Correcting for measurement error

Ton de Waal

Summer school 2023

Correcting for measurement error

- Based on work by former PhD student Laura Boeschoten

The situation

- We have several data sources that can be linked on unit level
- Same target variable is measured (with measurement error) in those data sources
- We aim to correct for measurement error
- We also aim to estimate accuracy of the corrected data

Possible methods

- Over-imputation
  - In over-imputation, imputation model is estimated and all values of target variable are imputed, including observed values
- Structural equation modelling (SEM)
  - SEM can be used to model each observed value as an imperfect measure of underlying latent (unobserved) variable
- Latent class modelling
  - Latent class modelling can be used to model each observed value of *categorical* variable as an imperfect measure of underlying latent (unobserved) variable with the true values

Examples

- Home-ownership: available in register and observed in survey
- (Un)employment: employment/unemployment available in register and observed in Labour Force Survey

Two elements in Laura's approach (MILC)

- Multiple Imputation
  - Uses bootstrap in background
- Latent Class modelling

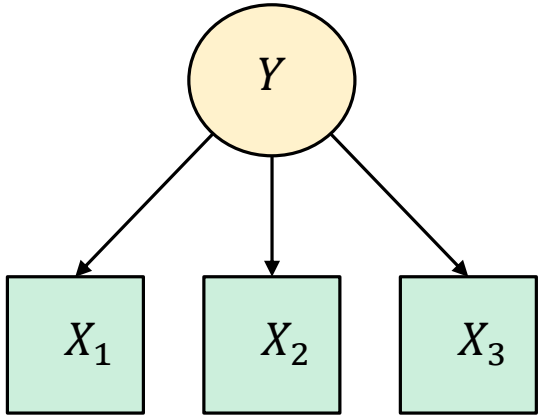
Model introduction

- Latent class analysis
  - Statistical modelling technique with many applications
  - In this workshop: focus on the use of latent class analysis for correcting for measurement errors in observed data
    - Requires multiple measurements for each unit

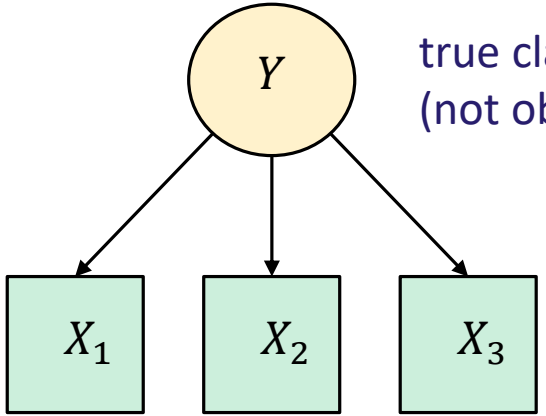
Model introduction

Latent class analysis (LCA)

- Measure something that cannot be measured directly (Y) by making use of (observed) indicators ($\boldsymbol{X}$)
- Traditional applications:
  - Personality traits
  - Disorder types (e.g. eating disorders)



Basic latent class model: parameters



true classification
(not observed)

observed classifications
(with measurement errors)

Notation:

- $\boldsymbol{X} = (X_1, X_2, X_3)$
- Generalize to s observed variables

Model parameters

- True class membership probabilities $\Pr(Y = y)$
- Error probabilities $\Pr(X_l = x_l | Y = y)$

Latent Class model for measurement error

- Let $\boldsymbol{X} = (x_1, x_2, \ldots, x_s)$ denote vector of observed categorical variables that measure same target variable (e.g., in s different datasets)
- True value with respect to variable of interest is represented by latent class variable Y

How does it work?

- Latent class model estimates its model parameters by looking at “majority votes”
  - If two out of three datasets (of equal quality) observe value a and third dataset another value b, value a is most likely to be correct
- Latent class model does the mathematics for you and helps with difficult decisions
  - For instance when value a is observed in two low-quality datasets and value b is observed in one high-quality dataset
  - You can add extra information to model, e.g. for longitudinal data that one dataset is generally more up-to-date than other dataset

Local independence assumption

- Distributions of the observed variables (X_1, …, X_s), are independent conditional on individuals' score on latent variable:

$$\Pr(\boldsymbol{X} = \boldsymbol{x}|Y = y) = \prod_{j=1}^{s} \Pr\left(X_j = x_j \middle| Y = y\right)$$

Mixture assumption

- Probability of obtaining a specific response pattern $\Pr(\boldsymbol{X} = \boldsymbol{x})$ is a weighted average of $\Pr(\boldsymbol{X} = \boldsymbol{x}|Y = y)$
- Interpretation: population has different groups (defined by Y), each with its own response pattern (measurement errors)
- LC model is given by

$$\Pr(\boldsymbol{X} = \boldsymbol{x}) = \sum_{y=1}^{L} \Pr(Y = y) \prod_{j=1}^{s} \Pr\left(X_j = x_j \middle| Y = y\right)$$

Posterior probabilities

- LC model can be used to estimate, for each unit in the data, probability of belonging to particular latent class, given its vector of observed values:

$$\Pr(Y = y|\boldsymbol{X} = \boldsymbol{x}) = \frac{\Pr(Y=y) \prod_{j=1}^{S} \Pr\left(X_j = x_j \middle| Y = y\right)}{\sum_{y'=1}^{L} \Pr(Y=y') \prod_{j=1}^{S} \Pr\left(X_j = x_j \middle| Y = y'\right)}$$

Conditioning on covariate

- You condition on covariate

$$\Pr(\boldsymbol{X} = \boldsymbol{x} | Q = q)$$

$$= \sum_{y=1}^{L} \Pr(Y = y | Q = q) \prod_{j=1}^{s} \Pr\left(X_j = x_j \middle| Y = y\right)$$

- You can also condition on several covariates
- $\Pr(\boldsymbol{X} = \boldsymbol{x} | Q = q, Z = z) =$
  $\sum_{y=1}^{L} \Pr(Y = y | Q = q, Z = z) \prod_{j=1}^{s} \Pr\left(X_j = x_j \middle| Y = y\right)$

Multiple imputation

- Create M completed data sets ($M \geq 2$) using appropriate imputation procedure
  - Each set consists of draws from predictive distribution of missing values
- Analyze each completed data set separately, treating all values as if they were observed
- Combine results by means of Rubin's pooling rules (Rubin, 1987): these rules allow one to estimate parameter of interest and its variance

How does it work?

- By creating several completed datasets and seeing how much variation there is between estimates based on those datasets, we get estimate for quality of these estimates

Rubin's pooling rules

- Suppose we want to estimate θ
  - $\hat{\theta}_m$= estimate of θ from m -th completed data set
  - V_m = estimate of variance of θ from m-th completed data set

Rubin's pooling rules

- MI estimate of scalar θ:

$$\bar{\theta}_M = \frac{1}{M}\sum_{m=1}^{M}\hat{\theta}_m$$

- MI estimate of variance

$$T_M = \bar{V}_M + \left(1 + \frac{1}{M}\right)\bar{B}_M$$

  - $\bar{V}_M = \frac{1}{M}\sum_{m=1}^{M} V_m$: within variance
  - $\bar{B}_M = \frac{1}{M-1}\sum_{m=1}^{M}(\hat{\theta}_m - \bar{\theta}_M)^2$: between variance

Multiple imputation: *proper* imputation

- In multiple imputation uncertainty in parameters of imputation model needs to be taken into account
- Imputation methods that take uncertainty in parameters of imputation model into account are called *proper*

Bootstrapping

- General approach to estimate measures of accuracy, for instance for (sampling) variance of
  - Population total or population mean
  - Model parameter
- Can be applied when sample of target population is available

Bootstrapping

- Typically used in two situations:
  - Data with an unknown distribution or one does not want to make an assumption about the distribution
  - Difficult to derive bias and variance from analytical formulas:
    - complex estimator
    - account for errors

Bootstrapping procedure

1. Repeatedly draw bootstrap samples with replacement from the original sample
2. For each bootstrap sample process your data in the same way as you did for the original sample to obtain estimate for target parameter
3. Estimate the variance and bias using results of step 2

How does it work?

- We want to estimate population parameters and/or their accuracy using samples from population
- We do not know true distribution of these parameters in population
- We do know distribution of parameters in observed sample
- Instead of drawing samples from population, we draw (bootstrap) samples from observed dataset

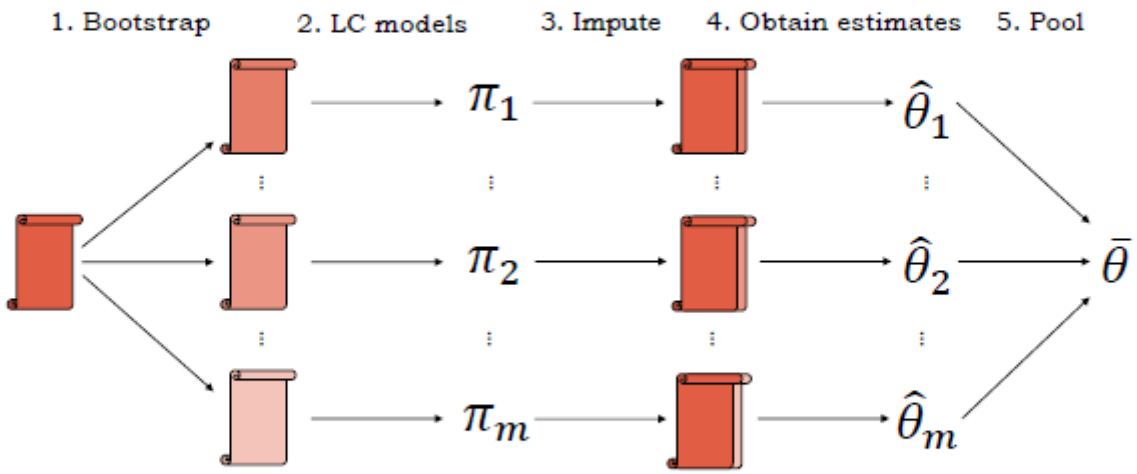
Bootstrapping

- Many different versions of bootstrapping
  - Non-parametric bootstrapping
  - Parametric bootstrapping
  - Pseudo-population bootstrapping
  - Bayesian bootstrapping
  - …

MILC (Multiple Imputation and LC analysis)

Link all data sets on unit level, and proceed with 5 steps:

1. Select M bootstrap samples from original dataset
2. Create LC model for every bootstrap sample
3. Multiply impute latent "true" variable Y for each bootstrap sample: M variables $(W_1, \dots, W_M)$ are created and imputed by drawing from posterior distribution $\Pr(Y = y|\boldsymbol{X} = \boldsymbol{x})$
4. Obtain estimates of interest from imputed variables
5. Pool the estimates using Rubin's rules for pooling (Rubin, 1987)


1. Bootstrap
2. LC models
3. Impute
4. Obtain estimates
5. Pool
π_1
π_2
π_m
$\hat{\theta}_1$
$\hat{\theta}_2$
$\hat{\theta}_m$
$\bar{\theta}$

MILC

- Bootstrapping is necessary to make imputation *proper*

MILC

- Probabilities for impossible combinations of values for target variable and background variable(s) are set to zero

Simulation study

- We created population containing five variables
  - Three dichotomous indicators (X_1; X_2; X_3) measuring latent dichotomous variable (Y)
  - Dichotomous covariate (Z) which has an impossible combination with a score of latent variable Y
  - Dichotomous covariate (Q)

Simulation study

- We considered three models
  - Conditional on Q only (“unconditional model”)
  - Conditional on Q and Z without preventing impossible combinations (“conditional model”)
  - Conditional on Q and Z with preventing impossible combinations (“restricted conditional model”)

Performance measures

- Interested in relation between imputed latent variable W and Z
  - Cell proportions of 2×2 table are denoted by: $W1 \times Z1$, $W2 \times Z1$, $W1 \times Z2$ and $W2 \times Z2$
  - $W1 \times Z2$ should contain 0 observations
  - We compare cell proportions of table $Y \times Z$ with cell proportions of $W \times Z$ from the samples
- We are also interested in relation between W and Q
  - We compare coefficient of logistic regression of Y on Q with logistic regression coefficient of W regressed on Q

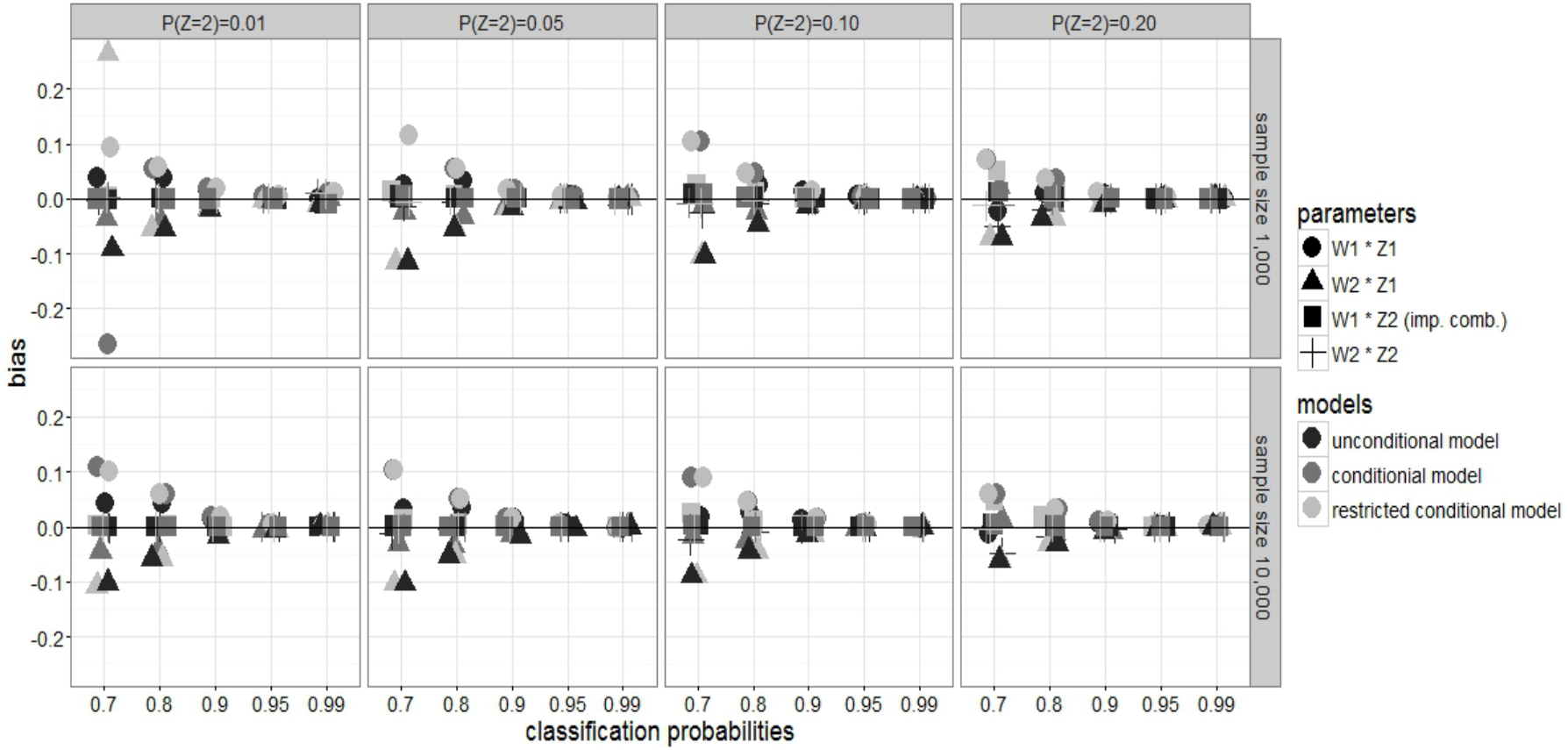
Performance measures

- Bias of estimates of interest: estimated bias is equal to difference between average estimate over all replications and true population value
- Coverage of 95% confidence interval: proportion of times that true population value falls within 95% confidence interval constructed around estimate over all replications
- Ratio of average standard error of estimate over standard deviation of 1,000 estimates

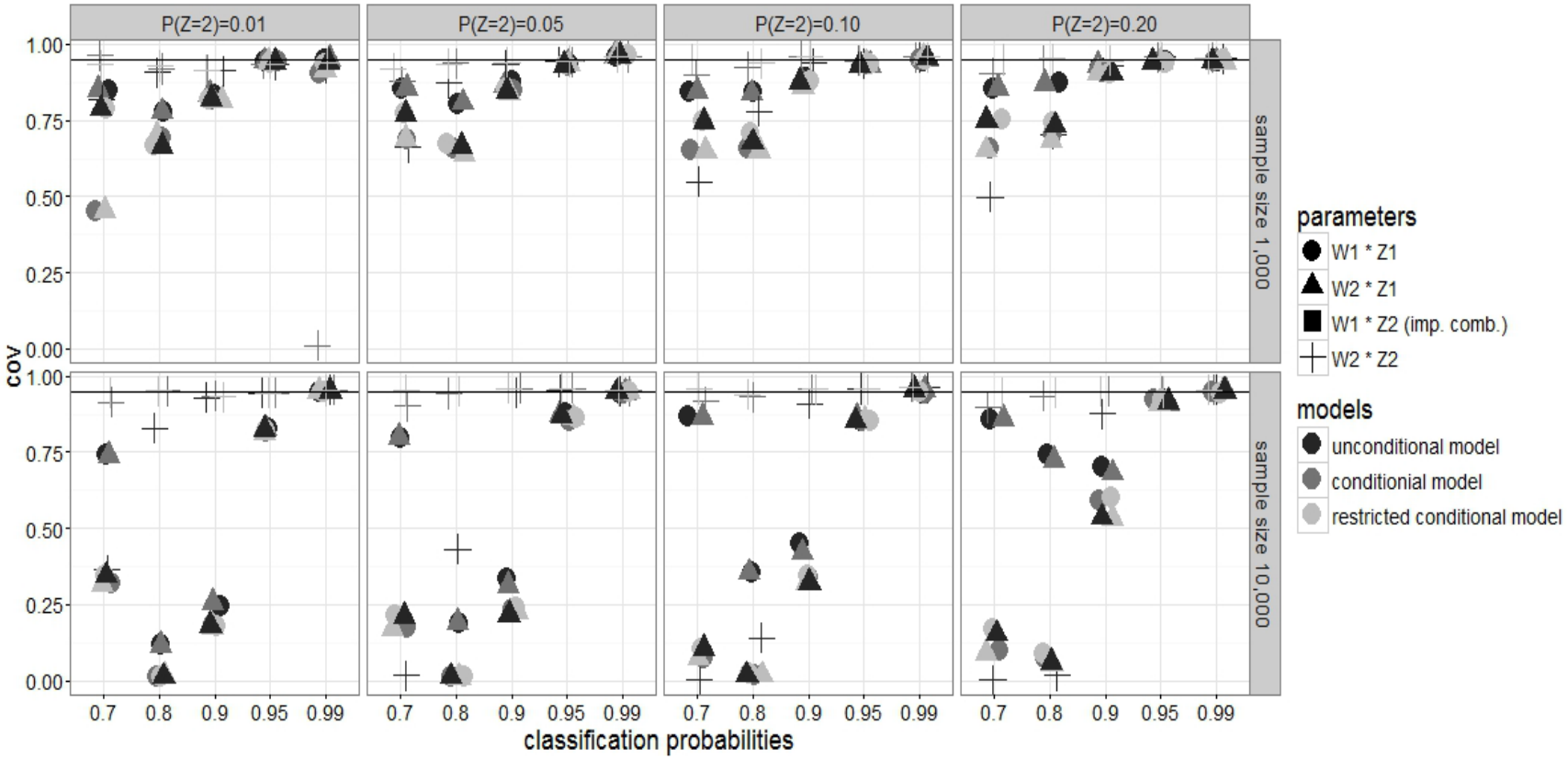
Simulation conditions

- Classification probabilities: 0.70; 0.80; 0.90; 0.95; 0.99
- $P(Z = 2)$: 0.01; 0.05; 0.10; 0.20
- Sample size: 1,000; 10,000
- Logit coefficients of Y regressed on Q of $\log(0.45/(1-0.45)) = -0.2007$, $\log(0.55/(1-0.55)) = 0.2007$ and $\log(0.65/(1-0.65)) = 0.6190$ corresponding to estimated odds ratio of 0.82, 1.22 and 1.86
  - Intercept fixed to 0
- Number of imputations: 5; 10; 20; 40

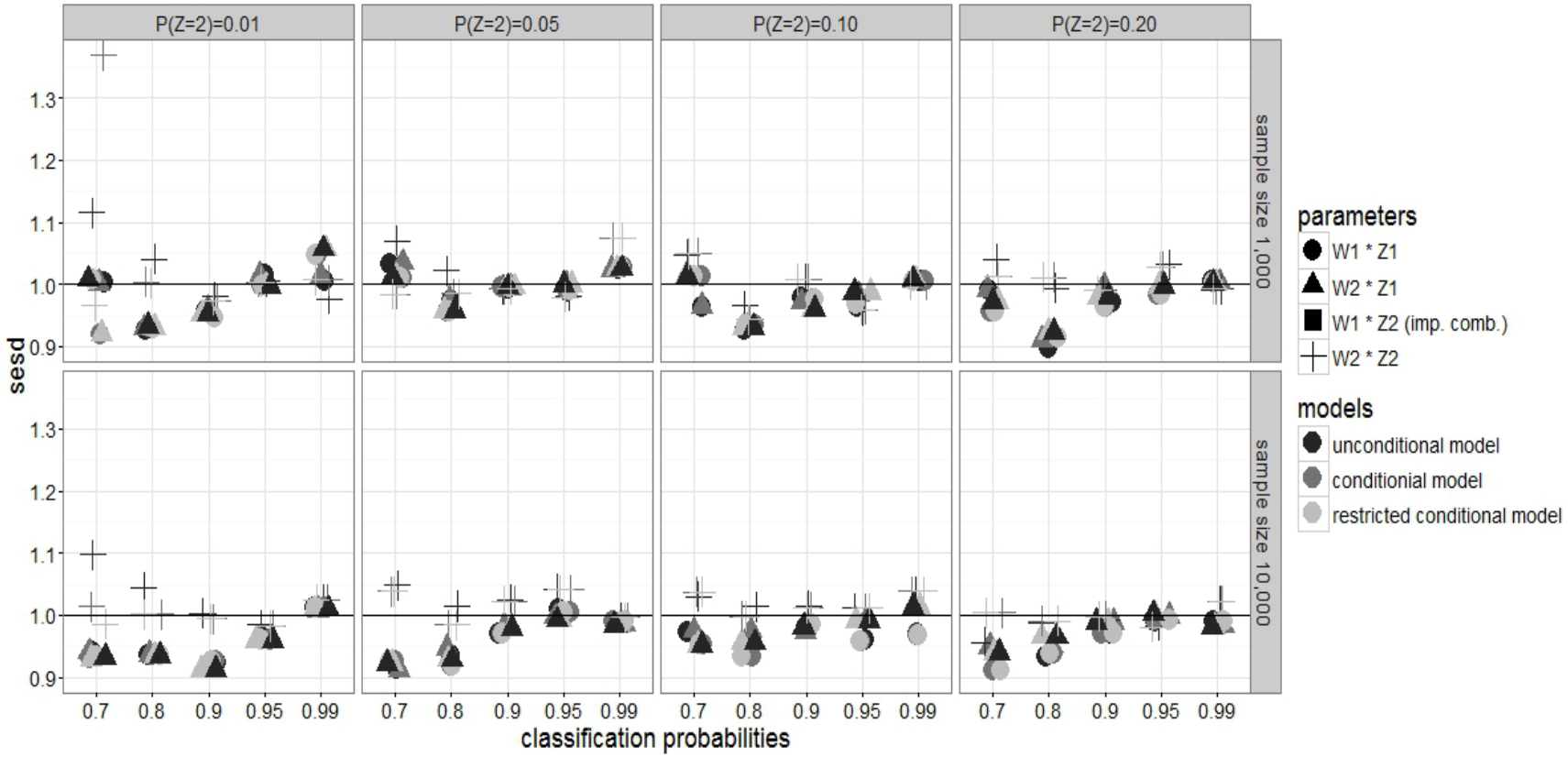
Bias of cell proportions



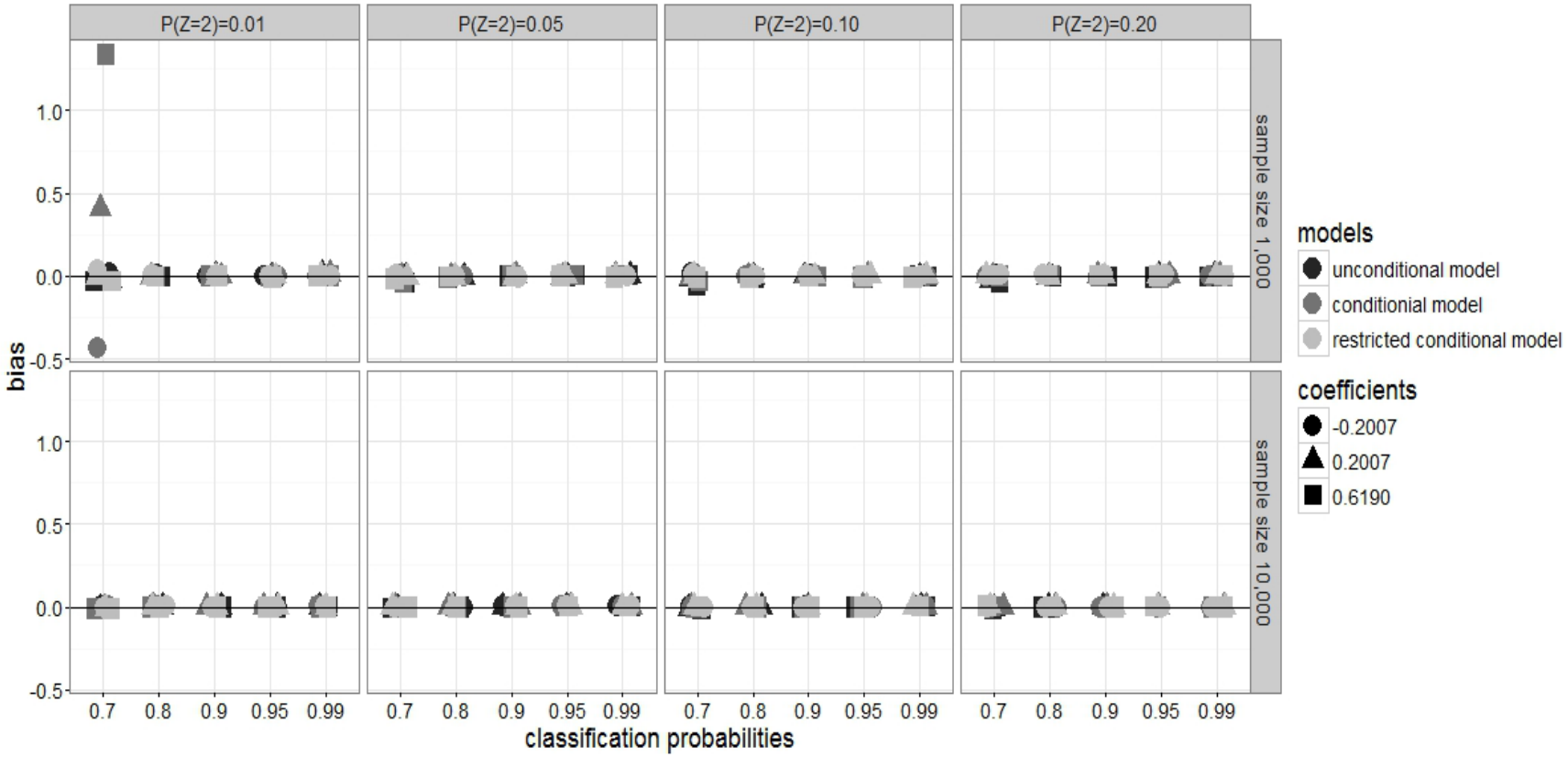
Coverage of 95% confidence interval



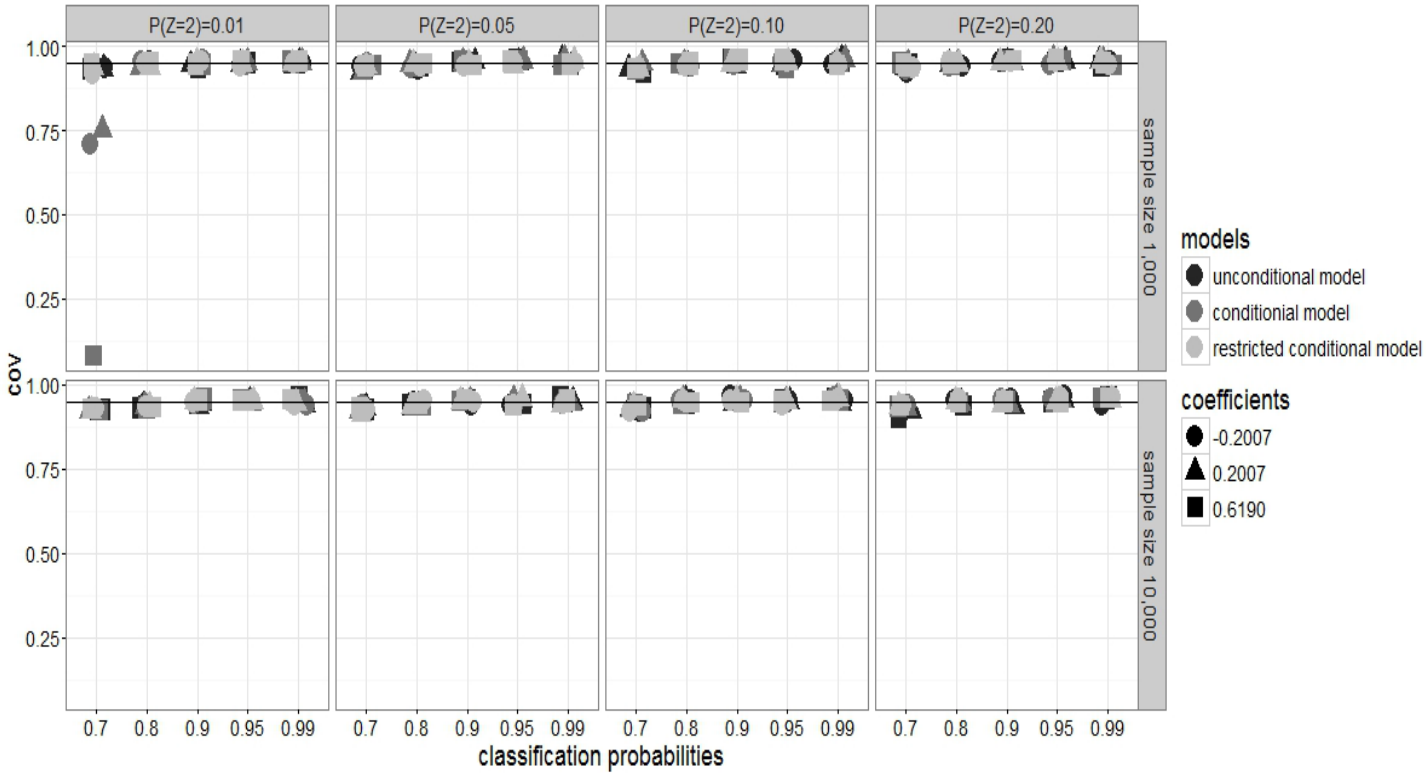
$se/sd(\hat{\theta})$



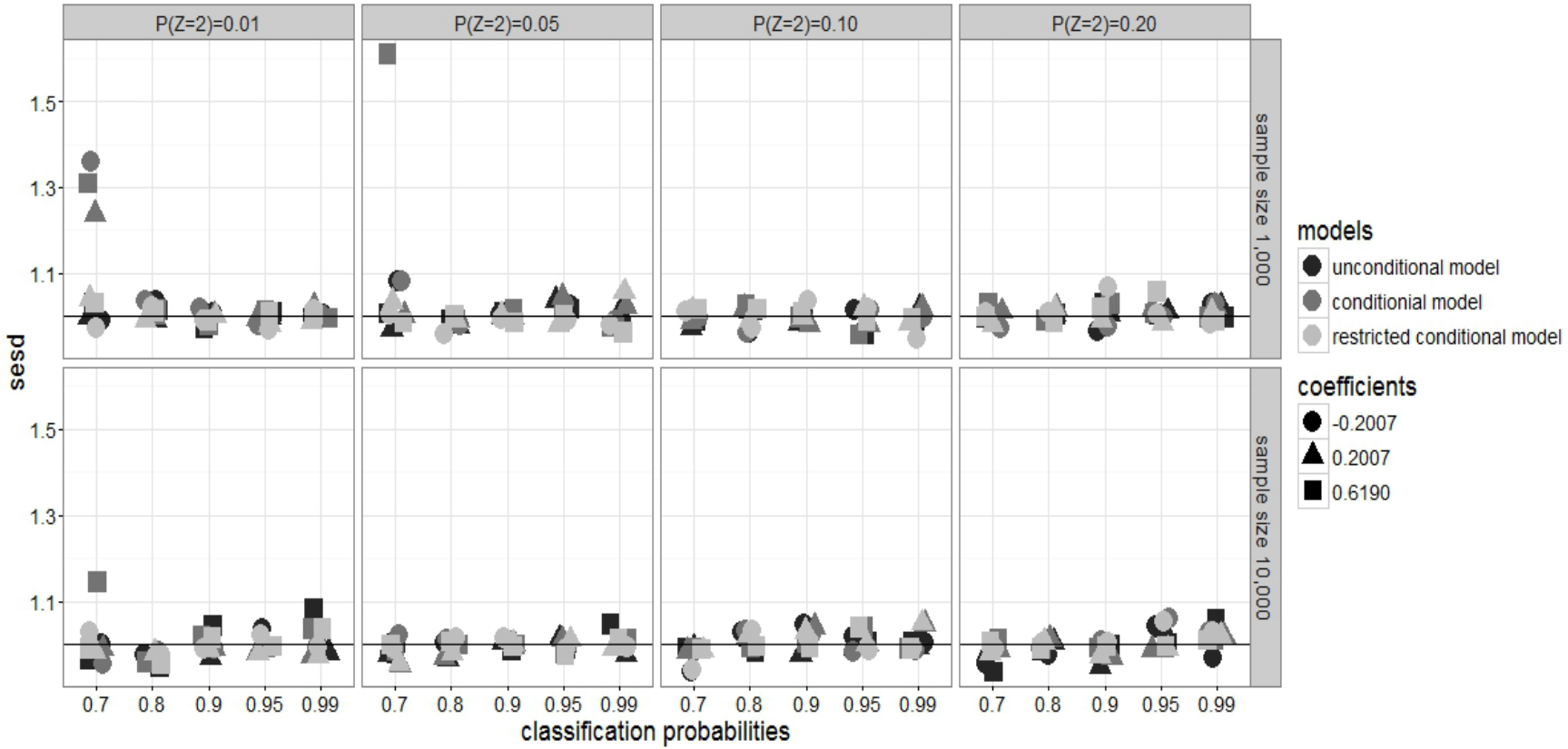
Bias of logistic regression coefficient



Coverage of 95% confidence interval



$se/sd(\hat{\theta})$



When can we apply method?

- In principle you need three or more independent data sources measuring the same target variables
- The more data sources, the better – especially if data sources are somewhat dependent
- With only two independent sources you need more information, e.g. subject-matter knowledge or from audit sample

References

- Over-imputation
  - Blackwell, M., Honaker, J., & King, G. (2017), A Unified Approach to Measurement Error and Missing Data: Overview and Applications. *Sociological Methods & Research* **46**(3), 303-341.
  - Blackwell, M., Honaker, J., & King, G. (2017), A Unified Approach to Measurement Error and Missing Data : Details and Extensions. *Sociological Methods & Research* **46**(3), 342-369.
- Structural equation modelling
  - Scholtus, S. & B.F.M. Bakker (2013) *Estimating the Validity of Administrative and Survey Variables by Means of Structural Equation Models*. CBS discussion paper. https://www.cbs.nl/nl-nl/achtergrond/2013/12/estimating-the-validity-of-administrative-and-survey-variables-through-structural-equation-modeling-a-simulation-study-on-robustness.

References

- Latent class analysis
  - Biemer, P.P. (2011), *Latent Class Analysis of Survey Error*. John Wiley & Sons, Hoboken.
  - Boeschoten, L., D. Oberski & T. de Waal (2017), Estimating Classification Errors under Edit Restrictions in Composite Survey-Register Data using Multiple Imputation Latent Class Modelling (MILC). *Journal of Official Statistics* **33**, 921-962.
  - Boeschoten, L. (2019), *Consistent Estimates for Categorical data Based on a Mix of Administrative Data sources and Surveys*. PhD thesis, Tilburg University.
  - Boeschoten, L. D. Filipponi & R. Varriale (2021), Combining Multiple Imputation and Latent Markov Modeling to Obtain Consistent Estimates of True Employment Status. *Journal of Survey Statistics and Methodology* **9**(3), 549-573, https://doi.org/10.1093/jssam/smz052
  - Pavlopoulos, D. & J.K. Vermunt (2015), Measuring Temporary Employment. Do Survey or Register Data Tell the Truth? *Survey Methodology* **41**, 197-214.