Supplementary technical appendix to "Inverse probability tilting for moment condition models with missing data": computational details on Monte Carlo experiments reported in the supplemental web appendix of the main paper

Daniel Egel, Bryan S. Graham, Cristine Campos de Xavier Pinto

May 9, 2011

1 Introduction

In the supplemental web appendix of the paper we compare the small sample properties of inverse probability tilting (IPT) with six leading missing data estimators:

1. Parametric imputation (PI) as in, for example, Rubin (1977)
2. Parametric inverse probability weighting (IPW) as in, for example, Wooldridge (2007)
3. The parametric augmented inverse probability weighting (AIPW) estimator of Robins, Rotnitzky and Zhao (1994)
4. The nonparametric IPW estimator of Hirano, Imbens and Ridder (2003) (HIR)
5. The nonparametric imputation estimator of Imbens, Newey and Ridder (2005) (INR)
6. The conditional expectation projection estimator of Chen, Hong and Tarozzi (2008) (CHT)

This document describes our implementation of each of these estimators and serves as an informal guide to the replication code we have placed online. The code used for the empirical application and the second set of 'higher-order' Monte Carlos is described in a separate documents.

For each estimator we outline the following:

1. Computation of the estimator (including procedure for estimating standard errors)
2. Calculation of the probability limit of the estimator
3. Calculation of the asymptotic 'standard error' of the estimator (i.e., the square root of its asymptotic variance divided by the square root of the sample size)

For the first goal we attempted to be as faithful as possible to the cited published sources for each estimator.

For the last two goals we followed the general recipe of treating each estimator as a (possibly sequential) GMM estimator. That is we solved its population estimating equations under the data generating process (DGP) corresponding to each design. This often involved analyzing an estimator under misspecification. We used standard GMM results to characterize the asymptotic variance of each estimator, again with each term in the standard formula evaluated under the design-specific DGP. The results of these calculations, which are of the (computer assisted) 'pencil and paper' type, appear in Columns 1 and 3 in Tables 5 to 8 in the supplemental web appendix. Here we provide the population estimating equations that were solved for the probability limits of each estimator and the form of the asymptotic sampling variances. These last two calculations were only performed for the first three parametric estimators (as well as IPT) since the remaining estimators are asymptotically unbiased with variances equal to the variance bound of Robins, Rotnitzky and Zhao (1994) and Hahn (1998).

Unless noted otherwise, all notation is as defined in the main text. In all cases we describe estimation for the case where the goal is to estimate the mean of a scalar 'outcome' variable that is missing at random (MAR). All references cited below are included in the bibliography of either the main paper or the supplemental web appendix.

2 Parametric imputation (PI)

Let

$$Y = t(X)'\Upsilon_{*} + U, \quad \mathbb{E}\left[UX\right] = 0,$$

where $t(X) = \left(1, h\left(X\right)'\right)'$ and $\Upsilon = \left(\varsigma, \Pi'\right)'$.

In step one we compute

$$\widehat{\Upsilon} = \left[\frac{1}{N}\sum\nolimits_{i=1}^{N} D_i t(X_i)t(X_i)'\right]^{-1} \times \left[\frac{1}{N}\sum\nolimits_{i=1}^{N} D_i t(X_i)Y_i\right],$$

while in a second step we calculate

$$\widehat{\gamma}_{PI} = \frac{1}{N}\sum\nolimits_{i=1}^{N} t(X_i)'\widehat{\Upsilon}.$$

This estimator is a sequential GMM estimator based on the pair of moment restrictions

$$\mathbb{E}\left[m\left(Z, \Upsilon_{*}, \gamma_{*}\right)\right] = \mathbb{E}\left[\left(\begin{array}{c} m_1\left(Z, \Upsilon_{*}\right) \\ m_2\left(Z, \Upsilon_{*}, \gamma_{*}\right) \end{array}\right)\right] = \mathbb{E}\left[\left(\begin{array}{c} Dt(X)\left(Y - t(X)'\Upsilon_{*}\right) \\ t(X)'\Upsilon_{*} - \gamma_{*} \end{array}\right)\right] = 0,$$

where a '*' subscript denotes the population solution and is used instead of a '0' subscript to emphasize the possibility of misspecification (i.e, that $t(X)'\Upsilon_{*}$ may not be the conditional mean of Y given X).

The Jacobian matrix is

$$M = \left[\begin{array}{cc} M_{1\Upsilon} & 0 \\ M_{2\Upsilon} & M_{2\gamma} \end{array}\right],$$

where, recalling that $\zeta = \mathbb{E}\left[h\left(X\right)\right]$,

$$M_{1\Upsilon} = \mathbb{E}\left[p\left(X\right)t(X)t(X)'\right], \quad M_{2\Upsilon} = \left(1, \zeta'\right)', \quad M_{2\gamma} = -1.$$

The covariance of the moment function is given by

$$\Omega = \left[\begin{array}{cc} \Omega_{11} & \Omega_{12} \\ \Omega_{12}' & \Omega_{22} \end{array}\right],$$

where

$$\Omega_{11} = \mathbb{E}\left[p\left(X\right)\left(Y - t(X)'\Upsilon_{*}\right)^2 t(X)t(X)'\right]$$
$$\Omega_{12} = \mathbb{E}\left[p\left(X\right)t(X)\left(Y - t(X)'\Upsilon_{*}\right)\left(t(X)'\Upsilon_{*} - \gamma_{*}\right)\right], \quad \Omega_{22} = \Upsilon_{*}'\mathbb{V}(t\left(X\right))\Upsilon_{*}.$$

Note that under correct specification we have $\Omega_{12} = 0$. The asymptotic variance V_{PI} is given by the bottom right element of $(M'\Omega^{-1}M)^{-1}$.

Estimated standard errors are computed using the above GMM formulation, replacing population quantities with the usual analog estimates (e.g., Newey and McFadden, 1994).

3 Inverse probability weighting (IPW)

Let $p(x, \widehat{\alpha})$ be a, possible misspecified (pseudo) maximum likelihood estimate of the propensity score. The IPW estimator solves

$$\frac{1}{N}\sum_{i=1}^{N}\frac{2D_i}{p(X_i,\widehat{\alpha})}\left(Y_i-\widehat{\gamma}_{IPW}\right)=0.$$

We estimate the sampling variance of $\widehat{\gamma}_{IPW}$ using the procedure suggested by Wooldridge (2008). Define $K=(k_1,\ldots,k_N)'$ and $L=(l_1,\ldots,l_N)'$ where

$$\begin{aligned} k_i &= -\frac{2D_i}{p(X_i,\widehat{\alpha})}\left(Y_i-\widehat{\gamma}_{IPW}\right) \\ l_i &= \frac{\partial p\left(X_i,\widehat{\alpha}\right)}{\partial\alpha}\left(\frac{D_i}{p\left(X_i,\widehat{\alpha}\right)}-\frac{1-D_i}{1-p\left(X_i,\widehat{\alpha}\right)}\right). \end{aligned}$$

The *estimated* sampling variance of $\widehat{\gamma}_{IPW}$ is given by

$$\widehat{V}_{IPW}=\widehat{A}^{-1}\widehat{B}\widehat{A}^{-1}$$

where

$$\begin{aligned} \widehat{A} &= \frac{1}{N}\sum_{i=1}^{N}\frac{2D_i}{p(X_i,\widehat{\alpha})} \\ \widehat{B} &= \frac{1}{N}(K'K)-\frac{1}{N}(K'L)(L'L)^{-1}\left(L'K\right). \end{aligned}$$

To solve for the probability limit of $\widehat{\gamma}_{IPW}$ we find the solution to

$$\mathbb{E}\left[\frac{D}{p(X,\alpha_*)}\left(Y-\gamma_*\right)\right]=0,$$

where α_* are the pseudo-true propensity score coefficients.

To calculate the IPW asymptotic sampling variance we again treat the estimator as a sequential GMM estimator. We start with the first order representation,

$$\sqrt{N}\left(\widehat{\alpha}-\alpha_*\right)=\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\Psi\left(Z_i,\alpha_*\right)+o_p\left(1\right)$$

where

$$\Psi\left(Z_i,\alpha_*\right)=-\mathbb{E}\left[\frac{\partial^2\log\left[f\left(\left.D\right|Z,\alpha_*\right)\right]}{\partial\alpha\partial\alpha'}\right]^{-1}\times\frac{\partial\log\left[f\left(\left.D_i\right|Z_i,\alpha_*\right)\right]}{\partial\alpha},$$

with $\ln f\left(\left.D_i\right|Z_i,\alpha_*\right)$ the i^{th} unit's contribution to the pseudo propensity score log-likelihood.

Notice that $\widehat{\gamma}_{IPW}$ solves

$$\frac{1}{N}\sum_{i=1}^{N}\frac{D_i}{p(X_i,\widehat{\alpha})}\left(Y_i-\widehat{\gamma}_{IPW}\right)=0,$$

so that if we use the mean value theorem to expand the left hand side around γ_* and α_*, we can show that

$$\sqrt{N}\left(\widehat{\gamma}_{IPW}-\gamma_*\right)=\mathbb{E}\left[\frac{D}{p\left(X,\alpha_*\right)}\right]^{-1}\times\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\Lambda\left(Z_i,\alpha_*,\gamma_*\right)+o_p\left(1\right),$$

where

$$\Lambda\left(Z_i,\alpha_*,\gamma_*\right)=\frac{D_i}{p\left(X,\alpha_*\right)}\left(Y_i-\gamma_*\right)+\mathbb{E}\left[\frac{D}{p\left(X,\alpha_*\right)^2}\frac{\partial p(X,\alpha_*)}{\partial\alpha}\left(Y-\gamma_*\right)\right]\times\Psi\left(Z_i,\alpha_*\right).$$

Using this first order representation of $\widehat{\gamma}_{IPW}$ we have

$$\sqrt{N}\left(\widehat{\gamma}_{IPW}-\gamma_0\right)\xrightarrow{D}\mathcal{N}\left(0,V_{IPW}\right),$$

where

$$V_{IPW}=\mathbb{E}\left[\frac{D}{p\left(X,\alpha_*\right)}\right]^{-1}\times\mathbb{E}\left[\Lambda\left(Z,\alpha_*,\gamma_*\right)\Lambda\left(Z,\alpha_*,\gamma_*\right)'\right]\times\mathbb{E}\left[\frac{D}{p\left(X,\alpha_*\right)}\right]^{-1}.$$

4 Augmented Inverse Probability Weighting (AIPW)

Our implementation of the AIPW estimator of Robins, Rotnitzky and Zhao (1994) consists of three steps:

Step 1: Calculate $p\left(X,\widehat{\alpha}\right)$ by (pseudo) maximum likelihood

Step 2: Estimate $.q(X,\widehat{\lambda},\gamma)=\widetilde{q}(X,\widehat{\lambda})-\gamma$ by least squares where

$$\widetilde{q}(X,\widehat{\lambda})=t(X_i)'\widehat{\lambda},\qquad\widehat{\lambda}=\left[\frac{1}{N}\sum\nolimits_{i=1}^{N}D_it(X_i)t(X_i)'\right]^{-1}\times\left[\frac{1}{N}\sum\nolimits_{i=1}^{N}D_it(X_i)Y_i\right].$$

Step 3: Find $\hat{\gamma}_{AIPW}$ satisfying

$$\frac{1}{N}\sum_{i=1}^{N}\left\{\frac{D_i}{p\left(X_i,\widehat{\alpha}\right)}\left(Y_i-\widehat{\gamma}_{AIPW}\right)-\frac{D_i-p\left(X_i,\widehat{\alpha}\right)}{p\left(X_i,\widehat{\alpha}\right)}\left(\widetilde{q}(X_i,\widehat{\lambda})-\widehat{\gamma}_{AIPW}\right)\right\}=0,$$

which is equivalent to

$$\widehat{\gamma}_{AIPW}=\frac{1}{N}\sum_{i=1}^{N}\frac{D_iY_i-\widetilde{q}(X_i,\widehat{\lambda})\left(D_i-p\left(X_i,\widehat{\alpha}\right)\right)}{p\left(X_i,\widehat{\alpha}\right)}.$$

We estimate the sampling variance of $\widehat{\gamma}_{AIPW}$ in each of two ways in the paper. The first estimator, which assumes that both the propensity score and CEF model are correct, is given by

$$\widehat{V}_{AIPW,1}=\frac{1}{N}\sum_{i=1}^{N}\left\{\frac{D_iY_i-\widetilde{q}(X_i,\widehat{\lambda})\left(D_i-p\left(X_i,\widehat{\alpha}\right)\right)}{p\left(X_i,\widehat{\alpha}\right)}-\widehat{\gamma}_{AIPW}\right\}^2.$$

The second estimator follows from standard results on sequential GMM. Note that each estimation step corresponds to solving the following three sample moment conditions

$$\frac{1}{N}\sum_{i=1}^{N} m_1(Z_i,\widehat{\alpha}) = \frac{1}{N}\sum_{i=1}^{N}\left\{\frac{D_i - p\left(X_i,\widehat{\alpha}\right)}{\left[1-p\left(X_i,\widehat{\alpha}\right)\right]p\left(X_i,\widehat{\alpha}\right)}\right\}\frac{\partial p(X_i,\widehat{\alpha})}{\partial \alpha} = 0$$

$$\frac{1}{N}\sum_{i=1}^{N} m_2(Z_i,\widehat{\alpha},\widehat{\lambda}) = \frac{1}{N}\sum_{i=1}^{N} D_i t(X_i)\left(Y_i - t(X_i)'\widehat{\lambda}\right) = 0$$

$$\frac{1}{N}\sum_{i=1}^{N} m_3(Z_i,\widehat{\alpha},\widehat{\lambda},\widehat{\gamma}) = \frac{1}{N}\sum_{i=1}^{N}\frac{D_i(Y_i-\widehat{\gamma})}{p\left(X_i,\widehat{\alpha}\right)} - \frac{\left[\widetilde{q}(X_i,\widehat{\lambda})-\widehat{\gamma}\right]'}{p\left(X_i,\widehat{\alpha}\right)}\left(D_i - p\left(X_i,\widehat{\alpha}\right)\right) = 0.$$

Defining $\theta = (\alpha, \lambda, \gamma)$ and

$$m(Z,\theta) = \left[\begin{array}{c} m_1(Z,\alpha) \\ m_2(Z,\alpha,\lambda) \\ m_3(Z,\alpha,\lambda,\gamma) \end{array}\right]$$

then the GMM variance estimator, $\widehat{V}_{AIPW,2}$, is equal to the lower-right-hand element of

$$= \widehat{M}^{-1}\widehat{\Lambda}\widehat{M}^{-1\prime}$$

where

$$\widehat{M} = \frac{1}{N}\sum_{i=1}^{N}\left.\frac{\partial m(Z_i,\widehat{\theta})}{\partial \theta}\right|_{\theta=\widehat{\theta}} \qquad \textit{and} \qquad \widehat{\Lambda} = \frac{1}{N}\sum_{i=1}^{N} m(Z_i,\widehat{\theta})m(Z_i,\widehat{\theta})'.$$

We use the GMM sequential representation to calculate the probability limit of $\widehat{\gamma}_{AIPW}$ for each design as well as its asymptotic standard error.

4.1 Hirano, Imbens and Ridder (2003)

For computation of the Hirano, Imbens and Ridder (2003) point estimates and standard errors we used our IPW program (see above). However we let the propensity score take the series logit form discussed by Hirano, Imbens and Ridder (2003) with the number of series terms, K, less than

$$K < N^{1/9},$$

where N is the sample size.

In the case of $N = 1,000$, which is the sample size that we used in our Monte Carlo experiments, this restriction would imply the use of less than 2.15 parameters including a constant. However, we rounded this number up so that estimation was done using three terms: a constant, X and X^2.

The HIR estimator is asymptotically (first order) unbiased with a sampling variance equal to the bound.

4.2 Imbens, Newey and Ridder (2005)

Following the notation of Imbens, Newey and Ridder (2005) as much as possible we let $r_t(X) = X^{t-1}$ and construct the $1 \times K$ vector for each observation

$$R_K(X_i) = (r_1(X_i), \ldots, r_K(X_i)),$$

so that $R_K(X_i)$ is a power series of degree K for observation i. Using only the N_1 observations with $D_i = 1$ we now define the $N_1 \times K$ matrix R_{K,N_1} as

$$R_{K,N_1} = \begin{pmatrix} R_K(X_1) \\ \vdots \\ R_K(X_{N_1}) \end{pmatrix}.$$

and similarly define Y_{N_1} as the $N_1 \times 1$ vector of the values of Y for these N_1 observations. We estimate $\hat{\tau}_K$ according to[1]

$$\hat{\tau}_K = \left(R'_{K,N_1} R_{K,N_1}\right)^{-1} R'_{K,N_1} Y_{N_1}.$$

$\hat{\tau}_K$ is then used to impute the counterfactual ATE over the complete set of observations:

$$\hat{\gamma}_K = \frac{1}{N} \sum_{i=1}^{N} R_K(X_i) \hat{\tau}_K.$$

We choose K to minimize estimated mean squared error (MSE):

$$\text{MSE} = \frac{1}{N} \left[\sigma_1^2 a'_{N_1} M_{K,N_1} a_{N_1} + \left(a'_{N_1} e_{N_1}\right)^2 - \sigma_1^2 a'_{N_1} A_{N_1} a_{N_1}\right],$$

where

$$\begin{aligned}
\sigma_1^2 &= \tfrac{e'_{N_1} e_{N_1}}{N-1} \\
e_{N_1} &= Y_{N_1} - X_{N_1} \left(X'_{N_1} X_{N_1}\right)^{-1} X'_{N_1} Y_{N_1} \\
M_{K_{N_1}} &= R_{K,N_1} \left(R'_{K,N_1} R_{K,N_1}\right)^{-1} R_{K,N_1} \\
A_{N_1} &= I - M_{K_{N_1}} \\
a_{N_1} &= \begin{bmatrix} \frac{1}{p(X_1, \widehat{\alpha})} \\ \vdots \\ \frac{1}{p\left(X_{N_1}, \widehat{\alpha}\right)} \end{bmatrix},
\end{aligned}$$

with $p\left(X_i, \widehat{\alpha}\right)$ equal to a parametric estimate of the propensity score (specifically a logit with an index linear in X).

1. Specifically we calculate $\hat{\gamma}_K$ for all possible positive integer values of K such that the minimum eigenvalue of $\left(R'_{K,N_1} R_{K,N_1}\right)$ is greater than $\frac{1}{2}$. The value of K that minimizes the MSE, K^*, is then selected as the optimal value so that

$$\hat{\gamma}_{INR} = \hat{\gamma}_{K^*}.$$

[1] Note that we are using a slight simplification of the procedure suggested by Imbens, Newey and Ridder (2005). While we use R_{K,N_1} directly in estimation the authors propose using instead $R_K(X_i) = \Omega_K^{-1/2} R_K(X_i)$ where $\Omega_K = \mathbb{E}[R_K(X) R_K(X)' | D = 1]$. While this normalization is useful for demonstrating the asymptotic properties of this estimator, it does not affect the point estimates.

An asymptotic variance estimator for $\hat{\gamma}_{INR}$ is not provided in the most recent draft of the Imbens, Newey and Ridder (2005) paper. Accordingly we estimate the variance of $\hat{\gamma}_{INR}$ by the sample variance of an estimate of the efficient influence function

$$\widehat{V}_{INR} = \frac{1}{N}\sum_{i=1}^{N}\left\{[\widehat{\mu}(X_i)]^2 + \left[\frac{D_i}{p\left(X_i,\widehat{\alpha}\right)}\left(Y_i - \widehat{\mu}(X_i)\right)\right]^2\right\},$$

where

$$\widehat{\mu}(X_i) = R_K(X_i)(R'_{K,N_1}R_{K,N_1})R'_{K,N}Y,$$

and we use the fact that the two parts of the influence function are asymptotically uncorrelated and also that $\gamma_0 = 0$.

The INR estimator is asymptotically (first order) unbiased with a sampling variance equal to the bound.

4.3 Chen, Hong and Tarozzi (2004, 2008)

Following the notation of Chen, Hong and Tarozzi (2004, 2008) as much as possible we, for some K, define the $K \times 1$ vector $q^K(X)$ as

$$q^K(X) = \left(q_1(X), \ldots, q_K(X)\right)',$$

where $\{q_l(X), l = 1, 2, \ldots\}$ denotes a sequence of basis functions (with K the degree of the sequence).[2] Using this vector we now define a $N_1 \times K$ matrix, Q_{N_1}, using all N_1 observations where $D = 1$

$$Q_{N_1} = \left(q^K(X_1), \ldots, q^K(X_{N_1})\right)'.$$

The first step of the estimation procedure involves computing

$$\widehat{\mathcal{E}}(X;\gamma) = \sum_{j=1}^{N} D_j \cdot (Y_j - \gamma)\, q^K(X_j) \cdot (Q'_{N_1}Q_{N_1})^{-1} \cdot q^K(X).$$

In the second step $\hat{\gamma}_{CHT}$ is chosen to minimize

$$\hat{\gamma}_{CHT} = \min_{\gamma \in \mathcal{G}} \left(\frac{1}{N}\sum_{i=1}^{N}\mathcal{E}(X_i,\beta)\right)'\left(\frac{1}{N}\sum_{i=1}^{N}\mathcal{E}(X_i,\beta)\right).$$

We estimate the variance of $\hat{\gamma}_{CHT}$ by

$$\widehat{V}_{CHT} = \frac{1}{N}\sum_{i=1}^{N}\left(\widehat{\mathcal{E}}(X_i;\hat{\gamma}_{CHT}) + \frac{D_i}{p\left(X_i,\widehat{\alpha}\right)}\left\{(Y_i - \hat{\gamma}_{CHT}) - \widehat{\mathcal{E}}(X_i;\hat{\gamma}_{CHT})\right\}\right)^2,$$

with $p\left(X_i,\widehat{\alpha}\right)$ equal to a parametric estimate of the propensity score (specifically a logit with an index linear in X).

The CHT estimator is asymptotically (first order) unbiased with a sampling variance equal to the bound.

[2] Typically this includes a constant as one of the terms.

5 Matlab Code

Accompanying this document are several Matlab programs that implement each of these estimators.

The code is organized so that each of the above estimators, and our IPT estimator, are in a separate folder. The folder 'DGP' contains files that implement the data generating process that we use for the Monte Carlo experiments. Finally, in the root directory there is a file called 'main.m' that implements the Monte Carlo experiment.[3]

Although we do not provide a description of each program in this document, it should be clear from the headers within each program what they do and how they are related to other programs.

[3] Note that the random number kernel has been set so that the results from the paper will be obtained if the program is run without any changes.