QualDep Models in TSCS

I. Introduction

A. TSCS considerations arise for limited/qualitative-dependent-variable (*QualDep*) models also…

B. Non-separability of stochastic from systematic components in these models renders the proper address of these considerations considerably more complicated.

C. Methods have been developed for binary, polychotomous, censored, truncated, ordered, count, duration, etc. models.

D. *Stimson's Law*: You can only solve one hard problem at a time, and solving it requires ignoring lots of other problems.

II. Binary Dependent-Variable Models

$$y^*_{i,t} = \mathrm{x}_{i,t}\beta + \epsilon_{i,t}$$

$$y_{i,t} = 1 \text{ if } y_{i,t} > 0$$

A. Typical Procedure for ML Analysis:

1. Distribution of *DepVar*; Bernouli: $\Pr(y_{it} = \{1,0\}) = p_{it}^{y_{it}} \times (1 - p_{it})^{1-y_{it}}$

2. Model parameters of interest; includes appropriate covariates, **X**, and appropriate functional form ("link function"). For binary (draw):

   a) Sigmoidal (S-Shaped) link, logit: $p = e^{\mathbf{x\beta}}\left(1 - e^{\mathbf{x\beta}}\right)^{-1}$

   b) Sigmoidal (S-Shaped) link, probit: $p = \boldsymbol{\Phi}(\mathbf{x\beta})$

   c) Many, many other sigmoidal functions; some useful ones relax symmetry, steepest at 0.5, etc.

3. Ensure—theoretically/substantively, by good/clever specification of **xβ**—or assume (or pray for)—conditional independence of obs, so

4. Joint likelihood of obs given model & data is product marginals; maximize log-likelihood for parameter estimates; -$\mathbf{H}^{-1}$ are vce.

B. Binary Models with Unit-Specific Effects

The basic set up of the repeated observations model for dichotomous dependent variables is similar to the standard models.

$$y_{it}^* = \beta \mathbf{x}_{it} + \alpha_i + u_{it}, \tag{8.1}$$

where

$$y_{it} = \begin{cases} 1 & \text{if } y_{it}^* > 0 \\ 0 & \text{if } y_{it}^* \leq 0 \end{cases}$$

and $i = 1, \ldots, N$ and $t = 1, \ldots, T$, and u_{it} is assumed to be iid with mean zero and variance σ_u^2.

- The choice we make about the distribution for the disturbance term matters a lot w/ dichotomous dep. vars; still based on our beliefs about the correlation between the explanatory variables and the individual specific effect, but also has implications for estimation approach.

- If $\alpha_i \perp \mathbf{x}_{it}$, estimate a random effects model, assuming $\alpha \sim IID(0, \sigma_\alpha^2)$.

- If correlation between α_i and $\mathbf{x}_{it}$, estimate a fixed effects model (again, α_i are fixed parameters to be estimated).

- If $T \to \infty$, then it is possible to get consistent estimates of $\boldsymbol{\beta}$ and α_i.

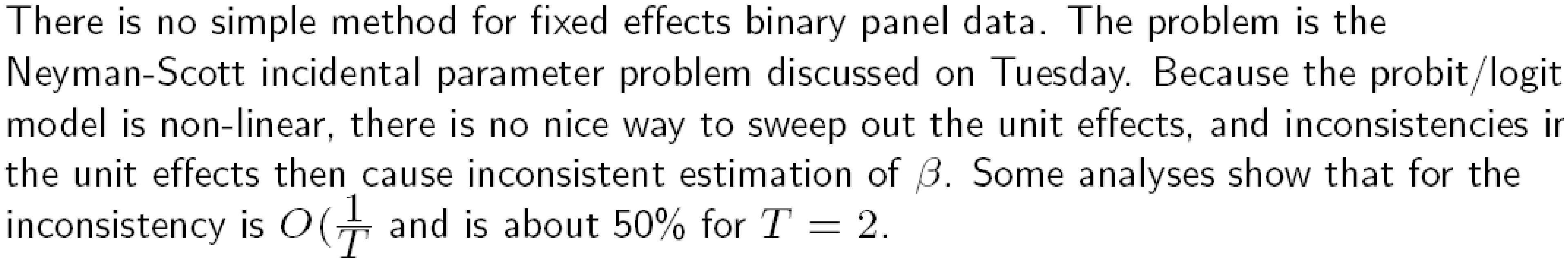
- However, if T is fixed and $N \to \infty$, then we have the incidental parameters problem—i.e., since the number of parameters increases with N, we cannot consistently estimate α_i for fixed T.

- Unfortunately, the inconsistency in α_i is transmitted to $\boldsymbol{\beta}$.

- The transformations that we perform in the linear regression case (viz., subtracting off within-group times means of the variables, differencing) are not valid with a qualitative limited dependent variable model b/c of nonlinearity of such models.

1. Fixed-Effect Logit (a.k.a., Conditional Logit):

a) Chamberlin’s *Clever strategy/insight*: Conditioning on the number of ones, i.e., on $\sum_{t=1}^{T} y_{it}$, works like fixed-effect transformation for binary, i.e., like FE/LSDV, because the FE are conditional means, and, in this context, those are probabilities, which are sums of y_{it}=1 (divided by T). Formally: $\sum_{t=1}^{T} y_{it}$ is a sufficient statistic for α_i.



There is no simple method for fixed effects binary panel data. The problem is the Neyman-Scott incidental parameter problem discussed on Tuesday. Because the probit/logit model is non-linear, there is no nice way to sweep out the unit effects, and inconsistencies ir the unit effects then cause inconsistent estimation of β. Some analyses show that for the inconsistency is $O(\frac{1}{T}$ and is about 50% for $T = 2$.

(1) That is, simple FE is biased in “sigmoidal Hurwicz/Nickell” fashion.

(2) Implications wider-ranging b/c sigmoidal function non-separability.

(3) However, bias also like Hurwicz/Nickell in that of order 1/T, meaning in reasonably long TSCS, may not be an issue worth fretting.

b) Formally, the likelihood for Chamberlin's Conditional Logit is:

$$L = \prod_{i=1}^{N} \Pr\left(y_{i1}, \ldots, y_{iT} | \sum_{t=1}^{T} y_{it} \right)$$

c) How it works/what it does, is easiest explained & seen in *T*=2 case:

$$L = \prod_{i=1}^{N} \Pr(y_{i1}) \Pr(y_{i2})$$

- Note that:

$$\Pr[y_{i1} = 0, y_{i2} = 0 | y_{i1} + y_{i2} = 0] = 1$$
$$\Pr[y_{i1} = 1, y_{i2} = 1 | y_{i1} + y_{i2} = 2] = 1$$

which means these probabilities add no information to the conditional log likelihood so we can ignore them.

- But

$$\Pr[y_{i1}=0,y_{i2}=1|y_{i1}+y_{i2}=1]=\frac{\Pr[y_{i1}=0,y_{i2}=1 \text{ and } y_{i1}+y_{i2}=1]}{\Pr[y_{i1}+y_{i2}=1]}$$

$$=\frac{\Pr[y_{i1}=0,y_{i2}=1 \text{ and } y_{i1}+y_{i2}=1]}{\Pr[y_{i1}=0,y_{i2}=1]+\Pr[y_{i1}=1,y_{i2}=0]}$$

d) Where the second denominator equals the first because the other two cases (both observations =1) has been discarded (conditioned away).

- If we assume that the data follow a logistic distribution then we can rewrite this as

$$\frac{\frac{1}{1+\exp(\alpha_i+\boldsymbol{\beta}'\mathbf{x}_{i1})}\frac{\exp(\alpha_i+\boldsymbol{\beta}'\mathbf{x}_{i2})}{1+\exp(\alpha_i+\boldsymbol{\beta}'\mathbf{x}_{i2})}}{\frac{1}{1+\exp(\alpha_i+\boldsymbol{\beta}'\mathbf{x}_{i1})}\frac{\exp(\alpha_i+\boldsymbol{\beta}'\mathbf{x}_{i2})}{1+\exp(\alpha_i+\boldsymbol{\beta}'\mathbf{x}_{i2})}+\frac{\exp(\alpha_i+\boldsymbol{\beta}'\mathbf{x}_{i1})}{1+\exp(\alpha_i+\boldsymbol{\beta}'\mathbf{x}_{i1})}\frac{1}{1+\exp(\alpha_i+\boldsymbol{\beta}'\mathbf{x}_{i2})}}$$

which simplifies to

$$\frac{\exp(\boldsymbol{\beta}'\mathbf{x}_{i2})}{\exp(\boldsymbol{\beta}'\mathbf{x}_{i1})+\exp(\boldsymbol{\beta}'\mathbf{x}_{i2})}$$

$$\frac{\exp(\boldsymbol{\beta}'\mathbf{x}_{i2})}{\exp(\boldsymbol{\beta}'\mathbf{x}_{i1}) + \exp(\boldsymbol{\beta}'\mathbf{x}_{i2})}$$

e) …is essentially a multinomial-logit form.

f) The other expression is: $e^{\mathbf{x}_{i1}\boldsymbol{\beta}} / \left(e^{\mathbf{x}_{i1}\boldsymbol{\beta}} + e^{\mathbf{x}_{i2}\boldsymbol{\beta}}\right)$

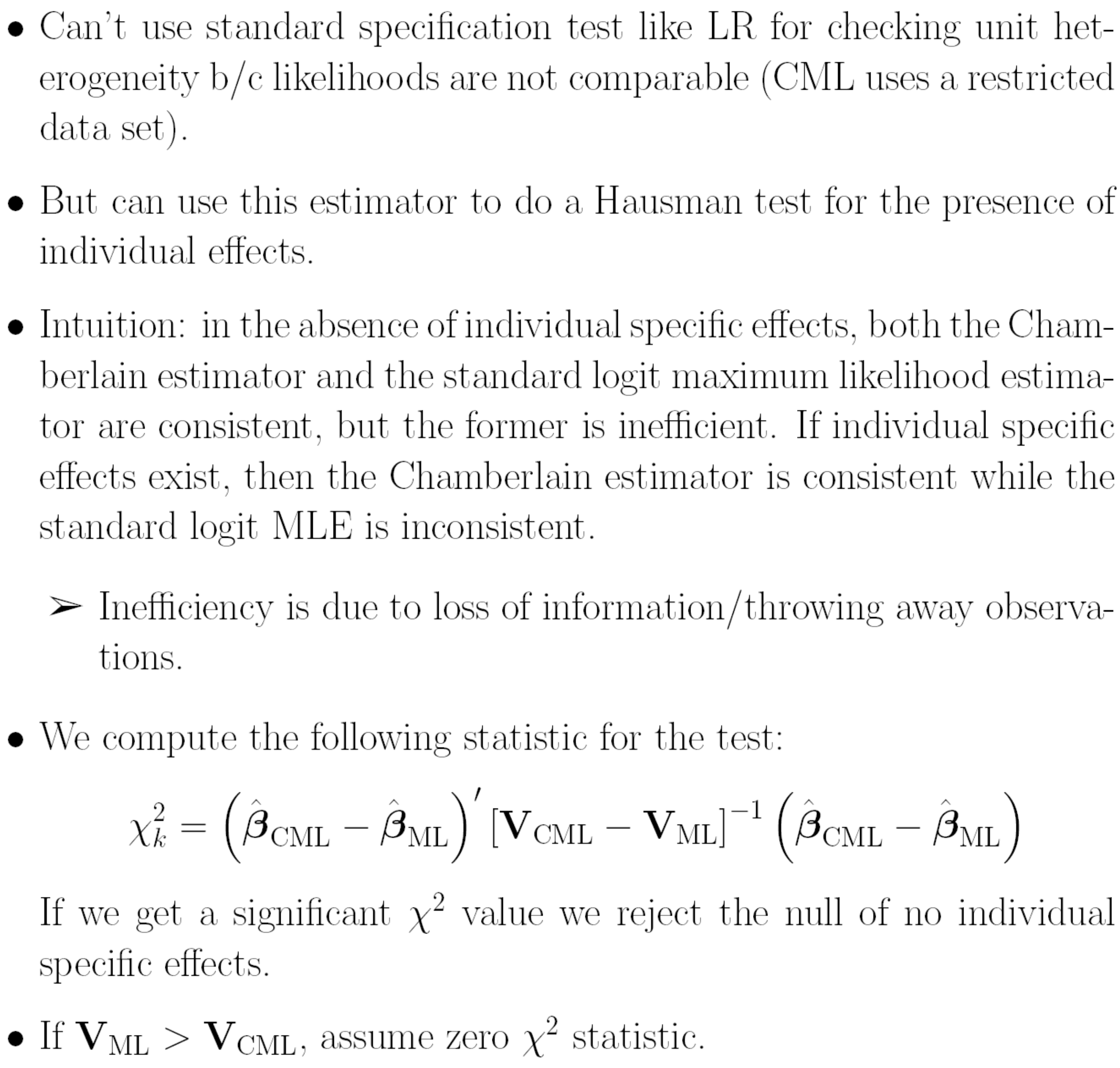
g) The likelihood to maximize takes these (rel'ly familiar) MNL forms.

- This can be extended to T of arbitrary size but the computations are excessive for $T > 10$.

h) i.e., number of orderings of 0's & 1's increases combinatorically in *T*.

i) Note: this at least analogous if not identical to the strategy & issues in network ERGM's (Exponential Random Graph Models).

j) Wawro’s Commentary:



- Can’t use standard specification test like LR for checking unit heterogeneity b/c likelihoods are not comparable (CML uses a restricted data set).

- But can use this estimator to do a Hausman test for the presence of individual effects.

- Intuition: in the absence of individual specific effects, both the Chamberlain estimator and the standard logit maximum likelihood estimator are consistent, but the former is inefficient. If individual specific effects exist, then the Chamberlain estimator is consistent while the standard logit MLE is inconsistent.

  ➢ Inefficiency is due to loss of information/throwing away observations.

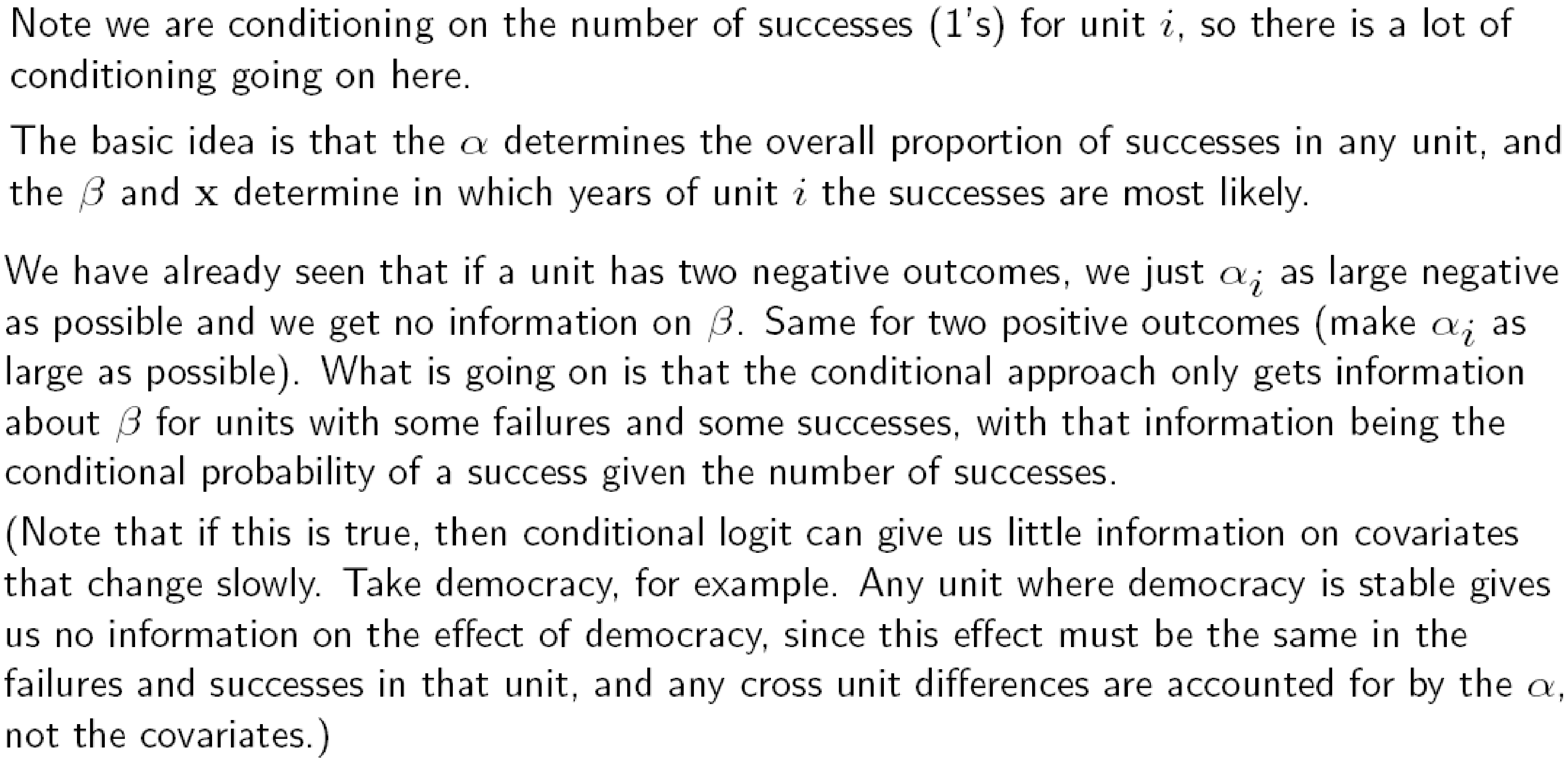
- We compute the following statistic for the test:

$$\chi_k^2 = \left(\hat{\boldsymbol{\beta}}_{\text{CML}} - \hat{\boldsymbol{\beta}}_{\text{ML}}\right)' \left[\mathbf{V}_{\text{CML}} - \mathbf{V}_{\text{ML}}\right]^{-1} \left(\hat{\boldsymbol{\beta}}_{\text{CML}} - \hat{\boldsymbol{\beta}}_{\text{ML}}\right)$$

  If we get a significant χ^2 value we reject the null of no individual specific effects.

- If $\mathbf{V}_{\text{ML}} > \mathbf{V}_{\text{CML}}$, assume zero χ^2 statistic.

k) Beck’s Commentary:



Note we are conditioning on the number of successes (1’s) for unit i, so there is a lot of conditioning going on here.

The basic idea is that the α determines the overall proportion of successes in any unit, and the β and $\mathbf{x}$ determine in which years of unit i the successes are most likely.

We have already seen that if a unit has two negative outcomes, we just α_i as large negative as possible and we get no information on β. Same for two positive outcomes (make α_i as large as possible). What is going on is that the conditional approach only gets information about β for units with some failures and some successes, with that information being the conditional probability of a success given the number of successes.

(Note that if this is true, then conditional logit can give us little information on covariates that change slowly. Take democracy, for example. Any unit where democracy is stable gives us no information on the effect of democracy, since this effect must be the same in the failures and successes in that unit, and any cross unit differences are accounted for by the α, not the covariates.)

NOW IF WE ASSUME THE P's ARE GENERATED STANDARD LOGIT, THIS SIMPLIFIES NICELY AND THE EFFECTS DISAPPEAR

Using the logit form, it is quite easy to write down these joint probabilities (done on Green p. 840), the α_i drop out of this equation and the conditional probability of the sequence (0,1) is just a logit. Nothing deep here - if you know you had one success out of two trials, the only information in the data about β is given by whether the first or second trial was positive, with that probability given to you by a logit BASED ONLY ON THE TWO OBSERVATIONS FOR UNIT i. (Thus conditional logit works for the same reason that with more than two outcomes you can do logit if you assume IIA.)

Note that conditional fixed effects logit is not exactly fixed effects logit, its qualities are asymptotic, and if you have a lot of units with all zeros or all one, you are in trouble. If you like fixed effects for continuous dv's, then this procedure inherits the good from that; if you don't like continuous fe's, it inherits the bad.

Note also that Ethan Katz has shown that as $T \rightarrow \infty$ that standard logit with fixed effects and conditional logit converge (this is well known), but in reality they are very close when $T > 15$. Thus only need Chamberlain for smallish T, for largish T just put in dummy variables for unit and run logit.

And of course, for fe logit, you have to believe that all the information in the data about the β, the parameters of interest, is contained in *when* observations in a unit are zero or one, *not* how many are zero or one. This seems to be throwing a lot of information away (just as in the discussion of Green, Kim and Yoon for dyadic BTSCS data).

2. Random-Effect Probit:

a) In probit, as Beck alluded, conditioning on fixed-effect $\sum_{t=1}^{T} y_{it}$ does not yield a familiar simplified expression (like C/FE-Logit's MNL form). Interestingly, though, RE simplifies better/more/more-easily in probit.

$$y_{it}^{*} = \mathbf{x}_{it}\boldsymbol{\beta} + \alpha_i + \varepsilon_{it} \text{ , with } \alpha_i \sim N\left(0, \sigma_{\alpha}^{2}\right)$$

$$y_{it} = \left\{1 \text{ if } y_{it}^{*} > 0 \text{ ; } 0 \text{ if } y_{it}^{*} < 0\right\}$$

b) Excluding the RE, we would just apply standard-ML probit, multiplying the *NT* marginal likelihoods to obtain the joint likelihood.

c) With random-effects—i.e., with the α_i being random variables—the joint likelihood is the product *N* inseparable *T*-dimensional likelihoods.

d) Note: this is same issue that arises in spatial (& multilevel) *QualDep* models also. The following strategy not work in space, though.

e) That's very hard (computationally intensive) to integrate. Applicable trick here (popularized by Butler and Moffitt, but known before) is to condition on α_i, which makes the *T* densities independent:

Writing $\nu_{i,t} = \alpha_i + \epsilon_{i,t}$, we then get

$$f(\nu_{i,1}, \nu_{i,2}, \ldots, \nu_{i,T}) = \int_{-\infty}^{\infty} f(\nu_{i,1}, \nu_{i,2}, \ldots, \nu_{i,T} | \alpha_i) f(\alpha_i) d\alpha_i$$

$$= \int_{-\infty}^{\infty} \prod_{t=1}^{T} f(\nu_{i,t} | \alpha_i) f(\alpha_i) d\alpha_i$$

f) This now just a product of 1-dimensional marginals, so feasible, but to get there, integrate over the α_i, so still rather intense. `xtprobit`, much slower than `probit`, exponentially so in *T*. *T*>10? impractical.

g) As usual, the REs must be *assumed* independent of **x**'s. Wawro *AJPS* (2001) discusses a correlated-RE logit. It makes some odd/strong assumptions of its own, of course, and is very hard to estimate.

Wawro offers a fuller discussion:

- Let

  $$\varepsilon_{it} = \alpha_i + u_{it}$$

  and assume $\alpha_i \sim N(0, \sigma_\alpha^2)$, $u_{it} \sim N(0, \sigma_u^2)$, and α_i and u_{it} are independent of each other. Then

  $$\text{var}[\varepsilon_{it}] = \sigma_u^2 + \sigma_\alpha^2 = 1 + \sigma_\alpha^2$$

  and

  $$\text{corr}[\varepsilon_{it}, \varepsilon_{is}] = \rho = \frac{\sigma_\alpha^2}{1 + \sigma_\alpha^2}$$

  for $t \neq s$. This implies $\sigma_\alpha^2 = \rho/(1 - \rho)$.

- We can write the probability associated with an observation as

  $$\Pr[y_{it}] = \int_{-\infty}^{q_{it}\boldsymbol{\beta}'\mathbf{x}_{it}} f(\varepsilon_{it}) d\varepsilon_{it} = \Phi[q_{it}\boldsymbol{\beta}'\mathbf{x}_{it}]$$

  where $q_{it} = 2y_{it} - 1$.

- Because of the α_i, the T observations for i are jointly normally distributed. The individual's contribution to the likelihood is

$$L_i = \Pr[y_{i1}, y_{i2}, \ldots y_{iT}]$$

$$= \int_{-\infty}^{q_{i1}\boldsymbol{\beta}'\mathbf{x}_{i1}} \int_{-\infty}^{q_{i2}\boldsymbol{\beta}'\mathbf{x}_{i2}} \cdots \int_{-\infty}^{q_{iT}\boldsymbol{\beta}'\mathbf{x}_{iT}} f(\varepsilon_{i1}, \varepsilon_{i2}, \ldots, \varepsilon_{iT}) d\varepsilon_{iT} \cdots d\varepsilon_{i2} d\varepsilon_{i1}$$

h) First, express the joint as conditionals times marginal (Bayes Law):

$$f(\varepsilon_{i1}, \ldots, \varepsilon_{iT}, \alpha_i) = f(\varepsilon_{i1}, \ldots, \varepsilon_{iT} | \alpha_i) f(\alpha_i)$$

i) Then integrate over the α_i:

$$f(\varepsilon_{i1}, \ldots, \varepsilon_{it}) = \int_{-\infty}^{\infty} f(\varepsilon_{i1}, \ldots, \varepsilon_{iT} | \alpha_i) f(\alpha_i) d\alpha_i$$

- Conditioned on α_i, the ε_is are independent:

$$f(\varepsilon_{i1}, \ldots, \varepsilon_{it}) = \int_{-\infty}^{\infty} \prod_{t=1}^{T} f(\varepsilon_{it}|\alpha_i) f(\alpha_i) d\alpha_i \qquad (8.2)$$

j) Now, write the univariate normals in the product and express as $g(\alpha_i)$:

$$L_i = \int_{-\infty}^{\infty} \frac{1}{\sigma_\alpha \sqrt{2\pi}} e^{-\frac{\alpha_i^2}{2\sigma_\alpha^2}} g(\alpha_i) d\alpha_i$$

k) Then we can change some notation and rearrange and get this into the form the computer actually uses to search for parameter estimates:

- Let $r_i = \frac{\alpha_i}{\sigma_\alpha \sqrt{2}}$, which implies $\alpha_i = \sigma_\alpha \sqrt{2} r_i = \theta r_i$ and $d\alpha_i = \theta dr_i$.
- Making the change of variable gives

$$L_i = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-r_i^2} g(\theta r_i) dr_i \tag{8.4}$$

- Working back to the probit model, we get i's contribution to the likelihood as

$$L_i = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-r_i^2} \left\{ \prod_{t=1}^{T} \Phi[q_{it}(\boldsymbol{\beta}'\mathbf{x}_{it} + \theta r_i)] \right\} dr_i \tag{8.5}$$

- Note that $\theta = \sqrt{\frac{2\rho}{1-\rho}}$.

Wawro's Commentary:

- Things to note:

  - ➢ The assumption that the α_i and $\mathbf{x}_{it}$ are uncorrelated is very restrictive. We are also assuming that the within-cross section correlation is the same across all time periods.

  - ➢ ρ can be interpreted as the proportion of the variance contributed by the unit effects.

  - ➢ We can test for unit heterogeneity by checking the statistical significance of ρ. One way to do this is with a likelihood ratio ratio test of the random effects probit and pooled probit models.

  - ➢ The standard way to evaluate the integral in the likelihood is by Gauss-Hermite quadrature. This raises some concerns about how the size of T and N affect the accuracy of the quadrature approximation, and some checks of the performance of the approximation are in order.

  - ➢ `Stata`'s `xtprobit` command can be used to estimate this model.

  - ➢ We could derive this model for the logistic distribution rather than the normal distribution.

8.8 Binary Time-Series Cross-Section (BTSCS) Data

- The methods above are appropriate when N is large and T is small. Beck, Katz, and Tucker ('98 *AJPS*) derive a method for when T is large.

- The method is based on the observation that BTSCS data is identical to grouped duration data. That is, we get to observe whether an event occurred or not only after the end of some discrete period (e.g., a year).

- Thus, we can use duration methods to correct for the problem of temporal dependence.

- Start from the hazard rate for the continuous time Cox proportional hazard model:

$$\lambda(t) = \exp(\boldsymbol{\beta}'\mathbf{x}_{it})\lambda_0(t)$$

- The survival function is given by

$$S(t) = \exp\left(-\int_0^t \lambda(\tau)d\tau\right)$$

- Assuming we get to observe only whether or not an event occurred between time $t_k - 1$ and t_k, we can write

$$\begin{aligned}\Pr\left(y_{it_k} = 1\right) &= 1 - \exp\left(-\int_{t_k-1}^{t_k} \lambda_i(\tau)d\tau\right) \\ &= 1 - \exp\left(-\int_{t_k-1}^{t_k} \exp(\boldsymbol{\beta}'\mathbf{x}_{it})\lambda_0(t)d\tau\right) \\ &= 1 - \exp\left(-\exp(\boldsymbol{\beta}'\mathbf{x}_{it})\int_{t_k-1}^{t_k} \lambda_0(t)d\tau\right)\end{aligned}$$

- Let

$$\begin{aligned}\alpha_{t_k} &= \int_{t_k-1}^{t_k} \lambda_0(t)d\tau \\ \kappa_{t_k} &= \ln(\alpha_{t_k})\end{aligned}$$

- Then

$$\begin{aligned}\Pr\left(y_{it_k} = 1\right) &= 1 - \exp\left(-\exp(\boldsymbol{\beta}'\mathbf{x}_{it})\alpha_{t_k}\right) \\ &= 1 - \exp\left(-\exp(\boldsymbol{\beta}'\mathbf{x}_{it} + \kappa_{t_k})\right)\end{aligned}$$

- This is a binary model with a complimentary log-log (cloglog) link. The cloglog link is identical to a logit link function when the probability of an event is small ($< 25\%$) and extremely similar when the probability of an event is moderate ($< 50\%$).

- For ease of application then, Beck, Katz, and Tucker recommend using the logistic analogue

$$\Pr\left(y_{it} = 1|\mathbf{x}_{it}\right) = \frac{1}{1 + \exp(-(\boldsymbol{\beta}'\mathbf{x}_{it} + \kappa_{t-t0}))}$$

  where κ_{t-t0} is a dummy variable marking the length of the sequence of zeros that precede the current observation. For example,

t	1	2	3	4	5	6	7	8	9
y	0	0	0	1	0	1	1	0	0
κ	κ_1	κ_2	κ_3	κ_4	κ_1	κ_2	κ_1	κ_1	κ_2

- The intuition behind why ordinary logit is inadequate for BTSCS data is that it doesn't allow for a nonconstant baseline hazard.

- Including the κ dummies allows duration dependence by allowing for a time-varying baseline hazard.

- To see how the κ dummies are interpretable as baseline probabilities or hazards, note

$$\Pr\left(y_{it} = 1|\mathbf{x}_{it} = 0, t0\right) = \frac{1}{1 + \exp(-\kappa_{t-t0})}$$

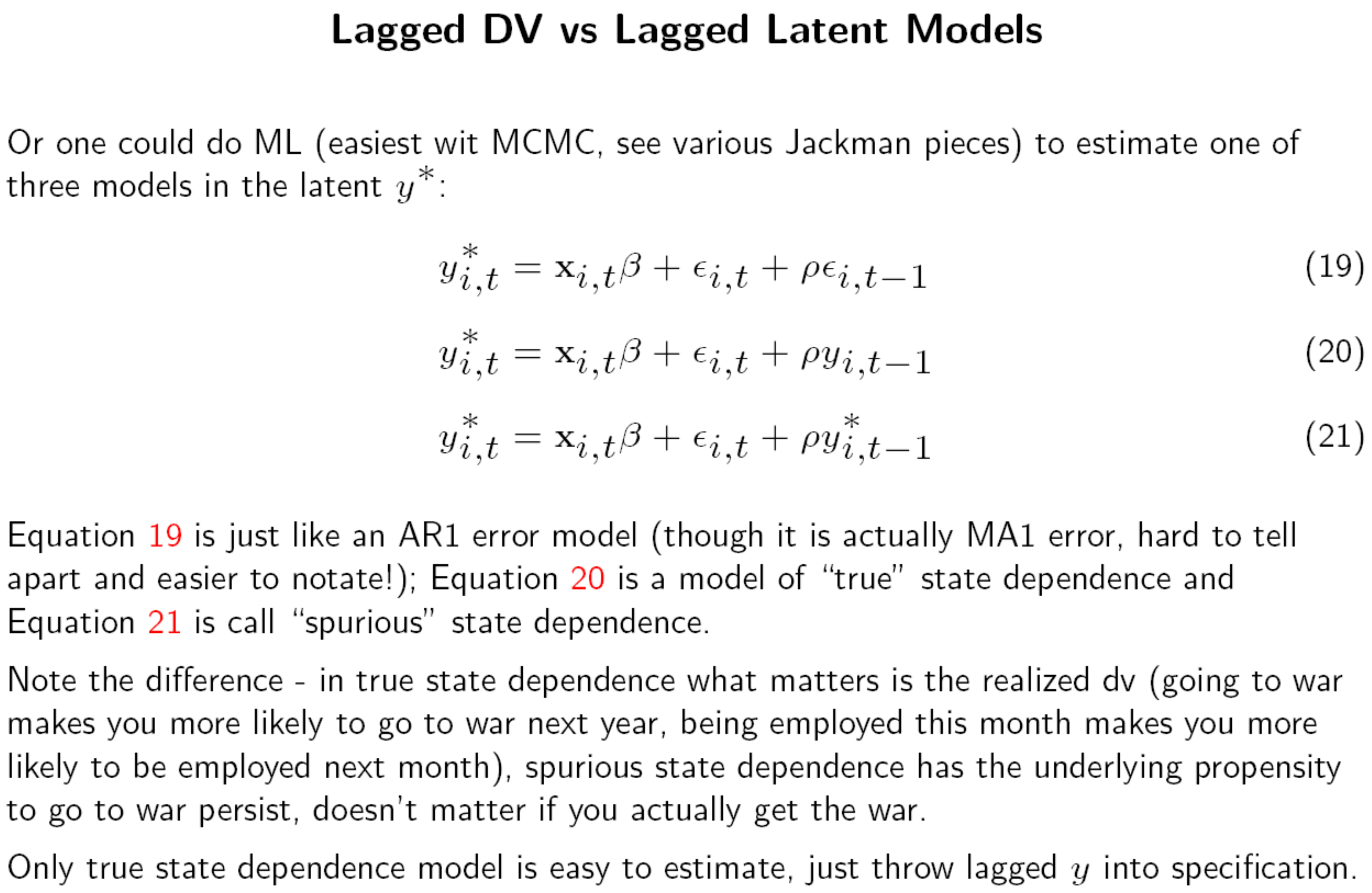
- The κ dummies are essentially time fixed effects that account for duration dependence. Thus when we estimate the model we need to create a matrix of dummies and concatenate it with the matrix of explanatory variables. For the example given above, this matrix would look like

$$\mathbf{K}_i = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}.$$

Note there are 4 columns because the longest spell is 4 periods long.

$$\Pr\left(y_{it} = 1 | \mathbf{x}_{it}\right) = \frac{1}{1 + \exp(-(\boldsymbol{\beta}'\mathbf{x}_{it} + \kappa_{t-t0}))}$$

Note: BKT time-dependence in BTSCS by time dummy/splines is a kludge…



Lagged DV vs Lagged Latent Models

Or one could do ML (easiest wit MCMC, see various Jackman pieces) to estimate one of three models in the latent y^*:

$$y^*_{i,t} = \mathbf{x}_{i,t}\beta + \epsilon_{i,t} + \rho\epsilon_{i,t-1} \tag{19}$$

$$y^*_{i,t} = \mathbf{x}_{i,t}\beta + \epsilon_{i,t} + \rho y_{i,t-1} \tag{20}$$

$$y^*_{i,t} = \mathbf{x}_{i,t}\beta + \epsilon_{i,t} + \rho y^*_{i,t-1} \tag{21}$$

Equation 19 is just like an AR1 error model (though it is actually MA1 error, hard to tell apart and easier to notate!); Equation 20 is a model of "true" state dependence and Equation 21 is call "spurious" state dependence.

Note the difference - in true state dependence what matters is the realized dv (going to war makes you more likely to go to war next year, being employed this month makes you more likely to be employed next month), spurious state dependence has the underlying propensity to go to war persist, doesn't matter if you actually get the war.

Only true state dependence model is easy to estimate, just throw lagged y into specification.

Others are doable, though. Like spatial-probit, except somewhat easier estimate because **W** for time, our **V**, is lower-triangular.