Advanced Empirical Economics I

Mario Larch

VWL VI: Chair of Empirical Economics
University of Bayreuth
mario.larch@uni-bayreuth.de

WS 2023/24
January 19, 2024

1. Introduction

2. Linear models
   - Ordinary Least Squares (OLS)
     - Identification
     - Consistency
     - Excursus: Asymptotic theory
     - Limit distribution
     - Asymptotic distribution
     - Heteroskedasticity-robust standard errors for OLS
     - Cluster-robust inference
   - Model misspecification
     - Inconsistency of OLS
     - Functional form misspecification
     - Endogeneity
     - Omitted variables
     - Pseudo-true values
     - Parameter heterogeneity

3. Statistical learning
   - Prediction vs. inference

- How do we learn f?
- Bias-Variance-Trade-Off
- Considerations in high dimensions
- Ridge regression

4 Maximum likelihood

- Poisson regression
- Classification
- An m-estimator
- Likelihood function
- Maximum likelihood estimator
  - Regularity conditions
  - Information matrix equality
  - Distribution of the ML estimator
- Quasi-maximum likelihood
- Marginal effects
- Poisson example
- Further examples

- The bivariate sample selection model
- The Roy model

5 Generalized Method of Moments (GMM)

- Examples
- Method of Moments estimator (MM)
- Generalized Method of Moments estimator (GMM)
    - Distribution of GMM estimator
    - Optimal weighting matrix
    - Optimal moment conditions
- Linear IV
    - Inconsistency of OLS
    - Instruments
    - Instrumental variables estimator
    - Wald estimator
    - IV estimation for multiple regression
    - Two-stage least squares
    - An example
    - Weak instruments
    - Inconsistency of IV estimators

Low precision and finite-sample bias

Non-linear IV

Further examples

6 The end

7 Appendix A: Generalized Least Squares (GLS) and Weighted least squares (WLS)

Generalized Least Squares (GLS) and Feasible GLS

Weighted Least Squares (WLS)

Introduction

Introduction

- Question for Students: Background and expectations.
- Focus: Methods and microeconometrics.
- But also: Applications.
- Organization:
  - Start: Monday, 11.12.2023.
  - End: Tuesday, 06.02.2024.
  - No lectures between 20.12.2023-07.01.2024.
  - Room: H21 (RWIIEG0.14).

Introduction

- This course provides a comprehensive treatment of mainly microeconometric methods, allowing to analyse individual-level data on the economic behaviour of individuals or firms using regression methods applied to cross-section and panel data.
- I will give a brief introduction to machine learning/statistical learning and relate it to what we have learned in the course.
- The linear regression model will be discussed, but basic knowledge is assumed. The course will use matrix algebra. A short refresher will be given in the tutorials.
- However, orientation toward the practitioner.

Introduction

- Main Reference: Cameron, A. Colin and Pravin K. Trivedi (2005), Microeconometrics - Methods and Applications, Cambridge University Press (http://cameron.econ.ucdavis. edu/mmabook/mma.html).
- Companion: Cameron, A. Colin and Pravin K. Trivedi (2022), Microeconometrics using STATA, second edition, Volume I + II, STATACorp LP (https://cameron.econ. ucdavis.edu/mus2/).
- Hansen, B. (2022), Econometrics, Princeton University Press (https://www.ssc.wisc.edu/ bhansen/econometrics/).
- Many books on R, for example Kleiber, C. and Achim Zeileis (2008), Applied Econometrics with R (Use R!), Springer.

Introduction

Tutorials (two):

- Group 1: Monday (16:15 (c.t.)-17:45) and Tuesday (18:00 (s.t.)-19:30), (start 11.12.).
- Group 2: Monday (18:00 (s.t.)-19:30) and Tuesday (16:15 (c.t.)-17:45), (start 11.12.).
- Room: PC-Pool (Monday: B9,01/PC-Pool (B9EG01), Tuesday: S56/PC-Pool (RWIEG1.0.00.117)).
- Both held by: Hanna Adam.
- Software: R (https://www.r-project.org/).

Introduction

Main empirical courses at our chair:

- Bachelor level:
  - Empirical Economics I: Introduction, data problems, OLS, Gauss-Markov-Theorem, heteroskedasticity, correlation versus causation.
  - Empirical Economics II: Stochastic processes, panel data estimators (SUR, diff-in-diff, fixed effects, random effects), time series econometrics (autocorrelation, ARMA, (P)ACF, forecasting).
- Master level:
  - Advanced Empirical Economics I: Estimation methods (linear and non-linear least squares, IV, MLE, GMM), applications.
  - Advanced Empirical Economics II: “Topic”-courses (e.g., time series econometrics, program evaluation methods, spatial econometrics, Bayesian econometrics, empirical international trade, ...).

Introduction

Are you familiar with the following concepts?

- Consistency.
- Bias.
- Limit distribution.
- Asymptotic distribution.
- Omitted variable bias.
- Information matrix.
- Quasi-Maximum likelihood.
- Central limit theorem.
- Law of large numbers.

Introduction

Occurring themes and problems:

- Data are often discrete or censored, in which case non-linear methods such as logit, probit, and Tobit models are used.
- Distributional assumptions for such data become critically important.
- Economic studies often aim to determine causation rather than merely measure correlation.
- Microeconomic data are typically collected using cross-section and panel surveys, censuses, or social experiments.

Introduction

Occurring themes and problems:

- It is not unusual that two or more complications occur simultaneously.
- Large data-sets (many observations, many explanatory variables).
- Microeconomic/Behavioural foundation, allowing to employ a structural approach.

Linear models

Ordinary Least Squares (OLS)

Linear models

- In modern microeconometrics the term regression refers to a bewildering range of procedures for studying the relationship between an outcome variable y and a set of regressors $\mathbf{x}$.
- The simplest example of regression is the OLS estimator in the linear regression model.
- After first defining the model and estimator, a quite detailed presentation of the asymptotic distribution of the OLS estimator is given.
- The exposition presumes previous exposure to a more introductory treatment.
- The model assumptions made here permit stochastic regressors and heteroskedastic errors and accommodate data that are obtained by exogenous stratified sampling.

Notation and conventions

Vectors are defined as column vectors and represented using lower-case bold. For example, for linear regression the regressor vector $\mathbf{x}$ is a $K \times 1$ column vector with jth entry x_j and the parameter vector $\boldsymbol{\beta}$ is a $K \times 1$ column vector with jth entry β_j, so

$$\underset{(K \times 1)}{\mathbf{x}} = \begin{bmatrix} x_1 \\ \vdots \\ x_K \end{bmatrix} \quad \text{and} \quad \underset{(K \times 1)}{\boldsymbol{\beta}} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_K \end{bmatrix}.$$

Notation and conventions

Then the linear regression model $y = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_K x_K + u$ is expressed as $y = \mathbf{x}'\boldsymbol{\beta} + u$. At times a subscript i is added to denote the typical ith observation. The linear regression equation for the ith observation is then

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + u_i.$$

The sample is one of N observations, $\{(y_i, \mathbf{x}_i), i = 1, ..., N\}$. Observations are usually assumed to be independent over i in the course.

Notation and conventions

Matrices are represented using upper-case bold. In matrix notation the sample is $(\mathbf{y}, \mathbf{X})$, where $\mathbf{y}$ is an $N \times 1$ vector with ith entry y_i and $\mathbf{X}$ is a matrix with ith row $\mathbf{x}_i'$, so

$$\underset{(N \times 1)}{\mathbf{y}} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} \quad \text{and} \quad \underset{(N \times K)}{\mathbf{X}} = \begin{bmatrix} \mathbf{x}_1' \\ \vdots \\ \mathbf{x}_N' \end{bmatrix}.$$

The linear regression model upon stacking all N observations is then

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u},$$

where $\mathbf{u}$ is an $N \times 1$ column vector with ith entry u_i.

Linear regression model

- In a standard cross-section regression model with N observations on a scalar dependent variable and several regressors, the data are specified as $(\mathbf{y}, \mathbf{X})$, where $\mathbf{y}$ denotes observations on the dependent variable and $\mathbf{X}$ denotes a matrix of explanatory variables.
- The general regression model with additive errors is written in vector notation as

$$\mathbf{y} = E\left[\mathbf{y}|\mathbf{X}\right] + \mathbf{u}, \tag{1}$$

where $E\left[\mathbf{y}|\mathbf{X}\right]$ denotes the conditional expectation of the random variable $\mathbf{y}$ given $\mathbf{X}$, and $\mathbf{u}$ denotes a vector of unobserved random errors or disturbances.

Linear regression model

- The right-hand side of this equation decomposes $\mathbf{y}$ into two components, one that is deterministic given the regressors and one that is attributed to random variation or noise.
- We think of $E[\mathbf{y}|\mathbf{X}]$ as a conditional prediction function that yields the average value, or more formally the expected value, of $\mathbf{y}$ given $\mathbf{X}$.
- A linear regression model is obtained when $E[\mathbf{y}|\mathbf{X}]$ is specified to be a linear function of $\mathbf{X}$.

Linear regression model

- y is referred to as the dependent variable or endogenous variable whose variation we wish to study in terms of variation in $\mathbf{x}$ and u.
- u is referred to as the error term or disturbance term in the population.
- $\mathbf{x}$ is referred to as regressors or predictors or covariates.
- Note, the sample equivalent of equation

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u}, \tag{2}$$

is

$$\mathbf{y} = \mathbf{X}\hat{\beta} + \hat{\mathbf{u}}, \tag{3}$$

where $\hat{\mathbf{u}}$ is the residual vector and $\hat{\beta}$ is the vector of the OLS estimates.

OLS estimator

- The OLS estimator is defined to be the estimator that minimizes the sum of squared errors

$$\sum_{i=1}^{N} \hat{u}_i^2 = \hat{\mathbf{u}}'\hat{\mathbf{u}} = \left(\mathbf{y} - \mathbf{X}\hat{\beta}\right)' \left(\mathbf{y} - \mathbf{X}\hat{\beta}\right). \tag{4}$$

In other words:

$$\min_{\hat{\beta}} S(\hat{\beta}) = \left(\mathbf{y} - \mathbf{X}\hat{\beta}\right)' \left(\mathbf{y} - \mathbf{X}\hat{\beta}\right). \tag{5}$$

- Expanding $S(\hat{\beta})$ gives:

$$\begin{aligned} \min_{\hat{\beta}} S(\hat{\beta}) &= \mathbf{y}'\mathbf{y} - \hat{\beta}'\mathbf{X}'\mathbf{y} - \mathbf{y}'\mathbf{X}\hat{\beta} + \hat{\beta}'\mathbf{X}'\mathbf{X}\hat{\beta} && (6) \\ &= \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}\hat{\beta} + \hat{\beta}'\mathbf{X}'\mathbf{X}\hat{\beta}. && (7) \end{aligned}$$

OLS estimator

- The necessary condition for a minimum is given by the first derivative with respect to $\hat{\beta}$ set equal to $\mathbf{0}$:

$$\frac{\partial S(\hat{\beta})}{\partial \hat{\beta}} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\hat{\beta} = \mathbf{0}. \tag{8}$$

- Solving for $\hat{\beta}$ yields the OLS estimator,

$$\hat{\beta}_{\text{OLS}} = (\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'\mathbf{y}. \tag{9}$$

OLS estimator

- If $\mathbf{X}'\mathbf{X}$ is of less than full rank, the inverse can be replaced by a generalized inverse.
- Then OLS estimation still yields the optimal linear predictor of y given $\mathbf{x}$ if squared error loss is used.
- But many different linear combinations of $\mathbf{x}$ will yield this optimal predictor.

Identification

Identification

- The OLS estimator can always be computed, provided that $\mathbf{X}'\mathbf{X}$ is non-singular.
- The more interesting issue is what $\hat{\beta}_{\text{OLS}}$ tells us about the data.
- We focus on the ability of the OLS estimator to permit identification of the conditional mean $E[\mathbf{y}|\mathbf{X}]$.

Identification

For the linear model the parameter β is identified if

1. $E[\mathbf{y}|\mathbf{X}] = \mathbf{X}\beta$.
2. $\mathbf{X}\beta^{(1)} = \mathbf{X}\beta^{(2)}$ if and only if $\beta^{(1)} = \beta^{(2)}$ (implies that $\mathbf{X}'\mathbf{X}$ is non-singular), i.e., that $\hat{\beta}_{\text{OLS}}$ is the unique solution of $\min_{\hat{\beta}} S(\hat{\beta})$.

Consistency

Consistency

- The properties of an estimator depend on the process that actually generated the data, the data generating process (dgp).
- We assume the dgp is $\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$.
- Then:

$$\begin{aligned}
\hat{\beta}_{\text{OLS}} &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} \\
&= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\left(\mathbf{X}\beta + \mathbf{u}\right) \\
&= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{X}\beta + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{u} \\
&= \beta + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{u}.
\end{aligned}$$

Excursus: Asymptotic theory

Excursus: Asymptotic theory

- Good, more accessible treatment: van der Vaart, A. W. (1998), Asymptotic Statistics, Cambridge University Press.
- Thorough discussion: White, H. (2000), Asymptotic Theory for Econometricians, Academic Press.
- Thorough discussion with focus on dynamic models: Prucha, I., B. Pötscher (1997), Dynamic Nonlinear Econometric Models: Asymptotic Theory, Springer, Berlin.

Excursus: Asymptotic theory

- In this excursus we consider the behaviour of a sequence of random variables b_N as $N \to \infty$.
- For estimation theory it is sufficient to focus on two aspects:
  1. Convergence in probability of b_N to a limit b, a constant or random variable that is very close to b_N in a probabilistic sense defined in the following.
  2. If the limit b is a random variable, we consider the limit distribution.
- Estimators are usually functions of averages or sums. Then it is easiest to derive limiting results by invoking results on the behaviour of averages, notably laws of large numbers and central limit theorems.

Excursus: Asymptotic theory

Convergence in probability

- Because of the intrinsic randomness of a sample we can never be certain that a sequence b_N, such as an estimator $\hat{\theta}$ (often denoted $\hat{\theta}_N$ to make clear that it is a sequence), will be within a given small distance of its limit, even if the sample is infinitely large.
- However, we can be almost certain.
- Different ways of expressing this near certainty correspond to different types of convergence of a sequence of random variables to a limit.
- The one most used in econometrics is convergence in probability.
- Others are: Mean-square convergence, almost sure convergence.

Excursus: Asymptotic theory

Convergence in probability

- Recall that a sequence of non-stochastic real numbers $\{a_N\}$ converges to a if, for any $\epsilon > 0$, there exists $N^* = N^*(\epsilon)$ such that, for all $N > N^*$:

$$|a_N - a| < \epsilon. \tag{10}$$

- Example: If $a_N = 2 + 3/N$, then the limit is $a = 2$ since $|a_N - a| = |2 + 3/N - 2| = |3/N| < \epsilon$ for all $N > N^* = 3/\epsilon$.

Excursus: Asymptotic theory

Convergence in probability

- When more generally we have a sequence of random variables we cannot be certain of being within ϵ of the limit, even for large N, because of intrinsic randomness.
- Instead, we require that the probability of being within ϵ is arbitrarily close to one.
- Thus we require:

$$\lim_{N \to \infty} Pr\left[|b_N - b| < \epsilon\right] = 1, \tag{11}$$

  for any $\epsilon > 0$.

Excursus: Asymptotic theory

Convergence in probability

- A formal definition is the following:

Definition: Convergence in probability

A sequence of random variables $\{b_N\}$ converges in probability to b if, for any $\epsilon > 0$ and $\delta > 0$, there exists $N^* = N^*(\epsilon, \delta)$ such that, for all $N > N^*$, $Pr[|b_N - b| < \epsilon] > 1 - \delta$.

- We write $\operatorname{plim} b_N = b$, where plim is shorthand for probability limit, or $b_N \xrightarrow{p} b$.

Excursus: Asymptotic theory

Consistency

- When the sequence $\{b_N\}$ is a sequence of parameter estimates $\hat{\theta}$, we have a large sample analogue of unbiasedness, consistency.
- A formal definition is the following:

Definition: Consistency

An estimator $\hat{\theta}$ is consistent for θ_0 if $\operatorname{plim} \hat{\theta} = \theta_0$.

Excursus: Asymptotic theory

Consistency

- Note that unbiasedness need not imply consistency.
- Unbiasedness states only that the expected value of $\hat{\theta}$ is θ_0, and it permits variability around θ_0 that need not disappear as the sample size goes to infinity.
- Also, a consistent estimator need not be unbiased.
- For example, adding $1/N$ to an unbiased and consistent estimator produces a new estimator that is biased but still consistent.
- Although the sequence of vector random variables $\{b_N\}$ may converge to a random variable b, in many econometric applications $\{b_N\}$ converges to a constant.
- For example, we hope that an estimator of a parameter will converge in probability to the parameter itself.

Excursus: Asymptotic theory

Consistency

Slutsky's Theorem

Let b_N be a finite-dimensional vector of random variables, and $g(\cdot)$ be a real-valued function continuous at a constant vector point $\mathbf{b}$. Then

$$\mathbf{b}_N \xrightarrow{p} \mathbf{b} \Rightarrow g(\mathbf{b}_N) \xrightarrow{p} g(\mathbf{b}).$$

- Slutsky's Theorem is one of the major reasons for the prevalence of asymptotic results versus finite-sample results in econometrics.
- It states a very convenient property that does not hold for expectations.
- For example, $\text{plim}(\mathbf{b}_N) = \text{plim}(b_{1N}, b_{2N}) = (b_1, b_2)$ implies $\text{plim}(b_{1N}b_{2N}) = b_1 b_2$, whereas $E[b_{1N}b_{2N}]$ generally differs from $E[b_{1N}]E[b_{2N}]$.

Excursus: Asymptotic theory

Laws of large numbers

- Laws of large numbers are theorems for convergence in probability in the special case where the sequence $\{b_N\}$ is a sample average, that is, $b_N = \bar{X}_N$, where

$$\bar{X}_N = \frac{1}{N}\sum_{i=1}^{N} X_i. \tag{12}$$

- Note that X_i here is general notation for a random variable, and in the regression context it does not necessarily denote the regressor variables.

Excursus: Asymptotic theory

Laws of large numbers

- A law of large numbers provides a much easier way to establish the probability limit of a sequence $\{b_N\}$ than the alternatives of brute-force use of the (ϵ, δ) definition.

Definition: Law of large numbers

A (weak) law of large numbers (LLN) specifies conditions on the individual terms X_i in $\bar{X}_N$ under which $(\bar{X}_N - E[\bar{X}_N]) \xrightarrow{p} 0$.

Excursus: Asymptotic theory

Laws of large numbers

- It can be helpful to think of a LLN as establishing that $\bar{X}_N$ goes to its expected value, even though strictly speaking it implies the weaker condition that $\bar{X}_N$ goes to the limit of its expected value, since the above condition implies that:

$$\text{plim}\, \bar{X}_N = \lim E[\bar{X}_N]. \tag{13}$$

- If the X_i have common mean μ, then this simplifies to $\text{plim}\, \bar{X}_N = \mu$.

Consistency

- To prove consistency we rewrite the expression for $\hat{\beta}_{\text{OLS}}$ as

$$\hat{\beta}_{\text{OLS}} \;=\; \beta + \left(N^{-1}\mathbf{X}'\mathbf{X}\right)^{-1} N^{-1}\mathbf{X}'\mathbf{u}.$$

- The reason for renormalization in the right-hand side is that $N^{-1}\mathbf{X}'\mathbf{X} = N^{-1}\sum_i \mathbf{x}_i\mathbf{x}_i'$ is an average that converges in probability to a finite non-zero matrix if $\mathbf{x}_i$ satisfies assumptions that permit a law of large numbers to be applied to $\mathbf{x}_i\mathbf{x}_i'$.

Consistency

- Then we may write

$$\text{plim}\,\hat{\beta}_{\text{OLS}} \;=\; \beta + \left(\text{plim}\,N^{-1}\mathbf{X}'\mathbf{X}\right)^{-1}\left(\text{plim}\,N^{-1}\mathbf{X}'\mathbf{u}\right),$$

  using Slutsky's Theorem (Theorem A.3).
- The OLS estimator is consistent for β (i.e., $\text{plim}\,\hat{\beta}_{\text{OLS}} = \beta$) if

$$\text{plim}\,N^{-1}\mathbf{X}'\mathbf{u} = \mathbf{0}.$$

- If a law of large numbers can be applied to the average $N^{-1}\mathbf{X}'\mathbf{u} = N^{-1}\sum_i \mathbf{x}_i u_i$ then a necessary condition for the previous expression to hold is that $E[\mathbf{x}_i u_i] = 0$.

Excursus: Asymptotic theory

Convergence in distribution

- Given consistency, the estimator $\hat{\theta}$ has a degenerate distribution that collapses on θ_0 as $N \to \infty$.
- We need to magnify or rescale $\hat{\theta}$ to obtain a random variable that has non-degenerate distribution as $N \to \infty$.
- Often the appropriate scale factor is $\sqrt{N}$, in which case we consider the behaviour of the sequence of random variables $b_N = \sqrt{N}\left(\hat{\theta} - \theta_0\right)$.

Excursus: Asymptotic theory

Convergence in distribution

- In general, the Nth random variable in the sequence b_N has an extremely complicated cumulative distribution function (cdf) F_N.
- Like any other function F_N, this may have a limit function where convergence is in the usual mathematical sense.

Definition: Convergence in distribution

A sequence of random variables $\{b_N\}$ is said to converge in distribution to a random variable b if $\lim_{N\to\infty} F_N = F$, at every continuity point of F, where F_N is the distribution of b_N, F is the distribution of b, and convergence is in the usual mathematical sense.

Excursus: Asymptotic theory

Convergence in distribution

- We write $b_N \xrightarrow{d} b$, and we call F the limit distribution of $\{b_N\}$.
- Convergence in probability implies convergence in distribution; that is, $b_N \xrightarrow{p} b$ implies $b_N \xrightarrow{d} b$.
- In general, the converse is not true.
- For example, let $b_N = X_N$, the Nth realization of $X \sim \mathcal{N}[\mu, \sigma^2]$.
- Then $b_N \xrightarrow{d} \mathcal{N}[\mu, \sigma^2]$, but $(b_N - b)$ has variance that does not disappear as $N \to \infty$, so b_N does not converge in probability to b.
- In the special case where b is a constant, however, $b_N \xrightarrow{d} b$ implies $b_N \xrightarrow{p} b$.
- In this case the limit distribution is degenerate, with all its mass at b.

Excursus: Asymptotic theory

Central limit theorems

- Central limit theorems are theorems on convergence in distribution when the sequence $\{b_N\}$ is a sample average.
- A central limit theorem provides a simpler way to obtain the limit distribution of a sequence $\{b_N\}$ than the alternatives such as brute-force use of convergence in distribution.
- From a law of large numbers, the sample average has a degenerate distribution as it converges to a constant, $\lim E[\bar{X}_N]$.
- So we scale $(\bar{X}_N - E[\bar{X}_N])$ by its standard deviation to construct a random variable with unit variance that may converge to a non-degenerate distribution.

Excursus: Asymptotic theory

Definition: Central limit theorem

Let

$$Z_N = \frac{\bar{X}_N - E[\bar{X}_N]}{\sqrt{V[\bar{X}_N]}}, \tag{14}$$

where $\bar{X}_N$ is a sample average. A central limit theorem (CLT) specifies the conditions on the individual terms X_i in $\bar{X}_N$ under which

$$Z_N \xrightarrow{d} \mathcal{N}(0, 1), \tag{15}$$

that is, under which Z_N converges in distribution to a standard normal random variable.

Excursus: Asymptotic theory

Product Limit Normal Rule

If a vector $\mathbf{a}_N \xrightarrow{d} \mathcal{N}[\boldsymbol{\mu}, \mathbf{A}]$ and a matrix $\mathbf{H}_N \xrightarrow{p} \mathbf{H}$, where $\mathbf{H}$ is positive definite, then

$$\mathbf{H}_N\mathbf{a}_N \xrightarrow{d} \mathcal{N}[\mathbf{H}\boldsymbol{\mu}, \mathbf{HAH}']. \tag{16}$$

Limit distribution

Limit distribution

- Given consistency, the limit distribution of $\hat{\beta}_{\text{OLS}}$ is degenerate with all the mass at β.
- To obtain the limit distribution we multiply $\hat{\beta}_{\text{OLS}}$ by $\sqrt{N}$, as this rescaling leads to a random variable that under standard cross-section assumptions (see slides 56 and 57) has non-zero yet finite variance asymptotically.
- Then we may write:

$$\sqrt{N}\left(\hat{\beta}_{\text{OLS}} - \beta\right) = \left(N^{-1}\mathbf{X}'\mathbf{X}\right)^{-1} N^{-1/2}\mathbf{X}'\mathbf{u}. \tag{17}$$

Limit distribution

- The proof of consistency assumed that $\operatorname{plim} N^{-1}\mathbf{X}'\mathbf{X}$ exists and is finite and non-zero.
- We assume that a central limit theorem can be applied to $N^{-1/2}\mathbf{X}'\mathbf{u}$ to yield a multivariate normal limit distribution with finite, non-singular covariance matrix.
- Applying the product rule for limit normal distributions (Theorem A.17), i.e., $\mathbf{H}_N = \left(N^{-1}\mathbf{X}'\mathbf{X}\right)^{-1}$ and $\mathbf{a}_N = N^{-1/2}\mathbf{X}'\mathbf{u}$ implies that the product in the right-hand side of (17) has a limit normal distribution.

Limit distribution

This leads to the following proposition, which permits regressors to be stochastic and does not restrict model errors to be homoskedastic.

Distribution of OLS estimator

Make the following assumptions:

1. The dgp is $\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$.
2. Data are independent over i with $E[\mathbf{u}|\mathbf{X}] = \mathbf{0}$ and $E[\mathbf{uu}'|\mathbf{X}] = \mathbf{\Omega} = \text{Diag}[\sigma_i^2]$.
3. The matrix $\mathbf{X}$ has full rank so that $\mathbf{X}\beta^{(1)} = \mathbf{X}\beta^{(2)}$ iff $\beta^{(1)} = \beta^{(2)}$.

Distribution of OLS estimator

Make the following assumptions:

4. The $K \times K$ matrix

$$\mathbf{M_{XX}} = \text{plim}\, N^{-1}\mathbf{X}'\mathbf{X} = \text{plim}\, \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i\mathbf{x}_i' = \lim \frac{1}{N}\sum_{i=1}^{N} E[\mathbf{x}_i\mathbf{x}_i'],$$

   exists and is finite non-singular.

5. The $K \times 1$ vector $N^{-1/2}\mathbf{X}'\mathbf{u} = N^{-1/2}\sum_{i=1}^{N} \mathbf{x}_i u_i \xrightarrow{d} \mathcal{N}[\mathbf{0}, \mathbf{M_{x\Omega x}}]$, where

$$\begin{aligned}\mathbf{M_{x\Omega x}} &= \text{plim}\, N^{-1}\mathbf{X}'\mathbf{u}\mathbf{u}'\mathbf{X} = \text{plim}\, \frac{1}{N}\sum_{i=1}^{N} u_i^2\mathbf{x}_i\mathbf{x}_i' \\ &= \lim \frac{1}{N}\sum_{i=1}^{N} E[u_i^2\mathbf{x}_i\mathbf{x}_i'].\end{aligned}$$

Limit distribution

Hence, the OLS estimator $\hat{\beta}_{\text{OLS}}$ is consistent for β and

$$\sqrt{N}(\hat{\beta}_{\text{OLS}} - \beta) \xrightarrow{d} \mathcal{N}[\mathbf{0}, \mathbf{M}_{\mathbf{xx}}^{-1}\mathbf{M}_{\mathbf{x}\Omega\mathbf{x}}\mathbf{M}_{\mathbf{xx}}^{-1}]. \tag{18}$$

Asymptotic distribution

Asymptotic distribution

- So far we have stated the limit distribution of $\sqrt{N}(\hat{\beta}_{\text{OLS}} - \beta)$, a rescaling of $\hat{\beta}_{\text{OLS}}$.
- Many practitioners prefer to see asymptotic results written directly in terms of the distribution of $\hat{\beta}_{\text{OLS}}$.
- This distribution is called an asymptotic distribution.
- The asymptotic distribution is interpreted as being applicable in large samples, meaning samples large enough for the limit distribution to be a good approximation but not so large that $\hat{\beta}_{\text{OLS}} \xrightarrow{p} \beta$ as then its asymptotic distribution would be degenerate.

Asymptotic distribution

- The asymptotic distribution is obtained from (18) by multiplication with $N^{-1/2}$ and addition of β.
- This yields the asymptotic distribution

$$\hat{\beta}_{\text{OLS}} \overset{a}{\sim} \mathcal{N}[\beta, N^{-1}\mathbf{M}_{\mathbf{xx}}^{-1}\mathbf{M}_{\mathbf{x\Omega x}}\mathbf{M}_{\mathbf{xx}}^{-1}], \tag{19}$$

  where the symbol $\overset{a}{\sim}$ means is “asymptotically distributed as.”
- The variance matrix in (19) is called the asymptotic variance matrix of $\hat{\beta}_{\text{OLS}}$ and is denoted $V[\hat{\beta}_{\text{OLS}}]$.

Asymptotic distribution

- Even simpler notation drops the limits and expectations in the definitions of $\mathbf{M_{xx}}$ and $\mathbf{M_{x\Omega x}}$ and the asymptotic distribution is denoted

$$\hat{\beta}_{\mathrm{OLS}} \overset{a}{\sim} \mathcal{N}[\beta, (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{\Omega}\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}], \tag{20}$$

and $V[\hat{\beta}_{\mathrm{OLS}}]$ is defined to be the variance matrix in (20).

Asymptotic distribution

- For implementation, the matrices $\mathbf{M_{xx}}$ and $\mathbf{M_{x\Omega x}}$ are replaced by consistent estimates $\widehat{\mathbf{M}}_{\mathbf{xx}}$ and $\widehat{\mathbf{M}}_{\mathbf{x\Omega x}}$.
- Then the estimated asymptotic variance matrix of $\hat{\beta}_{\text{OLS}}$ is

$$\hat{V}[\hat{\beta}_{\text{OLS}}] = N^{-1}\widehat{\mathbf{M}}_{\mathbf{xx}}^{-1}\widehat{\mathbf{M}}_{\mathbf{x\Omega x}}\widehat{\mathbf{M}}_{\mathbf{xx}}^{-1}. \tag{21}$$

- This estimate is called a sandwich estimate, with $\widehat{\mathbf{M}}_{\mathbf{x\Omega x}}$ sandwiched between $\widehat{\mathbf{M}}_{\mathbf{xx}}^{-1}$ and $\widehat{\mathbf{M}}_{\mathbf{xx}}^{-1}$.