Statistical Inference under Nonignorable Sampling and Nonresponse: An Empirical Likelihood Approach

Danny Pfeffermann

Hebrew University & Central Bureau of Statistics, Israel
University of Southampton, UK

DSMD Distinguished Seminar Series

Joint Paper with Dr. Moshe Feder
Why are survey data different from other data?

1. Samples drawn at random with known selection probabilities, may use the randomization distribution over all possible sample selections as basis of inference instead of the distribution underlying a population model. Combination of the two distributions also common.

2. However, samples often drawn with unequal selection probabilities, at least at some stages of the sampling process. When the sampling probabilities are correlated with the outcomes, the sampling is informative and the sample model is different from the population model. Ignoring an informative sampling scheme may result in large bias.
Why are survey data different from other data? (cont.)

3. Survey data almost inevitably subject to non-response, which likewise distorts the population model if response propensity related to the outcome of interest (MNAR nonresponse).

4. Sample data are often clustered due to the use of multi-stage cluster samples. Clusters are ‘natural units’, ⇒ observations within the same cluster correlated.

5. Data available to analysts is often masked (“swapped”, “contaminated”) and differs therefore from the true data.

Consider second and third issues in present talk.
Population and sample models

**Population model**: model holding for population outcomes, (census model); $f_p(y_i | x_i)$.

**Sample model**: model holding for the sample outcomes, (population model given sample selection). $f_s(y_i | x_i)$.

$$f_s(y_i | x_i) = f(y_i | x_i, I_i = 1) = \frac{\Pr(I_i = 1 | x_i, y_i)}{\Pr(I_i = 1 | x_i)} f_p(y_i | x_i)$$

$$= \frac{E_p(\pi_i | y_i, x_i)}{E_p(\pi_i | x_i)} f_p(y_i | x_i).$$

$I_i = 1$ if unit $i$ sampled, $\pi_i = \Pr(I_i = 1) =$ sampling probability.
Population and sample models (cont.)

\[ f_s(y_i | x_i) = f_p(y_i | x_i) \text{ iff } \Pr(I_i = 1 | y_i, x_i) = \Pr(I_i = 1 | x_i) \forall y_i. \]

If \( \exists y_i; \Pr(I_i = 1 | y_i, x_i) \neq \Pr(I_i = 1 | x_i) \) \( \implies \)

\[ \implies E_s(y_i | x_i) = E_p\{\[ \frac{\Pr(I_i = 1 | x_i, y_i)}{\Pr(I_i = 1 | x_i)} \times y_i \]| x_i\} \neq E_p(y_i | x_i). \]

Estimation of \( E_p(y_i | x_i) \) often main target of inference. Ignoring an informative sampling scheme can bias the inference badly.
How can we handle the difference between the two models?

Interested in population model!!! Consider first informative sampling but assume full response. We saw that if

\[ \Pr(i \in s \mid y_i, x_i) = \Pr(i \in s \mid x_i) \forall y_i \Rightarrow \text{sampling (response) ignorable.} \]

**Possibility 1:** include among model covariates all variables and interactions that are related to the outcome and might affect the sample selection or response probabilities (*Gelman, 2007, ...*).

Appealing but has several limitations:
Requires knowledge of population values of all variables $J_U$ determining sample selection and response. Often **not practical**.

Not operational when inclusion in the sample depends also on the outcome, **i.e.**, 
$Pr(I_i = 1|Y_U, X_U, J_U) \neq Pr(I_i = 1|X_U, J_U)$, e.g., **case-control studies, MNAR nonresponse**

Including the variables $J_U$ among the model covariates could **distort** the **scientific interpretation** of the model. Requires integrating them out of the model at a later stage;

$$f_p(y_s | x_s) = \int f_p(y_s | x_s, J_U)f_p(J_U | x_s)dJ_U.$$  

Requires modeling $f_p(J_U | x_s)$. **Complicated & often not feasible.**
**Proposition:** Replace $J_U$ by $W \rightarrow$ sampling weights (*Rubin*, 1985).

- Requires ‘knowledge’ of population response probabilities.
- The weights may not fully reflect the sampling design.
- Could again distort the model interpretation.
**Another solution:** (Feder, 2011).

Suppose first that \( X \) and \( J \) are known for every \( i \in U \).

1. **Impute** the missing population outcomes by use of the model,
   \[
   f_s(y \mid x, J_U) = f_p(y \mid x, J_U),
   \]

2. Fit population model \( f_p(y_i \mid x_i) \) using observed and **imputed** population values. (population data not affected by sampling).

   When covariates \( x_k \) and variables in \( J_U \) unknown for non-sampled units, impute them as well, e.g., by sampling from sample WR with probabilities \( \tilde{\pi}_i \propto (w_i - 1) \) (Pfeffermann & Sikov, 2011).

   **Simple** and has the advantage of not having to integrate
   \[
   f_p(y \mid x) = \int f_p(y \mid x, J_U)f_p(J_U \mid x)dJ_U.
   \]
Handling the difference between the two models (cont.)

Another approach in common use is to estimate the **census estimating equations (EE)** using the **sampling weights** $w_i = 1 / \pi_i$, and then solve the corresponding **estimated EE**.

**Census EE:** $\sum_{k=1}^{N} u(y_k, x_k; \theta) = 0 \rightarrow \text{estimate: } \sum_{i \in s} w_i u(y_i, x_i; \theta) = 0.$

(Binder, 1983, Godambe & Thompson, 1986, 2009, ...)

Yields in general **consistent** estimators under the combined **randomization-population** distribution, and estimation of the randomization variance generally simple. When EE defined by **likelihood score function**, called ‘**pseudo likelihood**’ estimation.
Limitations of probability weighting procedures

1. Limited mostly to **point estimation**. Probabilistic inference, *e.g.*, confidence intervals, commonly assumes large sample **normality**.

2. Inference uses **randomization distribution** and hence cannot condition on **selected sample**, like observed covariates or selected clusters (small areas...) in a multi-level model.

3. Probability-weighted estimators generally have **large variance**, particularly with small samples and large variation of the weights.

4. Use of randomization distribution does not lend itself to prediction problems, *e.g.*, prediction of the outcome for **non-sampled units** or prediction of area means for areas with no samples in **small area estimation** problems.
Modifications of probability weighting

1- For linear regression, Magee (1998) shows that under general assumptions any estimator \( \hat{\beta}_{MG}(a) = \left[ \sum_{i \in s} w_i a_i(\alpha)x_i'x_i \right]^{-1} \sum_{i \in s} w_i a_i(\alpha)x_i'y_i \) with positive weights \( a_i(\alpha) = a(x_i, \alpha) \) is consistent for \( \beta \) under the sample distribution. \( \alpha \) chosen to minimize a scalar variance criterion such as the determinant of asymptotic variance estimator,

\[
\text{A vâr}[\hat{\beta}_{MG}(a)] = \left[ \sum_{i \in s} w_i a_i(\alpha)x_i'x_i \right]^{-1} \sum_{i \in s} w_i^2 a_i^2(\alpha) \hat{\epsilon}_i^2 x_i'x_i' \left[ \sum_{i \in s} w_i a_i(\alpha)x_i'x_i \right]^{-1}
\]

\( \hat{\epsilon}_i = (y_i - x_i'\hat{\beta}_{pw}) \).

Estimator has asymptotically lower variance under the sample model than probability-weighted estimator.
2- Pfeffermann & Sverchkov (1999) consider population model,

\[ y_j = m(x_j; \theta) + \varepsilon_j, \quad E_p(\varepsilon_j \mid x_j) = 0, \quad E_p(\varepsilon_j^2 \mid x_j) = \sigma^2. \]

If \( \Pr(I_i = 1 \mid \pi_i, y_i, x_i) = \pi_i \Rightarrow E_p(y_i \mid x_i) = E_s(w_i y_i \mid x_i) / E_s(w_i \mid x_i) \) and

\[ \theta = \arg\min_{\theta \in \Theta} \frac{1}{n} \sum_{i \in s} E_s \{ q_i [y_i - m(x_i; \tilde{\theta})]^2 \mid x_i \}; \quad q_i = w_i / E_s(w_i \mid x_i). \]

The vector \( \theta \) estimated as,

\[ \hat{\theta}_Q = \arg\min_{\hat{\theta} \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \hat{q}_i [y_i - m(x_i; \tilde{\theta})]^2; \quad \hat{q}_i = w_i / \hat{E}_s(w_i \mid x_i). \]

Requires estimating \( E_s(w_i \mid x_i) \) but \( \hat{\theta}_Q \) consistent even if \( E_s(w_i \mid x_i) \) misspecified.
**Example:** Under Pop. linear regression, 
\[ \hat{\beta}_Q = \left[ \sum_{i \in s} \hat{q}_i x_i x_i' \right]^{-1} \sum_{i \in s} \hat{q}_i x_i y_i. \]

When the sample selection depends only on the covariates, the sampling process is **ignorable** \( \Rightarrow \) only necessary to account for net sampling effects on the conditional pdf \( f_p(y_i \mid x_i) \) by use of the weights, \( q_i = w_i / E_s(w_i \mid x_i) \).

The sampling weights \( w_i \) account for sampling effects on joint distribution of \( (y_i, x_i) \). As such, they tend to be more variable and the estimator \( \hat{\beta}_{PW} \) has often a larger variance.

Combination of last two modifications also possible; apply modification of *Magee* (1998) to the estimator \( \hat{\beta}_Q \).
Use of sample model for maximum likelihood estimation

Assuming full response, independence & \( \Pr(I_i = 1 \mid \pi_i, y_i, x_i) = \pi_i, \)

\[ L_s(\theta, \gamma; y_s, x_s) = \prod_{i \in s} \frac{\Pr(I_i = 1 \mid x_i, y_i; \gamma) f_p(y_i \mid x_i; \theta)}{\Pr(I_i = 1 \mid x_i; \gamma, \theta)} = \prod_{i \in s} \frac{E_s(w_i \mid x_i; \theta, \gamma) f_p(y_i \mid x_i; \theta)}{E_s(w_i \mid y_i, x_i; \gamma)}. \]

If outcomes independent under population model, they are ‘asymptotically independent’ under the sample model when increasing the population size but fixing the sample size.

Expectations in right hand side can be estimated in two steps:

1. Estimate \( \hat{E}_s(w_i \mid y_i, x_i) \) (possibly non-parametrically by weighted mean of weights for units with similar \((y, x)\), see later),

2. Integrate \( \hat{E}_s(w_i \mid x_i; \theta) = 1 / \int [E_s(w_i \mid y, x_i)]^{-1} f_p(y \mid x_i; \theta) dy. \)
**Example:** suppose that the population model is **logistic** with a discrete covariate and \( M \) values for the outcome \( y \). Estimate:

\[
\hat{E}_s(w_i \mid y_i = m, x_i = k) = \bar{w}_{mk} = \text{sample mean of weights in cell (} m, k \text{)}.
\]

\[
\hat{\pi}_{mk} = \hat{\Pr}_p(i \in s \mid y_i = m, x_i = k) = 1 / \hat{E}_s(w_i \mid y_i = m, x_i = k) = 1 / \bar{w}_{mk}.
\]

\[
\Rightarrow \Pr_s(y_i = m \mid x_i = k; \theta) \approx \frac{[\Pr_p(y_i = m \mid x_i = k; \theta) / \bar{w}_{mk}]}{\sum_{m^* = 1}^{M} [\Pr_p(y_i = m^* \mid x_i = k; \theta) / \bar{w}_{m^*k}]}.
\]

The sampling weights feature in the sample model, but this is **not** probability weighting. The sample model again **logistic**.
Use of ‘full likelihood’

\[
L_{\text{Full}}(\theta;\gamma; I_U, y_s, x_s, x_\delta) = \prod_{i \in s} \Pr(I_i = 1 \mid y_i, x_i; \gamma) f_p(y_i \mid x_i; \theta) \prod_{j \not\in s} \Pr(I_j = 0 \mid x_j; \theta, \gamma);
\]

\(I_U = \{I_1, \ldots, I_N\} \rightarrow\) vector of sample inclusion indicators and

\[
\Pr(I_j = 0 \mid x_j; \theta, \gamma) = \int \Pr(I_j = 0 \mid y_j, x_j, \gamma) f_p(y_j \mid x_j, \theta) dy_j \rightarrow \text{propensity score. (Gelman et al. (2003), Little (2004),...).}
\]

Incorporates the sampling probabilities for units outside the sample \(\Rightarrow\) employing more information, but requires knowledge of the covariates for all population units.
Advantages of the use of the sample model

1. Lends itself to standard model based inference; likelihood based methods, Bayesian inference or semi-parametric modelling.

2. The sample likelihood provides a coherent way of handling MNAR nonresponse when estimating population models.

3. Accommodates conditional inference given responding units.


5. Use of sample model enables testing whether the sampling process is ignorable. Pfeffermann and Sverchkov (2009) review test statistics proposed in the literature.
Variance estimation

Most estimators have general form,

\[ \hat{\beta}_t = \left( \sum_{i=1}^{n} w_i t_i x_i x'_i \right)^{-1} \sum_{i=1}^{n} w_i t_i x_i y_i = \left[ X'_s W_s T_s X_s \right]^{-1} \sum_{i=1}^{n} w_i t_i x_i y_i ; \]

\[ X' = [x_1, ..., x_n], \ W_s = \text{diag}[w_1, ..., w_n], \ T_s = \text{diag}[t_1, ..., t_n], \ t_i's \text{ define estimate.} \]

Randomization variance est.:

\[ \text{Var}_r (\hat{\beta}_t) = \left[ X'_s W_s T_s X_s \right]^{-1} \left[ \text{Var}_r \left( \sum_{i=1}^{n} w_i t_i x_i e_{it} \right) \left[ X'_s W_s T_s X_s \right]^{-1} \right] ; \ e_{it} = (y_i - x'_i \hat{\beta}_t). \]

Sample model est.: \( \text{Var}_{sm} (\hat{\beta}_t) = \left[ X'_s W_s T_s X_s \right]^{-1} \left( \sum_{i \in s} \left( w_i^2 t_i^2 e_{it}^2 x_i x'_i \right) \right) \left[ X'_s W_s T_s X_s \right]^{-1}. \)

For sample model MLE, use inverse information matrix.

Bootstrap est. \( \text{Var}_{BS} (\hat{\beta}) = \frac{1}{B} \sum_{b=1}^{B} (\hat{\beta}^{(b)} - \bar{\beta})(\hat{\beta}^{(b)} - \bar{\beta})'; \ \bar{\beta} = \frac{1}{B} \sum_{b=1}^{B} \hat{\beta}^{(b)}. \)
Simulation Results Comparing the Estimators

Pfeffermann (2011) presents simulation results comparing the performance of above procedures for the case of population linear regression (non-normal errors).

Main conclusions:

1. All the procedures produce unbiased point estimators (sample MLE of slope coefficient has bias of 2.5%).

2. OLS and Feder’s procedure have lowest randomization variance. Probability weighted estimator has highest variance.

3. **Standard** variance estimators fail to account for extra operations involved in production of corresponding point estimators. **Bootstrap** corrects this problem.
Handling Missing Not At Random (MNAR) Nonresponse

So far **assumed full response**. Suppose now that both the sampling design and the response mechanism are non-ignorable.

Let $R_i = 1(0)$ if unit $i$ responds (does not respond):

$$f_o(y_i | x_i) = f(y_i | x_i, I_i = 1, R_i = 1) = \frac{\Pr(R_i = 1 | y_i, x_i, I_i = 1)}{\Pr(R_i = 1 | x_i, I_i = 1)} f_s(y_i | x_i)$$

$$= \frac{\Pr(I_i = 1 | x_i, y_i) \Pr(R_i = 1 | y_i, x_i, I_i = 1)}{\Pr(R_i = 1 | x_i, I_i = 1) \Pr(I_i = 1 | x_i)} f_p(y_i | x_i)$$

$$f_o(y_i | x_i) = f_s(y_i | x_i) \text{ iff } \Pr(R_i = 1 | y_i, x_i, I_i = 1) = \Pr(R_i = 1 | x_i, I_i = 1) \forall y_i.$$
Handling Missing Not At random (MNAR) Nonresponse

Unlike the sample selection probabilities, the response probabilities are unknown, so no “response weights” available and the previous methods are no longer operational.

Requires modelling $\Pr(R_i = 1| y_i, x_i, I_i = 1; \gamma)$.

$$f_o(y_i | x_i) = \frac{\Pr(I_i = 1 | x_i, y_i) \Pr(R_i = 1 | y_i, x_i, I_i = 1)}{\Pr(R_i = 1 | x_i, I_i = 1) \Pr(I_i = 1 | x_i)} f_p(y_i | x_i) \Rightarrow$$

For parametric likelihood-based inference, need to model also $\Pr(I_i = 1 | y_i, x_i)$ and $f_p(y_i | x_i)$ (but see below).
Respondents’ likelihood

Let $R = \{i \in s \mid R_i = 1\}$ define the sample of respondents of size $r$.

$$L_0 = \prod_{i \in R} \frac{\Pr(I_i = 1 \mid y_i, x_i; \lambda) \Pr(R_i = 1 \mid y_i, x_i, I_i = 1; \gamma)}{\Pr(R_i = 1 \mid x_i, I_i = 1; \gamma, \theta) \Pr(I_i = 1 \mid x_i; \lambda, \theta)} f_p(y_i \mid x_i; \theta).$$

Does not require knowledge of $x$’s for nonresponding units.

Problems:

1. Can estimate $\Pr(I_i = 1 \mid y_i, x_i; \lambda) = E_R^{-1}(w_i \mid y_i, x_i; \lambda); w_i = (1 / \pi_i)$, by smoothing the weights around $(y_i, x_i)$, but need to model $\Pr(R_i = 1 \mid y_i, x_i, I_i = 1; \gamma)$ and $f_p(y_i \mid x_i; \theta)$, with no direct observations from either one of these distributions.

2. Likelihood not always identifiable and maximization often unstable.
Use of empirical likelihood

Proposed solution: Replace parametric likelihood by empirical likelihood.

Ass. 1: for each unit $i \in U$ exist values $u_i = (y_i, x_i', c_i', \tau_i, \rho_i)'$; $c_i \rightarrow$ vector of survey values with known Pop. Means $\bar{c}_U$;

$\tau_i = \Pr(I_i \mid y_i, x_i) = 1 / E_R(w_i \mid y_i, x_i)$; $\rho_i = \Pr(R_i = 1 \mid y_i, x_i, I_i = 1)$.

Ass. 2: population values generated from multinomial distribution with probabilities $p_i$ (Hartley & Rao, 1968). ($p_i$ is the ‘jump’ (point mass) of Pop. CDF $F_p(u_i)$ at $u_i$, Owen, 2001).

Ass. 3: pop. distribution has its support in sample of respondents.
Empirical likelihood (cont.)

Under proposed set up,

$$\Pr(u_i \mid i \in R) = \frac{p_i \rho_i}{\sum_{k \in R} p_k \rho_k},$$

and the likelihood is

$$EL = \prod_{i \in R} p_i^{(r)}; \quad p_i^{(r)} = \Pr(u_i \mid i \in R) = \frac{p_i \rho_i}{\sum_{k \in R} p_k \rho_k}.$$

Vector parameter of interest $p = (p_1, \ldots, p_r); \quad p_i \geq 0, \sum_{i=1}^r p_i = 1$. 
Maximization of empirical likelihood under constraints

Maximize $EL$ under the constraints,

$$\sum_{i \in R} p_i c_i = \bar{c}_U \iff \sum_{i \in R} p_i^{(r)} \tau_i^{-1} \rho_i^{-1} (c_i - \bar{c}_U) = 0. \quad (\bar{c}_U \to \text{known means of survey variables})$$

Special case: $\zeta_i = \tau_i \rho_i \Rightarrow \sum_{i \in R} p_i \zeta_i = \bar{\zeta}_U \equiv \frac{r}{N} \Rightarrow$

$$\sum_{i \in R} p_i^{(r)} \left[1 - \frac{r}{(N \tau_i \rho_i)} \right] = 0. \quad \text{(Important constraint)}$$

Model for response probabilities:

Example: $\rho_i(\gamma) = \text{Pr}(R_i = 1 \mid y_i, x_i; \gamma) = \text{logit}^{-1}[l(y_i, x_i; \gamma)];$

$l(y_i, x_i; \gamma)$ is a polynomial in $(y, x)$ with coefficients $\gamma$. Testable!!
MLE based on empirical likelihood

Estimate $\tau_i = \Pr(I_i = 1 \mid y_i, x_i) = \frac{1}{E_R(w_i \mid y_i, x_i)}$ by smoothing the weights around $(y_i, x_i)$ using kernel regression (see later).

(R function `npreg` in np package). (Not really MLE).

Probabilities $p_i^{(r)} = \Pr(u_i \mid i \in R) = \frac{p_i \hat{\tau}_i \rho_i(\gamma)}{\sum_{k \in R} p_k \hat{\tau}_k \rho_k(\gamma)}$ depend on $\gamma$ so

$p_i^{(r)} = p_i^{(r)}(\gamma)$ and need to maximize,

$$\max_{\gamma, p^{(r)}(\gamma)} \prod_{i=1}^r p_i^{(r)}(\gamma), \text{ s.t. } A(\gamma)p^{(r)}(\gamma) = 0; \quad p^{(r)}(\gamma) = [p_1^{(r)}(\gamma), \ldots, p_r^{(r)}(\gamma)].$$

Rows of $A(\gamma)$ consist of $\tau_i^{-1} \rho_i^{-1}(c_i - \overline{c}_U)$ and $[1 - r(N\tau_i \rho_i)^{-1}]$. 
MLE based on empirical likelihood (cont.)

Write $\max_{\gamma, p^{(r)}(\gamma)} \prod_{i=1}^{r} p_{i}^{(r)}(\gamma) = \max_{\gamma}[G(\gamma)];$

$G(\gamma) = \max_{p^{(r)}(\gamma)} \prod_{i=1}^{r} p_{i}^{(r)}(\gamma)$ (for given $\gamma$).

Maximization of $G(\gamma)$ by use of S function elm in R (Owen).

$G(\gamma)$ is the profile likelihood of $\gamma$.

MLE of $\gamma$: $\arg \max_{\gamma} G(\gamma)$ obtained by numerical optimization using routine optim in R, initialized by a grid search.

Once estimated $\gamma \Rightarrow \hat{\rho}_{i} = \rho_{i}(\hat{\gamma})$ and

$$\hat{p}_{i} = \hat{p}_{i}^{(r)}(\hat{\tau}_{i} \hat{\rho}_{i})^{-1} / \sum_{k=1}^{r} \hat{p}_{k}^{(r)}(\hat{\tau}_{k} \hat{\rho}_{k})^{-1}. $$
Estimation of population model parameters (if assumed)

For **EL** estimation we replaced *pdf* \( f_p(y_i \mid x_i; \theta) \) by probability \( p_i \).

Under mild conditions, \( \theta \) is the unique solution of estimating Eqs.

\[
W_p(\theta) = E_p \left( \frac{\partial \log f_p(y \mid x; \theta)}{\partial \theta} \mid x \right) = 0. \quad \Rightarrow
\]

Solve \( \hat{W}_p(\theta) = \sum_{j \in R} \hat{p}_j \left[ \frac{\partial \log f_p(y_j \mid x_j; \theta)}{\partial \theta} \right] = 0. \)

**Exp. 1**- Estimate \( \bar{Z}_U \) by solving \( \sum_{i=1}^{r} p_i (z_i - \mu_z) = 0 \Rightarrow \hat{\mu}_z = \sum_{i=1}^{r} p_i z_i. \)

**Exp. 2**- Let \( y_i = 1,0, \) \( \Pr(y_i = 1 \mid x_i) = \frac{\exp(x_i'\beta)}{[1 + \exp(x_i'\beta)]} = f_i(\beta), \)

Solve: \( \hat{W}_p(\theta) = \sum_{j \in R} \hat{p}_j [y_i - f_i(\beta)] x_i = 0 \) (used in simulations).
Simulation study

Generate **10,000** population values, \( x_j \sim \text{gamma}(2,2) \).

Generate **binary outcomes**, \( \Pr(y_j = 1 \mid x_j; \beta) = \logit^{-1}(-0.8 + 0.8x_j) \).

Select **sample**, \( I_j \sim \text{Ber}(\pi_j) \), \( \pi_j = \min(3500z_j^{-1} / \sum_{k=1}^{10000} z_k^{-1}, 0.9999) \).

\( z_j = \max[(x_j + 1.1)(2y_j + 1) + v_j, 0.01] \); \( v_j \sim \text{Uniform}(-0.2, 0.2) \).

**Calibration variables**, \( c_j = (1, x_j, y_j, x_jy_j, x_j^2, x_j^2y_j)' + \epsilon_j \); \( \epsilon_j \sim N(0, I) \).

**Response process**: \( \rho_i = \Pr(R_i = 1 \mid y_i, x_i) = \logit^{-1}(0.7 + 0.5x_i - 1.5y_i) \).

Repeat **300 times**. (x-values generated once).
**Estimation of** \[ \tau_i = \Pr(I_i = 1 | x_i, y_i) = E_R^{-1}(w_i | y_i, x_i) \]

Used **Nadaraya-Watson (1964)** kernel regression to estimate \( E_R(w_i | y_i, x_i) \) locally. For every \((x, y)\), compute weighted average of weights in neighbourhood around \((x, y)\), with the weight decreasing as the distance from \((x, y)\) increases.

Weights determined by **kernel function** \( K() \) with **bandwidth** \( h \).

**Example:** \( K_{h_x}(d) = \left(2\pi h^2\right)^{-1/2} \exp\left(-\frac{d^2}{2h^2}\right) \) (Gaussian kernel).

\[
\hat{E}_R(w_i | y_i, x_i) = \frac{\sum_{j=1}^{r} K_{h_x}(x_i - x_j) w_i}{\sum_{i=1}^{r} K_{h_x}(x_i - x_j)} ; \ y_i = 0, 1.
\]
Simulation results

Table 1. Mean estimates, standard errors and SQRT of mean of variance estimates,

\[ N = 10,000 \;; \; 3063 \leq n \leq 3954 \;; \; 2039 \leq r \leq 2636 \]

\( \gamma \)-coefficients: \( \gamma_0 = 0.7, \gamma_x = 0.5, \gamma_y = -1.5 \), 300 samples,

Correct model

<table>
<thead>
<tr>
<th>Mean estimates</th>
<th>Empirical Std</th>
<th>Av. Std estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{\gamma}_0 )</td>
<td>( \hat{\gamma}_x )</td>
<td>( \hat{\gamma}_y )</td>
</tr>
<tr>
<td>0.73</td>
<td>0.50</td>
<td>-1.54</td>
</tr>
</tbody>
</table>
Table 2. Mean estimates, standard errors and SQRT of mean of variance estimates,
\[ N = 10,000 \ ; \ 3063 \leq n \leq 3954 \ ; \ 2039 \leq r \leq 2636 \]
\( \beta \)-coefficients: \( \hat{\beta}_0 = -0.8, \hat{\beta}_1 = 0.8, \) 300 samples, **correct model**

<table>
<thead>
<tr>
<th>Method</th>
<th>Mean estimates</th>
<th>Empirical Std</th>
<th>Av. Std estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \hat{\beta}_0 )</td>
<td>( \hat{\beta}_1 )</td>
<td>( \hat{\beta}_0 )</td>
</tr>
<tr>
<td>FR UW</td>
<td>-1.90</td>
<td>0.80</td>
<td>0.08</td>
</tr>
<tr>
<td>FR PW</td>
<td>-0.80</td>
<td>0.80</td>
<td>0.08</td>
</tr>
<tr>
<td>MARUW</td>
<td>-2.67</td>
<td>0.96</td>
<td>0.11</td>
</tr>
<tr>
<td>MARPW</td>
<td>-1.56</td>
<td>0.96</td>
<td>0.11</td>
</tr>
<tr>
<td>CREL</td>
<td>-0.79</td>
<td>0.79</td>
<td>0.17</td>
</tr>
</tbody>
</table>

**FR**→ Full response, all sample data; **UW**→ Unweighted; **PW**→ Probability weighted; **MAR**→ Ignore response mechanism; **CREL**→ proposed method
Model testing

Can we test our model? why not?

We have observations from a model fitted to responding units.

Use classical goodness of fit test statistics.

Example: Hosmer and Lemeshow (1980).

Divide the range of x-values of responding units into $K$ groups, $R_x = G_1 \cup \ldots \cup G_K$; $\#(G_k) = r_g$. Let,

$$p_{gi} = Pr(y_{gi} = 1|I_{gi} = 1, R_{gi} = 1, x_{gi}); \; g = G_1, \ldots, G_K, \; i = 1, \ldots, r_g.$$

$$H_0: \forall g, \sum_{g=1}^{K} \frac{(\bar{y}_g - \bar{q}_g)^2}{\bar{q}_g (1-\bar{q}_g) / r_g} \sim \chi^2_{(K-2)}; \; \bar{y}_g = \frac{1}{r_g} \sum_{i=1}^{r_g} y_{gi}, \; \bar{q}_g = \frac{1}{r_g} \sum_{i=1}^{r_g} p_{gi}.$$
Testing Results

Distribution under correct model. $G=10$ equal size groups.

$est\ df = 8.355$
Model testing (cont.)

\[ EL = \prod_{i \in R} p_i^{(r)} ; \quad p_i^{(r)} = \Pr(u_i \mid i \in R) = \frac{p_i \tau_i \rho_i}{\sum_{k \in R} p_k \tau_k \rho_k} \]

Where can we go wrong?

The assumption that population observations are obtained with “probabilities” \( p_i \) is harmless.

We estimate \( \tau_i = \Pr(I_i = 1 \mid y_i, x_i) = \frac{1}{E_R(w_i \mid y_i, x_i)} \) nonparametrically.

Only parametric assumption is on \( \rho_i = \Pr(R_i = 1 \mid y_i, x_i) \).
Powers of test statistic

**Proportion of samples rejecting** \( H_0: \) \( \logit(\rho_i) = 0.7 + 0.5x - 1.5y. \)

\( G=10 \) equal size groups, 100 samples

**Correct model A:** \( \logit(\rho_i) = 0.7 + 0.5x - 1.5y + ax^2 \)

<table>
<thead>
<tr>
<th>Sig. level</th>
<th>( a=0 )</th>
<th>( a=-0.5 )</th>
<th>( a=-0.6 )</th>
<th>( a=-0.8 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \alpha = 0.05 )</td>
<td>0.06</td>
<td>0.32</td>
<td>0.60</td>
<td>0.75</td>
</tr>
<tr>
<td>( \alpha = 0.10 )</td>
<td>0.14</td>
<td>0.39</td>
<td>0.63</td>
<td>0.79</td>
</tr>
</tbody>
</table>

**Correct model B:** \( \logit(\rho_i) = 0.7 + 0.5x - 1.5y + bxy \)

<table>
<thead>
<tr>
<th>Sig. level</th>
<th>( b=0 )</th>
<th>( b=-1 )</th>
<th>( b=-1.5 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \alpha = 0.05 )</td>
<td>0.06</td>
<td>0.32</td>
<td>0.71</td>
</tr>
<tr>
<td>( \alpha = 0.10 )</td>
<td>0.14</td>
<td>0.39</td>
<td>0.74</td>
</tr>
</tbody>
</table>
Conclusions

1- Established and illustrated a general procedure for analyzing complex survey data subject to informative sampling and MNAR nonresponse, with “minimum” assumptions.

2- Easy to explain to users. \[ EL = \prod_{i \in R} \frac{p_i \tau_i \rho_i}{\sum_{k \in R} p_k \tau_k \rho_k}. \]

3- The model for the response probabilities is testable!!

4- Computations not trivial but can be implemented using available procedures in R, with some modifications.