UNBALANCED PANELS, SAMPLE SELECTION, AND ATTRITION

Econometric Analysis of Cross Section and Panel Data, 2e
MIT Press
Jeffrey M. Wooldridge

1. Introduction

- Unbalanced panel data sets often arise in practice. Estimating linear models with unbalanced panels is relatively easy (by POLS, RE, FE, and IV versions of these). The important question is: why are some time periods missing for some units?
- An important issue in the presence of unbalanced data: some estimators have advantages over others. For example, removing an unobserved effect allows more sample selection than RE or correlated RE approaches.

- In some cases, such as when an entire year is skipped for everyone, or when units are randomized out of a rotating panel, the sample selection can be assumed to be exogenous.
- In other cases, such as a wage offer function – as we saw for cross section data – selection might be fundamentally related to unobservables in the equation of interest.
- Panel data brings the additional complication of **attrition**, especially with disaggregated data (such as individuals, families, firms).

2. Pooled OLS Estimation with Unbalanced Panels

- We think about unbalanced panels as follows. Let $t = 1, \ldots, T$ be the time periods that describe the population of interest. In practice, these $t = 1$ is the earliest time a unit can appear, and $t = T$ is the latest time.
- A randomly drawn unit from the population results in "data" $\{(\mathbf{x}_{it}, y_{it}, s_{it}) : t = 1, \ldots, T\}$ where s_{it} is the selection indicator: $s_{it} = 1$ if we observe (all of) $(\mathbf{x}_{it}, y_{it})$, and zero otherwise.
- We will stay with large N (number of cross section draws), small T asymptotics.

• Start with the population model

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + v_{it}$$
$$E(\mathbf{x}_{it}'v_{it}) = \mathbf{0},\ t = 1, \ldots, T,$$

which we assume makes sense for all time periods. At this point, we leave the nature of v_{it} open.

• Because the s_{it} are random, the number of time periods observed for unit i, given by

$$T_i = \sum_{t=1}^{T} s_{it},$$

is properly viewed as random. T_i takes a value in $\{0, 1, \ldots, T\}$.

• Pooled OLS using the selected data, that is, the data where we observe $(\mathbf{x}_{it}, y_{it})$:

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

which we can write as

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

Consistency of POLS (T fixed, $N \to \infty$) follows from

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

and

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

• The rank condition essentially means the rank condition holds in the population and we do not select out “too little” of the population.

• The orthogonality condition is a kind of exogeneity requirement. If s_{it} is independent of $(\mathbf{x}_{it}, v_{it})$, that is, independent of $(\mathbf{x}_{it}, y_{it})$, then $E(s_{it}\mathbf{x}_{it}'v_{it}) = E(s_{it})E(\mathbf{x}_{it}'v_{it}) = \mathbf{0}$.

• If

$$E(v_{it}|\mathbf{x}_{it}, s_{it}) = 0$$

then $E(s_{it}\mathbf{x}_{it}'v_{it}) = \mathbf{0}$. Sufficient for the zero conditional mean is

$$E(v_{it}|\mathbf{x}_{it}, \mathbf{w}_{it}) = 0$$
$$s_{it} = h_t(\mathbf{x}_{it}, \mathbf{w}_{it})$$

• The condition $E(v_{it}|\mathbf{x}_{it}, \mathbf{w}_{it}) = 0$ means

$E(y_{it}|\mathbf{x}_{it}, \mathbf{w}_{it}) = E(y_{it}|\mathbf{x}_{it}) = \mathbf{x}_{it}\boldsymbol{\beta}$, so $\mathbf{w}_{it}$ is properly excluded from the conditional mean model.

• For example, if the conditional mean is dynamically complete,

$$E(y_{it}|\mathbf{x}_{it}, y_{i,t-1}, \mathbf{x}_{i,t-1}, \ldots y_{i1}, \mathbf{x}_{i1}) = E(y_{it}|\mathbf{x}_{it}) = \mathbf{x}_{it}\boldsymbol{\beta},$$

then s_{it} can depend on any elements in $(\mathbf{x}_{it}, y_{i,t-1}, \mathbf{x}_{i,t-1}, \ldots y_{i1}, \mathbf{x}_{i1})$.

• Selection is allowed to be correlated with v_{ir}, $r \neq t$, but this is not especially helpful if $v_{it} = c_i + u_{it}$.

• In general POLS inference should be made robust to arbitrary serial correlation and heteroskedasticity, as in the balanced case.

- In Stata, the command

```
reg y x1 ... xK, cluster(csid)
```

where “csid” is the cross section identifier, does not care if the panel is unbalanced. Of course, it is up to us to determine whether selection might be endogenous, that is, $E(s_{it}\mathbf{x}_{it}'v_{it}) \neq \mathbf{0}$.

- Because of the panel structure, simple tests for certain kinds of selection bias are possible, but they are indirect tests. By definition, s_{it} is always unity when data are actually used. So we cannot use s_{it} as an added regressor.

• But we can add functions of $\{s_{ir} : r = 1, \ldots, T\}$, such as T_i, and use a robust t test. Remember, this is effectively testing whether v_{it} is uncorrelated with s_{ir}, $r \neq t$, and consistency of POLS does not rely on their being uncorrelated.

• Can also use lags or leads, $s_{i,t-1}$ or $s_{i,t+1}$, and test their significance. (Lose the first or last time period in doing so.) But, again, consistency of POLS does not directly require v_{it} to be uncorrelated with s_{ir}, $r \neq t$, once $\mathbf{x}_{it}$ has been controlled for.

3. FE and RE Estimation with Unbalanced Panels

- Now explicitly include an additive, unobserved effect model for random draw i:

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

- Before we discuss FE and RE, when would POLS be consistent? Assuming the rank condition, we effectively need to assume

$$E(s_{it}\mathbf{x}_{it}'c_i) = 0, E(s_{it}\mathbf{x}_{it}'u_{it}),$$

which rules out selection as a function of the unobserved heterogeneity, c_i, or the idiosyncratic error, u_{it}.

Fixed Effects

• Fixed effects is now applied to the unbalanced sample. The time-demeaned data now uses different time periods for different i. Let

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

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

- The FE estimator is then

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

We can write

$$\begin{aligned} \ddot{y}_{it} &= (\mathbf{x}_{it}\boldsymbol{\beta} + c_i + u_{it}) - \left[\left(T_i^{-1} \sum_{r=1}^{T} s_{ir} x_{ir} \beta \right) + c_i + T_i^{-1} \sum_{r=1}^{T} s_{ir} u_{ir} \right] \\ &= \ddot{\mathbf{x}}_{it}\boldsymbol{\beta} + \ddot{u}_{it}, t = 1, \ldots, T, \end{aligned}$$

which looks like the balanced case.

• Algebra gives

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

• Consistency of FE on the selected sample follows from the POLS analysis:

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

and

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

• Note that $\ddot{\mathbf{x}}_{it}$ depends on $\mathbf{x}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT})$ and $\mathbf{s}_i = (s_{i1}, \ldots, s_{iT})$, so it is not enough to assume, say, s_{it} is independent of $(\mathbf{x}_{it}, u_{it})$.

• A sufficient condition is an extension of the usual strict exogeneity assumption:

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

• This rules out selection in any time period depending on the shocks in any time period. That is, this condition is generally violated if $Cov(s_{ir}, u_{it}) \neq 0$ for any (r, t) pair.

• Importantly, both conditions allow for s_{it} to depend on c_i in an unrestricted way.

- Using unbalanced panels with FE is straightforward in practice. The usual “xtreg” command in Stata allows for unbalanced panels and properly computes standard errors and test statistics.
- Note that any cross-sectional unit with only a single time period plays no role in the estimation; it drops out. Some say this leads to a “selection bias” using FE, but it does not provided the exogeneity condition holds. For example, some values of c_i may be associated with dropping out after one period, and FE removes the source of the selection bias.

- For the usual, nonrobust inference to be valid, one needs

$$E(\mathbf{u}_i\mathbf{u}_i'|\mathbf{x}_i, \mathbf{s}_i, c_i) = \sigma_u^2\mathbf{I}_T,$$

or something very close to it. (Obvious extension of FE.3.) The estimate of σ_u^2 is now

$$\hat{\sigma}_u^2 = \left[\sum_{i=1}^{N}(T_i - 1) - K\right]^{-1} \sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{\hat{u}}_{it}^2,$$

where the $\ddot{\hat{u}}_{it}$ are the usual FE residuals (computed only when $s_{it} = 1$).

- Since fixed effects is inconsistent if sample selection is not strictly exogenous, the sample selection indicator from other time periods should be insignificant at time t.
- We may be particular concerned with a shock today (change in u_{it}) makes it more likely the data are missing at $t + 1$.

• Suggests a simple test: add $s_{i,t+1}$ to the equation at time t (so that the last time period is lost), estimate the model by fixed effects (using the unbalanced panel), and compute the (robust) t statistic on $s_{i,t+1}$. This works quite generally, including for attrition when it is an absorbing state. We need $T \geq 3$ (and at least three time periods for some units).

• Can check for strict exogeneity of the covariates at the same time: add $(\mathbf{x}_{i,t+1}, s_{t+1})$ and use a joint test.

• In some cases, can use $s_{i,t-1}$, but this does not work in the attrition case because if $s_{it} = 1$ – that is, we use the time t observation for individual i – then $s_{i,t-1} = 1$, too.

• Other choices: $\sum_{r=1}^{t-1} s_{ir}$ and $\sum_{r=t+1}^{T} s_{ir}$; the latter works for attrition, too.

• To check for bias caused by selection in the context of a random coefficient model, add $1[T_i = 2] \cdot \mathbf{x}_{it}, \ldots, 1[T_i = T-1] \cdot \mathbf{x}_{it}$, estimate the augmented model by FE, and obtain a joint Wald test. (The $T_i = T$ group is the base group.) This is like a Chow test where the slopes are allowed to differ by the number of available time periods for each unit.

Random Effects

• Mechanically, RE is fairly simple to modify. As in the balanced case, it can be obtained as the POLS estimator on quasi-time-demeaned data. But now the fraction of the mean we remove depends on T_i:

$$\hat{\lambda}_i = 1 - \left\{ \frac{1}{[1 + T_i(\hat{\sigma}_c^2/\hat{\sigma}_u^2)]} \right\}^{1/2}.$$

Now define

$$\check{y}_{it} = y_{it} - \hat{\lambda}_i \bar{y}_i$$

where $\bar{y}_i = T_i^{-1} \sum_{r=1}^{T} s_{ir} y_{ir}$, and similarly for $\check{\mathbf{x}}_{it}$. Then, POLS of $\check{y}_{it}$ on $\check{\mathbf{x}}_{it}$ using the $s_{it} = 1$ data points.

• Ignore estimation of σ_c^2 and σ_u^2 in $\hat{\lambda}_i$:

$$y_{it} - \lambda_i \bar{y}_i = (\mathbf{x}_{it} - \lambda_i \bar{\mathbf{x}}_i)\boldsymbol{\beta} + (1 - \lambda_i)c_i + (u_{it} - \lambda_i \bar{u}_i).$$

• c_i is not eliminated, and so we need a stronger assumption about selection being unrelated to c_i, as in POLS case. But we also need strict exogeneity of selection, as in FE case. Sufficient are

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

which explicitly rules out selection that depends on c_i.

- If use RE on the unbalanced panel, should still obtain robust inference.
- Stata labels $\hat{\lambda}_i$ as "theta." The command

```
xtreg y x1 ... xK, re cluster(csid) theta
```

allows you to see the range of the $\hat{\lambda}_i$.

- Because time-constant variables can be included in RE, can add T_i to test for selection bias, and use interactions with $\mathbf{x}_{it}$, as with FE.

• Differencing methods across different pairs of time periods works under the FE assumptions. As a practical matter, helps to have T rows for each i in storing the data. (We will use this below for the special case of attrition.)

• If you use FD, lose more data than with FE: a time period is used with FE only if the previous time period is also available. (It does work well for attrition, where $s_{it} = 1$ means $s_{ir} = 1, r < t$.)

• For IV estimation, can show that fixed effects 2SLS (and any other IV method that first eliminates c_i) IV is consistent under $E(u_{it}|\mathbf{z}_i, \mathbf{s}_i, c_i) = 0, t = 1, \ldots, T$.

4. Tests for Selection Based on Inverse Mills Ratio

- Tests above for FE (and RE) do not allow for direct tests of whether s_{it} is correlated with u_{it}.
- Now consider the incidental truncation problem, where we can derive such tests. We assume that we always observe $\mathbf{x}_{it}$, but y_{it} is observed only when $s_{it} = 1$. Wage offer/LFP example.
- Write the linear unobserved effects model with sample selection as

$$y_{it1} = \mathbf{x}_{it1}\boldsymbol{\beta}_1 + c_{i1} + u_{it1}$$
$$s_{it2} = 1[\mathbf{x}_{it}\boldsymbol{\delta}_2 + \psi_2 + \bar{\mathbf{x}}_i\boldsymbol{\xi}_2 + v_{it2} \geq 0], t = 1, \ldots, T$$

where the second equation is a reduced form selection equation that uses the Chamberlain-Mundlak device.

• To make the approach believable, have imposed an exclusion restriction: something in $\mathbf{x}_{it}$ not in $\mathbf{x}_{it1}$, and that something should vary over time.

• It is possible to make the selection model more general. Would have a full set of time dummies in $\mathbf{x}_{it}$ (and $\mathbf{x}_{it1}$), but we might want more than the intercept to change with time. For example,

$$s_{it2} = 1[\mathbf{x}_{it}\boldsymbol{\delta}_{t2} + \psi_{t2} + \bar{\mathbf{x}}_i\boldsymbol{\xi}_{t2} + v_{it2} \geq 0]$$

or even

$$s_{it2} = 1[\psi_{t2} + \mathbf{x}_i\boldsymbol{\xi}_{t2} + v_{it2} \geq 0],$$

where $\mathbf{x}_i$ is $1 \times TK$ (without time period dummies).

• To estimate the selection model, we make a standard probit assumption

$$v_{it2}|\mathbf{x}_i \sim Normal(0,1), t = 1,\ldots,T,$$

so that, say,

$$P(s_{it2} = 1|\mathbf{x}_i) = \Phi(\mathbf{x}_{it}\boldsymbol{\delta}_2 + \psi_2 + \bar{\mathbf{x}}_i\boldsymbol{\xi}_2), t = 1,\ldots,T.$$

and we can estimate the parameters using pooled probit.

• For more flexibility, we can estimate the selection model separately for each t.

• Define the inverse Mills ratios:

$$\hat{\lambda}_{it2} \equiv \lambda(\mathbf{x}_{it}\hat{\boldsymbol{\delta}}_2 + \hat{\psi}_2 + \bar{\mathbf{x}}_i\hat{\boldsymbol{\xi}}_2) \text{ all } i, t.$$

Then a valid test is based on FE estimation using the selected sample, but adding $\hat{\lambda}_{it2}$ as an additional regressor. Under the null of no sample selection, we can use the usual t statistic, made robust to serial correlation and heteroskedasticity.

• Can interact $\hat{\lambda}_{it2}$ with time dummies to get at joint test with T degrees-of-freedom.

- (Potentially) Important: Adding the IMR to fixed effects estimation on the unbalanced panel does not generally produce a consistent estimate of β_1 if there is a sample selection problem. (See text, pages 582-583 for discussion.)
- As usual with small *T*, do not try to estimate the fixed effects inside the probit in forming the IMR.

5. Heckman Approaches to Correcting for Sample Selection Bias

- Approach: Use the Chamberlain-Mundlak device in the structural equation, too. Write the equation as

$$y_{it1} = \mathbf{x}_{it1}\boldsymbol{\beta}_1 + \psi_1 + \bar{\mathbf{x}}_i\boldsymbol{\xi}_1 + v_{it1},$$

where the composite error is $v_{it1} = a_{i1} + u_{it1}$ and $c_{i1} = \psi_1 + \bar{\mathbf{x}}_i\boldsymbol{\xi}_1 + a_{i1}$.

- Note that we have the time average of all elements of $\mathbf{x}_{it}$ in this equation.

- As usual, the C-M device is more restrictive than not specifying $D(c_{i1}|\mathbf{x}_i)$ at all. That is why using FE for the *test* is preferred.

• Assume

$$E(v_{it1}|\mathbf{x}_i, v_{it2}) = E(v_{it1}|v_{it2}) = \gamma_1 v_{it2}, t = 1, \ldots, T.$$

(Sufficient for the first equality is that (v_{it1}, v_{it2}) is independent of $\mathbf{x}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT})$, which is fairly standard.) Now we can write

$$y_{it1} = \mathbf{x}_{it1}\boldsymbol{\beta}_1 + \psi_1 + \bar{\mathbf{x}}_i\boldsymbol{\xi}_1 + \gamma_1 E(v_{it2}|\mathbf{x}_i, s_{it2}) + e_{it1}$$

where, by construction,

$$E(e_{it1}|\mathbf{x}_i, s_{it2}) = 0, t = 1, \ldots, T.$$

• We can find $E(v_{it2}|\mathbf{x}_i, s_{it2})$ from the probit model for s_{it2} (which has error v_{it2}). Call this function $h_{t2}(\mathbf{x}_i, s_{it2})$. Then we have, by definition,

$$y_{it1} = \mathbf{x}_{it1}\boldsymbol{\beta}_1 + \psi_1 + \bar{\mathbf{x}}_i\boldsymbol{\xi}_1 + \gamma_1 h_{t2}(\mathbf{x}_i, s_{it2}) + e_{it1},$$

$$E(e_{it1}|\mathbf{x}_i, s_{it2}) = 0, t = 1, \ldots, T.$$

• From earlier result, can applied pooled OLS to the augmented equation using the selected sample, to consistently estimate all the parameters.

• In particular, if we know $h_{t2}(\mathbf{x}_i, s_{it2})$ whenever $s_{it2} = 1$, then we can consistently estimate $\boldsymbol{\beta}_1, \psi_1, \boldsymbol{\xi}_1$ and γ_1 from the pooled OLS regression

$$y_{it1} \text{ on } 1, \mathbf{x}_{it1}, \bar{\mathbf{x}}_{ii}, h_{t2}(\mathbf{x}_i, 1)$$

using the observed data; that is, for all (i, t) with $s_{it2} = 1$. But

$$h_{t2}(\mathbf{z}_i, 1) = \lambda(\mathbf{x}_{it}\boldsymbol{\delta}_{t2} + \psi_{t2} + \bar{\mathbf{x}}_i\boldsymbol{\xi}_{t2})$$

is just the IMR, which we can estimate from the first-stage probits. (As specified, separately for each t.)

- So, we estimate the equation

$$y_{it1} = \mathbf{x}_{it1}\boldsymbol{\beta}_1 + \psi_1 + \bar{\mathbf{x}}_i\boldsymbol{\xi}_1 + \gamma_1\hat{\lambda}_{it2} + error_{it1}$$

by pooled OLS on the selected sample.

- Can use the delta method to correct for two-step estimation. Easier is the panel bootstrap. Again, we draw units when resampling, and we keep whatever time periods are available for each unit.

- In practice, using a valid t statistic for $\hat{\gamma}_1$ is likely to be similar to using the FE version of the test. Under the null, can use robust FE standard error.

- Can allow more flexibility by allowing a different coefficient on $\hat{\lambda}_{it2}$ for each t; just add interactions $d2_t \cdot \hat{\lambda}_{it2}, d3_t \cdot \hat{\lambda}_{it2}, \ldots, dT_t \cdot \hat{\lambda}_{it2}$. This is gotten by using $E(v_{it1}|v_{it2}) = \gamma_{t1} v_{it2}$.
- As in cross section case, can make the functional form more general:

$$E(v_{it1}|v_{it2}) = \gamma_{t1} v_{it2} + \eta_{t1}(v_{it2}^2 - 1)$$

or assume constant coeffients.

• Extending to the IV case is fairly straightforward. (Semykina and Wooldridge (2010, *Journal of Econometrics*)). The structural equation looks like

$$y_{it1} = \mathbf{x}_{it1}\boldsymbol{\beta}_1 + c_{i1} + u_{it1}$$

but now u_{it1} might be correlated with some elements of $\mathbf{x}_{it1}$. Plus, we now allow elements of $\mathbf{x}_{it1}$ to also be missing, provided we have instruments for them.

• Let $\mathbf{z}_{it}$ be the set of all exogenous variables that are always observed. (Some are exogenous regressors, if always observed, some are external instruments.)

- Ideally the IVs are time-varying to avoid making strong assumptions.

• We assume strict exogeneity of $\{\mathbf{z}_{it} : t = 1, \ldots, T\}$ conditional on c_{i1}, that is,

$$E(u_{it1}|\mathbf{z}_i) = 0, t = 1, \ldots, T.$$

• We allow correlation of $\{\mathbf{z}_{it} : t = 1, \ldots, T\}$ with c_{i1} via the C-M device, now expressed as

$$c_{i1} = \psi_1 + \bar{\mathbf{z}}_i \boldsymbol{\xi}_1 + a_{i1}$$
$$E(a_{i1}|\mathbf{z}_i) = 0.$$

Again, time averages of all exogenous variables are in $\bar{\mathbf{z}}_i$. But this excludes endogenous elements of $\mathbf{x}_{it1}$ or elements of $\mathbf{x}_{it1}$ we do not always observe.

• Under assumptions similar to those above, we can write

$$y_{it1} = \psi_1 + \mathbf{x}_{it1}\boldsymbol{\beta}_1 + \bar{\mathbf{z}}_i\boldsymbol{\xi}_1 + \gamma_1 E(v_{it2}|z_i, s_{it2}) + e_{it1}$$

where, by construction,

$$E(e_{it1}|\mathbf{z}_i, s_{it2}) = 0, t = 1, \ldots, T.$$

• This leads naturally to a 2SLS procedure: After obtaining the $\hat{\lambda}_{it2}$ from probits $P(s_{it2} = 1|\mathbf{z}_i) = \Phi(\mathbf{z}_{it}\boldsymbol{\delta}_{t2} + \psi_{t2} + \bar{\mathbf{z}}_i\boldsymbol{\xi}_{t2})$, apply pooled 2SLS to

$$y_{it1} = \mathbf{x}_{it1}\boldsymbol{\beta}_1 + \psi_1 + \bar{\mathbf{z}}_i\boldsymbol{\xi}_1 + \gamma_1\hat{\lambda}_{it2} + error_{it1}$$

using IVs $(1, \mathbf{z}_{it}, \bar{\mathbf{z}}_i, \hat{\lambda}_{it2})$.

- Again, can bootstrap the standard errors or use delta method.
- Under $H_0 : \gamma_1 = 0,$ can use a serial correlation/heteroskedasticity-robust t statistic.
- But, for testing $H_0 : \gamma_1 = 0$, it is better to use FE2SLS on the equation

$$y_{it1} = \mathbf{x}_{it1}\boldsymbol{\beta}_1 + \gamma_1\hat{\lambda}_{it2} + error_{it1}$$

using IVs $(\mathbf{z}_{it}, \hat{\lambda}_{it2})$. It maintains fewer assumptions under the null.

$$y_{it1} = \mathbf{x}_{it1}\beta_1 + \psi_1 + \bar{\mathbf{z}}_i\xi_1 + \gamma_1\hat{\lambda}_{it2} + error_{it1}$$

by FE-2SLS on the selected sample, using as instruments $(\mathbf{z}_{it}, \bar{\mathbf{z}}_i, \hat{\lambda}_{it2})$.

• Can also accomodate a Tobit selection equation. Consider the case of exogenous covariates always observed:

$$y_{it2} = \max(0, \psi_2 + \mathbf{x}_{it}\boldsymbol{\delta}_2 + \bar{\mathbf{x}}_i\boldsymbol{\xi}_2 + v_{it2}), t = 1, \ldots, T$$

or one of the extensions where the parameters can vary across t. Assume

$$v_{it2}|\mathbf{x}_i \sim Normal(0, \tau_2^2), t = 1, \ldots, T.$$

• If we also assume $E(v_{it1}|\mathbf{x}_i, v_{it2}) = E(v_{it1}|v_{it2}) = \gamma_1 v_{it2}$, then we get the same expectation $E(y_{it1}|x_i, v_{it2})$ as before. But now we can effectively observe v_{it2} whenever $s_{it2} = 1$. So, we use the basic fact that

$$E(y_{it1}|\mathbf{x}_i, v_{it2}, s_{it2}) = \mathbf{x}_{it1}\beta_1 + \psi_1 + \bar{\mathbf{x}}_i\xi_1 + \gamma_1 v_{it2}, t = 1, \ldots, T.$$

- Let $\hat{\psi}_2$, $\hat{\boldsymbol{\delta}}_2$, and $\hat{\boldsymbol{\xi}}_2$ be the Tobit estimates, and define

$$\hat{v}_{it2} = y_{it2} - \hat{\psi}_2 - \mathbf{x}_{it}\hat{\boldsymbol{\delta}}_2 - \bar{\mathbf{x}}_i\hat{\boldsymbol{\xi}}_2 \text{ if } s_{it2} = 1.$$

Then, we just use $\hat{v}_{it2}$ in place of $\hat{\lambda}_{it2}$:

$$y_{it1} = \mathbf{x}_{it1}\boldsymbol{\beta}_1 + \psi_1 + \bar{\mathbf{x}}_i\boldsymbol{\xi}_1 + \gamma_1\hat{v}_{it2} + error_{it1}$$

using the selected sampl. Of course, we can do a *t* test, as usual, to test for sample selection bias in this framework.

6. Heckman Corrections for Attrition

- Again start with the UE model

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + c_i + u_{it}, t = 1, \ldots, T, \tag{31}$$

and again let s_{it} be the selection indicator.

- Assume a random sample from the population at time $t = 1$. In other words, $s_{i1} \equiv 1$ for all i. With attrition, some units leave the sample in subsequent time periods.

- Assume that once a unit attrits from the sample, we we observe nothing about them; in other words, attrition is an *absorbing state*. (Can always arrange this by simply ignoring any information on returning units.)
- Conceptually, this setup can be problematical. It assumes that we are interested in a population defined at $t = 1$. But what if it is a population of firms, and some firms close or merge? What is the "right" population?

- Under attrition as an absorning state,

$$s_{it} = 1 \Rightarrow s_{ir} = 1, r < t.$$

- Now differencing is attractive as a way of removing c_i:

$$\Delta y_{it} = \Delta \mathbf{x}_{it}\boldsymbol{\beta} + \Delta u_{it}, t = 2, \ldots, T.$$

- We observe Δy_{it} and $\Delta \mathbf{x}_{it}$ whenever $s_{it} = 1$; we do not have to separately worry about $s_{i,t-1}$, as in cases with general missing data patterns.

• A natural starting point is to assume $\{\mathbf{x}_{it} : t = 1, \ldots, T\}$ is strictly exogenous, conditional on c_i (in the population, as always). At a minimum, $\Delta\mathbf{x}_{it}$ is uncorrelated with Δu_{it}.

• Let $\mathbf{w}_{it}$ be a set of variables that we always observe when $s_{i,t-1} = 1$ such that $\mathbf{w}_{it}$ is a good predictor of selection – in a sense soon to be made precise.

• Model the selection in time period t conditional on $s_{i,t-1} = 1$ as

$$s_{it} = 1[\mathbf{w}_{it}\boldsymbol{\delta}_t + v_{it} > 0]$$
$$v_{it}|(\mathbf{w}_{it}, \Delta\mathbf{x}_{it}, s_{i,t-1} = 1) \sim Normal(0, 1), t = 2, 3, \ldots, T.$$

• Importantly, at least some elements of $\mathbf{x}_{it}$ are not observed at time t for those who attrit in time t, so $\mathbf{x}_{it}$ is not (fully) contained in $\mathbf{w}_{it}$.

• Specify sequential probit models:

$$P(s_{it} = 1|\mathbf{w}_{it}, s_{i,t-1} = 1) = \Phi(\mathbf{w}_{it}\boldsymbol{\delta}_t), t = 2, \ldots, T.$$

• Use a sequence of probits starting with $t = 2$. For $t = 2,$ we use the entire sample to estimate a probit for still being in the sample in the second period. For $t = 3$, we estimate a probit for those units still in the sample as of $t = 2$. At $t = T$, we have the smallest group of observations because we only use units still in the sample as of $T - 1$.

• Where might the $\mathbf{w}_{it}$ come from? Since they have to be observed at time t for the entire subgroup with $s_{i,t-1} = 1$, $\mathbf{w}_{it}$ generally cannot contain variables dated at time t (unless some information is known at time t on people who attrit at time t). When the $\mathbf{x}_{it}$ are strictly exogenous, we can always include in $\mathbf{w}_{it}$ elements of $(\mathbf{x}_{i,t-1}, \mathbf{x}_{i,t-2}, \ldots, \mathbf{x}_{i1})$.

• The potential dimension of $\mathbf{w}_{it}$ grows as we move ahead through time, which is why a separate selection model should be estimated at each t.

- $y_{i,t-1}$ cannot be in $\mathbf{w}_{it}$ because $y_{i,t-1}$ is necessarily correlated with Δu_{it}.
- Nevertheless, if we assume dynamic completeness of the form

$$E(u_{it}|\mathbf{x}_i, y_{i,t-1}, \ldots, y_{i1}, c_i) = 0, t = 2, \ldots, T,$$

then elements from $(y_{i,t-2}, y_{i,t-3}, \ldots, y_{i1})$ can be in $\mathbf{w}_{it}$.

- Dynamic completeness in this context is very restrictive because if we impose strict exogeneity on $\mathbf{x}_{it}$ then it cannot include lagged y_{it}, and that makes dynamic completeness unlikely.

• In what sense do we need the $\mathbf{w}_{it}$ to be good predictors of attrition? A sufficient condition is, given $s_{i,t-1} = 1$,

$$(\Delta u_{it}, v_{it}) \text{ is independent of } (\Delta \mathbf{x}_{it}, \mathbf{w}_{it}).$$

• Δu_{it} independent of $(\Delta \mathbf{x}_{it}, \mathbf{w}_{it})$: true if $\mathbf{w}_{it}$ contains only lags of $\mathbf{x}_{it}$ because $\{\mathbf{x}_{it}\}$ is strictly exogenous.

• v_{it} independent of $(\Delta \mathbf{x}_{it}, \mathbf{w}_{it})$: can be very restrictive because $\Delta \mathbf{x}_{it}$ cannot be included in $\mathbf{w}_{it}$ in interesting cases (because $\mathbf{x}_{it}$ is not observed for everyone with $s_{i,t-1} = 1$). In other words, the (reduced form) selection equation cannot include $\Delta \mathbf{x}_{it}$, which means we are assuming

$$P(s_{it} = 1|\Delta \mathbf{x}_{it}, \mathbf{w}_{it}, s_{i,t-1} = 1) = P(s_{it} = 1|\mathbf{w}_{it}, s_{i,t-1} = 1)$$

• In practice, $\mathbf{w}_{it}$ includes $\mathbf{x}_{i,t-1}$, but we cannot allow attrition to depend on the changes $\Delta\mathbf{x}_{it}$.

• Important point: If

$$P(s_{it} = 1|\Delta\mathbf{x}_{it}, \Delta u_{it}, s_{i,t-1} = 1) = P(s_{it} = 1|\Delta\mathbf{x}_{it}, s_{i,t-1} = 1)$$

then the pooled OLS estimator from

$$\Delta y_{it} \text{ on } \Delta\mathbf{x}_{it}, \; i = 1, \ldots N; t = 2, \ldots, T \text{ using } s_{it} = 1$$

is consistent because $E(\Delta u_{it}|\Delta\mathbf{x}_{it}) = 0$. It would be better to ignore attrition!

• For the sake of argument, assume $(\Delta u_{it}, v_{it})$ is independent of $(\Delta \mathbf{x}_{it}, \mathbf{w}_{it})$, and also linearity:

$$E(\Delta u_{it} | v_{it}, s_{i,t-1} = 1) = \rho_t v_{it}, t = 2, \ldots, T.$$

Under the maintained assumptions we have

$$E(\Delta y_{it} | \Delta \mathbf{x}_{it}, \mathbf{w}_{it}, s_{it} = 1) = \Delta \mathbf{x}_{it} \boldsymbol{\beta} + \rho_t \lambda(\mathbf{w}_{it} \boldsymbol{\delta}_t), t = 2, \ldots, T.$$

• Two-step procedure: (1) Starting with $t = 2$, estimate a sequence of probit models for the group of units in the sample at time $t - 1$: probit of

$$s_{it} \text{ on } \mathbf{w}_{it} \text{ for the subsample with } s_{i,t-1} = 1.$$

The vector $\mathbf{w}_{it}$ grows as t increases. Obtain the inverse Mills ratios, $\hat{\lambda}_{it} \equiv \lambda(\mathbf{w}_{it}\hat{\boldsymbol{\delta}}_t)$

(2) Using the selected sample ($s_{it} = 1$), run the pooled OLS regression

$$\Delta y_{it} \text{ on } \Delta\mathbf{x}_{it}, d2_t\hat{\lambda}_{it}, d3_t\hat{\lambda}_{it}, \ldots, dT_t\hat{\lambda}_{it},$$

where allowing a different coefficient on $\hat{\lambda}_{it}$ in each time period is required because of the nature of the sequential procedure.

- A joint test of significance of the IMR terms (T – 1 restrictions) – made robust to serial correlation and heteroskedasticity – is a valid test of the null of no attrition bias.
- As usual, if there is evidence of attrition bias, the asymptotic variance matrix of $\hat{\boldsymbol{\beta}}$ needs to be adusted for the first-stage estimation of the $\hat{\boldsymbol{\delta}}_t$, possibly via bootstrapping.
- The fundamental problem here is that, because $\Delta\mathbf{x}_{it}$ cannot always be a subset of $\mathbf{w}_{it}$, a large difference between pooled OLS and the Heckman procedure does not mean the Heckman procedure is preferred.

• Can partly overcome the problems with the previous method and at the same time relax the strict exogeneity assumption (at least for some of the regressors).

• Still start with

$$\Delta y_{it} = \Delta \mathbf{x}_{it}\boldsymbol{\beta} + \Delta u_{it}, t = 2, \ldots, T.$$

• Now assume we have a vector of instrumental variables for $\Delta \mathbf{x}_{it}$; call this vector $\mathbf{z}_{it}$. The minimal requirement is that we observe $\mathbf{z}_{it}$ whenever $s_{it} = 1$.

• Like $\mathbf{w}_{it}$, $\mathbf{z}_{it}$ can include elements from $(\mathbf{x}_{i,t-1}, \ldots, \mathbf{x}_{i1})$, at least under the sequential exogeneity assumption

$$E(u_{it}|\mathbf{x}_{it}, \mathbf{x}_{i,t-1}, \ldots, \mathbf{x}_{i1}, c_i) = 0, t = 1, \ldots, T.$$

(Recall that this condition does allow for lagged dependent variables in $\mathbf{x}_{it}$). Now the key independence assumption is

$$(\Delta u_{it}, v_{it}) \text{ is independent of } (\mathbf{z}_{it}, \mathbf{w}_{it})$$

conditional on $s_{i,t-1} = 1$.

• Much more palatable than $(\Delta u_{it}, v_{it})$ conditionally independent of $(\Delta \mathbf{x}_{it}, \mathbf{w}_{it})$ because we can choose $\mathbf{z}_{it}$ to be a subset of $\mathbf{w}_{it}$, which we should do. Then it is sufficient that $(\Delta u_{it}, v_{it})$ is independent of $\mathbf{w}_{it}$ given $s_{i,t-1} = 1$.

• If we include in $\mathbf{z}_{it}$ some elements of $\Delta \mathbf{x}_{it}$, then the procedure can suffer from the same problems as before.

• On the other hand, if we are forced to use only elements of $(\mathbf{x}_{i,t-1}, \ldots, \mathbf{x}_{i1})$ in both $\mathbf{z}_{it}$ and $\mathbf{w}_{it}$, might have weak instruments (as in case without attrition).

- Really should have something in $\mathbf{w}_{it}$ that affects selection that is not needed in $\mathbf{z}_{it}$; it is unclear how to go about ensuring this in general.
- Can consistently estimate $\boldsymbol{\beta}$ by applying pooled 2SLS, on the selected sample, to

$$\Delta y_{it} = \Delta \mathbf{x}_{it}\boldsymbol{\beta} + \rho_2 d2_t\hat{\lambda}_{it} + \rho_3 d3_t\hat{\lambda}_{it} + \ldots + \rho_T dT_t\hat{\lambda}_{it} + error_{it}$$

with instruments $(\mathbf{z}_{it}, d2_t\hat{\lambda}_{it}, d3_t\hat{\lambda}_{it}, \ldots, dT_t\hat{\lambda}_{it})$.

7. IPW for Attrition

- Previous methods do not extend easily to nonlinear models, except in special cases.
- As in cross section case, IPW weighting uses different assumptions about the nature of selection. “Ignorability of selection” or “selection on observables.”
- Again, general M-estimation framework. The population problem is somewhat abstract, but applies to many cases. Let $\boldsymbol{\theta}_o$ uniquely solve the problem

$$\min_{\theta \in \Theta} \sum_{t=1}^{T} E[q_t(\mathbf{w}_{it}, \boldsymbol{\theta})].$$

- Again, think of drawing $\{(\mathbf{w}_{i1}, s_{i1}), (\mathbf{w}_{i2}, s_{i2}), \ldots, (\mathbf{w}_{iT}, s_{iT})\}$
- Least squares is simply $q_t(\mathbf{w}_{it}, \boldsymbol{\theta}) = [y_{it} - m(\mathbf{x}_{it}, \boldsymbol{\theta})]^2$. With linear model, may apply to level or FD data.
- General MLE allowed, too.
- Assume $s_{i1} = 1$ for all i, and $s_{it} = 1 \Rightarrow s_{i,t-1} = 1$ (attrition as an absorbing state).
- Estimation ignoring attrition means that we solve the estimation problem

$$\min_{\theta \in \Theta} N^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} s_{it} q_t(\mathbf{w}_{it}, \boldsymbol{\theta}).$$

Call this the unweighted M-estimator, $\hat{\boldsymbol{\theta}}_u$.

- For consistency of this estimator we need for $\boldsymbol{\theta}_o$ to uniquely solve

$$\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \sum_{t=1}^{T} E[s_{it} q_t(\mathbf{w}_{it}, \boldsymbol{\theta})].$$

When we break $\mathbf{w}_{it}$ into endogenous and exogenous variables, $\boldsymbol{\theta}_o$ does not generally solve the minimization problem over the selected subpopulation when s_{it} is "correlated" with y_{it} after conditioning on $\mathbf{x}_{it}$.

- For nonlinear least squares, or quasi-MLE in the LEF, where we have

$$E(y_{it}|\mathbf{x}_{it}) = m_t(\mathbf{x}_{it}, \boldsymbol{\theta}_o), t = 1, \ldots, T,$$

the unweighted estimator is consistent when

$P(s_{it} = 1|\mathbf{x}_{it}, y_{it}) = P(s_{it} = 1|\mathbf{x}_{it})$, but not usually otherwise.

- To allow for endogenous selection, adopt a weighting scheme.
- Generally, at time t, let $\mathbf{r}_{it}$ be a set of variables such that

$$P(s_{it} = 1|\mathbf{w}_{it}, \mathbf{r}_{it}) = P(s_{it} = 1|\mathbf{r}_{it}) \equiv p_{it} > 0, t = 1, \ldots, T.$$

(The "obvious" set of variables $\mathbf{r}_{it} = \mathbf{w}_{it}$ is not usually available since we will have to estimate the probabilities.)

• Note how we assume the probabilities are strictly positive. The IPW M-estimator, $\hat{\boldsymbol{\theta}}_w$, solves

$$\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} N^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}(s_{it}/p_{it})q_t(\mathbf{w}_{it},\boldsymbol{\theta}),$$

• As before, the key step to show consistency is to show the expected value of the weighted objective function equals the population expectation:

$$E[(s_{it}/p_{it})q_t(\mathbf{w}_{it},\boldsymbol{\theta})] = E[q_t(\mathbf{w}_{it},\boldsymbol{\theta})], t = 1,\ldots,T.$$

• This follows from iterated espectations IF we make the ignorability assumption (51):

$$
\begin{aligned}
E[(s_{it}/p_{it})q_t(\mathbf{w}_{it},\boldsymbol{\theta})] &= E\{E[(s_{it}/p_{it})q_t(\mathbf{w}_{it},\boldsymbol{\theta})|\mathbf{w}_{it},\mathbf{r}_{it}]\} \\
&= E\left\{\frac{E(s_{it}|\mathbf{w}_{it},\mathbf{r}_{it})}{p_{it}}q_t(\mathbf{w}_{it},\boldsymbol{\theta})\right\} \\
&= E\left\{\frac{P(s_{it}=1|\mathbf{r}_{it})}{p_{it}}q_t(\mathbf{w}_{it},\boldsymbol{\theta})\right\} \\
&= E\left\{\frac{p_{it}}{p_{it}}q_t(\mathbf{w}_{it},\boldsymbol{\theta})\right\} = E[q_t(\mathbf{w}_{it},\boldsymbol{\theta})].
\end{aligned}
$$

• Issues: (i) How do we choose the $\mathbf{r}_{it}$? (ii) How do we estimate the p_{it}? (iii) How do we do inference on $\hat{\boldsymbol{\theta}}_w$?

• Two standard ways to choose $\mathbf{r}_{it}$: (1) Let $\mathbf{z}_{i1}$ be a vector of variables that we observe for every cross sectional unit in the first time period. Then $\mathbf{r}_{it} = \mathbf{z}_{i1}$ for all $t \geq 2$. This requires the strong assumption

$$P(s_{it} = 1|\mathbf{w}_{it}, \mathbf{z}_{i1}) = P(s_{it} = 1|\mathbf{z}_{i1}), t = 2, \ldots, T.$$

• If this condition holds, estimation is fairly straightforward: In each time period, estimate flexible probit or logit models, of $P(s_{it} = 1|\mathbf{z}_{i1})$. Pooling does not make sense.

• (2) Build the p_{it} up in a sequential fashion. At time t, $\mathbf{z}_{it}$ is a set of variables observed for the subpopulation with $s_{i,t-1} = 1$. ($s_{i0} \equiv 1$ by convention). Let

$$\pi_{it} = P(s_{it} = 1|\mathbf{z}_{it}, s_{i,t-1} = 1), t = 2, \ldots, T.$$

Typically, z_{it} contains elements from $(w_{i,t-1}, \ldots, w_{i1})$, and perhaps variables dated at $t - 1$ or earlier that do not appear in the population model.

• Unfortunately, $\mathbf{z}_{it}$ rarely can depend on time-varying variables that are observed in period t (since we have to apply a binary response model for the sample with $s_{i,t-1} = 1$, and this includes units that have left the sample at time t).

• Question: How do we obtain p_{it} from the π_{it}? Not without some assumptions. Let $\mathbf{v}_{it} = (\mathbf{w}_{it}, \mathbf{z}_{it}), t = 2, \ldots, T$. An ignorability assumption that works is

$$\begin{aligned} P(s_{it} = 1 | \mathbf{v}_{i1}, \ldots, \mathbf{v}_{iT}, s_{i,t-1} = 1) \\ = P(s_{it} = 1 | \mathbf{z}_{it}, s_{i,t-1} = 1), t \geq 2. \end{aligned}$$

• That is, given the entire history $\mathbf{v}_i = (\mathbf{v}_{i1}, \ldots, \mathbf{v}_{iT})$, selection at time t (given being still in the sample at $t - 1$) depends only on $\mathbf{z}_{it}$; in practice,

this means only on variables observed at $t - 1$. In particular, changes in variables from $t - 1$ to t cannot affect the conditional selection probability.

• How do we use this (strong) assumption? By the law of conditional probability,

$$P(s_{it} = 1|\mathbf{v}_i) = P(s_{it} = 1|\mathbf{v}_i, s_{i,t-1} = 1)\cdots P(s_{i2} = 1|\mathbf{v}_i, s_{i1} = 1)P(s_{i1} = 1|v_i),$$

so, under the conditional ignorability assumption,

$$p_{it} \equiv P(s_{it} = 1|\mathbf{v}_i) = \pi_{it}\pi_{i,t-1}\cdots\pi_{i2}.$$

• Method:

(1) In each time period $t \geq 2$, estimate a binary response model for $P(s_{it} = 1|\mathbf{z}_{it}, s_{i,t-1} = 1)$, which means on the group still in the sample at $t - 1$. The fitted probabilities are the $\hat{\pi}_{it}$.

(2) Form $\hat{p}_{it} = \hat{\pi}_{it}\hat{\pi}_{i,t-1}\cdots\hat{\pi}_{i2}$. We are able to compute $\hat{p}_{it}$ only for units still in the sample at time $t - 1$.

(3) Solve the problem

$$\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} N^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}(s_{it}/\hat{p}_{it})q_t(\mathbf{w}_{it},\boldsymbol{\theta}),$$

• Using the sequential approach, and assuming the sequence of binary response models is correctly specified, can show that inference ignoring estimation of the p_{it} is conservative. So, can just obtain standard errors that allow serial correlation and do not assume an information matrix equality. (That is, a standard sandwich form.)

• In Stata, linear regression and probit would be

```
   regress y x1 x2 ... xK [pweight = 1/phat],
cluster(csid)
   probit y x1 x2 ... xK [pweight = 1/phat],
cluster(csid)
```

• Method suffers from similar drawback as Heckman approach based on pooled OLS. If $P(s_{it} = 1|\mathbf{w}_{it}) = P(s_{it} = 1|\mathbf{x}_{it})$ for $\mathbf{x}_{it}$ a subset of $\mathbf{w}_{it}$, $\boldsymbol{\theta}_o$ solves

$$\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} E[q_t(\mathbf{w}_{it}, \boldsymbol{\theta})|\mathbf{x}_{it}], \text{ all } \mathbf{x}_{it}, t \geq 1,$$

then the unweighted estimator is consistent:

$$E[s_{it}q_t(\mathbf{w}_{it}, \boldsymbol{\theta})] = E\{p_{it}(\mathbf{x}_{it})E[q_t(\mathbf{w}_{it}, \boldsymbol{\theta})|\mathbf{x}_{it}]\}$$

and $p_{it}(\mathbf{x}_{it}) \geq 0, E[q_t(\mathbf{w}_{it}, \boldsymbol{\theta}_o)|\mathbf{x}_{it}] \leq E[q_t(\mathbf{w}_{it}, \boldsymbol{\theta})|\mathbf{x}_{it}]$.

• Because the probability weights cannot depend on (all of) $\mathbf{x}_{it}$ at time t, IPW could cause inconsistency.

• Related to the previous point: It would be rare that we would apply IPW in the case of a model with completely specified dynamics. Why? Suppose, for example, we have a model of $E(y_{it}|\mathbf{x}_{it}, y_{i,t-1}, \ldots, \mathbf{x}_{i1}, y_{i0})$ and let $\boldsymbol{\theta}_o$ be the parameter vector. Then $\boldsymbol{\theta}_o$ would solve

$$\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} E\{[y_{it} - m_{it}(\boldsymbol{\theta})]^2|\mathbf{x}_{it}, y_{i,t-1}, \ldots, \mathbf{x}_{i1}, y_{i0}], \ t \geq 1,$$

and by iterated expectations, the same is true if we condition on any subset of $(\mathbf{x}_{it}, y_{i,t-1}, \ldots, \mathbf{x}_{i1}, y_{i0})$. But, in practice, the elements of z_{it} must come from $(y_{i,t-1}, \mathbf{x}_{i,t-1} \ldots, \mathbf{x}_{i1}, y_{i0})$, and so IPW would be unecessary.

• Further drawbacks: Do not currently know how to apply IPW estimators in "random effects" type models, with or without the Chamberlain device, or for dynamic models with unobserved effects. The sequential nature of the modeling seems crucial.

• With a large T, the $\hat{p}_{it} = \hat{\pi}_{it}\hat{\pi}_{i,t-1}\cdots\hat{\pi}_{i1}$ can become very small, which means a lot of weight is given to units still in the sample in the late time periods. Makes the IPW estimator sensitive to the $\hat{\pi}_{it}$.