# AN APPLICATION OF UNIT-LEVEL MODEL FOR FRACTIONS OF UNEMPLOYED

Tomas Rudys

Vilnius University

Fourth Baltic-Nordic Conference on Survey Statistics
24-28 August 2015, Helsinki, Finland

# Outline

1. Notation
2. Model
3. Simulation and results
4. Conclusions

**Aim of this talk**:
present simulation results on application of unit-level model for estimation of fractions of unemployed.
Simulation is based of real Lithuania LFS data.

# Notation

**Finite population** $\mathcal{U} = \{1, 2, ..., N\}$.
**Study variable** $y: \quad y_1, y_2, ..., y_N$

$$y_{ik} = \begin{cases} 1, & \text{if } k \text{ is unemployed,} \\ 0, & \text{otherwise,} \quad k = 1, 2, ..., N_i, \quad i = 1, 2, ...D. \end{cases}$$

$D = 57$ municipalities – small areas, of size $N_i$, $i = 1, 2, ..., D$.

**Parameter of interest** is fraction of unemployed in municipality $i$:

$$P_i = \tfrac{1}{N_i} \sum_{k \in \ munic_i} y_{ik}, \quad i = 1, 2, ..., D.$$

s $\subset \mathcal{U}$ is simple random sample of size $n$; $n_i$ – sample size in municipality.

# Logistic linear mixed model

Let $\mathbf{x}_{ik}$ denote the vector of fixed covariates, $k = 1, ..., N_i$, $i = 1, ..., D$. Assume two-stage model as folows [2]:

$$Y_{ik}|p_{ik} \sim \text{Bernoulli}(p_{ik})$$

$$logit(p_{ik}) = \mathbf{x}'_{ik}\boldsymbol{\beta} + v_i, \quad v_i \sim \mathcal{N}(0, \sigma_v^2),$$

where $\mathbf{x}_{ik}$ is vector of fixed model covariates, $i = 1, 2, ..., D$, $k = 1, ..., N_i$; $\boldsymbol{\beta}$ is vector of unknown model parameters;$v_i$ is random area effects independent from $\boldsymbol{\beta}$.

## Case I: Hierarchical Bayes (HB) version of logistic linear mixed model

Assume that mentioned model holds for the sample $\{(y_{ik}, \mathbf{x}_{ik}), k \in \mathbf{s}_i, i = 1, ...D\}$. A HB version of this model [2]:

(i) $y_{ik}|p_{ik} \sim$ Bernoulli$(p_{ik})$,

(ii) $logit(p_{ik}) = \mathbf{x}'_{ik}\boldsymbol{\beta} + v_i, \quad v_i \sim \mathcal{N}(0, \sigma_v^2)$,

(iii) $\boldsymbol{\beta}$ and $\sigma_v^2$ are mutually independent with

$$\boldsymbol{\beta} \sim \mathcal{N}(0, 1000),$$

$$\sigma_v^{-2} \sim \Gamma(a, b), \quad a > 0, \ b > 0.$$

## Case II: Logistic linear mixed model without random area effects

Consider the following model for $y_{ik}$, $k \in \mathbf{s}_i, i = 1, ...D$ [1]:

(i) $y_{ik}|p_{ik} \sim \text{Bernoulli}(p_{ik})$,

(ii) $logit(p_{ik}) = \mathbf{x}'_{ik}\boldsymbol{\beta}$,

(iii) $\boldsymbol{\beta} \sim \mathcal{N}(0, 1000)$.

# Estimation of model parameters

- MCMC algorithm Metropolis-within-Gibbs (suggested in Tierney 1994) was used for approximation of posterior distributions;
- R package *LaplacesDemon* [3] was used in simulations.

# Estimator of fraction

HB estimator of the finite population proportion $P_i$ [2]:

$$\widehat{P}_i^{HB} \approx \frac{1}{N_i}\left(\sum_{k\in \mathbf{s}_i} y_{ik} + \frac{1}{M}\sum_{j=1}^{M}\sum_{l\in \bar{\mathbf{s}}_i} \widehat{p}_{il}^{(j)}\right)$$

where

$$\widehat{p}_{il}^{(j)} = \frac{exp(\mathbf{x}_{il}'\widehat{\boldsymbol{\beta}}^{(j)})}{1+\mathbf{x}_{il}'\widehat{\boldsymbol{\beta}}^{(j)}},$$

$\bar{\mathbf{s}}_i$ is the set of nonsampled units in the $i$ th area. M is the number of MCMC samples obtained by MWG algorithm. Note that $\mathbf{x}_{il}$-values for $l \in \bar{\mathbf{s}}_i$ are needed.

# Case III: logistic regression model

Consider logistic regression model for $Y_{ik}$, $k = 1, ..., N_i$, $i = 1, ..., D$:

$$logit(\mathbf{E}(Y_{ik}|\mathbf{x}_{ik})) = logit(p_{ik}) = ln\left(\frac{p_{ik}}{1 - p_{ik}}\right) = \mathbf{x}'_{ik}\boldsymbol{\beta},$$

where

$$p_{ik} = \frac{exp(\mathbf{x}'_{ik}\boldsymbol{\beta})}{1 + exp(\mathbf{x}'_{ik}\boldsymbol{\beta})}.$$

Unknown model parameters $\boldsymbol{\beta}$ are estimated using the sample data and maximum likelihood method. Then the estimator of fraction for area $i$ is:

$$\widehat{P}_i^{LR} = \frac{1}{N_i} \sum_{k \in \mathcal{U}_i} \widehat{p}_{ik}, \quad i = 1, ..., D.$$

# Simulation study

Lithuanian Labor Force Survey data 2012.

Population size of $N = 22\,382$ individuals (15-74 years old).

Sample of size $n = 2\,000$ individuals.

Auxiliary variables:

- registered unemployed individuals,
- male,
- urban population,
- 55-74 years old population.

Estimates HT, LR1, LR2, LR3

for small area fractions $P_i$, $i = 1, ..., D$, $D = 57$ for
$R = 30$ simulation runs were calculated.

# Accuracy measures

Let $\widehat{P}_{ir}$, $i = 1, 2, ..., D$, denote small area estimates in the $r$th simulation run.

1. The root mean relative square error

$$
\begin{aligned}
rrmse_i &= \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\frac{\widehat{P}_{ir} - P_i}{P_i}\right)^2}, \\
rrmse &= \frac{1}{D}\sum_{i=1}^{D} rrmse_i.
\end{aligned}
$$

# Accuracy measures

2. The simulation standard error

$$se_i = \sqrt{\frac{1}{R-1}\sum_{r=1}^{R}\left(\widehat{P}_{ir} - \frac{1}{R}\sum_{r=1}^{R}\widehat{P}_{ir}\right)^2},$$

$$se = \frac{1}{D}\sum_{i=1}^{D} se_i.$$

3. The simulation bias

$$bias_i = \frac{1}{R}\sum_{r=1}^{R}\widehat{P}_{ir} - P_i,$$

$$bias = \frac{1}{D}\sum_{i=1}^{D} bias_i.$$

# Accuracy measures of the estimates

| Estimator | $rrmse$ | $se$ | $bias$ |
|---|---|---|---|
| HT | 0.7911 | 0.0609 | 0.0006 |
| LR1 | 0.4413 | 0.0105 | -0.0314 |
| LR2 | 0.4402 | 0.0103 | -0.0323 |
| LR3 | 0.4228 | 0.0071 | -0.0011 |

# Conclusions

Simulation results show:

- All error measures of the estimates obtained using models are smaller than in the case of Hotvitz-Thompson estimator.
- Estimates with logistic regression model, where unknown model parameters are estimated using maximum likelihood method has smaller root mean square relative error and standard error.

# References


1. H. J. Boonstra, B. Buelens, and M. Smeets. *Estimation of municipal unemployment fractions – a simulation study comparing different small area estimators*, Statistics Netherlands, Projectnr: DMH-205714 (2009).
2. J. N. K. Rao, *Small area estimation*, Hoboken: John Wiley & Sons (2003).
3. Statisticat LLC (2014). *LaplacesDemon: Complete environment for Bayesian Inference*. R package version 14.04.05, URL `http://www.bayesian-inference.com/software`.