Econometric Methods (Econometrics I)

Lecture 1:

Introduction and Statistical Prerequisites

Prof. Dr. Kai Carstensen

Kiel University

Winter Term 2023/2024

Outline of this lecture

1. Introduction: on estimating causal effects
2. Some math: vector and matrix derivatives
3. Review of basic OLS: estimation
4. Review of basic OLS: inference
5. Conditional expectations
6. Linear projections

Reference: Wooldridge, Chapters 1-2.

1. Introduction

What econometrics is good for

Economics suggests important relationships, often with policy implications, but virtually never suggests quantitative magnitudes of **causal effects**.

- How does another year of education change earnings?
- Does reducing class size cause an improvement in student performance?
- What is the effect of having children on labor supply?
- Does lowering the business property tax rate (in Germany, say, the *Gewerbesteuerhebesatz*) cause an increase in city economic activity?
- What is the effect on output growth of a 1 percentage point increase in interest rates by the ECB?
- What is the effect on GDP of a reduction in VAT?

Note: Econometrics is also good for forecasting when using time-series data. But in this course we concentrate on cross section data. The time series dimension is discussed in Econometrics II. And panel data in the respective specialization.

How use data to measure causal effects?

- Ideally, we would like to have data from a **randomized controlled experiment**.
  - **Controlled** means that there are a control group that receives no treatment and a treatment group that receives a treatment.
  - **Randomized** means that the treatment is assigned randomly.
- But almost always we only have observational (nonexperimental) data:
  - returns to education
  - wage statistics
  - survey data
- Most of the course deals with difficulties arising from using observational data to estimate causal effects:
  - confounding effects (omitted factors)
  - simultaneous causality
  - correlation does not imply causation
  - nonlinearities
  - etc

How to find a causal relationship?

- Use or develop a plausible theory that explains why a causing variable, w, should have an impact on another variable, y.
- Mimic an experiment: Make sure that "everything else" is either held constant or randomly distributed between those treated and those not treated.
- How can we hold other factors constant? We can condition on them. This means that effectively we compare treated and non-treated entities which are otherwise identical.
- Take into account that economic relationships are not deterministic. Hence, focus on the expected (average) relationship between y and x holding a vector of control variables, $\mathbf{c}$, constant: $\mathrm{E}(y|w,\mathbf{c})$.
- What we are typically interested in, is (for continuous variables) the partial effect

$$\frac{\partial \mathrm{E}(y|w,\mathbf{c})}{\partial w}$$

- This sounds simple, especially if $\mathrm{E}(y|w,\mathbf{c})$ is a linear function of w and $\mathbf{c}$, but there are many reasons why it may not. This is what this lecture is about.

Data structures

In this course, we assume **random sampling**. This means that we

1. specify a **population model** and
2. draw an **independently and identically distributed** (iid) random sample (individuals, firms, etc.) $i = 1, 2, \ldots, N$ from this population.

Data structures

Population model:

- Most often a model of the conditional expectation. For example,

$$\mathrm{E}(y|w, \mathbf{c}) = \beta_0 + \beta_1 w + \boldsymbol{\gamma}' \mathbf{c}.$$

- Think about the population. This can be very important. For example, when you are interested in how another year of education changes earnings, you must be specific: the active working population? including the currently jobless? only people without university degree? only people without vocational training? etc
- Note: You have to adapt sampling and measuring (what are the earnings of jobless people?) to your definition of the population.

Data structures

Random sampling:

- Often a sensible assumption for cross-sectional data.
- Implies that **all** variables (dependent variables and explanatory variables) are random.
- Hence, the assumption of fixed regressors sometimes still used in the baseline regression model does not apply.
- Moreover, the normality assumptions often used in basic econometrics need not hold: why should the variables be normally distributed?
- Consequence: finite-sample distributions of classical t-tests and F-tests are unknown.
- Thus, we resort to large-sample approximations using asymptotic analysis. This seems justified as most cross-sectional data sets are of moderate or even large size.

Note: the random sampling assumption might be more difficult to defend for panel and time series data. But this is delegated to the respective courses.

Example

Wage offer function for women: Suppose that the natural log of the wage offer, $wage^o$, is determined as

$$\log(wage^o) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 married + u \tag{1}$$

where

- *educ* is years of schooling, *exper* is years of labor market experience, and *married* is a binary variable indicating marital status,
- u, called the error term or disturbance, contains unobserved factors that affect the wage offer,
- interest lies in the unknown parameters, the β_j,
- random sampling depends on the population of interest (all working women? all women over age 18?)

Indexing

For deriving the properties of estimators, it is often useful to write the population model for a generic draw from the population. Instead of

$$\log(wage^o) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 married + u$$

we then write

$$\log(wage_i^o) = \beta_0 + \beta_1 educ_i + \beta_2 exper_i + \beta_3 married_i + u_i, \tag{2}$$

where i indexes the individual.

Stating assumptions in terms of u_i and $x_i \equiv (educ_i, exper_i, married_i)$ is the same as stating assumptions in terms of u and x.

2. Some Math

References

Any book on matrix algebra and calculus should have a section on matrix differentiation.

A brief but nice treatment can be found in: H. Lütkepohl (2007), New Introduction to Multiple Time Series Analysis, Springer-Verlag, Berlin, Appendix A.13.

Also feasible: W.H. Greene (2012), Econometric Analysis, 7th ed., Pearson, Appendix A.

Vector and matrix differentiation

Rule 1: Let $f(\boldsymbol{\beta})$ be a scalar function of the $(n \times 1)$ vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_n)'$. The column vector of first partial derivatives is

$$\underbrace{\frac{\partial f}{\partial \boldsymbol{\beta}}}_{n \times 1} := \begin{bmatrix} \partial f / \partial \beta_1 \\ \vdots \\ \partial f / \partial \beta_n \end{bmatrix}.$$

The row vector of first partial derivatives (often called gradient, symbol: $\nabla_{\boldsymbol{\beta}}$) is

$$\underbrace{\frac{\partial f}{\partial \boldsymbol{\beta}'}}_{1 \times n} = \nabla_{\boldsymbol{\beta}} f(\boldsymbol{\beta}) := \left[\frac{\partial f}{\partial \beta_1}, \ldots, \frac{\partial f}{\partial \beta_n} \right].$$

The $(n \times n)$ Hessian matrix of second partial derivatives is

$$\underbrace{\frac{\partial^2 f}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'}}_{n \times n} = \frac{\partial \frac{\partial f}{\partial \boldsymbol{\beta}}}{\partial \boldsymbol{\beta}'} = \nabla_{\boldsymbol{\beta}} \left[\nabla_{\boldsymbol{\beta}} f(\boldsymbol{\beta}) \right]' := \begin{bmatrix} \frac{\partial^2 f}{\partial \beta_1 \partial \beta_1} & \cdots & \frac{\partial^2 f}{\partial \beta_1 \partial \beta_n} \\ \vdots & & \vdots \\ \frac{\partial^2 f}{\partial \beta_n \partial \beta_1} & \cdots & \frac{\partial^2 f}{\partial \beta_n \partial \beta_n} \end{bmatrix}.$$

Example:

$$f(\boldsymbol{\beta}) = \beta_1^2 + \cdots + \beta_n^2 = \sum_{i=1}^{n} \beta_i^2 = \boldsymbol{\beta}'\boldsymbol{\beta}$$

$$\frac{\partial f}{\partial \boldsymbol{\beta}} = \frac{\partial \boldsymbol{\beta}'\boldsymbol{\beta}}{\partial \boldsymbol{\beta}} = \begin{bmatrix} \frac{\partial f}{\partial \beta_1} \\ \vdots \\ \frac{\partial f}{\partial \beta_n} \end{bmatrix} = \begin{bmatrix} 2\beta_1 \\ \vdots \\ 2\beta_n \end{bmatrix} = 2 \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_n \end{bmatrix} = 2\boldsymbol{\beta}$$

$$\frac{\partial^2 f}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = \frac{\partial \boldsymbol{\beta}'\boldsymbol{\beta}}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = \begin{bmatrix} 2 & 0 & \cdots & 0 \\ 0 & 2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 2 \end{bmatrix} = 2\boldsymbol{I}_n$$

Rule 2: Let

$$h(\boldsymbol{\beta}) = \begin{bmatrix} h_1(\boldsymbol{\beta}) \\ \vdots \\ h_m(\boldsymbol{\beta}) \end{bmatrix}$$

be a $(m \times 1)$ vector function that depends on the $(n \times 1)$ vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_n)'$.

Then the $(m \times n)$ matrix of first partial derivatives is

$$\frac{\partial h}{\partial \boldsymbol{\beta}'} := \begin{bmatrix} \frac{\partial h_1}{\partial \beta_1} & \cdots & \frac{\partial h_1}{\partial \beta_n} \\ \vdots & & \vdots \\ \frac{\partial h_m}{\partial \beta_1} & \cdots & \frac{\partial h_m}{\partial \beta_n} \end{bmatrix}.$$

Example:

$$h(\boldsymbol{\beta}) = \mathbf{X}\boldsymbol{\beta} = \begin{bmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & & \vdots \\ x_{m1} & \cdots & x_{mn} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_n \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} x_{1i}\beta_i \\ \vdots \\ \sum_{i=1}^{n} x_{mi}\beta_i \end{bmatrix}$$

$$\frac{\partial h}{\partial \boldsymbol{\beta}'} = \frac{\partial \mathbf{X}\boldsymbol{\beta}}{\partial \boldsymbol{\beta}'} = \begin{bmatrix} \frac{\partial \sum_{i=1}^{n} x_{1i}\beta_i}{\partial \beta_1} & \cdots & \frac{\partial \sum_{i=1}^{n} x_{1i}\beta_i}{\partial \beta_n} \\ \vdots & & \vdots \\ \frac{\partial \sum_{i=1}^{n} x_{mi}\beta_i}{\partial \beta_1} & \cdots & \frac{\partial \sum_{i=1}^{n} x_{mi}\beta_i}{\partial \beta_n} \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & & \vdots \\ x_{m1} & \cdots & x_{mn} \end{bmatrix} = \mathbf{X}$$

Rule 3 (Chain rule for vector differentiation): Let $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ be $(m \times 1)$ and $(n \times 1)$ vectors, respectively, and suppose $h(\boldsymbol{\alpha})$ is $(p \times 1)$ and $g(\boldsymbol{\beta})$ is $(m \times 1)$. Further assume $\boldsymbol{\alpha} = g(\boldsymbol{\beta})$ and define $f(\boldsymbol{\beta}) = h(g(\boldsymbol{\beta}))$. Then

$$\frac{\partial f(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'} = \frac{\partial h(g(\boldsymbol{\beta}))}{\partial \boldsymbol{\beta}'} = \frac{\partial h(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}'} \frac{\partial g(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'}.$$

Example:

$$h(\boldsymbol{\alpha}) = \boldsymbol{\alpha}'\boldsymbol{\alpha}, \quad \boldsymbol{\alpha} = g(\boldsymbol{\beta}) = y - X\boldsymbol{\beta}, \quad f(\boldsymbol{\beta}) = (y - X\boldsymbol{\beta})'(y - X\boldsymbol{\beta})$$

$$\frac{\partial f}{\partial \boldsymbol{\beta}'} = \frac{\partial h(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}'} \frac{\partial g(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'} = \frac{\partial \boldsymbol{\alpha}'\boldsymbol{\alpha}}{\partial \boldsymbol{\alpha}'} \frac{\partial (y - X\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'} = -2\boldsymbol{\alpha}'X = -2(y - X\boldsymbol{\beta})'X$$

3. Review of Basic OLS: Estimation

This is a review of Bachelor level econometrics. Those not familiar with it should join the prep course offered in the first half of the semester.

The population model we study is linear in its parameters,

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_K x_{Ki} + u_i$$

where

- $y, x_1, \ldots, x_K$ are observable random scalars,
- u is the unobservable random disturbance or error and
- $\beta_0, \beta_1, \beta_2, \ldots, \beta_K$ are the parameters to be estimated.

In many cases—especially when we write the model in vector form—we want to absorb the intercept into the slope parameters. Then we write

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_K x_{Ki} + u_i$$

and assume $x_{1i} = 1$.

The regression model in vector form:

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

where

$$\mathbf{x}_i \equiv (x_{1i}, \ldots, x_{Ki}) \qquad (1 \times K)$$
$$\boldsymbol{\beta} \equiv (\beta_1, \ldots, \beta_K)' \qquad (K \times 1).$$

Typically, in this formulation one regressor is a constant, e.g., $x_{1i} = 1$.

The regression model in matrix form for a sample of N observations:

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{U} \tag{3}$$

with the observation matrices

$$\mathbf{Y} \equiv (y_1, \ldots, y_N)' \qquad (N \times 1)$$
$$\mathbf{X} \equiv (\mathbf{x}_1', \ldots, \mathbf{x}_N')' \qquad (N \times K)$$

and the disturbances

$$\mathbf{U} \equiv (u_1, \ldots, u_N)' \qquad (N \times 1).$$

There is a sample of N observations $\mathbf{Y}$ and $\mathbf{X}$, where $\mathbf{X}$ has full column rank such that $\mathbf{X}'\mathbf{X}$ is invertible. We want to find a vector $\hat{\boldsymbol{\beta}}$ that minimizes the squared distance between the elements of $\mathbf{Y}$ and $\mathbf{X}\hat{\boldsymbol{\beta}}$. Hence, we have to minimize the objective function

$$S(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{N}(y_i - \mathbf{x}_i\hat{\boldsymbol{\beta}})^2 = \sum_{i=1}^{N}\hat{u}_i^2 = \hat{\mathbf{U}}'\hat{\mathbf{U}} = (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}).$$

First-order derivative:

$$\frac{\partial S(\hat{\boldsymbol{\beta}})}{\partial \hat{\boldsymbol{\beta}}'} = \frac{\partial S}{\partial \hat{\mathbf{U}}'}\frac{\partial \hat{\mathbf{U}}}{\partial \hat{\boldsymbol{\beta}}'} = \frac{\partial \hat{\mathbf{U}}'\hat{\mathbf{U}}}{\partial \hat{\mathbf{U}}'}\frac{\partial(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})}{\partial \hat{\boldsymbol{\beta}}'} = -2\hat{\mathbf{U}}'\mathbf{X} = -2(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'\mathbf{X}$$

First order condition:

$$-2(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{LS})'\mathbf{X} \stackrel{!}{=} 0 \quad \Rightarrow \quad \mathbf{Y}'\mathbf{X} - \hat{\boldsymbol{\beta}}_{LS}'\mathbf{X}'\mathbf{X} = 0 \quad \Rightarrow \quad \hat{\boldsymbol{\beta}}_{LS}' = \mathbf{Y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

OLS estimator:

$$\hat{\boldsymbol{\beta}}_{LS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

Does the OLS estimator find the minimum of the objective function $S(\hat{\beta})$?

Check the Hessian:

$$\frac{\partial^2 S(\hat{\boldsymbol{\beta}})}{\partial \hat{\boldsymbol{\beta}} \partial \hat{\boldsymbol{\beta}}'} = \frac{\partial}{\partial \hat{\boldsymbol{\beta}}'} \left(\frac{\partial S(\hat{\boldsymbol{\beta}})}{\partial \hat{\boldsymbol{\beta}}} \right) = \frac{\partial}{\partial \hat{\boldsymbol{\beta}}'} \left(-2\mathbf{X}'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \right)$$
$$= \frac{\partial}{\partial \hat{\boldsymbol{\beta}}'} \left(-2\mathbf{X}'\mathbf{Y} + 2\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} \right) = \frac{\partial (2\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}})}{\partial \hat{\boldsymbol{\beta}}'} = 2\mathbf{X}'\mathbf{X}.$$

The Hessian is (for any value of $\hat{\beta}_{LS}$) positive definite because the quadratic form $\mathbf{X}'\mathbf{X}$ is positive definite. Hence, the OLS estimator found a (local) minimum of the objective function.

Note that the "matrix expression"

$$\hat{\boldsymbol{\beta}}_{LS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

is equivalent to the "summation expression"

$$\hat{\boldsymbol{\beta}}_{LS} = \left(N^{-1} \sum_{i=1}^{N} \mathbf{x}_i' \mathbf{x}_i \right)^{-1} \left(N^{-1} \sum_{i=1}^{N} \mathbf{x}_i' y_i \right).$$

When the regressor list includes an intercept (what it almost always does),

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_K x_{Ki} + u_i = \beta_0 + \mathbf{x}_i \tilde{\boldsymbol{\beta}} + u_i,$$

we can derive an insightful expression for the OLS estimator if we separate the intercept β_0 from the other coefficients $\tilde{\boldsymbol{\beta}} = (\beta_1, \ldots, \beta_K)'$ and define $\mathbf{x}_i = (x_{1i}, \ldots, x_{Ki})$.

To this end, compute the average over all observations,

$$\bar{y} = \beta_0 + \bar{\mathbf{x}} \tilde{\boldsymbol{\beta}} + \bar{u},$$

and subtract it from the previous equation which yields the "de-meaned" equation

$$\underbrace{y_i - \bar{y}}_{\tilde{y}_i} = \underbrace{(\mathbf{x}_i - \bar{\mathbf{x}})}_{\tilde{\mathbf{x}}_i} \tilde{\boldsymbol{\beta}} + \underbrace{u_i - \bar{u}}_{\tilde{u}_i}.$$

Applying OLS to the de-meaned equation yields

$$\begin{aligned}\hat{\tilde{\beta}}_{LS} &= \left(N^{-1}\sum_{i=1}^{N}\tilde{\mathbf{x}}_i'\tilde{\mathbf{x}}_i\right)^{-1}\left(N^{-1}\sum_{i=1}^{N}\tilde{\mathbf{x}}_i'\tilde{y}_i\right)\\ &= \left(N^{-1}\sum_{i=1}^{N}(\mathbf{x}_i-\bar{\mathbf{x}})'(\mathbf{x}_i-\bar{\mathbf{x}})\right)^{-1}\left(N^{-1}\sum_{i=1}^{N}(\mathbf{x}_i-\bar{\mathbf{x}})'(y_i-\bar{y})\right)\\ &= \left(s_{\mathbf{x}}^2\right)^{-1}s_{\mathbf{x}',y}\end{aligned}$$

where

- $s_{\mathbf{x}}^2$ is the sample variance matrix of $\mathbf{x}$, and
- $s_{\mathbf{x}',y}$ is the sample covariance matrix between $\mathbf{x}'$ and y.

Define the OLS residuals as

$$\hat{\mathbf{U}} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{LS} = \mathbf{Y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = (\mathbf{I}_N - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{Y} = \mathbf{M}_\mathbf{X}\mathbf{Y},$$

where

$$\mathbf{M}_\mathbf{X} = \mathbf{I}_N - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

is a symmetric, idempotent matrix of rank $N - K$. Hence,

$$\mathbf{M}_\mathbf{X} = \mathbf{M}_\mathbf{X}' \qquad \text{and} \qquad \mathbf{M}_\mathbf{X}\mathbf{M}_\mathbf{X} = \mathbf{M}_\mathbf{X}.$$

For later use, the following reformulation is helpful

$$\hat{\mathbf{U}} = \mathbf{M}_\mathbf{X}\mathbf{Y} = \mathbf{M}_\mathbf{X}(\mathbf{X}\boldsymbol{\beta} + \mathbf{U}) = \mathbf{M}_\mathbf{X}\mathbf{X}\boldsymbol{\beta} + \mathbf{M}_\mathbf{X}\mathbf{U} = \mathbf{M}_\mathbf{X}\mathbf{U}$$

because $\mathbf{M}_\mathbf{X}\mathbf{X} = (\mathbf{I}_N - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{X} = \mathbf{X} - \mathbf{X} = 0$.

A widely used measure of fit, the R^2 is based on a variance decomposition which can be represented in terms of the de-meaned equation.

Defining the de-meaned observation matrices $\tilde{\mathbf{Y}}$ and $\tilde{\mathbf{X}}$, the first order condition implies

$$\tilde{\mathbf{X}}'(\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\hat{\tilde{\beta}}_{LS}) = 0 \qquad \Rightarrow \qquad \tilde{\mathbf{X}}'\hat{\mathbf{U}} = 0.$$

Define $\hat{\tilde{\mathbf{Y}}} = \tilde{\mathbf{X}}\hat{\tilde{\beta}}_{LS}$ and thus $\tilde{\mathbf{Y}} = \hat{\tilde{\mathbf{Y}}} + \hat{\mathbf{U}}$. Then the following decomposition holds:

$$\tilde{\mathbf{Y}}'\tilde{\mathbf{Y}} = (\hat{\tilde{\mathbf{Y}}} + \hat{\mathbf{U}})'(\hat{\tilde{\mathbf{Y}}} + \hat{\mathbf{U}}) = \hat{\tilde{\mathbf{Y}}}'\hat{\tilde{\mathbf{Y}}} + \hat{\mathbf{U}}'\hat{\mathbf{U}}$$

because the cross product is zero, $\hat{\tilde{\mathbf{Y}}}'\hat{\mathbf{U}} = \hat{\tilde{\beta}}'_{LS}\tilde{\mathbf{X}}'\hat{\mathbf{U}} = 0$.

Therefore, a measure of fit is

$$R^2 = \frac{\text{explained sum of squares}}{\text{total sum of squares}} = \frac{\hat{\tilde{\mathbf{Y}}}'\hat{\tilde{\mathbf{Y}}}}{\tilde{\mathbf{Y}}'\tilde{\mathbf{Y}}} = \frac{\tilde{\mathbf{Y}}'\tilde{\mathbf{Y}} - \hat{\mathbf{U}}'\hat{\mathbf{U}}}{\tilde{\mathbf{Y}}'\tilde{\mathbf{Y}}} = 1 - \frac{\hat{\mathbf{U}}'\hat{\mathbf{U}}}{\tilde{\mathbf{Y}}'\tilde{\mathbf{Y}}}.$$

In summation form, we obtain

$$R^2 = 1 - \frac{\hat{\mathbf{U}}'\hat{\mathbf{U}}}{\tilde{\mathbf{Y}}'\tilde{\mathbf{Y}}} = 1 - \frac{\sum_{i=1}^{N} \hat{u}_i^2}{\sum_{i=1}^{N} \tilde{y}_i^2} = 1 - \frac{\sum_{i=1}^{N} \hat{u}_i^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2}.$$

Using

$$\begin{aligned}\sum_{i=1}^{N} (y_i - \bar{y})^2 &= \sum_{i=1}^{N} (y_i^2 - 2\bar{y}y_i + \bar{y}^2) = \sum_{i=1}^{N} y_i^2 - 2\bar{y}\sum_{i=1}^{N} y_i + N\bar{y}^2 \\ &= \sum_{i=1}^{N} y_i^2 - 2N\bar{y}^2 + N\bar{y}^2 = \sum_{i=1}^{N} y_i^2 - N\bar{y}^2 = \mathbf{Y}'\mathbf{Y} - N\bar{y}^2\end{aligned}$$

the R^2 can also be written as

$$R^2 = 1 - \frac{\hat{\mathbf{U}}'\hat{\mathbf{U}}}{\mathbf{Y}'\mathbf{Y} - N\bar{y}^2}.$$

4. Review of Basic OLS: Inference

This is a review of Bachelor level econometrics. Those not familiar with it should join the prep course offered in the first half of the semester.

Classical (or Gauss-Markov) assumptions

In this section (and only in this section), we make the following (strict) assumptions:

- The K regressors $\mathbf{x}_i$ are not random but **fixed numbers**.
- The observation matrix $\mathbf{X}$ has full column rank K so that $\mathbf{X}'\mathbf{X}$ is invertible.
- The disturbances u_i, $i = 1, \ldots, N$, are independently and identically normally distributed with $\mathrm{E}[u_i] = 0$ and $\mathrm{Var}[u_i] = \sigma^2$:

$$u_i \sim \mathrm{Normal}(0, \sigma^2).$$

As all u_i, $i = 1, \ldots, N$, have mean zero, so has the vector $\mathbf{U}$: $\mathrm{E}[\mathbf{U}] = \mathbf{0}$.

The i.i.d. assumption on u_i implies that any two u_i and u_j, $i \neq j$, are uncorrelated:

$$\mathrm{Cov}[u_i, u_j] = E[u_i u_j] = 0.$$

Hence, the variance of the vector of disturbances is

$$\mathrm{Var}[\mathbf{U}] = \mathrm{E}\left[(\mathbf{U} - \mathrm{E}[\mathbf{U}])(\mathbf{U} - \mathrm{E}[\mathbf{U}])'\right] = \mathrm{E}\left[\mathbf{U}\mathbf{U}'\right]$$

$$= \mathrm{E}\left[\begin{pmatrix} u_1 \\ \vdots \\ u_N \end{pmatrix} \begin{pmatrix} u_1 & \cdots & u_N \end{pmatrix}\right] = \mathrm{E}\begin{pmatrix} u_1^2 & u_1 u_2 & \cdots & u_1 u_N \\ u_2 u_1 & u_2^2 & \cdots & u_2 u_N \\ \vdots & \vdots & \ddots & \vdots \\ u_N u_1 & u_N u_2 & \cdots & u_N^2 \end{pmatrix}$$

$$= \begin{pmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{pmatrix} = \sigma^2 \mathbf{I}_N$$

Recall that all u_i, $i = 1, \ldots, N$, are individually distributed as

$$u_i \sim \text{Normal}(0, \sigma^2).$$

Given that $\text{E}[\mathbf{U}] = \mathbf{0}$ and $\text{Var}[\mathbf{U}] = \sigma^2 \mathbf{I}_N$, the joint distribution of the vector of disturbances is

$$\mathbf{U} \sim \text{Normal}(\mathbf{0}, \sigma^2 \mathbf{I}_N).$$

To find the mean of the OLS estimator

$$\hat{\boldsymbol{\beta}}_{LS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

substitute the model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{U}$$

which yields

$$\hat{\boldsymbol{\beta}}_{LS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \mathbf{U}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{U} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{U}.$$

Now take expectations

$$\mathrm{E}[\hat{\boldsymbol{\beta}}_{LS}] = \mathrm{E}[\boldsymbol{\beta}] + \mathrm{E}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{U}] = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\,\mathrm{E}[\mathbf{U}] = \boldsymbol{\beta}$$

because the regressors are nonrandom and $\mathrm{E}[\mathbf{U}] = 0$.

Result: the OLS estimator is unbiased.

The variance of the OLS estimator is

$$
\begin{aligned}
\mathrm{Var}[\hat{\boldsymbol{\beta}}_{LS}] &= \mathrm{E}\left[\hat{\boldsymbol{\beta}}_{LS} - \mathrm{E}(\hat{\boldsymbol{\beta}}_{LS})\right]\left[\hat{\boldsymbol{\beta}}_{LS} - \mathrm{E}(\hat{\boldsymbol{\beta}}_{LS})\right]' \\
&= \mathrm{E}\left[\hat{\boldsymbol{\beta}}_{LS} - \boldsymbol{\beta}\right]\left[\hat{\boldsymbol{\beta}}_{LS} - \boldsymbol{\beta}\right]' \\
&= \mathrm{E}\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{U}\right]\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{U}\right]' \\
&= \mathrm{E}\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{U}\mathbf{U}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\right] \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\,\mathrm{E}[\mathbf{U}\mathbf{U}']\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2\mathbf{I}_N\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.
\end{aligned}
$$

Rule: Let $\mathbf{W}$ be a $N \times 1$ random vector with

$$\mathbf{W} \sim \text{Normal}(\mathbf{0}, \mathbf{V}),$$

and let $\mathbf{A}$ be a $K \times N$ matrix of full row rank K and $\mathbf{B}$ a K vector. Then the $K \times 1$ random vector

$$\mathbf{Z} = \mathbf{AW} + \mathbf{B}$$

is normally distributed with

$$\text{E}[\mathbf{Z}] = \text{E}[\mathbf{AW} + \mathbf{B}] = \mathbf{B}$$

$$\text{Var}[\mathbf{Z}] = \text{E}[(\mathbf{Z} - \mathbf{B})(\mathbf{Z} - \mathbf{B})'] = \text{E}[\mathbf{AWW}'\mathbf{A}'] = \mathbf{A}\,\text{E}[\mathbf{WW}']\mathbf{A}' = \mathbf{AVA}'.$$

Hence,

$$\mathbf{Z} \sim \text{Normal}(\mathbf{B}, \mathbf{AVA}').$$

To find the distribution of the OLS estimator, recall the expression

$$\hat{\boldsymbol{\beta}}_{LS} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{U}.$$

Since $\boldsymbol{\beta}$ and $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ are constant numbers, $\hat{\boldsymbol{\beta}}_{LS}$ is a linear function of the random vector $\mathbf{U}$.

Using the rule given on the previous slide, $\hat{\boldsymbol{\beta}}_{LS}$ is thus normally distributed with mean

$$\mathrm{E}[\hat{\boldsymbol{\beta}}_{LS}] = \boldsymbol{\beta}$$

and variance

$$\mathrm{Var}[\hat{\boldsymbol{\beta}}_{LS}] = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.$$

Result:

$$\hat{\boldsymbol{\beta}}_{LS} \sim \mathrm{Normal}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}).$$

Unfortunately, this distribution is not very helpful in practice because the variance σ^2 of the disturbances is unknown and must be estimated, inducing additional estimation uncertainty that must be accounted for.

The variance σ^2 is estimated as

$$\hat{\sigma}^2 = \frac{1}{N-K}\sum_{i=1}^{N}\hat{u}_i^2 = \frac{1}{N-K}\hat{\mathbf{U}}'\hat{\mathbf{U}}.$$

Due to the previous result $\hat{\mathbf{U}} = \mathbf{M}_\mathbf{X}\mathbf{U}$, this is equivalent to

$$\hat{\sigma}^2 = \frac{1}{N-K}\hat{\mathbf{U}}'\hat{\mathbf{U}} = \frac{1}{N-K}\mathbf{U}'\mathbf{M}_\mathbf{X}'\mathbf{M}_\mathbf{X}\mathbf{U} = \frac{1}{N-K}\mathbf{U}'\mathbf{M}_\mathbf{X}\mathbf{U}.$$

This is a quadratic form in normal random variables.

Rule: For standard normally distributed random variables, $\mathbf{W} \sim \text{Normal}(\mathbf{0}, \mathbf{I}_N)$, the quadratic form

$$\mathbf{W}'\mathbf{C}\mathbf{W},$$

where $\mathbf{C}$ is a symmetric, idempotent matrix of fixed numbers with rank r, is χ^2 distributed with r degrees of freedom.

To apply this rule, note that

$$\mathbf{U}/\sigma \sim \text{Normal}(\mathbf{0}, \mathbf{I}_N)$$

so that

$$\frac{\mathbf{U}'\mathbf{M}_\mathbf{X}\mathbf{U}}{\sigma^2} = \frac{N-K}{\sigma^2} \times \frac{\mathbf{U}'\mathbf{M}_\mathbf{X}\mathbf{U}}{N-K} = \frac{N-K}{\sigma^2} \times \hat{\sigma}^2 \sim \chi^2_{N-K}$$

because $\mathbf{M}_\mathbf{X}$ is a symmetric, idempotent matrix of fixed numbers with rank $N-K$.

Rule: If

(i) the random variable Z has a standard normal distribution,

(ii) the random variable W has a χ^2_r distribution, and

(iii) Z and W are independently distributed

then the random variable $Z/\sqrt{W/r}$ is t distributed with r degrees of freedom,

$$\frac{Z}{\sqrt{W/r}} \sim t_r.$$

The t-statistic for a single parameter β_i is

$$t = \frac{\hat{\beta}_{LS,i} - \beta_i}{\widehat{\text{se}(\hat{\beta}_{LS,i})}},$$

where $\widehat{\text{se}(\hat{\beta}_{LS,i})}$ is the i-th diagonal element of the root of the estimated variance of $\hat{\boldsymbol{\beta}}_{LS}$

$$\widehat{\text{Var}(\hat{\beta}_{LS})} = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}.$$

Note that this can be rewritten as

$$\widehat{\text{Var}(\hat{\beta}_{LS})} = \frac{\hat{\sigma}^2}{\sigma^2} \times \sigma^2(\mathbf{X}'\mathbf{X})^{-1} = \frac{\hat{\sigma}^2}{\sigma^2}\,\text{Var}(\hat{\beta}_{LS}).$$

Hence, the estimated variance of $\hat{\beta}_{LS,i}$ is

$$\widehat{\text{Var}(\hat{\beta}_{LS,i})} = \frac{\hat{\sigma}^2}{\sigma^2}\,\text{Var}(\hat{\beta}_{LS,i})$$

and the estimated standard error is thus

$$\widehat{\text{se}(\hat{\beta}_{LS,i})} = \frac{\hat{\sigma}}{\sigma}\sqrt{\text{Var}(\hat{\beta}_{LS,i})}.$$

Therefore, the t-statistic for a single parameter β_i can be written as

$$t_i = \frac{\hat{\beta}_{LS,i} - \beta_i}{\widehat{se(\hat{\beta}_{LS,i})}} = \frac{\hat{\beta}_{LS,i} - \beta_i}{\frac{\hat{\sigma}}{\sigma}\sqrt{Var(\hat{\beta}_{LS,i})}} = \frac{\left(\hat{\beta}_{LS,i} - \beta_i\right) / \sqrt{Var(\hat{\beta}_{LS,i})}}{\hat{\sigma}/\sigma}$$

$$= \frac{\left(\hat{\beta}_{LS,i} - \beta_i\right) / \sqrt{Var(\hat{\beta}_{LS,i})}}{\sqrt{\frac{\mathbf{U}'\mathbf{M}_\mathbf{X}\mathbf{U}}{\sigma^2}/(N-K)}} = \frac{Z}{\sqrt{W/(N-K)}}.$$

Note that

(i) the nominator $Z = \left(\hat{\beta}_{LS,i} - \beta_i\right) / \sqrt{Var(\hat{\beta}_{LS,i})}$ has a standard normal distribution,

(ii) $W = \frac{\mathbf{U}'\mathbf{M}_\mathbf{X}\mathbf{U}}{\sigma^2}$ has a χ^2_{N-K} distribution, and

(iii) Z and W are independently distributed (we do not show this here).

Hence, t_i is t-distributed with $N - K$ degrees of freedom.

Confidence intervals

Since the t statistic does not depend on any unknowns, we can use it to derive confidence intervals and coefficient tests.

A $(1-\alpha)100\%$ two-sided confidence interval for β_i can be derived from

$$\Pr\left(t_{N-K,\alpha/2} \leq t \leq t_{N-K,1-\alpha/2}\right) = 1-\alpha,$$

where $t_{N-K,\alpha/2}$ is the $\alpha/2$-quantile and $t_{N-K,1-\alpha/2}$ is the $1-\alpha/2$-quantile of the t distribution with $N-K$ degrees of freedom.

Note: the t distribution is symmetric, thus $t_{N-K,\alpha/2} = -t_{N-K,1-\alpha/2}$:

$$\Pr\left(-t_{N-K,1-\alpha/2} \leq t \leq t_{N-K,1-\alpha/2}\right) = 1-\alpha.$$

Confidence intervals

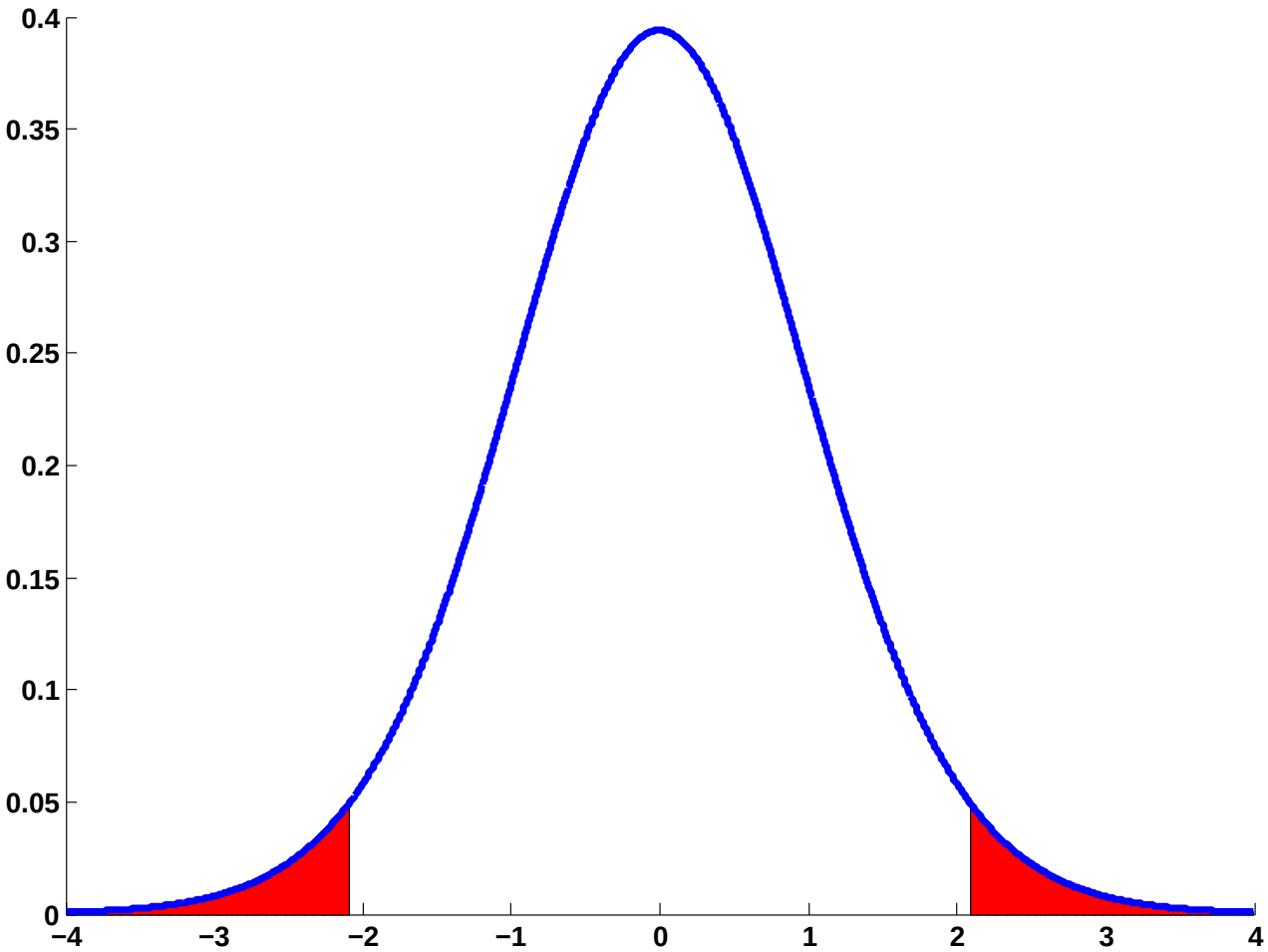


Figure: t distribution with 20 degrees of freedom

Confidence intervals

Use the symmetry and rearrange:

$$\Pr\left(-t_{N-K,1-\alpha/2} \leq (\hat{\beta}_i - \beta_i)/\widehat{\text{se}(\hat{\beta}_{LS,i})} \leq t_{N-K,1-\alpha/2}\right) = 1-\alpha$$

$$\Pr\left(\hat{\beta}_i - \widehat{\text{se}(\hat{\beta}_{LS,i})}t_{N-K,1-\alpha/2} \leq \beta_i \leq \hat{\beta}_i + \widehat{\text{se}(\hat{\beta}_{LS,i})}t_{N-K,1-\alpha/2}\right) = 1-\alpha.$$

So a $(1-\alpha)100\%$ two-sided confidence interval for β_i is given by the set

$$\left[\hat{\beta}_i - \widehat{\text{se}(\hat{\beta}_{LS,i})}t_{N-K,1-\alpha/2},\ \hat{\beta}_i + \widehat{\text{se}(\hat{\beta}_{LS,i})}t_{N-K,1-\alpha/2}\right].$$

It contains the true value of β_i with a probability of $(1-\alpha)100\%$.

Important: By "probability of $(1-\alpha)100\%$" we mean that if we could draw a large number of independent samples of $\mathbf{Y}$ from the population (remember: the $\mathbf{X}$ are fixed in this setting), in $(1-\alpha)100\%$ of the cases the confidence interval would contain the true value of β_i.

To test a two-sided hypothesis concerning a single parameter at the significance level of $\alpha 100\%$,

$$H_0 : \beta_i = \beta_{i,0} \quad \text{vs.} \quad H_1 : \beta_i \neq \beta_{i,0}$$

we can use the test statistic

$$t = \frac{\hat{\beta}_{LS,i} - \beta_{i,0}}{\widehat{\text{se}(\hat{\beta}_{LS,i})}}.$$

Under the null hypothesis, t is t_{N-K} distributed. For a test decision, use again the symmetry of the t distribution and compare the absolute value of the t statistic with the $(1 - \alpha/2)100\%$ quantile of the t_{N-K} distribution:

$$|t| > t_{N-K,1-\alpha/2} \quad \Rightarrow \quad \text{reject } H_0.$$

The p-value

For a two-sided hypothesis the p-value is the probability of observing a value of $\hat{\beta}_{LS,i}$ at least as different from β_i as the estimate actually computed from the data at hand ($\hat{\beta}^{act}_{LS,i}$) if the null hypothesis ($\beta_i = \beta_{i,0}$) is correct:

$$\begin{aligned} p\text{-value} &= \Pr_{H_0}\left[|\hat{\beta}_i - \beta_{i,0}| > |\hat{\beta}^{act}_i - \beta_{i,0}|\right] \\ &= \Pr_{H_0}\left[\left|\frac{\hat{\beta}_i - \beta_{i,0}}{\widehat{\text{se}}(\hat{\beta}_{LS,i})}\right| > \left|\frac{\hat{\beta}^{act}_i - \beta_{i,0}}{\widehat{\text{se}}(\hat{\beta}_{LS,i})}\right|\right] = \Pr_{H_0}\left[|t| > |t^{act}|\right]. \end{aligned}$$

Because under the null the t-statistic is t_{N-K} distributed, we have

$$p\text{-value} = \Pr(t < -|t^{act}|) + \Pr(t > |t^{act}|) = 2\Pr(t > |t^{act}|).$$

The p-value is routinely computed by econometrics packages such as Stata.

For any test, the null hypothesis is rejected if the p-value is smaller than the significance level α.

F tests

Consider a joint hypothesis that is linear in the coefficients and imposes q restrictions, where $q \leq K$. This hypothesis can be written in matrix notation as

$$\mathbf{R}\boldsymbol{\beta} = \mathbf{r},$$

where $\mathbf{R}$ is a $q \times K$ nonrandom matrix with full row rank and $\mathbf{r}$ is a $q \times 1$ nonrandom vector.

Example: After estimating the linear model

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i$$

(note: $x_{1i} = 1$) we want to test the joint hypothesis that (a) $\beta_1 = \beta_2$ and thus $\beta_1 - \beta_2 = 0$, and (b) $\beta_2 + \beta_3 = 1$. This gives rise to the restriction matrices

$$\begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

F tests

To test the joint restrictions, we may use an F test. The test statistic is

$$F = (\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})' \left[\mathbf{R}\widehat{\text{Var}(\hat{\beta}_{LS})}\mathbf{R}'\right]^{-1} (\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})/q$$

$$= \frac{(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})' \left[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\right]^{-1} (\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})/q}{\hat{\sigma}^2}.$$

Under the null hypothesis it is F distributed with q and $N - K$ degrees of freedom,

$$F \sim F_{q,N-K},$$

which is used to find the critical value.

Gauss-Markov Theorem

Theorem: In the linear regression model with regressor matrix $\mathbf{X}$, the OLS estimator $\hat{\boldsymbol{\beta}}$ is the minimum variance linear unbiased estimator of $\boldsymbol{\beta}$ if the Gauss-Markov assumptions hold.

Comments:

- Hence, if we restrict our class of estimators to those which are a linear function of $\mathbf{y}$ and unbiased, there is no better estimator than the OLS estimator in terms of estimation variance.
- The theorem can be extended to random regressors $\mathbf{X}$, see Greene (2012, p. 60-61).

5. Conditional Expectation

The role of conditional expectations in econometrics

In most cases, we are interested in the average relationship between y and x holding a vector of control variables, $\mathbf{c}$, constant.

This may be called a structural conditional expectation:

$$\mathrm{E}(y|x, \mathbf{c}).$$

Most OLS models are intended to estimate such a structural conditional expectation. Thus the importance of this concept.

Estimation is straightforward when you have a random sample.

But many complications can arise (which econometricians have to solve), in particular

- measurement error
- reverse causality
- omitted (unobserved) variables

Let us first review the concept of conditional expectation.

Definition

Let y be a random variable (*explained variable*), and let $\mathbf{x} \equiv (x_1, x_2, \ldots, x_K)$ be a $1 \times \mathbf{K}$ random vector of *explanatory variables*. Then the conditional expectation of y given $\mathbf{x}$ is

$$\mathrm{E}(y|x_1, \ldots, x_K) = \mathrm{E}(y|\mathbf{x}) = \int_{-\infty}^{\infty} y \, f_{y|\mathbf{x}}(y|\mathbf{x}) \, dy$$

when the random variables are continuous and

$$\mathrm{E}(y|x_1, \ldots, x_K) = \mathrm{E}(y|\mathbf{x}) = \sum_{y_j} y_j \, f_{y|\mathbf{x}}(y_j|\mathbf{x})$$

when the random variables are discrete.

Functional relationship

We are interested in the way $\mathsf{E}(y|\mathbf{x})$ depends on $\mathbf{x}$.

To make this explicit, we write it as the function $\mu(\mathbf{x})$, i.e.,

$$\mathsf{E}(y|x_1, \ldots, x_K) = \mathsf{E}(y|\mathbf{x}) = \mu(x_1, x_2, \ldots, x_K) = \mu(\mathbf{x}).$$

In many cases (such as OLS), we assume $\mu(\mathbf{x})$ is a function that depends on a finite number of (possibly unknown) parameters.

This yields a **parametric** model for $\mathsf{E}(y|\mathbf{x}) = \mu(\mathbf{x})$.

For $K = 2$ explanatory variables, here are some examples.

- linear in regressors and parameters:

$$\mathsf{E}(y|x_1, x_2) = \mu(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

- nonlinear in regressors but linear in parameters:

$$\begin{aligned} \mathsf{E}(y|x_1, x_2) &= \mu(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_2^2 \\ \mathsf{E}(y|x_1, x_2) &= \mu(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 \end{aligned}$$

- nonlinear in regressors and parameters

$$\mathsf{E}(y|x_1, x_2) = \mu(x_1, x_2) = \exp[\beta_0 + \beta_1 \log(x_1) + \beta_2 x_2], \qquad y \geq 0, x_1 > 0.$$

Random variable or numeric value?

Note that

$$\mathrm{E}\left(y|\mathbf{x}\right) = \mu(\mathbf{x})$$

is a random variable because $\mathbf{x}$ is a vector of random variables.

Sometimes it is necessary to concentrate on the case when $\mathbf{x}$ takes on a particular (nonrandom) value $\mathbf{x}_0$. Then

$$\mathrm{E}\left(y|\mathbf{x} = \mathbf{x}_0\right) = \mu(\mathbf{x}_0)$$

is a numeric value (a realization of a random variable).

In the following, we will often be a bit imprecise, mostly writing

$$\mathrm{E}\left(y|\mathbf{x}\right) = \mu(\mathbf{x})$$

even if a realization is meant. This should be clear from the context.

Partial effects

To study the effect of a causing variable on the dependent variable, one may calculate the partial effect of x_j on $\mathrm{E}(y|\mathbf{x})$ (or, to be somewhat imprecise, the partial effect of x_j on y):

$$\text{Partial effect:} \qquad \frac{\partial \mathrm{E}(y|\mathbf{x})}{\partial x_j} = \frac{\partial \mu(\mathbf{x})}{\partial x_j}.$$

The partial derivative is defined if $\mu(\cdot)$ is differentiable and x_j is continuous.

The change in $\mathrm{E}(y|\mathbf{x})$ when x_j is increased by a small amount, holding $x_1, \ldots, x_{j-1}, x_{j+1}, \ldots x_K$ constant, thus is approximately:

$$\Delta \mathrm{E}\,(y|\mathbf{x}) \approx \frac{\partial \mu(\mathbf{x})}{\partial x_j} \cdot \Delta x_j, \qquad \text{holding } x_1, \ldots, x_{j-1}, x_{j+1}, \ldots x_K \text{ fixed.} \tag{4}$$

Note: if x_j is discrete, partial effects are computed by comparing $\mathrm{E}\,(y|\mathbf{x})$ at different values of x_j holding the other variables fixed.

The partial effects in the purely linear model

$$\mathsf{E}\left(y|x_1,x_2\right)=\mu\left(x_1,x_2\right)=\beta_0+\beta_1x_1+\beta_2x_2$$

do not depend on the values of the regressors:

$$\frac{\partial \mathsf{E}(y|\mathbf{x})}{\partial x_i}=\beta_i \qquad \Rightarrow \qquad \mathsf{E}(dy|\mathbf{x})=\beta_i dx_i.$$

For small (but not infinitesimally small) changes in x_i we use the approximation

$$\mathsf{E}(\Delta y|\mathbf{x})\approx\beta_i\Delta x_i.$$

The partial effects in the log-linear model

$$\mathsf{E}\left(y|x_1, x_2\right) = \mu\left(x_1, x_2\right) = \beta_0 + \beta_1 \log(x_1) + \beta_2 \log(x_2)$$

depend on the value of the variable of interest:

$$\frac{\partial \mathsf{E}(y|\mathbf{x})}{\partial x_i} = \beta_i / x_i \qquad \Rightarrow \qquad \mathsf{E}(dy|\mathbf{x}) = \beta_i \frac{dx_i}{x_i}.$$

For small (but not infinitesimally small) changes in x_i we use the approximation

$$\mathsf{E}(\Delta y|\mathbf{x}) \approx \beta_i \frac{\Delta x_i}{x_i}.$$

Note that $100\frac{\Delta x_i}{x_i}$ is a percent change and β_i is called a semi-elasticity.

Interpretation: an increase in x_i by 1 percent ($\frac{\Delta x_i}{x_i} = 0.01$) leads to an expected change in y by $0.01\beta_i$ units.

The partial effects in the log-log model

$$\mathsf{E}\left(\log(y)|x_1, x_2\right) = \mu\left(x_1, x_2\right) = \beta_0 + \beta_1 \log(x_1) + \beta_2 \log(x_2)$$

depend on the values of the variable of interest and the lhs variable:

$$\frac{\partial \mathsf{E}(\log(y)|\mathbf{x})}{\partial x_i} = \beta_i / x_i \qquad \Rightarrow \qquad \mathsf{E}(d\log(y)|\mathbf{x}) = \mathsf{E}(dy/y|\mathbf{x}) = \beta_i \frac{dx_i}{x_i}.$$

For small (but not infinitesimally small) changes in x_i we use the approximation

$$\mathsf{E}\left(\frac{\Delta y}{y}|\mathbf{x}\right) \approx \beta_i \frac{\Delta x_i}{x_i}.$$

Note β_i is called an elasticity.

Interpretation: an increase in x_i by 1 percent ($\frac{\Delta x_i}{x_i} = 0.01$) leads to an expected change in y by β_i percent.

The partial effects in the interaction model

$$\mathsf{E}\left(y|x_1, x_2\right) = \mu\left(x_1, x_2\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

depend on the value of another variable, e.g.:

$$\frac{\partial \mathsf{E}(y|\mathbf{x})}{\partial x_1} = \beta_1 + \beta_3 x_2.$$

Sometimes x_2 is called a modifier of the partial effect of x_1 on y.

The error form of models of conditional expectations

When y is a random variable we would like to explain in terms of observable variables $\mathbf{x}$, it is useful to decompose y as

$$y = \mathrm{E}(y|\mathbf{x}) + u \tag{5}$$

with

$$\mathrm{E}(u|\mathbf{x}) = 0. \tag{6}$$

We can *always* write y as its conditional expectation, $\mathrm{E}(y|\mathbf{x})$, plus an **error term** or **disturbance term** that has *conditional* mean zero. In fact, thereby we define u.

Two important properties of u follow from $\mathrm{E}(u|\mathbf{x}) = 0$:

1. $E(u) = 0$ (by the law of iterated expectations, see below)
2. u is uncorrelated with any function of $\mathbf{x}$. In particular, it is uncorrelated with each of $x_1, \ldots, x_K$. Proof: Wooldridge, p. 31.

The error form allows us two write down population models in the usual way:

Consider again the conditional expectation

$$\mathsf{E}\left(y|x_1, x_2\right) = \mu\left(x_1, x_2\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$

The corresponding model in error form is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u, \qquad \mathsf{E}(u|x_1, x_2) = 0.$$

The law of iterated expectations

Simple form

Suppose y is a random variable and $\mathbf{x}$ is a random vector. The simple form of the law of iterated expectations (LIE) is

$$\mathrm{E}\left[\mathrm{E}(y|\mathbf{x})\right] = \mathrm{E}(y).$$

The proof can be found in any textbook on statistics.

The LIE can be used to directly prove that $\mathrm{E}(u|\mathbf{x}) = 0$ implies $\mathrm{E}(u) = 0$:

$$\mathrm{E}(u) = \mathrm{E}\left[\mathrm{E}(u|\mathbf{x})\right] = \mathrm{E}\left[0\right] = 0.$$

Suppose y is a random variable, $\mathbf{w}$ is a random vector and $\mathbf{x} = \mathbf{f}(\mathbf{w})$ is a function of $\mathbf{w}$ and thus also a random vector. For example, the vector $\mathbf{x}$ could be a subset of $\mathbf{w}$.

A general formulation of the LIE is

$$\mathrm{E}(y|\mathbf{x}) = \mathrm{E}\left[\mathrm{E}\left(y|\mathbf{w}\right)|\mathbf{x}\right]. \tag{7}$$

In addition, it holds that

$$\mathrm{E}(y|\mathbf{x}) = \mathrm{E}\left[\mathrm{E}\left(y|\mathbf{x}\right)|\mathbf{w}\right]. \tag{8}$$

Both statements are variants of the general rule: "The smaller information set always dominates."

Why is the LIE useful?

Let us apply the LIE to the case where $\mathbf{w} = \{\mathbf{x}, \mathbf{z}\}$. Then

$$\mathrm{E}\left[\mathrm{E}\left(y|\mathbf{x}, \mathbf{z}\right)|\mathbf{x}\right] = \mathrm{E}(y|\mathbf{x}).$$

This can be very helpful when our theoretical model includes variables $\mathbf{z}$ that are unobservable.

For example, consider the model

$$\mathrm{E}\left(y|x_1, x_2, z\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 z \tag{9}$$

but where z is unobserved.

How can we hope to estimate β_1 and β_2?

We need to find the conditional expectation solely with respect to x_1 and x_2.

By the LIE, and the linearity of the CE operator,

$$\begin{aligned} \mathsf{E}(y|x_1, x_2) &= \mathsf{E}(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 z \,|\, x_1, x_2) \\ &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 \mathsf{E}(z|x_1, x_2). \end{aligned} \tag{10}$$

So far, nothing is won. But now let us make a (hopefully economically justified) assumption about $\mathsf{E}(z|x_1, x_2)$. For example, it might be linear in x_1 and x_2:

$$\mathsf{E}(z|x_1, x_2) = \delta_0 + \delta_1 x_1 + \delta_2 x_2. \tag{11}$$

Then we can plug this into the previous equation and rearrange:

$$\begin{aligned} \mathsf{E}(y|x_1, x_2) &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (\delta_0 + \delta_1 x_1 + \delta_2 x_2) \\ &= (\beta_0 + \beta_3 \delta_0) + (\beta_1 + \beta_3 \delta_1) x_1 + (\beta_2 + \beta_3 \delta_2) x_2 \end{aligned}$$

This last expression is linear in (x_1, x_2) and estimable based on data for (x_1, x_2).

Without further restrictions, however, we cannot estimate β_1 and β_2. This is a variant of the *identification problem* in econometrics which we cover later.

Statistical independence and conditional expectation

Rule:

If $\mathrm{E}(u) = 0$ and u and $\mathbf{x}$ are independent, then $\mathrm{E}(u|\mathbf{x}) = 0$.

The converse is not true:

If $\mathrm{E}(u|\mathbf{x}) = 0$, then $\mathrm{E}(u) = 0$ and u and $\mathbf{x}$ are uncorrelated but possibly dependent (for example, their variances might be related).

6. Linear Projections

OLS versus linear projections

Recall that we use OLS to find the best linear relationship between data $\mathbf{Y}$ and $\mathbf{X}$, where best means that the squared distance

$$S(\hat{\boldsymbol{\beta}}) = (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

is minimized. Important: this is a linear relationship within a *sample*.

Finding a linear relationship between the random variable y and the $(1 \times K)$-dimensional random vector $\mathbf{x}$ is also interesting in *population*. This is achieved by minimizing the expected squared distance

$$\min_{b_0, \mathbf{b} \in \mathbb{R}^K} \mathrm{E}\left[(y - b_0 - \mathbf{x}\mathbf{b})^2\right]. \tag{12}$$

This is called a *linear projection* (or *minimum mean square linear predictor* or *least squares linear predictor*).

Hence, OLS is the sample analogue of linear prediction which is about the population.

Finding the parameters of the linear projection*

To minimize

$$\mathrm{E}\left[(y - b_0 - \mathbf{x}\mathbf{b})^2\right]$$

take the first derivatives,

$$\frac{\partial \mathrm{E}\left[(y - b_0 - \mathbf{x}\mathbf{b})^2\right]}{\partial b_0} = -2\,\mathrm{E}\left[y - b_0 - \mathbf{x}\mathbf{b}\right]$$
$$\frac{\partial \mathrm{E}\left[(y - b_0 - \mathbf{x}\mathbf{b})^2\right]}{\partial \mathbf{b}} = -2\,\mathrm{E}\left[(y - b_0 - \mathbf{x}\mathbf{b})\,\mathbf{x}'\right],$$

and equate them to zero,

$$\mathrm{E}\left[y - \beta_0 - \mathbf{x}\boldsymbol{\beta}\right] \stackrel{!}{=} 0$$
$$\mathrm{E}\left[(y - \beta_0 - \mathbf{x}\boldsymbol{\beta})\,\mathbf{x}'\right] \stackrel{!}{=} 0,$$

denoting the solutions to b_0 and $\mathbf{b}$ by β_0 and $\boldsymbol{\beta}$, respectively.

Finding the parameters of the linear projection*

For β_0 we obtain

$$\mathrm{E}[y] - \beta_0 - \mathrm{E}[\mathbf{x}]\boldsymbol{\beta} = 0 \quad \Rightarrow \quad \beta_0 = \mathrm{E}[y] - \mathrm{E}[\mathbf{x}]\boldsymbol{\beta}$$

and for $\boldsymbol{\beta}$, using the solution for β_0,

$$\begin{aligned}
0 &= \mathrm{E}\left[(y - \beta_0 - \mathbf{x}\boldsymbol{\beta})\,\mathbf{x}'\right] \\
0 &= \mathrm{E}\left[\mathbf{x}'y - \mathbf{x}'\beta_0 - \mathbf{x}'\mathbf{x}\boldsymbol{\beta}\right] \\
0 &= \mathrm{E}[\mathbf{x}'y] - \mathrm{E}[\mathbf{x}']\beta_0 - \mathrm{E}[\mathbf{x}'\mathbf{x}]\boldsymbol{\beta} \\
0 &= \mathrm{E}[\mathbf{x}'y] - \mathrm{E}[\mathbf{x}']\left(\mathrm{E}[y] - \mathrm{E}[\mathbf{x}]\boldsymbol{\beta}\right) - \mathrm{E}[\mathbf{x}'\mathbf{x}]\boldsymbol{\beta} \\
0 &= \mathrm{E}[\mathbf{x}'y] - \mathrm{E}[\mathbf{x}']\,\mathrm{E}[y] - \mathrm{E}[\mathbf{x}'\mathbf{x}]\boldsymbol{\beta} + \mathrm{E}[\mathbf{x}']\,\mathrm{E}[\mathbf{x}]\boldsymbol{\beta} \\
0 &= \left(\mathrm{E}[\mathbf{x}'y] - \mathrm{E}[\mathbf{x}']\,\mathrm{E}[y]\right) - \left(\mathrm{E}[\mathbf{x}'\mathbf{x}] - \mathrm{E}[\mathbf{x}']\,\mathrm{E}[\mathbf{x}]\right)\boldsymbol{\beta} \\
0 &= \mathrm{Cov}(\mathbf{x}', y) - \mathrm{Var}(\mathbf{x})\boldsymbol{\beta} \\
\boldsymbol{\beta} &= [\mathrm{Var}(\mathbf{x})]^{-1}\,\mathrm{Cov}(\mathbf{x}', y)
\end{aligned}$$

Interpretation of linear projections

Assume in population there is some possibly nonlinear relationship between y and $\mathbf{x}$.

Generally, we are interested in the conditional expectation

$$\mathrm{E}(y|\mathbf{x})$$

but it may be difficult to estimate or is just unknown. Therefore, we may focus on the linear projection

$$L(y|1,\mathbf{x}) = L(y|1,x_1,\ldots,x_K) = \beta_0 + \mathbf{x}\boldsymbol{\beta}$$

as a kind of linear approximation in population.

Note that the linear projection is identical to the conditional expectation whenever the conditional expectation is linear in $\mathbf{x}$.

OLS versus linear projection

Recall: The linear projection minimizes the expected squared deviations from the linear relationship between y and $\mathbf{x}$ while OLS minimizes the sample average of the squared deviations from the linear relationship between the observations y_i and $\mathbf{x}_i$, $i = 1, \ldots, N$. Hence, the OLS objective function is the sample analogue of the linear projection objective function.

Therefore, it is not surprising that the linear projection parameters

$$\beta_0 = \mathrm{E}[y] - \mathrm{E}[\mathbf{x}]\boldsymbol{\beta} \quad \text{and} \quad \boldsymbol{\beta} = [\mathrm{Var}(\mathbf{x})]^{-1}\,\mathrm{Cov}(\mathbf{x}', y)$$

and the OLS parameters

$$\hat{\beta}_{0,OLS} = \bar{y} - \bar{\mathbf{x}}\hat{\boldsymbol{\beta}}_{OLS} \quad \text{and} \quad \hat{\boldsymbol{\beta}}_{OLS} = [s_{\mathbf{x}}^2]^{-1} s_{\mathbf{x}',y}.$$

are in the same relationship:

The OLS parameters are the sample analogues of the linear projection parameters.

OLS versus linear projection

The relationship between OLS and linear projection becomes even more obvious when we define

$$\mathbf{x} = [1, x_1, \ldots, x_K].$$

Then the so-defined parameter vector $\boldsymbol{\beta}$ of the linear projection

$$L(y|\mathbf{x}) = L(y|1, x_1, \ldots, x_K) = \mathbf{x}\boldsymbol{\beta}$$

turns out to be

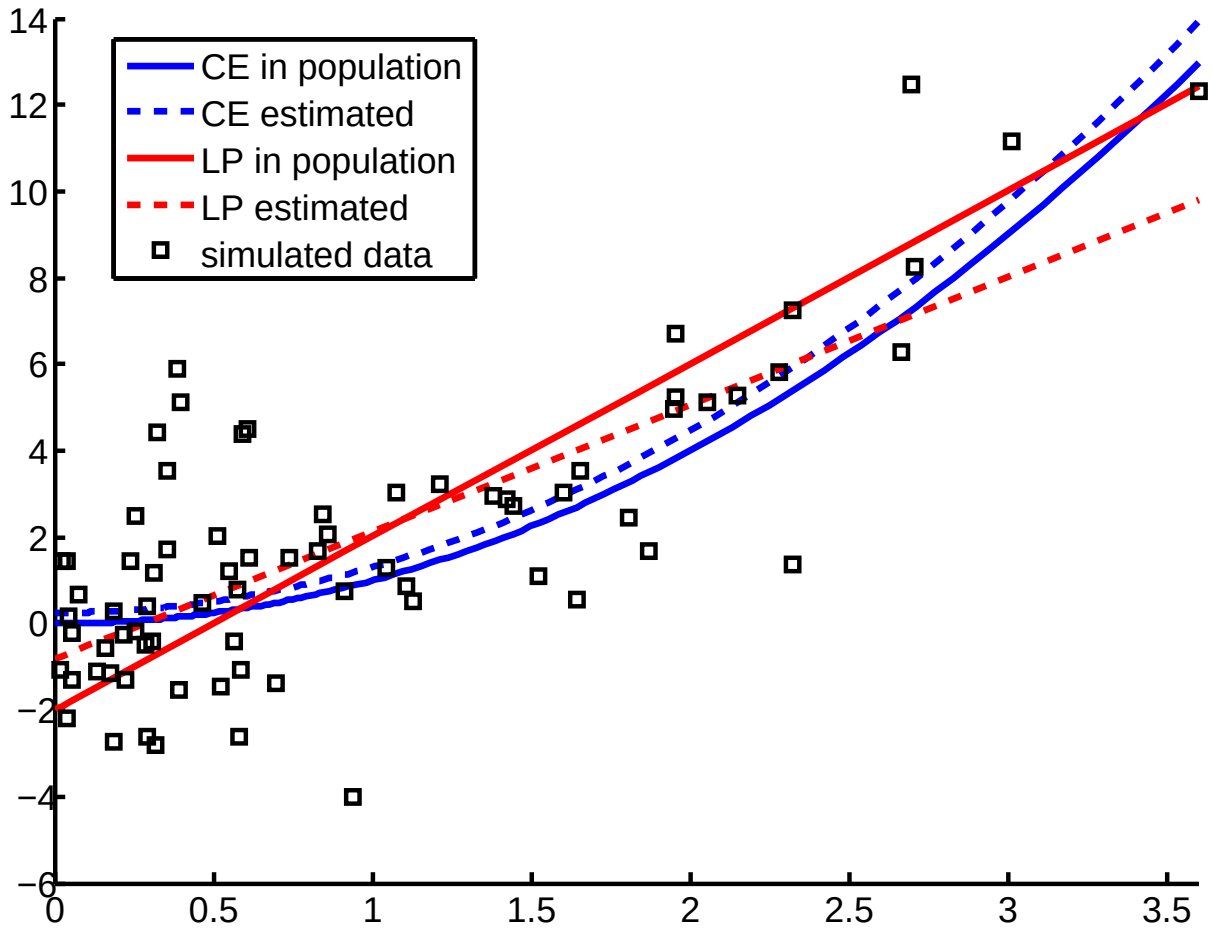
$$\boldsymbol{\beta} = [\mathrm{E}(\mathbf{x}'\mathbf{x})]^{-1}\,\mathrm{E}(\mathbf{x}'y).$$

The obvious sample counterpart is the OLS estimator

$$\hat{\boldsymbol{\beta}}_{LS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \left(N^{-1}\sum_{i=1}^{N}\mathbf{x}_i'\mathbf{x}_i\right)^{-1}\left(N^{-1}\sum_{i=1}^{N}\mathbf{x}_i'y_i\right).$$

Example

Model: $y = x^2 + u, x \sim exp(1), u \sim N(0,4) \quad \Rightarrow \quad E(y|x) = x^2, \quad L(y|1,x) = 4x - 2$



As for the conditional expectation, we can write the linear projection in error form

$$y = L(y|1, \mathbf{x}) + u = \beta_0 + \mathbf{x}\boldsymbol{\beta} + u.$$

By construction of the linear projection, the error u has the following properties:

1. It has mean zero: $E(u) = 0$.
2. It is uncorrelated with $\mathbf{x}$: $\text{Cov}(\mathbf{x}, u) = 0$.

Property #1:

Taking expectations of the linear projections in error form yields

$$\mathsf{E}(y) = \beta_0 + \mathsf{E}(\mathbf{x})\boldsymbol{\beta} + \mathsf{E}(u).$$

Comparing this with the first-order condition for β_0,

$$\beta_0 = \mathsf{E}[y] - \mathsf{E}[\mathbf{x}]\boldsymbol{\beta} \quad \Rightarrow \quad \mathsf{E}(y) = \beta_0 + \mathsf{E}(\mathbf{x})\boldsymbol{\beta}$$

shows that

$$\mathsf{E}(u) = 0.$$

The error form for linear projections

Proof*

Property #2:

Starting from the error form $y = \beta_0 + \mathbf{x}\boldsymbol{\beta} + u$ and subtracting its expectation $\mathrm{E}(y) = \beta_0 + \mathrm{E}(\mathbf{x})\boldsymbol{\beta}$ yields

$$y - \mathrm{E}(y) = [\mathbf{x} - \mathrm{E}(\mathbf{x})]\boldsymbol{\beta} + u.$$

Multiply from the left by $[\mathbf{x} - \mathrm{E}(\mathbf{x})]'$ and take expectations:

$$\underbrace{\mathrm{E}\left\{[\mathbf{x} - \mathrm{E}(\mathbf{x})]'[y - \mathrm{E}(y)]\right\}}_{\mathrm{Cov}(\mathbf{x}', y)} = \underbrace{\mathrm{E}\left\{[\mathbf{x} - \mathrm{E}(\mathbf{x})]'[\mathbf{x} - \mathrm{E}(\mathbf{x})]\right\}}_{\mathrm{Var}(\mathbf{x})}\boldsymbol{\beta} + \underbrace{\mathrm{E}\left\{[\mathbf{x} - \mathrm{E}(\mathbf{x})]'u\right\}}_{\mathrm{Cov}(\mathbf{x}', u)}$$

Using the definition of $\boldsymbol{\beta}$ yields:

$$\mathrm{Cov}(\mathbf{x}', y) = \mathrm{Var}(\mathbf{x})\underbrace{[\mathrm{Var}(\mathbf{x})]^{-1}\mathrm{Cov}(\mathbf{x}', y)}_{\boldsymbol{\beta}} + \mathrm{Cov}(\mathbf{x}', u)$$

and thus

$$\mathrm{Cov}(\mathbf{x}', y) = \mathrm{Cov}(\mathbf{x}', y) + \mathrm{Cov}(\mathbf{x}', u) \quad \Rightarrow \quad \mathrm{Cov}(\mathbf{x}', u) = 0.$$