IOSR Journal of Mathematics (IOSR-JM)
e-ISSN: 2278-5728, p-ISSN: 2319-765X. Volume 14, Issue 1 Ver. II (Jan. - Feb. 2018), PP 41-45
www.iosrjournals.org

Efficiency of Robust Regression Model in Sampling Designs Using Weighted Averages

Edwin Ayora[1], Romanus Odhiambo[2], George Orwa[3]
Researcher,Jomo Kenyatta University of Agriculture and Technology, P.O Box
Professor of statistics,Jomo Kenyatta University of Agriculture and Technology, P.O Box
Doctor of statistics,Jomo Kenyatta University of Agriculture and Technology, P.O Box
Corresponding Author: Edwin Ayora

Abstract: *A class of survey weighting methods provides consistent estimation of regression coefficients under unequal probability sampling. The minimization of the variance of the estimated coefficients within this class is considered. A series of approximations leads to a simple modification of the usual design weights. One type of application where unequal probabilities of selection arise is in the cross-national comparative surveys. In this case, the paper suggests the use of certain kind of within-country weight. This idea is investigated in an application to the data from African Social Survey, where a robust regression model is fitted with vote in an election as a dependent variable and with various variables of political science interest included as explanatory variables. The result shows that the use of the modified weights leads to considerable reduction in standard errors compared to design weights. Since robust regression model is unbiased, consistency then it follows the cramar Rao inequality assumption that variance is less in the estimated coefficients thus its most efficiency in measuring weights for within country data.*

Keywords: *Effeciency, variance, weighted averages, robust regression model.*

Date of Submission: 26-01-2018 Date of acceptance: 13-02-2018

I. Introduction

The rationale for the use of simple survey weights in a least square regression analysis is examined with respect to increasingly general specification of the population regression model. The appropriateness of the weighted regression estimate depends on which model is chosen. Suppose that a sample survey measures $p + 1$ variables on each of n individuals, so that the data consist of the $n \times 1$ matrix Y and the $n \times p$ matrix Xthen the least square estimates of the regression coefficients of Y on X is

$$\hat{\beta} = (X'X)^{-1}X'Y$$

However, the rows of Y and X often are not a simple random sample from the population. Differential sampling rates and differential response rates among various strata lead to different probabilities of selection for each individual. Kish (1965), discusses the computation of these probabilities for various sampling schemes, the differential sampling and response rates lead to the computation of weights for each case which attempts to give each sampling scheme the same relative importance in the sample that it has in the population.

The ordinary least squares estimator is a common choice of researchers, but under an informative design, the ordinary least square estimator is biased. The probability weighted estimator is consistent but may have a large variance. In a preliminary testing procedures, the paper tests for the importance of weights in estimation. If the null hypopaper is accepted, then the use of unweighted estimator is used. The paper incorporates the design weights into the estimation procedure.

In a simple random sample from the population, on unbiased estimator of the population parameter is the OLS, and an estimator of its variance is easy to calculate. However, in many surveys, the elements enter the sample with unequal probabilities. These weights are used to construct the probability weighted estimator. In more complex analyses such as regression the weighted estimator not only requires a more complicated calculation, but also often gives a larger variance than the unweighted version of the estimator.

Preliminary testing (pre-test) procedure are procedures in which a test of a model assumption is used to decide between two estimation procedures. Bancroft (1944), Huntsberger (1955) and Mosteller (1948) provide details about pre-test procedures. The pre-test procedure is characterized by a test statistic, T, calculated from the data set. The test T serves the purpose of determining the estimation method. If T is statistically significant at some significance level, chosen a priori, a given procedure will be used to estimate the parameter. Otherwise an alternative procedure will be used for calculating the parameter estimator.

II. Methodology

The outcome variable of interest yj is whether the respondent voted in the last national election held in their country. Electoral turnout is in decline in Africa, as elsewhere, and political scientists are interested in factors associated with turnout. Low turnout is of particular concern amongst young people and the analysed data for those aged 18-24, as in Fieldhouse et al. (2007)

III. Results and discussions

Y_j are independent, with a distribution depending on a k × 1 column vector θ of parameters, such that

$$E_m\left(\phi_j\left(Y_j; x_j, \theta\right)\right) = 0 \; for j = 1, \ldots, N$$

Where $\phi_j\left(Y_j; x_j, \theta\right)$ is a $k \times 1$ vector estimating function and $E_m(.)$ denotes expectation under the model.
The population level equation,
$\sum_{j=1}^{N} \varphi_{\boldsymbol{j}}(Y_{\boldsymbol{j}}; x_{\boldsymbol{j}}, \theta) = 0$
are unbiased estimating equation. Assuming that in some asymptotic framework a solution θ_U to these equation eventually exists uniquelly and under additional regularity conditions (Godambe and Thompson, 2009), θ_U converges in probability to θ.
A particular instance of such an estimating function is the unit level score function given by
$\theta_j\left(Y_j; x_j, \theta\right) = \log f_j(Y_j; x_j, \theta)$
where $f_j(Y_j; x_j, \theta)$ is the probability density or mass function for Y_j and θ_U is the "census" maximum likelihood estimator of θ which would apply if all populationvalues of y_j, x_j were observed.

For illustration, if a binary variable y_j, , taking values 0 and 1, obeys a logistic regression coefficients, it will result to
$\log \frac{f_j(1; x_j, \theta)}{f_j(0; x_j, \theta)} = x_j \theta$

$$\phi_j\left(Y_j; x_j, \theta\right) = Y_j - f_j(1; x_j, \theta) x_j^T$$

Where Tdenotes the transpose. Now, suppose that the y_j, x_j are only observed for units $j$in a sample drawn by a probability sampling scheme from U and let I_j , $j = 1, \ldots\ldots\ldots N$ be the sample indicators, where$I_j = 1$ if unit j is sampled and $I_j = 0$ if not. This research will be interested in the weighted estimator$\hat{\theta}_w$ which solves the sample estimating equations.
$\sum_{j=1}^{N} W_{\boldsymbol{j}} I_{\boldsymbol{j}} \phi_j(Y_{\boldsymbol{j}}; x_{\boldsymbol{j}}, \theta) = 0$
where w_j is a survey weight corresponding to condition 1.1 for the consistency of θu, these estimating equations are unbiased under the joint distribution induced by the design and the model$\hat{\theta}_w$ is consistent for θ if
$E_m E_p\left[W_j I_j \phi_j\left(Y_j; x_j, \theta\right)\right] = 0 \; for j = 1, \ldots, N$
Where $E_p(.)$ denotes expectation with respect to the sampling scheme. Two basic cases when condition 1.7 holds are

1. The w_j are constant, so that $\hat{\theta}_w$ is the unweighted estimator, and sampling is non-informative, that is $I_j \; and Y_j$ are independent (conditional on x_j) for eachj. This arises, in particular, when sample inclusion depends just on a set of design variables which are included in the vector x_j . Fuller (2009) reviews tests of this non-informative condition, including a test proposed by DuMouchel and Duncan (1983).

2. $w_j = d_j$ the design (Horvitz-Thompson) weight given by $d_j = \pi_j^{-1}$, where $\pi_j = E_p\left(I_j\right)$ is the inclussion probability of unit j. In each of these cases, the proof of 1.7 assumes model 1 holds. For example, to demonstrate that 1.7 holds in case 2 we write
$E_m E_p\left(d_j I_j \theta_j\right) = E_m(E_p\left(d_j I_j \theta_j\right)$ Where
$\phi_j = \phi_j\left(Y_j; x_j, \theta\right)$, and $E_p\left[d_j I_j\right] = 1$
Following a similar argument, (1.7) holds for the class of cases, generating (ii), defined by;
3. $w_j = d_j q_j$ Where$q_j = q\left(x_j\right)$ and $q(.)$ is an arbitrary function.
The class of weighted estimators defined by such weights is the one of primary interest in this paper and within which we consider minimizing the variance of linear combinations of the elements of $\hat{\theta}_w$ is
$Var_{mp}(\hat{\theta}_W) = J(\theta)^{-1} Var_{mp}(\sum_{j=1}^{N} W_{\boldsymbol{j}} I_{\boldsymbol{j}} \theta_{\boldsymbol{j}}) J(\theta)^{-1}$

Where$J(\theta) = E_m(\sum_{j=1}^{N} w_j \; I_j \; \frac{d\theta_j}{d\theta})$ and, when $w_j = d_j \, q_{xj}$, we can write,

$J(\theta) = \sum_{j=1}^{N} q_j \; E_m(\frac{d\theta_j}{d\theta})$

We would like to choose q_j so that the variance in (1.8) is minimized. For practical purposes, we consider it sufficient to minimize an approximation to this variance, since any weighted estimator in the class defined by $w_j = d_j q_j$ is consistent. We shall make a series of approximations to enable us to specify q_j as employed by Fuller (2009) of assuming Poisson sampling. Under this approximation, we may rewrite (1.8) as

$Var_{mp}(\hat{\theta}_w) \approx J(\theta)^{-1}\left[\sum_{j=1}^{N} Var_{mp}(d_j\ q_j \phi_j)\right]J(\theta)^{-1}$
Furthermore, we have

$$Var_{mp}(d_j q_j I_j \theta_j) = E_m\left[Var_p(d_j q_j I_j \theta_j)\right] + Var_m\left[E_p(d_j q_j I_j \theta_j)\right] = E_m\left[(d_j - 1)q_j^2 \theta_j \theta_j^T\right] + Var_m(q_j \theta_j)$$
$$= E_m(d_j q_j^2 \theta_j \theta_j^T)$$

Hence, when $w_j = d_j q(x_j)$ the asymptotic covariance matrix can be expressed as

$$Var_{mp}(\hat{\theta}_W) \approx [\sum_{j=1}^{N} q_j\ E_m(\frac{d\theta_j}{d\theta})]^{-1} \sum_{j=1}^{N} q_j^2\ E_m(d_j \theta_j \theta_j^T)\ [\sum_{j=1}^{N} q_j\ E_m(\frac{d\theta_j}{d\theta})]^{-1}$$

As a second simplification, we assume that $\phi_j(Y_{\boldsymbol{j}}; x_{\boldsymbol{j}}, \theta)$ is a score function so that
$E_m\left(\theta_j \theta_j^T\right) = -E_m\left(\frac{d\theta_j}{d\theta}\right) = H_j$, say, and also that we have a generalized linear model so that
$\theta = \beta$ is the vector of regression coefficients and $\theta_j\left(Y_j; x_j, \beta\right) = \lambda_j\ (Y_j; x_j, \beta) x_j^T$
where $\lambda_j(.)$ is a scalar function. Then we may write

$H_j = T_j^2 x_j^T x_j$, where $T_j^2 = E_m\left(\lambda_j^2\right)$, $\quad \lambda_i = \lambda_i\ (Y_j; x_j \beta)$ and

$$Var_{mp}(\hat{\beta}_w) \approx [\sum_{j=1}^{N} q_j\ T_j^2 x_j^T x_j]^{-1} \sum_{j=1}^{N} q_j^2\ v_j x_j^T x_j [\sum_{j=1}^{N} q_j T_j^2\ x_j]^{-1}$$

Where $V_j = E_m\left(\lambda_j \lambda_j^2\right)$. By analogy to the Gauss- Markov Theorem, the choice of q_j which minimizes the variance given by (1.9) of any linear combination of the elements of $\hat{\beta}_w$ is
$q_j^{opt} = q^{opt}(x_j) \alpha T_j^2 / V_j = E_m(d_j\ \lambda_j^2 / x_j)$
This generalizes an argument used by Fuller (2009) for the special case of heteroscedastic normal error linear regression with $k = 1$. We make the conditioning of x_j explicit on the right hand side of (1.10) to be clear that q^{opt} depends x_j . The quantity on the right hand side of (1.10) is not observed, but is estimable from auxiliary regressions of λ_j^2 and $d_j \lambda_j^2$ on x_j , where $\hat{\lambda}_j = \lambda_j\left(Y_j; x_j \hat{\beta}\right)$ and λ is a consistent estimator of β. These regressions and estimation of β could, for example, employ design weighted estimation. We do not pursue this idea further in this paper, however. Rather, we make the further approximation that d_j is uncorrelated with λ_j^2 (given x_j) so that expression (1.10) simplifies to
$q_j^{opt} \alpha \frac{1}{E_m(d_j / x_j)}$
The form of weighting in (1.11) is similar to semi parametric approach of pfeffermann and Sverchkov (1999), although they propose to take

$$q_j \alpha \frac{1}{E_{mp}(d_j / x_j\ , I_j = 1)}$$

Expression (1.11) can be yet further simplified by replacing $E_m\left(d_j / x_j\right)$ by the conditional expectation of d_j given a subset of the explanatory variable making up x_j . In practice, there is often just a single explanatory factor which is the determinant. Sources of variation in the π_j . In our cross sectional application, this is the country factor, i.e. acategorical variable with categories corresponding to countries. In this case, we may simplify set q_j to be equal within the categories of this factor and for a given category, to be reciprocal of the design weighted mean of d_j for sample units in the category. In the more general setting $E_m\left(d_j / x_j\right)$ in (1.11) may be estimted by design weighted regression.
Turning to standard error estimation and assuming that the finite population correction can be ignored, the asymptotic covariance matrix of $\hat{\theta}_w$ may be estimated consistently (Fuller, 2009) by $\hat{J}^{-1}\hat{V}\hat{J}^{-1}$, where $\hat{J} = \sum_{j=1}^{N} w_j I_j \frac{d\pi_j}{d\theta}$ evaluated to $\theta = \hat{\theta}_w$ and $\hat{V}$ is a consistent estimator of the covariance matrix of the Horvitz-Thompson estimator.

$\sum_{j=1}^{N} d_j\, I_j u_j$, Where $u_j = d_j^{-1} w_j \Pi_j \left(y_j, x_j, , \hat{\theta}\mathrm{w}\right)$is treated as a fixed vector of variables.
Thus, standard errors can be produced using a standard approach for fixed survey weights, ignoring the fact that the weights have been modified. In all of this paper it has been assumed that the model in (1) is correct. Godambe and Thompson (1986) argue that θ_u defined by (2) may still be of interest even if the model fails. They note that $\hat{\theta}_w$ is still design-Consistent forθ_u even under the model misspecification when design weights are used and they also demonstrate a minimum variance property of design weighting under the constraint that θ_u is estimated consistently. However, if model (1) fails, then θ_u, is not the only finite population parameter which can be defined and may be of interest. An arbitrary finite population parameter can be defined as the value of the parameter θ which indexes that version of the model which represents a good fit to the finite population values $y_j; x_j; j \in \mathrm{U}$ according to a specified criterion, whatever the truth of the model. The criterion for θ_u is that (2)holds. To consider an alternative finite population parameter, suppose the weights are of the form$w_j = d_j q_j$. Then $\hat{\theta}_w$ is consistent for the solution of
$\sum_{j=1}^{N} q_j\, \Pi_j (y_j; x_j, \theta)$=0

Assumed to exist uniquely, and this will not in general be the same as θ_u.Nevertheless, it is finite population parameter with (1.10) as the criterion and it is defined even if the model (1) fails. We suggest that whether the solution of (2) or (1.10) is of scientific interest depends on the application.

The modified weighting approach introduced will be used to analyse data from a particular cross-national survey. The basic setting where regression analysis is applied is data from several countries and where country is an explanatory variable in the model that is binary indicators of the different countries from part of x_i, such analyses have various purposes. One is to enable a quasi experimental evaluation ofthe relative impacts of different policies which are adopted in different countries, for example to compare the effects of different national tobacco control policies on smoker behaviour. Another is to enable a replication of some phenomenon of interest such as an election in the application in this paper. Cross national analysis may have broad comparative purposes, enabling the comparison of regression relationships across different national setting.

Sampling designs for cross national surveys are often subject to considerable variation in inclusion probabilities between countries since their principal aim is often comparative, it is common to set a minimum sample size or effective sample size in each country in order to achieve adequate precision of each national estimate. Population sizes of countries have a tremendous thousand foldrange; whereas sample sizes tend to be made more constant in order to obtain similar errors for national means. As a consequence, sampling fractions can vary greatly between countries and the country factor maybe viewed as an important design variable.This source of variation in sampling fractions between countries may also arise national surveys between sub national groups such as regions or jurisdictions with policy differences.

Sampling design of cross national surveys of relevance to this paper is that the design variables leading to unequal inclusion probabilities within countries will often differ between countries, since quite different kinds of sampling frames and field practices can be employed. As a consequence it will often be impractical as well as potentially scientifically inappropriate to include this design variables as explanatory variable in a pooled regression analysis of data across countries. It is still feasible that these design variables maybe associated with the outcome variable within the corresponding countries, after controlling for these explanatory variables which are included. Hence sample selection bias could occur if within country variable are excluded from the model. Some adjustment, such as weighting, may therefore oftenbe needed.

To apply the weight modification method introduced above, the weight $d_j q_j$, whered_j is the design weight and q_j is the function of the country factor assumed to beincluded as an explanatory variable. In equation 9, let $q_j = \frac{1}{d_{c(j)}}$ where dc denotes the design weighted mean of the design weights within country $C$and $C_j$the country to which unit j belongs.

The $dw_j = d_j q_j = \frac{d_j}{d_{c(j)}}$as the within country weight and $B_j = d_{c(j)}$as the between country weight.

The results of fitting the robust regression model, are presented in Table 1. Pseudo- maximum likelihood estimation is employed, solving (3.0.6) with ϕ_j defined by (3.0.5), where the w_j are either constant (unweighted estimation) or are design weights. The unweighted estimates are broadly similar to those in Fieldhouse et al. (2007), Table 4, although there are some differences, as may be expected since:

IV. Conclusion

To have internal consistency, the weights need to be the same for each response variable. Weights of the form d_j, d_{Bj}, d_{wj} depends on the response variable. For internal consistency is very important and weights that do not depend on y are preferred. One possibility for obtaining internal consistency is to obtain one set of shrinkage weights that is then used for all variables. Chambers and Rao and Singh proposed using ridge regression methods to shrink the weights. Where these methods depend only on the $x$variables and not on y.

Holt and Smith emphasize robustness as one of the virtues of single stage and multistage sampling. Robustness is of course the big concern in any model-based weighting scheme, particularly when nonresponse or under coverage occur since then one cannot check that the model holds for nonrespondents. The accuracy of any population estimate from this method, such as the estimated percentage of people who think that the government should improve the economy, health services and education depends entirely on the model underlying the weights scheme. If that model does not hold individuals outside the sample, then the population estimates have unknown quality. The tree of weights in Gelman's Figure 2 is a wonderful tool for studying the weights that result from various models. Figure 2 makes it clear that the big difference in the weight variability occurs in the examples studied when education categories are added to the weighting model. Other trees could be drawn when the factors for weighting are considered in a different order, or when robust regression methods are used to estimate the parameters.

Standard statistical methods for regression analysis remain valid when sample units have been selected according to the values of$x$. Such sample selection is common with survey data, as illustrated by the cross-national application in this paper, where x includes country identifiers and sample inclusion probabilities vary considerably by country. The problem is that, even though x may be dominant source of variation in the inclusion probabilities, there may be some residual variation which is associated with y and thus could lead to bias if standard methods are employed. The option of applying full design weighting may be heavy-handed when the residual variation and resulting bias are likely to be small, especially given the potential serious inflation of standard errors. In this paper the more modest optionof applying modified weights which still correct for bias arising from such residual variation but which avoid such serious inflation of standard errors. The modification of the weights may also protect against the model fitting being dominated by a small number of large countries.

References

[1]. Kish, L. (1994). Multipopulation Survey designs; five types with seven shared aspect.
[2]. DuMouchel,W.HandDuncan,G.J.(1983).Usingsamplesurveyweightsin multiple regression analysis of stratifiedsamples.
[3]. Pfeffermann, D. and Sverchkov, M. (1999). Parametric and Semiparametric estimationofregressionmodelsfittedtosurveydata.
[4]. Godambe,V.P.andThompson,M.E.(2009)EstimatingfunctionsandSurvey Sampling.InD.PfeffermannandC.R.Rao(Eds)SampleSurveys.
[5]. Fuller,W,A.(2009)SamplingStatistics,Wiley,Hoboken
[6]. Fieldhouse, E, Tranmer, M., and Rusell, A (2007). Electoral participation of youngpeopleinEurope.
[7]. Bancroft, T. A. (1944). On biases in estimation due to the use of preliminary tests of significances. Annals of Mathematical Statistics, 15, 190–204.
[8]. D. V. Huntsberger. A generalization of a preliminary testing procedure for pooling data. Annals of Mathematical Statistics, 26:734–743, 1955.
[9]. Frederick Mosteller. On pooling data. Journal of the American Statistical Association, 43:231–242, 1948.

Edwin Ayora "Efficiency of Robust Regression Model in Sampling Designs Using Weighted Averages." IOSR Journal of Mathematics (IOSR-JM), vol. 14, no. 1, 2018, pp. 41-45.