Coherent small area estimates for skewed business data

EESW 2013

Session 2, Advanced methods for processing and analysis
Chair: Maria Dolores Ugarte

Thomas Zimmermann   Ralf Münnich

Nuremberg, 9th September 2013

SAE and Business Data

- Small area methods are now in wide use
  - Geographical areas of interest
  - Domains of interest, e.g. NACE classes
- Business data characterized by outliers and skewed distributions $\rightarrow$ violation of assumptions
- Relationships between variables may be multiplicative
- Applying transformations may help to recover some of these assumptions
- Business surveys often based on designs with highly different weights
- Interaction between designs and models is of crucial importance
- Estimates for small areas should be coherent with estimates for aggregates

Estimators based on transformations

We may assume the following unit-level lognormal-mixed model (Berg and Chandra, 2012)

$$\log(y_{dj}) = \mathbf{x}_{dj}^T\beta + u_d + \varepsilon_{dj}, \quad d = 1, \ldots, D, \ j = 1, \ldots, N_d$$

where $\mathbf{x}_{dj}$ includes an intercept and the other components of it are appropriately transformed. $u_d \overset{i.i.d.}{\sim} N(0, \sigma_u^2)$ is the domain-specific random effect and $\varepsilon_{dj} \overset{i.i.d.}{\sim} N(0, \sigma_\varepsilon^2)$ the individual error term. The domain-specific random effect is assumed to be independent from the error term.

An optimal predictor

Minimizing the MSE under the unit-level lognormal mixed model yields the Empirical Bayes predictor

$$\hat{\theta}_d^{EBLOG} = \frac{1}{N_d} \left(\sum_{j \in s_d} y_{dj} + \sum_{j \notin s_d} \hat{y}_{dj}^{EBLOG} \right) \tag{1}$$

derived by Berg and Chandra (2012). The predictions for the non-sampled values ($j \notin s_d$) are given by :

$$\hat{y}_{dj}^{EBLOG} = \exp\left(\mathbf{x}_{dj}^T \hat{\boldsymbol{\beta}} + \hat{u}_d + 0.5 \hat{\sigma}_\varepsilon^2 (\hat{\gamma}_d / n_d + 1) \right) \tag{2}$$

with $\hat{\gamma}_d = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_\varepsilon^2 / n_d}$.

Area Level Lognormal Model

Assuming that the direct means are lognormally distributed, Slud and Maiti (2006) propose the following predictor:

$$\hat{\theta}_d^{ALLOG} = \exp\left(\bar{\mathbf{X}}_d^T \hat{\beta} + \hat{u}_d + 0.5\hat{\sigma}_u^2 \left(1 - \hat{\gamma}_d\right)\right) \quad (3)$$

Estimator (3) corrects for the presence of the random effect but ignores the variability of the parameter estimates.

Other Estimators

Design-based / Model-assisted Estimators

- Direct estimator, which is a weighted sample mean
- Generalized Regression Estimators:

$$\hat{\theta}_d^{GREG} = \frac{1}{\widehat{N}_d} \left[\sum_{k \in U_d} \hat{y}_k + \sum_{k \in s_d} w_k \left(y_k - \hat{y}_k \right) \right]$$

GREG Linear fixed-effects model used to predict $\hat{y}_k$

MLogGREG Predictions $\hat{y}_k^{EBLOG}$ are used

Benchmarked Estimators

We benchmark estimator (1) against the weighted sample total for the population to obtain the LOGBench predictor.

Dataset

- Our dataset is based on
  - the Italian register of enterprises (ASIA 2003)
  - and the survey of small and medium enterprises (PMI)
- We focus on the subset of small and medium enterprises $\rightarrow$ about 4.3 million entries
- Our variable of interest is the mean of labour costs in each domain
- Auxiliary information: Number of employees of each enterprise
- The original datasets were kindly provided by ISTAT

Setup

- Strata are cross-classifications of the first digit of the industry classification, Italian NUTS 1 areas and the classified size variable in terms of numbers of employees
- As most enterprises in the data set have less than 5 employees, we aggregate the size variable
  - Group 1 All enterprises with 1 – 5 employees
  - Group 2 Enterprises with 6 – 99 employees
- Stratum sizes vary between 799 and 364294
- Focus on SME: no *take-all stratum*
- Total sample size of $n = 67{,}989$
- $R = 10{,}000$ simulation runs

Domains

We consider two types of domains

1. **Planned domain structures**
   Domains as cross-classifications of NUTS 1 and the first digit of the industry classifcation
   $D = 45$ domains
   Domain sizes vary between 6340 and 398874
2. **Unplanned domain structures**
   Domains as cross-classifications of Italy's 20 regions and the first digit of the industry classifcation
   $D = 180$ domains
   Domain sizes range from 144 to 229873

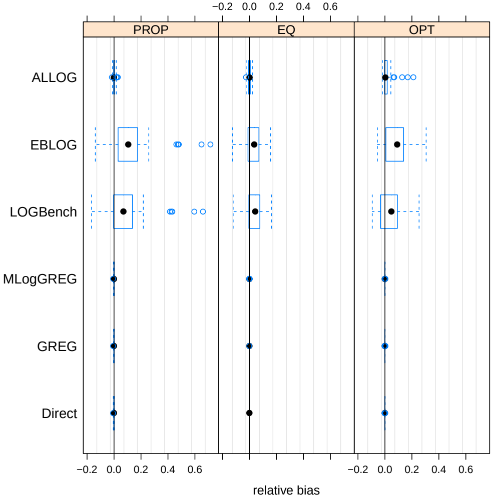
Gelman Factors and distribution of weights

Following Münnich and Burgard (2012) the **Gelman factor** is defined as the ratio of the largest to the smallest (design) weight:

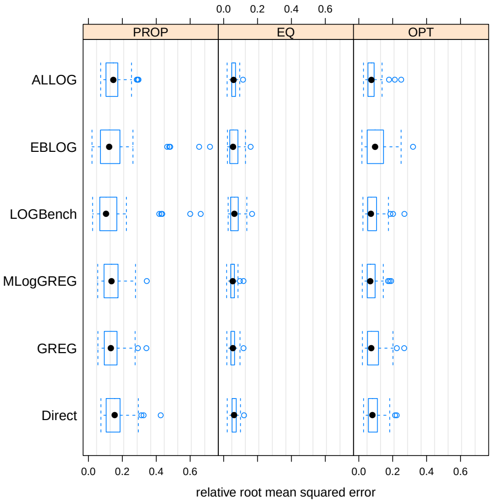
$$GF = \frac{\max\limits_{i=1,\ldots,N} \frac{1}{\pi_i}}{\min\limits_{i=1,\ldots,N} \frac{1}{\pi_i}}$$

Allocation	max/min	q95/q05	q75/q25
PROP	1.06	1.01	1.00
EQ	455.33	134.95	6.95
OPT	73.38	41.99	18.55

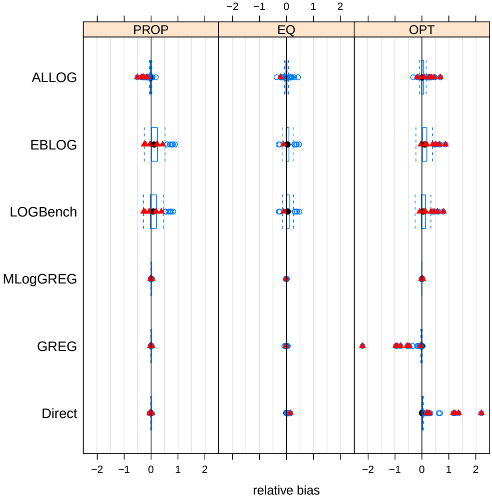
Relative Bias - planned domains



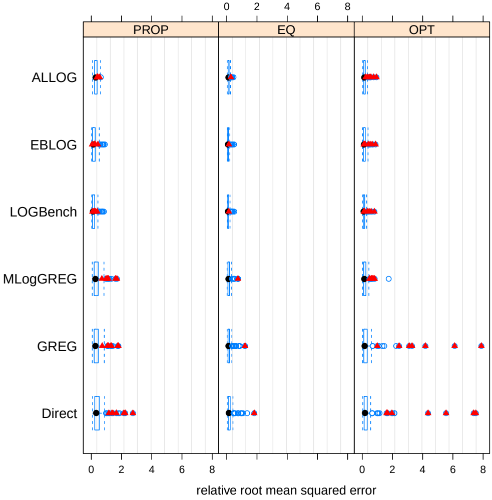
RRMSE - planned domains



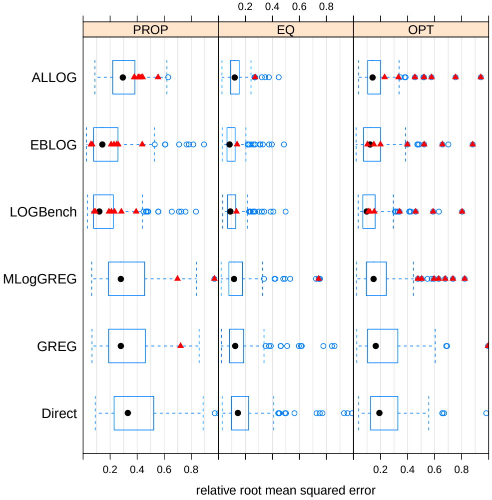
Relative Bias - unplanned domains



RRMSE - unplanned domains



RRMSE - unplanned domains



Summary and Outlook

- Model-assisted estimators best choice for planned (large) domain structurs
- For unplanned domain structures model-based estimators help to produce more reliable estimates
- In this application benchmarking is desirable for the estimation at domain level as well
- Incorporating design information may be beneficial for model-based estimators
- MSE estimation for log-transformed estimators is very computerintense

Acknowledgement

BLUE-ETS (Enterprise and Trade Statistics) is funded by the European Commission within the Seventh Framework Programme