① a) We have ε_{it} iid $\mathcal{N}(0,1)$ conditional on X

$$\text{in} \begin{cases} y_{it}^* = x_{it}'\beta + \varepsilon_{it} \\ y_{it} = \mathbb{1}_{y_{it}^* > 0} \end{cases}$$

then $\mathbb{P}(y_{it}=1 \mid X) = \mathbb{P}(y_{it}^* > 0 \mid X)$

$$= \mathbb{P}(y_{it}^* > 0 \mid x_{it})$$
$$= \mathbb{P}(\varepsilon_{it} > -x_{it}'\beta \mid x_{it})$$
$$= \Phi(x_{it}'\beta)$$

Thus this model can be fitted by MLE:

$$\hat{\beta} = \underset{b}{\text{argmax}} \sum_{i=1}^{n} \sum_{t=1}^{T_i} \left[y_{it} \ln \Phi(x_{it}'b) + (1-y_{it}) \ln\left(1-\Phi(x_{it}'b)\right) \right]$$

There are three main assumptions on ε_{it}:

① ε_{it} is independent over i,t. This means that the omitted variables should not display any time correlation, even for the same firm i. In particular, there is no time invariant unobserved heterogeneity. This is unlikely to be true as paying dividends may depend on unobservables of the firm (eg management practice) that are difficult to observe –

② We assume that ε_{it} are homoskedastic ie $V(\varepsilon_{it} \mid X) = 1$ for all i,t. This may not be reasonable if there is a lot of heterogeneity across firms.

③ We assume a specific pdf $\mathcal{N}(,)$ for ε_{it}.

Violations of ①, ② & ③ leads to inconsistent estimators if we use the probit model (pooled).

EC475 Mach 2010

① b) xtprob in Stata

Assume $\begin{cases} y_{it}^* = x_{it}'\beta + \varepsilon_{it} \\ y_{it} = \mathbb{1}_{y_{it}^* > 0} \\ \varepsilon_{it} = \alpha_i + \nu_{it}. \end{cases}$

We have a Balanced panel N observations of individuals for T periods.

The general formula is:

$$\mathbb{P}(y_i | X_i) = \mathbb{P}(y_{i1} \dots y_{iT} | X_i) = \underbrace{\int \dots \int}_{T} \dots \int_{a_{i1}}^{b_{i1}} f(\varepsilon_{i1}, \dots \varepsilon_{iT})\, d\varepsilon$$

where $\begin{cases} \text{if } y_{ij} = 0, & a_{ij} = -\infty, \quad b_{ij} = -x_{ij}'\beta \\ \text{if } y_{ij} = 1, & a_{ij} = -x_{ij}'\beta, \quad b_{ij} = +\infty \end{cases}$

This formula simplifies here if we assume that the observations (ν_{it}) are iid over t conditionnal on X_i, α_i. In xtprob, we assume $\nu_{it} \sim \mathcal{N}(0,1) \,|\, X_i, \alpha_i$

Then $\mathbb{P}(y_i | X_i) = \mathbb{E}\left(\mathbb{P}(y_i | X_i, \alpha_i)\right) = \int_{\alpha} \prod_{t=1}^{T} \mathbb{P}(y_{it} | X_i, \alpha_i) f_{\alpha_i}(\alpha) d\alpha$

We further assume that $\alpha_i | X_i \sim \mathcal{N}(0, \sigma_\alpha^2)$ iid over i.

Thus

$$\mathbb{P}(y_i | X_i) = \int_{-\infty}^{+\infty} \prod_{t=1}^{T} \mathbb{P}(y_{it} | X_i, \alpha_i) \frac{1}{\sigma_\alpha} \varphi\left(\frac{\alpha}{\sigma_\alpha}\right) d\alpha.$$

with $\mathbb{P}(y_{it} | X_i, \alpha_i) = \mathbb{P}(y_{it} | X_{it}, \alpha_i) = \Phi(x_{it}'\beta + \alpha_i)^{y_{it}} \times \left(1 - \Phi(x_{it}'\beta + \alpha_i)\right)^{1-y_{it}}$

The observations are independent over i (individuals) so we get: (conditional on X):

$$L = \prod_{i=1}^{N} \left(\int_{-\infty}^{+\infty} \prod_{t=1}^{T} \Phi(x_{it}'\beta + \alpha)^{y_{it}} \left(1 - \Phi(x_{it}'\beta + \alpha)\right)^{1-y_{it}} \frac{1}{\sigma_\alpha} \varphi\left(\frac{\alpha}{\sigma_\alpha}\right) d\alpha \right)$$

① c) Suppose now that we have:

$$\begin{cases} y_{it}^* = x_{it}'\beta + \varepsilon_{it} \\ y_{it} = \mathbb{1}\{y_{it}^* > 0\} \\ \varepsilon_{it} = \alpha_i + v_{it} \\ v_{it} = \rho v_{it-1} + \xi_{it} \end{cases} \qquad \xi_{it} \text{ iid. over } i,t \sim \mathcal{N}(0,1)$$

Then as before we can write:

$$\mathbb{P}(y_i|X_i) = \mathbb{P}(y_{i1} \dots y_{iT}|X_i) = \underbrace{\int\int_{a_j}^{b_j} \dots \int}_{T} f(\varepsilon_{i1}, \dots \varepsilon_{it})\, d\varepsilon.$$

where $\begin{cases} \text{if } y_{ij}=1, \; a_j = -x_{it}'\beta, \; b_j = +\infty \\ \text{if } y_{ij}=0, \; a_j = -\infty, \; b_j = -x_{it}'\beta \end{cases}$

Here this T dimensional integral does not simplify as the value of ε_{it} depends both of the fixed heterogeneity α_i and the past value of v_{it}, v_{it-1}. (

We need to use a simulation based procedure.

× The parameters of interest are: $\theta \begin{cases} \beta, \alpha \text{ the se of } \alpha_i, \\ \rho \text{ the autocorrelation factor of } v_{it}. \end{cases}$

× We could use SSTIL and define:

$$\hat{\theta} = \underset{\theta}{\text{argmax}} \frac{1}{N} \sum_{i=1}^{N} \ln \tilde{l}_i(\theta, R) = \underset{\theta}{\text{argmax}}\, \tilde{LL}(\theta)$$

where R is the number of simulations.

× We will need to iterate over our choices for θ, eg $\theta^{(n)}$

at $\theta^{(n)}$ we need to evaluate

$$\tilde{LL}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \ln \tilde{l}_i(\theta, R)$$

× ① We draw a T dimensional uniform random vectors R times $\Rightarrow$ $\underset{T\times 1}{\tilde{u}_{i1}}, \dots \underset{T\times 1}{\tilde{u}_{iR}}$

× ② Given that we know $\theta^{(n)}$ we define:

$$\tilde{\varepsilon}_i^r(\theta^{(n)}) = F_\varepsilon^{-1}(\tilde{u}_{iR}, \theta^{(n)})$$

thanks to our assumptions about the pdf of ε_i, F_ε.

× (3) Knowledge of $\tilde{\varepsilon}_i^r(\theta^{(m)})$ and suit gives us a simulated latent var $\tilde{y}_{it}^*(\theta^{(m)})$

and $\tilde{y}_{it}(\theta^{(m)}) = \mathbb{1}(\tilde{y}_{it}^* > 0)$

× (4) From our R simulations we can get the simulated likelihood $\tilde{l}_i(\theta^{(m)}, R)$ of a given vector of choices $y_i = (y_{i1} \cdots y_{iT})$ and thus compute $\tilde{LL}(\theta^{(m)})$

× (5) Keeping fixed our simulations (ie keeping the same R $T \times 1$ vectors $\tilde{u}_{ir}$). We can iterate (2), (3) and (4) to get the values of $\tilde{LL}(\theta^{(m)})$ and search for θ satisfying this objective criterion.

(1) d) unbalanced panel data, now we have T_i observations by individual i and T_i is not constant.

in a) this does not matter at all as we assumed ε_{it} iid over i and t cond on X. (if unbalancing exog).

in b) the likelihood of a given individual i becomes $\mathbb{P}(y_i|x_i) = \int_\alpha \prod_{t=1}^{T_i} \mathbb{P}(y_{it}|x_i, \alpha_i) f(\alpha) d\alpha$

if the process that determines the number of observations is exogenous

in c) same as in b we will need to adjust for ε_i $T_i \times 1$ of different sizes if the observations are not exogenous.

Here we have a panel of firms and balancing is unlikely to be exogenous (firms' closure), given that we are interested in the financial "health" of the firms.

ec475 ① d) we will need to model the process that determines the balancy or not if we think that it depends on the financial health of the firm. (5)

(e) S observations $y = X\beta + \varepsilon$

$$\begin{cases} A_1 \; r(X) = k \\ A_2 \; y_i = X_i'\beta + \varepsilon_i, \; E(\varepsilon_i) = 0 \\ A_3 \; E(\varepsilon_i | X) = 0 \\ A_4 \; E(\varepsilon\varepsilon' | X) = \sigma_\varepsilon^2 I_S \\ A_5 \; \varepsilon | X \sim \text{pdf}(0, \sigma_\varepsilon^2 I_S) \end{cases}$$

a) $\varepsilon | X \sim \mathcal{N}(0, \Omega)$ with $\Omega = \begin{pmatrix} \sigma_1^2 & \cdots & 0 \\ 0 & \cdots & \sigma_S^2 \end{pmatrix}$

The assumption A_4 is violated.
The OLS estimator is unbiased (and consistent under additional assumptions) and no longer BLUE because the disturbance term is not homoskedastic. If we know $\sigma_1^2, \ldots \sigma_S^2$ we could use GLS to obtain a BLUE estimator. If we do not know them we could use FGLS to obtain an asymptotically efficient estimator.

b) $y_t = \beta_1 + \beta_2 y_{t-1} + \cdots + \rho \underbrace{\varepsilon_{t-1} + \nu_t}_{\varepsilon_t}$

Then clearly A_3 is violated $E(\varepsilon_t | X) \neq 0$ as ε_t depends on ε_{t-1} as y_{t-1}.

2) b) as A_3 is invalid OLS is not unbiased. moreover A_3 s/u $E(\varepsilon_s x_s) = 0$ is also violated so OLS is not even consistent.

c) This violates A_2, we have in general:

$$y_i = F(x_i, \varepsilon_i, \beta)$$

where F is a non linear function.

Then if unobserved heterogeneity is additive

$$y_i = F(x_i, \beta) + \varepsilon_i$$

we could use NLLS or MLE methods.

In general, OLS will not estimate the true parameters β.

d) We have: $y_s^* = x_s \beta + \varepsilon_s$

and we observe $y_s = \begin{cases} y_s^* \text{ if } y_s^* \geq 1000 \\ \cdot \text{ otherwise.} \end{cases}$

This is a truncated sample.

Assuming $\varepsilon | X \sim N(0, \sigma_\varepsilon^2 I_s)$

we get $E(y_s^* | y_s^* \geq 1000, x_s) = x_s' \beta + E(\varepsilon_s | y_s^* \geq 1000, x_s)$

$E(\varepsilon_s | \varepsilon_s \geq 1000 - x_s'\beta, x_s)$

$$\sigma_\varepsilon^2 \left(\frac{\phi\left(\frac{1000 - x_s'\beta}{\sigma_\varepsilon^2}\right)}{1 - \Phi\left(\frac{1000 - x_s'\beta}{\sigma_\varepsilon^2}\right)} \right)$$

Again clearly A_2 is violated.

OLS will be biased and inconsistent for β. As in our sample, we need to model $E(y_s | x_s)$

$= E(y_s^* | x_s, y_s^* \geq 1000)$

and not $E(y_s^* | x_s)$.

We could use NLLS or MLE.