EXAMPLE: VARIANCE ESTIMATION FOR GREG UNDER SRS

GREG estimator is given by

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

(1) Direct GREG for planned domains

Assume **stratified SRSWOR** sampling (proportional allocation). Sample of n_d elements is drawn from the population of N_d elements in domain *d.* Domain sample sizes n_d are fixed by the sampling design. Design weights are $a_k = N_d / n_d$ for domain *d*.

Assisting model for direct GREG:

$y_k = \beta_{0d} + \beta_{1d} x_{1k} + ... + \beta_{Jd} x_{Jk} + \varepsilon_k = \mathbf{x}'_k \boldsymbol{\beta}_d + \varepsilon_k$ for $k \in U_d$, $d = 1,...,D$,

where

$\mathbf{x}_k = (1, x_{1k},...,x_{jk},...,x_{Jk})'$ is vector of auxiliary variable values known for every $k \in U$

$\boldsymbol{\beta}_d = (\beta_{0d}, \beta_{1d},...,\beta_{Jd})'$ is a vector of fixed effects defined for each domain separately

Model fitting by OLS. We obtain fitted values $\hat{y}_k = \mathbf{x}'_k \hat{\boldsymbol{\beta}}_d$ for $k \in U_d$.

Approximate variance estimator for direct GREG estimator for planned domains

$$\hat{V}_{srs}(\hat{t}_{dGREG}) = N_d^2 (1 - \frac{n_d}{N_d})(\frac{1}{n_d}) \sum_{k\in s_d} \frac{(e_k - \bar{e}_d)^2}{n_d - 1}$$

where residuals are $e_k = y_k - \hat{y}_k$, $k \in s_d$, and $\bar{e}_d = \sum_{k\in s_d} e_k / n_d$ is the mean of residuals in domain d $(d = 1,...,D)$.

(2) Indirect GREG for unplanned domains

Assume **SRSWOR sampling** with *n* elements drawn from the population of *N* elements. Domain sample sizes n_d are now random. Sampling fraction is n / N and design weights are $a_k = N / n$ for all $k \in U$.

Assisting model for indirect GREG:

$y_k = \beta_0 + \beta_1 x_{1k} + ... + \beta_J x_{Jk} + \varepsilon_k = \mathbf{x}'_k \boldsymbol{\beta} + \varepsilon_k$ for $k \in U$

where

$\mathbf{x}_k = (1, x_{1k},...,x_{jk},...,x_{Jk})'$ is vector of auxiliary variable values known for every $k \in U$

$\boldsymbol{\beta} = (\beta_0, \beta_1,...,\beta_J)'$ is a vector of fixed effects defined for the whole population

Model fitting by OLS. We obtain fitted values $\hat{y}_k = \mathbf{x}'_k \hat{\boldsymbol{\beta}}$ for $k \in U$.

By denoting $y_{dk} = I\{k \in U_d\} y_k$ and $e_{dk} = y_{dk} - \hat{y}_k$, $d = 1,...,D$, we obtain an **approximate variance estimator**:

$$\hat{V}_{srs}(\hat{t}_{dGREG}) = N^2 (1 - \frac{n}{N})(\frac{1}{n}) \sum_{k\in s} \frac{(e_{dk} - \bar{e}_d)^2}{n - 1}$$

Note that also elements outside the domain *d* contribute to the variance estimate, because $e_{dk} = -\hat{y}_k$ for elements $k \notin U_d$ and $k \in s$.

Next page: Illustration of HT and GREG estimation under more complex sampling design.

Table 3. Mean absolute relative error MARE (%) and mean coefficient of variation MCV (%) of direct HT and direct GREG estimators for **planned** domains.

	HT		**GREG**			
Auxiliary information	1 None		2 Domain sizes and domain totals of EMP		3 Domain sizes and domain totals of EMP and EDUC	
Domain sample size class	MARE %	MCV %	MARE %	MCV %	MARE %	MCV %
Minor $8 \le n_d \le 33$	11.5	11.9	5.8	7.7	6.4	6.8
Medium $34 \le n_d \le 45$	7.6	9.0	3.7	8.0	3.6	8.1
Major $46 \le n_d \le 277$	12.5	5.2	4.3	4.7	5.2	3.7

EXAMPLE: Direct HT and GREG estimation for planned domains
Sample: Stratified πPS (stratified WOR type PPS)
Strata: NUTS3, $D = 12$ domains, domain sample sizes n_d fixed

HT estimator:

$$\hat{t}_{dHT} = \sum\nolimits_{k \in s_d} a_k y_k \qquad \hat{V}_A\left(\hat{t}_{dHT}\right) = \frac{1}{n_d(n_d-1)} \sum_{k \in s_d} \left(n_d a_k y_k - \hat{t}_{dHT}\right)^2 \qquad (4)$$

Assisting models in GREG:

$y_k = \beta_{0d} + \beta_{1d}\text{EMP}_k + \varepsilon_k$ (column 2)

$y_k = \beta_{0d} + \beta_{1d}\text{EMP}_k + \beta_{2d}\text{EDUC}_k + \varepsilon_k$ (column 3)

Model fitting by WLS with weights $a_k = 1/\pi_k$. Fitted values are $\hat{y}_k = \mathbf{x}_k' \hat{\boldsymbol{\beta}}_d$ and residuals are $e_k = y_k - \hat{y}_k$.

Direct GREG estimator:

$$\hat{t}_{dGREG} = \sum\nolimits_{k \in U_d} \hat{y}_k + \sum\nolimits_{k \in s_d} a_k e_k = \sum\nolimits_{k \in s_d} a_k g_{dk} y_k$$

$$\hat{V}_2\left(\hat{t}_{dGREG}\right) = \sum_{k \in s_d} \sum_{l \in s_d} (a_k a_l - a_{kl}) g_{dk} e_k g_{dl} e_l \qquad (13)$$

g-weights are $g_{dk} = I_{dk} + I_{dk}\left(\mathbf{t}_{dx} - \hat{\mathbf{t}}_{dx}\right)' \hat{\mathbf{M}}_d^{-1} \mathbf{x}_k$, where $I_{dk} = I\{k \in U_d\}$

Table 4. Mean absolute relative error MARE (%) and mean coefficient of variation MCV (%) of direct HT and indirect GREG estimators for **unplanned** domains.

	HT		**GREG**	
Auxiliary information	1 None		2 Domain sizes and domain totals of EMP	
Domain sample size class	MARE %	MCV %	MARE %	MCV %
Minor $8 \le n_d \le 33$	11.5	28.3	7.6	9.0
Medium $34 \le n_d \le 45$	7.6	20.3	3.8	8.1
Major $46 \le n_d \le 277$	12.5	9.6	4.1	5.0

EXAMPLE: HT and indirect GREG estimation for unplanned domains
Sample: πPS (PPS-WOR)
Domains: NUTS3, $D = 12$ domains, domain sample sizes n_d random

HT estimator:

$$\hat{t}_{dHT} = \sum\nolimits_{k \in s_d} a_k y_k \qquad \hat{V}_U\left(\hat{t}_{dHT}\right) = \frac{1}{n(n-1)} \sum_{k \in s} \left(n a_k y_{dk} - \hat{t}_{dHT}\right)^2 \qquad (5)$$

Assisting model in GREG:

$Y_k = \beta_0 + \beta_1\text{EMP}_k + \varepsilon_k$ (column 2)

Model fitting by WLS with weights $a_k = 1/\pi_k$. Fitted values are $\hat{y}_k = \mathbf{x}_k' \hat{\boldsymbol{\beta}}$ and residuals are $e_k = y_k - \hat{y}_k$.

Indirect GREG estimator:

$$\hat{t}_{dGREG} = \sum\nolimits_{k \in U_d} \hat{y}_k + \sum\nolimits_{k \in s_d} a_k e_k = \sum\nolimits_{k \in s} a_k g_{dk} y_k$$

$$\hat{V}\left(\hat{t}_{dGREG}\right) = \sum_{k \in s} \sum_{l \in s} (a_k a_l - a_{kl}) g_{dk} e_k g_{dl} e_l \qquad (14)$$

g-weights are $g_{dk} = I_{dk} + \left(\mathbf{t}_{dx} - \hat{\mathbf{t}}_{dx}\right)' \hat{\mathbf{M}}^{-1} \mathbf{x}_k$, where $I_{dk} = I\{k \in U_d\}$