Introductory Econometrics: A Companion

Zhentao Shi

Contents

Chapter 1

Review of Probability

This version: September 7, 2016

1.1 Probability

1.1.1 Probability Space

- *Sample space* Ω is the collection of all possible outcomes.
- An *event* A is a subset of Ω.
- A σ-field, denoted by $\mathcal{F}$, is a collection of events such that: (i) $\emptyset \in \mathcal{F}$; (ii) if an event $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$; (iii) if $A_i \in \mathcal{F}$ for $i \in \mathbb{N}$, then $\bigcup_{i \in \mathbb{N}} A_i \in \mathcal{F}$.
- $(\Omega, \mathcal{F})$ is called a *measure space*.
- A function $\mu : \mathcal{F} \mapsto [0, \infty]$ is called a *measure* if it satisfies (i) $\mu(A) \geq 0$ for all $A \in \mathcal{F}$; (ii) if $A_i \in \mathcal{F}$, $i \in \mathbb{N}$, are mutually disjoint, then

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

- If $\mu(\Omega) = 1$, we call μ a *probability measure*. A probability measure is often denoted as P.

- $(\Omega, \mathcal{F}, P)$ is called a *probability space*.

1.1.2 Random Variable

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

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

  for any $B \in \mathcal{R}$, where $\mathcal{R}$ is the Borel σ-field on the real line. *Random variable* is an alternative name for a measurable function.

- $P_X : \mathcal{R} \mapsto [0,1]$ is also a probability measure if defined as $P_X(B) = P\left(X^{-1}(B)\right)$ for any $B \in \mathcal{R}$. This P_X is called the probability measure *induced* by the measurable function X .

- A measurable function is non-random; the randomness of the "random variable" is inherited from the underlying probability measure.

- Discrete random variable and continuous random variable.

1.1.3 Distribution Function

- (Cumulative) distribution function

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

- Properties of CDF: $\lim_{x\to-\infty} F(x) = 0$, $\lim_{x\to\infty} F(x) = 1$, non-decreasing, and right-continuous

$$\lim_{y\to x^+} F(y) = F(x).$$

- Probability density function (PDF): if there exists a function f such that for all x,

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

then f is called the PDF of X.

- Properties: $f(x) \geq 0$. $\int_a^b f(x)\, dx = F(b) - F(a)$

1.2 Expected Value

1.2.1 Integration

- X is called a *simple function* on a measurable space $(\Omega, \mathcal{F})$ if $X = \sum_i a_i 1\{A_i\}$ is a finite sum, where $a_i \in \mathbb{R}$ and $A_i \in \mathcal{F}$.

- Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and $a_i \geq 0$ for all i. The integral of X with respect to μ is

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

- Let X be a non-negative measurable function. The integral of X with

respect to μ is

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

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

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

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

1.2.2 Properties

- Elementary calculation: $E[X] = \sum_x xP(X = x)$ or $E[X] = \int xf(x)\,\mathrm{d}x$.
- $E[1\{A\}] = P(A)$.
- $E[X^r]$ is call the r-moment of X. Mean $\mu = E[X]$, variance $\mathrm{var}[X] = E\left[(X-\mu)^2\right]$, skewness $E\left[(X-\mu)^3\right]$ and kurtosis $E\left[(X-\mu)^4\right]$.

1.3 Multivariate Random Variable

- Bivariate random variable: $X : \Omega \mapsto \mathbb{R}^2$.
- Multivariate random variable $X : \Omega \mapsto \mathbb{R}^n$.

- Joint CDF: $F(x_1, \ldots, x_n) = P(X_1 \leq x_1, \ldots, X_n \leq x_n)$. Joint PDF is defined similarly.

- X and Y are *independent* if $P(X \in A, Y \in B) = P(X \in A) P(Y \in B)$ for all A and B.

1.3.1 Elementary Formulas

- conditional density $f(Y|X) = f(X,Y)/f(X)$

- marginal density $f(Y) = \int f(X,Y)\, dX$.

- conditional expectation $E[Y|X] = \int Y f(Y|X)\, dY$

- proof of law of iterated expectation

$$\begin{aligned} & E[E[Y|X]] = \int E[Y|X] f(X)\, dX \\ = & \int \left(\int Y f(Y|X)\, dY \right) f(X)\, dX = \int \int Y f(Y|X) f(X)\, dY dX \\ = & \int \int Y f(X,Y)\, dY dX = \int Y \left(\int f(X,Y)\, dX \right) dY = \int Y dY = E[Y]. \end{aligned}$$

- conditional probability, or Bayes' Theorem $P(A|B) = \frac{P(A,B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}$.

1.3.2 Law of Iterated Expectations

- Given a probability space $(\Omega, \mathcal{F}, P)$, a sub σ-algebra $\mathcal{G} \subset \mathcal{F}$ and a $\mathcal{F}$-measurable function X with $E|X| < \infty$, the *conditional expectation* $E[X|\mathcal{G}]$ is defined as a $\mathcal{G}$-measurable function such that $\int_A X dP = \int_A E[X|\mathcal{G}]\, dP$ for all $A \in \mathcal{G}$.

- Law of iterated expectations

$$E\left[E\left[Y|X\right]\right] = E\left[Y\right]$$

  is a trivial fact from the definition of the conditional expectation by taking $A = \Omega$.

- Properties of conditional expectations

  1. $E\left[E\left[Y|X_1, X_2\right]|X_1\right] = E\left[Y|X_1\right]$
  2. $E\left[E\left[Y|X_1\right]|X_1, X_2\right] = E\left[Y|X_1\right]$
  3. $E\left[h\left(X\right)Y|X\right] = h\left(X\right)E\left[Y|X\right]$

Chapter 2

Regression Model

This version: February 20, 2017

Notation: in this note, y is a scale random variable, and x is a $K \times 1$ random vector.

2.1 Conditional Expectation Model

A regression model can be written as

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

where $m(x) = E[y|x]$ is called the *conditional mean function*, and $\epsilon = y - m(x)$ is called the *regression error*. Such an equation holds for (y, x) that follows any joint distribution, as long as $E[y|x]$ exists. The error term ϵ satisfies these properties:

- $E[\epsilon|x] = 0$,

- $E[\epsilon] = 0$,
- $E[h(x)\epsilon] = 0$, where h is a function of x.

The last property implies that ϵ is uncorrelated with any function of x.

If we are interested in predicting y given x, then the conditional mean function $E[y|x]$ is "optimal" in terms of the *mean squared error* (MSE).

As y is not a deterministic function of x, we cannot predict it with certainty. In order to evaluate different methods of prediction, we must therefore propose a criterion for comparison. For an arbitrary prediction method $g(x)$, we employ a *loss function* $L(y, g(x))$ to measure how wrong is the prediction, and the expected value of the loss function is called the *risk* $R(y, g(x))$. The *quadratic loss function* is defined as

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

and the corresponding risk

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

is called the MSE.

Due to its operational convenience, MSE is one of the most widely used criterion. Under MSE, the conditional expectation function happens to be the best prediction method for y given x. In other words, the conditional mean function $m(x)$ minimizes the MSE.

We can take a guess-and-verify this claim of optimality. For an arbitrary

$g(x)$, the risk can be decomposed into three terms

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

The first term is irrelevant to $g(x)$. The second term $2E\left[\epsilon\left(m(x)-g(x)\right)\right] = 0$ is again irrelevant of $g(x)$. The third term, obviously, is minimized at $g(x) = m(x)$.

2.2 Linear Projection Model

As discussed in the previous section, we are interested in the conditional mean function $m(x)$. However, remind that

$$m(x) = E[y|x] = \int y f(y|x)\,\mathrm{d}y$$

is a complex function of x, as it depends on the joint distribution of (y, x).

A particular form of the conditional mean function is

$$m(x) = x'\beta,$$

a linear function of x.

Remark. The linear function is not as restrictive as one might thought. It can be used to generate some nonlinear (in random variables) effect if we

re-define x. For example, if

$$y = x_1\beta_2 + x_2\beta_2 + x_1x_2\beta_3 + e,$$

then $\frac{\partial}{\partial x_1} m(x_1, x_2) = \beta_1 + x_2\beta_3$, which is nonlinear in x_1, while it is still linear in the parameter β if we define a set of new regressors as $(\tilde{x}_1, \tilde{x}_2, \tilde{x}_3) = (x_1, x_2, x_1x_2)$.

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

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

Even though in general $m(x) \neq x'\beta$, the linear form $x'\beta$ is still useful as an approximation, as will be clear soon. Therefore, we may write the linear regression model, or the *linear projection model*, as

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

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

Remark. If a constant is included in x as a regressor, we have $E[e] = 0$.

The coefficient β in the linear projection model has a straightforward closed-form. Multiplying x on both sides and taking expectation, we have

$E[xy] = E[xx']\beta$. If $E[xx']$ is invertible, we can explicitly solve

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

Now we justify $x'\beta$ as an approximation to $m(x)$. Indeed, $x'\beta$ is the optimal *linear* predictor in terms of MSE; in other words,

$$\beta = \arg\min_{b \in \mathbb{R}^K} E\left[\left(y - x'b\right)^2\right]. \tag{2.1}$$

This fact can be verified by taking the first-order condition of the above minimization problem $\frac{\partial}{\partial \beta} E\left[\left(y - x'\beta\right)^2\right] = 2E\left[x\left(y - x'\beta\right)\right] = 0$.

In the meantime, $x'\beta$ is also the best *linear* approximation to $m(x)$. If we replace y in (2.1) by $m(x)$, we solve the minimizer as

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

Therefore β is also the best linear approximation to $m(x)$ in terms of MSE.

2.2.1 Subvector Regression

Sometimes we are interested in a subvector of β. For example, when we include an intercept and some variables in x, we are often more interested in the slope coefficients—the parameters associated with the random regressors—as they represent the size of effect of these explanatory factors. In such a regression

$$y = \beta_1 + x'\beta_2 + e,$$

we take an expectation to get $E[y] = \beta_1 + E[x]' \beta_2$. Take the difference of the two equations,

$$y - E[y] = (x - E[x])' \beta_2.$$

Therefore, we can explicitly solve β_2 as

$$\beta_2 = \left(E\left[(x - E[x])(x - E[x])'\right]\right)^{-1} E[(x - E[x])(y - E[y])] = (\text{var}(x))^{-1} \text{cov}(x, y),$$

This is a special case of the subvector regression.

To discuss the general case, we need to know *the formula of the inverse of a partitioned matrix*. If $Q = \begin{pmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{pmatrix}$ is a symmetric and positive definite matrix, then

$$Q^{-1} = \begin{pmatrix} (Q_{11} - Q_{12}Q_{22}Q_{21})^{-1} & -(Q_{11} - Q_{12}Q_{22}Q_{21})^{-1} Q_{12}Q_{22}^{-1} \\ -(Q_{22} - Q_{21}Q_{11}Q_{12})^{-1} Q_{21}Q_{11}^{-1} & (Q_{22} - Q_{21}Q_{11}Q_{12})^{-1} \end{pmatrix}.$$

We apply the above formulate to the expression of β, and we obtain $\beta_1 = A_{11\cdot 2}^{-1} A_{1y\cdot 2}$, where

$$\begin{aligned} A_{11\cdot 2} &= E\left[x_1 x_1'\right] - E\left[x_1 x_2'\right]\left(E\left[x_2 x_2'\right]\right)^{-1} E\left[x_2 x_1'\right] \\ A_{1y\cdot 2} &= E[x_1 y] - E\left[x_1 x_2'\right]\left(E\left[x_2 x_2'\right]\right)^{-1} E[x_2 y]. \end{aligned}$$

This is a brutal force approach for the explicit expression of the subvector β_1.

Alternatively, we can proceed in two steps. First, we run a multiple

regression[1]

$$\begin{aligned} x_1 &= x_2'\gamma + u \\ E[x_2 u] &= 0 \end{aligned}$$

so that the regressor error

$$u = x_1 - x_2'\gamma = x_1 - x_2' \left(E\left[x_2 x_2'\right]\right)^{-1} E\left[x_2 x_1'\right] = x_1 - E\left[x_1 x_2'\right] \left(E\left[x_2 x_2'\right]\right)^{-1} x_2.$$

We then run a simple regression of y on u, and the coefficient is

$$\theta = \left(E\left[uu'\right]\right)^{-1} E\left[u'y\right].$$

The nominator is

$$E\left[u'y\right] = E\left[x_1 y\right] - E\left[x_1 x_2'\right] \left(E\left[x_2 x_2'\right]\right)^{-1} E\left[x_2 y\right] = A_{1y\cdot 2}$$

and the denominator is

$$E\left[uu'\right] = E\left[\left(x_1 - E\left[x_1 x_2'\right] \left(E\left[x_2 x_2'\right]\right)^{-1} x_2\right)\left(x_1 - E\left[x_1 x_2'\right] \left(E\left[x_2 x_2'\right]\right)^{-1} x_2\right)'\right] = A_{11\cdot 2}.$$

It turns out $\beta_2 = \theta$.

While we can derive the expression of β_1 as a subvector of β, why do we come up with the two-step derivation? The latter makes clear that the coefficient represents the *partial effect* of the associate random variable.

[1] We do allow x_1 to be a vector. However, one may find it is easier to consider the special case that x_1 is a scalar random variable.

2.2.2 Omitted Variable Bias

We write the *long regression* as

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

and the *short regression* as

$$y = x_1'\gamma_1 + \gamma_2 + u.$$

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

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

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

and the short regression becomes

$$\tilde{y} = \tilde{x}_1'\gamma_1 + \tilde{u},$$

where *tilde* denotes the demeaned variable.

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

regression coefficient

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

Therefore, $\gamma_1 = \beta_1$ if and only if $E\left[\tilde{x}_1\tilde{x}_2'\right]\beta_2 = 0$, which demands either $E\left[\tilde{x}_1\tilde{x}_2'\right] = 0$ or $\beta_2 = 0$.

Obviously we prefer to run the long regression to attain β_1 if possible. However, sometimes x_2 is simply unobservable so the long regression is infeasible. When only the short regression is available, in some cases we are able to sign the bias, meaning that we know whether γ_1 is bigger or smaller than β_1.

Chapter 3

Least Squares

This version: February 20, 2017

Notation: y_i is a scalar, and x_i is a $K \times 1$ vector. Y is an $n \times 1$ vector, and X is an $n \times K$ matrix.

3.1 Algebra of Least Squares

3.1.1 OLS estimator

As we have learned from the linear project model, the parameter β

$$\begin{aligned} y_i &= x_i'\beta + e_i \\ E[x_i e_i] &= 0 \end{aligned}$$

can be written as $\beta = \left(E\left[x_i x_i'\right]\right)^{-1} E\left[x_i y_i\right]$.

While population is something imaginary, in reality we possess a sample of n observations. We thus replace the population mean $E\left[\cdot\right]$ by the sample

mean, and the resulting estimator is

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

This is one way to motivate the OLS estimator.

Alternatively, we can derive the OLS estimator from minimizing the sum of squared residuals

$$Q(\beta) = \sum_{i=1}^{n} \left(y_i - x_i'\beta \right)^2 = (Y - X\beta)'(Y - X\beta).$$

By the first-order condition

$$\frac{\partial}{\partial \beta} Q(\beta) = -2X'(Y - X\beta),$$

the optimality condition gives exactly the same $\widehat{\beta}$. Moreover, the second-order condition

$$\frac{\partial^2}{\partial \beta \partial \beta'} Q(\beta) = 2X'X$$

shows that $Q(\beta)$ is convex in β. ($Q(\beta)$ is strictly convex in β if $X'X$ is positive definite.)

Here we introduce some definitions and properties in OLS estimation.

- Fitted value: $\widehat{Y} = X\widehat{\beta}$.
- Projector: $P_X = X(X'X)^{-1} X$; Annihilator: $M_X = I_n - P_X$.
- $P_X M_X = M_X P_X = 0$.
- If $AA = A$, we call it an idempotent matrix. Both P_X and M_X are

idempotent.

- Residual: $\widehat{e} = Y - \widehat{Y} = Y - X\widehat{\beta} = M_X Y = M_X (X\beta + e) = M_X e$.
- $X'\widehat{e} = X M_X e = 0$.
- $\frac{1}{n}\sum_{i=1}^n \widehat{e}_i = 0$ if x_i contains a constant.

3.1.2 Goodness of Fit

The so-called R-square is the most popular measure of goodness-of-fit in the linear regression. R-square is well defined only when a constant is included in the regressors. Let $M_\iota = I_n - \frac{1}{n}\iota\iota'$, where ι is an $n \times 1$ vector of 1's. M_ι is the *demeaner*, in the sense that $M_\iota (z_1, \ldots, z_n)' = (z_1 - \overline{z}, \ldots, z_n - \overline{z})'$, where $\bar{z} = \frac{1}{n}\sum_{i=1}^n z_i$. For any X, we can decompose $Y = P_X Y + M_X Y = \widehat{Y} + \widehat{e}$. The total variation is

$$Y' M_\iota Y = \left(\widehat{Y} + \widehat{e}\right)' M_\iota \left(\widehat{Y} + \widehat{e}\right) = \widehat{Y}' M_\iota \widehat{Y} + 2\widehat{Y}' M_\iota \widehat{e} + \widehat{e}' M_\iota \widehat{e} = \widehat{Y}' M_\iota \widehat{Y} + \widehat{e}'\widehat{e}$$

where the last equality follows by $M_\iota \widehat{e} = \widehat{e}$ as $\frac{1}{n}\sum_{i=1}^n \widehat{e}_i = 0$, and $\widehat{Y}'\widehat{e} = Y' P_X M_X e = 0$. R-square is defined as $\widehat{Y}' M_\iota \widehat{Y} / Y' M_\iota Y$.

3.1.3 Frish-Waugh-Lovell Theorem

This theorem is the sample version of the subvector regression.

If $Y = X_1\beta_1 + X_2\beta_2 + e$, then $\widehat{\beta}_1 = \left(X_1' M_{X_2} X_1\right)^{-1} X_1' M_{X_2} Y$.

3.2 Statistical Properties of Least Squares

To talk about the statistical properties in finite sample, we impose the following assumptions.

1. The data $(y_i, x_i)_{i=1}^n$ is a random sample from the same data generating process $y_i = x_i'\beta + e_i$.

2. $e_i | x_i \sim N\left(0, \sigma^2\right)$.

3.2.1 Maximum Likelihood Estimation*

Under the normality assumption, $y_i | x_i \sim N\left(x_i'\beta, \gamma\right)$, where $\gamma = \sigma^2$. The *conditional* likelihood of observing a sample $(y_i, x_i)_{i=1}^n$ is

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

and the (conditional) log-likelihood function is

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

Therefore, the maximum likelihood estimator (MLE) coincides with the OLS estimator, and $\widehat{\gamma}_{\text{MLE}} = \widehat{e}'\widehat{e}/n$.

3.2.2 Finite Sample Distribution

We can show the finite-sample exact distribution of $\widehat{\beta}$. *Finite sample distribution* means that the distribution holds for any n; it is in contrast to *asymptotic distribution*, which holds only when n is arbitrarily large.

Since

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

we have the estimator $\widehat{\beta}|X \sim N\left(\beta, \sigma^2 \left(X'X\right)^{-1}\right)$, and

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

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

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

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

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

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

The numerator follows a standard normal, and the denominator follows $\frac{1}{n-K}\chi^2\left(n - K\right)$. Moreover, the numerator and the denominator are independent. As a result, $T_k \sim t\left(n - K\right)$.

3.2.3 Mean and Variance

Now we relax the normality assumption and statistical independence. Instead, we assume a random sample and

$$\begin{aligned} y_i &= x_i'\beta + e_i \\ E[e_i|x_i] &= 0 \qquad (3.1) \\ E\left[e_i^2|x_i\right] &= \sigma^2. \qquad (3.2) \end{aligned}$$

(3.1) is the *mean independence* assumption, and (3.2) is the *homoskedasticity* assumption.

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

These assumptions are about the first and second moment of e_i conditional on x_i. Unlike the normality assumption, they do not restrict the entire distribution of e_i.

- Unbiasedness:

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

  Unbiasedness does not rely on homoskedasticity.

- Variance:

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

3.2.4 Gauss-Markov Theorem*

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

There are numerous linearly unbiased estimators. For example, $\left(Z'X\right)^{-1}Z'y$ for $z_i = x_i^2$ is unbiased because $E\left[\left(Z'X\right)^{-1}Z'y\right] = E\left[\left(Z'X\right)^{-1}Z'\left(X\beta+e\right)\right] = \beta$.

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

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

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

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

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

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

where the last equality follows as

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

Therefore $A'A - \left(X'X\right)^{-1}$ is a positive semi-definite matrix. The variance of any $\tilde{\beta}$ is no smaller than the OLS estimator $\widehat{\beta}$.

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

Chapter 4

Large Sample Theory

This version: February 20, 2017

4.1 Asymptotics

Asymptotic theory is concerned about the behavior of statistics when the sample size is arbitrarily large. It is a useful approximation technique to simplify complicated finite-sample analysis.

4.1.1 Modes of Convergence

Convergence of a deterministic sequence means that for any $\varepsilon > 0$, there exists an $N(\varepsilon)$ such that for all $n > N(\varepsilon)$, we have $|z_n - z| < \varepsilon$. We say z is the limit of z_n, and write as $z_n \to z$.

In contrast to the convergence of a deterministic sequence, we are interested in the convergence of random variables. Since a random variable is "random", we must define clearly what "convergence" means. Several

modes of convergence are often encountered.

- Convergence almost surely*
- Convergence in probability: $\lim_{n\to\infty} P(\omega : |Z_n(\omega) - z| < \varepsilon) = 1$ for any $\varepsilon > 0$.
- Squared-mean convergence: $\lim_{n\to\infty} E\left[(z_n - z)^2\right] = 0$.

Example 1. z_n is a binary random variable: $z_n = \sqrt{n}$ with probability $1/n$, and $z_n = 0$ with probability $1 - 1/n$. Then $z_n \xrightarrow{p} 0$ but $z_n \overset{m.s.}{\nrightarrow} 0$.

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

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

Convergence in distribution is about *pointwise* convergence of CDF, not the random variables themselves.

Example 2. Let $x \sim N(0,1)$. If $z_n = x + 1/n$, then $z_n \xrightarrow{p} x$ and of course $z_n \xrightarrow{d} x$. However, if $z_n = -x + 1/n$, or $z_n = y + 1/n$ where $y \sim N(0,1)$ is independent of x, then $z_n \xrightarrow{d} x$ but $z_n \overset{p}{\nrightarrow} x$.

Cramér-Wold device handles convergence in distribution for random vectors? We say a sequence of K-dimensional random vectors (X_n) converge in distribution to X if we have $\lambda' X_n \xrightarrow{d} \lambda' X$ for any $\lambda \in \mathbb{R}^K$ with $\lambda'\lambda = 1$.

4.1.2 Law of Large Numbers[1]

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

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

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

- Chebyshev LLN: if $(z_1, \ldots, z_n)$ is a sample of i.i.d. observations, $E[z_1] = \mu$, and $\sigma^2 = \operatorname{var}[x_1] < \infty$ exists, then $\frac{1}{n}\sum_{i=1}^{n} z_i - \mu \xrightarrow{p} 0$.

Chebyshev LLN utilizes

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

Chebyshev inequality is a special case of

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

It is easy to verify Markov inequality.

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

[1]Though the results in this section hold for convergence almost surely, for simplicity we state them in terms of convergence in probability.

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

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

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

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

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

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

4.1.3 Central Limit Theorem

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

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} z_i \xrightarrow{d} N\left(0, \sigma^2\right). \tag{4.1}$$

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

Lindeberg-Levy CLT is the simplest CLT.

- If the sample is i.i.d., $E[x_1] = 0$ and $\text{var}\left[x_1^2\right] = \sigma^2 < \infty$, then (4.1) holds.

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

If $E\left[|x|^k\right] < \infty$ for a positive integer k, then

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

Under the assumption of Lindeberg-Levy CLT,

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

for all i, and by independence we have

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

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

- Lindeberg-Feller CLT: i.n.i.d., and *Lindeberg condition*: for any fixed $\varepsilon > 0$,
$$\frac{1}{s_n^2}\sum_{i=1}^{n}\int_{|x_i|>\varepsilon s_n} x_i^2 dP x_i \to 0$$
where $s_n = \left(\sum_{i=1}^{n}\sigma_i^2\right)^{1/2}$.
- Lyapunov CLT: i.n.i.d, finite $E\left[|x|^3\right]$.

4.1.4 Tools for Transformations

The original forms of LLN or CLT only deal with sample means. However, most of the econometric estimators of interest are functions of sample means. Therefore, we need tools to handle transformations.

- Small op: $x_n = o_p(r_n)$ if $x_n/r_n \xrightarrow{p} 0$.
- Big Op: $x_n = O_p(r_n)$ if for any $\varepsilon > 0$, there exists a $c > 0$ such that $P(x_n/r_n > c) < \varepsilon$.
- Continuous mapping theorem 1: If $x_n \xrightarrow{p} a$ and $f(\cdot)$ is continuous at a, then $f(x_n) \xrightarrow{p} f(a)$.

- Continuous mapping theorem 2: If $x_n \xrightarrow{d} x$ and $f(\cdot)$ is continuous almost surely on the support of x, then $f(x_n) \xrightarrow{d} f(x)$.
- Slutsky's Theorem: If $x_n \xrightarrow{d} x$ and $y_n \xrightarrow{p} a$, then
  - $x_n + y_n \xrightarrow{d} x + a$
  - $x_n y_n \xrightarrow{d} ax$
  - $x_n / y_n \xrightarrow{d} x/a$ if $a \neq 0$.
- Delta method: if $\sqrt{n}\left(\widehat{\theta} - \theta_0\right) \xrightarrow{d} N(0, \Omega)$, and $f(\cdot)$ is continuously differentiable at θ_0, then

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

4.2 Asymptotic Properties of OLS

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

4.2.1 Consistency

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

$$\widehat{\beta} - \beta = \left(X'X\right)^{-1} X'e = \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} x_i e_i. \tag{4.2}$$

The first term

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

and the second term

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

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

4.2.2 Asymptotic Normality

In finite sample, $\widehat{\beta}$ is a random variable. We have shown the distribution of $\widehat{\beta}$ under normality in the previous lecture. Without the restrictive normality assumption, how can we characterize the randomness of the OLS estimator? If we multiply $\sqrt{n}$ on both sides of (4.2), we have

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

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

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

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

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

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

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

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

4.2.3 Estimation of the Variance*

To show the finiteness of the variance, $\Sigma = E\left[x_i x_i' e_i^2\right]$. Let $z_i = x_i e_i$, so $\Sigma = E\left[z_i z_i'\right]$. Because of the Cachy-Schwarz inequality,

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

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

For the estimation of variance, homoskedastic,

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

The second term

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

The third term

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

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

For general heteroskedasticity,

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

The third term is bounded by

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

where the third inequality follows by $\left(a_1+\cdots+a_K\right)^2\leq K\left(a_1^2+\cdots+a_K^2\right)$.

The second term is bounded by

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

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

Chapter 5

Hypothesis Testing

This version: February 20, 2017

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

5.1 Hypothesis Testing

A *hypothesis* is a statement about the parameter space Θ. The *null hypothesis* Θ_0 is a subset of Θ of interest, typically suggested by some scientific theory. The *alternative hypothesis* $\Theta_1 = \Theta \backslash \Theta_0$ is the complement of Θ_0. *Hypothesis testing* is a decision whether to accept the null hypothesis or to reject it according to the observed evidence.

A *test function* is a mapping

$$\phi : \mathcal{X}^n \mapsto \{0, 1\},$$

Table 5.1: Decisions and Status

	accept H_0 (reject H_1)	reject H_0 (accept H_1)
H_0 true (H_1 false)	correct decision	Type I error
H_0 false (H_1 true)	Type II error	correct decision

size = P(reject $H_0|H_0$ true)
power = P(reject $H_0|H_0$ false)

where $\mathcal{X}$ is the sample space. We accept the null hypothesis if $\phi(\mathbf{x}) = 0$, or reject it if $\phi(\mathbf{x}) = 1$. The *acceptance region* is defined as $A_\phi = \{\mathbf{x} \in \mathcal{X}^n : \phi(\mathbf{x}) = 0\}$, and the *rejection region* is $R_\phi = \{\mathbf{x} \in \mathcal{X}^n : \phi(\mathbf{x}) = 1\}$. The *power function* of the test ϕ is

$$\beta_\phi(\theta) = P_\theta(\phi(\mathbf{X}) = 1) = E_\theta(\phi(\mathbf{X})).$$

The power function measures, at a given point, the probability that the test function rejects the null.

The *power* of ϕ at θ for some $\theta \in \Theta_1$ is defined as the value of $\beta_\phi(\theta)$. The *size* of the test ϕ is define as $\alpha = \sup_{\theta \in \Theta_0} \beta_\phi(\theta)$. Notice that the definition of power depends on a θ in the alternative, whereas that of size is independent of θ as it takes the supremum over the null. The *level* of the test ϕ is a value $\alpha \in (0, 1)$ such that $\alpha \geq \sup_{\theta \in \Theta_0} \beta_\phi(\theta)$, which is often used when it is difficult to attain the exact supremum. The *probability of committing Type I error* is $\beta_\phi(\theta)$ for some $\theta \in \Theta_0$. The *probability of committing Type II error* is $1 - \beta_\phi(\theta)$ for $\theta \in \Theta_1$; in other words, it is one minus the power at θ.

There has been a philosophical debate for decades about the hypothesis testing framework. At present the prevailing framework in statistics educa-

tion is the frequentist perspective. A frequentist views the parameter as a fixed constant, and they are conservative about the Type I error. Only if overwhelming evidence is demonstrated should a researcher reject the null.

Under the philosophy of protecting the null hypothesis, a desirable test should have a small level. Conventionally we take $\alpha = 0.01$, 0.05 or 0.1. There can be many tests of the correct size.

Example 3. A trivial test function, $\phi(\mathbf{X}) = 1\{0 \leq U \leq \alpha\}$, where U is a random variable from a uniform distribution on $[0,1]$, has correct size but no power. Another trivial test function $\phi(\mathbf{X}) = 1$ has the biggest power but incorrect size.

Usually, we design a test by proposing a test statistic $T_n : \mathcal{X}^n \mapsto \mathbb{R}^+$ and a critical value $c_{1-\alpha}$, and then define

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

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

Example 4. The concept of *level* is useful if we do not have information to derive the exact size of a test. If $(X_{1i}, X_{2i})_{i=1}^n$ are randomly drawn from some unknown joint distribution, but we only know that the marginal distribution is $X_{ji} \sim N(\theta_j, 1)$, for $j = 1, 2$. In order to test the joint hypothesis $\theta_1 = \theta_2 = 0$, we can construct a test function

$$\phi(\mathbf{X}_1, \mathbf{X}_2) = 1\left\{\left\{\sqrt{n}\left|\overline{X}_1\right| \geq c_{1-\alpha/4}\right\} \cup \left\{\sqrt{n}\left|\overline{X}_2\right| \geq c_{1-\alpha/4}\right\}\right\},$$

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

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

where the inequality follows by the Bonferroni inequality $P\left(A \cup B\right) \leq P\left(A\right) + P\left(B\right)$. Therefore, the level of $\phi\left(\mathbf{X}_1,\mathbf{X}_2\right)$ is α, but the exact size is unknown without the knowledge of the joint distribution. (Even if we know the correlation of X_{1i} and X_{2i}, putting two marginally normal distributions together does not make a jointly normal vector in general.)

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

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

uniformly over $\theta \in \Theta_1$.

Example 5. Suppose a random sample of size 6 is generated from

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

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

$\theta > 0$. All tests in

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

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

$$\beta_\phi(\theta) = P\left(\bar{X} \geq \frac{c}{\sqrt{6}} \right) = P\left(\sqrt{6}\left(\bar{X} - \theta\right) \geq c - \sqrt{6}\theta \right) = 1 - \Phi\left(c - \sqrt{6}\theta \right).$$

The test function

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

is the most powerful test in Ψ.

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

$$\sup_{\theta \in \Theta_0} P_\theta\left(T_n(\mathbf{x}) \leq T_n(\mathbf{X}) \right).$$

In the above expression, $T_n(\mathbf{x})$ is the realized value of the test statistic T_n, while $T_n(\mathbf{X})$ is the random variable generated by $\mathbf{X}$ under the null $\theta \in \Theta_0$. p-value is closely related to the corresponding test. When p-value is smaller than the specified test size α, the test rejects the null hypothesis.

p-value is a measure whether the data is consistent with the null hypothesis, or whether the evidence from the data is compatible with the null hypothesis. p-value is *not* the probability that the null hypothesis is true. Under the frequentist perspective, the null hypothesis is either true or false, with certainty. The randomness of a test comes only from sampling, not

from the hypothesis.

5.2 Confidence Interval

An *interval estimate* is a function $C : \mathcal{X}^n \mapsto \{\Theta' : \Theta' \subseteq \Theta\}$ that maps a point in the sample space to a subset of the parameter space. The *coverage probability* of an *interval estimator* $C(\mathbf{X})$ is defined as $P_\theta(\theta \in C(\mathbf{X}))$. The coverage probability is the frequency that the interval estimator captures the true parameter that generates the sample (From the frequentist perspective, the parameter is fixed while the region is random). It is *not* the probability that θ is inside the given region (From the Bayesian perspective, the parameter is random while the region is fixed conditional on $\mathbf{X}$.)

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

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

Find the coverage probability of the random interval

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

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

$$C(\mathbf{x}) = \{\theta \in \Theta : \phi_\theta(\mathbf{x}) = 0\},$$

then its coverage probability

$$P_\theta\left(\theta \in C\left(\mathbf{X}\right)\right) = 1 - P_\theta\left(\phi_\theta\left(\mathbf{X}\right) = 1\right) = 1 - \alpha.$$

5.3 Application in OLS

5.3.1 Wald Test

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

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

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

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

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

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

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

Example 7. In a linear regression

$$\begin{aligned} y &= x_i'\beta + e_i = \sum_{k=1}^{5} \beta_k x_{ik} + e_i. \\ E\left[e_i x_i\right] &= \mathbf{0}_5, \end{aligned} \tag{5.1}$$

where y is wage and

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

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

$$\sqrt{n}\widehat{\beta}_1 = \sqrt{n}\left(\widehat{\beta}_1 - \beta_1\right) = \sqrt{n}R\left(\widehat{\beta} - \beta\right) \xrightarrow{d} N\left(0, R\Omega R'\right) \sim N\left(0, \Omega_{11}\right), \tag{5.2}$$

where Ω_{11} is the $(1,1)$ (scalar) element of Ω. Therefore,

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

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

Example 7 is a test about a single coefficient, and the test statistic is essentially a t-statistic. Example 8 gives a test about a joint hypothesis.

Example 8. We want to simultaneously test $\beta_1 = 1$ and $\beta_3 + \beta_4 = 2$ in (5.1). The null hypothesis can be expressed in the general form $R\beta = r$, where the restriction matrix R is

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

and $r = (1, 2)'$.

Example 7 and 8 are linear restrictions. In order to test a nonlinear regression, we need the so-called *delta method*.

Theorem 9 (delta method)**.** *If* $\sqrt{n}\left(\widehat{\theta}-\theta_0\right) \xrightarrow{d} N\left(0,\Omega_{K\times K}\right)$*, and* $f:\mathbb{R}^K \mapsto \mathbb{R}^q$ *is a continuously differentiable function for some* $q \leq K$*, then*

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

Example 10. In the regression (5.1), the optimal experience level can be found by setting the first order condition with respective to experience to set, $\beta_3 + 2\beta_4 \text{experience}^* = 0$. We test the hypothesis that the optimal experience level is 20 years; in other words,

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

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

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

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

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

Since $\widehat{\beta} \xrightarrow{p} \beta_0$, by the continuous mapping theorem theorem, if $\beta_{0,4} \neq 0$, we

have $\frac{\partial}{\partial\beta}f\left(\widehat{\beta}\right) \xrightarrow{p} \frac{\partial}{\partial\beta}f\left(\beta_0\right)$. Therefore, the (nonlinear) Wald test is

$$W_n = n\left(f\left(\widehat{\beta}\right)-20\right)'\left(\frac{\partial f}{\partial\beta}\left(\widehat{\beta}\right)\widehat{\Omega}\frac{\partial f}{\partial\beta}\left(\widehat{\beta}\right)'\right)^{-1}\left(f\left(\widehat{\beta}\right)-20\right) \xrightarrow{d} \chi^2_1.$$

5.3.2 Lagrangian Multiplier Test*

Restricted least square

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

Turn it into an unrestricted problem

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

The first-order condition

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

Combine these two equations into a linear system,

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

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

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

We conclude that

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

$$\sqrt{n}\tilde{\lambda} \Rightarrow N\left(0, \left(RQ^{-1}R'\right)^{-1} RQ^{-1}\Omega Q^{-1}R' \left(RQ^{-1}R'\right)^{-1}\right).$$

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

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

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

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

5.3.3 Likelihood-Ratio test*

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

tion.

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

From the derivation of LM test, we have

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

Therefore

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

and

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

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

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

has variance

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

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

Chapter 6

Panel Data

This version: February 20, 2017

6.1 Panel Data

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

$$y_{it} = \beta_1 + x_{it}\beta_2 + u_{it},\ i = 1, \ldots, n;\ t = 1, \ldots, T \tag{6.1}$$

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

6.2 Fixed Effect

If $\operatorname{cov}(\alpha_i, x_{it}) = 0$, OLS is consistent for (6.1); otherwise the consistency breaks down. The fixed effect model allows α_i and x_{it} to be arbitrarily correlated. The trick to regain consistency is to eliminate $\alpha_i, i = 1, \ldots, n$. The rest of this section develops the consistency and asymptotic distribution of the *within estimator*, the default fixed-effect (FE) estimator. The within estimator transforms the data by subtracting all the observable variables by the corresponding group means. Averaging the T equations in (6.1) for the same i, we have

$$\overline{y}_i = \beta_1 + \overline{x}_i \beta_2 + \bar{u}_{it} = \beta_1 + \overline{x}_i \beta_2 + \alpha_i + \bar{\epsilon}_{it}. \tag{6.2}$$

where $\overline{y}_i = \frac{1}{T} \sum_{t=1}^{T} y_{it}$. Subtracting (6.2) from (6.1) gives

$$\tilde{y}_{it} = \tilde{x}_{it} \beta_2 + \tilde{\epsilon}_{it} \tag{6.3}$$

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

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

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

We know that OLS in (6.3) would be consistent if $\mathbb{E}[\tilde{\epsilon}_{it} | \tilde{x}_{it}] = 0$. Below we provide a sufficient condition, which is often called *strict exogeneity*.

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

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

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

Lemma (FE consistency)**.** *If FE.1 is satisfied, then $\widehat{\beta}_2^{FE}$ is consistent.*

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

Assumption (FE.2)**.** $\operatorname{var}(\epsilon_i|\mathbf{x}_i) = \sigma_\epsilon^2 I_T$.

Under FE.1 and FE.2, $\widehat{\sigma}_\epsilon^2 = \frac{1}{n(T-1)} \sum_{i=1}^n \sum_{t=1}^T \widehat{\tilde{\epsilon}}_{it}^2$ is a consistent estimator of σ_ϵ^2, where $\widehat{\tilde{\epsilon}} = \tilde{y}_{it} - \tilde{x}_{it}\widehat{\beta}_2^{FE}$. Note that the denominator is $n(T-1)$, not nT.

Theorem (FE asymptotic normality)**.** *If FE.1 and FE.2 are satisfied, then*

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

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

6.3 Random Effect

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

The model is again (6.1), while we assume

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

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

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

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

As mentioned before, FE estimation kills all time-invariant regressors. In contrast, RE allows time-invariant explanatory variables. Let us rewrite (6.1) as

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

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

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

where $\mathbf{S} = I_T \otimes S$. ("$\otimes$" denotes the Kronecker product.) In practice,

σ_α^2 and σ_ϵ^2 in S are unknown, so we seek consistent estimators. Again, we impose a simplifying assumption parallel to FE.2.

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

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

$$\begin{aligned}\widehat{\sigma}_u^2 &= \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\widehat{u}_{it}^2 \\ \widehat{\sigma}_\alpha^2 &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{T\left(T-1\right)}\sum_{t=1}^{T}\sum_{r=1}^{T}\sum_{r\neq t}\widehat{u}_{it}\widehat{u}_{ir}.\end{aligned}$$

Again, we claim the asymptotic normality.

Theorem (RE asymptotic normality)**.** *If RE.1 and RE.2 are satisfied, then*

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

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

Chapter 7

Generalized Method of Moments

This version: February 20, 2017

7.1 Generalized Method of Moments

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

7.1.1 Examples of Endogeneity

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

Example 11 (Dynamic Panel Model)**.** We know that the first-difference (FD) estimator is consistent for (static) panel data model. Nevertheless, the FD estimator encounters difficulty in a dynamic panel model

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

even if we assume

$$\mathbb{E}\left[\epsilon_{it} | \alpha_i, x_{i1}, \ldots, x_{iT}, y_{it-1}, y_{it-2}, \ldots, y_{i0}\right] = 0. \tag{7.1}$$

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

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

Under (7.1), $\mathbb{E}\left[\left(x_{it} - x_{it-1}\right)\left(\epsilon_{it} - \epsilon_{it-1}\right)\right] = 0$, but

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

Example 12 (Keynesian-Type Macro Equations)**.** This is a model borrowed from Hayashi (2000, p.193) but originated from Haavelmo (1943).

An econometrician is interested in learning β_2, the marginal propensity of consumption, in the Keynesian-type equation

$$C_i = \beta_1 + \beta_2 Y_i + u_i \tag{7.2}$$

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

$$Y_i = C_i + I_i,$$

where I_i is investment. We assume $\mathbb{E}[u_i|I_i] = 0$ as investment is determined in advance. OLS (7.2) will be inconsistent because in the reduced-form $Y_i = \frac{1}{1-\beta_2}(\beta_1 + u_i + I_i)$ implies $\mathbb{E}[Y_i u_i] = \mathbb{E}\left[u_i^2\right] / (1-\beta_2) \neq 0$. □

Example 13 (Classical Measurement Error)**.** Endogeneity also emerges when an explanatory variables is not directly observable but is replaced by a measurement with error. Suppose the true linear model is

$$y_i = \beta_1 + \beta_2 x_i^* + u_i, \tag{7.3}$$

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

$$x_i = x_i^* + v_i$$

with $\mathbb{E}[v_i|x_i^*, u_i] = 0$. Such a formulation of the measurement error is called the *classical measurement error*. When we substitute out the unobservable

x_i^* in (7.3), we have

$$y_i = \beta_1 + \beta_2 (x_i - v_i) + u_i = \beta_1 + \beta_2 x_i + e_i \tag{7.4}$$

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

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

OLS (7.4) would not deliver a consistent estimator. □

Example 14 (Demand-Supply System)**.** See Hansen's Chapter 15.

7.1.2 GMM in Linear Model

In this section we discuss GMM in a linear single structural equation. A structural equation is a model of economic interest. For example, (7.2) is a structural equation in which β_2 can be interpreted as the marginal propensity of consumption. Consider the following linear structural model

$$y_i = x_{1i}'\beta_1 + z_{1i}'\beta_2 + \epsilon_i, \tag{7.5}$$

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

variables, or simply instruments. Let $\beta = (\beta_1', \beta_2')'$ be a K-dimensional parameter of interest. From now on, we rewrite (7.5) as

$$y_i = x_i'\beta + \epsilon_i, \tag{7.6}$$

and we have a vector of instruments z_i.

Before estimating any structural econometric model, we must check identification. A model is *identified* if there is a one-to-one mapping between the distribution of the observed variables and the parameters. In other words, in an identified model any two parameter values β and $\tilde{\beta}$, $\beta \neq \tilde{\beta}$, cannot generate exactly the same distribution for the observable data. In the context of (7.6), identification requires that the true value β_0 is the only value on the parameters space that satisfies the moment condition

$$\mathbb{E}\left[z_i\left(y_i - x_i'\beta\right)\right] = 0_L. \tag{7.7}$$

The rank condition is sufficient and necessary for identification.

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

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

Theorem. *The parameter in (7.7) is identified if and only if the rank condition holds.*

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

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

Because $\operatorname{rank}\left(\mathbb{E}\left[z_i x_i'\right]\right) = K$, we would have $\mathbb{E}\left[z_i x_i'\right]\left(\beta_0 - \tilde{\beta}\right) = 0_L$ if and only if $\beta_0 - \tilde{\beta} = 0_K$, which violates $\tilde{\beta} \neq \beta_0$. Therefore β_0 is the unique value that satisfies (7.7).

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

Because identification is a prerequisite for structural estimation, from now on we always assume that the model is identified. When it is just-identified ($L = K$), by (7.7) we can express the parameter as

$$\beta = \left(\mathbb{E}\left[z_i x_i'\right]\right)^{-1}\mathbb{E}\left[z_i y_i\right]. \tag{7.8}$$

It follows by the principle of method of moments that

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

which is exactly the 2SLS when $L = K$. In the rest of this section, we focus on the over-identified case ($L > K$). When $L > K$, (7.7) involves more equations than the number of parameters, directly taking the inverse as in (7.8) is inapplicable.

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

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

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

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

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

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

Rearranging the above equation, we have

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

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

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

In practice, we use the sample moments to replace the corresponding pop-

ulation moments. The GMM estimator mimics its population formula.

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

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

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

or more concisely,

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

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

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

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

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

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

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

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

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

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

$$\sqrt{n}\left(\widehat{\beta} - \beta_0\right) \Rightarrow N\left(0_K, \left(\Sigma'W\Sigma\right)^{-1} \Sigma'W\Omega W\Sigma \left(\Sigma'W\Sigma\right)^{-1}\right). \tag{7.9}$$

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

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

The conclusion follows as $\frac{X'Z}{n} W \frac{Z'X}{n} \xrightarrow{\text{p}} \Sigma'W\Sigma$ and $\frac{X'Z}{n} W \frac{1}{\sqrt{n}} \sum z_i'\epsilon_i \Rightarrow \Sigma'W \times N\left(0, \Omega\right)$. □

It is clear from (7.9) that the GMM estimator's asymptotic variance depends on the choice of W. A natural question follows: can we optimally choose a W to make the asymptotic variance as small as possible? Here we claim the result without a proof.

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

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

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

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

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

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

   This second estimator is asymptotic efficient.

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

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

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

7.1.3 GMM in Nonlinear Model*

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

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

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

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

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

The GMM estimator is defined as

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

In these nonlinear models, a closed-form solution is in general unavailable,

while the asymptotic properties can still be established. We state these asymptotic properties without proofs.

Theorem. *If the model is identified, and*

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

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

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

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

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

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

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

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

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

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