

# Robustness to Parametric Assumptions in Missing Data Models

Bryan Graham  
NYU

Keisuke Hirano  
University of Arizona

April 2011

# Motivation

- We consider the classic missing data problem.
- In practice covariate dimension is often very high, and conventional asymptotics may be misleading.
- We consider a finite-sample setting where correct parametric specifications make a big difference.
- Adopt the Angrist and Hahn (2004) setup, where cells can be small or even empty.

# Outline

- Missing Data Model
- Stratified Sampling Setup
- Parametric Imputation
- Empirical Bayes
- Double Robustness and EB
- Monte Carlo
- Some work in progress

# MAR

- Random sample from a population
- Observe a covariate  $X$ , and if  $D = 1$ , observe  $Y$ .
- Interested in population mean of  $Y$ :

$$\theta = E[Y].$$

- Assume  $Y$  is missing at random (MAR):

$$Y \perp D | X.$$

- We are especially interested in cases where cells are small.

Let

$$\mu(x) = E[Y|X = x]$$

$$\sigma^2(x) = V[Y|X = x]$$

$$e(x) = Pr(D = 1|X = x) \quad \text{bounded away from 0.}$$

Semiparametric efficiency bound for estimating  $\theta$   
(Hahn, 1998):

$$VB = E [(\mu(X) - \theta)^2] + E \left[ \frac{\sigma^2(X)}{e(X)} \right].$$

Various efficient estimators proposed:

- Hahn (1998)
- Hirano, Imbens, Ridder (2003)
- Chen, Hong, Tarozzi (2008)

Typically involve NP estimation of  $\mu(x)$ ,  $e(x)$ , or both.

Typically identical if  $X$  discrete.

## Discrete Covariate/Small Cell Model

Following Angrist and Hahn (2004):

- Covariate takes values in  $\{x_1, \dots, x_K\}$ .
- $M_k$  individuals in each cell — fixed and known (stratified sampling). Let

$$p_k = \frac{M_k}{\sum_{j=1}^K M_j} \quad (\text{distribution of } X)$$

- Propensity score: in cell  $k$ , observe  $Y$  with probability  $e_k > 0$ .
- $n_k$  observed outcomes per cell, with

$$n_k \sim \text{Binomial}(e_k, M_k).$$

- In cell  $k$ , let

$$Y_{k1}, \dots, Y_{kn_k} \stackrel{\text{iid}}{\sim} (\mu_k, \sigma_k^2).$$

- Estimand:

$$\theta = \sum_{k=1}^K p_k \mu_k.$$

# Poststratification Estimator

- Let:

$$\bar{Y}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} Y_{ki}.$$

- Estimator:

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

- May work poorly if  $n_k$  small.
- For empty cells: drop cells (adjust  $p_k$ ), or combine cells.

# Parametric Imputation

- Suppose

$$\mu_k = x'_k \beta,$$

where  $\beta$  is a low-dimensional parameter.

- Then

$$E[\bar{Y}_k | n_1, \dots, n_K] = x'_k \beta,$$

$$V[\bar{Y}_k | n_1, \dots, n_K] = \sigma_k^2 / n_k.$$

- Let  $\hat{\beta} = \text{WLS of } \bar{Y}_k \text{ on } x_k \text{ (= OLS of underlying } y\text{'s on } x)$
- Parametric Imputation Estimator:

$$\hat{\theta}_{PI} = \sum_{k=1}^K p_k \{x'_k \hat{\beta}\}.$$

- Can handle empty cells, and will work well when model is correct.
- But may be sensitive to parametric specification.

# Empirical Bayes

- Based on Morris (1983) and Chamberlain (2009)
- Suppose (temporarily) that

$$\mu_k \sim N(x_k' \beta, \tau^2).$$

Note  $\tau^2 = 0$  corresponds to parametric imputation.

- Further suppose

$$\bar{Y}_k | \mu_k \sim N(\mu_k, v_k),$$

where  $v_k = \sigma_k^2 / n_k$ .

Marginal for cell means:

$$\bar{Y}_k \sim N(x'_k \beta, v_k + \tau^2).$$

Then posterior mean of  $\mu_k$  is

$$\mu_k^* = (1 - \gamma_k) \bar{Y}_k + \gamma_k (x'_k \beta),$$

where

$$\gamma_k = \frac{v_k}{v_k + \tau^2}.$$

Replace  $\beta, v_k, \tau^2$  with estimates, form

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

and

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

Small cells: if  $n_k = 0$  set  $\hat{\gamma}_k = 1$ , modify variance estimate if  $n_k = 1$ .

## Qualitative features of EB

- If all  $\hat{\gamma}_k$  near 1 (e.g.  $\hat{\tau}^2 \approx 0$ ), similar to parametric imputation
- If all  $\hat{\gamma}_k$  near 0 (e.g. if all  $n_k$  large), similar to poststratification
- More generally, weighting varies by cell based on  $v_k$  and  $\tau^2$ . Similar to adaptive bandwidth kernel.

## Double Robustness and EB

- **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.
- Double robustness: estimator is consistent if *one or both* DR1, DR2 hold.
- Protection against misspecification of  $\mu_k$ , *provided*  $G(x_k)$  is correct (and vice versa).

Bang and Robins (2005): can augment regression with inverse of propensity score.

Let  $\alpha_1^*, \alpha_2^*$  solve

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

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

Equivalent minimization problem:

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

If DR1 holds, (2) is solved by setting  $\alpha_1^* = \beta$  and  $\alpha_2^* = 0$ .  
Then

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

## Proof cont'd

If DR2 holds, first order conditions for (2) implies

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

Then since  $e_k = G(x_k)$ ,

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

□

Feasible version:

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

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

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

This is WLS of  $\bar{Y}_k$  on  $(x'_k, G^{-1}(x_k))'$ , with weights proportional to  $n_k$ .

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

## EB Extension of DR

- $\hat{\theta}_{DR}$  is parametric imputation estimator with an additional regressor.
- So we can develop a corresponding empirical Bayes extension  $\hat{\theta}_{EB2}$ .

## EB Extension of DR

- $\hat{\theta}_{DR}$  is parametric imputation estimator with an additional regressor.
- So we can develop a corresponding empirical Bayes extension  $\hat{\theta}_{EB2}$ .
- Triply robust?

## EB Extension of DR

- $\hat{\theta}_{DR}$  is parametric imputation estimator with an additional regressor.
- So we can develop a corresponding empirical Bayes extension  $\hat{\theta}_{EB2}$ .
- Triply robust?
- Caution: DR2 does not imply  $\tau^2 = 0$ . Could try to engineer alternative estimators of  $\gamma_k$ . (But see Monte Carlo evidence below.)

## Monte Carlo study

- Support of  $X$ :  $\{-J, \dots, 0, \dots, J\}$ .
- Equal size strata:  $M_k = M$  for all  $k$ .
- $\mu_k = x_k \beta$  (implies  $\theta = 0$ ).
- Propensity score: step function,  $e_k = .75$  if  $x_k < 0$ .  
Overall 1/2 probability of selection.
- $Y_{ki}$  iid  $N(\mu_k, \sigma^2)$ .

- Consider various values of  $J, M$  with overall sample size 3000.
- Choose  $\sigma^2$  so that (large sample) efficient estimator should have  $SE = .1$ .
- Specification for  $\mu_k$ : correct model and incorrect model (constant mean)
- Propensity score correctly specified

## Bias, Correct Parametric Mean

K	PS	PI	DR	EB	EBPS
5	-0.0043	-0.0047	-0.0043	-0.0047	-0.0043
15	0.0037	0.0021	0.0036	0.0022	0.0037
25	-0.0014	-0.0025	-0.0021	-0.0025	-0.0020
75	0.0016	-0.0003	0.0012	0.0002	0.0015
125	0.0032	0.0037	0.0033	0.0038	0.0034
375	<b>-0.2471</b>	0.0039	0.0032	0.0027	0.0024

$K = \#$  of cells

For  $K=375$ , median of 18 empty cells

## RMSE, Correct Parametric Mean

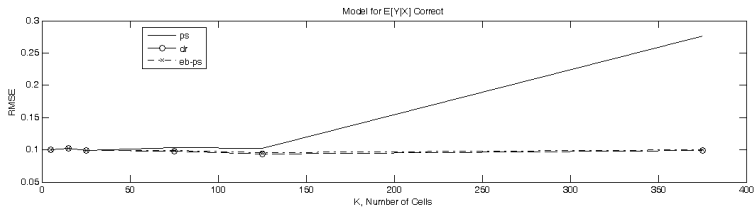
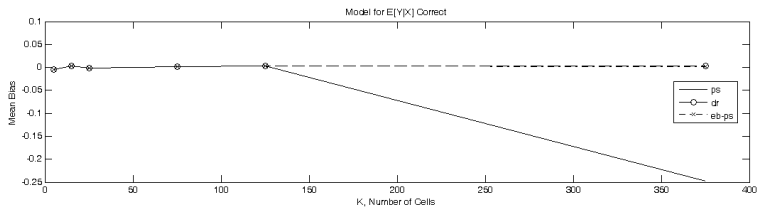
K	PS	PI	DR	EB	EBPS
5	0.0997	0.0981	0.0996	0.0980	0.0996
15	0.1023	0.0994	0.1022	0.0995	0.1022
25	0.0993	0.0960	0.0988	0.0961	0.0989
75	0.1030	0.0952	0.0982	0.0956	0.0985
125	0.1018	0.0902	0.0931	0.0926	0.0957
375	0.2766	0.0950	0.0990	0.0968	0.1004

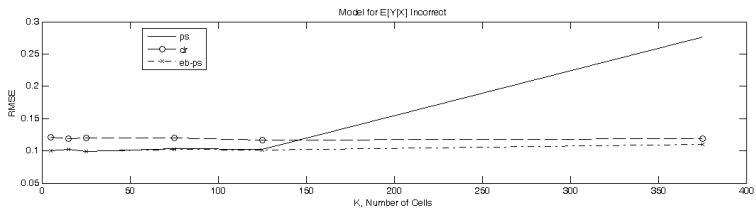
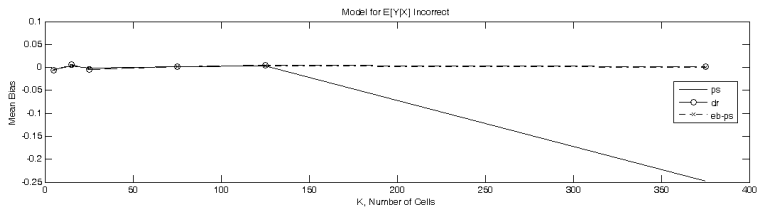
## Bias, Misspecified Mean

K	PS	PI	DR	EB	EBPS
5	-0.0043	<b>-1.9401</b>	-0.0054	<b>-0.0096</b>	-0.0043
15	0.0037	-2.2162	0.0059	-0.0145	0.0038
25	-0.0014	-2.2840	-0.0041	-0.0325	-0.0011
75	0.0016	-2.3360	0.0017	-0.0886	0.0017
125	0.0032	-2.3528	0.0045	-0.1394	0.0037
375	-0.2471	-2.3641	0.0018	-0.6982	0.0007

## RMSE, Misspecified Parametric Mean

K	PS	PI	DR	EB	EBPS
5	0.0997	1.9449	0.1213	0.1001	0.0998
15	0.1023	2.2201	0.1189	0.1034	0.1023
25	0.0993	2.2879	0.1196	0.1049	0.0995
75	0.1030	2.3395	0.1196	0.1360	0.1028
125	0.1018	2.3563	0.1170	0.1856	0.1016
375	0.2766	2.3677	0.1186	0.7651	0.1101





## Remarks

- PS is poor when cells are small.
- EB is similar to PI when parametric mean correctly specified.
- With misspecified mean but correct prop score:
  - PI poor, EB better
  - DR good, but EBPS even better – shrinking  $\hat{\tau}$  all the way to zero would not have helped!
- In general, EB-type estimators should be (nearly) admissible, but make different risk tradeoffs over parameter space.

## Extensions and Ongoing Work

- Continuous covariates: Gaussian process priors.
- Inference? Some existing literature on EB inference.
- Finite-sample MSE calculations and finite-sample decision theory.
- Asymptotics / limit experiments which preserve small/zero cells?

## Limits of experiments (very very preliminary)

Suppose  $(X_i, D_i, Y_i)$  are IID with

$$X_i \sim P \quad \text{with density } p(x) \text{ on } \mathcal{X},$$

$$\Pr(D_i = 1 | X_i = x) = e(x),$$

$$Y_i | D_i, X_i = x \sim F(y | x, \pi).$$

Parameters are  $\theta = (e(\cdot), \pi)$ , and we think of  $e(x)$  “close to zero.”

# Hellinger Transform

- Let  $\mathcal{E} = \{F_\theta; \theta \in \Theta\}$  be an experiment.
- Let  $\alpha = \{\alpha_\theta; \theta \in \Theta\}$  satisfy
  - $\alpha_\theta \geq 0$
  - $\sum_\theta \alpha_\theta = 1$
  - only a finite number of the  $\alpha_\theta > 0$ .

Define

$$\eta(\alpha) = \int \prod_{\theta} (dF_{\theta})^{\alpha_{\theta}}.$$

## Useful properties of Hellinger transform of $\mathcal{E}$ :

- Pointwise convergence of  $\eta(\alpha)$  is equivalent to weak convergence of experiments in the sense of Le Cam.
- In particular, it implies an asymptotic representation theorem.
- Hellinger transform of a product experiment is product of Hellinger transforms.

For simplicity, drop  $Y$  component and focus on  $(X, D)$ .

Local parametrization:

$$Pr(D_i = 1 | X_i = x) = \frac{e(x)}{n}, \quad i = 1, \dots, n.$$

Single-observation density:

$$f(x, d) = p(x) \left( \frac{e(x)}{n} \right)^d \left( 1 - \frac{e(x)}{n} \right)^{1-d}$$

wrt dominating measure  $\lambda(x, d) = \lambda_d(d)\lambda_x(x)$ .

Let  $\alpha$  have  $G$  non-zero components  $\alpha_1, \dots, \alpha_G$ .

Hellinger transform (single-obs):

$$\begin{aligned}\eta(\alpha) &= \int \int \prod_{g=1}^G f(x, d)^{\alpha_g} d\lambda(x, d) \\ &= \int \left[ \prod_g \left( \frac{e(x)}{n} \right)^{\alpha_g} + \prod_g \left( 1 - \frac{e(x)}{n} \right)^{\alpha_g} \right] dP(x) \\ &= E \left[ \frac{\prod_g e(X)^{\alpha_g}}{n} \right] + E \left[ \prod_g \left( 1 - \frac{e(X)}{n} \right)^{\alpha_g} \right].\end{aligned}$$

For  $n$  observations:

$$\begin{aligned}
 \eta_n(\alpha) &= \{\eta(\alpha)\}^n \\
 &= \left\{ E \left[ \frac{\prod_g e(X)^{\alpha_g}}{n} \right] + E \left[ \prod_g \left( 1 - \frac{e(X)}{n} \right)^{\alpha_g} \right] \right\}^n \\
 &\approx \left\{ \frac{E \left[ \prod_g e(X)^{\alpha_g} \right]}{n} + 1 - \frac{E \left[ \sum_g e(X) \alpha_g \right]}{n} \right\}^n \\
 &\rightarrow \exp \left( E \left[ \prod_g e(X)^{\alpha_g} \right] - E \left[ \sum_g e(X) \alpha_g \right] \right).
 \end{aligned}$$

Now, consider a different experiment:

$Z$  is a Poisson point process on  $\mathcal{X}$ , with intensity measure  $\nu(x)$ , indexed by  $e = e(x)$ , where

$$d\nu_e(x) = e(x)p(x)d\lambda_x(x).$$

By results in Le Cam and Yang (2000): Hellinger transform is

$$\begin{aligned} \eta(\alpha) &= \exp \left( \int \left[ \prod_{g=1}^G (d\nu_{e_g})^{\alpha_g} - \sum_{g=1}^G \alpha_g d\nu_{e_g} \right] \right) \\ &= \exp \left( E \left[ \prod_g e(X)^{\alpha_g} \right] - E \left[ \sum_g \alpha_g e(X) \right] \right). \end{aligned}$$

Hence the Poisson process experiment characterizes what can be done asymptotically for the “small probabilities” model.

Full model with  $Y$ : Poisson process on  $\mathcal{X} \times \mathcal{Y}$ ?