ECONOMETRICS

DAVID SALAZAR

ÍNDICE GENERAL

1

INTRODUCTION

1.1 CAUSAL RELATIONSHIPS AND CETERIS PARIBUS

Economic questions are almost always causality questions. What is the effect of raising taxes on entrepreneurship? ... That is, we wish to determine whether a change in one variable, say w, causes a change in another variable, say y.

The fundamental problem is that those variables are not in a void; w is probably correlated with other variables that also have an effect on y, and thus just observing a correlation of w with y is not enough. Then it enters the notion of ceteris paribus: holding all other relevant factors fixed, what's the effect of increasing w on y?. That is: we wish to estimate the partial effect of w on y: $\frac{\partial E(y|w,c)}{\partial w}$. To do this we use econometric methods.

After deciding on what factors should we control for, estimating the partial effect of w on y is relatively straightforward. The problem is deciding which variables should we use to control for, that is, which controls are relevant and which are not. Even then, we can discuss if the variables we're using are accurately measuring what we hope; or we only observe equilibrium values of systems that are simultaneously co-determined.

1.2 STOCHASTIC SETTING AND ASYMPTOTIC ANALYSIS

A conceptual setting that gives us a language to understand the empirical economic modeling.

For much of this book we adopt a random sampling assumption. More precisely, we assume that (1) a population model has been specified ($E(y|w,c)$ and a correspondent population of interest) and (2) an independent, identically distributed (i.i.d.) sample can be drawn from the population.

An important virtue of the random sampling assumption is that it allows us to separate the sampling assumption from the assumptions made on the population model. On the other hand, random sampling

entails that the explanatory variables are not fixed, as that would correspond on an experimental setting.

That is, all assumptions should be stated in terms of a population model (relationship between predictor and outcome variable, conditional independence, heteroskedasticity). In particular, it's useful to formulate the population model for a generic draw from the population. This makes YOU THINK about the crucial relationships in the population that will in turn determine how you must perform statistical inference.

1.3 CONDITIONAL EXPECTATIONS

1.3.1 *Role of Conditional expectations*

A substantial portion of research in econometric methodology can be interpreted as finding ways to estimate conditional expectations in the numerous settings that arise in economic applications. Conditional expectations that allow us to infer causality; these CE's are called structural conditional expectations. Once the necessary controls have been observed from a random sample of the population of interest, and all the relevant variables have been accurately measured, estimation of the structural conditional expectation is straightforward.

Possible problems:

- Measurement error.
- Variables are simultaneously determined.
- Not possible to get a random sample.
- Necessary controls cannot be directly observed.

Then, what we have is a set of identification assumptions under which we can recover the structural conditional expectation.

1.3.2 *Features of Conditional Expectations*

A conditional function is simply a function of what you're conditioning on; specifically, a function that tells you how the average of y changes as x changes. That is:

$$E(Y|X) = \mu(X)$$

To find it, simply take the average of y but now weighting by the conditional density $f(Y|X)$ instead of the unconditional density $f(Y)$. That is:

$$E(Y|X) = \int y f_{Y|X}(y|X)dy$$

Properties of the conditional expectation:

- If X and Y are independent, then: $E(Y|X) = E(Y)$
- $E(Yh(X)|X) = h(X)E(Y|X)$
- Linearity.
- Iterated Expectations: $E(E(Y|X)) = E(Y)$. Intuitively, the average (weighting by different values of X, as the conditional expectation is a function of X) of the average of Y as X changes is the average of Y.

1.3.3 *Conditional Expectations: Parametric Form*

$$y = E(Y|X) + u$$
$$E(u|X) = 0$$

That is, we can always decompose y as its conditional expectation plus an error term that has conditional mean zero.

The fact that the error term has a conditional mean of zero has the following implications:

- The error is also uncorrelated with any function of the explanatory variables.
- From the law of iterated expectations:

$$E(E(u|X)) = E(u) = 0$$

And equivalent way of stating $E(u|X) = 0$ is to say that we have the functional form of $E(Y|X)$ properly specified. That is, if we suppose that y is determined by the following simple model:

$$y = \beta_0 + \beta_1 x_1 + u$$

β_1 will only represent $\frac{\partial E(Y|X)}{\partial x_1}$ if $E(u|X) = 0$.

1.3.4 *Properties*

Remember: a conditional expectation is a function of what you're conditioning on, telling you how the average of the function changes as what you're conditioning on changes. You can think of it as the best prediction (defined by a specific cost function) of prediction based on what you're conditioning on. Then:

$$E(E(Y|X,Z)|X) = E(Y|X)$$

That is, if you're trying to predict the conditional expectation of Y given X, Z, then the best you can do is to use the conditional expectation $E(Y|X)$.

1.3.5 *Missing controls*

To make causal inferences in econometrics: to recover a structural conditional expectations. However, sometimes you cannot recover all the relevant controls: call those controls Z. Then, we can only recover $E(Y|X)$. The former derivation can helps us to relate our function of interest and our estimated function, and to check when can we recover one from the other. Then, let Y be determined as follows:

$$y = \beta_0 + \beta_1 x_1 + \gamma_1 z + u$$
$$E(y|x,z) = \beta_0 + \beta_1 x_1 + \gamma_1 z$$

Then, what we can recover is the following:

$$E(y|x) = E(E(y|x,z)|x) = \beta_0 + \beta_1 x_1 + E(\gamma_1 z|x)$$

1.3.6 *Linear in the parameters*

Any conditional expectation linear in parameters can be written as a conditional expectation linear in parameters and linear in some conditioning variables. That is, once you condition on $exper$, conditioning on $exper^2$ is unnecessary.

1.3.7 *Error Term*

If u is a random variable independent of the random vector x, then $E(u|x) = E(u)$. Thus, if $E(u) = 0$ and u, x are independent : $E(u|x) = 0$.

1.3.8 *Average Partial Effects and Unobserved Heterogeneity*

When we explicitly allow the expectation of the response variable, y, to depend on unobservables, usually called unobserved heterogeneity, we must be careful in specifying the partial effects of interest.

This particularly problematic when we have a nonlinear model with unobserved heterogeneity, such that the partial effect of interest depends on the unobserved heterogeneity. That is:

$$\frac{\partial E(y|x,q)}{\partial x} = f(x,q)$$

Then, if we cannot observe q, we cannot estimate the partial effect of interest. What can we do? Estimate the average partial effect! That is, average the partial effect across q (q's probability density function: $g(t)$.

$$APE = \int f(x,q)g(t)dt = E_q(\frac{\partial E(y|x,q)}{\partial x})$$

Thus, now we have a problem of estimating the density function of q ($g(t)$) without observing q. How can we do this? Enter the room a good proxy variable such that, conditional on the proxy variable, x, q are independent. That is:

$$g(t|x,w) = g(t|w)$$
$$E(y|x,q,w) = E(y|x,q)$$

The last equation just says that in predicting y, after observing (x,q), w is useless.

Then, we can show the following:

$$f(x,q) = E_w(\frac{\partial E(y|x,w)}{\partial x})$$

That is: we can recover the average partial effect of x on y by averaging over the partial effect of x, holding w constant.

2

BASIC ASYMPTOTIC THEORY

Asymptotic analysis is concerned with the various kinds of convergence of sequences of estimators as the sample size grows. That is, around what values the probability mass of the estimator tend to accumulate?

2.1 DEFINITIONS

Some properties that also apply to non-stochastic sequences of numbers.

2.1.1 *Convergence*

The sequence $\{a_N\}$ converges to a iff from some initial value, the difference between the sequence and the value is as small as required. That is, $\exists N_\epsilon$ such that $\forall N > N_\epsilon$, $|a - a_N| < \epsilon$. If we don't have to apply any transformation, we say the $a_N = o(1)$

2.1.2 *Bounded Sequences*

A sequence is bounded iff exists some value such that the sequence is always smaller than the value. That is: $\exists b$ such that $\forall N$, $|a_N| \leq b$. If we don't have to apply any transformation, we say the $a_N = O(1)$

2.1.3 *Relationship*

From the definition it's clear that if the series converges, then the series is bounded. But the converse is not true. Even more: if $a_N = o(1) \rightarrow a_N = O(1)$. That is, if the series converges without need of any transformation, then the untransformed sequence is also bounded.

2.2 DEFINITIONS IN PROBABILITY

2.2.1 *Convergence*

A sequence of random variables converges in probability to a iff the probability that the sequence deviates from the limit by an arbitrarily small amount goes to zero as $N \to \infty$. That is, $\forall \epsilon > 0$

$$P(|x_N - a| > \epsilon) \to 0, \text{ as } N \to \infty$$

That is, the probability mass of the random variable accumulates around (for any desired distance) the convergence limit as the sample size increases. If the sequence converges to zero, we write $x_N = o_p(1)$

2.2.2 *Bounded Random Variable*

A sequence of random variable is bounded in probability iff the probability that the sequence is, in absolute value, greater than the bounded can be made arbitrarily small from a specific sample values. That is:

$$P(|x_N| > b_\epsilon) < \epsilon, \ \forall \epsilon > 0, \ N > N_\epsilon$$

When the sequence of random variable is bounded in probability, we write $x_N = O_p(1)$

2.2.3 *Relationships*

If the sequence of random variables converges in probability, then the sequence of random variables is also bounded in probability. That is:

$$\text{if } x_N \xrightarrow{p} a, \text{ then } x_N = O_p(1)$$

Some other useful relationships. Let c_N be a non-stochastic sequence. Then:

- $c_N = O(1)$ iff $c_N = O_p(1)$
- $c_N = o(1)$ iff $c_N = o_p(1)$

2.2.4 *Transformations*

- $x_N = o_P(a_N)$ iff $\frac{x_N}{a_N} \xrightarrow{p} 0$.
- $x_N = O_P(a_N)$ iff $\frac{x_N}{a_N} = O_p(1)$.

2.3 RULES

The following rules will be useful:

- The sum of sequences of random variables that converge in probability to zero is also a sequence of random variables that converge to zero. That is:

$$o_p(1) + o_p(1) = o_p(1)$$

- The sum of sequences of random variables that are bounded in probability is also bounded in probability.

$$O_p(1) + O_p(1) = O_p(1)$$

- The multiplication of sequences of random variables that are bounded in probability is also bounded in probability.

$$O_p(1)O_p(1) = O_p(1)$$

- The multiplication of a sequence of a random variable that converges in probability to zero, by a sequence that is bounded in probability, still converges in probability to zero.

$$o_p(1)O_p(1) = o_p(1)$$

- The sum of a sequence of random variable that converges to zero plus a sequence of random variable that is bounded in probability it's also bounded in probability.

$$o_p(1) + O_p(1) = O_p(1)$$

- The multiplication of a sequences that converge in probability to zero also converges to zero.

$$o_p(1)o_p(1) = o_p(1)$$

In Matrix Form:

MATRIX TIMES VECTOR Let $Z = o_p(1)$ and $x_N = O_p(1)$. Then: $Zx_N = o_p(1)$

INVERSE If the square matrix Z: $Z \xrightarrow{p} A$, and A is invertible. Then:

$$Z_n^{-1} \text{ exists}$$

$$Z_n^{-1} \xrightarrow{p} A^{-1}$$

2.4 SLUTSKY'S THEOREM

Let $x_N \xrightarrow{p} c$ and let g be a continuous function. Then: $g(x_N) \xrightarrow{p} g(c)$

Slutsky's theorem is perhaps the most useful feature of the plim operator: it shows that the plim passes through nonlinear functions, provided they are continuous. The expectations operator does not have this feature, and this lack makes finite sample analysis diffcult for many estimators

2.5 CONVERGENCE IN DISTRIBUTION

Definition: $x_N \xrightarrow{d} x \sim Normal(\mu, \sigma^2)$

2.5.1 *Distribution to Bounded*

If a sequence of random variable converges in distribution to another random variable, then the series is bounded in probability. That is:

$$\text{if } x_N \xrightarrow{d} x, \text{ then } x_N = O_p(1)$$

The former argument will be very useful, as often is easier to check a sequence of random variable converges in distribution than to prove that is bounded in probability.

2.5.2 *Continuous Mapping Theorem*

Once we know the limiting distribution of x_N, we can know the limiting distribution of many interesting functions of the original sequence. That is, if g is continuous, then:

$$\text{if } x_N \xrightarrow{d} x, \text{ then } g(x_N) \xrightarrow{d} g(x)$$

In Matrix Form:

MATRIX TIMES VECTOR if $x_N \xrightarrow{d} x \sim MNormal(0, V)$ then $A'x_N \xrightarrow{d} z \sim MNormal(0, A'VA)$

2.6 ASYMPTOTIC EQUIVALENCE

If the subtraction of two sequences of random variables converges to zero, then the two of them have the same distribution. That is:

if $x_N - z_N = o_p(1)$ and $z_N \xrightarrow{d} z$ then $x_N \xrightarrow{d} z$

2.7 DELTA METHOD

Suppose we have an estimator such that it is $\sqrt{N}$ consistent. Thus:

$$\sqrt{N}(\hat{\theta} - \theta) \xrightarrow{d} N(0, \sigma^2)$$

Then, we can say that:

$$\hat{\theta} \xrightarrow{p} \theta$$

The detla method says that if we are estimating a function of our parameter, then:

$$\sqrt{N}(f(\hat{\theta}) - f(\theta)) \xrightarrow{d} N(0, f'(\theta)\sigma^2 f'(\theta))$$

By slutsky or the definition:

$$f(\hat{\theta}) \xrightarrow{p} f(\theta)$$

Parte I

LINEAR MODELS

3

SINGLE EQUATION LINEAR MODEL

In Econometrics, we're interested in structural models: models that represent a causal relationship, as opposed to a relationship that simply captures statistical associations. Sometimes the structural model is directly estimable; sometimes it isn't. In the latter case, our task will be to recover the structural model from an estimable model.

To do so, we start with a population model assuming a linear functional form:

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u$$

To obtain the population regression function we must make the conditional independence assumption.

$$E(u|x) = 0 \tag{A1.}$$

Then, by (**??**), the population regression model is the following:

$$E(y|x) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$$

Of course, Econometrics gets interesting when we cannot recover the population regression function from a *i.i.d*. sample from the population of interest. Then, not all the variables of our interest will be exogenous, i.e., uncorrelated with u. In applied econometrics, endogeneity usually arises in one of three ways:

OMITTED VARIABLES Omitted variables are an issue when we would like to control for one or more additional variables but, usually because of data unavailability, we cannot include them in a regression model. Thus, if the omitted variable is correlated with some of the other variables in the population model, these variables will be endogenous.

MEASUREMENT ERROR If in our population model there is a variable that we cannot accurately measure, we'll be introducing some error into our estimable model. Thus, depending on the as-

sumptions we make on the measurement error, our estimated variable may or may not be endogenous.

SIMULTANEITY Simultaneity arises when at least one of the explanatory variables is determined simultaneously along with y. For example, $x_k = f(y, \ldots)$. Thus, when x_k varies we don't know if we must attribute the corresponding observed change in y to x_k or if instead the change in x_k must be attributed to y. Conceptually, this is a more difficult situation to analyze, because we must be able to think of a situation where we could vary x_k exogenously, even though in the data that we collect y and x_k are generated simultaneously.

3.1 ASYMPTOTIC PROPERTIES OF OLS

A population model is specified as follows:

$$y = X\beta + u$$

Note that by definition we can make the $E(u) = 0$

ASSUMPTION # 1 $E(x'u) = 0$. This is an assumption of the joint distribution of both (x, u) in the population and has nothing to do with the relationships in the sample data. Interestingly:

$$cov(x, u) = E(x'u) - E(x)E(u) = E(x'u) = 0$$

Thus, this assumption is equivalent to saying: u has mean zero and is uncorrelated with each regressor.

ASSUMPTION # 2 rank $E(x'x) = k$.

With the former two assumptions, we can conclude that, under OLS, the parameter β is identified. In the context of models that are linear in the parameters under random sampling, identification of β simply means that β can be written in terms of population moments in observable variables. That is, let's start by deriving the population β

$$\begin{aligned} Y &= X\beta + u \\ X'Y &= X'X\beta + X'u \\ E(X'Y) &= E(X'X)\beta + E(X'u) \\ E(X'X)^{-1}E(X'Y) &= \beta \end{aligned}$$

Which has its sample analogous with the following:

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

Alternatively:

$$\begin{aligned} argmin(\hat{\beta})u'u &= argmin(\hat{\beta})(Y - X\hat{\beta})'(Y - X\hat{\beta}) \\ &= (Y' - \hat{\beta}'X')(Y - X\hat{\beta}) \\ &= Y'Y - Y'X\hat{\beta} - \hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta} \end{aligned}$$

Then note that $Y'X\hat{\beta} = (\hat{\beta}'X'Y)'$. And note that it's a scalar such that we can re-write the last equation as:

$$Y'Y - 2\hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta}$$

If we derive w.r.t. $\hat{\beta}$, we'll have:

$$\begin{aligned} -2X'Y + 2X'X\hat{\beta} &= 0 \\ 2X'X\hat{\beta} &= 2X'Y \\ X'X\hat{\beta} &= X'Y \\ (X'X)^{-1}X'X\hat{\beta} &= (X'X)^{-1}X'Y \\ \hat{\beta} &= (X'X)^{-1}X'Y \end{aligned}$$

3.1.1 *Consistency*

If we multiply and divide by N:

$$\begin{aligned} \hat{\beta} &= (N^{-1}X'X)^{-1}N^{-1}X'Y \\ \hat{\beta} &= (N^{-1}X'X)^{-1}N^{-1}X'(X\beta + u) \\ \hat{\beta} &= (N^{-1}X'X)^{-1}N^{-1}X'X\beta + (N^{-1}X'X)^{-1}N^{-1}X'u \\ \hat{\beta} &= \beta + (N^{-1}X'X)^{-1}(N^{-1}X'u) \\ \hat{\beta} - \beta &= (N^{-1}\sum^{N} x_i'x_i)^{-1}(N^{-1}\sum^{N} x_i'u_i) \end{aligned}$$

Let's analyze the right hand side terms:

$(N^{-1}\sum^{N} x_i'x_i)^{-1}$: by WLLN, $N^{-1}\sum^{N} x_i'x_i \xrightarrow{p} E(X'X)$. Then, by theorem: $(N^{-1}\sum^{N} x_i'x_i)^{-1} \xrightarrow{p} [E(X'X)]^{-1}$. Then, as the sequence $(N^{-1}\sum^{N} x_i'x_i)^{-1}$ converges in probability, it's also bounded in probability. Thus, $N^{-1}\sum^{N} x_i'x_i)^{-1} = O(1)$

$(N^{-1}\sum^N x_i'u_i)$: by WLLN, $(N^{-1}\sum^N x_i'u_i) \xrightarrow{p} E(X'u)$. By conditional independence assumption, $E(X'u) = 0$. Then, $(N^{-1}\sum^N x_i'u_i) = o(1)$.

Then, by theorem, $o(1)O(1) = o(1)$. Thus,

$$(N^{-1}\sum^N x_i'x_i)^{-1}(N^{-1}\sum^N x_i'u_i) \xrightarrow{p} 0$$

Thus, $\hat{\beta} - \beta \xrightarrow{p} 0$. Thus, $\hat{\beta}$ is a consistent estimator of β.

3.1.2 *Asymptotic Normality*

First, let's notice that the last expression can be written in the following way:

$$\hat{\beta} - \beta = (N^{-1}\sum^N x_i'x_i)^{-1}(N^{-1}\sum^N x_i'u_i)$$
$$\sqrt{N}(\hat{\beta} - \beta) = (N^{-1}\sum^N x_i'x_i)^{-1}(\frac{1}{\sqrt{N}}\sum^N x_i'u_i)$$

Then let's notice a couple of things about the expressions:

$(N^{-1}\sum^N x_i'x_i)^{-1}$ Por WLLN y teorema:$(N^{-1}\sum^N x_i'x_i)^{-1} \xrightarrow{p} [E(X'X)]^{-1}$. Luego, por equivalencia asintótica:

$$(N^{-1}\sum^N x_i'x_i)^{-1} - [E(X'X)]^{-1} = o_p(1)$$
$$(N^{-1}\sum^N x_i'x_i)^{-1} = o_p(1) + [E(X'X)]^{-1}$$

$(\frac{1}{\sqrt{N}}\sum^N x_i'u_i)$ Por supuesto de la muestra *i.i.d.*, $\{(x_i, u_i)\}$ es también *i.i.d.* Por supuesto de independencia condicional y la WLLN: $E((\frac{1}{\sqrt{N}}\sum^N x_i'u_i)) = 0$. Además, si suponemos homoscedasticidad, también podemos suponer que $Var(x_i, u_i) < \infty$. Luego, por Th. Lindeberg-Levy, $\{(x_i, u_i)\}$ cumple con las condiciones del CLT y por lo tanto:

$$(\frac{1}{\sqrt{N}}\sum^N x_i'u_i) \xrightarrow{d} M \sim N(0, B)$$

Con supuesto de errores esféricos:

$$B = Var(x_i, u_i) = E(u_i^2 x_i'x_i) = \sigma^2 E(x_i'x_i)$$

Por lo tanto,

$$\sqrt{N}(\hat{\beta}-\beta) = (o_p(1) + [E(X'X)]^{-1})^{-1}(\frac{1}{\sqrt{N}}\sum^{N} x_i'u_i)$$

$$\sqrt{N}(\hat{\beta}-\beta) = o_p(1)(N^{-1}\sum^{N} x_i'u_i) + [E(X'X)]^{-1}(\frac{1}{\sqrt{N}}\sum^{N} x_i'u_i)$$

$$\sqrt{N}(\hat{\beta}-\beta) = o_p(1) + [E(X'X)]^{-1}(\frac{1}{\sqrt{N}}\sum^{N} x_i'u_i)$$

$$x \xrightarrow{p} 0 \rightarrow x \xrightarrow{d} 0$$

Luego, por lo que demostramos de la distribución de $(\frac{1}{\sqrt{N}}\sum^{N} x_i'u_i)$ y teorema:

$$\sqrt{N}(\hat{\beta}-\beta) \xrightarrow{d} MV(0, [E(X'X)]^{-1}\sigma^2 E(x_i'x_i)[E(X'X)]^{-1})$$

$$\sqrt{N}(\hat{\beta}-\beta) \xrightarrow{d} MV(0, [E(X'X)]^{-1}\sigma^2)$$

Bottom line: under the assumption of full rank for X, conditional independence and spherical errors, our usual statistical inference is valid.

3.2 NON-SPHERICAL ERRORS?

Lately, it has become more popular to estimate b by OLS even when heteroskedasticity is sus- pected but to adjust the standard errors and test statistics so that they are valid in the presence of arbitrary heteroskedasticity. If the errors are homoskedastic, then:

$$\hat{Avar}(\hat{\beta}) = \hat{\sigma^2}(X'X)^{-1}$$

If the errors are not homoscedastic, then the proof of asymptotic normality is still valid, just that we need to estimate the variance differently. To do so, note that by the WLLN:

$$\frac{1}{N}\sum_i u_i^2 x_i'x_i \xrightarrow{p} B = Var(x_i, u_i) = E(u_i^2 x_i'x_i)$$

And that the following is a consistent estimator of B:

$$\hat{B} = \frac{1}{N}\sum_i \hat{u_i}^2 x_i'x_i$$

Then, we can estimate asymptotically correct variance for the estimators with the following:

$$\sqrt{N}(\hat{\beta}-\beta) \xrightarrow{d} MV(0, [E(X'X)]^{-1}\hat{B}[E(X'X)]^{-1})$$

$$A\hat{var}(\hat{\beta}) = (X'X)^{-1}\hat{B}(X'X)^{-1}$$

Remember, these standard errors are asymptotically valid in the presence of any kind of heteroskedasticity, including homoskedasticity.

4

INSTRUMENTAL VARIABLE ESTIMATION

4.1 ONE ENDOGENOUS ONE INSTRUMENT

Remember: in Econometrics Endogeneity often comes from three sources: omitted variable bias, simultaneity and measurement error in the dependent variable.

Thus, let's assume a population model like the following:

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_{k-1} x_{k-1} + \beta k x_k + u$$

Where all but x_k are exogenous variable. Thus, not only estimation of the former equation will result in an inconsistent estimation of β_k, but we cannot hope to recover the other coefficients.

The method of instrumental variables (IV) provides a general solution to the problem of an endogenous explanatory variable. To use the IV approach with x_k endogenous, we need an observable variable, z_1, not in equation (5.1) that satisfies two conditions:

EXOGENEITY $Cov(z_1, u) = 0$

RELEVANCE The instrument must be partially correlated with the endogenous variable, once all the other exogenous variables have been accounted for. That is:

$$x_k = \alpha_0 + \ldots + \theta z_1 + e$$

We must have that $\theta \neq 0$.

If the variable z_1 satisfies both conditions, we have an instrumental variable for x_k. **Remember: a regression works by predicting the dependent variable for constructed groups and computing average differences between these groups; in an instrumental variable estimation, we are comparing groups with different values for the instrumental variable. If these groups differ in their predicted average value, as we have assumed that the instrumental variable has**

no direct effect, neither is correlated with a relevant variable but the endogenous variable, then we can safely adjunct the effect to the endogenous variable x_k. That is: we are saying that we have found an exogenous variation of x_k.

Then, the original vector of parameters of interest can be recovered with the Wald Estimatior:

$$\begin{aligned} Y = X\beta + e,\ Z := (x_1, \ldots, x_{k-1}, z_1) \\ Z'Y = Z'X\beta + Z'e \\ E(Z'Y) = E(Z'X)\beta,\ \text{(Using exogeneity condition)} \\ (E(Z'X))^{-1}E(Z'Y) = (E(Z'X))^{-1}E(Z'X)\beta \\ \beta = (E(Z'X))^{-1}E(Z'Y) \end{aligned}$$

It can be shown that when $Z'Z$ is of full rank, then $E(Z'X)$ is invertible if and only if the instrument is relevant. Thus, given a random sample $\{(y, x, z)\}$:

$$\hat{\beta} = (\sum^{N} z_i' x_i)^{-1} (\sum^{N} z_i' y_i)$$

$$\hat{\beta} = (\frac{1}{N}\sum^{N} z_i' x_i)^{-1} (\frac{1}{N}\sum^{N} z_i' y_i) = (Z'X)^{-1}Z'Y$$

4.1.1 *Untestability of the Exogeneity condition*

Thus, we know our instrumental variable estimation will only work if the conditions of exogenity and relevance are both satisfied. Should we test the hell out of them? Of the relevance condition, yes! But the exogenity condition is untestable: to do so we would need to have a consistent estimator of the error with the correlation with the endogenous variable, but to have such an estimator we should have first the endogenous variable's coefficient. Thus, to directly test it with the residuals of our estimation would be to beg the question: a circular reasoning. Assuming that it holds, let's test it? No logic.

4.1.2 *Testing relevance*

As one is not willing to assume much in this model, then robust standard errors may be a good idea. In the current context, a simple t-test on θ should be enough. A rule of thumb is an F-test of at least 10. For many instruments, we should constraint their coefficients to zero and use F-test to compare the constrained and the unconstrained

model. The null hypothesis and the corresponding test statistic will be:

$$H_0 : \theta_1 = 0, \ldots . \theta_k = 0$$

$$H_a : \exists\ \theta_i \neq 0,\ i = 1, \ldots, k$$

$$\frac{\frac{R^2_{NC} - R^2_C}{M}}{\frac{R^2_C}{N-K-1}}$$

4.1.3 *Good contestable instrument*

Given that one condition is contestable and one is not, a good instrument should be contestable in the relevant condition. But the exogenity argument should be so good that no doubt remains about it.

4.1.4 *Local Average Treatment Effect*

Usually, we are interested in an average treatment effect. However, with instrumental variables we are only able to estimate the Local Average Treatment Effect (LATE): the average treatment for those individuals for whom the instrument was the determinant of their condition in the endogenous variable.

4.1.5 *Sources of Instruments*

Program Evaluation through randomization. Examples include job training programs and school voucher programs. Actual participation is almost always voluntary, and it may be endogenous because it can depend on unobserved factors that affect the response. However, it is often reasonable to assume that eligibility is exogenous. Because participation and eligibility are correlated, the latter can be used as an IV for the former.

A valid instrumental variable can also come from what is called a natural experiment. A natural experiment occurs when some (often unintended) feature of the setup we are studying produces exogenous variation in an otherwise endogenous explanatory variable.

4.2 MULTIPLE INSTRUMENTS: TWO STAGE LEAST SQUARES

Remember: the IV estimation is just using a single variable to induce an exogenous variation of the endogenous variable.

Let's assume we have $z_1, \ldots, z_M$ instruments such that we are confident the exogeneity condition holds for all of them.

If each of these has some partial correlation with x_k , we could have M different IV estimators. Actually, there are many more than this — more than we can count — since any linear combination of $x_1, \ldots, x_{k-1}, z_1, \ldots, z_M$ is uncorrelated with u. So which IV estimator should we use? There's a theorem that says that the most efficient estimator is the 2SLS estimator. Here's the intuition:

Out of all the possible linear combinations, we should use as instruments those that make the linear combination that is both exogenous and most highly correlated with the endogenous variable. Let $z := (1, x_1, \ldots, x_{k-1}, z_1, \ldots, z_M)$ with $L = K + M$. Then, the linear combination of z most highly correlated with x_k is given by the linear projection of x_k on z.

$$x_k = \delta_0 + \delta_1 x_1 + \ldots + \delta_{k-1} x_{k-1} + \theta_1 z_1 + \ldots + \theta_M z_M + e$$

$$x_k^* = \delta_0 + \delta_1 x_1 + \ldots + \delta_{k-1} x_{k-1} + \theta_1 z_1 + \ldots + \theta_M z_M$$

x_k^* is often interpreted as the part of x_k that is uncorrelated with u. If x_k is endogenous, it is because e is correlated with u.

If we could observe x_k^*, we would use it as an instrument for x_k and use the traditional IV estimator. However, we can estimate it using a random sample of z by OLS as long as we make the assumption that the exogenous variables are not perfectly correlated:

$$\hat{x_k} = \hat{\delta_0} + \hat{\delta_1} x_1 + \ldots + \hat{\delta_{k-1}} x_{k-1} + \hat{\theta_1} z_1 + \ldots + \hat{\theta_M} z_M$$

Then, if we use $z = \hat{x} = (x_1, \ldots, x_{k-1}, \hat{x_k})$ the IV estimator will be:

$$\hat{\beta} = (\sum^N \hat{x_i}' x_i)^{-1} (\sum^N \hat{x_i}' y_i)$$

The IV estimator itself can be written as an OLS estimator. To do so, note that:

$$\hat{X} = Z(Z'Z)^{-1} Z'X = PzX$$

Where Pz is both symmetric and idempotent. Thus:

$$\hat{X}'X = X'PzX = X'PzPzX = (PzX)'(PzX) = \hat{X}'\hat{X}$$

Thus, the IV estimator can be written as follows:

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

Thus, to find $\hat{\beta}$:

FIRST STAGE REGRESSION Regress x_k on $x_1, \ldots, x_{k-1}, z_1, \ldots, z_M$. Find $\hat{x_k}$.

SECOND STAGE REGRESSION Regress y on $x_1, \ldots, x_{k-1}, \hat{x_k}$ and thus you have your $\hat{\beta}$

In practice, we should avoid doing the procedure explicitly, as the computed standard errors in the second stage will be incorrect. Why? We need to estimate σ^2 from the sum of the residuals squared, but the relevant residuals are: $\hat{u_i} = y_i - x_i\beta$, and the residuals from the second stage regression are: $y_i - \hat{x_i}\beta$ The relevant condition here is that $E(Z'X)$ is of full rank, that is equivalent to saying that at least one of the instruments is partially correlated with the exogenous variable. Which is pretty intuitive, as we need at least one variable to induce an exogenous variation to adjunct to x_k. This condition has the same null hypothesis as a F-test in the first stage regression.

4.3 CONSISTENCY

Remember: 2SLS is one of the many consistent estimators, but it's the most efficient one.

Assume we have a random sample $\{(x, z, y)\}$. And z being all the exogenous variable in the population model of interest plus all the instruments we are going to use. Then, 2SLS is consistent under the following assumptions:

EXOGENEITY $E(Z'U) = 0$

RELVANCE $E(Z'X)$ must be of full rank. That is, z must be as related to z such that the matrix is full rank.

Also, our 2SLS estimator is the following:

$$\begin{aligned}
\hat{\beta} &= (\hat{X}'\hat{X})^{-1}(\hat{X}'Y) = (X'PzPzX)^{-1}(X'PzY) \\
&= (X'PzX)^{-1}(X'PzY) \\
&= (X'Z(Z'Z)^{-1}Z'X)^{-1}(X'Z(Z'Z)^{-1}Z'Y) \\
&= (X'Z(Z'Z)^{-1}Z'X)^{-1}(X'Z(Z'Z)^{-1}Z'(X\beta + u)) \\
&= \beta + (X'Z(Z'Z)^{-1}Z'X)^{-1}(X'Z(Z'Z)^{-1}Z'u)
\end{aligned}$$

Thus:

$$\hat{\beta} - \beta = (X'Z(Z'Z)^{-1}Z'X)^{-1}(X'Z(Z'Z)^{-1}Z'u)$$

The first term will be invertible as long as $E(Z'X)$ is invertible.

$$\hat{\beta} - \beta = ((\frac{1}{N}\sum^{N} x_i'z_i)(\frac{1}{N}\sum^{N} z_i'z_i)^{-1}(\frac{1}{N}\sum^{N} z_i'x_i))^{-1} \\ ((\frac{1}{N}\sum^{N} x_i'z_i)(\frac{1}{N}\sum^{N} z_i'z_i)^{-1}(\frac{1}{N}\sum^{N} z_i'u_i))$$

Let's study the first term:

$$(\frac{1}{N}\sum^{N} x_i'z_i) = (\frac{1}{N}\sum^{N} z_i'x_i) \xrightarrow{p} E(Z'X)$$

$$(\frac{1}{N}\sum^{N} z_i'z_i)^{-1} \xrightarrow{p} (E(Z'Z))^{-1}$$

Aplicando WLLN a cada término cada uno de ellos converge en probabilidad. Luego, por Slutsky:

$$((\frac{1}{N}\sum^{N} x_i'z_i)(\frac{1}{N}\sum^{N} z_i'z_i)^{-1}(\frac{1}{N}\sum^{N} z_i'x_i))^{-1} \\ \xrightarrow{p} (E(Z'X)(E(Z'Z))^{-1}E(Z'X))^{-1} = \\ (E(Z'X))^{-1}E(Z'Z)(E(Z'X))^{-1}$$

Luego, cada uno de ellos está acotado en probabilidad. Por Slutsky y lo anterior:

$$((\frac{1}{N}\sum^{N} x_i'z_i)(\frac{1}{N}\sum^{N} z_i'z_i)^{-1}(\frac{1}{N}\sum^{N} z_i'x_i))^{-1} = \\ O_p(1)O_p(1)O_p(1) = O_p(1)$$

Let's study the second term:

$$(\frac{1}{N}\sum^{N} x_i'z_i) \xrightarrow{p} E(Z'X)$$

$$(\frac{1}{N}\sum^{N} z_i'z_i)^{-1} \xrightarrow{p} (E(Z'Z))^{-1}$$

$$(\frac{1}{N}\sum^{N} z_i'u_i) \xrightarrow{p} E(Z'U) = 0$$

Los primeros términos convergen en probabilidad así que también están acotados en probabilidad. Luego, por Slutsky y lo anterior:

$$((\frac{1}{N}\sum^{N} x_i' z_i)(\frac{1}{N}\sum^{N} z_i' z_i)^{-1}(\frac{1}{N}\sum^{N} z_i' u_i)) = O_p(1)O_p(1)o_p(1) = o_p(1)$$

Por lo tanto:

$$\hat{\beta} - \beta = O_p(1)o_p(1) = 0$$

Luego, $\hat{\beta}$ es un estimador consistente.

4.4 ASYMPTOTIC NORMALITY

As usual, to derive asymptotic normality we need the continuous map theorem plus some assumptions about the second moment of $Z'U$. If we assume $E(u|z) = 0$ and $E(u^2|z) = \sigma^2 = Var(u^2|z)$. Then:

$$Var(Z'u) = E(u^2 Z'Z) = E(E(u^2 Z'Z|Z)) = \sigma^2 E(Z'Z)$$

$$\sqrt{N}(\hat{\beta} - \beta) = ((\frac{1}{N}\sum^{N} x_i' z_i)(\frac{1}{N}\sum^{N} z_i' z_i)^{-1}(\frac{1}{N}\sum^{N} z_i' x_i))^{-1} ((\frac{1}{N}\sum^{N} x_i' z_i)(\frac{1}{N}\sum^{N} z_i' z_i)^{-1}(\frac{1}{\sqrt{N}}\sum^{N} z_i' u_i))$$

El primer término se convierte tomando equivalencias asintóticas: (Nota: los $o_p(1)$ se operan y como estamos multiplicando por números que están acotados en probabilidad, todas las multiplicaciones también son $o_p(1)$. Luego, el resultado final es suma de $o_p(1)$ que es $o_p(1)$.)

$$((E(Z'X) + o_p(1))(E(Z'Z) + o_p(1))^{-1}(E(Z'X) + o_p(1)))^{-1} = (E(Z'X)(E(Z'Z))^{-1}E(Z'X))^{-1} + o_p(1)$$

El segundo término se convierte en lo siguiente debido a equivalencia asintótica:(Nota: los $o_p(1)$ se operan y como estamos multiplicando por números que están acotados en probabilidad, todas las multiplicaciones también son $o_p(1)$. Luego, el resultado final es suma de $o_p(1)$ que es $o_p(1)$.)

$$(E(Z'X) + o_p(1))(E(Z'Z) + o_p(1))^{-1}(\frac{1}{\sqrt{N}}\sum^{N} z_i' u_i)) = E(Z'X)(E(Z'Z))^{-1}(\frac{1}{\sqrt{N}}\sum^{N} z_i' u_i)) + o_p(1)$$

Por teorema mapa continuo y Linderberg-Levy para el término que todavía está en sumatoria, luego:

$$\begin{aligned}\sqrt{N}(\hat{\beta}-\beta) \xrightarrow{d} N(0, ((E(Z'X)(E(Z'Z))^{-1}E(Z'X))^{-1} \\ E(Z'X)(E(Z'Z))^{-1})\sigma^2 E(Z'Z) \\ (E(Z'X)(E(Z'Z))^{-1}E(Z'X))^{-1} \\ E(Z'X)(E(Z'Z))^{-1})'\end{aligned}$$

Operando:

$$\begin{aligned}\sqrt{N}(\hat{\beta}-\beta) \xrightarrow{d} \\ N(0, \sigma^2(E(Z'X)(E(Z'Z))^{-1}E(Z'X))^{-1})^{-1}\end{aligned}$$

4.4.1 *Asymptotic Efficiency*

Under the assumptions of exogeneity of instruments, their relevance, and homoskedastic errors, 2SLS estimator is efficient in the class of all instrumental variables estimators using instruments linear in z.

4.4.2 *Robust Standard Errors*

As with OLS, if we are not sure about homoskedastic errors, we can compute robust standard errors without assuming anything special about the convergence of $\sum^N u_i^2 z_i' z_i$.

Parte II

DISCRETE VARIABLES

5

MAXIMUM LIKELIHOOD

Maximum likelihood is a systematic approach to estimation. The approach is simple: given a family of distributions for our variable of interest, which of the distributions should we use? The one that maximizes the likelihood (the density) of observing our data. That is, our estimators will be those that maximize the likelihood of observing the data.

Of course, the main problem with maximum likelihood is that all its nice properties are conditional on having pinned down the correct joint distribution of the sample as a function of the parameters. Thus, MLE is not very robust to misspecification.

In a nutshell: In order to perform maximum likelihood analysis we need to specify, or derive from an underlying (structural) model, the density of y_i given x_i. We assume this density is known up to a finite number of unknown parameters, with the result that we have a parametric model of a conditional density. Then, we maximize the likelihood to find the parameters and we are done. Strictly speaking, we should specify the distribution of x_i too, but as long as this distribution does not depend on the parameters, our choice won't matter. Conditional on the specification, MLE estimation is consistent, asymptotically normal and asymptotically efficient.

6

BINARY RESPONSE MODELS

In binary response models, the variable to be explained, y, is a random variable taking on the values zero and one, which indicate whether or not a certain event has occurred. Our interest lies primarily in:

$$P(y = 1|x) = p(x)$$

That is, how the probability of some event depends on some other factors. Then, we are interested in the marginal effect of a variable, if the variable is continuous, or in the discrete change if the variable is discrete. That is:

$$\frac{\partial P(y = 1|x)}{\partial x}$$

$$P(y = 1|x = 1) - P(y = 1|x = 0)$$

6.1 LINEAR PROBABILITY MODELS

The linear probability model (LPM) for binary response y is specified as:

$$P(y = 1|x) = \beta_0 + \ldots + \beta_k x_k$$

Thus, the only change would be in the interpretation of the coefficients, as the change in the independent variable would cause a β increase in the probability of success. If x is a binary, the β simply measures the increase in probability of success, ceteris paribus.

Problems:

HETEROSKEDASTICITY Given the Bernoulli distribution of the response variable:

$$Var(y|x) = x\beta(1 - x\beta)$$

Thus, the variance depends of the x and there are heteroskedastic errors. Linear probability models should always be estimated with robust errors.

PREDICTION AND CONSTANT EFFECTS Given that we are assuming constant marginal effects, regardless of the starting point, it's possible to have probability predictions outside the interval for observations with extreme values. Thus, awkward and unreliable for prediction.

If the main purpose of estimating a binary response model is to approximate the partial effects of the explanatory variables, averaged across the distribution of x, then the LPM often does a very good job. But there is no guarantee that the LPM provides good estimates of the partial effects for a wide range of covariate values, especially for extreme values of x.

6.2 INDEX MODELS FOR BINARY RESPONSE: PROBIT AND LOGIT

We now study binary response models of the form:

$$P(y = 1|x) = G(x\beta) := p(x)$$

Where $0 < G(z) < 1,\ \forall z \in R$. Index models where G is a cdf can be derived more generally from an underlying latent variable model:

$$y^* = x\beta + \epsilon,\ y = I[y^* > 0]$$

If the distribution of ϵ is symmetric: $1 - G(-z) = G(z)$

Although the latent variable formulation highlights the direct, marginal effect of the variable on the latent variable, this effect is often of no interest. Our true interest is in the marginal effect of the variable on the probability of success:

$$\frac{\partial G(x\beta)}{\partial x_j} = g(x\beta)\beta_j$$

Therefore, the effect of x_j on the probability of success is not constant but depends on all the other independent variables. To interpret the effect of a discrete variable, the change in probability resulting from the variable being on. That is, a subtraction of probabilities.

If $G(x\beta)$ is a the CDF of a standard normal, then $g(x\beta)$ is:

$$g(x\beta) = \frac{1}{\sqrt{2\pi}} exp(\frac{-(x\beta)^2}{2})$$

If $G(x\beta)$ is a the CDF of a standard logistic function, then $g(x\beta)$ is:

$$G(x\beta) = \frac{exp(x\beta)}{1 + expx\beta)}$$

$$g(x\beta) = \frac{exp(x\beta)}{(1 + exp(x\beta))^2}$$

6.3 ERRORES TIPO I

Suponga el siguiente modelo de utilidad aleatoria con una regla de decisión de maximización de utilidad, donde los errores se distribuyen como error tipo I.

$$u_{1i} = X_1\beta + \epsilon_1$$
$$u_{2i} = X_2\beta + \epsilon_2$$

Luego:

$$P(D_{1i}|x_i\beta) = P(u_{1i} > u_{2i}) = P(x_{i1}\beta + \epsilon_1 > x_{i2}\beta + \epsilon_2) =$$
$$P(\epsilon_2 < x_{i1}\beta + \epsilon_1 - x_{2i}\beta)$$

Tenemos una distribución conjunta, luego por definición:

$$P(\epsilon_2 < x_{i1}\beta + \epsilon_1 - x_{2i}\beta) = \int_{\infty}^{\infty} P(\epsilon_2 < x_{i1}\beta + e_1 - x_{2i}\beta) f_{\epsilon_1}(e) de_1$$

Por definición de error tipo I:

$$P(\epsilon_2 < x_{i1}\beta + e_1 - x_{2i}\beta) = exp(-exp(x_{i1}\beta + e_1 - x_{2i}\beta))$$
$$f_{\epsilon_1}(e_1) = exp(-e_1 - exp(-e_1))$$

Luego:

$$P(\epsilon_2 < x_{i1}\beta + \epsilon_1 - x_{2i}\beta) = \int_{\infty}^{\infty} exp(-exp(x_{i1}\beta + e_1 - x_{2i}\beta))exp(-e - exp(-e_1))$$

$$\int_{\infty}^{\infty} exp(-exp(x_{i1}\beta + e_1 - x_{2i}\beta) - e - exp(-e_1))de_1$$

$$\int_{\infty}^{\infty} exp(-exp(e_1)exp(x_{i1}\beta - x_{2i}\beta) - e_1 - exp(-e_1))de_1$$

$$\int_{\infty}^{\infty} exp(-exp(e_1)(exp(x_{i1}\beta - x_{2i}\beta) + 1) - e_1)de_1$$

Si hacemos el cambio de variable:

$$e^{\lambda} = \frac{e^{x_{i1}\beta}}{e^{x_{2i}\beta}} + 1$$

$$\int_{\infty}^{\infty} exp(-e_1 - exp(e_1)exp(\lambda))de_1$$

Sumo y resto $exp(\lambda)$:

$$\int_{\infty}^{\infty} exp(-(e-\lambda) - \lambda - exp(-(e-\lambda)))de_1$$

$$exp(-\lambda)\int_{\infty}^{\infty} exp(-(e-\lambda) - exp(-(e-\lambda))de_1$$

Que por definición de error tipo I y su función de densidad:

$$exp(-\lambda)1 = (-exp(x_{i1}\beta - x_{2i}\beta) - 1)$$

7

MULTINOMIAL

The first model we cover applies when a unit?s response or choice depends on individual characteristics of the unit?but not on attributes of the choices.

Let y be a random variable taking J possible values: $y = \{0, 1, \ldots, J\}$, and a x a set of conditioning variables. Thus, we are interested in the marginal effect on the different response probabilities. As there are J+1 probabilities, we only have to estimate J probabilities for each individual.

7.1 MULTINOMIAL LOGIT

The multinomial logit thus have the following distribution:

$$P(y = j|x\beta) = \frac{e^{x\beta}}{1 + \sum_j^J e^{x\beta}}$$

$$P(y = 0|x\beta) = \frac{1}{1 + \sum_j^J e^{x\beta}}$$

Thus, the conditional distribution of y is given by:

$$f(y|x\beta) = \Pi_{j=0}^J P(y = j|x\beta)^{I(y=j)}$$

$$log(f(y|x\beta)) = \sum_{j=0}^{J} I(y = j) log(P(y = j|x\beta)$$

Then, the log likelihood will be given by:

$$L(y|x\beta) = \sum_{i=1}^{N} \sum_{j=0}^{J} I(y = j) log(P(y = j|x\beta)$$

The log likelihood function is globally concave. However, the parameters can only be identified as the deviations from the effect for a base category.

The fitted probabilities can be used for prediction purposes: for each observation i, the outcome with the highest estimated probability is the predicted outcome. This can be used to obtain a percent correctly predicted, by category if desired.

7.2 PROBABILISTIC CHOICE MODEL

$$y_{ij}^* = x_{ij}\beta + a_{ij}$$

Then, the rule of maximization will yield $y_i = argmax(y_{11}^*, \ldots, y_{1J}^*)$

If the a_{ij} are independent and identically distributed as extreme value type I.

$$P(y_i = j|x\beta) = \frac{e^{x_{ij}}}{\sum_{j=1}^{J} e^{x_{ij}\beta}}$$

The CL and MNL models have similar response probabilities, but they di?er in some important respects. In the MNL model, the conditioning variables do not change across alternative: for each i, xi contains variables specific to the individual but not to the alternatives. This model is appropriate for problems where characteristics of the alternatives are unimportant or are not of interest, or where the data are simply not available. For example, in a model of occupational choice, we do not usually know how much someone could make in every occupation. What we can usually collect data on are things that affect individual productivity and tastes, such as education and past experience. The MNL model allows these characteristics to have different effects on the relative probabilities between any two choices.

7.3 INDEPENDENCE FROM IRRELEVANT ALTERNATIVES

That relative probabilities for any two alternatives depend only on the attributes of those two alternatives. This is called the independence from irrelevant alternatives (IIA) assumption because it implies that adding another alternative or changing the characteristics of a third alternative does not affect the relative odds between alternatives j and h.

Parte III

PANEL DATA

In Panel Data, we observe each unit at least twice. That is, we can observe how the variables evolve for one individual. Then, we can evaluate the effect of increasing some variable, controlling for individual constant heterogeneity.

8

RANDOM AND FIXED EFFECTS

8.1 OMITTED VARIABLES PROBLEM

Let $(x_1, \ldots, x_k)$ be observable random variables and c an unobservable random variable. Assuming a linear model for the conditional expectation:

$$E(y|x, c) = x\beta + c$$

Due to the nature of c, we cannot estimate this structural conditional expectation unless we use a proxy for c, or use an instrumental variable for the variable that we are interested. However, if we can observe the same cross section units at different points in time — that is, if we can collect a panel data set — then other possibilities arise. How? With Panel Data and let's assume the unobservable effect is time constant; that is:

$$E(y_t|x_t, c) = x_t\beta + c, \quad t = 1, \ldots, T$$

The assumption that c is constant over time, and has a constant partial effect over time, is crucial to the following analysis. An unobserved, time-constant variable is called an unobserved effect in panel data analysis. When t represents different time periods for the same individual, the unobserved effect is often interpreted as capturing features of an individual, such as cognitive ability, motivation, or early family upbringing, that are given and do not change over time. Similarly, if the unit of observation is the firm, c contains unobserved firm characteristics—such as managerial quality or structure—that can be viewed as being (roughly) constant over the period in question.

That is, given how we represent the uncertainty around our prediction, we assume that once we know the independent variables and the corresponding unobservable, we could reliably estimate the conditional effect of all the variables. In error form:

$$y_t = x_t\beta + c + u_t$$

Then, by definition:

$$E(x_t' u_t) = 0$$

If, additionally, we say $E(x_t' c) = 0$, then Pooled OLS is unbiased and consistent. Otherwise, we need better. These methods will try to eliminate the unobservable effect from the conditional expectation, thus allowing us to consistently estimate the parameters of interest in the conditional expectation. However, we need to be careful with two additional things:

- Due to the transformations, we'll need different orthogonality assumptions that do not follow from the standard orthogonality condition.
- The transformations will erase any variable that is constant through time; thus, we won't be able to estimate the effect for any of these. This outcome is not surprising: if c is allowed to be arbitrarily correlated with the elements of x_t, the effect of any variable that is constant across time cannot be distinguished from the effect of c. Therefore, we can consistently estimate β_j only if there is some variation in x_{tj} over time.

8.2 ASSUMPTIONS

The basic unobserved effects model (UEM) can be written, for a randomly drawn cross section observation i, as:

$$y_{it} = x_{it}\beta + c_i + u_{it}$$

In addition to unobserved effect, there are many other names given to ci in applications: unobserved component, latent variable, and unobserved heterogeneity are common. The u_{it} are called the idiosyncratic errors or idiosyncratic disturbances because these change across t as well as across i.

Two frameworks:

RANDOM EFFECTS: a random effects framework is synonymous with zero correlation between the observed explanatory variables and the unobserved effect: $cov(c_i, x_{ijt}) = 0 \; \forall j = 1, \ldots, k$

FIXED EFFECTS: it means that one is allowing for arbitrary dependence between the unobserved effect c_i and the observed explanatory variables x_{it}.

8.2.1 *Orthogonality*

We need a new condition for orthogonality. We can say that:

$$E(y_{it}|x_{i1},\dots,x_{iT},c_i) = E(y_{it}|x_{it},c_i) = x_{it}\beta$$

That is: means that, once x_{it} and c_i are controlled for, x_{is} has no partial effect on y_{it} for $s \neq t$. When this assumption holds, we say that the x are strictly exogenous conditional on the unobserved effect c_i. If we allow the following error-form formulation:

$$y_{it} = x_{it}\beta + c_i + u_{it}$$

The strictly exogenous x become:

$$E(u_{it}|x_{i1},\dots,x_{iT},c_i) = 0,\ t = 1,\dots,T$$

Which in turn implies that explanatory variables in each time period are uncorrelated with the idiosyncratic error in each time period:

$$E(x'_{it}u_{is}) = 0, \forall s,t$$

To conclude: in Panel Data, we must ask ourselves two sets of questions:

(1) Is the unobserved effect, c_i, uncorrelated with x_{it} for all t? (2) Is the strict exogeneity assumption (conditional on c_i) reasonable?.

Therefore, the strict exogeneity assumption never holds in unobserved effects models with lagged dependent variables.

8.3 ESTIMATION METHODS

8.3.1 *Pooled*

Suppose the following linear model:

$$y_{it} = x_{it}\beta + c_i + u_{it} = x_{it}\beta + v_{it}$$

Pooled estimation will be consistent if we are willing to argue the following:

- $E(x'_{it}u_{it}) = 0\ \forall t$
- $E(x'_{it}c_i) = 0\ \forall t$

Even if the former assumption holds, the composite errors will be serially correlated due to the presence of c_i in each time period. Therefore, inference using pooled OLS requires robust variance matrix estimator and robust test statistics.

8.4 RANDOM EFFECTS

As with pooled OLS, a random e?ects analysis puts ci into the error term. In fact, random e?ects analysis imposes more assumptions than those needed for pooled OLS: strict exogeneity in addition to orthogonality between c_i and x_{it}. Stating the assumption in terms of conditional means, we have:

$$X_i := (x_{i1}, x_{2i}, \ldots, x_{Ti})$$

$$Y_i = X_i\beta + v_i$$

$$\Omega := E(v_i v_i')$$

1. *a)* $E(u_{it}|x_i, c_i) = 0 \; \forall t$
   b) $E(c_i|x_i) = E(c_i) = 0$
2. GLS assumption: full rank $(X_i'\Omega X_i)$
3. *a)* $E(u_i u_i') = \sigma_u^2 I$
   b) $E(c_i^2) = \sigma_c^2$

Under these 3 assumptions, RE is efficient in the class of estimators consistent under $E(v_i|x_i) = 0$.

Why do we maintain Assumption 1.a when it is more restrictive than needed for a pooled OLS analysis? The random effects approach exploits the serial correlation in the errors due to c_i, but we must transform the model to do so; transformation that changes our assumptions. Additionally:

- $E(u_{it}^2) = \sigma_u^2 \; \forall t$
- $E(u_{it}u_{is}) = 0 \; \forall t \neq s$

That is, assuming that the idiosyncratic errors are homoskedastic. Note that both assumptions are the same as: $E(u_i u_i') = \sigma_u^2 I$ Then, we can derive the var-cov of the composite errors:

The variance, then, is the same regardless of the t:

$$E(v_{it}^2) = E(c_i^2) + 2E(c_i u_{it}) + E(u_{it}^2) = \sigma_c^2 + \sigma_u^2$$

The covariance, then, is the same regardless of the t:

$$E(v_{it}v_{is}) = E[(c_i + u_{it})(c_i + u_{is})] = E(c_i^2) = \sigma_c^2$$

Unlike standard models for serial correlation in time series settings, the random effects assumption implies strong persistence in the unobservables over time, due, of course, to the presence of c_i.

Then:

$$\beta_{RE} = (\sum_{i}^{N} X_i' \hat{\Omega^{-1}} X_i)^{-1} (\sum_{i}^{N} X_i' \hat{\Omega^{-1}} Y_i)$$

Thus, the problem of efficient estimation reduces to finding $\hat{\Omega^{-1}}$. To do so, we need both consistent estimators of $E(v_{it}^2)$,$E(v_{it}v_{is}) = E(c_i^2)$. Note first that to find $E(v_{it}^2)$ is the same as having a consistent estimation of the model and finding the Mean Squared error of the model. Thus, we can perform Pooled, which is consistent, and find it. For the covariances, we do the same.

In Pooled, write the model with the composite errors. In RE, assumptions from RE world, write the model in stacked form and make assumptions about the idiosyncratic errors.

8.5 FIXED EFFECTS FRAMEWORK

$$y_{it} = x_{it}\beta + c_i + u_{it}$$

The RE approach to estimating b effectively puts ci into the error term, under the assumption that ci is orthogonal to x_{it}, and then accounts for the implied serial correlation in the composite error $v_{it} = c_i + u_{it}$ using a GLS analysis. In many applications the whole point of using panel data is to allow for c_i to be arbitrarily correlated with the x_{it}. A fixed effects analysis achieves this purpose explicitly.

Assumptions:

1. $E(u_{it}|x_i, c) = 0 \ \forall t$. Strict exogeneity of $\{x_{it}\}$conditional on the unobserved effect. The difference with RE is that we allow any arbitrary expression for $E(c_i|x_i)$.As we suggested in Section 10.1, this robustness comes at a price: without further assumptions, we cannot include time-constant factors in x_{it}. The reason is simple: if ci can be arbitrarily correlated with each element of x_{it}, there is no way to distinguish the effects of time-constant observables from the time-constant unobservable c_i. When analyzing individuals, factors such as gender or race cannot be included in $x_i t$. In panel data analysis the term "time-varying explanatory variables"means that each element of $x_i t$ varies over time for some cross section units.

8.5.1 *Within Transformation*

First, average the equation for each cross-section unit across time. Such that:

$$\bar{y_{it}} = \bar{x_{it}}\beta + \bar{c_i} + \bar{u_{it}}$$

Subtracting this equation from the equation:

$$y_{it} - \bar{y_{it}} = (x_{it} - \bar{x_{it}})\beta + u_{it} - \bar{u_{it}}$$

The time demeaning of the original equation has removed the individual specific effect c_i. With c_i out of the picture, it is natural to think of estimating equation by pooled OLS. Before investigating this possibility, we must remember that equation is an estimating equation: the interpretation of β comes from the (structural) conditional expectation $E(y_{it}|x_i, c_i) = E(y_{it}|x_{it}, c_i) = x_{it}\beta + c_i$

To check if we can consistently estimate the equation, we need to check the key assumption in Pooled OLS:

$$E((u_{it} - \bar{u_i})(x_{it} - \bar{x_i})) = 0,\ t = 1, \ldots, T$$

But the fixed effects assumption entails even more: it entails that the strict exogeneity holds in the within model:

$$E((u_{it} - \bar{u_i})|(x_{i1} - \bar{x_i}), \ldots, (x_{iT} - \bar{x_i})) = 0$$