# TRUMP: Tuned Regression Unbiased Mean Predictor

**Sarjinder Singh and Stephen A. Sedory**
Department of Mathematics
Texas A&M University-Kingsville
**Kingsville, TX 78363**

## Abstract

In this paper, we introduce a new Tuned Regression Unbiased Mean Predictor (TRUMP) which we show can be adjusted to have smaller variance than the linear regression predictor due to Hansen, Hurwitz and Madow (1953) especially when there is heteroscedasticity of a form that depends on a parameter, which we call here the Hillary Campaign Coefficient. In that case the proposed new TRUMP model can be made more efficient than the Best Linear Unbiased Predictor (BLUP) when it is based on an appropriate choice of a parameter called the TRUMP Care coefficient. We extend the work to chain-type TRUMP Cuts. Some highlights of the work from Singh and Sedory (2017b) are presented here.

**Key words:** Calibration, Jackknifing, TRUMP Cuts, TRUMP care Coefficient, Chain-Type TRUMP Cuts, First Basic Information (FBI).

## 1. Introduction

Let $y_i$ and $x_i$, $i=1,2,...,N$, be the values of the study variable and auxiliary variable, respectively, of the i$^{th}$ unit in the population $\Omega$. Here we consider the problem of estimating the population mean

$$\bar{Y} = N^{-1}\sum_{i=1}^{N} y_i \tag{1.1}$$

by assuming that the population mean

$$\bar{X} = N^{-1}\sum_{i=1}^{N} x_i \tag{1.2}$$

of the auxiliary variable is known.

Let $(y_i, x_i)$, $i=1,2,...,n$, be the values of the study variable and auxiliary variable of the i$^{th}$ unit in the sample $s$ drawn using the simple random and with replacement sampling (SRSWR) scheme.

Let

$$\bar{y}_n = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{1.3}$$

and

$$\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1.4}$$

be the sample means for the study variable and the auxiliary variable respectively.

Assuming linear regression where the regression line need not pass through the origin, the well known mean predictor model is given by:

$$y_i = \alpha + \beta x_i + e_i \tag{1.5}$$

with the assumptions:

$$E_m(e_i \mid x_i) = 0 \tag{1.6}$$

$$E_m(e_i^2 \mid x_i) = V_m(e_i \mid x_i) = \sigma^2 v(x_i) \tag{1.7}$$

and

$$E_m(e_i e_j \mid x_i x_j) = C_m(e_i, e_j \mid x_i, x_j) = 0 \tag{1.8}$$

where $E_m$, $V_m$, $C_m$ denote the model expectation, variance and covariance, and $v(x_i)$ is any function of $x_i$ determining the heteroscedasticity in a population.

It will be helpful to point out that the estimators $\hat{\alpha}$ and $\hat{\beta}$ are to be obtained by minimizing, with respect to $\alpha$ and $\beta$, the weighted error sum of squares (WSSE), given by

$$\text{WSSE} = \sum_{i=1}^{n} \{v(x_i)\}^{-1} e_i^2 = \sum_{i=1}^{n} \{v(x_i)\}^{-1} \left[y_i - \alpha - \beta x_i\right]^2 \tag{1.9}$$

Then the Best Linear Unbiased Predictor (BLUP), in the case of heteroscedasticity, is given by

$$\bar{y}_{reg} = \hat{\alpha} + \hat{\beta}\,\bar{X} \tag{1.10}$$

where

$$\hat{\alpha} = \frac{\left(\sum_{i=1}^{n} \frac{y_i}{v(x_i)}\right)\left(\sum_{i=1}^{n} \frac{x_i^2}{v(x_i)}\right) - \left(\sum_{i=1}^{n} \frac{x_i}{v(x_i)}\right)\left(\sum_{i=1}^{n} \frac{x_i y_i}{v(x_i)}\right)}{\left(\sum_{i=1}^{n} \frac{1}{v(x_i)}\right)\left(\sum_{i=1}^{n} \frac{x_i^2}{v(x_i)}\right) - \left(\sum_{i=1}^{n} \frac{x_i}{v(x_i)}\right)^2} \tag{1.11}$$

and

$$\hat{\beta} = \frac{\left(\sum_{i=1}^{n} \frac{1}{v(x_i)}\right)\left(\sum_{i=1}^{n} \frac{x_i y_i}{v(x_i)}\right) - \left(\sum_{i=1}^{n} \frac{x_i}{v(x_i)}\right)\left(\sum_{i=1}^{n} \frac{y_i}{v(x_i)}\right)}{\left(\sum_{i=1}^{n} \frac{1}{v(x_i)}\right)\left(\sum_{i=1}^{n} \frac{x_i^2}{v(x_i)}\right) - \left(\sum_{i=1}^{n} \frac{x_i}{v(x_i)}\right)^2} \tag{1.12}$$

Note that the estimators $\hat{\alpha}$ and $\hat{\beta}$ are Best Linear Unbiased Estimators (BLUE) of $\alpha$ and $\beta$ if the model is Homoscedastic. If there is heteroscedasticity, then $\hat{\alpha}$ and $\hat{\beta}$ are still unbiased estimators of $\alpha$ and $\beta$, but do not follow the Gauss Markov Theorem for minimum variance. This result from the Gauss-Markov theorem leaves room for obtaining improved estimators of $\alpha$ and $\beta$ which could further improve Best Linear Unbiased Predictor (BLUP) in the presence of heteroscedasticity.

Singh and Sedory (2017a) have proposed a Tuned Ratio Unbiased Mean Predictor (TRUMP) given by

$$\bar{y}_{\text{TRUMP}} = \left( \frac{\sum_{j=1}^{n} \bar{y}_n(j)_{TC}\, \bar{x}_n(j)_{TC}}{\sum_{j=1}^{n} \{\bar{x}_n(j)_{TC}\}^2} \right) \bar{X} \tag{1.13}$$

where

$$\bar{y}_n(j)_{TC} = \frac{n^g\, y_j - \bar{y}_n}{n^g - 1} \tag{1.14}$$

are called TRUMP Cuts (TC). The TC is obtained by calibrating the jth sampled observation $y_j$ by $n^g$, and then subtracting the sampled mean value $\bar{y}_n$. The quantity $g \neq 0$ is called the TRUMP Care Coefficient and its value depends on past experience (or otherwise), which we call the First Basic Information (FBI).

For example, if $g = -1$, then

$$\bar{y}_n(j)_{TC} = \frac{y_j - n\bar{y}_n}{1-n} = \frac{n\bar{y}_n - y_j}{n-1} = \bar{y}_n(j) \tag{1.15}$$

which leads to the usual jackknifing due to Quenouille (1956) and which was first used by Tukey (1958) to estimate the variance. Likewise, for the auxiliary variable, the TC is defined as

$$\bar{x}_n(j)_{TC} = \frac{n^g\, x_j - \bar{x}_n}{n^g - 1} \tag{1.16}$$

Singh and Sedory (2017a) considered the minimization of the model variance component of the variance of their proposed estimator $\bar{y}_{\text{TRUMP}}$ given by

$$V(\bar{y}_{\text{TRUMP}}) = \frac{\sigma^2}{n}\left(\frac{n^{2g+1}+1-2n^g}{(n^g-1)^2}\right)E_p\sum_{j=1}^{n}\left[\left\{(n-1)^2\overline{w}_n(j)-(n-2)\right\}^2\right]$$
$$+\frac{(1-2n^g)\sigma^2}{n(n^g-1)^2}E_p\left[\sum_{j\neq j'=1}^{n}\left\{(n-1)^2\overline{w}_n(j)-(n-2)\right\}\left\{(n-1)^2\overline{w}_n(j')-(n-2)\right\}\right] \qquad (1.17)$$

In the next section, we extend the work of Singh and Sedory (2017a) to the case of the Tuned Regression Unbiased Mean Predictor (TRUMP) and present only a few highlights of their detailed work in Singh and Sedory (2017b).

## 2. TRUMP: Tuned Regression Unbiased Mean Predictor

Consider a sample $s$ of $n$ observations taken by the simple random and with replacement (SRSWR) design where the observed values are $(y_i, x_i)$, $i = 1,2,...,n$. Following Singh, Sedory, Rueda, Arcos and Arnab (2016), we now consider a new estimator of the population mean $\bar{Y}$ defined as:

$$\bar{y}^*_{\text{TRUMP}} = \sum_{j\in s}\left[(n-1)^2\overline{w}_n(j)-(n-2)\right]\bar{y}_n(j)_{TC} \qquad (2.1)$$

where

$$\bar{y}_n(j)_{TC} = \frac{n^g\, y_j - \bar{y}_n}{n^g - 1} \qquad (2.2)$$

are called TRUMP Cuts (TC), and $g$ is called the TRUMP Care Coefficient, and its value depends on the availability of First Basic Information (FBI).

Assume the preliminary jackknifed weights are given by

$$\overline{w}_n(j) = \frac{1-w_j}{n-1} \qquad (2.3)$$

which we wish to tune further to get a better predictor of the population mean from the sample, $s$, and where the $w_j$ are any arbitrarily chosen weights, which may or may not be known. The linear model in (1.5) can be written as:

$$\frac{(n^g y_j - \bar{y}_n)}{n^g - 1} = \frac{\alpha(n^g-1)}{(n^g-1)} + \beta\left(\frac{n^g x_j - \bar{x}_n}{n^g-1}\right) + \frac{(n^g e_j - \bar{e}_n)}{(n^g-1)} \qquad (2.4)$$

or

$$\bar{y}_n(j)_{TC} = \alpha + \beta\, \bar{x}_n(j)_{TC} + \bar{e}_n(j)_{TC} \tag{2.5}$$

Under the model assumptions, we have

$$E_m\left(\bar{e}_n(j)_{TC} \mid x_j\right) = 0 \tag{2.6}$$

$$V_m\left(\bar{e}_n(j)_{TC} \mid x_j\right) = \frac{\sigma^2\left[n^{g+1}(n^{g+1}-2)v(x_j) + \sum_{j=1}^{n} v(x_j)\right]}{n^2(n^g-1)^2} \tag{2.7}$$

and

$$C_m\left(\bar{e}_n(j)_{TC}, \bar{e}_n'(j')_{TC} \mid x_j, x_{j'}\right) = \frac{\sigma^2}{n^2(n^g-1)^2}\left[\sum_{j=1}^{n} v(x_j) - n^{g+1}\{v(x_j) + v(x_{j'})\}\right] \tag{2.8}$$

The expressions in (2.7) and (2.8) clearly indicates that the TRUMP Care Coefficient "$g$" has some role in the estimation process, and consequently the First Basic Information (FBI) could again be helpful. Following Singh and Sedory (2017), finally we get the Tuned Regression Unbiased Mean Predictor (TRUMP) as:

$$\bar{y}^*_{\text{TRUMP}} = \alpha + \beta\, \bar{X} + \sum_{j=1}^{n}\left[(n-1)^2\, \bar{w}_n(j) - (n-2)\right]\bar{e}_n(j)_{TC} \tag{2.9}$$

The variance of the proposed tuned unbiased regression predictor $\bar{y}^*_{\text{TRUMP}}$ with the TRUMP Cuts model is given by

$$\begin{aligned} V(\bar{y}^*_{\text{TRUMP}}) &= \frac{\sigma^2}{n^2(n^g-1)^2} E_p \sum_{j=1}^{n} \Psi_j \left\{(n-1)^2 \bar{w}_n(j) - (n-2)\right\}^2 \\ &+ \frac{\sigma^2}{n^2(n^g-1)^2} E_p \sum_{j \neq j'=1}^{n} \Psi^{\circ}_{jj'} \left\{(n-1)^2 \bar{w}_n(j) - (n-2)\right\}\left\{(n-1)^2 \bar{w}_n(j') - (n-2)\right\} \end{aligned} \tag{2.10}$$

where

$$\Psi_j = \left\{n^{g+1}(n^{g+1}-2)v(x_j) + \sum_{j=1}^{n} v(x_j)\right\} \tag{2.11}$$

and

$$\Psi^{\circ}_{jj'} = \sum_{j=1}^{n} v(x_j) - n^{g+1}\{v(x_j) + v(x_{j'})\} \tag{2.12}$$

Now we suggest obtaining the new tuned weights in (2.1) such that the model variance component in (2.10) is minimum subject to the two tuned calibration constraints as used in the associated Lagrange's function :

$$L = \frac{\sigma^2}{n^2(n^g-1)^2}\sum_{j=1}^{n}\Psi_j\{(n-1)^2\overline{w}_n(j)-(n-2)\}^2$$
$$-2\lambda_1\left[\sum_{j=1}^{n}\overline{w}_n(j)-1\right]-2\lambda_2\left[\sum_{j=1}^{n}\overline{w}_n(j)\overline{x}_n(j)_{TC}-\frac{\overline{X}-n(2-n)\overline{x}_n}{(n-1)^2}\right] \quad (2.13)$$

On using the optimal values of $\lambda_1$ and $\lambda_2$, the TRUMP weights are then given by:

$$\overline{w}_n(j) = \frac{(n-2)}{(n-1)^2} + \frac{\left\{\frac{1}{\Psi_j}\sum_{j=1}^{n}\frac{\{\overline{x}_n(j)_{TC}\}^2}{\Psi_j} - \frac{\overline{x}_n(j)_{TC}}{\Psi_j}\sum_{j=1}^{n}\frac{\overline{x}_n(j)_{TC}}{\Psi_j}\right\}}{(n-1)^2\left[\left(\sum_{j=1}^{n}\frac{1}{\Psi_j}\right)\sum_{j=1}^{n}\frac{\{\overline{x}_n(j)_{TC}\}^2}{\Psi_j} - \left\{\sum_{j=1}^{n}\frac{\overline{x}_n(j)_{TC}}{\Psi_j}\right\}^2\right]}$$

$$+\frac{\overline{X}\left[\left(\sum_{j=1}^{n}\frac{1}{\Psi_j}\right)\left(\frac{\overline{x}_n(j)_{TC}}{\Psi_j}\right) - \frac{1}{\Psi_j}\sum_{j=1}^{n}\frac{\overline{x}_n(j)_{TC}}{\Psi_j}\right]}{(n-1)^2\left[\left(\sum_{j=1}^{n}\frac{1}{\Psi_j}\right)\sum_{j=1}^{n}\frac{\{\overline{x}_n(j)_{TC}\}^2}{\Psi_j} - \left\{\sum_{j=1}^{n}\frac{\overline{x}_n(j)_{TC}}{\Psi_j}\right\}^2\right]} \quad (2.14)$$

For this choice of TRUMP weights, the proposed Tuned Regression Unbiased Mean Predictor $\overline{y}^*_{\text{TRUMP}}$ in (2.1) becomes:

$$\overline{y}^*_{\text{TRUMP}} = \left(\frac{\sum_{j=1}^{n}\frac{\overline{y}_n(j)_{TC}}{\Psi_j}\sum_{j=1}^{n}\frac{\{\overline{x}_n(j)_{TC}\}^2}{\Psi_j} - \left(\sum_{j=1}^{n}\frac{\overline{x}_n(j)_{TC}}{\Psi_j}\right)\left(\sum_{j=1}^{n}\frac{\overline{y}_n(j)_{TC}\overline{x}_n(j)_{TC})}{\Psi_j}\right)}{\left(\sum_{j=1}^{n}\frac{1}{\Psi_j}\right)\left(\sum_{j=1}^{n}\frac{\{\overline{x}_n(j)_{TC}\}^2}{\Psi_j}\right) - \left(\sum_{j=1}^{n}\frac{\overline{x}_n(j)_{TC}}{\Psi_j}\right)^2}\right)$$
$$+\left(\frac{\left(\sum_{j=1}^{n}\frac{1}{\Psi_j}\right)\left(\sum_{j=1}^{n}\frac{\overline{y}_n(j)_{TC}\overline{x}_n(j)_{TC}}{\Psi_j}\right) - \left(\sum_{j=1}^{n}\frac{\overline{x}_n(j)_{TC}}{\Psi_j}\right)\left(\sum_{j=1}^{n}\frac{\overline{y}_n(j)_{TC}}{\Psi_j}\right)}{\left(\sum_{j=1}^{n}\frac{1}{\Psi_j}\right)\left(\sum_{j=1}^{n}\frac{(\overline{x}_n(j)_{TC})^2}{\Psi_j}\right) - \left(\sum_{j=1}^{n}\frac{\overline{x}_n(j)_{TC}}{\Psi_j}\right)^2}\right)\overline{X} \quad (2.15)$$

which is clearly a Tuned Regression Unbiased Mean Predictor (TRUMP) under the TRUMP Cuts model;

$$\overline{y}_n(j)_{TC} = \alpha + \beta\,\overline{x}_n(j)_{TC} + \overline{e}_n(j)_{TC} \quad (2.16)$$

The estimators of $\alpha$ and $\beta$ in the TRUMP Cut model (2.16) are, respectively, given by

$$\hat{\alpha}_{TC} = \frac{\sum_{j=1}^{n} \frac{\bar{y}_n(j)_{TC}}{\Psi_j} \sum_{j=1}^{n} \frac{\{\bar{x}_n(j)_{TC}\}^2}{\Psi_j} - \left( \sum_{j=1}^{n} \frac{\bar{x}_n(j)_{TC}}{\Psi_j} \right) \left( \sum_{j=1}^{n} \frac{\bar{x}_n(j)_{TC}\bar{y}_n(j)_{TC}}{\Psi_j} \right)}{\left( \sum_{j=1}^{n} \frac{1}{\Psi_j} \right) \left( \sum_{j=1}^{n} \frac{\{\bar{x}_n(j)_{TC}\}^2}{\Psi_j} \right) - \left( \sum_{j=1}^{n} \frac{\bar{x}_n(j)_{TC}}{\Psi_j} \right)^2} \tag{2.17}$$

and

$$\hat{\beta}_{TC} = \frac{\left( \sum_{j=1}^{n} \frac{1}{\Psi_j} \right) \left( \sum_{j=1}^{n} \frac{\bar{y}_n(j)_{TC}\bar{x}_n(j)_{TC}}{\Psi_j} \right) - \left( \sum_{j=1}^{n} \frac{\bar{x}_n(j)_{TC}}{\Psi_j} \right) \left( \sum_{j=1}^{n} \frac{\bar{y}_n(j)_{TC}}{\Psi_j} \right)}{\left( \sum_{j=1}^{n} \frac{1}{\Psi_j} \right) \left( \sum_{j=1}^{n} \frac{\{\bar{x}_n(j)_{TC}\}^2}{\Psi_j} \right) - \left( \sum_{j=1}^{n} \frac{\bar{x}_n(j)_{TC}}{\Psi_j} \right)^2} \tag{2.18}$$

which are obtained by minimizing weighted error sum of squares given by:

$$\text{WSSE} = \sum_{j=1}^{n} \Psi_j^{-1} \{\bar{e}_n(j)_{TC}\}^2 = \sum_{j=1}^{n} \Psi_j^{-1} [\bar{y}_n(j)_{TC} - \alpha - \beta \bar{x}_n(j)_{TC}]^2 \tag{2.19}$$

In other words, the proposed Tuned Regression Unbiased Mean Predictor (TRUMP), in case of heteroscedasticity, is given by

$$\bar{y}^*_{\text{TRUMP}} = \hat{\alpha}_{TC} + \hat{\beta}_{TC}\, \bar{X} \tag{2.20}$$

## 3. What is behind TRUMP?

Note that the TRUMP Cuts are linear transformations on both the dependent and independent variables. It is a bitter truth that in the presence of homoscedasticity the linear regression estimator, $\bar{y}_{reg}$, cannot be improved upon under the proposed TRUMP Cuts. It is fortunate that in the presence of heteroscedasticity, the estimator of $\alpha$ given by

$$\hat{\alpha} = \frac{\left( \sum_{i=1}^{n} \frac{y_i}{v(x_i)} \right) \left( \sum_{i=1}^{n} \frac{x_i^2}{v(x_i)} \right) - \left( \sum_{i=1}^{n} \frac{x_i}{v(x_i)} \right) \left( \sum_{i=1}^{n} \frac{x_i y_i}{v(x_i)} \right)}{\left( \sum_{i=1}^{n} \frac{1}{v(x_i)} \right) \left( \sum_{i=1}^{n} \frac{x_i^2}{v(x_i)} \right) - \left( \sum_{i=1}^{n} \frac{x_i}{v(x_i)} \right)^2} \tag{3.1}$$

is not invariant, although the estimator of the regression coefficient $\beta$ given by

$$\hat{\beta} = \frac{\left(\sum_{i=1}^{n} \frac{1}{v(x_i)}\right)\left(\sum_{i=1}^{n} \frac{x_i y_i}{v(x_i)}\right) - \left(\sum_{i=1}^{n} \frac{x_i}{v(x_i)}\right)\left(\sum_{i=1}^{n} \frac{y_i}{v(x_i)}\right)}{\left(\sum_{i=1}^{n} \frac{1}{v(x_i)}\right)\left(\sum_{i=1}^{n} \frac{x_i^2}{v(x_i)}\right) - \left(\sum_{i=1}^{n} \frac{x_i}{v(x_i)}\right)^2} \tag{3.2}$$

is invariant, under the proposed TRUMP cut transformation.

In particular the proposed TRUMP Cuts model is fortunate in the sense that the estimators $\hat{\alpha}_{TC}$ and $\hat{\beta}_{TC}$ can be written as:

$$\hat{\alpha}_{TC} = \frac{1}{(n^g - 1)}\left[n^g \hat{\alpha}^* - \left(\bar{y}_n - \hat{\beta}^* \bar{x}_n\right)\right] \tag{3.3}$$

with

$$\hat{\alpha}^* = \frac{\sum_{j=1}^{n} \frac{y_j}{\Psi_j} \sum_{j=1}^{n} \frac{x_j^2}{\Psi_j} - \left(\sum_{j=1}^{n} \frac{x_j}{\Psi_j}\right)\left(\sum_{j=1}^{n} \frac{x_j y_j}{\Psi_j}\right)}{\left(\sum_{j=1}^{n} \frac{1}{\Psi_j}\right)\left(\sum_{j=1}^{n} \frac{x_j^2}{\Psi_j}\right) - \left(\sum_{j=1}^{n} \frac{x_j}{\Psi_j}\right)^2} \tag{3.4}$$

and

$$\hat{\beta}_{TC} = \hat{\beta}^* = \frac{\left(\sum_{j=1}^{n} \frac{1}{\Psi_j}\right)\left(\sum_{j=1}^{n} \frac{y_j x_j}{\Psi_j}\right) - \left(\sum_{j=1}^{n} \frac{x_j}{\Psi_j}\right)\left(\sum_{j=1}^{n} \frac{y_j}{\Psi_j}\right)}{\left(\sum_{j=1}^{n} \frac{1}{\Psi_j}\right)\left(\sum_{j=1}^{n} \frac{x_j^2}{\Psi_j}\right) - \left(\sum_{j=1}^{n} \frac{x_j}{\Psi_j}\right)^2} \tag{3.5}$$

where $\hat{\alpha}^*$ and $\hat{\beta}^*$ are the estimators of $\alpha$ and $\beta$ obtained by minimizing the weighed sum of squares $(WSSE_1)$ defined as:

$$WSSE_1 = \sum_{j=1}^{n} \Psi_j^{-1}\left[y_j - \alpha - \beta x_j\right]^2 \tag{3.6}$$

On comparing (3.1) with (3.3) and (3.2) with (3.5) one can see that the estimators $\hat{\alpha}_{TC}$ differs from $\hat{\alpha}$ by two factors, the TRUMP Care Coefficient ( $g$ ) and $\Psi_j$, and $\hat{\beta}_{TC}$ differs from $\hat{\beta}$ only by factor $\Psi_j$. Recall that the value of $\Psi_j$ is given by

$$\Psi_j = n^{g+1}(n^{g+1} - 2)v(x_j) + \sum_{j=1}^{n} v(x_j) \tag{3.7}$$

Thus both the value of the TRUMP care coefficient $(g)$ and the heteroscedasticity (H) are playing a role in making TRUMP more efficient or less efficient. Further note that if $\hat{\alpha}^* = \bar{y}_n - \hat{\beta}^* \bar{x}_n$ (for $\Psi_j = \text{constant}$), then $\hat{\alpha}_{TC} = \hat{\alpha}^*$, that is $\hat{\alpha}_{TC}$ also becomes invariant under TRUMP Cuts.

If $v(x_j) = (n^g - 1)^2$ and $g = \ln(2)/\ln(n) - 1$, then $\Psi_j = \sigma^2/n$, that is the proposed TRUMP is as efficient as the linear regression estimator in the absence of such heteroscedasticity. In case of heteroscedasticity of the form below, the value of $\Psi_j$ depends on the value of the TRUMP Care Coefficient $g$. In particular, consider a heteroscedasticity function of the form:

$$v(x_j) = x_j^{(H+1)/2}, \tag{3.8}$$

where $H$ is called the Hillary Campaign Coefficient. It is clear that for a given value of $H$ the value of the TRUMP Care Coefficient $g$ could be adjusted such that the proposed TRUMP shows efficiency over the weighted least square regression estimator, and hence the First Basic Information (FBI) could be helpful.

A clever idea can be to set the value of TRUMP Care Coefficient as:

$$g = \frac{\log(2)}{\log(n)} - 1, \tag{3.9}$$

Such choice of the value of $g$ makes TRUMP model almost free from heteroscedasticity and almost as efficient as the linear regression estimator. Now for a given value of the Hillary Campaign Coefficient $H$, we adjust the value of TRUMP Care Coefficient $g$ such that the proposed TRUMP performs better. In other words, the TRUMP Care Coefficient could help in controlling the effect of Heteroscedasticity (i.e., the Hillary Campaign Effects) when estimating the y-intercept and the regression coefficient in the weighted linear regression model.

In the next section, we perform a simulation study to investigate the performance of the proposed TRUMP.

## 4. Which family is supporter of TRUMP?

To discover some families of distribution that support the use of TRUMP, we did a simulation study in which we generated bivariate data sets from the model:

$$y_i = m(x_i) + e_i\, v(x_i) \tag{4.1}$$

where the choice of $m(x_i)$ can form different types of models. Similar to Bredit, Opsomer and Sancheg-Borrego (2016), we consider three choices of mean functions which lead to the Linear, Bump and Jump models given by:

$$\text{LINEAR: } m(x_i) = 1 + 2(x_i - \mu_X) \tag{4.2}$$

$$\text{BUMP: } m(x_i) = 1 + 2(x_i - \mu_X) + e^{-200(x_i - \mu_X)^2} \tag{4.3}$$

and

$$\text{JUMP: } m(x_i) = 3 + 4(x_i - \mu_X) I_{x_i \le \mu_X} + 0.1.I_{x_i > \mu_X} \tag{4.4}$$

where $\mu_x = N^{-1}\sum_{i=1}^{N} x_i$ is the population mean of a variable generated from a Gamma distribution and $I_{x_i \le \mu_x}$ is an indicator function taking a value of 1 or 0 depending on whether $x_i \le \mu_x$ or not. Further we generated $x_i \sim G(a,b)$ and $e_i \sim N(0,1)$. We consider the three estimators in comparison, which we redefine as follows:

$$\hat{\theta}_0 = \bar{y}_n \quad \text{(Sample mean)} \tag{4.5}$$

$$\hat{\theta}_1 = \hat{\alpha} + \hat{\beta}\,\bar{X} \quad \text{(Weighted regression predictor)} \tag{4.6}$$

and

$$\hat{\theta}_2 = \hat{\alpha}_{TC} + \hat{\beta}_{TC}\,\bar{X} \quad \text{(TRUMP)} \tag{4.7}$$

Note that we have kept same naive control $\hat{\theta}_0 = \bar{y}_n$ while comparing the weighted least square predictor and the proposed TRUMP in various heteroscedastic cases.

For different sample sizes, $n$, we computed the percent relative efficiency of the jth predictor $\hat{\theta}_j$ over the sample mean predictor $\hat{\theta}_0$ as:

$$\mathrm{RE}(j) = \frac{\sum_{k=1}^{NITR}\left(\hat{\theta}_{0|k} - \bar{Y}\right)^2}{\sum_{k=1}^{NITR}\left(\hat{\theta}_{j|k} - \bar{Y}\right)^2} \times 100\% \tag{4.8}$$

where $NITR$ stands for the number of iterations.

The value of percent relative efficiency (RE) of the proposed TRUMP over the linear regression estimator is computed as

$$\mathrm{RE} = \frac{\sum_{k=1}^{NITR}\left(\hat{\theta}_{1|k} - \bar{Y}\right)^2}{\sum_{k=1}^{NITR}\left(\hat{\theta}_{2|k} - \bar{Y}\right)^2} \times 100\% \tag{4.9}$$

In the simulation study, we generated $N = 1050$ random values from the Gamma distribution with $a = 3.6$, and $b = 1.5$. Such a choice of parameters results in bivariate data with correlation coefficients values of $\rho_{xy}$ and $\rho_{m(x_i)y}$ depending on the nature of heteroscedasticity. Heteroscedasticity is determined by the value of the Hillary Campaign Coefficient ($H$). It is likely that the higher the value of $H$ in the function

$$v(x_i) = x_i^{\frac{(H+1)}{2}} \tag{4.10}$$

the lower the value of the correlation coefficient between the study variable, auxiliary variable and/or the mean function. The choice $H = -1$ makes the model homoscedastic.

From the population of $N = 1050$ units, we selected $NITR = 10{,}000$ samples each of sizes $n = 20$ (say) and then adjusted the value of TRUMP Care Coefficient ($g$) for a given value of $H$ from the three models, although we discuss only the linear model. In the linear model, we set the value of Hillary Campaign Coefficient ($H$) to be 0.5, 1.0, 1.5, 2.0, and 2.5, then using simulation study we found these values of the TRUMP Care Coefficient ($g$) such that the proposed TRUMP should be more efficient than the weighted linear regression estimator. For the case of linear model, the results are presented in Table 4.1.

**Table 4.1**. RE(1), RE(2) and TRUMP Care Coefficient $g$ values for the linear model.

| | | Mean | Std | Min | Max | Min | Max |
|---|---|---|---|---|---|---|---|
| $n$ | RE(1) | RE(2) | RE(2) | RE(2) | RE(2) | $g$ | $g$ |
| | $H = 0.5$, $\rho_{xy} = 0.77679$ | | | | | | |
| 20 | 241.04 | 244.85 | 2.84 | 241.45 | 249.94 | -0.4186 | -0.0686 |
| 25 | 239.42 | 243.65 | 2.85 | 240.11 | 248.07 | -0.4347 | -0.0347 |
| 30 | 247.54 | 252.01 | 3.18 | 248.26 | 257.97 | -0.4462 | -0.0462 |
| 35 | 247.23 | 251.44 | 3.15 | 247.76 | 258.22 | -0.5550 | -0.0550 |
| 40 | 246.57 | 250.48 | 3.17 | 246.70 | 257.52 | -0.6121 | -0.0621 |
| | $H = 1.0$, $\rho_{xy} = 0.68108$ | | | | | | |
| 20 | 185.32 | 185.32 | 0.00 | 185.32 | 185.32 | 1.3814 | 1.7314 |
| 25 | 183.56 | 183.56 | 0.00 | 183.56 | 183.56 | 1.2653 | 1.7153 |
| 30 | 189.11 | 189.11 | 0.00 | 189.11 | 189.11 | 1.1538 | 1.7038 |
| 35 | 187.98 | 187.98 | 0.00 | 187.98 | 187.98 | 1.1450 | 1.6950 |
| 40 | 187.81 | 187.81 | 0.00 | 187.81 | 187.81 | 0.9379 | 1.6879 |
| | $H = 1.5$, $\rho_{xy} = 0.56867$ | | | | | | |
| 20 | 158.04 | 160.87 | 5.63 | 158.08 | 183.50 | 0.0314 | 1.7314 |
| 25 | 156.11 | 158.35 | 4.60 | 156.14 | 178.74 | 0.0653 | 1.7153 |
| 30 | 160.25 | 162.76 | 5.43 | 160.27 | 187.37 | 0.0538 | 1.7038 |
| 35 | 158.86 | 161.53 | 6.10 | 158.87 | 189.81 | 0.0450 | 1.6950 |
| 40 | 159.16 | 162.00 | 6.68 | 159.18 | 193.19 | 0.0379 | 1.6879 |
| | $H = 2.0$, $\rho_{xy} = 0.45498$ | | | | | | |
| 20 | 151.62 | 157.30 | 9.76 | 151.73 | 193.58 | 0.0814 | 1.7314 |
| 25 | 149.39 | 155.02 | 10.10 | 149.47 | 191.73 | 0.0653 | 1.7153 |
| 30 | 153.02 | 158.76 | 10.12 | 153.08 | 192.00 | 0.0538 | 1.7038 |
| 35 | 151.60 | 157.22 | 9.98 | 151.66 | 194.18 | 0.0450 | 1.6950 |
| 40 | 152.49 | 157.75 | 9.87 | 152.53 | 198.30 | 0.0379 | 1.6879 |
| | $H = 2.5$, $\rho_{xy} = 0.35497$ | | | | | | |
| 20 | 162.27 | 172.68 | 15.33 | 162.51 | 222.62 | 0.0814 | 1.7314 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 25 | 159.59 | 169.11 | 15.26 | 159.74 | 222.32 | 0.0653 | 1.7153 |
| 30 | 163.47 | 173.30 | 16.34 | 163.61 | 227.10 | 0.1038 | 1.7038 |
| 35 | 162.18 | 172.17 | 17.15 | 162.30 | 228.67 | 0.0950 | 1.6950 |
| 40 | 163.91 | 174.12 | 17.74 | 164.01 | 230.27 | 0.0879 | 1.6879 |

If the value of the Hillary Campaign Coefficient is $H = 0.5$, then for a sample of 20 units, the value of RE(1) is 241.04, whereas the average value of RE(2) is 244.85 with a standard deviation of 2.84, the minimum value of RE(2) is 241.45 and maximum value of RE(2) is 249.94 as the value of the TRUMP Care Coefficient $g$ changes -0.4186 to -0.0686 with a step of 0.05. Thus the First Basic Information (FBI) about the value of Hillary Campaign Coefficient $H$ could help to adjust the TRUMP Care Coefficient $g$ such that the proposed TRUMP can perform better. It may be worth pointing out that in case of the linear model, the minimum value of RE(2) remains higher than the RE(1) value for all choices of the sample sizes taken and all values of $H$ and $g$ considered. In other words, if the study and auxiliary variables are following a linear trend but have heteroscedasticity then the proposed TRUMP is recommended to search for a TRUMP Care Coefficient that would lead to efficient results. Figure 4.1 shows the behavior of RE(1) and RE(2) for different values of $H$ in the range 0.5 to 2.5 with a step of 0.5 where varying the sample size and the value of the TRUMP care coefficient. The small vertical lines in the left panel of Figure 4.1 are due to changes in sample size for a given value of $H$. The tall vertical lines in the right panel of the Figure 4.1 are consequences of change both the sample size and the value of the TRUMP care coefficient. The cause of reduction in RE(1) value as the value of $H$ increases from 0.5 to 2.0, may or may not be true only to the value of the correlation coefficient $\rho_{xy}$ between the study variable and the auxiliary variable. In case of linear model, $\rho_{xy} = \rho_{m(x_i)y}$ where $m(x_i)$ is the mean function for the linear model.

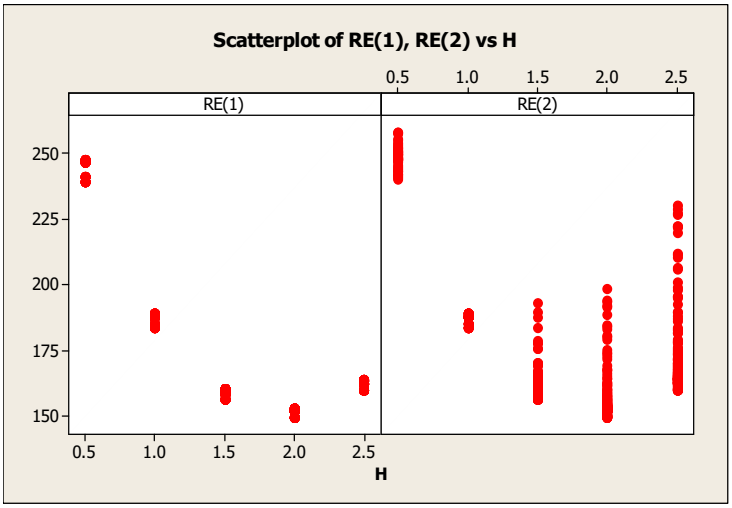


**Fig.4.1.** Effect of Hillary Campaign Coefficient with linear model

A summary of the RE values obtained for different values of $H$ and $n$ are reported in Table 4.2.

**Table 4.2**. RE values for the linear model.

| $n$ | Mean | Std | Min | Max |
|---|---|---|---|---|
| | $H = 0.5$ | | | |
| 20 | 101.58 | 1.18 | 100.17 | 103.69 |
| 25 | 101.77 | 1.19 | 100.29 | 103.61 |
| 30 | 101.81 | 1.29 | 100.29 | 104.21 |
| 35 | 101.70 | 1.27 | 100.21 | 104.45 |
| 40 | 101.59 | 1.29 | 100.05 | 104.44 |
| | $H = 1.0$ | | | |
| 20 | 100.00 | 0.00 | 100.00 | 100.00 |
| 25 | 100.00 | 0.00 | 100.00 | 100.00 |
| 30 | 100.00 | 0.00 | 100.00 | 100.00 |
| 35 | 100.00 | 0.00 | 100.00 | 100.00 |
| 40 | 100.00 | 0.00 | 100.00 | 100.00 |
| | $H = 1.5$ | | | |
| 20 | 101.79 | 3.56 | 100.02 | 116.11 |
| 25 | 101.43 | 2.94 | 100.02 | 114.49 |
| 30 | 101.56 | 3.39 | 100.01 | 116.93 |
| 35 | 101.68 | 3.84 | 100.01 | 119.48 |
| 40 | 101.78 | 4.20 | 100.01 | 121.38 |
| | $H = 2.0$ | | | |
| 20 | 103.75 | 6.44 | 100.07 | 127.68 |
| 25 | 103.77 | 6.76 | 100.05 | 128.35 |
| 30 | 103.75 | 6.61 | 100.04 | 125.61 |
| 35 | 103.70 | 6.59 | 100.03 | 128.08 |
| 40 | 103.45 | 6.47 | 100.03 | 130.04 |
| | $H = 2.5$ | | | |
| 20 | 106.41 | 9.45 | 100.15 | 137.19 |
| 25 | 105.98 | 9.56 | 100.11 | 139.32 |
| 30 | 106.02 | 9.99 | 100.09 | 138.93 |
| 35 | 106.16 | 10.57 | 100.07 | 140.99 |
| 40 | 106.23 | 10.82 | 100.06 | 140.48 |

From Figure 4.2 and Figure 4.3, for value of $g$ between 0 and 1 shows that applying the proposed TRUMP in the presence of heteroscedasticity could be helpful. From Table 4.2, one can see that in the case of linear trend, if the value of $H = 1$, which means $v(x_i) = x_i$, then the proposed TRUMP and the weighted linear regression are equally efficient for certain range of $g$. As soon as the value of $H$ becomes higher, then RE values show an increasing trend.

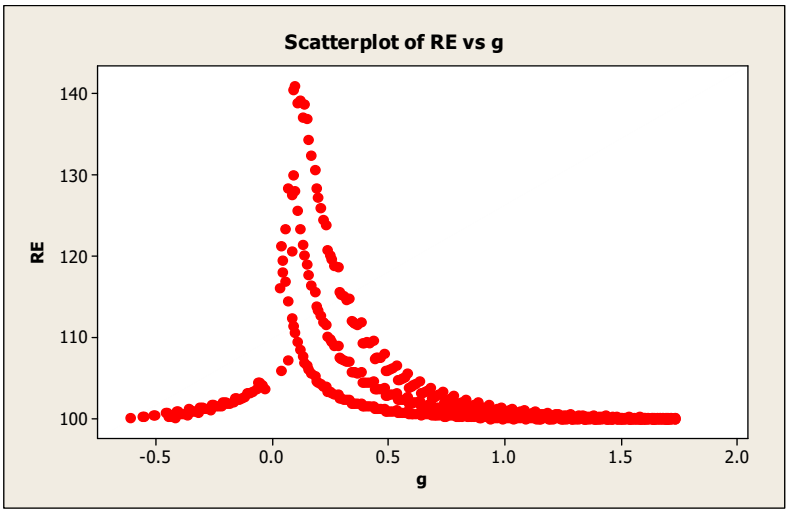


**Fig. 4.2.** Effect of TRUMP Care Coefficient with linear model

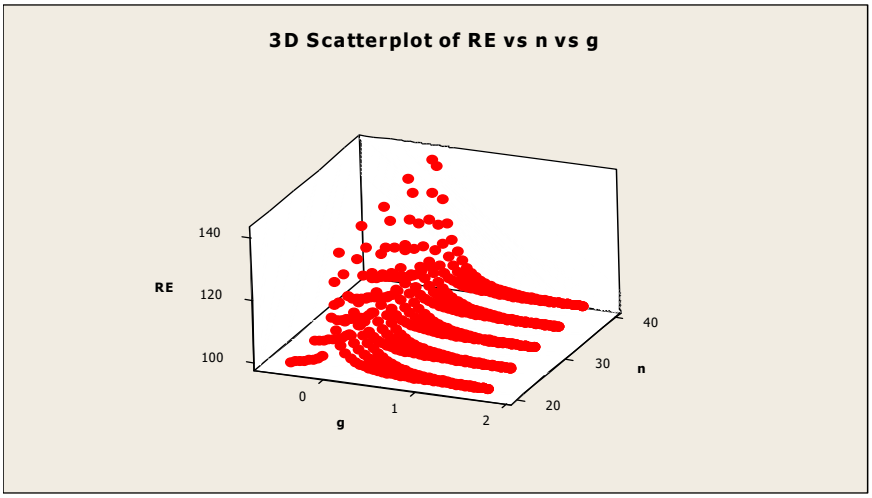


**Fig. 4.3.** Effect of TRUMP Care Coefficient and sample size with linear model

Similar results are observed for in case of a Bump Model and a Jump Model.

**Remark 4.1**. Adjustment of $(\frac{H+1}{2})$ has been made such that there should not be too much heteroscedasticity in a population, but one could always change to other function, such as:

$$v(x_i) = x_i^H, \quad v(x_i) = [\log(x_i)]^H \text{ and } v(x_i) = |x_i - \mu_x|^H \text{ etc.}$$

which we will be exploring in future studies.

In the next section, we consider the study of Tuned Regression Unbiased Mean Predictor (TRUMP) with Chain Type TRUMP Cuts.

## 5. TRUMP with Chain-Type TRUMP Cuts

Here in brief, we introduce the Chain-Type TRUMP Cuts as follows. The first-TRUMP Cuts based on the $j_1$th unit are defined as:

$$\bar{y}_n(j_1)_{TC} = \frac{n^{g_1} y_{j_1} - \bar{y}_n}{n^{g_1} - 1} \tag{5.1}$$

where $g_1 \neq 0$ is now called the first-TRUMP Care coefficient.

Recall that if $g_1 = -1$ then the first-TRUMP Cuts take the form

$$\bar{y}_n(j_1)_{TC} = \frac{n\bar{y}_n - y_{j_1}}{n-1} = \bar{y}_n(j_1) \tag{5.2}$$

which is called the method of jackknifing due to Quenouille (1956).

Now we define second-TRUMP Cuts based on the $j_1$th and $j_2(\neq j_1)$th units as:

$$\bar{y}_n(j_2 \mid j_1)_{TC} = \frac{(n-1)^{g_2} y_{j_2} - \bar{y}_n(j_1)_{TC}}{(n-1)^{g_2} - 1} \tag{5.3}$$

where $g_2 \neq 0$ is now called the second-TRUMP Care coefficient.

Now if $g_1 = g_2 = -1$, then

$$\bar{y}_n(j_2 \mid j_1)_{TC} = \frac{n\bar{y}_n - y_{j_2} - y_{j_1}}{(n-2)} = \bar{y}_n(j_2 \mid j_1) \tag{5.4}$$

which is again the method of jackknifing two distinct units $j_1$ and $j_2$ from a sample due to Quenouille (1956).

Under Chain-Type TRUMP Cuts, it is shown in Singh and Sedory (2017b) that the proposed Second Tuned Regression Unbiased Mean Predictor $\bar{y}_{\text{TRUMP}}^{Second*}$ takes the form:

$$\bar{y}_{\text{TRUMP}}^{Second*} =$$

$$\frac{\left(\sum_{j_1=1}^{n}\sum_{j_2\neq j_1=1}^{n} \frac{\bar{y}_n(j_2 \mid j_1)_{TC}}{\Psi(j_2 \mid j_1)}\right)\sum_{j_1=1}^{n}\sum_{j_2\neq j_1=1}^{n} \frac{\{\bar{x}_n(j_2 \mid j_1)_{TC}\}^2}{\Psi(j_2 \mid j_1)} - \left(\sum_{j_1=1}^{n}\sum_{j_2\neq j_1=1}^{n} \frac{\bar{x}_n(j_2 \mid j_1)_{TC}}{\Psi(j_2 \mid j_1)}\right)\left(\sum_{j_1=1}^{n}\sum_{j_2\neq j_1=1}^{n} \frac{\bar{y}_n(j_2 \mid j_1)_{TC}\bar{x}_n(j_2 \mid j_1)_{TC}}{\Psi(j_2 \mid j_1)}\right)}{\left(\sum_{j_1=1}^{n}\sum_{j_2\neq j_1=1}^{n} \frac{1}{\Psi(j_2 \mid j_1)}\right)\left(\sum_{j_1=1}^{n}\sum_{j_2\neq j_1=1}^{n} \frac{\{\bar{x}_n(j_2 \mid j_1)\}^2}{\Psi(j_2 \mid j_1)}\right) - \left(\sum_{j_1=1}^{n}\sum_{j_2\neq j_1=1}^{n} \frac{\bar{x}_n(j_2 \mid j_1)}{\Psi(j_2 \mid j_1)}\right)^2}$$

$$+\left(\frac{\left(\sum_{j_1=1}^{n}\sum_{j_2\neq j_1=1}^{n} \frac{1}{\Psi(j_2 \mid j_1)}\right)\left(\sum_{j_1=1}^{n}\sum_{j_2\neq j_1=1}^{n} \frac{\bar{y}_n(j_2 \mid j_1)_{TC}\bar{x}_n(j_2 \mid j_1)_{TC}}{\Psi(j_2 \mid j_1)}\right) - \left(\sum_{j_1=1}^{n}\sum_{j_2\neq j_1=1}^{n} \frac{\bar{x}_n(j_2 \mid j_1)_{TC}}{\Psi(j_2 \mid j_1)}\right)\left(\sum_{j_1=1}^{n}\sum_{j_2\neq j_1=1}^{n} \frac{\bar{y}_n(j_2 \mid j_1)_{TC}}{\Psi(j_2 \mid j_1)}\right)}{\left(\sum_{j_1=1}^{n}\sum_{j_2\neq j_1=1}^{n} \frac{1}{\Psi(j_2 \mid j_1)}\right)\left(\sum_{j_1=1}^{n}\sum_{j_2\neq j_1=1}^{n} \frac{\{\bar{x}_n(j_2 \mid j_1)_{TC}\}^2}{\Psi(j_2 \mid j_1)}\right) - \left(\sum_{j_1=1}^{n}\sum_{j_2\neq j_1=1}^{n} \frac{\bar{x}_n(j_2 \mid j_1)_{TC}}{\Psi(j_2 \mid j_1)}\right)^2}\right)\bar{X} \tag{5.5}$$

where $\Psi(j_2 \mid j_1)$ can be had from Singh and Sedory (2017b).

In the next section, we would be interested in working on the issue of whether creating chain-type TRUMP Cuts is more painful than useful.

### 5.1 Are Chain-Type TRUMP Cuts Painful?

The second-TRUMP Cuts estimator $\bar{y}_{TRUMP}^{Second^*}$, in addition to depending on the values of two TRUMP Care coefficients $g_1$ and $g_2$, also depends upon $\Psi(j_2 \mid j_1) = f(v(x_{j_1}), v(x_{j_2}))$ which is a joint function of the Hillary Campaign Coefficients. There could be two different **H**illary Campaign Coefficient $H_1$ and $H_2$ such as

$$v(x_{j_1}) = x_{j_1}^{(H_1+1)/2} \text{ and } v(x_{j_2}) = x_{j_2}^{(H_2+1)/2} \tag{5.6}$$

Thus the second-TRUMP Cuts model (Singh and Sedory, 2017b) has the flexibility of making use of FBI about two Hillary Campaign Coefficients, and of working efficiently. For simplicity, keeping $H_1 = H_2 = H$, there is still the flexibility of adjusting two TRUMP Care coefficients $g_1$ and $g_2$, yet the computational work may be painful and time consuming for the computer when doing simulations. Searching for a pair of TRUMP Care coefficients $g_1$ and $g_2$ that make for a more efficient estimator can be a daunting enterprise.

### 5.2 Does performance Boom with Chain-Type TRUMP Cuts?

For demonstration purposes, we kept the same data set as in the case of Jump Model with $H = 0.5$, $n = 20$, $\rho_{xy} = 0.60983$, and $\rho_{m(x)y} = 0.75008$. To make things clear to the reader, we computed the percent relative efficiency of the BLUP:

$$\hat{\theta}_1^* = \hat{\alpha} + \hat{\beta}\,\bar{X} \quad \text{(Weighted regression predictor)} \tag{5.7}$$

and second-TRUMP Cuts predictor,

$$\hat{\theta}_2^* = \bar{y}_{TRUMP}^{Second*} \tag{5.8}$$

with respect to the sample mean predictor

$$\hat{\theta}_0^* = \bar{y}_n \tag{5.9}$$

as:

$$\text{RE}(\hat{\theta}_j^*) = \frac{\sum_{k=1}^{NITR}\left(\hat{\theta}_{0|k}^* - \bar{\bar{Y}}\right)^2}{\sum_{k=1}^{NITR}\left(\hat{\theta}_{j|k}^* - \bar{\bar{Y}}\right)^2} \times 100\% = RE(j), \text{ say} \tag{5.10}$$

where $NITR = 10{,}000$ denotes the number of iterations. Note that the value of $RE(1)$ is free from the values of the TRUMP Care coefficients $g_1$ and $g_2$, and in this study its computed value remains 129.42%. The value of $RE(2)$ changes in the same range from 129.58% to 135.96% for the fixed choice of $g_1$ as the value $g_2$ changes. A pictorial presentation of the results for four choices of $g_1 = -6.5$, -5.5, -4.5 and -3.5, and for a searched range of $g_2$ from -4.0 to -1.30, and retaining only those cases where $RE(2) > RE(1)$, is shown in Figure 5.1.

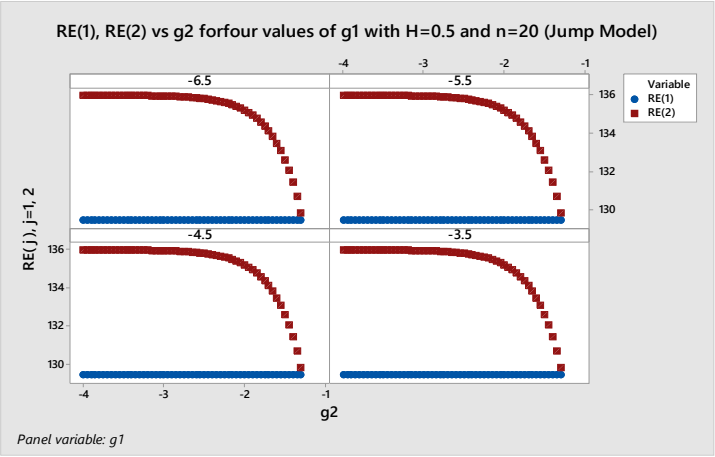


**Fig. 5.1** RE(1) values and RE(2) values.

A close inspection of Figure 5.1 reveals that it is not an illusion that the percent relative efficiency may "boom" where one has the flexibility of using multiple TRUMP Care coefficients. Thus use of additional TRUMP Care Coefficients seems helpful in practice. It may be worth pointing out here that a carefully chosen single TRUMP Care coefficient can also perform very well.

## Acknowledgements

At the end, we acknowledge the use of R Core Team (2012) package in the simulation study. All opinions are authors' own, and do not represent any institute or organization.

## References

Breidt, F. J., Opsomer, J.D. and Sanchez-Borrego, I. (2016). Nonparametric variance estimation under fine stratification: An alternative to collapsed strata. *J. Amer. Statist. Assoc.*, 111 (514), 822-833.

Hansen, M. H., Hurwitz, W. N. and Madow, W. G. (1953). *Sample Survey Methods and Theory*. New York, John Wiley and Sons, 456--464.

Quenouille, M.H. (1956). Notes on bias in estimation. *Biometrika*, 43, 353-360.

R Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.ISBN 3-900051-07-0, URL http://www.R-project.org/

Singh, S. and Sedory, S.A. (2017a). TRUMP: Tuned Ratio Unbiased Mean Predictor. *Presented at the Joint Statistical Meeting 2017, Baltimore, Maryland, USA.*

Singh, S. and Sedory, S.A. (2017b). TRUMP: Tuned Ratio Unbiased Mean Predictor. *Working monograph.*

Singh, S., Sedory, S.A., Rueda, M.M, Arcos, A., and Arnab R. (2016). *A new concept for tuning design weights in survey sampling: Jackknifing in Theory and Practice*. Elsevier: London.

Tukey, J.W. (1958). Bias and confidence in not-quite large samples (abstract). *Ann. Math. Statist*., 29, 61-75.