Robustness to Parametric Assumptions in Missing Data Models

By BRYAN S. GRAHAM AND KEISUKE HIRANO*

Draft: 17 January 2011

Suppose we have a random sample from a population of interest. For each sampled unit we observe the covariate X, which we assume is discrete with support $\{x_1, \ldots, x_K\}$. For some units, we also observe the variable Y. Let $D = 1$ if we observe Y, and $D = 0$ otherwise. We are interested in the population mean of Y, $\theta = \mathbb{E}[Y] = \sum_{k=1}^{K} p_k \mu_k$, where $\mu_k = \mathbb{E}[Y|X = x_k]$ and $p_k = \Pr(X = x_k)$.

We assume that Y is missing at random (MAR): $Y \perp D|X$. Suppose also that the propensity score $e_k = \Pr(D = 1|X = x_k)$ is bounded away from zero (a support condition). Then in large samples, there will be at least some units with Y observed for each possible value of X, so that $\mathbb{E}[Y|X = x_k, D = 1]$ is identified. Since $\mu_k = \mathbb{E}[Y|X = x_k, D = 1]$ under MAR, we have

$$\theta = \sum_{k=1}^{K} p_k \mathbb{E}[Y|X = x_k, D = 1].$$

Let M_k equal the number of sampled units with $X = x_k$ (i.e., in cell k), and let $\hat{p}_k = \left[\sum_{j=1}^{K} M_j\right]^{-1} M_k$. The poststratification estimator for θ is

$$\hat{\theta}_{PS} = \sum_{k=1}^{K} \hat{p}_k \overline{Y}_k,$$

where $\overline{Y}_k$ is the average of Y across those units with $D = 1$ and $X = x_k$ (i.e., the complete-case k cell mean or the sample analog of $\mathbb{E}[Y|X = x_k, D = 1]$).

When M_k is large for all $k = 1, \ldots, K$ the poststratification estimator $\hat{\theta}_{PS}$ works well in practice and attains the semiparametric variance bound for θ derived by Jinyong Hahn (1998). Unfortunately in many applications it is common for K to be large and M_k to be small (at least for some values of k). In such settings the problem of empty cells, where Y is unavailable for all sampled units with $X = x_k$, may be severe (Paul R. Rosenbaum 1987).

In settings with small cells, there may be substantial gains from imposing restrictions on the means μ_k, but there is also a danger of misspecification. We explore ways to increase the robustness of parametric imputation estimators. First, we develop a simple empirical Bayes estimator, which combines parametric and unadjusted estimates of μ_k in a data-driven way. Second, we consider ways to use knowledge of the propensity score to help guard against misspecification of μ_k, using a double robust estimator and an empirical Bayes approach. This does not contradict the efficiency bound analysis of Hahn (1998), which is relevant for settings where M_k is large for all k.

I. Sampling framework and estimators

Following Joshua Angrist and Jinyong Hahn (2004) we consider a stratified random sampling scheme. Let N be the total sample size with M_k chosen such that $M_k/N = p_k$ for all k (i.e., we assume that p_k, which characterizes the marginal distribution of X, is known). Within each cell, the probability of observing the outcome Y is e_k, so that the number of observed outcomes is $n_k \sim \text{Binomial}(e_k, M_k)$.

Conditional on n_k, the observed outcomes $Y_{k1}, \ldots, Y_{kn_k}$ are i.i.d. (and independent across cells) with mean μ_k and variance σ_k^2.

The poststratification estimator for θ, modi-

* Graham: Department of Economics, New York University, bryan.graham@nyu.edu. Hirano: Department of Economics, University of Arizona, hirano@u.arizona.edu. Hirano acknowledges support from the National Science Foundation under grant SES-0962488.

fied to take into account that the p_k are known, is

$$\hat{\theta}_{PS} = \sum_{k=1}^{K} p_k \overline{Y}_k,$$

where $\overline{Y}_k = n_k^{-1} \sum_{i=1}^{n_k} Y_{ki}$. The nonparametric imputation estimator of Hahn (1998), and the estimated propensity score weighting estimator of Keisuke Hirano, Guido W. Imbens and Geert Ridder (2003) (modified appropriately for the missing data problem considered here), are both equal to $\hat{\theta}_{PS}$ in the discrete covariate case. This estimator may perform poorly if some cells have a small number of complete observations. If some cells are empty (i.e., $n_k = 0$), then the estimator must be modified, for example by dropping empty cells or combining cells in some way.

An alternative is to posit a restricted model for the cell means:

$$\mu_k = x_k' \beta, \tag{1}$$

where β is a low-dimensional parameter. (We could also easily handle specifications of the form $\mu_k = t(x_k)'\beta$ for a known function t.) Then

$$\mathbb{E}[\overline{Y}_k | n_1, \ldots, n_K] = x_k' \beta,$$
$$\mathbb{V}(\overline{Y}_k | n_1, \ldots, n_K) = \sigma_k^2 / n_k,$$

and, conditional on $(n_1, \ldots, n_K)$, the $(\overline{Y}_1, \ldots, \overline{Y}_K)$ will be mutually independent. We could estimate β by a weighted least squares (WLS) regression of the $\overline{Y}_k$ on x_k, with weights proportional to n_k. (This is equivalent to an ordinary least squares (OLS) regression of all the observed Y_{ki} on $X_{ki} = x_k$.) Then the parametric imputation estimator is

$$\hat{\theta}_{PI} = \sum_{k=1}^{K} p_k (x_k' \hat{\beta}).$$

The parametric estimator would typically do better when the assumption on the means (1) holds, and could be used even if some cells are empty. However, if (1) does not hold, then $\hat{\theta}_{PI}$ may be severely biased. Our goal is to develop estimators that improve upon the poststratification estimator when cell sizes are small, but are not as sensitive to misspecification as the parametric imputation estimator.

Following Carl N. Morris (1983) and Gary Chamberlain (2009), we consider an empirical Bayes approach. (See also David S. Lee and David Card (2008) for a closely related approach.) In the following statements, we implicitly condition on $n_1, \ldots, n_K$. Suppose that

$$\overline{Y}_k | \mu_k \overset{\text{ind}}{\sim} \mathcal{N}(\mu_k, v_k), \quad k = 1, \ldots, K,$$

where $v_k = \sigma_k^2 / n_k$, and

$$\mu_k \overset{\text{ind}}{\sim} \mathcal{N}(x_k' \beta, \tau^2), \quad k = 1, \ldots, K.$$

This reduces to (1) when $\tau^2 = 0$. Under this setup the marginal distribution of the cell averages $\overline{Y}_1, \ldots, \overline{Y}_k$ is

$$\overline{Y}_k \overset{\text{ind}}{\sim} \mathcal{N}(x_k' \beta, v_k + \tau^2). \tag{2}$$

Let $\gamma_k = v_k / (v_k + \tau^2)$. The posterior for μ_k, treating β, v_k, and τ^2 as known, is

$$\mu_k \overset{\text{ind}}{\sim} N(\mu_k^*, v_k(1 - \gamma_k)),$$

where

$$\mu_k^* = (1 - \gamma_k) \overline{Y}_k + \gamma_k (x_k' \beta).$$

This suggests the (infeasible) estimator

$$\hat{\theta}_{EB0} = \sum_{k=1}^{K} p_k \left[(1 - \gamma_k) \overline{Y}_k + \gamma_k (x_k' \beta) \right].$$

To construct a feasible version, let $\hat{\beta}$ be the least squares estimator as in the imputation estimator. Let $\hat{\tau}^2$ be the pseudo maximum likelihood estimator of τ in (2), taking as given the regression estimates[1] $\hat{\beta}$ and the following estimates of the v_k:

$$\hat{v}_k = \frac{\hat{\sigma}_k^2}{n_k},$$

where the $\hat{\sigma}_k^2$ are the within-cell sample variances of the y_{ki}. We then form

$$\hat{\gamma}_k = \frac{\hat{v}_k}{\hat{v}_k + \hat{\tau}_k^2},$$

[1] We could estimate β and τ jointly by pseudo maximum likelihood, but for our extensions below, this form is somewhat more convenient. Another alternative is to carry out full Bayesian hierarchical inference.

and

$$\hat{\theta}_{EB1} = \sum_{k=1}^{K} p_k \left[(1 - \hat{\gamma}_k)\overline{Y}_k + \hat{\gamma}_k (x_k' \hat{\beta}) \right].$$

Although we motivated the estimator by a Gaussian hierarchical model, the estimator has a number of appealing properties that do not depend on normality. When cell sizes M_k are large, so that n_k is also large when the support condition holds, the $\hat{\gamma}_k$ will be close to zero, and the estimator will be similar to the poststratification estimator $\hat{\theta}_{PS}$. On the other hand, if the parametric model is close to being correct, and $\hat{\tau}^2$ is close to zero, the estimator will be similar to the parametric imputation estimator.

However, for intermediate values of $\hat{\gamma}_k$, the estimator is not a simple weighted average of $\hat{\theta}_{PS}$ and $\hat{\theta}_{PI}$. Instead, within each cell we take a weighted average of $\overline{Y}_k$ and $x_k'\hat{\beta}_k$, with the weights depending on the value of $\hat{\tau}^2$ and on the $\hat{v}_k$. Thus the estimator is similar to a kernel-type smoothing estimator with an adaptive bandwidth: when n_k is large $\hat{\theta}_{EB1}$ typically places more weight on the nonparametric estimate $\overline{Y}_k$ relative to the parametric estimate $x_k'\hat{\beta}_k$.

The estimator needs to be modified in order to deal with empty or nearly empty cells. If $n_k = 0$, then $\overline{Y}_k$ is not defined. In that case, we set $\hat{\gamma}_k = 1$, so that the estimator uses the parametric model to impute the cell mean. If $n_k = 1$, then the variance estimate $\hat{\sigma}_k^2 = 0$. For such cells we instead use the average estimated variance among the cells with $n_k \geq 2$ in order to obtain the shrinkage term $\hat{\gamma}_k$. The parameter τ^2 is estimated using only the cells with $n_k \geq 2$.

II. Double robustness

James Robins and coauthors have proposed an alternative approach to robustifying estimators based on parametric mean restrictions. In the double robust (DR) approach, the empirical researcher posits a model for the means, and a model for the propensity score (in our notation, the e_k). A DR estimator is one that is consistent for the parameters of interest provided at least one of the two parametric restrictions are satisfied. Various DR estimators have been proposed, including James M. Robins, Andrea Rotnitzky and Lue Ping Zhao (1994), Keisuke Hirano and Guido W. Imbens (2001), Heejung Bang and James M. Robins (2005), Jeffrey M. Wooldridge (2007), Weihua Cao, Anastasios A. Tsiatis and Marie Davidian (2009), and Bryan S. Graham, Christine Campos de Xavier Pinto and Daniel Egel (2010).

Suppose we have two possible parametric restrictions:

ASSUMPTION DR1: *$\mu_k = x_k'\beta$ for all k.*

ASSUMPTION DR2: *$e_k = G(x_k)$ for all k, where G is a known function.*

Bang and Robins (2005) show that a DR estimator can be constructed by augmenting a regression with the inverse of the (parametric) propensity score. In our setup, this can be implemented through the following weighted linear projection problem: choose α_1^*, α_2^* to solve

(3)
$$\min_{\alpha_1,\alpha_2} \sum_{k=1}^{K} p_k e_k \mathbb{E}\left[\left(\overline{Y}_k - x_k'\alpha_1 - G^{-1}(x_k)\alpha_2 \right)^2 \right].$$

The results in Bang and Robins (2005) imply the following result, which we prove for completeness:

PROPOSITION 1: *If DR1, DR2, or both hold, then*

$$\theta = \sum_{k=1}^{K} p_k \left[x_k'\alpha_1^* + G^{-1}(x_k)\alpha_2^* \right].$$

PROOF:

The minimization problem (3) is equivalent to the problem

(4)
$$\min_{\alpha_1,\alpha_2} \sum_{k=1}^{K} p_k e_k \left(\mu_k - x_k'\alpha_1 - G^{-1}(x_k)\alpha_2 \right)^2.$$

First, suppose DR1 holds. Then clearly (4) is solved by setting $\alpha_1^* = \beta$ and $\alpha_2^* = 0$. Then

$$\sum_{k=1}^{K} p_k \left[x_k'\alpha_1^* + G^{-1}(x_k)\alpha_2^* \right] = \sum_{k=1}^{K} p_k \left[x_k'\beta \right] = \sum_{k=1}^{K} p_k \mu_k = \theta.$$

Next, suppose DR2 holds. The first order conditions for (4) implies

$$\sum_{k=1}^{K} p_k \frac{e_k}{G(x_k)} \left(\mu_k - x_k'\alpha_1^* - G^{-1}(x_k)\alpha_2^* \right) = 0.$$

Hence if $e_k = G(x_k)$ for all k,

$$\sum_{k=1}^{K} p_k \mu_k = \sum\nolimits_{k=1}^{K} p_k \left[x_k' \alpha_1^* + G^{-1}(x_k)\alpha_2^* \right].$$

To construct a feasible version of this estimator, let

$$\hat{e}_k = \frac{n_k}{M_k}.$$

Then $p_k \hat{e}_k \propto n_k$, so we could solve

$$\min_{\alpha_1, \alpha_2} \sum_{k=1}^{K} n_k \left(\overline{Y}_k - x_k' \alpha_1 - G^{-1}(x_k)\alpha_2 \right)^2.$$

This is WLS of $\overline{Y}_k$ on $(x_k', G^{-1}(x_k))'$, with weights proportional to n_k, and is equivalent to OLS of the observed Y_{ki} on $(X_{ki}', G^{-1}(X_{ki}))'$. Let $\hat{\alpha}_1$ and $\hat{\alpha}_2$ denote these estimates. The Bang and Robins DR estimator is

$$\hat{\theta}_{DR} = \sum_{k=1}^{K} p_k \left[x_k' \hat{\alpha}_1 + G^{-1}(x_k)\hat{\alpha}_2 \right].$$

An empirical Bayes extension can be based on the marginal model

$$\overline{Y}_k \overset{\text{ind}}{\sim} N(x_k'\alpha_1 + G^{-1}(x_k)\alpha_2, v_k + \tau^2).$$

We can form the empirical Bayes estimate exactly as before, after augmenting the regressor vector with the term $G^{-1}(x_k)$. Note however that under Assumption DR2, we will *not* necessarily have $\tau^2 = 0$. This suggests that it may be useful to consider alternative estimators for τ^2, which shrink the estimate to zero when the data indicate that the propensity score restriction is close to being satisfied. We defer such extensions to future work.

III. Monte Carlo study

We carry out a simple simulation study to compare the various estimators. Suppose the covariate cells are

$$\{x_1, \ldots, x_K\} = \{-J, \ldots, 0, \ldots, J\},$$

so that $K = 2J + 1$, and $M_k = M$ for all k, so that $p_k = 1/K$. We specify $\mu_k = x_k \beta$, which implies that $\theta = 0$. The propensity score is

$$e_k = \begin{cases} 0.75 & \text{if } x_k < 0 \\ (.5K - .75J)/(K - J) & \text{if } x_k \geq 0 \end{cases}$$

This gives an overall probability of $1/2$ of observing the outcome. The outcomes Y_{ki} are independently drawn from a normal distribution with mean μ_k and variance σ^2 (the variance is constant across cells). Under this model, the variance bound for estimating θ is

$$VB = \sum_{k=1}^{K} p_k \frac{\sigma^2}{e_k}.$$

(See Theorem 5 of Xiaohong Chen, Han Hong and Alessandro Tarozzi (2004) and Section 5.2 of Guido W. Imbens and Jeffrey M. Wooldridge (2009).)

We consider six designs with J chosen such that $K = 5, 15, 25, 75, 125$, and 375; $N = 3000$ across all designs such that the common cell sizes are $M = 600, 200, 120, 40, 24$ and 8. We choose σ^2 based on K and M to set $VB = 30$. This implies that an efficient estimator should have a standard deviation of 0.1 in large samples. We also choose β so that the variance of μ_k is equal to 30 in each design. For each design we perform 1000 Monte Carlo replications.

We apply the estimators developed above under two parametric specifications (where required). In the first, μ_k is correctly assumed to be linear in x_k. In the second, μ_k is erroneously assumed to be constant over x_k. To conserve space we only report the latter sets of results in detail.

These results are reported in Table 1. Each row of the Table corresponds to an estimator, with columns denoting the different designs. The entries show mean bias for each estimator/design as well as its standard deviation across Monte Carlo replications (in parentheses).

The sampling distribution of the post-stratification estimator $\hat{\theta}_{PS}$ is well approximated by conventional asymptotic approximations for designs where the number cells K is small and cell size M is reasonably large. However, when $K = 375$ (so that $M = 8$), the presence of empty cells induces substantial bias and inflates variance.

Not surprisingly, the parametric imputation, double robust, and empirical Bayes estimators all perform well when they incorporate a correctly-specified conditional mean model (results not shown). When the conditional mean model is incorrect, as in Table 1, their properties diverge. The parametric imputation estimator is biased when it is based on an incorrectly specified conditional mean model. The double robust estimator exhibits low bias. Although in our experiments this estimator is also based on an incorrect conditional mean model, it does utilize the true propensity score. It sampling distribution, however, is relatively more dispersed than that of the post stratification estimator. The empirical Bayes estimator moderately outperforms the parametric imputation estimator across all designs. However for K large / M small it also exhibits substantial bias. Incorporating the true propensity score into the marginal model eliminates this bias. Importantly, the sampling distribution of this estimator is less dispersed than that of the double robust estimator with a standard deviation 15 to 20 percent smaller.

IV. Conclusion

In many applications the number of discrete covariate cells is large relative to the sample size. In such situations many cells may contain few, or even no, observations of the outcome of interest Y. Using a parametric model to impute cell means is one approach to solving this empty cell problem. We have outlined an alternative approach to estimating cell means and associated population average parameters. In cells with many observed outcomes our approach is nonparametric; in cells with few such observations it is essentially parametric; while in intermediate cases we combine a nonparametric and parametric estimate of the cell mean. Incorporating the propensity score into our parametric imputation model appears to help guard against misspecification.

In further work it would be useful to explore other ways of choosing the γ_k and to formally characterize the large sample properties of our estimator. Of particular interest are asymptotic sequences which allow K to grow with N, as first suggested by Angrist and Hahn (2004).

REFERENCES

Angrist, Joshua, and Jinyong Hahn. 2004. "When to Control for Covariates? Panel Asymptotics for Estimates of Treatment Effects." *Review of Economics and Statistics*, 86(1): 58–72.

Bang, Heejung, and James M. Robins. 2005. "Doubly Robust Estimation in Missing Data and Causal Inference Models." *Biometrics*, 61(4): 962–973.

Cao, Weihua, Anastasios A. Tsiatis, and Marie Davidian. 2009. "Improving Efficiency and Robustness of the Doubly Robust Estimator for a Population Mean with Incomplete Data." *Biometrika*, 96(3): 723–734.

Chamberlain, Gary. 2009. "Bayesian Aspects of Treatment Choice." Working paper, Harvard University.

Chen, Xiaohong, Han Hong, and Alessandro Tarozzi. 2004. "Semiparametric Efficiency in GMM Models of Nonclassical Measurement Errors, Missing Data and Treatment Effects." Cowles Foundation Discussion Paper 1644.

Graham, Bryan S., Christine Campos de Xavier Pinto, and Daniel Egel. 2010. "Inverse Probability Tilting for Moment Condition Models with Missing Data." Working paper, NYU.

Hahn, Jinyong. 1998. "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects." *Econometrica*, 66(2): 315–331.

Hirano, Keisuke, and Guido W. Imbens. 2001. "Estimation of Causal Effects using Propensity Score Weighting: An Application to Data on Right Heart Catheterization." *Health Services & Outcomes Research Methodology*, 2: 259–278.

Hirano, Keisuke, Guido W. Imbens, and Geert Ridder. 2003. "Efficient Estimation of Average Treatment Effects using the Estimated Propensity Score." *Econometrica*, 71(4): 1161–1189.

Imbens, Guido W., and Jeffrey M. Wooldridge. 2009. "Recent Developments in the Econometrics of Program Evaluation." *Journal of Economic Literature*, 47(1): 5–86.

Lee, David S., and David Card. 2008. "Regression Discontinuity Inference with Specification

K	5	15	25	75	125	375
$\hat{\theta}_{PS}$	-0.0043	0.0037	-0.0014	0.0016	0.0032	-0.2471
	(0.0996)	(0.1023)	(0.0994)	(0.1030)	(0.1018)	(0.1243)
$\hat{\theta}_{PI}$	-1.9401	-2.2162	-2.2840	-2.2360	-2.3528	-2.3641
	(0.1359)	(0.1317)	(0.1336)	(0.1294)	(0.1283)	(0.1292)
$\hat{\theta}_{DR}$	-0.0054	0.0059	-0.0041	0.0017	0.0045	0.0018
	(0.1212)	(0.1188)	(0.1196)	(0.1196)	(0.1170)	(0.1187)
$\hat{\theta}_{EB1}$	-0.0096	-0.0145	-0.0325	-0.0886	-0.1394	-0.6982
	(0.0997)	(0.1025)	(0.0997)	(0.1031)	(0.1226)	(0.3131)
$\hat{\theta}_{EB2}$	-0.0043	0.0038	-0.0011	0.0017	0.0037	0.0007
	(0.0997)	(0.1023)	(0.0996)	(0.1028)	(0.1016)	(0.1101)

TABLE 1—MONTE CARLO RESULTS FOR INCORRECTLY SPECIFIED CONDITIONAL MEAN MODEL

Error." *Journal of Econometrics*, 142(2): 655–674.

Morris, Carl N. 1983. "Parametric Empirical Bayes Inference: Theory and Applications." *Journal of the American Statistical Association*, 78(381): 47–55.

Robins, James M., Andrea Rotnitzky, and Lue Ping Zhao. 1994. "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed." *Journal of the American Statistical Association*, 89(427): 846–856.

Rosenbaum, Paul R. 1987. "Model-based Direct Adjustment." *Journal of the American Statistical Association*, 82(398): 387–394.

Wooldridge, Jeffrey M. 2007. "Inverse Probability Weighted Estimation for General Missing Data Problems." *Journal of Econometrics*, 141(2): 1281–1301.