Robust inference in two-phase sampling with application to unit nonresponse

David Haziza and Jean-François Beaumont

Université de Montréal and Statistics Canada

International Total Survey Error Workshops 2011

Quebec, Canada

June 21, 2011
Outline of the presentation

1. Introduction
2. Measuring the influence: the conditional bias
3. Robust estimators
4. Application to unit nonresponse
5. Simulation study
6. Concluding remarks
Influential units

- Unusual observations with possibly large design weights
Influential units

- Unusual observations with possibly large design weights
- Many survey statistics are sensitive to the presence of influential units
Influential units

- Unusual observations with possibly large design weights
- Many survey statistics are sensitive to the presence of influential units
- Including or excluding an influential unit in the calculation of these statistics can have a dramatic impact on their magnitude.
Influential units

- Unusual observations with possibly large design weights
- Many survey statistics are sensitive to the presence of influential units
- Including or excluding an influential unit in the calculation of these statistics can have a dramatic impact on their magnitude.
- The occurrence of outliers is common in business surveys because the distributions of variables (e.g., revenue, sales, etc.) are highly skewed (heavy right tail)

David Haziza and Jean-François Beaumont
Robust inference in two-phase sampling
June 21, 2011
Influential units

- Unusual observations with possibly large design weights
- Many survey statistics are sensitive to the presence of influential units
- Including or excluding an influential unit in the calculation of these statistics can have a dramatic impact on their magnitude.
- The occurrence of outliers is common in business surveys because the distributions of variables (e.g., revenue, sales, etc.) are highly skewed (heavy right tail)
- Influential units are legitimate observations
Influential units

- Unusual observations with possibly large design weights
- Many survey statistics are sensitive to the presence of influential units
- Including or excluding an influential unit in the calculation of these statistics can have a dramatic impact on their magnitude.
- The occurrence of outliers is common in business surveys because the distributions of variables (e.g., revenue, sales, etc.) are highly skewed (heavy right tail)
- Influential units are legitimate observations
- The impact of influential units can be minimized by using a good sampling design: for example, stratified sampling with a take-all stratum
Influential units

- Unusual observations with possibly large design weights
- Many survey statistics are sensitive to the presence of influential units
- Including or excluding an influential unit in the calculation of these statistics can have a dramatic impact on their magnitude.
- The occurrence of outliers is common in business surveys because the distributions of variables (e.g., revenue, sales, etc.) are highly skewed (heavy right tail)
- Influential units are legitimate observations
- The impact of influential units can be minimized by using a good sampling design: for example, stratified sampling with a take-all stratum
Influential units

- Even with a good sampling design, influential units may still be selected in the sample (e.g., stratum jumpers)
Even with a good sampling design, influential units may still be selected in the sample (e.g., stratum jumpers).

In the presence of influential units, survey statistics are (approximately) unbiased but they can have a very large variance.
Influential units

- Even with a good sampling design, influential units may still be selected in the sample (e.g., stratum jumpers).
- In the presence of influential units, survey statistics are (approximately) unbiased but they can have a very large variance.
- Reducing the influence of large values produces stable but biased estimators.
Even with a good sampling design, influential units may still be selected in the sample (e.g., stratum jumpers).

In the presence of influential units, survey statistics are (approximately) unbiased but they can have a very large variance.

Reducing the influence of large values produces stable but biased estimators.

Treatment of influential units: trade-off between bias and variance
Even with a good sampling design, influential units may still be selected in the sample (e.g., stratum jumpers)

In the presence of influential units, survey statistics are (approximately) unbiased but they can have a very large variance.

Reducing the influence of large values produces stable but biased estimators

Treatment of influential units: trade-off between bias and variance
Two-phase designs

- $U$: finite population of size $N$
Two-phase designs

- $U$: finite population of size $N$
- $s_1$: first-phase sample, of size $n_1$
Two-phase designs

- $U$: finite population of size $N$
- $s_1$: first-phase sample, of size $n_1$
- $s_2$: second-phase sample, of size $n_2$, selected from $s_1$
Two-phase designs

- \( U \): finite population of size \( N \)
- \( s_1 \): first-phase sample, of size \( n_1 \)
- \( s_2 \): second-phase sample, of size \( n_2 \), selected from \( s_1 \)
- \( l_{1i} \): first-phase sample selection indicator for unit \( i \)
Two-phase designs

- $U$: finite population of size $N$
- $s_1$: first-phase sample, of size $n_1$
- $s_2$: second-phase sample, of size $n_2$, selected from $s_1$
- $I_{1i}$: first-phase sample selection indicator for unit $i$
- $I_{2i}$: second-phase sample selection indicator for unit $i$
Two-phase designs

- $U$: finite population of size $N$
- $s_1$: first-phase sample, of size $n_1$
- $s_2$: second-phase sample, of size $n_2$, selected from $s_1$
- $l_{1i}$: first-phase sample selection indicator for unit $i$
- $l_{2i}$: second-phase sample selection indicator for unit $i$
- Vectors of indicators: $l_1 = (l_{11}, \cdots, l_{1N})'$ and $l_2 = (l_{21}, \cdots, l_{2N})'$
- First-phase inclusion probability for unit $i$: $\pi_{1i} = P(l_{1i} = 1)$
Two-phase designs

- $U$: finite population of size $N$
- $s_1$: first-phase sample, of size $n_1$
- $s_2$: second-phase sample, of size $n_2$, selected from $s_1$
- $l_{1i}$: first-phase sample selection indicator for unit $i$
- $l_{2i}$: second-phase sample selection indicator for unit $i$
- Vectors of indicators: $\mathbf{l}_1 = (l_{11}, \cdots, l_{1N})'$ and $\mathbf{l}_2 = (l_{21}, \cdots, l_{2N})'$
- First-phase inclusion probability for unit $i$: $\pi_{1i} = P(l_{1i} = 1)$
- Second-phase inclusion probability for unit $i$: $\pi_{2i}(\mathbf{l}_1) = P(l_{2i} = 1|\mathbf{l}_1; l_{1i} = 1)$
Two-phase designs

- **U**: finite population of size \( N \)
- \( s_1 \): first-phase sample, of size \( n_1 \)
- \( s_2 \): second-phase sample, of size \( n_2 \), selected from \( s_1 \)
- \( l_{1i} \): first-phase sample selection indicator for unit \( i \)
- \( l_{2i} \): second-phase sample selection indicator for unit \( i \)
- Vectors of indicators: \( \mathbf{l}_1 = (l_{11}, \ldots, l_{1N})' \) and \( \mathbf{l}_2 = (l_{21}, \ldots, l_{2N})' \)
- First-phase inclusion probability for unit \( i \): \( \pi_{1i} = P(l_{1i} = 1) \)
- Second-phase inclusion probability for unit \( i \): \( \pi_{2i}(\mathbf{l}_1) = P(l_{2i} = 1|\mathbf{l}_1; l_{1i} = 1) \)
Two-phase sampling

\[ I_{1i} = 0, I_{2i} = 0 \]

\[ I_{1i} = 1, I_{2i} = 1 \]

\[ I_{1i} = 1, I_{2i} = 0 \]
Invariance

- A two-phase sampling design possesses the invariance property if 
  \[ P(I_2|I_1) = P(I_2) \]
**Invariance**

- A two-phase sampling design possesses **the invariance property** if $P(I_2|I_1) = P(I_2)$
- Invariance $\Rightarrow \pi_2(i|I_1) = \pi_2(i)$
- Example of invariance: simple random sampling without replacement in both phases and both $n_1$ and $n_2$ are fixed prior to sampling
Invariance

- A two-phase sampling design possesses the invariance property if $P(I_2|I_1) = P(I_2)$
- Invariance $\Rightarrow \pi_{2i}(I_1) = \pi_{2i}$
- Example of invariance: simple random sampling without replacement in both phases and both $n_1$ and $n_2$ are fixed prior to sampling
- Example of non-invariance:
  - simple random sampling without replacement in the first phase
Invariance

A two-phase sampling design possesses the invariance property if
\[ P(I_2|I_1) = P(I_2) \]

Invariance \( \Rightarrow \pi_{2i}(I_1) = \pi_{2i} \)

Example of invariance: simple random sampling without replacement in both phases and both \( n_1 \) and \( n_2 \) are fixed prior to sampling

Example of non-invariance:
- simple random sampling without replacement in the first phase
- proportional-to-size sampling in the second phase. That is,

\[
\pi_{2i}(I_1) = n_2 \frac{x_i}{\sum_{i \in s_1} x_i},
\]

where \( x \) is a size variable available for all \( i \in s_1 \)
Invariance

- A two-phase sampling design possesses the invariance property if
  \[ P(I_2|I_1) = P(I_2) \]
- Invariance \( \Rightarrow \pi_{2i}(I_1) = \pi_{2i} \)
- Example of invariance: simple random sampling without replacement in both phases and both \( n_1 \) and \( n_2 \) are fixed prior to sampling
- Example of non-invariance:
  - simple random sampling without replacement in the first phase
  - proportional-to-size sampling in the second phase. That is,
    \[ \pi_{2i}(I_1) = n_2 \frac{x_i}{\sum_{i \in s_1} x_i} \]
    where \( x \) is a size variable available for all \( i \in s_1 \)
- In the remaining, we assume that the two-phase design satisfies the invariance property
Point estimation

- Goal: estimate a population total of a variable of interest $y$, 

$$Y = \sum_{i \in U} y_i$$
Point estimation

- Goal: estimate a population total of a variable of interest $y$,

$$Y = \sum_{i \in U} y_i$$

- $y$-values: available only for $i \in s_2$
Point estimation

- Goal: estimate a population total of a variable of interest $y$,

$$ Y = \sum_{i \in U} y_i $$

- $y$-values: available only for $i \in s_2$

- Complete data estimator: Double expansion estimator

$$ \hat{Y}_{DE} = \sum_{i \in s_2} \frac{y_i}{\pi_1 i \pi_2 i} = \sum_{i \in s_2} \frac{y_i}{\pi^*_i} $$
Point estimation

- Goal: estimate a population total of a variable of interest \( y \),
  \[
  Y = \sum_{i \in U} y_i
  \]

- \( y \)-values: available only for \( i \in s_2 \)

- Complete data estimator: Double expansion estimator
  \[
  \hat{Y}_{DE} = \sum_{i \in s_2} \frac{y_i}{\pi_1 i \pi_2 i} = \sum_{i \in s_2} \frac{y_i}{\pi^*_i}
  \]

- \( \hat{Y}_{DE} \) is design-unbiased for \( Y \); that is,
  \[
  E_1 E_2 (\hat{Y}_{DE} | I_1) = Y
  \]
Total error

The total error of $\hat{Y}_{DE}$:

$$\hat{Y}_{DE} - Y = (\hat{Y}_E - Y) + (\hat{Y}_{DE} - \hat{Y}_E)$$

(1)

where $\hat{Y}_E = \sum_{i \in s_1} \pi_{1i}^{-1} y_i$ is the estimator one would have used in a single-phase sampling design.
The total error of $\hat{Y}_{DE}$:

\[
\hat{Y}_{DE} - Y = (\hat{Y}_E - Y) + (\hat{Y}_{DE} - \hat{Y}_E)
\]

where $\hat{Y}_E = \sum_{i \in s_1} \pi_{1i}^{-1} y_i$; is the estimator one would have used in a single-phase sampling design.

An influential unit may have an impact on both the first phase and the second phase errors.
Total error

- The total error of $\hat{Y}_{DE}$:

$$\hat{Y}_{DE} - Y = (\hat{Y}_E - Y) + (\hat{Y}_{DE} - \hat{Y}_E)$$  \hspace{1cm} (1)

where $\hat{Y}_E = \sum_{i \in s_1} \pi_1^{-1} y_i$ is the estimator one would have used in a single-phase sampling design

- An influential unit may have an impact on both the first phase and the second phase errors

- How to measure the influence (or impact) of a unit on both errors?
The total error of $\hat{Y}_{DE}$:

$$\hat{Y}_{DE} - Y = (\hat{Y}_E - Y) + (\hat{Y}_{DE} - \hat{Y}_E)$$

(1)

where $\hat{Y}_E = \sum_{i \in s_1} \pi_{1i}^{-1} y_i$ is the estimator one would have used in a single-phase sampling design.

- An influential unit may have an impact on both the first phase and the second phase errors.
- How to measure the influence (or impact) of a unit on both errors? Single phase sampling: the conditional bias; Moreno-Rebollo, Munoz-Reyez and Munoz-Pichardo (1999), Beaumont, Haziza and Ruiz-Gazen (2011).
Total error

- The total error of $\hat{Y}_{DE}$:

$$\hat{Y}_{DE} - Y = (\hat{Y}_E - Y) + (\hat{Y}_{DE} - \hat{Y}_E)$$

where $\hat{Y}_E = \sum_{i \in s_1} \pi_1^{-1} y_i$ is the estimator one would have used in a single-phase sampling design

- An influential unit may have an impact on both the first phase and the second phase errors

- How to measure the influence (or impact) of a unit on both errors? Single phase sampling: the conditional bias; Moreno-Rebollo, Munoz-Reyez and Munoz-Pichardo (1999), Beaumont, Haziza and Ruiz-Gazen (2011).

We distinguish between three cases:

- $i \in s_2$: sampled unit
- $i \in s_1 - s_2$: sampled in first phase but not in the second phase
- $i \in U - s_1$: nonsampled unit
Measuring the influence: the conditional bias

- We distinguish between three cases:
  - \( i \in s_2 \): sampled unit
  - \( i \in s_1 - s_2 \): sampled in first phase but not in the second phase
  - \( i \in U - s_1 \): nonsampled unit

- We can only reduce the influence of the sampled units (i.e., the units belonging to \( s_2 \))
Measuring the influence: the conditional bias

- We distinguish between three cases:
  - \( i \in s_2 \): sampled unit
  - \( i \in s_1 - s_2 \): sampled in first phase but not in the second phase
  - \( i \in U - s_1 \): nonsampled unit

- We can only reduce the influence of the sampled units (i.e., the units belonging to \( s_2 \))
- Nothing can be done for the other units at the estimation stage
Measuring the influence: the conditional bias

- We distinguish between three cases:
  - $i \in s_2$: sampled unit
  - $i \in s_1 - s_2$: sampled in first phase but not in the second phase
  - $i \in U - s_1$: nonsampled unit

- We can only reduce the influence of the sampled units (i.e., the units belonging to $s_2$)

- Nothing can be done for the other units at the estimation stage

- Influence of sampled unit $i \in s_2$:

$$B_{i}^{DE}(l_{1i} = 1, l_{2i} = 1) = E_1 E_2(\hat{Y}_{DE} - Y|l_{1i} = 1, l_{2i} = 1) = E_1(\hat{Y}_{E} - Y|l_{1i} = 1) + E_1 E_2(\hat{Y}_{DE} - \hat{Y}_{E}|l_{1i} = 1, l_{2i} = 1)$$
Measuring the influence: the conditional bias

- We distinguish between three cases:
  - \( i \in s_2 \): sampled unit
  - \( i \in s_1 - s_2 \): sampled in first phase but not in the second phase
  - \( i \in U - s_1 \): nonsampled unit

- We can only reduce the influence of the sampled units (i.e., the units belonging to \( s_2 \))

- Nothing can be done for the other units at the estimation stage

- Influence of sampled unit \( i \in s_2 \):

\[
B_{i}^{DE}(l_{1i} = 1, l_{2i} = 1) = E_1E_2(\hat{Y}_{DE} - Y|l_{1i} = 1, l_{2i} = 1)
= E_1(\hat{Y}_{E} - Y|l_{1i} = 1)
+ E_1E_2(\hat{Y}_{DE} - \hat{Y}_{E}|l_{1i} = 1, l_{2i} = 1)
\]
Measuring the influence: the conditional bias

- Arbitrary two-phase design:

\[
B_{i}^{DE}(l_{1i} = 1, l_{2i} = 1) = \sum_{j \in U} \left( \frac{\pi_{1ij}}{\pi_{1i}\pi_{1j}} - 1 \right) y_{j}
\]

Influence of unit \( i \) on the first-phase error

\[
+ \sum_{j \in U} \frac{\pi_{1ij}}{\pi_{1i}\pi_{1j}} \left( \frac{\pi_{2ij}}{\pi_{2i}\pi_{2j}} - 1 \right) y_{j}
\]

Influence of unit \( i \) on the second-phase error

\[
= \sum_{j \in U} \left( \frac{\pi_{ij}^{*}}{\pi_{i}^{*}\pi_{j}^{*}} - 1 \right) y_{j}
\]

Total influence of unit \( i \)
Measuring the influence: the conditional bias

- **SRSWOR/SRSWOR**: \( \pi_i^* = \frac{n_1}{N} \times \frac{n_2}{n_1} = \frac{n_2}{N} \)

\[
B_i^{DE}(l_1 = 1, l_2 = 1) = \frac{N}{(N-1)} \left( \frac{N}{n_1} - 1 \right)(y_i - \bar{Y}) \\
+ \frac{N}{(N-1)} \left( \frac{n_1}{n_2} - 1 \right)(y_i - \bar{Y}) \\
= \frac{N}{(N-1)} \left( \frac{n_2}{n_2} - 1 \right)(y_i - \bar{Y})
\]
Measuring the influence: the conditional bias

**SRSWOR/SRSWOR:** \( \pi_i^* = \frac{n_1}{N} \times \frac{n_2}{n_1} = \frac{n_2}{N} \)

\[
B_{DE}^i (l_1i = 1, l_2i = 1) = \frac{N}{(N - 1)} \left( \frac{N}{n_1} - 1 \right) (y_i - \bar{Y}) \\
+ \frac{N}{(N - 1)} \frac{N}{n_1} \left( \frac{n_1}{n_2} - 1 \right) (y_i - \bar{Y}) \\
= \frac{N}{(N - 1)} \left( \frac{N}{n_2} - 1 \right) (y_i - \bar{Y})
\]

**Poisson sampling/Poisson sampling:**

\[
B_{DE}^i (l_1i = 1, l_2i = 1) = \left( \frac{1}{\pi_{1i}} - 1 \right) y_i + \frac{1}{\pi_{1i}} \left( \frac{1}{\pi_{2i}} - 1 \right) y_i \\
= \left( \frac{1}{\pi_i^*} - 1 \right) y_i
\]
Measuring the influence: the conditional bias

Arbitrary design/Poisson sampling:

\[ B_i^{DE}(I_{1i} = 1, I_{2i} = 1) = \sum_{j \in U} \left( \frac{\pi_{1ij}}{\pi_{1i}\pi_{1j}} - 1 \right) y_j + \pi_{1i}^{-1} (\pi_{2i}^{-1} - 1) y_i \]
Measuring the influence: the conditional bias

- Arbitrary design/Poisson sampling:

\[ B_{i}^{DE}(l_{1i} = 1, l_{2i} = 1) = \sum_{j \in U} \left( \frac{\pi_{1ij}}{\pi_{1i}\pi_{1j}} - 1 \right) y_{j} + \pi_{1i}^{-1} (\pi_{2i}^{-1} - 1) y_{i} \]

- Conditional bias:
  - unknown ⇒ must be estimated
Measuring the influence: the conditional bias

- **Arbitrary design/Poisson sampling:**

\[
B^D_i(l_{1i} = 1, l_{2i} = 1) = \sum_{j \in U} \left( \frac{\pi_{1ij}}{\pi_{1i}\pi_{1j}} - 1 \right) y_j + \pi_{1i}^{-1} (\pi_{2i}^{-1} - 1) y_i
\]

- **Conditional bias:**
  - unknown \(\Rightarrow\) must be estimated
  - can be interpreted as a **contribution of each unit** (sampled or nonsampled) to the total error
Measuring the influence: the conditional bias

- Arbitrary design/Poisson sampling:

\[ B_i^{DE}(l_1i = 1, l_2i = 1) = \sum_{j \in U} \left( \frac{\pi_{1ij}}{\pi_{1i}\pi_{1j}} - 1 \right) y_j + \pi_{1i}^{-1} (\pi_{2i}^{-1} - 1) y_i \]

- Conditional bias:
  - unknown \(\Rightarrow\) must be estimated
  - can be interpreted as a contribution of each unit (sampled or nonsampled) to the total error
  - take fully account of the sampling design: an unit may be highly influential under a given sampling design but may have little or no influence under another sampling design
Measuring the influence: the conditional bias

- Arbitrary design/Poisson sampling:

\[
B_{i}^{DE}(l_{1i} = 1, l_{2i} = 1) = \sum_{j \in U} \left( \frac{\pi_{1ij}}{\pi_{1i} \pi_{1j}} - 1 \right) y_{j} + \pi_{1i}^{-1} (\pi_{2i}^{-1} - 1) y_{i}
\]

- Conditional bias:
  - unknown \(\Rightarrow\) must be estimated
  - can be interpreted as a contribution of each unit (sampled or nonsampled) to the total error
  - take fully account of the sampling design: an unit may be highly influential under a given sampling design but may have little or no influence under another sampling design
  - If \(\pi_{i}^{*} = 1\), then \(B_{i}^{DE}(l_{1i} = 1, l_{2i} = 1) = 0\)
Measuring the influence: the conditional bias

- Arbitrary design/Poisson sampling:

\[ B_{i}^{DE}(l_{1i} = 1, l_{2i} = 1) = \sum_{j \in U} \left( \frac{\pi_{1ij}}{\pi_{1i}\pi_{1j}} - 1 \right) y_{j} + \pi_{1i}^{-1} (\pi_{2i}^{-1} - 1) y_{i} \]

- Conditional bias:
  - unknown \(\Rightarrow\) must be estimated
  - can be interpreted as a contribution of each unit (sampled or nonsampled) to the total error
  - take fully account of the sampling design: an unit may be highly influential under a given sampling design but may have little or no influence under another sampling design
  - If \(\pi_{i}^{*} = 1\), then \(B_{i}^{DE}(l_{1i} = 1, l_{2i} = 1) = 0\)
A robust version of the double expansion estimator

Following Beaumont, Haziza and Ruiz-Gazen (2011), we obtain

$$\hat{Y}_{DE}^R = \hat{Y}_{DE} - \sum_{i \in s_2} \hat{B}_{i}^{DE}(l_{1i} = 1, l_{2i} = 1) + \sum_{i \in s_2} \psi \left\{ \hat{B}_{i}^{DE}(l_{1i} = 1, l_{2i} = 1) \right\}$$
A robust version of the double expansion estimator

- Following Beaumont, Haziza and Ruiz-Gazen (2011), we obtain

\[
\hat{Y}^R_{DE} = \hat{Y}_{DE} - \sum_{i \in s_2} \hat{B}_i^{DE}(l_1i = 1, l_2i = 1) + \sum_{i \in s_2} \psi \left\{ \hat{B}_i^{DE}(l_1i = 1, l_2i = 1) \right\}
\]

- Example of \(\psi\)-function:

\[
\psi(t) = \begin{cases} 
  c & \text{if } t > c \\
  t & \text{if } |t| \leq c \\
  -c & \text{if } t < -c 
\end{cases}
\]

- \(c\): tuning constant
A robust version of the double expansion estimator

Following Beaumont, Haziza and Ruiz-Gazen (2011), we obtain

$$\hat{Y}_{DE}^R = \hat{Y}_{DE} - \sum_{i \in s_2} \hat{B}_i^{DE}(l_1i = 1, l_2i = 1) + \sum_{i \in s_2} \psi \left\{ \hat{B}_i^{DE}(l_1i = 1, l_2i = 1) \right\}$$

Example of $\psi$-function:

$$\psi(t) = \begin{cases} 
  c & \text{if } t > c \\
  t & \text{if } |t| \leq c \\
  -c & \text{if } t < -c 
\end{cases}$$

$c$: tuning constant

Special case: single-phase sampling; i.e., $l_2i = 1$ for all $i \Rightarrow \hat{Y}_{DE}^R$ reduces to the robust estimator proposed by Beaumont, Haziza and Ruiz-Gazen (2011).
Unit nonresponse

- $s_2$: set of respondents
Unit nonresponse

- $s_2$: set of respondents
- $n_2$: number of responding units (random)
Unit nonresponse

- $s_2$: set of respondents
- $n_2$: number of responding units (random)
- $l_{2i}$: response indicator for unit $i$
Unit nonresponse

- $s_2$: set of respondents
- $n_2$: number of responding units (random)
- $l_{2i}$: response indicator for unit $i$
- $\pi_{2i}$: unknown response probability for unit $i$.
Unit nonresponse

- $s_2$: set of respondents
- $n_2$: number of responding units (random)
- $l_{2i}$: response indicator for unit $i$
- $\pi_{2i}$: unknown response probability for unit $i$.
- We assume sampled units respond independently of one another (similar to Poisson sampling in the second phase)
Unit nonresponse

- $s_2$: set of respondents
- $n_2$: number of responding units (random)
- $I_{2i}$: response indicator for unit $i$
- $\pi_{2i}$: unknown response probability for unit $i$.
- We assume sampled units respond independently of one another (similar to Poisson sampling in the second phase)
- Propensity score adjusted estimator, assuming the $\pi_{2i}$’s are known:

\[
\hat{Y}_{PSA} = \sum_{i \in s_2} \frac{y_i}{\pi_1 i \pi_{2i}}
\]
Unit nonresponse

- $s_2$: set of respondents
- $n_2$: number of responding units (random)
- $l_{2i}$: response indicator for unit $i$
- $\pi_{2i}$: unknown response probability for unit $i$.
- We assume sampled units respond independently of one another (similar to Poisson sampling in the second phase)
- Propensity score adjusted estimator, assuming the $\pi_{2i}$'s are known:
  \[ \tilde{Y}_{PSA} = \sum_{i \in s_2} \frac{y_i}{\frac{\pi_{1i}}{\pi_{1i} \pi_{2i}}} \]

  **Influence of a responding unit:**

  \[ B_{PSA}^{1}(l_{1i} = 1, l_{2i} = 1) = \sum_{j \in U} \left( \frac{\pi_{1ij}}{\pi_{1i} \pi_{1j}} - 1 \right) y_j + \frac{\pi_{1i}^{-1}}{\pi_{2i}^{-1} - 1} y_i \]

  - Influence of unit $i$ on the sampling error
  - Influence of unit $i$ on the nonresponse error
In practice, the response probability $\pi_{2i}$ is unknown.
Nonresponse model

- In practice, the response probability $\pi_{2i}$ is unknown
- Parametric nonresponse model: $\pi_{2i} = m(x_i, \alpha)$,
Nonresponse model

- In practice, the response probability $\pi_{2i}$ is unknown
- **Parametric nonresponse model:** $\pi_{2i} = m(x_i, \alpha)$, where
  - $m(.)$ is a known function
  - $x_i$ is a vector of auxiliary variables available for all the sampled units (respondents and nonrespondents)
  - $\alpha$ is a vector of unknown parameters

**Example:** logistic regression model

$$
\pi_{2i} = \frac{\exp(x_i'\alpha)}{1 + \exp(x_i'\alpha)}
$$

Estimated response probability for unit $i$: 

$$
\hat{\pi}_{2i} = m(x_i, \hat{\alpha})
$$

**Special case:**

- $x_i$ is a vector of weighting class indicators
  - weight adjustment by the inverse of the within-class response rate
Nonresponse model

- In practice, the response probability $\pi_{2i}$ is unknown
- **Parametric nonresponse model**: $\pi_{2i} = m(x_i, \alpha)$, where
  - $m(.)$ is a known function
  - $x_i$ is a vector of auxiliary variables available for all the sampled units (respondents and nonrespondents)
  - $\alpha$ is a vector of unknown parameters
- **Example: logistic regression model**

$$
\pi_{2i} = \frac{\exp (x_i' \alpha)}{\exp (1 + x_i' \alpha)}
$$
Nonresponse model

- In practice, the response probability $\pi_{2i}$ is unknown
- **Parametric nonresponse model:** $\pi_{2i} = m(x_i, \alpha)$, where
  - $m(.)$ is a known function
  - $x_i$ is a vector of auxiliary variables available for all the sampled units (respondents and nonrespondents)
  - $\alpha$ is a vector of unknown parameters
- **Example: logistic regression model**
  $$\pi_{2i} = \frac{\exp(x_i'\alpha)}{\exp(1 + x_i'\alpha)}$$
- Estimated response probability for unit $i$: $\hat{\pi}_{2i} = m(x_i, \hat{\alpha})$
Nonresponse model

- In practice, the response probability $\pi_{2i}$ is unknown

**Parametric nonresponse model:** $\pi_{2i} = m(x_i, \alpha)$, where
  - $m(.)$ is a known function
  - $x_i$ is a vector of auxiliary variables available for all the sampled units (respondents and nonrespondents)
  - $\alpha$ is a vector of unknown parameters

**Example: logistic regression model**

$$\pi_{2i} = \frac{\exp(x'_i \alpha)}{\exp(1 + x'_i \alpha)}$$

- Estimated response probability for unit $i$: $\hat{\pi}_{2i} = m(x_i, \hat{\alpha})$

**Special case:** $x_i$ is a vector of weighting class indicators $\Rightarrow$ weight adjustment by the inverse of the within-class response rate
Nonresponse model

- In practice, the response probability $\pi_{2i}$ is unknown
- **Parametric nonresponse model**: $\pi_{2i} = m(x_i, \alpha)$, where
  - $m(.)$ is a known function
  - $x_i$ is a vector of auxiliary variables available for all the sampled units (respondents and nonrespondents)
  - $\alpha$ is a vector of unknown parameters
- **Example**: logistic regression model
  \[
  \pi_{2i} = \frac{\exp(x_i'\alpha)}{\exp(1 + x_i'\alpha)}
  \]
- Estimated response probability for unit $i$: $\hat{\pi}_{2i} = m(x_i, \hat{\alpha})$
- **Special case**: $x_i$ is a vector of weighting class indicators $\Rightarrow$ weight adjustment by the inverse of the within-class response rate
Nonresponse model

- Propensity score adjusted estimator: 
  \[ \hat{Y}_{PSA} = \sum_{i \in s_2} \frac{y_i}{\pi_1 \hat{\pi}_2} \]
Nonresponse model

- Propensity score adjusted estimator: \( \hat{Y}_{PSA} = \sum_{i \in s_2} \frac{y_i}{\pi_1 \hat{\pi}_{2i}} \)
- One can show that

\[
\hat{Y}_{PSA} - \hat{Y}_L = O_p(n^{-1}),
\]

where \( \hat{Y}_L \) is the linearized version of \( \hat{Y}_{PSA} \).
Nonresponse model

- Propensity score adjusted estimator: \( \hat{Y}_{\text{PSA}} = \sum_{i \in s_2} \frac{y_i}{\pi_1 \hat{\pi}_2} \)
- One can show that

\[ \hat{Y}_{\text{PSA}} - \hat{Y}_L = O_p(n^{-1}), \]

where \( \hat{Y}_L \) is the linearized version of \( \hat{Y}_{\text{PSA}} \).
- Asymptotic conditional bias of a responding unit:

\[ B^L_i(l_{1i} = 1, l_{2i} = 1) = E_1 E_2(\hat{Y}_L - Y | l_1, l_{1i} = 1, l_{2i} = 1) \]
Nonresponse model

- Propensity score adjusted estimator: \( \hat{Y}_{PSA} = \sum_{i \in s_2} \frac{y_i}{\pi_1 \hat{\pi}_2} \)
- One can show that

\[ \hat{Y}_{PSA} - \hat{Y}_L = O_p(n^{-1}), \]

where \( \hat{Y}_L \) is the linearized version of \( \hat{Y}_{PSA} \).

- Asymptotic conditional bias of a responding unit:

\[ B_i^L(l_1i = 1, l_2i = 1) = E_1 E_2(\hat{Y}_L - Y|l_1, l_1i = 1, l_2i = 1) \]

- Robust version of \( \hat{Y}_{PSA} \)

\[ \hat{Y}_{PSA}^R = \hat{Y}_{PSA} - \sum_{i \in s_2} \hat{B}_i^{PSA}(l_1i = 1, l_2i = 1) + \sum_{i \in s_2} \psi \left\{ \hat{B}_i^{PSA}(l_1i = 1, l_2i = 1) \right\} \]
Simulation study

- We generated a population of size $N = 10000$ with two variables: $y$ and $x$
Simulation study

- We generated a population of size $N = 10000$ with two variables: $y$ and $x$
- $x \sim Gamma$
Simulation study

- We generated a population of size $N = 10000$ with two variables: $y$ and $x$
- $x \sim \text{Gamma}$
- Mixture model: $y_i = \delta_i \times (100 + xi + 5\epsilon_i) + (1 - \delta_i) \times (400 + xi + 50\epsilon_i)$
We generated a population of size \( N = 10000 \) with two variables: \( y \) and \( x \).

\[ x \sim \text{Gamma} \]

Mixture model:

\[ y_i = \delta_i \times (100 + xi + 5\epsilon_i) + (1 - \delta_i) \times (400 + xi + 50\epsilon_i) \]

\( \epsilon_i \sim N(0, 1) \)
Simulation study

- We generated a population of size $N = 10000$ with two variables: $y$ and $x$
- $x \sim \text{Gamma}$
- Mixture model: $y_i = \delta_i \times (100 + xi + 5\epsilon_i) + (1 - \delta_i) \times (400 + xi + 50\epsilon_i)$
- $\epsilon_i \sim N(0, 1)$
- 5% contamination: i.e., $P(\delta_i = 1) = 0.95$
Simulation study

- We generated a population of size $N = 10000$ with two variables: $y$ and $x$
- $x \sim \text{Gamma}$
- Mixture model: $y_i = \delta_i \times (100 + xi + 5\epsilon_i) + (1 - \delta_i) \times (400 + xi + 50\epsilon_i)$
- $\epsilon_i \sim N(0, 1)$
- 5% contamination: i.e., $P(\delta_i = 1) = 0.95$
- Select $R = 10000$ samples, of size $n = 500$, according to simple random sampling without replacement
Simulation study

- We generated a population of size \( N = 10000 \) with two variables: \( y \) and \( x \)
- \( x \sim \text{Gamma} \)
- Mixture model: 
\[
y_i = \delta_i \times (100 + xi + 5\epsilon_i) + (1 - \delta_i) \times (400 + xi + 50\epsilon_i)
\]
- \( \epsilon_i \sim N(0, 1) \)
- 5\% contamination: i.e., \( P(\delta_i = 1) = 0.95 \)
- Select \( R = 10000 \) samples, of size \( n = 500 \), according to simple random sampling without replacement
- Generate nonresponse: Bernoulli trials with probability \( \pi_{2i} \), where
\[
\pi_{2i} = \frac{1}{\exp(\alpha_0 + \alpha_1 x_i)}
\]
- Global response rate: 70\%
Simulation study

- We computed: $\hat{Y}_{PSA}$ and $\hat{Y}_{PSA}^R$
Simulation study

- We computed: \( \hat{Y}_{PSA} \) and \( \hat{Y}_{PSA}^R \)
- \( \hat{\pi}_{2i} \): estimated using a logistic regression model with \( x \) as a predictor
Simulation study

- We computed: \( \hat{Y}_{PSA} \) and \( \hat{Y}_{PSA}^R \)
- \( \hat{\pi}_{2i} \): estimated using a logistic regression model with \( x \) as a predictor
- Monte Carlo measures:
  - Monte Carlo percent Relative Bias:
    
    \[
    RB(\hat{Y}) = \frac{1}{10000} \sum_{t=1}^{10000} (\hat{Y}_t - Y)
    \]
Simulation study

- We computed: \( \hat{Y}_{PSA} \) and \( \hat{Y}_{PSA}^R \)
- \( \hat{\pi}_{2i} \): estimated using a logistic regression model with \( x \) as a predictor
- Monte Carlo measures:
  - Monte Carlo percent Relative Bias:
    \[
    RB(\hat{Y}) = \frac{1}{10000} \sum_{t=1}^{10000} (\hat{Y}_t - Y)
    \]
  - Relative Efficiency with respect to the nonrobust estimator:
    \[
    RE(\hat{Y}_{PSA}^R) = \frac{MSE(\hat{Y}_{PSA}^R)}{MSE(\hat{Y}_{PSA})}
    \]
- Note: \( \hat{Y}_{PSA} \) has negligible bias
Relative bias of the robust estimator (5% contamination)
Relative efficiency with respect to the nonrobust estimator (5% contamination)
Concluding remarks

- Conditional bias: measure of influence that takes account of the sampling design, the parameter to be estimated and the estimator.
Concluding remarks

- Conditional bias: measure of influence that takes account of the sampling design, the parameter to be estimated and the estimator.
- If the invariance property does not hold, it is still possible to assess the influence of a sampled unit and construct robust estimators.

Results can be extended to the case of calibration estimators, which are important in the unit nonresponse context since weight adjustment procedures by the inverse of the estimated response probabilities are generally followed by some form of calibration. Further investigations are required:
- Choice of the tuning constant.
- MSE estimation: reverse framework for variance estimation?
Concluding remarks

- Conditional bias: measure of influence that takes account of the sampling design, the parameter to be estimated and the estimator.
- If the invariance property does not hold, it is still possible to assess the influence of a sampled unit and construct robust estimators.
- Results can be extended to the case of calibration estimators ⇒ important in the unit nonresponse context since weight adjustment procedures by the inverse of the estimated response probabilities are generally followed by some form of calibration.
Concluding remarks

- Conditional bias: measure of influence that takes account of the sampling design, the parameter to be estimated and the estimator.
- If the invariance property does not hold, it is still possible to assess the influence of a sampled unit and construct robust estimators.
- Results can be extended to the case of calibration estimators ⇒ important in the unit nonresponse context since weight adjustment procedures by the inverse of the estimated response probabilities are generally followed by some form of calibration.
- Requires further investigations:
  - Choice of the tuning constant
  - MSE estimation: reverse framework for variance estimation?