CRE METHODS FOR UNBALANCED PANELS

Correlated Random Effects Panel Data Models

IZA Summer School in Labor Economics
May 13-19, 2013
Jeffrey M. Wooldridge
Michigan State University

1. Introduction
2. Linear Model with Additive Heterogeneity
3. Linear Model with Correlated Random Slopes
4. A Modeling Approach for Nonlinear Models
5. Estimating Average Partial Effects
6. Application to Probit/Fractional Response

1. Introduction

- In linear model with additive heterogeneity, unbalanced panels cause no serious issues provided certain assumptions about selection hold.
- FE is more robust than RE in the sense that with RE selection must be assumed uncorrelated with heterogeneity as well as with idiosyncratic shocks. FE allows arbitrary correlation between selection and c_i.

- For nonlinear models, RE easily adapts to unbalanced panels, but the selection assumptions are strong.
- Conditional MLE allows unbalanced panels when it applies (logit, Poisson).
- Treating the c_i as parameters to estimate – so called “fixed effects” – does not require a balanced panel (but still has an incidental parameters problem with small T).

- Drawback to correlated RE approach compared with CMLE or FE: Not clear how to allow for unbalanced panels when we want heterogeneity to be correlated with covariates and selection.
- Want to be able to handle nonlinear CRE models.
- Without specifically modeling the selection rule – say, using a Heckman-type approach – all methods assume selection is independent of shocks. We can test this assumption.

2. Linear Model with Additive Heterogeneity

- We still draw a random sample from the cross section, with units indexed by i. But now we may not observe a complete set of time series observations.
- Model this situation using a sequence of selection indicators, $\{s_{i1}, \ldots, s_{iT}\}$, where $s_{it} = 1$ if time period t for unit i can be used in estimation. Usually this means that we observe all elements of $(\mathbf{x}_{it}, y_{it})$.

- We only use an (i, t) pair when a full set of data is observed, as happens when software is used for RE and FE on unbalanced panels. So we focus on complete-case estimators without imputation.
- We still write the population model for a random draw i as

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + c_i + u_{it}, \; t = 1, \ldots, T,$$

where $\mathbf{x}_{it}$ can generally include a fully set of time dummies, or other aggregate time variables.

• Along with a rank condition, a sufficient (but not necessary) condition for consistency of FE on the unbalanced panel is

$$E(u_{it}|\mathbf{x}_i, c_i, \mathbf{s}_i) = 0, \; t = 1, \ldots, T,$$

where $\mathbf{x}_i = (\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{iT})$ and $\mathbf{s}_i = (s_{i1}, s_{i2}, \ldots, s_{iT})$.

• Allows selection in any time period to be correlated with $(\mathbf{x}_i, c_i)$, but selection in all time periods must be unrelated to the idiosyncratic shocks.

• The time-demeaned data now uses different time periods for different i. Let

$$\ddot{y}_{it} = y_{it} - T_i^{-1} \sum_{r=1}^{T} s_{ir} y_{ir} = y_{it} - \bar{y}_i$$

$$\ddot{\mathbf{x}}_{it} = \mathbf{x}_{it} - T_i^{-1} \sum_{r=1}^{T} s_{ir} \mathbf{x}_{ir} = \mathbf{x}_{it} - \bar{\mathbf{x}}_i$$

where $T_i = \sum_{r=1}^{T} s_{ir}$ is random.

- The FE estimator is then

$$\hat{\boldsymbol{\beta}} = \left(N^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} s_{it} \ddot{\mathbf{x}}_{it}' \ddot{\mathbf{x}}_{it} \right)^{-1} \left(N^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} s_{it} \ddot{\mathbf{x}}_{it}' \ddot{y}_{it} \right)$$

$$= \boldsymbol{\beta} + \left(N^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} s_{it} \ddot{\mathbf{x}}_{it}' \ddot{\mathbf{x}}_{it} \right)^{-1} \left(N^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} s_{it} \ddot{\mathbf{x}}_{it}' u_{it} \right)$$

- Consistency (fixed T, $N \to \infty$) follows if

$$rank \left[\sum_{t=1}^{T} E(s_{it} \ddot{\mathbf{x}}_{it}' \ddot{\mathbf{x}}_{it}) \right] = K$$

$$E(s_{it} \ddot{\mathbf{x}}_{it}' u_{it}) = \mathbf{0}, \; t = 1, \ldots, T.$$

- In the balanced case we know that if we estimate the equation

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + \psi + \bar{\mathbf{x}}_i\boldsymbol{\xi} + v_{it}$$

by pooled OLS or RE we get the FE estimate for $\boldsymbol{\beta}$.

- Conveniently, the same result holds for unbalanced case, with a caveat: the time averages are defined only for the complete-case observations.

- In other words, estimate the equation

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + \psi + \bar{\mathbf{x}}_i\boldsymbol{\xi} + v_{it}$$

by pooled OLS using the $s_{it} = 1$ observations. The coefficient vector $\hat{\boldsymbol{\beta}}$ is identical to the fixed effects.

• Unlike in the balanced case, any aggregate time variables, including time dummies, should be part of $\mathbf{x}_{it}$, and their time averages must be included in $\bar{\mathbf{x}}_i$ to get the FE estimates on the other variables.

• Wooldridge (2010, working paper) shows a more general result. Recall that the RE estimator can be obtained from a pooled OLS regression. On an unbalanced panel, it is

$$y_{it} - \theta_i \bar{y}_i \text{ on } \mathbf{x}_{it} - \theta_i \bar{\mathbf{x}}_i, \ (1 - \theta_i)\bar{\mathbf{x}}_i, \ (1 - \theta_i)\mathbf{z}_i \text{ if } s_{it} = 1,$$

where

$$\theta_i = 1 - [\sigma_u^2/(\sigma_u^2 + T_i\sigma_c^2)]^{1/2}$$

varies because T_i varies.

Algebraic Fact: Let $\tilde{\boldsymbol{\beta}}$ be the vector $(K \times 1)$ of coefficients on $\mathbf{x}_{it} - \theta_i \bar{\mathbf{x}}_i$ in the POLS regression above. Then $\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{FE}$, the fixed effects estimate on the unbalanced panel.

- Note that $\mathbf{z}_i$ can contain any time-constant variables, including functions of T_i, or interactions of the form $T_i \cdot \bar{\mathbf{x}}_i$, or allow a different set of coefficients on $\bar{\mathbf{x}}_i$ for each different T_i. The coefficients on $\bar{\mathbf{x}}_i$ may changed widely across T_i, yet the estimate on $\mathbf{x}_{it}$ is still the FE estimate.

- The algebraic equivalence still justifies the robust, regression-based Hausman test.
- Write a model with time-constant variables $\mathbf{z}_i$ as

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + \mathbf{z}_i\boldsymbol{\gamma} + c_i + u_{it}, \; t = 1, \ldots, T,$$

where $\mathbf{z}_i$ includes a constant.

- Use the Mundlak equation

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + \bar{\mathbf{x}}_i\boldsymbol{\xi} + \mathbf{z}_i\boldsymbol{\gamma} + a_i + u_{it}$$

and estimate this by RE.

- The regression based Hausman test is just a (robust) Wald test of $H_0 : \boldsymbol{\xi} = \mathbf{0}$ after RE estimation of the augmented equation.
- Remember that FE (and RE) assume that selection in any time period is not correlated with u_{it}. Might worry that a shock today affects being in the sample in the future.
- Wooldridge (2010, MIT Press): Using FE on the unbalanced panel, estimate the equation

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + \eta s_{i,t+1} + c_i + u_{it}, \; t = 1, \ldots, T-1$$

and test $H_0 : \eta = 0$ using a robust test.

• If we use RE we can (and should) test if c_i is correlated with functions of T_i. For example, define dummies

$$d_{ir} = 1[T_i = r], r = 1, \ldots, T-1$$

and add these to the usual RE estimation.

• Significance of $\mathbf{d}_i = (d_{i1}, \ldots, d_{i,T-1})$ casts serious doubt on the usual RE analysis. FE does not care.

- We can jointly test the assumption that

$$E(c_i|\mathbf{x}_{i1},\ldots,\mathbf{x}_{iT},s_{i1},\ldots,s_{iT}) = E(c_i)$$

by estimating the equation

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + \psi + \bar{\mathbf{x}}_i\boldsymbol{\xi} + \mathbf{d}_i\boldsymbol{\gamma} + v_{it}$$

by RE using the unbalanced panel and jointly testing $H_0 : \boldsymbol{\xi} = \mathbf{0}, \boldsymbol{\gamma} = \mathbf{0}$.

3. Linear Model with Correlated Random Slopes

- Now consider a model with all slopes heterogeneous:

$$E(y_{it}|\mathbf{x}_i, a_i, \mathbf{b}_i) = a_i + \mathbf{x}_{it}\mathbf{b}_i,$$

so, in the population, $\{\mathbf{x}_{it} : t = 1, \ldots, T\}$ is strictly exogenous conditional on $(a_i, \mathbf{b}_i)$.

- Write $a_i = \alpha + c_i$, $\mathbf{b}_i = \boldsymbol{\beta} + \mathbf{d}_i$ and

$$y_{it} = \alpha + \mathbf{x}_{it}\boldsymbol{\beta} + c_i + \mathbf{x}_{it}\mathbf{d}_i + u_{it}$$

where $E(u_{it}|\mathbf{x}_i, a_i, \mathbf{b}_i) = E(u_{it}|\mathbf{x}_i, c_i, \mathbf{d}_i) = 0$ for all t.

• Selection may be related to $(\mathbf{x}_i, a_i, \mathbf{b}_i)$ but not the idiosyncratic shocks:

$$E(y_{it}|\mathbf{x}_i, a_i, \mathbf{b}_i, \mathbf{s}_i) = E(y_{it}|\mathbf{x}_i, a_i, \mathbf{b}_i)$$

or

$$E(u_{it}|\mathbf{x}_i, a_i, \mathbf{b}_i, \mathbf{s}_i) = 0, \ t = 1, \ldots, T.$$

• With enough time periods, we could obtain $\hat{a}_i$, $\hat{\mathbf{b}}_i$ ("fixed effects") and then average these. But need $T_i > K + 1$.

- Unfortunately, if selection is correlated with $\mathbf{b}_i$, there are no intuitive robustness results for the usual FE estimator.
- How might we use a CRE approach?

- Multiply through by selection:

$$s_{it}y_{it} = s_{it}\alpha + s_{it}\mathbf{x}_{it}\boldsymbol{\beta} + s_{it}c_i + s_{it}\mathbf{x}_{it}\mathbf{d}_i + s_{it}u_{it}$$

- Because we only use observations with $s_{it} = 1$, we handle the heterogeneity by conditioning on the entire history of selection and the values of the covariates if selected. That is, $\{(s_{it}, s_{it}\mathbf{x}_{it}) : t = 1, \ldots, T\}$. If $s_{it} = 0$ the observation is not used; if $s_{it} = 1$, the observation is used, and we observe $\mathbf{x}_{it}$.

- If we condition on the larger information set $\{(\mathbf{x}_{i1}, s_{i1}), (\mathbf{x}_{i2}, s_{i2}), \ldots, (\mathbf{x}_{iT}, s_{iT})\}$ and heterogeneity depends only on the covariates, we would be left with an equation that is not estimable unless the covariates are always observed.

• Write $\mathbf{h}_i \equiv \{\mathbf{h}_{it} : t = 1, \ldots, T\} \equiv \{(s_{it}, s_{it}\mathbf{x}_{it}) : t = 1, \ldots, T\}$. Then, extending Mundlak (1978) and Chamberlain (1982), we work with

$$E(s_{it}y_{it}|\mathbf{h}_i) = s_{it}\alpha + s_{it}\mathbf{x}_{it}\boldsymbol{\beta} + s_{it}E(c_i|\mathbf{h}_i) + s_{it}\mathbf{x}_{it}E(\mathbf{d}_i|\mathbf{h}_i)$$

and then make assumptions concerning $E(c_i|\mathbf{h}_i)$ and $E(\mathbf{d}_i|\mathbf{h}_i)$.

• Alternatively, we could eliminate c_i using the within transformation, and focus just on $E(\mathbf{d}_i|\mathbf{h}_i)$.

- A simple approach is to model the expectations as exchangeable functions of $\{\mathbf{h}_{it} : t = 1, \ldots, T\}$ – extending the balanced case considered by Altonji and Matzkin (2005).
- From the model with constant slopes, natural to start with

$$\mathbf{w}_i \equiv (T_i, \bar{\mathbf{x}}_i).$$

- Fairly natural extension of Mundlak (1978).

- A flexible specification with $g_{ir} \equiv 1[T_i = r]$:

$$E(c_i|T_i, \bar{\mathbf{x}}_i) = \sum_{r=1}^{T} \psi_r (g_{ir} - \rho_r) + \sum_{r=1}^{T} g_{ir} \cdot (\bar{\mathbf{x}}_i - \boldsymbol{\mu}_r)\boldsymbol{\xi}_r$$

$$E(\mathbf{d}_i|T_i, \bar{\mathbf{x}}_i) = \sum_{r=1}^{T} (g_{ir} - \rho_r)\boldsymbol{\kappa}_r + \sum_{r=1}^{T} g_{ir} \cdot [(\bar{\mathbf{x}}_i - \boldsymbol{\mu}_r) \otimes \mathbf{I}_K]\boldsymbol{\eta}_r,$$

where

$$\rho_r = P(T_i = r) = E\{1[T_i = r]\}$$

$$\boldsymbol{\mu}_r \equiv E(\bar{\mathbf{x}}_i|T_i = r)$$

• As a practical matter, the formulation above is identical to running separate regressions for each T_i:

$$y_{it} \text{ on } 1,\ \mathbf{x}_{it},\ \bar{\mathbf{x}}_i,\ (\bar{\mathbf{x}}_i - \hat{\boldsymbol{\mu}}_r) \otimes \mathbf{x}_{it},\ \text{for } s_{it} = 1$$

where $\hat{\boldsymbol{\mu}}_r = N_r^{-1}\left(\sum_{i=1}^{N} 1[T_i = r]\bar{\mathbf{x}}_i\right)$ and N_r is the number of observations with $T_i = r$. The coefficient on $\mathbf{x}_{it}$, $\hat{\boldsymbol{\beta}}_r$, is the APE given $T_i = r$. We can average these across r to obtain the overall APE.

- Notice that we lose the $T_i = 1$ observations – just as in a standard fixed effects estimation.
- We could use grouping based on T_i so that the APEs include the entire population.

• In any formulation, including the basic FE estimation, have a simple test for dynamic selection bias – that is, for the null $E(y_{it}|\mathbf{x}_i, a_i, \mathbf{b}_i, \mathbf{s}_i) = E(y_{it}|\mathbf{x}_i, a_i, \mathbf{b}_i)$. Assumes that our model for $E(c_i, \mathbf{d}_i|\{(s_{it}, s_{it}\mathbf{x}_{it}) : t = 1, \ldots, T\}$ is correctly specified. Then, no other functions of $\{(s_{it}, s_{it}\mathbf{x}_{it}) : t = 1, \ldots, T\}$ should appear in $E(s_i y_{it}|\mathbf{h}_i)$.

• Assuming we have some T_i with $T_i \geq 3$, add as extra regressors at time t the variables $(s_{i,t+1}, s_{i,t+1}\mathbf{x}_{i,t+1})$. We can compute a fully robust (to serial correlation and heteroskedasticity) Wald test of joint significance. Produces a joint test of ignorable selection and strictly exogenous covariates.

4. A Modeling Approach for Nonlinear Models

- Interested in the conditional distribution

$$D(\mathbf{y}_{it}|\mathbf{x}_{it}, \mathbf{c}_i),$$

or maybe just an expectation.

- Assume strictly exogenous covariates conditional on $\mathbf{c}_i$ *and* ignorable selection:

$$D(\mathbf{y}_{it}|\mathbf{x}_i, \mathbf{c}_i, \mathbf{s}_i) = D(\mathbf{y}_{it}|\mathbf{x}_{it}, \mathbf{c}_i), \; t = 1, \ldots, T.$$

• Focus on methods that only exploit assumptions about the marginal distributions, $D(\mathbf{y}_{it}|\mathbf{x}_{it}, \mathbf{c}_i)$, not joint distributions. Average partial effects are generally identified without restricting conditional dependence.

• Let $g_t(\mathbf{y}_t|\mathbf{x}_{it}, \mathbf{c}_i; \boldsymbol{\gamma})$ be a parametric density conditional on $(\mathbf{x}_{it}, \mathbf{c}_i)$.

• As with the linear model, specify models for

$$D(\mathbf{c}_i|\{(s_{it}, s_{it}\mathbf{x}_{it}) : t = 1, \ldots, T\}).$$

• Let $\mathbf{w}_i$ be a vector of known functions of $\{(s_{it}, s_{it}\mathbf{x}_{it}) : t = 1, \ldots, T\}$ that act as sufficient statistics, so that

$$D(\mathbf{c}_i|\{(s_{it}, s_{it}\mathbf{x}_{it}) : t = 1, \ldots, T\}) = D(\mathbf{c}_i|\mathbf{w}_i)$$

• Because $D(\mathbf{y}_{it}|\mathbf{x}_{it}, \mathbf{c}_i, s_{it} = 1) = D(\mathbf{y}_{it}|\mathbf{x}_{it}, \mathbf{c}_i)$ and we have a density for the latter, we can obtain the density given $(\mathbf{x}_{it}, \mathbf{w}_i, s_{it} = 1)$ as

$$f_t(\mathbf{y}_t|\mathbf{x}_{it}, \mathbf{w}_i; \boldsymbol{\gamma}, \boldsymbol{\delta}) = \int_{\mathbb{R}^M} g_t(\mathbf{y}_t|\mathbf{x}_{it}, \mathbf{c}; \boldsymbol{\gamma}) h(\mathbf{c}|\mathbf{w}_i; \boldsymbol{\delta}) d\mathbf{c}$$

where $h(\mathbf{c}|\mathbf{w}_i; \boldsymbol{\delta})$ is a parametric density for $D(\mathbf{c}_i|\mathbf{w}_i)$ and M is the dimension of $\mathbf{c}_i$.

• The partial log-likelihood function for the entire sample is

$$\sum_{i=1}^{N} \sum_{t=1}^{T} s_{it} \log[f_t(\mathbf{y}_{it}|\mathbf{x}_{it}, \mathbf{w}_i; \boldsymbol{\gamma}, \boldsymbol{\delta})].$$

• Need a robust sandwich estimator for serial correlation (and maybe other misspecifications).

• Same basic arguments carry through if we focus on estimating conditional means. We start with $E(y_{it}|\mathbf{x}_{it}, \mathbf{c}_i) = m_t(\mathbf{x}_{it}, \mathbf{c}_i)$ and obtain $E(y_{it}|\mathbf{x}_{it}, \mathbf{w}_i)$ by integrating $m_t(\mathbf{x}_{it}, \mathbf{c}_i)$ with respect to the density of $\mathbf{c}_i$ given $\mathbf{w}_i$ (which, again, is valid for the mean when $s_{it} = 1$). Or, we just assert a model for $E(y_{it}|\mathbf{x}_{it}, \mathbf{w}_i)$.

• Can then use a host of quasi-log likelihoods on the selected sample, including the Bernoulli for fractional responses, the gamma for nonnegative (continuous) responses, and the Poisson for nonnegative (count) responses.

5. Estimating Average Partial Effects

- In most nonlinear models, the parameters $\boldsymbol{\gamma}$ appearing in $f_t(\mathbf{y}_t|\mathbf{x}_t, \mathbf{c}; \boldsymbol{\gamma})$ provide only part of the story for the effect of $\mathbf{x}_t$ on $\mathbf{y}_t$. Using pooled methods we often cannot fully identify $\boldsymbol{\gamma}$ or the distribution of $\mathbf{c}_i$.
- Even in the unbalanced case we can use the Blundell and Powell (2003) approach provided a vector $\mathbf{w}_i$ suitably proxies for correlation between $\mathbf{c}_i$ and $\{(s_{it}, s_{it}\mathbf{x}_{it}) : t = 1, \ldots, T\}$.

- Let $q_t(\mathbf{x}_t, \mathbf{w}; \boldsymbol{\theta})$ denote the mean associated with $f_t(\mathbf{y}_t|\mathbf{x}_t, \mathbf{w}; \boldsymbol{\theta})$. Then

$$ASF(\mathbf{x}_t) = E_{\mathbf{w}_i}[q_t(\mathbf{x}_t, \mathbf{w}_i; \boldsymbol{\theta})],$$

that is, we can obtain the ASF by averaging out the observed vector of sufficient statistics, $\mathbf{w}_i$, from $E(y_{it}|\mathbf{x}_t, \mathbf{w}_i, s_{it} = 1)$ rather than averaging out $\mathbf{c}_i$ from $E(y_{it}|\mathbf{x}_t, \mathbf{c}_i)$.

- Consistent estimation is then easy:

$$\widehat{ASF}(\mathbf{x}_t) = N^{-1} \sum_{i=1}^{N} q_t(\mathbf{x}_t, \mathbf{w}_i; \hat{\boldsymbol{\theta}})$$

- Panel bootstrap is still justified with unbalanced panel.

- Tricky to estimate an overall average partial effect. With a balanced panel, it is natural to average $\widehat{APE}_{tj}(\mathbf{x}_{it}, \mathbf{w}_i)$ across the distribution of $(\mathbf{x}_{it}, \mathbf{w}_i)$, and then possibly across t, too. But if selection s_{it} depends on $\mathbf{x}_{it}$, averaging across the selected sample does not consistently estimate $E_{(\mathbf{x}_{it},\mathbf{w}_i)}[APE_{tj}(\mathbf{x}_{it}, \mathbf{w}_i)]$.
- Presumably we still have an idea of useful values to plug in for $\mathbf{x}_t$.

6. Application to Probit/Fractional Response

• The probit model, either for true binary responses or fractional responses, is natural for applying the previous approach. The model is

$$P(y_{it} = 1|\mathbf{x}_i, c_i) = P(y_{it} = 1|\mathbf{x}_{it}, c_i) = \Phi(\mathbf{x}_{it}\boldsymbol{\beta} + c_i), \, t = 1, \ldots, T$$

where $\mathbf{x}_{it}$ can include time dummies or other aggregate time variables.

• Once we specify $P(y_{it} = 1|\mathbf{x}_i, c_i)$ and assume that selection is conditionally ignorable for all t, that is,

$$P(y_{it} = 1|\mathbf{x}_i, c_i, \mathbf{s}_i) = P(y_{it} = 1|\mathbf{x}_i, c_i),$$

all that is left is to specify a model for $D(c_i|\mathbf{w}_i)$ for suitably chosen functions $\mathbf{w}_i$ of $\{(s_{it}, s_{it}\mathbf{x}_{it}) : t = 1, \ldots, T\}$.

- A specification linear in $\bar{\mathbf{x}}_i$ but with intercept and slopes different for each T_i is

$$E(c_i|\mathbf{w}_i) = \sum_{r=1}^{T} \psi_r 1[T_i = r] + \sum_{r=1}^{T} 1[T_i = r] \cdot \bar{\mathbf{x}}_i \boldsymbol{\xi}_r$$

- At a minimum, should let the variance of c_i change with T_i:

$$Var(c_i|\mathbf{w}_i) = \exp\left(\tau + \sum_{r=1}^{T-1} 1[T_i = r]\omega_r \right)$$

• If we also maintain that $D(c_i|\mathbf{w}_i)$ is normal, then we obtain the following response probability for $s_{it} = 1$ (with normalization that does not affect estimation of APEs):

$$P(y_{it} = 1|\mathbf{x}_{it}, \mathbf{w}_i, s_{it} = 1) = \Phi\left[\frac{\mathbf{x}_{it}\boldsymbol{\beta} + \sum_{r=1}^{T} \psi_r g_{ir} + \sum_{r=1}^{T} g_{ir} \cdot \bar{\mathbf{x}}_i \boldsymbol{\xi}_r}{\exp\left(\sum_{r=2}^{T} g_{ir}\omega_r\right)^{1/2}} \right]$$

so that the denominator is unity when all ω_r are zero. (Recall $g_{ir} = 1[T_i = r]$.)

• No difficulty in adding $g_{ir} \cdot \bar{\mathbf{x}}_i$ for $r = 1, \ldots, T$ to the variance function.

• Computational bonus: Estimable by so-called “heteroskedastic probit” software, where the explanatory variables at time t are $(1, \mathbf{x}_{it}, g_{i1}, \ldots, g_{iT}, g_{i1} \cdot \bar{\mathbf{x}}_i, \ldots, g_{iT} \cdot \bar{\mathbf{x}}_i)$ and the explanatory variables in the variance are simply the dummy variables $(g_{i2}, \ldots, g_{iT})$, or also add $g_{i1} \cdot \bar{\mathbf{x}}_i, \ldots, g_{iT} \cdot \bar{\mathbf{x}}_i$.

• The APEs are easy to obtain from the estimated ASF:

$$\widehat{ASF}(\mathbf{x}_t) = N^{-1}\sum_{i=1}^{N}\Phi\left[\frac{\mathbf{x}_t\hat{\boldsymbol{\beta}} + \sum_{r=1}^{T}\hat{\psi}_r g_{ir} + \sum_{r=1}^{T} g_{ir}\cdot\bar{\mathbf{x}}_i\hat{\boldsymbol{\xi}}_r}{\exp\left(\sum_{r=2}^{T} g_{ir}\hat{\omega}_r\right)^{1/2}}\right]$$

where the coefficients with "^" are from the pooled heteroskedastic probit estimation.

• The functions of $(T_i, \bar{\mathbf{x}}_i)$ are averaged out, leaving the result a function of $\mathbf{x}_t$. If, say, x_{tj} is continuous, its APE is estimated as

$$\widehat{APE}_{tj}(\mathbf{x}_t) = \hat{\beta}_j\left\{N^{-1}\sum_{i=1}^{N}\phi\left[\frac{\mathbf{x}_t\hat{\boldsymbol{\beta}} + \sum_{r=1}^{T}\hat{\psi}_r g_{ir} + \sum_{r=1}^{T} g_{ir}\cdot\bar{\mathbf{x}}_i\hat{\boldsymbol{\xi}}_r}{\exp\left(\sum_{r=2}^{T} g_{ir}\hat{\omega}_r\right)^{1/2}}\right]\right\}.$$

• The above procedure applies, without change, if y_{it} is a fractional response; that is, $0 \leq y_{it} \leq 1$. The original model is $E(y_{it}|\mathbf{x}_{it}, c_i) = \Phi(\mathbf{x}_{it}\boldsymbol{\beta} + c_i)$ and APEs are on the mean response.

• Could allow $\hat{\boldsymbol{\beta}}$ to change with each T_i (losing $T_i = 1$). Then, just estimate an APE for each $T_i \geq 2$ and average the results. Just a separate probit with explanatory variables $(1, \mathbf{x}_{it}, \bar{\mathbf{x}}_i)$.

• Or, allow heteroskedasticity as a function of $\bar{\mathbf{x}}_i$, with ASF

$$\widehat{ASF}(\mathbf{x}_t) = N^{-1} \sum_{r=2}^{T} \sum_{i=1}^{N} g_{ir} \Phi\left[\frac{\mathbf{x}_t \hat{\boldsymbol{\beta}}_r + \hat{\psi}_r + \bar{\mathbf{x}}_i \hat{\boldsymbol{\xi}}_r}{\exp(\bar{\mathbf{x}}_i \hat{\boldsymbol{\lambda}}_r/2)} \right].$$

- Same ideas apply to ordered probit and even multinomial logit, as well as Tobit and count data models.
- With count data, Poisson FE estimator has very nice robustness properties, including for sample selection – $E(y_{it}|\mathbf{x}_i, c_i, \mathbf{s}_i) = E(y_{it}|\mathbf{x}_{it}, c_i) = \exp(\mathbf{x}_{it}\boldsymbol{\beta} + c_i)$ is sufficient – but only for the scalar, additive heterogeneity case.

- Scope for CRE approach in more complicated models, such as

$$E(y_{it}|\mathbf{x}_i, \mathbf{c}_i) = \exp(a_i + \mathbf{x}_{it}\mathbf{b}_i).$$

- CMLE approach makes no sense with all coefficients heterogeneous (and it is not clear a CMLE exists, anyway).
- FE approach can be implemented with lots of large T_i ($T_i > K + 1$), but how well? CRE approach is available but with restrictions on $D(\mathbf{c}_i|\{(s_{it}, s_{it}\mathbf{x}_{it}) : t = 1, \ldots, T\})$.

EXAMPLE: Effects of Spending on School Test Pass Rates

```
. use meap94_98
. tab year

1992=school |
  yr 1991-2 |      Freq.     Percent        Cum.
------------+-----------------------------------
       1994 |      1,012       14.15       14.15
       1995 |      1,205       16.85       31.01
       1996 |      1,635       22.87       53.87
       1997 |      1,635       22.87       76.74
       1998 |      1,663       23.26      100.00
------------+-----------------------------------
      Total |      7,150      100.00

. egen tobs = sum(1), by(schid)
. tab tobs

  number of |
       time |
    periods |      Freq.     Percent        Cum.
------------+-----------------------------------
          3 |      1,512       21.15       21.15
          4 |      1,028       14.38       35.52
          5 |      4,610       64.48      100.00
------------+-----------------------------------
      Total |      7,150      100.00

. gen tobs3 = tobs == 3
. gen tobs4 = tobs == 4
```

```
. egen lavgrexpb = mean(lavgrexp), by(schid)
. egen lunchb = mean(lunch), by(schid)
. egen lenrolb = mean(lenrol), by(schid)
. egen y95b = mean(y95), by(schid)
. egen y96b = mean(y96), by(schid)
. egen y97b = mean(y97), by(schid)
. egen y98b = mean(y98), by(schid)
. gen tobs4 = tobs == 4
. gen tobs3 = tobs == 3
. gen tobs3_lavgrexp = tobs3*lavgrexp
. gen tobs4_lavgrexp = tobs4*lavgrexp
```

```
. xtreg math4 lavgrexp lunch lenrol y95 y96 y97 y98, fe cluster(schid)

Fixed-effects (within) regression               Number of obs      =      7150
Group variable: schid                           Number of groups   =      1683

R-sq:  within  = 0.3602                         Obs per group: min =         3
       between = 0.0292                                        avg =       4.2
       overall = 0.1514                                        max =         5

                                                F(7,1682)          =    431.08
corr(u_i, Xb)  = 0.0073                         Prob > F           =    0.0000

                              (Std. Err. adjusted for 1683 clusters in schid)
------------------------------------------------------------------------------
             |               Robust
       math4 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    lavgrexp |   6.288376   2.431317     2.59   0.010     1.519651     11.0571
       lunch |  -.0215072   .0390732    -0.55   0.582    -.0981445      .05513
      lenrol |  -2.038461   1.789094    -1.14   0.255    -5.547545    1.470623
         y95 |    11.6192   .5358469    21.68   0.000     10.56821     12.6702
         y96 |   13.05561   .6910815    18.89   0.000     11.70014    14.41108
         y97 |   10.14771   .7326314    13.85   0.000     8.710745    11.58468
         y98 |   23.41404   .7669553    30.53   0.000     21.90975    24.91833
       _cons |   11.84422   25.16643     0.47   0.638    -37.51659    61.20503
-------------+----------------------------------------------------------------
     sigma_u |   15.84958
     sigma_e |  11.325028
         rho |  .66200804   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

```
. xtreg math4 lavgrexp tobs3_lavgrexp tobs4_lavgrexp lunch lenrol
          y95 y96 y97 y98, fe cluster(schid)

Fixed-effects (within) regression               Number of obs      =      7150
Group variable: schid                           Number of groups   =      1683

R-sq:  within  = 0.3611                         Obs per group: min =         3
       between = 0.0120                                        avg =       4.2
       overall = 0.0003                                        max =         5

                                                F(9,1682)          =    337.10
corr(u_i, Xb)  = -0.8983                        Prob > F           =    0.0000

                                 (Std. Err. adjusted for 1683 clusters in schid)
------------------------------------------------------------------------------
               |               Robust
         math4 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
---------------+--------------------------------------------------------------
      lavgrexp |   3.501465   2.637181     1.33   0.184    -1.671038    8.673967
tobs3_lavgrexp |   8.048717   4.452136     1.81   0.071     -.683593    16.78103
tobs4_lavgrexp |   9.103049   5.641861     1.61   0.107    -1.962758    20.16886
         lunch |  -.0292364   .0387205    -0.76   0.450    -.1051817     .046709
        lenrol |  -2.169307   1.802788    -1.20   0.229    -5.705251    1.366636
           y95 |   12.01813   .5312958    22.62   0.000     10.97606     13.0602
           y96 |   13.56065   .6962891    19.48   0.000     12.19497    14.92634
           y97 |   10.60934   .7423073    14.29   0.000     9.153396    12.06528
           y98 |   23.84989   .7789923    30.62   0.000       22.322    25.37779
         _cons |    10.6043   24.93789     0.43   0.671    -38.30827    59.51686
---------------+--------------------------------------------------------------
       sigma_u |  41.080099
       sigma_e |  11.319318
           rho |  .92943391   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

```
. test tobs3_lavgrexp tobs4_lavgrexp

 ( 1)  tobs3_lavgrexp = 0
 ( 2)  tobs4_lavgrexp = 0

       F(  2,  1682) =    2.67
            Prob > F =    0.0694
```

```
. * Now fractional response.

. replace math4 = math4/100
(7150 real changes made)

. replace lunch = lunch/100
(7146 real changes made)

.
. capture program drop frac_het

.
. program frac_het
  1.         version 12
  2.         args llf xb zg
  3.         quietly replace `llf' = $ML_y1*log(normal(`xb'*exp(-`zg'))) ///
>      + (1 - $ML_y1)*log(1 - normal(`xb'*exp(-`zg')))
  4. end

.
. ml model lf frac_het (math4 = lavgrexp lunch lenrol y95 y96 y97 y98 lavgrexpb ///
>   lunchb lenrolb y95b y96b y97b y98b tobs3 tobs4) (tobs3 tobs4, nocons), ///
>   vce(cluster distid)

. ml max
```

```
                                                Number of obs   =       7150
                                                Wald chi2(16)   =    1690.63
Log pseudolikelihood = -4414.8409               Prob > chi2     =     0.0000

                              (Std. Err. adjusted for 467 clusters in distid)
------------------------------------------------------------------------------
             |               Robust
       math4 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
eq1          |
    lavgrexp |   .1142198    .100819     1.13   0.257    -.0833819    .3118215
       lunch |  -.1396103    .135914    -1.03   0.304    -.4059968    .1267763
      lenrol |   -.067624   .0736354    -0.92   0.358    -.2119468    .0766988
         y95 |   .3241894   .0212756    15.24   0.000     .2824899    .3658889
         y96 |   .3724917   .0286602    13.00   0.000     .3163187    .4286647
         y97 |   .2830853   .0301771     9.38   0.000     .2239392    .3422313
         y98 |   .7162732   .0352366    20.33   0.000     .6472108    .7853357
   lavgrexpb |   .1622914   .1338626     1.21   0.225    -.1000745    .4246574
      lunchb |  -.0126246   .0014318    -8.82   0.000    -.0154309   -.0098183
     lenrolb |  -.0029272   .0766622    -0.04   0.970    -.1531823    .1473279
        y95b |   .8794288     .76738     1.15   0.252    -.6246084    2.383466
        y96b |   .7270724   .2274786     3.20   0.001     .2812225    1.172922
        y97b |   .6338092   .6669099     0.95   0.342    -.6733102    1.940929
        y98b |   .2733774   .6390411     0.43   0.669    -.9791201    1.525875
       tobs3 |    .022217    .076674     0.29   0.772    -.1280612    .1724953
       tobs4 |    .088465   .1223424     0.72   0.470    -.1513217    .3282518
       _cons |  -1.856404   .9018976    -2.06   0.040    -3.624091   -.0887172
-------------+----------------------------------------------------------------
eq2          |
       tobs3 |   .2007713   .0875105     2.29   0.022     .0292538    .3722888
       tobs4 |   .5504922   .1154134     4.77   0.000     .3242861    .7766983
------------------------------------------------------------------------------
```

```
. glm math4 lavgrexp lunch lenrol y95 y96 y97 y98 lavgrexpb lunchb lenrolb
    y95b y96b y97b y98b tobs3 tobs4, fam(bin) link(probit) cluster(schid)

note: math4 has noninteger values

Generalized linear models                         No. of obs      =      7150
Optimization     : ML                             Residual df     =      7133

                                (Std. Err. adjusted for 1683 clusters in schid)
------------------------------------------------------------------------------
             |               Robust
       math4 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    lavgrexp |   .1227899   .0669842     1.83   0.067    -.0084967    .2540764
       lunch |  -.0831614   .1047519    -0.79   0.427    -.2884712    .1221485
      lenrol |  -.0556512   .0490405    -1.13   0.256    -.1517689    .0404665
         y95 |   .3186249   .0143788    22.16   0.000     .2904429    .3468068
         y96 |   .3647386   .0189796    19.22   0.000     .3275393    .4019379
         y97 |   .2860664   .0201033    14.23   0.000     .2466647    .3254681
         y98 |   .6760248   .0217182    31.13   0.000     .6334578    .7185917
   lavgrexpb |   .1658169     .08903     1.86   0.063    -.0086787    .3403124
      lunchb |  -.0113902   .0010958   -10.39   0.000    -.0135381   -.0092424
     lenrolb |   .0202697   .0531842     0.38   0.703    -.0839694    .1245088
        y95b |   .9325259   .3529265     2.64   0.008     .2408026    1.624249
        y96b |   .5439736   .1438847     3.78   0.000     .2619647    .8259826
        y97b |   .6807815   .2587424     2.63   0.009     .1736557    1.187907
        y98b |   .2624711    .338214     0.78   0.438    -.4004161    .9253584
       tobs3 |  -.0431248    .044767    -0.96   0.335    -.1308666     .044617
       tobs4 |  -.0771368   .0413601    -1.87   0.062     -.158201    .0039274
       _cons |  -2.194583   .5328879    -4.12   0.000    -3.239024   -1.150142
------------------------------------------------------------------------------
```

```
. margins, dydx(lavgrexp)

Average marginal effects                        Number of obs   =       7150
Model VCE    : Robust

Expression   : Predicted mean math4, predict()
dy/dx w.r.t. : lavgrexp

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    lavgrexp |    .043285   .0236081     1.83   0.067    -.0029861    .0895561
------------------------------------------------------------------------------
```