New Developments in Econometrics
Lecture 18: Missing Data

Jeff Wooldridge

Cemmap Lectures, UCL, June 2009

1. When Can Missing Data be Ignored?
2. Inverse Probability Weighting
3. Imputation
4. Heckman-Type Selection Corrections

1. **When Can Missing Data be Ignored**?

- Linear model with IVs:

$$y_i = \mathbf{x}_i\boldsymbol{\beta} + u_i, \tag{1}$$

where $\mathbf{x}_i$ is $1 \times K$, instruments $\mathbf{z}_i$ are $1 \times L$, $L \geq K$. Let s_i is the selection indicator, $s_i = 1$ if we can use observation i. With $L = K$, the "complete case" estimator is

$$\hat{\boldsymbol{\beta}}_{IV} = \left(N^{-1}\sum_{i=1}^{N} s_i\mathbf{z}_i'\mathbf{x}_i\right)^{-1}\left(N^{-1}\sum_{i=1}^{N} s_i\mathbf{z}_i'y_i\right) \tag{2}$$

$$= \beta + \left(N^{-1}\sum_{i=1}^{N} s_i\mathbf{z}_i'\mathbf{x}_i\right)^{-1}\left(N^{-1}\sum_{i=1}^{N} s_i\mathbf{z}_i'u_i\right). \tag{3}$$

- For consistency, *rank* $E(\mathbf{z}_i'\mathbf{x}_i|s_i = 1) = K$ and

$$E(s_i\mathbf{z}_i'u_i) = \mathbf{0}, \tag{4}$$

which is implied by

$$E(u_i|\mathbf{z}_i, s_i) = 0. \tag{5}$$

Sufficient for (5) is

$$E(u_i|\mathbf{z}_i) = 0,\ \ s_i = h(\mathbf{z}_i) \tag{6}$$

for some function $h(\bullet)$.

- Zero covariance assumption in the population, $E(\mathbf{z}_i' u_i) = 0$, is not sufficient for consistency when $s_i = h(\mathbf{z}_i)$.
- If $\mathbf{x}_i$ contains elements correlated with u_i, we cannot select the sample based on those endogenous elements even though we are instrumenting for them.
- Special case is when $E(y_i|\mathbf{x}_i) = \mathbf{x}_i\boldsymbol{\beta}$ and selection s_i is a function of $\mathbf{x}_i$.

- Nonlinear models/estimation methods:

  Nonlinear Least Squares: $E(y|\mathbf{x}, s) = E(y|\mathbf{x})$.

  Least Absolute Deviations: $Med(y|\mathbf{x}, s) = Med(y|\mathbf{x})$

  Maximum Likelihood: $D(\mathbf{y}|\mathbf{x}, s) = D(\mathbf{y}|\mathbf{x})$ or $D(s|\mathbf{y}, \mathbf{x}) = D(s|\mathbf{x})$.

- All of these allow selection on **x** but not generally on **y**. For estimating $\mu = E(y_i)$, unbiasedness and consistency of the sample mean computed using the selected sample requires $E(y|s) = E(y)$.

- Panel data: if we model $D(\mathbf{y}_t|\mathbf{x}_t)$, and s_t is the selection indicator, the sufficient condition to ignore selection is

$$D(s_t|\mathbf{x}_t,\mathbf{y}_t) = D(s_t|\mathbf{x}_t),\ t = 1,\ldots,T. \tag{7}$$

Let the true conditional density be $f_t(\mathbf{y}_{it}|\mathbf{x}_{it},\boldsymbol{\gamma})$. Then the partial log-likelihood function for a random draw i from the cross section can be written as

$$\sum_{t=1}^{T} s_{it} \log f_t(\mathbf{y}_{it}|\mathbf{x}_{it},\mathbf{g}) \equiv \sum_{t=1}^{T} s_{it} l_{it}(\mathbf{g}). \tag{8}$$

Can show under (7) that

$$E[s_{it}l_{it}(\mathbf{g})|\mathbf{x}_{it}] = E(s_{it}|\mathbf{x}_{it})E[l_{it}(\mathbf{g})|\mathbf{x}_{it}]. \tag{9}$$

- If $\mathbf{x}_{it}$ includes $\mathbf{y}_{i,t-1}$, (7) allows selection on $\mathbf{y}_{i,t-1}$, but not on "shocks" from $t-1$ to t.
- Similar findings for NLS, quasi-MLE, quantile regression.
- Methods to remove time-constant, unobserved heterogeneity: for a random draw i,

$$y_{it} = \eta_t + \mathbf{x}_{it}\boldsymbol{\beta} + c_i + u_{it}, \tag{10}$$

with IVs $\mathbf{z}_{it}$ for $\mathbf{x}_{it}$. Random effects IV methods (unbalanced panel):

$$E(u_{it}|\mathbf{z}_{i1},\ldots,\mathbf{z}_{iT},s_{i1},\ldots,s_{iT},c_i) = 0,\ t = 1,\ldots,T \tag{11}$$

$$E(c_i|\mathbf{z}_{i1},\ldots,\mathbf{z}_{iT},s_{i1},\ldots,s_{iT}) = E(c_i) = 0. \tag{12}$$

Selection in any time period cannot depend on u_{it} or c_i.

- FE on unbalanced panel: can get by with just (11). Let $\ddot{y}_{it} = y_{it} - T_i^{-1}\sum_{r=1}^{T} s_{ir}y_{ir}$ and similarly for and $\ddot{\mathbf{x}}_{it}$ and $\ddot{\mathbf{z}}_{it}$, where $T_i = \sum_{r=1}^{T} s_{ir}$ is the number of time periods for observation i. The FEIV estimator is

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

Weakest condition for consistency is $\sum_{t=1}^{T} E(s_{it}\ddot{\mathbf{z}}_{it}'u_{it}) = 0$.

- One important violation of (11) is when units drop out of the sample in period $t+1$ because of shocks (u_{it}) realized in time t. This generally induces correlation between $s_{i,t+1}$ and u_{it}.

- A simple variable addition test is to estimate the auxiliary model

$$y_{it} = \eta_t + \mathbf{x}_{it}\boldsymbol{\beta} + \rho s_{i,t+1} + c_i + u_{it}$$

by FE2SLS, where $s_{i,t+1}$ acts as its own instrument, and test $\rho = 0$. Lose a time period, so need $T \geq 3$ initially.

- Similar to test of strict exogeneity of covariates.

- Consistency of FE (and FEIV) on the unbalanced panel under breaks down if the slope coefficients are random and one ignores this in estimation. The error term contains the term $\mathbf{x}_i\mathbf{d}_i$ where $\mathbf{d}_i = \mathbf{b}_i - \boldsymbol{\beta}$.
- Simple test based on the alternative

$$E(\mathbf{b}_i|\mathbf{z}_{i1},\ldots,\mathbf{z}_{iT},s_{i1},\ldots,s_{iT}) = E(\mathbf{b}_i|T_i). \tag{13}$$

Add interaction terms of dummies for each possible sample size (with $T_i = T$ as the base group):

$$1[T_i = 2]\mathbf{x}_{it}, 1[T_i = 3]\mathbf{x}_{it}, ..., 1[T_i = T-1]\mathbf{x}_{it}. \tag{14}$$

Estimate equation by FE or FEIV. (In latter case, IVs are $1[T_i = r]\mathbf{z}_{it}$.)

- Can use FD in basic model, too, which is very useful for attrition problems (later). Generally, if

$$\Delta y_{it} = \varphi_t + \Delta \mathbf{x}_{it}\boldsymbol{\beta} + \Delta u_{it},\ t = 2,\ldots,T \tag{15}$$

and, if $\mathbf{z}_{it}$ is the set of IVs at time t, we can use

$$E(\Delta u_{it}|\mathbf{z}_{it}, s_{it}) = 0 \tag{16}$$

as being sufficient to ignore the missingess. Again, can add $s_{i,t+1}$ to test for attrition.

- Nonlinear models with unosberved effects are more difficult to handle. Certain conditional MLEs (logit, Poisson) can accomodate selection that is arbitrarily correlated with the unobserved effect.

2. Inverse Probability Weighting

Weighting with Cross-Sectional Data

- When selection is not on conditioning variables, can try to use probability weights to reweight the selected sample to make it representative of the population. Suppose y is a random variable whose population mean $\mu = E(y)$ we would like to estimate, but some observations are missing on y. Let $\{(y_i, s_i, \mathbf{z}_i) : i = 1, \ldots, N\}$ indicate independent, identically distributed draws from the population, where z_i is always observed (for now).

- Missingness is “ignorable” or “selection on observables” assumption:

$$P(s = 1|y, \mathbf{z}) = P(s = 1|\mathbf{z}) \equiv p(\mathbf{z}) \tag{17}$$

where $p(\mathbf{z}) > 0$ for all possible values of $\mathbf{z}$. Consider

$$\tilde{\mu}_{IPW} = N^{-1} \sum_{i=1}^{N} \left(\frac{s_i}{p(\mathbf{z}_i)} \right) y_i, \tag{18}$$

where s_i selects out the observed data points. Using (17) and iterated expectations, can show $\hat{\mu}_{IPW}$ is consistent (and unbiased) for y_i. (Same kind of estimate used for treatment effects.)

- Sometimes $p(z_i)$ is known, but mostly it needs to be estimated. Let $\hat{p}(z_i)$ denote the estimated selection probability:

$$\hat{\mu}_{IPW} = N^{-1} \sum_{i=1}^{N} \left(\frac{s_i}{\hat{p}(\mathbf{z}_i)} \right) y_i. \tag{19}$$

Can also write as

$$\hat{\mu}_{IPW} = N_1^{-1} \sum_{i=1}^{N} s_i \left(\frac{\hat{\rho}}{\hat{p}(\mathbf{z}_i)} \right) y_i \tag{20}$$

where $N_1 = \sum_{i=1}^{N} s_i$ is the number of selected observations and $\hat{\rho} = N_1/N$ is a consistent estimate of $P(s_i = 1)$.

- A different estimate is obtained by solving the least squares problem

$$\min_m \sum_{i=1}^{N} \left(\frac{s_i}{\hat{p}(\mathbf{z}_i)} \right) (y_i - m)^2.$$

- Horowitz and Manski (1998) study estimating population means using IPW. HM focus on bounds in estimating $E[g(y)|\mathbf{x} \in A]$ for conditioning variables $\mathbf{x}$. Problem with certain IPW estimators based on weights that estimate $P(s = 1)/P(s = 1|\mathbf{z})$: the resulting estimate of the mean can lie outside the natural bounds. One should use $P(s = 1|\mathbf{x} \in A)/P(s = 1|\mathbf{x} \in A, \mathbf{z})$ if possible. Unfortunately, cannot generally estimate the proper weights if x is sometimes missing.

- The HM problem is related to another issue. Suppose

$$E(y|\mathbf{x}) = \alpha + \mathbf{x}\boldsymbol{\beta}. \tag{21}$$

Let **z** be a variables that are always observed and let $p(\mathbf{z})$ be the selection probability, as before. Suppose at least part of x is not always observed, so that **x** is not a subset of **z**. Consider the IPW estimator of α, $\boldsymbol{\beta}$ solves

$$\min_{a,\mathbf{b}} \sum_{i=1}^{N} \left(\frac{s_i}{\hat{p}(\mathbf{z}_i)} \right) (y_i - a - \mathbf{x}_i \mathbf{b})^2. \tag{22}$$

- The problem is that if

$$P(s = 1|\mathbf{x}, y) = P(s = 1|\mathbf{x}), \tag{23}$$

the IPW is generally inconsistent because the condition

$$P(s = 1|\mathbf{x}, y, \mathbf{z}) = P(s = 1|\mathbf{z}) \tag{24}$$

is unlikely. On the other hand, if (23) holds, we can consistently estimate the parameters using OLS on the selected sample.

- If $\mathbf{x}$ always observed, case for weighting is much stronger because then $\mathbf{x} \subset \mathbf{z}$. If selection is on $\mathbf{x}$, this should be picked up in large samples in flexible estimation of $P(s = 1|\mathbf{z})$.

- If selection is exogenous and **x** is always observed, is there a reason to use IPW? Not if we believe $E(y|\mathbf{x}) = \alpha + \mathbf{x}\boldsymbol{\beta}$ along with the homoskedasticity assumption $Var(y|\mathbf{x}) = \sigma^2$. Then, OLS is efficient and IPW is less efficient. IPW can be more efficient with heteroskedasticity (but WLS with the correct heteroskedasticity function would be best).

- Still, one can argue for weighting under (23) as a way to consistently estimate the linear projection. Write

$$L(y|1,x) = \alpha^* + \mathbf{x}\boldsymbol{\beta}^* \tag{25}$$

where $L(\cdot|\cdot)$ denotes the linear projection. Under under $P(s = 1|\mathbf{x},y) = P(s = 1|\mathbf{x})$, the IPW estimator is consistent for $\boldsymbol{\theta}^* = (\alpha^*,\boldsymbol{\beta}^{*\prime})'$. The unweighted estimator has a probabilty limit that depends on $p(\mathbf{x})$.

- Parameters in LP show up in certain treatment effect estimators, and are the basis for the "double robustness" result of Robins and Ritov (1997) in the case of linear regression.
- The double robustness result holds for certain nonlinear models, but must choose model for $E(y|\mathbf{x})$ and the objective function appropriately; see Wooldridge (2007). (For binary or fractional response, use logistic function and Bernoulli quasi-log likelihood (QLL). For nonnegative response, use exponential function with Poisson QLL.)

- Return to the IPW regression estimator under $P(s=1|\mathbf{x},y,\mathbf{z})=P(s=1|\mathbf{z})=G(\mathbf{z},\boldsymbol{\gamma})$, with

$$E(u)=0, E(\mathbf{x}'u)=0, \tag{26}$$

for a parametric function $G(\bullet)$ (such as flexible logit), and $\hat{\gamma}$ is the binary response MLE. The asymptotic variance of $\hat{\boldsymbol{\theta}}_{IPW}$, using the estimated probability weights, is

$$Avar\sqrt{N}(\hat{\boldsymbol{\theta}}_{IPW}-\boldsymbol{\theta})=[E(\mathbf{x}_i'\mathbf{x}_i)]^{-1}E(\mathbf{r}_i\mathbf{r}_i')[E(\mathbf{x}_i'\mathbf{x}_i)]^{-1}, \tag{27}$$

where $\mathbf{r}_i$ is the $P\times 1$ vector of population residuals from the regression $(s_i/p(\mathbf{z}_i))\mathbf{x}_i'u_i$ on $\mathbf{d}_i'$, and $\mathbf{d}_i$ is the $M\times 1$ score for the MLE used to obtain $\hat{\boldsymbol{\gamma}}$.

- Variance in (27) is always smaller than the variance if we knew $p(\mathbf{z}_i)$.

Leads to a simple estimate of $Avar(\hat{\boldsymbol{\theta}}_{IPW})$:

$$\left(\sum_{i=1}^{N}(s_i/\hat{G}_i)\mathbf{x}_i'\mathbf{x}_i\right)^{-1}\left(\sum_{i=1}^{N}\hat{\mathbf{r}}_i\hat{\mathbf{r}}_i'\right)\left(\sum_{i=1}^{N}(s_i/\hat{G}_i)\mathbf{x}_i'\mathbf{x}_i\right)^{-1} \tag{28}$$

If selection is estimated by logit with regressors $\mathbf{h}_i = \mathbf{h}(\mathbf{z}_i)$,

$$\hat{\mathbf{d}}_i = \mathbf{h}_i'(s_i - \Lambda(\mathbf{h}_i\hat{\boldsymbol{\gamma}})), \tag{29}$$

where $\Lambda(a) = \exp(a)/[1 + \exp(a)]$.

- Illustrates an interesting finding of RRZ (1995): Can never do worse for estimating the parameters of interest, $\boldsymbol{\theta}$, and usually do better, when adding irrelevant functions to a logit selection model in the first stage. The Hirano, Imbens, and Ridder (2003) estimator keeps expanding $\mathbf{h}_i$.
- Adjustment in (27) carries over to general nonlinear models and estimation methods. Ignoring the estimation in $\hat{p}(\mathbf{z})$, as is standard, is asymptotically conservative. When selection is exogenous in the sense of $P(s = 1|\mathbf{x}, y, \mathbf{z}) = P(s = 1|\mathbf{x})$, the adjustment makes no difference.

- Nevo (2003) studies the case where population moments are $E[\mathbf{r}(\mathbf{w}_i, \boldsymbol{\theta})] = \mathbf{0}$ and selection depends on elements of $\mathbf{w}_i$ not always observed. Use information on population means $E[\mathbf{h}(\mathbf{w}_i)]$ such that $P(s = 1|\mathbf{w}) = P(s = 1|h(\mathbf{w}))$ and use method of moments. For a logit selection model,

$$E\left[\frac{s_i}{\Lambda(\mathbf{h}(\mathbf{w}_i)\boldsymbol{\gamma})}\mathbf{r}(\mathbf{w}_i, \boldsymbol{\theta})\right] = 0 \tag{30}$$

$$E\left[\frac{s_i\mathbf{h}(\mathbf{w}_i)}{\Lambda(\mathbf{h}(\mathbf{w}_i)\boldsymbol{\gamma})}\right] = \bar{\boldsymbol{\mu}}_h \tag{31}$$

where $\bar{\boldsymbol{\mu}}_h$ is known. Equation (31) generally identifies $\boldsymbol{\gamma}$, and $\hat{\boldsymbol{\gamma}}$ can be used in a second step to choose $\hat{\boldsymbol{\theta}}$ in a weighted GMM procedure.

Attrition in Panel Data

- Inverse probability weighting can be applied to the attrition problem in panel data. Many estimation methods can be used, but consider MLE. We have a parametric density, $f_t(y_t|\mathbf{x}_t, \boldsymbol{\theta})$, and let s_{it} be the selection indicator. Pooled MLE on on the observed data:

$$\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sum_{i=1}^{N}\sum_{t=1}^{T} s_{it} \log f_t(y_{it}|\mathbf{x}_{it}, \boldsymbol{\theta}), \tag{32}$$

which is consistent if $P(s_{it} = 1|y_{it}, \mathbf{x}_{it}) = P(s_{it} = 1|\mathbf{x}_{it})$. If not, maybe we can find variables $\mathbf{r}_{it}$, such that

$$P(s_{it} = 1|y_{it}, \mathbf{x}_{it}, \mathbf{r}_{it}) = P(s_{it} = 1|\mathbf{r}_{it}) \equiv p_{it} > 0. \tag{33}$$

- The weighted MLE is

$$\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \sum_{i=1}^{N} \sum_{t=1}^{T} (s_{it}/p_{it}) \log f_t(y_{it}|\mathbf{x}_{it}, \boldsymbol{\theta}). \tag{34}$$

Under (33), $\hat{\theta}_{IPW}$ is generally consistent because

$$E[(s_{it}/p_{it})q_t(\mathbf{w}_{it}, \boldsymbol{\theta})] = E[q_t(\mathbf{w}_{it}, \boldsymbol{\theta})] \tag{35}$$

where $q_t(\mathbf{w}_{it}, \boldsymbol{\theta}) = \log f_t(y_{it}|\mathbf{x}_{it}, \boldsymbol{\theta})$.

- How do we choose $\mathbf{r}_{it}$ to make (33) hold (if possible)? RRZ (1995) propose a sequential strategy,

$$\pi_{it} = P(s_{it} = 1|\mathbf{z}_{it}, s_{i,t-1} = 1), t = 1, \ldots, T. \tag{36}$$

Typically, $\mathbf{z}_{it}$ contains elements from $(\mathbf{w}_{i,t-1}, \ldots, \mathbf{w}_{i1})$.

- How do we obtain p_{it} from the π_{it}? Not without some strong assumptions. Let $\mathbf{v}_{it} = (\mathbf{w}_{it}, \mathbf{z}_{it}), t = 1, \ldots, T$. An ignorability assumption that works is

$$P(s_{it} = 1 | \mathbf{v}_i, s_{i,t-1} = 1) = P(s_{it} = 1 | \mathbf{z}_{it}, s_{i,t-1} = 1). \tag{37}$$

That is, given the entire history $\mathbf{v}_i = (\mathbf{v}_{i1}, \ldots, \mathbf{v}_{iT})$, selection at time t depends only on variables observed at $t - 1$. RRZ (1995) show how to relax it somewhat in a regression framework with time-constant covariates. Using (37), can show that

$$p_{it} \equiv P(s_{it} = 1 | \mathbf{v}_i) = \pi_{it}\pi_{i,t-1} \cdots \pi_{i1}. \tag{38}$$

- So, a consistent two-step method is: (i) In each time period, 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 $\hat{\pi}_{it}$. Form $\hat{p}_{it} = \hat{\pi}_{it}\hat{\pi}_{i,t-1} \cdot\cdot\cdot \hat{\pi}_{i1}$. (ii) Replace p_{it} with $\hat{p}_{it}$ in (34), and obtain the weighted pooled MLE.
- As shown by RRZ (1995) in the regression case, it is more efficient to estimate the p_{it} than to use know weights, if we could. See RRZ (1995) and Wooldridge (2002) for a simple regression method for adjusting the score.

- IPW for attrition suffers from a similar drawback as in the cross section case. Namely, if $P(s_{it} = 1|\mathbf{w}_{it}) = P(s_{it} = 1|\mathbf{x}_{it})$ then the unweighted estimator is consistent. If we use weights that are not a function of $\mathbf{x}_{it}$ in this case, the IPW estimator is generally inconsistent.
- Related to the previous point: would rarely apply IPW in the case of a model with completely specified dynamics. Why? If we have a model for $D(y_{it}|\mathbf{x}_{it}, y_{i,t-1},\ldots,\mathbf{x}_{i1}, y_{i0})$ or $E(y_{it}|\mathbf{x}_{it}, y_{i,t-1},\ldots,\mathbf{x}_{i1}, y_{i0})$, then our variables affecting attrition, $\mathbf{z}_{it}$, are likely to be functions of $(y_{i,t-1},\mathbf{x}_{i,t-1},\ldots,\mathbf{x}_{i1}, y_{i0})$. If they are, the unweighted estimator is consistent. For misspecified models, we might still want to weight.

3. **Imputation**

- So far, we have discussed when we can just drop missing observations (Section 1) or when the complete cases can be used in a weighting method (Section 2). A different approach to missing data is to try to fill in the missing values, and then analyze the resulting data set as a complete data set. Little and Rubin (2002) provide an accessible treatment to *imputation* and *multiple imputation* methods, with lots of references to work by Rubin and coauthors.

- Imputing missing values is not always valid. Most methods depend on a *missing at random* (MAR) assumption. When data are missing on the response variable, y, MAR is essentially the same as $P(s = 1|y, \mathbf{x}) = P(s = 1|\mathbf{x})$. *Missing completely at random* (MCAR) is when s is independent of $\mathbf{w} = (\mathbf{x}, y)$.

- MAR for general missing data patterns. Let $\mathbf{w}_i = (\mathbf{w}_{i1}, \mathbf{w}_{i2})$ be a random draw from the population. Let $r_i = (r_{i1}, r_{i2})$ be the "retention" indicators for $\mathbf{w}_{i1}$ and $\mathbf{w}_{i2}$, so $r_{ig} = 1$ implies $\mathbf{w}_{ig}$ is observed. MCAR is that $\mathbf{r}_i$ is independent of $\mathbf{w}_i$. The MAR assumption is that $P(r_{i1} = 0, r_{i2} = 0|\mathbf{w}_i) = P(r_{i1} = 0, r_{i2} = 0) \equiv \pi_{00}$ and so on.

- MAR is more natural with monotone missing data problems; we just saw the case of attrition. If we order the variables so that if $\mathbf{w}_{ih}$ is observed the so is $\mathbf{w}_{ig}$, $g < h$. Write

$f(\mathbf{w}_1,\ldots,\mathbf{w}_G) = f(\mathbf{w}_G|\mathbf{w}_{G-1},\ldots,\mathbf{w}_1)$

$\cdot f(\mathbf{w}_{G-1}|\mathbf{w}_{G-1},\ldots,\mathbf{w}_1)\cdots f(\mathbf{w}_2|\mathbf{w}_1)f(\mathbf{w}_1)$. Partial log likelihood:

$$\sum_{g=1}^{G} r_{ig} \log f(\mathbf{w}_{ig}|\mathbf{w}_{i,g-1},\ldots,\mathbf{w}_{i1},\boldsymbol{\theta}), \qquad (39)$$

where we use $r_{ig} = r_{ig}r_{i,g-1}\cdots r_{i2}$. Under MAR,

$$E(r_{ig}|\mathbf{w}_{ig},\ldots,\mathbf{w}_{i1}) = E(r_{ig}|\mathbf{w}_{i,g-1},\ldots,\mathbf{w}_{i1}). \qquad (40)$$

(39) is the basis for filling in data in monotonic MAR schemes.

- Simple example of imputation. Let $\mu_y = E(y)$, but data are missing on some y_i. Unless $P(s_i = 1|y_i) = P(s_i = 1)$, the complete-case average is not consistent for μ_y. Suppose that the selection is ignorable conditional on $\mathbf{x}$:

$$E(y|\mathbf{x}, s) = E(y|\mathbf{x}) = m(\mathbf{x}, \boldsymbol{\beta}). \tag{41}$$

NLS using selected sample is consistent for $\boldsymbol{\beta}$. Obtain a fitted value, $m(\mathbf{x}_i, \hat{\boldsymbol{\beta}})$, for any unit it the sample. Let $\hat{y}_i = s_i y_i + (1 - s_i) m(\mathbf{x}_i, \hat{\boldsymbol{\beta}})$ be the imputed data. Imputation estimator:

$$\hat{\mu}_y = N^{-1} \sum_{i=1}^{N} \{s_i y_i + (1 - s_i) m(\mathbf{x}_i, \hat{\boldsymbol{\beta}})\}. \tag{42}$$

- From $plim(\hat{\mu}_y) = E[s_i y_i + (1 - s_i) m(\mathbf{x}_i, \boldsymbol{\beta})]$ we can show consistency of $\hat{\mu}_y$ because under (41),

$$E[s_i y_i + (1 - s_i) m(\mathbf{x}_i, \boldsymbol{\beta})] = E[m(\mathbf{x}_i, \boldsymbol{\beta})] = \mu_y. \tag{43}$$

- Danger in using imputation methods: we might be tempted to treat the imputed data as real random draws. Generally leads to incorrect inference because of inconsistent variance estimation. (In linear regression, easy to see that estimated variance is too small.)
- Little and Rubin (2002) call (43) the method of “conditional means.” In Table Table 4.1 they document the downward bias in variance estimates.

- LR propose adding a random draw to $m(\mathbf{x}_i, \hat{\boldsymbol{\beta}})$ – assuming that we can estimate $D(y|\mathbf{x})$. If we assume $D(u_i|\mathbf{x}_i) = Normal(0, \sigma_u^2)$, draw $\check{u}_i$ from a Normal$(0, \hat{\sigma}_u^2)$, distribution, where $\hat{\sigma}_u^2$ is estimated using the complete case nonlinear regression residuals, and then use $m(\mathbf{x}_i, \hat{\boldsymbol{\beta}}) + \check{u}_i$ for the missing data. Called the "conditional draw" method of imputation (special case of stochastic imputation).
- Generally difficult to quantity the uncertainty from single-imputation methods, where a single imputed values is obtained for each missing variable. Can bootstrap the entire estimation/imputation steps, but this is computationally intensive.

- Multiple imputation is an alternative. Its theoretical justification is Bayesian, based on obtaining the posterior distribution – in particular, mean and variance – of the parameters conditional on the observed data. For general missing data patterns, the computation required to impute missing values is intensive, and involves simulation methods of estimation. See also Cameron and Trivedi (2005).
- General idea: rather than just impute one set of missing values to create one “complete” data set, create several imputed data sets. (Often the number is fairly small, such as five or so.) Estimate the parameters of interest using each imputed data set, and average to obtain a final parameter estimate and sampling error.

- Let $\mathbf{W}_{mis}$ denote the matrix of missing data and $\mathbf{W}_{obs}$ the matrix of observations. Assume that MAR holds. MAR used to estimate $E(\boldsymbol{\theta}|\mathbf{W}_{obs})$, the posterier mean of $\boldsymbol{\theta}$ given $\mathbf{W}_{obs}$. But by iterated expectations,

$$E(\boldsymbol{\theta}|\mathbf{W}_{obs}) = E[E(\boldsymbol{\theta}|\mathbf{W}_{obs}, \mathbf{W}_{mis})|\mathbf{W}_{obs}]. \tag{44}$$

If $\hat{\boldsymbol{\theta}}_d = E(\boldsymbol{\theta}|\mathbf{W}_{obs}, \mathbf{W}_{mis}^{(d)})$ for imputed data set d, then approximate $E(\boldsymbol{\theta}|\mathbf{W}_{obs})$ as

$$\bar{\boldsymbol{\theta}} = D^{-1} \sum_{d=1}^{D} \hat{\boldsymbol{\theta}}_d. \tag{45}$$

- Further, we can obtain a “sampling” variance by estimating $Var(\theta|\mathbf{W}_{obs})$ using

$$\begin{aligned} Var(\theta|\mathbf{W}_{obs}) &= E[Var(\boldsymbol{\theta}|\mathbf{W}_{obs},\mathbf{W}_{mis})|\mathbf{W}_{obs}] \\ &\quad + Var[E(\boldsymbol{\theta}|\mathbf{W}_{obs},\mathbf{W}_{mis})|\mathbf{W}_{obs}], \end{aligned} \tag{46}$$

which suggests

$$\begin{aligned} \widehat{Var}(\theta|\mathbf{W}_{obs}) &= D^{-1}\sum_{d=1}^{D}\hat{\mathbf{V}}_d \\ &\quad + (D-1)^{-1}\sum_{d=1}^{D}(\hat{\boldsymbol{\theta}}_d-\bar{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}_d-\bar{\boldsymbol{\theta}})' \\ &\equiv \bar{\mathbf{V}}+\mathbf{B}. \end{aligned} \tag{47}$$

- For small number of imputations, a correction is usually made, namely, $\bar{\mathbf{V}} + (1 + D)^{-1}\mathbf{B}$. assuming that one trusts the MAR assumption and the underlying distributions used to draw the imputed values, inference with multiple imputations is fairly straightforward. D need not be very large so estimation using nonlinear models is relatively easy, given the imputed data.
- Use caution when applying to models with missing conditioning variables. Suppose $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)$, we are interested in $D(y|\mathbf{x})$, data are missing on y and $\mathbf{x}_2$, and selection is a function of $\mathbf{x}_2$. Using the complete cases will be consistent. Imputation methods would not be, as they require $D(s|y, \mathbf{x}_1, \mathbf{x}_2) = D(s|\mathbf{x}_1)$.

4. Heckman-Type Selection Corrections

- With random slopes in the population, get a new twist on the usual Heckman procedure.

$$y_1 = a_1 + \mathbf{x}_1\mathbf{b}_1 \equiv \alpha_1 + \mathbf{x}_1\boldsymbol{\beta}_1 + u_1 + \mathbf{x}_1\mathbf{e}_1$$

where $u_1 = a_1 - \alpha_1$ and $\mathbf{e}_1 = \mathbf{b}_1 - \boldsymbol{\beta}_1$. Let $\mathbf{x}$ be the full set of exogenous explanatory variables with $\mathbf{x}_1$ a strict subset of $\mathbf{x}$.

- Assume selection follows a standard probit:

$$\begin{aligned} y_2 &= [\eta_2 + \mathbf{x}\boldsymbol{\delta}_2 + v_2 > 0] \\ D(v_2|\mathbf{x}) &= Normal(0,1) \end{aligned}$$

- Also, $(u_1, \mathbf{e}_1, v_2)$ independent of $\mathbf{x}$ with $E(u_1, \mathbf{e}_1 | v_2)$ linear in v_2. Then

$$E(y_1|\mathbf{x}, v_2) = \alpha_1 + \mathbf{x}_1\boldsymbol{\beta}_1 + \rho_1 v_2 + \mathbf{x}_1 v_2 \boldsymbol{\psi}_1$$

and so

$$E(y_1|\mathbf{x}, y_2 = 1) = \alpha_1 + \mathbf{x}_1\boldsymbol{\beta}_1 + \rho_1\lambda(\eta_2 + \mathbf{x}\boldsymbol{\delta}_2) + \lambda(\eta_2 + \mathbf{x}\boldsymbol{\delta}_2) \cdot \mathbf{x}_1\boldsymbol{\psi}_1$$

- The twist, compared with the usual Heckman procedure, is to add the interactions $\hat{\lambda}_{i2} \cdot \mathbf{x}_{i1}$, where $\hat{\lambda}_{i2} = \lambda(\hat{\eta}_2 + \mathbf{x}_i\hat{\boldsymbol{\delta}}_2)$ are the estimated inverse Mills ratios from the probit, to the usual two-step method:

$$y_{i1} \text{ on } 1,\ \mathbf{x}_{i1},\ \hat{\lambda}_{i2},\ \hat{\lambda}_{i2} \cdot \mathbf{x}_{i1} \ \text{ using } y_{i2} = 1$$

- Bootstrapping is convenient for inference. Full MLE, where $(u_1, \mathbf{e}_1, v_2)$ is multivariate normal, would be substantially more difficult.
- Can test joint significance of $(\hat{\lambda}_{i2}, \hat{\lambda}_{i2} \cdot \mathbf{x}_{i1})$ to test null of no selection bias – no need to adjust for first-stage estimation.
- Be careful with functional form. Interactions might be significant because population model is not a true conditional mean.

- Back to constant slopes but endogenosu explanatory variable. Advantages of applying IV methods when data are missing on explanatory variables in addition to the response variable. Briefly, a variable that is exogenous in the population model need not be in the selected subpopulation. (Example: wage-benefits tradeoff.)

$$
\begin{aligned}
y_1 &= \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + u_1 && (48)\\
y_2 &= \mathbf{z}_2\boldsymbol{\delta}_2 + v_2 && (49)\\
y_3 &= 1[\mathbf{z}\boldsymbol{\delta}_3 + v_3 > 0]. && (50)
\end{aligned}
$$

- Assume (a) $(\mathbf{z}, y_3)$ is always observed, (y_1, y_2) observed when $y_3 = 1$; (b) $E(u_1|\mathbf{z}, v_3) = \gamma_1 v_3$; (c) $v_3|\mathbf{z} \sim Normal(0,1)$; (d) $E(\mathbf{z}_2' v_2) = \mathbf{0}$ and $\boldsymbol{\delta}_{22} \neq \mathbf{0}$ where $\mathbf{z}_2\boldsymbol{\delta}_2 = \mathbf{z}_1\boldsymbol{\delta}_{21} + \mathbf{z}_{21}\boldsymbol{\delta}_{22}$.
- Then we can write

$$y_1 = \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + g(\mathbf{z}, y_3) + e_1 \tag{51}$$

where $e_1 = u_1 - g(\mathbf{z}, y_3) = u_1 - E(u_1|\mathbf{z}, y_3)$. Selection is exogenous in (51) because $E(e_1|\mathbf{z}, y_3) = 0$. Because y_2 is not exogenous, we estimate (51) by IV, using the selected sample, with IVs $[\mathbf{z}_2, \lambda(\mathbf{z}\boldsymbol{\delta}_3)]$ because $g(\mathbf{z}, 1) = \lambda(\mathbf{z}\boldsymbol{\delta}_3)$.

- The two-step estimator is (i) Probit of y_3 on $\mathbf{z}$ to (using all observations) to get $\hat{\lambda}_{i3} \equiv \lambda(\mathbf{z}_i\hat{\boldsymbol{\delta}}_3)$; (ii) IV (2SLS if overidentifying restrictions) of y_{i1} on $\mathbf{z}_{i1}, y_{i2}, \hat{\lambda}_{i3}$ using IVs $(\mathbf{z}_{i2}, \hat{\lambda}_{i3})$.
- If y_2 is always observed, tempting to obtain the fitted values $\hat{y}_{i2}$ from the reduced form y_{i2} on $\mathbf{z}_{i2}$, and then use OLS of y_{i1} on $\mathbf{z}_{i1}, \hat{y}_{i2}, \hat{\lambda}_{i3}$ in the second stage. But this effectively puts $\alpha_1 v_2$ in the error term, so we would need $u_1 + \alpha_2 v_2$ to be normally (or something similar). Rules out discrete y_2. The procedure just outlined uses the linear projection $y_2 = \mathbf{z}_2\boldsymbol{\pi}_2 + \eta_2\lambda(\mathbf{z}\boldsymbol{\delta}_3) + r_3$ in the selected population, and does not care whether this is a conditional expectation.

- In theory, can set $\mathbf{z}_2 = \mathbf{z}$, although that usually means lots of collinearity in the (implicit) reduced form for y_2 in the selected sample.
- Choosing $\mathbf{z}_1$ a strict $\mathbf{z}_2$ and $\mathbf{z}_2$ a strict ssubset of $\mathbf{z}$ enforces discipline. Namely, we should have an exogenous variable that would be valid as an IV for y_2 in the absense of sample selection, and at least one more variable (in $\mathbf{z}$) that mainly affects sample selection.

- If an explanatory variable is not always observed, ideally can find an IV for it and treat it as endogenous even if it is exogenous in the population. Generally, the usual Heckman approach (like IPW and imputation) is hard to justify in the model $E(y|\mathbf{x}) = E(y|\mathbf{x}_1)$ if $\mathbf{x}_1$ is not always observed. The first-step would be estimation of $P(s = 1|\mathbf{x}_2)$ where $\mathbf{x}_2$ is always observed. But then we would be assuming $P(s = 1|\mathbf{x}) = P(s = 1|\mathbf{x}_2)$, effectively an exclusion restriction on a reduced form.
- Lecture notes discuss linear unobserved effects models with endogenous explanatory variables and attrition: modify cross-sectional IV procedure.