HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

SAE course Spring 2015
Risto Lehtonen, University of Helsinki

EXTENDED FAMILY OF GREG ESTIMATORS

GREG estimator

$$\hat{t}_{dGREG} = \sum_{k \in U_d} \hat{y}_k + \sum_{k \in s_d} a_k (y_k - \hat{y}_k), \quad d = 1, \ldots, D$$

EXAMPLE: Continuous study variable *y,* sample data *s*
Assisting model: P-model:

$$y_k = \mathbf{x}_k' \boldsymbol{\beta} + \varepsilon_k, \qquad k \in U \tag{1}$$

where $\mathbf{x}_k = (1, x_{1k}, \ldots, x_{Jk})'$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_J)'$

WLS estimator of finite population counterpart $\mathbf{B}$ of vector $\boldsymbol{\beta}$

$$\hat{\mathbf{B}} = \left(\sum_{k \in s} a_k \mathbf{x}_k \mathbf{x}_k' \right)^{-1} \sum_{k \in s} a_k \mathbf{x}_k y_k$$

Fitted values: $\hat{y}_k = \mathbf{x}_k' \hat{\mathbf{B}}$ for all $k \in U$
Residuals: $e_k = y_k - \hat{y}_k$ for all $k \in s$

NOTE: The P-model (1) does not involve components that account for the possible population heterogeneity (differences between the domains)

Alternative model formulations to account for the domain differences?

Assisting models

GREG estimators with assisting models from the family of generalized linear mixed models (GLMM)

(1) Generalized linear fixed-effects model

$$E_m(y_k) = f(\mathbf{x}_k;\boldsymbol{\beta})$$

where $f(\cdot;\boldsymbol{\beta})$ describes the functional form

Continuous study variable *y*: linear model
Binary y-variable: binomial logistic model
Count variable *y*: Poisson regression model (log)

$\boldsymbol{\beta}$ is the vector of model parameters
E_m refers to expectation with respect to model

The model is fitted to sample data $\{(y_k, \mathbf{x}_k);\ \ k \in s\}$ by accounting for the sampling design properties (unequal probability sampling, stratification, weights)

We obtain estimates $\hat{\mathbf{B}}$ of parameter vector $\mathbf{B}$, where $\mathbf{B}$ is again the finite population substitute to the superpopulation parameter $\boldsymbol{\beta}$

Fitted values $\hat{y}_k = f(\mathbf{x}_k;\hat{\mathbf{B}})$ are calculated for all $k \in U$ by using estimates $\hat{\mathbf{B}}$ and auxiliary data $\mathbf{x}_k$

(2) Generalized linear mixed model

$$E_m(y_k \mid \mathbf{u}_d) = f(\mathbf{x}_k'(\boldsymbol{\beta} + \mathbf{u}_d)) \tag{2}$$

where $f(\cdot;\boldsymbol{\beta} + \mathbf{u}_d)$ describes the functional form
$\mathbf{u}_d$ denotes the vector of domain level random effects

Fitted values $\hat{y}_k = f(\mathbf{x}_k'(\hat{\mathbf{B}} + \hat{\mathbf{u}}_d))$ are calculated for all $k \in U$ by using estimates $\hat{\mathbf{B}}$ and $\hat{\mathbf{u}}_d$ and auxiliary data $\mathbf{x}_k$

(a) Linear mixed model for continuous y-variable

$$\begin{aligned} E_m(y_k \mid \mathbf{u}_d) &= \mathbf{x}_k'(\boldsymbol{\beta} + \mathbf{u}_d) \\ &= (\beta_0 + u_{0d}) + (\beta_1 + u_{1d})\, x_{1k} + \ldots + (\beta_J + u_{Jd})\, x_{Jk} \end{aligned} \tag{3}$$

where $\mathbf{u}_d = (u_{0d}, u_{1d}, \ldots, u_{Jd})'$ is vector of domain level random effects (random intercepts and slopes)

EXAMPLE: Model with random intercept term u_{0d} and random slope term u_{1d} associated to variable x_{1k}

$$E_m(y_k \mid \mathbf{u}_d) = (\beta_0 + u_{0d}) + (\beta_1 + u_{1d})\, x_{1k} + \beta_2\, x_{2k}$$

NOTE: The corresponding fixed-effects P-model:

$$E_m(y_k) = \beta_0 + \beta_1\, x_{1k} + \beta_2\, x_{2k}$$

What is the main difference in these two models?

(b) **Logistic mixed model** for binary y-variable

Binomial logistic mixed model

$$E_m(y_k \mid \mathbf{u}_d) = P\{y_k = 1 \mid \mathbf{u}_d\} = \frac{\exp(\mathbf{x}_k'(\boldsymbol{\beta} + \mathbf{u}_d))}{1 + \exp(\mathbf{x}_k'(\boldsymbol{\beta} + \mathbf{u}_d))} \qquad (4)$$

where the study variable *y* is **binary**

EXAMPLE: 0: Unemployed
1: Employed

NOTE: The study variable y can be **polytomous**
Multinomial logistic mixed model

EXAMPLE: 1: Employed
2: Unemployed
3: Not in labour force

Lehtonen, R., C.-E. Särndal, and A. Veijanen (2003). The effect of model choice in estimation for domains, including small domains. *Survey Methodology* **29**, 33-44.

Lehtonen, R., C.-E. Särndal, and A. Veijanen (2005). Does the model matter? Comparing model-assisted and model-dependent estimators of class frequencies for domains. *Statistics in Transition* **7**, 649-673.

NOTE: Fixed-effects model corresponding to (4)

$$E_m(y_k) = P\{y_k = 1\} = \frac{\exp(\mathbf{x}_k'\boldsymbol{\beta})}{1 + \exp(\mathbf{x}_k'\boldsymbol{\beta})} \qquad (5)$$

LGREG: GREG estimator assisted by the model (5)

Lehtonen, R. and A. Veijanen (1998). Logistic generalized regression estimators. *Survey Methodology* **24**, 51-55.