

# Generalized regression and model-calibration estimation for domains: Accuracy comparison

Risto Lehtonen<sup>1</sup>, Carl-Erik Särndal<sup>2</sup> and Ari Veijanen<sup>3</sup>

<sup>1</sup> University of Helsinki, Finland  
e-mail: [risto.lehtonen@helsinki.fi](mailto:risto.lehtonen@helsinki.fi)

<sup>2</sup> Université de Montréal, Canada  
e-mail: [carl-erik.sarndal@rogers.com](mailto:carl-erik.sarndal@rogers.com)

<sup>3</sup> Statistics Finland, Finland  
e-mail: [ari.veijanen@stat.fi](mailto:ari.veijanen@stat.fi)

## 1 Preliminaries and key questions

In this paper, we discuss generalized regression (GREG) estimation and model-calibration estimation for population subgroups or domains under unequal probability sampling. The classical GREG estimator of Särndal, Swensson and Wretman (1992) uses a fixed-effects linear assisting model. A multinomial logistic model is introduced as an assisting model for GREG in Lehtonen and Veijanen (1998). Logistic GREG has been examined further for domain estimation in Lehtonen, Särndal and Veijanen (2003, 2005), Lehtonen and Veijanen (2008) and Myrskylä (2007).

Model calibration (MC), which also provides a design-based method, was introduced by Wu and Sitter (2001) and was further discussed in Wu (2003), Lehtonen, Myrskylä, Särndal and Veijanen (2007), Särndal (2007) and Lehtonen, Särndal and Veijanen (2008). A key property of MC is that the weights are calibrated to the population total of the predictions derived via an assumed model. For comparability with the GREG approach, we use a logistic model. Under this model, GREG and MC require an access to unit-level auxiliary information. Both GREG and MC provide nearly design unbiased methods.

We extend in this paper the model-calibration method to domain estimation. We present results on the accuracy of logistic GREG and MC estimators of domain totals of a binary response variable. The results are based on Monte Carlo experiments where repeated probability proportional-to-size samples were drawn from an artificially generated finite population.

The different combinations of the level of model calibration (at the population level, at the domain level) and model type (common models, models with domain-specific parameters) are illustrated in Table 1. In this paper, we examine the performance of GREG and MC with different parametrization of the logistic regression model. For a fair comparison with GREG, we consider the case where MC is at the domain level. This is because we assume that the domains are identifiable (domain membership is known for every population element). The estimators MC-P-C and MC-P-D, which use model calibration at the population level, are thus excluded. We ask: Does MC outperform GREG for certain model choices– or vice versa?

**Table 1.** GREG and MC estimators by the level of model calibration and model type

Level of model calibration	Model type	
	C: Common model formulation for all domains	D: Model formulation with domain-specific parameters
P: Population level	MC-P-C	MC-P-D
D: Domain level	MC-D-C	MC-D-D
N: None	GREG-N-C	GREG-N-D

## 2 Methods

This section contains a skeleton of the methodology.

Notation

$U = \{1, 2, \dots, k, \dots, N\}$  Population (fixed, finite)

$U_1, \dots, U_d, \dots, U_D$  Domains of interest (non-overlapping, identifiable)

### Sampling design

Systematic  $\pi$ PS with sample size  $n$

$s$  Sample from  $U$

$s_d = s \cap U_d$  Random part of  $s$  falling in domain  $d$

$\pi_k = n \frac{x_{1k}}{\sum_{k \in U} x_{1k}}$  Inclusion probability for  $k \in U$  in  $\pi$ PS with  $x_1$  as the size variable

$a_k = 1/\pi_k$  Sampling weight for  $k \in s$

We observe values  $y_k$  of binary response variable  $y$  for  $k \in s$

### Models

(1) Common model formulation for all domains

Estimators MC-P-C, MC-D-C and GREG-N-C

The logistic model is given by

$$E_m(y_k) = \frac{\exp(\mathbf{x}'_k \boldsymbol{\beta})}{1 + \exp(\mathbf{x}'_k \boldsymbol{\beta})}$$

where

$\mathbf{x}_k = (1, x_{1k}, \dots, x_{pk})'$ , known for every  $k \in U$

$\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)'$

$\beta_j$  are fixed effects common for all domains,  $j = 0, \dots, p$

(2) Model formulation with domain-specific intercepts

Estimators MC-P-D, MC-D-D and GREG-N-D

The logistic model is given by

$$E_m(y_k) = \frac{\exp(\mathbf{x}'_k \boldsymbol{\beta})}{1 + \exp(\mathbf{x}'_k \boldsymbol{\beta})}$$

where

$\mathbf{x}_k = (I_{1k}, \dots, I_{Dk}, x_{1k}, \dots, x_{pk})'$ , known for every  $k \in U$

$I_{dk} = 1$  if  $k \in U_d$ ,  $I_{dk} = 0$  otherwise,  $d = 1, \dots, D$

$\boldsymbol{\beta} = (\beta_{01}, \dots, \beta_{0D}, \beta_1, \dots, \beta_p)'$

$\beta_{0d}$  are domain-specific intercepts,  $d = 1, \dots, D$

$\beta_j$  are common slopes,  $j = 1, \dots, p$

## Target parameters and GREG and MC estimators

Target parameters  $Y_d = \sum_{U_d} y_k$ ,  $d = 1, \dots, D$

Domain totals of binary response variable  $y$

### GREG estimators GREG-N-C and GREG-N-D

$$\hat{Y}_{dGREG} = \sum_{k \in U_d} \hat{y}_k + \sum_{k \in s_d} a_k \hat{e}_k, \quad d = 1, \dots, D,$$

where weights are  $a_k = 1/\pi_k$  and residuals are  $\hat{e}_k = y_k - \hat{y}_k$

Fitted values  $\hat{y}_k = \frac{\exp(\mathbf{x}'_k \hat{\boldsymbol{\beta}})}{1 + \exp(\mathbf{x}'_k \hat{\boldsymbol{\beta}})}$  are calculated for every  $k \in U$

NOTE: Domain membership need to be known for every  $k \in U$

### Model calibration estimators

#### (1) Model calibration at the population level MC-P-C and MC-P-D

MC estimators are of the form  $\hat{Y}_{dMC} = \sum_{k \in s_d} w_k y_k$

Calibration: Find weights  $w_k$  to satisfy calibration equation

$$\sum_{k \in s} w_k \mathbf{z}_k = \sum_{k \in U} \mathbf{z}_k = \mathbf{z}_U \quad \text{where } \mathbf{z}_k = (1, \hat{y}_k)'$$

Solution: Minimize

$$\sum_{k \in s} \frac{(w_k - a_k)^2}{a_k} - \lambda'(\sum_{k \in s} w_k \mathbf{z}_k - \mathbf{z}_U)$$

under calibration constraints

$$\sum_{k \in s} w_k = N \quad \text{and} \quad \sum_{k \in s} w_k \hat{y}_k = \sum_{k \in U} \hat{y}_k$$

NOTE: Domain membership need not to be known for every  $k \in U$

#### (2) Model calibration at the domain level MC-D-C and MC-D-D

MC estimators are of the form  $\hat{Y}_{dMC} = \sum_{k \in s_d} w_{dk} y_k$

Calibration: Find weights  $w_{dk}$  to satisfy calibration equation

$$\sum_{k \in s_d} w_{dk} \mathbf{z}_k = \sum_{k \in U_d} \mathbf{z}_k = \mathbf{z}_{U_d} \quad \text{where } \mathbf{z}_k = (1, \hat{y}_k)'$$

Solution: Minimize

$$\sum_{k \in s_d} \frac{(w_{dk} - a_k)^2}{a_k} - \lambda'(\sum_{k \in s_d} w_{dk} \mathbf{z}_k - \mathbf{z}_{U_d})$$

under calibration constraints

$$\sum_{k \in s_d} w_{dk} = N_d \quad \text{and} \quad \sum_{k \in s_d} w_{dk} \hat{y}_k = \sum_{k \in U_d} \hat{y}_k$$

NOTE: Domain membership need to be known for every  $k \in U$

## Monte Carlo simulation

Artificial finite population

One million elements

$D = 100$  domains, size of domain proportional to  $\exp(u)$ ,  $u \sim U(0, 2.9)$

Population generating model: Logistic mixed model

The binary random variable  $Y_k$  was defined by  $P\{Y_k = 1\} = \frac{\exp(\eta_k)}{1 + \exp(\eta_k)}$

with  $\eta_k = (u_{0d} - 5) + (1 + u_{1d})x_{1k} + (1 + u_{2d})x_{2k}$

where

$x_1$  size variable in  $\pi$ PS,  $x_1 \sim U(1, 11)$

$x_2 \sim U(-5, 5)$ , independent of  $x_1$

Random effects  $N(0, \sigma_u^2)$ ,  $\sigma_{u_0}^2 = 9$ ,  $\sigma_{u_i}^2 = 0.125$ ,  $\text{Corr}(u_0, u_i) = -0.5$ ,  $i = 1, 2$

Binary  $y_k$  : If a random number from  $U(0, 1)$  was smaller than the computed value  $P\{Y_k = 1\}$ ,

then  $y_k = 1$ , otherwise  $y_k = 0$

Correlations:  $\text{corr}(y, x_1) = 0.41$

$\text{corr}(y, x_2) = 0.32$

$\text{corr}(x_1, x_2) = -0.001$

Sampling

$K = 1000$  independent with-replacement samples drawn with  $\pi$ PS

Size variable:  $x_1$

Sample size  $n = 10,000$  elements

Accuracy measure

Relative root mean squared error

$$\text{RRMSE}(\hat{Y}_d) = \sqrt{\frac{1}{K} \sum_{v=1}^K (\hat{Y}_d(s_v) - Y_d)^2 / Y_d}, \quad d = 1, \dots, D$$

RRMSE figures are averaged over domain sample size classes:

Minor (20-69)      47 domains

Medium (70-119)   19 domains

Major (120-)      34 domains

## 3 Results

Results on the accuracy of GREG and MC estimators are in Table 2. Now, we consider the case where MC is at the domain-level. The results indicate that for the common model type, the accuracy of MC is better than that of GREG. GREG and MC indicate similar accuracy when model calibration is at the domain level and the underlying model contains domain-specific intercept terms.

The structure of the linear part of the model affects accuracy. Accuracy tends to improve for both estimators when using models that incorporate domain-specific intercepts. Usually, accuracy is better for estimators whose model includes the  $\pi$ PS size variable  $x_1$ . Best accuracy is for models, which include both  $x_1$  and  $x_2$ .

**Table 2.** Accuracy of GREG and MC estimators by model type

Model type	Linear part of model	Estimator of domain total	Mean RRMSE (%)		
			Minor (20-69)	Medium (70-119)	Major (120-)
Common models	$\beta_0 + \beta_1 x_{1k}$	MC-D-C	26.4	12.6	13.2
		GREG-N-C	28.9	13.4	16.3
Common intercept and common slopes	$\beta_0 + \beta_2 x_{2k}$	MC-D-C	26.1	13.2	13.0
		GREG-N-C	28.2	13.8	14.6
	$\beta_0 + \beta_1 x_{1k} + \beta_2 x_{2k}$	MC-D-C	22.1	9.9	10.8
		GREG-N-C	24.8	10.9	14.3
Domain-specific models	$\beta_{0d} + \beta_1 x_{1k}$	MC-D-D	26.2	12.5	13.2
		GREG-N-D	24.9	12.4	13.1
Domain-specific intercepts and common slopes	$\beta_{0d} + \beta_2 x_{2k}$	MC-D-D	25.9	13.2	12.9
		GREG-N-D	26.0	13.2	12.9
	$\beta_{0d} + \beta_1 x_{1k} + \beta_2 x_{2k}$	MC-D-D	20.3	8.9	9.9
		GREG-N-D	19.8	8.9	9.8

## 4 Conclusions

The comparison of GREG and MC methods for domain estimation shows that GREG is more sensitive to the model choice than MC. If the explanatory power of the assisting model of GREG is “weak”, model calibration can improve accuracy. This can happen if calibration is at the domain level. But if the explanatory power of the model is “strong”, model calibration does not necessarily improve accuracy. These findings are important for practical purposes. Other variants of model calibration estimators and their relation to the GREG family estimators are topics of further research.

## References

- Lehtonen, R. and Veijanen, A. (1998). Logistic generalized regression estimators. *Survey Methodology* 24, 51-55.
- Lehtonen, R., Särndal, C.-E. and Veijanen, A. (2003). The effect of model choice in estimation for domains, including small domains. *Survey Methodology*, 29, 33-44.
- Lehtonen, R., Särndal, C.-E. and Veijanen, A. (2005). Does the model matter? Comparing model-assisted and model-dependent estimators of class frequencies for domains. *Statistics in Transition*, 7, 649-673.
- Lehtonen, R., Myrskylä, M., Särndal, C.-E. and Veijanen, A. (2007). Estimation for domains and small areas under unequal probability sampling. Invited paper, the SAE2007 Conference, Pisa, September 2007. (CD rom).
- Lehtonen, R. and Veijanen, A. (2008). Design-based methods of estimation for domains and small areas. Chapter 31 in C.R. Rao and D. Pfeffermann (Eds.) *Handbook of Statistics*, Vol 29, Sample Surveys: Theory, Methods and Inference. New York: Elsevier. (in press).
- Lehtonen, R., Särndal C.-E. and Veijanen, A. (2008). Generalized regression and model-calibration estimation for domains. Invited paper, NORDSTAT 2008 Conference, Vilnius, June 2008.
- Myrskylä, M. (2007). Generalized regression estimation for domain class frequencies. Helsinki: Statistics Finland, *Research Reports* 247.
- Särndal, C.-E. (2007). The calibration approach in survey theory and practice. *Survey Methodology* 33, 99-119.
- Särndal, C.-E., Swensson, B. and Wretman, J.H. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- Wu, C. (2003). Optimal calibration estimators in survey sampling. *Biometrika*, 90, 937-951
- Wu, C. and Sitter, R.R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96, 185-193