Finite Population Estimation Under Generalized Linear Model Assistance

Luz Marina Rondón
Universidad Nacional de Colombia

Luis Hernando Vanegas
Universidad Nacional de Colombia

Cristiano Ferraz
Universidad Federal de Pernambuco

Reporte Interno de Investigación No. 17
Departamento de Estadística
Facultad de Ciencias
Universidad Nacional de Colombia
Bogotá, COLOMBIA
Abstract

Finite population estimation is the overall goal of sample surveys. When information regarding auxiliary variables are available, one may take advantage of general regression estimators (GREG) to improve sample estimate precision. GREG estimators may be derived when the relationship between interest and auxiliary variables is represented by a normal linear model. However, in some scenarios, such as estimating class frequencies or counting process totals, Bernoulli or Poisson responses models are more suitable for describing the relationship between interest and auxiliary variables than normal linear model ones. This paper focuses on the general case for which the relationship between interest variables and the available auxiliary ones may be suitably described by a generalized linear model. The variable of interest’s finite population distribution is viewed in such scenario as if generated by an exponential family distribution, which includes Bernoulli, Poisson, Gamma and inverse Gaussian distributions. The resulting estimator is a generalized linear model regression estimator (GEREG). Its general form and basic statistical properties are presented and studied analytically and, empirically, through of Monte Carlo experiments. Three applications are presented in which the GEREG estimator shows better performance than GREG.

Key words: Auxiliary information, generalized linear models, model-assisted estimation, pseudomaximum likelihood.

1 Introduction

Finite population estimation is sample surveys’ general goal. Whenever interest relies on estimating means, totals and percentages, parameters are estimated using the Horvitz-Thompson estimator.

* Corresponding author.
Email address: lmrondonp@unal.edu.co (Luz Marina Rondon).

Reporte interno de investigacion, Universidad Nacional de Colombia. 4 March 2011
When information regarding auxiliary variables is available, however, one may take advantage of general regression estimators (GREG) to improve sample estimate precision. Such GREG estimators may be derived in a model assisted estimation approach (Särndal, Swensson and Wretman (1992)) when the relationship between interest and auxiliary variables is represented by a normal linear model. The list of authors working on the subject is broad. Isaki and Fuller (1982), Wright (1983), Fuller (2002) and Särndal, Swensson and Wretman (1992). Lehtonen and Veijanen (1998) may perhaps be the first to argue that in some instances (such as when estimating class frequencies) models for multinomial responses are more suitable for describing the relationship between interest and auxiliary variables than normal linear model ones. LGREG estimator properties have been presented and studied, this being a multinomial logistic model assisted estimator. Breidt and Opsomer (2000) focused on a nonparametric regression modeling assisted approach, proposing and studying the properties of local polynomial regression estimators. Estevao and Särndal (2004) studied several applications of domain estimation using calibration. Duchesne (2003) has illustrated the efficiency of inference based on estimator assisted by logistic regression regarding normal regression by using Monte Carlo experiments. Lehtonen, Särndal and Veijanen (2003) compared the performance of different estimators that using auxiliary information in the estimation stage, including those assisted by the logistic regression model. Lehtonen, Särndal and Veijanen (2005) studied the importance of model specification for estimating the total of a polytomous interest variable for a number of large or small domains. Li (2008) introduced the Box-Cox technique into the generalized regression estimator, which can be especially suitable to deal with highly skewed continuous and positive interest variables, where normal response models may not be appropriate. This paper focuses on the general case for which the relationship between interest variables and the available auxiliary ones may be suitably described by a generalized linear model. Finite population distribution of the interest variable is viewed in this scenario as if generated by an exponential family distribution, which includes Bernoulli, Poisson, Gamma and inverse Gaussian distributions; the resulting estimator is a generalized linear model regression estimator (GEREG). Its general form and basic statistical properties are presented and studied. $U = \{1, 2, \ldots, N\}$ was thus chosen as being a finite population. A sample $S$, size $n$, was obtained from $U$ using $p(\cdot)$ sampling design. $\pi_k = P(k \in S)$ was denoted as being the first-order inclusion probability for the $k$-th population element. Likewise, $\pi_{kl} = P(k, l \in S)$ represented the second-order inclusion probability for the elements $k$ and $l$. $y_k$ was denoted as being the value of the interest variable which could be discrete or continuous, while $x_k = (x_{k1}, \ldots, x_{kJ})^T$ was the auxiliary information vector for the $k$-th element. The paper is organized as follows. In Section 2 we introduced generalized linear model in finite population. In Section 3 we presented the GEREG estimator and its main properties were studied. Besides, in that section was presented a simulation study that illustrates the performance of the GEREG estimator. Section 4 we presented
three applications of the proposed estimator.

2 Generalized Linear Model in Finite Population

The literature concerning generalized linear models is vast; for instance, McCullagh and Nelder (1989), Dobson (2001) and Fox (2008). \(Y_1, \ldots, Y_n\) are considered the values of an interest variable \(Y\) measured in \(n\) elements in a GLM setup. \(Y\)'s are supposed to be independent random variables with probability distribution from the exponential family. Let \(Y_k\) be random variable \(Y\) for element \(k\). Its density function may be expressed as

\[
f(y; \theta_k, \phi_k) = \exp\{\phi_k[y \theta_k - b(\theta_k)] + c(y, \phi_k)\},
\]

where \(c(\cdot)\) is a known function, \(\theta_k\) is the canonical parameter, \(E(Y_k) = \mu_k = b'(\theta_k)\), \(\text{Var}(Y_k) = \phi_k^{-1} V(\mu_k)\), with \(V_k = V(\mu_k) = \partial \mu_k / \partial \theta_k\) the variance function and \(\phi_k^{-1} > 0\) the dispersion parameter. The generalized linear models are defined by (1) and by the following systematic component

\[
g(\mu_k) = \eta_k = \sum_{j=1}^J \beta_j x_{kj} = x_k^T \beta,
\]

where \(x_k = (x_{k1}, \ldots, x_{kJ})^T\) is a vector of \(J\) explanatory variables for element \(k\), \(\beta = (\beta_1, \ldots, \beta_J)^T\) a vector having unknown parameters, and \(g(\cdot)\) a monotone differentiable function, called the link function. When \(g(\cdot)\) is defined in such a way that \(\theta_k = \eta_k\) for every \(k\), then \(g(\cdot)\) is called the canonical link function. It has been assumed in this work that if \(k\) and \(l\) such that \(\phi_k \neq \phi_l\), then \(\phi_k \propto \delta_k, k = 1, \ldots, n\), with \(\delta_k\) being a known quantity for every \(k\). Particular cases of such setup include normal, Bernoulli, Poisson, Gamma and the inverse normal response models. Table 1 shows \(b(\theta), \theta, \phi\) and \(V(\mu)\) values for main exponential family distributions.

<table>
<thead>
<tr>
<th>Distribution</th>
<th>(b(\theta))</th>
<th>(\theta)</th>
<th>(\phi)</th>
<th>(V(\mu))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>(\theta^2/2)</td>
<td>(\mu)</td>
<td>(1/\sigma^2)</td>
<td>1</td>
</tr>
<tr>
<td>Poisson</td>
<td>(e^\theta)</td>
<td>(\log \mu)</td>
<td>1</td>
<td>(\mu)</td>
</tr>
<tr>
<td>Bernoulli</td>
<td>(\log(1 + e^\theta))</td>
<td>(\log{\mu/(1 - \mu)})</td>
<td>1</td>
<td>(\mu(1 - \mu))</td>
</tr>
<tr>
<td>Gamma</td>
<td>(-\log(-\theta))</td>
<td>(-1/\mu)</td>
<td>(1/(CV)^2)</td>
<td>(\mu^2)</td>
</tr>
<tr>
<td>N. Inversa</td>
<td>(-\sqrt{-2\theta})</td>
<td>(-1/2\mu^2)</td>
<td>(\phi)</td>
<td>(\mu^3)</td>
</tr>
</tbody>
</table>

The maximum-likelihood method may be used to estimate GLM parameters. The estimation method can use sampling weights in the finite population context. For instance, \(\hat{\mu}_k^U\) and \(\hat{\mu}_k^S\) are estimators of \(\mu_k\); however, the first one
uses information about \( U \), while the second one just uses information about \( S \). Similarly, there are differences between \( \beta, \hat{\beta}_U, \hat{\beta}_S \) and \( \hat{\beta}_\pi^S \), since \( \beta \) is the super-population parameter vector while \( \hat{\beta}_U \) is the \( \beta \) estimator based on \( U \). There are two options for estimating \( \beta \) with information from sample \( S \): the first, from which \( \hat{\beta}_S = \arg\max_\beta L_S(\beta) \) is obtained by assigning equal weighting to individuals; the second one, from which the pseudomaximum likelihood estimator \( \hat{\beta}_\pi^S \) is obtained, takes into account sample design assigning weights for individuals according to first-order inclusion probabilities. (see Table 2).

**Table 2**

<table>
<thead>
<tr>
<th>Estimation of ( \beta ) and ( \mu_k )</th>
</tr>
</thead>
<tbody>
<tr>
<td>With information about ( U )</td>
</tr>
<tr>
<td>------------------------------------------</td>
</tr>
<tr>
<td>( \hat{\beta}<em>U = \arg\max</em>\beta L_U(\beta) )</td>
</tr>
<tr>
<td>( \hat{\beta}<em>\pi^S = \arg\max</em>\beta L_\pi^S(\beta) )</td>
</tr>
<tr>
<td>( \hat{\mu}_k^U = g^{-1}(x_k^T\hat{\beta}_U) )</td>
</tr>
<tr>
<td>( \hat{\mu}<em>k^\pi^S = g^{-1}(x_k^T\hat{\beta}</em>\pi^S) )</td>
</tr>
</tbody>
</table>

The loglikelihood function for sample \( S(S \subset U) \) considering sample weighting, often called \( \pi \)-weighted loglikelihood or pseudo-likelihood, can be written in the following way

\[
L_\pi^S(\beta) = \sum_{k \in S} \frac{\phi_k}{\pi_k} [y_k - b(\beta; x_k)],
\]

then, \( \hat{\beta}_\pi^S = \arg\max_\beta L_\pi^S(\beta) \) and \( \hat{\mu}_k^\pi_S = g^{-1}(x_k^T\hat{\beta}_\pi^S) \) are the pseudomaximum likelihood estimators (see, for instance, Skinner and Holt (1989); Lehtonen and Pahkinen (2004)) of \( \beta \) and \( \mu_k \), respectively. If \( \pi_k = \pi \) for \( k \in U \), then \( \hat{\beta}_S \) and \( \hat{\beta}_\pi^S \) are equivalents. The estimator \( \hat{\beta}_\pi^S \) can also be found by solving equation \( U_\pi^S(\hat{\beta}_\pi^S) = 0 \). The function \( U_\pi^S(\beta) = \partial L_\pi^S(\beta)/\partial \beta = (U_\pi^S(\beta), \ldots, U_\pi^S(\beta))^T \) is the score function for \( \beta \) (see, for example, Cox and Hinkley (1974)), where

\[
U_\pi^S(\beta) = \sum_{k \in S} \frac{\phi_k}{\pi_k} (y_k - \mu_k) \frac{\delta \theta_k}{\delta \eta_k} x_{kj}, \quad j = 1, \ldots, J.
\]

The pseudomaximum likelihood estimator is equivalent to the weighted least squares estimator for normal response models having canonical link (identity), which can be expressed as

\[
\hat{\beta}_S^\pi = (X_S^T W_S X_S)^{-1} X_S^T W_S Y_S,
\]

where \( X_S = (x_1, \ldots, x_n)^T \), \( Y_S = (y_1, \ldots, y_n)^T \) and \( W_S = \text{diag}\{\phi_1/\pi_1, \ldots, \phi_n/\pi_n\} \).

The loglikelihood function for the population can be written in the following
way
\[ L_U(\beta) = \sum_{k \in U} \phi_k [y_k \theta(\beta; \mathbf{x}_k) - b(\theta(\beta; \mathbf{x}_k))], \]

then, \( \hat{\beta}_U = \arg \max_{\beta} L_U(\beta) \) and \( \hat{\mu}_k = g^{-1}(x_k^T \hat{\beta}_U) \) are the maximum likelihood estimators of \( \beta \) and \( \mu_k \), respectively. The score function for \( \beta \) in the population \( U \) is given by
\[ U_{Uj}(\beta) = \sum_{k \in U} \phi_k (y_k - \mu_k) \frac{\delta \theta_k}{\delta \eta_k} x_{kj}, \quad j = 1, \ldots, J. \] (6)

3 Generalized linear model regression estimator (GEREG)

Considering estimating a total population \( t_y = \sum_{k \in U} y_k \). The GEREG for \( t_y \) may be written as
\[ \hat{t}_G = \sum_{k \in U} \hat{\mu}_k^S + \sum_{k \in S} \frac{(y_k - \hat{\mu}_k^S)}{\pi_k}, \] (7)
where \( \xi \), the model explaining the relationship between the response and the auxiliary variables, may be expressed as
\[
\begin{align*}
E_\xi(y_k) &= \mu_k; \\
\text{Var}_\xi(y_k) &= \phi_k^{-1} V(\mu_k) \\
g(\mu_k) &= \sum_{j=1}^J \beta_j x_{kj} = \mathbf{x}_k^T \beta
\end{align*}
\] (8)

where \( \beta \) is a vector having unknown parameters, \( g(\cdot) \) is the link function, \( \mathbf{x}_k = (x_{k1}, \ldots, x_{kJ})^T \) is the auxiliary information vector for the \( k \)-th population element and \( E_\xi(\cdot) \) and \( \text{Var}_\xi(\cdot) \) are the expectation and variance in model \( \xi \), respectively. In expression (7) \( \hat{\mu}_k^S \) is given by \( g^{-1}(x_k^T \hat{\beta}_S^T) \), with \( \hat{\beta}_S^T \), the \( \beta \) estimator based in \( S \), takes sampling weighting into account. We assumed that, if \( k \) and \( l \) so that \( \phi_k \neq \phi_l \), then \( \phi_k \propto \delta_k \), with \( \delta_k \) being a known quantity for every \( k \in U \). If \( \xi \) considers homogeneity in the dispersion parameter (\( \phi_k = \phi \)), then we assumed that \( \phi_k = 1 \) for every \( k \in U \). The \( \hat{t}_G \) estimator requires knowledge about auxiliary information vector \( \mathbf{x}_k \) for every \( k \in U \), however, if the link function in the \( \xi \) model is identity, \( \hat{t}_G \) only requires knowledge about \( \mathbf{x}_k \) for every \( k \in S \) and total population vector \( t_x = (tx_1, \ldots, tx_J)^T \), where \( tx_j = \sum_{k \in U} x_{kj}, \ j = 1, \ldots, J. \)
3.1 Alternative expression for GEREG

The second term in (7) is an adjustment. Theorem 1 provides conditions in which the term adjustment can be eliminated, thus simplifying the expression for the GEREG estimator. This theorem also allows writing $t_y$ using the fitted values of the model based on $U$.

**Theorem 1.** If the GEREG estimator $\hat{t}_G$ is based on a $\xi$ model where it has been verified that

\[ \sum_{j=1}^{J} a_j h_j = 1 \]

for $h_j = (h_{1j}, \ldots, h_{Nj})^T$ with $h_{kj} = \phi_k \frac{\partial \theta_k}{\partial \eta_k} x_{kj}$ and $a = (a_1, \ldots, a_j)^T$ a constant vector, then

\[ (1) \sum_{k \in U} (y_k - \hat{\mu}_U^k) = 0, \]

i.e, the total of $y$ can be written as $t_y = \sum_{k \in U} \hat{\mu}_U^k$;

\[ (2) \sum_{k \in S} (y_k - \hat{\mu}_S^k)/\pi_k = 0, \]

i.e, the estimator $\hat{t}_G$ can be written as $\hat{t}_G = \sum_{k \in U} \hat{\mu}_S^k$.

According to theorem 1, the smaller the difference between $\hat{\beta}_S$ and $\hat{\beta}_U$ the smaller also the difference between $\hat{t}_G$ and $t_y$. In the terms of finite population consistency, the estimator $\hat{\beta}_S$ is consistent for $\hat{\beta}_U$, therefore, $\hat{t}_G$ is consistent for $t_y$. Moreover, if $\pi_k = \pi$ for $k \in U$, from theorem 1 we have $\sum_{k \in S} y_k = \sum_{k \in S} \hat{\mu}_S^k$.

**Corollary 1.** If GEREG estimator $\hat{t}_G$ is based on an $\xi$ model where it has been verified that there is

(1) Homogeneity in the dispersion parameter, i.e, $\phi_k = \phi$ for $k \in U$;

(2) A systematic component having an intercept, i.e, there is $\beta_j$ in $\beta$ so that $\partial \eta_k/\partial \beta_j = C$ for $k \in U$, with $C$ a constant;

(3) A systematic component having a canonical link, i.e, $\theta_k = \eta_k$ for $k \in U$;

then, the results (1) and (2) of theorem 1 are satisfied.

The corollary 1 provides, from the model formulation point of view, clearer conditions than the 1 theorem so that the term adjustment term in the GEREG estimator can be eliminated.

**Corollary 2.** If the GEREG estimator $\hat{t}_G$ is based on an $\xi$ model with normal response, canonical link (identity) and $a = (a_1, \ldots, a_j)^T$ is a constant vector so that, for $k \in U$ it has been verified that

\[ \phi_k^{-1} = \sum_{j=1}^{J} a_j x_{kj}, \]

then, the results (1) and (2) of 1 theorem are satisfied.

The result in the corollary 2 is analogous to that found in Särndal, Swensson and Wretman (1992). Then, the 1 theorem is an extension of the result found in Särndal, Swensson and Wretman (1992) for generalized linear models.
3.2 Expectation and variance

The following theorem provides approximate expressions for the expectation and variance of the GEREG estimator, taking into account the sampling design and the regression model formulated.

**Theorem 2.** If \( \hat{t}_G \) is given by \( \sum_{k \in U} \hat{\mu}_k^S \) and \( \xi \) considers the canonical link we have

\[
\begin{align*}
(1) \quad & E(\hat{t}_G) \approx t_y; \\
(2) \quad & \text{Var}(\hat{t}_G) \approx \Gamma_U^T \Sigma_U \Gamma_U; \\
(3) \quad & \hat{\text{Var}}(\hat{t}_G) = \Gamma_S^T \Sigma_S \Gamma_S,
\end{align*}
\]

where \( \Gamma_U = (X_U^T W_U^* X_U)^{-1} X_U^T W_U^* \phi^* \), \( W_U^* = \text{diag}\{\phi_1 V^U_1, \ldots, \phi_N V^U_N\} \), \( \Gamma_S = (X_U^T W_S^* X_U)^{-1} X_S^T W_S^* \phi^* \), \( W_S^* = \text{diag}\{\phi_1 V^S_1, \ldots, \phi_N V^S_N\} \), \( \Sigma_U = \{\sigma_{ij}^U\} \) and \( \Sigma_S = \{\sigma_{ij}^S\} \), with \( \sigma_{ij}^U = \sum_{k \in U} \sum_{l \in U} \Delta_{kl} \frac{x_{ki} E_k}{\pi_k} \frac{x_{lj} E_l}{\pi_l} \), \( \sigma_{ij}^S = \sum_{k \in S} \sum_{l \in S} \Delta_{kl} \frac{x_{ki} E_k}{\pi_k} \frac{x_{lj} E_l}{\pi_l} \), \( \phi^* = (\phi_1^{-1}, \ldots, \phi_N^{-1})^T \), \( V^U = V(\hat{\mu}_U^*) \), \( V^S = V(\hat{\mu}_S^*) \), \( E_k = y_k - \hat{\mu}_U^* \) and \( e_k = y_k - \hat{\mu}_S^* \).

**Corollary 3.** Under the conditions of the Corollary 1, we have that

\[
\begin{align*}
(1) \quad & E(\hat{t}_G) \approx t_y; \\
(2) \quad & \text{Var}(\hat{t}_G) \approx \sum_{k \in U} \sum_{l \in U} \Delta_{kl} \frac{E_k E_l}{\pi_k \pi_l}; \\
(3) \quad & \hat{\text{Var}}(\hat{t}_G) = \sum_{k \in S} \sum_{l \in S} \Delta_{kl} \frac{E_k E_l}{\pi_k \pi_l}.
\end{align*}
\]

**Corollary 4.** If the GEREG estimator \( \hat{t}_G \) is based on an \( \xi \) model with normal response, canonical link (identity) and \( a = (a_1, \ldots, a_j)^T \) is a constant vector so that, for \( k \in U \) it has been verified that

\[
\phi_{k}^{-1} = \sum_{j=1}^{J} a_j x_{kj},
\]

then, the results (1) and (2) of 3 corollary are satisfied.

Corollary 3 and 4 shows that when the \( \xi \) model considers canonical link, intercept and homogeneity in the dispersion parameter, the expressions for \( \text{Var}(\hat{t}_G) \) and \( \hat{\text{Var}}(\hat{t}_G) \) become simplified, i.e, stated that the estimator \( \hat{t}_G \) is approximately unbiased for \( t_y \) and, an approximation of the variance of \( \hat{t}_G \) can be calculate applying the formula for the variance of Horvitz-Thompson estimator to residuals \( E_k \).
3.3 The models role

The role of model $\xi$ in the estimating $t_y$ can be observed by assessing the impact on $E_k$ of small perturbations in $y_k$. Pregibon (1981) proposed an amount measuring the influence of $y_k$ on $\hat{\mu}^{U_k}$ in generalized linear models, this being an extension of the measurement $\partial \hat{\mu}^{U_k} / \partial y_k$ applied to normal linear models. Then, when $\xi$ considers canonical link and homogeneity in dispersion parameter, then quantity $\gamma_k = 1 - \hat{h}_{kk}$ can be used to assessing the influence of $y_k$ on $\hat{\mu}^{U_k}$ in generalized linear models, this being an extension of the measurement $\partial \hat{\mu}^{U_k} / \partial y_k$ applied to normal linear models. Small $\gamma_k$ values indicate that an change in $y_k$ leads to $\hat{\mu}^{U_k}$ similarly changing, suggesting that the $k$-th population element has a large weighting on estimating $\hat{\beta}^U$ and on $E_k$ magnitude. Likewise, large $\gamma_k$ values indicate that an change in $y_k$ does not similarly lead to an change in $\hat{\mu}^{U_k}$, suggesting that the $k$-th population element has small weighting in estimating $\hat{\beta}^U$ and on $E_k$ magnitude. For example, for a model having an intercept, one explanatory variable an d homogeneity in the dispersion parameter, we have the following:

i) Canonical link

$$\gamma_k = 1 - \frac{\hat{V}_k^U}{\sum_{i \in U} V_i^U} - \frac{(x_k - \bar{x}_U)^2 \hat{V}_k^U}{\sum_{i \in U} (x_i - \bar{x}_U)^2 V_i^U}$$

ii) Identity link

$$\gamma_k = 1 - \frac{(1/\hat{V}_k^U)}{\sum_{i \in U} (1/\hat{V}_i^U)} - \frac{(x_k - \bar{x}_U)^2 (1/\hat{V}_k^U)}{\sum_{i \in U} (x_i - \bar{x}_U)^2 (1/\hat{V}_i^U)}$$

where $\bar{x}_U$ is the population mean of explanatory variable $x$. In both cases considered above, $\gamma_k \in (0, 1)$ and, for a fixed value of $x_k$, the $\gamma_k$ value depends on the variance function, in direct proportion when the link is canonical and, inversely proportional when link is identity. Thus, the $\xi$ model determines the structure of weighting or importance for elements in $U$, which affects the values of $E_k$ on Var($\hat{t}_G$).

3.4 Relative efficiency

The relative efficiency of $\hat{t}_G$ compared to the Horvitz-Thompson estimator using $p(\cdot)$ sampling design can be evaluated with $\Delta_{p(\cdot)} = \text{Var}(\hat{t}_G)_{p(\cdot)}/\text{Var}(\hat{t}_{HT})_{p(\cdot)}$ measurement. For example, for sampling designs such as simple Random Sampling (SRS) and Bernoulli sampling (BER) the $\Delta_{p(\cdot)}$ measurement does not depend on sample size. The following result led us to observe the impact of the model’s goodness of fit on $\hat{t}_G$ relative efficiency.
Theorem 3. Under the conditions of the corollary 1, if the $\hat{t}_g$ is used under

i) Simple Random Sampling (SRS) we have

$$\Delta_{SRS} = \frac{\text{Var}(\hat{t}_g)_{SRS}}{\text{Var}(\hat{t}_{HT})_{SRS}} \approx 1 - R_U^2$$

ii) Bernoulli Sampling (BER) we have

$$\Delta_{BER} = \frac{\text{Var}(\hat{t}_g)_{BER}}{\text{Var}(\hat{t}_{HT})_{BER}} \approx \frac{1 - R_U^2}{1 + \frac{CV_y^2}{1-1/N}}$$

where $R_U$ is Pearson’s linear correlation coefficient calculated between $y_k$’s and $\hat{\mu}_U^k$’s and CV is the variation coefficient of $y$ in the population $U$.

This result shows that when the model’s goodness of fit increases then the $\hat{t}_g$ efficiency compared to the Horvitz-Thompson estimator also increases.

It is also possible to observe that, using comparable sample sizes, we have $\text{Var}(\hat{t}_g)_{BER}/\text{Var}(\hat{t}_g)_{SRS} \approx 1 - 1/N$. Besides, 3 theorem means that it is possible to show that under SRS sampling, the Horvitz-Thompson estimator has efficiency equal to $\hat{t}_g$ when using an $f^*$ sampling fraction, given by

$$f^* \approx \frac{f}{f + (1 - f)(1 - R_U^2)},$$

where $f = n/N$ is the sampling fraction used by $\hat{t}_g$. For instance, when the GEREG estimator uses a 0.2 sampling fraction under SRS sampling, the $\xi$ model satisfies corollary 1 conditions and has a goodness of fit described by $R_U^2 = 0.95$, the Horvitz-Thompson estimator requires a sampling fraction around four times greater than the GEREG estimator to obtaining the same efficiency.

3.5 Particular cases

This section considers estimators assisted by models satisfying the conditions of theorem 1 for different response distributions.

3.5.1 Normal response with constant variance

The estimator (9) is obtained through the least squares method in the statistical literature, having no distributional assumptions. The same estimator was obtained in this paper using a regression model having normal response
and canonical link. According to corollary 1, this estimator is given by

\[ \hat{\mu}_g = \sum_{k \in U} \hat{\mu}_k^S = \sum_{k \in U} x_k^T \hat{\beta}_S^\pi = N \hat{\beta}_S^\pi_1 + \sum_{j=2}^J \hat{\beta}_{SJ}^\pi t_{xj}, \]  

(9)

with \( \hat{\beta}_S^\pi = (\hat{\beta}_S^\pi_1, \ldots, \hat{\beta}_S^\pi_J)^T \). The model \( \xi \) that assists (9) can be written as

\[
\begin{align*}
E_{\xi}(Y_k) &= \mu_k; \\
\text{Var}_{\xi}(Y_k) &= \sigma^2 \\
\mu_k &= \beta_1 + \sum_{j=2}^J \beta_j x_{kj}
\end{align*}
\]

From (5) \( \hat{\beta}_S^\pi \) can be written as \( \hat{\beta}_S^\pi = (X_S^T W_S X_S)^{-1} X_S^T W_S Y_S \) with \( X_S = (x_1, \ldots, x_n)^T \), \( x_k = (1, x_{k2}, \ldots, x_{kJ})^T \) and \( W_S = \text{diag}\{1/\pi_1, \ldots, 1/\pi_n\} \). According to corollary 3 under SRS sampling, variance (9) approximation and its corresponding estimator can be written as

\[ \text{Var}(\hat{t}_g)_{\text{SRS}} \approx N(1 - R_S^2) \frac{1}{f} S_{yU}^2 \quad \text{and} \quad \hat{\text{Var}}(\hat{t}_g)_{\text{SRS}} = N(1 - R_S^2) \frac{1}{f} S_{yS}^2 \]

where \( S_{yU}^2 = \frac{1}{N-1} \sum_U (y_k - \bar{y}_U)^2 \), \( S_{yS}^2 = \frac{1}{n-1} \sum_S (y_k - \bar{y}_S)^2 \) and \( R_S \) is the Pearson’s linear correlation coefficient calculated between \( y_k \)'s and \( \hat{\mu}_k^S \)'s. Similarly, from corollary 3 with BER sampling, variance (9) approximation and its corresponding estimator can be written as

\[ \text{Var}(\hat{t}_g)_{\text{BER}} \approx \frac{N(1 - R_S^2)}{(1 - 1/N)^{-1}} \frac{1}{\pi} S_{yU}^2 \quad \text{and} \quad \hat{\text{Var}}(\hat{t}_g)_{\text{BER}} = \frac{n \cdot N(1 - R_S^2)}{N \pi (1 - 1/n)^{-1}} \frac{1}{\pi} S_{yS}^2 \]

where \( N \pi \) is expected sample size.

### 3.5.2 Normal response with variance heterogeneity

The following estimator was assisted by a model having normal response, variance heterogeneity and which satisfies the conditions of corollary 2

\[ \hat{t}_g = \sum_{k \in U} \hat{\mu}_k^S = \sum_{k \in U} \hat{\beta}_S^\pi x_k = \hat{\beta}_S^\pi x_t = \frac{t_x}{t_y} \hat{t}_y, \]  

(10)

with \( \hat{t}_x \) and \( \hat{t}_y \) being the Horvitz-Thompson estimators for \( t_x \) and \( t_y \), respectively. The estimator (10) is called Ratio estimator (see for instance, Särndal, Swensson and Wretman (1992)). The \( \xi \) model that assisted (10) can be written as

\[
\begin{align*}
E_{\xi}(Y_k) &= \mu_k; \\
\text{Var}_{\xi}(Y_k) &= \sigma^2 x_k, \quad x_k > 0 \\
\mu_k &= \beta x_k
\end{align*}
\]
From (5) \( \hat{\beta}_S^\pi \) can be written as \( \hat{\beta}_S^\pi = (X_S^T W_S X_S)^{-1} X_S^T W_S Y_S \) with \( W_S = \text{diag}\{1/x_1 \pi_1, \ldots, 1/x_n \pi_n\} \). According to theorem 2 under SRS sampling, approximation of variance (10) and its corresponding estimator are given by

\[
\text{Var}(\hat{t}_G)_{\text{SRS}} = N \frac{1 - f}{f} \left[ S^2_{yU} + \frac{S_{xU} S_{yU} \bar{y}_U^2}{x_U^2} \left( \frac{S_{xU}}{S_{yU}} - 2R_U \right) \right]
\]

and

\[
\hat{\text{Var}}(\hat{t}_G)_{\text{SRS}} = N \frac{1 - f}{f} \left[ S^2_{yS} + \frac{S_{xS} S_{yS} \bar{y}_S^2}{x_S^2} \left( \frac{S_{xS}}{S_{yS}} - 2R_S \right) \right]
\]

Similarly, from theorem 2 under BER sampling, (10) variance approximation and its corresponding estimator can be written as

\[
\text{Var}(\hat{t}_G)_{\text{BER}} = \frac{N}{(1 - 1/N)^{1/\pi}} \left[ S^2_{yU} + \frac{S_{xU} S_{yU} \bar{y}_U^2}{x_U^2} \left( \frac{S_{xU}}{S_{yU}} - 2R_U \right) \right]
\]

and

\[
\hat{\text{Var}}(\hat{t}_G)_{\text{BER}} = \frac{n}{N \pi (1 - 1/n)^{1/\pi}} \left[ S^2_{yS} + \frac{S_{xS} S_{yS} \bar{y}_S^2}{x_S^2} \left( \frac{S_{xS}}{S_{yS}} - 2R_S \right) \right]
\]

One can easily show that, using SRS and BER sampling designs, \( \hat{t}_G \) is more efficient than the Horvitz-Thompson estimator when \( 2R_U > S_{xU}/S_{yU} \).

### 3.5.3 Bernoulli response

Suppose \( y_k \), the response for element \( k \) in a particular population is a dichotomous variable, assuming value one if the element presents a characteristic of interest, and zero, otherwise. The \( t_y \) parameter of interest can be written as \( NP \), where \( N \) is the population size and \( P \) is the proportion of population elements having the characteristic of interest. The estimator for \( NP \), assisted by a Bernoulli regression model \( \xi \) satisfying the conditions of corollary 1 can be written as

\[
\hat{t}_G = \sum_{k \in U} \hat{\mu}_k = \sum_{k \in U} \left[ 1 + \exp \left( -\hat{\beta}_1 - \sum_{j=2}^J \hat{\beta}_j x_{kj} \right) \right]^{-1}
\]

where \( \xi \), the model assisting in the estimation, can be described as

\[
\begin{align*}
E_\xi(Y_k) = \mu_k &= P_\xi[y_k = 1|x_k] \\
\text{Var}_\xi(Y_k) = \mu_k(1 - \mu_k) \\
\log \left[ \frac{\mu_k}{1 - \mu_k} \right] &= \beta_1 + \sum_{j=2}^J \beta_j x_{kj}
\end{align*}
\]

As discussed in section 3.3, for a given \( x_k \), the closest \( \hat{\mu}_k \) is 0.5, the greater the weighting or importance of element \( k \) in the estimation. According to
corollary 3 under SRS sampling, the approximated variance of (11) and its corresponding estimator can be expressed as

$$\text{Var}(\hat{t}_G)_{SRS} \approx \frac{N(1 - R^2_U)}{1 - 1/N} \frac{1 - f}{f} P(1 - P)$$

and

$$\hat{\text{Var}}(\hat{t}_G)_{SRS} = \frac{N(1 - R^2_S)}{1 - 1/n} \frac{1 - \hat{P}}{1 - \hat{P}},$$

where $\hat{P} = \frac{1}{n} \sum_S y_k$ is the sample proportion of elements having the characteristic of interest. On the other hand, in BER sampling, the approximated variance of (11) and its corresponding estimator can be expressed as

$$\text{Var}(\hat{t}_G)_{BER} \approx N(1 - R^2_U) \frac{1 - \pi}{\pi} P(1 - P)$$

and

$$\hat{\text{Var}}(\hat{t}_G)_{BER} = \frac{n}{N \pi} N \pi N(1 - R^2_S) \frac{1 - \pi}{\pi} \hat{P}(1 - \hat{P}).$$

### 3.5.4 Poisson response

When the variable of interest represent a counting process, estimating population totals can be assisted by a Poisson regression model. Such an estimator, which satisfies the conditions of corollary 1, can be expressed by

$$\hat{t}_G = \sum_{k \in U} \hat{\mu}_k = \sum_{k \in U} \exp \left( \hat{\beta}_S^1 + \sum_{j=2}^{J} \hat{\beta}_S^j x_{kj} \right) = \hat{\kappa}^\pi \prod_{j=2}^{J} \hat{\kappa}_j^{\pi x_{kj}}, \quad (12)$$

with $\hat{\kappa}_j^\pi = \exp(\hat{\beta}_S^j)$. Where $\xi$, the model assisting the estimation, is given by

$$\begin{cases}
E_{\xi}(Y_k) = \mu_k \\
\text{Var}_{\xi}(Y_k) = \mu_k \\
\log(\mu_k) = \beta_1 + \sum_{j=2}^{J} \beta_j x_{kj}
\end{cases}$$

As discussed in section 3.3, for a given $x_k$, the higher $\hat{\mu}_k$, the higher the relevance (weighting) of element $k$. According to corollary 3 in SRS sampling, the approximated variance of (12) and its corresponding estimator can be expressed as

$$\text{Var}(\hat{t}_G)_{SRS} \approx N(1 - R^2_U) \frac{1 - f}{f} \frac{1}{f} S^2_{yU} \quad \text{and} \quad \hat{\text{Var}}(\hat{t}_G)_{SRS} = N(1 - R^2_S) \frac{1 - f}{f} S^2_{ys}. $$
On the other hand, in BER sampling, the approximated variance of (12) and its corresponding estimator can be expressed as

$$\text{Var}(\hat{t}_G)_{BER} \approx \frac{N(1 - R^2_U)}{(1 - 1/N)^{-1}} \frac{1 - \pi}{\pi} S^2_{gy}$$

and

$$\hat{\text{Var}}(\hat{t}_G)_{BER} = \frac{n}{N\pi} \frac{N(1 - R^2_S)}{(1 - 1/n)^{-1}} \frac{1 - \pi}{\pi} S^2_{gs}$$

### 3.5.5 Gamma response

When the variable of interest is skewed continuous and positive (i.e. income) the estimation can be assisted by the Gamma model. The estimator of $t_y$ assisted by a Gamma response regression model and satisfying the conditions of corollary 1 can be expressed in the following way

$$\hat{t}_G = \sum_{k \in U} \hat{\mu}_k^S = \sum_{k \in U} \left( \hat{\beta}_{S1}^\pi + \sum_{j=2}^J \hat{\beta}_{Sj}^\pi x_{kj} \right)^{-1},$$

(13)

where $\xi$, the model that assisting the estimation, can be written as

$$\begin{cases}
E_\xi(Y_k) = \mu_k \\
\text{Var}_\xi(Y_k) = \mu_k^2 \\
\mu_k^{-1} = \beta_1 + \sum_{j=2}^J \beta_j x_{kj}
\end{cases}$$

As discussed in section 3.3, for a given $x_k$, the higher $\hat{\mu}_k$, the higher the relevance (weighting) of element $k$. According to corollary 3 in SRS sampling, the approximated variance of (13) and its corresponding estimator is given by

$$\text{Var}(\hat{t}_G)_{SRS} \approx N(1 - R^2_U) \frac{1 - f}{f} S^2_{gy}$$

and

$$\hat{\text{Var}}(\hat{t}_G)_{SRS} = N(1 - R^2_S) \frac{1 - f}{f} S^2_{gs}$$

On the other hand, in BER sampling, the approximated variance of (13) and its corresponding estimator can be expressed as

$$\text{Var}(\hat{t}_G)_{BER} \approx \frac{N(1 - R^2_U)}{(1 - 1/N)^{-1}} \frac{1 - \pi}{\pi} S^2_{gy}$$

and

$$\hat{\text{Var}}(\hat{t}_G)_{BER} = \frac{n}{N\pi} \frac{N(1 - R^2_S)}{(1 - 1/n)^{-1}} \frac{1 - \pi}{\pi} S^2_{gs}$$

### 3.5.6 Stratified sampling

i) Estimating population totals

For a stratified SRS design in which the model assisting estimation in each stratum satisfies the conditions of corollary 1, the estimator for $t_y$ is

$$\hat{t}_G = \sum_{h=1}^H \hat{t}_{Gh} = \sum_{h=1}^H \sum_{k \in U_h} \hat{\mu}_k^S,$$
where \( \hat{t}_{Gh} \) is the estimator of the total of \( y \) in the stratum \( h \). According to corollary 3, the variance estimator expression is given by

\[
\text{Var}(\hat{t}_G)_{SRS} = \sum_{h=1}^{H} \text{Var}(\hat{t}_{Gh})_{SRS} \approx \sum_{h=1}^{H} N_h (1 - R_{U_h}^2) \frac{1 - f_h}{f_h} S_{yU_h}^2,
\]

with \( N_h \) being the size of the stratum \( h \), \( f_h \) the sampling fraction in the stratum \( h \), \( S_{yU_h}^2 \) the variance of \( y \) in the stratum \( h \) and \( R_{U_h} \) is Pearson’s linear correlation coefficient calculated between \( y_k \)’s and \( \hat{\mu}_U^k \)’s in the stratum \( h \). The estimator of the variance expression is given by:

\[
\hat{\text{Var}}(\hat{t}_G)_{SRS} = \sum_{h=1}^{H} \hat{\text{Var}}(\hat{t}_{Gh})_{SRS} = \sum_{h=1}^{H} N_h (1 - R_{S_h}^2) \frac{1 - \pi_h}{\pi_h} S_{yS_h}^2,
\]

where \( S_{yS_h}^2 \) is the variance of \( y \) the sample of stratum \( h \) and \( R_{S_h} \) is Pearson’s linear correlation coefficient calculated between \( y_k \)’s and \( \hat{\mu}_S^k \)’s in the sample from stratum \( h \). With BER sampling, the total population estimator expression is the same as for the SRS case. The variance estimator expression can be written as

\[
\text{Var}(\hat{t}_G)_{BER} = \sum_{h=1}^{H} \text{Var}(\hat{t}_{Gh})_{BER} \approx \sum_{h=1}^{H} (N_h - 1)(1 - R_{U_h}^2) \frac{1 - \pi_h}{\pi_h} S_{yU_h}^2,
\]

with \( N_h \pi_h \) being expected sample size in stratum \( h \). The variance estimator expression is given by

\[
\hat{\text{Var}}(\hat{t}_G)_{BER} = \sum_{h=1}^{H} \hat{\text{Var}}(\hat{t}_{Gh})_{BER} = \sum_{h=1}^{H} \frac{(n_h - 1)}{\pi_h} (1 - R_{S_h}^2) \frac{1 - \pi_h}{\pi_h} S_{yS_h}^2,
\]

where \( n_h \) is the sample size in stratum \( h \).

\[\text{ii) Estimating population means}\]

For a stratified SRS design in which the model assisting estimation in each stratum satisfies the conditions of corollary 1, the population mean estimator is given by

\[
\hat{m}_G = \sum_{h=1}^{H} a_h \hat{m}_{Gh} = \sum_{h=1}^{H} a_h \sum_{k \in U_h} \frac{1}{N_h} \hat{\mu}_h^k,
\]

with \( \hat{m}_{Gh} \) being the estimator of mean in stratum \( h \) and \( a_h = N_h / (\sum N_h) \) is the proportion of individuals from population in stratum \( h \). According to corollary 3, the variance estimator expression is given by

\[
\text{Var}(\hat{m}_G)_{SRS} = \sum_{h=1}^{H} a_h^2 \text{Var}(\hat{m}_{Gh})_{SRS} \approx \sum_{h=1}^{H} a_h^2 \frac{1 - R_{U_h}^2}{N_h} \frac{1 - f_h}{f_h} S_{yU_h}^2,
\]

14
The variance estimator expression is given by

\[ \hat{\text{Var}}(\hat{m}_G)_{\text{SRS}} = \sum_{h=1}^{H} a_h^2 \hat{\text{Var}}(\hat{m}_{Gh})_{\text{SRS}} = \sum_{h=1}^{H} a_h^2 \frac{1 - \hat{R}^2_h}{N_h} \frac{1}{f_h} \sigma_{y|h}^2 \]

With BER sampling, the mean population estimator expression is the same as for the SRS case. The variance estimator expression is given by

\[ \hat{\text{Var}}(\hat{m}_G)_{\text{BER}} = \sum_{h=1}^{H} a_h^2 \hat{\text{Var}}(\hat{m}_{Gh})_{\text{BER}} \approx \sum_{h=1}^{H} a_h^2 \frac{N_h - 1}{N_h^2} (1 - R^2_{U_h}) \frac{1}{\pi_h} \sigma_{y|h}^2 \]

The variance estimator expression is given by

\[ \hat{\text{Var}}(\hat{m}_G)_{\text{BER}} = \sum_{h=1}^{H} a_h^2 \hat{\text{Var}}(\hat{m}_{Gh})_{\text{BER}} = \sum_{h=1}^{H} a_h^2 \frac{(n_h - 1)}{N_h^2 \pi_h} (1 - R^2_{S_h}) \frac{1}{\pi_h} \sigma_{y|h}^2 \]

3.6 Simulation study

A simulation study was carried out for comparing the behavior of the estimators described above in different scenarios.

3.6.1 Bernoulli response

Twelve populations consisting of \( N = 1000 \) individuals each were generated where one auxiliary variable, \( x \), was considered, generated by following chi-square distribution, having 1 degree of freedom. The value of \( \hat{\mu}_k^U \) for the element \( k \) from each population was calculated as \( \hat{\mu}_k^U = \left[ 1 + \exp(-\hat{\beta}_U_{1} - \hat{\beta}_U_{2} x_k) \right]^{-1} \), and interest variable, \( y_k \), was randomly generated following Bernoulli distribution with \( \hat{\mu}_k^U \) as parameter. In each population, the \( \hat{\beta}_U = (\hat{\beta}_U_{1}, \hat{\beta}_U_{2})^T \) values were chosen so that they ensured values for \( R^2_{U} = 0.5, 0.65 \) and \( P = 0.15, 0.35, 0.65 \) and 0.85. 10000 independent samples were obtained for each population having \( n = 40 \) and \( n = 50 \) individuals, according to SRS scheme. These samples were used for calculating estimates of \( P = N^{-1} \sum y \) and estimator’s variance as well as 95% confidence intervals (based on normality) for \( P \) using two estimators assisted by regression models having normal and Bernoulli responses, denoted as \( \hat{P}_N \) and \( \hat{P}_B \), respectively. These GEREG estimators satisfied the conditions of corollary 1. Performance measurements were calculated for each estimator as: i) relative bias of \( \hat{P} \); ii) relative efficiency of \( \hat{P} \) regarding Horvitz-Thompson estimator; iii) relative bias of \( \hat{\text{Var}}(\hat{P}) \); and iv) 95% confidence interval coverage rate for \( P \). Table 3 presents the results. It can be observed that the relative bias of \( \hat{P}_B \) was always less than that for the \( \hat{P}_N \); however, relative bias was very small for both estimators. Regarding relative efficiency, the values in the Table were always lower than 100%, indicating
that the $\hat{P}_N$ and $\hat{P}_B$ estimators were more efficient than the Horvitz-Thompson one. However, greater efficiency was always provided for the estimator assisted by the model having Bernoulli response. The relative bias of $\hat{V}(P)$ was relatively large, mainly for the estimator assisted by the model having Bernoulli response. The 95% confidence interval coverage rates for $P$ for both estimators were always lower than nominal values; nevertheless, taking into account that the sample size was small, coverage rate values were good. Besides, it can be observed that $\hat{P}_B$ performed better than $\hat{P}_N$, mainly when the value of the population proportion $P$ was large.

Table 3
Performance of $\hat{P}_N$ and $\hat{P}_B$ estimating the population proportion $P$

<table>
<thead>
<tr>
<th>$R_U^2$</th>
<th>$n$</th>
<th>$P$</th>
<th>$\frac{P - \hat{P}_N}{\hat{V}(P)}$</th>
<th>$\frac{P - \hat{P}_N}{\hat{V}(\hat{P}_N)}$</th>
<th>CR</th>
<th>$\frac{P - \hat{P}_B}{\hat{V}(P)}$</th>
<th>$\frac{P - \hat{P}_B}{\hat{V}(\hat{P}_B)}$</th>
<th>CR</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.50</td>
<td>40</td>
<td>0.15</td>
<td>1.51</td>
<td>65.18</td>
<td>8.39</td>
<td>91.81</td>
<td>0.86</td>
<td>54.11</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.35</td>
<td>0.99</td>
<td>58.29</td>
<td>5.99</td>
<td>93.55</td>
<td>0.12</td>
<td>52.32</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.65</td>
<td>0.74</td>
<td>65.93</td>
<td>11.23</td>
<td>92.52</td>
<td>0.13</td>
<td>53.75</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.85</td>
<td>0.30</td>
<td>76.18</td>
<td>3.51</td>
<td>92.01</td>
<td>0.30</td>
<td>54.60</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>0.15</td>
<td>1.03</td>
<td>65.57</td>
<td>8.12</td>
<td>92.16</td>
<td>0.77</td>
<td>54.29</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.35</td>
<td>0.10</td>
<td>60.04</td>
<td>7.95</td>
<td>93.30</td>
<td>0.08</td>
<td>53.75</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.65</td>
<td>0.72</td>
<td>63.14</td>
<td>6.58</td>
<td>93.62</td>
<td>0.08</td>
<td>51.97</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.85</td>
<td>0.30</td>
<td>74.83</td>
<td>3.52</td>
<td>92.52</td>
<td>0.19</td>
<td>52.81</td>
</tr>
<tr>
<td>0.65</td>
<td>40</td>
<td>0.15</td>
<td>1.84</td>
<td>59.21</td>
<td>10.73</td>
<td>91.49</td>
<td>0.84</td>
<td>41.79</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.35</td>
<td>0.13</td>
<td>48.17</td>
<td>8.20</td>
<td>93.27</td>
<td>0.03</td>
<td>37.31</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.65</td>
<td>0.96</td>
<td>58.66</td>
<td>10.85</td>
<td>92.49</td>
<td>0.12</td>
<td>37.76</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.85</td>
<td>0.35</td>
<td>72.45</td>
<td>3.65</td>
<td>93.21</td>
<td>0.30</td>
<td>37.88</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>0.15</td>
<td>1.22</td>
<td>57.08</td>
<td>6.61</td>
<td>92.64</td>
<td>0.76</td>
<td>38.88</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.35</td>
<td>0.25</td>
<td>48.37</td>
<td>8.05</td>
<td>93.36</td>
<td>0.11</td>
<td>37.23</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.65</td>
<td>0.75</td>
<td>57.93</td>
<td>9.03</td>
<td>93.13</td>
<td>0.01</td>
<td>37.69</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.85</td>
<td>0.24</td>
<td>69.75</td>
<td>0.39</td>
<td>93.46</td>
<td>0.20</td>
<td>36.78</td>
</tr>
<tr>
<td>0.75</td>
<td>40</td>
<td>0.15</td>
<td>2.70</td>
<td>57.40</td>
<td>12.75</td>
<td>90.43</td>
<td>0.04</td>
<td>32.76</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.35</td>
<td>0.45</td>
<td>42.59</td>
<td>5.66</td>
<td>93.52</td>
<td>0.23</td>
<td>27.58</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.65</td>
<td>0.99</td>
<td>55.48</td>
<td>10.63</td>
<td>92.79</td>
<td>0.07</td>
<td>32.62</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.85</td>
<td>0.38</td>
<td>72.24</td>
<td>4.16</td>
<td>92.63</td>
<td>0.29</td>
<td>32.86</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>0.15</td>
<td>1.42</td>
<td>54.30</td>
<td>6.21</td>
<td>92.82</td>
<td>0.12</td>
<td>29.38</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.35</td>
<td>0.22</td>
<td>43.57</td>
<td>8.09</td>
<td>93.34</td>
<td>0.02</td>
<td>28.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.65</td>
<td>0.80</td>
<td>54.38</td>
<td>8.28</td>
<td>93.10</td>
<td>0.05</td>
<td>31.68</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.85</td>
<td>0.22</td>
<td>68.89</td>
<td>0.78</td>
<td>93.34</td>
<td>0.17</td>
<td>30.34</td>
</tr>
</tbody>
</table>

3.6.2 Poisson response

Nine populations consisting of $N = 1000$ individuals each were generated where one auxiliary variable, $x$, was considered, generated by following chi-square distribution with 1 degree of freedom. The $\mu_k^U$ value for element $k$ from each population was calculated as $\mu_k^U = \exp(\beta_{U1} + \beta_{U2}x_k)$, and interest variable, $y_k$, was randomly generated following Poisson distribution with $\mu_k^U$ as parameter. The $\beta_U$ values in each population were chosen so that they ensured that values of $R_U^2 = 0.5, 0.65$ and 0.75 and $m = N^{-1}t_y = 1.5, 3.0$ and 5.0. 10000 independent samples were obtained, having population sizes of $n = 40$ and $n = 50$ elements, according to SRS scheme. These samples were
used for calculating estimates of \( m \) and estimator’s variance as well as 95% confidence intervals (based on normality) for \( m \) using two estimators assisted by regression models having normal and Poisson responses, denoted as \( \hat{m}_N \) and \( \hat{m}_P \), respectively. These GEREG estimators satisfied the conditions of corollary 1. The same measurements of performance as in the Bernoulli case were calculated for each estimator. Table 4 presents the results. It can be observed that the relative bias of \( \hat{m}_P \) was always less than that for the \( \hat{m}_N \) case. However, relative bias was very small for both estimators. Regarding relative efficiency, the values in the Table were always lower than 100%, indicating that the \( \hat{m}_N \) and \( \hat{m}_P \) estimators were more efficient than the Horvitz-Thompson one. However, greater efficiency always pertained to the estimator assisted by the model having Poisson response. The relative bias of \( \text{Var}(\hat{m}) \) was relatively large, mainly for the estimator assisted by the model having Poisson response. For both estimators, the 95% confidence interval coverage rates for \( m \) were always lower than the nominal values; nevertheless, taking into account that the sample size was small, coverage rate values were good. Besides, it can be observed that \( \hat{m}_P \) performed better than \( \hat{m}_N \), mainly when population mean \( m \) value was small.

### Table 4

Performance of \( \hat{m}_N \) and \( \hat{m}_P \) estimating the population mean \( m \)

| \( R^2_U \) | \( n \) | \( m \) | \( \hat{m}_N \) | \( \text{Measures(\%) \Big|}_{\hat{m}_N} \) | \( \hat{m}_P \) | \( \text{Measures(\%) \Big|}_{\hat{m}_P} \) |
|---|---|---|---|---|---|---|
| (\( \bar{m} \)) | (\( \text{V}(\hat{m}) \)) | (\( \text{V}(\hat{m}) \)) | (\( \text{CR} \)) | (\( \bar{m} \)) | (\( \text{V}(\hat{m}) \)) | (\( \text{V}(\hat{m}) \)) | (\( \text{CR} \)) |
| 0.50 | 40 | 1.5 | 1.21 | 60.94 | 8.37 | 91.98 | 0.62 | 55.58 | 16.65 | 91.25 |
| | | 3.0 | 0.78 | 56.91 | 8.75 | 92.34 | 0.38 | 55.01 | 14.71 | 91.47 |
| | | 5.0 | 0.44 | 54.13 | 8.01 | 92.83 | 0.19 | 52.06 | 11.13 | 92.39 |
| | 50 | 1.5 | 0.98 | 60.80 | 6.07 | 92.34 | 0.42 | 55.12 | 14.48 | 91.75 |
| | | 3.0 | 0.54 | 56.36 | 5.96 | 93.04 | 0.30 | 53.51 | 10.63 | 92.50 |
| | | 5.0 | 0.30 | 55.11 | 8.49 | 92.91 | 0.18 | 52.66 | 11.10 | 92.58 |
| 0.65 | 40 | 1.5 | 2.00 | 58.41 | 14.77 | 89.18 | 1.35 | 49.11 | 16.42 | 88.98 |
| | | 3.0 | 1.43 | 51.69 | 10.44 | 91.05 | 0.37 | 37.24 | 15.45 | 91.27 |
| | | 5.0 | 0.80 | 45.99 | 11.18 | 91.37 | 0.28 | 37.23 | 14.95 | 90.99 |
| | 50 | 1.5 | 1.65 | 54.64 | 9.55 | 90.50 | 1.01 | 43.68 | 17.17 | 89.33 |
| | | 3.0 | 1.08 | 51.02 | 7.72 | 91.61 | 0.21 | 35.85 | 10.55 | 92.26 |
| | | 5.0 | 0.76 | 45.68 | 9.27 | 92.26 | 0.12 | 36.97 | 12.76 | 92.26 |
| 0.75 | 40 | 1.5 | 3.22 | 58.22 | 14.16 | 84.41 | 0.65 | 34.80 | 14.59 | 85.49 |
| | | 3.0 | 1.85 | 49.67 | 14.06 | 88.63 | 0.59 | 30.20 | 18.12 | 88.74 |
| | | 5.0 | 1.17 | 43.07 | 12.33 | 90.36 | 0.01 | 26.79 | 17.01 | 91.51 |
| | 50 | 1.5 | 2.52 | 58.47 | 10.11 | 86.11 | 0.49 | 32.03 | 17.59 | 87.22 |
| | | 3.0 | 1.41 | 48.60 | 10.57 | 89.92 | 0.50 | 28.96 | 13.10 | 89.98 |
| | | 5.0 | 1.12 | 43.02 | 10.63 | 91.02 | 0.16 | 26.23 | 13.83 | 92.08 |

### 3.6.3 Gamma response

Nine populations consisting of \( N = 1000 \) individuals each were generated, where one auxiliary variable, \( x \) was considered, being generated following chi-square distribution with 1 degree of freedom. The value of \( \hat{\mu}_k^U \) for e element \( k \) for each population was calculated as \( \hat{\mu}_k^U = (\hat{\beta}_{U1} + \hat{\beta}_{U2} x_k)^{-1} \) and interest
variable, $y_k$, was randomly generated following Gamma distribution with $\hat{\mu}_k^U$ and $\phi$ as parameters. In each population, the values of $\hat{\beta}_U$ and $\phi$ were chosen so that they ensured the values of $R^2_U = 0.5, 0.65$ and $0.75$ and $m = 5, 10$ and $15$. 10000 independent samples were obtained from each population, $n = 40$ and $n = 50$ element sizes, according to SRS scheme. These samples were used for calculating estimates of $m$ and estimator’s variance as well as 95% confidence intervals (based on normality) for $m$ using two estimators assisted by regression models having normal and Gamma responses, denoted as $\hat{m}_N$ and $\hat{m}_G$, respectively. These GEREG estimators satisfied the conditions of corollary 1. The same performance measurements as in the Bernoulli and Poisson cases were calculated for each estimator. Table 5 presents the results.

It can be observed that the relative bias of $\hat{m}_G$ was always less than for $\hat{m}_N$. However, relative bias was very small for both estimators. Regarding relative efficiency, the values in the Table were always lower than 100%, indicating that the $\hat{m}_N$ and $\hat{m}_G$ estimators were more efficient than the Horvitz-Thompson one. However, the greater efficiency was always given by the estimator assisted by the model having gamma response. The relative bias of $\text{Var}(\hat{m})$ was relatively small. For both estimators, 95% confidence interval coverage rates for $m$ were always lower than nominal values; nevertheless, taking into account that sample size was small, coverage rate values were good. Besides, it can be observed that $\hat{m}_G$ performed better than $\hat{m}_N$, mainly when populational mean $m$ value was small.

**Table 5**

Performance of $\hat{m}_N$ and $\hat{m}_G$ estimating the population mean $m$

<table>
<thead>
<tr>
<th>$R^2_U$</th>
<th>$n$</th>
<th>$m$</th>
<th>$\hat{m}_N$</th>
<th>$\hat{m}_G$</th>
<th>CR</th>
<th>$\frac{m}{m - 1}$</th>
<th>$\frac{\text{V}(\hat{m})}{\text{V}(\hat{m}_{HT})}$</th>
<th>$\frac{\text{V}(\hat{m})}{\text{V}(\hat{m}^2)}$</th>
<th>$\frac{\text{V}(\hat{m})}{\text{V}(\hat{m})}$</th>
<th>CR</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.50</td>
<td>40</td>
<td>5</td>
<td>1.29</td>
<td>69.75</td>
<td>8.62</td>
<td>92.02</td>
<td>0.19</td>
<td>53.58</td>
<td>4.82</td>
<td>93.22</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10</td>
<td>0.55</td>
<td>66.39</td>
<td>11.10</td>
<td>92.04</td>
<td>0.03</td>
<td>55.73</td>
<td>5.66</td>
<td>93.36</td>
</tr>
<tr>
<td></td>
<td></td>
<td>15</td>
<td>0.54</td>
<td>61.45</td>
<td>8.81</td>
<td>92.20</td>
<td>0.09</td>
<td>52.28</td>
<td>4.23</td>
<td>93.30</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>5</td>
<td>0.94</td>
<td>67.97</td>
<td>7.62</td>
<td>92.24</td>
<td>0.18</td>
<td>52.76</td>
<td>4.52</td>
<td>93.23</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10</td>
<td>0.51</td>
<td>63.59</td>
<td>6.45</td>
<td>93.04</td>
<td>0.04</td>
<td>53.27</td>
<td>0.84</td>
<td>94.34</td>
</tr>
<tr>
<td></td>
<td></td>
<td>15</td>
<td>0.35</td>
<td>58.96</td>
<td>5.37</td>
<td>93.36</td>
<td>0.06</td>
<td>51.15</td>
<td>3.56</td>
<td>93.84</td>
</tr>
<tr>
<td>0.65</td>
<td>40</td>
<td>5</td>
<td>1.76</td>
<td>72.40</td>
<td>14.26</td>
<td>90.66</td>
<td>0.12</td>
<td>41.65</td>
<td>4.96</td>
<td>93.08</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10</td>
<td>1.15</td>
<td>58.86</td>
<td>12.61</td>
<td>91.28</td>
<td>0.02</td>
<td>37.61</td>
<td>2.93</td>
<td>93.96</td>
</tr>
<tr>
<td></td>
<td></td>
<td>15</td>
<td>0.80</td>
<td>53.42</td>
<td>10.63</td>
<td>92.06</td>
<td>0.09</td>
<td>37.60</td>
<td>7.10</td>
<td>93.32</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>5</td>
<td>1.38</td>
<td>71.60</td>
<td>12.98</td>
<td>91.90</td>
<td>0.11</td>
<td>41.39</td>
<td>4.67</td>
<td>93.18</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10</td>
<td>0.87</td>
<td>59.87</td>
<td>12.23</td>
<td>92.08</td>
<td>0.03</td>
<td>37.45</td>
<td>0.81</td>
<td>94.10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>15</td>
<td>0.54</td>
<td>51.95</td>
<td>10.02</td>
<td>92.36</td>
<td>0.08</td>
<td>37.58</td>
<td>1.75</td>
<td>93.78</td>
</tr>
<tr>
<td>0.75</td>
<td>40</td>
<td>5</td>
<td>2.19</td>
<td>76.46</td>
<td>12.65</td>
<td>90.40</td>
<td>0.24</td>
<td>34.61</td>
<td>8.66</td>
<td>92.18</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10</td>
<td>1.47</td>
<td>63.31</td>
<td>18.95</td>
<td>90.42</td>
<td>0.10</td>
<td>32.57</td>
<td>6.74</td>
<td>92.66</td>
</tr>
<tr>
<td></td>
<td></td>
<td>15</td>
<td>1.23</td>
<td>58.05</td>
<td>19.50</td>
<td>89.54</td>
<td>0.12</td>
<td>29.98</td>
<td>4.12</td>
<td>93.30</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>5</td>
<td>1.62</td>
<td>76.49</td>
<td>11.76</td>
<td>91.26</td>
<td>0.16</td>
<td>36.12</td>
<td>6.36</td>
<td>92.22</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10</td>
<td>1.08</td>
<td>61.48</td>
<td>15.09</td>
<td>90.76</td>
<td>0.10</td>
<td>32.84</td>
<td>6.39</td>
<td>92.98</td>
</tr>
<tr>
<td></td>
<td></td>
<td>15</td>
<td>0.84</td>
<td>55.31</td>
<td>13.96</td>
<td>91.06</td>
<td>0.06</td>
<td>29.49</td>
<td>4.08</td>
<td>93.64</td>
</tr>
</tbody>
</table>

18
4 Applications

4.1 Academic Performance Index (API)

The academic performance index (API) has been used annually as performance measurement for every school in California. The API is based on a standardized test applied to students. Data from 2000 was analyzed. The data set contains 6194 observations and 37 variables and was obtained from the Survey package for the statistical software R (www.r-project.org). The goal was to estimate the API average for the population of schools using a stratified sample design. The strata were formed by the levels of Stype variable (elementary (E), middle (M) and high school (H)). SRS and BER sampling scheme performance was investigated in each strata. The \( \hat{m}_{\text{G}} \) estimator discussed in section 3.5.6 was applied; the estimation in each stratum was assisted by models having normal and Gamma responses as described in (9) and (13). The percentage of parents not graduating from high school in each school not.hsg was the auxiliary variable. Strata sizes were given by \( N_1 = 4421, N_2 = 1018 \) and \( N_3 = 755 \). Strata sample sizes were 100, 50 and 50, respectively for the first, second and third stratum. Such numbers corresponded to expected sample sizes when using a BER sampling scheme.

Table 6 shows \( \hat{m}_{\text{G}} \) relative efficiency (\( \Delta_{p(i)} \)) regarding the Horvitz-Thompson estimator for the two regression models and the two sample schemes. It can be seen that the Gamma assisted \( \hat{m}_{\text{G}} \) estimator was more efficient than the normal assisted one in every strata. Normal assisted estimator variance was approximately 1.11 times greater than the variance for the Gamma model assisted estimator for both sampling schemes. Figure 1 shows the relationship between interest and the auxiliary variables. The \( R^2_U \) values for the normal model’s goodness of fit were 0.51, 0.54 and 0.61 for E, M and H, respectively. The \( R^2_U \) values for Gamma model were 0.55, 0.60 and 0.66 for E, M and H.

<table>
<thead>
<tr>
<th>Sampling scheme</th>
<th>Model assisting</th>
<th>School type</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Elemental</td>
<td>Middle</td>
</tr>
<tr>
<td>SRS</td>
<td>Normal</td>
<td>49.11</td>
<td>46.48</td>
</tr>
<tr>
<td></td>
<td>Gamma</td>
<td>44.65</td>
<td>40.32</td>
</tr>
<tr>
<td>BER</td>
<td>Normal</td>
<td>1.70</td>
<td>1.49</td>
</tr>
<tr>
<td></td>
<td>Gamma</td>
<td>1.64</td>
<td>1.41</td>
</tr>
</tbody>
</table>
4.2 Quality of life in Colombia

The unsatisfied basic necessities indicator (UBN) was calculated for the 1118 Colombian municipalities, based on the household and population census carried out by the National Administrative Department of Statistics - DANE (www.dane.gov.co). The data set was related to information from 2005. The UBN is an indicator of whether the population’s basic needs are been provided for, concerning i) inadequacy housing, ii) overpopulated households, iii) households having inadequated services, iv) high economic dependence households and v) families were children do not attend school. The goal was to estimate the proportion of municipalities (excluding the capital Bogota) presenting a higher than 25% UBN. A stratified sample design was used with strata based on a Distance variable defined as the distance, in kilometers, between each municipality and the closest city having more than 1 million inhabitants. Three strata were considered: one having a distance of less that 150 km (<150 km), one having 150 to 200 Km distance (150-200 km) and one having a distance of over 200 km (>200 km). SRS and BER scheme performance was investigated in each strata. The \( \hat{m}_G \) estimator discussed in the
section 3.5.6 was applied, the estimation in each strata was assisted by models having normal and Bernoulli responses as described in (9) and (11). The variable Density, the average number of people living in an area of 1 km$^2$ of the city was the auxiliary variable. Stratum sizes were $N_1 = 308$, $N_2 = 395$ and $N_3 = 414$, respectively. Stratum sample sizes were 46, 59 and 62, respectively, for the first, second and third stratum. Such sizes were expected values when using BER sampling.

Table 7 shows the relative efficiency ($\Delta_p(c)$) of the $\hat{m}_c$ estimator compared to the Horvitz-Thompson estimator, for both regression models and sample schemes. It can be seen that the Bernoulli assisted $\hat{m}_c$ estimator was more efficient than the normal assisted one in every strata. The variance of the normal assisted estimator was approximately 1.18 times greater than the variance of the Bernoulli model assisted estimator for both sampling schemes. Figure 2 shows the relationship between interest and auxiliary variables. The $R^2_U$ values for the normal model’s goodness of fit were 0.06, 0.27 and 0.13 for $<150$, 150-200 and $>200$, respectively. The $R^2_U$ values for the Bernoulli model were 0.24, 0.40 and 0.20 for $<150$, 150-200 and $>200$, respectively.

Additionally, we wanted to estimate the average number of UBN components higher than 10%. In this case, possible interest variable values were 0, 1, 2, 3, 4 and 5, thereby motivating an estimator assisted by a Poisson regression model. Table 8 shows the relative efficiency ($\Delta_p(c)$) of the $\hat{m}_c$ estimator compared to the Horvitz-Thompson estimator for both regression models and sample schemes. It can be seen that the Poisson assisted $\hat{m}_c$ estimator was more efficient than the normal assisted one in all strata. Normal assisted estimator variance was about 1.17 times greater than variance for the Poisson model assisted estimator for both sampling schemes. Figure 3 shows the relationship between interest and auxiliary variables. The $R^2_U$ values for the normal model’s goodness of fit were 0.06, 0.16 and 0.10 for $<150$, 150-200 and $>200$, respectively. The $R^2_U$ values for the Poisson model were 0.24, 0.27 and 0.21 for $<150$, 150-200 and $>200$, respectively.

<table>
<thead>
<tr>
<th>Sampling scheme</th>
<th>Model assisting</th>
<th>Distance (Kms)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Normal</td>
<td>$&lt; 150$</td>
<td>72.76</td>
</tr>
<tr>
<td></td>
<td>Bernoulli</td>
<td>93.09</td>
<td>150 - 200</td>
</tr>
<tr>
<td>SRS</td>
<td></td>
<td>&gt; 200</td>
<td>6.60</td>
</tr>
<tr>
<td>BER</td>
<td>Normal</td>
<td>21.68</td>
<td>6.99</td>
</tr>
<tr>
<td></td>
<td>Bernoulli</td>
<td>21.25</td>
<td>6.86</td>
</tr>
</tbody>
</table>
Fig. 2. Scatter plots between interest and auxiliary variables for the *Quality of life in Colombia* application when the interest variable is dichotomous

Table 8
Relative efficiency (percentage) of $\hat{m}_G$ regarding to Horvitz-Thompson estimator for the *Quality of life in Colombia* application when the interest variable is a counting process

<table>
<thead>
<tr>
<th>Sampling scheme</th>
<th>Model assisting</th>
<th>Distance (Kms)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>&lt; 150</td>
<td>150 - 200</td>
</tr>
<tr>
<td>SRS</td>
<td>Normal</td>
<td>93.88</td>
<td>83.84</td>
</tr>
<tr>
<td></td>
<td>Poisson</td>
<td>76.42</td>
<td>73.23</td>
</tr>
<tr>
<td>BER</td>
<td>Normal</td>
<td>23.22</td>
<td>13.01</td>
</tr>
<tr>
<td></td>
<td>Poisson</td>
<td>22.21</td>
<td>12.66</td>
</tr>
</tbody>
</table>

Appendix

A Theorem 1

*Proof.* .
Fig. 3. Scatter plots between interest and auxiliary variables for the *Quality of life in Colombia* example when the interest variable is a counting process

**Distance less than 150 kms**

- Components of UBN more than 10%
- Population density (persons/km²)

**Distance between 150 and 200 kms**

- Components of UBN more than 10%
- Population density (persons/km²)

**Distance more than 200 kms**

- Components of UBN more than 10%
- Population density (persons/km²)

**Total**

- Components of UBN more than 10%
- Population density (persons/km²)

**Mathematical Details**

i) $\hat{\beta}_U$ estimate satisfies $U_U^T(\hat{\beta}_U) = 0$, i.e., $U_{Uj}^T(\hat{\beta}_U) = h_j^T E = 0$, $j = 1, \ldots, J$, where $h_j = (h_{kj}, \ldots, h_{Nj})^T$ and $E = (E_1, \ldots, E_N)^T$, with $h_{kj} = \phi_k \frac{\partial \theta_k}{\partial \eta_k} x_{kj}$.

If exist $a = (a_1, \ldots, a_J)^T$ such that $\sum_{j=1}^{J} a_j h_j = 1_N$, then, $\sum_{k \in U} (y_k - \mu_k^U) = 1_N^T E = \left(\sum_{j=1}^{J} a_j h_j\right)^T E = \sum_{j=1}^{J} a_j h_j^T E$. From $U_{Uj}^T(\hat{\beta}_U) = 0$ above, $j = 1, \ldots, J$, we have $\sum_{j=1}^{J} a_j h_j^T E = 0$, therefore, $t_y = \sum_{k \in U} \hat{\mu}_k^U$.

ii) $\hat{\beta}_S^\pi$ estimate satisfies $U_S^\pi(\hat{\beta}_S^\pi) = 0$, i.e., $U_{Sj}^\pi(\hat{\beta}_S^\pi) = h_{Sj}^\pi e^\pi = 0$, $j = 1, \ldots, J$, where $h_{Sj}$ and $e^\pi$ are $n$-dimension column vectors with elements $h_{kj}$ and $e_k/\pi_k$, respectively, with $k \in S$. If exist $a$ such that $\sum_{j=1}^{J} a_j h_j = 1_N$, then, $\sum_{j=1}^{J} a_j h_{Sj} = 1_n$. Therefore, $\sum_{k \in S} \frac{(y_k - \hat{\mu}_k^S)}{\pi_k} = 1_n^T \pi \ e^\pi = \left(\sum_{j=1}^{J} a_j h_{Sj}\right)^T e^\pi = \sum_{j=1}^{J} a_j h_{Sj}^T e^\pi$. From $U_{Sj}^\pi(\hat{\beta}_S^\pi) = 0$ above, $j = 1, \ldots, J$, we have $\sum_{j=1}^{J} a_j h_{Sj}^\pi e^\pi = \sum_{j=1}^{J} a_j h_{Sj}^\pi e^\pi$. 

23
0, then, \( \hat{t}_G = \sum_{k \in U} \hat{\mu}^S_k \).

\[\square\]

### B Corollary 1

**Proof.** Homogeneity in the dispersion parameter, systematic component having intercept and canonical link led to \( h_{kj} = \phi x_{kj} \), where \( h_{k1} = \phi \). Then, with \( a_1 = \phi^{-1} \) and \( a_j = 0 \) for \( j \neq 1 \), we have \( \sum_{j=1}^{J} a_j h_{j} = 1_n \). Therefore, using theorem 1 we could proved that \( t_y = \sum_{k \in U} \hat{\mu}^U_k \) and \( \hat{t}_G = \sum_{k \in U} \hat{\mu}^S_k \).

\[\square\]

### C Corollary 2

**Proof.** Normal response, systematic component having canonical link and dispersion parameter expressed as \( \phi^{-1} = \sum_{j=1}^{J} a_j x_{kj} \), led to \( h_{kj} = x_{kj} \). We could verified that \( \sum_{j=1}^{J} a_j h_{j} = 1_N \). Therefore, using theorem 1 we could proved that \( t_y = \sum_{k \in U} \hat{\mu}^U_k \) and \( \hat{t}_G = \sum_{k \in U} \hat{\mu}^S_k \).

\[\square\]

**Lemma 1.** If the \( \hat{t}_G \) is based on an \( \xi \) model where it has been verified that there is homogeneity in the dispersion parameter, systematic component having intercept and canonical link, we have \( \hat{\mu}^T(\mathbf{Y}_U - \hat{\mu}) \approx 0 \) with \( \hat{\mu} = (\hat{\mu}^U_1, \ldots, \hat{\mu}^U_N)^T \)

**Proof.** Expanding \( \hat{\mu}^U_k \) by Taylor series of first-order around of \( x_k^o = (1, x_{k2}^o, \ldots, x_{kJ}^o)^T \) we have the following

\[ \hat{\mu}^U_k = \mu(x_k, \hat{\beta}_U) \approx \hat{\mu}^o_k + \sum_{j=2}^{J} V(\hat{\mu}^o_k) \hat{\beta}_U(x_{kj} - x_{kj}^o), \]

where \( \hat{\mu}^o_k = \mu(x_k^o, \hat{\beta}_U) \). For example, with \( x_{kj}^o = N^{-1} t_{x_j} = \bar{x}_{U_j}, \ j = 2, \ldots, J, \) we have that \( x_k^o = x^o \) and therefore \( \hat{\mu}^o_k = \hat{\mu}^o \), then

\[ \hat{\mu}^U_k \approx \hat{\mu}^o + \sum_{j=2}^{J} V(\hat{\mu}^o) \hat{\beta}_U(x_{kj} - \bar{x}_{U_j}) \]

\[ = \sum_{j=1}^{J} m_j x_{kj}, \]

24
where \( m_j = V(\hat{\mu}^*)\hat{\beta}_{Uj} \) for \( j \neq 1 \) and \( m_1 = \hat{\mu}^* - \sum_{j=2}^{J} V(\hat{\mu}^*)\hat{\beta}_{Uj} \). In matrix form \( \hat{\mu} \) can be written as

\[
\hat{\mu} \approx X_{\mu} m,
\]

with \( m = (m_1, \ldots, m_J)^T \). \( \hat{\beta}_{U} \) estimate satisfies \( U_U(\hat{\beta}_{U}) = 0 \). As \( \hat{t}_G \) is based on an \( \xi \) model where it has been verified that there is homogeneity in the dispersion parameter, systematic component having intercept and canonical link, \( U_U(\beta) \) can be expressed as \( U_U(\beta) = X_U^T(Y_U - \mu) \), therefore, we have

\[
X_U^T(Y_U - \hat{\mu}) = 0.
\]

D Theorem 2

Proof. According Pregibon (1981) in the convergence of the iterative process to solving \( U_S(\hat{\beta}_S^\pi) = 0 \), \( \hat{\beta}_S^\pi \) estimate can be expressed as

\[
\hat{\beta}_S^\pi = (X_S^T\hat{V}_S X_S)^{-1}X_S^T\hat{V}_S \hat{Y}_S^T
\]

\[
\approx (X_S^T\hat{V}_U X_S)^{-1}X_S^T\hat{V}_U \hat{Y}_U^T, \tag{D.1}
\]

where \( \hat{V}_S = \text{diag}\{\hat{V}_1^S \phi_1/\pi_1, \ldots, \hat{V}_n^S \phi_n/\pi_n\} \), \( \hat{V}_U = \text{diag}\{\hat{V}_1^U \phi_1/\pi_1, \ldots, \hat{V}_n^U \phi_n/\pi_n\} \), \( \hat{Y}_S = (\hat{y}_1^S, \ldots, \hat{y}_n^S)^T \), \( \hat{Y}_U = (\hat{y}_1^U, \ldots, \hat{y}_n^U)^T \) with \( \hat{y}_k^S = \{\eta_k + V_k^{-1}(y_k - \mu_k)\} \beta_S^\pi = \hat{\beta}_S^\pi \) and \( \hat{y}_k^U = \{\eta_k + V_k^{-1}(y_k - \mu_k)\} \beta = \hat{\beta}_U \). From result 5.10.1 of Särndal, Swensson and Wretman (1992) we have that \( \beta_S^\pi \) (as in the expression D.1) is an approximately unbiased estimator of \( \hat{\beta}_U \) (i.e. \( E(\beta_S^\pi - \hat{\beta}_U) \approx 0 \)) with the approximate variance-covariance matrix given by

\[
(X_U^T W_U^* X_U)^{-1}\Sigma_U (X_U^T W_U^* X_U)^{-1}
\]

Expanding the GEREG estimator \( \hat{t}_G \) by Taylor series of first-order around of \( \beta_U \) we have the following

\[
\hat{t}_G \approx \sum_{k \in U} \hat{\mu}_k^U + \sum_{j=1}^{J} \sum_{k \in U} \hat{V}_k^U x_{kj} (\hat{\beta}_S^\pi - \hat{\beta}_U)
\]

\[
\approx t_y + 1_N^T \hat{V}_U X_U (\hat{\beta}_S^\pi - \hat{\beta}_U).
\]

Then, we can write
i) \( E(\hat{t}_G) \approx t_y + \mathbf{1}_N^T \hat{\mathbf{V}}_U \mathbf{X}_U E(\hat{\beta}_S^* - \hat{\beta}_U) = t_y \)

ii) \( \text{Var}(\hat{t}_G) \approx \mathbf{1}_N^T \hat{\mathbf{V}}_U \mathbf{X}_U \text{Var}(\hat{\beta}_S^* - \hat{\beta}_U) \mathbf{X}_U^T \hat{\mathbf{V}}_U \mathbf{1}_N = \Gamma_U^T \Sigma_U \Gamma_U \),

and the \( \text{Var}(\hat{t}_G) \) estimator can be obtained replacing the unknown quantities by its estimators in \( \text{Var}(\hat{t}_G) \), i.e, \( \hat{\mathbf{V}}_k^U \) by \( \hat{\mathbf{V}}_S^k \) and \( \sigma_{ij}^U \) by \( \sigma_{ij}^S \), well

iii) \( \hat{\text{Var}}(\hat{t}_G) = \Gamma_S^T \Sigma_S \Gamma_S. \)

\[ \square \]

E  Corollary 3

**Proof.** As \( \hat{t}_G \) is based in an \( \xi \) model where it has been verified that there is homogeneity in the dispersion parameter and systematic component with intercept and canonical link we have

\[
\Gamma_U = (\mathbf{X}_U^T \hat{\mathbf{V}}_U \mathbf{X}_U)^{-1} \mathbf{X}_U^T \hat{\mathbf{V}}_U \mathbf{1}_N = (1, 0, \ldots, 0)_{(J \times 1)},
\]

then, \( \text{Var}(\hat{t}_G) \) is the \( (1, 1) \) element of \( \Gamma_U \), i.e, \( \text{Var}(\hat{t}_G) \approx \sigma_{11}^U \) and \( \hat{\text{Var}}(\hat{t}_G) \approx \sigma_{11}^S \), where \( x_{k1} = \phi_k = 1 \) for \( k \in U \).

\[ \square \]

F  Theorem 3

**Proof.** Using the Pearson’s linear correlation coefficient definition we can write

\[
1 - R_U^2 = 1 - \frac{\left[ \sum_{k \in U} y_k \hat{\mu}_k^U - N \bar{y}_U \bar{\mu}_U \right]^2}{\left( \sum_{k \in U} y_k^2 - N \bar{y}_U^2 \right) \left( \sum_{k \in U} (\hat{\mu}_k^U)^2 - N \bar{\mu}_U^2 \right)},
\]

Where \( \bar{\mu}_U = N^{-1} \sum_{k \in U} \hat{\mu}_k^U \). As the conditions of corollary 1 are satisfied we have \( \sum_{k \in U} (y_k - \hat{\mu}_k^U) = 0 \), then, \( \bar{y}_U = \bar{\mu}_U \) and we can write the following

\[
1 - R_U^2 = 1 - \frac{\left[ \sum_{k \in U} y_k \hat{\mu}_k^U - N \bar{y}_U \bar{\mu}_U \right]^2}{\left( \sum_{k \in U} y_k^2 - N \bar{y}_U^2 \right) \left( \sum_{k \in U} (\hat{\mu}_k^U)^2 - N \bar{\mu}_U^2 \right)}
\]

26
From lemma 1 we have $\hat{\mu}^T(Y_U - \hat{\mu}) \approx 0$, i.e., $\sum_{k \in U} y_k \hat{\mu}_k^U \approx \sum_{k \in U} (\hat{\mu}_k^U)^2$, therefore

$$1 - R^2_U \approx 1 - \frac{\left[ \sum_{k \in U} \hat{\mu}_k^{U2} - N \bar{y}_U^2 \right]^2}{\left( \sum_{k \in U} y_k^2 - N \bar{y}_U^2 \right) \left( \sum_{k \in U} (\hat{\mu}_k^U)^2 - N \bar{y}_U^2 \right)}$$

$$= 1 - \frac{\sum_{k \in U} y_k^2 - 2 \sum_{k \in U} y_k \hat{\mu}_k^U + \sum_{k \in U} \hat{\mu}_k^{U2}}{\sum_{k \in U} y_k^2 - N \bar{y}_U^2}$$

$$= \frac{\sum_{k \in U} (y_k - \hat{\mu}_k^U)^2}{\sum_{k \in U} (y_k - \bar{y}_U)^2} = \Delta_{AAS}.$$


