Economics 850

James G. MacKinnon

September, 2024

Introduction

- ECON 850 is the first course of a two-course sequence in econometrics intended for Ph.D. students.
- Classes: Tuesday 10:00–11:20, Thursday 8:30–9:50. Dunning 213 (Hand-Purvis Conference Room).
- It is assumed that all students have taken a serious masters-level econometrics course.
- Familiarity with basic concepts of mathematical statistics would also be very helpful.
- There will be extensive use of matrix algebra, including projection matrices, and the geometry of vector spaces.
- Although considerable time will be devoted to asymptotic theory, it will not be developed in a fully rigorous way. That will be done in ECON 851.

- The first two-thirds of the course will be based on the first six chapters of the never-to-be-finished second edition of *Econometric Theory and Methods* by R. Davidson and J. G. MacKinnon.
- Every student will be provided with PDF copy.
- The remainder of the course will use material from the first edition of *ETM*, from the 1993 book *Estimation and Inference in Econometrics*, and from slides based on "Cluster-robust inference: A guide to empirical practice" (*Journal of Econometrics*, 2023).
- Both books may be legally downloaded as PDF files.
- All students are assumed to be familiar with Stata and/or R. Assignments could probably also be done in a matrix language such as Matlab, Octave, or Ox, but it would be more work.

Course Outline

1. Introductory material based on Chapter 1 of ETM2.
2. The Geometry of Least Squares—Chapter 2 of ETM2.
3. Basic Properties of OLS—Chapter 3 of ETM2.
4. Introduction to Asymptotic Theory—Chapter 3 of ETM2.
5. Hypothesis Testing—Chapter 4 of ETM2.
6. Confidence Intervals—Chapter 5 of ETM2.
7. Bootstrap Methods—parts of Chapter 6 of ETM2.
8. Methods for Clustered Data—the Guide.
9. Generalized Least Squares—Chapter 7 of ETM.
10. Instrumental Variables—Chapter 8 of ETM.
11. Nonlinear Least Squares—Chapter 6 of ETM + supplement

- There will be four assignments, which collectively will account for 20% of the final mark. These assignments will make extensive use of the computer.
- The midterm examination will be worth 20% of the final mark. The date has not yet been determined.
- The final examination will be worth 60% of the final mark.
- The identity of the T.A. is not yet known.
- Tutorial: TBA
- T.A. Office Hours: TBA
- `http://qed.econ.queensu.ca/pub/faculty/mackinnon/econ850/`

Some Properties of PDFs and CDFs

- Random variables may be discrete (binary, counts) or continuous.
- A discrete random variable X takes on values $x_1, x_2, \ldots$, each with probability $p(x_i)$, such that $\sum_i p(x_i) = 1$.
- Number of possible values of i may be finite (just 2 for binary r.v.) or countably infinite (for count r.v.).
- A continuous random variable X can take on real values. The realized value of X is often denoted x.
- The distribution of X is described by a cumulative distribution function, or CDF: $F(x) = \Pr(X \leq x)$.
- $0 \leq F(x) \leq 1$.
- $F(x)$ tends to 0 as $x \to -\infty$.
- $F(x)$ tends to 1 as $x \to +\infty$.
- $F(x)$ must be a weakly increasing function of x.

- The probability that $x = X$ is always zero.
- If $a < b$, then $\Pr(X \leq b) = \Pr(X \leq a) + \Pr(a < X \leq b)$.
- Therefore, $\Pr(a \leq X \leq b) = F(b) - F(a)$.
- If $b = a$, then we get $F(a) - F(a) = 0$.

The probability density function, or PDF, is just the derivative of the CDF: $f(x) \equiv F'(x)$. Evidently,

$$\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = \int_{-\infty}^{\infty} F'(x)\,\mathrm{d}x = F(\infty) - F(-\infty) = 1. \tag{1}$$

More generally,

$$\int_{a}^{b} f(x)\,\mathrm{d}x = \Pr(a \leq X \leq b) = F(b) - F(a). \tag{2}$$

But if we set $b = a$, we just get $F(a) - F(a) = 0$.

- It is evident that $f(x) \geq 0$, because $F(x)$ is non-decreasing.
- $f(x)$ is *not* bounded above by unity, because the value of a PDF at a point x is not a probability.
- The PDF of the standard normal distribution is

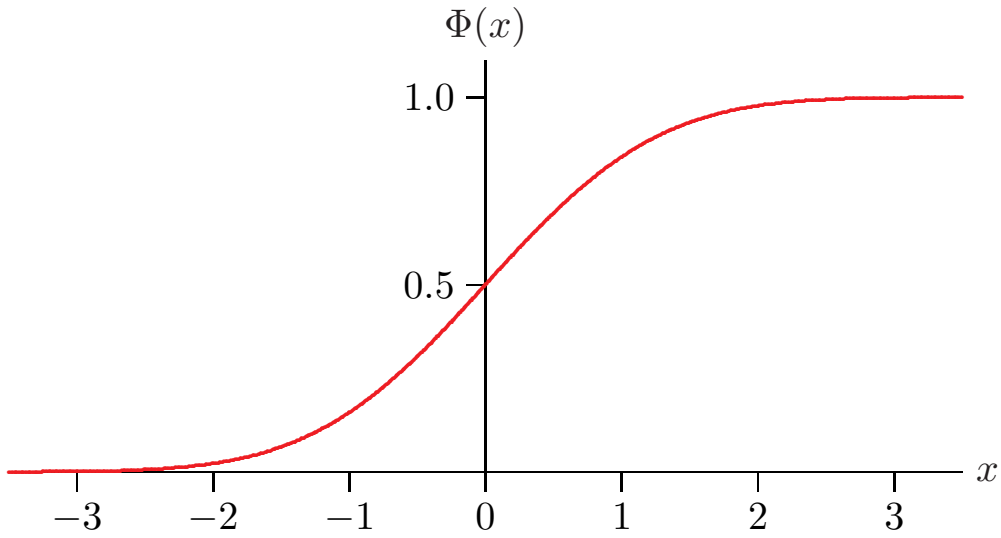
$$\phi(x) = (2\pi)^{-1/2} \exp\left(-\tfrac{1}{2}x^2\right). \tag{3}$$

- The CDF of the standard normal distribution is

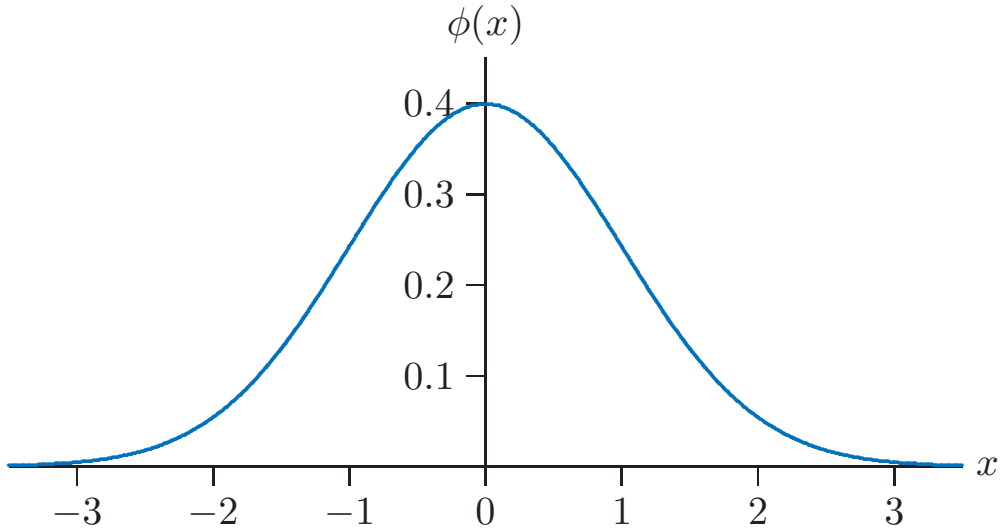
$$\Phi(x) = \int_{-\infty}^{x} \phi(y)\, \mathrm{d}y. \tag{4}$$

  This has no closed-form solution.
- The maximum of the PDF $\phi(x)$ is at $x = 0$, where the slope of the CDF $\Phi(x)$ is steepest.

Standard Normal CDF:



Standard Normal PDF:



- For a continuous random variable, the **population mean** is

$$\mu \equiv \mathrm{E}(x) \equiv \int_{-\infty}^{\infty} x f(x)\, \mathrm{d}x. \tag{5}$$

  Since x can range from $-\infty$ to ∞, this integral may well diverge. Not every continous random variable has a mean!

- The k^{th} uncentered moment of x is

$$m_k(x) \equiv \int_{-\infty}^{\infty} x^k f(x)\, \mathrm{d}x. \tag{6}$$

- The k^{th} central moment of the distribution of x is

$$\mu_k \equiv \mathrm{E}(x-\mu)^k = \int_{-\infty}^{\infty} (x-\mu)^k f(x)\, \mathrm{d}x. \tag{7}$$

  Central moments are invariant to μ.

- The second central moment is the **variance**, $\text{Var}(x) = \sigma^2$.
- The square root of the variance, σ, is called the **standard deviation**.
- Estimates of standard deviations of parameter estimates are called **standard errors**.
- If $\bar{x}$ is the **sample mean** of x_i, $i = 1, \ldots, N$, then the **sample standard deviation** is

$$\text{s.d.}(x) = \left(\frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2 \right)^{1/2}. \tag{8}$$

Under the assumption that the x_i are uncorrelated,

$$\text{s.e.}(\bar{x}) = \frac{1}{\sqrt{N}} \text{s.d.}(x). \tag{9}$$

Joint Distributions

- A continuous, bivariate random variable (x_1, x_2) has the distribution function

$$F(x_1, x_2) = \Pr\big((X_1 \leq x_1) \cap (X_2 \leq x_2)\big). \tag{10}$$

  Thus the joint CDF $F(x_1, x_2)$ is the joint probability that both $X_1 \leq x_1$ and $X_2 \leq x_2$.

- The **joint density function** is

$$f(x_1, x_2) = \frac{\partial^2 F(x_1, x_2)}{\partial x_1 \partial x_2}. \tag{11}$$

- Like all densities, this joint PDF integrates to one:

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x_1, x_2)\, \mathrm{d}x_1 \mathrm{d}x_2 = 1. \tag{12}$$

- The joint CDF is related to the joint PDF by

$$F(x_1, x_2) = \int_{-\infty}^{x_2} \int_{-\infty}^{x_1} f(y_1, y_2)\, \mathrm{d}y_1 \mathrm{d}y_2. \tag{13}$$

- X_1 and X_2 are said to be **independent** if $F(x_1, x_2)$ is the product of the **marginal CDFs** of x_1 and x_2:

$$F(x_1, x_2) = F(x_1, \infty)F(\infty, x_2) = F(x_1)F(x_2). \tag{14}$$

- The marginal density of x_1 is

$$f(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2)\, \mathrm{d}x_2 = \frac{\partial F(x_1, \infty)}{\partial x_1}. \tag{15}$$

  Thus $f(x_1)$ is obtained by integrating x_2 out of the joint density.

- If x_1 and x_2 are independent, so that (14) holds, then

$$f(x_1, x_2) = f(x_1)f(x_2). \tag{16}$$

- Suppose that A and B are any two events. Then $\Pr(A \mid B)$ and is defined implicitly by the equation

$$\Pr(A \cap B) = \Pr(B)\ \Pr(A \mid B). \tag{17}$$

  Evidently, $\Pr(B) \neq 0$, since we cannot condition on B when $\Pr(B) = 0$.

- Equation (17) underlies all of Bayesian statistics.

$$\Pr(A \cap B) = \Pr(A)\Pr(B \mid A) \tag{18}$$

  is just (17) with A and B interchanged.

- Equations (17) and (18) imply that

$$\Pr(A \mid B) = \frac{\Pr(B \mid A)\,\Pr(A)}{\Pr(B)}. \tag{19}$$

- For Bayesian estimation, the sample $\boldsymbol{y}$ plays the role of B, and the parameter vector $\boldsymbol{\theta}$ plays the role of A. Thus we have

$$f(\boldsymbol{\theta} \mid \boldsymbol{y}) = \frac{f(\boldsymbol{y} \mid \boldsymbol{\theta}) f(\boldsymbol{\theta})}{f(\boldsymbol{y})}, \tag{20}$$

  where the $f(\cdot)$ denote densities. This is one version of **Bayes' Rule**.
- In words, the **posterior density** is equal to the **likelihood** times the **prior density**, divided by the unconditional density of $\boldsymbol{y}$. If we ignore the denominator, then

$$\textbf{posterior} \propto \textbf{prior} \times \textbf{likelihood}.$$

- The **conditional density**, or **conditional PDF**, of X_1 for a given value x_2 is

$$f(x_1 \mid x_2) = \frac{f(x_1, x_2)}{f(x_2)}. \tag{21}$$

- If we let y denote x_1 and x denote x_2, then the **conditional expectation** of y given x is

$$\mathrm{E}(y \mid x) = h(x), \tag{22}$$

where $h(x)$ could be any sort of function. It is a (deterministic) function that gives us $\mathrm{E}(y)$ for every possible value of x.

- A very simple example is the **regression function**

$$\mathrm{E}(y) = \beta_1 + \beta_2 x. \tag{23}$$

Notice that there is no "error term" or "disturbance" here.

- The **Law of Iterated Expectations** is very useful. It tells us that

$$\mathrm{E}\big(\mathrm{E}(X_1 \mid X_2)\big) = \mathrm{E}(X_1). \tag{24}$$

  In words, the *unconditional* expectation of X_1 is equal to the expectation of the *conditional* expectation.

- Any deterministic function of a conditioning variable x_2 is its own conditional expectation. Thus

$$\mathrm{E}(X_2 \mid X_2) = X_2 \quad \text{and} \quad \mathrm{E}(X_2^2 \mid X_2) = X_2^2. \tag{25}$$

  Similarly,

$$\mathrm{E}\big(X_1 h(X_2) \mid X_2\big) = h(X_2)\mathrm{E}(X_1 \mid X_2) \tag{26}$$

  for any deterministic function $h(\cdot)$.

- An important special case arises when $\mathrm{E}(X_1 \mid X_2) = 0$. In that case, for any function $h(\cdot)$, $\mathrm{E}(X_1 h(X_2)) = 0$, because

$$\begin{aligned} \mathrm{E}\big(X_1 h(X_2)\big) &= \mathrm{E}\big(\mathrm{E}(X_1 h(X_2) \mid X_2)\big) \\ &= \mathrm{E}\big(h(X_2)\mathrm{E}(X_1 \mid X_2)\big) \\ &= \mathrm{E}\big((h(X_2)0\big) = 0. \end{aligned} \tag{27}$$

  The first two equalities follow from (24), the Law of Iterated Expectations, and (26), respectively.

  Since $\mathrm{E}(X_1 \mid X_2) = 0$, the third equality then follows immediately.

  This result will prove to be useful when we discuss estimation of regression models based on the method of moments.

The Specification of Regression Models

- Because $\mathrm{E}(u_i \mid x_i) = 0$,

$$\mathrm{E}(y_i \mid x_i) = \beta_1 + \beta_2 x_i + \mathrm{E}(u_i \mid x_i) = \beta_1 + \beta_2 x_i. \tag{28}$$

- Suppose that we estimate the model (28) when in fact

$$y_i = \beta_1 + \beta_2 x_i + \beta_3 x_i^2 + v_i. \tag{29}$$

  Then

$$\mathrm{E}(u_i \mid x_i) = \mathrm{E}\big(\beta_3 x_i^2 + v_i \mid x_i\big) = \beta_3 x_i^2, \tag{30}$$

  which must be nonzero unless $x_i = 0$.

- Should the observations in a sample be indexed by $t = 1, \ldots, T$ or $i = 1, \ldots, N$?
- ETM mostly uses $t = 1, \ldots, n$, but my slides will use $i = 1, \ldots, N$ except for time series.

- The **information set** is the set of potential explanatory variables, denoted Ω_i, which is what we condition on. Instead of (28),

$$\mathrm{E}(y_i \mid \Omega_i) = \beta_1 + \beta_2 x_i. \tag{31}$$

- **Exogenous** and **endogenous** variables.
- **Disturbances** rather than **error terms**.
- These are often assumed to be **independent and identically distributed**, or **IID**.
- **Serial correlation** can arise when observations are ordered by time. Then $\mathrm{E}(u_t \mid u_s) \neq 0$, perhaps only when $|t - s|$ is small.
- **Heteroskedasticity** means that $\mathrm{Var}(u_i)$ is not constant. It may depend on $\boldsymbol{X}_i$, or it may depend on lagged values of $\mathrm{Var}(u_i)$.
- **Clustering** implies that $\mathrm{Cov}(u_{gi}\, u_{gj}) \neq 0$. Here the sample is divided into clusters indexed by g.

Equation (31), by itself, is not a **complete specification**. If a model is completely specified, we can simulate it. For the regression model (28):

- Fix the sample size, N.
- Choose β_1 and β_2, the parameters of the **deterministic specification**.
- Obtain the N values x_i, $i = 1, \ldots, N$, of the explanatory variable.
- Evaluate $\beta_1 + \beta_2 x_i$ for $i = 1, \ldots, N$.
- Choose the distribution of the disturbances, if necessary specifying parameters such as mean and variance.
- Use a **random-number generator**, or **RNG**, to generate values of u_i.
- Form the **simulated values** y_i by adding the disturbances to the values of the regression function.
- For a **dynamic model** like $y_t = \beta_1 + \beta_2 y_{t-1} + u_t$, the data need to be generated recursively.

Alternative models for the mean of y_i conditional on x_i:

$$y_i = \beta_1 + \beta_2 x_i + \beta_3 x_i^2 + u_i \tag{32}$$

$$y_i = \gamma_1 + \gamma_2 \log x_i + u_i \tag{33}$$

$$y_i = \delta_1 + \delta_2 \frac{1}{x_i} + u_i. \tag{34}$$

These are all linear models. A nonlinear (but rarely sensible) model is

$$y_i = \mathrm{e}^{\beta_1} x_{i2}^{\beta_2} x_{i3}^{\beta_3} + u_i. \tag{35}$$

A better model is

$$y_i = \mathrm{e}^{\beta_1} x_{i2}^{\beta_2} x_{i3}^{\beta_3} \mathrm{e}^{v_i}. \tag{36}$$

If we take logarithms of both sides, we get

$$\log y_i = \beta_1 + \beta_2 \log x_{i2} + \beta_3 \log x_{i3} + v_i, \tag{37}$$

which is a **loglinear regression model**.

Method-of-Moments Estimation

The **method of moments**, or **MM**, replaces population quantities by sample analogs.

Suppose there is just one parameter (β_1, the population mean) to estimate. The sample mean of the disturbances is

$$\frac{1}{N}\sum_{i=1}^{N} u_i = \frac{1}{N}\sum_{i=1}^{N}(y_i - \beta_1). \tag{38}$$

Equating this to 0 yields

$$\frac{1}{N}\sum_{i=1}^{N} y_i - \beta_1 = 0. \tag{39}$$

The MM estimate $\hat{\beta}_1$ is just the mean of the observed values:

$$\hat{\beta}_1 = \frac{1}{N}\sum_{i=1}^{N} y_i. \tag{40}$$

For a simple linear regression model with two parameters, (39) becomes

$$\frac{1}{N}\sum_{i=1}^{N}(y_i - \beta_1 - \beta_2 x_i) = 0. \tag{41}$$

We need one more equation to solve for $\hat{\beta}_1$ and $\hat{\beta}_2$.

We use the fact that $\mathrm{E}(u_i \mid x_i) = 0$. By the law of iterated expectations,

$$\mathrm{E}(x_i u_i) = \mathrm{E}\big(\mathrm{E}(x_i u_i \mid x_i)\big) = \mathrm{E}\big(x_i \mathrm{E}(u_i \mid x_i)\big) = 0. \tag{42}$$

Thus we can supplement (41) by the following equation:

$$\frac{1}{N}\sum_{i=1}^{N} x_i(y_i - \beta_1 - \beta_2 x_i) = 0. \tag{43}$$

Equations (41) and (43) can be written as

$$\beta_1 + \left(\frac{1}{N}\sum_{i=1}^{N} x_i\right)\beta_2 = \frac{1}{N}\sum_{i=1}^{N} y_i \tag{44}$$

$$\left(\frac{1}{N}\sum_{i=1}^{N} x_i\right)\beta_1 + \left(\frac{1}{N}\sum_{i=1}^{N} x_i^2\right)\beta_2 = \frac{1}{N}\sum_{i=1}^{N} x_i y_i. \tag{45}$$

After multiplication by N, these equations become

$$\begin{bmatrix} N & \sum_{i=1}^{N} x_i \\ \sum_{i=1}^{N} x_i & \sum_{i=1}^{N} x_i^2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{N} y_i \\ \sum_{i=1}^{N} x_i y_i \end{bmatrix}. \tag{46}$$

But (46) is just a special case of

$$\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}^\top \boldsymbol{y}, \tag{47}$$

where $\boldsymbol{X}^\top \boldsymbol{X}$ is the matrix of sums of squares and cross-products of every regressor with every other regressor.

Thus we obtain the famous formula for the **ordinary least squares**, or **OLS**, estimator:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}. \tag{48}$$

In general, of course, there are k moment conditions, one for each regressor:

$$\boldsymbol{X}^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{0}. \tag{49}$$

Here we treat the constant term as a column of 1s within $\boldsymbol{X}$.

We could also obtain $\hat{\boldsymbol{\beta}}$ by minimizing the sum of squared residuals

$$\begin{aligned} \text{SSR}(\boldsymbol{\beta}) &= \sum_{i=1}^{N} (y_i - \boldsymbol{X}_i \boldsymbol{\beta})^2 \\ &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\ &= \boldsymbol{y}^\top \boldsymbol{y} - 2\boldsymbol{y}^\top \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta}. \end{aligned} \tag{50}$$

The first-order conditions are

$$-2\boldsymbol{X}^\top(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})=\boldsymbol{0}. \tag{51}$$

These can be rewritten as

$$\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta}=\boldsymbol{X}^\top\boldsymbol{y}. \tag{52}$$

The solution is evidently

$$\hat{\boldsymbol{\beta}}=(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}. \tag{53}$$

Here, as before, we have explicitly assumed that $\boldsymbol{X}$ has full rank k. Otherwise, $\boldsymbol{X}^\top\boldsymbol{X}$ would not be invertible.