**What's New in Econometrics**?                                    **NBER**, **Summer 2007**
**Lecture 12**, **Wednesday**, **August 1st**, **11-12.30 pm**
# Missing Data

These notes discuss various aspects of missing data in both pure cross section and panel data settings. We begin by reviewing assumptions under which missing data can be ignored without biasing estimation or inference. Naturally, these assumptions are tied to "exogenous" sampling.

We then consider three popular solutions to missing data: inverse probability weighting, imputation, and Heckman-type selection corrections. The first to methods maintain "missing at random" or "ignorability" or "selection on observables" assumptions. Heckman corrections, whether applied to cross section data or panel data, linear models or (certain) nonlinear models, allow for "selection on unobservables." Unfortunately, their scope of application is limited. An important finding is that all methods can cause more harm than good if selection is on conditioning variables that are unobserved along with response variables.

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

It is easy to obtain conditions under which we can ignore the fact that certain variables for some observations, or all variables for some observations, have missing values. Start with a linear model with possibly endogenous explanatory variables:

$$y_i = x_i\beta + u_i, \tag{1.1}$$

where $x_i$ is $1 \times K$ and the instruments $z_i$ are $1 \times L$, $L \geq K$. We model missing data with a selection indicator, drawn with each $i$. The binary variable $s_i$ is defined as $s_i = 1$ if we can use observation $i$, $s_i = 0$ if we cannot (or do not) use observation $i$. In the $L = K$ case we use IV on the selected sample, which we can write as

$$\hat{\beta}_{IV} = \left( N^{-1}\sum_{i=1}^{N} s_i z_i' x_i \right)^{-1} \left( N^{-1}\sum_{i=1}^{N} s_i z_i' y_i \right) \tag{1.2}$$

$$= \beta + \left( N^{-1}\sum_{i=1}^{N} s_i z_i' x_i \right)^{-1} \left( N^{-1}\sum_{i=1}^{N} s_i z_i' u_i \right) \tag{1.3}$$

For consistency, we essentially need

$$rank\ E(s_i z_i' x_i) = K \tag{1.4}$$

and

$$E(s_i z_i' u_i) = 0, \tag{1.5}$$

which holds if $E(z_i'u_i|s_i) = 0$, which in turn is implied by

$$E(u_i|z_i, s_i) = 0. \tag{1.6}$$

Sufficient for (1.6) is

$$E(u_i|z_i) = 0, \quad s_i = h(z_i) \tag{1.7}$$

for some function $h(\cdot)$. Note that the zero covariance assumption, $E(z_i'u_i) = 0$, is not sufficient for consistency when $s_i = h(z_i)$. A special case is when $E(y_i|x_i) = x_i\beta$ and selection $s_i$ is a function of $x_i$. Provided the selected sample has sufficient variation in $x$, can consistently estimate $\beta$ by OLS on the selected sample.

We can use similar conditions for nonlinear models. What is sufficient for consistency on the selected sample?

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

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

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

All of these allow selection on $x$ but not generally on $y$ (or unobservables that affect $y$).

In the statistics literature, just using the data for which we observe all of $(y_i, x_i, z_i)$ (or just $(y_i, x_i)$ without instruments) is called the "complete case method." In cases where we model some feature of $D(y|x)$, it is clear that the richer is $x$, the more likely ignoring selection will not bias the results. In the case of estimating unconditional moments, say $\mu = E(y_i)$, unbiasedness and consistency of the sample on the selected sample requires $E(y|s) = E(y)$.

Similar conditions can be obtained for panel data. For example, if we model $D(y_t|x_t)$, and $s_t$ is the indicator equal to one if $(x_t, y_t)$ is observed, then the condition sufficient to ignore selection is

$$D(s_t|x_t, y_t) = D(s_t|x_t), \ t = 1, \dots, T. \tag{1.8}$$

If, for example, $x_t$ contains $y_{t-1}$, then selection is allowed to depend on the lagged response under (1.8). To see that (1.8) suffices, 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

$$\sum_{t=1}^{T} s_{it} \log f_t(y_{it}|x_{it}, g) \equiv \sum_{t=1}^{T} s_{it} l_{it}(g). \tag{1.9}$$

Except for ensuring identifiability of $\gamma$, it suffices to show that $E[s_{it}l_{it}(\gamma)] \geq E[s_{it}l_{it}(g)]$ for all $g \in \Gamma$ (the parameter space). But by a well-known result from MLE theory – the Kulback-Leibler information inequality – $\gamma$ maximizes $E[l_{it}(g)|x_{it}]$ for all $x_{it}$. But

$$E[s_{it}l_{it}(g)|x_{it}] = E\{E[s_{it}l_{it}(g)|y_{it},x_{it}]|x_{it}\} = E\{E(s_{it}|y_{it},x_{it})l_{it}(g)|x_{it}\}$$
$$= E\{E(s_{it}|x_{it})l_{it}(g)|x_{it}\} = E(s_{it}|x_{it})E[l_{it}(g)|x_{it}],$$

where we used $E(s_{it}|y_{it},x_{it}) = E(s_{it}|x_{it})$ from (1.8). 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 $g \in \Gamma$. Taking expectations of this inequality and using iterated expectations gives the result. Thus, we have shown that $\gamma$ maximizes the expected value of each term in the summand in (1.9) – often not uniquely – and so it also maximizes the expected value of the sum. For identification, we have to assume it is the unique maximizer, as is usually the case of the model is identified in an unselected population and the selection scheme selects out "enough" of the population. One implication of this finding is that selection is likely to be less of a problem in dynamic models where lags of $y$ and lags of other covariates appear, because then selection is allowed to be an arbitrary function of them. But, what is ruled out by (1.8) is selection that depends on idiosyncratic shocks to $y$ between $t - 1$ and $t$.

Methods to remove time-constant, unobserved heterogeneity deserve special attention. Suppose we have the linear model, written for a random draw $i$,

$$yi_t = \eta_t + x_{it}\beta + c_i + u_{it}. \tag{1.10}$$

Suppose that we have instruments, say $z_{it}$, for $x_{it}$, including the possibility that $z_{it} = x_{it}$. If we apply random effects IV methods on the unbalanced panel, sufficient for consistency (fixed $T$) are

$$E(u_{it}|z_{i1},\ldots,z_{iT},s_{i1},\ldots,s_{iT},c_i) = 0, \ t = 1,\ldots,T \tag{1.11}$$

and

$$E(c_i|z_{i1},\ldots,z_{iT},s_{i1},\ldots,s_{iT}) = E(c_i) = 0, \tag{1.12}$$

along with a suitable rank condition. Somewhat weaker conditions suffice, but the general point is that selection must be strictly exogenous with respect to the idiosyncratic errors as well as the unobserved effect, $c_i$. If we use the fixed effects estimator on the unbalanced panel, we can get by with the first assumption, but, of course, all the instruments and selection to be arbitrarily correlated with $c_i$. To see why, consider the just identified case and define, say, $\ddot{y}_{it} = y_{it} - T_i^{-1}\sum_{r=1}^{T} s_{ir}y_{ir}$ 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 $i$ (properly viewed as random). The FEIV estimator is

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

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

Because $\ddot{z}_{it}$ is a function of $(z_{i1}, \ldots, z_{iT}, s_{i1}, \ldots, s_{iT})$, (1.11) implies $\sum_{t=1}^{T} E(s_{it} \ddot{z}_{it}' u_{it}) = 0$ (as do weaker assumptions). There is a set of second moment assumptions that makes the usual, nonrobust inference procedures valid, but these impose homoskedasticity and serial independence of the $u_{it}$ conditional on $(z_i, s_i, c_i)$.

There are some simple ways to test for selection bias in panel data applications. One important violation of (1.11) is when units drop out of the sample in period $t + 1$ because of shocks realized in time $t$. This generally induces correlation between $s_{i,t+1}$ and $u_{it}$. A simple test in the FE environment is to simply add $s_{i,t+1}$ to the equation at time $t$, and estimate the resulting model by fixed effects (or FEIV). A simple $t$ test can be used (probably fully robust). Of course the test entails dropping the last time period, and it need not have power for detecting correlation between $s_{it}$ and $u_{it}$ – that is, contemporaneous selection.

The consistency of FE (and FEIV) on the unbalanced panel under (1.11) breaks down if the slope coefficients are random but one ignores this in estimatin. That is, replace $\beta$ with $b_i$ but still use the FE estimator. Then the error term contains the term $x_i d_i$ where $d_i = b_i - \beta$. If selection is a function of $d_i$, then the usual FE estimator will be inconsistent. (Recall that the FE estimator, on balanced panels, has some robustness to random slopes.) A simple test is to allow $d_i$ to be correlated with selection through the number of available time periods, $T_i$. The idea is to consider alternatives with

$$\mathrm{E}(b_i | z_{i1}, \ldots, z_{iT}, s_{i1}, \ldots, s_{iT}) = E(b_i | s_{i1}, \ldots, s_{iT}) = E(b_i | T_i). \tag{1.13}$$

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

$$1[T_i = 2] x_{it}, \ 1[T_i = 3] x_{it}, \ \ldots, \ 1[T_i = T - 1] x_{it} \tag{1.14}$$

to the equation and estimate it by FE. Significance of these interaction terms indicates that random slopes are correlated with the available time periods, and suggests one might have to remove those random slopes (if possible).

If we first difference instead to remove $c_i$ – a method that has important advantages for

4

attrition problems – we can apply the pooled IV results:

$$\Delta y_{it} = \varphi_t + \Delta x_{it} + \Delta u_{it}, \ t = 2, \ldots, T \tag{1.15}$$

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

$$E(\Delta u_{it}|z_{it}, s_{it}) = 0 \tag{1.16}$$

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

### 2.1. Weighting with Cross-Sectional Data

A general solution to solving missing data problems when selection is not exogenous is based on probability weights. To illustrate, for simplicity, 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 a vector that, for now, we assume is always observed. Suppose we assume the "selection on observables" assumption

$$P(s = 1|y, z) = P(s = 1|z) \equiv p(z) \tag{2.1}$$

where $p(z) > 0$ for all possible values of $z$. Then we can solve the missing data problem by weighting the observed data points by $1/p(z_i)$:

$$\tilde{\mu}_{IPW} = N^{-1} \sum_{i=1}^{N} \left( \frac{s_i}{p(z_i)} \right) y_i, \tag{2.2}$$

where note that $s_i$ selects out the observed data points. It is easy to show, using iterated expectations, that $\hat{\mu}_{IPW}$ is not only consistent for $y_i$, it is unbiased, too. (This same kind of estimator arises in treatment effect estimation.) Of course, except in special cases, we must estimate $p(z_i)$; when $z_i$ is always observed along with $s_i$, flexible binary response models such as logit or probit, or nonparametric methods, can be used. Let $\hat{p}(z_i)$ denote the estimated selection probability (also called the propensity score). Then an operational estimator is

$$\hat{\mu}_{IPW} = N^{-1} \sum_{i=1}^{N} \left( \frac{s_i}{\hat{p}(z_i)} \right) y_i. \tag{2.3}$$

As written, this estimator assumes we know the size of the random sample, $N$, which is not

necessarily the case for some sampling schemes, such as variable probability sampling. We can also write $\hat{\mu}_{IPW}$ as

$$\hat{\mu}_{IPW} = N_1^{-1}(N_1/N) \sum_{i=1}^{N} \left( \frac{s_i}{\hat{p}(z_i)} \right) y_i = N_1^{-1} \sum_{i=1}^{N} s_i \left( \frac{\hat{\rho}}{\hat{p}(z_i)} \right) y_i \tag{2.4}$$

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 $\hat{\rho}/\hat{p}(z_i)$, which can be greater or less than unity. (By iterated expectatins, $\rho = E[p(z_i)]$.) Equation (2.4) shows that $\hat{\mu}_{IPW}$ is a weighted average of the observed data points with weights $\hat{\rho}/\hat{p}(z_i)$. Yet a different estimator is obtained by solving the least squares problem

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

which results in

$$\check{\mu}_{IPW} = \left( \sum_{h=1}^{N} \frac{s_h}{\hat{p}(z_h)} \right)^{-1} \left( \sum_{i=1}^{N} \left( \frac{s_i}{\hat{p}(z_i)} \right) y_i \right), \tag{2.5}$$

which is a different version a weighted.average.

Horowitz and Manski (1998) have considered the problem of estimating population means using IPW. Their main focus is on establishing bounds that do not rely on potentially strong, untestable assumptions such as the unconfoundedness assumption in (2.1). But they also note a particular problem with certain IPW estimators even when the conditioning variable, $x$, is always observed. They consider estimation of the mean $E[g(y)|x \in A]$ for some set $A$. If we define $d_i = 1[x_i \in A]$ then the problem is to estimate $E[g(y)|d = 1]$. HM point out that, if one uses the weights commonly reported with survey data – weights that do not condition on the event $d = 1$ – then the IPW estimate of the mean can lie outside the logically possible values of $E[g(y)|d = 1]$. HM note that this problem can be fixed by using probability weights $P(s = 1|d = 1)/P(s = 1|d = 1, z)$ unfortunately, this choice is not possible when data on $x$ can also be missing.

Failure to condition on $d = 1$ when computing the probability weights when interest lies in $E[g(y)|d = 1]$ is related to a general problem that arises in estimating models of conditional means when data are missing on $x$. To see why, suppose the population regression function is linear:

$$E(y|x) = \alpha + x\beta. \tag{2.6}$$

Let $z$ be a variables that are always observed and let $p(z)$ be the selection probability, as before. Now, suppose that at least part of $x$ is not always observed, so that $x$ is not a subset of $z$. This means that some elements of $x$ cannot appear in $p(z)$ because $p(z)$ normally has to be estimated using the data on $(s_i, z_i)$ for all $i$. The IPW estimator of $\beta$ solves

$$\min_{a,b} \sum_{i=1}^{N} \left( \frac{s_i}{\hat{p}(z_i)} \right) (y_i - a - x_i b)^2. \tag{2.7}$$

Here is the problem: suppose that selection is exogenous in the sense that

$$P(s = 1|x,y) = P(s = 1|x). \tag{2.8}$$

Then we saw in Section 1 that using least squares on the selected sample results in a consistent estimator of $\theta = (\alpha, \beta')'$, which is also $\sqrt{N}$-asymptotically normal. What about the weighted estimator? The problem is that if (2.8) holds, and $z$ does not include $x$, then it is very unlikely that

$$P(s = 1|x,y,z) = P(s = 1|z). \tag{2.9}$$

In other words, the key unconfoundedness assumption fails, and the IPW estimator of $\theta$ is generally inconsistent. We actually introduce inconsistency by weighting when a standard unweighted regression on the complete cases would be consistent. In effect, the IPW estimator uses weights that are functions of the wrong variables.

If $x$ is always observed and can (and should) be included in $z$, then weighting is much more attractive. Typically, $z$ might contain lagged information, or interview information that would not be included in $x$. If it turns out that selection is a function only of $x$, flexible estimation of the model $P(s = 1|z)$ will pick that up in large sample sizes.

If $x$ is always observed and we know that $P(s = 1|x,y) = P(s = 1|x)$, is there any reason to weight by $1/p(x)$? If $E(y|x) = \alpha + x\beta$ and $Var(y|x)$, weighting is asymptotically inefficient. If $E(y|x) = \alpha + x\beta$ but $Var(y|x)$ is heteroskedastic, then weighting may or may not be more efficient than not. (The efficient estimator would be the WLS estimator that appropriately accounts for $Var(y|x)$, different issue than probability weighting.) But both weighting and not weighting are consistent. The advantage of weighting is that, if the population "model" is in fact just a linear projection, the IPW estimator consistently estimates that linear projection while the unweighted estimator does not. In other words, if we write

$$L(y|1,x) = \alpha^* + x\beta^* \tag{2.10}$$

where $L(\cdot|\cdot)$ denotes the linear projection, then under $P(s = 1|x,y) = P(s = 1|x)$, the IPW estimator is consistent for $\theta^*$. The unweighted estimator has a probabilty limit that depends on $p(x)$.

One reason to be interested in the LP is that the parameters of the LP show up in certain treatment effect estimators. The notes on treatment effects contained a discussion of a "double robustness" result due to Robins and Ritov (1997); see also Wooldridge (2007). The idea is this. In treatment effect applications, the ATE requires estimation of $E(y_g)$ for the two counterfactual outcomes, $g = 0, 1$). The LP has the property that $E(y_1) = \alpha_1^* + E(x)\beta_1^*$, and so, if we consistently estimate $\alpha_1^*$ and $\beta_1^*$ then we can estimate $E(y_1)$ by averaging across $x$. A similar statement holds for $y_0$. Now, the IPW estimator identifies $\alpha_1^*$ and $\beta_1^*$ if the model for $p(x)$ is correctly specified. But if $E(y_1|x) = \alpha_1 + x\beta_1$ then the IPW estimator is consistent for $\alpha_1$ and $\beta_1$ even if $p(x)$ is misspecified. And, of course, $E(y_1) = \alpha_1 + E(x)\beta_1$. So, regardless of whether we are estimating the conditional mean parameters or the LP parameters, we consistently estimate $E(y_1)$. The case where the IPW estimator does not consistently estimate $E(y_1)$ is when $E(y_1|x)$ is not linear and $p(x)$ is misspecified.

The double robustness result holds for certain nonlinear models, too, although one must take care in combining the conditional mean function with the proper objective function – which, in this case, means quasi-log-likelihood function. The two cases of particular interest are the logistic response function for binary or fractional responses coupled with the Bernoulli QLL, and the exponential response function coupled with the Poisson QLL.

Returning to the IPW regression estimator that solves (2.7), suppose we assume the ignorability assumption (2.9),

$$E(u) = 0, E(x'u) = 0,$$

$$p(z) = G(z, \gamma)$$

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

$$Avar\sqrt{N}\,(\hat{\theta}_{IPW} - \theta) = [E(x_i'x_i)]^{-1}E(r_ir_i')[E(x_i'x_i)]^{-1}, \tag{2.11}$$

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 $\hat{\gamma}$. The asymptotic variance of $\hat{\theta}_{IPW}$ is

easy to estimate:

$$\left(\sum_{i=1}^{N}[s_i/G(z_i,\hat{\gamma})]x_i'x_i\right)^{-1}\left(\sum_{i=1}^{N}\hat{r}_i\hat{r}_i'\right)\left(\sum_{i=1}^{N}[s_i/G(z_i,\hat{\gamma})]x_i'x_i\right)^{-1},\tag{2.12}$$

or, if $x_i$ is always observed, the terms $s_i/G(z_i,\hat{\gamma})$ can be dropped in the outer parts of the sandwich. In the case that $d_i$ is the score from a logit model of $s_i$ on functions, say, $h(z_i)$, $\hat{d}_i$ has the simple form

$$\hat{d}_i = h_i'(s_i - \Lambda(h_i\hat{\gamma})),\tag{2.13}$$

where $\Lambda(a) = \exp(a)/[1 + \exp(a)]$ and $h_i = h(z_i)$. This illustrates a very interesting finding of Robins, Rotnitzky, and Zhou (1995) and related to the Hirano, Imbens, and Ritter (2003) efficient estimator for means using IPW estimators. Suppose that, for a given set of functions $h_{i1}$, the logit model is correctly specified in the sense that there is a $\gamma_1$ such that $P(s_i = 1|z_i) = \Lambda(h_{i1}\gamma_1)$. Now suppose we take some additional functions of $z_i$, say $h_{i2} = h_2(z_i)$, and add them to the logit. Then, asymptotically, the coefficients on $h_{i2}$ are zero, and so the adjustment to the asumptotic variance comes from regressing $(s_i/\Lambda(h_{i1}\gamma_1)x_i'u_i$ on $(h_{i1}, h_{i2})[s_i - \Lambda(h_{i1}\gamma_1)]$. Now, notice that, even though the coefficients on $h_{i2}$ are zero in the logit model, the score vector depends on $(h_{i1}, h_{i2})$. Therefore, the residual variance from regressing $(s_i/\Lambda(h_{i1}\gamma_1)x_i'u_i$ on $(h_{i1}, h_{i2})[s_i - \Lambda(h_{i1}\gamma_1)]$ is generally smaller than that from using the correct logit model, which is obtained from regressing on $h_{i1}[s_i - \Lambda(h_{i1}\gamma_1)]$. By overspecifying the logit model for $s_i$, we generally reduce the asymptotic variance of the IPW estimator. And the process does not stop there. We can keep adding functions of $z_i$ to the logit to reduce the asymptotic variance of the estimator of the IPW estimator. In the limit, if we have chosen the sequence of functions so that they approximate any well-behaved function, then we achieve asymptotic efficiency. This is precisely what the HIR estimator does by using a logit series estimator for the propensity score.

Wooldridge (2007) shows that the adjustment to the asymptotic variance in (2.12) carries over to general nonlinear models and estimation methods. One consequence is that ignoring the estimation in $\hat{p}(z)$ – as commercial software typically does when specifying probability weights – results in conservative inference. But the adjustment to obtain the correct asymptotic variance is fairly straightforward.

Nevo (2003) explicitly considers a generalized method of moments framework, and shows how to exploit known population moments to allow selection to depend on selected elements

of the data vector $w$. (Hellerstein and Imbens (1998) use similar methods to improve estimation when population moments are known.) In particular, Nevo assumes that, along with the moment conditions $E[r(w, \theta)] = 0$, the population moments of the vector $h(w)$, say $\mu_h$, are known. Under the assumption that selection depends on $h(w)$, that is, $P(s = 1|w) = P(s = 1|h(w))$, Nevo obtains an expanded set of moment conditions that can be used to estimate $\theta$ and the parameters $\gamma$ in the selection equation. Suppose we use a logit model for $P(s = 1|h(w))$. Then

$$E\left[ \frac{s_i}{\Lambda(h(w_i)\gamma)} r(w_i, \theta) \right] = 0 \tag{2.14}$$

and

$$E\left[ \frac{s_i h(w_i)}{\Lambda(h(w_i)\gamma)} \right] = \mu_h. \tag{2.15}$$

Equation (2.15) generally identifies $\gamma$, and then this $\hat{\gamma}$ can be used in a second step to choose $\hat{\theta}$ to make the weighted sample moments

$$N^{-1} \sum_{i=1}^{N} \left[ \frac{s_i}{\Lambda(h(w_i)\hat{\gamma})} r(w_i, \hat{\theta}) \right] \tag{2.16}$$

as close to zero as possible. Because (2.15) adds as many moment restrictions as parameters, the GMM estimator using both sets of moment conditions is equivalent to the two-step estimator just described.

Another situation where the missing data problem can be solved via weighting is when data have been censored due to a censored duration. The response variable of interest may be the duration, or it may be a variable observed only if a duration or survival time is observed. Let $y$ be a univariate response and $x$ a vector of conditioning variables, and suppose we are interested in estimating $E(y|x)$. A random draw $i$ from the population is denoted $(x_i, y_i)$. Let $t_i > 0$ be a duration and let $c_i > 0$ denote a censoring time (where $t_i = y_i$ is allowed). Assume that $(x_i, y_i)$ is observed whenever $t_i \leq c_i$, so that $s_i = 1(t_i \leq c_i)$. Under the assumption that $c_i$ is independent of $(x_i, y_i, t_i)$,

$$P(s_i = 1|x_i, y_i, t_i) = G(t_i), \tag{2.17}$$

where $G(t) \equiv P(c_i \geq t)$. In order to use inverse probability weighting, we need to observe $t_i$ whenever $s_i = 1$, which simply means that $t_i$ is uncensored. Plus, we need only observe $c_i$ when $s_i = 0$. of $c_i$. As shown in Wooldridge (2007), it is more efficient to estimate $G(\cdot)$ using

the density of $\min(c_i, t_i)$ given $t_i$. Generally, let $h(c, \gamma)$ denote a parametric model for the density of the censoring times, $c_i$, and let $G(t, \gamma)$ be the implied model for $P(c_i \geq t)$. The log likelihood is

$$\sum_{i=1}^{N} \{(1 - s_i) \log[h(c_i, \gamma)] + s_i \log[G(t_i, \gamma)]\}, \tag{2.18}$$

which is just the log-likelihood for a standard censored estimation problem but where $t_i$ (the underlying duration) plays the role of the censoring variable. As shown by Lancaster (1990 for grouped duration data, where $h(c, \gamma)$ is piecewise constant, the solution to (2.18) gives a survivor function identical to the Kaplan-Meier estimator but where the roles of $c_i$ and $t_i$ are reversed; that is, we treat $t_i$ as censoring $c_i$. The linear regression model has a long history, and has been studied recently by Honoré, Khan, and Powell (2002). See also Rotnitzky and Robins (2005) for a survey of how to obtain semiparametrically efficient estimators. The Koul-Susarla-van Ryzin (1981) estimator is an IPW least squares estimator, but their proposals for inference are very difficult to implement. As shown by Wooldridge (2007), this is another instance where estimating the selection probability by MLE is more efficient than using the known probability (if you could). Plus, obtaining the smaller variance matrix involves only a multivariate regression of the weighted score for the second stage problem – OLS, NLS, MLE, or IV – on the score for the first-stage Kaplan-Meier estimation. This simple procedure is valid when the distribution of $c_i$ is taken to be discrete. Other authors undertake the asymptotics allowing for an underlying continuous censoring time, which makes estimating asymptotic variances considerably more difficult.

## 2.2 Attrition in Panel Data

Inverse probability weighting can be applied to solve, in some cases, the attrition problem in panel data. For concreteness, consider maximum pooled maximum likelihood, where we model a density $f_t(y_t|\mathbf{x}_t)$ for any conditioning variables $\mathbf{x}_t$. These need not be strictly exogenous or always observed. Let $f_t(y_t|\mathbf{x}_t, \theta)$ be the parametric model, and let $s_{it}$ be the selection indicator. We assume that attrition is absorbing, so $s_{it} = 1 \Rightarrow s_{ir} = 1$ for $r < t$. The estimator that ignores attrition solves

$$\max_{\theta \in \Theta} \sum_{i=1}^{N} \sum_{t=1}^{T} s_{it} \log f_t(y_{it}|\mathbf{x}_{it}, \theta), \tag{2.19}$$

which is consistent if $P(s_{it} = 1|y_{it}, \mathbf{x}_{it}) = P(s_{it} = 1|\mathbf{x}_{it})$. This follows by showing

$E[P(s_{it} = 1|\mathbf{x}_{it})E[\log f_t(y_{it}|\mathbf{x}_{it}, \boldsymbol{\theta})|\mathbf{x}_{it}]$, and using the fact that the true value of $\theta$ maximizes $E[\log f_t(y_{it}|\mathbf{x}_{it}, \boldsymbol{\theta})|\mathbf{x}_{it}]$ for all $t$, and $P(s_{it} = 1|\mathbf{x}_{it}) \geq 0$. But, if selection depends on $y_{it}$ even after conditioning on $\mathbf{x}_{it}$, the unweighted estimator is generally inconsistent. If $\mathbf{w}_{it} = (\mathbf{x}_{it}, y_{it})$, then perhaps we can find variables $\mathbf{r}_{it}$, such that

$$P(s_{it} = 1|\mathbf{w}_{it}, \mathbf{r}_{it}) = P(s_{it} = 1|\mathbf{r}_{it}) \equiv p_{it} > 0, t = 1, \ldots, T. \tag{2.20}$$

(The "obvious" set of variables $\mathbf{r}_{it} = \mathbf{w}_{it}$ is not usually available since we will have estimate the probabilities.) If we could observe the $p_{it}$, we could use the weighted MLE,

$$\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \sum_{i=1}^{N} \sum_{t=1}^{T} (s_{it}/p_{it}) \log f_t(y_{it}|\mathbf{x}_{it}, \boldsymbol{\theta}), \tag{2.21}$$

which we call $\hat{\theta}_w$. The estimator $\hat{\theta}_w$ is generally consistent because

$$E[(s_{it}/p_{it})q_t(\mathbf{w}_{it}, \boldsymbol{\theta})] = E[q_t(\mathbf{w}_{it}, \boldsymbol{\theta})], t = 1, \ldots, T,. \tag{2.22}$$

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

How do we choose $\mathbf{r}_{it}$ to make (2.20) hold (if possible)? A useful strategy, considered by RRZ, is to 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

$$\pi_{it} = P(s_{it} = 1|z_{it}, s_{i,t-1} = 1), t = 1, \ldots, T. \tag{2.23}$$

Typically, $\mathbf{z}_{it}$ contains elements from $(\mathbf{w}_{i,t-1}, \ldots, \mathbf{w}_{i1})$, 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 $t$!) Given the monotone nature of selection, we can estimate models for $\pi_{it}$ sequential when the $\mathbf{z}_{it}$ are observed for every unit in the sample at time $t-1$.

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

$$P(s_{it} = 1|\mathbf{v}_{i1}, \ldots, \mathbf{v}_{iT}, s_{i,t-1} = 1) = P(s_{it} = 1|\mathbf{z}_{it}, s_{i,t-1} = 1), t \geq 1. \tag{2.24}$$

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 $z_{it}$; in practice, this means only on variables observed at $t-1$. This is a strong assumption; RRZ (1995) show how to relax it somewhat in a regression framework with time-constant covariates. Using this assumption, we can show that

$$p_{it} \equiv P(s_{it} = 1|\mathbf{v}_i) = \pi_{it}\pi_{i,t-1} \cdots \pi_{i1}. \tag{2.25}$$

In the general framework, we have $\mathbf{r}_{it} = (\mathbf{z}_{it}, \ldots, \mathbf{z}_{i1})$ but, because of the ignorability assumption, it is as if we can take $\mathbf{r}_{it} = [(\mathbf{w}_{i1}, \mathbf{z}_{i1}), \ldots, (\mathbf{w}_{iT}, \mathbf{z}_{it})]$.

So, a consistent two-step method is:

(1) 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 $\hat{p}_{it} = \hat{\pi}_{it} \hat{\pi}_{i,t-1} \cdots \hat{\pi}_{i1}$. Note that we are able to compute $\hat{p}_{it}$ only for units in the sample at time $t - 1$.

(2) Replace $p_{it}$ with $\hat{p}_{it}$ in (2.21), and obtain the weighted M-estimator.

Consistency is straightforward – standard two-step estimation problem – if we have the correct functional form and the ignorability of selection assumption holds. 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; it is similar to that used for the cross section case, but just pooled across $t$.

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: weighting uncessesarily causes inconsistency.

Related to the previous point is that it would be rare to apply IPW in the case of a model with completely specified dynamics. Why? Suppose, for example, we have a model of $E(y_{it} | x_{it}, y_{i,t-1}, \ldots, x_{i1}, y_{i0})$. If our variables affecting attrition, $z_{it}$, are functions of $(y_{i,t-1}, \ldots, x_{i1}, y_{i0})$ – as they often must be – then selection is on the basis of conditioning variables, and so the unweighted estimator is also consistent. RRZ (1995) explicitly cover regressions that do not have correct dynamics.

## 3. Imputation

Section 1 discussed conditions under which dropping observations with any missing data results in consistent estimators. Section 2 showed that, under an unconfoundedness assumption, inverse probability weighting can be applied to the complete cases to recover population parameters. One problem with using IPW for models that contain covariates is that the weighting may actually hurt more than it helps if the covariates are sometimes missing and selection is largely a function of those covariates.

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. Imputation methods, and multiple imputation use

either means, fitted values, values or averages from "similar" observations, or draws from posterior distributions to fill in the missing values. Little and Rubin (2002) provides an accessible treatment with lots of references to work by Rubin and coauthors.

Naturally, such procedures cannot always be valid. 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 the unconfoundedness assumption $P(s = 1|y,x) = P(s = 1|x)$. (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. For example, in a bivariate case, 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 $w_i$, so $D(r_i|w_i) = D(r_i)$. The MAR assumption is that implies

$P(r_{i1} = 0, r_{i2} = 0|w_i) = P(r_{i1} = 0, r_{i2} = 0) \equiv \pi_{00}, P(r_{i1} = 1, r_{i2} = 0|w_{i1}),$

$P(r_{i1} = 0, r_{i2} = 1|w_{i2}),$ and then

$P(r_{i1} = 1, r_{i2} = 1|w_i) = 1 - \pi_{00} - P(r_{i1} = 1, r_{i2} = 0|w_{i1}) - P(r_{i1} = 0, r_{i2} = 1|w_{i2}).$ 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, which sometime apply in panel data situations with attrition. Order the $w_{ig}$ so that if $w_{ih}$ is observed the so is $w_{ig}$, $g < h$. Then the retention indicators satisfy $r_{ig} = 1 \Rightarrow r_{i,g-1} = 1$. Under MAR, the joint density $f(w_1,\ldots,w_G)$ is easy to estimate. 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

$$\sum_{g=1}^{G} r_{ig} \log f(w_{ig}|w_{i,g-1},\ldots,w_{i1},\theta), \tag{3.1}$$

where $f(w_1|w_0,\theta) \equiv f(w_1|w_0,\theta)$, and it suffices to multiply only by $r_{ig}$ because $r_{ig} = r_{ig}r_{i,g-1}\cdots r_{i2}.$ Under MAR,

$$E(r_{ig}|w_{ig},\ldots,w_{i1}) = E(r_{ig}|w_{i,g-1},\ldots,w_{i1}), \tag{3.2}$$

and so by (3.2),

$$E[r_{ig} \log f(w_{ig}|w_i^{(g-1)}\theta)|w_i^{(g-1)}] = E(r_{ig}|w_i^{(g-1)})E[\log f(w_{ig}|w_i^{(g-1)}\theta)|w_i^{(g-1)}]. \tag{3.3}$$

The first term on the RHS of (3.3) is $E(r_{ig}|w_i^{(g-1)}) = P(r_{ig} = 1|w_i^{(g-1)}) \geq 0$ and the true value of

$\theta$ maximizes the second part by the conditional Kullback-Leibler information inequality (for example, Wooldridge (2002, Chapter 13)). Therefore, the parameters of the conditional densities are generally identified, provided the missing data problem is not too severe.

Before briefly describing how multiple imputation works, a simple example helps illustrate the general idea behind imputation. Suppose $y$ is a random variable in a population with mean $\mu_y$, but data are missing on some $y_i$ random drawn from the population. Let $s_i$ be the binary selection indicator, and let $\mathbf{x}_i$ be a set of observed covariates. So, a random draw consists of $(\mathbf{x}_i, y_i, s_i)$ but where $y_i$ is missing if $s_i = 0$. As we discussed earlier, unless $s$ is independent of $y$ – that is, the data are MCAR – the complete-case sample average,

$$\tilde{\mu}_y = \left( \sum_{i=1}^{N} s_i \right)^{-1} \sum_{i=1}^{N} s_i y_i, \tag{3.4}$$

is not unbiased or consistent for $\mu_y$; its probability limit is, of course, $E(y|s = 1)$.

Suppose, however, that the selection is ignorable conditional on $\mathbf{x}$:

$$E(y|\mathbf{x}, s) = E(y|\mathbf{x}) = m(\mathbf{x}, \boldsymbol{\beta}), \tag{3.5}$$

where $m(\mathbf{x}, \boldsymbol{\beta})$ is, for simplicity, a parametric function. As we discussed in Section 1, nonlinear least squares, and a variety of quasi-MLEs, are consistent for $\boldsymbol{\beta}$ using the selected sample. Now, 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 $\mu_y$ is

$$\hat{\mu}_y = N^{-1} \sum_{i=1}^{N} \{s_i y_i + (1 - s_i) m(\mathbf{x}_i, \hat{\boldsymbol{\beta}})\}. \tag{3.6}$$

The plim of $\hat{\mu}_y$ is easy to find by replacing $\hat{\boldsymbol{\beta}}$ with $\boldsymbol{\beta}$ and sample average with the population average:

$$
\begin{aligned}
E[s_i y_i + (1 - s_i) m(\mathbf{x}_i, \boldsymbol{\beta})] &= E[E(s_i y_i | \mathbf{x}_i, s_i)] + E[(1 - s_i) m(\mathbf{x}_i, \boldsymbol{\beta})] \\
&= E[s_i E(y_i | \mathbf{x}_i, s_i)] + E[(1 - s_i) m(\mathbf{x}_i, \boldsymbol{\beta})] \\
&= E[s_i m(\mathbf{x}_i, \boldsymbol{\beta})] + E[(1 - s_i) m(\mathbf{x}_i, \boldsymbol{\beta})] \\
&= E[m(\mathbf{x}_i, \boldsymbol{\beta})] = \mu_y.
\end{aligned} \tag{3.7}
$$

(Of course, we could average the $m(\mathbf{x}_i, \hat{\boldsymbol{\beta}})$ across all $i$, but that would throw away some information on the $y_i$ that we observe.)

If $D(y|\mathbf{x}, s) = D(y|\mathbf{x})$ then we can use MLE on the complete cases, obtain estimates of the

parameters, say $\hat{\theta}$, and then use $m(\mathbf{x}_i, \hat{\beta})$ as above, where $m(\mathbf{x}, \beta)$ is the mean function implied by the model for $D(y|\mathbf{x})$. For example, $y$ could be a corner solution response and then we use a Tobit or some flexible extension for $D(y|\mathbf{x})$.

One danger in using even simple imputation methods like the one just covered is that we will ignore the errors in the imputed values. $\hat{y}_i$ differs from $y_i$ for two reasons. First, if we write

$$y_i = m(\mathbf{x}_i, \beta) + u_i, \tag{3.8}$$

then, even if we knew $\beta$, the error would be $u_i$. (In effect, we are replace $y_i$ with its conditional expectation.) Having to estimate $\beta$ further introduces estimation error. Analytical formulas can be worked out, but bootstrapping a standard error or confidence interval for $\hat{\mu}_y$ is also straightforward: we would draw observation indices at random, without replacement, and perform the imputation steps on each new bootstrap sample.

As an example of how just using the imputed values as if they were real data, suppose we run a linear regression using the complete data and obtain $\mathbf{x}_i\hat{\beta}$. Again defining $\hat{y}_i = s_i y_i + (1 - s_i)\mathbf{x}_i\hat{\beta}$, suppose we use the imputed data set to reestimate $\beta$. It is well known that we just get $\hat{\beta}$ back again. However, our estimated error variance will be too small because every residual for an imputed data point is identically zero. It follows that, while $SSR/(N_1 - K)$ is generally unbiased for $\sigma_u^2$ (under the Gauss-Markov assumptions), where $N_1$ is the number of complete cases, $SSR/(N - K)$ has a downward bias.

The previous method ignores the random error in (3.4); Little and Rubin (2002) call it the method of "conditional means." Generally, as they show in Table 4.1, the method of conditional means results in downward bias in estimating variances. Instead, LR propose adding a random draw to $m(\mathbf{x}_i, \hat{\beta})$ to impute a value. Of course, this entails have a distribution from which to draw the $u_i$. If we assume that $u_i$ is independent of $\mathbf{x}_i$ and normally distributed, then we can draw, say, $\check{u}_i$ from a Normal$(0, \hat{\sigma}_u^2)$, distribution, where $\hat{\sigma}_u^2$ is estimated using the complete case nonlinear regression residuals. This procedure works well for estimating $\sigma_y^2$ in the case where linear regression is used and $(\mathbf{x}_i, y_i)$ is jointly normal. LR refer to this as the "conditional draw" method of imputation, which is a special case of stochastic imputation.

Little and Rubin argue that the conditional draw approach, at least in the jointly normal case, works well when a covariate is missing. Suppose that $\mathbf{x} = (x_1, x_2)$ and data are missing on $x_2$ but not $(x_1, y)$. One possibility for imputing $x_{i2}$ when it is missing is to regress $x_{i2}$ on $x_{i1}$ using the complete cases, and then use fitted values, or conditional draws, to impute $x_{i2}$. LR show that the method of conditional draws (not conditional means) works well when $y$ is

included along with $x_1$ in obtained the estimated conditional means from the complete-case regression.

Unfortunately, except in simple cases, it is difficult to quantity the uncertainty from single-imputation methods, where one imputed values is obtained for each missing variable. One possibility, which has been studied in the statistics literature, is to bootstrap the entire estimation method – assuming, of course, that the imputations eliminates the nonresponse bias (so that missing at random holds). In the example of conditional draws above, the imputation procedure is simply included in any subsequent estimation, and bootstrap samples are obtained over and over again. On each bootstrap replication, say $b$, an estimate of the parameters using the complete cases, $\hat{\theta}_{complete}^{(b)}$ is obtained (which would be the beta hats and error variance estimate in the regression case), missing data values are imputed using conditional draws, and then an estimate of $\theta$ using the imputed data, $\hat{\theta}_{imputed}^{(b)}$, can be obtained. Of course, this can be computationally intensive for nonlinear estimation problems.

An alternative is the method of multiple imputation. Its justification is Bayesian, and 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. But the idea is easily illustrated using the above example. Rather than just impute one set of missing values to create one "complete" data set, created several imputed data sets. (Often the number is fairly small, such as five or so.) Then, estimate the parameters of interest using each imputed data set, and then use an averaging to obtain a final parameter estimate and sampling error.

Briefly, let $\mathbf{W}_{mis}$ denote the matrix of missing data and $\mathbf{W}_{obs}$ the matrix of observations. Assume that MAR holds. Then multiple imputation is justified as a way to estimate $E(\theta|\mathbf{W}_{obs})$, the posterier mean of $\theta$ given $\mathbf{W}_{obs}$. But by iterated expectations,

$$E(\theta|\mathbf{W}_{obs}) = E[E(\theta|\mathbf{W}_{obs}, \mathbf{W}_{mis})|\mathbf{W}_{obs}]. \tag{3.9}$$

Now, if we can obtain estimates $\hat{\theta}_d = E(\theta|\mathbf{W}_{obs}, \mathbf{W}_{mis}^{(d)})$ for imputed data set $d$, then we can approximate $E(\theta|\mathbf{W}_{obs})$ as

$$\bar{\theta} = D^{-1} \sum_{d=1}^{D} \hat{\theta}_d, \tag{3.10}$$

which is just the average of the parameter estimates across the imputed samples.

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

$$Var(\theta|\mathbf{W}_{obs}) = E[Var(\theta|\mathbf{W}_{obs},\mathbf{W}_{mis})|\mathbf{W}_{obs}] + Var[E(\theta|\mathbf{W}_{obs},\mathbf{W}_{mis})|\mathbf{W}_{obs}], \tag{3.11}$$

which suggests

$$\widehat{Var}(\theta|\mathbf{W}_{obs}) = D^{-1}\sum_{d=1}^{D}\hat{\mathbf{V}}_d + (D-1)^{-1}\sum_{d=1}^{D}(\hat{\theta}_d - \bar{\theta})(\hat{\theta}_d - \bar{\theta})' \tag{3.12}$$

$$\equiv \bar{\mathbf{V}} + \mathbf{B},$$

where $\bar{\mathbf{V}}$ is the average of the variance estimates across imputed samples and $\mathbf{B}$ is the between-imputation variance. For small a small number of imputations, a correction is usually made, namely, $\bar{\mathbf{V}} + (1 + D)^{-1}\mathbf{B}$. Therefore, assume that one trusts the MAR assumption, and the underlying distributions used to draw the imputed values, inference with multiple imputations is fairly straightforward. Because $D$ need not be very large, estimation of nonlinear models using multiple imputations is not computationally prohibitive (once one has the imputed data, of course).

Like weighting methods, imputation methods have an important shortcoming when applied to estimation of models with missing conditioning variables. Suppose again that $\mathbf{x} = (\mathbf{x}_1,\mathbf{x}_2)$, we are interested in some feature of the conditional distribution $D(y|\mathbf{x})$, data are missing on $y$ and $\mathbf{x}_2$ – say, for the same units – and selection is a function of $\mathbf{x}_2$. Then, as we discussed in Section 1, MLE using the complete cases is consistent, asymptotically normal, and inference is standard. What about imputation methods? Because they generally rely on MAR, they would require that $D(s|y,\mathbf{x}_1,\mathbf{x}_2) = D(s|\mathbf{x}_1)$. Because this is false in this example, MI cannot be expected to produce convincing imputations.

Imputation for the monotone case discussed above is relatively straightforward under MAR, and MAR is at least plausible. Because the conditional densities are identified, imputation can proceed sequentially: given $w_{i1}$ and $\hat{\theta}$, missing values on $w_{i2}$ can be imputed by drawing from $f_2(\cdot|w_{i1},\hat{\theta})$. Then, $w_{i3}$ can be imputed by drawing from $f(\cdot|\hat{w}_{i2},w_{i1},\hat{\theta})$, where $\hat{w}_{i2}$ may or may not be imputed. And so on.

## 4. Heckman-Type Selection Corrections

### 4.1. Corrections with Instrumental Variables

Here we briefly cover the well-known Heckman selection correction with endogenous explanatory variables in a linear model. The model is

$$y_1 = \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + u_1 \tag{4.1}$$

$$y_2 = \mathbf{z}\boldsymbol{\delta}_2 + v_2 \tag{4.2}$$

$$y_3 = 1[\mathbf{z}\boldsymbol{\delta}_3 + v_3 > 0]. \tag{4.3}$$

where $\mathbf{z}$ is $1 \times L$ with first element unity (and also in $\mathbf{z}_1$). As usually, $L_1 < L$ for identification. The key point to be made here is, depending on how the Heckman correction is carried out in this case, (4.2) can just be a linear projection – in which case the nature of $y_2$ is unrestricted – or, effectively, $v_2$ must be normal and independent of $\mathbf{z}$. Intuitively, we need two elements in $\mathbf{z}$ not also in $z_1$: loosely, one to induce exogenous variation in $y_2$ and the other to induce exogenous variation in selection. If we 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) $v_3|\mathbf{z} \sim \text{Normal}(0, 1)$; (d) $E(\mathbf{z}'v_2) = \mathbf{0}$ and $\boldsymbol{\delta}_{22} \neq \mathbf{0}$, then we can write

$$y_1 = \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + g(\mathbf{z}, y_3) + e_1 \tag{4.4}$$

where $e_1 = u_1 - g(\mathbf{z}, y_3) = u_1 - E(u_1|\mathbf{z}, y_3)$. So, selection is exogenous in (4.4) because $E(e_1|\mathbf{z}, y_3) = 0$. Because $y_2$ is not exogenous, we estimate (4.4) by IV, using the selected sample, where the instruments are $(\mathbf{z}, \lambda(\mathbf{z}\boldsymbol{\delta}_3))$ because $g(\mathbf{z}, 1) = \lambda(\mathbf{z}\boldsymbol{\delta}_3)$. So, 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 the selected sample and instruments $(\mathbf{z}_i, \hat{\lambda}_{i3})$.

If $y_2$ is always observed, it is 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 normal (or something similar); it would not be consistent for discrete $y_2$, for example. Implicitly, the reduced form estimated by the proper two-step procedure is, on the selected sample, $y_2 = \mathbf{z}\boldsymbol{\pi}_2 + \eta_2\lambda(\mathbf{z}\boldsymbol{\delta}_3) + r_3$. But this is just a linear projection; generally, the rank condition on the selected sample should hold if $\mathbf{z}$ causes sufficient variation in $y_2$ and $y_3$ in the population.

This example raises another point: even if $y_2$ is exogenous in the full population, one should generally treat it as endogenous in the selected subsample. Why? Because $y_2$ cannot be included in the first-stage probit if it is not always observed, so consistency of the Heckman procedure would require $P(y_3 = 1|\mathbf{z}_1, y_2) = P(y_3 = 1|\mathbf{z}_1)$, a tenuous assumption. Unless we have an instrument for $y_2$, simply treating it as exogenous in the second stage after omitting it

from the first is tantamount to imposing an exclusion restriction on a reduced form.

In addition to the linear model, with or without endogenous variables, Heckman-type corrections are available for a limited set of nonlinear models. Terza (1998) contains the approach for exponential functions with exogenous explanatory variables, where the selection equation follows a probit; see also Wooldridge (2002, Chapter 19). A selection correction is also fairly easy to implement in probit models, too; see Wooldridge (2002, Chapter 17). As in trying to account for endogenous explanatory variables in such models, merely inserting an estimated inverse Mills ratio inside, say, an exponential model, or probit model, or Tobit model. One can always base a test on a variable-addition approaches, but they cannot be shown to solve the selection problem.

A very similar issue arises when using Heckman's method to correct for attrition in panel data (when selection on observables does not hold). With attrition as an absorbing state, it is common to estimate models in first differences to remove additive heterogeneity, say

$$\Delta y_{it} = \Delta x_{it}\beta + \Delta u_{it}, t = 2, \ldots, T. \tag{4.5}$$

We assume $s_{it} = 1 \Rightarrow s_{ir} = 1, r < t$. Let $w_{it}$ be a set of variables that we always observe when $s_{i,t-1} = 1$ such that $w_{it}$ is a good predictor of selection – in a sense soon to be made precise. We model the selection in time period $t$ conditional on $s_{i,t=1} = 1$ as

$$s_{it} = 1[w_{it}\delta_t + v_{it} > 0] \tag{4.6}$$

$$v_{it}|(w_{it}, s_{i,t-1} = 1) \sim Normal(0, 1), t = 2, 3, \ldots, T. \tag{4.7}$$

Since attrition is an absorbing state, $s_{it-1} = 0$ implies $s_{it} = 0$. This leads to a probit model for $s_{it}$ conditional on $s_{i,t-1} = 1$ :

$$P(s_{it} = 1|w_{it}, s_{i,t-1} = 1) = \Phi(w_{it}\delta_t), t = 2, \ldots, T. \tag{4.8}$$

Naturally, we need to estimate $\delta_t$, which we do as a sequence of probits. 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 $t = 2$. And so on. When we reach $t = T$, we have the smallest group of observations because we only use units still in the sample as of $T - 1$. Where might the $w_{it}$ come from? Since they have to be observed at time $t$ for the entire subgroup with $s_{i,t-1} = 1$, $w_{it}$ 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 $x_{it}$ are strictly exogenous, we can always include in $w_{it}$ elements of $(x_{i,t-1}, x_{i,t-2}, \ldots, x_{i1})$. Note that the potential dimension of $w_{it}$ grows as we move ahead through time. Unfortunately, $y_{i,t-1}$ cannot

be in $w_{it}$ because $y_{i,t-1}$ is necessarily correlated with $\Delta u_{it}$. But, if we assume that

$$E(u_{it}|x_i,y_{i,t-1},\ldots,y_{i1},c_i) = 0, t = 2,\ldots,T, | \tag{4.9}$$

then elements from $(y_{i,t-2},y_{i,t-3},\ldots,y_{i1})$ can be in $w_{it}$. If we start with a model where $x_{it}$ is strictly exogenous, as in standard panel data models, assumption (4.9) is very strong because in such models $u_{it}$ tends to be serially correlated, and therefore correlated with lagged $y_{it}$ in general. Still, since we are allowing for $c_i$, it might be that the errors $\{u_{it}\}$ are serially uncorrelated.

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

$$(\Delta u_{it}, v_{it}) \text{ is independent of } (\Delta x_{it}, w_{it}). \tag{4.10}$$

Now, $\Delta u_{it}$ is independent of $(\Delta x_{it}, w_{it})$ holds if $w_{it}$ contains only lags of $x_{it}$ because we assume $x_{it}$ is strictly exogenous. Unfortunately, $v_{it}$ is independent of $(\Delta x_{it}, w_{it})$ can be very restrictive because $\Delta x_{it}$ cannot be included in $w_{it}$ in interesting cases (because $x_{it}$ is not observed for everyone with $s_{i,t-1} = 1$). Therefore, when we apply a sequential Heckman method, we must omit at least some of the explanatory variables in the first-stage probits. If attrition is largely determined by changes in the covariates (which we do not see for everyone), using pooled OLS on the FD will be consistenty, whereas the Heckman correction would actually cause inconsistency.

As in the cross section case, we can "solve" this problem by using instrumental variables for any elements of $\Delta x_{it}$ not observed at time $t$. Assume sequential exogeneity, that is

$$E(u_{it}|x_{it},x_{i,t-1},\ldots,x_{i1},c_i) = 0, t = 1,\ldots,T. \tag{4.11}$$

(Recall that this condition does allow for lagged dependent variables in $x_{it}$). We now replace (4.10) with

$$(\Delta u_{it}, v_{it}) \text{ is independent of } (z_{it}, w_{it}) \tag{4.12}$$

conditional on $s_{i,t-1} = 1$. Choosing $z_{it}$ to be a subset of $w_{it}$ is attractive, because then (4.12) $E(\Delta u_{it}|z_{it},w_{it},v_{it},s_{i,t-1} = 1) = E(\Delta u_{it}|w_{it},v_{it},s_{i,t-1} = 1)$, in which case (4.12) holds if $(\Delta u_{it},v_{it})$ is independent of $w_{it}$ given $s_{i,t-1} = 1$. Then, after a sequence of probits (where, in each time period, we use observations on all units available in the previous time periods), we can apply pooled 2SLS, say, on the selected sample, to the equation

$$\Delta y_{it} = \Delta x_{it}\beta + \rho_2 d2_t\hat{\lambda}_{it} + \rho_3 d3_t\hat{\lambda}_{it} + \ldots + \rho_T dT_t\hat{\lambda}_{it} + error_{it}. \tag{4.13}$$

with instruments $(z_{it}, d2_t \hat{\lambda}_{it}, d3_t \hat{\lambda}_{it}, \ldots, dT_t \hat{\lambda}_{it})$. Because $\hat{\lambda}_{it}$ depends on $w_{it}$, it is critical to have an element in $w_{it}$ moving around selection separately from its correlation with $\Delta x_{it}$.

One can also test and correct for selection bias for any pattern of missing data on the response variable (or, generally, on endogenous explanatory variables). The key is that data are always observed on variables taken to be strictly exogenous, conditional on unobserved heterogeneity. Semykina and Wooldridge (2006) work through the details for the model

$$y_{it} = x_{it}\beta + c_i + u_{it} \tag{4.14}$$
$$E(u_{it}|z_i, c_i) = 0,$$

where $z_i = (z_{i1}, \ldots, z_{iT})$, so that some elements of $x_{it}$ are possibly endogenous, but the instruments, $z_{it}$, are strictly exogenous but allowed to be correlated with $c_i$. A simple test for correlation between $s_{it}$ and the idiosyncratic error – which, recall from Section 1, causes inconsistency in the FE-IV estimator, is available using Heckman's approach. In the first stage, estimate a pooled probit, or separate probit models, on $z_{it}$ and, say, the time averages, $\bar{z}_i$. Obtain estimated inverse Mills ratios. Then, estimate the equation

$$y_{it} = x_{it}\beta + \rho\hat{\lambda}_{it} + c_i + error_{it} \tag{4.15}$$

by FEIV, and use a standard (but robust) test of $\rho = 0$. This allows for endogeneity of $x_{it}$ under $H_0$, and so is a pure selection bias test. Or, the $\hat{\lambda}_{it}$ can be interacted with year dummies. The usefulness of this test is that it maintains only $E(E(u_{it}|z_i, s_i, c_i) = 0$ under $H_0$. Unfortunately, as a correction procedure, it generally does not lead to consistent estimators. (See Semykina and Wooldridge (2006).) As it turns out, a procedure that does produce consistent estimates under certain assumptions is just to add the time-average of the instruments, $\bar{z}_i$, to (4.15) and use pooled IV, where $\bar{z}_i$ and $\hat{\lambda}_{it}$ act as their own instruments.

## References

(To be added.)