Working Paper 2.

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(The Hague, Netherlands, 24-26 April 2017)

Correcting for misclassification under edit restrictions in combined survey-register data using Multiple Imputation Latent Class modelling (MILC)

Prepared by Laura Boeschoten, Tilburg University and Statistics Netherlands, Netherlands

I. Introduction

1. National Statistical Institutes (NSIs) often use large datasets to estimate population tables on many different aspects of society. A way to create these rich datasets as efficiently and cost effectively as possible is by utilizing already available register data. When more information is required than already available, registers can be supplemented with survey data (De Waal, 2015). Composite data containing both surveys and registers were for example used when constructing the different population tables for the 2011 Dutch Census (Schulte Nordholt et al., 2014).

2. However, caution is advised as both surveys and registers can contain classification errors. These can be detected when the separate datasets within the composite dataset contain variables measuring the same attribute, or when combinations of scores on different variables are in practice impossible.

3. To estimate the number of classification errors in a composite dataset and to simultaneously impute a new variable which takes the uncertainty caused by these classification errors into account, we developed a method which combines Multiple Imputation (MI) and Latent Class (LC) analysis (Boeschoten et al., 2016). With this method it is possible to obtain estimates that are consistent and that take edit rules into account, which is especially useful within official statistics since cells in cross tables that represent a combination of scores that is in practice impossible, should contain zero observations (De Waal, 2015).

4. However, if a researcher is interested in estimating a cross table between variable that has been imputed by the MILC method and another variable, this variable should be taken into account in the LC model as a covariate. A problem is that it is not always possible incorporate this covariate in the LC model at the moment the model was constructed. Reasons for this can be that not all necessary information was available at the time the LC model was constructed, or that the researcher was not interested in the relation with this covariate at that time. However, not incorporating this covariate in the LC model leads to biased results when this cross table is estimated.

5. When LC analysis is applied, a solution to this problem is found by using the ‘three-step’ approach. With this approach, a variable containing LC assignments is related to covariates that were not taken into account in the initial LC model. This is done by taking the relation between the assigned LCs and the `true' LCs into account (Bakk et al., 2013).

6. In this paper, we illustrate how we incorporated the three-step approach into the MILC method. In the next section, we provide some background information on the three-step approach and on the MILC method. In the methodology section we illustrate how we incorporated the three-step approach into the MILC method. Next, we discuss the setup and first results of a simulation study which investigates the performance of the MILC method in combination with the three-step approach.

II. Background

A. The MILC method

7. With the MILC method, we estimate the number of classification errors in a composite dataset and simultaneously incorporates edit restrictions. Further, a new variable is imputed which takes uncertainty caused by the classification errors as well as the edit rules into account and that can be used to obtain consistent estimates. On the left side of Figure 1, a graphical overview of the procedure of the MILC method is given. The method starts with the original composite dataset comprising *L* measures of the same attribute of interest. In the **first** step, *m* bootstrap samples are taken from the original composite dataset.

8 In the **second** step, an LC model is estimated for every bootstrap sample. Here we use the *L* indicator variables ($Y_1,\ldots,Y_L$) of the latent property *X* (which has *C* categories and a specific category is denoted by *x* ($x=1,\ldots,C$)). Covariates of interest (here denoted by *Q* and *Z*) are also incorporated in the LC model. The LC model for response pattern $P(Y=y|Q=q,Z=z)$ is then estimated as:

$$P(Y=y|Q=q,Z=z)=\sum_{x=1}^{C}P(X=x|Q=q,Z=z)\prod_{l=1}^{L}P(Y_l=y_l|X=x) \quad (1)$$

(Vermunt & Magidson, 2004). Edit restrictions can be implemented within the LC model, leading to constrained parameter estimation. An example of such a restriction can be:

$$P(X=1|Z=2)=0\,. \quad (2)$$

9. In the **third** step, *m* new empty variables are created in the original dataset. The *m* empty variables are imputed using the posterior membership probabilities obtained from the corresponding *m* LC models:

$$P(X=x|Y=y,Q=q,Z=z)=\frac{P(X=x|Q=q,Z=z)\prod_{l=1}^{L}P(Y_l=y_l|X=x)}{\sum_{x=1}^{C}P(X=x|Q=q,Z=z)\prod_{l=1}^{L}P(Y_l=y_l|X=x)}\,. \quad (3)$$

These represent the probability that a unit is member of a latent class given the combination of scores on the indicators.

10. In the **fourth** step, estimates of interest are obtained from the *m* variables and in the **fifth** step, the estimates are pooled using Rubin's rules for pooling (Rubin, 1987, p.76). The pooled estimate is obtained by:

$$\hat{\theta}=\frac{1}{m}\sum_{y=1}^{Y}\hat{\theta_i}\quad . \quad (4)$$

The total variance is estimated as

$$VAR_{total}=\overline{VAR}_{within}+VAR_{between}+\frac{VAR_{between}}{m}, \quad (5)$$

where $\overline{VAR}_{within}$ is the within imputation variance calculated by

$$\overline{VAR}_{within}=\frac{1}{m}\sum_{i=1}^{m}VAR_{within_i}\ , \quad (6)$$

and $VAR_{between}$ is estimated by:

$$VAR_{between}=\frac{1}{m-1}\sum_{i=1}^{m}(\hat{\theta_i}-\hat{\theta})(\hat{\theta_i}-\hat{\theta})'\,. \quad (7)$$

Besides the uncertainty caused by missing or conflicting data represented by the spread of parameter estimate values, $VAR_{between}$ also contains parameter uncertainty, which was introduced by the bootstrap performed in the first step of the MILC method (Van der Palm et al, 2016).

B. The three step approach

11. Within the MILC method, an LC model is specified incorporating both the measurement model for the latent variable as well as the structural model describing the relation of the latent variable with covariates. With the three step approach, these two models are estimated in separate steps. In the *first step*, the measurement model for the relationship between the latent variable and its indicators is built:

$$P(Y = y) = \sum_{x=1}^{C} P(X = x) \prod_{l=1}^{L} P(Y_l = y_l | X = x)\,, \tag{8}$$

from which posterior membership probabilities can be obtained:

$$P(X = x | Y = y) = \frac{P(X=x)\prod_{l=1}^{L} P(Y_l = y_l | X = x)}{\sum_{x=1}^{C} P(X=x)\prod_{l=1}^{L} P(Y_l = y_l | X = x)}\,. \tag{9}$$

These represent the probability that a unit is member of a latent class given the combination of scores on the indicators. When drawing from these probabilities, a new imputed variable can be created, *W*. Creating imputed variable *W* can be considered as the *second step*.

12. In the *third step*, the predicted class membership variable *W* is used in further analysis with other variables. Although we investigate the relation between imputed variable *W* and covariates (which we denote by *Z*), we are actually interested in the relation between the `true' latent variable *X* and *Z*. *W* is not exactly equal to *X* (there is some classification error) and this should be taken into account. This can be done by using information about how the *X*×*Z* distribution is related to the *W*×*Z* distribution. Therefore, we specify an LC model where we use *W* as an indicator of *X* and define the form of the *X*-*Z* distribution:

$$P(W = w, Z = z) = \sum_{x=1}^{C} P(X = x, Z = z) P(W = w | X = x) \tag{10}$$

We assume here that *Z* is independent of *Y* given *X*, and we see that the *W* and *Z* distribution are weighted sums of the entries in the *X* and *Z* distribution, where the weights are the misclassification probabilities *P(W=w|X=x)*. The relationship between *W* and *Z* can be obtained by adjusting the relationship between *W* and *Z* for the misclassification probabilities *P(W=w|X=x)* (Vermunt, 2010). *P(W=w|X=x)* gives the probability of a certain class assignment conditional on the true class and is a quantification of the overall quality of the classification obtained from the LC model in the first step (Bolck et al., 2004). The larger the probability for *w=x*, the better the classification. Using the LC parameters, which can be found by maximising the likelihood, this quantity can be obtained as follows:

$$P(W = w | X = x) = \sum_{y=1}^{Y} \frac{P(Y=y)P(X=x|Y=y)P(W=w|Y=y)}{P(X=x)} \tag{11}$$

Adjusting the relationship between *W* and *Z* in this manner is also known as the Maximum Likelihood (ML) approach.

13. Another option is to use the Bolck-Croon-Hagenaars (BCH) approach [6]. With the BCH approach, the relationship described in equation 3 is re-expressed as follows:

$$P(X = x, Z = z) = \sum_{w=1}^{W} P(W = w, Z = z) d^{*}_{wx}\,, \tag{12}$$

where d^{*}_{wx} represents an element of the inverted *x*×*x* matrix *D* with elements *P(W=w|X=x)*. We weight the *W*-*Z* distribution by the inverse of the classification errors to obtain the distribution we are interested in.

III. Methodology

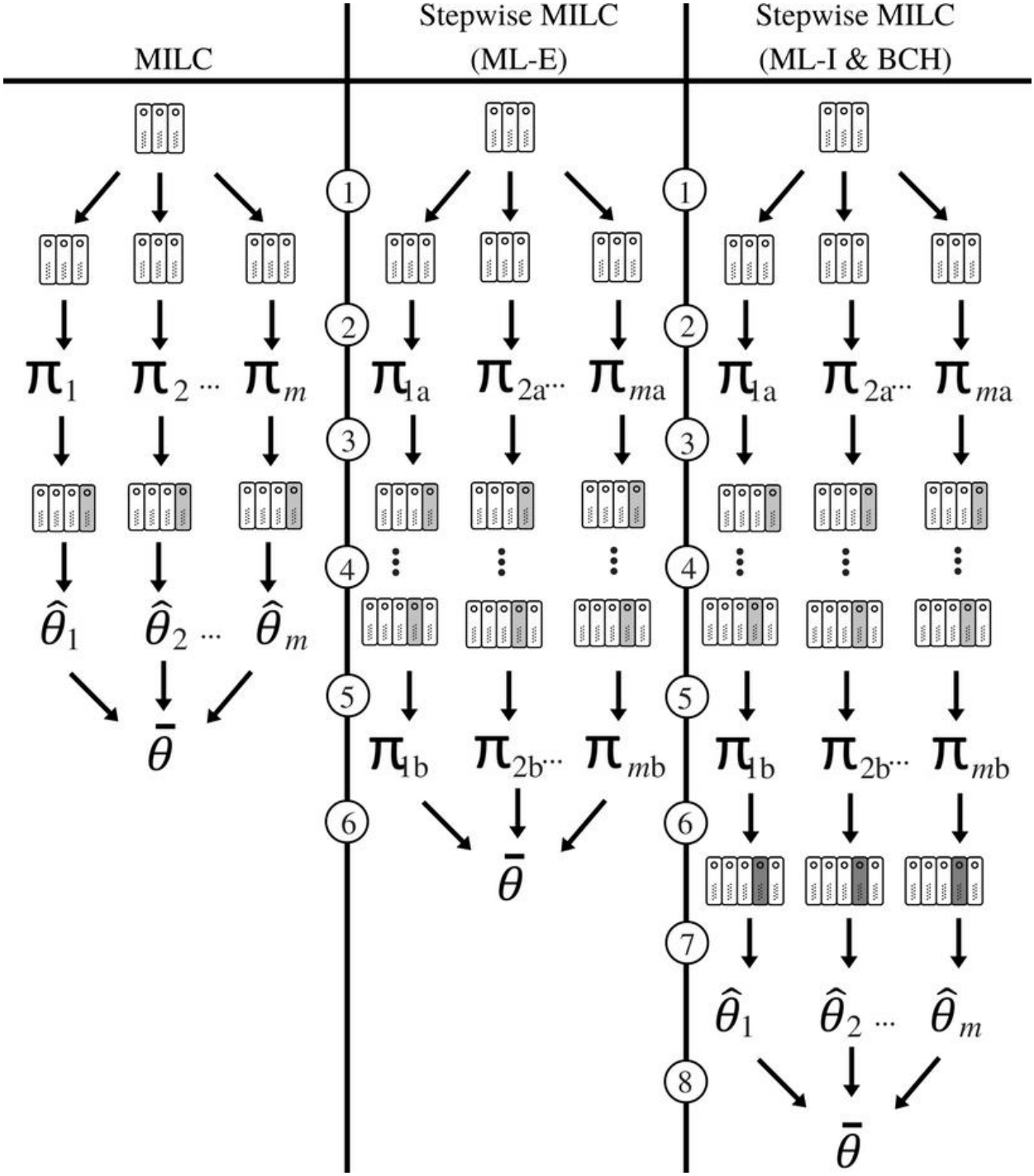

Figure 1 Systematic overview of the MILC method, and three alternative versions to incorporate the three-step approach into the MILC method: ML-E, ML-I and BCH.

14. In the middle and right column of Figure 1, two alternative approaches for incorporating the three step approach into the MILC method are shown. For both approaches, in the **first** step, *m* bootstrap samples are taken from the composite dataset. This step is not different from the original MILC method. In the **second** step, an LC model is estimated for every bootstrap sample. Different from the original MILC method here is that this LC model contains only the measurement model for the relationship between the latent variable and its indicators, as was described by equation 8. In the **third** step, *m* new empty variables are created in the original dataset and are imputed using the posterior membership probabilities obtained from the corresponding *m* LC models.

15. At this point, it is important to also save the posterior membership probabilities and the classification probabilities as described by equation 9 and 11. In the **fourth** step, newly obtained covariate information is included in the composite dataset. In the **fifth** step, an LC model is specified where the imputed variable W is used as an indicator of the latent 'true' variable X. This is done either with the ML approach (as described in equations 10 and 11) or with the BCH approach (as described in equation 12). If the updated parameter estimates are sufficient to answer the research questions of interest, these can be obtained directly from the LC output and be pooled using Rubin's rules (as described by step **six** in the middle column of Figure 1).

16. It is also possible to update the imputations and to acquire a 'new' W variable, this is described by the right column in Figure 1. The **sixth** step will then be to create new imputations using the posterior membership probabilities from the LC models created in step **five**. Estimates of interest can then be obtained in the **seventh** step, and in the **eight** step, these can be pooled again using the pooling rules defined by Rubin.

IV. Simulation Study

A. Simulation Setup

17. We use a composite dataset with three measures (Y_1, Y_2, Y_3) of the property of interest (X). We also have two covariates, Q and Z. We are interested in the bias and the coverage of the 95% confidence interval of the logit coefficients of X regressed on Q. Furthermore, Z is a restriction covariate, i.e. a covariate for which some combinations with X are not possible in practice. In our case, we assume that $W_1 \times Z_2$ is impossible. Therefore we are interested in how often the combination of scores $W_1 \times Z_2$ occurs, which is in practice impossible.

18. We compare four different methods to estimate these relations. As a baseline, we investigate the logit coefficient of Q and the frequency of $W_1 \times Z_2$ after applying the MILC method when only the indicators Y_1, Y_2 and Y_3 are incorporated in the LC model (1). Next, we investigate the estimates when the three step approach is incorporated into the MILC method. Here we investigate the estimates obtained directly from the new LC model used to apply the ML approach (2). We use the posterior membership probabilities obtained from the LC model where the ML approach was applied to impute a new variable W and we investigate the estimates obtained using the new W (3). At last, we investigate the estimates when W was imputed by making use of posterior membership probabilities estimated using the BCH approach (4).

19. Furthermore, we manipulated properties of the composite dataset to empirically evaluate the performance of the four methods described. The main properties of this simulation study are summarized as follows:

- Different classification probabilities of the three dichotomous indicators of X (Y_1, Y_2, Y_3): 0.70; 0.80; 0.90; 0.95; 0.99.
- Different logit coefficients of X regressed on Q: 0.00; 0.50; 1.00; 2.00.
- Different values for $P(Z=2)$, where $W_1 \times Z_2$ is the restricted cell which should contain zero observations. $P(Z=2)$=0.05; 0.10; 0.20.
- Sample size is 200.

B. Simulation Results

20. In Figure 2, we see the bias of the logit coefficient of latent variable X regressed on covariate Q under the different simulation conditions. Here, we compare 'no correction', where covariate Q was not taken into account by the LC model for X, with three different methods to apply the three-step approach. When investigating the figure, it is immediately clear that when using no correction, bias is introduced when estimating the relation between X and Q. Bias is larger when the classification probabilities are low, and the bias is also larger when the coefficients are larger. When the correction methods are applied, we also detect bias when the classification probabilities are low, and this bias is also more severe for larger

coefficients, although it is much smaller compared to the bias when no correction is applied. From the correction methods, BCH produces the largest bias when the classification probabilities are *0.70*, but for higher classification probabilities, the bias is the smallest for BCH.

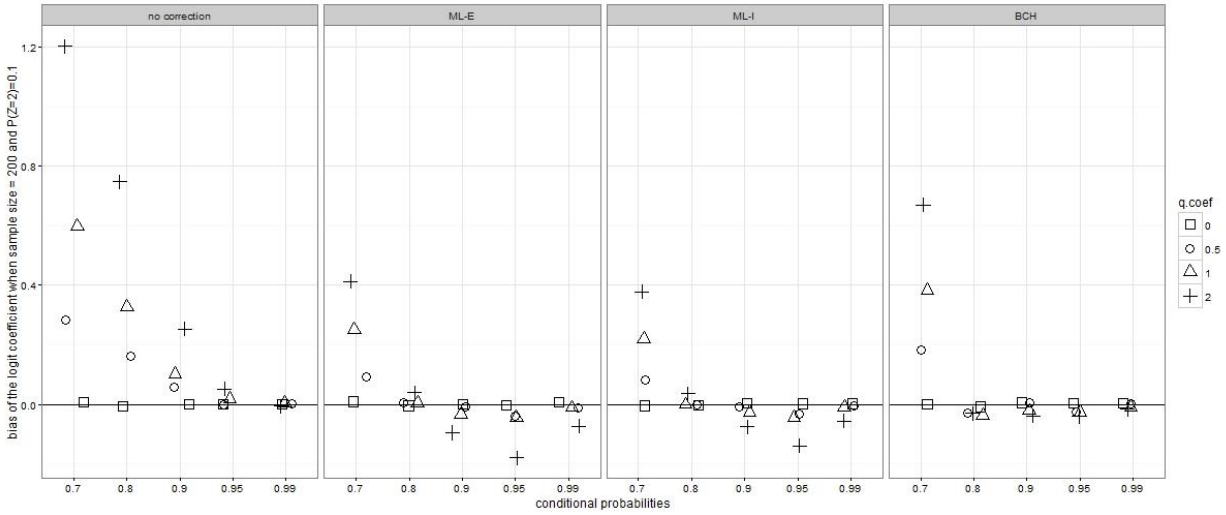

Figure 2 Bias of the logit coefficient of latent variable *X* regressed on covariate *Q* in situations with different values for the classification probabilities and values for the coefficient itself. Sample size is 200 and *P(Z=2)=0.10*.

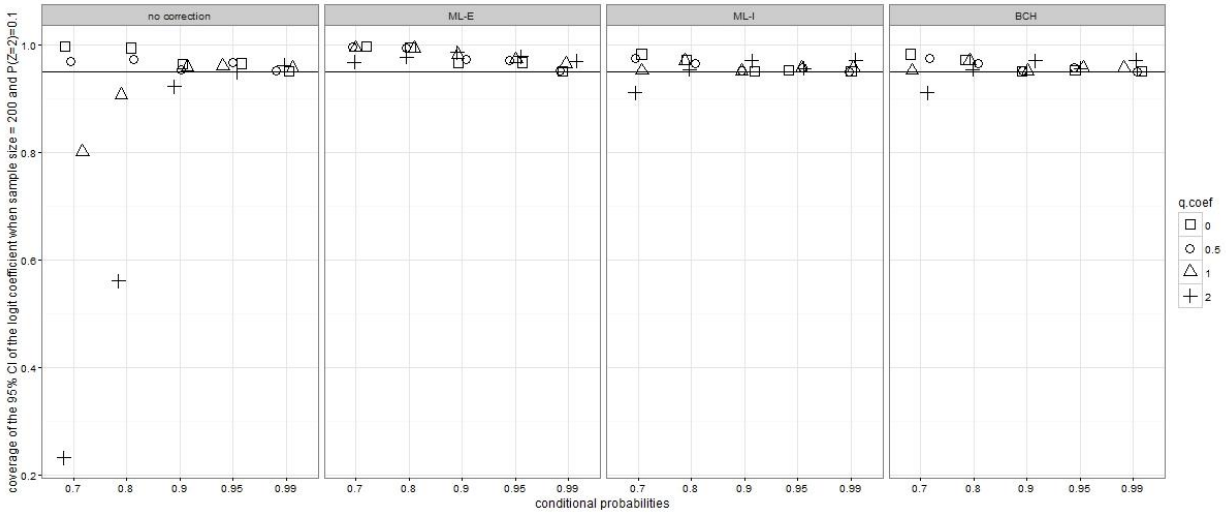

Figure 3 Coverage of the 95% confidence interval of the logit coefficient of latent variable *X* regressed on covariate *Q* in situations with different values for the classification probabilities and values for the coefficient itself. Sample size is 200 and *P(Z=2)=0.10*.

21. In Figure 2, we see the coverage of the 95% confidence interval of the logit coefficient of latent variable *X* regressed on covariate *Q* under the different simulation conditions. Here, we compare 'no correction', where covariate *Q* was not taken into account by the LC model for *X*, with three different methods to apply the three-step approach. When investigating the figure, we see that we deal with undercoverage when no correction is applied when the coefficient is large and the classification probabilities are low (*0.70* and *0.80*). When we apply the different bias correction methods, we see that often a small amount of overcoverage is produced. Overall, the coverage for ML-I (using the ML approach and impute a new variable *W* using the newly obtained posterior membership probabilities) is closest to the nominal rate of *0.95*.

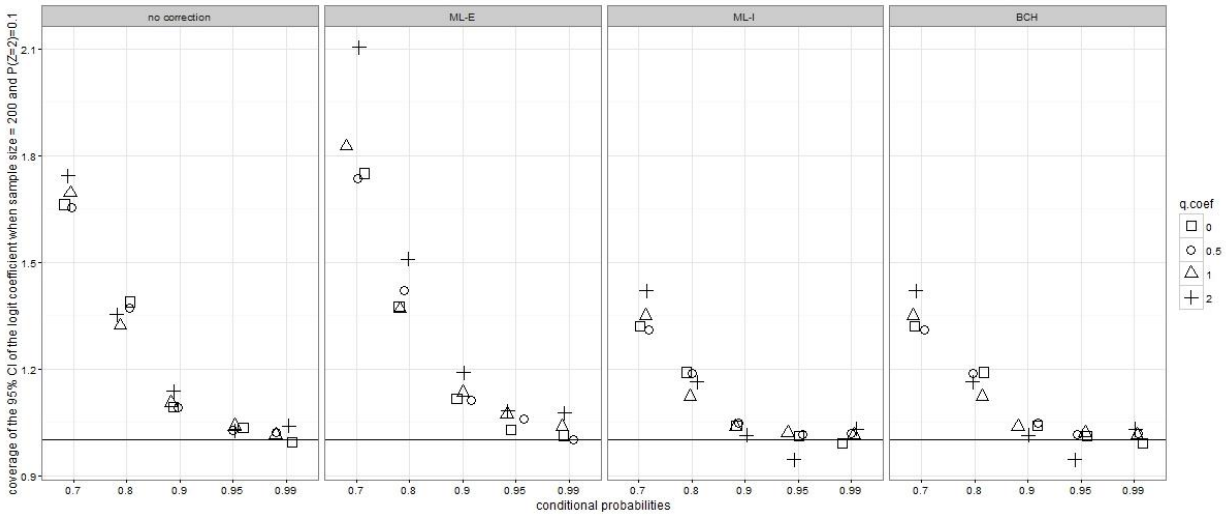

Figure 4 Ratio of the average standard error over the estimates over the standard deviation of the estimates of the logit coefficient of latent variable *X* regressed on covariate *Q* in situations with different values for the classification probabilities and values for the coefficient itself. Sample size is 200 and *P(Z=2)=0.10*.

22. In Figure 4, it can be seen that the ratio of the average standard error over the estimates of the standard deviation of the estimates is smallest over all classification probabilities for the ML-I and BCH methods. ML-E apparently does not improve the standard errors compared to applying no correction. This is probably caused by the fact that sampling variance is not incorporated in the third step of the three-step approach here, which is the case with ML-I and BCH. Furthermore, the standard errors become closer to the nominal rate of 1 when the classification probabilities increase.

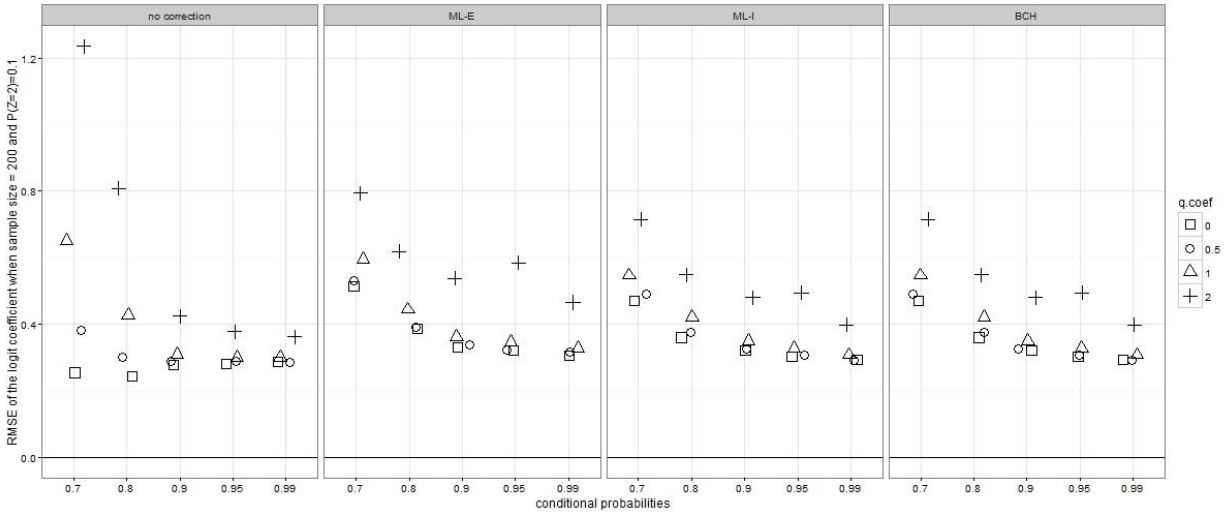

Figure 5 Root mean squared error of the logit coefficient of latent variable *X* regressed on covariate *Q* in situations with different values for the classification probabilities and values for the coefficient itself. Sample size is 200 and *P(Z=2)=0.10*.

23. In Figure 5, results in terms of root mean squared error are shown. We see that when no correction is applied, the root mean squared error is especially high when the classification probabilities are low, and decrease as the classification probabilities increase. As the classification probabilities increase (and the quality of the data increases), the RMSE values when no correction is applied become lower compared to when correction methods are used, which makes sense because we investigate bias due to measurement error here, if there is no measurement error, there is no bias as well, and applying a correction method will therefore not be necessary. ML-I is the correction method which has the lowest RMSE values overall.

24. In Figure 6, we see the frequency of $W_1 \times Z_2$, which is a combination of scores that cannot happen in practice. To make sure that combinations of scores like these do not occur within our dataset, we make

use of edit restrictions. In Figure 6, we see the frequencies when $P(Z_2)=0.10$, the logit coefficient of X regressed on Q is *2* and the sample size is *200*. When no corrections are applied, the covariates are not incorporated in the LC model and the edit restriction is therefore also not incorporated. We can see this clearly from the histogram, because the impossible combination is always found, more often when the classification probabilities are lower. With ML-E, the impossible combination is under no conditions found in the dataset, while with ML-I, the impossible combination can be found a small number of times. At first sight, this seems unexpected since we explicitly incorporated the edit restriction in the LC model. However, when specifying the model in Latent Gold (Vermunt & Magidson, 2013), we specified Bayes constants to obtain more efficient results for the logit coefficients shown in Figures 2 to 5. Bayes constants help to prevent boundary issues, which is often very useful. However, an edit restriction is a boundary itself. When using ML instead of Bayesian estimation in Latent Gold will result in exactly zero observations here. For BCH, we see a small number of observations when the classification probabilities are *0.70*, and no observations for all other conditions.

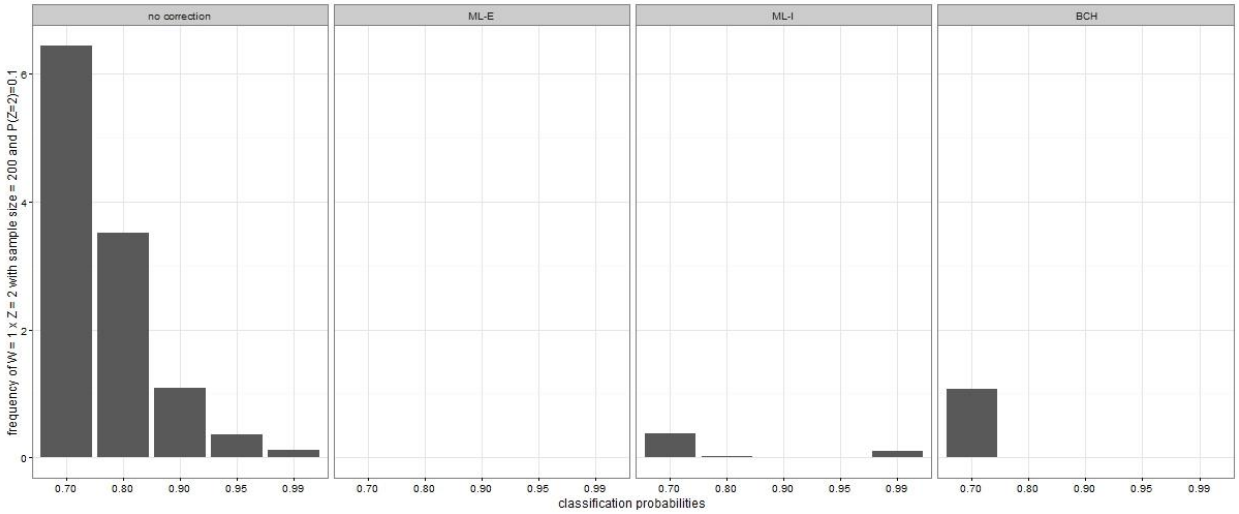

Figure 6 Frequency of $W_1 \times Z_2$ in situations with different values for the classification probabilities. The results shown are produced by a model where the logit coefficient of X regressed on Q is *2* and $P(Z_2)=0.10$.

V. Conclusions

23. The simulation results have shown that applying no correction leads to biased estimates of logit coefficients and undercoverage of the 95% confidence interval when the classification probabilities are below *0.90*. All correction methods perform much better compared to when no correction method is used. Furthermore, edit restrictions are also not incorporated in the model without correction and possibly leads to large numbers of scores that cannot be obtained in practice.

24. In the presentation we plan to present whether comparable results are obtained when the sample size is larger and when other values for $P(Z=2)$ are used. Furthermore, we are planning to apply the different correction methods to a dataset from Statistics Netherlands.

25. Previous research has shown that researchers can estimate the number of classification errors in a composite dataset and simultaneously incorporate edit restrictions using the MILC method. Since a new variable is imputed which takes uncertainty caused by the classification errors as well as the edit rules into account, this imputed variable can be used to obtain consistent estimates. We extended this method in such a way that edit restrictions, or relations in covariates in general, can now be incorporated into the model later on and consistent estimates can still be obtained.

VI. References

Z. Bakk, F.B. Tekle, and J.K. Vermunt, Estimating the association between latent class membership and external variables using bias adjusted three-step approaches, Sociological Methodology, vol.43, 1 (2013) 272-311.

L. Boeschoten, D. Oberski and T. de Waal, Estimating classification error under edit restrictions in combined survey-register data, CBS discussion paper (2016)

A. Bolck, M. Croon, and J.A. Hagenaars, Estimating latent structure models with categorical variables: One-step versus three-step estimators. Political Analysis, 12 (2004) 3-27.

T. de Waal, Obtaining numerically consistent estimates from a mix of administrative data and surveys. Statistical Journal of the IAOS (2015), 1-13.

D.B. Rubin, Multiple imputation for nonresponse in surveys, John Wiley and Sons, Vol. 81 (1987).

E. Schulte Nordholt, J. Van Zeijl & L. Hoeksma, Dutch Census 2011, analysis and methodology. The Hague/Heerlen (2014).

D.W. van der Palm, L.A. van der Ark & J.K. Vermunt, Divisive latent class modelling as a density estimation method for categorical data. Journal of Classification (2016) 1-21

J.K. Vermunt, Latent class modelling with covariates: Two improved three-step Approaches. Political Analysis, 18 (2010) 450-469.

J.K. Vermunt & J. Magidson, Latent class analysis. The sage encyclopedia of social sciences research methods (2004) 549-553.

J.K. Vermunt & J. Magidson, Technical guide for Latent GOLD 5.0: Basic, advanced, and syntax. Statistical Innovations Inc., Belmont, MA (2013).