STATISTICS IN TRANSITION new series, Spring 2014
Vol. 15, No. 2, pp. 183–196

ON THE USE OF SAMPLING WEIGHTS AND SAMPLE DISTRIBUTION WHEN ESTIMATING REGRESSION MODELS UNDER INFORMATIVE SAMPLING

Adulhakeem A. H. Eideh[1]

ABSTRACT

In this paper we show that the use of sampling weights when estimating regression models with survey data discussed by Magee, Robb and Burbidge (1998), and the use of sample distribution in fitting regression models with survey data proposed by Pfeffermann and Sverchkov (1999) are coincide methods dealing with the same statistical problem.

Key words: sample likelihood, first order inclusion probability, two-step maximum likelihood method.

1. Introduction

Some recent work has considered the definition of the sample distribution under informative sampling. When the sample selection probabilities depend on the values of the model response variable, even after conditioning on auxiliary variables, the sampling mechanism becomes informative and the selection effects need to be accounted for in the inference process. Pfeffermann, Krieger and Rinott (1998) propose a general method of inference on the population distribution (model) under informative sampling that consists of approximating the parametric distribution of the sample measurements. The sample distribution is defined as the distribution of the sample measurements given the selected sample. Under informative sampling, this distribution is different from the corresponding population distribution, although for several examples the two distributions are shown to be in the same family and only differ in some or all the parameters. The authors discuss and illustrate a general approach of approximating the marginal sample distribution for a given population distributions and first order sample selection probabilities. For more discussion on analysis of complex survey data, see Chambers and Skinner (2003), Skinner, Holt, and Smith (1989), Skinner (1994), Magee, Robb and Burbidge (1998),

[1] Department of Mathematics, College of Science and Technology, Al-Quds University, Abu-Dies campus, Palestine, P.O. Box 20002, Jerusalem. E-mail: msabdul@science.alquds.edu.

Eideh (2003, 2007, 2008, 2009, 2010, 2011, 2012a, 2012b), Eideh and Nathan (2006, 2009), Pfeffermann, Krieger and Rinott (1998), Pfeffermann and Sverchkov (1999, 2003), and Sverchkov and Pfeffermann (2004).

In this paper we will show that the use of sampling weights when estimating regression models with survey data discussed by Magee, Robb and Burbidge (1998), and the use of sample distribution in estimating regression models with survey data discussed by Pfeffermann, Krieger and Rinott (1998) and Pfeffermann and Sverchkov (1999) are coincide methods dealing with same statistical problem.

The plan of this paper is as follows. In Section 2 we consider probability weighting. In Section 3 we discuss pseudo-likelihood estimation. Section 4 deals with the use of sampling weights when estimating regression models with survey data. Section 5 introduces the use of sample distribution when estimating regression models with survey data. We conclude with a brief discussion in Section 6.

2. Probability weighting

Let $U = \{1,...,N\}$ denote a finite population consisting of N units. Let y be the target or study variable of interest and let y_i be the value of y for the ith population unit. At this stage the values y_i are assumed to be fixed unknown quantities. Suppose that an estimate is needed for the population total of y, $T = \sum_{i \in U} y_i$. A probability sample s is drawn from U according to a specified sampling design. The sample size is denoted by n . The sampling design induces inclusion probabilities for the different units of U . Let $\pi_i = \Pr(i \in s)$ be the first order inclusion probability of the ith population unit. The Horvitz-Thompson estimator or probability-weighted (PW) estimator of the population total of y, $T = \sum_{i \in U} y_i$ is given by:

$$\hat{T} = \sum_{i \in s} w_i y_i$$

where $w_i = 1/\pi_i$ is the sampling weight of unit $i \in U$, that is we weigh each sample observation i by the sampling weight, w_i . This estimator is design-unbiased, that is $E_D\left(\sum_{i \in s} w_i y_i\right) = T$, where E_D denotes the expectation under repeated sampling. For more discussion on probability weighting, see Sarndal, Swensson, and Wretman (1992).

3. Pseudo-likelihood estimation

We now consider the population values $y_1,..., y_N$ as random variables, which are independent realizations from a distribution with probability density function (pdf) $f_p(y_i \mid \theta)$, indexed by a vector of parameters θ. We now consider the estimation of the superpopulation parameter, θ, rather than the prediction of the (random variable) total T. Let

$$l(\theta \mid y_1,..., y_N) = \sum_{i=1}^{N} \log f_p(y_i \mid \theta)$$

be the census log-likelihood. The census maximum likelihood estimator of θ solves the population likelihood equations:

$$U(\theta) = \sum_{i=1}^{N} \frac{\partial \{\log f_p(y_i \mid \theta)\}}{\partial \theta} = 0$$

Following Binder (1983), the pseudo-maximum likelihood (PML) estimator is the solution of: $\hat{U}(\theta) = 0$, where $\hat{U}(\theta)$ is a sample estimator of the function $U(\theta)$. For example, the probability-weighted estimator of $U(\theta)$ is such an estimator:

$$\hat{U}_w(\theta) = \sum_{i \in s} w_i \frac{\partial \{\log f_p(y_i \mid \theta)\}}{\partial \theta}, \text{ where } w_i = 1/\pi_i .$$

That is, when the explicit form of the population likelihood is not available, we weight instead the sample likelihood and solve the weighted equations.

4. On the use of sampling weights when estimating regression models with survey data

Magee, Robb and Burbidge (1998), from now on (MRB1998), argue that when the population regression coefficient is of interest, the use of sampling weights can be desirable in regression models with complex survey data. A two-step maximum likelihood estimator is proposed as an alternative to ordinary least square and weighted least squares.

Before dealing with the problem, and defining the sample distribution mathematically, let us introduce the following notations: f_p and $E_p(\cdot)$ denote the pdf and the mathematical expectation of the population distribution, respectively, and f_s and $E_s(\cdot)$ denote the pdf and the mathematical expectation of the sample distribution.

4.1. Population model

We now consider the population values (x_i, y_i), $i = 1, \ldots, N$ as random variables, which are independent realizations from a distribution with probability density function $f_p(x, y \mid \theta)$, indexed by a vector parameter θ.

4.2. Sampling scheme

We consider a sampling design with selection probabilities $\pi_i = \Pr(i \in s)$, and sampling weight $w_i = 1/\pi_i$; $i = 1, \ldots, N$. The π_i, s may depend on the population values (x, y) as well as on other factors unknown to the researches, call these factors Z. Assume that $\pi_i \sim h(\pi | x, y, \gamma)$ where γ is a parameter indexing h. Thus, we now consider the population values (x_i, y_i, π_i) , $i = 1, \ldots, N$, as random variables, which are independent realizations from a distribution with probability density function (pdf):

$$f_p(x, y, \pi | \theta, \gamma) = f_p(x, y | \theta) \times h_p(\pi | \gamma)$$

The researcher has a sample of n observations (x_i, y_i, π_i) , $i \in s$. Each $i \in U$ is included in s with probability π_i.

The parameter of interest is the regression coefficient $\boldsymbol{\beta} = (\beta_0, \beta_1)$:

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

where $E_p(u_i x_i) = 0$, $i = 1, \ldots, N$.

4.3. Two-step maximum likelihood (ml) estimators

We consider an estimator that uses structure on the population probability density function imposed by modelling the process that generates the π_i, s. Assume that:

$$f_p(x, y, \pi | \theta, \gamma) = f_p(x, y | \theta) \times h_p(\pi | \gamma)$$

can be described as:

$$y = \beta_0 + \beta_1 x + u \tag{1}$$

where $u \underset{p}{\sim} N\left(0, \sigma_y^2\right)$, and

$$\pi^* = \ln \pi = \gamma_0 + \gamma_1 x + \gamma_2 xy + v \tag{2}$$

where $v \underset{p}{\sim} N\left(0, \sigma_{\pi^*}^2\right)$. Also, assume that u and v are independent of each other and of x.

4.4. Sample model

The probability density function of y given x_i, in the sample is given by:

$$\begin{aligned} f_s(y|x_i) &= f_p(y|x_i, i \in s) \\ &= \frac{f_p(y|x_i) \times \Pr(i \in s|x_i, y)}{\int f_p(y|x_i) \times \Pr(i \in s|x_i, y) dy} \end{aligned} \tag{3}$$

Under the conditions of equation (1), we have:

$$(y|x_i, i \in s) \sim N\left(\beta_0 + x_i\left(\beta_1 + \gamma_2 \sigma_y^2\right), \sigma_y^2\right) \tag{4}$$

Similarly, the probability density function of π^* given (x_i, y_i), in the sample is given by:

$$\begin{aligned} h_s(\pi^*|x_i, y_i) &= h_p(\pi^*|x_i, y_i, i \in s) \\ &= \frac{h_p(\pi^*|x_i, y_i) \times \Pr(i \in s|x_i, y_i, \pi^*)}{\int h_p(\pi^*|x_i, y_i) \times \Pr(i \in s|x_i, y_i, \pi^*) d\pi^*} \end{aligned} \tag{5}$$

From equation (2), we have:

$$\Pr(i \in s|x_i, y_i, \pi^*) = \pi = \exp(\pi^*)$$

Thus, under the conditions of equation (2), we obtain:

$$(\pi^*|x_i, y_i, i \in s) \sim N\left(\gamma_0^* + \gamma_1 x_i + \gamma_2 x_i y_i, \sigma_{\pi^*}^2\right) \tag{6}$$

where $\gamma_0^* = \gamma_0 + \sigma_{\pi^*}^2$.

The two-step maximum likelihood method can be performed as follows:

First step:

Estimate of γ_2 can be obtained from ordinary least squares (OLS) estimation of:

$$\pi_i^* = \gamma_0^* + \gamma_1 x_i + \gamma_2 x_i y_i + error_i\,,\ i = 1,\ldots,n \tag{7}$$

Second step:

Estimation of $\boldsymbol{\beta} = (\beta_0, \beta_1)$ based on equation (4). A consistent estimator of $\boldsymbol{\beta} = (\beta_0, \beta_1)$ can be obtained from the OLS estimation, (or ML estimation, because of normality), of

$$y_i = \beta_0 + x_i \beta_1^* + error_i\,,\ i = 1,\ldots,n$$

where $\beta_1^* = \beta_1 + \hat{\gamma}_2 \sigma_y^2$, which are given by:

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1^* \overline{x}$$

$$\beta_1^* = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2}$$

Thus,

$$\hat{\beta}_1 = \hat{\beta}_1^* - \hat{\gamma}_2 \hat{\sigma}_y^2 \tag{8}$$

5. The use of sample distribution when estimating regression models with survey data

In recent articles by Krieger and Pfeffermann (1997), Pfeffermann, Krieger, and Rinott (1998), from now on (PKR1998) and Pfeffermann and Sverchkov (1999), the authors introduced an analytic likelihood-based inference from complex survey data under informative sampling. Their basic idea is to derive the distribution of the sample data by modelling the population distribution and the conditional expectation of the first order sample inclusion probabilities. Once this sample distribution is extracted, standard likelihood-based inferential methods can be used to obtain estimates of the parameters of the population model under consideration.

The sample distribution refers to the superpopulation distribution of the sample measurements, as induced by the population model and the sample selection scheme with the selected sample of units held fixed. In order to describe the fundamental idea behind this approach, we assume full response. Let

$\mathbf{x}_i = \left(x_{i1},..., x_{ip}\right)'$, $i \in U$ be the values of a vector of auxiliary variables, $x_1,..., x_p$, and $\mathbf{z} = \left\{z_1,..., z_N\right\}$ be the values of known design variables, used for the sample selection process not included in the model under consideration. In what follows, we consider a sampling design with selection probabilities $\pi_i = \Pr(i \in s)$, and sampling weight $w_i = 1/\pi_i$; $i = 1,..., N$. In practice, the π_i's may depend on the population values $(\mathbf{x}, \mathbf{y}, \mathbf{z})$. We express this dependence by writing: $\pi_i = \Pr(i \in s \,|\, \mathbf{x}, \mathbf{y}, \mathbf{z})$ for all units $i \in U$. Since $\pi_1,..., \pi_N$ are defined by the realizations $(\mathbf{x}_i, y_i, \mathbf{z}_i), i = 1,..., N$, therefore, they are random realizations defined on the space of possible populations. The sample S consists of the subset of U selected at random by the sampling scheme with inclusion probabilities $\pi_1,..., \pi_N$. Denote by $\mathbf{I} = \left(I_1,..., I_N\right)'$ the N by one sample indicator (vector) variable, such that $I_i = 1$ if unit $i \in U$ is selected to the sample and $I_i = 0$ if otherwise. The sample s is defined accordingly as $s = \{i \,|\, i \in U, I_i = 1\}$ and its complement by $c = \bar{s} = \{i \,|\, i \in U, I_i = 0\}$. We assume probability sampling, so that $\pi_i = \Pr(i \in s) > 0$ for all units $i \in U$.

5.1. Population model

We now consider the population values $y_1,..., y_N$ as random variables, which are independent realizations from a distribution with probability density function (pdf) $f_p(y_i \,|\, \theta)$, indexed by a vector of parameters θ. Assume that the population pdf depends on known values of the auxiliary variables $\mathbf{x}_i$, so that $y_i \sim f_p(y_i \,|\, \mathbf{x}_i, \theta)$.

5.2. Sample model

We consider a sampling design with selection probabilities $\pi_i = \Pr(i \in s)$ be the first order inclusion probability of the ith population unit, and the sampling weights $w_i = 1/\pi_i$ is the sampling weight of unit $i \in U$. In practice, the π_i's may depend on the population values $(\mathbf{x}, \mathbf{y}, \mathbf{z})$. We express this dependence by writing:

$$\pi_i = \Pr(i \in s \,|\, \mathbf{x}, \mathbf{y}, \mathbf{z}) \quad \text{for all units } i \in U$$

According to Krieger and Pfeffermann (1997), the (marginal) sample pdf of y_i is defined as:

$$f_s(y_i \mid \mathbf{x}_i, \theta, \gamma) = f_p(y_i \mid \mathbf{x}_i, \theta, \gamma, i \in s)$$
$$= \frac{E_p(\pi_i \mid \mathbf{x}_i, y_i, \gamma) \times f_p(y_i \mid \mathbf{x}_i, \theta)}{E_p(\pi_i \mid \mathbf{x}_i, \theta, \gamma)} \quad (9)$$

where θ is the parameter indexing the population distribution, and γ is the informativeness parameter indexing:

$$E_p(\pi_i \mid \mathbf{x}_i, \theta, \gamma) = \int E_p(\pi_i \mid \mathbf{x}_i, y_i, \gamma) \times f_p(y_i \mid \mathbf{x}_i, \theta) dy_i$$

Note that $E_p(\pi_i \mid \mathbf{y}_i) = E_{\mathbf{z}_i \mid \mathbf{y}_i} E_p(\pi_i \mid \mathbf{y}_i, \mathbf{z}_i)$, so that $\mathbf{z}_i$ is integrated out in equation (9). See Eideh and Nathan (2006).

The question that arises is how we can identify and estimate $E_p(\pi_i \mid y_i, \mathbf{x}_i)$ based only on the sample data $\{y_i, \mathbf{x}_i, w_i;\ i \in s\}$. Pfeffermann and Sverchkov (1999) proved the following relationships: for vector of random variables $(y_i, \mathbf{x}_i)$, the following relationships hold:

$$E_s(w_i \mid y_i, x_i) = \{E_p(\pi_i \mid y_i, x_i)\}^{-1} \quad (10a)$$

$$E_p(y_i \mid \mathbf{x}_i) = \{E_s(w_i \mid \mathbf{x}_i)\}^{-1} E_s(w_i y_i \mid \mathbf{x}_i) \quad (10b)$$

$$E_s(w_i) = \{E_p(\pi_i)\}^{-1} \quad (10)$$

5.3. Estimation

Having derived the sample distribution, (PKR1998) proved that if the population measurements y_i are independent, then as $N \to \infty$ (with n fixed) the sample measurements are asymptotically independent, so we can apply standard inference procedures to complex survey data by using the marginal sample distribution for each unit. Based on the sample data $\{y_i, \mathbf{x}_i, w_i;\ i \in s\}$, (PKR1998) proposed a two-step estimation method.

Step one:

Estimate the informativeness parameters γ using equation (10a), using regression analysis. Denote the resulting estimate of γ by $\tilde{\gamma}$.

Step two:

Substitute $\tilde{\gamma}$ in the sample log-likelihood function, and then maximize the resulting sample log-likelihood function with respect to the population parameters, θ:

$$\begin{aligned} l_{rs}(\theta,\tilde{\gamma}) &= l_{srs}(\theta) - \sum_{i=1}^{n} \log E_p(\pi_i \mid \mathbf{x}_i, \theta, \tilde{\gamma}) \\ &= l_{srs}(\theta) + \sum_{i=1}^{n} \log E_s(w_i \mid \mathbf{x}_i, \theta, \tilde{\gamma}) \end{aligned} \tag{11}$$

where $l_{rs}(\theta,\tilde{\gamma})$ is the sample log-likelihood after substituting $\tilde{\gamma}$ in the sample log-likelihood function, and where

$$l_{srs}(\theta) = \sum_{i=1}^{n} \log\{f_p(y_i \mid \mathbf{x}_i, \theta)\}$$

is the classical log-likelihood obtained by ignoring the sample design.

5.4. Illustration

5.4.1. Population model

Assume the following population model:

$$y_i = \beta_0 + \beta_1 x_i + u_i \tag{12}$$

where $u_i \underset{p}{\sim} N(0,\sigma_y^2)$ and $E_p(u_i x_i) = 0$, so that $y_i \underset{p}{\sim} N(\beta_0 + \beta_1 x_i, \sigma_y^2)$, $i = 1,\ldots,N$.

Now assume that:

$$E_p(\pi_i \mid \mathbf{x}_i, y_i, \gamma) = \exp(\gamma_0 + \gamma_1 x_i + \gamma_2 x_i y_i),\ i = 1,\ldots,N \tag{13}$$

We interpret this exponential inclusion probability model approximation (13) in the spirit of probability proportional to size sampling scheme as follows. Let the size measure be:

$$d_i = \exp(\gamma_{0a} + \gamma_1 x_i + \gamma_2 x_i y_i + v_i)$$

where $E_p(v_i) = 0$ and $V_p(v_i) = \sigma_{\pi^*}^2$.

Let

$$\pi_i = \frac{nd_i}{T_d},\ T_d = \sum_{i=1}^{N} d_i$$

Assume N is large enough so that the difference between $N\bar{d}$ and $E(N\bar{d}) = N\mu_d$ can be ignored, so that $\pi_i = nd_i / T_d \cong nd_i / N\mu_d$, or $\pi_i \propto d_i$. Furthermore, since

$$\begin{aligned}\pi_i &= \frac{nd_i}{T_d} \cong \frac{nd_i}{N\mu_d} \\ &= \frac{n}{N\mu_d}\exp(\gamma_{0a} + \gamma_1 x_i + \gamma_2 x_i y_i + v_i) \\ &= \exp(\gamma_0 + \gamma_1 x_i + \gamma_2 x_i y_i + v_i)\end{aligned}$$

where $\gamma_0 = \gamma_{0a} + \ln\left(\frac{n}{N\mu_d}\right)$, therefore

$$\pi_i^* = \ln \pi_i = \gamma_0 + \gamma_1 x_i + \gamma_2 x_i y_i + v_i \tag{14}$$

where $E_p(v_i) = 0$ and $V_p(v_i) = \sigma^2_{\pi^*}$.

Under these assumptions and using Taylor series approximation, we can show that:

$$E_p(\pi_i \mid \mathbf{x}_i, y_i, \boldsymbol{\gamma}) = \exp(\gamma_0 + \gamma_1 x_i + \gamma_2 x_i y_i)$$

where $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \gamma_2)$.

Comment 1: See the similarity between (14) and (2).

5.4.2. Sample model

By substituting (12) and (13) in (9), we have:

$$y_i \underset{s}{\sim} N\left(\beta_0 + x_i(\beta_1 + \gamma_2\sigma_y^2), \sigma_y^2\right) \tag{15}$$

Comment 2: Note that (4) and (15) are similar.

5.4.3. Two-step estimation

First step:

Estimate the informativeness parameter γ_2 using (13) and (10) as follows:

$$E_s\left(w_i \mid \mathbf{x}_i, y_i, \boldsymbol{\gamma}\right) = \exp\left(-\left(\gamma_0 + \gamma_1 x_i + \gamma_2 x_i y_i\right)\right) \tag{16}$$

Using Taylor series approximation, we have $E \ln Y \cong \ln E(Y)$, so that

$$\begin{aligned} \ln E_s\left(w_i \mid \mathbf{x}_i, y_i, \boldsymbol{\gamma}\right) &= E_s\left(\ln w_i \mid \mathbf{x}_i, y_i, \boldsymbol{\gamma}\right) \\ &= -\left(\gamma_0 + \gamma_1 x_i + \gamma_2 x_i y_i\right) \end{aligned} \tag{17}$$

Hence,

$$\ln w_i = -\ln \pi_i = -\left(\gamma_0 + \gamma_1 x_i + \gamma_2 x_i y_i\right)$$

or

$$\ln \pi_i = \gamma_0 + \gamma_1 x_i + \gamma_2 x_i y_i + error_i,\ i \in s \tag{18}$$

Therefore, estimation of γ_2, denoted by $\tilde{\gamma}_2$, can be obtained from OLS estimation of (18), or you can use nonlinear regression model.

Second step:

Estimates of $\boldsymbol{\beta} = \left(\beta_0, \beta_1\right)$ can be obtained by using OLS estimation (or ML estimation method, because of linearity) of the following regression model:

$$y_i = \beta_0 + x_i \beta_1^* + error_i\ i = 1, \ldots, n \tag{19}$$

where $\beta_1^* = \beta_1 + \tilde{\gamma}_2 \sigma_y^2$, which are given by:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1^* \bar{x}$$

$$\beta_1^* = \frac{\sum_{i=1}^n \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sum_{i=1}^n \left(x_i - \bar{x}\right)^2}$$

So that,

$$\hat{\beta}_1 = \hat{\beta}_1^* - \hat{\gamma}_2 \hat{\sigma}_y^2$$

which are similar to (8).

6. Conclusion

In this paper we investigated two methods on the use of sampling weights when fitting regression model to survey data under informative probability sampling design. We showed that the only difference between the method proposed by MRB1998 and the method proposed by PKR1998 is in estimating the informativeness parameter γ_2. In MRB1998 method the estimator of γ_2 is

ML estimator, while in PKR1998 method the estimator of γ_2 is only the OLS. The intercepts γ_0^* and γ_0 have the same functional form.

The MRB1998 consider the estimator that uses more structure on the population density imposed by modelling the process generating the first order inclusion probabilities, and in their paper they consider only one model, see equation (2); while PKR1998 incorporate the sampling weights via the conditional expectation of first order inclusion probabilities given the response variable, and they consider only two models. Subsequently, Eideh (2003) proposed logit and probit models. In this paper we justified that the models that generate the first order inclusion probabilities are similar, see equations 7, 14 and 18.

Furthermore, in the last decade survey statisticians have been using the sample distribution for analysis of survey data under informative probability sampling design in several applications, in particular: prediction of finite population total under single stage sampling and two-stage sampling; fitting multilevel modelling; fitting time series models; small area estimation; estimating generalized linear models. Also, they have proposed tests of informativeness of sampling design and the test of ignorability of nonresponse in surveys, which is not the case for MRB1998. Hence, the use of sample distribution in analysis of survey data applies. Consequently, it is suggested to use the PKR1998.

We hope that this investigation will encourage further theoretical, empirical and practical research in these directions.

Acknowledgements

The author is grateful to the referees for their valuable comments.

REFERENCES

BINDER, D. A., (1983). On the variances of asymptotically normal estimators from complex surveys. International Statistical Review, 51, 279–292.

CHAMBERS, R., SKINNER, C., (2003). Analysis of Survey Data. New York: John Wiley.

EIDEH, A. H., (2003). Estimation for Longitudinal Survey Data under Informative Sampling, PhD Thesis, Department of Statistics, Hebrew University of Jerusalem.

EIDEH, A. H., (2007). A Correction Note on Small Area Estimation. International Statistical Review. 75, 122–123.

EIDEH A. H., (2008). Estimation and Prediction of Random Effects Models for Longitudinal Survey Data under Informative Sampling. Statistics in Transition – New Series. Volume 9, Number 3, December 2008, pp. 485–502.

EIDEH A. H., (2009). On the use of the Sample Distribution and Sample Likelihood for Inference under Informative Probability Sampling. DIRASAT (Natural Science), Volume 36 (2009), Number 1, pp. 18–29.

EIDEH, A. H., (2010). Analytic Inference of Complex Survey Data under Informative Probability Sampling. Proceedings of the Tenth Islamic Countries Conference on Statistical Sciences (ICCS-X), Volume I. The Islamic Countries Society of Statistical Sciences, Lahore: Pakistan, (2010). Edited by: Zeinab Amin and Ali S. Hadi. The American University in Cairo: pp. 507–536.

EIDEH, A. H., (2011). Informative Sampling on Two Occasions: Estimation and Prediction. Pakistan Journal of Statistics and Operation Research (PJSOR). Pak.j.stat.oper.res. VII No.2, 2011, pp. 283–303.

EIDEH, A. H., (2012a). Fitting Variance Components Model and Fixed Effects Model for One-Way Analysis of Variance to Complex Survey Data. Communications in Statistics – Theory and Methods, 41, pp. 3278–3300.

EIDEH, A. H., (2012b). Estimation and Prediction under Nonignorable Nonresponse via Response and Nonresponse Distributions. Journal of the Indian Society of Agriculture Statistics, 66(3) 2012, pp. 359-380.

EIDEH, A. H., NATHAN, G. (2006). Fitting Time Series Models for Longitudinal Survey Data under Informative Sampling. Journal of Statistical Planning and Inference, 136, 9, pp. 3052–3069. [Corrigendum, 137 (2007), p 628].

EIDEH, A. H., NATHAN, G., (2009). Two-Stage Informative Cluster Sampling with application in Small Area Estimation. Journal of Statistical Planning and Inference, 139, pp. 3088–3101.

KRIEGER, A. M, PFEFFERMANN, D., (1997). Testing of distribution functions from complex sample surveys. Journal of Official Statistics. 13: pp. 123–142.

MAGEE, L., ROBB, A. L., BURBIDGE, J. B., (1998). On the use of sampling weights when estimating regression models with survey data. Journal of Econometrics 84, 251–271.

PFEFFERMANN, D., KRIEGER, A. M, RINOTT, Y., (1998). Parametric distributions of complex survey data under informative probability sampling. Statistica Sinica 8: 1087–1114.

PFEFFERMANN, D., SVERCHKOV, M., (1999). Parametric and semi-parametric estimation of regression models fitted to survey data. Sankhya, 61, B, 166–186.

PFEFFERMANN, D., SVERCHKOV, M., (2003). Fitting Generalized Linear Models under Informative Probability Sampling. In: Analysis of Survey Data. (Eds. R. Chambers and C. J. Skinner). New York: Wiley, pp. 175–195.

SARNDAL, C-E., SWENSSON, B., WRETMAN, J., (1992). Model assisted survey sampling, New York: Springer.

SKINNER, C. J., (1994). Sample models and weights. American Statistical Association Proceedings of the Section on Survey Research Methods, 133–142.

SKINNER, C. J., HOLT, D., SMITH, T. M. F (Eds.), (1989). Analysis of Complex Surveys, New York: Wiley.

SVERCHKOV, M., PFEFFERMANN, D., (2004). Prediction of finite population totals based on the sample distribution. Survey Methodology, 30, 79–92.