Correcting for misclassification under edit restrictions in combined survey-register data

Laura Boeschoten\textsuperscript{12}, Daniel Oberski\textsuperscript{3}, Ton de Waal\textsuperscript{12} \& Marcel Croon\textsuperscript{1}

1 Tilburg University
2 Statistics Netherlands
3 Utrecht University

April 2017
We developed a method that corrects for misclassification in combined datasets and simultaneously handles edit restrictions: MILC.
Key Messages

- We developed a method that corrects for misclassification in combined datasets and simultaneously handles edit restrictions: MILC.
- The method is based on a combination of Latent Class (LC) analysis and Multiple Imputation (MI).
Key Messages

- We developed a method that corrects for misclassification in combined datasets and simultaneously handles edit restrictions: MILC.
- The method is based on a combination of Latent Class (LC) analysis and Multiple Imputation (MI).
- Variables can be added to the LC model at a later stage.
Key Messages

- We developed a method that corrects for misclassification in combined datasets and simultaneously handles edit restrictions: MILC.
- The method is based on a combination of Latent Class (LC) analysis and Multiple Imputation (MI).
- Variables can be added to the LC model at a later stage.
- The LC model can be extended in many ways.
Why use the MILC method?

- Combined sources are often used by researchers.
Why use the MILC method?

- Combined sources are often used by researchers.
- Both surveys and registers always contain some misclassification.
Why use the MILC method?

- Combined sources are often used by researchers.
- Both surveys and registers always contain some misclassification.
- The causes for misclassification in registers and in surveys can be very different.
Why use the MILC method?

- Combined sources are often used by researchers.
- Both surveys and registers always contain some misclassification.
- The causes for misclassification in registers and in surveys can be very different.
- Edit restrictions can be applied simultaneously.
Step 1: Combined survey-register dataset

- Misclassification in surveys and registers can be detected by variables in each dataset measuring the same attribute.
- These variables can be considered indicators measuring a latent (unobserved) ‘true’ variable.
Step 2: Bootstrap samples

- $m$ bootstrap samples are drawn from the original combined dataset.
- The bootstrap samples are taken to incorporate parameter uncertainty.
- $m = 5$ is generally enough to obtain valid inference.
Step 3: The LC model

- For each bootstrap sample, an LC model is constructed to measure latent variable $X$ using indicators $Y$.
- The number of categories of $X$ is equal to the number of categories in the indicators.
- Covariates $Q$ are also included.
- We specify the model in Latent GOLD, which uses ML for parameter estimation.
Step 3: Incorporating edit restrictions into the LC model

If $Q$ is a covariate and the combination of scores $X = 1$ and $Q = 2$ is impossible in practice, the corresponding edit restriction is:

$$P(X = 1|Q = 2) = 0$$
Last steps...

4. Multiple imputations
5. Obtain estimates of interest
6. Pool the estimates using Rubin’s rules
When can we use MILC

- Your data needs to be of a certain quality.
- Latent GOLD output gives you the entropy $R^2$, which is a value between 0 and 1 that indicates how well you can predict class membership based on the observed variables.
- For valid inference, you need an entropy $R^2$ of 0.90 or higher.
The three-step approach

- The imputed variable \( (W) \) is used in further analysis with other variables.
The three-step approach

- The imputed variable \((W)\) is used in further analysis with other variables.
- Although we investigate the relation between \(W\) and covariates \(Q\), we are actually interested in the relation between latent ‘true’ variable \(X\) and \(Q\).
The three-step approach

- The imputed variable \((W)\) is used in further analysis with other variables.
- Although we investigate the relation between \(W\) and covariates \(Q\), we are actually interested in the relation between latent ‘true’ variable \(X\) and \(Q\).
- \(W\) is not equal to \(X\), there is some misclassification.
Two methods to update the LC model

1. ML approach
By using $W$ as an indicator for $X$, a new LC model is created:

$$P(W = w, Q = q) = \sum_{x=1}^{C} P(X = x, Q = q)P(W = w|X = x)$$  \hspace{1cm} (1)
Two methods to update the LC model

1. ML approach
   By using $W$ as an indicator for $X$, a new LC model is created:

   $$
   P(W = w, Q = q) = \sum_{x=1}^{C} P(X = x, Q = q)P(W = w|X = x)
   $$

   (1)

2. BCH approach:
   We re-express:

   $$
   P(X = x, Q = q) = \sum_{w=1}^{W} P(W = w, Q = q)d_{wx}^*
   $$

   (2)

   where we weight the $W - Q$ distribution by the inverse of the classification errors.
Combining MILC and three-step

2. Bootstrap samples.
3. The LC model.
4. Multiple imputations.
5. **Acquire new information.**
6. **Updating the LC model.**
7. Creating new imputations.
8. Obtain estimates of interest.
Further research

1. Assumptions:
   - Observed indicators are independent of each other.
   - Misclassification is independent of covariates.
   - Covariates are free of error.
   - Edit rules applied are ‘hard edits’.
Further research

2. Missing data:

- Item non-response
- Unit non-response
Thank you!