

Latent variable models for finite population inference

Maria Giovanna Ranalli

Dip. di Scienze Politiche
Università degli Studi di Perugia, Italy

January 31, 2024



Acknowledgments and Disclaimer

I am particularly indebted for the research I am going to present to

- Giorgio E. Montanari and Alina Matei
- Andrea Neri, Enrico Fabrizi, Gaia Bertarelli, Monica Pratesi ...
- Regione Umbria, Istat, Banca d'Italia for sharing data and problems

Outline

- 1 Introduction
 - Estimation framework
 - Latent class and latent trait models
- 2 Nonresponse
 - (Generalized) calibration
 - Response propensity estimation
- 3 Response: latent constructs
 - Disability
 - Other examples
- 4 Conclusions and further research

Inferential framework

- Inference for a finite population $U = \{1, \dots, k, \dots, N\}$
- Focus is on **descriptive inference** vs. analytic inference
- Var. of interest: $\mathbf{y}_k = (y_{1k}, \dots, y_{mk}, \dots, y_{Mk})$
- Population total $t_y = \sum_U \mathbf{y}_k$
- $s \subseteq U$ of size n is drawn using a sampling design $p(s)$
- $\pi_k = Pr(k \in s)$
- Horvitz-Thompson estimator: $\hat{t}_y^{\text{HT}} = \sum_s d_k \mathbf{y}_k$, $d_k = \pi_k^{-1}$ is unbiased for t_y .

Inferential framework

- Inference for a finite population $U = \{1, \dots, k, \dots, N\}$
- Focus is on **descriptive inference** vs. analytic inference
- Var. of interest: $\mathbf{y}_k = (y_{1k}, \dots, y_{mk}, \dots, y_{Mk})$
- Population total $\mathbf{t}_y = \sum_U \mathbf{y}_k$
- $s \subseteq U$ of size n is drawn using a sampling design $p(s)$
- $\pi_k = Pr(k \in s)$
- Horvitz-Thompson estimator: $\hat{\mathbf{t}}_y^{\text{HT}} = \sum_s d_k \mathbf{y}_k$, $d_k = \pi_k^{-1}$ is **unbiased** for \mathbf{t}_y .

... but ...

Hamlet to Horatio, W. Shakespeare

*“There are more things in heaven and earth, Horatio,
Than are dreamt of in your philosophy”*

Nonresponse: adjusting for nonignorable unit nonresponse using latent variables

- Latent class models to form response homogeneity groups for generalized calibration
- Latent trait models to estimate the response propensity

... but ...

Hamlet to Horatio, W. Shakespeare

*“There are more things in heaven and earth, Horatio,
Than are dreamt of in your philosophy”*

Variable of interest: multivariate context

- not directly observable – disability, social integration, educational poverty, well-being
- measured with error – integration of survey and admin data

... but ...

Hamlet to Horatio, W. Shakespeare

*“There are more things in heaven and earth, Horatio,
Than are dreamt of in your philosophy”*

Granularity: survey used to obtain estimates of descriptive statistics for the whole population U and for **subpopulations** U_i of interest, sample s_i may have small dimension $n_i \rightarrow$ **small area estimation** problem

Outline

- 1 Introduction
 - Estimation framework
 - Latent class and latent trait models
- 2 Nonresponse
 - (Generalized) calibration
 - Response propensity estimation
- 3 Response: latent constructs
 - Disability
 - Other examples
- 4 Conclusions and further research

Latent variable models for multivariate data

- Observed variables are imperfect measures of some latent variable
- Different types of latent variable models according to whether the observed and latent variables are categorical or continuous

	Latent	
Manifest/Observed	Continuous	Categorical
Continuous	Factor Analysis	Mixture Models
Categorical	Latent Trait Models	Latent Class models

- Categorical → clustering
- Continuous → mapping

Latent variable models for multivariate data

- Observed variables are imperfect measures of some latent variable
- Different types of latent variable models according to whether the observed and latent variables are categorical or continuous

	Latent	
Manifest/Observed	Continuous	Categorical
Continuous	Factor Analysis	Mixture Models
Categorical	Latent Trait Models	Latent Class models

- Categorical → clustering
- Continuous → mapping

Latent class analysis

- Individuals can be classified into mutually exclusive and exhaustive **latent classes**, based on their pattern of answers on a set of **manifest categorical variables**
- True class membership is unknown for each individual.
- Latent class analysis is a model-based method for clustering (or classification)
- The final number of classes is not usually predetermined prior to analysis

Latent class analysis

- Individuals can be classified into mutually exclusive and exhaustive **latent classes**, based on their pattern of answers on a set of **manifest categorical variables**
- True class membership is unknown for each individual.
- Latent class analysis is a model-based method for clustering (or classification)
- The final number of classes is not usually predetermined prior to analysis

Latent class models: categorical latent variable

[Lazarsfeld and Henry, 1968, Goodman, 1974, Bartholomew et al., 2011]

- Let the vector of L manifest variables observed on unit k be

$$\omega_k = (\omega_{1k}, \dots, \omega_{\ell k}, \dots, \omega_{Lk}),$$

and $\mathbf{h} = (h_1, \dots, h_L)$ is a possible response pattern.

- Manifest variables can be binary or polytomous

Latent class model

- ϑ_k : latent class variable
- c : particular latent class
- C : number of latent classes

$$P(\boldsymbol{\omega}_k = \mathbf{h}) = \sum_{c=1}^C P(\vartheta_k = c) P(\boldsymbol{\omega}_k = \mathbf{h} | \vartheta_k = c),$$

- Conditional Independence Assumption

$$P(\boldsymbol{\omega}_k = \mathbf{h} | \vartheta_k = c) = \prod_{\ell=1}^L P(\omega_{\ell k} = h_{\ell} | \vartheta_k = c).$$

Latent class model

- ϑ_k : latent class variable
- c : particular latent class
- C : number of latent classes

$$P(\boldsymbol{\omega}_k = \mathbf{h}) = \sum_{c=1}^C P(\vartheta_k = c) P(\boldsymbol{\omega}_k = \mathbf{h} | \vartheta_k = c),$$

- Conditional Independence Assumption

$$P(\boldsymbol{\omega}_k = \mathbf{h} | \vartheta_k = c) = \prod_{\ell=1}^L P(\omega_{\ell k} = h_{\ell} | \vartheta_k = c).$$

Latent class prediction → classification

- Covariates may influence the probabilities $P(\vartheta_k = c)$
- The parameters of a latent class model are usually estimated by means of the EM algorithm
- C may be chosen using model selection criteria like AIC, BIC
- The classification problem of assigning respondents to latent classes may be solved with a simple **Bayes rule**
- **Posterior probabilities**

$$P(\vartheta_k = c | \omega_k = \mathbf{h}) = \frac{P(\vartheta_k = c)P(\omega_k = \mathbf{h} | \vartheta_k = c)}{\sum_{c=1}^C P(\vartheta_k = c)P(\omega_k = \mathbf{h} | \vartheta_k = c)}$$

Latent class prediction → classification

- Covariates may influence the probabilities $P(\vartheta_k = c)$
- The parameters of a latent class model are usually estimated by means of the EM algorithm
- C may be chosen using model selection criteria like AIC, BIC
- The classification problem of assigning respondents to latent classes may be solved with a simple **Bayes rule**
- **Posterior probabilities**

$$P(\vartheta_k = c | \boldsymbol{\omega}_k = \mathbf{h}) = \frac{P(\vartheta_k = c)P(\boldsymbol{\omega}_k = \mathbf{h} | \vartheta_k = c)}{\sum_{c=1}^C P(\vartheta_k = c)P(\boldsymbol{\omega}_k = \mathbf{h} | \vartheta_k = c)}$$

Latent trait models: continuous latent variable

e.g. [Bartholomew et al., 2002]

- Manifest **binary** variables ω_k (only for ease of notation)
- $q_{\ell k} = P(\omega_{\ell k} = 1 | \theta_k)$
- The latent trait model is defined as

$$\text{logit}(q_{\ell k}) = \beta_{\ell 0} + \beta_{\ell 1} \theta_k,$$

where $\beta_{\ell 0}$ and $\beta_{\ell 1}$ are model parameters.

- $\theta_k \rightarrow$ ability
- $\beta_{\ell 0} \rightarrow$ difficulty
- $\beta_{\ell 1} \rightarrow$ discrimination
- This is also known as the **Two parameter logistic Rasch model**
- **Simple Rasch model**: $\text{logit}(q_{\ell k}) = \beta_{\ell 0} + \beta_1 \theta_k$

Latent trait models: continuous latent variable

e.g. [Bartholomew et al., 2002]

- Manifest **binary** variables ω_k (only for ease of notation)
- $q_{\ell k} = P(\omega_{\ell k} = 1 | \theta_k)$
- The latent trait model is defined as

$$\text{logit}(q_{\ell k}) = \beta_{\ell 0} + \beta_{\ell 1} \theta_k,$$

where $\beta_{\ell 0}$ and $\beta_{\ell 1}$ are model parameters.

- $\theta_k \rightarrow$ ability
- $\beta_{\ell 0} \rightarrow$ difficulty
- $\beta_{\ell 1} \rightarrow$ discrimination
- This is also known as the **Two parameter logistic Rasch model**
- **Simple Rasch model**: $\text{logit}(q_{\ell k}) = \beta_{\ell 0} + \beta_1 \theta_k$

Latent trait models: continuous latent variable

e.g. [Bartholomew et al., 2002]

- Manifest **binary** variables ω_k (only for ease of notation)
- $q_{\ell k} = P(\omega_{\ell k} = 1 | \theta_k)$
- The latent trait model is defined as

$$\text{logit}(q_{\ell k}) = \beta_{\ell 0} + \beta_{\ell 1} \theta_k,$$

where $\beta_{\ell 0}$ and $\beta_{\ell 1}$ are model parameters.

- $\theta_k \rightarrow$ ability
- $\beta_{\ell 0} \rightarrow$ difficulty
- $\beta_{\ell 1} \rightarrow$ discrimination
- This is also known as the **Two parameter logistic Rasch** model
- **Simple Rasch** model: $\text{logit}(q_{\ell k}) = \beta_{\ell 0} + \beta_1 \theta_k$

Computation of θ_k

- Estimation of θ_k via ML methods.
- Usually $\theta_k \sim N(0, 1)$.
- The likelihood of ω_k is given by

$$f(\omega_k) = \int f(\omega_k | \theta_k) h(\theta_k) d\theta_k,$$

where

$$f(\omega_k | \theta_k) = \prod_{\ell=1}^L f_{\ell}(\omega_{\ell k} | \theta_k) = \prod_{\ell=1}^L q_{\ell k}^{\omega_{\ell k}} (1 - q_{\ell k})^{1 - \omega_{\ell k}}$$

Assumptions

- ① *Conditional independence*: item responses are independent given the latent variable
- ② *Monotonicity*: as the latent variable θ_k increases, the probability of response to an item increases or stays the same across intervals of θ_k .
- ③ *Unidimensionality*: a single latent variable fully explains the correlation among the manifest variables.
 - Covariates to relax CIA
 - Model selection and diagnostics
 - Multidimensional latent variable: $\theta_k = (\theta_{1k}, \dots, \theta_{jk}, \dots, \theta_{Jk})$,
 $J \ll L$

Assumptions

- ① *Conditional independence*: item responses are independent given the latent variable
- ② *Monotonicity*: as the latent variable θ_k increases, the probability of response to an item increases or stays the same across intervals of θ_k .
- ③ *Unidimensionality*: a single latent variable fully explains the correlation among the manifest variables.
 - Covariates to relax CIA
 - Model selection and diagnostics
 - Multidimensional latent variable: $\theta_k = (\theta_{1k}, \dots, \theta_{jk}, \dots, \theta_{Jk})$,
 $J \ll L$

Outline

- 1 Introduction
 - Estimation framework
 - Latent class and latent trait models
- 2 Nonresponse
 - (Generalized) calibration
 - Response propensity estimation
- 3 Response: latent constructs
 - Disability
 - Other examples
- 4 Conclusions and further research

Motivation

- The Survey on Household Income and Wealth (SHIW) is an official survey conducted by the Central Bank of Italy since 1962
- The main focus is to collect detailed information on household **income, wealth and total expenditure**
- Probabilistic sample of about 8,000 Italian households
- Unit response rate 0.56, item nonresponse negligible. **Unit nonresponse** is unlikely to be at random
- Several variables of interest: some are sensitive and possibly subject to **measurement error**

Unit nonresponse

- Respondents set: $r \subseteq s$; $r = \{k \in s | R_k = 1\}$
- Two-phase setting
- Response probability: $p_k = P(R_k = 1 | k \in s)$
- Double expansion estimator: $\hat{t}_y^{2E} = \sum_r d_k p_k^{-1} \mathbf{y}_k$
- Auxiliary information: $\mathbf{x}_k = (x_{1k}, \dots, x_{qk})$ is known for $k \in r$ and population total t_x is known OR can be unbiasedly estimated with $\hat{t}_x^{HT} = \sum_s d_k \mathbf{x}_k$.
- t_x^* denotes the population total of \mathbf{x} or its Horvitz-Thompson estimator accordingly.

Unit nonresponse

- Respondents set: $r \subseteq s$; $r = \{k \in s | R_k = 1\}$
- Two-phase setting
- Response probability: $p_k = P(R_k = 1 | k \in s)$
- Double expansion estimator: $\hat{t}_y^{2E} = \sum_r d_k p_k^{-1} \mathbf{y}_k$
- **Auxiliary information**: $\mathbf{x}_k = (x_{1k}, \dots, x_{qk})$ is known for $k \in r$ and population total t_x is known OR can be unbiasedly estimated with $\hat{t}_x^{HT} = \sum_s d_k \mathbf{x}_k$.
- t_x^* denotes the population total of \mathbf{x} or its Horvitz-Thompson estimator accordingly.

Calibration tools, see e.g. [Särndal and Lundström, 2005]

Benchmark constraints

It pursues the construction of a single set of weights w_k modifying the basic d_k 's, while satisfying

$$\sum_r w_k \mathbf{x}_k = \mathbf{t}_x^* \quad (1)$$

Adjustments are assumed to be a function of an unknown (but estimable) linear combination of auxiliary variables, i.e.

$$w_k = d_k F(\mathbf{x}_k \boldsymbol{\gamma}).$$

Once an estimate for $\boldsymbol{\gamma}$ is obtained from (1),

$$\hat{\mathbf{t}}_y^{\text{CAL}} = \sum_r d_k F(\mathbf{x}_k \hat{\boldsymbol{\gamma}}) \mathbf{y}_k.$$

Calibration tools, see e.g. [Särndal and Lundström, 2005]

Benchmark constraints

It pursues the construction of a single set of weights w_k modifying the basic d_k 's, while satisfying

$$\sum_r w_k \mathbf{x}_k = \mathbf{t}_x^* \quad (1)$$

Adjustments are assumed to be a function of an unknown (but estimable) linear combination of auxiliary variables, i.e.

$$w_k = d_k F(\mathbf{x}_k \boldsymbol{\gamma}).$$

Once an estimate for $\boldsymbol{\gamma}$ is obtained from (1),

$$\hat{\mathbf{t}}_y^{\text{CAL}} = \sum_r d_k F(\mathbf{x}_k \hat{\boldsymbol{\gamma}}) \mathbf{y}_k.$$

Generalized calibration [Deville, 2000, Kott, 2006]

Classical calibration

$$w_k = d_k F(\mathbf{x}_k \boldsymbol{\gamma}) \quad \rightarrow \quad p_k^{-1} = F(\mathbf{x}_k \boldsymbol{\gamma})$$

Generalized calibration

$$w_k = d_k F(\mathbf{z}_k \boldsymbol{\gamma}) \quad \rightarrow \quad p_k^{-1} = F(\mathbf{z}_k \boldsymbol{\gamma})$$

- \mathbf{x} : q calibration variables
- \mathbf{z} : p model variables
- constraints: $\sum_r w_k \mathbf{x}_k = \sum_r d_k F(\mathbf{z}_k \boldsymbol{\gamma}) \mathbf{x}_k = \mathbf{t}_x^*$

$$\hat{\mathbf{t}}_y^{\text{GCAL}} = \sum_r d_k F(\mathbf{z}_k \hat{\boldsymbol{\gamma}}) \mathbf{y}_k$$

Generalized calibration [Deville, 2000, Kott, 2006]

Classical calibration

$$w_k = d_k F(\mathbf{x}_k \boldsymbol{\gamma}) \quad \rightarrow \quad p_k^{-1} = F(\mathbf{x}_k \boldsymbol{\gamma})$$

Generalized calibration

$$w_k = d_k F(\mathbf{z}_k \boldsymbol{\gamma}) \quad \rightarrow \quad p_k^{-1} = F(\mathbf{z}_k \boldsymbol{\gamma})$$

- \mathbf{x} : q calibration variables
- \mathbf{z} : p model variables
- constraints: $\sum_r w_k \mathbf{x}_k = \sum_r d_k F(\mathbf{z}_k \boldsymbol{\gamma}) \mathbf{x}_k = \mathbf{t}_x^*$

$$\hat{\mathbf{t}}_y^{\text{GCAL}} = \sum_r d_k F(\mathbf{z}_k \hat{\boldsymbol{\gamma}}) \mathbf{y}_k$$

Generalized calibration and Nonignorable nonresponse

- Since z_k need only be known for $k \in r$, **elements of y** can be used as model variables [Deville, 2000, Kott and Chang, 2010]
- [Lesage, Haziza and D'Haultfeuille, 2019] warn against **variance amplification** for generalized calibration
- Several variables are likely to affect nonresponse, some may be **unobservable** (like willingness to respond, attitudes) and/or may be affected by response (measurement) error (like income and wealth)

Generalized calibration and Nonignorable nonresponse

- Since z_k need only be known for $k \in r$, **elements of y** can be used as model variables [Deville, 2000, Kott and Chang, 2010]
- [Lesage, Haziza and D'Haultfeuille, 2019] warn against **variance amplification** for generalized calibration
- Several variables are likely to affect nonresponse, some may be **unobservable** (like willingness to respond, attitudes) and/or may be affected by response (measurement) error (like income and wealth)

Generalized calibration with latent variables

- Selection of a plausible set of model variables z
- Latent class and latent trait models
- y variables can be included among the manifest variables
- Reduction of the dimensionality of the model vector (smaller variability of the set of weights)



R., M. G., Matei, A., Neri, A. (2023)

Generalised calibration with latent variables for the treatment of unit nonresponse in sample surveys

Statistical Methods & Applications, 32(1), 169–195.

Definition of the model variable vector

- $\mathbf{z}_k = \hat{\boldsymbol{\theta}}_k$, or
- $\mathbf{z}_k = (\hat{\boldsymbol{\theta}}_k, \mathbf{z}_k^0)$ (where \mathbf{z}_k^0 represent model variables different from $\hat{\boldsymbol{\theta}}_k$) where the estimate of the latent variable $\boldsymbol{\theta}_k$ is obtained, either using latent trait or latent class models.
- In the latter case $\hat{\boldsymbol{\theta}}_k = (\hat{\theta}_{1k}, \dots, \hat{\theta}_{Ck})$ is the indicator variable vector such that $\theta_{ck} = 1$ if unit k belongs to latent class c
- Classification of units in latent classes provides an alternative and data-driven way of building **Response Homogeneity Groups** to deal with nonresponse.

Definition of the model variable vector

- $\mathbf{z}_k = \hat{\boldsymbol{\theta}}_k$, or
- $\mathbf{z}_k = (\hat{\boldsymbol{\theta}}_k, \mathbf{z}_k^0)$ (where \mathbf{z}_k^0 represent model variables different from $\hat{\boldsymbol{\theta}}_k$) where the estimate of the latent variable $\boldsymbol{\theta}_k$ is obtained, either using latent trait or latent class models.
- In the latter case $\hat{\boldsymbol{\theta}}_k = (\hat{\theta}_{1k}, \dots, \hat{\theta}_{Ck})$ is the indicator variable vector such that $\theta_{ck} = 1$ if unit k belongs to latent class c
- Classification of units in latent classes provides an alternative and data-driven way of building **Response Homogeneity Groups** to deal with nonresponse.

Properties

Variance estimation accounts for the two-phases

- The two-phase approach: $\hat{V}_{2p}(\hat{t}_{ym}^{\text{GCAL}}) = \hat{V}_{\text{sam}} + \hat{V}_{\text{nr}}$
- The reverse approach [Shao and Steel, 1999]:
 $\hat{V}_{\text{rev}}(\hat{t}_{ym}^{\text{GCAL}}) = \hat{V}_1 + \hat{V}_2$
- Jackknife [Kott, 2006]: $\hat{V}_{\text{jack}}(\hat{t}_{ym}^{\text{GCAL}})$

Simulation studies

- More efficient and less biased than classical calibration that uses only auxiliary information x as model variables.
- Evidence of the variance amplification when using y directly
- Latent variable approach reduces the variance

Properties

Variance estimation accounts for the two-phases

- The two-phase approach: $\hat{V}_{2p}(\hat{t}_{ym}^{GCAL}) = \hat{V}_{sam} + \hat{V}_{nr}$
- The reverse approach [Shao and Steel, 1999]:
 $\hat{V}_{rev}(\hat{t}_{ym}^{GCAL}) = \hat{V}_1 + \hat{V}_2$
- Jackknife [Kott, 2006]: $\hat{V}_{jack}(\hat{t}_{ym}^{GCAL})$

Simulation studies

- More efficient and less biased than classical calibration that uses only auxiliary information x as model variables.
- Evidence of the variance amplification when using y directly
- Latent variable approach reduces the variance

The application on SHIW data

- Variable of particular interest: **average yearly individual net wealth** (= income + financial wealth – liabilities)
- Previous research based on the SHIW data shows that nonresponse is nonignorable and depends on the true wealth.
- True wealth is not observed either for respondents because of measurement error.

The manifest variables $\omega_k \rightarrow 12$ entries

- Individual observed wealth class (ordinal variable with **five** levels)
 - **Six** dummy indicators for the ownership of a secondary dwelling, of bonds, of agricultural and of non-agricultural land, of other non-residential buildings and for the household living in a deluxe dwelling
 - Number of total call attempts needed to make the interview (ranging between 1 and 4)
- ⇒ Model selection brings a good classification in **5 latent classes**.

Description of the latent classes (RHG) $\hat{\theta}_k \rightarrow 5$ entries

Very rich : high financial and non-financial wealth, living in luxury residence, **difficult** to contact/interview (about 16 % of population);

Well-off : medium-high wealth, living in luxury residence, **easy** to contact/interview (about 30%);

Average : average wealth, owners of some lands and non residential buildings, **easy** to contact/interview (about 16%);

Below average : low wealth, **easy** to contact/interview (about 25%);

Very poor : low wealth, almost zero financial wealth, **difficult** to contact/interview (about 13%).

Calibration Variables $\mathbf{x}_k \rightarrow 18$ entries

Two sources of auxiliary information.

- ① National statistical office (demographic)
 - age (5 classes)
 - gender
 - education (3 levels)
 - nationality (italian/foreigner)
 - job status (employed/unemployed/inactive)
 - geographical area (north/centre/south)
- ② Department of the Treasury (administrative records of real estate owners)
 - value of the owned dwelling (5 classes)

Estimators Compared

- (1) the Hajek estimator (no nonresponse adjustment);
- (2) a two phase estimator in which response probabilities are estimated via a logistic model that uses covariates known also for nonrespondents (details are in [Neri and Ranalli, 2011]);
- (3) classical calibration using the 18 x calibration variables;

Generalized calibration using as model variables z :

- (4) the 5 latent classes $\hat{\theta}_k$,
- (5) the 12 manifest variables ω_k ,
- (6) 5 individual observed wealth classes (classes of y_k).

Results

Table: Estimated mean of the net wealth, estimated jackknife standard error and % coefficient of variation, standard deviation of the set of final weights.

	Est. Mean	Std Err Jackknife	%CV	w_k St Dev
(1) Hajek estimator	252,407	7,780	3.08	2,466
(2) Two phase estimator	296,012	8,862	2.99	3,349
(3) Classical calibration	282,379	6,733	2.38	2,550
Generalized calibration:				
(4) Model var. – latent cl.	307,762	7,844	2.55	2,697
(5) Model var. – manifest var.	329,457	26,458	8.03	8,306
(6) Model var. – classes of y	319,805	8,394	2.62	3,075

Outline

- 1 Introduction
 - Estimation framework
 - Latent class and latent trait models
- 2 Nonresponse
 - (Generalized) calibration
 - Response propensity estimation
- 3 Response: latent constructs
 - Disability
 - Other examples
- 4 Conclusions and further research

Response probabilities/propensities p_k 's are unknown

Double expansion estimator: $\hat{t}_y^{2E} = \sum_r d_k p_k^{-1} \mathbf{y}_k$

Estimates for p_k may come from

- implicit models: post-stratification, (generalized) calibration
- explicit models: logistic models for R_k on $k \in s$
- $\text{logit}(p_k) = \mathbf{x}'_k \boldsymbol{\alpha}$ (Kim & Kim, 2007)
- $\text{logit}(p_k) = \alpha_0 + \alpha_1 y_{kj}$

Responding to a survey is an **attitude**

- Let θ_k be a measure of the “will to respond”
- $\text{logit}(p_k) = \alpha_0 + \alpha_1 \theta_k$



Matei, A., & R, M. G. (2015).

Dealing with non-ignorable nonresponse in survey sampling: a latent modeling approach

Survey Methodology, 41(1), 145–165.

Response probabilities/propensities p_k 's are unknown

Double expansion estimator: $\hat{t}_y^{2E} = \sum_r d_k p_k^{-1} \mathbf{y}_k$

Estimates for p_k may come from

- implicit models: post-stratification, (generalized) calibration
- explicit models: logistic models for R_k on $k \in s$
- $\text{logit}(p_k) = \mathbf{x}'_k \boldsymbol{\alpha}$ (Kim & Kim, 2007)
- $\text{logit}(p_k) = \alpha_0 + \alpha_1 y_{kj}$

Responding to a survey is an **attitude**

- Let θ_k be a measure of the “will to respond”
- $\text{logit}(p_k) = \alpha_0 + \alpha_1 \theta_k$



Matei, A., & R, M. G. (2015).

Dealing with non-ignorable nonresponse in survey sampling: a latent modeling approach

Survey Methodology, 41(1), 145–165.

Response probabilities/propensities p_k 's are unknown

Double expansion estimator: $\hat{t}_y^{2E} = \sum_r d_k p_k^{-1} \mathbf{y}_k$

Estimates for p_k may come from

- implicit models: post-stratification, (generalized) calibration
- explicit models: logistic models for R_k on $k \in s$
- $\text{logit}(p_k) = \mathbf{x}'_k \boldsymbol{\alpha}$ (Kim & Kim, 2007)
- $\text{logit}(p_k) = \alpha_0 + \alpha_1 y_{kj}$

Responding to a survey is an **attitude**

- Let θ_k be a measure of the “will to respond”
- $\text{logit}(p_k) = \alpha_0 + \alpha_1 \theta_k$



Matei, A., & R, M. G. (2015).

Dealing with non-ignorable nonresponse in survey sampling: a latent modeling approach

Survey Methodology, 41(1), 145–165.

How to measure θ_k ?

It may be related strongly on the type and on the subject of the survey (e.g. surveys on sensitive issues: sexual attitudes, politics, income, drug abuse...).

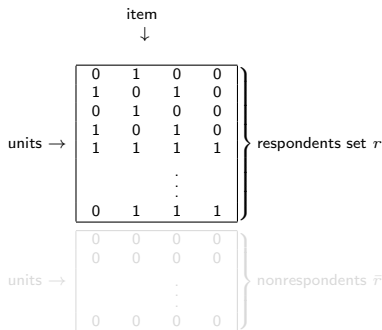
Chambers and Skinner

Analysis of Survey Data – Wiley, 2003, p.278:

“[...] from a theoretical perspective the difference between unit and item nonresponse is unnecessary. Unit nonresponse is just an extreme form of item nonresponse.”

Key idea: item responses as manifest variables

- Item nonresponse affects L questionnaire items
- For each item ℓ , $\ell = 1, \dots, L$, and each unit k , $k \in r$, $\omega_{\ell k} = 1$ if unit k answers to item ℓ and 0 otherwise



Key idea: item responses as manifest variables

- Item nonresponse affects L questionnaire items
- For each item ℓ , $\ell = 1, \dots, L$, and each unit k , $k \in r$, $\omega_{\ell k} = 1$ if unit k answers to item ℓ and 0 otherwise

		item					
		↓					
		0	1	0	0		
		1	0	1	0		
		0	1	0	0		
		1	0	1	0		
		1	1	1	1		
				⋮			
		0	1	1	1		
units →						}	respondents set r
		0	0	0	0		
		0	0	0	0		
				⋮			
		0	0	0	0		
units →						}	nonrespondents \bar{r}

The proposed method

- Response indicators to the variables are assumed to be related to an assumed underlying continuous scale which determines a latent variable used to estimate the response probabilities

$$\omega_{\ell k} \quad \text{for } \ell = 1, \dots, L \text{ and } k \in s \quad \rightarrow \quad \theta_k \quad \rightarrow \quad p_k$$

- $\text{logit}(p_k) = \alpha_0 + \alpha_1 \hat{\theta}_k$ using data $(R_k, \hat{\theta}_k)$ for $k \in s$.
- \hat{p}_k can be used in a double expansion fashion

$$\hat{\mathbf{t}}_y = \sum_r d_k \hat{p}_k^{-1} \mathbf{y}_k$$



Findings

- The approach reduces nonresponse bias in the case of nonignorable nonresponse
- Requires that unit nonresponse and item nonresponse are driven by the same factor
- Can be used also when auxiliary information, at the sample or at the population level, is not available.
- It can be used **anytime** item nonresponse is present on a set of variables of interest to obtain a **useful covariate**
- It can be extended to **ordinal** variables such as

$$\omega_{\ell k} = \begin{cases} 0 & \text{if } k \text{ refuses to respond} \\ 1 & \text{if } k \text{ doesn't know} \\ 2 & \text{if } k \text{ responds} \end{cases}$$

Findings

- The approach reduces nonresponse bias in the case of nonignorable nonresponse
- Requires that unit nonresponse and item nonresponse are driven by the same factor
- Can be used also when auxiliary information, at the sample or at the population level, is not available.
- It can be used **anytime** item nonresponse is present on a set of variables of interest to obtain a **useful covariate**
- It can be extended to **ordinal** variables such as

$$\omega_{\ell k} = \begin{cases} 0 & \text{if } k \text{ refuses to respond} \\ 1 & \text{if } k \text{ doesn't know} \\ 2 & \text{if } k \text{ responds} \end{cases}$$

Findings

- The approach reduces nonresponse bias in the case of nonignorable nonresponse
- Requires that unit nonresponse and item nonresponse are driven by the same factor
- Can be used also when auxiliary information, at the sample or at the population level, is not available.
- It can be used **anytime** item nonresponse is present on a set of variables of interest to obtain a **useful covariate**
- It can be extended to **ordinal** variables such as

$$\omega_{\ell k} = \begin{cases} 0 & \text{if } k \text{ refuses to respond} \\ 1 & \text{if } k \text{ doesn't know} \\ 2 & \text{if } k \text{ responds} \end{cases}$$

Outline

- 1 Introduction
 - Estimation framework
 - Latent class and latent trait models
- 2 Nonresponse
 - (Generalized) calibration
 - Response propensity estimation
- 3 Response: latent constructs
 - Disability
 - Other examples
- 4 Conclusions and further research

Motivation: Disability – need for care

- Estimating the number of people in condition of **severe disability** that requires intensive care is very important for regional governments that are (in Italy) responsible for health policy.
- → Estimating the number of people with different levels of **functional disability** within health districts of a particular administrative region (Umbria).

The variable of interest is latent

Observed variables are imperfect measures of some latent variable

$$\omega_k = (\omega_{1k}, \dots, \omega_{lk}, \dots, \omega_{Lk}) = \mathbf{y}_k$$

Data

- In the Italian survey on Health Conditions and Appeal to Medicare there is a set of $L = 9$ items (3/4 response categories) that survey
 - 1 difficulties in movements;
 - 2 difficulties in everyday activities and tasks;

The variable of interest is latent

Observed variables are imperfect measures of some latent variable

$$\omega_k = (\omega_{1k}, \dots, \omega_{lk}, \dots, \omega_{Lk}) = \mathbf{y}_k$$

Data

- In the Italian survey on Health Conditions and Appeal to Medicare there is a set of $L = 9$ items (3/4 response categories) that survey
 - 1 difficulties in movements;
 - 2 difficulties in everyday activities and tasks;

Items 1 – difficulties in movements

ITEM	Categories
DIST = longest walkable distance	1 = More than 200 m. 2 = Less than 200 m. 3 = Only few steps
STAIR = going up and down the stairs	1 = Yes 2 = With some effort 3 = With a lot of effort 4 = No
STOOP = stooping down	Same

Items 2 – difficulties in everyday activities and tasks

ITEM	Categories
BED = getting in and out of bed	1 = No effort 2 = With some effort 3 = With the help of others
CHAIR = sitting and standing	Same
DRESS = getting dressed and undressed	Same
BATH = taking a bath or a shower	Same
WASH = washing ones face and hands	Same
EAT = eating cutting ones food	Same

Measurement issues and estimation



Montanari, G. E., R. M. G., Eusebi, P. (2011)

Latent variable modeling of disability in people aged 65 or more
Statistical Methods & Applications, 20, 49–63.

- Disability has different dimensions (functional, mental, physical...)
- Latent class and latent trait models to obtain a measure of disability that is mostly connected with the need for care
- **Two-step** approach:
 - ① Classification of units is obtained
 - ② A Horvitz-Thompson estimator is computed to obtain estimates of the amount of a (sub)population belonging to each class

The small area estimation side of the problem

Survey provides reliable estimates at NUTS2 level (Region)

[HD]	<i>Sample</i>			<i>Population</i>		
	50 – 64	65 – 74	≥ 75	50 – 64	65 – 74	≥ 75
[11]	32	28	23	14300	8576	8584
[12]	52	28	30	10367	6440	6853
[21]	117	57	50	34375	20065	19145
[22]	27	16	16	10652	6363	6197
[23]	28	20	20	10059	6606	7069
[24]	26	15	14	10589	6485	6869
[31]	14	12	6	2107	1328	1722
[32]	41	29	38	9356	5534	6235
[33]	88	56	57	18584	10945	12451
[41]	120	56	52	25989	15875	15949
[42]	36	26	20	10401	6440	6776
[43]	39	28	23	8347	5327	5986

A two-step solution?

- Latent class memberships are used as a known dependent variable in a small area model (e.g. multinomial mixed effects model as in Ghosh et al., JASA, 1998)
- × When using latent variable estimates in a regression model, the association between the real value of the latent variable and the covariates is underestimated (Mesbah, 2004)
- × Propagation of the errors from step 1 to step 2
- × MSE for a small area estimate from step 2 ??

A one-step solution



Fabrizi, E., Montanari, G. E., R, M. G. (2016).

A hierarchical latent class model for predicting disability small area counts from survey data.

Journal of the Royal Statistical Society: Series A, 179(1), 103-131.

Latent class models & small area estimation model

Let $\mathbf{y}_{ik} = (y_{ik1}, \dots, y_{ikL})$ be the vector of responses for unit k in small area i , and \mathbf{h} a possible answer pattern. Then

$$P(\mathbf{y}_{ik} = \mathbf{h}) = \sum_{c=1}^C P(\vartheta_{ik} = c) P(\mathbf{y}_{ik} = \mathbf{h} | \vartheta_{ik} = c)$$

$$\log \frac{P(\vartheta_{ik} = c)}{P(\vartheta_{ik} = 1)} = \alpha_{0c} + \mathbf{x}'_{ik} \boldsymbol{\alpha}_{1c} + f_c(z_{ik}) + v_{ic},$$

for $c = 2, \dots, C$,

where v_{ic} are random effects accounting for area-specific heterogeneity not accounted for by the regressors, $v_{ic} \sim N(0, \sigma_{vc}^2)$.

Model fit, selection, and diagnostics

- Hierarchical Bayes approach
- Number of latent classes?
- Model (prior) specification
- MCMC output to obtain a measure of uncertainty

Conditional probabilities $P(\mathbf{y}_{ik} = \mathbf{h} | \vartheta_{ik} = c)$: classes 1, 2

Without difficulties (75.8%)

	Bath	Bed	Chair	<i>Dist</i>	Dress	Eat	<i>Stair</i>	<i>Stoop</i>	Wash
1	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.99	1.00
2	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.00
3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4							0.00	0.00	

With difficulties in movements(11.5%)

	Bath	Bed	Chair	<i>Dist</i>	Dress	Eat	<i>Stair</i>	<i>Stoop</i>	Wash
1	0.81	0.83	0.93	0.66	0.94	0.99	0.18	0.13	0.98
2	0.15	0.16	0.07	0.33	0.05	0.01	0.72	0.74	0.01
3	0.04	0.01	0.01	0.01	0.01	0.01	0.07	0.13	0.01
4							0.02	0.01	

Conditional probabilities $P(\mathbf{y}_{ik} = \mathbf{h} | \vartheta_{ik} = c)$: classes 1, 2

Without difficulties (75.8%)

	Bath	Bed	Chair	<i>Dist</i>	Dress	Eat	<i>Stair</i>	<i>Stoop</i>	Wash
1	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.99	1.00
2	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.00
3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4							0.00	0.00	

With difficulties in movements(11.5%)

	Bath	Bed	Chair	<i>Dist</i>	Dress	Eat	<i>Stair</i>	<i>Stoop</i>	Wash
1	0.81	0.83	0.93	0.66	0.94	0.99	0.18	0.13	0.98
2	0.15	0.16	0.07	0.33	0.05	0.01	0.72	0.74	0.01
3	0.04	0.01	0.01	0.01	0.01	0.01	0.07	0.13	0.01
4							0.02	0.01	

Conditional probabilities: classes 3 and 4

With difficulties in movements and daily tasks (4.0%)

	Bath	Bed	Chair	<i>Dist</i>	Dress	Eat	<i>Stair</i>	<i>Stoop</i>	Wash
1	0.07	0.19	0.47	0.53	0.14	0.86	0.12	0.10	0.76
2	0.58	0.71	0.52	0.42	0.75	0.09	0.50	0.40	0.22
3	0.35	0.10	0.01	0.06	0.11	0.05	0.35	0.39	0.02
4							0.03	0.11	

Partial dependency (4.2%)

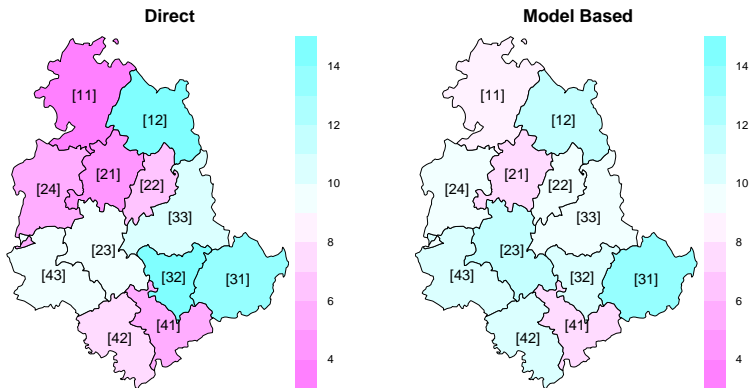
	Bath	Bed	Chair	<i>Dist</i>	Dress	Eat	<i>Stair</i>	<i>Stoop</i>	Wash
1	0.04	0.07	0.07	0.09	0.08	0.56	0.03	0.03	0.70
2	0.25	0.60	0.79	0.45	0.59	0.40	0.07	0.11	0.25
3	0.71	0.33	0.13	0.46	0.33	0.04	0.42	0.29	0.05
4							0.48	0.57	

Conditional probabilities: class 5

Full dependency, intensive need for care (4.9%)

	Bath	Bed	Chair	<i>Dist</i>	Dress	Eat	<i>Stair</i>	<i>Stoop</i>	Wash
1	0.02	0.02	0.02	0.03	0.02	0.13	0.02	0.01	0.04
2	0.02	0.06	0.13	0.06	0.04	0.38	0.02	0.02	0.30
3	0.97	0.93	0.85	0.91	0.94	0.49	0.06	0.05	0.66
4							0.91	0.91	

“Direct” and Model based estimates of the percentage of pop belonging to class 4 or 5 for each Health District



Outline

- 1 Introduction
 - Estimation framework
 - Latent class and latent trait models
- 2 Nonresponse
 - (Generalized) calibration
 - Response propensity estimation
- 3 Response: latent constructs
 - Disability
 - Other examples
- 4 Conclusions and further research

Social integration

- The survey on Integration of the Second Generation
- The questionnaire investigates many different dimensions of social inclusion
 - Use of native and local languages.
 - Relationship with schoolmates and teachers.
 - Relationship with friends, free time and social habits.
 - Composition of the family and household conditions.
- The aim is to study social and educational integration of foreign students in Italy by citizenship
- 9 items and two latent classes. Two-step approach with an area-level SAE model



Giovinazzi, F., Cocchi, D. (2022).

Social Integration of Second Generation Students in the Italian School System
Social Indicators Research 160, 287–307

Economic Well-being

- EU-SILC survey
- Multidimensionality of well-being indicators
- A subset of these indicators
 - severe material deprivation
 - equivalized disposable income
 - housing ownership
 - housing density
- Measure economic well-being indicators at a subregional level, NUTS3 and Local Administrative Units (LAU) 2, in Italy
- **Factor analysis**. Two-step approach with a unit-level SAE model



Moretti, A., Shlomo, N., Sakshaug, J. W. (2021)
Small area estimation of latent economic well-being,
Sociological Methods and Research, 50(4), 1660-1693.

Educational Poverty

- Deprivation of the ability to learn, experiment, develop and freely nourish skills, talents, and aspirations
- Multidimensionality of EP
- Composite index available from ISTAT
- 33 binary items AVQ survey
- The aim is to develop and measure EP indicators at a Regional and Degree of Urbanization level
- Latent trait model. Two-step approach with an area-level SAE model
- ? Multidimensional latent trait model. One-step approach with a unit-level SAE model



Bertarelli, G., R., M.G., , Pratesi, M. (202x)
Small Area Estimation of Educational Poverty with latent variable models,
 Work in progress

Educational Poverty

- Deprivation of the ability to learn, experiment, develop and freely nourish skills, talents, and aspirations
- Multidimensionality of EP
- Composite index available from ISTAT
- 33 binary items AVQ survey
- The aim is to develop and measure EP indicators at a Regional and Degree of Urbanization level
- Latent trait model. Two-step approach with an area-level SAE model
- ? Multidimensional latent trait model. One-step approach with a unit-level SAE model



Bertarelli, G., R., M.G., , Pratesi, M. (202x)

Small Area Estimation of Educational Poverty with latent variable models,

Work in progress

Employment status

- Estimation of the employment status of the Italian resident population
- Three sources: census survey data, labour force survey data and administrative information.
- None can be considered a benchmark
- Hidden Markov Models (longitudinal extension of latent class models)



Boeschoten, L., Filipponi, D., Varriale, R. (2021)

Combining multiple imputation and Hidden Markov Modeling to obtain consistent estimates of employment status,
Journal of Survey Statistics and Methodology, 9(3), 549-573.

Outline

- 1 Introduction
 - Estimation framework
 - Latent class and latent trait models
- 2 Nonresponse
 - (Generalized) calibration
 - Response propensity estimation
- 3 Response: latent constructs
 - Disability
 - Other examples
- 4 Conclusions and further research

Conclusions and further research

- Treatment of non-standard finite population estimation settings
- Nonresponse
- Unobservable responses
 - Composite indicators
 - Data integration (e.g. administrative and survey data)
- Measure of accuracy of final estimates
- Frequentist vs Bayesian
- Integration of design features
- Two-step vs One-step

Conclusions and further research

- Treatment of non-standard finite population estimation settings
- Nonresponse
- Unobservable responses
 - Composite indicators
 - Data integration (e.g. administrative and survey data)
- Measure of accuracy of final estimates
- Frequentist vs Bayesian
- Integration of design features
- Two-step vs One-step

Conclusions and further research

- Treatment of non-standard finite population estimation settings
- Nonresponse
- Unobservable responses
 - Composite indicators
 - Data integration (e.g. administrative and survey data)
- Measure of accuracy of final estimates
- Frequentist vs Bayesian
- Integration of design features
- Two-step vs One-step

References I



Bartholomew, D. J., Knott, M., and Moustaki, I. (2011).
Latent Variable Models and Factor Analysis. A unified Approach. 3rd Edition.
Wiley.



Bartholomew, D. J., Steele, F., Moustaki, I., Galbraith, J. I. (2002).
The Analysis and Interpretation of Multivariate Data for Social Scientists.
Chapman and Hall.



Biemer, P. P. (2011).
Latent class analysis of survey error.
John Wiley & Sons.



Boeschoten, L., Filipponi, D., Varriale, R. (2021)
*Combining multiple imputation and Hidden Markov Modeling to obtain
consistent estimates of employment status,*
Journal of Survey Statistics and Methodology, 9(3), 549-573.

References II



Chang, T. and Kott, P. S. (2008).

Using calibration weighting to adjust for nonresponse under a plausible model.
Biometrika, 95:555–571.



Deville, J.-C. (2000).

Generalized calibration and application to weighting for non-response.
In *Compstat - Proceedings in Computational Statistics: 14th Symposium Held in Utrecht, The Netherlands*, pages 65–76, New York. Springer.



Deville, J.-C. and Särndal, C.-E. (1992).

Calibration estimators in survey sampling.
Journal of the American Statistical Association, 87:376–382.



Fabrizi, E., Montanari, G. E., Ranalli, M. G. (2016).

A hierarchical latent class model for predicting disability small area counts from survey data.
Journal of the Royal Statistical Society: Series A, 179(1), 103-131.

References III



Folsom, R. E. and Singh, A. C. (2000).

The generalized exponential model for design weight calibration for extreme values, nonresponse and poststratification.

In Section on Survey Research Methods, ASA, 598–603.



Fuller, W. A., Loughin, M. M., and Baker, H. D. (1994).

Regression weighting in the presence of nonresponse with application to the 1987-1988 Nationwide Food Consumption Survey.

Survey Methodology, 20:75–85.



Giovinazzi, F., Cocchi, D. (2022).

Social Integration of Second Generation Students in the Italian School System
Social Indicators Research 160, 287–307








Goodman, L. A. (1974).

Exploratory latent structure analysis using both identifiable and unidentifiable models.

Biometrika, 61:215–231.

References IV

-  Kott, P. S. (2006).
Using calibration weighting to adjust for nonresponse and coverage errors.
Survey Methodology, 32:133–142.
-  Kott, P. S. and Chang, T. (2010).
Using calibration weighting to adjust for nonignorable unit nonresponse.
JASA, 105:1265–1275.
-  Lazarsfeld, P. and Henry, N. (1968).
Latent structure analysis.
Boston, Houghton Mifflin.
-  Lesage, E., Haziza, D., and D'Haultuille, X. (2019).
A cautionary tale on instrumental calibration for the treatment of nonignorable unit nonresponse in surveys.
Journal of the American Statistical Association, 114(526):906–915.
-  Montanari, G. E., R, M. G., Eusebi, P. (2011)
Latent variable modeling of disability in people aged 65 or more
Statistical Methods & Applications, 20, 49–63.

References V



Neri, A. and Ranalli, M. G. (2011).

To misreport or not to report? The measurement of household financial wealth.
Statistics in Transition, 12-2:281–300.



Ranalli, M. G., Matei, A., Neri, A. (2023)

Generalised calibration with latent variables for the treatment of unit nonresponse in sample surveys
Statistical Methods & Applications, 32(1), 169–195.



Särndal, C.-E. and Lundström, S. (2005).

Estimation in Surveys with Nonresponse.
Wiley, New York.



Shao, J. and Steel, P. (1999).

Variance estimation for survey data with composite imputation and nonnegligible sampling fractions.
JASA, 94:254–265.