ERSA Training Workshop
Lecture 5: Estimation of Binary Choice Models with Panel Data

Måns Söderbom

Friday 16 January 2009

1 Introduction

The methods discussed thus far in the course are well suited for modelling a **continuous**, **quantitative** variable - e.g. economic growth, the log of value-added or output, the log of earnings etc. Many economic phenomena of interest, however, concern variables that are not continuous or perhaps not even quantitative.

- What characteristics (e.g. parental) affect the likelihood that an individual obtains a higher degree?

- What are the determinants of the decision to export?

- What determines labour force participation (employed vs not employed)?

- What factors drive the incidence of civil war?

In this lecture we discuss how to model binary outcomes, using panel data. We will look at some empirical applications, including a dynamic model of exporting at the firm-level. The core reference is Chapter 15 in Wooldridge. We will also discuss briefly how tobit and selection models can be estimated with panel data.

2 Recap: Binary choice models without individual effects

Whenever the variable that we want to model is **binary**, it is natural to think in terms of probabilities, e.g.

- 'What is the probability that an individual with such and such characteristics owns a car?'

- 'If some variable X changes by one unit, what is the effect on the probability of owning a car?'

- When the dependent variable y_{it} is binary, it is typically equal to one for all observations in the data for which the event of interest has happened ('success') and zero for the remaining observations ('failure').

- We now review methods that can be used to analyze what factors 'determine' changes in the probability that y_{it} equals one.

2.1 The Linear Probability Model

Consider the linear regression model

$$\begin{aligned} y_{it} &= \beta_1 + \beta_2 x_{2it} + ... + \beta_K x_{Kit} + c_i + u_{it} \\ y_{it} &= \boldsymbol{x}_{it}\boldsymbol{\beta} + c_i + u_{it}, \end{aligned}$$

where y is a binary response variable, $\boldsymbol{x}_{it}$ is a $1 \times K$ vector of observed explanatory variables (including a constant), $\boldsymbol{\beta}$ is a $K \times 1$ vector of parameters, c_i is an unobserved time invariant individual effect, and u_{it} is a zero-mean residual uncorrelated with all the terms on the right-hand side.

- Assume strict exogeneity holds - the residual u_{it} is uncorrelated with all x-variables over the entire time period spanned by the panel (see earlier lectures on this course).

- Since the dependent variable is binary, it is natural to interpret the expected value of y as a probability. Indeed, under random sampling, the **unconditional** probability that y equals one is equal to the unconditional expected value of y, i.e. $E(y) = \mathsf{Pr}(y = 1)$.

- The **conditional** probability that y equals one is equal to the conditional expected value of y:

$$\Pr\left(y_{it} = 1 | \boldsymbol{x}_{it}, c_i\right) = E\left(y_{it} | \boldsymbol{x}_{it}, c_i; \boldsymbol{\beta}\right).$$

So if the model above is correctly specified, we have

$$\begin{aligned} \Pr\left(y_{it} = 1 | \boldsymbol{x}_{it}, c_i\right) &= \boldsymbol{x}_{it}\boldsymbol{\beta} + c_i, \\ \Pr\left(y_{it} = 0 | \boldsymbol{x}_{it}, c_i\right) &= 1 - \left(\boldsymbol{x}_{it}\boldsymbol{\beta} + c_i\right). \end{aligned} \tag{1}$$

- Equation (1) is a **binary response model**. In this particular model the probability of success (i.e. $y = 1$) is a **linear** function of the explanatory variables in the vector $\boldsymbol{x}$. Hence this is called a linear probability model (LPM).

- We can therefore use a linear regression model to estimate the parameters, such as OLS or the within estimator. Which particular linear estimator we

should use depends on the relationship between the observed explanatory variables and the unobserved individual effects - see the earlier lectures in the course for details.

[EXAMPLE 1: Modelling the decision to export in Ghana's manufacturing sector. To be discussed in class.]

2.1.1 Weaknesses of the Linear Probability Model

- One undesirable property of the LPM is that we can get predicted "probabilities" either less than zero or greater than one. Of course a probability by definition falls within the (0,1) interval, so predictions outside this range are meaningless and somewhat embarrassing.

- A related problem is that, conceptually, it does not make sense to say that a probability is **linearly** related to a continuous independent variable for all possible values. If it were, then continually increasing this explanatory variable would eventually drive $P(y = 1|x)$ above one or below zero.

- A third problem with the LPM, is that the residual is heteroskedastic. The easiest way of solving this problem is to obtain estimates of the standard errors that are robust to heteroskedasticity.

- A fourth and related problem is that the residual is not normally distributed. This implies that inference in small samples cannot be based on the usual suite of normality-based distributions such as the t test.

2.1.2 Strengths of the Linear Probability Model

- Easy to estimate, easy to interpret results. Marginal effects, for example, are straightforward:

$$\frac{\Delta \Pr\left(y_{it} = 1 | \boldsymbol{x}_{it}, c_i\right)}{\Delta x_{j,it}} = \beta_j$$

- Certain econometric problems are easier to address within the LPM framework than with probits and logits - for instance using instrumental variables whilst controlling for fixed effects.

[EXAMPLE 2: Miguel, Satyanath and Sergenti, JPE, 2004: Modelling the likelihood of civil war in Sub-Saharan Africa allowing for fixed effects and using instruments. To be discussed in class.]

2.2 Logit and Probit Models for Binary Response

- The two main problems with the LPM were: nonsense predictions are possible (there is nothing to bind the value of Y to the (0,1) range); and linearity doesn't make much sense conceptually. To address these problems we can use a nonlinear binary response model.

- For the moment we assume there are no unobserved individual effects. Under this assumption, we can use standard cross-section models to estimate the parameters of interest, even if we have panel data. Of course, the assumption that there are no unobserved individual effects is very restrictive, and in subsequent sections we discuss various ways of relaxing this assumption.

- We write our nonlinear binary response model as

$$\begin{aligned} \Pr(y=1|\boldsymbol{x}) &= G(\beta_1+\beta_2 x_2+...+\beta_K x_K) \\ \Pr(y=1|\boldsymbol{x}) &= G(\boldsymbol{x\beta}), \end{aligned} \tag{2}$$

  where G is a function taking on values strictly between zero and one: $0<G(z)<1$, for all real numbers z (individual and time subscripts have been omitted here).

- This is an **index model**, because $\Pr(y=1|x)$ is a function of the vector x only through the **index**

$$\boldsymbol{x\beta}=\beta_1+\beta_2 x_2+...+\beta_k x_k,$$

  which is a scalar. Notice that $0<G(\boldsymbol{x\beta})<1$ ensures that the estimated response probabilities are strictly between zero and one, which thus addresses the main worries of using LPM.

- G is a **cumulative density function** (cdf), monotonically increasing in the index z (i.e. $\boldsymbol{x\beta}$), with

$$\begin{aligned} \Pr(y = 1|\boldsymbol{x}) &\rightarrow 1 \text{ as } \boldsymbol{x\beta} \rightarrow \infty \\ \Pr(y = 1|\boldsymbol{x}) &\rightarrow 0 \text{ as } \boldsymbol{x\beta} \rightarrow -\infty. \end{aligned}$$

  It follows that G is a non-linear function, and hence we cannot use a linear regression model for estimation.

- Various non-linear functions for G have been suggested in the literature. By far the most common ones are the logistic distribution, yielding the **logit** model, and the standard normal distribution, yielding the **probit** model.

- In the logit model,

$$G(\boldsymbol{x\beta}) = \frac{\exp(\boldsymbol{x\beta})}{1 + \exp(\boldsymbol{x\beta})} = \Lambda(\boldsymbol{x\beta}),$$

which is between zero and one for all values of $\boldsymbol{x\beta}$ (recall that $\boldsymbol{x\beta}$ is a scalar). This is the cumulative distribution function (CDF) for a logistic variable.

- In the probit model, G is the standard normal CDF, expressed as an integral:

$$G\left(\boldsymbol{x\beta}\right) = \Phi\left(\boldsymbol{x\beta}\right) \equiv \int_{-\infty}^{\boldsymbol{x\beta}} \phi\left(v\right) dv,$$

where

$$\phi\left(v\right) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{v^2}{2}\right),$$

is the standard normal density. This choice of G also ensures that the probability of 'success' is strictly between zero and one for all values of the parameters and the explanatory variables.

The logit and probit functions are both increasing in $\boldsymbol{x\beta}$. Both functions increase relatively quickly at $\boldsymbol{x\beta} = 0$, while the effect on G at extreme values of $\boldsymbol{x\beta}$ tends to zero. The latter result ensures that the **partial effects** of changes in explanatory variables are **not constant**, a concern we had with the LPM.

A latent variable framework

- As we have seen, the probit and logit models resolve some of the problems with the LPM model. The key, really, is the specification

$$\Pr(y = 1|x) = G(\boldsymbol{x}\boldsymbol{\beta}),$$

  where G is the cdf for either the standard normal or the logistic distribution, because with any of these models we have a functional form that is easier to defend than the linear model. This, essentially, is how Wooldridge motivates the use of these models.

- The traditional way of introducing probits and logits in econometrics, however, is not as a response to a functional form problem. Instead, probits and logits are traditionally viewed as models suitable for estimating parameters of interest when the dependent variable is not fully observed.

- Let y^* be a continuous variable that we do not observe - a **latent variable** - and assume y^* is determined by the model

$$\begin{aligned} y^* &= \beta_1 + \beta_2 x_2 + ... + \beta_K x_K + e \\ &= \boldsymbol{x}\boldsymbol{\beta} + e, \end{aligned} \tag{3}$$

where e is a residual, assumed uncorrelated with $\boldsymbol{x}$ (i.e. $\boldsymbol{x}$ is not endogenous). While we do not observe y^*, we do observe the discrete choice made by the individual, according to the following choice rule:

$$\begin{aligned} y &= 1 \text{ if } y^* > 0 \\ y &= 0 \text{ if } y^* \leq 0. \end{aligned}$$

Why is y^* unobserved? Think about y^* as representing net utility of, say, buying a car. The individual undertakes a cost-benefit analysis and decides to purchase the car if the net utility is positive. We do not observe (because we cannot measure) the 'amount' of net utility; all we observe is

the actual outcome of whether or not the individual does buy a car. (If we had data on y^* we could estimate the model (3) with OLS as usual.)

- Now, we want to model the probability that a 'positive' choice is made (e.g. buying, as distinct from not buying, a car). By definition,

$$\Pr(y = 1|x) = \Pr(y^* > 0|x),$$

hence

$$\Pr(y = 1|x) = \Pr(e > -\boldsymbol{x\beta}),$$

which results in the logit model if e follows a logistic distribution, and the probit model if e is follows a (standard) normal distribution:

$$\begin{aligned} \Pr(y = 1|x) &= \Lambda(\boldsymbol{x\beta}) \text{ (logit)} \\ \Pr(y = 1|x) &= \Phi(\boldsymbol{x\beta}) \text{ (probit)} \end{aligned}$$

(integrate and exploit symmetry of the distribution to arrive at these expressions).

2.2.1 The likelihood function

- Probit and logit models are estimated by means of **Maximum Likelihood (ML)**. That is, the ML estimate of $\boldsymbol{\beta}$ is the particular vector $\hat{\boldsymbol{\beta}}^{ML}$ that gives the greatest likelihood of observing the outcomes in the sample $\{y_1, y_2, ...\}$, conditional on the explanatory variables $\boldsymbol{x}$.

- By assumption, the probability of observing $y_i = 1$ is $G(\boldsymbol{x}\boldsymbol{\beta})$ while the probability of observing $y_i = 0$ is $1 - G(\boldsymbol{x}\boldsymbol{\beta})$. It follows that the probability of observing the entire sample is

$$L(y|\boldsymbol{x};\boldsymbol{\beta}) = \prod_{i \in l} G(\boldsymbol{x}_i\boldsymbol{\beta}) \prod_{i \in m} [1 - G(\boldsymbol{x}_i\boldsymbol{\beta})],$$

  where l refers to the observations for which $y = 1$ and m to the observations for which $y = 0$.

- We can rewrite this as

$$L\left(y|\boldsymbol{x};\boldsymbol{\beta}\right)=\prod_{i=1}^{N}G\left(\boldsymbol{x}_i\boldsymbol{\beta}\right)^{y_i}\left[1-G\left(\boldsymbol{x}_i\boldsymbol{\beta}\right)\right]^{(1-y_i)},$$

because when $y=1$ we get $G\left(\boldsymbol{x}_i\boldsymbol{\beta}\right)$ and when $y=0$ we get $\left[1-G\left(\boldsymbol{x}_i\boldsymbol{\beta}\right)\right]$.

- The **log** likelihood for the sample is

$$\ln L\left(y|\boldsymbol{x};\boldsymbol{\beta}\right)=\sum_{i=1}^{N}\left\{y_i\ln G\left(\boldsymbol{x}_i\boldsymbol{\beta}\right)+\left(1-y_i\right)\ln\left[1-G\left(\boldsymbol{x}_i\boldsymbol{\beta}\right)\right]\right\}.$$

The MLE of β **maximizes** this log likelihood function.

- If G is the logistic CDF then we obtain the logit log likelihood:

$$\begin{aligned}
\ln L\left(y|\boldsymbol{x};\boldsymbol{\beta}\right) &= \sum_{i=1}^{N}\left\{y_i \ln \Lambda\left(\boldsymbol{x}_i\boldsymbol{\beta}\right) + \left(1-y_i\right)\ln\left[1-\Lambda\left(\boldsymbol{x}_i\boldsymbol{\beta}\right)\right]\right\} \\
\ln L\left(y|\boldsymbol{x};\boldsymbol{\beta}\right) &= \sum_{i=1}^{N}\left\{y_i \ln\left(\frac{\exp\left(\boldsymbol{x}_i\boldsymbol{\beta}\right)}{1+\exp\left(\boldsymbol{x}_i\boldsymbol{\beta}\right)}\right) + \left(1-y_i\right)\ln\left(\frac{1}{1+\exp\left(\boldsymbol{x}_i\boldsymbol{\beta}\right)}\right)\right\},
\end{aligned}$$

- If G is the standard normal CDF we get the probit log likelihood:

$$\ln L\left(y|\boldsymbol{x};\boldsymbol{\beta}\right) = \sum_{i=1}^{N}\left\{y_i \ln \Phi\left(\boldsymbol{x}_i\boldsymbol{\beta}\right) + \left(1-y_i\right)\ln\left[1-\Phi\left(\boldsymbol{x}_i\boldsymbol{\beta}\right)\right]\right\}.$$

- The maximum of the sample log likelihood is found by means of certain algorithms (e.g. Newton-Raphson) but we don't have to worry about that here.

2.2.2 Interpretation: Partial effects

In most cases the main goal is to determine the effects on the response probability $\mathsf{Pr}(y = 1|x)$ resulting from a change in one of the explanatory variables, say x_j.

Case I: The explanatory variable is continuous.

- When x_j is a continuous variable, its partial effect on $\mathsf{Pr}(y = 1|x)$ is obtained from the partial derivative:

$$\begin{aligned} \frac{\partial \mathsf{Pr}(y = 1|x)}{\partial x_j} &= \frac{\partial G(\boldsymbol{x\beta})}{\partial x_j} \\ &= g(\boldsymbol{x\beta}) \cdot \beta_j, \end{aligned}$$

where

$$g(z) \equiv \frac{dG(z)}{dz}$$

is the **probability density function** associated with G.

- Because the density function is non-negative, the partial effect of x_j will always have **the same sign** as β_j.

- Notice that the partial effect depends on $g(\boldsymbol{x}\boldsymbol{\beta})$; i.e. for different values of $x_1, x_2, ..., x_k$ the partial effect will be different.

- **Example:** Suppose we estimate a probit modelling the probability that a manufacturing firm in Ghana does some exporting as a function of firm

size. For simplicity, abstract from other explanatory variables. Our model is thus:

$$\Pr\left(\text{exports} = 1|size\right) = \Phi\left(\beta_0 + \beta_1 size\right),$$

where size is defined as the natural logarithm of employment. The probit results are

	coef.	t-value
β_0	-2.85	16.6
β_1	0.54	13.4

Since the coefficient on $size$ is positive, we know that the marginal effect must be positive. Treating $size$ as a continuous variable, it follows that the marginal effect is equal to

$$\begin{aligned} \frac{\partial \Pr\left(\text{exports} = 1|size\right)}{\partial size} &= \phi\left(\beta_0 + \beta_1 \cdot size\right)\beta_1 \\ &= \phi\left(-2.85 + 0.54 \cdot size\right)0.54, \end{aligned}$$

where $\phi(.)$ is the standard normal density function:

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-z^2/2\right).$$

We see straight away that the marginal effect depends on the size of the firm. In this particular sample the mean value of log employment is 3.4 (which corresponds to 30 employees), so let's evaluate the marginal effect at $size = 3.4$:

$$\begin{aligned} & \frac{\partial \Pr(\text{exports} = 1 | size = 3.4)}{\partial size} \\ = & \frac{1}{\sqrt{2\pi}} \exp\left(-\left(-2.85 + 0.54 \cdot 3.4\right)^2/2\right) 0.54 \\ = & \ 0.13, \end{aligned}$$

Hence, evaluated at log employment $= 3.4$, the results imply that an increase in log size by a small amount Δ raises the probability of exporting by 0.13Δ.

The Stata command 'mfx compute' can be used to obtain marginal effects, with standard errors, after logit and probit models.

Case II: The explanatory variable is discrete. If x_j is a discrete variable then we should not rely on calculus in evaluating the effect on the response probability. To keep things simple, suppose x_2 is binary. In this case the partial effect from changing x_2 from zero to one, holding all other variables fixed, is

$$G\left(\beta_1 + \beta_2 \cdot 1 + ... + \beta_K x_K\right) - G\left(\beta_1 + \beta_2 \cdot 0 + ... + \beta_K x_K\right).$$

Again this depends on all the values of the other explanatory variables and the values of all the other coefficients.

Again, knowing the **sign** of β_2 is sufficient for determining whether the effect is positive or not, but to find the **magnitude** of the effect we have to use the formula above.

The Stata command 'mfx compute' can spot dummy explanatory variables. In such a case it will use the above formula for estimating the partial effect.

3 Binary choice models for panel data

We now turn to the issue of how to estimate probit and logit models allowing for unobserved individual effects. Using a latent variable framework, we write the panel binary choice model as

$$\begin{aligned} y_{it}^{*} &= \boldsymbol{x}_{it}\boldsymbol{\beta} + c_i + u_{it}, \\ y_{it} &= 1\left[y_{it}^{*} > 0\right], \end{aligned} \tag{4}$$

and

$$\Pr\left(y_{it} = 1 | \boldsymbol{x}_{it}, c_i\right) = G\left(\boldsymbol{x}_{it}\boldsymbol{\beta} + c_i\right),$$

where $G\left(.\right)$ is either the standard normal CDF (probit) or the logistic CDF (logit).

- Recall that, in linear models, it is easy to eliminate c_i by means of first differencing or using within transformation.

- Those routes are **not** open to us here, unfortunately, since the model is nonlinear (e.g. differencing equation (4) does not remove c_i).

- Moreover, if we attempt to estimate c_i directly by adding $N-1$ individual dummy variables to the probit or logit specification, this will result in severely biased estimates of $\boldsymbol{\beta}$ unless T is large. This is known as the **incidental parameters problem:** with T small, the estimates of the c_i are inconsistent (i.e. increasing N does not remove the bias), and, unlike the linear model, the inconsistency in c_i has a 'knock-on effect' in the sense that the estimate of $\boldsymbol{\beta}$ becomes inconsistent too.

3.1 Incidental parameters: A classical example

Consider the logit model in which $T = 2$, β is a scalar, and x_{it} is a time dummy such that $x_{i1} = 0, x_{i2} = 1$. Thus

$$\begin{aligned}
\Pr\left(y_{it} = 1 | x_{i1}, c_i\right) &= \frac{\exp\left(\beta \cdot 0 + c_i\right)}{1 + \exp\left(\beta \cdot 0 + c_i\right)} \equiv \Lambda\left(\beta \cdot 0 + c_i\right), \\
\Pr\left(y_{it} = 1 | x_{i2}, c_i\right) &= \frac{\exp\left(\beta \cdot 1 + c_i\right)}{1 + \exp\left(\beta \cdot 1 + c_i\right)} \equiv \Lambda\left(\beta \cdot 1 + c_i\right).
\end{aligned}$$

Suppose we attempt to estimate this model with N dummy variables included to control for the individual effects. There would thus be $N + 1$ parameters in the model: $c_1, c_2, ..., c_i, ...c_N, \beta$. Our parameter of interest is β.

However, it can be shown that, in this particular case,

$$p \lim_{N \to \infty} \hat{\beta} = 2\beta.$$

That is, the probability limit of the logit dummy variable estimator - for this admittedly very special case - is double the true value of β. With a bias of 100% in very large (infinite) samples, this is not a very useful approach. This form of inconsistency also holds in more general cases: unless T is large, the logit dummy variable estimator will not work.

- So how can we proceed? I will discuss three common approaches: the traditional random effects (RE) probit (or logit) model; the conditional fixed effects logit model; and the Mundlak-Chamberlain approach.

3.2 The traditional random effects (RE) probit

Model:

$$\begin{aligned} y_{it}^{*} &= \boldsymbol{x}_{it}\boldsymbol{\beta} + c_i + u_{it}, \\ y_{it} &= 1\left[y_{it}^{*} > 0\right], \end{aligned}$$

and

$$\Pr\left(y_{it} = 1 | \boldsymbol{x}_{it}, c_i\right) = G\left(\boldsymbol{x}_{it}\boldsymbol{\beta} + c_i\right),$$

The key assumptions underlying this estimator are:

- c_i and $\boldsymbol{x}_{it}$ are independent

- the $\boldsymbol{x}_{it}$ are strictly exogenous (this will be necessary for it to be possible to write the likelihood of observing a given series of outcomes as the product of individual likelihoods).

- c_i has a **normal** distribution with zero mean and variance σ_c^2 (note: homoskedasticity).

- $y_{i1}, ..., y_{iT}$ are independent conditional on $(\boldsymbol{x}_i, c_i)$ - this rules out serial correlation in y_{it}, conditional on $(\boldsymbol{x}_i, c_i)$. This assumption enables us to write the likelihood of observing a given series of outcomes as the product of individual likelihoods. The assumption can easily be relaxed - see eq. (15.68) in Wooldridge (2002).

- Clearly these are restrictive assumptions, especially since endogeneity in the explanatory variables is ruled out. The only advantage (which may strike you as rather marginal) over a simple pooled probit model is that the RE model allows for serial correlation in the unobserved factors determining y_{it}, i.e. in $(c_i + u_{it})$.

- However, it is fairly straightforward to extend the model and allow for correlation between c_i and $\boldsymbol{x}_{it}$ - this is precisely what the Mundlak-Chamberlain approach achieves, as we shall see below.

- Clearly, if c_i had been observed, the likelihood of observing individual i would have been

$$\prod_{t=1}^{T} \left[\Phi\left(\boldsymbol{x}_{it}\boldsymbol{\beta} + c_i\right)\right]^{y_{it}} \left[1 - \Phi\left(\boldsymbol{x}_{it}\boldsymbol{\beta} + c_i\right)\right]^{(1-y_{it})},$$

and it would have been straightforward to maximize the sample likelihood conditional on $\boldsymbol{x}_{it}, c_i, y_{it}$.

- Because the c_i are unobserved, however, they cannot be included in the likelihood function. As discussed above, a dummy variables approach cannot be used, unless T is large. What can we do?

- Recall from basic statistics (Bayes' theorem for probability densities) that, in general,

$$f_{x|y}(x, y) = \frac{f_{xy}(x, y)}{f_y(y)},$$

where $f_{x|y}(x, y)$ is the conditional density of X given $Y = y$; $f_{xy}(x, y)$ is the joint distribution of random variables X, Y; and $f_y(y)$ is the marginal

density of Y. Thus,

$$f_{xy}(x,y) = f_{x|y}(x,y) f_y(y).$$

- Moreover, the marginal density of X can be obtained by integrating out y from the joint density

$$f_x(x) = \int f_{xy}(x,y)\, dy = \int f_{x|y}(x,y) f_y(y)\, dy.$$

- Clearly we can think about $f_x(x)$ as a likelihood contribution. For a linear model, for example, we might write

$$f_\varepsilon(\varepsilon) = \int f_{\varepsilon c}(\varepsilon, c)\, dc = \int f_{\varepsilon|c}(\varepsilon, c) f_c(c)\, dc,$$

where $\varepsilon_{it} = y_{it} - (\boldsymbol{x}_{it}\boldsymbol{\beta} + c_i)$.

- In the context of the traditional RE probit, we **integrate out** c_i from the likelihood as follows:

$$L_i\left(y_{i1},...,y_{iT}|\boldsymbol{x}_{i1},...,\boldsymbol{x}_{iT};\boldsymbol{\beta},\sigma_c^2\right) =$$

$$\int \prod_{t=1}^{T} \left[\Phi\left(\boldsymbol{x}_{it}\boldsymbol{\beta}+c\right)\right]^{y_{it}} \left[1-\Phi\left(\boldsymbol{x}_{it}\boldsymbol{\beta}+c\right)\right]^{(1-y_{it})} \left(1/\sigma_c\right)\phi\left(c/\sigma_c\right)dc.$$

- In general, there is no analytical solution here, and so numerical methods have to be used. The most common approach is to use a **Gauss-Hermite quadrature** method, which amounts to approximating

$$\int \prod_{t=1}^{T} \left[\Phi\left(\boldsymbol{x}_{it}\boldsymbol{\beta}+c\right)\right]^{y_{it}} \left[1-\Phi\left(\boldsymbol{x}_{it}\boldsymbol{\beta}+c\right)\right]^{(1-y_{it})} \left(1/\sigma_c\right)\phi\left(c/\sigma_c\right)dc$$

as

$$\pi^{-1/2} \sum_{m=1}^{M} w_m \prod_{t=1}^{T} \left[\Phi\left(\boldsymbol{x}_{it}\boldsymbol{\beta} + \sqrt{2}\sigma_c g_m\right)\right]^{y_{it}} \left[1 - \Phi\left(\boldsymbol{x}_{it}\boldsymbol{\beta} + \sqrt{2}\sigma_c g_m\right)\right]^{(1-y_{it})} , \tag{5}$$

where M is the number of nodes, w_m is a prespecified weight, and g_m a prespecified node (prespecified in such a way as to provide as good an approximation as possible of the normal distribution).

- For example, if $M = 3$, we have

w_m	g_m
0.2954	-1.2247
1.1826	0.0000
0.2954	1.2247

in which case (5) can be written out as

$$0.1667 \prod_{t=1}^{T} \left[\Phi\left(\boldsymbol{x}_{it}\boldsymbol{\beta} - 1.731\sigma_c\right)\right]^{y_{it}} \left[1 - \Phi\left(\boldsymbol{x}_{it}\boldsymbol{\beta} - 1.731\sigma_c\right)\right]^{(1-y_{it})}$$
$$+0.6667 \prod_{t=1}^{T} \left[\Phi\left(\boldsymbol{x}_{it}\boldsymbol{\beta}\right)\right]^{y_{it}} \left[1 - \Phi\left(\boldsymbol{x}_{it}\boldsymbol{\beta}\right)\right]^{(1-y_{it})}$$
$$+0.1667 \prod_{t=1}^{T} \left[\Phi\left(\boldsymbol{x}_{it}\boldsymbol{\beta} + 1.731\sigma_c\right)\right]^{y_{it}} \left[1 - \Phi\left(\boldsymbol{x}_{it}\boldsymbol{\beta} + 1.731\sigma_c\right)\right]^{(1-y_{it})}.$$

In practice a larger number of nodes than 3 would of course be used (the default in Stata is $M = 12$). Lists of weights and nodes for given values of M can be found in the literature.

- To form the sample log likelihood, we simply compute weighted sums in this fashion for each individual in the sample, and then add up all the

individual likelihoods expressed in natural logarithms:

$$\log L = \sum_{i=1}^{N} \log L_i \left(y_{i1}, ..., y_{iT} | \boldsymbol{x}_{i1}, ..., \boldsymbol{x}_{iT}; \boldsymbol{\beta}, \sigma_c^2 \right).$$

Marginal effects at $c_i = 0$ can be computed using standard techniques. This model can be estimated in Stata using the **xtprobit** command.

[EXAMPLE 3: Modelling exports in Ghana using probit and allowing for unobserved individual effects. Discuss in class].

Whilst perhaps elegant, the above model does **not** allow for a correlation between c_i and the explanatory variables, and so does not achieve anything in terms of addressing an endogeneity problem. We now turn to more useful models in that context.

3.3 The "fixed effects" logit model

Now return to the panel logit model:

$$\Pr\left(y_{it}=1|\boldsymbol{x}_{it},c_i\right)=\Lambda\left(\boldsymbol{x}_{it}\boldsymbol{\beta}+c_i\right).$$

- One important advantage of this model over the probit model is that will be possible to obtain a consistent estimator of $\boldsymbol{\beta}$ without making any assumptions about how c_i is related to $\boldsymbol{x}_{it}$ (however, you need strict exogeneity to hold; cf. within estimator for linear models).

- This is possible, because the logit functional form enables us to eliminate c_i from the estimating equation, once we condition on what is sometimes referred to as a "minimum sufficient statistic" for c_i.

To see this, assume $T = 2$, and consider the following **conditional** probabilities:

$$\Pr\left(y_{i1} = 0, y_{i2} = 1 | x_{i1}, x_{i2}, c_i, y_{i1} + y_{i2} = 1\right),$$

and

$$\Pr\left(y_{i1} = 0, y_{i2} = 1 | x_{i1}, x_{i2}, c_i, y_{i1} + y_{i2} = 1\right).$$

The key thing to note here is that we condition on $y_{i1} + y_{i2} = 1$, i.e. that y_{it} **changes** between the two time periods. For the logit functional form, we have

$$\begin{aligned}\Pr\left(y_{i1} + y_{i2} = 1 | x_{i1}, x_{i2}, c_i\right) = & \frac{\exp\left(\boldsymbol{x}_{i1}\boldsymbol{\beta} + c_i\right)}{1 + \exp\left(\boldsymbol{x}_{i1}\boldsymbol{\beta} + c_i\right)} \frac{1}{1 + \exp\left(\boldsymbol{x}_{i2}\boldsymbol{\beta} + c_i\right)} \\ & + \frac{1}{1 + \exp\left(\boldsymbol{x}_{i1}\boldsymbol{\beta} + c_i\right)} \frac{\exp\left(\boldsymbol{x}_{i2}\boldsymbol{\beta} + c_i\right)}{1 + \exp\left(\boldsymbol{x}_{i2}\boldsymbol{\beta} + c_i\right)},\end{aligned}$$

or simply

$$\Pr\left(y_{i1} + y_{i2} = 1 | x_{i1}, x_{i2}, c_i\right) = \frac{\exp\left(\boldsymbol{x}_{i1}\boldsymbol{\beta} + c_i\right) + \exp\left(\boldsymbol{x}_{i2}\boldsymbol{\beta} + c_i\right)}{\left[1 + \exp\left(\boldsymbol{x}_{i1}\boldsymbol{\beta} + c_i\right)\right]\left[1 + \exp\left(\boldsymbol{x}_{i2}\boldsymbol{\beta} + c_i\right)\right]}.$$

Furthermore,

$$\Pr\left(y_{i1}=0, y_{i2}=1 | x_{i1}, x_{i2}, c_i\right) = \frac{1}{1+\exp\left(\boldsymbol{x}_{i1}\boldsymbol{\beta}+c_i\right)} \frac{\exp\left(\boldsymbol{x}_{i2}\boldsymbol{\beta}+c_i\right)}{1+\exp\left(\boldsymbol{x}_{i2}\boldsymbol{\beta}+c_i\right)},$$

hence, conditional on $y_{i1}+y_{i2}=1$,

$$\Pr\left(y_{i1}=0, y_{i2}=1 | x_{i1}, x_{i2}, c_i, y_{i1}+y_{i2}=1\right)$$

$$= \frac{\exp\left(\boldsymbol{x}_{i2}\boldsymbol{\beta}+c_i\right)}{\exp\left(\boldsymbol{x}_{i1}\boldsymbol{\beta}+c_i\right)+\exp\left(\boldsymbol{x}_{i2}\boldsymbol{\beta}+c_i\right)}$$

$$\Pr\left(y_{i1}=0, y_{i2}=1 | x_{i1}, x_{i2}, y_{i1}+y_{i2}=1\right) = \frac{\exp\left(\Delta\boldsymbol{x}_{i2}\boldsymbol{\beta}\right)}{1+\exp\left(\Delta\boldsymbol{x}_{i2}\boldsymbol{\beta}\right)}$$

- The key result here is that the c_i are **eliminated**. It follows that

$$\Pr\left(y_{i1}=1, y_{i2}=0 | x_{i1}, x_{i2}, y_{i1}+y_{i2}=1\right) = \frac{1}{1+\exp\left(\Delta\boldsymbol{x}_{i2}\boldsymbol{\beta}\right)}.$$

- Remember:

1. These probabilities condition on $y_{i1} + y_{i2} = 1$

2. These probabilities are independent of c_i.

Hence, by maximizing the following **conditional** log likelihood function

$$\log L = \sum_{i=1}^{N} \left\{ d_{01i} \ln \left(\frac{\exp\left(\Delta \boldsymbol{x}_{i2} \boldsymbol{\beta}\right)}{1 + \exp\left(\Delta \boldsymbol{x}_{i2} \boldsymbol{\beta}\right)} \right) + d_{10i} \ln \left(\frac{1}{1 + \exp\left(\Delta \boldsymbol{x}_{i2} \boldsymbol{\beta}\right)} \right) \right\},$$

we obtain consistent estimates of $\boldsymbol{\beta}$, regardless of whether c_i and $\boldsymbol{x}_{it}$ are correlated.

- The trick is thus to condition the likelihood on the outcome series (y_{i1}, y_{i2}), and in the more general case $(y_{i1}, y_{i2}, ..., y_{iT})$. For example, if $T = 3$, we can condition on $\sum_t y_{it} = 1$, with possible sequences $\{1, 0, 0\}, \{0, 1, 0\}$ and $\{0, 0, 1\}$, or on $\sum_t y_{it} = 2$, with possible sequences $\{1, 1, 0\}, \{1, 0, 1\}$ and $\{0, 1, 1\}$. Stata does this for us, of course. This estimator is requested in Stata by using **xtlogit** with the **fe** option.

[EXAMPLE 4: Modelling exports in Ghana using a "fixed effects" logit. To be discussed in class].

Note that the logit functional form is crucial for it to be possible to eliminate the c_i in this fashion. It won't be possible with probit. So this approach is not really very general. Another awkward issue concerns the interpretation of the results. The estimation procedure just outlined implies we do not obtain estimates of c_i, which means we can't compute marginal effects.

3.4 Modelling the random effect as a function of x-variables

The previous two methods are useful, but arguably they don't quite help you achieve enough:

- the traditional random effects probit/logit model requires strict exogeneity and zero correlation between the explanatory variables and c_i;

- the fixed effects logit relaxes the latter assumption but we can't obtain consistent estimates of c_i and hence we can't compute the conventional marginal effects in general.

We will now discuss an approach which, in some ways, can be thought of as representing a middle way. Start from the latent variable model

$$\begin{aligned} y_{it}^{*} &= \boldsymbol{x}_{it}\boldsymbol{\beta} + c_i + e_{it}, \\ y_{it} &= 1_{[y_{it}^{*}>0]}. \end{aligned}$$

Consider writing the c_i as an **explicit function** of the x-variables, for example as follows:

$$c_i = \psi + \bar{\boldsymbol{x}}_i\boldsymbol{\xi} + a_i, \tag{6}$$

or

$$c_i = \phi + \boldsymbol{x}_i\boldsymbol{\tau} + b_i \tag{7}$$

where $\bar{\boldsymbol{x}}_i$ is an average of $\boldsymbol{x}_{it}$ over time for individual i (hence time invariant); $\boldsymbol{x}_i$ contains $\boldsymbol{x}_{it}$ for all t; a_i is assumed uncorrelated with $\bar{\boldsymbol{x}}_i$; b_i is assumed uncorrelated with $\boldsymbol{x}_i$. Equation (6) is easier to implement and so we will focus on this (see Wooldridge, 2002, pp. 489-90 for a discussion of the more general specification).

- Assume that $var\left(a_i\right) = \sigma_a^2$ is constant (i.e. there is homoskedasticity) and that e_i is normally distributed - the model that then results is known as **Chamberlain's random effects probit model**. You might say (6) is restrictive, in the sense that functional form assumptions are made, but at least it allows for non-zero correlation between c_i and the regressors $\boldsymbol{x}_{it}$.

- The probability that $y_{it} = 1$ can now be written as

  $$\mathsf{Pr}\left(y_{it} = 1|\boldsymbol{x}_{it}, c_i\right) = \mathsf{Pr}\left(y_{it} = 1|\boldsymbol{x}_{it}, \bar{\boldsymbol{x}}_i, a_i\right) = \mathbf{\Phi}\left(\boldsymbol{x}_{it}\boldsymbol{\beta} + \psi + \bar{\boldsymbol{x}}_i\xi + a_i\right).$$

  You now see that, after having added $\bar{\boldsymbol{x}}_i$ to the RHS, we arrive at the traditional random effects probit model:

$$L_i\left(y_{i1}, ..., y_{iT}|\boldsymbol{x}_{i1}, ..., \boldsymbol{x}_{iT}; \boldsymbol{\beta}, \sigma_a^2\right) = \int \prod_{t=1}^{T} \left[\mathbf{\Phi}\left(\boldsymbol{x}_{it}\boldsymbol{\beta} + \psi + \bar{\boldsymbol{x}}_i\xi + a\right)\right]^{y_{it}}$$

$$\times \left[1 - \mathbf{\Phi}\left(\boldsymbol{x}_{it}\boldsymbol{\beta} + \psi + \bar{\boldsymbol{x}}_i\xi + a\right)\right]^{(1-y_{it})} \left(1/\sigma_a\right) \phi\left(a/\sigma_a\right) da.$$

- Effectively, we are adding $\bar{\boldsymbol{x}}_i$ as control variables to allow for some correlation between the random effect c_i and the regressors.

- If $\boldsymbol{x}_{it}$ contains **time invariant** variables, then clearly they will be collinear with their mean values for individual i, thus preventing separate identification of β-coefficients on time invariant variables.

- We can easily compute marginal effects at the mean of c_i, since

$$E\left(c_i\right) = \psi + E\left(\bar{\boldsymbol{x}}_i\right)\boldsymbol{\xi}$$

- Notice also that this model nests the simpler and more restrictive traditional random effects probit: under the (easily testable) null hypothesis that $\xi = 0$, the model reduces to the traditional model discussed earlier.

[EXAMPLE 5: To be discussed in class].

3.5 Relaxing the normality assumption for the unobserved effect

The assumption that c_i (or a_i) is normally distributed is potentially strong. One alternative is to follow Heckman and Singer (1984) and adopt a **non-parametric** strategy for characterizing the distribution of the random effects. The premise of this approach is that the distribution of c can be approximated by a discrete multinomial distribution with Q points of support:

$$\Pr(c = C_q) = P_q,$$

$0 \leq P_q \leq 1$, $\sum_q P_q = 1$, $q = 1, 2, ..., Q$, where the C_q, and the P_q are parameters to be estimated.

Hence, the estimated "support points" (the C_q) determine possible realizations for the random intercept, and the P_q measure the associated probabilities. The

likelihood contribution of individual i is now

$$L_i\left(y_{i1},...,y_{iT}|\boldsymbol{x}_{i1},...,\boldsymbol{x}_{iT};\boldsymbol{\beta},\sigma_c^2\right) =$$

$$\sum_q^Q P_q \prod_{t=1}^T \left[\Phi\left(\boldsymbol{x}_{it}\boldsymbol{\beta}+C_q\right)\right]^{y_{it}} \left[1-\Phi\left(\boldsymbol{x}_{it}\boldsymbol{\beta}+C_q\right)\right]^{(1-y_{it})}.$$

Compared to the model based on the normal distribution for c_i, this model is clearly quite flexible.

In estimating the model, one important issue refers to the number of support points, Q. In fact, there are no well-established theoretically based criteria for determining the number of support points in models like this one. Standard practice is to increase Q until there are only marginal improvements in the log likelihood value. Usually, the number of support points is small - certainly below 10 and typically below 5.

Notice that there are many parameters in this model. With 4 points of support, for example, you estimate 3 probabilities (the 4th is a 'residual' probability resulting from the constraint that probabilities sum to 1) and 3 support points (one is omitted if - as typically is the case - $\boldsymbol{x}_{it}$ contains a constant). So that's 6 parameters compared to 1 parameter for the traditional random effects probit based on normality. That is the consequence of attempting to estimate the **entire distribution of** c.

Unfortunately, implementing this model is often difficult:

- Sometimes the estimator will not converge.

- Convergence may well occur at a local maximum.

- Inverting the Hessian in order to get standard errors may not always be possible.

So clearly the additional flexibility comes at a cost. Whether that is worth incurring depends on the data and (perhaps primarily) the econometrician's preferences. We used this approach in the paper "Do African manufacturing firms learn from exporting?", JDS, 2004, and we obtained some evidence this approach outperformed one based on normality.

Allegedly, the Stata program **gllamm** can be used to produce results for this type of estimator.*

*http://www.gllamm.org/

3.6 Dynamic Unobserved Effects Probit Models

Earlier in the course you have seen that using lagged dependent variables as explanatory variables complicates the estimation of standard linear panel data models. Conceptually, similar problems arise for nonlinear models, but since we don't rely on differencing the steps involved for dealing with the problems are a little different.

Consider the following dynamic probit model:

$$\begin{aligned} y_{it}^{*} &= \rho y_{i,t-1} + \boldsymbol{z}_{it}\boldsymbol{\delta} + c_i + u_{it}, \\ y_{it} &= 1\left[y_{it}^{*} > 0\right], \end{aligned}$$

and

$$\mathsf{Pr}\left(y_{it} = 1 | \boldsymbol{x}_{it}, c_i\right) = \Phi\left(\rho y_{i,t-1} + \boldsymbol{z}_{it}\boldsymbol{\delta} + c_i\right),$$

where $\boldsymbol{z}_{it}$ are strictly exogenous explanatory variables (what follows below is applicable for logit too). With this specification, the outcome y_{it} is allowed to depend on the outcome in $t-1$ as well as unobserved heterogeneity. Observations:

- The unobserved effect c_i is correlated with $y_{i,t-1}$ by definition

- The coefficient ρ is often referred to as the **state dependence** parameter. If $\rho \neq 0$, then the outcome $y_{i,t-1}$ **influences** the outcome in period t, y_{it}.

- If $var\left(c_i\right) > 0$, so that there is **unobserved heterogeneity**, we cannot use a pooled probit to test $H_0 : \rho = 0$. The reason is that under $var\left(c_i\right) > 0$, there will be serial correlation in the y_{it}.

In order to distinguish state dependence from heterogeneity, we need to allow for both mechanisms at the same time when estimating the model. If c_i had been observed, the likelihood of observing individual i would have been

$$\prod_{t=1}^{T} \left[\Phi\left(\rho y_{i,t-1} + \boldsymbol{z}_{it}\boldsymbol{\delta} + c_i\right)\right]^{y_{it}} \left[1 - \Phi\left(\rho y_{i,t-1} + \boldsymbol{z}_{it}\boldsymbol{\delta} + c_i\right)\right]^{(1-y_{it})}.$$

As already discussed, unless T is very large, we cannot use the dummy variables approach to control for unobserved heterogeneity. Instead, we will integrate out c_i using similar techniques to those discussed for the nondynamic model.

- However, estimation is more involved because $y_{i,t-1}$ is not uncorrelated with c_i.

Now, we observe in the data the series of outcomes $(y_{i0}, y_{i1}, y_{i2}, ..., y_{iT})$. Suppose for the moment that y_{i0} is actually independent of c_i. Clearly this

is not a very attractive assumption, and we will relax it shortly. Under this assumption, however, the likelihood contribution of individual i takes the form

$$\begin{aligned} f\left(y_{i1}, y_{i2}, ..., y_{iT}, c_i\right) &= f_{y(T)|y(T-1),c_i}\left(y_{iT}, y_{i,T-1}, c_i\right) \\ &\times f_{y(T-1)|y(T-2),c_i}\left(y_{i,T-1}, y_{i,T-2}, c_i\right) \\ &\times f_{y2|y1,c_i}\left(y_{i2}, y_{i1}, c_i\right) \\ &\times f_{y1|y0,c_i}\left(y_{i1}, y_{i0}, c_i\right) \\ &\times f_{y0}\left(y_{i0}\right), \end{aligned}$$

and so we can integrate out c_i in the usual fashion:

$$\begin{aligned} f\left(y_{i1}, y_{i2}, ..., y_{iT}\right) &= f_{y0}\left(y_{i0}\right) \int f_{y(T)|y(T-1),c}\left(y_{iT}, y_{i,T-1}, c\right) \\ &\times f_{y(T-1)|y(T-2),c}\left(y_{i,T-1}, y_{i,T-2}, c\right) \times ... \\ &... \times f_{y2|y1,c}\left(y_{i2}, y_{i1}, c\right) \times f_{y1|y0,c}\left(y_{i1}, y_{i0}, c\right) f_c\left(c\right) dc. \end{aligned} \tag{8}$$

The dependence of y_{i1} on c_i in the likelihood contribution $f_{y2|y1,c}\left(y_{i2}, y_{i1}, c\right)$ is captured by the term $f_{y1|y0,c}\left(y_{i1}, y_{i0}, c\right)$, the dependence of y_{i2} on c_i in the

likelihood contribution $f_{y3|y2,c}(y_{i3}, y_{i2}, c)$ is captured by the term $f_{y2|y1,c}(y_{i1}, y_{i0}, c)$, and so on.

- Consequently the right-hand side of (8) really does result in $f(y_{i1}, y_{i2}, ..., y_{iT})$, i.e. a likelihood contribution that is not dependent on c_i.

- However, key for this equality to hold is that there is no dependence between y_{i0} and c_i - otherwise I would not be allowed to move the density of y_{i0} out of the integral.

Suppose now I do not want to make the very strong assumption that y_{i0} is

actually independent of c_i. In that case, I am going to have to tackle

$$\begin{aligned} f\left(y_{i1}, y_{i2}, ..., y_{iT}\right) &= \int f_{y(T)|y(T-1),c}\left(y_{iT}, y_{i,T-1}, c\right) \\ &\times f_{y(T-1)|y(T-2),c}\left(y_{i,T-1}, y_{i,T-2}, c\right) \\ &\times f_{y2|y1,c}\left(y_{i2}, y_{i1}, c\right) \times f_{y1|y0,c}\left(y_{i1}, y_{i0}, c\right) \\ &\times f_{y0|c}\left(y_{i0}, c\right) f_c\left(c\right) dc. \end{aligned}$$

The dynamic probit version of this equation is

$$\begin{aligned} L_i\left(y_{i1}, ..., y_{iT}|\boldsymbol{x}_{i1}, ..., \boldsymbol{x}_{iT}; \boldsymbol{\beta}, \sigma_c^2\right) &= \int \prod_{t=1}^{T} \left[\Phi\left(\rho y_{i,t-1} + \boldsymbol{z}_{it}\boldsymbol{\delta} + c\right)\right]^{y_{it}} \\ &\times \left[1 - \Phi\left(\rho y_{i,t-1} + \boldsymbol{z}_{it}\boldsymbol{\delta} + c\right)\right]^{(1-y_{it})} \\ &f_{y0|c,z(i)}\left(y_{i0}, \boldsymbol{z}_i, c\right)\left(1/\sigma_c\right)\phi\left(c/\sigma_c\right) dc. \end{aligned}$$

Basically I have an endogeneity problem: in $f_{y0|c,z(i)}\left(y_{i0}, \boldsymbol{z}_i, c\right)$, the regressor y_{i0} is correlated with the unobserved random effect. This is usually called

the **initial conditions problem**. Clearly as T gets large the problem posed by the initial conditions problem becomes less serious (smaller weight of the problematic term), but with T small it can cause substantial bias.

3.6.1 Heckman's (1981) solution

Heckman (1981) suggested a solution. He proposed dealing with $f_{y0|c,z(i)}\left(y_{i0}, \boldsymbol{z}_i, c\right)$ by adding an equation that explicitly models the dependence of y_{i0} on c_i and $\boldsymbol{z}_i$. It's conceivable, for example, to assume

$$\Pr\left(y_{i0}|\boldsymbol{z}_i, c_i\right) = \Phi\left(\eta + \boldsymbol{z}_i\boldsymbol{\pi} + \gamma c_i\right),$$

where $\eta, \boldsymbol{\pi}, \gamma$ are to be estimated jointly with the $\rho, \boldsymbol{\delta}$ and $\boldsymbol{\delta}$. The key thing to notice here is the presence c_i. Clearly, if $\gamma \neq 0$, then c_i is correlated with the initial observation y_{i0}.

Now write the dynamic probit likelihood contribution of individual i as

$$L_i\left(y_{i1}, ..., y_{iT}|\boldsymbol{x}_{i1}, ..., \boldsymbol{x}_{iT}; \boldsymbol{\beta}, \sigma_c^2\right) =$$

$$\int \prod_{t=1}^{T} \left[\Phi\left(\rho y_{i,t-1} + \boldsymbol{z}_{it}\boldsymbol{\delta} + c\right)\right]^{y_{it}} \left[1 - \Phi\left(\rho y_{i,t-1} + \boldsymbol{z}_{it}\boldsymbol{\delta} + c\right)\right]^{(1-y_{it})}$$
$$\left[\Phi\left(\eta + \boldsymbol{z}_i\boldsymbol{\pi} + \gamma c\right)\right]^{y_{i0}} \left[1 - \Phi\left(\eta + \boldsymbol{z}_i\boldsymbol{\pi} + \gamma c\right)\right]^{(1-y_{i0})} \left(1/\sigma_c\right) \phi\left(c/\sigma_c\right) dc.$$

A maximum likelihood estimator based on a sample likelihood function made up of such individual likelihood contributions will be consistent, under the assumptions made above.

The downside of this procedure is that you have to code up the likelihood function yourself. I have written a SAS program that implements this estimator (heckman81_dprob) - one day I might translate this into Stata code...

3.6.2 Wooldridge's (2005) solution

An alternative approach, which is **much** easier to implement than the Heckman (1981) estimator, has been proposed by Wooldridge. It goes like this.

- Rather than writing y_{i0} as a function of c_i and $\boldsymbol{z}_i$ (Heckman, 1981), we can write c_i as a function of y_{i0} and $\boldsymbol{z}_i$:

$$c_i = \psi + \xi_0 y_{i0} + \boldsymbol{z}_i \boldsymbol{\xi} + a_i,$$

  where $a_i \sim Normal\left(0, \sigma_a^2\right)$ and independent of $y_{i0}, \boldsymbol{z}_i$.

- Notice that the relevant likelihood contribution

$$L_i\left(y_{i1}, ..., y_{iT} | \boldsymbol{x}_{i1}, ..., \boldsymbol{x}_{iT}; \boldsymbol{\beta}, \sigma_c^2\right) =$$

$$\int \prod_{t=1}^{T} \left[\Phi\left(\rho y_{i,t-1} + \boldsymbol{z}_{it}\boldsymbol{\delta} + c\right)\right]^{y_{it}} \left[1 - \Phi\left(\rho y_{i,t-1} + \boldsymbol{z}_{it}\boldsymbol{\delta} + c\right)\right]^{(1-y_{it})}$$
$$f_{y0|c,z(i)}\left(y_{i0}, \boldsymbol{z}_i, c\right)\left(1/\sigma_c\right)\phi\left(c/\sigma_c\right)dc.$$

can be expressed alternatively as

$$L_i\left(y_{i1}, ..., y_{iT}|\boldsymbol{x}_{i1}, ..., \boldsymbol{x}_{iT}; \boldsymbol{\beta}, \sigma_c^2\right) =$$

$$\int \prod_{t=1}^{T} \left[\Phi\left(\rho y_{i,t-1} + \boldsymbol{z}_{it}\boldsymbol{\delta} + c\right)\right]^{y_{it}} \left[1 - \Phi\left(\rho y_{i,t-1} + \boldsymbol{z}_{it}\boldsymbol{\delta} + c\right)\right]^{(1-y_{it})}$$
$$f_{c|y0,z(i)}\left(y_{i0}, \boldsymbol{z}_i, c\right)\left(1/\sigma_c\right)\phi\left(c/\sigma_c\right)dc,$$

or, given the specification now adopted for c,

$$L_i\left(y_{i1}, ..., y_{iT}|\boldsymbol{x}_{i1}, ..., \boldsymbol{x}_{iT}; \boldsymbol{\beta}, \sigma_a^2\right) =$$

$$\int \prod_{t=1}^{T} \left[\Phi\left(\rho y_{i,t-1} + \boldsymbol{z}_{it}\boldsymbol{\delta} + \psi + \xi_0 y_{i0} + \boldsymbol{z}_i\boldsymbol{\xi} + a\right)\right]^{y_{it}}$$
$$\left[1 - \Phi\left(\rho y_{i,t-1} + \boldsymbol{z}_{it}\boldsymbol{\delta} + \psi + \xi_0 y_{i0} + \boldsymbol{z}_i\boldsymbol{\xi} + a\right)\right]^{(1-y_{it})}$$
$$f_a(a)\,(1/\sigma_a)\,\phi(a/\sigma_a)\,da.$$

Hence, because a, is (assumed) uncorrelated with $\boldsymbol{z}_i$ and y_{i0}, we can use standard random effects probit software to estimate the parameters of interest. This approach also allows us, of course, to test for state dependence ($H_0 : \rho = 0$) whilst allowing for unobserved heterogeneity (if we ignore heterogeneity, we basically cannot test convincingly for state dependence).

- Notice that Wooldridge's method is very similar in spirit to the Mundlak-Chamberlain methods introduced earlier.

[EXAMPLE 6. To be discussed in class.]

4 Extension I: Panel Tobit Models

The treatment of tobit models for panel data is very similar to that for probit models. We state the (non-dynamic) unobserved effects model as

$$y_{it} = \max\left(0, \boldsymbol{x}_{it}\boldsymbol{\beta} + c_i + u_{it}\right),$$

$$u_{it}|\boldsymbol{x}_{it}, c_i \sim Normal\left(0, \sigma_u^2\right).$$

We cannot control for c_i by means of a dummy variable approach (incidental parameters problem), and no tobit model analogous to the "fixed effects" logit exists. We therefore consider the random effects tobit estimator (Note: Honoré has proposed a "fixed effects" tobit that does not impose distributional assumptions. Unfortunately it is hard to implement. Moreover, partial effects cannot be estimated. I therefore do not cover this approach. See Honoré's web page if you are interested).

4.1 Traditional RE tobit

For the traditional random effects tobit model, the underlying assumptions are the same as those underlying the traditional RE probit. That is,

- c_i and $\boldsymbol{x}_{it}$ are independent
- the $\boldsymbol{x}_{it}$ are strictly exogenous (this will be necessary for it to be possible to write the likelihood of observing a given series of outcomes as the product of individual likelihoods).
- c_i has a normal distribution with zero mean and variance σ_c^2

- $y_{i1}, ..., y_{iT}$ are independent conditional on $(\boldsymbol{x}_i, c_i)$, ruling out serial correlation in y_{it}, conditional on $(\boldsymbol{x}_i, c_i)$. This assumption can be relaxed.

Under these assumptions, we can proceed in exactly the same way as for the traditional RE probit, once we have changed the log likelihood function from probit to tobit. Hence, the contribution of individual i to the sample likelihood is

$$L_i\left(y_{i1}, ..., y_{iT} | \boldsymbol{x}_{i1}, ..., \boldsymbol{x}_{iT}; \boldsymbol{\beta}, \sigma_c^2\right) =$$

$$\int \prod_{t=1}^{T} \left[1 - \Phi\left(\frac{\boldsymbol{x}_{it}\boldsymbol{\beta} + c}{\sigma_u}\right)\right]^{1[y_i=0]}$$
$$\left[\phi\left(\left(y_{it} - \boldsymbol{x}_{it}\boldsymbol{\beta} - c\right)/\sigma_u\right)/\sigma_u\right]^{1[y_i=1]} \left(1/\sigma_c\right) \phi\left(c/\sigma_c\right) dc.$$

This model can be estimated using the **xttobit** command in Stata.

4.2 Modelling the random effect as a function of x-variables

The assumption that c_i and $\boldsymbol{x}_{it}$ are independent is unattractive. Just like for the probit model, we can adopt a Mundlak-Chamberlain approach and specify c_i as a function of observables, eg.

$$c_i = \psi + \bar{\boldsymbol{x}}_i \boldsymbol{\xi} + a_i.$$

This means we rewrite the panel tobit as

$$y_{it} = \max\left(0, \boldsymbol{x}_{it}\boldsymbol{\beta} + \psi + \bar{\boldsymbol{x}}_i\boldsymbol{\xi} + a_i + u_{it}\right),$$

$$u_{it}|\boldsymbol{x}_{it}, a_i \sim Normal\left(0, \sigma_u^2\right).$$

From this point, everything is analogous to the probit model (except of course the form of the likelihood function, which will be tobit and not probit) and so

there is no need to go over the estimation details again. Bottom line is that we can use the xttobit command and just add individual means of time varying x-variables to the set of regressors. Partial effects of interest evaluated at the mean of c_i are easy to compute, since

$$E\left(c_i\right) = \psi + E\left(\bar{\boldsymbol{x}}_i\right)\boldsymbol{\xi}.$$

4.3 Dynamic Unobserved Effects Tobit Models

Model:

$$y_{it} = \max\left(0, \rho y_{i,t-1} + \boldsymbol{z}_{it}\boldsymbol{\delta} + c_i + u_{it}\right),$$

$$u_{it}|\boldsymbol{z}_{it}, y_{i,t-1}, ..., y_{i0}, c_i \sim Normal\left(0, \sigma_u^2\right).$$

Notice that this model is most suitable for corner solution outcomes, rather than censored regression (see Wooldridge, 2002, for a discussion of this distinction) - this is so because the lagged variable is observed $y_{i,t-1}$, not latent $y^*_{i,t-1}$. The discussion of the dynamic RE probit applies in the context of the dynamic RE tobit too. The main complication compared to the nondynamic model is that there is an initial conditions problem: y_{i0} depends on c_i. Fortunately, we can use Heckman's (1981) approach or (easier) Wooldridge's approach. Recall

that the latter involves assuming

$$c_i = \psi + \xi_0 y_{i0} + \boldsymbol{z}_i \boldsymbol{\xi} + a_i,$$

so that

$$y_{it} = \max\left(0, \rho y_{i,t-1} + \boldsymbol{z}_{it}\boldsymbol{\delta} + \psi + \xi_0 y_{i0} + \boldsymbol{z}_i \boldsymbol{\xi} + a_i + u_{it}\right).$$

We thus add to the set of regressors the initial value y_{i0} and the entire vector $\boldsymbol{z}_i$ (note that these variables will be "time invariant" here), and then estimate the model using the xttobit command as usual. Interpretation of the results and computation of partial effects are analogous to the probit case.

5 Sample selection panel data models

Model:

$$y_{it} = \boldsymbol{x}_{it}\boldsymbol{\beta} + c_i + u_{it}, \qquad \text{(Primary equation)}$$

where selection is determined by the equation

$$s_{it} = \left\{ \begin{array}{ll} 1 & \text{if } \boldsymbol{z}_{it}\boldsymbol{\gamma} + d_i + v_{it} \geq 0 \\ 0 & \text{otherwise} \end{array} \right\}. \qquad \text{(Selection equation)}$$

Assumptions regarding unobserved effects and residuals are as for the RE tobit-

- If selection bias arises because c_i is correlated with d_i, then estimating the main equation using a fixed effects or first differenced approach on the selected sample will produce consistent estimates of $\boldsymbol{\beta}$.

- However, if $corr\left(u_{it}, v_{it}\right) \neq 0$, we can address the sample selection problem using a panel Heckit approach. Again, the Mundlak-Chamberlain approach is convenient - that is,

  - Write down specifications for c_i and d_i and plug these into the equations above

  - Estimate T different selection probits (i.e. do not use xtprobit here, use pooled probit). Compute T inverse Mills ratios.

  - Estimate

$$y_{it} = \boldsymbol{x}_{it}\boldsymbol{\beta} + \boldsymbol{x}_i\boldsymbol{\phi} + D_1\rho_1\hat{\lambda}_1 + ... + D_T\rho_T\hat{\lambda}_T + e_{it},$$

    on the selected sample. This yields consistent estimates of $\boldsymbol{\beta}$, provided the model is correctly specified.