} Q' \Omega^{-1} \Omega \Omega^{-1} Q (Q' \Omega^{-1} Q)^{-1} \\ &= (Q' \Omega^{-1} Q)^{-1}. \end{aligned}$$

We will show that $(CQ)^{-1} C \Omega C' (Q' C')^{-1} - (Q' \Omega^{-1} Q)$ is positive semidefinite. To do this, we write both terms in sandwich forms:

$$\begin{aligned} (Q' \Omega^{-1} Q)^{-1} &= (CQ)^{-1} C \Omega^{1/2} \times (\Omega^{1/2} Q (Q' \Omega^{-1} Q)^{-1} Q' \Omega^{-1/2}) \times \Omega^{1/2} C' (CQ)^{-1}; \\ (CQ)^{-1} C \Omega C' (Q' C')^{-1} &= (CQ)^{-1} C \Omega^{1/2} \times \Omega^{1/2} C' (Q' C')^{-1}. \end{aligned}$$

Note that since Ω is positive definite, $\Omega^{1/2}$ exists. Writing $R := \Omega^{-1/2} Q$, we then have

$$\begin{aligned} &(CQ)^{-1} C \Omega C' (Q' C')^{-1} - (Q' \Omega^{-1} Q) \\ &= (CQ)^{-1} C \Omega^{1/2} \left(I_{l+1} - R(R'R)^{-1} R' \right) \Omega^{1/2} C' (CQ)^{-1} \\ &= (CQ)^{-1} C \Omega^{1/2} M_R \Omega^{1/2} C' (CQ)^{-1} \geq 0, \end{aligned}$$

since M_R is positive semidefinite. □

Thus:

Proposition 15.18. *If $\hat{\Omega} \xrightarrow{p} \Omega := \mathbb{E}[u^2 z z']$, we say*

$$\hat{\beta}_{\text{OGMM}} = (X' Z \hat{\Omega}^{-1} Z' X)^{-1} X' Z \hat{\Omega}^{-1} Z' Y$$

is a (feasible) optimal GMM estimator. Note that

$$\sqrt{n}(\hat{\beta}_{\text{OGMM}} - \beta) \xrightarrow{d} \mathcal{N}\left(0, (Q' \Omega^{-1} Q)^{-1}\right).$$

The only problem that remains now is to get a consistent estimate of $\mathbb{E}[u^2 z z']$. What residuals do we use?

15.6 GMM Under Conditional Homoskedasticity, Two Stage Least Squares

Assumption 15.19 (Conditional Homoskedasticity).

$$\mathbb{E}[u^2|z] = \mathbb{E}[u^2] = \sigma^2.$$

Since in this case we have $\mathbb{E}[u^2 z z'] = \sigma^2 \mathbb{E}[z z']$, we set

$$C_{\text{OGMM}} = \frac{\mathbb{E}[z z']^{-1} \mathbb{E}[z x']}{\sigma^2}.$$

This is precisely the coefficients in the first stage regression of x on z . A feasible optimal GMM estimator is then given by

$$\begin{aligned} \hat{\beta}_{\text{OGMM}} &= (X'Z[\sigma^2 Z'Z]^{-1}ZX)^{-1}X'Z[\sigma^2 Z'Z]^{-1}Z'Y \\ &= (X'P_ZX)^{-1}X'P_ZY. \end{aligned}$$

This is the **two-stage least squares** estimator because it performs the previous task of reducing the number of moments by first regressing the columns X on Z using OLS.

Remark 15.20. The 2SLS estimator is still consistent for β even if conditional homoscedasticity does not hold, though it is not asymptotically optimal. We will leverage this fact to do optimal GMM estimation under heteroskedasticity. ☕

Remark 15.21. For inference, 2SLS estimator depends only on second moments, while the OGMM estimator depends on fourth moments and can thus be more noisy in finite sample, even when it is asymptotically optimal. ☕

The “first stage” regression is

$$X = Z\Pi + V,$$

where Π is a $(l+1) \times (k+1)$ matrix of parameters. It finds the $k+1$ linear combinations of the $l+1$ instruments that are closest to X in the Euclidean norm. The projection of each column of X on to Z is given by $P_ZX = Z\hat{\Pi}$. This is the sample analog of the choice

$$c_Z = [\mathbb{E}(zz')^{-1}\mathbb{E}(zx')]'\ z.$$

Note that for the included instruments X_j , we have $P_ZX_j = X_j$ because X_j is one of the columns of Z .

In the “second stage”, the exogenous and endogenous regressors X are replaced by the exogenous regressors and the projection of the endogenous regressors onto Z .

The original regression model is $Y = X\beta + U$. The model we actually estimate is $Y = P_ZX\beta + \epsilon$. Estimating this second stage regression by OLS produces

$$\hat{\beta}_{\text{2SLS}} = (X'P_ZX)^{-1}X'P_ZY = \hat{\beta}_{\text{OGMM}}.$$

Remark 15.22. If $l = k$, then since c is square and invertible, we have

$$\hat{\beta} = (\hat{C}Z'X)^{-1}\hat{C}Z'Y = (Z'X)^{-1}Z'Y = \hat{\beta}_{IV}.$$

All GMM estimators (and in particular the (2SLS) estimator) is the same as the IV estimator. ☕

The asymptotic distribution of the 2SLS estimator is

$$\sqrt{n}(\hat{\beta}_{2SLS} - \beta) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2 [Q' \mathbb{E}(zz')^{-1} Q]^{-1}\right).$$

Proposition 15.23. Let $\hat{U} := Y - X\hat{\beta}_{2SLS}$. A consistent estimator of σ^2 is given by

$$\hat{\sigma}^2 := \frac{\hat{U}'\hat{U}}{n}.$$

Proof. Since $\hat{U} = U - X(\hat{\beta}_{2SLS} - \beta)$,

$$\frac{\hat{U}'\hat{U}}{n} = \frac{U'U}{n} + o_p(1).$$

□

Our results may be summarized as follows:

Proposition 15.24 (Two Stage Least Squares). *Under homoskedasticity, $\hat{\beta}_{2SLS} := (X'P_ZX)^{-1}X'P_ZY$ is an asymptotically optimal GMM estimator and*

$$\sqrt{n}(\hat{\beta}_{2SLS} - \beta) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2 [Q' \mathbb{E}(zz')^{-1} Q]^{-1}\right),$$

where we recall the notation $Q = \mathbb{E}[zx']$. The asymptotic variance can be consistently estimated by

$$\hat{V} := n\hat{\sigma}^2(X'P_ZX)^{-1}.$$

15.7 GMM Under Heteroskedasticity

Under heteroskedasticity, the variance does not simplify. A consistent estimate of Ω is given by

$$\hat{\Omega} := \frac{1}{n} \sum \hat{u}_i z_i z_i'$$

where $\hat{u}_i := y_i - x_i'\hat{\beta}_{2SLS}$. The proof is identical to the heteroskedasticity case when considering OLS estimation. The result follows because $\hat{\beta}_{2SLS}$ is a \sqrt{n} -consistent estimator of β .

While $\hat{\beta}_{2SLS}$ is not asymptotically optimal, it does allow for consistent estimation of Ω because it depends only on Z , X , and Y .

Under heteroskedasticity, the optimal GMM estimator is

$$\hat{\beta}_{OGMM} = (X'Z\hat{\Omega}^{-1}Z'X)^{-1}X'Z\hat{\Omega}^{-1}Z'Y,$$

and

$$\sqrt{n}(\hat{\beta}_{\text{OGMM}}) \xrightarrow{d} \mathcal{N}(0, V),$$

where $V = (Q' \Omega^{-1} Q)^{-1}$ can be consistently estimated by

$$\hat{V} := \left(\frac{X'Z}{n} \hat{\Omega} \frac{Z'X}{n} \right)^{-1}.$$

Remark 15.25. Although $\hat{\beta}_{2\text{SLS}}$ is not asymptotically optimal, it does allow for consistent estimation of Ω because it depends only Z, Z, Y . Its finite sample performance is also not affected by the need to estimate Ω . ☕

Example 15.26. Consider the model

$$y = \beta x + u, \quad \mathbb{E}[u|x] = 0,$$

where x and y are scalar random variables. By the law of iterated expectations, we have $\mathbb{E}[uf(x)] = 0$ for any function f of x . Each choice of x produces a method of moments estimator of β as the solution to $\mathbb{E}[(y - \beta x)f(x)] = 0$. In particular, we have $\mathbb{E}[xu] = \mathbb{E}[x^2u] = 0$ and may thus use $z = (x, x^2)$ as instruments to estimate β . Here x is an included instrument, and x^2 is an excluded instrument.

Under conditional homoscedasticity, an optimal GMM estimator is the 2SLS estimator. In the first stage we regress all covariates on the instruments:

$$x = \gamma_0 x + \gamma_1 x^2 + v, \quad \mathbb{E}[v(x, x^2)] = 0.$$

But this simply gives $\gamma_0 = 1$ and $\gamma_1 = 0$. A perfect fit is obtained, thus the OLS estimator is asymptotically optimal.

Next consider the heteroskedasticity case. Let Z be the $n \times 2$ matrix of the observations of z and X be the $n \times 1$ matrix of the observations of x . We have

$$\hat{\beta}_{\text{OGMM}} = (X'Z\hat{\Omega}^{-1}Z'X)^{-1}X'Z\hat{\Omega}^{-1}Z'Y,$$

where $\hat{\Omega}$ is a consistent estimator of $\Omega = \mathbb{E}[u^2 zz']$. A choice of $\hat{\Omega}$ is

$$\hat{\Omega} = \frac{1}{n} \sum_{i=1}^n \hat{u}_i z_i z_i',$$

where $\hat{u}_i = y_i - x_i' \hat{\beta}_{\text{OLS}}$. ☕

16 Appendix A: A List of Theorems

Proposition 16.1.

- \mathbb{E} is linear.
- If $X \leq Y$ with probability 1, then $\mathbb{E}X \leq \mathbb{E}Y$.

Theorem 16.2 (Jensen's Inequality). *If X is such that $\mathbb{E}X$ and $\mathbb{E}g(X)$ exist and g is convex, then*

$$g(\mathbb{E}X) \leq \mathbb{E}g(X)$$

where the inequality is strict if g is strictly convex and X is not constant.

Proposition 16.3 (Properties of Conditional Expectation).

- *Linearity.*
- *(Law of iterated expectation) $\mathbb{E}(Y) = \mathbb{E}(\mathbb{E}(Y|X))$.*
- *(Taking out what is known) $\mathbb{E}(f(X) + g(X)Y|X) = f(X) + g(X)\mathbb{E}(Y|X)$.*

Proposition 16.4 (Law of total variance).

$$\mathbb{V}(Y) = \mathbb{E}(\mathbb{V}(Y|X)) + \mathbb{V}(\mathbb{E}(Y|X)),$$

where

$$\mathbb{V}(Y|X) := \mathbb{E}\{[Y - \mathbb{E}(Y|X)]^2|X\} = \mathbb{E}(Y^2|X) - \mathbb{E}(Y|X)^2.$$

Proposition 16.5.

$$\begin{aligned} X \text{ is independent of } Y &\implies X \text{ is mean independent of } Y \\ &\implies \mathbb{C}(X, Y) = 0. \end{aligned}$$

Proposition 16.6 (Existence of Moments). *Suppose $\mathbb{E}(|X|^k) < \infty$ for some $k > 0$. Then for $0 < r < k$, $\mathbb{E}(|X|^r) < \infty$.*

Theorem 16.7 (Chebychev's Inequality). *Suppose X^r is a non-negative integrable random variable for some $r > 0$. Then for any $\delta > 0$, we have*

$$\mathbb{P}(X \geq \delta) \leq \frac{\mathbb{E}(X^r)}{\delta^r}.$$

Lemma 16.8. $Y = 0$ almost surely if and only if $\mathbb{E}Y^2 = 0$.

Theorem 16.9 (Cauchy-Schwarz Inequality). *If $\mathbb{E}(X^2)$ and $\mathbb{E}(Y^2)$ exist, then*

$$|\mathbb{E}(XY)|^2 \leq \mathbb{E}(X^2)\mathbb{E}(Y^2)$$

with equality if and only if $X = aY$ almost surely for some constant a .

Corollary 16.10. *The correlation is bounded between -1 and 1 , with equality if and only if $X - \mathbb{E}X = b(Y - \mathbb{E}Y)$ for some constant b , which holds if and only if $X = a + bY$ for some constants a, b .*

Theorem 16.11 (Holder's Inequality). *If X and Y are random variables, then*

$$\mathbb{E}(|XY|) \leq \mathbb{E}(|X|^p)^{1/p} \mathbb{E}(|Y|^q)^{1/q}$$

for any $p, q > 0$ such that $1/p + 1/q = 1$.

Proposition 16.12. *Let X be a random vector such that $\mathbb{V}(X)$ exists. If A is a constant matrix and b a constant vector, then*

$$\mathbb{V}(AX + b) = A \mathbb{V}(X) A'.$$

Proposition 16.13. *Note that $\mathbb{E}(XX')$ is always positive semidefinite. Moreover, if it is invertible, then it is positive definite.*

Proposition 16.14. Suppose X is a $(k \times 1)$ random vector and $\mathbb{E}(XX')$ exists. Then $\mathbb{E}(XX')$ is invertible if and only if there is no perfect collinearity in X .

Proposition 16.15. If X_n converges in r -th mean to X , then $X_n \xrightarrow{p} X$.

Theorem 16.16 (Weak Law of Large Numbers). Suppose $\{X_i\}_{i \geq 1}$ is an iid sequence of random variables with $\mathbb{E}(X_i) = \mu$. Then, $\bar{X}_n \xrightarrow{p} \mu$.

Theorem 16.17 (WLLN for Moments). If $\mathbb{E}(X_i^k) < \infty$, then

$$\frac{1}{n} \sum_i X_i^k \xrightarrow{p} \mathbb{E}(X^k).$$

Proposition 16.18. Let X_n be a sequence of $(k \times 1)$ random vectors. Then,

- $X_n \xrightarrow{p} X$ if and only if $X_{n,i} \xrightarrow{p} X_i$ for $i = 1, \dots, k$.
- $X_n \rightarrow X$ in r -th mean if and only if $X_{n,i} \rightarrow X_i$ in r -th mean for $i = 1, \dots, k$.

Corollary 16.19. The weak law of large numbers for random vectors.

Theorem 16.20 (Continuous Mapping Theorem). Let $g : \mathbb{R}^k \rightarrow \mathbb{R}^n$ be continuous on $S \subset \mathbb{R}^k$ with $\mathbb{P}(X \in S) = 1$. Then the following hold:

- (i) if $X_n \xrightarrow{p} X$, then $g(X_n) \xrightarrow{p} g(X)$.
- (ii) If $X_n \xrightarrow{d} X$, then $g(X_n) \xrightarrow{d} g(X)$.

Theorem 16.21 (Slutsky's Theorem). Suppose $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$ for some constant c . Then,

$$X_n + Y_n \xrightarrow{d} X + c, \quad X_n Y_n \xrightarrow{d} Xc, \quad X_n/Y_n \xrightarrow{d} X/c \text{ provided } c \neq 0.$$

Proposition 16.22. Let $A_n \in \mathbb{R}^{P \times K}$ be a sequence of matrices converging in probability to a constant matrix A . Let B_n be a sequence of $(K \times 1)$ random vectors such that $B_n \xrightarrow{d} \mathcal{N}(\mu, \Sigma)$. Then,

$$A_n B_n \xrightarrow{d} A \mathcal{N}(\mu, \Sigma) \sim \mathcal{N}(A\mu, A\Sigma A').$$

Theorem 16.23 (Central Limit Theorem). Let $\{X_i\}_{i \geq 1}$ be an iid sequence of $(K \times 1)$ random vectors with mean μ and finite variance matrix Σ . Then

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

Theorem 16.24 (Delta Method). Let $\{X_n\}_{n \geq 1}$ be a sequence of $(K \times 1)$ random vectors and suppose

$$n^r (X_n - c) \xrightarrow{d} X$$

for some $r > 0$ and constant vector c . Let $g : \mathbb{R}^k \rightarrow \mathbb{R}^d$ be differentiable at the point c . Then,

$$n^r (g(X_n) - g(c)) \xrightarrow{d} Dg(c)X.$$

In particular, if $X \sim \mathcal{N}(0, \Sigma)$, then

$$n^r (g(X_n) - g(c)) \xrightarrow{d} \mathcal{N}(0, Dg(c)\Sigma Dg(c)').$$

Theorem 16.25. Suppose $\mathbb{E}(y^2) < \infty$ and $\mathbb{E}(x_j^2) < \infty$ for each $j = 1, \dots, k$. The function $g(x) := \mathbb{E}(y|x)$ is the best predictor of y given x under square loss. That is,

$$\mathbb{E}(y|x) \in \arg \min_g \mathbb{E}[(y - g(x))^2].$$

Proposition 16.26. If X is full column rank, then the OLS estimator is given by

$$\hat{\beta} = (X'X)^{-1}X'Y.$$

Theorem 16.27 (Projection Theorem). *Let $y \in \mathbb{R}^n$ and let S be any nonempty subspace of \mathbb{R}^n . There exists a unique point \hat{y} such that $\|y - \hat{y}\|$ is minimized over S . A necessary and sufficient condition for \hat{y} is that $y - \hat{y}$ is orthogonal to every vector in S .*

Proposition 16.28. *M_X projects a vector on to the $n - k$ dimensional vector space orthogonal to the column space of X .*

Proposition 16.29 (Projection Matrices, Elementary Properties).

- P_X is symmetric and idempotent.
- $P_X M_X = M_X P_X = 0$, since $(P_X Y)' M_X Y = Y' P_X M_X Y = 0$.
- $P_X X = X$, $M_X X = 0$.
- For every Y , $Y = P_X Y + M_X Y = \hat{Y} + \hat{U}$. Thus $\|Y\|^2 = \|P_X Y\|^2 + \|M_X Y\|^2$.

Proposition 16.30. *Assuming MLR.1–4, we have:*

- The OLS estimator is unbiased.
- Let $\Omega := \mathbb{V}(U|X) = \mathbb{E}(U'U|X)$. Then,

$$\mathbb{V}(\hat{\beta}|X) = (X'X)^{-1} X' \Omega X (X'X)^{-1}.$$

Thus, assuming also MLR.5, we have $\Omega = \sigma^2 I_n$ and so

$$\mathbb{V}(\hat{\beta}|X) = \sigma^2 (X'X)^{-1}.$$

Theorem 16.31 (Gauss-Markov). *Under MLR.1–5, the OLS estimator is the **best linear unbiased estimator**. That is, it achieves the smallest variance in the class of linear estimators⁶ that are also unbiased conditional on X . More precisely, let $\tilde{\beta}$ be the OLS estimator and let $\tilde{\beta} = A(X)Y$ satisfy $\mathbb{E}(\tilde{\beta}|X) = \beta$. We have then that $\mathbb{V}(\tilde{\beta}|X) - \mathbb{V}(\hat{\beta}|X)$ is positive semidefinite.*

Corollary 16.32. *Let r be an arbitrary $k \times 1$ vector. The Gauss-Markov theorem implies that $r' \hat{\beta}$ is the best linear unbiased estimator of $r' \beta$, since*

$$\mathbb{V}(r' \tilde{\beta}|X) - \mathbb{V}(r' \hat{\beta}|X) = r' [\mathbb{V}(\tilde{\beta}|X) - \mathbb{V}(\hat{\beta}|X)] r \geq 0.$$

In particular, taking $r = e_j$, we have $\mathbb{V}(\hat{\beta}_j|X) \leq \mathbb{V}(\tilde{\beta}_j|X)$.

Proposition 16.33. *Consider the model*

$$y = x' \beta + u.$$

And assume $\{y_i, x_i\}_{i=1}^n$ is iid.

- Suppose $\mathbb{E}(ux) = 0$ and $\mathbb{E}(xx')$ is invertible (so that $n^{-1} \sum x_i x_i'$ is invertible with probability approaching 1). Then, the OLS estimator is consistent; $\hat{\beta} \xrightarrow{p} \beta$.

Note that if we are doing prediction, $\mathbb{E}(ux) = 0$ is true by construction. When dealing with causality, however, this is an assumption we need to argue.

- If $\mathbb{V}(xu)$ also exists, then $\hat{\beta}$ is asymptotically normal. Specifically, we have $\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(0, \Sigma)$, where $\Sigma := \mathbb{E}(xx')^{-1} \mathbb{V}(xu) \mathbb{E}(xx')^{-1}$.

Proposition 16.34 (Asymptotic Covariance Under Homoskedasticity).

$$\Sigma = \sigma^2 \mathbb{E}(xx')^{-1}.$$

Proposition 16.35 (Consistent Estimators).

- The estimator

$$\hat{\sigma}^2 := \frac{1}{n - k} \sum \hat{u}_i^2 = \frac{SSR}{n - k}.$$

is consistent and unbiased.

⁶A linear estimator of β_j is a linear combinations of the $\{y_i\}$, with coefficients depending on $\{x_i\}$.

(ii) The estimator

$$\hat{\Sigma} := \hat{\sigma}^2 \left(\frac{1}{n} \sum x_i x_i' \right)^{-1}$$

is consistent.

Lemma 16.36. Let $\{Z_i\}_{i \geq 1}$ be a sequence of identically distributed random vectors such that $\mathbb{E}(\|Z_i\|^r) < \infty$. Then,

$$\frac{\max_{1 \leq i \leq n} \|Z_i\|}{n^{1/r}} \xrightarrow{p} 0.$$

Proposition 16.37. The estimator

$$\hat{\Sigma} := \left(\frac{1}{n} \sum x_i x_i' \right)^{-1} \left(\frac{1}{n} \sum \hat{u}_i^2 x_i x_i' \right) \left(\frac{1}{n} \sum x_i x_i' \right)^{-1}$$

is consistent. Note that the estimator can equivalently be written as

$$\hat{\Sigma} \equiv n(X'X)^{-1}X'\hat{\Omega}X(X'X)^{-1},$$

where

$$\hat{\Omega} := \begin{pmatrix} \hat{u}_1^2 & & & 0 \\ & \hat{u}_2^2 & & \\ & & \ddots & \\ 0 & & & \hat{u}_n^2 \end{pmatrix}.$$

Proposition 16.38. Consider the model $Y = X_1\beta_1 + X_2\beta_2 + U$. Let $\tilde{X}_2 := M_{X_1}X_2$ be the vector of residuals from a regression of (each column of X_2 on X_1). Denote $\tilde{Y} := M_{X_1}Y = \tilde{U}$ similarly. Then, the OLS estimator $\hat{\beta}_2$ equals:

- The OLS estimator of Y on \tilde{X}_2 .
- The OLS estimator of \tilde{Y} on \tilde{X}_2 .

In particular, we have the formula

$$\hat{\beta}_2 = (X_2'M_{X_1}X_2)^{-1}X_2'M_{X_1}Y.$$

Proposition 16.39 (Long and Short Regression). Suppose x_1 and x_2 has dimension d_1 and d_2 . Consider the short regression

$$y = x_1'\beta_{1s} + u_s, \quad \mathbb{E}[xu_s] = 0$$

and the long regression

$$y = x_1'\beta_{1l} + x_2'\beta_{2l} + u_l, \quad \mathbb{E}[xu_l] = 0.$$

Then,

$$\beta_{1a} = \beta_{1l} + \alpha\beta_{2l},$$

where α is a $d_1 \times d_2$ dimensional matrix, the j th column of which is the population regression coefficient vector from regressing the j th component of x_2 onto x_1 .

Theorem 16.40 (Yule-Frisch-Waugh-Lovell). β_2 is both the best linear predictor of \tilde{y} given \tilde{x}_2 and the best linear predictor of y given \tilde{x}_2 .

Proposition 16.41. Under homoskedasticity, the estimator

$$\hat{\sigma}^2 := \frac{SSR}{n-k} \sim \frac{\sigma^2}{n-k} \cdot \chi_{n-k}^2$$

is unbiased, consistent, and independent of $\hat{\beta}$.

Proposition 16.42. A positive definite and symmetric matrix A has a square root $A^{1/2}$ with inverse $A^{-1/2} = (A^{-1})^{1/2}$.

Proposition 16.43 (Testing Multiple Linear Restrictions). *If \hat{V} is a consistent estimator of the asymptotic variance of $\hat{\beta}$, then*

$$\begin{aligned} n \cdot (R\hat{\beta}_n - R\beta)' (R\hat{V}_n R')^{-1/2} (R\hat{V}_n R')^{-1/2} (R\hat{\beta}_n - R\beta) \\ = n \cdot (R\hat{\beta}_n - R\beta)' (R\hat{V}_n R')^{-1} (R\hat{\beta}_n - R\beta) \xrightarrow{d} \chi_P^2. \end{aligned}$$

Proposition 16.44. *We have the decomposition*

$$\text{ATE} = \text{ATT} \cdot \mathbb{P}(D = 1) + \text{ATU} \cdot \mathbb{P}(D = 0).$$

Proposition 16.45. *If, in addition to unconfoundedness, we have $D_i \perp\!\!\!\perp x_i$, then the naive comparison is the ATE.*

Proposition 16.46. *Suppose the model is $y = \beta_0 + \beta_1 x + v$, where x is a scalar. Suppose z satisfies Assumption 14.2. Then, the IV estimator for β_1*

$$\hat{\beta}_1^{\text{IV}} = \frac{\sum (z_i - \bar{z}) y_i}{\sum (z_i - \bar{z}) x_i}$$

is consistent and asymptotically normal. If $\mathbb{E}[v|z] = 0$ and $\mathbb{V}(v|z) = \sigma^2$, then

$$\sqrt{n}(\hat{\beta}_1^{\text{IV}} - \beta_1) \xrightarrow{d} \mathcal{N}\left(0, \frac{\sigma^2}{\mathbb{V}(x) \text{Corr}(x, z)^2}\right)$$

Proposition 16.47. *Assumption 15.3 is equivalent to assuming linear conditional means and constant treatment effects conditional on w .*

Proposition 16.48. *Assumption 15.1 and Assumption 15.3 imply $\mathbb{E}(u|z, w) = 0$. In particular, we have instrumental validity, $\mathbb{E}(zu) = 0$.*

Proposition 16.49. *In the case of exact identification, the IV estimator*

$$\hat{\beta}^{\text{IV}} := \left(\frac{1}{n} \sum z_i x_i' \right)^{-1} \left(\frac{1}{n} \sum z_i y_i \right) = (Z'X)^{-1} Z'Y$$

coincides with the Two Stage Least Squares (2SLS) estimator

$$\hat{\beta}_{2\text{SLS}} := (X' P_Z X)^{-1} X' P_Z Y.$$

Moreover, both estimators are consistent and asymptotically normal.

Proposition 16.50. *Suppose there is no perfect collinearity in z and $\mathbb{E}[zz'] < \infty$. Then, $\mathbb{E}[zx']$ is full rank if and only if $\mathbb{E}[zz']^{-1} \mathbb{E}[zx']$ is full rank.*

Proposition 16.51 (Consistency and Asymptotic Normality of GMM estimators). *Let $C \in \mathbb{R}^{(k+1) \times (l+1)}$ and suppose $\hat{C} \xrightarrow{P} C$. The estimator based on \hat{C} is consistent and asymptotically normal.*

Proposition 16.52 (Optimal Choice of C). *Assume $\Omega := \mathbb{E}(u^2 zz')$ is invertible and let $Q := \mathbb{E}(zx')$. Then,*

$$C_{\text{OGMM}} := Q' \Omega^{-1} = \mathbb{E}(zx') [\mathbb{E}(u^2 zz')]^{-1}$$

minimizes the asymptotic variance of the GMM estimator.

Proposition 16.53. *If $\hat{\Omega} \xrightarrow{P} \Omega := \mathbb{E}[u^2 zz']$, we say*

$$\hat{\beta}_{\text{OGMM}} = (X' Z \hat{\Omega}^{-1} Z' X)^{-1} X' Z \hat{\Omega}^{-1} Z' Y$$

is a (feasible) optimal GMM estimator. Note that

$$\sqrt{n}(\hat{\beta}_{\text{OGMM}} - \beta) \xrightarrow{d} \mathcal{N}\left(0, (Q' \Omega^{-1} Q)^{-1}\right).$$

Proposition 16.54. Let $\hat{U} := Y - X\hat{\beta}_{2\text{SLS}}$. A consistent estimator of σ^2 is given by

$$\hat{\sigma}^2 := \frac{\hat{U}'\hat{U}}{n}.$$

Proposition 16.55 (Two Stage Least Squares). Under homoskedasticity, $\hat{\beta}_{2\text{SLS}} := (X'P_ZX)^{-1}X'P_ZY$ is an asymptotically optimal GMM estimator and

$$\sqrt{n}(\hat{\beta}_{2\text{SLS}} - \beta) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2 [Q' \mathbb{E}(zz')^{-1}Q]^{-1}\right),$$

where we recall the notation $Q = \mathbb{E}[zx']$. The asymptotic variance can be consistently estimated by

$$\hat{V} := n\hat{\sigma}^2(X'P_ZX)^{-1}.$$

17 Appendix B: Common Distributions

$X \sim$	$\text{supp } \mathbb{P}(X = \cdot)$	$\mathbb{P}(X = x)$	\mathbb{E}	\mathbb{V}
Binomial(n, p)	$\{0, 1, \dots, n\}$	$\binom{n}{x} p^x (1-p)^{n-x}$	np	$np(1-p)$
Geometric(p)	\mathbb{N}	$(1-p)^{x-1} p$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Poisson(λ)	$\mathbb{N} \cup \{0\}$	$\frac{\lambda^x e^{-\lambda}}{x!}$	λ	λ

$X \sim$	$\text{supp } f$	f	\mathbb{E}	\mathbb{V}
Uniform(a, b)	$[a, b]$	$\frac{1}{b-a}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
$\mathcal{N}(\mu, \sigma^2)$	$(-\infty, \infty)$	$\frac{1}{\sqrt{2\pi}\sigma^2} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2
Exponential(λ) = Gamma($1, \lambda$)	$[0, \infty)$	$\lambda e^{-\lambda x}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Gamma(α, β)	$(0, \infty)$	$\frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)}$	$\frac{\alpha}{\beta}$	$\frac{\alpha}{\beta^2}$
Beta(α, β)	$(0, 1)$	$\frac{x^{\alpha-1} (1-x)^{\beta-1}}{B(\alpha, \beta)}$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

Definition 17.1. Let $Z_i \stackrel{\text{iid}}{\sim} \mathcal{N}(0, 1)$.

- Then

$$\sum Z_i^2 \sim \chi_k^2 = \text{Gamma}\left(\frac{k}{2}, \frac{1}{2}\right).$$

- If $W \sim \chi_n^2$ and $Z \perp W$, then

$$\frac{Z}{\sqrt{W/k}} \sim t_k.$$

- If $W_1 \sim \chi_{k_1}^2$, $W_2 \sim \chi_{k_2}^2$, and $W_1 \perp W_2$, then

$$\frac{W_1/k_1}{W_2/k_2} \sim F_{k_1, k_2}.$$

Note that $\text{supp } F = \mathbb{R}_+$, and $\mathbb{E}[F] = k_2/(k_1 + k_2) \approx 1$.

Proposition 17.2 (Approximations of Binomial).

- If $np_n \rightarrow \lambda$, then $\text{Binomial}(n, p) \rightarrow \text{Poisson}(\lambda)$.
- Using CLT, $\text{Binomial}(n, p) \approx \mathcal{N}(np, np(1-p))$.

Proposition 17.3 (\mathcal{N}). Let $X_i \stackrel{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ and put $S^2 := \frac{1}{n-1} \sum (X_i - \bar{X})^2$. Then

- $\mathbb{E}[S^2] = \sigma^2$. $\bar{X} \perp S^2$.
- $\sqrt{n} \frac{\bar{X} - \mu}{\sigma} \sim \mathcal{N}(0, 1)$.
- $(n-1) \frac{S^2}{\sigma^2} = \frac{1}{\sigma^2} \sum (X_i - \bar{X})^2 \sim \chi_{n-1}^2$.
- $\sqrt{n} \frac{\bar{X} - \mu}{S} \sim t_{n-1}$.

Proposition 17.4 (Multivariate Normal). Let $(X, Y) \sim \mathcal{N}_2$.

- $\mathbb{C}(X, Y) = 0$ if and only if $X \perp Y$.

- The MLEs are given by

$$\hat{\mu}_X = \bar{X}, \quad \hat{\sigma}_Y^2 =^{-1} \sum (X_i - \bar{X})^2, \quad R := \hat{\rho} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}.$$

R is known as the Pearson correlation coefficient.

- $(X, Y) \sim \mathcal{N}_2$ if and only if $aX + bY \sim \mathcal{N}$ for any $a, b \in \mathbb{R}$.
- The distribution of X conditional on $Y = y$ is given by

$$X|Y = y \sim \mathcal{N}\left(\mu_X + \rho \frac{\sigma_X}{\sigma_Y}(y - \mu_Y), \sigma_X^2(1 - \rho^2)\right).$$

Let $\mathbf{X} \sim \mathcal{N}_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. If $\boldsymbol{\Sigma}$ is positive definite, then \mathbf{X} has density

$$f(\mathbf{X}) = \frac{1}{(2\pi)^{k/2} \det(\boldsymbol{\Sigma})} \exp\left(-\frac{1}{2}(\mathbf{X} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{X} - \boldsymbol{\mu})\right).$$

Proposition 17.5 (Exponential Distribution, “Memoryless”).

$$\mathbb{P}(T \leq x + y | T > x) = \mathbb{P}(T \leq y).$$

Proposition 17.6 (Gamma Distribution). Let $X \sim \text{Gamma}(\alpha, \beta)$.

- (i) If $X_i \stackrel{\text{iid}}{\sim} \text{Gamma}(\alpha_i, \beta)$, then $\sum X_i \sim \text{Gamma}(\sum \alpha_i, \beta)$.
- (ii) If $\alpha > 1$, then $\mathbb{E}[1/X] = \frac{\beta}{\alpha-1}$.
- (iii) $\beta X \sim \text{Gamma}(\alpha, 1)$.

Index

- k -th central moment of X , 5
- k -th moment of X , 5

- absolutely continuous, 4
- adjusted R^2 , 20
- asymptotic distribution, 16
- asymptotic pivot, 15
- average treatment effect, 43
- average treatment effect on the treated, 43
- average treatment effect on the untreated, 43

- best linear predictor, 8
- best linear unbiased estimator, 22, 68
- best predictor of Y under square loss, 7
- binning estimator, 7

- classical measurement error, 53
- clustered standard errors, 48
- coefficient of determination, 20
- convergences in probability, 10
- converges in r -th mean, 10
- converges in distribution, 11

- design matrix, 18

- empirical distribution, 11
- endogenous, 51, 56
- Exclusion, 55
- Exogeneity, 55
- exogenous, 51, 56

- generalized least squares, 21
- GMM estimator, 60

- heteroskedasticity robust standard error, 34
- Homoskedasticity, 21

- Instrument relevance, 57
- instrumental validity, 56, 57, 70
- instrumental variables, 57
- IV estimator, 57, 70

- mean independent, 5
- method of moments estimator, 15

- omitted variable bias, 29
- optimal GMM estimator, 62, 70
- order condition, 57

- perfect collinearity, 8
- pivot, 15
- population R^2 , 20
- prediction error, 8
- projection matrix, 19

- rank condition, 57
- Relevance, 55
- relevance, 60
- residual, 8
- residual maker, 19

- saturated, 39
- standard error, 33

- test statistic, 15
- Two Stage Least Squares, 57, 70
- two stage least squares, 60
- two-stage least squares, 63

- unconfounded, 46