Microeconometrics

Lecture Notes

Michele Pellizzari [1]

PRELIMINARY DRAFT. DO NOT QUOTE. COMMENTS WELCOME

May, 2008

[1]These notes are largely based on (my own interpretation of) the textbook by Jeffrey M. Wooldridge. 2002. *Econometric Analysis of Cross Section and Panel Data*. MIT Press. For some parts, I also borrowed heavily from (i) Arellano M. 2003. *Panel Data Econometrics*. Oxford University Press. (especially for part 3 on panel data) and (ii) Cameron A.C. and P.K. Trivedi. 2005. *Microeconometrics. Methods and Applications*. Cambridge University Press. Obviously, I am solely responsible for any error, omission or imprecision. Contact details: Michele Pellizzari, IGIER-Bocconi, via Roentgen 1, 20136-Milan (Italy). michele.pellizzari@unibocconi.it

Preface

These lecture notes have been prepared for the course of Microeconometrics that I have been teaching at Bocconi University since the academic year 2004-2005. The course is taught in the second semester of the first year of a two-years master program in Economics and Social Sciences. The level of complexity of these lecture notes is supposed to be adequate for master students who have already taken had at least one basic econometrics course. The students are also supposed to know all the necessary material from statistics and mathematics. Some knowledge of the most applied fields in economics, like labour and development, is helpful to grab the relevance of the topics covered in these notes.

The course, and consequently also the lecture notes, is divided into three main parts. In the first part, I review some of the basic econometric concepts in *linear cross-sectional models*, covering topics like OLS, IV, GLS and system estimation. Within this first part, I also spend some time revising the basic elements of asymptotic theory, that are used throughout the entire course to assess the properties of several estimation methods.

The second part is devoted entirely to *limited dependent variable models*, including dichotomous and multinomial choice models, censoring, truncation and selection. These models are discussed mostly within a maximum likelihood framework, whose basic features are discussed at the beginning of this second part.

Finally, the third part of the course contains an introduction to the *econometrics of panel data*. For the sake of time, I focus almost exclusively on static linear models with unobserved heterogeneity and the presentation is centered around fixed-effects models.

There is nothing new or original in these notes. They are essentially my own interpretation of the material in two well-known textbooks:

- Cameron A.C. and P.K. Trivedi. 2005. *Microeconometrics. Methods and Applications*. Cambridge University Press.
- Wooldridge J.M. 2002. *Econometric Analysis of Cross Section and Panel Data*. MIT Press.

For the third part on panel data, I follow somewhat closely the line of presentation in another textbook:

- Arellano M. 2003. *Panel Data Econometrics*. Oxford University Press.

I find the sequence and the line of presentation of the material in this book extremely clear and coherent. However, Arellano (2003) is written at a

technical level that is slightly higher than the one I expect from my students. So, in these lecture notes I maintain his sequence of presentation of the material but the level of complexity is more similar to Cameron and Trivedi (2005) or Wooldridge (2002).

Obviously, I am solely responsible for any error, omission or imprecision in these notes. The authors of the books mentioned above bear absolutely no responsibility for what is written in this document.

The choice of which topics to cover and the presentation of which textbook to follow for any given topic is entirely specific to the course, the institution and the students these lecture notes are targeted to.

During the course, the theoretical lectures are complemented by applied computer sessions and problem sets that allow students to put their hands on real world data using the techniques discussed in the lectures. In the applied computer sessions we use Stata as a statistical software. In these lecture notes I make occasional references to things the students should see either in the problem sets or in the computer sessions.

Let me emphasize once again that the material in these notes is not novel. In fact, I was initially against my writing of them. I noticed that students pay a lot more attention during the lectures when they know that they can only count on their own notes to prepare the exam. Moreover, for those who will eventually end up doing empirical analysis in their professions, either in the academia or outside, it is very important that they learn how to use one (or more) econometric textbook(s), as they will often have to go back to it for reference at many points in their careers. Naturally, the availability of lecture notes discourages students from using (or even buying) the textbook. For all these reasons I see very few advantages in providing these lecture notes. However, students have been requesting them so insistently in the past few years that I started wondering whether all these beliefs of mine were correct. I am still doubtful about the true effect of these lecture notes on my students' learning so I am taking this first version of them as some sort of experiment (and one that I might try to seriously evaluate at some point).

Finally, let me thank all the students who have taken the course so far. They have been the true inspiration for these notes. As a graduate student I have always felt that I was learning more from my classmates than from my professors. Now, every time the course ends I always feel that I have learnt more from the students than what they have learnt from me.

Contents

Part 1:
The Econometrics of Linear models

1 The single equation linear model

One of the most important uses of econometrics is the empirical test of the predictions of economic theory. So, let us start from the theory and suppose we have a model that predicts a certain relationship between an endogenous variable y and a set of exogenous factors X:[1]

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_K x_K \tag{1.1}$$

For example, suppose y is earnings and X includes things like age, education, gender, etc. This linear specification, although very simple, is still quite general for economic models that can often be linearized around some equilibrium point (and what we probably see in the data are equilibrium outcomes or small deviations from the equilibrium).

If we want to bring this model to the data (for example, to test it or simply to produce estimates of its parameters) we need to augment it with an error term, that we call u:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_K x_K + u \tag{1.2}$$

To continue with our example, we don't really think that wages are mechanically determined by the factors we include in X (age, education, gender, etc.). In particular, we do not expect all persons with exactly the same age, education, gender, etc. to earn exactly the same wage. In fact, unless we really think that the model perfectly represents the data, there will always be deviations between what would be predicted by the model and the real world observations. The error term is meant to represent these deviations and thus reconcile the theoretical model with the data.

We will generally call equation 1.1 the *theoretical model* and equation 1.2 the *population model*, that is the theoretical model augmented with the error term will represent the data for any individual in the population that we want to study (and this is why we call it the population model).

Obviously, it is very rare in any empirical application to be able to work with the entire population of interest. In our example, we won't typically have access to data on the universe of all wage earners in the country (although sometimes this may happen). More frequently, we apply our population model to a (hopefully representative) sample of the population of interest. We will usually indicate with N the size of the sample and index

[1] X is a row vector of explanatory variables: $X = (x_1 \ x_2 \ \ldots x_K)$.

with i the observations in this sample. We can then introduce the third definition of our model, the *empirical model*:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_K x_{Ki} + u_i \quad \forall i = 1, \ldots, N \tag{1.3}$$

The distinction between the population and the empirical model (equation 1.3) may appear trivial but it is in fact very important. The population model is a description of the process that generates the data whereas the empirical model is an operational version of it. In particular, given a sample of N observations the empirical model can be calculated, i.e. we can replace y_i and X_i with actual numbers and perform calculations.

1.1 Where does the error term come from?

Let us look a bit more in details at the nature of the error term. We describe briefly below three very common factors that give rise to the error term in empirical models.

1. **Specification error**. In equation 1.1 we assumed that the relationship between y and X was simply linear. This is obviously a very specific example. More generally, a theoretical model may describe the relationship between variables according to a generic function (often the model does not even specify the function completely but only some of its characteristics like the sign of the partial derivivatives). In other words, a more general specification of our theoretical model could be the following:

$$y = f(x_1, x_2 \ldots, x_K) \tag{1.4}$$

   In principle we could estimate empirically a model of this type once we choose a specific functional form (not necessarily linear) for $f(\cdot)$.[2] However, the linear specification (which we are going to use extensively for this first part of the course) is often taken as a simple approximation of more sophisticated or perhaps unknown functional forms. If we impose a linear specification on equation 1.4, one simple way of

[2] We would need to apply an estimation method known as Non-Linear Least Squares (NLS), which essentially extends the usual OLS methodology to models where the relationship between y and X is not linear. We are not going to cover NLS in this course.

rewriting it is:

$$\begin{aligned} y &= X\beta + \underbrace{[f(X) - X\beta]}_{u=specification\ error} \\ &= X\beta + u \end{aligned} \tag{1.5}$$

where $X = (1\ x_1\ x_2 \dots x_K)$ is the vector of all explanatory variables and $\beta = (\beta_0\ \beta_1\ \beta_2 \dots \beta_K)$ the corresponding vector of parameters. In equation 1.5 the term $u = [f(X) - X\beta]$. Thus, one important source of error in econometrics is due to the specification of the model that we choose. For operational purposes we need to specify our model with a specific functional form and unless we are sure that that specific function is the true one, our model will contain a specification error.

2. **Measurement error**. This is one of the most common sources of problems in econometrics applications.[3] Abstract from any possible specification error (although measurement and specification errors can and often do coexist) and suppose that your empirical model is perfectly specified. Unfortunately, though, you observe one or more of your variables with some error. For example, the dependent variable y may be measured with error so that what we really observe is not y but a noisy version of it, call it y^*:

$$y^* = y + \epsilon \tag{1.6}$$

where ϵ is a random error of measurement.[4] In this case, the model that we can effectively bring to the data needs to be specified in terms of y^*, which is observable, rather than y, which we cannot observe. So, let's substitute equation 1.6 into the theoretical model of equation 1.1 to obtain a model written in terms of y^*:

$$y^* = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_K x_K + \epsilon \tag{1.7}$$

As you can see, by this simple substitution we have now obtained a model that looks totally similar to the empirical model of equation 1.2 (other than for the name of the error term) where the error term now clearly has a specific nature, that is measurement. There could also

[3] We will see more on this issue later on.

[4] ϵ may be due to errors in coding the data or errors made by survey respondents in answering questions and so on.

be measurement error in one (or more) of the explanatory variables. Suppose, for instance, that x_1 is measured with error, so that what we observe is a noisy version of it, call it x_1^*:

$$x_1^* = x_1 + \eta \tag{1.8}$$

Again, the operational version of the model has to be written in terms of the observable variables, so we substitute equation 1.8 into equation 1.2 to get:

$$\begin{aligned} y &= \beta_0 + \beta_1(x_1^* - \eta) + \beta_2 x_2 + \cdots + \beta_K x_K = \\ &= \beta_0 + \beta_1 x_1^* + \beta_2 x_2 + \cdots + \beta_K x_K + \underbrace{(-\beta_1 \eta)}_{\text{error term}} \end{aligned} \tag{1.9}$$

And we obtain a model where the error term arises as a form of measurement error. Obviously, there can be measurement error in both the dependent and the independent variables as well as in more than just one explanatory variable.

3. **Omitted variables**. Another important source of error is omitting relevant explanatory variables from the model. This is also a very pervasive problem in econometrics and we will return to it several times throughout the course. It may be seen also as a problem of specification error in the sense that the specification of the model is incorrect. This time, however, the problem is not with the functional form but rather with some explanatory variables that have been omitted from the definition of X. Abstract from problems of measurement and functional form assumptions and suppose, for example, that there is a variable x_{K+1} which is indeed a determinant of y but that we have omitted from our model, either because we wanted to simplify our analysis or perhaps because we forgot about it.[5] So the true model would be:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_K x_K + \underbrace{(\beta_{K+1} x_{K+1})}_{\text{error term}} \tag{1.10}$$

and the error term now includes the omitted variable.

[5] Or, more worryingly, because it cannot be observed.

1.2 OLS Identification assumptions

For the most part of this course our primary interest will be the production of *correct* estimates of the parameters of the model that will be considered ($\beta = (\beta_0\ \beta_1 \dots \beta_K)'$ in the specific model considered in this section). We will be a lot more precise about what we mean by *correct* later on.[6]

The simplest and most popular method to produce such estimates is known as the *Ordinary Least Squares* (OLS) method. The typical econometric textbook lists a long series of assumptions required by OLS to produce estimates that satisfy certain desirable characteristics. Let us simplify things enormously and consider what I think are the most important of these assumptions [7]:

OLS Assumption 1 $E(u) = 0$

OLS Assumption 2 $Cov(x_k, u) = 0 \qquad \forall k = 1, 2, \dots, K$

Assumption 1 is sort of free if our model includes a constant. To show this simple fact, suppose that the expected value of the error term of our model is equal to some constant c different from zero, $E(u) = c \neq 0$. Then, we could rewrite the model by adding and subtracting the constant c:

$$y = (\beta_0 + c) + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_K x_K + \underbrace{(u - c)}_{\text{new error term}} \tag{1.11}$$

Now, the new error term, call it $\widetilde{u} = u - c$, has zero mean. In other words, if our model includes a constant the error term can always be considered to have zero mean because, if by any chance its expected value is different from zero, such non-zero mean will be captured by the constant term. The only drawback of this approach is that if the error term has a non-zero expected value it is not going to be possible to estimate the constant of the model (β_0) separately from the mean of the error term (c). All we can estimate is the constant term of the new model of equation 1.11, that is $\beta_0 + c$.

Assumption 2 is instead the crucial identification assumption. It is by imposing Assumption 2 that we are able to produce estimates of the parameters of the model (we will see how in a second). If assumption 2 is not satisfied for one or more regressors we say the those regressors are *endogenous*. Namely, if $Cov(x_j, u|X_{-j}) \neq 0$ then x_j is said to be endogenous and identification (usually of all the β_k's parameters) is not possible.[8]

[6] We will usually require estimators to be consistent.

[7] This simplification is in fact so gigantic that you should not circulate these notes too much for the sake of your teacher's reputation!

[8] X_{-j} is the vector of all regressors excluding j.

But how can a variable be endogenous? As assumption 2 suggests, it depends on whether it is correlated with the error term and this is why we spent some time discussing the nature of the error term in the previous section. The short paragraphs below show how the presence of omitted variables and/or measurement error may induce endogeneity.

- **Measurement error**. Measurement is a very common source of endogeneity problems. Take the simple example of equation 1.9 where the measurement error is in one of the regressors. In this case, there is a mechanic correlation between the variable measured with error - x_1^* - and the error term which includes $-\beta_1\eta$ (plus probably some other stuff which we do not consider here for simplicity). In fact, using the definiton of x_1^* from equation 1.8 it is easy to show that:

$$Cov(x_1^*, -\beta_1\eta) = Cov(x_1 + \eta, -\beta_1\eta) \neq 0 \tag{1.12}$$

  Even in the simple case in which the error term is completely non-systematic, that is assuming $E(\eta|x_1) = E(\eta) = 0$, the covariance above is different from zero (namely equal to $-\beta_1^2 Var(\eta)$ in this particular instance).[9] We will come back to the consequences of measurement error later on. What is important to notice here is that it induces endogeneity even in the simplest possible case.

- **Omitted variables**. Obviously, if the omitted variable(s) is correlated with one or more of the other regressors then the model will be affected by endogeneity. In the example of equation 1.10, if $Cov(x_j, x_{K+1}|X_{-j}) \neq 0$, then x_j will be endogenous and this fact will impede identification. As we pointed out in the previous section, omitted variables are ubiquitous in economics and the problem becomes particularly cumbersome when the omitted variable is hard or even impossible to observe. In fact, the most common reason why some relevant variables are omitted from an empirical model is that they simply cannot be observed. In other cases, we might have forgotten about them or thought they were not important. These cases are simple because we can merely add these variables to the regression and solve the problem. Obviously, if the omitted variables cannot be observed things are a lot more complicated. A leading example of omitted variable bias which we will use again later on in the course is the following. Suppose we want to run

[9] This particular case of a non-systematic measurement error is often labeled *classical measurement error*.

a wage regression with education as an explanatory variable. We are particularly interested in the estimation of the effect of education on wages (the so called *returns to education*):

$$w = \beta_0 + \beta_1 educ + (\beta_2 ability + u) \tag{1.13}$$

where w is a measure of individual wages and *educ* represents years of schooling. In this regression an obvious omitted variable is *ability*, that is some indicator of the intellectual capabilities of individuals (e.g. how smart one is). Ability is obviously a relevant determinant of wages as smarter people typically earn more, however ability is very hard to observe and is omitted.[10] Omitting ability from equation 1.13 is problematic because it is likely that education and ability are strongly correlated as smarter individuals also study longer. A huge literature - which we will analyze in some details later on - has explored this issue and suggested possible solutions.[11]

1.3 Estimation of the parameters

In this section we discuss the simplest methods to produce estimates of the parameters of a simple linear model. For simplicity, let us consider a univariate model with a constant[12]:

$$y = \alpha + \beta x + u \tag{1.14}$$

The first thing we need to do is imposing some assumptions (that's always the initial step). Here we impose the OLS assumptions that we already discussed in the previous section and that can be rewritten very simply for the specific model of equation 1.14:

1. $E(u) = 0$
2. $Cov(x, u) = 0$

[10] Sometimes we find datasets with IQ scores or other observable proxies for ability but these instances are rather rare.

[11] A couple of classical references on this topic are: Card, D. 1993. *"Using Geographic Variation in College Proximity to Estimate the Return to Schooling"*. NBER Working Paper N. 4483; Angrist, J.D. and A.B. Krueger. 1991. *"Does Compulsory School Attendance Affect Schooling and Earnings?"* The Quarterly Journal of Economics, vol. 106(4), 979-1014.

[12] You should know this but let me still add this note for clarity. A univariate model is a model with just one explanatory variable other than the constant.

Notice that using assumption 1, assumption 2 can be rewritten as:

$$Cov(x,u) = E[(x-\overline{x})(u-\overline{u})] = E[(x-\overline{x})u] = E(xu) - \overline{x}E(u) = E(xu) = 0$$

where $\overline{u} = E(u)$ and similarly for $\overline{x} = E(x)$. So, let us work with the two assumptions stated in the following format:

1. $E(u) = 0$
2. $E(xu) = 0$

Replacing equation 1.14 into these two assumptions yields the following:

$$\begin{aligned} E(u) &= E(y - \alpha - \beta x) = 0 \\ &= E(y) - \alpha - \beta E(x) = 0 \\ \alpha &= E(y) - \beta E(x) \end{aligned} \tag{1.15}$$

and

$$\begin{aligned} E(xu) &= E\left[x(y - \alpha - \beta x)\right] \\ &= E(xy) - \alpha E(x) - \beta E(x^2) = 0 \end{aligned} \tag{1.16}$$

Replacing equation 1.15 into equation 1.16 yields:

$$\begin{aligned} [E(xy) - E(y)E(x)] &- \beta\left[E(x^2) - E(x)^2\right] = 0 \\ Cov(x,y) &- \beta Var(x) = 0 \\ \beta &= \frac{Cov(x,y)}{Var(x)} \end{aligned} \tag{1.17}$$

Equation 1.17 is very important. It states that β, the coefficient we are usually interested in, can be written as a function of moments in the population. In this particular case it can be written as the ratio between the covariance of x and y and the variance of x.

Remember that our aim here is to produce an estimator of β that satisfies some desirable properties. Intuitively, the most obvious way to proceed to produce an estimator of β is to replace the population moments on the right hand side of equation 1.17 with their sample analog that can be computed from the data:

$$\widehat{\beta} = \frac{\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{N}(x_i - \overline{x})^2} \tag{1.18}$$

It turns out that this simple estimator satisfies one very important and desirable property: it is a consistent estimator.[13]

$$\widehat{\beta} \xrightarrow{p} \frac{Cov(x,y)}{Var(x)} = \beta \tag{1.19}$$

This result is obtained thanks to the *analogy principle* that essentially states that, if the sample of data is representative of the population of interest, then its sample moments converge in probability to the corresponding population moments.

So now we know that at least $\widehat{\beta}$ is a consistent estimator for β. But how efficient is it? How fast does it converge to the true β? For this we will need to know the distribution of $\widehat{\beta}$. Wait until the next section for this and let us now generalize the derivation of this simple estimator to multivariate models, i.e. models with more than just one explanatory variable.

K explanatory variables

Define

$$\underset{1\times K}{X} = (x_1 \; x_2 \ldots x_K)$$

the vector of K explanatory variables with the first variable $x_1 = 1$ being the constant. Similarly, define

$$\underset{K\times 1}{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{pmatrix}$$

the corresponding $(K \times 1)$ vector of coefficients.

The multivariate version of the population model 1.14 can be written as:

$$\underset{1\times 1}{y} = \underset{1\times K}{X}\underset{K\times 1}{\beta} + \underset{1\times 1}{u} \tag{1.20}$$

In this setting, assumptions 1 and 2 can be written jointly as:

$$\underset{K\times 1}{E(X'u)} = E\left[\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_K \end{pmatrix} u\right] = E\begin{bmatrix} u \\ x_2 u \\ \vdots \\ x_K u \end{bmatrix} = 0 \tag{1.21}$$

[13] This is also something that you should know very well but just to refresh ideas, a consistent estimator is a statistics whose probability to be different from the number (or vector) it is supposed to estimate goes to zero as the number of observations (N) goes to infinity. We will go back to these issues later in section 1.4.

(recall that $x_1 = 1$).

Equation 1.21 is in fact a list of K conditions which we can use to estimate the K coefficients in β as follows:

$$\begin{aligned} E(X'u) = E\left[X'(y - X\beta)\right] &= E(X'y) - E(X'X)\beta = 0 \\ E(X'X)\beta &= E(X'y) \\ \beta &= E(X'X)^{-1}E(X'y) \end{aligned} \tag{1.22}$$

As in equation 1.17, also equation 1.22 shows that β can be written as a function of observable population moments. Consequently, according to the *analogy principle*, a simple consistent estimator of β can be produced by replacing the population moments in equation 1.22 with their corresponding sample moments:

$$\widehat{\beta} = \left[N^{-1}\sum_{i=1}^{N} X_i'X_i\right]^{-1}\left[N^{-1}\sum_{i=1}^{N} X_i'y_i\right] \tag{1.23}$$

Another way of writing equation 1.23, which you might have encountered in previous econometrics courses, is by defining the following matrices:

$$\underset{N\times K}{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{pmatrix} = \begin{pmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,K} \\ x_{2,1} & x_{2,2} & \dots & x_{2,K} \\ \vdots & \vdots & \vdots & \vdots \\ x_{N,1} & x_{N,2} & \dots & x_{N,K} \end{pmatrix}$$

and

$$\underset{N\times 1}{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}$$

so that we can express $\widehat{\beta}$ as:

$$\widehat{\beta} = (X'X)^{-1}X'Y \tag{1.24}$$

Why Ordinary Least Squares?

Most of you should have seen the OLS estimator $\widehat{\beta}$ either in the formulation of equation 1.18 or equation 1.23 or 1.24 but by now you may be wondering why on earth it is called *Ordinary Least Squares* estimator. In the previous derivation we have never taken the *least* of any *square* whatever.

The name derives simply from the fact that the same estimator can be obtained by minimizing the sum of the squared residuals of the model, i.e. the difference between the true value of y and its prediction from the model. To see this, consider the simple univariate version of the model as in equation 1.14 and write the error term as:

$$u_i = y_i - \alpha - \beta x_i \tag{1.25}$$

The idea underlying OLS is to find the estimators of α and β that make the *@@ the sum of the squared* residual*s* as close to zero as possible. In some sense this also means finding the estimators that use the information in x and y in the best possible way. So, an alternative definition of the OLS estimators of α and β is the following:

$$(\widehat{\alpha}, \widehat{\beta}) = \underset{\alpha,\beta}{min} \sum_{i=1}^{N} [y_i - \alpha - \beta x_i]^2 \tag{1.26}$$

The first order conditions for this problem are:

$$-2 \sum_{i=1}^{N} [y_i - \alpha - \beta x_i] = 0 \tag{1.27}$$

$$-2 \sum_{i=1}^{N} x_i [y_i - \alpha - \beta x_i] = 0 \tag{1.28}$$

Notice that equations 1.27 and 1.28 are in fact the sample analogs of assumptions 1 and 2. Equation 1.27 simply states that the (sample) average of the residuals should be equal to zero, just like assumption 1 states that the (population) mean of the residual should be zero. Similarly, equation 1.28 states that the (sample) covariance of x and u should be zero, just like assumption 2. Not surprisingly, then, by solving this simple system of 2 equations in 2 unknowns one gets exactly the same expressions for $\widehat{\alpha}$ and $\widehat{\beta}$ that we obtained before.

1.4 Brief aside on asymptotic theory

Most of you should have already seen the things we discuss in this section in previous econometrics courses. It is still perhaps useful to review some of these concepts to refresh ideas and also to harmonize our languages so that we all know what we are talking about and why. In this discussion I will try to give you the intuition of the statistical concepts that we need for this

course but I will also avoid much of the technicalities and formalities that you typically see in a statistics course.[14].

Earlier on we said that we wanted to produce *correct* estimators for some parameters of interest and so far we only specified that we want such estimators to be *consistent*. Moreover, we also left the question of the distribution of the estimator unanswered. That is a very important question that we really want to address because otherwise we would not be able to run tests of our results (for example, we may want to test if our β of interest is significantly different from zero) and to do that we need to know the distribution of our estimator.

In this section we will review briefly the basic statistical tools and concepts needed to derive *asymptotic* consistency and the *asymptotic* distribution of estimators. But before we move to reviewing these concepts, it is perhaps useful to clarify why we focus on the *asymptotic* properties of estimators and what that means exactly.

Why do we care about the asymptotic properties of the estimators?

The estimators that we produce in econometrics are essentially statistics meant to contain information about some underlying true parameters that exist - or are assumed to exist - in a given population. For example, in the simple case of the linear univariate model $\widehat{\beta}$ is an estimator that tells us something about the true parameter β that we assumed to be part of the relationship between two given variables x and y in the population.

Usually, such estimators are produced with data from the population of interest. Sometimes we are lucky and the sample of data is very large, some other times the sample is instead very small. The first good reason to be interested in the asymptotic properties of our estimators is that such properties are independent of the size of the sample. A consistent estimator is consistent regardless of whether it is produced using a large or a small sample. In fact, asymptotic theory studies the characteristics of statistics produced with samples of finite size as the size of such samples converges to infinity.

The alternative to considering the asymptotic properties of an estimator is looking at its *small sample* properties, i.e. the characteristics of the estimator in the particular sample used to produce it.

[14]Such technicalities and formalities are nonetheless extremely important! So, if you feel that the material in this section is completely new or very hard, you may want to go back and revise some of your notes from statistics.

The second important reason to look at the asymptotic properties of our estimators is simply that sometimes their small sample features (like *unbiasedness*, for example) are very difficult or even impossible to analyze. This is particularly true for estimators that are non-linear functions of the random variables of the model. In this course we will focus almost exclusively on the asymptotic properties of our estimators because we will usually work with non-linear estimators and because one of the advantages of microeconometrics is that we usually have access to very large datasets where the small sample properties of any statistics can be safely approximated with their asymptotic ones.

Convergence in probability

Let us start with a simple math definition:

Definition 1 (Convergence of a sequence of *non-random* numbers) *A sequence of* ***non random*** *numbers* $\{a_N; N = 1, 2, \dots\}$ *converges to a if:*

$$\begin{array}{c} \forall \epsilon > 0 \ \exists N_\epsilon \text{ such that:} \\ N > N_\epsilon \ \Rightarrow \ |a_N - a| < \epsilon \end{array}$$

and we write

$$a_N \to a \text{ as } N \to \infty$$

This definition should be familiar to all of you. It is simply the definition of convergence of a sequence of non-random numbers. Non-random numbers are the topic of mathematics, so you should have seen definition 1 in some of your math courses.

Convergence in probability is simply the translation of definition 1 into the realm of random numbers, the main object of statistics.

Definition 2 (Convergence in probability) *A sequence of* ***random*** *numbers* $\{x_N; N = 1, 2, \dots\}$ *converges to a constant a if:*

$$\forall \epsilon > 0 \ \lim_{N \to \infty} Pr\{|x_N - a| > \epsilon\} = 0$$

and we write

$$x_N \xrightarrow{p} a \text{ or } plim(x_N) = a$$

Definition 2 is simply the equivalent of definition 1 for random numbers. In fact, for random numbers definition 1 simply cannot be applied. A random number by definition can never be equal to a constant unless its distribution is degenerated to that constant with zero variance, which makes it a non-random number. To translate the definition of convergence to random numbers we need definition 2 which simply states that a sequence of random numbers converges to a constant if in the limit the probability of the event that the random number is far from the constant goes to zero, where far is defined for any distance ϵ.

Using definition 2 we can then give a proper definition of consistent estimator:

Definition 3 (Consistent estimator) *Let $\{\widehat{\theta}_N : N = 1, 2, \dots\}$ be a sequence of estimators of the $K \times 1$ vector θ, where N indexes sample size. $\widehat{\theta}$ is a consistent estimator of θ if plim $\widehat{\theta} = \theta$.*

Essentially, definition 3 says that if an estimator is consistent we are guaranteed that if we were to compute it on larger and larger samples, the probability that it is far from its estimands would become smaller and smaller. That seems like a nice desirable property for an estimator.

But, how can we calculate the probability limit of a random variable? The following famous theorem is of great help:

Theorem 1 (The weak law of large numbers) *Let $\{w_i : i = 1, 2, \dots N\}$ be a sequence of $(G \times 1)$ i.i.d. random vectors, such that $E(|w_{ig}|) < \infty \; \forall g = 1, 2, \dots, G$, then*

$$N^{-1} \sum_{i=1}^{N} w_i \xrightarrow{p} E(w_i)$$

You must have seen theorem 1 somewhere in your previous courses and I hope you can now appreciate its power and beauty.[15] It essentially says that if you simply have i.i.d. random numbers with finite means, the sample average of such random numbers will converge in probability to their true population mean. To bring this to our standard econometrics set up, what we usually have is a vector of data for a sample of N observations. For example, our w_i might be the vector $(y_i, x_{i1}, x_{i2}, \dots, x_{iK})$ of the dependent variable and all the explanatory variables. The weak law of large numbers says that, if the observations are i.i.d (and they certainly are if the sample is a random extraction from the population), then the sample averages of each

[15] I agree there are more beautiful things in life but still...

variable are consistent estimators for the population averages. As you can certainly imagine, theorem 1 is the fundament of what we called the *analogy principle* that we used earlier to construct $\widehat{\beta}$. In fact, the analogy principle is an extension of the weak law of large to functions of random variable. Theorem 1, in fact, simply says that, under its assumptions, sample averages converge in probability to population averages. The analogy principle needs more than that, it needs functions of sample averages to converge to the same functions of population averages. Fortunately, the following theorem guarantees precisely that:

Theorem 2 (Slutsky's theorem) *If $w_N \xrightarrow{p} w$ and if g is a continuous function, then $g(w_N) \xrightarrow{p} g(w)$.*

Importantly, theorem 2 also motivates why we often prefer to look at asymptotic consistency rather than *unbiasedness*, a property of estimators that you have certainly seen in previous courses.

Definition 4 (Unbiased estimator) *The random vector $\widehat{\beta}$ is an **unbiased** estimator of the vector of constants β if $E(\widehat{\beta}) = \beta$.*

In fact, the expectation operator does not work through non linear functions. In particular, using the same notation of theorem 2, if $E(w_N) = w$ then $E(g(w_N)) = g(w)$ only if g is a *linear* function (on top of being continuous, which is still required). This means that for estimators that are non linear functions of the random variables of our model we would not be able to check unbiasedness. And most of these estimators that we will see in his course are in fact non linear.

Convergence in distribution

Often our interest in an estimator lies not only in the knowledge of its magnitude but also in the possibility of running statistical tests, for example about the estimator being different from zero or from another estimator. To do that we need to know the distribution of the estimator because that is the basis for constructing any test statistics.

The following definition helps us identifying the distribution of an estimator:

Definition 5 (Convergence in distribution) *A sequence of random variables $\{x_N; N = 1, 2, \dots\}$ converges in distribution to the continuous random variable x if and only if:*

$$\lim_{N\to\infty} F_N(\xi) = F(\xi) \quad \forall \xi \in R$$

where F_N is the cdf of x_N and F the cdf of x. And we write

$$x_N \xrightarrow{d} x$$

So, if the x_N's are estimators of some underlying parameters produced on samples of various sizes N and if we know that they converge in distribution to a random variable x with some known distribution F, then we can also say that the asymptotic distribution of any of these x_N's is F and we write it as follows:

$$x_N \overset{a}{\sim} F$$

Similarly to how we proceeded with consistency, we now need a theorem to make definition 5 operational, a theorem that allows to identify the asymptotic distribution of a given estimator. Here is the theorem we need:

Theorem 3 (Central Limit Theorem) *Let $\{w_i : i = 1, 2, \dots N\}$ be a sequence of $(G \times 1)$ i.i.d. random vectors, such that $E(w_{ig}^2) < \infty \ \forall g = 1, 2, \dots, G$ and $E(w_i) = 0$, then*

$$N^{-1/2} \sum_{i=1}^{N} w_i \xrightarrow{d} N(0, B)$$

where $B = Var(w_i) = E(w_i w_i')$

This theorem is even more powerful and more beautiful than the weak law of large numbers! It essentially says that, under conditions that are very easily met by most available datasets (the observations are i.i.d. with finite variance), the sample mean (adjusted by the division by $N^{-1/2}$ instead of simply N^{-1}) is asymptotically distributed according to a *normal*. So regardless of the actual (small-sample) distribution of w_i, the asymptotic distribution is normal (which is in fact one of the reasons why the normal is actually called normal!).

1.5 Asymptotic properties of OLS

Let us now apply the results of section 1.4 to our simple OLS estimator $\widehat{\beta}$ to show that it is consistent and asymptotically normal. Let us start with consistency.

Consistency of OLS (and unbiasedness)

Consider the multivariate version of the model as described in equation 1.20 and rewrite the OLS estimator of β as follows:

$$\begin{aligned}\widehat{\beta} = (X'X)^{-1}X'Y &= (X'X)^{-1}X'(X\beta + u) \\ &= \beta + (X'X)^{-1}X'u \\ &= \beta + \left(N^{-1}\sum_{i=1}^{N} x_i'x_i\right)^{-1}\left(N^{-1}\sum_{i=1}^{N} x_i'u_i\right)\end{aligned}$$

Hence:

$$plim\widehat{\beta} = \beta + plim\left[\left(N^{-1}\sum_{i=1}^{N} x_i'x_i\right)^{-1}\left(N^{-1}\sum_{i=1}^{N} x_i'u_i\right)\right]$$

Let us now make use of theorem 1 to derive the following probability limits:

1. $$plim N^{-1}\sum_{i=1}^{N} x_i'x_i = E(x_i'x_i) = A$$

   and assume that A is a finite and semipositive matrix (which implies that the assumptions of both theorem 1 and 3 as regards the moments of the random variables being finite are satisfied).[16]

2. $$plim N^{-1}\sum_{i=1}^{N} x_i'u_i = E(x_i'u_i) = 0$$

   as implied by assumption 1.21.

Hence, thanks to theorem 2 we can simply write:

$$plim\ \widehat{\beta} = \beta + A^{-1}\cdot 0 = \beta$$

which shows that $\widehat{\beta}$ is a consistent estimator for β.

Note incidentally that assumption 1.21 is not sufficient to guarantee unbiasedness of $\widehat{\beta}$. To see this, write the expected value of $\widehat{\beta}$ as follows:

$$E(\widehat{\beta}) = \beta + E\left[(X'X)^{-1}X'u\right]$$

[16] This assumption is in fact equivalent to the more standard rank assumption that you typically find in most textbooks: $rank\ E(X'X) = K$.

where to obtain unbiasedness we need the term $E\left[(X'X)^{-1}X'u\right]$ to be equal to zero. However, unlike the *plim*, the expectation operator does not work through non linear functions and $(X'X)^{-1}X'u$ is a very non-linear function of the x_i's and the u_i's and cannot be written as $E\left[(X'X)^{-1}\right]E(X'u)$. To obtain unbiasedness we need the following stronger assumption:

$$E(u|X) = 0 \tag{1.29}$$

which guarantees that u is uncorrelated with any function of X and thus also with $(X'X)^{-1}X'$. Under assumption 1.29 we can then first derive the expected value of $\widehat{\beta}$ conditional on X:

$$\begin{aligned} E(\widehat{\beta}|X) &= \beta + E\left[(X'X)^{-1}X'u|X\right] \\ &= \beta + (X'X)^{-1}X'E(u|X) = \beta \end{aligned}$$

and then, by simple iterated expectation:

$$E(\widehat{\beta}) = E_X\left[E(\widehat{\beta}|X)\right] = E_X(\beta) = \beta$$

which shows that $\widehat{\beta}$ is unbiased, but only under assumption 1.29 and not the weaker assumption 1.21.

Asymptotic normality of OLS

To derive the asymptotic distribution of $\widehat{\beta}$ it is useful to start from the following expression:

$$\begin{aligned} \sqrt{N}(\widehat{\beta} - \beta) &= \left[N^{-1}\sum_{i=1}^{N} x_i'x_i\right]^{-1}\left[N^{-1}\sum_{i=1}^{N} x_i'u_i\right]N^{1/2} \\ &= \left[N^{-1}\sum_{i=1}^{N} x_i'x_i\right]^{-1}\left[N^{-1/2}\sum_{i=1}^{N} x_i'u_i\right] \end{aligned} \tag{1.30}$$

Since $\left[N^{-1}\sum_{i=1}^{N} x_i'x_i\right]$ converges in probability to the finite semipostive matrix A (as we have assumed earlier), some simple lemmas of theorems 1 and 3 guarantee that the asymptotic distribution of the expression in equation 1.30 is the same as the asymptotic distribution of $A^{-1}\left[N^{-1/2}\sum_{i=1}^{N} x_i'u_i\right]$.[17]

[17] We are not going into the details of these lemmas and how they work. You find all the details in the textbook.

Let us focus, then, on the expression $\left[N^{-1/2}\sum_{i=1}^{N} x_i'u_i\right]$ which satisfies all the assumptions of the central limit theorem (theorem 3): it is the sum, normalized by $N^{-1/2}$, of a sequence of i.i.d. random vectors with zero mean (and this is guaranteed by assumption 1.21) and finite variance. Then, by direct application of the central limit theorem:

$$N^{-1/2}\sum_{i=1}^{N} x_i'u_i \xrightarrow{d} N(0, B) \tag{1.31}$$

where $B = Var(x_i'u_i)$. With the result in expression 1.30, it is then easy to derive the following asymptotic distribution:

$$\sqrt{N}(\widehat{\beta} - \beta) \xrightarrow{d} N(0, A^{-1}BA^{-1}) \tag{1.32}$$

Let us make things simple and assume *homoskedasticity* and no *serial correlation*, i.e. $Var(u_i|x_i) = E(u_i^2|x_i) = \sigma^2$, so that:[18]

$$\begin{aligned} B = Var(x_i'u_i) &= E(x_i'u_iu_ix_i) \\ &= E(u_i^2x_i'x_i) = \sigma^2E(x_i'x_i) = \sigma^2A \end{aligned}$$

This simplifies expression 1.32 to:

$$\sqrt{N}(\widehat{\beta} - \beta) \xrightarrow{d} N(0, \sigma^2A^{-1}) \tag{1.33}$$

and makes it easy to proceed to the following derivation:

$$\begin{aligned} (\widehat{\beta} - \beta) &\xrightarrow{d} N(0, N^{-1}\sigma^2A^{-1}) \\ \widehat{\beta} &\xrightarrow{d} N(\beta, N^{-1}\sigma^2A^{-1}) \\ \widehat{\beta} &\xrightarrow{d} N(\beta, \sigma^2\frac{E(X'X)^{-1}}{N}) \end{aligned} \tag{1.34}$$

which defines the asymptotic distribution of the OLS estimator.[19]

[18] By homoskedasticity we mean that all errors have the same (finite) variance: $Var(u_i) = Var(u_j)\ \forall i, j$. By serial correlation we mean the presence of some type of correlation across error terms: $Cov(u_i, u_j) \neq 0$ for some $i \neq j$. Very frequently, these two features (homoskedasticity and lack of serial correlation) are assumed together. For simplicity, unless specified differently, when we assume homoskedasticity we will also mean to assume lack of serial correlation.

[19] Also for this last derivation we have used some lemmas and corollaries of the main theorems 1 and 3 which we are not discussing in details here. All the details are in the textbook.

We now know a lot about $\widehat{\beta}$, in particular we know its probability limit and its asymptotic distribution. However, in order to be able to really use the distribution in expression 1.34 we need one additional bit. In fact, in that expression the asymptotic variance-covariance matrix of $\widehat{\beta}$ depends on a parameter σ^2 and a matrix $E(X'X)$ that need to be estimated. The simplest way of obtaining consistent estimators for these unknowns is, again, to apply the analogy principle and estimate them using their sample counterparts:[20]

$$\widehat{\sigma}^2 = N^{-1}\sum_{i=1}^{N}\widehat{u}_i^2 = N^{-1}\sum_{i=1}^{N}(y_i - x_i\widehat{\beta})^2 \tag{1.35}$$

$$\widehat{A} = \widehat{E(X'X)} = N^{-1}\sum_{i=1}^{N}x_i'x_i \tag{1.36}$$

Finally, how can we proceed in case we are not willing to assume homoskedasticity of the error term? Nothing particularly complicated. In that case the asymptotic variance-covariance matrix of $\widehat{\beta}$ does not simplify to $\sigma^2\frac{E(X'X)^{-1}}{N}$ and remains equal to $\frac{A^{-1}BA^{-1}}{N}$ which can be easily estimated with $\widehat{A}$ as defined in equation 1.36 and the following:

$$\widehat{B} = N^{-1}\sum_{i=1}^{N}(\widehat{u}_i^2 x_i'x_i) \tag{1.37}$$

In this case we can express the (estimated) asymptotic variance-covariance matrix of $\widehat{\beta}$ as follows:

$$\widehat{AVar}(\widehat{\beta}) = \left(\sum_{i=1}^{N}x_i'x_i\right)^{-1}\left(\sum_{i=1}^{N}\widehat{u}_i^2x_i'x_i\right)\left(\sum_{i=1}^{N}x_i'x_i\right)^{-1} \tag{1.38}$$

which is also called the *Huber-White robust variance-covariance matrix.*[21]

Note, however, that whether the errors of the model are homoskedastic or heteroskedastic is only a matter of *efficiency* and not of consistency. In other words, if the errors are homoskedastic, $\widehat{\beta}$ will have a smaller variance (i.e. more efficient) but consistency is guaranteed under any assumption

[20] In previous econometrics courses you might have used a slightly different estimator for $\widehat{\sigma}^2$: $\widehat{\sigma}^2 = (N-K)^{-1}\sum_{i=1}^{N}\widehat{u}_i^2$. This alternative estimator is not only consistent but also unbiased. For consistency, however, the correction for the degrees of freedom of the model is not necessary.

[21] The notation $AVar$ to indicate the asymptotic variance of a random number is pretty standard and we will make use of it several times later on in the course.

about the variance-covariance structure of the error term as long as assumption 1.21 is satisfied. Knowing the distribution of $\widehat{\beta}$ it is then easy to derive the distribution of various test statistics like the t-test or the F-test. Moreover, the usual standard errors that are commonly reported in tables of econometric results are nothing but the square-roots of the elements on the main diagonal of the variance-covariance matrix of the estimator.

1.6 Additional topics in single-equation linear models

Omitted variable bias

We have already seen in section 1.1 that variables omitted from the specification of the model are one of the possible sources of error. In this section we look more in details at what are the conditions that make such omission more or less problematic.

For simplicity, suppose the true model is simply:

$$y_i = \beta_0 + \beta_1 x_i + \gamma q_i + u_i \tag{1.39}$$

with $Cov(u_i, x_i) = 0$ and $Cov(u_i, q_i) \neq 0$. In specifying the model for estimation, however, we omit the variable q_i. This may happen for many reasons: we may simply forget about it or wrongly think that it is not important or - and this is the most common and most problematic case - q_i might simply be not observable in our data. Then, the model that we end up estimating has a composite error term that includes q_i:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{1.40}$$

where $\epsilon_i = \gamma q_i + u_i$.

What happens if we run OLS on the model in equation 1.40? The OLS estimator will be computed as follows:

$$\widehat{\beta}_1 = \frac{\widehat{Cov(y_i, x_i)}}{\widehat{Var(x_i)}} = \frac{N^{-1}\sum_{i=1}^{N}(y_i - \overline{y})(x_i - \overline{x})}{N^{-1}\sum_{i=1}^{N}(x_i - \overline{x})^2} \tag{1.41}$$

By the analogy principle, this estimator converges in probability to the following:

$$\begin{aligned} \widehat{\beta}_1 \quad &\xrightarrow{p} \quad \frac{Cov(y_i, x_i)}{Var(x_i)} = \frac{Cov\left[(\beta_0 + \beta_1 x_i + \gamma q_i + u_i)x_i\right]}{Var(x_i)} \\ &= \quad \frac{\beta_1 Var(x_i) + \gamma Cov(q_i, x_i) + Cov(u_i, x_i)}{Var(x_i)} \\ &= \quad \beta_1 + \gamma \frac{Cov(q_i, x_i)}{Var(x_i)} \end{aligned} \tag{1.42}$$

Expression 1.42 shows that $\widehat{\beta}_1$ is consistent only if either $\gamma = 0$ or $Cov(q_i, x_i) = 0$. In the first case ($\gamma = 0$) omitting q_i from the model was in fact the right choice to make becasue it does not appear there. In the second case ($Cov(q_i, x_i) = 0$) the omitted variable is in fact part of the model but since it is uncorrelated with x_i its omission does not affect the estimation of the parameter β_1.

This result is very important. It tells us what I consider one of the most important rules in the practice of econometrics. If we are interested in the consistent estimation of some parameters of interest (β_1 in the example above) we should not throw into the model all possible regressors. We should instead include only those that are likely to be correlated with our variables of interest (and that are also likely to influence the outcome y). Any other variable, even if it explains a lot of the variation in y, is not necessary and can be safely omitted. Things are obviously different if the purpose of our study is making predictions.

Measurement error

In section 1.1 we discussed measurement issues as a potential source of error in the model. But, what are exactly the implications of measurement error for the estimation of the parameters of the model? We will consider two separate cases. First, we look at the implications of having measurement error in the dependent variable of the model and, then we will assume that one (or more) of the regressors are measured with error.

Measurement error in the dependent variable. Call y^* the true dependent variable so that the correct specification of the model is:

$$y^* = X\beta + u \tag{1.43}$$

Also assume that the model 1.43 satisfies all the nice OLS assumptions that we stated earlier. In particular, $E(u|X) = 0$. Unfortunately, however, we only observe a noisy measure of y^* which we call simply y and define as follows:

$$y = y^* + \epsilon \tag{1.44}$$

where ϵ is the error of measurement. For the time being, we do not impose any specific assumption on ϵ (other than its not being observable). So, in practice, we would like to estimate model 1.43 but that specification is not usable because y^* is not observable. Then, the only thing we can do is

replace y^* in model 1.43 and see what that implies for the estimation. Model 1.43 written in a form that can be used for estimation is the following:

$$y = X\beta + (\epsilon + u) \tag{1.45}$$

Now, the only reason why the estimation of the parameters in this model might be problematic is correlation between the error of measurement ϵ and some of the regressors X. One typical assumption about the error of measurement is its randomness, i.e. its being uncorrelated with anything else in the model. This assumption (which we will see more in details later) is very reasonable if the error arises from things like misreporting, errors in variable coding, et. So, unless there are particular reasons to believe that the measurement error is somehow correlated with some of the regressors, the only implication of using a mismeasured dependent variable is the larger variation of the error term of the model which, in turn, implies less efficient estimates, i.e. estimators with a larger standard errors.

Measurement error in one explanatory variable. The second instance of measurement error is in the explanatory variables. As we will see, this is typically a much more serious problem. Consider a simple case of a univariate model of the following type:

$$y = \alpha + \beta x^* + u \tag{1.46}$$

which satisfies the assumption $E(u|x^*) = 0$ and could therefore be consistently estimated with OLS. Unfortunately, the variable x^* is not observable. Instead, we observe a noisy measure of it:

$$x = x^* + \epsilon \tag{1.47}$$

Again, let us replace equation 1.47 into model 1.46 to obtain a specification of the model that can be estimated with the observable variables:

$$y = \alpha + \beta x + (u - \beta\epsilon) \tag{1.48}$$

where $(u - \beta\epsilon)$ is the new composite error term of the model. In other words, equation 1.48 describes the model that would be estimated if we run a regression of y on the observable x, which is essentially the only reasonable thing we can do. In such regression, the error term will include also the error of measurement.[22] The key to understand where the problem arises is the

[22]In all these models (1.43, 1.45, 1.46 and 1.48) one possible interpretation of the error term u is that it includes all other possible sources of error that are not the type of measurement error considered. For example, in model 1.46 or 1.48, u may arise from measurement error in the dependent variable y or from specification error.

endogeneity of x in model 1.48 which is necessarily correlated with the error term. To see that, consider the simplest possible case in which the error of measurement is totally random:[23]

$$Cov(x^*, \epsilon) = 0 \tag{1.49}$$
$$Cov(u, \epsilon) = 0 \tag{1.50}$$

Even in this sort of *best-case scenario*, however, the regressor x is endogenous in model 1.48 because it is correlated with part of the error term, namely with ϵ:

$$Cov(x, \epsilon) = Cov[(x^* + \epsilon), \epsilon] = \sigma_\epsilon^2$$

where σ_ϵ^2 is the variance of the error of measurement. So, it should be clear that the OLS estimator of β produced using a mis-measured regressor is not going to be consistent. Let us look in details at the bias of such estimator:

$$\begin{aligned} \widehat{\beta} \quad &\xrightarrow{p} \quad \frac{Cov(x, y)}{Var(x)} = \frac{Cov[(x^* + \epsilon), (\alpha + \beta x^* + u)]}{Var(x^*) + Var(\epsilon)} \\ &= \quad \beta \frac{Var(x^*)}{Var(x^*) + Var(\epsilon)} < \beta \end{aligned}$$

This result tells us something important: if we estimate parameters using mis-measured explanatory variables what we obtain is an under-estimate of the true parameter. This is not of great consolation but at least we know the direction of the bias. In the literature the bias arising from this type of measurement is called *attenuation bias*.

2 Instrumental variable estimation

Instrumental variable estimation is the classical solution to the problem of *endogeneity*. A variable is considered to be *endogenous* in a model when it is correlated with the error term, so that assumption 1.21 fails. For example, consider the following multivariate model:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{K-1} x_{K-1} + \beta_K x_K + u \tag{2.1}$$

where the following assumptions holds:

- $E(u) = 0$;

[23] *@@ questa nota è uguale alla 9: la toglierei!:* Measurement error in the explanatory variable that satisfies the conditions 1.49 and 1.50 is called *classical measurement error*.

- $Cov(x_j, u) = 0 \ \forall j = 1, 2, \ldots, K-1$;
- but $Cov(x_K, u) \neq 0$.

In the previous sections we have seen that problems of endogeneity may arise from various sources: omitted variables, measurement error, etc. Moreover, notice that the problem is particularly worrisome due to the fact that endogeneity of one regressor typically prevents consistent estimation of all the other parameters of the model.[24]

The instrumental variable (IV) approach to solve endogeneity is based on the idea of finding an external variable z_1, the instrument, that satisfies the following two important properties:

IV Assumption 1

$Cov(z_1, u) = 0$

which essentially means that the instrument should not be endogenous itself, i.e. it should be *exogenous*;

IV Assumption 2

$Cov(z_1, x_K | x_1, \ldots, x_{K-1}) \neq 0$*, and possibly this correlation should be large (positive or negative does not matter).*

The conditioning of the other exogenous variables of the model guarantees that the correlation between the instrument and the endogenous variable is not spuriously driven by other regressors. The importance of this detail will be clear later.

Finding such instrumental variable is the most difficult part of the entire procedure. There is no predetermined procedure that can be followed to find a good instrument. It's all about being smart and creative. And convincing. In fact, as we will see later, while Assumption 1 can be tested empirically, there is no way to test Assumption 2 (other than having an alternative instrument) and you will just have to find an idea for an instrument that is so smart and convincing that people reading your results will have nothing to criticize! In this sense, you can see IV as the channel through which creativity enters the realm of econometrics and statistics, which makes everything a lot more romantic but also a lot more complicated!

To give you an example of a good instrument, here is one that is often used in studies of family economics. Suppose you want to study the wage penalty that young mothers pay when they re-enter the labour market after

[24] You are going to show this in a problem set.

pregnancy. So, suppose you have data on a sample of working women and your main equation has wages on the left-hand-side and on the right-hand-side an indicator for whether the woman had a baby in the previous 6-12 months plus a set of controls. You are obviously (and rightly) worried that the motherhood indicator is endogenous in this equation. Your brilliant idea for the instrument is the gender of previous children in the family. There is ample evidence that the likelihood of having a second baby is higher if the first baby was a girl. This is particularly true in developing countries but it also applies to industrialized ones. Even more robust is the higher likelihood of an additional pregnancy if in the family there are only children of the same sex (if you have two boys you're more likely to go for a third children than if you had a boy and a girl). These types of variables seem fairly exogenous as (generally) one cannot do much to choose the gender of one's children (although selective abortion is an issue in some developing countries. A leading paper on this issue is Oster, E. 2005. *"Hepatitis B and the Case of the Missing Women."* Journal of Political Economy, vol. 113(6), 1163-1216.).

Once, you have found your brilliant idea for an instrument, things become easy. In the original model there are K parameters to be estimated. To do that we can use the following set of K moment conditions:

$$\begin{aligned} E(x_1 u) &= 0 \\ E(x_2 u) &= 0 \\ &\vdots \\ E(x_{K-1} u) &= 0 \\ E(z_1 u) &= 0 \end{aligned}$$

which can be written jointly as:

$$E(z'u) = 0 \tag{2.2}$$

where $z = (x_1, x_2, \ldots, x_{K-1}, z_1)$ is the vector that includes all the exogenous variables of the model: all the x's, excluding x_K that is endogenous, and the instrument z_1.

Using assumption 2.2 it is easy to show that the vector of parameters is in fact identified:

$$\begin{aligned} E(z'u) &= E[z'(y - x\beta)] \\ &= E(z'y) - E(z'x)\beta \\ \beta &= E(z'x)^{-1} E(z'y) \end{aligned} \tag{2.3}$$

Now, we can derive a consistent estimator of β by simply applying the analogy principle to equation 2.3:

$$\widehat{\beta}_{IV} = \left[N^{-1} \sum_{i=1}^{N} z_i' x_i \right]^{-1} \left[N^{-1} \sum_{i=1}^{N} z_i' y_i \right] \tag{2.4}$$

The painful discussion of asymptotic theory in section 1.4 now comes handy because it guarantees that $\widehat{\beta}_{IV}$ is a consistent and asymptotically normal estimator of β.

Obviously, this estimation method extends directly to cases in which the number of endogenous variables is greater than one, in which case we will have to find (at least) one instrument for each of them. For example, if in the previous model also x_{K-1} were endogenous we would have to find an additional instrument z_2 such that $E(z_2 u) = 0$ and $Cov(z_2, x_{K-1}) \neq 0$. Then, we would simply redefine the vector of all exogenous variables as $z = (x_1, x_2, \ldots, x_{K-2}, z_2, z_1)$ and proceed as before to compute $\widehat{\beta}_{IV}$.

Models of the type discussed in this section where the number of instruments is exactly equal to the number of endogenous variables are called *just identified*. In the following section we consider *over-identified* models, that is models where the number of instruments exceeds the number of endogenous variables.

2.1 Multiple instruments: two-stages least squares (2SLS)

In some fortunate cases you are so smart that you find more than just one instrument for each (or some) of the endogenous variable. In these cases the model is called *over-identified*, meaning that there is more than just one way to compute a consistent estimator for the parameters.

Let us keep things simple and consider a model with just one endogenous variable, the same model of the previous section, but several instruments $z_1, \ldots, z_M$, all of them satisfying the conditions to be valid instruments:

$$\begin{aligned} Cov(z_h, u) &= 0 \ \forall h = 1, \ldots, M \\ Cov(x_K, z_h | x_1, \ldots, x_{K-1}) &\neq 0 \ \forall h = 1, \ldots, M \end{aligned}$$

In principle, with all these instruments we could construct up to M different IV estimators. Actually, a lot more. In fact, any linear combination of two or more of the M instruments is also a valid instrument. So the potential set of $\widehat{\beta}_{IV}$ that we could construct is very large and the question is which one to choose.

Remember that one of the properties of a good instrument is that it should be strongly correlated with the endogenous variable.[25] Hence, it seems reasonable to choose as instrument the one particular linear combination of all instruments that maximizes the correlation with the endogenous variable. But how do we find such linear combination? The simplest way to do that is to run a OLS regression of the endogenous variable x_K on all the instruments:

$$x_K = \vartheta_1 z_1 + \cdots + \vartheta_M z_M + \delta_1 x_1 + \cdots + \delta_{K-1} x_{K-1} + e \tag{2.5}$$

The estimated ϑ's and δ's obtained from such regression will be used as the coefficients of the linear combination:

$$\widehat{x}_K = \widehat{\vartheta}_1 z_1 + \cdots + \widehat{\vartheta}_M z_M + \widehat{\delta}_1 x_1 + \cdots + \widehat{\delta}_{K-1} x_{K-1} \tag{2.6}$$

Now we can proceed as if we had only one instrument, $\widehat{x}_K$. Notice that in equation 2.5 I have included also all the other exogenous variables in the model. We will discuss the reason for this in the next session. For now make a little act of faith and accept that this is the right way to proceed. Let me also further clarify why the OLS coefficients of equation 2.5 are actually the coefficients of the linear combination of the instruments (and the other exogenous variables) that maximize correlation with x_K. Remember that the OLS method minimizes the squared residuals. In other words, the OLS method looks for the coefficients that make the right-hand-side of equation 2.5 the most similar to the left-hand-side, which essentially amounts to maximizing the covariance between the two sides (as you make them as similar as possible).[26] To conclude, this new IV estimator, which is called *Two Stages Least Squares* (2SLS), can be derived using $\widehat{x}_K$ as a single instrument for x_K and applying the same procedure as in section 2.

Define $\widehat{z}$ the vector of exogenous variables analogously to z in section 2: $\widehat{z} = (x_1, x_2, \ldots, x_{K-1}, \widehat{x}_K)$. And compute the 2SLS estimator as:

$$\widehat{\beta}_{2SLS} = \left[N^{-1} \sum_{i=1}^{N} \widehat{z}_i' x_i \right]^{-1} \left[N^{-1} \sum_{i=1}^{N} \widehat{z}_i' y_i \right] \tag{2.7}$$

[25] We will clarify the reason why such condition is important in the next section when we discuss the issue of *weak instruments*.

[26] Another, simpler way of saying this is that $\widehat{x}_K$ is the *linear projection* of x_K on all the exogenous variables, i.e. all instruments as well as all the exogenous regressors. If you do not remember exactly what a linear projection is, you can find more details in paragraph 2.3 in the textbook (page 24).

Again, thanks to the few things about asymptotic theory we discussed in section 1.4, we automatically know that $\widehat{\beta}_{2SLS}$ is consistent and asymptotically normal.[27]

But why did we call this estimator two-stages-least-squares? That's because it can be obtained by a simple two-step procedure:

1. **First stage.** Regress each endogenous variable on all exogenous ones, i.e. all instruments and all exogenous regressors, and obtain predicted values;

2. **Second stage.** In the main model, replace the endogenous variables with their predictions from the first stage regressions and run OLS.

The resulting OLS estimator from the second stage regression is in fact $\widehat{\beta}_{2SLS}$.[28]

Finally, notice that when the model is just-identified, the simple IV and the two-stages procedure lead exactly to the same estimator. To see this equality, notice that the OLS estimator from the second stage regression can be written in full matrix notation as:

$$\widehat{\beta}_{2SLS} = (\widehat{Z}'\widehat{Z})^{-1}(\widehat{Z}'Y) \tag{2.8}$$

Also, remember that $\widehat{Z}$ is the matrix of predictions from the first stage regression and can thus be expressed as:[29]

$$\widehat{Z} = Z(Z'Z)^{-1}Z'X \tag{2.9}$$

If we now replace this expression into the full-matrix notation of $\widehat{\beta}_{2SLS}$ we

[27] Finding its exact asymptotic variance-covariance matrix is a bit more complex than usual because we should take into account the fact that in computing this estimator we are using one variable that is itself an estimate (and thus has some additional variation). All standard statistical packages compute 2SLS with correct standard errors and we skip this derivation. However, you should keep in mind that whenever in an estimation procedure you use a variable that is itself an estimate you should worry about the computation of the standard errors. You can read more about this in the textbook, section 6.1 page 155.

[28] The standard errors of this second stage regression, however, will have to be adjusted to account for the fact that one (or more) of the regressors are estimates.

[29] The prediction from the first stage is simply Z times the estimated set of coefficients, whose expression is in fact $(Z'Z)^{-1}Z'X$. The matrix notation is useful because it automatically takes into account that only one of the elements in Z is actually estimated while all the others are simply repeated.

obtain exactly the estimator of equation 2.7 in full-matrix notation:

$$\begin{aligned}
\widehat{\beta}_{2SLS} &= (\widehat{Z}'\widehat{Z})^{-1}(\widehat{Z}'Y) \\
&= (\underbrace{X'Z(Z'Z)^{-1}Z'}_{\widehat{Z}'}\underbrace{Z(Z'Z)^{-1}Z'X}_{\widehat{Z}})^{-1}(\widehat{Z}'Y) \\
&= (\underbrace{X'Z(Z'Z)^{-1}Z'}_{\widehat{Z}'}X)^{-1}(\widehat{Z}'Y) \\
&= (\widehat{Z}'X)^{-1}(\widehat{Z}'Y)
\end{aligned}$$

2.2 Additional (but important!) notes on IV

Why do we put all the exogenous variables in the first stage regression?

To clarify this point, let us consider a very simple example of a model with just two regressors, one of which is endogenous:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u \tag{2.10}$$

with $E(u) = 0$, $Cov(x_1, u) = 0$ but $Cov(x_2, u) \neq 0$. Also suppose that we have a valid instrument for x_2, a variable z_1 such that $Cov(z_1, u) = 0$ and $Cov(z_1, x_2) \neq 0$.

Now, consider what happens if we omit x_1 from the first-stage regression, i.e. if we run the first-stage regression only on the instrument. We still want to allow the possibility that x_1 enters the specification of x_2 so let us write the first stage regression as follows:

$$x_2 = \vartheta_0 + \vartheta_1 z_1 + (\delta_1 x_1 + e) = \vartheta_0 + \vartheta_1 z_1 + v \tag{2.11}$$

where v is a composite error term equal to $\delta_1 x_1 + e$. If the two regressors x_1 and x_2 are unrelated to each other then δ_1 would be equal to zero.

So, if we run the first stage regression without x_1 the prediction that we obtain is $\widetilde{x}_2 = \widetilde{\vartheta}_0 + \widetilde{\vartheta}_1 z_1$.[30] The residual of this regression is $\widetilde{v} = x_2 - \widetilde{x}_2$ and, by the analogy principle, it converges in probability to the composite error term $v = \delta_1 x_1 + e$.

So, when we replace x_2 in equation 2.10 with $\widetilde{x}_2$ we obtain the following:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 \widetilde{x}_2 + (\beta_2 \widetilde{v} + u) \tag{2.12}$$

[30] We use $\widetilde{\cdot}$ instead of $\widehat{\cdot}$ to differentiate this analysis from the one developed in section 2.1.

which shows that unless β_2 or δ_1 are equal to zero, x_1 will be correlated with the error term of the second-stage regression and will thus be endogenous and impede identification of all the parameters in the model. In fact, the error term of the second stage regression is $\beta_2\widetilde{v} + u$ and is asymptotically equal to $\beta_2(\delta_1 x_1 + e) + u$ which is by definition correlated with x_1.

Notice that the instances which make the omission of x_1 in the first-stage regression irrelevant are rather peculiar. If δ_1 is equal to zero, that means that x_1 and x_2 are uncorrelated with each other (conditional on the instrument) which makes the inclusion of x_1 in the main model also irrelevant for the consistent estimation of β_2. If, instead, β_2 is equal to zero than the true model simply does not include x_2 which eliminates any problem with endogeneity from the very start.

Weak instruments: why should the instrument be *highly* correlated with the endogenous variable?

Consider a simple model with just one regressors that is endogenous:

$$y = \beta_0 + \beta_1 x_1 + u \tag{2.13}$$

and suppose there is one valid instrument z available to construct an IV estimator. We know that that estimator converges in probability to the following expression:

$$\begin{aligned} \widehat{\beta}_{IV} \quad &\xrightarrow{p} \quad \frac{Cov(z,y)}{Cov(z,x)} = \frac{Cov[z,(\beta_0 + \beta_1 x_1 + u)]}{Cov(z,x)} \\ &= \quad \beta_1 + \frac{Cov(z,u)}{Cov(z,x)} \end{aligned}$$

If the instrument is valid, then asymptotically $Cov(z,u) = 0$ and the probability limit of $\widehat{\beta}_{IV}$ is simply β. However, this result is correct only as $N \to \infty$ while in small samples the $Cov(z,u)$ will never be exactly equal to zero due to sampling variation.[31] If the instrument is valid, then, $\widehat{\beta}_{IV}$ will be certainly consistent but might be subject to a small sample bias (just like all estimators). Notice, however, that if the instrument is *weak*, that is only weakly correlated with the endogenous variable, then $Cov(z,x)$ will

[31] Notice that when we talk about *small samples* we do not necessarily mean samples of small size. We simply intend to refer to non-asymptotic properties. In this terminology, a small sample is any sample with N smaller than ∞.

be small and the small sample bias might in fact become very large even if $Cov(z, u)$ is also small.[32]

This problem is particularly worrisome since several studies have shown that weak instruments can induce potentially very large biases also with relatively big samples (up to 100,000 observations).[33] As a rule of thumb, an instrument is considered weak if the t statistics in the first-stage regression is smaller than 10-15. If you have more than one instrument you should look at the F statistics for the test of joint significance of all instruments (excluding the other exogenous regressors).

Testing endogeneity: the Hasuman test

Is it possible to test whether a regressor is endogenous? The answer to this question is yes, however, such test can only be performed once an instrument for the potentially endogenous regressor is found. And we know that finding instruments is the most complicated part of the entire process so in some sense the test for endogeneity comes a bit too late. In principle we would like to know whether a regressor is endogenous or not before wasting our precious time looking for a nice idea for an instrument. It is only when we have found one that the endogeneity test can effectively be performed.[34]

The test, called *Hausman test* after the name of Jerry Hausman who first proposed it, is based on the following idea. Suppose you have a model where you suspect one (or more) regressors to be endogenous and you have found the necessary valid instruments to construct an IV estimator. Now, if the regressor(s) are really endogenous, the OLS estimator will be biased while the IV estimator will be consistent:

$$\begin{aligned} H_1 \quad &: \quad E(u|x) \neq 0 \text{ endogeneity} \\ \widehat{\beta}_{OLS} \quad &\xrightarrow{p} \quad \beta + bias \\ \widehat{\beta}_{IV} \quad &\xrightarrow{p} \quad \beta \\ \widehat{\beta}_{IV} - \widehat{\beta}_{OLS} \quad &\xrightarrow{p} \quad bias \neq 0 \end{aligned}$$

[32] How strong the correlation between the instrument and the endogenous variable should be is a very subtle issue. In principle, the instrument should capture all the variation in the endogenous variable that is not correlated with the error term and thus induces endogeneity. For this reason we do not want such correlation to be too high. At the same time, however, the instrument should be strong enough to avoid weak-instrument bias in small samples. There is no technical solution to this trade-off and you will have to evaluate the goodness of your instrument in the light of the specific setting case by case.

[33] See Staiger, D. and J.H. Stock. *"Instrumental Variables Regression with Weak Instruments."* Econometrica, vol.65, 557-586.

[34] As you can guess from this short preamble, I am not a great fan of the Hausman test.

Under the alternative hypothesis that the model is not affected by endogeneity, both estimators are consistent:

$$\begin{aligned} H_0 \quad &: \quad E(u|x) = 0 \text{ no endogeneity} \\ \widehat{\beta}_{OLS} \quad &\xrightarrow{p} \quad \beta \\ \widehat{\beta}_{IV} \quad &\xrightarrow{p} \quad \beta \\ \widehat{\beta}_{IV} - \widehat{\beta}_{OLS} \quad &\xrightarrow{p} \quad 0 \end{aligned}$$

The idea, then, is to test hypothesis H_0 by testing that the difference between the OLS and the IV estimator is asymptotically equal to zero. To this end, we construct the quadratic form of such difference:

$$H = (\widehat{\beta}_{IV} - \widehat{\beta}_{OLS})'[Var(\widehat{\beta}_{IV} - \widehat{\beta}_{OLS})]^{-1}(\widehat{\beta}_{IV} - \widehat{\beta}_{OLS}) \overset{a}{\sim} \chi^2_K \tag{2.14}$$

and, since we know that both $\widehat{\beta}_{IV}$ and $\widehat{\beta}_{OLS}$ are asymptotically normal, the quadratic form will be asymptotically distributed according to a χ^2 distribution with K degrees of freedom.[35]

The computation of the Hausman test, however, poses one little problem. If you look at equation 2.14 you notice that, having produced $\widehat{\beta}_{IV}$ and $\widehat{\beta}_{OLS}$ we can directly compute the difference but we have no clue about how to calculate the variance of the difference. All we obtain from the estimation of the two estimators are the variance-covariance matrices of each of them, i.e. $AVar(\widehat{\beta}_{IV})$ and $AVar(\widehat{\beta}_{OLS})$, but we know nothing about their covariance. So how do we compute the variance of the difference?

Fortunately, a simple theorem tells us how to do this. You find the proof of the theorem in the appendix while here we only sketch the intuition which goes as follows: under H_0 we know that $\widehat{\beta}_{OLS}$ is not only consistent but also *efficient*, i.e. it is the linear estimator with the smallest possible variance. Using this fact, it is possible to show that:

$$Var(\widehat{\beta}_{IV} - \widehat{\beta}_{OLS}) = Var(\widehat{\beta}_{IV}) - Var(\widehat{\beta}_{OLS}) \tag{2.15}$$

With this formula we can directly compute the Hausman test since both $Var(\widehat{\beta}_{IV})$ and $Var(\widehat{\beta}_{OLS})$ are already known from the estimation of $\widehat{\beta}_{IV}$ and $\widehat{\beta}_{OLS}$.

Over-identification test

When the model is *over-identified*, i.e. when we have more instruments for each endogenous variable, we may not want to use all of them. In fact, there

[35] Remember that K is the number of parameters of the model, i.e. the dimensionality of the vector β.

is a trade-off (that we are not going to analyse in details) between the power of the first stage regression and the efficiency of the IV estimator: the more instruments we use the more powerful the first-stage regression will be, in the sense that it will explain a larger and larger fraction of the variance of the endogenous variable, but also the more instruments we use the larger the variance of the estimator, i.e. the less efficient it will be.

To give you an extreme example, imagine to have just one endogenous variable and two instruments. From what you have learned so far, the best thing to do in such case is simply to use both instruments in a 2SLS procedure. However, suppose that the two instruments are almost perfectly collinear (if they were perfectly collinear, you would not be able to run the first order equation) so that, conditional on one of them, there is very little additional information to be exploited from the other. In such case, you would expect the estimators produced using either one or the other of the two instruments to be asymptotically identical (and probably very similar also in small samples). However, it is easy to guess that the one produced using both instruments will be the least efficient: the use of two instruments reduces the available degrees of freedom without adding much information.

So, how do we choose which instrument(s) to keep in case of over-identification? The common practice is to keep those that appear to be most significant in the first stage regression. However, one could also construct a formal test to compare the estimators produced with two different subsets of instruments. If the test shows that the two estimates are asymptotically identical, then there is no need to use all instruments jointly.

We are not going to see over-identification tests in details.

3 Linear Systems of Equations and Generalized Least Squares

So far we have considered only the estimation of one single equation. However, sometimes the econometric model consists of several equations which may or may not have the same explanatory variables. A general specification of a *system of equations* is the following:

$$\begin{cases} y_1 &= x_1\beta_1 &+ u_1 \\ \vdots &= \vdots & \\ y_G &= x_G\beta_G &+ u_G \end{cases} \tag{3.1}$$

where each x_g is a $(1 \times K_g)$ vector and each β_g is a $(K_g \times 1)$ vector for all $g = 1, \ldots, G$. This specification allows the x_g's to be all different and also

of different sizes, i.e. the regressors in each equation may or may not be the same and even the number of regressors may differ across equations.

Before going into the technicalities, let us discuss briefly why we look at this particular issue. What is particular about systems of equations? In principle really nothing. If each equation satisfies the necessary properties then they can all be consistently estimated using usual methods (OLS, IV, 2SLS) one by one. There are however at least three cases in which we may want to estimate several equations jointly.

The first case is when there are some cross-equation restrictions on the parameters. For example, it might be that from theory or from the data generating process or from the sampling scheme we know that a certain relationship between the parameters of some equations are satisfied. For example, we might know that by definition some coefficients of the first equation must be identical in the second equation or perhaps that one coefficient in one equation should be the opposite of another in another equation. Obviously, if we were to estimate the equations separately, this type of information could not be exploited.

The second motive for estimating several equations jointly is the suspect that for one reason or another, although there might not be any cross-equation restriction, the error terms of such equations are correlated. For example, suppose you have expenditure data on various items for a sample of individuals or households, then you may want to specify one equation for each item (food, clothing, durables, et.) and the error terms of those equations are very likely to include some unobservable variables like preferences towards certain types of good or the elasticities of substitution between goods that are not observable. And, since at least some of these unobservables will be the same in all equations, the error terms will very likely be correlated with each other. In such situation, if these omitted unobservable variables do not generate endogeneity (i.e. if they are uncorrelated with the other observable regressors) we could still estimate each equation separately. However, by doing so we would not be using the information contained in the correlation across the error terms which instead can be exploited when estimating the equations jointly.

Finally, the third instance is one in which the equations are truly simultaneous, that is the dependent variable of one (or more) equation enter as explanatory variable in another (or several others). We will not cover *simultaneous equation models* in this course as they involve a series of identification problems that go beyond our interests.

So, the types of systems of equations that we look at are motivated by the desire to exploit pieces of information (cross-equation restrictions or

correlation across the error terms) that would be lost if we were to estimate each equation separately. This also means that, compared to an equation-by-equation procedure, system estimation improves efficiency but both methods lead to consistent estimates under the same assumptions.

In general, in this course we are a lot more interested in consistency than efficiency because in microeconometrics we usually have access to large datasets where the number of observations is high, which already guarantee a relatively high efficiency of the estimates.[36] Therefore we may often prefer a less efficient but simpler or more robust estimation method to a more complicated one that leads to more efficient estimates. With many observations the efficiency gain might be limited.

So why is it that we devote this whole section to the analysis of a method that can only improve the efficiency of our estimates? This is partly because system estimates are frequent and you may want to know how they are produced. Second, with these notions you should also be able to look at simultaneous equations yourselves. But probably most importantly, because the best method used to estimate equations jointly is pretty famous and can be applied in many other settings. Such method is called *Generalized Least Squares* and you should have all seen it in some previous econometrics courses.

3.1 System OLS

Let us now look in details at how to estimate the system of equations 3.1. The first (and essentially the only) thing to do is to rewrite the system in compact matrix notation by simply stacking all equations together:

$$\underbrace{Y}_{(\mathbf{G}\times\mathbf{1})} = \underbrace{X}_{(\mathbf{G}\times\mathbf{K})} \underbrace{\beta}_{(\mathbf{K}\times\mathbf{1})} + \underbrace{u}_{(\mathbf{G}\times\mathbf{1})} \tag{3.2}$$

where $K = K_1 + \cdots + K_G$ is the number of parameters to be estimated and the bold matrices have the following meaning and dimensions:[37]

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_G \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_G \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_G \end{pmatrix} \quad \mathbf{u} = \begin{pmatrix} u_1 \\ \vdots \\ u_G \end{pmatrix}$$

[36] This point is pretty obvious but if you want to see it in formulas, take a look at equation 1.34 where it is clear that as N increases the asymptotic variance of the OLS estimator decreases.

[37] Notice that K is the number of parameters to be estimated and therefore takes into account possible cross-equation restrictions on the parameters of the type discussed above.

Once the model is organised in this compact form, it can be treated as a simple (although bigger) OLS. Identification is still guaranteed by the usual orthogonality condition between each equation's regressors and error terms. Here, there are G of these conditions and they can be jointly written as:

$$E\begin{pmatrix} X_1'u_1 \\ \vdots \\ X_G'u_G \end{pmatrix} = E(\mathbf{X}'\mathbf{u}) = 0 \tag{3.3}$$

Notice that condition 3.3 simply requires each set of regressors X_g to be uncorrelated with the error term of its own equation u_g. In particular, we are not excluding the possibility of correlation across equations, i.e. X_g could be correlated with u_h where $h \neq g$.

It is easy to show that condition 3.3 guarantees identification of the entire vector of parameters β:

$$\begin{aligned} E(\mathbf{X}'\mathbf{u}) &= E[\mathbf{X}'(\mathbf{Y} - \mathbf{X}\beta)] = 0 \\ \beta &= [E(\mathbf{X}'\mathbf{X})]^{-1}E(\mathbf{X}'\mathbf{Y}) \end{aligned} \tag{3.4}$$

Equation 3.4 shows that β is identified as it can be written in terms of population moments of observable variables. Simple application of the analogy principle allows to derive the OLS estimator for the system of equations 3.2 as follows:

$$\widehat{\beta}_{SOLS} = \left(N^{-1}\sum_{i=1}^{N}\mathbf{X_i}'\mathbf{X_i}\right)^{-1}\left(N^{-1}\sum_{i=1}^{N}\mathbf{X_i}'\mathbf{Y_i}\right) \tag{3.5}$$

We indexed this estimator SOLS to indicate that it is a *System OLS* estimator.

By simple application of the analogy principle, we know that $\widehat{\beta}_{SOLS}$ is consistent and asymptotically normal:

$$\begin{aligned} \widehat{\beta}_{SOLS} &\xrightarrow{p} \beta \\ \widehat{\beta}_{SOLS} &\xrightarrow{d} N\left(\beta, AVar(\widehat{\beta}_{SOLS})\right) \end{aligned}$$

To complete our knowledge of this estimator it only remains to compute an estimator for $AVar(\widehat{\beta}_{SOLS})$. The simplest thing to do is the following:

$$\widehat{AVar(\widehat{\beta}_{SOLS})} = \left(\sum_{i=1}^{N}\mathbf{X_i}'\mathbf{X_i}\right)^{-1}\left(\sum_{i=1}^{N}\mathbf{X_i}'\widehat{\mathbf{u}}_\mathbf{i}\widehat{\mathbf{u}}_\mathbf{i}'\mathbf{X_i}\right)\left(\sum_{i=1}^{N}\mathbf{X_i}'\mathbf{X_i}\right)^{-1} \tag{3.6}$$

where - as you (hopefully) might suspect - $\widehat{\mathbf{u_i}} = \mathbf{Y_i} - \mathbf{X_i}\widehat{\beta}_{\mathbf{SOLS}}$. Again by mere application of the analogy principle, $\widehat{AVar(\widehat{\beta}_{SOLS})}$ is a consistent estimator for $AVar(\widehat{\beta}_{SOLS})$.

If the model is *homoskedastic* (all the u_g's have the same variance σ^2) then equation 3.6 simplifies to:

$$\widehat{AVar(\widehat{\beta}_{SOLS})} = \sigma^2 \left(\sum_{i=1}^{N} \mathbf{X_i'X_i} \right)^{-1} \tag{3.7}$$

Let us discuss briefly the advantages of this model over the separate estimation of each equation by OLS. We have already said that, SOLS allows to incorporate cross-equation restrictions on the parameters directly into the model. In the presence of such restrictions, then, $\widehat{\beta}_{SOLS}$ will differ from the set of β's estimated on each single equation. However, in the absence of such restrictions $\widehat{\beta}_{SOLS}$ will be *numerically identical* to the β's estimated on each single equation by OLS. In such case, the advantage of SOLS can be seen only in the variance-covariance matrix, which will be more efficient because it takes into account possible correlations across the error terms of different equations.[38] A system of linear equations with no cross-equation restrictions on the parameters but that allows potential correlation across the error terms of different equations is called *Seemingly Unrelated Regressions Model*. The name suggests the fact that in the absence of explicit cross-correlation restrictions on the parameters, the equations of the system seem to be totally unrelated although in reality they might be related through the error terms. But, since the error terms are unobservable, such relationships do not appear explicitly. Thus, the name seemingly unrelated. Finally, if there are no restrictions on the parameters and the error terms are also uncorrelated with each other, SOLS and OLS equation-by-equation lead exactly to the same results, i.e. numerically identical estimates of both the parameters and the variance-covariance matrix.

Finally, notice that if there is correlation across the error terms of different equations, although $\widehat{\beta}_{SOLS}$ is always consistent and at least as efficient as the estimators produced by OLS equation by equation, it is not *BLUE* (best linear unbiased estimator), i.e. it is not the linear estimator with the smallest possible variance.[39] In fact, we know from previous econometrics courses that the BLUE property requires homoskedasticity and absence of

[38] Obviously, such advantage is lost if one assumes homoskedasticity which excludes correlation across error terms.

[39] If you do not remember these things you should go back to your notes form previous econometrics courses and quickly take a look at them.

serial correlation. In the next section, we discuss a different estimation method that exploits more effectively the information included in the covariance matrix of the error terms and produces the best linear estimator for the parameters of a system of potentially correlated linear equations.

3.2 Generalized Least Squares (GLS)

The idea of this method is very simply. We want to find a transformation of the model such that the new transformed model satisfies all the standard OLS assumptions to obtain a consistent *and efficient* estimator. More formally, we need a transformation matrix $\mathbf{T}$ such that by pre-multiplying the model by $\mathbf{T}$:

$$\mathbf{TY} = \mathbf{TX}\beta + \mathbf{Tu} \tag{3.8}$$

we obtain a model that is homoskedastic and with no serial correlation:

$$E(\mathbf{Tuu'T}) = \mathbf{I}$$

where c is some constant. Looking at this expression it should be obvious that the most intuitive candidate for $\mathbf{T}$ is $[E(\mathbf{uu'})]^{-1/2}$. We call this matrix $\mathbf{\Omega^{-1/2}}$. In fact:

$$E(\mathbf{\Omega^{-1/2}uu'\Omega^{-1/2}}) = \mathbf{\Omega^{-1/2}}E(\mathbf{uu'})\mathbf{\Omega^{-1/2}} = \mathbf{I}$$

Thus, the transformed model is:

$$\left[\mathbf{\Omega^{-1/2}Y}\right] = \left[\mathbf{\Omega^{-1/2}X}\right]\beta + \left[\mathbf{\Omega^{-1/2}u}\right] \tag{3.9}$$

Our specific choice of $\mathbf{\Omega}$ guarantees that the model 3.9 is indeed homoskedastic and with no serial correlation. However, the transformation might have induced endogeneity of some of the regressors. Remember, in fact, that the only assumption we made for identification of the SOLS model is exogeneity equation by equation.

This means that the regressors of one equation cannot be correlated with the error term of the same equation but they could still be correlated with the error terms of other equations in the model. If that is the case (if some X's of some equations are correlated with the error terms of other equations), the model 3.9 might indeed be affected by endogeneity since now the error term of any equation is a transformed version of the original one and such transformation is a complex function of all the other error terms of the model.

To guarantee exogeneity of the model 3.9 we need to assume that each set of regressors X_g is uncorrelated with each error term $u_h \; \forall g, h$. A compact notation for this large set of conditions is:

$$E\left(\mathbf{X} \otimes \mathbf{u}\right) = 0 \tag{3.10}$$

where $\otimes$ is the Kronecker product operator.[40] Under assumption 3.10, the transformed model is easily identified as usual:

$$\beta = \left(\mathbf{X}^{*\prime}\mathbf{X}^{*}\right)^{-\mathbf{1}} \mathbf{X}^{*\prime}\mathbf{Y}^{*} = \left(\mathbf{X}'\mathbf{\Omega}^{-\mathbf{1}}\mathbf{X}\right)^{-\mathbf{1}} \left(\mathbf{X}'\mathbf{\Omega}^{-\mathbf{1}}\mathbf{Y}\right) \tag{3.11}$$

where for notational simplicity (and convention) we have redefined $\mathbf{X}^{*} = \mathbf{\Omega}^{-\mathbf{1/2}}\mathbf{X}$ and $\mathbf{Y}^{*} = \mathbf{\Omega}^{-\mathbf{1/2}}\mathbf{Y}$. Applying once more our old good analogy principle, we can then derive the *GLS* estimator of β as:

$$\widehat{\beta}_{\mathbf{GLS}} = \left[\sum_{\mathbf{i=1}}^{\mathbf{N}} \mathbf{X}_{\mathbf{i}}'\mathbf{\Omega}^{-\mathbf{1}}\mathbf{X}_{\mathbf{i}}\right]^{-\mathbf{1}} \left[\sum_{\mathbf{i=1}}^{\mathbf{N}} \mathbf{X}_{\mathbf{i}}'\mathbf{\Omega}^{-\mathbf{1}}\mathbf{Y}_{\mathbf{i}}\right] \tag{3.12}$$

which will be consistent and efficient.

But, is this estimator really computable? Or, in other words, does equation 3.11 really guarantee that β is identified? Not really. In fact, in order to compute $\widehat{\beta}_{\mathbf{GLS}}$ one needs to know $\mathbf{\Omega}$, the variance-covariance matrix of the error terms, which are unobservable! This also means that without knowledge of $\mathbf{\Omega}$, β is not identified.

To compute $\widehat{\beta}_{\mathbf{GLS}}$ we need (consistent) estimates of $\mathbf{\Omega}$ or, equivalently, of $\mathbf{u}$. We can obtain such estimates by estimating the model by SOLS in a first stage, obtain the SOLS residuals $\widehat{\mathbf{u}} = \mathbf{Y} - \mathbf{X}\widehat{\beta}_{\mathbf{SOLS}}$ and use them to compute a consistent estimator of $\mathbf{\Omega}$:

$$\widehat{\mathbf{\Omega}} = \mathbf{N}^{-\mathbf{1}} \sum_{\mathbf{i=1}}^{\mathbf{N}} \widehat{\mathbf{u}}_{\mathbf{i}}\widehat{\mathbf{u}}_{\mathbf{i}}' \tag{3.13}$$

With this $\widehat{\mathbf{\Omega}}$ we can finally compute the GLS estimator of β, that in this case is called *Feasible GLS (FGLS)*:

$$\widehat{\beta}_{\mathbf{FGLS}} = \left[\sum_{\mathbf{i=1}}^{\mathbf{N}} \mathbf{X}_{\mathbf{i}}'\widehat{\mathbf{\Omega}}^{-\mathbf{1}}\mathbf{X}_{\mathbf{i}}\right]^{-\mathbf{1}} \left[\sum_{\mathbf{i=1}}^{\mathbf{N}} \mathbf{X}_{\mathbf{i}}'\widehat{\mathbf{\Omega}}^{-\mathbf{1}}\mathbf{Y}_{\mathbf{i}}\right] \tag{3.14}$$

[40] The Kronecker product is an ugly but useful matrix operation that applies the standard matrix multiplication of each element of $\mathbf{X}$ with each element of $\mathbf{u}$.

Finally, note that $\widehat{\beta}_{FGLS}$ is generally more efficient than $\widehat{\beta}_{SOLS}$ but obviously less than $\widehat{\beta}_{GLS}$ that is however not computable. To improve efficiency of the FGLS estimator, the procedure described above can iterated. Once the FGLS estimator has been calculated, we can compute a new vector of estimated residuals, a new estimated $\mathbf{\Omega}$ and a new $\widehat{\beta}_{FGLS}$ and we can repeat this procedure until the new $\widehat{\beta}_{FGLS}$ is close enough to the previous one (convergence).

To conclude, notice that, although we have presented the GLS and FGLS estimators in the context of system estimation where their role emerges somewhat naturally, they can be applied more generally to all models that are heteroskedastic and/or serially correlated and therefore not fully efficient. And, in fact, you should have seen these methods somewhere in your previous econometrics courses in the context of the simple single equation linear model.

Appendix: Proof of $Var(\widehat{\beta}_{IV} - \widehat{\beta}_{OLS}) = Var(\widehat{\beta}_{IV}) - Var(\widehat{\beta}_{OLS})$

In principle:

$$Var(\widehat{\beta}_{IV} - \widehat{\beta}_{OLS}) = Var(\widehat{\beta}_{IV}) + Var(\widehat{\beta}_{OLS}) - 2Cov(\widehat{\beta}_{IV}, \widehat{\beta}_{OLS}) \quad (3.15)$$

and the problem with this expression is that $Cov(\widehat{\beta}_{IV}, \widehat{\beta}_{OLS})$ is generally unknown. In this case, however, we can make use of the fact that under H_0 $\widehat{\beta}_{OLS}$ is not only significant but also efficient and show that

$$Cov(\widehat{\beta}_{IV}, \widehat{\beta}_{OLS}) = Var(\widehat{\beta}_{OLS}) \quad (3.16)$$

Replacing this result into equation 3.15 immediately yields $Var(\widehat{\beta}_{IV} - \widehat{\beta}_{OLS}) = Var(\widehat{\beta}_{IV}) - Var(\widehat{\beta}_{OLS})$.

We will see first a simpler version of the proof of equation 3.16 that has only one small bug and later a more complete one.

Simpler version. Consider a third estimator of β that is a (convex) linear combination of $\widehat{\beta}_{OLS}$ and $\widehat{\beta}_{IV}$:[41]

$$\widehat{\widehat{\beta}} = \lambda\widehat{\beta}_{OLS} + (1-\lambda)\widehat{\beta}_{IV} \quad (3.17)$$

such that $\widehat{\widehat{\beta}} = \widehat{\beta}_{OLS}$ when $\lambda = 1$. Now compute the variance of $\widehat{\widehat{\beta}}$:

$$Var(\widehat{\widehat{\beta}}) = \lambda^2 Var(\widehat{\beta}_{OLS}) + (1-\lambda)^2 Var(\widehat{\beta}_{IV}) + 2\lambda(1-\lambda)Cov(\widehat{\beta}_{IV}, \widehat{\beta}_{OLS}) \quad (3.18)$$

Since $\widehat{\beta}_{OLS}$ is efficient under H_0, it must be that this variance is minimized at $\lambda = 1$. Then, let us compute the derivative of equation 3.18 and set it equal to zero with $\lambda = 1$:[42]

$$\begin{aligned} \frac{\partial Var(\widehat{\widehat{\beta}})}{\partial \lambda} &= 2\lambda Var(\widehat{\beta}_{OLS}) - 2(1-\lambda)Var(\widehat{\beta}_{IV}) + \\ &+ 2(1-2\lambda)Cov(\widehat{\beta}_{OLS}, \widehat{\beta}_{IV}) \end{aligned} \quad (3.19)$$

At $\lambda = 1$, equation 3.19 becomes:

$$2Var(\widehat{\beta}_{OLS}) - 2Cov(\widehat{\beta}_{OLS}, \widehat{\beta}_{IV}) = 0 \quad (3.20)$$

[41] The linear combination needs to be convex ($0 \leq \lambda \leq 1$) in order for this third estimator to be consistent under H_0.

[42] Here is the little bug. $\lambda = 1$ is a corner solution in this minimization problem so we cannot really simply take the derivative and set it equal to zero at this point. Later we generalize by looking at both the first and the second order conditions of the minimization problem.

which immediately shows equality 3.16.

Complete version. Start from equation 3.18 and consider the correct conditions for minimizing the variance of $\widehat{\widehat{\beta}}$ over the closed interval $\lambda \in [0, 1]$:

$$\frac{\partial Var(\widehat{\widehat{\beta}})}{\partial \lambda} \leq 0 \tag{3.21}$$

$$\frac{\partial Var(\widehat{\widehat{\beta}})}{\partial \lambda \partial \lambda} \geq 0 \tag{3.22}$$

$$\tag{3.23}$$

At $\lambda = 1$ the first order condition becomes:

$$Cov(\widehat{\beta}_{OLS}, \widehat{\beta}_{IV}) \geq Var(\widehat{\beta}_{OLS}) \tag{3.24}$$

while the second order condition is:

$$Var(\widehat{\beta}_{OLS}) + Var(\widehat{\beta}_{IV}) - 2Cov(\widehat{\beta}_{OLS}, \widehat{\beta}_{IV}) \geq 0 \tag{3.25}$$

Now, suppose by contradiction that the first order condition 3.24 is satisfied with strict inequality and $Cov(\widehat{\beta}_{OLS}, \widehat{\beta}_{IV}) > Var(\widehat{\beta}_{OLS})$. Then, the second condition 3.25 cannot be satisfied given that under H_0 $Var(\widehat{\beta}_{OLS}) < Var(\widehat{\beta}_{IV})$. Consequently, the two conditions 3.24 e 3.25 can only be jointly satisfied if $Cov(\widehat{\beta}_{OLS}, \widehat{\beta}_{IV}) = Var(\widehat{\beta}_{OLS})$, as required.

Part 2:
Maximum Likelihood and Limited Dependent Variable Models

4 The maximum likelihood estimation

Maximum likelihood is an estimation method based on a very simple and intuitive principle. It essentially consists in finding the values of the parameters of the model that maximize the probability of observing the particular sample at hand.

In order to be able to do so one needs to make enough assumptions to write a closed form expression for the probability that the actual sample is observed from the original population. Such closed-form formula is typically a function of data and parameters and one reasonable way to estimate the parameters is simply to take the ones that maximize the formula, i.e. the probability of observing the sample at hand. A couple of examples will clarify things.

Example1: The Probit model. Suppose you have some theory that states that a given *latent* variable y^* is determined by some observable variables $\mathbf{x}$ according to the following model:[1]

$$y^* = \mathbf{x}\beta + u \tag{4.1}$$

where $\mathbf{x}$ is a row vector of K explanatory variables and β the corresponding column vector of K parameters. y^*, however, is not observable. What is observed is instead y, which is defined as follows:

$$y = \begin{cases} 1 & \text{if } y^* > 0 \\ 0 & \text{if } y^* \leq 0 \end{cases} \tag{4.2}$$

You also have access to a sample of N i.i.d. observations of the vector $(y_i, \mathbf{x}_i)$ $\forall i = 1, \ldots, N$. Now, the first thing we need in order to be able to write an expression for the probability of observing the sample at hand is an assumption about the distribution of the random variables in the population. For this specific example, assume that $u \sim N(0, 1)$. Then, we can write the

[1] The term *latent variable model* refers more generally to econometric models that are developed by assuming the existence of an unobservable variable (the latent variable) used to justify the identification assumptions of the empirical model.

probability of observing a $y_i = 1$, conditional on $\mathbf{x}_i$, as follows:[2],[3]

$$\begin{aligned} Pr(y_i = 1|\mathbf{x}_i, \beta) &= Pr(\mathbf{x}_i\beta + u_i > 0|\mathbf{x}_i) = \\ &= Pr(u_i > -\mathbf{x}_i\beta|\mathbf{x}_i, \beta) = \\ &= 1 - \Phi(-\mathbf{x}_i\beta) = \Phi(\mathbf{x}_i\beta) \end{aligned} \tag{4.3}$$

Similarly, the probability of observing a $y_i = 0$ is:

$$Pr(y_i = 0|\mathbf{x}_i, \beta) = 1 - \Phi(\mathbf{x}_i\beta) \tag{4.4}$$

We can then combine equation 4.3 and 4.4 into a single expression for the probability function of y_i:[4]

$$p(y_i|\mathbf{x}_i, \beta) = [\Phi(\mathbf{x}_i\beta)]^{y_i} [1 - \Phi(\mathbf{x}_i\beta)]^{1-y_i} \tag{4.5}$$

Finally, we can use equation 4.5, which describes the probability of observing a specific observation i, to write the probability of observing the entire set of N observations, i.e. the probability of observing y_1 and y_2 and...y_N. This is particularly easy in the case of i.i.d. observations:[5]

$$L(\beta) = \prod_{i=1}^{N} p(y_i|\mathbf{x}_i, \beta) \tag{4.6}$$

Equation 4.6 is called the *Likelihood function* of the sample while equation 4.5, which is the bit of the likelihood function that refers to observation i, is called the *likelihood contribution* of i.[6] Statistically, they are both probability functions while from a mathematical standpoint they are functions of known data points (the y_i's and the $\mathbf{x}_i$'s) and unknown parameters (β). So, once we have data for y_i and $\mathbf{x}_i$, equation 4.6 is merely a function of β, with the simple twist that the statistical interpretation of such function

[2]Notice that what we consider here is not exactly the probability function of y but rather the probability function conditional on the explanatory variables $\mathbf{x}$. In fact, the textbook talks about *conditional maximum likelihood estimation*. We will discuss this important detail later on.

[3]As it is customary in many textbooks, we will write $f(\cdot)$ and $F(\cdot)$ to indicate generic density and generic cumulative density functions. We use $\varphi(\cdot)$ and $\Phi(\cdot)$ to indicate the density and the cumulative density of the *standard normal distribution*.

[4]With $p(\cdot)$ we will indicate the generic probability function of discrete random variables.

[5]We will maintain the i.i.d. assumption almost always throughout the course.

[6]The distinction between the entire likelihood function and the single likelihood contributions may appear irrelevant at this stage but it may become extremely important in models where, unlike the case considered here, the functional form of the likelihood contribution may vary across observations.

is the probability of observing the sample. The method of maximum likelihood consists in estimating β with that particular set of parameters that maximizes equation 4.6. Often, for a number of theoretical and computational reasons that we are not going to analyze, it is preferable to maximize the (natural) logarithm of the likelihood function so that a common way of specifying the maximum likelihood estimator is the following:[7]

$$\widehat{\beta}_{ML} = \underset{\beta \in \Theta}{\operatorname{argmax}}\ logL(\beta) = \underset{\beta \in \Theta}{\operatorname{argmax}} \sum_{i=1}^{N} \ln p(y_i|\mathbf{x}_i, \beta) \tag{4.7}$$

For generality, in equation 4.7 we also limit the set of available β's to an unspecified set Θ. This allows to impose various types of constraints in case we may want to do so.

Example 2: Linear regression. Also a simple linear regression can be estimated with maximum likelihood. Consider the following model:

$$y_i = \mathbf{x}_i\beta + u \tag{4.8}$$

where, on top of the usual OLS assumptions, we also assume a specific distribution for the error term. One popular choice is normality: $u \sim N(0, \sigma^2)$, so that the conditional distribution of $y|\mathbf{x}$ is $N(\mathbf{x}\beta, \sigma^2)$. We can now follow the same procedure adopted for the probit model that consists in obtaining a closed-form expression of the probability of observing the sampled data and maximizing it with respect to the parameters. In this case, however, the observed data are in continuous format (i.e. the observable random variable y is continuous while the observable random variable in the probit model is discrete as it can only take value 1 or 0) so we cannot write the probability function but rather the density function:

$$\begin{aligned} f(y_i|\mathbf{x}_i, \beta, \sigma) &= f\left(u_i = y_i - \mathbf{x}_i\beta|\mathbf{x}_i\right) = \\ &= \varphi\left(\frac{y_i - \mathbf{x}_i\beta}{\sigma}\right) \end{aligned} \tag{4.9}$$

Equation 4.9 is the likelihood contribution of a generic observation i. Under the usual assumption that the N sampled observations are i.i.d., the full log likelihood function can be written as follows:

$$L(\beta, \sigma) = \sum_{i=1}^{N} \ln \varphi\left(\frac{y_i - \mathbf{x}_i\beta}{\sigma}\right) \tag{4.10}$$

[7] Obviously, given that the logarithm is a monotonic transformation, the β that maximizes the log-likelihood function also maximizes the likelihood.

Notice that in this case, since we have assumed a general variance σ^2 for the distribution of the error term, there is one additional parameter to be estimated, i.e. the β's and σ. The specific functional form of this model allows to derive closed-form expressions of the maximum likelihood estimators of β and σ.[8] Notice, in fact, that:

$$\begin{aligned}\ln\varphi\left(\frac{y_i-\mathbf{x}_i\beta}{\sigma}\right) &= \ln\left[\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{y_i-\mathbf{x}_i\beta}{\sigma}\right)^2\right)\right]= \\ &= -\ln\sigma-\frac{1}{2}\ln 2\pi-\frac{1}{2\sigma^2}\left(y_i-\mathbf{x}_i\beta\right)^2\end{aligned}$$

so that:

$$L(\beta,\sigma)=-N\ln\sigma-\frac{N}{2}\ln 2\pi-\frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(y_i-\mathbf{x}_i\beta\right)^2 \tag{4.11}$$

To obtain the maximum likelihood estimators of β and σ, we simply maximize equation 4.11 with respect to each parameter:

$$\frac{1}{\sigma^2}\sum_{i=1}^{N}\mathbf{x}_i'\left(y_i-\mathbf{x}_i\beta\right) = 0 \tag{4.12}$$

$$-\frac{N}{\sigma}+\frac{1}{\sigma^3}\sum_{i=1}^{N}\left(y_i-\mathbf{x}_i\beta\right)^2 = 0 \tag{4.13}$$

Combining these two first order conditions one gets:

$$\widehat{\beta}_{ML} = \left[\sum_{i=1}^{N}\mathbf{x}_i'\mathbf{x}_i\right]^{-1}\left[\sum_{i=1}^{N}\mathbf{x}_i'y_i\right] \tag{4.14}$$

$$\widehat{\sigma}^2_{ML} = N^{-1}\sum_{i=1}^{N}\left(y_i-\mathbf{x}_i\beta\right)^2 \tag{4.15}$$

which are exactly the same estimators obtained by OLS![9] Under normality, the OLS estimator is also a maximum likelihood estimator. If we had assumed a different distribution of the error term, we would have obtained different estimators. Also notice that it is not always the case, like in this

[8] Which was not the case with the probit model.

[9] If the derivation of the first order conditions 4.12 and 4.13 is unclear to you, you can find a rather complete discussion of optimization in the context of matrix algebra in chapter 2.9 in Green, W.H. 2000. *"Econometric Analysis"*. Fourth edition. Prentice Hall. (or some more recent edition).

example, that the maximum likelihood estimator of a model can be written in a closed-form expression nor that it is a linear function of the error term (like the OLS estimator, as we showed in the first part of the course). In fact, in the previous example, the maximum likelihood estimator of the probit model does not have a closed form solution nor it is a linear function of the error term.

4.1 (slightly) More formal maximum likelihood theory

The two examples seen so far should have clarified the basic principle of maximum likelihood. Essentially, this method assumes a specific distribution on the data and derives the parameters of such distribution by picking those that maximize the probability (computed on the basis of the assumed distribution) of observing the specific data at hand.

Before we now move to some more technical details, let me clarify what I consider the most important feature of this estimation method. As we will see in a minute, maximum likelihood estimators have some very desirable properties: they are consistent and efficient. Somehow the best estimators one could ask for. These nice features, however, come at the cost of very strong assumptions necessary to obtain them. Namely, the assumption about the distribution of the data.

Compare what you know about OLS with what you know about maximum likelihood. With OLS we only need an assumption about the conditional mean of the distribution of y (or the error term u) conditional on $\mathbf{x}$ ($E(u|\mathbf{x}) = 0$) to obtain consistency. This relatively mild assumption does not guarantee efficiency.

With maximum likelihood we make a much stronger assumption. We make an assumption about the entire distribution of y (conditional on $\mathbf{x}$) rather than only on one of the moments of such distribution. In exchange for this stronger assumption, we get the best possible estimator one could think of (consistent and efficient).

The problem is that, if the distributional assumption is incorrect (something we will never be able to test rigorously given that it is an identification assumption), then the ML estimator looses not only efficiency but also consistency! This is the main trade-off of maximum likelihood that you should always consider in your analysis.

Let us now formalize and generalize a bit more the procedure described in the examples above. We will not go into the derivation of the asymptotic properties of maximum likelihood estimators but we will simply state them.

In general, we have access to a i.i.d. sample of N observations $(y_i, \mathbf{x}_i)$ for $i = 1, ..., N$. The first step to apply maximum likelihood consists in assuming a particular distribution of the data that depends on the parameters of interest:

ML Assumption 1 $y|\mathbf{x} \sim$ *i.i.d.* $F(\cdot|\theta)$

Two things are worth noticing about assumption 1. First of all, F is the assumed distribution of y conditional on $\mathbf{x}$ whereas, if we were to follow the line of reasoning of our previous presentation, we should have assumed a distribution for the joint data observation $(y, \mathbf{x})$. In this course we will almost always work with *conditional likelihoods* rather than *full likelihoods*. Later in this section we will clarify the meaning of this choice.

Second, the expression $F(\cdot|\theta)$ specifies that the distribution depends on a set of unknown parameters θ that are the object of the estimation. We indicate this feature of the distribution by conditioning on the fixed parameters θ. Note that this is simply a notational shortcut to emphasize the fact that $F(\cdot|\theta)$ is a function of the parameters. From a statistical standpoint conditioning on a set of constants is totally irrelevant.

With assumption 1 we can write the log-likelihood contribution of a generic observation i:

$$\ell_i(\theta) = lnf(y_i|\mathbf{x}_i, \theta) \tag{4.16}$$

where f is the density function of F. The log-likelihood function is constructed by simply summing the N individual log-likelihood contributions:

$$L(\theta) = \sum_{i=1}^{N} \ell_i(\theta) = \sum_{i=1}^{N} lnf(y_i|\mathbf{x}_i, \theta) \tag{4.17}$$

Then, the maximum likelihood estimator of θ is defined as follows:

$$\widehat{\theta}_{ML} = \arg\max_{\theta \in \Theta} L(\theta) \tag{4.18}$$

Under assumption 1 and some additional technical assumptions, $\widehat{\theta}_{ML}$ is *consistent, asymptotically normal and asymptotically efficient*, among all asymptotically normal consistent estimators. The additional assumptions needed to obtain such results are the following:

1. Θ is *compact* set;[10]

[10] A compact set is *closed* and *bounded*.

2. f is continuous and twice differentiable over Θ;

3. the density function $f(y|\theta)$ is such that, if $\theta_1, \theta_2 \in \Theta$, $f(y|\theta_1) = f(y|\theta_2)$ if and only if $\theta_1 = \theta_2$.

While assumptions 1 and 2 are mere technical details, the third assumption is an important identification condition (together with the key assumption 1). Loosely speaking, it guarantees that the likelihood function is never *flat* with respect to θ. If the density could assume the same value at the same point y for different values of the parameters, then there could be more that one set of parameter values that maximize the likelihood function. In other words, this third assumption assures that the solution of the maximization problem in 4.18 is unique.

And what is the asymptotic distribution of $\widehat{\theta}_{ML}$? We already know that it is normal but the mean and the variance-covariance matrix are defined as follows:

$$\widehat{\theta}_{ML} \overset{a}{\sim} N\left[\theta, -E\left(\frac{\partial^2 L(\theta)}{\partial\theta\partial\theta'}\right)^{-1}\right] \tag{4.19}$$

The formula of the asymptotic variance-covariance matrix of $\widehat{\theta}_{ML}$ should remind you something you know from statistics. It is the negative of the inverse of the Hessian matrix. In statistics this is called the *Cramer-Rao lower bound* and represents the smallest possible variance an estimator can reach. This obviously implies that $\widehat{\theta}_{ML}$ is efficient.

Conditional vs. Full Maximum Likelihood

In most of our previous discussion we have specified the model in terms of the distribution of y conditional on the explanatory variables $\mathbf{x}$ whereas the intuitive (and also the less intuitive) arguments were based on the actual joint distribution of y and $\mathbf{x}$. How do these two things square together?

Notice that the joint density of y and $\mathbf{x}$ can be written as:

$$f(y, \mathbf{x}|\theta) = f(y|\mathbf{x}, \theta) f(\mathbf{x}|\theta)$$

If the distribution of $\mathbf{x}$ does not depend on the same set of parameters of the conditional distribution of $y|\mathbf{x}$ (i.e. $f(\mathbf{x}|\theta) = f(\mathbf{x})$), then the value of θ that maximizes $f(y|\mathbf{x}, \theta)$ also maximizes $f(y, \mathbf{x}|\theta)$. This is the usual assumption underlying the use of *conditional maximum likelihood* rather than *full maximum likelihood*.

Further, notice that if $\mathbf{x}$ is endogenous in the model under analysis (i.e. if it is correlated with the residual of $E(y|\mathbf{x})$), then the above assumption

does not hold and the distribution of $\mathbf{x}$ necessarily depends on the same parameters of the distribution of $y|\mathbf{x}$ (other than in very special cases).

In models where $\mathbf{x}$ is endogenous, conditional maximum likelihood does not lead to consistent estimates and a full maximum likelihood approach is needed, which requires the specification of the joint distribution of y and $\mathbf{x}$.

5 Binary response models: Probit and Logit

Binary response models are models where the dependent variable y only takes two values (1 and 0 or whatever other two numbers or categories). The probit model that we saw as an example of maximum likelihood in the previous section is probably the most popular binary response model.

Let us now go back to that example and discuss the probit model a bit more in details. The model is based on the assumption that there exists an underlying unobservable (latent) variable y^*:

$$y^* = \mathbf{x}\beta + u \tag{5.1}$$

such that the observable y is:

$$y = \begin{cases} 1 & \text{if} \quad y^* > 0 \\ 0 & \text{if} \quad y^* \leq 0 \end{cases} \tag{5.2}$$

As already mentioned, to estimate this model with maximum likelihood the crucial assumption that one needs to impose is some distribution on the error term. The probit model is characterized by normality of the error term (if we were to make a different assumption the model would not be a probit!).

To make our discussion a bit more general than before, let us assume that the error is distributed normally, with mean zero and generic variance σ^2:

$$u \sim N(0, \sigma^2) \tag{5.3}$$

Then, as we did previously, we can write the probability of observing

$y = 1$ and $y = 0$ as follows:[11]

$$\begin{aligned} Pr(y = 1|\mathbf{x}) &= Pr(\mathbf{x}\beta + u > 0|\mathbf{x}) = \\ &= Pr(u > -\mathbf{x}\beta|\mathbf{x}) = Pr(\frac{u}{\sigma} > -\frac{\mathbf{x}\beta}{\sigma}|\mathbf{x}) = \\ &= \Phi\left(\frac{\mathbf{x}\beta}{\sigma}\right) \qquad (5.4) \\ Pr(y = 0|\mathbf{x}) &= 1 - \Phi\left(\frac{\mathbf{x}\beta}{\sigma}\right) \qquad (5.5) \end{aligned}$$

Using equations 5.4 and 5.5 we can write the log-likelihood contribution of a generic individual i as follows:

$$\ell_i(\beta, \sigma) = y_i ln\Phi\left(\frac{\mathbf{x}_i\beta}{\sigma}\right) + (1 - y_i)ln\left[1 - \Phi\left(\frac{\mathbf{x}_i\beta}{\sigma}\right)\right] \qquad (5.6)$$

and the complete log-likelihood function as:

$$L(\beta, \sigma) = \sum_{i=1}^{N}\left\{y_i ln\Phi\left(\frac{\mathbf{x}_i\beta}{\sigma}\right) + (1 - y_i)ln\left[1 - \Phi\left(\frac{\mathbf{x}_i\beta}{\sigma}\right)\right]\right\} \qquad (5.7)$$

Maximization of equation 5.7 with respect to β and σ leads to the maximum likelihood estimators for such parameters. However, is it really possible to find distinct estimators for β and σ? Look at the functional form of the objective function 5.7 and notice that β and σ always appear as a ratio of one another. This implies that the two parameters cannot be separately identified. All we can identify is their ratio.

In other words, suppose that $\widehat{\beta}$ and $\widehat{\sigma}$ are the true values of β and σ that maximize the likelihood function 5.7. Then, also $\widetilde{\beta} = c\widehat{\beta}$ and $\widetilde{\sigma} = c\widehat{\sigma}$ maximize it, for any constant c. You can also see this result by taking the first order conditions for maximization of equation 5.7 with respect to both parameters and verify that they are identical. In other words, the optimization procedure leads to just one equation for two parameters.

Hence, in the probit model only the ratio $\frac{\beta}{\sigma}$ is identifiable. This means that whatever assumption we make on the variance of the distribution of the error term will be irrelevant because that parameter will not be identifiable. As a convention, in the probit model we usually assume the problem away by setting $\sigma = 1$, as we did in the previous section when we presented the probit model as an example of maximum likelihood.

[11] In most of our discussion below, we will omit for simplicity the conditioning on the coefficients on the model β.

Under such normalization, the maximum likelihood estimator of β is defined as follows:

$$\widehat{\beta}_{ML} = \arg\max_{\beta} L(\beta, 1) \tag{5.8}$$

As mentioned earlier on, the probit model is only one binary response model. Another popular one is the *logit* model, which follows exactly the same line of thought of the probit by simply assumes a slightly different distribution for the error term, the logistic distribution.[12]

5.1 Additional issues in binary response models

Interpretation of the parameters

In a linear model the parameters are very easily interpretable in terms of partial derivatives. For example, in the very simple model $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_K x_K + u$ the parameters are simply the partial derivatives of the dependent variable y with respect to each single regressor x_k:

$$\frac{\partial y}{\partial x_k} = \beta_k$$

where x_k is a generic regressor with $k = 1, ..., K$.

In non-linear models like the probit (or the logit) things are more complicated. In such models the parameters can be directly interpreted as marginal effects only with respect to the unobservable latent variable y^*. However, such variable is often just a statistical artifact that we use to justify the model and what we are really interested in is the effect of changes in the regressors on the probability that the observable dependent variable y takes value 1 (or 0). For a generic binary response model, such probability can we written as follows:

$$Pr(y = 1|\mathbf{x}) = F(\mathbf{x}\beta) = F(\beta_0 + \beta_1 x_1 + \cdots + \beta_K x_K) \tag{5.10}$$

and what we would like to know is the partial derivative of this expression with respect to each of the regressors.

Obviously, such derivatives are not simply equal to the parameters of the model:

$$\frac{\partial Pr(y = 1|\mathbf{x})}{\partial x_k} = f(\mathbf{x}\beta)\beta_k \tag{5.11}$$

[12] Just to refresh your statistics, the cdf of the logistic distribution is:

$$\Lambda(z) = \frac{e^z}{1 + e^z} \tag{5.9}$$

Equation 5.11 shows that in binary choice models (but this will be a general feature of almost all maximum likelihood models that we will see) the marginal effects are not constant but vary with the characteristics of the single observation, with the $\mathbf{x}$'s.

When presenting results from the estimation of these models it is usually good practice to present the marginal effects computed at some meaningful point of the distribution of the regressors. For example, one solution that is often taken by many researchers consists in computing and presenting the marginal effects at the mean of the $\mathbf{x}$'s:

$$\frac{\partial Pr(y=1|\overline{\mathbf{x}})}{\partial x_k} = f(\overline{\mathbf{x}}\beta)\beta_k$$

where $\overline{\mathbf{x}}$ is the vector of the means of the regressors.

Notice, however, that the coefficients of the model already tell us something about the marginal effects. In particular, they tell two things: first, the sign of the marginal effect is always the same as the sign of the coefficient ($f(\mathbf{x}\beta)$ is a density function and it is therefore always positive); second, they tell the relative importance of the marginal effects of two different regressors:

$$\frac{\partial Pr(y=1|\mathbf{x})/\partial x_k}{\partial Pr(y=1|\mathbf{x})/\partial x_j} = \frac{\beta_k}{\beta_j}$$

For example, if $\beta_k > \beta_j$ than a marginal change in x_k has a larger effect on $Pr(y=1|\mathbf{x})$ than a marginal change in x_j.

This discussion extends rather easily to computing the effects of changes in some non-continuous regressor, like a dummy variable. In those cases, we cannot formally take derivatives but need to compute the effects (which will not be *marginal* in the most technical sense) by taking the difference between the probability that $y = 1$ conditional on the dummy (or other categorical variable) taking one or another value.

Goodness of fit: the *Pseudo-*R^2

In linear model the R^2 is a very popular measure of the goodness of fit of the model. It essentially measures the fraction of variance of the dependent variable that is captured by the model and it is constructed as follows:

$$R^2 = \frac{\sum_{i=1}^{N}\left(\widehat{y}_i - \overline{y}\right)^2}{\sum_{i=1}^{N}\left(y_i - \overline{y}\right)^2} \tag{5.12}$$

where $\widehat{y}_i = \mathbf{x}_i\widehat{\beta}$ is the prediction of y as obtained from the estimated model and $\overline{y} = N^{-1}\sum_{i=1}^{N} y_i$ is the sample mean of y.

Computing a similar measure for a binary choice model is meaningless. In fact, the prediction that can be obtained from the estimated model is either a prediction about the unobservable latent variable $\widehat{y}_i^* = \mathbf{x}_i\widehat{\beta}$ or the predicted probability that the observed y is equal to some value (say 1), $\widehat{Pr(y_i = 1|\mathbf{x}_i)} = F(\mathbf{x}_i\widehat{\beta})$[13]. The problem is that none of these possible predictions can be meaningfully compared with an empirical counterpart.

The latent variable y_i^* is not observable so we don't know what to compare the predicted $\widehat{y}_i^*$ with. The same for the predicted probability $\widehat{Pr(y_i = 1|\mathbf{x}_i)}$ whose empirical counterpart is also not observable. What we observe is the outcome y_i and not its probability function.

Hence, we need to come up with some other idea to measure how well the model fits the data. In the literature several of these measures have been proposed, here we only mention two and focus our attention on one of them which is by far the most commonly used.

The first measure is the *Percentage of Correctly Predicted Outcomes* and is constructed by arbitrarily setting a threshold to the predicted probability $\widehat{Pr(y_i = 1|\mathbf{x}_i)}$, usually 0.5, and generating a predicted outcome $\widehat{y}_i$ as follows:

$$\widehat{y}_i = \begin{cases} 1 & \text{if} \quad \widehat{Pr(y_i = 1|\mathbf{x}_i)} \geq 0.5 \\ 0 & \text{if} \quad \widehat{Pr(y_i = 1|\mathbf{x}_i)} < 0.5 \end{cases} \tag{5.13}$$

The percentage of correctly predicted outcomes is then computed as the percentage of observations that have $y_i = \widehat{y}_i$. This is a very simple and easily interpretable measure but, as you can imagine, it is also heavily dependent on the arbitrarily chosen threshold to assign the predicted outcome to a 1 or a 0. For this reason it is much less popular that the second measure, the so-called *Pseudo-*R^2.

The Pseudo-R^2 is based on the idea of comparing the maximized likelihood of the model with the maximized likelihood of a minimal model where the only explanatory variable is a constant. Remember that the likelihood is essentially the probability that, given the assumption on the distribution and the functional form of the random variables in the model, the specific dataset at hand is in fact observed.

In this sense, then, a good model, a model that fits well the data, is one that delivers a high value of the likelihood. Now, suppose that our specific model is really bad, meaning that none of the regressors explains anything

[13]@@ Matter of editing: in what follows you always write $\widehat{Pr(y_i = 1|\mathbf{x}_i)}$, meaning *"predicted probability"*. The notation however is not clear: the *widehat* is not sufficiently wide. I would type it: $\widehat{Pr}(y_i = 1|\mathbf{x}_i)$.

about the data. If this is the case then the likelihood of our model should not be that different from the likelihood of a minimal model where we take out all the regressors and leave only the constant.

To formalize this intuition and construct the Pseudo-R^2 call L^* the maximized value of the log-likelihood of our model and L_0 the maximized log-likelihood of a model where the only regressor is a constant. The Pseudo-R^2 is then defined as follows:

$$\text{Pseudo-}R^2 = 1 - \frac{L^*}{L_0} \tag{5.14}$$

This measure has some nice properties that make it easy to interpret it.

- $0 \leq \text{Pseudo-}R^2 < 1$. For the Pseudo-$R^2$ to be always positive we need $\frac{L^*}{L_0} \leq 1$ which is guaranteed if $|L^*| \leq |L_0|$, which is always true. Remember, in fact, that both L^* and L_0 are logarithms of probabilities and are therefore always negative by definition. Moreover, L^* can only be larger or at a minimum equal to L_0 given that it includes a constant plus other regressors. So, the complete model explains at least as much as the model with just the constant. And, since we are talking about negative numbers, $L^* \geq L_0$ implies $|L^*| \leq |L_0|$, which guarantees that Pseudo-$R^2 \geq 0$. The Pseudo-R^2 is also always smaller than 1. For that to be the case we simply need $\frac{L^*}{L_0} > 0$ which is obviously always true given than both the numerator and the denominator of this ratio are negative.

- $\frac{\partial \text{Pseudo-}R^2}{\partial L^*} > 0$. This second important property guarantees that the Pseudo-R^2 increases with the (maximized) likelihood of the model, which makes it a meaningful measure of its goodness of fit. The property is easily verified by noting that $\frac{\partial \text{Pseudo}-R^2}{\partial L^*} = -\frac{1}{L_0}$, which is a positive number given than L_0 is negative.

The linear probability model

An alternative simpler specification of binary choice models that is becoming more and more popular is the so-called *linear probability model*. This is essentially a simple linear regression where we use the dichotomous observable outcome y as a dependent variable, as if we forgot that such variable is discrete and distributed on just two possible values.

To compare this model with the probit (or logit), it is useful to describe it as a linear specification of the probability of observing $y = 1$:

$$Pr(y = 1|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_K x_K = \mathbf{x}\beta \tag{5.15}$$

This specification is obviously wrong. In fact, the left-hand-side of this equation is a probability and it is therefore constrained between 0 and 1. On the right-hand-side, instead, there is nothing that imposes this restriction. This implies, for example, that the prediction of the model $\mathbf{x}_i\widehat{\beta}$ can well be above 1 or below zero, thus making its interpretation complex. Moreover, the model is intrinsically heteroskedastic. To see this, notice that essentially the model that is estimated can be written as follows:

$$y = \begin{cases} 1 \\ 0 \end{cases} = \mathbf{x}\beta + u \tag{5.16}$$

which implies the same specification of equation 5.15:

$$\begin{aligned} E(y|\mathbf{x}) &= 1 \times Pr(y = 1|\mathbf{x}) + 0 \times Pr(y = 0|\mathbf{x}) \\ &= Pr(y = 1|\mathbf{x}) = \mathbf{x}\beta \end{aligned}$$

The error term of this model is itself a dichotomous variable:

$$u = \begin{cases} 1 - \mathbf{x}\beta & \text{if} \quad y = 1 \quad \text{with probability } Pr(y = 1|\mathbf{x}) = \mathbf{x}\beta \\ -\mathbf{x}\beta & \text{if} \quad y = 0 \quad \text{with probability } Pr(y = 0|\mathbf{x}) = 1 - \mathbf{x}\beta \end{cases}$$

with mean zero:

$$\begin{aligned} E(u|\mathbf{x}) &= (1 - \mathbf{x}\beta) \times Pr(y = 1|\mathbf{x}) - \mathbf{x}\beta \times Pr(y = 0|\mathbf{x}) \\ &= (1 - \mathbf{x})\mathbf{x}\beta - \mathbf{x}\beta(1 - \mathbf{x}) = 0 \end{aligned}$$

and variance equal to:[14]

$$\begin{aligned} Var(u|\mathbf{x}) &= E(u^2)|\mathbf{x}) = (1 - \mathbf{x}\beta)^2\mathbf{x}\beta + (\mathbf{x}\beta)^2(1 - \mathbf{x}\beta) \\ &= (1 - \mathbf{x}\beta)\mathbf{x}\beta \end{aligned}$$

which is not constant and varies with the observation's characteristics thus showing that the model is necessarily heteroskedastic.

Heteroskedasticity, however, is not a big problem. We know how to deal with it using GLS instead of OLS to estimate the parameters. Moreover, heteroskedasticity per se only affects the efficiency of the estimates and not their consistency.

In fact, despite the problems listed above, the estimates of a linear probability model are consistent under the usual assumption that the error term

[14] Alternatively, recall that the variance of a dichotomous variable that takes one value with probability p and the other with probability $1-p$ is simply $p(1-p)$. Using this result it is immediate to show that $Var(u|\mathbf{x}) = (1 - \mathbf{x}\beta)\mathbf{x}\beta$.

and the regressors are uncorrelated with each other, $E(\mathbf{x}_i' u) = 0$. And this is the first and most important advantage of this model: it is consistent under milder assumptions than the probit (or logit) that instead requires rather arbitrary assumptions about the entire distribution of the error term.

A second feature of the linear probability model that is often regarded as an advantage (although I don't really see it as such a big advantage!) is the easiness of interpretation of its results. In fact, contrary to the probit/logit models the coefficients of the linear probability model directly represent the marginal effects of the regressors on the dependent variable.

Finally, the last important advantage of the linear probability model, which is probably the most important factor that has influenced its recent popularity, is the fact that it lends itself a lot more easily than the probit/logit to the application of panel data methods. To fully understand this point you will have to wait a little bit until we discuss panel data econometrics.

6 Multinomial response models

In this section we study a particular class of models designed for the analysis of multinomial response data. These are data that can take a finite set of values and are a natural extension of the simpler binary choice data. Data of this type are typical, for example, in surveys that offer respondents a finite set of possible answers.

6.1 Multinomial probit model

Let us start with a simple example where the data we want to analyze include just three possible categories that are mutually exclusive, i.e. each observation is coded into just one category and all observations are coded. For example, we could have a sample of workers who are classified in three different occupational categories: blue collars, white collars, managers.

To describe the decision process that leads observations into the three categories it is again useful to consider three latent variables:

$$\begin{aligned} y_0^* &= x_0\beta_0 + u_0 \\ y_1^* &= x_1\beta_1 + u_1 \\ y_2^* &= x_2\beta_2 + u_2 \end{aligned}$$

where x_0, x_1 and x_2 are vectors of observable explanatory variables. A simple economic interpretation of such latent variables is, for example, the utility level that each agent i would obtain in each of the three choices. Consistently with this interpretation and as it is usually the case for such latent variables, y_0^*,y_1^* and y_2^* are not observable. What we observe, instead, is simply a set of zero-one indicators for each category:

$$\begin{aligned} y_0 &= \begin{cases} 1 & \text{if alternative 0 is chosen} \\ 0 & \text{otherwise} \end{cases} \\ y_1 &= \begin{cases} 1 & \text{if alternative 1 is chosen} \\ 0 & \text{otherwise} \end{cases} \\ y_2 &= \begin{cases} 1 & \text{if alternative 2 is chosen} \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

Accordingly with the utility interpretation of the latent variables, it is customary to define the process that leads observations into the alternative categories as follows:

$$\begin{aligned} y_0 &= \begin{cases} 1 & \text{if } y_0^* > y_1^* \text{ and} y_0^* > y_2^* \\ 0 & \text{otherwise} \end{cases} \\ y_1 &= \begin{cases} 1 & \text{if } y_1^* > y_0^* \text{ and} y_1^* > y_2^* \\ 0 & \text{otherwise} \end{cases} \\ y_2 &= \begin{cases} 1 & \text{if } y_2^* > y_0^* \text{ and} y_2^* > y_1^* \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

In other words, we assume that the observations (agents) choose the alternative that delivers the highest utility. This interpretation and the consequent definition of the decision process above is extremely common in economics applications however, depending on the type of data to be analyzed and the assumptions made on the underlying data generating process, other settings may be possible. What is general in the multinomial models approach is the assumption about a particular process that generates the observed data pattern whatever that process might be.

Going back to our 3-categories example, define the probabilities of observing a generic observation i into each of the three categories as follows:

$$P_0 = Pr(y_0 = 1, y_1 = 0, y_2 = 0|x) \tag{6.1}$$
$$P_1 = Pr(y_0 = 0, y_1 = 1, y_2 = 0|x) \tag{6.2}$$
$$P_2 = Pr(y_0 = 0, y_1 = 0, y_2 = 1|x) \tag{6.3}$$

Then:

$$\begin{aligned} P_0 &= 1 - P_1 - P_2 && (6.4)\\ P_1 &= Pr(x_1\beta_1 + u_1 > x_0\beta_0 + u_0, x_1\beta_1 + u_1 > x_2\beta_2 + u_2|x) \\ &= Pr(u_0 - u_1 < x_1\beta 1 - x_0\beta_0, u_2 - u_1 < x_1\beta 1 - x_2\beta_2|x) && (6.5)\\ P_2 &= Pr(x_2\beta_2 + u_2 > x_0\beta_0 + u_0, x_2\beta_2 + u_2 > x_1\beta_1 + u_1|x) \\ &= Pr(u_0 - u_2 < x_2\beta 2 - x_0\beta_0, u_1 - u_2 < x_2\beta 2 - x_1\beta_1|x) && (6.6) \end{aligned}$$

where $x = (x_0\ x_1\ x_2)$. Equation 6.4 exploits the assumption that the three choices are mutually exclusive.

As you might have guessed by now, we can now make some distributional assumptions about the random errors u_0, u_1 and u_2 and derive explicit expressions for the probabilities in 6.4, 6.5 and 6.6. We will then be able to write the likelihood function of the model and derive the maximum likelihood estimators for β_0, β_1 and β_2.

So, let us start with the distributional assumptions and impose joint normality of all three error terms:

$$\begin{pmatrix} u_0 \\ u_1 \\ u_2 \end{pmatrix} \sim N \left[0, \begin{pmatrix} \sigma_0^2 & \sigma_{01} & \sigma_{02} \\ & \sigma_1^2 & \sigma_{12} \\ & & \sigma_2^2 \end{pmatrix} \right] \tag{6.7}$$

To simplify the notation, define $\eta_1 = u_0 - u_1$ and $\eta_2 = u_2 - u_1$ and $V_2(\theta_1) = x_1\beta_1 - x_0\beta_0$ and $V_1(\theta_1) = x_1\beta_1 - x_2\beta_2$, where $\theta_1 = (\beta_0'\ \beta_1')'$ and $\theta_2 = (\beta_1'\ \beta_2')'$. Using these new terms and the distributional assumptions to rewrite the probability in 6.5 as follows:

$$\begin{aligned} P_1 &= Pr(\eta_1 < V_1(\theta_1), \eta_2 < V_2(\theta_2)|x) = \\ &= \int_{-\infty}^{V_1(\theta_1)} \int_{-\infty}^{V_2(\theta_2)} f(\eta_1, \eta_2) d\eta_1 d\eta_2 && (6.8) \end{aligned}$$

where $f(\eta_1, \eta_2)$ is the joint probability density function of the random variables η_1 and η_2 that, given the distributional assumption 6.7, are jointly normally distributed with mean zero, variances equal to:

$$\begin{aligned} Var(\eta_1) &= \sigma_0^2 + \sigma_1^2 - 2\sigma_{01} = \sigma_{\eta_1}^2 \\ Var(\eta_2) &= \sigma_1^2 + \sigma_2^2 - 2\sigma_{12} = \sigma_{\eta_2}^2 \end{aligned}$$

and covariance equal to $Cov(\eta_1, \eta_2) = \sigma_{02} - \sigma_{01} - \sigma_{12} + \sigma_1^2 = \sigma_{\eta_1\eta_2}$.[15]

[15] Obviously, this is a rather general specification. To simplify things and reduce the

Similarly, by redefining $\eta_3 = u_0 - u_2$ and $\eta_4 = u_1 - u_2$ and $V_3(\theta_3) = x_2\beta_2 - x_0\beta_0$ and $V_4(\theta_4) = x_2\beta_2 - x_1\beta_1$, where $\theta_3 = (\beta_0' \ \beta_2')'$ and $\theta_4 = (\beta_1' \ \beta_2')'$, the probability in 6.6 can be rewritten as:

$$\begin{aligned} P_2 &= Pr(\eta_3 < V_3(\theta_3), \eta_4 < V_4(\theta_4)|x) = \\ &= \int_{-\infty}^{V_3(\theta_3)} \int_{-\infty}^{V_4(\theta_4)} f(\eta_3, \eta_4) d\eta_3 d\eta_4 \end{aligned} \tag{6.9}$$

where the joint distribution of η_3 and η_4 can be easily calculated as above for η_1 and η_2.

Finally, the probability 6.1 can be simply computed as the residual event $P_{i0} = 1 - P_{i1} - P_{i2}$ thanks to the assumption that the alternative choices are mutually exclusive. With these results it is now easy to construct the log-likelihood contribution of a single observation as well as the complete log-likelihood function:

$$\ell_i(\beta, \Omega) = 1(y_{i0} = 1) \ln P_{i0} + 1(y_{i1} = 1) \ln P_{i1} + 1(y_{i2} = 1) \ln P_{i2} \tag{6.10}$$

$$L(\beta, \Omega) = \sum_{i=1}^{N} \ell_i(\beta, \Omega) \tag{6.11}$$

where $\beta = (\beta_0' \ \beta_1' \ \beta_2')'$ and $\Omega = E(uu')$ with $u = (u_0 \ u_1 \ u_2)'$. As usual, then, the maximum likelihood estimator of β and Ω is obtained by maximization of the likelihood function 6.11.

The model we just described is called *Multinomial probit* and it is characterized by the normality assumption on the error structure. The model is very general and flexible: the alternatives can be more than just 3, the errors could be correlated or uncorrelated with whatever correlation structure and the alternatives could even not be mutually exclusive. However, despite these many nice features, the multinomial probit is not particularly popular due to its computational complexity.

Notice, in fact, that even in this simple 3-alternative model, the likelihood function involves double integrals. In general, with a generic number J of mutually exclusive choices the likelihood function involves integral of the $J-1$th order.[16] This makes the optimization procedure computationally very intense. Already with more than 5 alternatives a standard modern

number of parameters to be estimated one may assume homoskedasticity (i.e. $\sigma_0 = \sigma_1 = \sigma_2$) and/or independence between the three original error terms u_0, u_1 and u_2. Notice, however, that even under such assumption the composite error terms η_1 and η_2 will be correlated so that the expression of P_{i1} (but also P_{i2}) will involve a double integral.

[16] If the choices are not mutually exclusive then the order of integration is the same as the number of alternatives.

desktop computer can take several days to complete the optimization procedure!

Moreover, the computation of the marginal effects is also extremely complicated. In particular, the sign of the marginal effects is not necessarily the same as the sign of the coefficients (as it was instead the case with the probit model). For example, from our previous discussion:

$$\frac{\partial P_1}{\partial x_k} = \frac{\partial \left[\int_{-\infty}^{V_{i1}(\theta_1)} \int_{-\infty}^{V_{i2}(\theta_2)} f(\eta_1, \eta_2) d\eta_1 d\eta_2 \right]}{\partial x_k}$$

whose sign is clearly undetermined.

How can we overcome the computational complexity of the multinomial probit model? In the next sections we will present two classes of models that address the issue in different ways. The first possibility consists in changing the distributional assumptions hoping to find a particular distribution for the error terms that leads to a likelihood function with a more tractable functional form. This is the approach taken by the *multinomial logit model* in its various specifications. Alternatively, one could make different assumptions about the underlying decision process. Obviously, the extent to which this is possible depends very much on the type of data and credibility of the assumptions imposed on them. Models that follow this approach are, among other, the various specifications of *ordered response models* and *nested models*.

In the following sections we present some of these models.

6.2 Multinomial logit models

The structure of the *multinomial logit model* is similar to the multinomial probit with the important difference that the error terms are assumed to have a specific distribution that allows to derive a very simple and tractable likelihood function.

Suppose each observation can be coded into just one of $J + 1$ possible alternatives.[17] Also assume that the process that leads observations into a specific alternative is driven by a set of latent variables $y_j^* \ \forall j = 0, \ldots, J$. In economic terms y_j^* can be interpreted as the utility obtainable from alternative j.

[17] Let us start counting the alternatives from 0 so that the total number of alternatives is $J + 1$. The reason for this notation will become clear later on when we will need to take one alternative as a reference category.

Define the latent variable as follows:

$$y_j^* = x_j\beta_j + u_j \tag{6.12}$$

where x_j is a vector of explanatory variables that may vary across alternatives, i.e. both the number and the type of explanatory variables may be different across the j alternatives. Consistently with the utility interpretation of y^* and similarly to the multinomial probit, we assume that observations end up in the alternative with the highest y^*.

As usual, y_j^* is not observable. What can be observed in the data is simply the alternative associated to each observation. While in the multinomial probit it was convenient to write the observable information in a set of J dichotomous variables, in this case it is more useful to define one single observable variable y as follows:

$$y = j \text{ if alternative } j \text{ is chosen} \tag{6.13}$$

The multinomial logit models (in their different specifications) are all characterized by the assumption that the error terms are distributed *i.i.d.* according to a *type I extreme value distribution*:

$$u_j \sim \begin{cases} f(u) &= e^{-u-e^{-u}} \\ F(u) &= e^{-e^{-u}} \end{cases} \tag{6.14}$$

This ugly distribution is in fact convenient because it leads to the following specification of the probability of a generic alternative j being chosen:[18],[19]

$$\begin{aligned} Pr(u_j > u_h \ \forall h \neq j|x) &= Pr(y = j|x) = \\ &= \frac{e^{x_j\beta_j}}{\sum_{h=0}^{J} e^{x_h\beta_h}} \end{aligned} \tag{6.15}$$

Notice, however, that the multinomial logit specification not only assumes a different distribution of the errors but it also requires them to be uncorrelated across alternatives. We will discuss this point more at length in section 6.2.

By replacing the observable x_{ih} $\forall h = 0, \ldots, J$ and $\forall i = 1, \ldots, N$ in equation 6.15 we obtain the likelihood contribution of each single observation:

$$P_i(\theta) = Pr(y_i = j|x_i) = \frac{e^{x_{ij}\beta_j}}{\sum_{h=0}^{J} e^{x_{ih}\beta_h}} \tag{6.16}$$

[18] In what follows we define $x = (x_1 \ \ldots x_J)$.

[19] The derivation of expression 6.15 is quite complex and we are not going to analyze it in details here.

where $\theta = (\beta_0', \ldots, \beta_J')'$ is the vector of all the parameters of the model to be estimated. The estimation procedure is the usual maximum likelihood optimization where the log-likelihood function can be easily written as:

$$L(\theta) = \sum_{i=1}^{N} \ln P_i(\theta) \tag{6.17}$$

The model in 6.17 is the most flexible of several possible versions of the multinomial logit and it is usually called *McFadden* or *mixed multinomial logit.*[20]

To fully understand the differences between the various versions of the multinomial logit (we will see three of them), it is important to recognize that the model 6.17 can include three different types of explanatory variables:

1. explanatory variables that vary only across observations i and are constant across alternatives j (e.g. in a model of transportation modes where i are travelers and j are modes of transportation, variables of this type might be age or education of travelers);

2. explanatory variables that vary only across alternatives j and are constant across observations i (e.g. characteristics of the transportation modes, color of the trains, comfort of the cars);

3. explanatory variables that vary both across observations i and across alternatives j (e.g. time to destination, price of the ticket);

The mixed multinomial logit model is a model with explanatory variables of type 2 and/or type 3. In fact, variables of type 1 cannot be identified in the mixed model. To see this in a relatively simple way, suppose that the set of explanatory variables is the same for all alternatives and that such explanatory variables can be divided into two groups: one group z_{ij} contains only regressors of type 2 or type 3, while the other group m_i contains only regressors of type 1. Then the individual likelihood contribution 6.16 becomes:

$$Pr(y_i = j|x_i) = \frac{e^{z_{ij}\delta_j + m_i\gamma_j}}{\sum_{h=0}^{J} e^{z_{ij}\delta_j + m_i\gamma_j}} \tag{6.18}$$

[20]Be careful with the labeling of these models as there is a good deal of confusion in the literature. Different authors and different textbooks give different names to the various models. Here we adopt one notation and naming that is no more common than others, so focus on the actual characteristics of the models rather than on their names.

where δ_j and γ_j are the coefficients of the two types of regressors.

In this model, the coefficients γ_j cannot be identified. Suppose, in fact, that the true γ is γ^*, which means that γ^* maximizes the likelihood 6.17. Now, consider an alternative set of parameters $\zeta = \gamma^* + q$ where q is a simple constant. Notice that the model looks the same regardless of whether we compute it at γ^* or at ζ:

$$\begin{aligned} Pr(y_i = j|x_i) &= \frac{e^{z_{ij}\delta_j + m_i(\gamma_j^* + q)}}{\sum_{h=0}^{J} e^{z_{ij}\delta_j + m_i(\gamma_j^* + q)}} = \\ &= \frac{e^{qm_i} \cdot e^{z_{ij}\delta_j + m_i\gamma_j^*}}{\sum_{h=0}^{J} e^{qm_i} \cdot e^{z_{ij}\delta_j + m_i\gamma_j^*}} = \\ &= \frac{e^{z_{ij}\delta_j + m_i\gamma_j^*}}{\sum_{h=0}^{J} e^{z_{ij}\delta_j + m_i\gamma_j^*}} \end{aligned} \tag{6.19}$$

So, both γ^* and ζ maximize the likelihood of the model but only γ^* is the true set of parameters, which implies that it cannot be estimated with this model.[21]

The other two models of the multinomial logit class that we consider differ on the assumptions imposed on the regressors and the parameters.

Classical multinomial logit model

How can you estimate a multinomial logit model where all the explanatory variables are of type 1? Such model would look like the following:

$$Pr(y_i = j|x_i) = \frac{e^{x_i\beta_j}}{\sum_{h=0}^{J} e^{x_i\beta_h}} \tag{6.20}$$

where we eliminated the subscript j from the set of regressors to indicate that the explanatory variables are constant across alternatives and only vary across observations.

As we saw earlier in equation 6.19, the coefficients of this model cannot be identified. The problem is brutally solved by a conventional assumption that sets the coefficients of the first alternative (alternative 0) to zero:

$$\beta_0 = 0 \tag{6.21}$$

[21] In mathematical terminology, the maximization problem of equation 6.17 is undetermined.

Under this assumption, the model becomes:

$$P_i(\theta) = Pr(y_i = j \neq 0|x_i) = \frac{e^{x_i\beta_j}}{1 + \sum_{h=1}^{J} e^{x_i\beta_h}} \tag{6.22}$$

and it is identified and called *classical multinomial logit model*.

Obviously, the normalization of $\beta_0 = 0$ does not come for free and the parameters of the model 6.20 remain essentially unidentified. In fact, the parameters in 6.22 estimate something slightly different and need to be interpreted accordingly.[22] To see this, consider the probability of the reference alternative 0:

$$Pr(y_i = 0|x_i) = \frac{1}{1 + \sum_{h=1}^{J} e^{x_i\beta_h}} \tag{6.23}$$

so that the probability of a generic alternative j over the probability of the reference alternative zero is:

$$\frac{Pr(y_i = j|x_i)}{Pr(y_i = 0|x_i)} = e^{x_i\beta_j} \tag{6.24}$$

which implies that the coefficients of the classical multinomial model should be interpreted as the *percentage variation in the probability of alternative j relative to the reference alternative 0* due to a marginal change in each regressor:

$$\beta_{jk} = \frac{\partial \ln [Pr(y_i = j|x_i)/Pr(y_i = 0|x_i)]}{\partial x_k} \tag{6.25}$$

Conditional logit model

Now, consider a model where the explanatory variables are always of type 2 and type 3 but where the coefficients are constant across the alternatives:

$$Pr(y_i = j|x_i) = \frac{e^{x_{ij}\beta}}{\sum_{h=0}^{J} e^{x_i\beta}} \tag{6.26}$$

This model is named *Conditional logit* and shares the same feature of the mixed logit: it cannot identify the parameters of regressors of type 1 that vary only across individuals. Obviously, since the set of coefficients β is the same across all alternatives, the conditional logit model also requires that the set of explanatory variables is the same across alternatives. The conditional logit model is essentially a mixed model with constant coefficients across alternatives.

[22] To be formally precise, one should actually call the parameters in 6.22 differently from the ones in 6.20 because they are in fact different things.

The Independence of Irrelevant Alternatives (IIA) assumption

All the multinomial logit models reviewed in this section incorporate an important assumption about the probability structure.

As noted earlier on, the nicely tractable functional form of the multinomial logit derives from a specific distributional assumption. Such assumption, however, contains two important elements.

The first is the actual choice of the distribution, i.e. the type I extreme value instead, for example, of the normal. This part of the assumption is not particularly problematic given that it is as arbitrary as any other distribution. The second element is the i.i.d. feature of the joint distribution of the error terms. Notice that in the multinomial probit model we let all errors to be correlated with each other in a totally unrestricted form. The parameters of the variance-covariance matrix of the error terms were in fact parameters to be estimated.

In the multinomial logit, instead, we make the much stronger assumption that such matrix is perfectly diagonal, i.e. the error terms are uncorrelated with each other.[23] In some sense, the tractability of the multinomial logit models comes at the (non trivial) cost of imposing this much stronger assumption.

One important implication of this assumption is that the relative probability of choosing one alternative over another is independent of what other alternatives are available:

$$\frac{Pr(y_i = j|x_i)}{Pr(y_i = k|x_i)} = \frac{e^{x_{ij}\beta_j}}{e^{x_{ik}\beta_k}} \tag{6.27}$$

Notice that this relative probability does not depend on anything that refers to any of the other available alternatives. This is a feature of the model that might not be appropriate in many empirical applications.

[23] Be careful here not to confuse the correlation of the errors across alternatives and across individuals. Throughout our discussion, we have always maintained the assumption that the errors are i.i.d. across individuals (otherwise we would not be able to write the likelihood function as a simple product of the individual likelihood contributions). In the specific case of the multinomial models, however, each observation is characterized by several error terms, one for each possible alternative. In the multinomial probit we let these error terms be correlated across alternatives (but always uncorrelated across individuals). In the multinomial logit, instead, we assume them to be uncorrelated across alternatives (as well as across individuals).

6.3 Ordered response models

As we mentioned earlier on, another alternative to reduce the complexity of the analysis of multinomial data is to change the assumptions about the data generating process, the process that leads observations into the various categories. How and to what extent this is possible depends on the actual process under consideration.

In this section, we analyze a class of models that can be used when the multinomial data can be ordered according to some meaningful criterium. This is for example the case of survey data where respondents answer on a scale: questions on one's level of satisfaction (1=extremely unsatisfied, 5=very satisfied) or on political orientation (1=very left wing; 10=very right wing) are typical cases.

The advantage of this type of data over more general instanced of multinomial data that cannot be meaningfully ordered is the fact that they can be easily described by making use of one single latent variable, y^*:

$$y^* = x\beta + u \tag{6.28}$$

The observable data is simply $y = \{0, 1, 2, \ldots, J\}$ *@@ the total number of available categories is* $J + 1$. Ordered response models assume that the observable data are determined according to the following rule based on the latent variable y^*:

$$y = \begin{cases} 0 & \text{if } y^* \leq \alpha_1 \\ 1 & \text{if } \alpha_1 < y^* \leq \alpha_2 \\ \vdots & \\ J & \text{if } y^* > \alpha_J \end{cases} \tag{6.29}$$

where $\alpha_1, \ldots, \alpha_J$ are additional parameters of the model that need to be estimated.[24]

Under this assumption, it is relatively easy to derive the probability that any given observation i is observed in each of the @@ $J + 1$ categories:

$$\begin{aligned} Pr(y_i = 0|x_i) &= Pr(x_i\beta + u \leq \alpha_1|x_i) = \\ &= Pr(u \leq \alpha_1 - x_i\beta|x_i) \\ Pr(y_i = 1|x_i) &= Pr(\alpha_1 - x_i\beta < u \leq \alpha_2 - x_i\beta|x_i) \\ &\vdots \\ Pr(y_i = J|x_i) &= Pr(u > \alpha_J - x_i\beta|x_i) \end{aligned}$$

[24] In some cases you might even know the values of such parameters and don't need to estimate them.

At this point, in order to be able to write these probabilities in closed form and then apply maximum likelihood we need to impose some distributional assumption on the error term. The *ordered probit model* is characterized by the normality assumption. Alternatively, one could pick other distributions and a very popular one is also the logistic, which gives rise to the so-called *ordered logit model*.

The important feature of all these models, however, is the fact that, contrary to the previous (unordered) multinomial probit and logit models, here there is only one error term to be considered so that all the issues connected with the possibility that the random errors of different choices were correlated are now completely irrelevant.

Let us stick to the multinomial probit and derive the likelihood of the model under the (standard) normality assumption, $u \sim N(0,1)$:

$$\begin{aligned} Pr(y_i = 0|x_i) &= \Phi(\alpha_1 - x_i\beta) \\ Pr(y_i = 1|x_i) &= \Phi(\alpha_2 - x_i\beta) - \Phi(\alpha_1 - x_i\beta) \\ &\vdots \\ Pr(y_i = J|x_i) &= 1 - \Phi(\alpha_J - x_i\beta) \end{aligned}$$

The individual log-likelihood contribution can then be expressed as follows:

$$\begin{aligned} \ell_i(\alpha,\beta) &= 1(y_i = 0)\ln\Phi(\alpha_1 - x_i\beta) + 1(y_i = 1)\ln\left[\Phi(\alpha_2 - x_i\beta) - \Phi(\alpha_1 - x_i\beta)\right] \\ &\quad + \cdots + 1(y_i = J)\ln\left[1 - \Phi(\alpha_J - x_i\beta)\right] \end{aligned} \tag{6.30}$$

where α is the vector of all the auxiliary parameters $\alpha_1 \ldots \alpha_J$. And the complete log likelihood is simply the sum of the log-likelihood contributions for all observations in the sample:

$$L(\alpha,\beta) = \sum_{i=1}^{N} \ell_i(\alpha,\beta) \tag{6.31}$$

Maximization of equation 6.31 leads to the maximum likelihood estimators of α and β.

A final important note on marginal effects is necessary.[25] Like all other non linear models, the coefficients β of the ordered probit do not coincide with the marginal effects of the probabilities of observing individuals in any of the @@ $J+1$ categories. In particular, if we were interested in the marginal effects of a generic explanatory variable x_k on the probabilities of

[25] Although all of them apply identically to ordered logit as well

observing individuals in any of the categories, such effects can be calculated as follows:

$$\begin{aligned}
\frac{\partial Pr(y_i = 0|x_i)}{\partial x_{ik}} &= -\beta_k \varphi(\alpha_1 - x_i\beta) \\
\frac{\partial Pr(y_i = j|x_i)}{\partial x_{ik}} &= -\beta_k \left[\varphi(\alpha_{j-1} - x_i\beta) - \varphi(\alpha_j - x_i\beta)\right] \\
&\quad \forall j = 1, \ldots, J-1 \\
\frac{\partial Pr(y_i = J|x_i)}{\partial x_{ik}} &= \beta_k \varphi(\alpha_J - x_i\beta)
\end{aligned}$$

So, the sign of the coefficients unambiguously determines the sign of the marginal effects only for the probability of the first and the last option (and in opposite directions) while the sign of the marginal effects of all intermediate choices remains ambiguous and depends on the α's. Notice, however, that the β's can always be interpreted as the marginal effects of the regressors on the latent variable y^*, which is particularly useful when the latent variable can be given some easily interpretable meaning and it is not a mere modeling device.

6.4 Nested choice models

Another possibility to change the assumptions about the data generating process in a convenient way are *nested models*. We are not going to look at them in details and we just present the intuition in this brief section.

Nested models can be applied when the decision process lends itself to de divided in a sequence of decisions. For example, suppose we have data on car purchases of a sample of individuals and that there are a total of 6 car models, A, B, C, D, E and F. Also suppose that these 6 models can be grouped in 3 categories: sport cars (A and B), station wagons (C and D) and off-roads (E and F). We could then assume that people first choose what type of car they want (i.e. sport, station or off-road) and then, conditional on the type, they choose the model.

There are obviously many decisions that can be split in this fashion and in all these cases nested models can be applied. The most common distributional assumption for these models is the logistic distribution which gives rise to the *nested logit model*.

Nested models are relatively easy to handle computationally (especially in their logistic version) and do not require the strong independence assumption that is needed for the multinomial logit models.

7 Models for incomplete observations: censoring and truncation

In this section we present models for the analysis of data where the dependent variable is not completely observed. In particular, we are going to consider two cases: *censoring* and *truncation*. Let us start by looking at what censored and truncated data are.

Censored data. Censored data arise when a given variable is censored at some point in the distribution. For example, if y^* is the true uncensored variable, then a version of y^* that is censored from below (or *bottom coded*) at a certain level a can be defined as follows:

$$y = \begin{cases} y^* & \text{if} \quad y^* > a \\ a & \text{if} \quad y^* \leq a \end{cases} \tag{7.1}$$

Similarly for censoring from above (or *top coding*) at a given level b:

$$y = \begin{cases} y^* & \text{if} \quad y^* < b \\ b & \text{if} \quad y^* \geq b \end{cases} \tag{7.2}$$

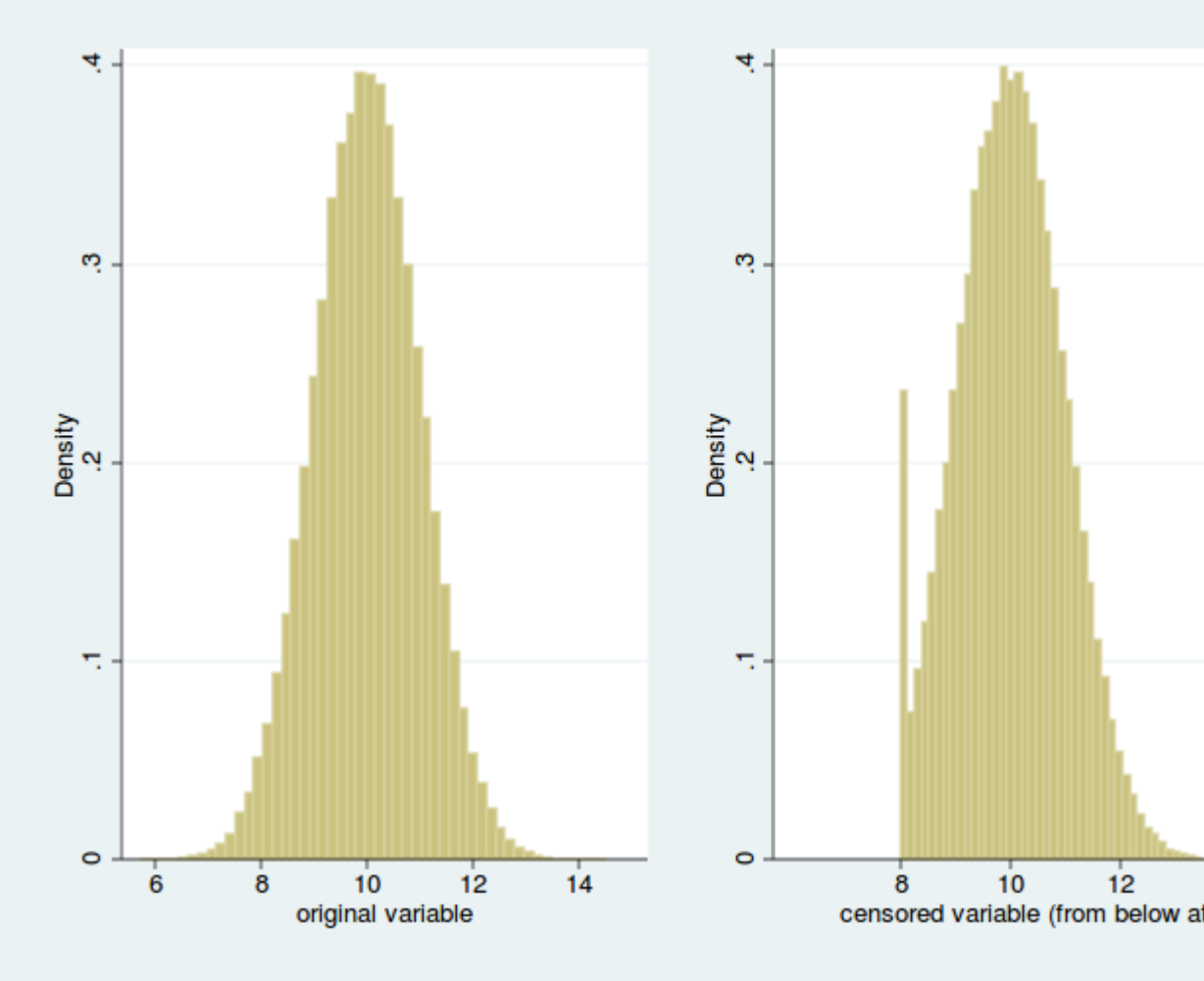
Obviously, censoring can also take place from both below and above. Censored data can arise for two rather distinct reasons. First, data could be artificially top or bottom coded at some levels. A typical example is income data that are very frequently top coded at some level, i.e. those earning very high incomes are all coded as earning a given maximum level (say 500,000 euros per year). Such censoring is usually imposed to prevent identification of the top earners in accordance with existing privacy legislations.

In some other instances censored data arise naturally from the specific problem under consideration. This is, for example, the case of charity donations where lots of persons simply decide not to donate anything and the distribution of donations typically shows a mass point at zero and a continuous distribution for positive values. In these cases of natural censoring, the uncensored variable y^* as we defined it above simply does not exist (or perhaps it is only a latent variable, like the propensity to donate in the charity example) and the true variable is already censored.

The figure below shows how a normally distributed variable (with mean 10 and variance 1), displayed in the left panel, looks like under censoring from below at 8 (in the right panel).

Truncated data. Truncated data are very similar to censored ones with the only exception that the observations beyond the censoring thresholds

Figure 1: Censoring from below

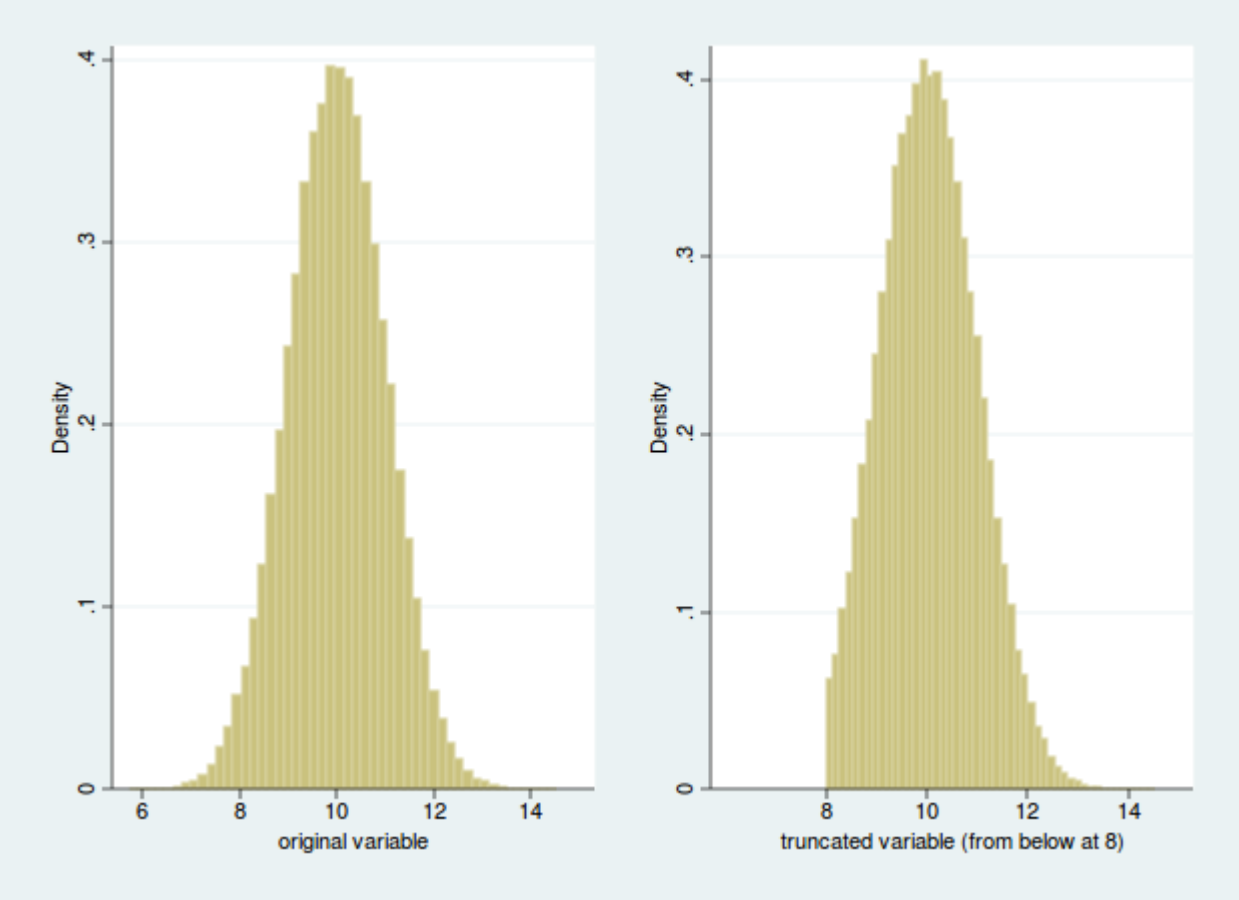


(either below or above) are not censored but simply erased from the sample or set to missing. Following the same examples 7.1 and 7.2, a similar but truncated (from both bottom and above) example would look as follows:

$$y = \begin{cases} y^* & \text{if} \quad a < y^* < b \\ . & \text{otherwise} \end{cases} \tag{7.3}$$

Similarly, figure 7 shows the difference between the same normally distributed variable of figure 7 and its truncated version (from bottom at 8).

Figure 2: Truncation from below



In some cases we are forced to treat a variable as truncated even though its distribution might be fully observable simply because some of the explanatory variables X are not observable at some points of the distribution

of y. Imagine, for example, a situation in which you want to study the relationship between earnings and firm size but firm size is not available for those earning above 500,000 euros a year. In this case, although the actual distribution of earnings might not be truncated, the distribution of those observations that you can use for the analysis - i.e. those that have valid earnings and firm size information - is in fact truncated.

7.1 The implications of censoring and truncation for OLS

What happens if we simply run OLS regression with a censored or truncated dependent variable? To see the implications of censoring and truncation on the OLS method, consider the following model:

$$y^* = X\beta + u \tag{7.4}$$

where y^* is the complete variable (i.e. not censored nor truncated). Also assume that the model satisfies the standard OLS assumptions:

$$E(u) = 0 \tag{7.5}$$
$$E(X'u) = 0 \tag{7.6}$$

These assumptions imply that if y^* could be observed, the OLS estimator of β would be nicely consistent.[26] Unfortunately, y^* is not fully observable. Let us consider what happens if we run OLS on a censored or truncated version of y^*.

7.2 The implications of censoring for OLS

The best way to understand the implications of censoring for OLS (for truncation we will use the same methodology) is to look at the functional form of the *regression function*. OLS assumes that the functional form of the conditional mean of the dependent variable given the regressors is simply linear. Such conditional mean is also called the regression function. Given the linear model of equation 7.4, the regression function is in fact simply linear:

$$E(y^*|X) = X\beta \tag{7.7}$$

If we run OLS using a censored version of y^* we are essentially assuming that also the conditional mean of the censored dependent variable given the

[26] We keep assuming that the consistent (and efficient) estimation of β is the main objective of our analysis.

regressors is linear. So, let us assume a specific censoring - for simplicity assume censoring from below at zero - and look at the conditional mean under the assumption that the uncensored model is the one in 7.4 and that it satisfies the assumptions 7.5 and 7.6.

Call y the censored variable and define it as follows:

$$y = \begin{cases} y^* & \text{if} \quad y^* > 0 \\ 0 & \text{if} \quad y^* \leq 0 \end{cases} \tag{7.8}$$

Then, the conditional mean of y given X can be computed as follows:

$$\begin{aligned} E(y|X) &= Pr(y = 0|X) \times 0 + Pr(y > 0|X)E(y|X, y > 0) && (7.9) \\ &= Pr(y > 0|X)E(y|X, y > 0) \\ &= Pr(u > -X\beta)\left[X\beta + E(u|u > -X\beta)\right] && (7.10) \end{aligned}$$

Equation 7.10 shows clearly that the conditional mean of y in a regression on the explanatory variables X is far from linear.

An alternative way of interpreting equation 7.10 is the following:

$$E(y|X) = X\beta + [Pr(u > -X\beta)E(u|u > -X\beta) - (1 - Pr(u > -X\beta))X\beta] \tag{7.11}$$

Equation 7.11 shows that if we estimate a OLS regression of y on X we are essentially omitting one variable, the complex term in square brackets. In fact, that term is part of the regression function but we omit it. Moreover, since it is a complex function of X and the error term u, it is also correlated with x itself and it is therefore not an innocuous omitted variable, i.e. one that is uncorrelated with the explanatory variables and can be safely left in the error term.[27]

This discussion should have convinced you that running OLS on a model with a censored dependent variable does not guarantee consistent estimates of the parameters.

7.3 The implications of truncation for OLS

A very similar story applies to truncated data. Consider a case where the truncated variable is defined as follows:

$$y = \begin{cases} y^* & \text{if} \quad y^* > 0 \\ . & \text{if} \quad y^* \leq 0 \end{cases} \tag{7.12}$$

[27] It might be possible to find a particular distributional assumption for u such that the omitted term is uncorrelated with X but that must be a very particular case.

In this case the regression function can be computed as:

$$\begin{aligned} E(y|X) &= E(y^*|X, y^* > 0) \\ &= E(X\beta + u|X, X\beta + u > 0) \\ &= X\beta + E(u|X, u > -X\beta) \end{aligned} \tag{7.13}$$

Similarly to equation 7.11, also equation 7.13 shows that the conditional mean of the truncated dependent variable is equal to $X\beta$ plus an additional term that is a (usually non-linear) function of both X and u.

If we do not specify this term, which is what happens if we simply run OLS of y on X, then the term $E(u|X, u > -X\beta)$ is omitted and goes into the error term. And just like in the censoring example above, such term is in general correlated with X and thus impedes identification of the parameters.[28]

Hence, also in the case of truncation, simply running OLS of the truncated dependent variable on the regressors does not produce consistent estimates of the parameters.

7.4 Consistent estimation of *censored* data models: the Tobit model

In equation 7.10 we showed that when the dependent variable is censored at zero the regression function is

$$E(y|X) = Pr(u > -X\beta)\,[X\beta + E(u|u > -X\beta)] \tag{7.14}$$

Without further assumptions about the distribution of u we would not be able to move ahead. The Tobit model allows to consistently estimate the parameters of the model 7.4 under the assumption 7.5 and 7.6 and the additional assumption that the error term of the model is normally distributed:

$$u \sim N(0, \sigma^2) \tag{7.15}$$

Under normality we can derive the density of $y|X$ and apply maximum likelihood. In doing that we need to be careful and notice that this time the form of the likelihood contribution is not going to be identical for all observations in the sample. In particular, for censored observations, i.e. those with $y_i = 0$, the likelihood is:

$$Pr(y_i = 0|X_i) = 1 - \Phi(X_i\beta/\sigma) \tag{7.16}$$

[28] We say "in general" because one might actually be able to find a specific distributional assumption on u that guarantees that $E(u|X, u > -X\beta)$ does not depend on X.

The likelihood contribution of the non-censored observations, i.e. those with $y_i > 0$, is instead:

$$f(y_i|X, y_i > 0) = f(y_i^*|X, y_i^* > 0) \tag{7.17}$$

Notice that, while the computation of equation 7.16 is straightforward, equation 7.17 is more complicated. How can we get a closed form expression for the truncated density $f(y_i^*|X, y_i^* > 0)$? The simplest and still most formally correct way to proceed is to consider the truncated cumulative and then take its derivative:

$$\begin{aligned} F(c|y^* > 0) &= Pr(y^* < c|y^* > 0) = \frac{Pr(y^* < c, y^* > 0)}{Pr(y^* > 0)} \\ &= \frac{Pr(0 < y^* < c)}{Pr(y^* > 0)} = \frac{F(c) - F(0)}{1 - F(0)} \end{aligned} \tag{7.18}$$

Now, let us simply compute $f(c|X, y^* > 0)$ by taking the derivative of $F(c|y^* > 0)$:[29]

$$\begin{aligned} f(c|X, y^* > 0) &= \frac{\partial F(c|y^* > 0)}{\partial c} \\ &= \frac{\partial \left[\frac{F(c)-F(0)}{1-F(0)}\right]}{\partial c} = \\ &= \frac{f(c)}{1 - F(0)} \end{aligned} \tag{7.19}$$

where, under assumption 7.15, $f(c) = \frac{1}{\sigma}\phi(\frac{c-X\beta}{\sigma})$ and $1 - F(0) = \Phi(\frac{X\beta}{\sigma})$.

Notice, however, that equation 7.19 describes the density of a random variable that is distributed normally but truncated at zero. In other words, the density in equation 7.19 integrates to 1 over the range $(0, \infty)$. In the case of a censored model, however, the zeros are observed so that the density of the entire model should integrate to 1 over the range of all possible continuous positive values **and** the values concentrated in the mass point at zero.

To account for the mass point, the density of the continuous values should be weighted by the probability of observing values larger than zero, that is:

$$\begin{aligned} \Pr(y > 0|X) &= \Pr(X\beta + u > 0|X) = \Pr(u > -X\beta|X) \\ &= 1 - \Phi(-X\beta/\sigma) = \Phi(X\beta/\sigma) \end{aligned}$$

[29] In this formula we use f and F to indicate generic density and cumulative functions although we have already assumed our model to be normal. This is because even under normality of y^* $f(c|X, y^* > 0)$ is not the standard normal density but rather its truncated version.

In the end, the likelihood contribution of observations with $y_i > 0$ can be written as:

$$f(y_i|X_i, y > 0) = \Phi(X_i\beta/\sigma) f(y_i|X_i, y^* > 0) = \frac{1}{\sigma}\phi\left(\frac{y_i - X_i\beta}{\sigma}\right) \tag{7.20}$$

Using equations 7.16 and 7.20 it is now easy to derive the log-likelihood contribution of a generic observation i as follows:

$$\ell(\beta, \sigma) = 1(y_i = 0)\ln\left[1 - \Phi(X_i\beta/\sigma)\right] + 1(y_i > 0)\ln\left[\frac{1}{\sigma}\phi\left(\frac{y - X\beta}{\sigma}\right)\right] \tag{7.21}$$

and then the complete log-likelihood function:

$$L(\beta, \sigma) = \sum_{i=1}^{N}\left\{1(y_i = 0)\ln\left[1 - \Phi(X_i\beta/\sigma)\right] + 1(y_i > 0)\ln\left[\frac{1}{\sigma}\phi\left(\frac{y - X\beta}{\sigma}\right)\right]\right\} \tag{7.22}$$

This model is known under the name *Tobit*.

Maximization of equation 7.22 leads to maximum likelihood estimators of β and σ. Notice that, contrary to most previous models, now β and σ can be identified separately and there is no need to normalize the variance of the distribution.

7.5 Consistent estimation of *truncated* data models: the truncated regression model

Having derived the Tobit model, it is now very easy to derive the correct likelihood function for truncated data. A truncated model is simply a censored one where the censored observations are missing. So, the density of the model is simply the density described in equation 7.19 without any further adjustment. Truncated models are also generally estimated under normality so that the density of the model can be written as follows:

$$\begin{aligned} f(y|X) &= f(y^*|X, y^* > 0) = \frac{f(y)}{1 - F(0)} \\ &= \frac{\frac{1}{\sigma}\phi\left(\frac{y - X\beta}{\sigma}\right)}{\Phi(X\beta/\sigma)} \end{aligned} \tag{7.23}$$

Using this expression, then, the individual log-likelihood contribution is:

$$\ell_i(\beta, \sigma) = -\ln\sigma + \ln\phi\left(\frac{y_i - X_i\beta}{\sigma}\right) - \ln\left[\Phi(X\beta/\sigma)\right] \tag{7.24}$$

and the complete log-likelihood is simply:

$$L(\beta, \sigma) = -N \ln \sigma + \sum_{i=1}^{N} \left\{ \ln \phi \left(\frac{y_i - X_i \beta}{\sigma} \right) - \ln \left[\Phi(X\beta/\sigma) \right] \right\} \tag{7.25}$$

7.6 Additional notes on censored and truncated data models

Let us conclude this section on censored and truncated data with a couple of comments on these models.

The first important thing to notice is that censoring is 'better' than truncation, in the sense that censored data contain more information about the true underlying distribution of the data than their truncated analogs. The reason should be obvious. With censored data the censored observations are available (i.e. the X's are observable) and are simply censored at some given level (zero in the examples above). With truncated data, the truncated observations are simply not available.

The second important note concerns the marginal effects. In the case of censored and truncated data the type of marginal effects of main interest really depends on the specific analysis. In principle, in fact, we may be interested in the effect of changes in the explanatory variables on the conditional mean of either the true dependent variable y^* or the censored/truncated dependent variable y. And our discussion in section 7.1 shows that the two conditional means are very different. If we want to look at y^*, then the conditional mean is simply $E(y^*|X) = X\beta$ and the coefficients already directly represent marginal effects. If we are interested in the censored/truncated distribution, then, we know that, depending on whether the data are censored or truncated, the conditional mean looks as follows:

$$\text{Censoring:} \quad E(y|X) = Pr(u > -X\beta) \left[X\beta + E(u|u > -X\beta) \right] \tag{7.26}$$

$$\text{Truncation:} \quad E(y|X) = X\beta + E(u|u > -X\beta) \tag{7.27}$$

Usually, when the censoring/truncation arises naturally from the structure of the data, like in the example on charity donations, the most interesting effects to analyze are perhaps those on the censored/truncated mean of the data. On the other hand, when censoring/truncation arises from some statistical artifact of an underlying true distribution, like in the example of data anonymization for privacy purposes, we are typically interested in the effects of the explanatory variables on the true uncensored/untruncated data.

Notice, additionally, that in order to be able to compute the marginal effects on either 7.26 or 7.27, we need to write $E(u|u > -X\beta)$ in some

explicit format. This is only possible under normality of u. In particular, it is only when u is normal that it is possible to derive an explicit functional form for the truncated mean $E(u|u > -X\beta)$. In fact, the following general property holds only for the normal distribution:

$$\text{if } z \sim N(\mu, \sigma^2) \;\Rightarrow\; E(z|z > c) = \mu + \sigma \frac{\varphi(\frac{c-\mu}{\sigma})}{1 - \Phi(\frac{c-\mu}{\sigma})} \tag{7.28}$$

Using the property in 7.28 we can write the following:

$$\begin{aligned} E(u|u > -X\beta) &= \sigma \frac{\varphi(\frac{-X\beta}{\sigma})}{1 - \Phi(\frac{-X\beta}{\sigma})} \\ &= \sigma \frac{\varphi(\frac{X\beta}{\sigma})}{\Phi(\frac{X\beta}{\sigma})} \\ &= \sigma\lambda\left(X\beta/\sigma\right) \end{aligned} \tag{7.29}$$

where $\lambda(c) = \frac{\varphi(c)}{\Phi(c)}$ is called the *inverse Mills ratio.*

Now, combining equations 7.26 or 7.27 with 7.29 yields:

$$\text{Censoring:} \quad E(y|X) = \Phi(X\beta/\sigma)X\beta + \sigma\phi\left(X\beta/\sigma\right) \tag{7.30}$$

$$\text{Truncation:} \quad E(y|X) = X\beta + \sigma\lambda\left(X\beta/\sigma\right) \tag{7.31}$$

and the marginal effects can be easily calculated.[30]

8 Sample selection: the Heckman model

So far we have always assumed that the sample at hand was a random draw from the population of interest. Unfortunately, there is plenty of cases in which this is not really the case and the data that can be used for the analysis are a non-random selection of the population.

To understand the implications of selection, let us consider the following general model:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_K x_K + u \tag{8.1}$$

that satisfies the standard exogeneity assumption $E(u|X) = 0$, where $X = (1\ x_1 \ldots x_K)$ is the vector of explanatory variables of the model.[31] The only problem with this model is selection.

[30] You will have to do it in a problem set.

[31] For simplicity, let us make the exogeneity assumption in its stronger conditional mean independence form. Most of the analysis in this section holds also under the weaker $E(X'u) = 0$.

For example, suppose that a random sample of N observations was originally extracted from the population but information on the variables of the model was missing for some of them so that the model can effectively be run only on a selected set of N. Define an indicator s_i which is equal to 1 for those observations that can be used for the estimation and zero for the others:

$$s_i = \begin{cases} 1 & \text{if } \{y_i, X_i\} \text{ exists} \\ 0 & \text{if } \{y_i, X_i\} \text{ does not exist or is incomplete} \end{cases} \tag{8.2}$$

With this set-up it is easy to write the OLS estimator of $\beta = (\beta_0\, \beta_1 \ldots \beta_K)'$ as follows:

$$\begin{aligned} \widehat{\beta}_{OLS} &= \left[\sum_{i=1}^{N} s_i X_i' X_i\right]^{-1} \left[\sum_{i=1}^{N} s_i X_i' y_i\right] \\ &= \beta + \left[\sum_{i=1}^{N} s_i X_i' X_i\right]^{-1} \left[\sum_{i=1}^{N} s_i X_i' u_i\right] \end{aligned} \tag{8.3}$$

Equation 8.3 clearly shows that the OLS estimator computed on a selected sample is consistent under the usual conditional mean independence assumption only if $E(sX'u) = 0$, which is true only if $E(u|s) = 0$ (on top of $E(u|X) = 0$). In other words, for OLS to produce consistent estimates on a selected sample we need the mean of the error term to be independent not only from the regressors but also from the selection process.

Let us now consider some examples of selection that are common in applied work.

Random selection. Suppose that s is a simple random Bernoulli variable with parameter p:

$$s \sim Bernoulli(p) \tag{8.4}$$

This type of selection arises, for example, when we work with a dataset that is very large, too large for our computer to perform calculations in a reasonable time. In this case, we may want to use a random subsample of the data in order to maintain its representativeness but reduce computational complexity. A simple way of selecting such random subsample is precisely to define a random Bernoulli indicator that takes value 1 for the selected observations and zero for the others. The parameter of the Bernoulli determines which fraction of the original data we select. With $p = 0.5$ we select half of the original sample. By definition, this type of selection is exogenous and

satisfies the condition $E(u|s) = 0$. Thus, performing OLS on a randomly selected sample does not pose any specific problem.

Deterministic selection. Suppose now that selection is based on a deterministic rule $g(\cdot)$ of the exogenous variables:

$$s = g(X) \tag{8.5}$$

This is, for example, the case of selection of the sample based on age or gender. In many cases we work only with individuals in a certain age range or of a certain gender. By definition, if $E(u|X) = 0$ and s is a deterministic function of X, then also $E(u|s) = 0$.[32] Hence, in this case, too, selection does not pose any particular problem for the consistent estimation of β with OLS.[33] Notice, finally, that deterministic selection may occur also when the determinist variables that define the selection process are not in the set of explanatory variables of the model. What is important is that they are exogenous.

Selection based on the dependent variable (or truncation). Truncated data are data that arise from sample selection based on the dependent variable. In fact, if we define s as follows:

$$s = \begin{cases} 1 & \text{if} \quad a_1 < y < a_2 \\ 0 & \quad \text{otherwise} \end{cases} \tag{8.6}$$

then, the distribution of the selected observations is truncated from below at a_1 and from above at a_2. By definition, this type of selection is not exogenous since $E(u|y)$ cannot be equal to zero as y is itself a function of u. In fact, we already showed in section 7.1 that OLS on truncated data leads to inconsistent estimates and in section 7.5 we discussed how to solve the problem.

[32] Notice that this implication is not necessarily guaranteed if we use the weaker assumption $E(X'u) = 0$.

[33] Notice, however, that if the true value of β varies across observations, then what we estimate on the selected sample is the parameter for the specific selected population. For example, suppose we want to estimate the returns to education in a wage regression and we run the estimation only on males. There is plenty of reasons to believe that the returns to education differ between gender (they are typically lower for women) so, if we run the model only on males we estimate consistently the returns for males which is a different parameter from the average returns to education that we would estimate if we were to use both males and females for the estimation.

Endogenous selection. Endogenous selection arises whenever $E(u|s) \neq 0$.[34] This may happen in many circumstances. For example, in survey data where people are asked about their incomes it is often the case that people at the tails of the wage distribution refuse to answer: the very poor because they are ashamed of confessing to a stranger their actual economic conditions, the very rich because they might be afraid that their high income makes their answer identifiable in the data. In this situation, we only observe income data for those who actually answered the question who might be a very non-randomly selected group of the original sample of interviewed individuals.

Another important example, that is in fact the one specific application that motivated the original developers of sample selection models, is based on a model of *wages and labour market participation*. Suppose that individuals differ in their productivity as well as in their preferences for work. More productive workers will receive higher wage offers from potential employers. Call w_i^0 the wage offer received by individual i. Workers with stronger preferences for work have lower reservation wages, i.e. the minimum wage at which a person is willing to work. Call w_i^r the reservation wage of individual i. Define both w^0 and w^r as linear functions of completely or partially different explanatory variables:

$$w_i^0 = X_{i1}\beta_1 + u_{i1} \tag{8.7}$$
$$w_i^r = X_{i2}\beta_2 + u_{i2} \tag{8.8}$$

Also assume that both these equations satisfy the conditional mean independence assumption:

$$E(u_1|X_1) = 0 \tag{8.9}$$
$$E(u_2|X_2) = 0 \tag{8.10}$$

Our interest is primarily in the estimation of the parameters of the wage offers equation β_1. Obviously, people work only if their wage offer is larger than their reservation wage:

$$\begin{array}{lll} w_i^0 \geq w_i^r & \Rightarrow i \text{ works} \\ w_i^0 < w_i^r & \Rightarrow i \text{ is inactive/unemployed} \end{array} \tag{8.11}$$

What is typically observed in the data is actual wages for individuals who work and no wages for those who do not work. In other words, we only

[34] So, also truncation can be interpreted as a case of endogenous selection and, in fact, it can also be solved using the methodology that we discuss here.

observe w_i^0 if $w_i^0 \geq w_i^r$. We can then define the selection rule as follows:

$$\begin{aligned} s_i &= 1\left(w_i^0 \geq w_i^r\right) \\ &= 1\left(X_{i1}\beta_1 - X_{i2}\beta_2 + u_{i1} - u_{i2} \geq 0\right) \\ &= 1\left(Z_i\delta + v_i \geq 0\right) \end{aligned} \tag{8.12}$$

where $Z_i = (X_{i1}\ X_{i2})$, $\delta = (\beta_1\ \beta_2)'$ and $v_i = u_{i1} - u_{i2}$. The model can then be described by the following two equations, the wage equation and the participation equation:

$$w_i^0 = X_{i1}\beta_1 + u_{i1} \tag{8.13}$$

$$s_i = 1\left(Z_i\delta + v_i \geq 0\right) \tag{8.14}$$

In this specific example, selection is clearly endogenous since the random component of the selection equation v_i is correlated with the error term of the wage equation u_{i1} by construction. In fact, $v_i = u_{i1} - u_{i2}$. Notice in particular that v_i and u_{i1} would be correlated even if u_{i1} and u_{i2} were not.[35]

8.1 Consistent estimation of models with sample selection

In this section we present a general methodology to estimate models using a non-randomly selected sample. As a general preamble, this will only be possible if we have information about how the selection was carried out and at least some variables are available for the non-selected observations as well.

Let us first generalize the previous labour market participation example into a more general model with endogenous selection. The model is composed of two equations: a *main equation*, whose parameters are our main interest; and a *selection equation* that describes the selection process.

$$y_i = X_i\beta + u_i \tag{8.15}$$

$$s_i = 1\left(Z_i\delta + v_i > 0\right) \tag{8.16}$$

We will discuss the estimation of this model under the following assumptions:

- s_i and Z_i are always observed for all the N sampled observations;
- y_i and X_i are observed only when $s_i = 1$;[36]

[35] Note additionally that v_i and u_{i1} may be correlated also through the explanatory variables if $E(u_{i1}|X_{i2}) \neq 0$ and/or $E(u_{i2}|X_{i1}) \neq 0$.

[36] We can relax this and allow either y_i or X_i to be always observable. What is important is that the main equation can only be run on the selected sample of observations with $s_i = 1$.

- $E(u|X,Z) = E(v|X,Z) = 0$;
- $v \sim N(0,1)$;[37]
- $E(u|v) = \gamma v$, with $\gamma \neq 0$. This assumption states on the one hand that u is not mean independent from v and, additionally, it also imposes a specific linear form to such conditional mean.[38]

Now, consider the regression function of y on the selected observations:

$$\begin{aligned} E(y|X, s=1) &= X\beta + E(u|X, s=1) \\ &= X\beta + E(u|X, v > -Z\delta) \end{aligned}$$

Using the above assumptions we can now develop this expression further. $E(u|v) = \gamma v$ implies that u can be simply written as $u = \gamma v + \xi$ where ξ is a non-systematic error with zero mean. Hence:

$$\begin{aligned} E(y|X, s=1) &= X\beta + E(u|X, v > -Z\delta) \\ &= X\beta + E(\gamma v + \xi|X, v > -Z\delta) \\ &= X\beta + \gamma E(v|v > -Z\delta) \end{aligned}$$

Now, we can exploit the property of the truncated normal distribution 7.28 that we have already used in section 7.6 and obtain:

$$\begin{aligned} E(y|X, s=1) &= X\beta + \gamma E(v|v > -Z\delta) \\ &= X\beta + \gamma \frac{\phi(-Z\delta)}{1 - \Phi(-Z\delta)} \\ &= X\beta + \gamma \frac{\phi(Z\delta)}{\Phi(Z\delta)} \\ &= X\beta + \gamma\lambda(Z\delta) \end{aligned} \tag{8.17}$$

where $\lambda(\cdot)$ is the *inverse Mills ratio*.

Equation 8.17 is crucial to understand the spirit of the estimation procedure that we are going to describe now. Equation 8.17 shows that the true regression function of y given X on the set of selected observations is not simply $X\beta$ but it includes an additional term $\gamma\lambda(Z\delta)$. Then, if we simply

[37] This assumption can be relaxed to $v \sim N(0, \sigma^2)$ but normality is paramount.

[38] Many textbooks present this model by assuming normality for both u and v. In reality, normality of u is not really required. What is needed is that the functional form of the conditional mean of u given v is linear. If both u and v are normally distributed, the conditional mean is in fact linear with $\gamma = \frac{Cov(u,v)}{Var(v)}$.

run OLS of y on X using the selected observations, the true functional form of the model is the one described in equation 8.17 and the term $\gamma\lambda(Z\delta)$ is left in the error term. Since in most cases X and Z overlap at least partially - i.e. some of the explanatory variables in X also appear in Z - this omission induces endogeneity of X and impedes identification of the parameters β.[39]

One simple solution to this identification problem is to include the term $\lambda(Z\delta)$ in the regression and estimate γ as an additional parameter. This is the intuition of a famous paper by James Heckman who came up with a simple procedure to consistently estimate the parameters of this model.[40] The title of the paper is already suggestive: *"Sample selection bias as a specification error"*.

But, how can we compute the omitted variable $\lambda(Z\delta)$? In reality, we already know most of the ingredients needed to compute it: we know its exact functional form (thanks to the normality assumption) and we already know the values of the Z's for each observation (thanks to the assumption that Z_i is always observable). The only missing ingredient is the vector of parameters δ.

We can, however, obtain consistent estimates of δ by simply running a probit (or logit or linear probability) model of s on Z:

$$\Pr(s = 1|Z) = \Phi(Z\delta) \tag{8.18}$$

Notice that this model can only be run if s and Z are available for all the observations, both the selected and the non selected. Otherwise, the dichotomous dependent variable of the probit would simply be constant.

Once we have consistent estimates of δ, call it $\widehat{\delta}$, we can produce a consistent estimate of the inverse Mills ratio for each observation in the sample:

$$\widehat{\lambda}_i = \lambda(Z_i\widehat{\delta}) \tag{8.19}$$

and use it to run OLS on our main equation:

$$y_i = X_i\beta + \gamma\widehat{\lambda}_i + u_i \tag{8.20}$$

to obtain consistent estimates of β and γ, call them $\widehat{\beta}$ and $\widehat{\gamma}$. This procedure is commonly known as the *Heckman procedure* or *Heckman selection model*.

[39] It is very hard to think of an application in which X and Z are completely different sets of variables. Typically, in applied work one faces the opposite problem that X and Z are identical. In section 8.2 we discuss these issues in some more details.

[40] Heckman, J.J. (1979). *"Sample selection bias as a specification error"*. Econometrica, vol. 47, 153-161.

Computing the correct standard errors of these estimates is a bit more complicated. The reason is, once again, that in equation 8.1 one regressor is an estimate itself and this affects the form of the variance-covariance matrix of the estimated parameters. We are not going to discuss the details of how to compute the correct standard errors of a selection model. Someone has done it for us and most (all) statistical softwares adjust the estimates accordingly.

8.2 Additional notes on the *Heckman selection model*

In this section we discuss a few specificities of the Heckman model.

The first important note concerns the relationship between X and Z. While it is rather difficult to think of a specific application where they are completely different sets of variables, the problem one often faces in applied work is that they tend to be identical. Let us now consider in some more details what are the implications of having X and Z completely separated and completely identical.

If X and Z are sets of completely different variables, i.e. none of the explanatory variables in X appear in Z and viceversa, then by looking at equation 8.17 it should be clear that omitting the term $\lambda(Z\delta)$ from the equation does not induce endogeneity of X. In this sense, then, the parameters β could be consistently estimated by simply running OLS of y on X using only the selected observations. The only one parameter that would be inconsistent in this case is the constant, unless $E[\lambda(Z\delta)] = 0$.

If X and Z are instead completely identical, then we might have a problem of multicollinearity. In fact, if $X = Z$ it is possible to show that the inverse Mills ratio can be reasonably approximated with a linear function:

$$E(y|X) \approx X\beta + a + bZ\delta = X(\beta + b\delta) + a \tag{8.21}$$

where a and b are simple constants. Equation 8.21 shows that when $X = Z$ the model is close to be perfectly collinear and identification is only guaranteed by the fact that the inverse Mills ratio is a non-linear function. Collinearity typically leads to excessively large standard errors.

In general, then, it is good practice to estimate selection models in the same spirit of instrumental variables, with $Z = X + Z_1$. In other words, the vector of explanatory variables of the selection equation should include all the explanatory variables of the main model plus some others, the Z_1, which are usually called *excluded variables* or *exclusion restrictions*.

This is really very similar to the 2SLS approach, where the first stage regression is run with all the explanatory variables of the main model plus

the instruments. In fact, if one or more of the X's were excluded from the selection equation they could induce endogeneity in the main equation if they were instead part of the true selection process and we excluded them, just like omitting some of the explanatory variables of the main model from the first stage equation in a 2SLS model can lead to inconsistency.

Also, if no excluded variables Z_1 were included so that $X = Z$ the selection model would be almost perfectly collinear, just like 2SLS would be perfectly collinear and thus un-identifiable without instruments. And just like with instruments, the search for valid excluded variables requires the same creativity and imagination of the search for valid instruments.

Finally, note that the Heckman procedure automatically produces estimates of γ that can be used to readily test the hypohesis of endogenous selection. If $\widehat{\gamma}$ turns out not to be significantly different from zero, then it means that the selection process was exogenous to the main equation and we could have produced consistent estimates of β by simply running OLS of y on X using the selected observations.

Part 3:
Introduction to the Econometrics of Panel Data

9 What are panel data and why are they so useful?

Panel data are data where the same observation is followed over time (like in time series) and where there are many observations (like in cross-sectional data). In this sense, panel data combine the features of both time-series and cross-sectional data and methods.

Following a popular convention, we will call N the number of observations (or the cross-sectional dimension of the data) and T the number of time periods (or the time or longitudinal dimension of the data). Although it is customary to think of the longitudinal dimension of the dataset along the time horizon, panel data are essentially all those datasets that span (at least) two dimensions. Usually these are (i) a cross section of observation (ii) over a sequence of time periods, but all other options are possible: a set of N firms each having T establishments in different locations; a set of N countries each having T regions, and so on. For simplicity, in our discussion we will keep thinking about a cross section of N units observed repeatedly over T periods.

Generally speaking, there exist two type of panel datasets. *Macro panels* are characterised by having a relatively large T and a relatively small N. A typical example is a panel of countries where the variables are macro data like the unemployment rate or the growth rate. These variables are usually available for a long period of time and for a limited set of countries. *Micro panels*, instead, usually cover a large set of units N for a relatively short number of periods T. Here, the typical example is a survey sample (similar to the ones that we have seen in the previous parts of the course) that is repeated over time by interviewing the same persons several times.

Another important classification is between *balanced* and *unbalanced* panels. A balanced dataset is one in which all the N observations are followed for the same number of periods T. In an unbalanced dataset each observation might be available for a different number of periods so that the time dimension is potentially different for different observations.

For reasons that will become clear later, the treatment of macro-panels is somewhat easier than that of micro-panels. In fact, most of the techniques that are specific for panel data have been developed in the context of microeconometrics, although they are now common practice among the macro-econometricians as well.

The availability of more and more panel datasets has been welcomed with great enthusiasm by all kinds of social scientists, from economists to demographers, from sociologists to political scientists. The reasons for such enthusiasm can be broadly grouped into two uses of these data. First, panel

data offer a very effective solution to one of the most frequent and complicated identification problem in empirical studies, that is the presence of *unobservable omitted variables*. As we have seen in the previous parts of this course, omitted variables can easily lead to inconsistency if they are correlated with any of the explanatory variables of the model. This problem can be easily solved if the omitted variables can be observed and one can get data on them. In which case it is sufficient to explicitly include them in the set of explanatory variables. Obviously, the problem is much more serious if the omitted variable is not observable and therefore it cannot be included among the regressors. As we will see later on, panel data offer a very simple and effective solution to unobservable omitted variables in all cases in which such unobservables are constant over time (or over the longitudinal dimension of the data).

The second use of panel data that makes them so precious to all social scientists is the study of the *dynamics of cross-sectional processes*. For example, if one wants to study the determinants of transitions in and out of poverty or in and out of unemployment one needs panel data to identify who transits from which state to which, from poverty to non-poverty and vice-versa or from unemployment to employment or to inactivity. In a simple cross-section such processes cannot be studied for the simple fact that they cannot be observed.

In our presentation of the econometrics of panel data we are going to focus almost exclusively on balanced micro panels with N large and T small and on methods to address the problem of time-invariant unobservable heterogeneity. We will also concentrate only on linear models.

10 The static linear model with unobserved heterogeneity: introduction and intuition

Let us start our discussion with a simple linear model of the following type:

$$y = X\beta + \eta + v \tag{10.1}$$

that satisfies the following assumption:

$$E(v|X, \eta) = 0 \tag{10.2}$$

This is a model that is very similar to the simplest linear model that we saw in the first part of this course with the only addition of the term η. Assumption 10.2 guarantees that, if both X and η were observable and

included in the specification of the model as explanatory variables, we could obtain consistent estimates of the parameters β by mere application of OLS.

The problem arises because η is unobservable. Then, the only thing we can do is regress y on X and leave η in the error term:

$$y = X\beta + (\eta + v) = X\beta + u \tag{10.3}$$

where $u = \eta + v$. In this regression the correct identification of β is impossible if $Cov(u, X) \neq 0$. Given assumption 10.2, such correlation can be non-zero only due to correlation between η and X.

In other words, if the unobserved (and, therefore, omitted) term η is correlated with some of the other regressors X, then identification of β via simple OLS is impossible.

As we have discussed at length in previous parts of the course, one common solution to the omitted variable problem is instrumental variables. We would have to find a variable z that is uncorrelated with the error term (both with η and with v) and at the same time correlated with the endogenous variable(s) X. However, we also know that finding such instrumental variable(s) can be very difficult.

The beauty of panel data is that they offer a very simple and convenient alternative solution. Let us look briefly at the intuition here and then do things more formally in the next section.

Consider the same model of equation 10.1 but now assume that we have data for the same process in two different time periods:

$$y_1 = X_1\beta + \eta + v_1 \tag{10.4}$$

$$y_2 = X_2\beta + \eta + v_2 \tag{10.5}$$

where we have assumed that η is constant over time (in fact, it does not have a subscript) and the coefficients β are also constant over time. Now, transform the model by taking the difference between the two periods:

$$\begin{aligned} \Delta y &= y_2 - y_1 = (X_2 - X_1)\beta + (v_2 - v_1) \\ &= \Delta X\beta + \Delta v \end{aligned} \tag{10.6}$$

The key trick is that by taking differences we have eliminated the unobserved heterogeneity term η from the model. And the new differenced equation 10.6 satisfies all the necessary assumptions to guarantee consistency of the OLS estimator. In particular, given assumption 10.2, $E(\Delta v) = 0$ and $Cov(\Delta v, \Delta X) = 0$. Thus, we can simply apply OLS to the differenced equation and obtain consistent estimates of β.

This is in essence the intuition of how panel data help addressing the problem of unobserved heterogeneity. Before moving to a more formal discussion of the methodology, let us make a few additional comments.

First of all, it is important to notice that the idea of taking first differences of a given cross-sectional model over time periods may easily backfire. In particular, this approach allows to produce estimates of the coefficients only of those regressors that vary over time. Suppose that one or more of the variables in X were constant over time (like gender in a simple individual level regression), then when we take differences these observable regressors will be canceled out, just like η. As a consequence, we will not be able to estimate the coefficients on those variables.

The second important comment concerns the purpose of our analysis. In most of this course, we have maintained the assumption that the main purpose of the analysis is the consistent estimation of the parameters of the model. However, another common reason to do econometrics is the production of reliable forecasts of the dependent variable outside the estimation sample. By taking differences we are able to control for unobserved heterogeneity but this comes at the cost of dropping from the model all the explanatory variables that are fixed over time (as well as the unobserved heterogeneity term). And these time-invariant explanatory variables may, in fact, explain a lot of the variation of the dependent variable so that prediction based exclusively on those regressors that vary over time might be very imprecise.

Finally, note that the key condition for the consistency of OLS on the differenced equation 10.6 is $Cov(\Delta v, \Delta X) = 0$, which requires a particular interpretation of assumption 10.2 to be valid. Taken literally, assumption 10.2 simply states that X_1 is uncorrelated with v_1 and X_2 is uncorrelated with v_2 or, more precisely, that $E(v_1|X_1) = 0$ and $E(v_2|X_2) = 0$. For consistency of OLS on the differenced equation, we need slightly more than that. In particular, we need X_1 to be uncorrelated with both v_1 and v_2 and similarly, also X_2 to be uncorrelated with both v_1 and v_2.[1] When the explanatory variables satisfy these conditions, i.e. when they are uncorrelated with the error term at any time, they are said to be *strictly exogenous*.

In the following sections we are going to see more formally how to produce consistent and efficient estimators of the parameters of a linear model with unobserved heterogeneity. Section 11 considers the case in which such heterogeneity (the η term in our example above) is correlated with one or more of the observable regressors, while in section 12 we look at the alterna-

[1] Or, more precisely, $E(v_1|X_1, X_2) = 0$ and $E(v_2|X_1, X_2) = 0$.

tive situation when the unobserved heterogeneity term is uncorrelated with all other observable regressors.

11 Fixed-effects estimation

In this section we review four different methodologies that produce exactly the same consistent estimator of the parameters of a linear model with unobserved heterogeneity that is consistent even when such heterogeneity is correlated with one or more of the observable regressors. Such estimator is usually called *fixed-effects estimator* and the four methods are first-differences (section 11.1), deviations from the mean (section 11.2), orthogonal deviations (section 11.3) and dummy variable estimation (section 11.4).

11.1 The first-difference estimator

Let us now address more formally the construction of a consistent estimator of the parameters of a model similar to the one in equation 10.1 by means of taking differences between the observations of two adjacent periods. Our unit of observation will be a vector of y's and x's for a given individual i at different times t:

$$\{y_{i1}, y_{i2}, \ldots, y_{iT}, x_{i1}, x_{i2}, \ldots, x_{iT}\} \text{ with } i = 1, \ldots, N$$

The most general specification of the model that we consider is the following:

$$y_{it} = x_{it}\beta + \eta_i + v_{it} \quad \forall i = 1, \ldots, N \text{ and } \forall t = 1, \ldots, T \tag{11.1}$$

where for notational simplicity we indicate with x_{it} a generic vector of K time-varying explanatory variables for individual i at time t. It is also convenient to rewrite the model in stacked format as follows:

$$\begin{aligned} y_i &= X_i\beta + \eta_i \cdot \iota_T + v_i \\ \underset{T\times 1}{\begin{pmatrix} y_{i1} \\ \vdots \\ y_{iT} \end{pmatrix}} &= \underset{T\times K}{\begin{pmatrix} x_{i1} \\ \vdots \\ x_{iT} \end{pmatrix}} \underset{K\times 1}{\begin{pmatrix} \beta_1 \\ \vdots \\ \beta_K \end{pmatrix}} + \underset{T\times 1}{\begin{pmatrix} \eta_i \\ \vdots \\ \eta_i \end{pmatrix}} + \underset{T\times 1}{\begin{pmatrix} v_{i1} \\ \vdots \\ v_{iT} \end{pmatrix}} \end{aligned} \tag{11.2}$$

where we indicated with ι_T a simple column vector of T ones, $\iota_T = (1, \ldots, 1)'$. The specification of the model in 11.2 is useful to state succinctly the key identification assumption:

$$E(v_i | X_i, \eta_i) = 0 \tag{11.3}$$

which implies that X_i is *strictly exogenous* and η_i is uncorrelated with the random component of the model v_i. For the sake of simplicity, we also make an assumption about the conditional variance-covariance matrix of the error term:

$$Var(v_i|X_i, v_i) = \sigma^2 I_T = \begin{pmatrix} \sigma^2 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma^2 \end{pmatrix} \tag{11.4}$$

Suppose for generality that $T \geq 3$, then the model transformed in first-differences (i.e. differences taken between each time observation and its first adjacent time observation) can be written as follows:

$$\begin{aligned} \Delta y_{i2} &= \Delta x_{i2}\beta + \Delta v_{i2} \\ &\vdots \\ \Delta y_{iT} &= \Delta x_{iT}\beta + \Delta v_{iT} \end{aligned} \tag{11.5}$$

which can also be written in compact notation as:

$$Dy_i = DX_i + Dv_i \tag{11.6}$$

where D is a *transformation matrix* that transforms the model from levels into first differences:

$$\underset{(T-1)\times T}{D} = \begin{pmatrix} -1 & 1 & 0 & \dots & \dots & 0 \\ 0 & -1 & 1 & 0 & \dots & 0 \\ \vdots & & \ddots & \ddots & & 0 \\ 0 & \dots & \dots & \dots & -1 & 1 \end{pmatrix} \tag{11.7}$$

Notice that, while the original model has T equations for each individual i, the first-differenced model looses one equation, as for the first time observation it is impossible to take first difference. The differenced model, thus, has only $T - 1$ equations.

A consistent estimator of β can be obtained by simply applying OLS to the model in equation 11.6:

$$\begin{aligned} \widehat{\beta} &= \left[\sum_{i=1}^{N} (DX_i)' DX_i\right]^{-1} \left[\sum_{i=1}^{N} (DX_i)' Dy_i\right] \\ &= \left[\sum_{i=1}^{N} X_i' D' DX_i\right]^{-1} \left[\sum_{i=1}^{N} X_i' D' Dy_i\right] \end{aligned} \tag{11.8}$$

Under the assumptions we made so far, this estimator is certainly consistent. However, it is not efficient.

In fact, even if the variance-covariance matrix of the errors in levels (the v_i's) is assumed to be *scalar*, taking first-differences induces serial correlation.[2] To see that, notice that two generic adjacent first-differenced errors can be written as follows:

$$\begin{aligned} \Delta v_{it} &= v_{it} - v_{it-1} \\ \Delta v_{it-1} &= v_{it-1} - v_{it-2} \end{aligned}$$

Both these errors have the same variance:

$$Var(\Delta v_{it}) = Var(\Delta v_{it-1}) = 2\sigma^2$$

so that homoskedasticity is maintained in the first-differenced model. However, the two errors are obviously correlated with each other:

$$\begin{aligned} Cov\left(\Delta v_{it}, \Delta v_{it-1}\right) &= Cov(v_{it} - v_{it-1}, v_{it-1} - v_{it-2}) \\ &= -\sigma^2 \end{aligned}$$

while all other higher order covariances are zero. The variance-covariance matrix of the new first-differenced errors can then be written as:

$$Var(DV_i|X_i) = \sigma^2 DD' = \sigma^2 \underbrace{\begin{pmatrix} 2 & -1 & 0 & \dots & 0 \\ -1 & 2 & -1 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & -1 & 2 & -1 \\ 0 & \dots & 0 & -1 & 2 \end{pmatrix}}_{(T-1)\times(T-1)} \tag{11.9}$$

which is a symmetric matrix with all 2's on the main diagonal, -1's one position off the main diagonal on both sides and all other elements equal to zero.

We have now discovered that the model in first-differences is affected by serial correlation. As we know, this implies that the OLS estimator, although still consistent, is not efficient. In other words, it is possible to construct another estimator that is always consistent but more efficient than OLS. We also know how to produce this estimator. The methodology that allows to produce consistent and efficient estimators in the presence of serial

[2] Remember the definition of a *scalar* matrix, that is a diagonal matrix with all identical elements on the main diagonal

correlation is GLS. The GLS estimator of the first-differenced model can thus be written as follows:

$$\widehat{\beta}_{FD} = \left[\sum_{i=1}^{N} X_i' D'(DD')^{-1} D X_i\right]^{-1} \left[\sum_{i=1}^{N} X_i' D'(DD')^{-1} D y_i\right] \tag{11.10}$$

We will call this estimator the *first-difference estimator* and indicate it with $\widehat{\beta}_{FD}$. The first-differences estimator is thus simply GLS applied on the model in first differences. Note in particular, that $\widehat{\beta}_{FD}$ is a *pure GLS* estimator and not a *feasible GLS*, as it is usually the case. As you probably remember, GLS is usually impossible to apply directly because it requires knowledge of the specific form of the variance-covariance matrix of the error term, which we called Ω in our first discussion of the GLS method. In this specific case Ω is $D'D$, a simple matrix of 2's and -1's. Hence, under homoskedasticity of the model in levels, we can easily derive the exact form of the variance-covariance matrix of the model in first differences and thus apply GLS directly with no need to compute its feasible version.

11.2 The within-group or deviations-from-the-means estimator

Take a closer look at the formula of the first-difference estimator in equation 11.10 and focus your attention on the term $D'(DD')^{-1}D$. Call this $T \times T$ matrix $Q = D'(DD')^{-1}D$ and rewrite the first-difference estimator as:

$$\widehat{\beta}_{FD} = \left[\sum_{i=1}^{N} X_i' Q X_i\right]^{-1} \left[\sum_{i=1}^{N} X_i' Q y_i\right] \tag{11.11}$$

It is relatively easy to show that Q can be rewritten as follows:

$$Q = D'(DD')^{-1}D = I_T - \frac{1}{T}\iota_T \iota_T' = \begin{pmatrix} 1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 1 \end{pmatrix} - \frac{1}{T}\begin{pmatrix} 1 & \dots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \dots & 1 \end{pmatrix}$$

Notice that Q can be interpreted as a new transformation matrix that transforms each observation of the model in levels into its deviation from the within-individual time average:

$$Qy_i = \begin{pmatrix} y_{i1} - \frac{1}{T}\sum_{t=1}^{T} y_{it} \\ y_{i2} - \frac{1}{T}\sum_{t=1}^{T} y_{it} \\ \vdots \\ y_{iT} - \frac{1}{T}\sum_{t=1}^{T} y_{it} \end{pmatrix}$$

Similarly, the entire model transformed by the Q matrix can be written as follows:

$$\begin{aligned} Qy_i &= QX_i\beta + Qv_i \\ (y_{i1} - \overline{y}_i) &= (x_{i1} - \overline{x}_i)\beta + (v_{i1} - \overline{v}_i) \\ (y_{i2} - \overline{y}_i) &= (x_{i2} - \overline{x}_i)\beta + (v_{i2} - \overline{v}_i) \\ \vdots &\ \vdots\ \vdots \\ (y_{iT} - \overline{y}_i) &= (x_{iT} - \overline{x}_i)\beta + (v_{iT} - \overline{v}_i) \end{aligned} \tag{11.12}$$

where $\overline{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{it}$, $\overline{x}_i = \frac{1}{T}\sum_{t=1}^{T} x_{it}$ and $\overline{v}_i = \frac{1}{T}\sum_{t=1}^{T} v_{it}$.

What is important to notice is that also this transformation eliminates the unobserved heterogeneity term η from the model. In fact, the time average of η_i is η_i itself and therefore, once taking deviations, it simply cancels away. Additionally, under the original assumption 11.3 the resulting transformed model of equation 11.12 satisfies the OLS requirements for consistency. Before computing the OLS estimator of model 11.12, however, notice that Q has a couple of interesting properties. First of all, it is symmetric:

$$Q' = D'(DD')^{-1}D = Q$$

Additionally, it is an *idempotent* matrix, i.e. a matrix that remains equal to itself if it is multiplied by itself:

$$Q'Q = D'(DD')^{-1}DD'(DD')^{-1}D = D'(DD')^{-1}D = Q$$

With these results we can easily compute the OLS estimator on the model in deviations from the time averages and show that it is equal to the first-difference estimator:

$$\begin{aligned} \widehat{\beta}_{WG} &= \left[\sum_{i=1}^{N} X_i'Q'QX_i\right]^{-1}\left[\sum_{i=1}^{N} X_i'Q'Qy_i\right] \\ &= \left[\sum_{i=1}^{N} X_i'QX_i\right]^{-1}\left[\sum_{i=1}^{N} X_i'Qy_i\right] \end{aligned} \tag{11.13}$$

This estimator is called the *within-group* estimator and it is numerically identical to the first-difference estimator (recall that $Q = D'(DD')^{-1}D$).[3]

[3]The name 'within-group' comes from the fact that this estimator is constructed using only variation in both the dependent and the independent variables within individuals over time. Later in section 13 we will see a 'between-group' estimator that instead uses only variation across individuals.

The result that $\widehat{\beta}_{FD} = \widehat{\beta}_{WG}$ should come as a big surprise to you. $\widehat{\beta}_{FD}$ is produced as GLS on a model of $T-1$ equations and it is efficient. $\widehat{\beta}_{WG}$ is produced as OLS on a model of T equations and, at least at first sight, it should not be efficient. In fact, similarly to first-differencing, taking deviations from the time averages eliminates η from the model but at the same time induces serial correlation in the model. Under homoskedasticity of the model in levels, the variance-covariance matrix of the model in deviations from the means can be written as follows:

$$Var(Qv_i) = QVar(v_i)Q = \sigma^2 QQ' = \sigma^2 Q \tag{11.14}$$

which is not a scalar matrix. Hence, $\widehat{\beta}_{WG}$ should not be efficient because the most efficient linear estimator of β should be the GLS estimator on the model in deviations from the means. However, $\widehat{\beta}_{WG}$ is identical to $\widehat{\beta}_{FD}$, which is efficient. The solution to this puzzle is easy to see. Let us compute the GLS estimator on the model in deviations from the means, which should be the efficient one:

$$\begin{aligned} \widehat{\beta}_{GLS} &= \left[\sum_{i=1}^{N} X_i' Q'(Q)^{-1} Q X_i\right]^{-1} \left[\sum_{i=1}^{N} X_i' Q'(Q)^{-1} Q y_i\right] \\ &= \left[\sum_{i=1}^{N} X_i' Q X_i\right]^{-1} \left[\sum_{i=1}^{N} X_i' Q y_i\right] = \widehat{\beta}_{WG} \end{aligned} \tag{11.15}$$

In other words, given the particular structure of the model in deviations from the means, the OLS and the GLS estimators of that model coincide.

Finally, note that when you ask Stata to estimate a fixed-effects model with the command *xtreg* it effectively applies this model, i.e. OLS on the model transformed in deviations from the individual time means.

11.3 The orthogonal-deviations estimator

Go back to the first-differences estimator of equation 11.10. Recall that it was computed as GLS on the model in first-differences. Also recall that GLS is nothing but OLS applied on a transformed version of the model that recovers serial independence of the errors.

What is, then, the transformation that we implicitly applied to the model in first differences when we did GLS? If you think about our old discussion of generalized least squares, you should remember that such transformation was simply the variance-covariance matrix of the errors to the power of

$-1/2$. And the variance-covariance matrix of the model in first-differences is DD', as shown in equation 11.9.[4]

To summarize, to produce the first-differences estimator we have first transformed the original model in levels using the matrix D, then transformed the resulting model again using the GLS transformation $(DD')^{-1/2}$ and finally applied OLS. Hence, the overall transformation that gets us from the original model in levels to the one that we use for the OLS estimation is $(DD')^{-1/2}D$. Let us call this matrix $A = (DD')^{-1/2}D$.

With a bit of patience and after a bit of algebra, it is possible to show that:

$$Ay_i = \begin{cases} \left[y_{i1} - \frac{1}{T-1}\left(y_{i2} + \cdots + y_{iT}\right)\right] c_1 \\ \left[y_{i2} - \frac{1}{T-2}\left(y_{i3} + \cdots + y_{iT}\right)\right] c_2 \\ \vdots \\ \left[y_{iT-1} - y_{iT}\right] c_{T-1} \end{cases}$$

where $c_t^2 = \frac{T-t}{T-t+1}$. Matrix A essentially transforms the original model in levels into *forward orthogonal deviations*, i.e. into differences with the average of all future values, adjusted by a constant term c_t that guarantees homoskedasticity.

In other words, an alternative way of producing the fixed-effects estimator is applying OLS to the model transformed in forward orthogonal deviations, a transformation that gets rid of the unobserved heterogeneity term η_i and at the same time avoids introducing serial correlation.[5] In fact, the variance covariance matrix of the model transformed in forward orthogonal deviations is scalar:

$$Var(Av_i) = \sigma^2(DD')^{-1/2}DD'(DD')^{-1/2} = \sigma^2 I_{T-1}$$

11.4 The dummy-variable estimator

The fourth (and last) method to obtain the fixed-effects estimator of β approaches the problem of unobserved heterogeneity rather differently from the previous methods. Rather than thinking about fancy transformations of the model that get rid of η_i, now we try and simply estimate it. In fact,

[4] More correctly, The variance-covariance matrix of the model in first differences is $\sigma^2 DD'$. However, if in the formula for an estimator we multiply both the term powered to the -1 and the term powered to 1 by the scalar σ^2 it will obviously cancel away.

[5] The same result can be obtained with *backward orthogonal deviations*, that is differences with the mean of previous values.

one alternative way of thinking about η_i is as a coefficient to be estimated rather than as an unobserved term to be eliminated.

Go back to the original specification of the model:

$$y_i = X_i\beta + \iota_T\eta_i + v_i$$

and now stack again all the N observations together into a unique model:

$$\begin{array}{rcl} y & = & X\beta + C\eta + v \\ \underset{NT\times 1}{\begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}} & = & \underset{NT\times K}{\begin{pmatrix} X_1 \\ \vdots \\ X_N \end{pmatrix}} \underset{K\times 1}{\beta} + \underset{NT\times N}{\begin{pmatrix} \iota_T & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \iota_T \end{pmatrix}} \underset{N\times 1}{\begin{pmatrix} \eta_1 \\ \vdots \\ \eta_N \end{pmatrix}} + \underset{NT\times 1}{\begin{pmatrix} v_1 \\ \vdots \\ v_N \end{pmatrix}} \end{array} \tag{11.16}$$

In this specification, the matrix

$$C = \begin{pmatrix} \iota_T & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \iota_T \end{pmatrix}$$

contains a set of dummies, one for each individual i in the sample while the vector

$$\eta = \begin{pmatrix} \eta_1 \\ \vdots \\ \eta_N \end{pmatrix}$$

is the vector of coefficients of those dummies. And just like the other coefficients of the model, the vector η can be estimated. Additionally, once the model is interpreted in this fashion, there is no identification problem to be fixed. There is no unobserved heterogeneity term that is omitted and left in the error term. So, we can get nice consistent and efficient estimates of both β and η by applying OLS to the model in equation 11.16.

To do that, it is easier to rewrite the model as follows:

$$y = \begin{pmatrix} X & C \end{pmatrix} \begin{pmatrix} \beta \\ \eta \end{pmatrix} + v = W\delta + v \tag{11.17}$$

where $W = \begin{pmatrix} X & C \end{pmatrix}$ is the matrix of regressors and $\delta = \begin{pmatrix} \beta \\ \eta \end{pmatrix}$ the vector of coefficients. Then, the OLS estimator of δ is simply:

$$\widehat{\delta} = \widehat{\begin{pmatrix} \beta \\ \eta \end{pmatrix}} = (W'W)^{-1}(W'y) \tag{11.18}$$

To check whether this method produces the same estimates of β of the previous methods (first-differences, deviations from the means and orthogonal deviations), we need to extract from $\widehat{\delta}$, which is an estimator of the vector $(\beta\ \eta)'$ *(@@ I would write this as* $(\beta'\ \eta')'$*)*, only the estimates of β. How can we do that? It is a relatively easy thing to do if you recall something from previous econometrics courses, namely *partitioned regression*.[6]

Using some simple results from partitioned regression and after some manipulations, it can be shown that:

$$\begin{aligned}\widehat{\beta}_{DV} &= \left[X'\left(I_{NT} - C(C'C)^{-1}C'\right)X\right]^{-1}\left[X'\left(I_{NT} - C(C'C)^{-1}C'\right)y\right] \\ &= \left[X'\left(I_{NT}\otimes Q\right)X\right]^{-1}\left[X'\left(I_{NT}\otimes Q\right)y\right] \qquad (11.19)\end{aligned}$$

where $\otimes$ indicates the *Kronecker product*.[7] After some additional algebra, it is also possible to show that:

$$\begin{aligned}\widehat{\beta}_{DV} &= \left[X'\left(I_{NT}\otimes Q\right)X\right]^{-1}\left[X'\left(I_{NT}\otimes Q\right)y\right] \\ &= \left[\sum_{i=1}^{N} X_i'QX_i\right]^{-1}\left[\sum_{i=1}^{N} X_i'Qy_i\right] \qquad (11.20)\end{aligned}$$

which is the within-group estimator as we defined it in equation 11.13 and, as a consequence, it is also equal to $\widehat{\beta}_{FD}$ and $\widehat{\beta}_{OD}$, where $\widehat{\beta}_{OD}$ is the orthogonal-deviations estimator.

Notice that, unlike all previous methods, now we also have estimates of the individual effects, which can be written as follows:

$$\widehat{\eta}_i = \frac{1}{T}\sum_{t=1}^{T}\left(y_{it} - x_{it}\widehat{\beta}_{DV}\right) \qquad (11.21)$$

Equation 11.21 clearly shows that the $\widehat{\eta}_i$'s are estimated using exclusively the time variation within individuals. And it could not be otherwise, given that they are estimates of effects that are fixed over time for each single individual. The smaller T, the fewer observations would be available for the estimation of the η_i's and the less precise such estimates will be. In the limit, if $T = 1$, then, the η_i's cannot be estimated. Moreover, this also implies that their asymptotic properties are valid as T tends to infinity while

[6] We are not going to revise partitioned regression here.

[7] The Kronecker product is an ugly matrix operator that we are not going to review here in details. It essentially multiplies each element of a matrix for the entire other matrix, thus resulting in a block matrix. I can guarantee that it is not particular fun (nor particularly instructive), but you can try and show that, indeed, $I_{NT} - C(C'C)^{-1}C' = I_{NT}\otimes Q$.

all the estimators of β that we have seen so far are consistent as NT goes to infinity. In fact, NT is the size of the actual sample that is used to estimate them.

To some of you this last dummy-variable method might seem the simplest and most straightforward of all. So, why is it that we had to come up with all the fancy transformations of the model that we discussed above? In the end, they all made our lives miserable because they were more or less difficult to handle and allowed us to estimate only β. This dummy-variable thing is simpler and gives more results: estimates of both β and η.

The problem with the dummy-variable estimator is very simple to spot and still very pervasive. In micro applications N is large and T is small. And since there is one η for each $i = 1, \ldots, N$, this means that there will be many η's to be estimated, often in the order of thousands, hundreds of thousands or more. Estimating a model with that many parameters is simply not feasible from a computational point of view. Additionally, if T is small, the resulting estimates will be very imprecise. Thus, it seems like a good idea to save computing time and power, forget about estimating the η's and transform the model in order to get rid of them.

If N is small and T is large, however, the dummy variable estimator is a very appealing option. And, in fact, macro econometricians use it all the time. This also explains why it has been mostly the micro-econometricians who developed specific methods to deal with panel data. The macro guys solved the problem of unobserved heterogeneity much more easily by simply plugging into their models a few dummies and estimating them by standard methods.

11.5 Estimating σ^2 in panel data

In the previous sections we have discussed different methodologies to produce the fixed-effects estimator. However, we do not know yet how to compute its variance-covariance matrix. Notice that the variance-covariance matrix of $\widehat{\beta}_{FE}$ can be written as:[8,9]

$$Var(\widehat{\beta}_{FE}) = \sigma^2 \left(\sum_{i=1}^{N} X_i' Q X_i \right)^{-1} \tag{11.22}$$

[8] For the analysis in this section we will work under the assumption that $\widehat{\beta}_{FE}$ has been constructed using the method of deviations from the means.

[9] We will generically indicate the fixed-effects estimator as $\widehat{\beta}_{FE}$ regardless of the method used to produce it.

To be able to compute such variance-covariance matrix, we need to produce an estimate of σ^2, the variance of the error term of the model. All the other elements of equation 11.22 are known. Since $\widehat{\beta}_{FE}$ is essentialy a simple OLS estimator on the model in deviations from the means, the most intuitive estimator of σ^2 is perhaps the variance of the residuals of the model in deviations from the mean. By mere application of the analogy principle, such estimator will converge in probability to the variance of the model in deviations from the means. Unfortunately, however, the variance of the model in deviations from the means is not σ^2:[10]

$$\begin{aligned} Var\left(\widetilde{v}_{it}\right) &= Var\left(v_{it}-\overline{v}_i\right) = E\left[\left(v_{it}-\frac{1}{T}\sum_{s=1}^{T}v_{is}\right)^2\right] = \\ &= \sigma^2+\frac{1}{T^2}T\sigma^2-2E\left(v_{it}\frac{1}{T}\sum_{s=1}^{T}v_{is}\right) \\ &= \sigma^2+\frac{\sigma^2}{T}-\frac{2}{T}\sigma^2=\sigma^2\frac{T-1}{T} \end{aligned} \tag{11.23}$$

Equation 11.23 shows that the variance of error term in deviations from the means is $\sigma^2\frac{T-1}{T}$ and not just σ^2. Thus, the variance of the estimated residuals $\widehat{\widetilde{v}}_{it}$, where

$$\widehat{\widetilde{v}}_{it}=\widetilde{y}_{it}-\widetilde{x}_{it}\widehat{\beta}_{FE} \tag{11.24}$$

converges in probability to $\sigma^2\frac{T-1}{T}$ and it is therefore not a consistent estimator of σ^2:[11]

$$\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{T}\sum_{t=1}^{T}\widehat{\widetilde{v}}_{it}^2\right)\xrightarrow{p}\frac{T-1}{T}\sigma^2 \tag{11.25}$$

The problem is easily solved by adjusting the inconsistent estimator of equation 11.25 as follows:

$$\frac{1}{N(T-1)}\sum_{i=1}^{N}\sum_{t=1}^{T}\widehat{\widetilde{v}}_{it}^2\xrightarrow{p}\sigma^2 \tag{11.26}$$

The formula in equation 11.26 suggests a couple of additional comments. First of all, it seems like we were taking an average over NT observations

[10] Let us indicate with $\widetilde{-}$ the variables in deviations from the means.

[11] Notice that as $T\to\infty$ also $\sigma^2\frac{T-1}{T}\to\sigma^2$ and $\left(\frac{1}{T}\sum_{t=1}^{T}\widehat{\widetilde{v}}_{it}^2\right)$ becomes a consistent estimator of σ^2. Here, we are mostly interested in estimators whose asymptotic properties are valid as $NT\to\infty$, regardless of whether it is N or T that grows larger. Generally, the largest dimension of a typical micro-dataset is N.

but we effectively divide the sum of such NT elements by $N(T-1)$. If you think about it, however, when you run OLS on the model in deviations from the means you are using NT observations but one for each individual is totally redundant. In fact, both the method of first-differences and of orthogonal deviations use $T-1$ observations for each individual i and all three methods produce exactly the same estimator. In deviations from the means one observations for each individual i is redundant because it contains no additional information. To understand this simple point, consider a set of three numbers (e.g. 3,7 and 8) and suppose that you are told only two of them and their average (e.g. 3 and 8 and the average, which is 6). Then, you could univocally reconstruct what is the third number in the sequence (e.g. $6 \times 3 - 3 - 8 = 7$). Similarly, in the model in deviations from the means one time observation for each individual is redundant because, once we know the other $T-1$, the Tth observation does not contain any additional information and can be fully reconstructed from the others.[12] In the dummy variable model things are only slightly different. It is true that in that case we use all NT observations and that all of them are necessary but it is also true that in that model we need to estimate N additional parameters, i.e. one fixed-effect for each individual, and $NT - N = N(T-1)$.

The second important comment is more practical. All modern statistical packages have built-in routines to do fixed-effects estimation and such routines obviously produce the correct standard errors. However, if for some reason you have to run fixed-effect manually, for example by applying OLS to a model that you have previously transformed in deviations from the individual time means, the resulting standard errors will be inconsistent because they will be produced using a standard OLS estimator of σ, i.e. the one in equation 11.25, and not the correct one of equation 11.26. The smaller T, the larger the bias.

12 Random-Effects estimation

In this section we discuss how to estimate consistently and efficiently the parameters of the model in the less problematic case when the unobserved heterogeneity term is **uncorrelated** with all other observable regressors. The most efficient estimator under such assumption is a simple GLS on the model in levels and it is normally called *random-effects estimator*.

[12] In fact, as we will see in the computer room, if you eliminate one time observation for each individual in a panel you obtain exactly the same fixed-effect estimator as with the entire sample.

Let us proceed with order and consider the same model of section 11.1:

$$\begin{aligned} y_{it} &= x_{it}\beta + \eta_i + v_{it} \\ &= x_{it}\beta + u_{it} \quad \forall i = 1, \ldots, N \text{ and } \forall t = 1, \ldots, T \end{aligned} \tag{12.1}$$

where $u_{it} = \eta_i + v_{it}$ and where we make the following assumption:

$$E(u_{it}|x_{it}) = 0 \tag{12.2}$$

Assumption 12.2 implies that the unobserved heterogeneity term η_i is uncorrelated with both v_{it}, like in fixed-effects models, and with x_{it}. Hence, omitting it from the specification of the model and leaving it in the error term, as in equation 12.1, does not impede identification of β via OLS.

The OLS estimator of β from the model in equation 12.1 is the following:

$$\widehat{\beta}_{POLS} = \left[\sum_{i=1}^{N} X_i' X_i\right]^{-1} \left[\sum_{i=1}^{N} X_i' y_i\right] \tag{12.3}$$

where $X_i = (x_{i1} \; x_{i2} \; \ldots \; x_{iT})'$ and $y_i = (y_{i1} \; y_{i2} \; \ldots \; y_{iT})'$. We call this estimator *pooled OLS (POLS)* and it is totally analogous to the system OLS (SOLS) estimator that we discussed in the first part of the course. In SOLS there were several equations for the same observation and also here there are several equations for the same observation with the only small difference that now each equation is nothing but the same identical model repeated several times at different time points.

But, is $\widehat{\beta}_{POLS}$ the best possible estimator we could produce on this model under assumption 12.2? Obviously not. In fact, even under the simplest possible assumption about the variance-covariance matrix of the error term, the model is necessarily affected by serial correlation. η_i appears in all the error terms of the same individual i at all times, therefore the error terms of the different time observations of the same individual will necessarily be correlated.

To see this more formally, make the simplest possible assumptions about the error terms:

$$Var(v_{it}) = \sigma_v^2 \quad \forall i \text{ and } \forall t \tag{12.4}$$

$$Cov(v_{it}, v_{is}) = 0 \quad \forall i \text{ and } \forall t \neq s \tag{12.5}$$

$$Var(\eta_i) = \sigma_\eta^2 \quad \forall i \tag{12.6}$$

and then notice that:

$$\begin{aligned} Var(u_{it}) &= \sigma_\eta^2 + \sigma_v^2 \quad \forall i \text{ and } \forall t \\ Cov(u_{it}, u_{is}) &= \sigma_\eta^2 \quad \forall i \text{ and } \forall t \neq s \end{aligned}$$

or, more compactly:

$$\begin{aligned} Var(u_i) &= E(u_i u_i') = E\left[\begin{pmatrix} u_{i1} \\ u_{i2} \\ \vdots \\ u_{iT} \end{pmatrix} (u_{i1}\ u_{i2} \dots u_{iT}) \right] \\ &= \begin{bmatrix} (\sigma_\eta^2 + \sigma_v^2) & \dots & \sigma_\eta^2 \\ \vdots & \ddots & \vdots \\ \sigma_\eta^2 & \dots & (\sigma_\eta^2 + \sigma_v^2) \end{bmatrix} = \Omega \end{aligned} \tag{12.7}$$

Let us call this matrix Ω.

We know that in these cases OLS still produces consistent estimates (so $\widehat{\beta}_{POLS}$ is consistent) but GLS is the most efficient estimation method. The GLS estimator of model 12.1 is called *random effects estimator* and it can be written as follows:

$$\widehat{\beta}_{RE} = \left[\sum_{i=1}^{N} X_i' \Omega^{-1} X_i \right]^{-1} \left[\sum_{i=1}^{N} X_i' \Omega^{-1} y_i \right] \tag{12.8}$$

Notice, however, that, contrary to the GLS estimator that we produced on the model in first differences in section 11.1, the estimator in equation 12.8 is not feasible, meaning that without knowledge of Ω we cannot compute it. And Ω is a symmetric matrix whose elements on the main diagonal are all identical and equal to $\sigma_\eta^2 + \sigma_v^2$ and all the elements off the main diagonal are equal to σ_η^2. Hence, to make the estimator feasible we need to construct consistent estimators of σ_v^2 and σ_η^2. This is where different types of random effect estimators differ, depending on how these estimated Ω is constructed. Here we are only going to see one solution, the most popular that is known as the *Balestra-Nerlove estimator*.

The Balestra-Nerlove random-effects estimator uses the following estimator of σ_v^2:

$$\widehat{\sigma}_v^2 = \frac{1}{N(T-1)} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(\widetilde{y}_{it} - \widetilde{x}_{it} \widehat{\beta}_{FE} \right)^2 \tag{12.9}$$

which is exactly the same estimator of the variance of the random error term that we discussed in section 11.5.

As for the variance of η, the most intuitive way to do it would be using the estimated $\widehat{\eta}_i$ from section 11.4 and compute their variance. This approach, although essentially correct, has two important drawbacks. First, as we mentioned earlier, estimating a dummy-variable model can be very complex with micro-panels where N is large and T is small for simple computational reasons. Moreover, even if we manage to produce estimates of the η_i's, such estimates will be very imprecise if T is small.

Let's then look for an alternative and start by noticing the following:

$$\begin{aligned} Var(\overline{u}_i) &= Var\left(\frac{1}{T}\sum_{t=1}^{T} u_{it}\right) = Var\left[\frac{1}{T}\sum_{t=1}^{T}(\eta_i + v_{it})\right] \\ &= Var\left(\eta_i + \frac{1}{T}\sum_{t=1}^{T} v_{it}\right) = \sigma_\eta^2 + \frac{\sigma_v^2}{T} \end{aligned} \tag{12.10}$$

Equation 12.10 essentially tells us that if we were able to get a consistent estimate of $Var(\overline{u}_i)$ we could simply estimate σ_η^2 as:

$$\widehat{\sigma}_\eta^2 = \widehat{Var(\overline{u}_i)} - \frac{\sigma_v^2}{T} \tag{12.11}$$

Notice that $\overline{u}_i$ is the time average of the error term of the model in levels but it can also be interpreted as the residual of the *between-group model*, i.e. a model where we consider only one observation for each individual which is equal to the time average of all the observations:

$$\begin{aligned} \overline{y}_i &= \frac{1}{T}\sum_{t=1}^{T} y_{it} = \left(\frac{1}{T}\sum_{t=1}^{T} x_{it}\right)\beta + \eta_i + \left(\frac{1}{T}\sum_{t=1}^{T} v_{it}\right) \\ &= \overline{x}_i\beta + \eta_i + \overline{u}_i \end{aligned} \tag{12.12}$$

The OLS estimator of β from the between-group model of equation 12.12 is:

$$\widehat{\beta}_{BG} = \left[\sum_{i=1}^{N} \overline{x}_i'\overline{x}_i\right]^{-1}\left[\sum_{i=1}^{N} \overline{x}_i'\overline{y}_i\right] \tag{12.13}$$

This estimator is called between-group for the obvious reason that, contrary to the within-group, it only exploits variation between observations and it forgets about variation within observations over time.

Once we have computed $\widehat{\beta}_{BG}$, we can easily estimate $\overline{u}_i$ as follows:

$$\widehat{\overline{u}}_i = \overline{y}_i - \overline{x}_i \widehat{\beta}_{BG} \tag{12.14}$$

and finally get our consistent estimate of σ_η^2 without the need of estimating the dummy-variable model:

$$\begin{aligned} \widehat{\sigma}_\eta^2 &= \left(\frac{1}{N} \sum_{i=1}^{N} \widehat{\overline{u}}_i^2 \right) - \frac{\widehat{\sigma}_v^2}{T} \\ &= \left[\frac{1}{N} \sum_{i=1}^{N} \left(\overline{y}_i - \overline{x}_i \widehat{\beta}_{BG} \right)^2 \right] - \frac{\widehat{\sigma}_v^2}{T} \end{aligned} \tag{12.15}$$

where $\widehat{\sigma}_v^2$ is defined in equation 12.9.

With $\widehat{\sigma}_v^2$, as defined in equation 12.9, and $\widehat{\sigma}_\eta^2$, as defined in equation 12.15, we can then construct a consistent estimator of Ω and use it to do feasible GLS on the model in levels.

To conclude, it is important to notice that the only advantage of $\widehat{\beta}_{RE}$ over $\widehat{\beta}_{POLS}$ is efficiency. While both estimators are consistent, $\widehat{\beta}_{RE}$ is more efficient.

13 Comparing Fixed- and Random-Effects: the Hausman test (again)

Similarly to the test of endogeneity in the context of instrumental variable, the Hausman test can be applied to panel data to test test for the presence of unobserved heterogeneity.

In fact, note that, under the assumption that η_i is uncorrelated with the other observable regressors of the model x_{it}, both $\widehat{\beta}_{FE}$ and $\widehat{\beta}_{RE}$ are consistent but $\widehat{\beta}_{RE}$ is also efficient. Additionally, $\widehat{\beta}_{FE}$ is consistent also if that assumption fails, i.e. if η_i and x_{it} were correlated. This is totally analogous to the reasoning used to motivate the Hausman test in the context of instrumental variables, where $\widehat{\beta}_{OLS}$ was consistent and efficient under the assumption of no endogeneity while $\widehat{\beta}_{IV}$ was consistent but not efficient regardless of endogeneity.

Hence, The Hausman test for panel data is identical to the one for instrumental variables. It tests the assumption

$$H_0 : \quad Cov\left(x_{it}, \eta_i\right) = 0 \tag{13.1}$$

and it is defined as follows:

$$H = \left(\widehat{\beta}_{RE} - \widehat{\beta}_{FE}\right)' \left[Var\left(\widehat{\beta}_{RE} - \widehat{\beta}_{FE}\right)\right]^{-1} \left(\widehat{\beta}_{RE} - \widehat{\beta}_{FE}\right) \overset{a}{\sim} \chi^2_K \quad (13.2)$$

This statistics is distributed asymptotically according to a χ^2 distribution with K degrees of freedom, where K is the number of coefficients in β. As we learnt in the first part of the course, it is easy to show that:

$$Var\left(\widehat{\beta}_{RE} - \widehat{\beta}_{FE}\right) = Var\left(\widehat{\beta}_{FE}\right) - Var\left(\widehat{\beta}_{RE}\right) \quad (13.3)$$

thus making the computation of the test much easier.