What's New in Econometrics?
Lecture 12
Missing Data

Jeff Wooldridge
NBER Summer Institute, 2007

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:

$$y_i = x_i\beta + u_i, \tag{1}$$

where x_i is $1 \times K$, instruments 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

$$\hat{\beta}_{IV} = \left(N^{-1} \sum_{i=1}^{N} s_i z_i' x_i \right)^{-1} \left(N^{-1} \sum_{i=1}^{N} s_i z_i' y_i \right) \tag{2}$$

$$= \beta + \left(N^{-1} \sum_{i=1}^{N} s_i z_i' x_i \right)^{-1} \left(N^{-1} \sum_{i=1}^{N} s_i z_i' u_i \right) \tag{3}$$

For consistency, $rank\, E(z_i' x_i | s_i = 1) = K$ and

$$E(s_i z_i' u_i) = 0, \tag{4}$$

which is implied by

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

Sometimes we can add a function of z_i to the equation that forces (5) to be true. Sufficient for (5) is

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

for some function $h(\cdot)$.

- Zero covariance assumption in the population, $E(z_i' u_i) = 0$, is not sufficient for consistency when $s_i = h(z_i)$. Special case is when

$E(y_i|x_i) = x_i\beta$ and selection s_i is a function of x_i.

- Nonlinear models/estimation methods:

Nonlinear Least Squares: $E(y|x,s) = E(y|x)$.

Least Absolute Deviations: $Med(y|x,s) = Med(y|x)$

Maximum Likelihood: $D(y|x,s) = D(y|x)$ or $D(s|y,x) = D(s|x)$.

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

- Panel data: if we model $D(y_t|x_t)$, and s_t is the selection indicator, the sufficient condition to ignore selection is

$$D(s_t|x_t,y_t) = D(s_t|x_t),\ t = 1,\ldots,T. \tag{7}$$

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

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

Can show under (7) that

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

By the Kullback-Leibler information inequality,

$E[l_{it}(\gamma)|x_{it}] \geq E[l_{it}(g)|x_{it}]$ for all $g \in \Gamma$ (parameter space). Because

$E(s_{it}|x_{it}) = P(s_{it} = 1|x_{it}) \geq 0$, it follows that

$E[s_{it} l_{it}(\gamma)|x_{it}] \geq E[s_{it} l_{it}(g)|x_{it}]$ for all $g \in \Gamma$. Apply LIE again to

conclude γ maximizes the expected value of (8). We cannot just initially appeal to genearl MLE results; (8) is not a proper log-likelihood function.

- If x_{it} includes, say, $y_{i,t-1}$, then (7) allows selection to depend on $y_{i,t-1}$, but not on "shocks" from $t-1$ to t.
- Similar findings for nonlinear least squares, quasi-MLE, quantile regression.
- Methods to remove time-constant, unobserved heterogeneity: suppose we have the linear model, written for a random draw i,

$$y_{it} = \eta_t + x_{it}\beta + c_i + u_{it}, \tag{10}$$

with instruments z_{it} for x_{it}. Random effects IV methods on the

unbalanced panel use

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

and

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

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

- FE on unbalanced panel means we can get by with just the first assumption. Let $\ddot{y}_{it} = y_{it} - T_i^{-1}\sum_{r=1}^{T} s_{ir}y_{ir}$ and similarly for and $\ddot{x}_{it}$ and $\ddot{z}_{it}$, where $T_i = \sum_{r=1}^{T} s_{ir}$ is the number of time periods for observation i.

The FEIV estimator is

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

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

- 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}. To test, just add $s_{i,t+1}$ to the equation at time t, estimate the model by fixed effects (or FEIV), and compute (robust) t test.
- Consistency of FE (and FEIV) on the unbalanced panel under (11) breaks down if the slope coefficients are random and one ignores this in estimation. (Earlier: FE and FEIV still can produce consistent

estimators in balanced case.) The error term contains the term $x_i d_i$ where $d_i = b_i - \beta$. Simple test based on the alternative

$$\mathrm{E}(b_i|z_{i1}, \ldots, z_{iT}, s_{i1}, \ldots, s_{iT}) = E(b_i|T_i). \tag{13}$$

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

$$1[T_i = 2]x_{it},\ 1[T_i = 3]x_{it},\ \ldots,\ 1[T_i = T-1]x_{it}. \tag{14}$$

Estimate equation by FE or FEIV.

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

$$\Delta y_{it} = \varphi_t + \Delta x_{it} + \Delta u_{it},\ t = 2, \ldots, T \tag{15}$$

and, if z_{it} is the set of IVs at time t, we can use

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

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

- Not suprisingly, nonlinear models with unosberved effects are considerably more difficult to handle, although 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 populatin. 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 $\{(y_i, s_i, z_i) : i = 1, \ldots, N\}$ indicate independent, identically distributed draws from the population, where z_i is always observed (for now). "Selection on observables" assumption

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

where $p(z) > 0$ for all possible values of z. Consider

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

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 (variable probability stratified sampling), but mostly it needs to be estimated. (And, even for VP sampling, it *should* be estimated if possible.) Let $\hat{p}(z_i)$ denote the estimated selection probability:

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

Can also write as

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

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 $P(s_i = 1)$. The weights reported to account for missing data are often $\hat{\rho}/\hat{p}(z_i)$.

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

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

- Horowitz and Manski (1998) have considered the problem of estimating population means using IPW. They focus on bounds in estimating $E[g(y)|x \in A]$ for conditioning variables, x. But they also note a problem with certain IPW estimators based on weights that estimate $P(s = 1)/P(s = 1|d = 1, z)$: the resulting estimate of the mean

can lie outside the natural bounds (when $g(y)$ is bounded). One should use $P(s = 1|x \in A)/P(s = 1|x \in A, z)$ if possible (which are not the included sampling weights). Unfortunately, cannot generally estimate the proper weights if x is sometimes missing.

- The HM problem is related to another issue. Suppose

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

Let z be a variables that are always observed and let $p(z)$ 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, β solves

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

The problem is that if

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

the IPW is generally inconsistent because the condition

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

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

- If x is always observed, case for weighting is much stronger because then $x \subset z$. If selection is on x, this should be picked up in large

samples in the estimation of $P(s = 1|z)$.

- If (23) holds and x is always observed, is there a reason to use IPW?

Not if we believe (21) along with the homoskedasticity assumption $Var(y|x) = \sigma^2$. 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

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

where $L(\cdot|\cdot)$ denotes the linear projection. Under under $P(s = 1|x, y) = P(s = 1|x)$, the IPW estimator is consistent for θ^*. The

unweighted estimator has a probabilty limit that depends on $p(x)$.

- 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 $E(y|x)$ 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 $P(s = 1|y, z) = P(s = 1|z) = G(z, \gamma)$, with

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

for a parametric function $G(\cdot)$ (such as flexible logit), and $\hat{\gamma}$ is the binary response MLE. As shown by Robins, Rotnitzky, and Zhou (1995) and Wooldridge (2007), the asymptotic variance of $\hat{\theta}_{IPW}$, using the estimated probability weights, is

$$Avar\sqrt{N}(\hat{\theta}_{IPW} - \theta) = [E(x_i'x_i)]^{-1}E(r_i r_i')[E(x_i'x_i)]^{-1}, \tag{27}$$

where r_i is the $P \times 1$ vector of population residuals from the regression $(s_i/p(z_i))x_i'u_i$ on d_i', where d_i is the $M \times 1$ score for the MLE used to obtain $\hat{\gamma}$. This is always smaller than the variance if we knew $p(z_i)$. Leads to a simple estimate of $Avar(\hat{\theta}_{IPW})$:

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

If selection is estimated by logit with regressors $h_i = h(z_i)$,

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

where $\Lambda(a) = \exp(a)/[1 + \exp(a)]$ and $h_i = h(z_i)$.

- Illustrates an interesting finding of RRZ (1995), related to the Hirano, Imbens, and Ritter (2003) efficient estimator for means using IPW estimators. Suppose for functions $h_{i1} = h_1(z_i)$, the logit model is correctly specified: $P(s_i = 1|z_i) = \Lambda(h_{i1}\gamma_1)$. Now take additional functions, $h_{i2} = h_2(z_i)$, and add them to the logit. Asymptotically, the

coefficients on h_{i2} are zero, so the adjustment to variance of $\hat{\theta}_{IPW}$ comes from regressing $[s_i/\Lambda(h_{i1}\gamma_1)] \cdot x_i'u_i$ on $[s_i - \Lambda(h_{i1}\gamma_1)] \cdot (h_{i1}, h_{i2})$. This reduces the residual variance relative to just using h_{i1}, so $\hat{\theta}_{IPW}$ using (h_{i1}, h_{i2}) generaly more efficient than using the "correct" functions, h_{i1}. HIR estimator keeps expanding h_i.

- Wooldridge (2007): adjustment in (27) carries over to general nonlinear models and estimation methods. Ignoring the estimation in $\hat{p}(z)$, as is standard, is asymptotically conservative. When selection is exogenous in the sense of $P(s = 1|x, y, z) = P(s = 1|x)$, the adjustment makes no difference.
- As a particular example, consider VP sampling. It one uses the

known sampling probabilities (probability of keeping an observation that falls into a given stratum), this is less efficient than using the frequencies estimated from the data. (These require knowing how many times each stratum was sampled.) When the latter are used, the adjustment is to subtract off within-stratum means in computing the sampling variation in the score:

$$\widehat{Avar(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})} = \left[\sum_{i=1}^{M} \mathbf{x}_i' \mathbf{x}_i / \hat{p}_{g_i} \right]^{-1}$$

$$\cdot \left\{ \sum_{g=1}^{G} \hat{p}_g^{-2} \left[\sum_{i=1}^{M_g} (\mathbf{x}_{gi}' \hat{u}_{gi} - \overline{\mathbf{x}_g' \hat{u}_g})(\mathbf{x}_{gi}' \hat{u}_{gi} - \overline{\mathbf{x}_g' \hat{u}_g})' \right] \right\} \tag{30}$$

$$\cdot \left[\sum_{i=1}^{M} \mathbf{x}_i' \mathbf{x}_i / \hat{p}_{g_i} \right]^{-1}$$

absorbing the intercept into $\mathbf{x}_i$. If we drop $\overline{\mathbf{x}_g' \hat{u}_g}$ from the middle, we get the usual sandwich estimator for weighted least squares, which is larger than (30). Generally, the adjustment in (30) is the sourced of variance reduction using knowledge of stratum membership (with and without clustered data, too).

- Nevo studies the case where the population moments are $E[r(w_i, \theta)] = 0$ and the selection probability depends on elements of

w_i that are not always observed, and uses information on population means $E[h(w_i)]$ such that $P(s = 1|w) = P(s = 1|h(w))$ to obtain an expanded set of moment conditions for GMM estimation. So, if we use a logit model for selection,

$$E\left[\frac{s_i}{\Lambda(h(w_i)\gamma)} r(w_i, \theta) \right] = 0 \tag{31}$$

and

$$E\left[\frac{s_i h(w_i)}{\Lambda(h(w_i)\gamma)} \right] = \mu_h \tag{32}$$

where μ_h is known. Equation (32) generally identifies γ, and then this $\hat{\gamma}$ can be used in a second step to choose $\hat{\theta}$ in a weighted GMM

procedure.

- IPW can be used when data are missing due to a censored duration, t_i, where c_i is the censoring time. The needed probabilities turn out to be $G(t_i)$ where $G(t) \equiv \mathrm{P}(c_i \geq t)$ is the survivor function for the censoring values. This can be estimated using Kaplan-Meier estimator with roles of c_i and t_i are reversed. See Rotnitzky and Robins (2005) for a survey of how to obtain semiparametrically efficient estimators in linear regression. Holds for lots of nonlinear models, too.

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, $f_t(y_t|\mathbf{x}_t,\theta)$, and let s_{it} be the selection indicator. We already discussed just using pooled OLS on on the observed data:

$$\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\log f_t(y_{it}|\mathbf{x}_{it},\boldsymbol{\theta}), \tag{33}$$

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

$$P(s_{it}=1|\mathbf{w}_{it},\mathbf{r}_{it}) = P(s_{it}=1|\mathbf{r}_{it}) \equiv p_{it} > 0 \tag{34}$$

where $\mathbf{w}_{it} = (\mathbf{x}_{it},y_{it})$. The weighted MLE is

$$\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sum_{i=1}^{N}\sum_{t=1}^{T}(s_{it}/p_{it})\log f_t(y_{it}|\mathbf{x}_{it},\boldsymbol{\theta}). \tag{35}$$

Under (34), $\hat{\theta}_{IPW}$ is generally consistent because

$$E[(s_{it}/p_{it})q_t(\mathbf{w}_{it},\boldsymbol{\theta})] = E[q_t(\mathbf{w}_{it},\boldsymbol{\theta})] \tag{36}$$

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 (34) hold (if possible)? RRZ (1995) propose a sequential strategy,

$$\pi_{it} = P(s_{it} = 1|z_{it}, s_{i,t-1} = 1), t = 1,\ldots,T. \tag{37}$$

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

$$P(s_{it} = 1|\mathbf{v}_i, s_{i,t-1} = 1) = P(s_{it} = 1|\mathbf{z}_{it}, s_{i,t-1} = 1). \tag{38}$$

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 this assumption, we can show that

$$p_{it} \equiv P(s_{it} = 1|\mathbf{v}_i) = \pi_{it}\pi_{i,t-1} \cdot \cdot \cdot \pi_{i1}. \tag{39}$$

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 (35), 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 $D(y_{it}|x_{it}, y_{i,t-1}, \ldots, x_{i1}, y_{i0})$ or $E(y_{it}|x_{it}, y_{i,t-1}, \ldots, x_{i1}, y_{i0})$, then our variables affecting attrition, z_{it}, are likely to be functions of $(y_{i,t-1}, \ldots, x_{i1}, y_{i0})$. 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 cannot always be valid, of course. Most methods depend on a *missing at random* (MAR) assumption. When data are missing on only one variable – say, the response variable, y – MAR is essentially the same as $P(s = 1|y, x) = P(s = 1|x)$. The assumption *missing completely at random* (MCAR) is when s is independent of $w = (x, y)$.
- MAR can be defined for general missing data patterns. Let $w_i = (w_{i1}, w_{i2})$ be a random draw from the population, where data can be missing on either variable. Let $r_i = (r_{i1}, r_{i2})$ be the "retention"

indicators for w_{i1} and w_{i2}, so $r_{ig} = 1$ implies w_{ig} is observed. The MCAR assumption is that r_i is independent of w_i. The MAR assumption is that $P(r_{i1} = 0, r_{i2} = 0|w_i) = P(r_{i1} = 0, r_{i2} = 0) \equiv \pi_{00}$, $P(r_{i1} = 1, r_{i2} = 0|w_{i1})$, and $P(r_{i1} = 0, r_{i2} = 1|w_{i2})$. Even with just two variables, the restrictions imposed by MAR are not especially appealing, unless, of course, we have good reason to just assume MCAR.

- MAR is more natural with monotone missing data problems; we just saw the case of attrition. If we order the variables so that if w_{ih} is observed the so is w_{ig}, $g < h$. Write $f(w_1, \ldots, w_G) = f(w_G|w_{G-1}, \ldots, w_1) \cdot f(w_{G-1}|w_{G-1}, \ldots, w_1) \cdots f(w_2|w_1)f(w_1)$. Given parametric models, we

can write partial log likelihood as

$$\sum_{g=1}^{G} r_{ig} \log f(w_{ig}|w_{i,g-1},\ldots,w_{i1},\theta), \tag{40}$$

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

$$E(r_{ig}|w_{ig},\ldots,w_{i1}) = E(r_{ig}|w_{i,g-1},\ldots,w_{i1}). \tag{41}$$

As we showed in the attrition case, partial MLE based on (40) is consistent and $\sqrt{N}$-asymptotically normal in general. This 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 $\mathbf{x}$:

$$E(y|\mathbf{x}, s) = E(y|\mathbf{x}) = m(\mathbf{x}, \boldsymbol{\beta}), \tag{42}$$

where $m(\mathbf{x}, \boldsymbol{\beta})$ is a parametric function. From Section 1, NLS using the selected sample is consistent for $\boldsymbol{\beta}$. Because we observe $\mathbf{x}_i$ for all i, we can obtained fitted values, $m(\mathbf{x}_i, \hat{\boldsymbol{\beta}})$, 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. Then an imputation estimator of μ_y is

$$\hat{\mu}_y = N^{-1} \sum_{i=1}^{N} \{s_i y_i + (1 - s_i) m(\mathbf{x}_i, \hat{\boldsymbol{\beta}})\}. \tag{43}$$

From $\text{plim}(\hat{\mu}_y) = E[s_i y_i + (1 - s_i) m(\mathbf{x}_i, \boldsymbol{\beta})]$ we can show consistency of $\hat{\mu}_y$ because, by (42) and iterated expectations,

$$E[s_i y_i + (1 - s_i) m(\mathbf{x}_i, \boldsymbol{\beta})] \;\; = E[m(\mathbf{x}_i, \boldsymbol{\beta})] \;\; = \mu_y. \tag{44}$$

- 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 their Table Table 4.1 the document the downward bias in variance estiimates.

- Instead, LR propose adding a random draw to $m(\mathbf{x}_i, \hat{\boldsymbol{\beta}})$ to impute a value – assuming, of course, that we can estimate $D(y|\mathbf{x})$. If we assume

that $D(u_i|\mathbf{x}_i) = Normal(0, \sigma_u^2)$, 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, which is a special case of stochastic imputation.

- Generally difficult to quantity the uncertainty from single-imputation methods, where one imputed values is obtained for each missing variable. One possibility is to bootstrap the entire estimation/imputation steps. Can be computationally intensive because imputation needs to be done for each bootstrap sample.
- 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 quite complicated, and involves simulation methods of estimation. LR and Cameron and Trivedi (2005) provide discussion.

- 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 then use an averaging 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,

$$E(\boldsymbol{\theta}|\mathbf{W}_{obs}) = E[E(\boldsymbol{\theta}|\mathbf{W}_{obs}, \mathbf{W}_{mis})|\mathbf{W}_{obs}]. \tag{45}$$

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

$$\bar{\boldsymbol{\theta}} = D^{-1} \sum_{d=1}^{D} \hat{\boldsymbol{\theta}}_d, \tag{46}$$

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

$$Var(\theta|\mathbf{W}_{obs}) = E[Var(\boldsymbol{\theta}|\mathbf{W}_{obs}, \mathbf{W}_{mis})|\mathbf{W}_{obs}] + Var[E(\boldsymbol{\theta}|\mathbf{W}_{obs}, \mathbf{W}_{mis})|\mathbf{W}_{obs}], \tag{47}$$

which suggests

$$\begin{aligned}\widehat{Var}(\theta|\mathbf{W}_{obs}) &= D^{-1}\sum_{d=1}^{D}\hat{\mathbf{V}}_d \\ &\quad + (D-1)^{-1}\sum_{d=1}^{D}(\hat{\boldsymbol{\theta}}_d - \bar{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}_d - \bar{\boldsymbol{\theta}})' \\ &\equiv \bar{\mathbf{V}} + \mathbf{B},\end{aligned} \tag{48}$$

where $\bar{\mathbf{V}}$ is the average of the variance estimates across imputed samples and $\mathbf{B}$ is the between-imputation variance. 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.

- Like weighting methods, imputation methods shortcomings when applied to estimation of models with missing conditioning variables. If $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)$, we are interested in $D(y|\mathbf{x})$, data are missing on y and $\mathbf{x}_2$ – say, for the same units – 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

● The lecture notes discuss 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.)

$$y_1 = \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + u_1 \tag{49}$$

$$y_2 = \mathbf{z}\boldsymbol{\delta}_2 + v_2 \tag{50}$$

$$y_3 = 1[\mathbf{z}\boldsymbol{\delta}_3 + v_3 > 0]. \tag{51}$$

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) $v_3|\mathbf{z}$ ~Normal$(0,1)$; (d) $E(\mathbf{z}'v_2) = \mathbf{0}$ and

$\boldsymbol{\delta}_{22} \neq \mathbf{0}$, then we can write

$$y_1 = \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + g(\mathbf{z}, y_3) + e_1 \tag{52}$$

where $e_1 = u_1 - g(\mathbf{z}, y_3) = u_1 - E(u_1|\mathbf{z}, y_3)$. Selection is exogenous in (52) because $E(e_1|\mathbf{z}, y_3) = 0$. Because y_2 is not exogenous, we estimate (52) by IV, using the selected sample, with IVs $(\mathbf{z}, \lambda(\mathbf{z}\boldsymbol{\delta}_3))$ 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 instruments $(\mathbf{z}_i, \hat{\lambda}_{i3})$.

- If y_2 is always observed, tempting to obtain the fitted values $\hat{y}_{i2}$ from the reduced form y_{i2} on $\mathbf{z}_i$, 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 $y_2 = \mathbf{z}\boldsymbol{\pi}_2 + \eta_2 \lambda(\mathbf{z}\boldsymbol{\delta}_3) + r_3$ in the selected population, and does not care whether this is a conditional expectation.

- Should have at least two elements in $\mathbf{z}$ not in $\mathbf{z}_1$: one to exogenously vary y_2, one to exogenously vary selection, y_3.
- 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.