Estimation and Imputation under Nonignorable Nonresponse

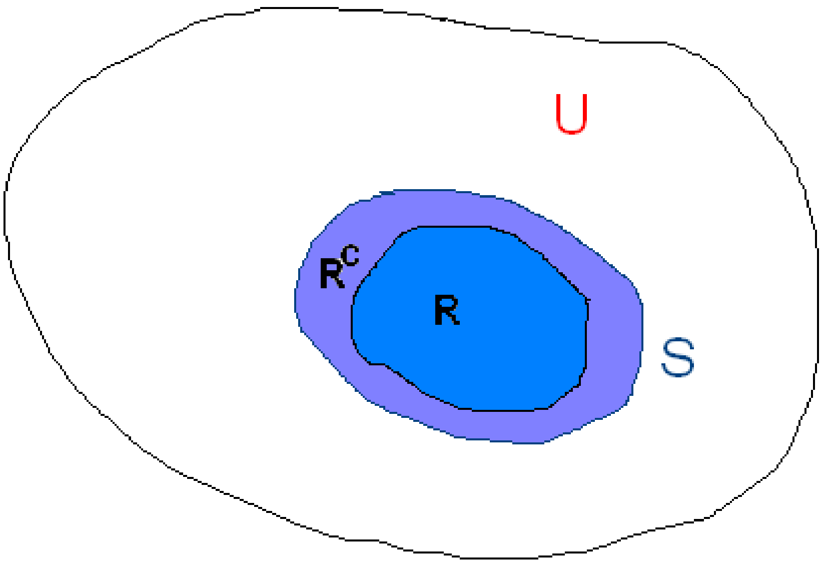
General set up

Population U with measurements $(Y_i, X_i = (X_{1i}, ..., X_{pi}))$.

Population outcomes are independent realizations from distribution with pdf $f_p(Y_i \mid X_i; \theta)$.

Sample S **of size** n **selected with known probabilities** $\pi_i = P(i \in S)$.

Subsample $R = \{1, ..., n_r\}$ **of Respondents with unknown probabilities to respond.**



Objectives of the research

Estimate unknown model parameters

Impute missing values

Estimate population means

Assumption: **the population and sample distributions of $Y_i \mid X_i$ are the same.**

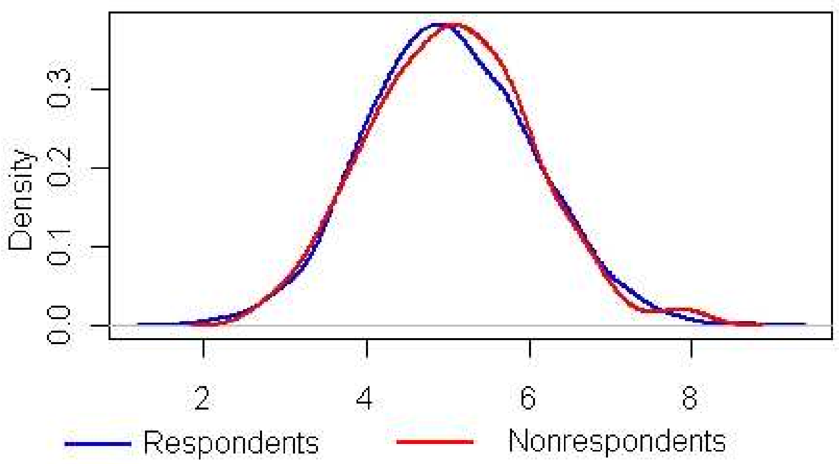
$$f_p(Y_i \mid X_i;\theta) = f(Y_i \mid X_i, i \in S;\theta) = f_S(Y_i \mid X_i;\theta)$$

Examples of Nonresponse

1. MCAR (Missing completely at random)

Y_1	$R_1 = 1$
Y_2	$R_2 = 1$
Y_r	$R_r = 1$
?	$R_{r+1} = 0$
?	
?	$R_n = 0$

$$f(Y_i \mid R_i = 1; \theta) = f(Y_i \mid R_i = 0; \theta) = f(Y_i; \theta)$$



Examples of Nonresponse (cont.)

Estimation:
based on the respondent's observations

Imputation:

1. $y_j^* = \overline{Y}_R = \frac{1}{r}\sum_{i=1}^{r} y_i$

2. Random draws from $\hat{f}(Y_i)$

Population mean estimator: $\overline{Y}_R$

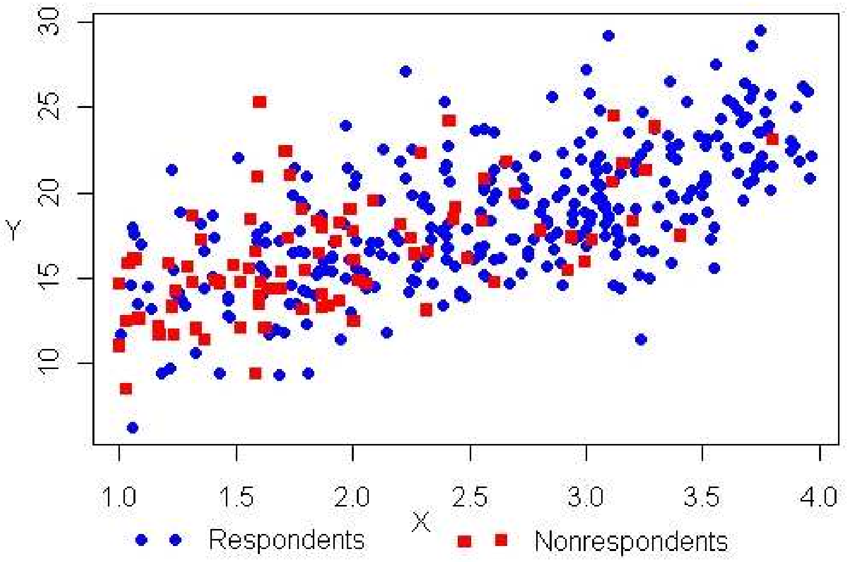
Examples of Nonresponse (cont.)

2. MAR (Missing at Random). Assume that

$$f(Y_i \mid X_i, R_i = 1; \theta) = f(Y_i \mid X_i, R_i = 0; \theta) = f(Y_i \mid X_i; \theta)$$

Y_1	X_1	$R_1 = 1$
Y_2	X_2	$R_2 = 1$
Y_r	X_r	$R_r = 1$
?	X_{r+1}	$R_{r+1} = 0$
?		
?	X_n	$R_n = 0$

Estimation:
based on the respondent's observations



Examples of Nonresponse (cont.)

If $Y_i \mid X_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$ **then**

Imputation:

1. $y_j^* = \hat{\beta}_0 + \hat{\beta}_1 X_j$

2. Random draws from $\hat{f}(Y_j \mid X_j)$

Population mean estimator: $\hat{\beta}_0 + \hat{\beta}_1 \overline{X}_R$

Examples of Nonresponse (cont.)

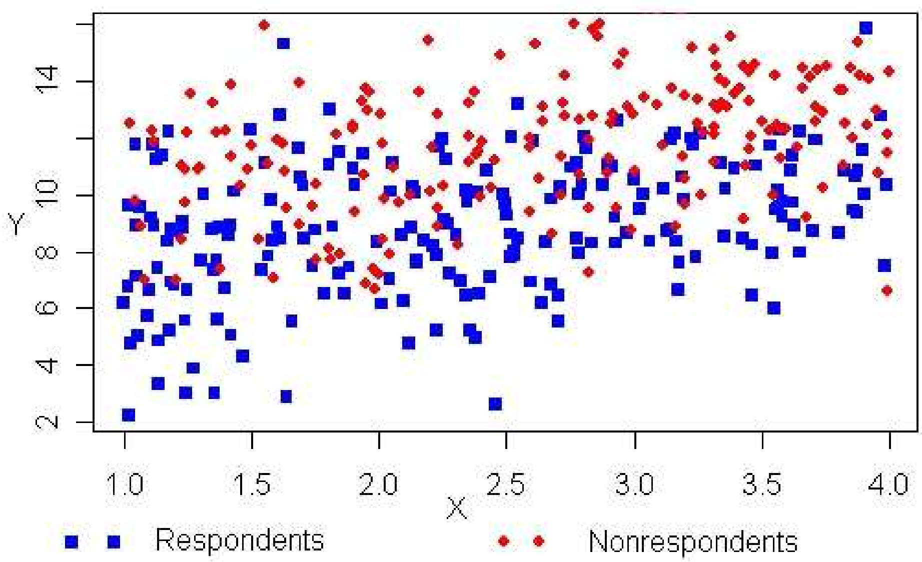
3. Not Missing at random (NMAR)

$$f_R(Y_i \mid X_i) = f(Y_i \mid X_i, i \in S, R_i = 1) = \frac{\Pr(R_i = 1 \mid Y_i, X_i, i \in S)}{\Pr(R_i = 1 \mid X_i, i \in S)} f_S(Y_i \mid X_i)$$

where

$$\Pr(R_i = 1 \mid X_i, i \in S) = \int f_S(Y_i \mid X_i) \Pr(R_i = 1 \mid Y_i, X_i, i \in S) dY_i$$

and $f_S(Y_i \mid X_i)$ **is the sample *pdf* under complete response. In this research we assume that the sample *pdf* and the population *pdf* are the same.**



Existing Approaches

Full likelihood (Selection models)

$$f(Y_i, R_i \mid X_i, \theta, \gamma) = \Pr(R_i \mid Y_i, X_i, \gamma) f_S(Y_i \mid X_i, \theta)$$

where

$f_S(Y_i \mid X_i, \theta)$ **defines the sample *pdf* (model);**

$\Pr(R_i \mid Y_i, X_i, \gamma)$ **models the response process;**

θ **and** γ **denote the unknown parameters of the two models respectively.**

The full likelihood:

$$L = \prod_{i=1}^{r} \Pr(R_i = 1 \mid Y_i, X_i; \gamma) f_S(Y_i \mid X_i; \theta) \prod_{i=r+1}^{n} \Pr(R_i = 0 \mid X_i; \theta, \gamma),$$

where

$$\Pr(R_i = 0 \mid X_i; \theta, \gamma) = 1 - \Pr(R_i = 1 \mid X_i; \theta, \gamma) = 1 - \int \Pr(R_i = 1 \mid Y_i, X_i; \gamma) f_S(Y_i \mid X_i; \theta) dY_i$$

Drawback: **knowledge of nonrespondents' covariates is required.**

Existing Approaches (cont.)

Greenlees *et al.* (1982) assume $f_S(Y_i \mid X_i, \theta)$ **normal and** $\Pr(R_i \mid Y_i, X_i, \gamma)$ **logistic.**

Beaumont (2000) drops normality assumption for the regression residuals. Requires knowledge of nonrespondents' covariates.

Tang *et al.* (2003) does not require parametric assumption on $\Pr(R_i \mid Y_i, X_i, \gamma)$**, but assumes that it is a function of** Y_i**. Requires knowledge of nonrespondents' covariates.**

Proposed approach

Distribution of the responding unit:

$$f_R(Y_i \mid X_i) = \frac{\Pr(R_i = 1 \mid Y_i, X_i, i \in S)}{\Pr(R_i = 1 \mid X_i, i \in S)} f_s(Y_i \mid X_i)$$

where

$$\Pr(R_i = 1 \mid X_i, i \in S) = \int f_s(Y_i \mid X_i) \Pr(R_i = 1 \mid Y_i, X_i, i \in S) dY_i$$

Respondents likelihood

$$L_{\text{Resp}} = \prod_{i=1}^{r} f(Y_i \mid X_i, R_i = 1, i \in S; \theta, \gamma) = \prod_{i=1}^{r} \frac{\Pr(R_i = 1 \mid Y_i, X_i, i \in S; \gamma) f_s(Y_i \mid X_i; \theta)}{\Pr(R_i = 1 \mid X_i, i \in S; \theta, \gamma)}$$

L_{resp} **does not require knowledge of covariates** X_i **of the nonresponding units or modeling of distribution of sampled** X_i**.**

Proposed approach (cont.)

Calibration constraints: assume that $X^{pop}=(X_1^{pop},\dots,X_p^{pop})$ **are known from a census or administrative records.**

Parameter γ **could be estimated from the equations which match the Horwitz-Thompson estimators for the population means to their known values, as follows:**

$$\sum_{i=1}^{r} w_i \frac{X_{ki}}{\pi(Y_i,X_i;\gamma)} = X_k^{pop}, k=1,\dots,q$$

where $\omega_i=\frac{1}{P(i\in S)}$ $,i=1,\dots,r$ **denotes the sampling weights,** $\pi(Y_i,X_i)=\Pr(R_i=1|Y_i,X_i,i\in S)$**.**

$$E\sum_{i=1}^{r}\frac{w_i X_{ki}}{\pi(Y_i,X_i;\gamma)} = E\sum_{i=1}^{N}\frac{w_i X_{ki} R_i I_i}{\pi(Y_i,X_i;\gamma)} = \sum_{i=1}^{N}\frac{w_i X_{ki}}{\pi(Y_i,X_i;\gamma)} ER_i I_i =$$

$$\sum_{i=1}^{N}\frac{w_i X_{ki}}{\pi(Y_i,X_i;\gamma)} EE(R_i I_i \mid I_i) = \sum_{i=1}^{N}\frac{w_i X_{ki}}{\pi(Y_i,X_i;\gamma)} E(I_i(P(R_i=1\mid i\in S))) =$$

$$\sum_{i=1}^{N}\frac{w_i X_{ki}}{\pi(Y_i,X_i;\gamma)} \pi(Y_i,X_i;\gamma)P(i\in S) = X_k^{pop}$$

Proposed approach (cont.)

Respondents Likelihood with calibration constraints

$$L_{\text{Resp}} = \prod_{i=1}^{r} f(Y_i \mid X_i, R_i = 1, i \in S; \theta, \gamma) = \prod_{i=1}^{r} \frac{\Pr(R_i = 1 \big| Y_i, X_i, i \in S; \gamma) f_s(Y_i \big| X_i; \theta)}{\Pr(R_i = 1 \big| X_i, i \in S; \theta, \gamma)}$$

If $\pi(Y_i, X_i; \gamma)$ **is a function of** $\gamma_0 + \gamma_1 X_{1i} + \ldots \gamma_q X_{qi} + \gamma_{q+1} Y_i$**,** γ **can be estimated from the equations:**

$$\sum_{i=1}^{r} w_i \frac{1}{\pi(Y_i, X_i; \gamma)} = N$$

$$\sum_{i=1}^{r} w_i \frac{X_{ki}}{\pi(Y_i, X_i; \gamma)} = X_k^{pop}, k = 1, \ldots, q$$

$$\sum_{i=1}^{r} w_i \frac{(Y_i - E_s(Y_i \mid X_i))}{\pi(Y_i, X_i; \gamma)} = 0$$

We propose to use the equations $l(\theta, \gamma) = \frac{\partial L_{resp}(\theta, \gamma)}{\partial \theta} = 0$ **and calibration constraints** $h(\theta, \gamma) = 0$ **in order to estimate unknown parameters** θ **and** γ**.**

Proposed approach (cont.)

The respondents' likelihood for Generalized Linear Sample Models (GLM):

$$f_s(Y_i \left| X_i ; \beta , \phi \right.) = e^{a(\phi)[Y_i \sum_{s=0}^{p} \beta_s X_{si} - g(\sum_{s=0}^{p} \beta_s X_{si}) + d(Y_i)] + \eta(\phi, Y_i)}$$

Taking the derivatives of the log-likelihood with respect to β and ϕ, we obtain the following equations:

$$\sum_{i=1}^{r} \left(Y_i - E_R \left(Y_i \mid X_i ; \beta, \phi, \gamma \right) \right) X_{ki} = 0,\ k = 0, \ldots, p$$

$$\sum_{i=1}^{r} \left(d(Y_i) - E_R \left(d(Y_i) \mid X_i ; \beta, \phi, \gamma \right) \right) = 0$$

Proposed approach (cont.)

Let $\theta^{(0)}$ denote initial values for the vector θ indexing the sample *pdf* $f_s(Y_i|X_i;\theta)$.

***Step j*: For given $\hat{\theta}^{(j)}$ from iteration *j*, set $\theta = \hat{\theta}^{(j)}$ and solve the calibration constraints $h(\theta,\gamma)=0$ as a function of the unknown parameters γ indexing the model $\pi(Y_i, X_i;\gamma)$ for the response probabilities. This step yields estimators $\hat{\gamma}^{(j+1)}$.**

***Step j+1*: Solve the equations $l(\theta,\gamma)=0$ with respect to θ, with γ equal to $\gamma^{(j+1)}$. This step yields new estimators $\hat{\theta}^{(j+1)}$.**

Continue the iterations until convergence.

Some theoretical properties

Theorem

Let $\hat{\xi}' = (\hat{\theta}', \hat{\gamma}')$ define the estimator obtained by application of the algorithm. Suppose that:

I) The population (sample) model belongs to the family of generalized linear models,

II) $0 < \pi(y_i, v_i; \gamma) < 1$, with bounded first derivatives with respect to γ.

III) The functions $l(\theta, \gamma)$ and $h(\theta, \gamma)$ are continuous and twice differentiable with respect to (θ, γ) in a compact neighborhood of the solution $\xi_0' = (\theta_0', \gamma_0')$.

IV) The matrices $\frac{\partial l(\theta,\gamma)}{\partial \theta}, \frac{\partial h(\theta,\gamma)}{\partial \gamma}$ are nonsingular in a neighborhood of the true vector parameter $\tilde{\xi} = (\tilde{\theta}', \tilde{\gamma}')'$.

Then, as $N \to \infty, n \to \infty$ such that $(N/n) < \infty$ the estimator $\hat{\xi}' = (\hat{\theta}', \hat{\gamma}')$ converges in probability to the solution $\xi_0' = (\theta_0', \gamma_0')$.

We show also that under some

regularity conditions the estimator

$\hat{\xi} = (\hat{\theta}', \hat{\gamma}')'$ **is consistent for** $\tilde{\xi}$

and

$$\sqrt{\boldsymbol{n}}(\hat{\xi} - \tilde{\xi}) \xrightarrow{D} N[0, V(\tilde{\xi})] \text{ !}$$

Imputation of missing values and Estimation of population totals

When covariates are unknown for nonrespondents

$$\hat{\bar{Y}}_{(1)} = \frac{1}{N}\sum_{i=1}^{r}\frac{w_i Y_i}{\hat{\pi}(Y_i, X_i)}$$

When covariates are known for nonrespondents

$\hat{\bar{Y}}_{(2)} = \frac{1}{N}\sum_{i=1}^{n} w_i Y_i^*$; $Y_i^* = Y_i$ **if** $i \in R$, $Y_i^* = Y_i^{imp}$ **if** $i \in R^c$.

Imputation of missing values and Estimation of population totals

If all covariates are observed, the imputed values, Y_i^{imp}, can be computed either as,

$$Y_i^{imp} = E_{R^C}(Y_i \mid X_i) = E(Y_i \mid X_i, i \in R^C),$$

or by generating at random observations from the conditional *pdf* $f_{R^C}(Y_i \mid X_i)$

$$f_{R^C}(Y_i \mid X_i) = f(Y_i \mid X_i, R_i = 0) = \frac{\Pr(R_i = 0 \mid Y_i, X_i, i \in S) f_s(Y_i \mid X_i)}{\Pr(R_i = 0 \mid X_i, i \in s)} =$$

$$\frac{[1 - \pi(Y_i, X_i)] f_s(Y_i \mid X_i)}{[1 - \pi(X_i)]}$$

Imputation of missing covariates

We assume instead that

$$\Pr(X_i = x_i \mid R_i = 1, i \in S) = \frac{1}{r} \qquad \forall x_i \in R$$

(equal probability of $1/r$ for each vector covariate observed for the responding units).

Under this assumption we can estimate $P_{X|0}(x_i)$ by the empirical probability function,

$$\hat{P}_{X|0}(x_i) = \Pr(X_i = x_i \mid R_i = 0, i \in S) = \frac{[1-\pi(x_i)]}{\pi(x_i)[\sum_1^r (1/\pi(X_i)) - r]}.$$

It can be easily shown that

$$\Pr(R_i = 1 \mid i \in S) = \frac{r}{\sum_1^r (1/\pi(X_i))},$$

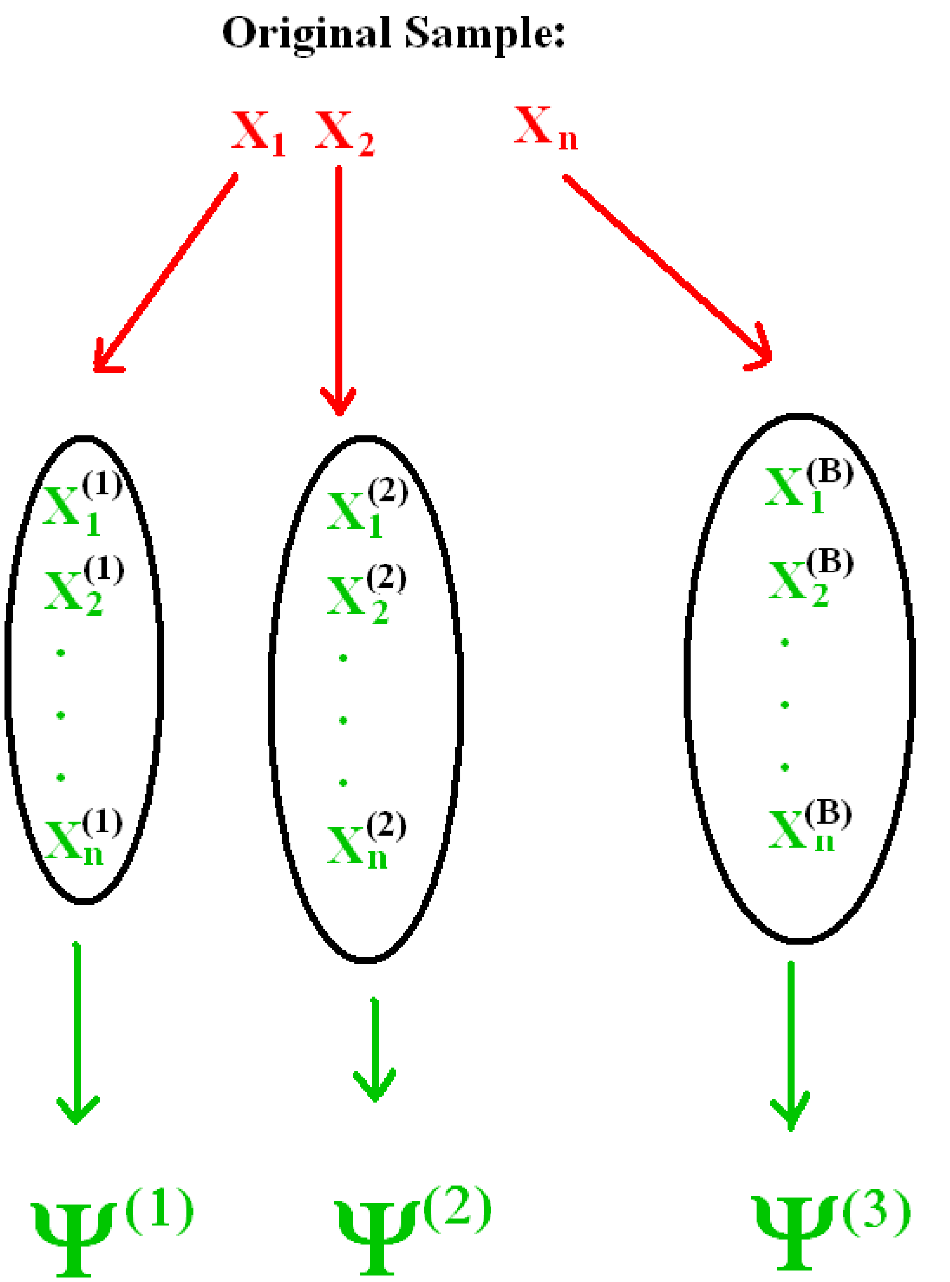
guaranteeing

$$\sum_{x_i} \Pr(X_i = x_i \mid R_i = 0, i \in S) = 1.$$

Calculation of Variance: Bootstrap

Suppose we have a sample $X_1, X_2 ..., X_n \sim F$ **and we wish to compute the variance of a statistic** $\psi(X_1, X_2 ..., X_n)$

When the theoretical distribution of a statistic of interest is complicated or unknown, Bootstrap allows estimation of the sample distribution of almost any statistic using only very simple methods.


Original Sample:
X1 X2 Xn
X1(1)
X2(1)
Xn(1)
X1(2)
X2(2)
Xn(2)
X1(B)
X2(B)
Xn(B)
Ψ(1)
Ψ(2)
Ψ(3)

Empirical study

Household Expenditure Survey in Israel, 2005.

Households (HH) sampled with equal probabilities by two-stage sampling, first sampling localities and then HH within the sampled localities. The 60 largest localities (out of 171) sampled with certainty. Remaining localities and HHs sampled systematically.

Target outcome variable:

"HH income per standard person".

Response rate:

Initially 37%. After several recalls 90%.

Covariates unknown for nonresponding HH after last recall.

Empirical study (cont)

In this study we restrict to HH where the head of the HH is an employee, aged 25-64, born in Israel and working in the last 3 months preceding the survey.

Responding HH: HH that responded to the first questionnaire.

Nonresponding HH: HH that responded on one of the recalls.

Data available for responding and nonresponding units. Sample size n=1717, Resp. HH r=629, Nonresp. HH n-r=1088.

Empirical study (cont.)

Sample model:

$$Y_i = X_i'\beta + \varepsilon_i, \ \ \varepsilon_i \sim N(0, \sigma_\varepsilon^2)$$

Response probabilities,

$$P(R_i = 1 \mid Y_i, X_i) = \frac{1}{1 + e^{-(\delta Y_i + X_i'\gamma)}},$$

where Y_i is the log income per standard person in household i and $X_i = \left[1, x_{i1}, \ldots, x_{ip}\right]'$ is the vector of covariates for the household.

Fitting the sample model with 17 covariates to all the sampled HH (n=1717) yields a good fit with

$R^2 = 0.6$

Empirical study (cont.)

Gender	Head of the household is female.
Age	Age of the head of the household
District1	Household located in Jerusalem, Tel-Aviv, Haifa, Ramat-Gan or Holon.
District2	Household located in Zefat, Kinneret, Akko, Emek Yizrael or the Golan heights.
District3	Household located in Hadera, Sharon or Petah Tiqwa.
District4	Household located in Ramla.
District5	Household located in Rehovot.
District6	Household located in Ashqelon or Be'er-Sheva.
District7	Household located in Yehuda or Shomron.
Hours	Number of monthly working hours of head of household
Earners	Number of earners.
HHsize	Number of standard persons in the household.
School10	Number of school years of head of the household is less than 10.
School12	Number of school years of head of the household is between 10 and 12.
School15	Number of school years of head of the household is more than 12 with nonacademic education.
School16	Number of school years of head of the household is more than 12 with academic education.
Occupation0	Head of household employed as academic professional.
Occupation1	Head of household employed as associate professional or technician.
Occupation2	Head of household employed as a manager.
Occupation3	Head of household employed as a clerical worker.
Occupation4	Head of household employed as an agent, sales worker or service worker.
Occupation5678	Head of household employed as a skilled worker.
Occupation9	Head of household employed as an unskilled worker.

Empirical study (cont.)

Fitting the response model with the same covariates (+logY) to all the sampled HH (n=1717) shows that logY and many of the covariates are highly insignificant.

Since covariates are unknown for the nonrespondents, the nonresponse is not ignorable if the distributions of some of the significant covariates are different in subsamples of respondents and nonrespondents.

Percentage of HH by size						
HH Size	**1**	**2**	**3**	**4**	**5**	**6+**
Resp.	**6.18**	**13.63**	**19.33**	**26.94**	**20.60**	**13.31**
Nonresp.	**12.39**	**18.99**	**17.34**	**24.40**	**17.34**	**9.54**

Empirical study (cont.)

Response model fitted based on all sampled HH (Respondents and "Nonrespondents"), and based only on responding HH.

Coeff.	Cons.	logY	Gender	Dist.43	Dist.44	Dist.53	HHsize
All HH	0.91	-.21	-0.20	0.88	-0.58	-0.77	0.10
Respond.	1.38	-.21	-0.26	0.91	-0.59	-0.79	0.12

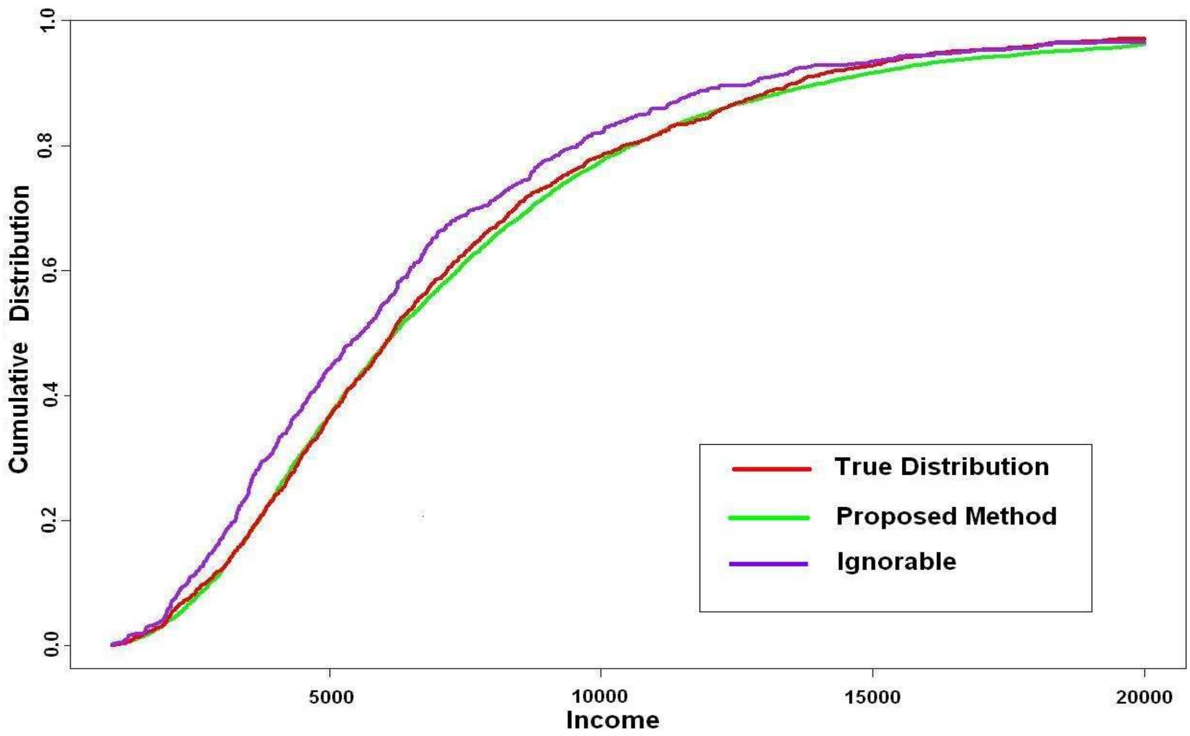
Sample model fitted based on all sampled HH (Respondents and "Nonrespondents"), and based only on responding HH.

Coeff.	Cons.	Gender	Age	Dist. 21	Dist. 41	Dist. 42	Dist. 43
All HH	7.32	-0.13	0.02	-0.18	0.17	0.13	0.17
Respond.	7.22	-0.14	0.02	-0.10	0.15	0.10	0.16

Coeff.	Dist.44	Dist. 51	Dist.52	Earners	HHsize	Occ.0	Occ.1
All HH	0.18	0.23	0.09	0.24	-0.14	0.44	0.22
Respond.	0.17	0.28	0.15	0.26	-0.13	0.45	0.24

Empirical study (cont.)

Empirical cumulative distributions of incomes



Empirical study (cont.)

Prediction of the population mean income

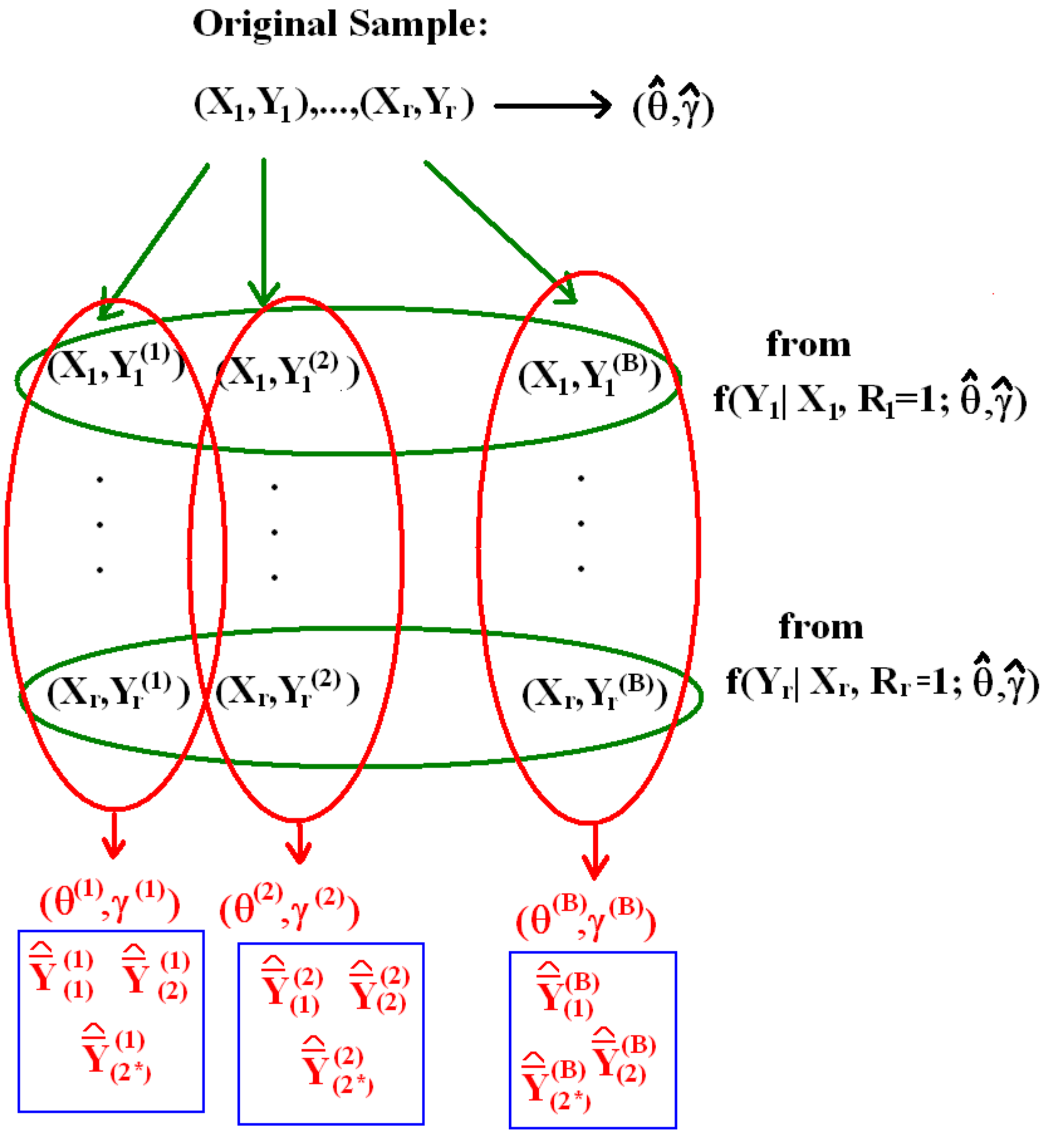
$$\hat{\bar{Y}}_{(1)} = \frac{1}{N}\sum_{i=1}^{r}\frac{w_i Y_i}{\hat{\pi}(Y_i, X_i)}$$

$$\hat{\bar{Y}}_{(2)} = \frac{1}{N}\sum_{i=1}^{n} w_i Y_i^{*}$$ **(unknown covariates)**

$$\hat{\bar{Y}}_{(2*)} = \frac{1}{N}\sum_{i=1}^{n} w_i Y_i^{*}$$ **(known covariates)**

where $Y_i^{*} = Y_i$ **if** $i \in R$, $Y_i^{*} = Y_i^{imp}$ **if** $i \in R^c$.

Empirical study (cont.)



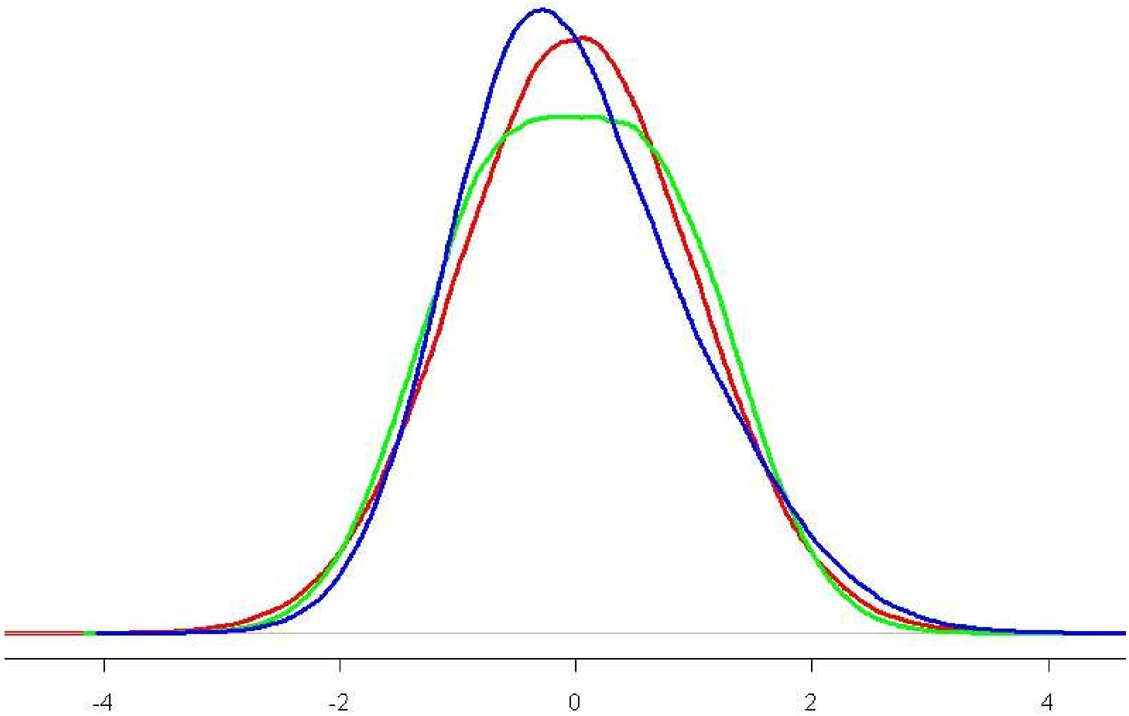
Empirical study (cont.)

Estimation of sample mean of income (True $\bar{Y} = 7215.06$).
Conditional S.E. 500 bootstrap samples.

Estimator	Estimate		Standard Error
	Original sample	Mean over bootstrap	
$\hat{\bar{Y}}_{(1)}$	**7332.30**	**7299.17**	**147.38**
$\hat{\bar{Y}}_{(2)}$	**7311.06**	**7297.09**	**146.58**
$\hat{\bar{Y}}_{(2^*)}$	**7272.26**	**7265.53**	**140.81**

Empirical study (cont.)

Testing goodness of fit



Empirical study (cont.)

Testing goodness of fit

Test	Skewed Distribution				Flat Distribution			
	Significance level				Significance level			
	0.01	0.025	0.05	0.10	0.01	0.025	0.05	0.10
KS	0.832	0.892	0.936	0.960	0.245	0.549	0.637	0.775
AD	0.936	0.964	0.984	0.988	0.588	0.725	0.784	0.853
CM	0.924	0.948	0.980	0.988	0.490	0.696	0.765	0.843
$C_{(3)}$	0.876	0.932	0.956	0.984	0.000	0.000	0.020	0.088
$C_{(4)}$	0.112	0.188	0.264	0.356	0.480	0.647	0.716	0.823

Thank you!