Introduction to Econometrics

(Outline of lectures)

Andrea Ichino
University of Bologna and Cepr

April 29, 2009

Contents

1 Introduction

The scope of econometrics:

- To design and estimate statistical models of relationships between socio-economic variables.
- To establish under what conditions these relationships have a causal interpretation.

Some examples from Wooldridge-Chapter 1. and ... your own research work:

- Education and earnings
- Law enforcement and city crime levels
- Fertilizer and Crop Yield
- Minimum wage and unemployment
- Job training and productivity
- ...

1.1 The tool-box of econometrics

i. A well defined question and the population for which it matters.

ii. The ideal experiment we would like to run in order to answer the question.

iii. A feasible strategy to address the question in the absence of the ideal experiment.

iv. An accessible sample of data from the population of interest:
   - Cross-sectional data
   - Time-series data
   - Panel data
   - Examples from Wolrdridge-Chapter 1

v. The model of statistical inference (Rubin, 1991): how to infer from the sample the population relationship between variables in which we are interested.

1.2 The econometric sequence at the LMEC

This initial course is devoted to the most standard tools of econometrics.

- The simple regression model;
- Multiple regression analysis.

Then the sequence splits between:

- *Time-series-econometrics*:
  two courses devoted to the study of models for time series data and panel data “with large t and small n”.
- *Micro-econometrics*:
  two courses devoted to the study of models for cross-sectional and panel data “with small t and large n”.

The last course in the micro-econometric sequence is specifically dedicated to methods for the identification and estimation of causal relationships.

2 The simple regression model

Consider:

- an outcome variable y: e.g. *labor earnings*;
- a variable x which we consider as a possible determinant of y in which we are interested: e.g. *years of education*;
- a variable e which describes all the other determinants of y that we do not observe.

The general notation for the model that relates y, x and e is

$$y = f(x, e) \tag{1}$$

We are interested in the relationship between x and y in the population, which we can study from two perspectives:

i. To what extent knowing x allows us to "predict something" about y.

ii. Whether Δx "causes" Δy given a proper definition of causality.

2.1 Regression and the Conditional Expectation Function

We deviate slightly from Wooldridge and follow Angrist and Pischke (2008) to show that *Regression*, independently of causality, is a useful tool within the first perspective because of its link with the *Conditional Expectation Function*.

We can always decompose 1 in the following way:

$$y = E(y|x) + \epsilon \tag{2}$$

where $E(y|x)$ is the Conditional Expectation Function (CEF) of y given x, and $\epsilon = y - E(y|x)$ is by construction:

- mean independent of x:

$$E(\epsilon|x) = E(y - E(y|x)|x) = E(y|x) - E(y|x) = 0 \tag{3}$$

- is uncorrelated with any function of x, i.e. for any h:

$$E(h(x)\epsilon) = E(h(x)E(\epsilon|x)) = 0 \tag{4}$$

Here is an example of the CEF of labor earnings given education in the US.

Properties of the Conditional Expectation Function

The CEF provides useful information because of some interesting properties.

Property 1. *Let $m(x)$ be any function of x. The CEF solves*

$$E(y|x) = arg\min_{m(.)} E\left[(y - m(x))^2\right] \tag{5}$$

The CEF minimizes the Mean Square Error of the prediction of Y given X.

Property 2.

$$\begin{aligned} V(y) &= V(E(y|x)) + V(\epsilon) \\ &= V(E(y|x)) + E(V(y|x)) \end{aligned} \tag{6}$$

The variance of y can be decomposed in the variances of the CEF and of ϵ.

Exercise: prove the two properties.

2.2 The Population Regression Function

We do not know the CEF but we can show that the Population Regression Function (PRF) is a “good” approximation to the true CEF.

The PRF is the linear function

$$y_p = \beta_0 + \beta_1 x \tag{7}$$

such that β_0 and β_1 minimize the square of the residual distance $u = y - y_p$ in the population, i.e. the “distance” between y and the PRF line itself:

$$(\beta_0, \beta_1) = \text{arg}\min_{b_0, b_1} E\left[(y - b_0 - b_1 x)^2\right] \tag{8}$$

The First Order conditions of problem 8 are:

$$\begin{aligned} E\left[x(y - b_0 - b_1 x)\right] &= 0 \\ E\left[(y - b_0 - b_1 x)\right] &= 0 \end{aligned} \tag{9}$$

The parameters of the Population Regression Function

The solutions are:

$$\beta_1 = \frac{E[x(y-\beta_0)]}{E(x^2)} = \frac{Cov(y,x)}{V(x)} \tag{10}$$

$$\beta_0 = E(y) - \beta_1 E(x) \tag{11}$$

Note that by definition of β_0 and β_1:

$$y = y_p + u = \beta_0 + \beta_1 x + u \tag{12}$$

and

$$E(xu) = E[x(y - \beta_0 - \beta_1 x)] = 0 \tag{13}$$

In words, the PRF is the linear function of x that makes the residuals u uncorrelated with x in the population.

Properties of the Population Regression Function

The PRF is linked to the CEF by some interesting properties:

Property 3. *If the CEF is linear then the PRF is the CEF. This happens, specifically:*

- *when y and x are jointly normally distributed;*
- *in a fully saturated model (to be defined below in the context of multiple regression)*

Property 4. *The PRF is the best linear predictor of y in the sense that it minimizes the Mean Square Error of the prediction.*

Property 5. *The PRF is the best linear approximation to the CEF in the sense that it minimizes the Mean Square Error of the approximation.*

Exercise: prove these properties.

Parenthesis: an informative exercise

Take any dataset and assume that this is your entire population

Define the variables of interest y and x.

Estimate the linear regression of y on x.

Compute $\bar{y} = E(y|x)$ and estimate the linear regression $\bar{y}$ on x.

Compare the results of the two estimations and comment on your findings.

In which sense the properties of the CEF and the PRF are relevant for your findings?

Could this result be useful whenever data providers do not want to release individual observations?

What have we accomplished so far by this way of reasoning?

If we are simply interested in predicting y given x it would be useful to know the correspondent CEF because of its properties.

We do not know the CEF but the PRF is the best linear approximation to the CEF and the best linear predictor of y given x .

If we had data for the entire population we could then use the PRF, which we can characterize precisely, to predict y given x.

Usually, however, we have (at best) a random sample of the population.

We now have to show that the Sample Regression Function (SRF) is a “good” estimate of the Population Regression Function according to some criteria.

This is an inference problem.

“Repetita juvant”: again on the orthogonality condition

By saying that our goal is to estimate the PRF and that the PRF is defined as:

$$y_p = \beta_0 + \beta_1 x \tag{14}$$

where the parameters satisfy by construction:

$$(\beta_0, \beta_1) = arg \min_{b_0, b_1} E\left[(y - b_0 - b_1 x)^2\right] \tag{15}$$

the condition

$$E(xu) = E[x(y - \beta_0 - \beta_1 x)] = 0 \tag{16}$$

is not a necessary assumption for regression to make sense (as in standard econometrics): it follows instead from the definition β_0 and β_1 and, as we will see below, it ensures that:

- The OLS-MM estimator is by definition “consistent for the PRF”;
- and unbiased in some important special cases.

Note: at this stage the PRF does not have a causal interpretation, which requires a definition of causality and assumptions that will be discussed in Section 2.7.

2.3 From the Sample Regression Function to the Population Regression Function

Now suppose that we have a random sample of the population

Definition 1. *If* $\{z_1 ... z_i ... z_n\}$ *are independent draws from a population with density function f(z,θ), then* $\{z_1 ... z_i ... z_n\}$ *is a random sample from the population defined by f(z,θ). Note that each draw is a random variable.*

Exercise: make sure that you understand the meaning of random sampling.

We want to know whether the sample analogs of

$$\beta_1 = \frac{Cov(y,x)}{V(x)} \quad \text{and} \quad \beta_0 = E(y) - \beta_1 E(x) \tag{17}$$

which (denoting sample averages with $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$) are:

$$\hat{\beta}_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad \text{and} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \tag{18}$$

can be considered as "good" estimators of β_1 and β_0 under some criteria to be defined.

Why should we focus on the sample analog of β_1 (or β_0)?

An “estimator” is a function (a “recipe”) of the sample of data which originates an “estimate” (a “cake”) when the actual sample draws (the ”ingredients”) are combined in the way suggested by the estimator.

The “quality” of the estimate (the “cake”) depends on the properties of the estimator (the “recipe”) and on the caracteristics of the actual sample (the “ingredients”).:

Before analysing the properties of $\hat{\beta}_1$ let’s consider three justifications for thinking about this specific recipe among the many we could have considered.

The “cakes” we would like to obtain are the parameters β_0 and β_1 of the PRF defined as the fitted line that minimizes residuals from y in the population.

We are asking whether the slope $\hat{\beta}_1$ of the sample fitted line (SRF) “approaches” the “cake we would like to have”, which is β_1. (Same for β_0)

2.3.1 The “Method of Moment” justification of $\hat{\beta}_0$ and $\hat{\beta}_1$

The “Methods of Moments” constructs estimators on the basis of restrictions concerning moments of the population that should be satisfied also in the sample (under random sampling), given the definition of the parameters to be estimated.

The definition of the PRF parameters that we have given implies that the following two moment conditions should hold in the data

$$E(u) = E\left[y - \beta_0 - \beta_1 x\right] = 0 \tag{19}$$

$$E(xu) = E\left[x(y - \beta_0 - \beta_1 x)\right] = 0 \tag{20}$$

Given random sampling (i.e. if the sample is a scaled down but perfect image of the population), these two conditions should hold also in the sample.

The moment conditions in the sample

The analogs of the population moment conditions in the sample are:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^{n} x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \tag{21}$$

With simple algebra in Wooldridge-Chapter2 one can derive the Method of Moment estimators for β_1 and β_0:

$$\hat{\beta}_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad \text{and} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \tag{22}$$

Pay attention to an important necessary condition for the construction of these estimators

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 > 0 \tag{23}$$

What does this mean for your research question and your empirical work?

2.3.2 The "Least Squares" justification of $\hat{\beta}_0$ and $\hat{\beta}_1$

An equivalent justification of $\hat{\beta}_0$ and $\hat{\beta}_1$ is that they should be chosen in a way such that the SRF minimizes the sum of squared residuals in the sample, i.e. the distance between the sample observations and the SRF itself.

The PRF minimizes the sum of squared residual in the population, which suggests that it might be good if the SRF achieves the same result in the sample

The Ordinary Least Square estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are constructed as

$$(\hat{\beta}_0, \hat{\beta}_1) = arg \min_{\hat{b}_0, \hat{b}_1} \sum_{i=1}^{n} \left[(y_i - \hat{b}_0 - \hat{b}_1 x_i)^2 \right] \tag{24}$$

It is easy to check that the FOCs of this problem are *de facto* identical to 21:

$$\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \quad \text{and} \quad \sum_{i=1}^{n} x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \tag{25}$$

The OLS estimators

Since the OLS conditions 25 and the MM conditions 21 are the same, they deliver the same estimators:

$$\hat{\beta}_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad \text{and} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}, \tag{26}$$

The second order conditions of the minimization problem 24 are satisfied.

The way to do it (see Woodridge-Appendix 2A) is to add and subtract $\hat{\beta}_0 + \hat{\beta}_1 x_i$ within the squared parentheses in the minimand 24 to get

$$\sum_{i=1}^{n}\left[(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) + (\hat{\beta}_0 - \hat{b}_0) + (\hat{\beta}_1 x_i - \hat{b}_1 x_i)\right]^2 \tag{27}$$

Developping the square one can show that the minimum occurs for $\hat{b}_0 = \hat{\beta}_0$ and $\hat{b}_1 = \hat{\beta}_1$.

2.3.3 The “Maximum Likelihood” justification of $\hat{\beta}_0$ and $\hat{\beta}_1$ (for future reference)

There is a third way to justify the $\hat{\beta}_0$ and $\hat{\beta}_1$ estimators based on the logic of Maximum Likelihood (ML).

This justification requires the assumption that y is distributed normally.

Thanks to this distributional assumption, in addition to the MM and OLS desirable properties that we will discuss below, $\hat{\beta}_0$ and $\hat{\beta}_1$ acquire also the properties of ML estimators.

We will discuss the additional properties of ML estimators later.

Now we just want to show that $\hat{\beta}_0$ and $\hat{\beta}_1$ can also be interpreted as ML estimates, under the assumption of normality.

The likelihood function

Consider the model

$$y_i = \beta_0 + \beta_1 x_i + u_i \tag{28}$$

and suppose that :

$$u_i \sim f(u_i|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(u_i)^2}{2\sigma^2}} \tag{29}$$

which implies

$$y_i \sim f(y_i|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}} \tag{30}$$

Assuming an observed set of independent sample draws, the likelihood function is defined as:

$$L(y|x, \beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}} \tag{31}$$

Given a sample of observations y_i and x_i, L is the probability of observing the sample given the parameters β_0, β_1 and σ^2.

The “Recipe” of maximum likelihood estimation

The ML estimator (the “recipe”) chooses $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\sigma}^2$ as the values of β_0, β_1 and σ^2 that maximize the likelihood, given the observed sample.

$$\{\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2\} = \text{arg} \min_{\beta_0,\beta_1,\sigma^2} L(y|x, \beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-\beta_0-\beta_1 x_i)^2}{2\sigma^2}} \tag{32}$$

Computations simplify if we maximize the log likelihood:

$$Log[\ L\ (y|x, \beta_0, \beta_1, \sigma^2)] = \sum_{i=1}^{n} log \left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-\beta_0-\beta_1 x_i)^2}{2\sigma^2}} \right] \tag{33}$$

$$= -\frac{N}{2} log(2\pi) - \frac{N}{2} log(\sigma^2) - \frac{1}{2} \sum_{i=1}^{n} \frac{(y_i - \beta_0 - \beta_1 x_i)^2}{\sigma^2} \tag{34}$$

First Order Conditions for β_0 and β_1

Maximization of the log likelihood with respect to β_0 and β_1 implies that:

$$(\hat{\beta}_0, \hat{\beta}_1) = arg \min_{\beta_0, \beta_1} \sum_{i=1}^{n} \left[(y_i - \beta_0 - \beta_1 x_i)^2 \right] \tag{35}$$

The FOC's are identical for the ML, MM and OLS problems (see 21 and 25):

$$\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \quad \text{and} \quad \sum_{i=1}^{n} x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \tag{36}$$

Solving the FOC we get the same estimator:

$$\hat{\beta}_1 = \frac{\frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2} \quad \text{and} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \tag{37}$$

And Second Order Conditions can be checked as in the OLS problem.

We defer the analysis of the additional properties of ML estimators to later.

2.4 Algebraic and geometric properties of the OLS-MM estimators $\hat{\beta}_0$ and $\hat{\beta}_1$

- The Sample Regression Function is the set of the *fitted values*

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \tag{38}$$

- The estimated sample residuals $\hat{u} = y - \hat{y}$ satisfy:

$$\sum_{i=1}^{n} \hat{u}_i = 0 \tag{39}$$

$$\sum_{i=1}^{n} x_i \hat{u}_i = 0 \tag{40}$$

$$\sum_{i=1}^{n} (\hat{y}_i - \bar{y}) \hat{u}_i = 0 \tag{41}$$

- A geometric interpretation (see the figure drawn in class) of the OLS-MM orthogonal decomposition

$$y = \hat{y} + \hat{u} \tag{42}$$

A decomposition of the total variation of y_i

The OLS-MM estimator decomposes the total variation of y_i into a component explained by x_i and a residual unexplained component.

$$SST = \text{Total Sum of Squares } = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{43}$$

$$SSE = \text{Explained Sum of Squares } = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \tag{44}$$

$$SSR = \text{Residual Sum of Squares } = \sum_{i=1}^{n} \hat{u}^2 \tag{45}$$

$$SST = SSE + SSR \tag{46}$$

The proof is easy, developping the square in SST and using 41.

2.5 Goodness of fit and the R-squared

Assuming variability in the sample ($SST \neq 0$), the R-Squared is defined as

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST} \tag{47}$$

which takes values between 0 and 1.

The R-squared measures the proportion of the total variation of y that is explained by x.

It is also a measure of the goodness of fit of the model.

While a low R-squared may appear to be a “bad sign”, we will show later that x may still be a very important determinant of y even if the R-squared is low.

2.6 Three desirable statistical properties of the OLS-MM estimators $\hat{\beta}_0$ and $\hat{\beta}_1$

One can think of several properties that an estimator (a "recipe") should have in order to produce satisfactory estimates ("cakes").

At this stage we focus on three of these possible properties.

Note that the estimate is a random variable, because it is a function of the sample observations which are random variables.

The desirable properties are:

i. Unbiasedness;

ii. Consistency;

iii. Efficiency.

2.6.1 Are $\hat{\beta}_0$ and $\hat{\beta}_1$ unbiased for β_0 and β_1 ?

An estimator of some population parameter is unbiased when its expected value is equal to the population parameter that it should estimate.

The crucial population parameter of interest is the slope of the PRF.

We want to prove that:

$$E(\hat{\beta}_1|\{x_i\}) \equiv E\left(\frac{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})}{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}|\{x_i\}\right) = \frac{Cov(y,x)}{V(x)} \equiv \beta \quad (48)$$

To prove this result we need 4 assumptions, three of which have already been introduced.

Angrist and Pischke (2008) implicitly note, however, that unbiasedness is not so crucial and we should care for consistency, which (as we will see) does not require the fourth assumption.

The necessary assumptions for the proof

SLR 1: In the population, y is related to x and u as:

$$y = \beta_0 + \beta_1 x + u \tag{49}$$

SLR 2: The n observations y_i and x_i are a random sample of the population and the residual u_i is defined by:

$$y_i = \beta_0 + \beta_1 x_i + u_i \tag{50}$$

SLR 3: The observations $\{x_1, ..., x_n\}$ are not all equal

SLR 4: The residual u is mean-independent of x:

$$E(u|x) = 0 \tag{51}$$

Note that the definition of β_0 and β_1 in the PRF implies

$$E(ux) = 0 \tag{52}$$

but 52 does not imply 51 (while 51 implies 52).

In Section 2.6.2 we will show that SLR 4 is not needed for consistency, for which 52 is enough.

What is the “deep” meaning of SLR 4

Suppose that y is earnings, x is years of education and u is the effect of unobservable genetic ability A (and nothing else matters for earnings):

$$u = \gamma A \tag{53}$$

The assumption that

$$E(u|x) = 0 \tag{54}$$

means that the expected effect of genetic ability on earnings is the same *at each given level of education*.

The assumption is not satisfied in cases like the following:

- All subjects have the same ability A, but ability has a stronger effect on earnings at higher education levels: $\gamma > 0$ grows with x;
- A unit of ability A has the same effect γ on earnings for everybody, but subjects with higher education have more ability: $A > 0$ and grows with x.

Proof of the unbiasedness of the OLS-MM estimator $\hat{\beta}_1$

Note first that SLR 3 is needed otherwise $\hat{\beta}_1$ would not exist.

It is then useful to consider the following general result which is easy to verify for any random variables z_i and w_i:

$$\sum_{i=1}^{n}(z_i - \bar{z})(w_i - \bar{w}) = \sum_{i=1}^{n} z_i(w_i - \bar{w}) = \sum_{i=1}^{n}(z_i - \bar{z})w_i \tag{55}$$

Note that this holds also when $z_i = w_i$.

Using 55, the fact that $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$, and SLR 1 and SLR 2 to substitute for y_i , we can rewrite $\hat{\beta}_1$ as:

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{i=1}^{n} y_i(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(\beta_0 + \beta_1 x_i + u_i)(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ &= \beta_1 + \frac{\sum_{i=1}^{n}(u_i)(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\end{aligned} \tag{56}$$

Proof of the unbiasedness of the OLS-MM estimator $\hat{\beta}_1$ (cont.)

Substituting 56 in 48 and defining the Total Sum of Squared deviation from the mean of x as

$$SST_x = \sum_{i=1}^{n}(x_i - \bar{x})^2 : \tag{57}$$

we obtain:

$$\begin{aligned} E(\hat{\beta}_1|\{x_i\}) &= E\left(\beta_1 + \frac{\sum_{i=1}^{n}(u_i)(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}|\{x_i\}\right) \\ &= \beta_1 + \frac{1}{SST_x}\left(\sum_{i=1}^{n} E[u_i(x_i - \bar{x})|\{x_i\}]\right) \\ &= \beta_1 + \frac{1}{SST_x}\left(\sum_{i=1}^{n}(x_i - \bar{x})E(u_i|\{x_i\})\right) \\ &= \beta_1 \end{aligned} \tag{58}$$

The last equality holds because of SLR 4 and random sampling.

A “situation” in which SLR 4 holds

Consider a population in which no one has taken any education and earnings are a constant plus the random effect of genetic ability.

$$y = \beta_0 + u \tag{59}$$

where (without loss of generality) $E(u) = 0$.

Extract two random samples from this population and give two different levels of education x_1 and x_2 to the two groups.

Since the two random samples are “representative images” of the population

$$E(u) = E(u|x = x_1) = E(u|x = x_2) = 0 \tag{60}$$

Randomized controlled experiments deliver the assumption SLR 4 that we need.

This is analogous to what Wooldridge characterizes as a situation in which x_i is $\textit{fixed in repeated samples}$.

The assumption also holds obviously in the case of non-stochastic x.

A special case: the CEF is the PRF

If y and x are jointly normally distributed:

$$E(y|x) = \beta_0 + \beta_1 x \tag{61}$$

and in this case the CEF is the PRF because the CEF is linear. In this case, by definition, $u = y - E(y|x)$ is such that:

$$E(u|x) = E(y - E(y|x)|x) = E(y|x) - E(y|x) = 0 \tag{62}$$

When the CEF is linear, SLR 4 is no longer an assumption, because the population parameters β_0 and β_1 that we want to estimate are the ones that ensure that this condition is satisfied.

This is the assumption of Galton's study of the intergenerational transmission of height, in which the word "Regression" was first used. In the regression:

$$h_s = \alpha + \beta h_f + \epsilon \tag{63}$$

where h_j is the height of generation j, Galton estimated that $\beta < 1$ which implies that the child of tall parents will not be as tall as they are, i.e. without new random shocks "height would regress to the mean" across generations.

Proof of the unbiasedness of the OLS-MM estimator $\hat{\beta}_0$ (cont.)

The proof of unbiasedness of $\hat{\beta}_0$ is straightforward. Taking the sample average of 50 we get that

$$\bar{y} = \beta_0 + \beta_1 \bar{x} + \bar{u} \tag{64}$$

Then, using 26

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \beta_0 + (\beta_1 - \hat{\beta}_1)\bar{x} + \bar{u} \tag{65}$$

And therefore:

$$\begin{aligned} E(\hat{\beta}_0|x) &= \beta_0 + E(\beta_1 - \hat{\beta}_1)\bar{x}|x) + E(\bar{u}|x) \\ &= \beta_0 \end{aligned} \tag{66}$$

because $E(\hat{\beta}_1|x) = E(\beta_1|x)$ and $E(\bar{u}|x) = 0$.

2.6.2 Are $\hat{\beta}_0$ and $\hat{\beta}_1$ consistent for β_0 and β_1 ?

An estimator of a population parameter is consistent when the estimates it produces can be made arbitraily close to the population parameter by increasing the sample size.

Formally, we say that $\hat{\beta}_1$ is consistent for β_1 if it *converges in probability* to β_1.

$$\text{plim } \hat{\beta}_1 = \beta_1 \tag{67}$$

and similarly for $\hat{\beta}_0$.

To prove consistency we need to use the

Proposition 1. *The Law of Large Numbers: Sample moments converge in probability to the corresponding population moments.*

For example, the probability that the sample mean is close to the population mean can be made as high as one likes by taking a large enough sample.

Properties of probability limits

Consider a random sample $z_1 ... z_n$

Property 1. *For any sample moment θ_n and continuous function $h(.)$:*

$$plim\ \theta_n = \theta \tag{68}$$

implies

$$plim\ h(\theta_n) = h(\theta) \tag{69}$$

Property 2. *Given two sample moments θ_n and ξ_n with*

$$plim\ \theta_n = \theta \tag{70}$$

$$plim\ \xi_n = \xi \tag{71}$$

we have,

$$plim\ (\theta_n + \xi_n) = \theta + \xi \tag{72}$$

$$plim\ (\theta_n \xi_n) = \theta\xi \tag{73}$$

$$plim\ \left(\frac{\theta_n}{\xi_n}\right) = \frac{\theta}{\xi} \tag{74}$$

Consistency of the OLS-MM estimator

Using 56 we can write:

$$\begin{aligned}\text{plim } \hat{\beta}_1 &= \text{plim } \left(\beta_1 + \frac{\sum_{i=1}^{n}(u_i)(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right) \\ &= \beta_1 + \frac{\text{plim } \left(\sum_{i=1}^{n}(u_i)(x_i - \bar{x})\right)}{\text{plim } \left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)} \\ &= \beta_1 + \frac{Cov(x,u)}{Var(x)} = \beta_1\end{aligned} \tag{75}$$

where the last equality derives from 16: x and u are uncorrelated because of the way we defined the PRF and the parameters that we want to estimate.

As a result the OLS-MM estimator may not be unbiased for the PRF (if $E(u|x) = 0$ does not hold) but is by definition consistent for the PRF.

For consistency we need only SLR 1 - SLR 3, but keep in mind that if the PRF does not have a causal interpretation (see below in Section 2.7), OLS-MM is consistent only for the PRF not for the causal effect of x on y.

2.6.3 Are $\hat{\beta}_0$ and $\hat{\beta}_1$ the "most efficient" estimators for β_0 and β_1 ?

A third desirable property of an estimator is *efficiency* which requires that the estimator has a small variance, possibly the smallest in a given class of estimators.

Remember that since the estimate is a function of random variables (the sample observations), it is itself a random variable.

We have seen that under assumptions SLR 1 - SLR 4,

$$E(\hat{\beta}_1|x) = \beta_1 \quad \text{and} \quad E(\hat{\beta}_0|x) = \beta_0 \tag{76}$$

We know want to find

$$V(\hat{\beta}_1|x) \quad \text{and} \quad V(\hat{\beta}_0|x) \tag{77}$$

The simplest context in which these variances can be computed is the one of *homoscedasticity*

Homoscedasticity

SLR 5: The error u is said to be homoscedastic if it has the same variance given any value of the explanatory variable x:

$$V(u|x) = \sigma^2 \tag{78}$$

It is important to realize that SLR 5:

- is not needed to prove unbiasedness
- it is just introduced at this stage to simplify the calculation of the variance of the estimator, but we will later remove it because it is unlikely to hold in most applications.

What we can say at this stage is that under SLR1 - SLR5:

$$E(y|x) = \beta_0 + \beta_1 x \quad \text{and} \quad V(y|x) = \sigma^2 \tag{79}$$

which is the situation described in Figure 2.8 of Wooldridge.

The variance of $\hat{\beta}_1$ under homoscedasticity

Using 56 we can express the variance of $\hat{\beta}_1$ (see problem 2.10 for the $\hat{\beta}_0$) as

$$\begin{aligned}
V(\hat{\beta}_1|x) &= V\left(\beta_1 + \frac{\sum_{i=1}^{n}(u_i)(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}|x\right) \text{ (β_1 is a constant)} & (80)\\
&= \left(\frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)^2 V\left(\sum_{i=1}^{n}(u_i)(x_i - \bar{x})|x\right) \text{ (conditioning on x)}\\
&= \left(\frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)^2 \sum_{i=1}^{n}(x_i - \bar{x})^2 V\left(u_i)|x\right) \text{ (indep., random i)}\\
&= \left(\frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)\sigma^2 \text{ (homoschedasticity)}\\
&= \frac{\sigma^2}{SST_x}
\end{aligned}$$

The variance of $\hat{\beta}_1$ is smaller, the smaller is the variance of the unobserved component and the larger is the sample variance of the explanatory variable x.

How can we estimate σ^2

Given a sample we have SST_x but we still need an estimate of σ^2. Consider:

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + u_i \quad (81) \\ y_i &= \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{u}_i \quad (82) \end{aligned}$$

Note that

$$\hat{u}_i - u_i = -(\hat{\beta}_0 - \beta_0) - (\hat{\beta}_1 - \beta_1)x_i \quad (83)$$

which implies that the estimated residual $\hat{u}_i$ is in general different than the unobservable component u_i. Taking the sample average of 83 we get:

$$\bar{u} = (\hat{\beta}_0 - \beta_0) + (\hat{\beta}_1 - \beta_1)\bar{x} \quad (84)$$

where $\bar{u}$ is the sample average of the u_i (note that the sample average of $\hat{u}_i$ is zero). Adding 84 to 83 we get:

$$\hat{u}_i = (u_i - \bar{u}) - (\hat{\beta}_1 - \beta_1)(x_i - \bar{x}) \quad (85)$$

Since $\sigma^2 = E(u_i^2)$ it would seem natural to construct an estimator $\hat{\sigma}^2$ building around $\sum_{i=1}^{n}(\hat{u}_i^2)$.

An unbiased estimator for σ^2

Using 85:

$$\begin{aligned} E(\sum_{i=1}^{n} \hat{u}_i^2) &= E[\sum_{i=1}^{n}(u_i - \bar{u})^2] + E[(\hat{\beta}_1 - \beta_1)^2 \sum_{i=1}^{n}(x_i - \bar{x})^2] \qquad (86) \\ &- 2E[(\hat{\beta}_1 - \beta_1) \sum_{i=1}^{n} u_i(x_i - \bar{x})] \\ &= (n-1)\sigma^2 + \sigma^2 - 2\sigma^2 = (n-2)\sigma^2 \end{aligned}$$

Hence and unbiased estimator of σ^2 is:

$$\hat{\sigma} = \frac{1}{n-2} \sum_{i=1}^{n} \hat{u}_i^2 \qquad (87)$$

The intuition for the $n-2$ is that there are only $n-2$ degrees of freedom in the OLS residuals since

$$\sum_{i=1}^{n} \hat{u}_i = 0 \quad \text{and} \quad \sum_{i=1}^{n} x_i \hat{u}_i = 0 \qquad (88)$$

2.6.4 Asymptotic efficiency

If the sample size is large enough, in parallel to consistency we may be interested in the asymptotic distribution (specifically the variance) of the OLS-MM estimator.

It is possible to prove that under the assumptions SLR 1 - SLR 5

$$\sqrt{n}(\hat{\beta}_1 - \beta_1) \sim \text{Normal}\left(0, \frac{\sigma^2}{V(x)}\right) \tag{89}$$

Moreover, it is also possible to show that the asymptotic variance is the smallest in the class of linear estimators.

2.6.5 The Gauss-Markov Theorem

Under the assumptions:

SLR 1: In the population y is a linear function of x.

SLR 2: The n observations y_i and x_i are a random sample of the population.

SLR 3: The observations $\{x_1, ..., x_n\}$ are not all equal.

SLR 4: The residual u is mean-independent of x.

SLR 5: The error u is homoschedastic.

The OLS-MM estimator is the Best Linear Unbiased Estimators (BLUE) and has the smallest asymptotic variance in the class of linear estimators for the parameters in

$$y = \beta_0 + \beta_1 x + u \tag{90}$$

Note that SLR 5 is needed only for efficiency.

The proof is easier in the context of the matrix derivation of the OLS-MM estimator which we will discuss below.

2.7 Causality and Regression: a brief introduction for future reference

So far we have characterized the Population Regression Function as a linear approximation to the Conditional Expectation Function.

OLS-MM is an estimator of the PRF with some desirable properties.

Given a specific sample, the Sample Regression Function estimated with OLS-MM is a “good” estimate of the PRF-CEF.

It is not an estimate of the causal effect of x on y unless the CEF-PRF itself can be interpreted in a causal sense.

We want to briefly introduce what it means to give a causal interpretation to the PRF-CEF and what this implies for the regression.

A more detailed and exhaustive analysis of the problem of Causal Inference is left for the third course of the LMEC microeconometrics sequence.

What is needed for a "causal" interpretation of the PRF

For each subject in the population there exist two "potential wage levels" depending on whether one goes to college (high education) or not (low education):

$$\begin{aligned} y_h &= \mu_h + \nu \\ y_l &= \mu_l + \nu \end{aligned} \tag{91}$$

where $E(\nu) = 0$. Only one of these outcomes realizes and is effectively observed.

The "causal effect" of college attendance on earnings for a subject is defined as the difference between the two potential outcomes (Holland 1986):

$$\tau = y_h - y_l = \mu_h - \mu_l \tag{92}$$

This population parameter is not identified for a given subject because nobody is observed in both the two potential "treatment" situations.

Let $x = 1$ denote college attendance while $x = 0$ indicates lower education. The observed wage level y is given by:

$$y = y_l(1 - x) + y_h x \tag{93}$$

From potential to observed outcomes

We want to know if and under what conditions the parameter β_1 of the PRF

$$y = \beta_0 + \beta_1 x + u \tag{94}$$

identifies the average causal effect of college attendance on earnings in the population.

Substituting 91 in 93 the causal relationship between x and y is:

$$y = \mu_l + (\mu_h - \mu_l)x + \nu \tag{95}$$

which looks promising, but we need to show that, given how we defined β_1 in the PRF (see equation 8), it follows that:

$$\beta_1 = \mu_h - \mu_l \tag{96}$$

In other words we need to show that in this context, if

$$(\beta_0, \beta_1) = arg \min_{b_0, b_1} E\left[(y - b_0 - b_1 x)^2\right] \tag{97}$$

then 96 holds.

A useful general result: regression when x is a dummy

We have seen that the solution to problem 97 is

$$\beta_1 = \frac{Cov(y,x)}{V(x)} = \frac{E(yx) - E(y)E(x)}{E(x^2) - (E(x))^2} \tag{98}$$

Note that β_1 is a population parameter (not an estimator).

Since x is a dummy, $V(x) = p(1-p)$ where $p = Pr(x = 1)$, while the numerator of 98 is:

$$\begin{aligned} E(yx) - E(y)E(x) &= E(y|x=1)p - pE(y) \qquad (99) \\ &= E(y|x=1)p - p[E(y|x=1)p + E(y|x=0)(1-p)] \\ &= E(y|x=1)p(1-p) - E(y|x=0)p(1-p) \end{aligned}$$

and therefore

$$\beta_1 = \frac{Cov(y,x)}{V(x)} = E(y|x=1) - E(y|x=0) \tag{100}$$

The corresponding OLS-MM estimator obtained by substituting sample averages on the right hand side of 100 is called "Wald estimator".

Is β_1 a causal parameter?

Substituting 95 in 100, we get

$$\begin{aligned}\beta_1 &= E(y|x=1) - E(y|x=0) \\ &= E(\mu_h + \nu|x=1) - E(\mu_l + \nu|x=0) \\ &= \mu_h - \mu_l + [E(\nu|x=1) - E(\nu|x=0)] \\ &= \tau + [E(\nu|x=1) - E(\nu|x=0)]\end{aligned} \tag{101}$$

where the term in brackets is called *Selection Bias* (SB) and captures all the (pre-treatment) unobservable differences between college graduates and other subjects, which are not attributable to college attendance.

The PRF and β_1 have a causal interpretation if the SB is zero, i.e. “treated” and “non-treated” subjects would be identical in the absence of treatment. This may happen:

- in a randomized controlled experiment;
- when for other reasons not controlled by the researcher, exposure to treatment is random in the population.

The more general result when x is not dummy

In the more general situation in which x is not a dummy

$$\begin{aligned} \beta_1 &= \frac{Cov(y,x)}{V(x)} = \frac{Cov[(\mu_l + \tau x + \nu), x]}{V(x)} \\ &= \tau + \frac{Cov(\nu, x)}{V(x)} \end{aligned} \tag{102}$$

and the PRF is causal when, in the population, the treatment x is uncorrelated with unobservable pre-treatment characteristics ν of subjects.

The interpretation is the same as in the “binary x” case.

Causality is a feature of the relationship between x and y, and can be identified only when subjects are randomly exposed to x.

When random exposure of subjects to x occurs in the population of interest, we can interpret the PRF as a causal relationship.

Another way to put it

Compare:

$$y = \beta_0 + \beta_1 x + u \tag{103}$$
$$y = \mu_l + \tau x + \nu \tag{104}$$

We know that by definition β_0 and β_1 in 103 imply

$$Cov(x, u) = E(xu) = 0 \tag{105}$$

but nothing guarantees that the u which derive from the definition of the PRF parameters and satisfies 105, coincide with ν.

Only when x and ν are that

$$Cov(x, \nu) = E(x\nu) = 0 \tag{106}$$

i.e. we have random exposure of subjects to x in the population, then

$$\nu = u \qquad \text{and} \qquad \beta_1 = \tau \tag{107}$$

and the PRF can be interpreted causally.

Consistency and causality

Following the A-P approach, the OLS-MM estimator is consistent for the PRF by definition of the population parameters it aims to estimate because

$$Cov(x, u) = E(xu) = 0 \tag{108}$$

follows from the definition of β_1 and β_0 and is not an assumption.

But "consistency" simply means that the SRF can be made arbitrarily close to the PRF by increasing the sample size.

Thus, consistency of OLS-MM implies nothing about causality. Only if

$$Cov(x, \nu) = E(x\nu) = 0 \tag{109}$$

the PRF is a causal relationship, in which case the OLS-MM is consistent for the causal effect of x on y in the population.

If we are not interested in unbiasedness (and why should we) we can forget of:

$$E(u|x) = 0 \tag{110}$$

2.8 Summary

- The causal effect of x and y requires comparing counterfactuals and cannot be identified for a specific subject.
- If we have a population in which exposure to x is random, then the PRF identifies the average causal effect of x on y in the population.
- But even if exposure to x is not random, we can still define and be interested in the PRF, which is the MMSE approximation to the unknown CEF.
- The PRF defines its parameters in a way such that the population residuals are uncorrelated with x, but this does not ensure a causal interpretation.
- However this definition of the PRF guarantees that we can say something about the PRF (and the CEF) with a random sample of the population.
- Given a specific sample, the OLS-MM estimator provides the Best linear Unbiased Estimates of the PRF parameters (independently of causality) if the SLR 1 - SLR 5 assumptions of Gauss Markov hold.
- SLR 1 - SLR 3 are enough for OLS-MM to be consistent for the PRF.

An example of an interesting, but not causal, PRF

Suppose that the olive oil y produced by a tree in my field depends on the daily rainfall x during spring, which changes from tree to tree because of wind.

Rainfall is arguably random and I am interested in the causal relationship

$$y = \mu + \tau x + \nu \tag{111}$$

where ν captures other determinants of a trees' product y.

Under 30% of the trees (my random sample) I have a device that gives a daily rainfall measure $\tilde{x}$ of the rain falling on the tree, with a random error η:

$$\tilde{x} = x + \eta \tag{112}$$

The relationship between $\tilde{x}$ and y is

$$y = \mu + \tau\tilde{x} - \tau\eta + \nu = \mu + \tau\tilde{x} + e \tag{113}$$

and is not causal because

$$Cov(\tilde{x}, e) = Cov(\tilde{x}, -\tau\eta + \nu) = -\tau V(\eta) \tag{114}$$

Can the PRF of y on x be interpreted causally?

Consider the population regression:

$$y = \beta_0 + \beta_1 \tilde{x} + u \tag{115}$$

where β_0 andβ_1 are defined to ensure that $Cov(\tilde{x}, u) = 0$, which implies:

$$\begin{aligned} \beta_1 &= \frac{Cov(y, \tilde{x})}{V(\tilde{x})} \qquad (116) \\ &= \frac{Cov(\mu + \tau\tilde{x} + e, \tilde{x})}{V(\tilde{x})} = \tau + \frac{Cov(e, \tilde{x})}{V(\tilde{x})} \\ &= \tau - \tau\frac{V(\eta)}{V(x) + V(\eta)} \end{aligned}$$

The PRF is not causal, because regression is not capable to distinguish between variation in x which have an effect and variation in η which have no effect.

Therefore the PRF provides a downward biased measure of the causal effect of x on y and the size of the bias depends on the "noise-to-signal" ratio:

$$\frac{V(\eta)}{V(x) + V(\eta)}$$

Is it still interesting to estimate the PRF?

I cannot use β_1 from the PRF to say what would happen if I artificially increase the quantity of rainfall on my tree.

Indeed, if I used it I would underestimate the causal effect.

But I can still use the PRF as the best predictor of a tree's product given my (imperfect) rainfall measurement.

If I need to decide in advance how many olive oil bottles I should buy, the PRF gives me the best prediction given the available information.

With a random sample of rainfall measures and related olive oil output, I can estimate the SRF which would be consistent for the PRF.

The consistency of $\hat{\beta}_1$ for β_1 would still be desirable for prediction purposes, even if $\hat{\beta}_1$ would not be consistent for τ.

3 Multiple Regression Analysis

Even if a non causal PRF may be interesting, our main goal is and should be to estimate a PRF that is also causal.

We now consider cases in which it is reasonable to make the *Conditional Independence Assumption*.

This assumption says that controlling for a set of observable variables, the PRF has a causal interpretation for the main effect of interest.

In the following section we want to understand:

- the meaning of this assumption;
- how it relates to multiple regression.

An example: the effect of children's sex on parental time with children, controlling for the number of children.

In the last course of the microeconometric sequence we will consider other assumptions that allow to estimate consistently causal parameters.

3.1 The Conditional Independence Assumption (CIA) and Regression

Consider a causal model like 95 that we derived in Section 2.7:

$$y = \mu + \tau_1 x_1 + \nu \tag{117}$$

where y is earnings and x_1 is years of education.

Suppose that $\nu = \tau_2 x_2 + \omega$ where x_2 is genetic ability.

Then β_1 of the PRF of y on x_1 is

$$\begin{aligned}
\beta_1 &= \frac{Cov(y, x_1)}{V(x_1)} \qquad (118) \\
&= \frac{Cov(\mu + \tau_1 x_1 + \nu, x_1)}{V(x_1)} = \tau_1 + \frac{Cov(\nu, x_1)}{V(x_1)} \\
&= \tau_1 + \tau_2 \frac{Cov(x_1, x_2)}{V(x_1)} + \frac{Cov(x_1, \omega)}{V(x_1)}
\end{aligned}$$

which, given $\tau_2 \neq 0$, is equal to the causal parameter τ_1 of equation 117 only if

$$Cov(x_1, x_2) = Cov(x_1, \omega) = 0.$$

A solution if x_2 is observable

The Conditional Independence Assumption says that for a given value of $x_2 = k$

$$Cov(x_1, x_2 | x_2 = k) = 0 \tag{119}$$

which is obvious because now x_2 is a fixed number, and

$$Cov(x_1, \omega | x_2 = k) = 0 \tag{120}$$

which is less obvious: it is actually the crucial part of the assumption.

If we take a sub-group of the population with a given level of ability $x_2 = k$ and we estimate the PRF for this population sub-group, the PRF is causal.

When education x_1 is a binary variable, the assumption is easier to interpret:

$$Cov(x_1, \omega | x_2 = k) = 0 = [E(\omega | x_1 = 1, x_2 = k) - E(\omega | x_1 = 0, x_2 = k)]$$

which says that among individual with ability $x_2 = k$ there is no "selection bias" in the choice between education levels. In other words education is chosen randomly for given ability.

The population multiple regression function

Consider the population regression of y on both x_1 and x_2:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u \tag{121}$$

where

$$(\beta_0, \beta_1, \beta_2) = arg \min_{b_0, b_1, b_2} E\left[(y - b_0 - b_1 x_1 - b_2 x_2)^2\right] \tag{122}$$

i.e. where the population parameters are defined to minimize the square of the difference between y and the PMRF itself.

We want to show that if the CIA holds β_1 is the causal effect of x_1 on y

And the same is true symmetrically if we are interested in the effect of x_2.

Therefore, if given a random sample we can estimate consistently the PMRF, we can obtain consistent estimates of the causal parameters of interest.

The coefficients of the PMRF

The First Order Conditions for problem 122 are:

$$E\left[x_1(y-\beta_0-\beta_1 x_1-\beta_2 x_2)\right] = 0 \tag{123}$$
$$E\left[x_2(y-\beta_0-\beta_1 x_1-\beta_2 x_2)\right] = 0 \tag{124}$$
$$E\left[(y-\beta_0-\beta_1 x_1-\beta_2 x_2)\right] = 0 \tag{125}$$

The first two conditions are symmetric: let's focus on 123.

Consider the simple linear PRF of x_1 on x_2. We can always write:

$$x_1 = \hat{x}_1 + \hat{r}_1 \tag{126}$$

which we can substitute in 123 to get

$$E\left[(\hat{x}_1+\hat{r}_1)(y-\beta_0-\beta_1 x_1-\beta_2 x_2)\right] = E\left[(\hat{x}_1+\hat{r}_1)u\right] = 0 \tag{127}$$

By the definition of the PRF, $E(\hat{x}_1 u) = 0$, since $\hat{x}_1$ is a linear function of x_2. Moreover $E(\hat{r}_1 x_2) = 0$ given 126 and $E(\hat{r}_1 \beta_0) = 0$. Thus 127 becomes:

$$E\left[\hat{r}_1(y-\beta_1 x_1)\right] = 0 \tag{128}$$

The coefficients of the PMRF (cont.)

Substituting 126 in 128 we get:

$$E\left[\hat{r}_1(y - \beta_1(\hat{x}_1 + \hat{r}_1))\right] = 0 \tag{129}$$

Again because $E(\hat{r}_1\hat{x}_1) = E(\hat{r}_1 x_2) = 0$ we are left with

$$E\left[\hat{r}_1(y - \beta_1\hat{r}_1)\right] = 0 \tag{130}$$

which finally gives

$$\beta_1 = \frac{E(\hat{r}_1 y)}{E(\hat{r}_1^2)} = \frac{Cov(\hat{r}_1, y)}{V(\hat{r}_1)} \tag{131}$$

The PRF coefficient β_1 is equal to the covariance between y and the residuals of the PRF of x_1 on x_2, divided by the variance of these residuals.

We now want to show that if the CIA is satistified

$$\beta_1 = \tau_1 \tag{132}$$

and the PRF of y on x_1 and x_2 has a causal interpretation for the effect of x_1.

Similar results holds for β_2.

The coefficients of the PMRF under the CIA

Substitute the causal model 117 in the numerator of 131:

$$\begin{aligned} E(\hat{r}_1 y) &= E(\hat{r}_1(\mu + \tau_1 x_1 + \tau_2 x_2 + \omega)) \\ &= \tau_1 E(\hat{r}_1^2) + \tau_2 E(\hat{r}_1 x_2) + E(\hat{r}_1 \omega) \end{aligned} \tag{133}$$

Note that:

$$E(\hat{r}_1 x_2) = 0 \tag{134}$$

and

$$E(\hat{r}_1 \omega) = E(E(\hat{r}_1 \omega | x_2)) = E(Cov(x_1, \omega | x_2)) = 0 \tag{135}$$

where the second equation is satisfied if the CIA 120 holds.

Therefore:

$$\beta_1 = \frac{E(\hat{r}_1 y)}{E(\hat{r}_1^2)} = \frac{\tau_1 E(\hat{r}_1^2)}{E(\hat{r}_1^2)} = \tau_1 \tag{136}$$

If the CIA holds the PMRF can be interpreted causally for x_1.

The same may (but does not have to) be true symmetrically for x_2.

Summary

We have shown that if we are interested in the causal effect of x_1 on y the CIA may represent a solution.

The CIA says that the other variables $\{x_2, ..., x_k\}$ that we observe are detailed and exaustive enough to guarantee that if two subjects are equal in terms of these variables the value of x_1 is effectively assigned randomly to them.

The randomness of the assignment of x_1 given $\{x_2...x_k\}$ is what permits a causal interpretation of β_1.

In what follows in this course we assume that the CIA holds symmetrically for all variables, and therefore all the parameters of the PMRF can be interpreted causally.

In the final course of the LMEC econometric sequence we will discuss alternative solutions when the CIA cannot be assumed to hold.

3.2 Interpretation of the partial Multiple Regression coefficient

Extending the analysis to many covariates x, consider:

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + u \tag{137}$$

$$(\beta_0, ..., \beta_k) = arg \min_{b_0,...,b_k} E\left[(y - b_1 x_1 - ... - b_k x_k)^2\right] \tag{138}$$

the generic parameter β_j (for $j > 0$) is

$$\beta_j = \frac{E(\hat{r}_j y)}{E(\hat{r}_j^2)} = \frac{Cov(\hat{r}_j, y)}{V(\hat{r}_j)} \tag{139}$$

This parameter measures the effect on y of the component of x_j that is orthogonal to the other x variables. In fact this parameter can be obtained by:

- regressing x_j on all the others x variables;
- taking the residuals of this regression $\hat{r}_j$;
- considering the simple PRF of y on the single variable $\hat{r}_j$;
- $\hat{r}_j$ captures the part of x_j that is orthogonal to the other x variables.

3.3 From the SMRF to the PMRF in matrix form

As for the case of the simple linear regression, we now suppose to have a random sample of observations on y and $x_1, ...x_k$ and we ask:

- whether we can extend the OLS-MM estimator;
- whether the OLS-MM estimator continues to have good properties.

Since we have multiple covariates it is convenient to use matrix notation.

$$Y = X\beta + U \tag{140}$$

where

- Y is the $n \times 1$ column vector of observations on the outcome y_i.
- X is the $n \times (k + 1)$ matrix of observations x_{ij} on the jth covariate.
- U is the $n \times 1$ column vector of observations u_i.
- β is the $(k + 1) \times 1$ column vector of the parameters to be estimated.

Note that X includes a column with all elements equal to 1 and the corresponding parameter is the constant β_0.

The basic set of necessary assumption

MLR 1: The population regression function is linear in the parameters:

$$Y = X\beta + U \tag{141}$$

MLR 2: The n observations on Y and X are a random sample of the population, so that

$$y_i = X_i\beta + u_i \tag{142}$$

where X_i is the ith row of X.

MLR 3: There is no perfect collinearity, i.e no variable in X is constant (in addition to the constant term ...) and there is no exact linear dependency between any set of variables in X. Thus X has full rank equal to (k+1).

MLR-3 is crucial and sometimes may generate unexpected problems.

It is a generalized version of SLR-3 in the simple regression case.

Example: consider the case of a regression of earnings on dummies for gender. Why X cannot contain a constant and both gender dummies?

The OLS-MM estimator in matrix form

Under these assumptions, the OLS-MM estimator solves the following problem

$$\hat{\beta} = arg\min_{b} U'U = arg\min_{b}[Y - Xb]'[Y - Xb] \tag{143}$$

where b is a $(k+1) \times 1$ column vector of possible parameter values.

There are $k+1$ FOC for this problem which we can write as

$$\frac{\partial U'U}{\partial b} = X'[Y - X\hat{\beta}] = 0 \tag{144}$$

or

$$X'X\hat{\beta} = X'Y \tag{145}$$

which give the OLS-MM estimator in matrix form

$$\hat{\beta} = (X'X)^{-1}X'Y \tag{146}$$

where the full rank of X makes $X'X$ invertible.

Algebraic properties of OLS in matrix form

The fitted values are

$$\hat{Y} = X\hat{\beta} \tag{147}$$

and the estimated residuals are

$$\hat{U} = Y - \hat{Y} = Y - X\hat{\beta} \tag{148}$$

Therefore the first order condition 144 can also be written as

$$X'\hat{U} = 0 \tag{149}$$

and since the first row of X' is a row of ones (the constant), the sum of the OLS residuals is zero.

3.4 Unbiasedness of the OLS-MM estimator of the PMRF

The proof of unbiasedness is similar to the simple regression case;

$$\begin{aligned}\hat{\beta} &= (X'X)^{-1}X'Y \\ &= (X'X)^{-1}X'(X\beta + U) \\ &= (X'X)^{-1}X'X\beta + (X'X)^{-1}X'U \\ &= \beta + (X'X)^{-1}X'U\end{aligned} \tag{150}$$

Taking the expectation

$$\begin{aligned}E(\hat{\beta}|X) &= \beta + (X'X)^{-1}X'E(U|X) \\ &= \beta\end{aligned} \tag{151}$$

which follows from the assumption:

MLR 4: Conditioning on the entire matrix X each u_i has zero mean

$$E(U|X) = 0 \tag{152}$$

Think about the meaning of this assumption in a times series context with lag and lead variables.

3.4.1 Omitted variable bias and inclusion of irrelevant regressors

Suppose that we have omitted a variable Z which we think should be included for the CIA to hold. Thus:

$$U = Z\gamma + V \tag{153}$$

The expected value of the estimator for β is:

$$\begin{aligned} E(\hat{\beta}|X) &= \beta + (X'X)^{-1}X'E[U|X] \\ &= \beta + (X'X)^{-1}E[X'Z|X]\gamma + (X'X)^{-1}X'E[V|X] \\ &= \beta + (X'X)^{-1}X'E[Z|X]\gamma \end{aligned} \tag{154}$$

The omission of Z generates a bias if

- the mean of Z is not independent of X;
- Z has a non-zero effect γ on the outcome.

The sign of the bias is easy to determine if X and Z include only one variable each. Not obvious otherwise.

3.5 Variance of the OLS-MM estimator of the PMRF

We now derive the variance of the OLS-MM estimator under the simple case of homoschedasticity

MLR 5: The variance-covariance matrix of the unobservable component is

$$Var(U|X) = E(UU'|X) = \sigma^2 I_n \tag{155}$$

where I_n is the $n \times n$ identity matrix.

Note that this assumption (which we have already seen in the simple regression case) has two important components:

- The variance of u_i should not depend on any variable x_j.
- The covariance between u_t and u_s should be zero for any t and s. This component:
  - typically does not hold in time series because of serial correlation;
  - it is traditionally assumed to hold because of random sampling in a cross-sectional context; but recently authors understand that in most applications it cannot be assumed to hold even in a cross section (see below).

Variance-covariance matrix

Given 146 and 151:

$$\begin{aligned} Var(\hat{\beta}|X) &= E[(\hat{\beta}-\beta)(\hat{\beta}-\beta)'|X] \\ &= E[(X'X)^{-1}X'UU'X(X'X)^{-1}|X] \\ &= (X'X)^{-1}X'E[UU'|X]X(X'X)^{-1} \\ &= (X'X)^{-1}X'\sigma^2 I_n X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1} \end{aligned} \tag{156}$$

which is a $(k+1)\times(k+1)$ matrix.

The OLS-MM estimator is more precise:

- the smaller is the variance of the unobservable components.
- the larger is the total variation in the observable regressors X.
- the smaller is the collinearity among the observable regressors in X.

What does this mean for strategies that we can adopt to increase precision of OLS-MM?

An alternative useful way to write the variance of the OLS-MM estimator

Following Wooldridge (Appendix to Chapter 3), the variance of the jth parameter can be written as

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1 - R_j^2)} \tag{157}$$

where

- $SST_j = \sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$ is the total sample variation of the regressor x_j.
- R_j^2 is the R-squared of the regression of x_j on the other regressors.

This expression emphasizes the three components of the variance of the OLS-MM estimator:

- variance of the unobservable components;
- variance of the regressors;
- multicollinearity between the regressors.

Is it always a good idea to include more regressors?

An unbiased estimator of σ^2

We want to show that

$$\hat{\sigma}^2 = \frac{1}{n-k-1}\hat{U}'\hat{U} \tag{158}$$

is unbiased for σ^2. Note that for $k = 1$ this is the same estimator that we have studied for the simple linear regression case.

$$\begin{aligned} \hat{U} &= Y - X\hat{\beta} \\ &= Y - X(X'X)^{-1}X'Y \\ &= MY = M(X\beta + U) \\ &= MU \end{aligned} \tag{159}$$

Where $M = I - X(X'X)^{-1}X'$ is a symmetric and idempotent matrix:

- $M' = M$
- $M'M = M$
- $MX = 0$
- $MY = MU$

An unbiased estimator of σ^2 (cont.)

$$\begin{aligned} E[\hat{U}'\hat{U}|X] &= E[U'M'MU|X] && (160) \\ &= E[tr(U'MU)|X] && \text{because a scalar is equal to its trace} \\ &= E[tr(MUU')|X] && \text{because of the property of the trace} \\ &= tr(ME[UU'|X]) \\ &= tr(M)\sigma^2 \\ &= (n-k-1)\sigma^2 \end{aligned}$$

which proves the result. The last equality follows because

$$\begin{aligned} tr(M) &= tr(I_n) - tr(X(X'X)^{-1}X') && (161) \\ &= tr(I_n) - tr((X'X)^{-1}X'X) \\ &= tr(I_n) - tr(I_{k+1}) \\ &= n-k-1 \end{aligned}$$

In a sample of size n that we use to estimate $k+1$ parameters β, we are left with only $n-k-1$ “degrees of freedom” to estimate σ^2.

3.6 The Gauss-Markov theorem

Under the assumptions

MLR 1: The population regression function is linear in the parameters:

$$Y = X\beta + U \tag{162}$$

MLR 2: The n observations on Y and X are a random sample of the population

$$y_i = X_i\beta + u_i \tag{163}$$

MLR 3: There is no collinearity and X has full rank equal to (k+1).

MLR 4: Conditioning on the entire matrix X each u_i has zero mean

$$E(U|X) = 0 \tag{164}$$

MLR 5: The variance-covariance matrix of the unobservable component is

$$Var(U|X) = E(UU'|X) = \sigma^2 I_n \tag{165}$$

The OLS-MM estimator $\hat{\beta}$ is the best linear unbiased estimator.

Proof of the Gauss Markov theorem

Consider a generic alternative linear unbiased estimator

$$\tilde{\beta} = A'Y \tag{166}$$

where A is a $n \times (k+1)$ matrix. Linearity in Y implies that A is a function of X but cannot be a function of Y. Since $\tilde{\beta}$ is unbiased it must be the case that:

$$\begin{aligned} E(\tilde{\beta}|X) &= A'X\beta + A'E(U|X) \qquad\qquad (167) \\ &= A'X\beta \qquad\qquad \text{because } E(U|X) = 0 \\ &= \beta \end{aligned}$$

and therefore $A'X = I_{k+1}$ and $\tilde{\beta}$ characterizes the class of linear (in Y) unbiased estimators.

The variance of $\tilde{\beta}$ is:

$$\begin{aligned} Var(\tilde{\beta}|X) &= E[(\tilde{\beta} - \beta)(\tilde{\beta} - \beta)'|X] \qquad\qquad (168) \\ &= E[A'UU'A|X] \\ &= \sigma^2(A'A) \end{aligned}$$

Proof of the Gauss Markov theorem (cont)

$$\begin{aligned} Var(\tilde{\beta}|X) - Var(\hat{\beta}|X) &= \sigma^2[A'A - (X'X)^{-1}] \qquad (16 \\ &= \sigma^2[A'A - A'X(X'X)^{-1}X'A] \text{ because } A'X = I_{k} \\ &= \sigma^2 A'[I_n - X(X'X)^{-1}X']A \\ &= \sigma^2 A'MA \end{aligned}$$

Since M is symmetric and idempotent, $A'MA$ is positive semidefinite for any conformable A, which proves the result.

The OLS-MM estimator $\hat{\beta}$ has the smallest variance in the class of linear unbiased estimators.

3.7 Partialling out: again on the interpretation of the PMRF

The matrix

$$H = Z(Z'Z)^{-1}Z' \tag{170}$$

is called a"projection matrix" because if you premultiply any vector Y by H, the result is the projection of the vector Y on the space spanned by Z.

Numerically it gives the least square prediction of Y given Z (see graphical interpretation of OLS).

$$Y_Z = HY = Z(Z'Z)^{-1}Z'Y = Z\hat{\psi} \tag{171}$$

for the PRF

$$Y = Z\psi + V \tag{172}$$

Note that H is symmetric and idempotent.

Projections

Consider the population regression:

$$Y = X\beta + U = W\delta + Z\gamma + U \tag{173}$$

where W is the main variable of interest and Z contains a set of other control variables.

Consider the two projections

$$\begin{aligned} Y_Z &= HY = Z(Z'Z)^{-1}Z'Y = Z\tilde{\gamma} && (174) \\ W_Z &= HW = Z(Z'Z)^{-1}Z'W = Z\tilde{\rho} && (175) \end{aligned}$$

Consider the residuals from these two projections that we denote as

$$\begin{aligned} \tilde{Y} &= Y - Y_Z && (176) \\ \tilde{W} &= W - W_Z && (177) \end{aligned}$$

What happens if we regress $\tilde{Y}$ on $\tilde{W}$?

Partialling out matrices

Consider now the symmetric idempotent matrix M:

$$M = I - H = I - Z(Z'Z)^{-1}Z' \tag{178}$$

If you premultiply any vector by M you obtain the least square estimated residuals of the regression of the vector on Z (see graphical analysis).

Specifically:

$$\begin{aligned}
\tilde{Y} &= Y - Y_Z && (179)\\
&= MY = Y - Z(Z'Z)^{-1}Z'Y && (180)\\
\tilde{W} &= W - W_Z && (181)\\
&= MW = W - Z(Z'Z)^{-1}Z'W && (182)\\
\tilde{U} &= U - U_Z && (183)\\
&= MU = U - Z(Z'Z)^{-1}Z'U && (184)\\
\tilde{Z} &= Z - Z_Z && (185)\\
&= MZ = Z - Z(Z'Z)^{-1}Z'Z = 0 && (186)
\end{aligned}$$

Partialling out matrices (cont.)

Let's now premultiply the PMRF 173 by M:

$$\begin{aligned} MY &= MW\delta + MZ\gamma + MU \\ \tilde{Y} &= \tilde{W}\delta + \tilde{U} \end{aligned} \tag{187}$$

which explains why M is called a "partialling out" matrix. Note that this PRF satisfies Gauss-Markov.

Consider the OLS-MM estimator of 187

$$\begin{aligned} \hat{\delta} &= (\tilde{W}'\tilde{W})^{-1}\tilde{W}'\tilde{Y} \\ &= (W'M'MW)^{-1}W'M'MY \\ &= (W'MW)^{-1}W'MY \end{aligned} \tag{188}$$

It is obtained by regressing Y on the component of W which is orthogonal to Z and is numerically identical to the OLS-MM estimator of δ that we would obtain by estimating directly 173.

Also the standard error is numerically identical: $Var(\hat{\delta}) = \sigma^2(W'MW)^{-1}$.

3.8 Good and bad habits concerning control variables

It is important to realize that it may not always be a good idea to add controls in a regression, specificaly controls that are themselves causally affected by the main variable of interest.

We know that it is a good idea to control for omitted variables, when they are needed to ensure the CIA. If the causal PRF is

$$Y = X\beta + Z\gamma + U \tag{189}$$

and we run

$$Y = X\beta + V \tag{190}$$

we get a biased and inconsistent estimate

$$E(\hat{\beta}) = \beta + (X'X)^{-1}E[X'Z]\gamma \tag{191}$$

If we have observations on Z we should include them in the regression.

It is a good idea to include Z even if $E[X'Z] = 0$, in which case the goal is not to avoid a bias but to increase efficiency.

Controlling to increase precision

Consider a random experiment in which a training program X is randomly assigned to estimate its effect on future earnings Y. The causal PRF is

$$Y = X\beta + U \tag{192}$$

Consider a set of predetermined demografic characteristics D, which by random assignment of X are not correlated with X, but have a causal effect on Y.

If we run the PMRF

$$Y = X\beta + D\gamma + V \tag{193}$$

the OLS estimator for β is:

$$\hat{\beta} = (X'MX)^{-1}X'MY \tag{194}$$

where $M = I - D(D'D)^{-1}D'$. Note that $MX = X$ because $D(D'D)^{-1}D'X = 0$: D and X are not correlated. But

$$\sigma_U^2 = Var(U) = \gamma^2 Var(D) + Var(V) > Var(V) = \sigma_V^2$$

and therefore β is estimated more precisely using 193.

A first case of misleading control variable

Now suppose that D is instead the occupation chosen by the subject after training: white and blue collars.

The training program increases the chance of a white collar occupation.

Note that X is randomly assigned in the population, but not within the occupational group!

If we estimate

$$Y = X\beta + U \tag{195}$$

we get an unbiased and consistent estimate of β which is the overall causal effect of training, including the effect that runs through the occupational choice.

In this case, it would be a bad idea to run

$$Y = X\beta + D\gamma + V \tag{196}$$

unless the efficiency gain were huge.

A first case of misleading control variable (cont.)

If we did run 196, we would get

$$\hat{\beta} = (X'MX)^{-1}X'MY \neq (X'X)^{-1}X'Y \qquad (197)$$

To understand the bias note that 196 is equivalent to comparing trained and not trained for given occupation, i.e. in the case of $D_0 = D_1 = 1$ (here and below subscripts denote the potential earnings and assignments to training):

$$\begin{aligned} E(Y|X=1, D=1) \; - \; & E(Y|X=0, D=1) \qquad \text{(198} \\ = \; & E(Y_1|X=1, D_1=1) - E(Y_0|X=0, D_0=1) \\ = \; & E(Y_1|D_1=1) - E(Y_0|D_0=1) \\ = \; & E(Y_1 - Y_0|D_1=1) + [E(Y_0|D_1=1) - E(Y_0|D_0=1 \end{aligned}$$

where the second equality derives from the joint independence of Y_1, D_1, Y_0, D_0 from X.

The bias is represented by the selection effect $[E(Y_0|D_1 = 1) - E(Y_0|D_0 = 1)]$ which reflects the fact that composition of the pool of white collar workers has changed because of training even in the counterfactual case of no training.

A second case of misleading control variable

Let's now go back to the case in which the true causal PRF is

$$Y = \alpha + X\beta + Z\gamma + U \tag{199}$$

where Z is predetermined ability, X is education and Y is earnings, but we can observe only a measure $\tilde{Z}$ of Z taken after education has occurred (e.g. IQ):

$$\tilde{Z} = \pi_0 + X\pi_1 + Z\pi_2 + e \tag{200}$$

Substituting 200 in 199 we get

$$Y = \left(\alpha - \gamma\frac{\pi_0}{\pi_2}\right) + \left(\beta - \gamma\frac{\pi_1}{\pi_2}\right) X + \frac{\gamma}{\pi_2}\tilde{Z} + U \tag{201}$$

And the OLS-MM estimator would be biased and inconsistent for the causal parameters of interest.

Depending on assumptions, in this case we could still say something on β.

But the point is that timing is crucial in the choice of appropriate control variables.

4 Inference and hypothesis testing

We are now interested in testing hypothesis concerning the parameters of the PRF, using the estimator that we have constructed and analysed in the previous sections

Here are some examples of hypotheses that we may want to test

- $\beta_j = 0$;
- $\beta_j = q$ where q is any real number;
- $\beta_j \leq q$ where q is any real number, including 0;
- $\beta_j = \beta_h$;
- $\beta_j^2 - 2\beta_j\beta_i = 0$
- $r(\beta) = q$ where $r(.)$ is any function of the parameters.

To test these hypotheses using the theory of Classical Hypothesis Testing, we need to make assumptions on the distribution of the OLS-MM estimator $\hat{\beta}$.

4.1 Small sample distribution of the OLS-MM estimator $\hat{\beta}$

If we are not in a condition to use large sample asymptotic properties of OLS-MM, the only solution is to make small sample distributional assumptions on the unobservable component U.

The Classical Linear Model Assumption is Normality:

MLR 6: In the population U is independent of X and is distributed normally with zero mean and variance $\sigma^2 I_n$

$$U \sim \text{Normal}(0, \sigma^2 I_n) \tag{202}$$

Note that this implies

$$Y \sim \text{Normal}(X\beta, \sigma^2 I_n) \tag{203}$$

Discussion of the small sample assumption of Normality.

From the distribution of U to the distribution of $\hat{\beta}$

Since we know from 150 that

$$\hat{\beta} = \beta + (X'X)^{-1}X'U \tag{204}$$

using 202 it is easy to see that

$$\hat{\beta} \sim \text{Normal}(\beta, \sigma^2(X'X)^{-1}) \tag{205}$$

And for a single PRF parameter we have that the standardized distribution

$$\frac{\hat{\beta}_j - \beta}{sd(\hat{\beta}_j)} = \frac{\hat{\beta}_j - \beta}{\frac{\sigma}{\sqrt{SST_j(1-R_j^2)}}} \sim \text{Normal}(0, 1) \tag{206}$$

In practice, we do not know σ and we have to use its estimate $\hat{\sigma} = \frac{\hat{U}'\hat{U}}{n-k-1}$ so that:

$$\frac{\hat{\beta}_j - \beta}{\hat{sd}(\hat{\beta}_j)} = \frac{\hat{\beta}_j - \beta}{\frac{\hat{\sigma}}{\sqrt{SST_j(1-R_j^2)}}} \sim t_{n-k-1} \tag{207}$$

where t_{n-k-1} denotes a "t distribution" with $n-k-1$ degrees of freedom.

4.2 Small sample testing of an hypothesis

The general logic of classical hypothesis testing can be summarized as follows:

- Define the "null hypothesis" H_0 on the parameter that we want to test.
- Construct a "test statistic" (based on the estimator) and characterize its distribution under H_0.
- Compute the value of the test statistic in the specific sample at our disposal.
- Using the theoretical distribution of the test statistic establish the probability of observing the value that we have actually obtained for the test statistic if H_0 is true.
- If this probability is "sufficiently small" reject H_0.
- The "significance" of the test is the threshold level of probability that we consider sufficiently low to conclude that it is unlikely that the test statistics that we have observed could have originated under H_0.
- The "p-value" of the test is the smallest significance level at which H_0 would actually be rejected given the sample. Note that the p-value is a probability

$H_0 : \beta_j = 0$ against the one sided alternative $H_1 : \beta_j > 0$

The simplest testable hypothesis is that X_j has positive effect on Y

$$H_0 : \beta_j = 0 \quad \text{against} \quad H_1 : \beta_j > 0 \tag{208}$$

The test statistic for this hypothesis and its distribution under H_0 are

$$t_{\hat{\beta}_j} = \frac{\hat{\beta}_j}{\hat{sd}(\hat{\beta}_j)} \sim t_{n-k-1} \tag{209}$$

We reject H_0 if in our sample

$$t_{\hat{\beta}_j} = \frac{\hat{\beta}_j}{\hat{sd}(\hat{\beta}_j)} > c \tag{210}$$

where the critical level $c > 0$ is such that (see Wooldridge Figure 4.2)

$$Pr(\tau > c|H_0) = s \quad \text{with} \quad \tau \sim t_{n-k-1} \tag{211}$$

and s is the significance level (e.g. $s = 0.01$ or $s = 0.05$). The p-value is:

$$p = Pr(\tau > t_{\hat{\beta}_j} = \frac{\hat{\beta}_j}{\hat{sd}(\hat{\beta}_j)}|H_0) \tag{212}$$

$H_0 : \beta_j = 0$ against the one sided alternative $H_1 : \beta_j < 0$

Similarly we can test that X_j has a negative effect on Y

$$H_0 : \beta_j = 0 \quad \text{against} \quad H_1 : \beta_j < 0 \tag{213}$$

The test statistic for this hypothesis and its distribution unde H_0 are

$$t_{\hat{\beta}_j} = \frac{\hat{\beta}_j}{\hat{sd}(\hat{\beta}_j)} \sim t_{n-k-1} \tag{214}$$

We reject H_0 if in our sample

$$t_{\hat{\beta}_j} = \frac{\hat{\beta}_j}{\hat{sd}(\hat{\beta}_j)} < -c \tag{215}$$

where the critical level $-c < 0$ is such that (see Wooldridge Figure 4.3)

$$Pr(\tau < -c | H_0) = s \quad \text{with} \quad \tau \sim t_{n-k-1} \tag{216}$$

and s is the significance level (e.g. $s = 0.01$ or $s = 0.05$). The p-value is:

$$p = Pr(\tau < t_{\hat{\beta}_j} = \frac{\hat{\beta}_j}{\hat{sd}(\hat{\beta}_j)} | H_0) \tag{217}$$

$H_0 : \beta_j = 0$ **against the two sided alternative** $H_1 : \beta_j \neq 0$

More generally we can test that X_j has a non zero effect on Y

$$H_0 : \beta_j = 0 \quad \text{against} \quad H_1 : \beta_j \neq 0 \tag{218}$$

The test statistic for this hypothesis and its distribution under H_0 are again

$$t_{\hat{\beta}_j} = \frac{\hat{\beta}_j}{\hat{sd}(\hat{\beta}_j)} \sim t_{n-k-1} \tag{219}$$

We reject H_0 if in our sample

$$|t_{\hat{\beta}_j}| = \left| \frac{\hat{\beta}_j}{\hat{sd}(\hat{\beta}_j)} \right| > c \tag{220}$$

where the critical level c is such that (see Wooldridge Figure 4.4)

$$Pr(|\tau| > c|H_0) = 0.5s \quad \text{with} \quad \tau \sim t_{n-k-1} \tag{221}$$

and s is the significance level (e.g. $s = 0.01$ or $s = 0.05$). The p-value is:

$$p = 2Pr(\tau > |t_{\hat{\beta}_j}| = \left| \frac{\hat{\beta}_j}{\hat{sd}(\hat{\beta}_j)} \right| |H_0) \tag{222}$$

$H_0 : \beta_j = k$ **against the two sided alternative** $H_1 : \beta_j \neq k$

In this case we test that the effect of X_j has a specific size:

$$H_0 : \beta_j = k \quad \text{against} \quad H_1 : \beta_j \neq k \tag{223}$$

The test statistic for this hypothesis and its distribution under H_0 are again

$$t_{\hat{\beta}_j} = \frac{\hat{\beta}_j - k}{\hat{sd}(\hat{\beta}_j)} \sim t_{n-k-1} \tag{224}$$

We reject H_0 if in our sample

$$|t_{\hat{\beta}_j}| = \left| \frac{\hat{\beta}_j - k}{\hat{sd}(\hat{\beta}_j)} \right| > c \tag{225}$$

where the critical level c is such that (see Wooldridge Figure 4.5)

$$Pr(|\tau| > c | H_0) = \frac{1}{2} s \quad \text{with} \quad \tau \sim t_{n-k-1} \tag{226}$$

and s is the significance level (e.g. $s = 0.01$ or $s = 0.05$). The p-value is:

$$p = 2Pr(\tau > |t_{\hat{\beta}_j}| = \left| \frac{\hat{\beta}_j - k}{\hat{sd}(\hat{\beta}_j)} \right| |H_0) \tag{227}$$

4.2.1 Confidence intervals

Consider the interval $\{-\lambda_\Phi, \lambda_\Phi\}$ defined by the equation:

$$Pr\left(-\lambda_\Phi < \frac{\hat{\beta}_j - \beta_j}{\hat{sd}(\hat{\beta}_j)} < \lambda_\Phi\right) = \Phi \tag{228}$$

The limits $\{-\lambda_\Phi, \lambda_\Phi\}$ can be computed using the fact that $\frac{\hat{\beta}_j - \beta}{\hat{sd}(\hat{\beta}_j)} \sim t_{n-k-1}$. Rearranging 228:

$$Pr\left(\hat{\beta}_j - \lambda_\Phi \hat{sd}(\hat{\beta}_j) < \beta < \hat{\beta}_j + \lambda_\Phi \hat{sd}(\hat{\beta}_j)\right) = \Phi \tag{229}$$

which says that with proability Φ the true value of the parameter β belong to the interval $\{\hat{\beta}_j \pm \lambda_\Phi \hat{sd}(\hat{\beta}_j)\}$. In large sample, when the t distribution approximates normal distribution a realiable approximation of the 95% confidence interval is

$$Pr\left(\hat{\beta}_j - 1.96\hat{sd}(\hat{\beta}_j) < \beta < \hat{\beta}_j + 1.96\hat{sd}(\hat{\beta}_j)\right) = 0.95 \tag{230}$$

which means that with 95% probability the parameter is within two standard deviations from the estimate.

4.2.2 Testing hypothesis about linear combinations of parameters

There are situations in which we are interested in testing a slightly more complicated hypothesis:

$$H_0 : \beta_j = \beta_k \quad \text{against} \quad H_1 : \beta_j \neq \beta_k \tag{231}$$

The test statistic for this hypothesis and its distribution under H_0 are again

$$t_{\hat{\beta}_j,\hat{\beta}_k} = \frac{\hat{\beta}_j - \hat{\beta}_k}{\hat{sd}(\hat{\beta}_j - \hat{\beta}_k)} \sim t_{n-k-1} \tag{232}$$

and we could follow the usual procedure to test the hypothesis

What is slighlty more problematic in this case is the computation of

$$\hat{sd}(\hat{\beta}_j - \hat{\beta}_k) = \sqrt{[\hat{sd}(\hat{\beta}_j]^2 + [\hat{sd}(\hat{\beta}_k]^2 - 2\hat{Cov}(\hat{\beta}_j, \hat{\beta}_k)} \tag{233}$$

Given that $Var(\hat{\beta}|X) = \hat{\sigma}^2(X'X)^{-1}$ we have all the ingredients to compute the test statistics. But there is a simpler alternative.

Rearranging the PRF to test linear combination of hypotheses

Consider the population regression:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u \tag{234}$$

and suppose that we want to to test

$$H_0 : \beta_1 = \beta_2 \quad \text{against} \quad H_1 : \beta_1 \neq \beta_2 \tag{235}$$

If we add and subtract $\beta_2 x_1$ in 234, we get:

$$\begin{aligned} y &= \beta_0 + (\beta_1 - \beta_2)x_1 + \beta_2(x_2 + x_1) + u \qquad (236) \\ y &= \beta_0 + \theta x_1 + \beta_2(x_2 + x_1) + u \end{aligned}$$

and we can now test with the standard procedure:

$$H_0 : \theta = 0 \quad \text{against} \quad H_1 : \theta \neq 0 \tag{237}$$

Note that the estimates of the coefficients on x_2 in 234 and on $(x_2 + x_1)$ in 236 must be numerically identical.

4.2.3 Testing multiple linear restrictions: the F test

Consider the unrestricted regression in matrix form

$$Y = X_1\beta_1 + X_2\beta_2 + U_{ur} \tag{238}$$

where

- X_1 is a $n \times k_1 + 1$ matrix;
- β_1 is $k_1 + 1$ vector of parameters;
- X_2 is a $n \times k_2$ matrix;
- β_2 is k_2 vector of parameters;

and suppose that we want to test the following joint hypothesis on the β_2 parameters:

$$H_0 : \beta_2 = 0 \quad \text{against} \quad H_1 : \beta_2 \neq 0 \tag{239}$$

In which sense and why testing the joint hypothesis is different than the testing the k_2 separate hypotheses on the β_2 parameters?

The F test statistics

Consider the restricted regression

$$Y = X_1\beta_1 + U_r \tag{240}$$

and the unrestricted PRF 238. A natural starting point to construct a test statistic for the joint hypothesis is to see by how much the Sum of Squared Residuals (SSR) increases going from the restricted to the unrestricted PRF

The F statistic is built around this idea:

$$F = \frac{\frac{(SSR_r - SSR_{ur})}{k_2}}{\frac{SSR_{ur}}{n-k-1}} \sim F_{k_2,n-k-1} \tag{241}$$

where k_2 is the number of restrictions (the dimension of X_2) and k is the total number of parameters.

The F statistic is distributed accordint to an F distribution because it can be shown to be the ratio of two χ^2 distributions.

The F test statistics (cont.)

Note that the numerator of F is always positive and it is larger, the larger the reduction of SSR delivered by the unrestricted PRF.

We reject H_0, if our sample gives

$$|F| \geq c \tag{242}$$

where the critical level c is such that

$$Pr(f > c|H_0) = s \quad \text{with} \quad f \sim F_{k_2,n-k-1} \tag{243}$$

and s is the significance level (e.g. $s = 0.01$ or $s = 0.05$). The p-value is:

$$p = Pr(f > F|H_0) \tag{244}$$

Note that the F statistics can be construed not only for exclusion restrictions but also for more complicated linear restrictions, as long as we can specify the restricted and unrestricted PRF.

The “R-squared” form of the F test

In some cases it may be convenient to exploit the fact that

$$SSR_r = (1 - R_r^2) \tag{245}$$

$$SSR_{ur} = (1 - R_{ur}^2) \tag{246}$$

and therefore the F statistics can be expressed as a function of the R-squared of the restricted and unrestricted distribution:

$$F = \frac{\frac{(R_{ur}^2 - R_r^2)}{k_2}}{\frac{1 - R_{ur}^2}{n-k-1}} \sim F_{k_2, n-k-1} \tag{247}$$

This form of the test is completely equivalent but more convenient for computational purposes.

The F statistics and the overall significance of a regression

Most packages report the F test for the joint hypothesis that all the regressors have no effect:

$$H_0 : \beta = 0 \quad \text{against} \quad H_1 : \beta \neq 0 \tag{248}$$

In this case the restricted PRF is

$$y = \beta_0 + U_r \tag{249}$$

and the F test is

$$F = \frac{\frac{(R^2)}{k}}{\frac{1-R^2}{n-k-1}} \sim F_{k,n-k-1} \tag{250}$$

because the R-squared of the restricted PRF is zero.

The information of the F test statistics in this case is evidently the same of the R-squared statistic, but it is framed in a way that allows for a test on the significance of all the regressors.

4.2.4 Power of a test (intuition for future reference)

The significance of a test measures the probability of rejecting H_0 when it is true, i.e. the probability of “Type I” decision errors.

But to evaluate the usefulness of a test we need also to worry about “Type II” errors: i.e. failing to reject H_0 when some specific alternative is in fact true.

The “power of a test” is 1 minus the probability of Type II errors, i.e. the probability of not rejecting the alternative when it is true

Computing the power of a test requires defining a specific alternative and the distribution of the test statistic under H_1

In chosing between different possible tests we want the one that has more power, for any given level of significance.

4.3 Large sample distribution of the OLS-MM estimator $\hat{\beta}$

The advantage of a large sample is that we do not need to make any distributional assumption on the outcome Y.

In particular we do not have to assume MLR 6: normality of $U|X$.

This is particularly important from a methodological/philosophical point of view because it allows us to use all the machinery of regression analysis also in cases where normality is clearly a wrong assumption:

- Discrete dependent variables
- Limited dependent variables
- "Conditional on positive" models
- Duration analysis
- Count data analysis

Thanks to large samples, econometrics becomes considerably simpler!

4.3.1 Summary of the asymptotic theory results that we need

To derive the asymptotic distribution of the OLS-MM estimator we need the following results

- The Law of Large Numbers
- The Central Limit Theorem
- Slutsky's Theorem
- The Continuous Mapping Theorem
- The Delta Method

For further details and proofs see Angrist Pischke (2008) and Knight (2000).

Law of Large Numbers

Theorem 1. The Law of Large Numbers:
Sample moments converge in probability to the corresponding population moments.

In other words, the probability that the sample mean (or any other moment) is close to the population mean can be made as high as you like by taking a large enough sample.

Central Limit Theorem

Theorem 2. The Central Limit Theorem:
Sample moments are asymptotically Normally distributed after subtracting the corresponding population moment and multiplying by the square root of the sample size. The covariance matrix is given by the variance of the underlying random variable.

For example, in the case of the sample mean:

$$\sqrt{n}\left(\frac{\sum_{i=1}^{n} w_i}{n} - \mu\right) \quad \overset{d}{\rightarrow} \quad \text{Normal}(0, B) \tag{251}$$

where w_i is an i.i.d. random sample and $B = Var(w_i)$.

Note that without the multiplication by $\sqrt{n}$, the standardized moment would converge to zero.

In other words, in large enough samples, appropriately standardized sample moments are approximately Normally distributed.

Slutsky's Theorem

Theorem 3. Slutsky's Theorem

Part 1:
Let a_n *be a statistic with a limiting distribution and let* b_n *be a statistic with probability limit* b*. Then* $a_n + b_n$ *and* $a_n + b$ *have the same limiting distribution.*

Part 2:
Let a_n *be a statistic with a limiting distribution and let* b_n *be a statistic with probability limit* b*. Then* $a_n b_n$ *and* $a_n b$ *have the same limiting distribution.*

The continuous mapping theorem and Delta Method

Theorem 4. The continuous mapping theorem
The probability limit of $h(b_n)$ is $h(b)$ if $Plimb_n = b$ and $h(.)$ is continuous.

Theorem 5. The Delta Method
The asymptotic distribution of $h(b_n)$ is Normal with covariance matrix $\nabla h(b)'\Omega\nabla h(b)$, if $Plim\, b_n = b$, $h(.)$ is continuously differentiable at b with gradient $\nabla h(b)$, and b_n has asymptotic normal distribution with the covariance matrix Ω.

In other words, consider a vector-valued random variable that is asymptotically Normally distributed.

Most scalar functions of this random variable are also asymptotically Normally distributed, with covariance matrix given by a quadratic form with the covariance matrix of the random variable on the inside and the gradient of the function evaluated at the probability limit of the random variable on the outside.

The estimator under consideration: $\hat{\beta}$

Given the PRF $Y = X\beta + U$:

$$\begin{aligned} \hat{\beta} &= (X'X)^{-1}X'Y \\ &= \beta + (X'X)^{-1}X'U \\ &= \beta + \left(\frac{1}{n}\sum_{i=1}^{n} X_i' X_i\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} X_i' u_i\right) \end{aligned} \tag{252}$$

where X_i is the $1 \times k+1$ vector of the regressors observed for subject i.

The asymptotic distribution of $\hat{\beta}$ is the same as the distribution of

$$\sqrt{n}(\hat{\beta} - \beta) = \left(\frac{1}{n}\sum_{i=1}^{n} X_i' X_i\right)^{-1} \frac{1}{\sqrt{n}} \left(\sum_{i=1}^{n} X_i' u_i\right) \tag{253}$$

which can be determined applying the asymptotic results stated above to the sample moments on the right hand side of 253.

Consistency of $\hat{\beta}$

Consistency derives from the application to (252) of

- the Continuous Mapping Theorem;
- the Law of Large Numbers.

Exploiting the fact that probability limits pass through continuous functions and substituting population moment to sample moment, $\hat{\beta}$ converges in probability to:

$$\begin{aligned} \hat{\beta} &\xrightarrow{p} \beta + \left(E(X_i'X_i)\right)^{-1}\left(E(X_i'u_i)\right) \\ &= \beta \end{aligned} \tag{254}$$

where the last equality holds because $E(X_i'u_i) = 0$ by definition of the PRF.

Asymptotic distribution of $\hat{\beta}$

Applying Slutsky's Theorem to 253, $\sqrt{N}(\hat{\beta}-\beta)$ has the same distribution of

$$\left(E(X_i'X_i)\right)^{-1}\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}X_i'u_i\right) \tag{255}$$

Since

$$\frac{1}{n}\sum_{i=1}^{n}X_i'u_i \quad \xrightarrow{p} \quad E(X_i'u_i)=0 \tag{256}$$

then

$$\sqrt{N}\left(\frac{1}{n}\sum_{i=1}^{n}X_i'u_i\right) \quad \xrightarrow{d} \quad \text{Normal}(0,E(X_i'X_iu_i^2)) \tag{257}$$

because it is a root-n blown up and centered sample moment, for which we can use the Central Limit Theorem.

Note that $E(X_i'X_iu_i^2)$ is a $(k+1)\times(k+1)$ matrix.

Asymptotic distribution of $\hat{\beta}$ (cont.)

It then follows that:

$$\hat{\beta} \quad \xrightarrow{d} \quad \text{Normal}(\beta, [E(X_i'X_i)^{-1}][E(X_i'X_i)u_i^2][E(X_i'X_i)^{-1}]) \tag{258}$$

where note that $[E(X_i'X_i)^{-1}][E(X_i'X_i)u_i^2][E(X_i'X_i)^{-1}]$ is again a $(k+1) \times (k+1)$ matrix.

It is important to realize that to derive this result we have not assumed homoscedasticity.

We have only assumed to have identically and independently distributed random sample observations, which is necessary for CLT and LLN to hold.

These asymptotic standard errors are called "Robust", or "Huber - Eicker - White" standard errors (White (1980)) and provide accurate hypothesis tests in large sample with minimal assumptions.

Asymptotic distribution of $\hat{\beta}$ (cont.)

If we are willing to assume homoschedasticity then

$$E(u_i^2|X) = \sigma^2 \tag{259}$$

and the "Robust" variance covariance matrix in 258 simplifies to

$$\begin{aligned} [E(X_i'X_i)^{-1}][E(X_i'X_i)u_i^2][E(X_i'X_i)^{-1}] &= \\ [E(X_i'X_i)^{-1}][E(X_i'X_iE(u_i^2|X)][E(X_i'X_i)^{-1}] &= \\ \sigma^2[E(X_i'X_i)^{-1}][E(X_i'X_i)][E(X_i'X_i)^{-1}] &= \\ \sigma^2[E(X_i'X_i)^{-1}] & \end{aligned} \tag{260}$$

and

$$\hat{\beta} \quad \xrightarrow{d} \quad \text{Normal}(\beta, \sigma^2[E(X_i'X_i)^{-1}]) \tag{261}$$

We defer Section 5.2 a discussion of when and why we should use Robust Standard Errors, and we focus here on how to perform inference and hypothesis testing in large samples.

4.4 Large sample testing of an hypothesis

The classical hypothesis testing procedure that we have described for small samples extends to large samples with one important caveat.

Consider the test statistics:

$$t_{\hat{\beta}_j} = \frac{\hat{\beta}_j}{\hat{sd}(\hat{\beta}_j)} \tag{262}$$

where $\hat{sd}(\hat{\beta}_j)$ is now the sample counterpart of the asymptotic variance-covariance matrices in 258 or 261.

This test statistics is not distributed "exactly" like a Student's t_{n-k-1} because the numerator is not exactly normal but only approximately normal.

Practically this is not really a problem because in large samples the Student's t_{n-k-1} is almost not distinguishable from a Normal.

This is why we do not have to tell STATA whether we are in "large" or "small' samples!

4.5 A general format of a test: The Wald test

Given the PRF

$$Y = X\beta + U \tag{263}$$

let's now consider the most general formulation of an hypothesis concerning β:

$$H_0 : r(\beta) = q \quad \text{against} \quad H_1 : r(\beta) \neq q \tag{264}$$

where $r(.)$ is any function of the parameters and $r(\beta) - q$ is a $\rho \times 1$ vector, if ρ is the number of restrictions.

So H_0 and H_1 are systems of ρ equations if there are ρ restrictions.

Example :

$$H_0 : r(\beta) = R\beta = q \quad \text{against} \quad H_1 : r(\beta) = R\beta \neq q \tag{265}$$

where R is a $\rho \times k + 1$ matrix which charaterize the ρ restrictions on the parameters that we would like to test.

Exercise on the specification of a set of restrictions

Suppose that you are estimating the log of a Cobb Douglas production function in which output depends on labor and capital and you want to test:

- constant returns to scale;
- the return to one unit of labor is twice the return to one unit of capital;
- there exist neutral technological progress/regress.

What is R for these restrictions?

The logic of the Wald test

The logic of the test is that if the restrictions are valid the quantity $r(\hat{\beta}) - q$ should be close to 0 while otherwise it should be far away from 0.

The Wald form of the test statistic that captures this logic is

$$W = [r(\hat{\beta}) - q]'[Var(r(\hat{\beta}) - q)]^{-1}[r(\hat{\beta}) - q] \qquad (266)$$

In other words we want to evaluate how far away from 0 is $r(\hat{\beta}) - q$ after normalizing it by its average variability. Note that W is a scalar.

If $r(\hat{\beta}) - q$ is normally distributed, under H_0

$$W \sim \chi^2_\rho \qquad (267)$$

where the number of degrees of freedom ρ is the number of restrictions to be tested.

The difficulty in computing the test statistics is how to determine the variance at the denominator.

The variance of the Wald statistics

Using the Delta Method in a setting in which $h(\hat{\beta}) = r(\hat{\beta}) - q$

$$Var[r(\hat{\beta}) - q] = \left[\frac{\partial r(\hat{\beta})}{\partial \hat{\beta}}\right] [Var(\hat{\beta})] \left[\frac{\partial r(\hat{\beta})}{\partial \hat{\beta}}\right]' \tag{268}$$

where note that $\left[\frac{\partial r(\hat{\beta})}{\partial \hat{\beta}}\right]$ is a $\rho \times k + 1$ matrix and therefore $Var[r(\hat{\beta}) - q]$ is a $\rho \times \rho$ matrix.

Going back to the example in which $r(\hat{\beta}) - q = R\hat{\beta} - q$

$$Var[R\hat{\beta} - q] = R[Var(\hat{\beta})]^{-1}R' \tag{269}$$

and the Wald test is

$$W = [R\hat{\beta} - q]'[RVar(\hat{\beta})R']^{-1}[R\hat{\beta}) - q] \tag{270}$$

and $Var(\hat{\beta})$ is in practice estimated by substituting the sample counterparts of the asymptotic variance-covariance matrices in 258 or 261, depending on what we want to assume about homoschedasticity.

Exercise: Wald test and simple restrictions

Consider again the unrestricted regression in matrix form

$$Y = X_1\beta_1 + X_2\beta_2 + U_{ur} \tag{271}$$

where

- X_1 is a $n \times 2$ matrix including the constant;
- β_1 is dimension 2 vector of parameters;
- X_2 is a $n \times 1$ matrix;
- β_2 is dimension 1 vector of parameters;

and suppose that we want to test the following joint hypothesis on the β_2 parameters:

$$H_0 : \beta_2 = 0 \quad \text{against} \quad H_1 : \beta_2 \neq 0 \tag{272}$$

What is R in this case?

Exercise: Wald test and simple restriction (cont.)

It is easy to verify that in this case the Wald test is

$$\begin{aligned} W &= [R\hat{\beta} - q]'[RVar(\hat{\beta})R']^{-1}[R\hat{\beta}) - q] \\ &= \frac{\hat{\beta}_2^2}{Var(\hat{\beta}_2)} \end{aligned} \tag{273}$$

which is the square of a standard t-test, and is distributed as a χ^2 distribution

Other large sample testing procedures

The Wald test is a general form of a large sample test that requires the estimation of the unrestricted model.

There are cases in which this may be difficult or even impossible.

In the context of Maximum Likelihood estimation, an alternative large sample testing procedure is the Lagrange Multiplier test, which requires instead only the estimation of the restricted model.

A third alternative is the Likelihood Ratio test, which requires instead the estimation of both the restricted and the unrestricted models.

These other general sample testing procedure are left for future discussion in the context of Maximum Likelihood Estimation.

4.6 A Lagrange Multiplier test in the context of linear regression

In the simple context of linear regression we can define a LM test for multiple exclusion restrictions without having to relate it to Maximum Likelihood (but the name comes from that context!)

Consider again the unrestricted regression in matrix form

$$Y = X_1\beta_1 + X_2\beta_2 + U_{ur} \tag{274}$$

where

- X_1 is a $n \times k_1 + 1$ matrix including the constant;
- β_1 is dimension $k_1 + 1$ vector of parameters;
- X_2 is a $n \times k_2$ matrix;
- β_2 is dimension k_2 vector of parameters;

and suppose that we want to test the following joint hypothesis on the β_2 parameters:

$$H_0 : \beta_2 = 0 \quad \text{against} \quad H_1 : \beta_2 \neq 0 \tag{275}$$

A Lagrange Multiplier test in the context of linear regression (cont.)

Suppose to estimate the restricted PRF

$$Y = X_1\beta_{r1} + U_r \tag{276}$$

where the subscript r indicates that the population parameters and unobservables of this restricted equation may differ from the corresponding one of the unrestricted PRF.

It is intuitive to hypothesize that in the auxiliary regression

$$\hat{U}_r = X_1\gamma_1 + X_2\gamma_2 + V \tag{277}$$

if the restrictions in the primary PRF are valid then

$$H_0 : \beta_2 = 0 \quad \Rightarrow \quad \gamma_2 = 0 \tag{278}$$

A Lagrange Multiplier test in the context of linear regression (cont.)

Let the R-squared of the auxiliary regression 277 be R_U^2 and consider the statistics

$$LM = nR_U^2 \tag{279}$$

If the restrictions are satisfied, this statistics should be close to zero because;

- X_1 is by construction orthogonal to U_R and therefore $\gamma_1 = 0$;
- and $\gamma_2 = 0$ if the restrictions are satisfied.

Since, given k_2 exclusion restrictions:

$$LM = nR_U^2 \sim \chi^2_{k_2} \tag{280}$$

we can use the Classical testing procedure to test H_0.

5 Non-standard standard errors

We have considered instances in which the assumption of homoscedasticity

$$E(U^2|X) = \sigma^2 I_n \tag{281}$$

appears not plausible, but we have deferred so far a proper analysis of the consequences and the solutions to the violations of this assumption.

We are interested in two kinds of violations, that we study separately:

i. Heteroscedasticity:

$$E(u_i^2|X) = \sigma_i^2 \neq \sigma_j^2 = E(u_j^2|X) = \tag{282}$$

but $E(u_i u_j) = 0$ for $i \neq j$.

ii. Serial correlation and clustering

$$E(u_i u_j) \neq 0 \tag{283}$$

for some or all $i \neq j$ but $\sigma_i^2 = \sigma_j^2$ for $i \neq j$.

5.1 Heteroscedasticity

We should first realize that heteroscedasticity is likely to be the "rule" rather than the "exception".

Suppose that the CEF is non-linear and we approximate the CEF with a linear PRF. Note that

$$\begin{aligned} E[(y_i - X_i\beta)^2|X_i] &= E\{[y_i - E(y_i|X_i) + E(y_i|X_i) - X_i\beta]^2|X_i\} \quad (284) \\ &= E[y_i - E(y_i|X_i)|X_i]^2 + [E(y_i|X_i) - X_i\beta]^2 \\ &= V[y_i|X_i] + [E(y_i|X_i) - X_i\beta]^2 \end{aligned}$$

which indicates that even if $V[y_i|X_i]$ is constant, we can have heteroscedasticity because the linear PRF may be a better or a worse approximation of the non-linear CEF at different values of X_i.

If the CEF is non-linear, it is almost sure that the residual of a linear PRF will display heteroscedasticity.

Heteroscedasticity is a problem also with a linear CEF

Consider this emblematic case:

$$y_i = \begin{cases} 1 & \text{with probability} \quad Pr(y_i = 1) = P \\ 0 & \text{with probability} \quad Pr(y_i = 0) = 1 - P \end{cases} \tag{285}$$

Note that:

$$E(y_i) = 1P + 0(1 - P) = P \tag{286}$$

Assume that the CEF for this model is linear and therefore:

$$E(Y|X) = X\beta \tag{287}$$

Using this assumption:

$$\begin{aligned} Y &= E(Y|X) + (Y - E(Y|X)) \\ &= X\beta + \epsilon \end{aligned} \tag{288}$$

where

$$\epsilon = \begin{cases} 1 - X\beta & \text{with probability} \quad P \\ -X\beta & \text{with probability} \quad 1 - P \end{cases} \tag{289}$$

Advantages and disadvantages of the LPM

The LPM is:

- computationally simple and
- imposes little structure on the data

but its PRF is clearly heteroscedastic because

$$\begin{aligned} E(\epsilon) &= (1 - X\beta)P + (-X\beta)(1 - P) \\ &= (1 - X\beta)X\beta + (-X\beta)(1 - X\beta) = 0. \end{aligned} \tag{290}$$

but the variance is given by:

$$\begin{aligned} E(\epsilon^2) &= (1 - X\beta)^2 X\beta + (-X\beta)^2(1 - X\beta) \\ &= (1 - X\beta)X\beta \end{aligned} \tag{291}$$

Observations for which $P_i = X_i\beta$ is close to 1 or 0 have relatively low variance while observations with $P_i = X_i\beta$ close to .5 have relatively high variance.

Testing for heteroscedasticity and its consequences

Traditional econometrics offers several tests for heteroscedasticity (see some in Wooldridge Chapter 8).

However, for the reasons outlined above it is safer to start from the assumption that your model is heteroscedastic.

Fortunately the consequences of heteroscedasticity are not dramatic in most cases because it

- is irrelevant for unbiasedness of $\hat{\beta}$;
- is irrelevant for consistency of $\hat{\beta}$;

The only disturbing consequence of heteroscedasticity is that the variance-covariance matrix of $\hat{\beta}$ is biased if it is computed assuming homoscedasticity which is not true.

If this happens our inference and hypothesis testing are wrong.

The traditional solution: Generalized Least Squares

If we have reasons to assume that we know the functional form of the heteroscedasticity which affects the error term the solution consists in transforming the heteroscedastic PRF into an homoscedastic one. Consider the regression

$$y_i = X_i\beta + u_i \tag{292}$$

and assume that

$$Var(u_i|X_i) = h(X_i)\sigma^2 = h_i\sigma^2 \tag{293}$$

Consider now the transformed model

$$\frac{y_i}{\sqrt{h_i}} = \frac{X_i}{\sqrt{h_i}}\beta + \frac{u_i}{\sqrt{h_i}} \tag{294}$$

$$\tilde{y}_i = \tilde{X}_i\beta + \tilde{u}_i \tag{295}$$

The residual in the transformed PRF is no longer heteroscedastic.

$$Var(\tilde{u}_i|X_i) = \frac{1}{h_i}Var(u_i|X_i) = \sigma^2 \tag{296}$$

Generalized Least Squares (cont.)

Using matrix notation, assume that for the PRF:

$$Y = X\beta + U \tag{297}$$

the variance-covariance matrix of residuals is

$$E(UU'|X) = \sigma^2\Omega \tag{298}$$

where Ω is a $n \times n$ matrix. Note that this assumption include not only heteroscedasticity but also non zero covariance between disturbances of different observations, a topic on which we come back below.

It is possible to define an invertible square matrix P such that

$$\begin{aligned} PP' &= \Omega \\ (PP')^{-1} &= P'^{-1}P^{-1} = \Omega^{-1} \end{aligned} \tag{299}$$

Generalized Least Squares (cont.)

If we now premultiply the original PRF by P^{-1}

$$\begin{aligned} P^{-1}Y &= P^{-1}X\beta + P^{-1}U \\ \tilde{Y} &= \tilde{X}\beta + \tilde{U} \end{aligned} \tag{300}$$

we obtain an homoscedastic model because

$$\begin{aligned} E(\tilde{U}\tilde{U}'|X) &= E(P^{-1}UU'P'^{-1}) \\ &= P^{-1}E(UU'|X)P'^{-1} \\ &= \sigma^2 P^{-1}\Omega P'^{-1} \\ &= \sigma^2 P^{-1}PP'P'^{-1} \\ &= \sigma^2 I_n \end{aligned} \tag{301}$$

and the OLS-MM estimator for this model is called GLS (Aitken) estimator

$$\hat{\beta}_{GLS} = [X'\Omega^{-1}X]^{-1}X'\Omega^{-1}Y \tag{302}$$

Generalized Least Squares (cont.)

It is easy to show that the GLS estimator is unbiased

$$E(\hat{\beta}_{GLS}|X) = \beta + [X'\Omega^{-1}X]^{-1}X'\Omega^{-1}E(U|X) = \beta \qquad (303)$$

if the the OLS-MM estimator is unbiased, and of course it ensures the correct estimation of the variance-covariance matrix for inference and hypothesis testing

The problem is that the matrix Ω need to be estimated, if it is not known, in order to make the GLS transformation feasible, which typically means:

- Run the original PRF to obtain inefficient but consistent estimates of the residuals;
- Use these estimated residuals to estimate Ω;
- Apply the GLS transformation using $\hat{\Omega}$.

But this procedure, gives only consistent estimates because Ω is estimated.

Nowadays, the use of “Robust standard errors” is preferred to GLS.

Generalized Least Squares and Linear Probability Models

In cases like the LPM, it is not even advisable to use the Feasible GLS transformation.

Predicted probabilities

$$\hat{P}_i = X_i \hat{\beta}$$

may lie outside the [0,1] range

This may produce non-sense probabilities for forecasting purposes and negative estimated variances so that GLS cannot be implemented.

Moreover estimates are also very sensitive to extreme realizations of the X variables.

In cases like this Robust Standard Errors are the only solution.

Robust Standard Errors

We derived the asymptotic distribution of $\hat{\beta}$ without assuming homoscedasticity

$$\hat{\beta} \quad \xrightarrow{d} \quad \text{Normal}(\beta, [E(X_i'X_i)^{-1}][E(X_i'X_i)U_i^2][E(X_i'X_i)^{-1}]) \tag{304}$$

and this variance-covariance matrix gives standard errors that are robust to deviations from homoscedasticity.

These Standard Errors are typically larger than the ones derived under homoscedasticity, but for practical purposes, we should not expect differences larger than 30%. Angrist and Pischke suggest that larger differences are a sign of more serious problems (e.g. programming errors).

Note also that in small sample they may be smaller than the ones derived under homoschedasticity . Again when this happens it may be a sign of more serious problems.

See Angrist and Pischke for simulations and a discussion.

5.2 Clustering and the Moulton problem

The clustering of observations may originate much more serious problems than heteroschedasticity.

Suppose that we are interested in the model

$$y_{ig} = \beta_0 + \beta_1 x_g + u_{ig} \tag{305}$$

where:

- y_{ig} is the outcome of subject i in group j: for example the textscore of a students in a class.
- x_g is a determinant of the outcome that changes only across groups: for example class size.
- u_{ig} is the unobservable component which is likely to be correlated across subjects within the same group.

Even if classize where randomly assigned (e.g. Krueger (1999)), we cannot assume independence of observations within groups.

Modeling dependence across observations

Following Angrist and Pischke we can for example assume that

$$E(u_{ig}u_{jg}) = \rho\sigma_u^2 > 0 \tag{306}$$

where ρ is the intra-class correlation coefficient and σ^2 is the residual variance.

This covariance structure may originate from the assumption that

$$u_{ig} = v_g + \eta_{ig} \tag{307}$$

where both these components are assumed to be homoschedastic in order to emphasize the correlation problem.

Moulton (1986) shows that this error structure can increase standard errors by a considerable amounts and should not be neglected.

Equal group size and constant within group regressors

Consider first the simpler case in which all groups are of the same size n and the regressor is fixed within groups.

Moulton (1986) shows that the ratio between the variance that takes into account the correlation across observations, $Var(\hat{\beta})$, and the conventional variance $Var_c(\hat{\beta})$ which assumes zero correlation is

$$\frac{Var(\hat{\beta}_1)}{Var_c(\hat{\beta}_1)} = 1 + (n-1)\rho \tag{308}$$

The square root of this term is called "the Moulton factor" in the literature:

- obviously no problem if $\rho = 0$ or $n = 1$;
- if $\rho = 1$ observations are duplicated but contain no additional information;
- with $n = 100$ and and $\rho = .1$ the Moulton factor is around 3.

The problem is even more serious for the realistic case in which the group sizes change and the regressors change within groups.

Variable group size and regressors within group regressors

In the more general case

$$y_{ig} = \beta_0 + \beta_1 x_{ig} + u_{ig} \tag{309}$$

the ratio is

$$\frac{Var(\hat{\beta}_1)}{Var_c(\hat{\beta}_1)} = 1 + \left[\frac{V(n_g)}{\bar{n}} + \bar{n} - 1\right] \rho_x \rho \tag{310}$$

where:

- n_g is the size of group g;
- $\bar{n}$ is the average group size;
- ρ_x is the intragroup correlation of x_{ig} defined as:

$$\rho_x = \frac{\sum_g \sum_{i \neq k} (x_{ig} - \bar{x})(x_{kg} - \bar{x})}{Var(x_{ig}) \sum_g n_g(n_g - 1)} \tag{311}$$

  which measure the extent to which the factor x changes in a similar way for individuals in the same group. The extreme case of $\rho_x = 1$ is the case of a factor that does not change within group.

What makes the Moulton problem more serious

Clearly the Moulton problem is more serious when

- the correlation ρ_x of the regressor within group is larger;
- the group size is more variable;
- the average group size is larger, which, for given sample size means that
- the number of clusters is smaller.

For a given sample size, fewer clusters imply that there is less independent information in the data, and this reduces the precision of thr estimates.

For a given number of clusters, it does not pay much to increase sample size within groups unless the correlation of the regressor within groups is really low.

With constant regressors within group, adding more observations within groups with the same number of clusters does not help.

An illustration of the problem using Kruger (1999)

The study estimates the effects of class size on children's percentile test scores, finding $\hat{\beta} = -0.62$ with a conventional robust standard error of 0.09.

The parameters to evaluate the Moulton factor are

- $\rho_x = 1$ because class size does not change within a class;
- $V(n_g) = 17.1$ because class size changes across between classes;
- the intraclass correlation of residuals is $\rho = .31$;
- the average class size $\bar{n} = 19.4$

which gives

$$\frac{Var(\hat{\beta}_1)}{Var_c(\hat{\beta}_1)} \approx 7 \tag{312}$$

and a Moulton factor of $\sqrt{7} = 2.62$.

The correct standard error is therefore 0.24, almost three times larger.

Some solutions to the Moulton problem

- *Clustered standard errors*

  Liang and Zeger (1986) generalize the robust Variance-Covariance matrix of White (1986) to take into account intra-group correlation:

$$Var(\hat{\beta}) = (X'X)^{-1}[\sum_g (X_g'\hat{\Psi}_g X_g)](X'X)^{-1} \tag{313}$$

$$\hat{\Psi}_g = a\hat{U}_g\hat{U}_g' = \begin{bmatrix} \hat{u}_{1g}^2 & \hat{u}_{1g}\hat{u}_{2g} & \cdots & \hat{u}_{1g}\hat{u}_{ng} \\ \hat{u}_{2g}\hat{u}_{1g} & \hat{u}_{2g}^2 & \cdots & \hat{u}_{2g}\hat{u}_{ng} \\ \vdots & \vdots & \vdots & \vdots \\ \hat{u}_{ng}\hat{u}_{1g} & & \cdots & \hat{u}_{ng}^2 \end{bmatrix}. \tag{314}$$

  X_g is the matrix of regressors for group g and a is a degrees of freedom adjustment factor

  This estimator is consistent when the number of groups increases, but is not consistent when group size increases and the number of groups is fixed.

  With a small number of clusters this solution may not be reliable.

- *Parametric solution*
  Use equation 310 to correct manually the standard errors; The Stata commands "loneway" gives the necessary ingredients.

- *Group averages*
  Use group averages, i.e. estimate:

$$\bar{y}_g = \beta_0 + \bar{x}_g \beta_1 + \bar{u}_g \tag{315}$$

  where upper bars denote the within cluster averages of the corresponding variables.

  Standard errors are asymptotically consistent with respect to the number of groups (not group size)

  Group size plays a useful for fixed number of clusters because when it increases group averages are closer to be normally distributed.

  In particular $\bar{u}_g$ is close to normal, improving the small sample properties of the regression.

- *Group averages with micro-covariates*
  If the model is

$$y_{ig} = \beta_0 + X_g\beta_1 + W_{ig}\delta + u_{ig} \tag{316}$$

where W_{ig} is a vector of covariates that changes within clusters, one can proceed in two steps

i. In step 1 estimate:

$$y_{ig} = \mu_g + W_{ig}\delta + u_{ig} \tag{317}$$

where μ_g are dummies for the different clusters. Note that given 307 and 316

$$\mu_g = \beta_0 + X_g\beta_1 + v_g \tag{318}$$

ii. In step 2 estimate

$$\hat{\mu}_g = \beta_0 + X_g\beta_1 + \{v_g + (\mu_g - \hat{m}u_g)\} \tag{319}$$

which needs GLS, for efficiency, using the inverse of the estimated variance of the group level residual $v_g + (\mu_g - \hat{m}u_g)$ as weight. (See Angrist Pischke for the problems that this may cause with few clusters).

A final note on the problem of clustering

Interestingly because of this problem, the large sample size of typical microeconometric studies may not help much!

Clustering is a pervasive and serious problem in microeconometrics.

A sort of ... revenge of macroeconometricians!

6 Miscellaneous topics from Wooldridge

- Effects of data scaling on OLS statistics.
- Use of logs.
- Polynomial models.
- Goodness of fits and selection of regressors.
- Dummy variables as regressors.
- Interactions (in general and with dummmy variables
- Testing for differences across groups
- Measurement error
- Missing data
- Outliers and influential observations
- Non random sampling

7 References

Angrist, Joshua D. and Pischke, Steve (2008), “Mostly Harmless Econometrics: An Empiricist’s Companion” Princeton University Press, forthcoming.

Green, William (200?), “Econometric Analysis” MacMillan.

Knight, ??? (2000), “Mathematical statistics” ????.

Angrist, Joshua D., Guido W. Imbens and Donald B. Rubin (1996), “Identification of Causal Effects Using Instrumental Variables”, *Journal of the American Statistical Association* 91, 444–472.

Holland, Paul W. (1986), “Statistics and Causal Inference”, *Journal of the American Statistical Association* 81, 945–970.

Krueger, Alan B. (1999) “Experimental Estimates of Education Production Functions.” *Quarterly Jour- nal of Economics*, 114:497-532.

Liang, K. and Scott L. Zeger (1986, “Longitudinal Data Analysis Using Generalized Linear Models ” *Bio- metrika* 73, 13-22.

Mood, Alexander, Franklin Graybill and Duane Boes, (1974) “Introduction to the theory of Statistics”, McGraw Hill, 3d. Edition.

Moulton, Brent. 1986. “Random Group Eects and the Precision of Regression Estimates”, *Journal of Econometrics*, 32, pp. 385-397.

Rubin, Donald B. (1991), “Practical Implications of Modes of Statistical Inference for Causal Effects and the Critical Role of the Assignment Mechanism.” *Biometrics*, 47:1213-34.

White, Halbert. (1980). “A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity” *Econometrica*, 48:817-38.

Wooldridge, Jeffrey M. (2007), “Introductory Econometrics: A modern approach” Thomson South Western.