

Longitudinal (Panel and Time Series Cross-Section) Data - Day 3 - Binary Dependent Variables

Nathaniel Beck
Department of Political Science
University of California, San Diego
La Jolla, CA 92093
beck@ucsd.edu
<http://weber.ucsd.edu/~nbeck>

Summer, 2001

Models with Binary Dependent Variables

So far have assumed continuous dependent variable.

Now assume binary dependent variable.

There exist some extensions to more complicated limited dv's, but these are not well worked out, and, when they exist, are direct extensions of binary analyses.

Typically in this work we ignore many of the complications of the last few days

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

Many analysts have ignored all problems, panel and otherwise, and have just done logit.

(Or probit, by and large one can freely interchange, will make clear where cannot)

Logit more common, so use that if have free choice

BEGIN WITH TSCS DATA (BTSCS!) THEN IN SECOND PART OF CLASS DO PANEL DATA

What goes wrong with BTSCS data

What is the problem?

Just as in standard TSCS, observations from same unit are not independent. It is usually easiest to use a standard latent variables setup to think about issues. Here is how to look at some dynamics for a BTSCS model (will perhaps become clearer tomorrow).

In Beck and Katz (1997) we run some simulations of the following form:

$$y_{i,t}^* = \mathbf{x}_{i,t}\beta + \epsilon_{i,t} \quad (1a)$$

$$y_{i,t} = 1 \text{ if } y_{i,t}^* > 0 \quad (1b)$$

where

$$x_{i,t} = \rho_x x_{i,t-1} + \nu_{i,t} \quad (2)$$

$$\epsilon_{i,t} = \rho_\epsilon \epsilon_{i,t-1} + \mu_{i,t} \quad (3)$$

varying the two ρ 's.

We found that with very high ρ (both) the probit standard errors underestimated variability by almost 50% in some cases, and that with small T 's the mean squared error of the estimated $\hat{\beta}$ could be quite high.

(Under such conditions, Poirier and Ruud's classic article showed that ordinary probit is consistent but not efficient, and se's are wrong, analogous to ignoring serial correlation in standard time series).

Huber grouped standard errors

One solution to unit interdependence is the Huber robust standard errors discussed earlier.

Our simulations for dynamics showed this worked well in producing correct standard errors.

Since this solution seems costless, it can and probably always should be used.

Effects - back to TSCS data!

While as we shall see, effects are very problematic for binary panel data, but all of the technical problems disappear with large T . Thus one could conceivably use fixed effects for BTSCS data.

(Random effects will be hard, because you have to model the effects and the error process so that one can get decent solutions. Later we will see random effects probit, which probably will work for BTSCS data, though it was designed for panels.)

Thus we can just estimate a standard logit or probit where the latent form is

$$y_{i,t}^* = \mathbf{x}_{i,t}\beta + f_i + \epsilon_{i,t} \quad (4)$$

and the rest is as in any logit/probit.

This has all the good and bad features of fixed effects models discussed on Tues.

It also has a similar test that all the f_i are equal, though this time is a LR test rather than an F -test.

This method could work if the covariates are not all constant or slowly changing.

As with continuous dv's, fixed effects have costs and benefits, and these should be assessed.

In the controversy with Green et al. in IO, the issue is not the general suitability of fixed effects, but rather how they perform in a specific context, IR dyad-year data.

Back to Event History

Much BTSCS data in political science is IR data generated by dyad-year data sets. These are observations on pairs of nations over many (50-100) years, with the dependent variable being something like whether the dyad was in conflict in that year.

Depending on how we choose our dyads, we might have between 800 and 2000 dyads, each observed for 50 years. How should we analyze this to take account of the unit dependence.

One way to proceed is to note that BTSCS dyad-year data looks like event history data, with each 1 marking a failure and the time between 1's, that is, the number of 0's, being the time until or between failures. This works just fine so long as the number of 1's is not large, so we have a good distribution of failure times.

We could thus simply convert the binary TSCS data into event history data and use the methods we have already learned. The Cox semi-parametric approach is most appropriate (unless we have some idea about the shape of the hazard function).

Thinking about binary TSCS data as event history data gives us certain insights.

- The simple probit/logit approach is equivalent to the assumption of no duration dependence in event history analysis. Thus if we are willing to assume no duration dependence and use an exponential, then the simple probit/logit model is fine. But normally we test for duration dependence, no such test for logit/probit. (Re-examine Berry and Berry.)
- While it looks like we have $N \times T$ binary TSCS observations, this is the same as N duration observations. Thus even if $N \times T$ is large, we may not have all that much information. This is of particular importance in evaluating some critiques of the democratic peace analyses.

- While we think of probit/logit as having troubles with rare events, such rare events are the lifeblood of event history analysis. Non-rare events yield very short durations, which are difficult; rare events yield a wide variety of durations, which is good.
- Binary TSCS data in event form may have more than one event per unit. In the probit/logit setup, we assume that second and subsequent events can be modeled just like first events; events history modelers realize that life is more complex. Thus we have both duration dependence for time until failure, and complexities involved in modeling subsequent failures for the same unit.

Discrete Time Duration Models

Sticking with one event per unit for now, we can model the time until an event in binary TSCS data by a discrete time duration model. These are in many ways simpler than continuous time models.

Let us assume we have time measured in equal discrete intervals, $0, 1, \dots, t, \dots$. We only observe whether someone dies in the interval $(t - 1, t]$, (open on the left, closed on the right) and will assume this is a death at t . Then we need discrete time analogues of the survivor and hazard function. The survivor function will simply be a step function, with steps at $1, 2, \dots, t, \dots$.

Let y_i be the duration for the i 'th unit; y_i is a discrete random variable, with support at the positive integers.

Using the same notation as for continuous time (note that we have to be careful here differentiating $y > t$ from $y \geq t$ since we are in discrete time)

$$\begin{aligned}
 S(t) &= P(y > t) \\
 &= P(y > t | y > t - 1)P(y > t - 1) \\
 &= \prod_{i=0}^{t-1} P(y > t - i | y > t - i - 1)
 \end{aligned} \tag{5}$$

where $S(0) = P(y > 0) = 1$. We still have

$$F(t) = 1 - S(t). \tag{6}$$

Since F is discrete, we have an associated discrete density with support on the positive integers,

$$f(t) = F(t) - F(t - 1). \tag{7}$$

Here the density is a probability; it is the unconditional probability of dying in the interval $(t - 1, t]$.

We can define the discrete hazard analogously to the continuous time hazard, though we can now use probability interpretations. Letting $h(t)$ be the hazard of dying at time t (or in the interval $(t - 1, t]$ given that one survived until $t - 1$, we have

$$h(t) = \frac{f(t)}{S(t - 1)}. \quad (8)$$

Since $1 - h(t)$ is the conditional probability of surviving at t given survival through $t - 1$, substituting in Equation 5, we get

$$S(t) = \prod_{i=0}^{t-1} [1 - h(t - i)] \quad (9)$$

Note that we could have unequal interval lengths which just makes the notation more complicated (and is of little practical interest).

We also could start with a continuous time model, and then use integrated hazards to produce discrete quantities (so that the discrete hazard is the integrated hazard from $t - 1$ to t and so forth). This is the way Sueyoshi (JAE, 1996) proceeds. But it is not obvious that there is a practical gain to this other than to show the obvious underlying unity of discrete and continuous time models. Note that a continuous time Weibull will induce a discrete time “Weibull” and so forth. Note also that the Cox proportional hazards model is really a discrete time model.

Estimation of discrete duration models via logit

If we estimate a binary dependent variable model with dependent variable $y_{i,t}$ being whether unit i failed in the interval $(t - 1, t]$ (by probit, logit or any other binary model), as a function of covariates $\mathbf{x}_{i,t}$, we are estimating a model for $h(t)$. (If the dependent variable is scored as 1 for non-failure, then we have a model for $1 - h(t)$). Thus if we start with a logit or probit model we induce a survival model (or, as Sueyoshi or Alt, King and Signorino do, we can start with a survival model which induces the binary model).

If we estimate a simple binary dependent variable model (for shorthand I say “logit”), with the $x_{i,t}$ as independent variables, we are assuming the hazard rate is time invariant (that is,

$$h_{i,t} = h(\mathbf{x}_{i,t}) \quad (10)$$

so that the relationship between the covariates and the conditional probability of failure is the same at all time points) or that the failure process is memoryless.

These are the same set of assumptions that led to the exponential. While the simple independent logit is not the exact discrete time analogue of the exponential, the two models are very similar.

To allow for duration dependency we would need to estimate a binary model (with $y_{i,t}$ being one for the failure of unit i in the interval $(t - 1, t]$ which has

$$h_{i,t} = h_t(\mathbf{x}_{i,t}). \quad (11)$$

But this is too general. A possible compromise to allow for different intercepts at each time point, so the hazard changes with time. Thus we might have

$$h_{i,t} = a_t + h(\mathbf{x}_{i,t}) \quad (12)$$

which would be estimated by putting in a period dummy in the logit. But these are perhaps too jagged, so we might be better off using

$$h_{i,t} = s(t) + h(\mathbf{x}_{i,t}) \quad (13)$$

where s is some function of time. In my own work I model s as a “smoothing spline” but a natural spline or even a polynomial in time would probably do. Note that one just adds this term to the basic logit, so easy enough to do.

One could also just use the time dummies, which is easier if a bit less elegant.

Remember, *the time variable is time since the last “event,” not the particular period of the observation.*

Sometimes the time dummies or the spline indicate that we don’t need to include time in the specification (using standard tests on the coefficients of all the time variables). At that point we can assume no duration dependence and use ordinary logit.

Or the spline may look linear, in which case we can just use time in the logit (again remembering that time is time since the last event).

A more formal derivation

So far the approach has been informal. We can formalize it by thinking of the binary TSCS data as having come from a *grouped time* duration model, where grouping is by year.

Start with a continuous time Cox proportional hazards model, so

$$h_i(t) = h_0(t)e^{\mathbf{x}_{i,t}\beta}. \quad (14)$$

Letting $S(t)$ be the probability of surviving beyond t , by the math of hazard rates we have basic identity that

$$S(t) = \exp\left(-\int_0^t h(\tau)d\tau\right). \quad (15)$$

We only observe whether or not an event occurred between time $t_k - 1$ and t_k (assuming annual data) and are interested in the probability of this event, $P(y_{i,t_k} = 1)$. This probability is one minus the probability of surviving beyond t_k given survival up to $t_k - 1$. Assuming no prior events (so $t_0 = 0$), we then get

$$\begin{aligned} P(y_{i,t_k} = 1) &= 1 - \exp\left(-\int_{t_k-1}^{t_k} h_i(\tau)d\tau\right) \\ &= 1 - \exp\left(-\int_{t_k-1}^{t_k} e^{\mathbf{x}_{i,t_k}\beta} h_0(\tau)d\tau\right) \\ &= 1 - \exp\left(-e^{\mathbf{x}_{i,t_k}\beta} \int_{t_k-1}^{t_k} h_0(\tau)d\tau\right) \end{aligned}$$

(Note that \mathbf{x} is indexed by t_k not τ because we assume that the independent variables are only measured for an entire interval and not for every instant in the interval $t_k - 1$ to t_k .) Since the baseline hazard is unspecified, we can just treat the integral of the baseline hazard as an unknown constant. Defining

$$\begin{aligned} \alpha_{t_k} &= \int_{t_k-1}^{t_k} h_0(\tau)d\tau \text{ and} \\ \kappa_{t_k} &= \log(\alpha_{t_k}) \end{aligned}$$

we then have

$$\begin{aligned} P(y_{i,t_k} = 1) &= 1 - \exp\left(-e^{\mathbf{x}_{i,t_k}\beta} \alpha_{t_k}\right) \\ &= 1 - \exp\left(-e^{\mathbf{x}_{i,t_k}\beta + \kappa_{t_k}}\right) \end{aligned}$$

This is exactly a binary dependent variable model with a cloglog link and the κ (time dummy) terms added.

There is almost no difference in practice between estimated a logit model and a cloglog model. The κ terms are the temporal dummies, which by a smoothness argument we can model as a spline in time.

Further Complications

Thinking about the war data as event history data leads to thinking about other issues. Dyads can fight a number of wars. Here time is coded so that the end of a war (which can last more than one year) starts the clock anew.

But no events history analyst would stop without thinking that perhaps the modeling of subsequent events is different from the first event: the hazards, conditional on a prior event, is different from the hazard conditional on no prior event (in general). Durations of second hospitalizations are longer than durations of first hospitalizations!

This is difficult to model. One solution is to add a variable to the hazard function which counts the number of previous failures. or perhaps to select only those cases up to the first failure and see if that model differs from the overall model.

Multiple spells is a very hot issue now in biostatistics, particularly in modeling hospitalizations for AIDS patients.

Fixed effects in dyad-year data

Green et al, in IO (2001) note that fixed effects would solve a lot of the same problems. How do they work for the dyad-year data.

Note that we must drop from the analysis every unit which never conflicts.

Suppose $y_{i,t} = 0 \forall t$ for some i . Then clearly maximum likelihood will just choose the parameters so the $P(y_{i,t})$ is as small as possible. But since we have fixed effects, just choose f_i as large a negative number ($-\infty$ if you like) as possible. By choosing f_i as large negative as we can, we can drive down the $P(y_{i,t})$ to zero. And note that the data on the i 'th unit do not then affect the estimation of the β .

In the dyad-year data, most dyads never conflict. This means we are throwing out 90% of our data, not good. (Basically, the fixed effects model says that if US and Canada do not fight, it is because of the US-Canada dyad name, not the fact they are democracies.)

Thinking of dyad-year data as event history, would we do an event history analysis with a different intercept for each individual?????

Sometimes fixed effects do make sense, but think about what they are saying before you use them!!!

Unit rather than dyad fixed effects

There are intermediate solutions. We could take each dyad and assign it a fixed effects.

Thus, instead of thinking of a dyad as dyad i and using Equation 4, think of a dyad as dyad ij where each partner is indexed. We could then estimate

$$y_{ij,t}^* = \mathbf{x}_{ij,t}\beta + f_i + f_j + \epsilon_{ij,t} \quad (18)$$

Note, the single dyadic fixed effect just assumes the possibility of an interaction between the fixed effect for i and j whereas Equation 18 assumes an additive effect only. The saving in df's is enormous (a factor of 10 or more in standard data).

We should not simply use fixed effects off the shelf because they are easy to get off the shelf. When using dyadic data, think about dyads.

(The double fixed effects are not for free; they are still highly correlated with slowly moving variables like democracy and kill time invariant variables like geography.)

Lagged DV vs Lagged Latent Models

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

$$y_{i,t}^* = \mathbf{x}_{i,t}\beta + \epsilon_{i,t} + \rho\epsilon_{i,t-1} \quad (19)$$

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

$$y_{i,t}^* = \mathbf{x}_{i,t}\beta + \epsilon_{i,t} + \rho y_{i,t-1}^* \quad (21)$$

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 called "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.

The model with the lagged latent corresponds to a time series model in the latent variable (with a lagged dependent variable) which is just transformed into the 0-1 observed value each period, but where that transformation is only a measurement issue and has no consequences for the behavior of the system.

Markov Transition Matrices

A Markov process (first order) assumes that whether or not you are 0 or 1 at time t is a function only of where you were at time $t - 1$ and covariates.

In the true state dependence model, we have (in the latent)

$$Pr(y_{i,t} = 1 | y_{i,t-1} = 1) = \text{logit}(\mathbf{x}_{i,t}\beta + \rho)$$

whereas

$$P(y_{i,t} = 1 | y_{i,t-1} = 0) = \text{logit}(\mathbf{x}_{i,t}\beta)$$

so the two logit equations are parallel (in the latent space).

Transition model

This is a very strong assumption. A weaker one is

$$P(y_{i,t} = 1 | y_{i,t-1} = 0) = \text{Probit}(\mathbf{x}_{i,t}\beta) \quad (22)$$

$$P(y_{i,t} = 1 | y_{i,t-1} = 1) = \text{Probit}(\mathbf{x}_{i,t}\alpha) \quad (23)$$

which can be written more compactly as

$$P(y_{i,t} = 1) = \text{Probit}(\mathbf{x}_{i,t}\beta + y_{i,t-1}\mathbf{x}_{i,t}\gamma) \quad (24)$$

where

$$\gamma = \alpha - \beta. \quad (25)$$

This is just a fully interactive model, conditioning on lagged y . It says that the independent variables produce the dependent variable with parameter β when the system was in state $y = 0$ last period but with parameter α when system was in state $y = 1$ last period.

While the equation is written with the same independent variables for both states, no reason they could not differ, that is, the process governing transitions from 0 to 1 could be completely unrelated to the process governing transitions from 1 to 0 (or, equivalently, 1 to 1).

The transition model and event history

With continuous dependent variable, the estimate of the fully interactive model (the analogue of Equation 24) is not identical to estimating each of the subset regressions (the analogues of Equations 22 and 23). This is because the full estimation estimates a single estimate of σ^2 whereas the two subset regressions estimate two different σ^2 .

But note for any binary dependent variable, we assume for identification that $\sigma^2 = 1$. Thus we get *identical* estimates of the parameters of Equation 24 whether we estimate two subset regressions or one big regression (remembering that to drop the first observation in any unit, since we do not know what state it was in last period).

But note that Equation 22 is just the duration *independent* form of our event history methods. Thus we can see the event history methods as generalizing this subset logit.

If we have enough event history information (we do not with disputes, they do not last long enough), we could also estimate the event history generalization of Equation 23.

Thus the event history methods subsume the transition model.

Binary Panel Data - Random Effects Probit

This does not look like event history data (or we observe such short histories that our methods will not work very well).

Not a huge number of Binary Panel studies in PS.

Can do random effects Probit, since the maths works out nicely

The model is just like Equation 1a for fixed effects

$$y_{i,t}^* = \mathbf{x}_{i,t}\beta + \alpha_i + \epsilon_{i,t} \quad (26)$$

where α and ϵ are independent and $\alpha \sim N(0, \sigma_\alpha^2)$.

If this WERE independent probit, the loglike would just be the sum of the loglikes. But here all the probabilities for the T obs on unit i are related, so best we can do is break the problem down into the sum of N loglikes, each of which is a T -fold integral. Doable if T is very small, but gets tough.

The trick (popularized by Butler and Moffitt, but known before) is to condition on α_i ; conditional on this, the T densities are independent.

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

$$\begin{aligned} f(\nu_{i,1}, \nu_{i,2}, \dots, \nu_{i,T}) &= \int_{-\infty}^{\infty} f(\nu_{i,1}, \nu_{i,2}, \dots, \nu_{i,T} | \alpha_i) f(\alpha_i) d\alpha_i \\ &= \int_{-\infty}^{\infty} \prod_{t=1}^T f(\nu_{i,t} | \alpha_i) f(\alpha_i) d\alpha_i \end{aligned}$$

where the second term follows from the conditional independence of the ν conditional on α_i .

This is usual rewritten with tedious algebra to make it more suitable for numeric optimization, but you can already see that the problem now involves only a simple integral, and so is not much more difficult than ordinary probit.

Implemented in Stata in xtprobit. Slower than probit by quite a bit, but still doable. As T gets bigger, gets slower (and less accurate), unknown how large T can be and still feasible (10???)

As usual, the random effects must be ASSUMED independent of the independent variables. See Wawro, AJPS, 2001 for a discussion of correlated random effects logit, though this model makes some odd assumptions of its own (or depends on very strong identifying assumptions) and is quite hard to estimate.

Fixed Effects Conditional Logit

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 in 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 $T = 2$.

Chamberlain has proposed a solution, which conditions on the fixed effects in a better way. It only works (or has a nice functional form) for logit. It turns out to be easy to do with standard logit programs, and is implemented in Stata in clogit.

While the details get tedious (especially the notation), the basic idea is simple. A sufficient statistic for the unit effect, α_i is $\sum_{t=1}^T y_{i,t}$. This is because, for whatever estimate of β we have, we can adjust α_i to match the unit proportion of observed positive outcomes.

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 x determine in which years of unit i the successes are most likely.

To see this, suppose we know that 40% of unit 1's outcomes were successes. Take whatever estimates for β and whatever values of the ind vars. you like. Now just pick an α_1 so as to make the overall average P(success) in unit 1 40%. We can clearly move α_1 around to produce any average probability of success we like (from 0 to 1), so obviously the best "estimator" of α_1 is given by whatever value produces an average probability of success of 40%. While this is not a consistent estimate of α_1 (since T is fixed and is only based on T obs), it does allow us to estimate β conditional on all the α_i .

As is typical in examples, work with $T = 2$. Everything holds for larger T but the notation and counting get tedious.

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.)

So the only interesting case is one success (out of two), that is, fail, succeed or succeed, fail. These are totally symmetric. The conditional probability of fail, succeed given one success is

$$P(0, 1|1\text{success}) = \frac{P(0, 1\text{and}1\text{ success})}{P(1\text{ success})} = \frac{P(0, 1)}{P(0, 1) + P(1, 0)}$$

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.)

Thus in the likelihood, if you observe (0, 1) for unit i , the contribution is that logit determined probability, if (1, 0) it is just one minus that probability.

Situation, and why Stata does Chamberlain and conditional multinomial logit in the same routine, is perhaps easier to see for $T = 3$.

Suppose unit i has 1 success, and, in particular, the string (1, 0, 0) for the three time periods. Now conditioning on only getting one success, and three possible periods for that

success, the contribution to the likelihood is just the multinomial logit probability of getting a success at time period 1 given that you could have gotten it at time periods 1, 2 or 3. Thus clogit is just doing a series of MNL computations.

At that point, the α_i are like individual specific attributes, (as are any time invariant covariates), time varying covariates are like alternative specific attributes. (This may become clearer next week.) This is why the inconsistency of the estimates of α does not cause problems for the estimates of β .)

What if you condition on two successes. Then say you observe (1, 1, 0). You could have observed either (1, 1, 0) or (1, 0, 1) or (0, 1, 1) so again is just a slightly more complicated MNL problem.

You can see that with bigger T there is lots for Stata to keep track of, but fortunately this is not a problem for Stata programmers, not us. If you wanted to program yourself, building up from standard MNL, you would have lots of tedious bookkeeping to do.

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).

GEE

A popular approach is based on the Generalized Linear Model (GLM), and is known as the "General Estimating Equation."

See Zorn, AJPS 2001. Like Huber, this is a way of avoiding modeling the real situation, but it might work okay in practice (and may well be better than doing nothing).

The method is based on pseudo-ML. It assumes we have the mean function right (that is y is still the probit or logit of $x\beta$). But the variance function takes the interdependence of the observations into account.

Thus it, like Huber, is a kludge, but a kludge is probably better than nothing at all.

But if you estimate GEE with ar1 errors, look at the estimate of the serial correlation and see if it makes sense (doesn't for disputes, see Beck, forthcoming).