Progressive Data Modeling

Zahoor Ahmad* and Li-Chun Zhang†

Abstract

Currently a substantial initiative is taken at the National Statistical Offices to make full use of administrative data in statistical production. For example, several studies have previously been carried out at the United Kingdom Office of National Statistics (ONS), such as forecasting value-added-tax (VAT) turnover at the unit-level, adjusting VAT register totals towards the existing Monthly Business Survey (MBS) -based turnover estimates etc. The VAT data are said to be progressive in the sense that VAT reports (or observations) of a particular time period t of interest may arrive at various points long after t. For timely prediction of the VAT turnover total before all the data have arrived, a critical issue is when the timeliness of VAT reporting is related to VAT turnover i.e. informative reporting. In this work we develop new approaches for handling informative reporting, drawing on the relevant techniques for informative sampling and informative nonresponse. We study approaches to modelling the potential informative report, methods for estimation, including maximum likelihood and estimation equations, and illustrate the methodology.

Key Words: Progressive data; existent population; reporting population; reporting probability; informative reporting; estimating equations.

1. Introduction

For some decades now, alongside survey sampling and population census, administrative registers have been an important data source for official statistics. They provide frames and valuable auxiliary information for sample surveys and censuses. Systems of inter-linked statistical registers (i.e. registers for statistical uses) have been developed on the basis of various available administrative registers to produce a wide range of purely register-based statistics, see Statistics Denmark (1995); Statistics Finland (2004); Wallgren and Wallgren (2007) for detail.

There is currently a considerable drive at the National Statistical Offices to exploit the potentials of administrative data in statistical production. For instance, several investigations have previously been carried out at ONS, such as forecasting VAT turnover at the unit-level, adjusting VAT register totals towards the existing MBS-based turnover estimates etc.

1.1 Progressive Data

Many administrative data sources, unlike sample surveys and censuses, do not always have a closing date, after which the data become static and can only be altered in editing. Reporting and registration delays and corrections can occur a long time after the statistical reference date, whether by allowance or negligence. See e.g. Hedlin et al. (2006) for delayed introduction of birth units in the UK BR, Linkletter and Sitter (2007) for delays in Natural Gas Production reports in Texas, and

*Post Graduate Student, Department of Social Statistics & Demography, Faculty of Social, Human and Mathematical Sciences, University of Southampton, Southampton, SO17 1BJ, UK. Email: Z.Ahmad@soton.ac.uk

†Professor, Department of Social Statistics & Demography, Faculty of Social, Human and Mathematical Sciences, University of Southampton, Southampton, SO17 1BJ, UK.
Email: L.Zhang@soton.ac.uk

Zhang and Fosen (2012) for delays in the Norwegian Employer/Employee Register. National accounts are based on multiple and complex data sources and typically undergo several routine revisions as more and better source data are incorporated in to the final estimates. The progressive data in the multiple sources of the National Account induces sometimes up to seven revisions over time. Depending on the situations, input data delays and changes may cause coverage errors or measurement errors, or both, in the integrated data.

For formal definition of progressive data, let t be the *reference* time point of interest and $t + d$ the *measurement* time point, for $d \geq 0$. Let U_t and y_t be the target population and value at t, respectively. For a unit i, let $I_{i(t;t+d)} = 1$ if the unit i belongs to U_t according to information available at time $t+d$ and $I_{i(t;t+d)} = 0$ otherwise, and let $y_{i(t;t+d)}$ be the observed value for t at $t+d$, provided $I_{i(t;t+d)} = 1$. The data are said to be *progressive* if, for $d \neq d' > 0$, we *can* have

$$I_{i(t;t+d)} \neq I_{i(t;t+d')} \quad \text{or} \quad y_{i(t;t+d)} \neq y_{i(t;t+d')},$$

which lead to coverage errors or measurement errors, respectively, or both. Progressiveness is a distinct feature of administrative data sources compared to sample surveys, unless one is determined to overlook all delays and changes after a certain period. One can refer to the VAT register as an example of longitudinal progressive data set based on two characteristics that are most relevant in the present context. It is longitudinal because various measurements such as turnover are recorded for different time points. It is progressive because the measurement of a given reference point is not fixed over time, due to delays in reporting and changes to the previously recorded values - this is a distinct feature that does not figure in traditional sample survey theory.

For modelling purposes, non-reporting/nonresponse is regarded as resulting from a random mechanism, which may be related to the outcome variables. If it can be shown that no such relationship exists, the nonresponse mechanism is MCAR. Even if such a relationship does exist, it may still be due to a mechanism which is MAR, whereby the probability of the response is, given the observed outcome and covariates, does not depend on the missing outcomes. If the nonresponse probabilities depend also on the values of the missing data, then they define an informative missing or NMAR data mechanism.

In survey sampling, the sampling probabilities are typically known to the analyst fitting the model, at least for the sampled units, the response probabilities are generally unknown and need to be modelled under nonignorable nonresponse. Ignoring an informative sample or nonignorable nonresponse and thus assuming implicitly that the model holding for the observed outcomes is the same as the target population model may yield large biases and erroneous inference. The books edited by Kasprzyk et al. (1989), Skinner et al. (1989) and Chambers and Skinner (2003) contain many discussions and illustrations of the effect of ignoring informative sampling or nonignorable nonresponse. See also Pfeffermann (1993), Pfeffermann (1996), Pfeffermann and Sverchkov (2009) and Pfeffermann and Sikov (2011) for further discussions and examples, with many other more recent references.

We often encounter missing data in longitudinal studies. Missing data occur whenever one or more of the sequences of measurements from individuals are incomplete, in the sense that the desired measurements are not available, or otherwise not taken. The missingness in the longitudinal data often depends on the unobserved value of the outcome at a given assessment time, that is, the missing data are often nonignorable. When data are nonignorably missing, it is necessary to model

the missing data mechanism for valid statistical inferences. Analysis of missing data has been considered by many authors in the literature (e.g., Diggle and Kenward (1994); Ibrahim et al. (1999); Ibrahim et al. (2001); Molenberghs and Verbeke (2001); Sinha et al. (2010); Sinha et al. (2011); Statistics Finland (2004); Verbeke and Molenberghs (2005); Wu et al. (2009); Xie (2008); Yi and Cook (2002); and many others). Little (1995) discusses techniques for modelling the data and the missing data mechanism simultaneously, and presents a number of examples to describe likelihood-based inferences via maximum likelihood or Bayesian approaches. Little and Rubin (2002) review methods for analysing data with various types of missing data mechanisms.

Like any observational data, the VAT reports are intrinsically associated with uncertainty. New and different frameworks are needed for administrative data, because the nature of uncertainty (or sources of errors) is different from that arising in survey sampling. VAT data are said to be progressive in the sense that VAT reports (or observations) of a particular time period t of interest may arrive at various points long after t, whether by arrangement or not. For timely prediction of the VAT turnover total before all the data have arrived, a critical issue is when the timeliness of VAT reporting is related to VAT turnover i.e. informative non-reporting. In this work we develop new approaches for handling informative non-reporting/missing that are discussed below.

1.2 Prediction Framework for Longitudinal Progressive Data

Zhang and Pritchard (2013) extended the following prediction framework of Valliant et al. (2000) for progressive data. Let $U_{e(t;t+d)}$ be a known universe of *existent* units. For instance, in repeated statistical production, one may include in $U_{e(t;t+d)}$ all the units that have previously been included in the population, i.e. $U_{e(t;t+d)} = \{i; \sum_{j=0}^{\infty} I_{i(t-j;t+d)} > 1\}$. The existent universe U_e admits a bipartition, denoted by $U_{e+} \cup U_{e-}$, where U_{e+} contains the units that actually belong to the target universe U_t, i.e. $I_{it} = 1$, and U_{e-} those that do not, i.e. $I_{it} = 0$. Put $U_{0(t;t+d)} = U_t \backslash U_{e+(t;t+d)}$ i.e. the target units that are not included in the existent universe. One may refer to $U_{0(t;t+d)}$ as the *birth delays* and U_{e-} the *death delays*. Let $Y_t = \sum_{i \in U_t} y_{it}$ be the target total of interest. A general expression of a prediction-based estimator of Y_t at $t + d$ can be given as

$$\hat{Y}_{(t;t+d)} = \sum_{i \in U_{e(t;t+d)}} \hat{I}_{it} \hat{y}_{it} + \sum_{i \in U_{0(t;t+d)}} \hat{y}_{it} \tag{1}$$

Zhang and Pritchard (2013) applied the prediction framework (1) to VAT register data in UK. The prediction approach (1) requires modelling of $\{I_{it},\ y_{it};\ i \in U_{i(t;t+d)}\}$ and $Y_{0(t;t+d)}$, with or without conditioning on historic y- and I-values and other relevant auxiliaries. Zhang and Pritchard (2013) notice potential connections of modelling progressive data to the literature on estimation in the presence of nonresponse and informative sampling.

Before discussing estimation/modelling approaches, it is necessary to explain different populations in the case of longitudinal progressive data. As we have an existent population U_{et} at time t that is known at time $t + d$ but not all the units are in the target population U_t. The U_{et} contains both active (potentially reporting units) and non-active units. The units which are active belongs to the target population U_t. Denote by $U_{(e+)t}$ the units $i \in U_{et}$ that are active ($I_{it} = 1$), and by U_{rt} the sub-population of active units that report ($\delta_{it} = 1 | I_{it} = 1$). Not every $i \in U_t$

belongs to U_{et} due to so-called birth delays, which are non-existing units at time $t+d$, but later turn out to belong to U_t nevertheless. Similarly, $U_{et}\backslash U_{(e+)t}$ may be called death delays. Hence U_{rt} is a subset of $U_{(e+)t}$ and $U_{(e+)t}$ is a subset of U_{et}. Also $U_e\backslash U_{e+}$ is a subset of $U_e\backslash U_r$. By this way we have two mechanisms to deal with while discussing the estimation approaches; first is selection mechanism (from U_{et} to $U_{(e+)t}$) and second is reporting/response mechanism (from $U_{(e+)t}$ to U_{rt}).

Currently in the following two estimation/modelling approaches, we considered the second phase i.e. the inference from U_{rt} to $U_{(e+)t}$ under informative non-reporting. The modelling/estimation approaches proposed for dealing with report-selection mechanism are equally applicable to longitudinal studies aimed to account for informative selection and non-reporting.

2. Modelling/Estimation approaches

For prediction and estimation of model parameters, presently we have provided two approaches. The first approach is based on the likelihood method in which we follow the idea of Pfeffermann (2011) and developed the Bayes-Based Likelihood (BBL) method for prediction as well as for estimation of model parameters in the case of longitudinal data under informative non-reporting/non-response and currently assuming that all units are selected. Later on this assumption will be relaxed. BBL approach can be complicated because we need to assume density of active population and there can be a potential non-identifiability of reporting model. Also it is computationally heavy. An alternative to BBL approach, we developed finite population estimating equations approach. The EE approach is free of distributional assumptions of the outcome variable and it is computationally easy. Also in this approach an intuitive empirical estimator of reporting/response probability is used that is based on the historic reporting indicators rather assuming an explicit/parametric form of model for reporting indicator. For EE approach we have to have historic data. As the BBL is not the full-likelihood, hence not fully efficient either; one might still expect it to compare favourably to the EE approach.

2.1 Bayes-Based Maximum Likelihood Approach

The case where the missing data are not MAR (NMAR) can be treated by postulating a parametric model for the distribution of the outcomes before non-response and a model for the response mechanism. These two models define a parametric model for the joint distribution of the outcomes and response indicators, and therefore the parameters of these models can be estimated by maximization of the likelihood based on this joint distribution. See, Greenlees et al. (1982), Rubin (1987), Little (1993), Beaumont (2000), Little and Rubin (2002) and Qin et al. (2002). In particular, most of the approaches assume that the model covariates are known also for the non-respondents, which is often not the case. Pfeffermann (2011) provided a Bayes-based conditional likelihood approach to deal with non-ignorable non-response mechanism when covariates are unknown for the non-respondents. Pfeffermann and Sikov (2011) review approaches proposed in the literature to deal with NMAR non-response, but these approaches are quite limited. To develop Bayes-based maximum likelihood approach for longitudinal studies we proceed as follows.

Let y_{it} denote the value of an outcome variable Y at time t, associated with unit i belonging to the active population $U_{(e+)t}$. Let x_{it} denote a vector of auxiliary

covariates including historic y-value associated with unit i. As defined above U_{rt} be the reported population with reported outcomes and covariates, and let $U_{\bar{r}t}$ be the unreported/delayed population for which at least the outcomes are not reported (missing) at reference time point t.

Following the idea of complex survey modelling under informative nonreponse given by Pfeffermann (2011), we developed the model for reporting population U_{rt} when we have the *pdf* of active population $U_{(e+)t}$ and conditional reporting probability model. Let $\delta_{it} = 1$ if $i \in U_{rt}$ and $\delta_{it} = 0$ if $i \in U_{\bar{r}t}$. The conditional *pdf* of the outcome y_{it} given that unit i is in the reporting population is

$$f_{U_{rt}}\left(y_{it}|x_{it}\right) = f\left(y_{it}|x_{it}, \delta_{it}=1\right) = \frac{\Pr\left(\delta_{it}=1|y_{it}, x_{it}\right)}{\Pr\left(\delta_{it}=1|x_{it}\right)} f_{U_{(e+)t}}\left(y_{it}|x_{it}\right), \qquad (2)$$

where $\Pr\left(\delta_{it}=1|y_{it}, x_{it}\right)$ is reporting probability model and $f_{U_{(e+)t}}\left(y_{it}|x_{it}\right)$ is *pdf* of active population $U_{(e+)t}$. The $\Pr\left(\delta_{it}=1|x_{it}\right) = \int \Pr\left(\delta_{it}=1|y_{it}, x_{it}\right) f_{U_{(e+)t}}\left(y_{it}|x_{it}\right) dy_{it}$. By (2), if the active population outcome and the reports are independent between the units, and the covariates are only known for the reporting units, one can estimate the parameter θ governing the active population model and the parameters γ governing the model for reporting probabilities by maximizing the reporting population likelihood

$$L_{U_{rt}} = \prod_{i=1}^{r} f\left(y_{it}|x_{it}, \delta_{it}=1; \theta, \gamma\right) = \prod_{i=1}^{r} \frac{\Pr\left(\delta_{it}=1|y_{it}, x_{it}; \gamma\right) f_{U_{(e+)t}}\left(y_{it}|x_{it}; \theta\right)}{\Pr\left(\delta_{it}=1|x_{it}; \theta, \gamma\right)}.$$

The model for reporting population can be fitted under informative non-reporting if we know the reporting probability model $\Pr\left(\delta_{it}=1|y_{it}, x_{it}; \gamma\right)$ and the density of active population $f_{U_{(e+)t}}\left(y_{it}|x_{it}; \theta\right)$. In literature, different response probability models like linear, exponential, logit and probit models have been used. One of these models can be used in (2).

As in (2), on right hand side there is a product of two functions. It is possible to have a problem of non-identifiability. Identifiable model can be obtained using logistic model instead of exponential by imposing the condition that at least one covariate should differ among covariates used for reporting model and density of the active population (see Pfeffermann and Landsman (2011)). However, in practice, the covariates featuring in the population model need not be the same as the covariates featuring in the model of the conditional response probabilities. Feder and Pfeffermann (2015) adopted empirical likelihood approach, which helps to avoids the identifiability issue.

In finite population estimating equating equation (EE) approach discussed in next chapter, we assumed a model that provides stationarity in reporting probabilities so that the empirical estimator of reporting probability i.e. ratio of reporting and existing history can be used as an estimator of unknown reporting probabilities, the detail is given in Section 2.2.1. To compare Bayes-based likelihood (BBL) approach with EE approach, we assumed following reporting probability model for BBL approach,

$$\Pr(\delta_{it}=1|y_{it}, x_{it}) = \pi_{it} = \left[1 + exp\left\{\eta(y_{it}, x_{it}) + d_{it}\right\}\right]^{-1}, \qquad (3)$$

where η is a known function depending on y_{it} and x_{it}. Possible models for d_{it} are

(i). IID with mean 0 and variance σ^2 (or σ_t^2), e.g. $d_{it} \sim N(0, \sigma^2)$.

(ii). $u_i + v_{it}$, where u_i is an individual random effect and suppose $u_i \sim N(0, \sigma_u^2)$, and v_{it} is IID with mean 0 and variance σ_v^2, e.g. $v_i \sim N(0, \sigma_v^2)$. Then $Corr(d_{it}, d_{is}) = \sigma_u^2/(\sigma_u^2 + \sigma_v^2)$.

(iii). an $AR(1)$ over t for fixed i, e.g. $d_{it} = \rho d_{i,t-1} + e_{it}$, where ρ is a fixed scalar and $e_{it} \sim N(0, \sigma_e^2)$ then $d_{it} \sim N(0, \sigma_e^2/(1-\rho^2))$ and $Corr(d_{it}, d_{is}) = \rho^{(s-t)}$. It is assumed that $|\rho| < 1$ to guarantee stationarity.

For sensitivity analysis one can postulate other η than (3).
From reporting probability model (3), we have another random variable d_{it} along with y_{it} and δ_{it}, the joint density of these three random variables can be written as

$$f(y_{it}, d_{it}, \delta_{it}|x_{it}) = \Pr(\delta_{it}|y_{it}, x_{it}, d_{it}) f(y_{it}|x_{it}, d_{it}) h(d_{it}) = \Pr(\delta_{it}|y_{it}, d_{it}) f(y_{it}|x_{it}) h(d_{it}). \tag{4}$$

Assuming independence between y_{it} and d_{it}, the *pdf* of y_{it} given that unit i is in the reporting population, is

$$f_{U_{rt}}(y_{it}|x_{it}, \delta_{it} = 1) = \frac{\int_{d_{it}} \Pr(\delta_{it} = 1|y_{it}, x_{it}, d_{it}) f_{U_{(e+)t}}(y_{it}|x_{it}) h_{U_{(e+)t}}(d_{it}) d_{d_{it}}}{\int_{d_{it}} \int_{y_{it}} \Pr(\delta_{it} = 1|y_{it}, x_{it}, d_{it}) f_{U_{(e+)t}}(y_{it}|x_{it}) h_{U_{(e+)t}}(d_{it}) d_{y_{it}} d_{d_{it}}}. \tag{5}$$

To estimate the parameters of the reporting population distribution, by (5), if the active population outcome and the reports are independent between the units, and the covariates are only known for the reporting units, one can estimate the parameter θ governing the active population model and the parameters γ governing the model for reporting probabilities by maximizing the reporting population likelihood,

$$\begin{aligned} L_{U_{rt}} &= \prod_{i \in U_{rt}} f_{U_{rt}}(y_{it}|x_{it}, \delta_{it} = 1; \theta, \gamma) \\ &= \prod_{i \in U_{rt}} \frac{\int_{d_{it}} \Pr(\delta_{it} = 1|y_{it}, x_{it}, d_{it}, \gamma) f_{U_{(e+)t}}(y_{it}|x_{it}; \theta) h_{U_{(e+)t}}(d_{it}; \sigma_d^2) d_{d_{it}}}{\int_{d_{it}} \int_{y_{it}} \Pr(\delta_{it} = 1|y_{it}, x_{it}, d_{it}, \gamma) f_{U_{(e+)t}}(y_{it}|x_{it}, \theta) h_{U_{(e+)t}}(d_{it}; \sigma_d^2) d_{y_{it}} d_{d_{it}}}. \end{aligned} \tag{6}$$

2.1.1 Prediction without covariates using BBL

In simulation study, the BBL approach is used for prediction of population mean without x. For this, suppose the parameters of the distribution of y_{it} are θ and σ^2, as e.g. the normal and log-normal distribution then the likelihood function (6) takes the following form,

$$\begin{aligned} L_{U_{rt}} &= \prod_{i \in U_{rt}} f_{U_{rt}}(y_{it}|\delta_{it} = 1; \theta, \sigma^2, \gamma) \\ &= \prod_{i \in U_{rt}} \frac{\int_{d_{it}} \Pr(\delta_{it} = 1|y_{it}, d_{it}, \gamma) f_{U_{(e+)t}}(y_{it}; \theta, \sigma^2) h_{U_{(e+)t}}(d_{it}; \sigma_d^2) d_{d_{it}}}{\int_{d_{it}} \int_{y_{it}} \Pr(\delta_{it} = 1|y_{it}, d_{it}, \gamma) f_{U_{(e+)t}}(y_{it}, \theta, \sigma^2) h_{U_{(e+)t}}(d_{it}; \sigma_d^2) d_{y_{it}} d_{d_{it}}}. \end{aligned} \tag{7}$$

For prediction of population mean, suppose for reporting units the estimated density is $\hat{f}_{U_{rt}}(y_{it}|\delta_{it} = 1; \hat{\theta}, \hat{\gamma})$ and for non-reporting units we can write

$$\begin{aligned} &\hat{f}_{U_{\bar{r}t}}(y_{it}|\delta_{it} = 0; \hat{\theta}, \hat{\gamma}) \\ &= \frac{\int_{d_{it}} [1 - \Pr(\delta_{it} = 1|y_{it}, d_{it}, \hat{\gamma})] f_{U_{(e+)t}}(y_{it}; \hat{\theta}) h_{U_{(e+)t}}(d_{it}; \sigma_d^2) d_{d_{it}}}{\int_{d_{it}} \int_{y_{it}} [1 - \Pr(\delta_{it} = 1|y_{it}, d_{it}, \hat{\gamma})] f_{U_{(e+)t}}(y_{it}, \hat{\theta}) h_{U_{(e+)t}}(d_{it}; \sigma_d^2) d_{y_{it}} d_{d_{it}}}. \end{aligned} \tag{8}$$

Now the population mean at time t can be estimated as

$$\begin{aligned}\hat{\bar{Y}} &= \frac{1}{N}\sum_{i=1}^{r} y_{it} + \frac{1}{N}\sum_{i=r+1}^{N} \hat{E}_{\bar{r}}\left(y_{it}|\delta_{it}=0\right) \\ &= \frac{r}{N}\bar{y}_{rt} + \frac{1}{N}\sum_{i=r+1}^{N}\int_{\bar{r}} y_{it} f_{U_{\bar{r}t}}(y_{it}|\delta_{it}=0;\hat{\theta},\hat{\gamma})d_{y_{it}}. \end{aligned} \tag{9}$$

2.1.2 Prediction with covariates using BBL

The BBL approach is also used to estimate the model parameters in simulation study. Suppose, we have a linear regression model of y with regression coefficients β and variance σ^2, then the likelihood function (6) can be written as

$$\begin{aligned} L_{U_{rt}} &= \prod_{i\in U_{rt}} f_{U_{rt}}(y_{it}|x_{it},\delta_{it}=1;\beta,\sigma^2,\gamma) \\ &= \prod_{i\in U_{rt}} \frac{\int_{d_{it}} \Pr(\delta_{it}=1|y_{it},x_{it},d_{it},\gamma) f_{U_{(e+)t}}(y_{it}|x_{it};\beta,\sigma^2) h_{U_{(e+)t}}(d_{it};\sigma_d^2) d_{d_{it}}}{\int_{d_{it}}\int_{y_{it}} \Pr(\delta_{it}=1|y_{it},x_{it},d_{it},\gamma) f_{U_{(e+)t}}(y_{it}|x_{it},\beta,\sigma^2) h_{U_{(e+)t}}(d_{it};\sigma_d^2) d_{y_{it}} d_{d_{it}}}. \end{aligned} \tag{10}$$

For estimation of population mean θ, suppose for reporting units the estimated density is $\hat{f}_{U_{rt}}(y_{it}|x_{it},\delta_{it}=1;\hat{\theta},\hat{\gamma})$ and for non-reporting units we can write

$$\begin{aligned} &\hat{f}_{U_{\bar{r}t}}(y_{it}|x_{it},\delta_{it}=0;\hat{\theta},\hat{\gamma}) \\ &= \frac{\int_{d_{it}} \left[1-\Pr(\delta_{it}=1|y_{it},x_{it},d_{it},\hat{\gamma})\right] f_{U_{(e+)t}}(y_{it}|x_{it};\hat{\theta}) h_{U_{(e+)t}}(d_{it};\sigma_d^2) d_{d_{it}}}{\int_{d_{it}}\int_{y_{it}} \left[1-\Pr(\delta_{it}=1|y_{it},x_{it},d_{it},\hat{\gamma})\right] f_{U_{(e+)t}}(y_{it}|x_{it},\hat{\theta}) h_{U_{(e+)t}}(d_{it};\sigma_d^2) d_{y_{it}} d_{d_{it}}}. \end{aligned} \tag{11}$$

Now the population mean at time t can be estimated as

$$\begin{aligned}\hat{\bar{Y}} &= \frac{1}{N}\sum_{i=1}^{r} y_{it} + \frac{1}{N}\sum_{i=r+1}^{N} \hat{E}_{\bar{r}}\left(y_{it}|x_{it},\delta_{it}=0\right) \\ &= \frac{r}{N}\bar{y}_{rt} + \frac{1}{N}\sum_{i=r+1}^{N}\int_{\bar{r}} y_{it} f_{U_{\bar{r}t}}(y_{it}|x_{it},\delta_{it}=0;\hat{\theta},\hat{\gamma})d_{y_{it}}, \end{aligned} \tag{12}$$

where x_{it} assumed known for non-reporting units in existing population.

2.2 Finite Population Estimating Equations Approach

We have seen that using BBL approach, a potential non-identifiability can occur while maximizing reporting model and it will become more difficult when only historic response values will be used as covariates. Also this approach is computationally heavy because we cannot explicitly obtain the estimators from reporting likelihood when the logistic model is used for reporting probability. We need to use numerical optimization routines to obtain the parameter estimates and an other potential issue can be the initial values to initiate these optimization algorithms.

For an alternative to BBL approach, we developed finite population estimating equations approach for finite population prediction and estimating the model parameters. We use as the estimator of reporting probability the observed historic

reporting rate for each existent unit on its own. Estimating equations for the parameter of interest, e.g. finite population mean or regression coefficients, are built on these estimated reporting probabilities. Even though asymptotically (under certain conditions) the MLE will attain the Cramer-Rao lower bound (which is the smallest variance). Moreover, MLE estimators are based on the assumption that the distribution is known (else the estimator is misspecified), however an estimating equation can be free of such assumptions and explicit/parametric form of $\Pr(\delta_{it} \mid y_{it}, x_{it})$ (see Godambe (1991a)).

As discussed earlier that currently we are only dealing with active population $U_{(e+)t}$ to reporting population U_{rt}. The combined business of U_{et} to U_{rt} will be treated later. Finite population estimating equation approach discussed in this section, however, is applicable to missing data problems and not restricted to progressive data setting. In this approach we allow the $y's$ to be fixed and the basis of inference is a model for δ_{it} given in (13).

2.2.1 Model

Let $\mathbf{y}_i = (y_{i0}, y_{i1}, ..., y_{iT})$ and $\mathbf{x}_i = (x_{i0}, x_{i1}, ..., x_{iT})$, where y_{it} represents an outcome variable of interest at time t and x_{it} is a vector of covariates at time t including the historic $y's$.

To model response variable in the presence of informative non-reporting, the reporting probabilities are unknown unlike informative sampling. We need to model the reporting probabilities and then the estimates of this model parameters are required. But here we want to make use of existing and reporting history of each unit to estimate its individual reporting probability and this non-parametric estimate will be used as an estimate of unknown reporting probabilities. For informative non-reporting, the reporting probabilities need to depend on response variable above their dependence on covariates. In longitudinal nature of data the reporting probabilities for each time period that depends on respective response variable must have a trend over time and it will become non-stationary time series. But the estimator that we want to use for reporting probability is of stationary nature. Stationary of a time series means that the mean, variance and autocorrelation structure do not change over time and no periodic fluctuations.

We postulate *informative but stationary and individual* reporting. For example, the actual reporting at a business could depend on the accounting system, the personal responsible for the reporting, etc. all of which can potentially be related to the size of the business and hence possibly the response variable y of interest, beyond whatever x that is available. Meanwhile, there is bound to be some stability over a limited time period. For such a scenario, we can define model for unknown reporting probabilities for all previous time periods as well as for current time period as follows. Let δ_{it} be the response indicator for $t = 0, 1, ..., T$, put

$$Pr\left(\delta_{it} = 1\right) = \pi_{it} = \left[1 + exp\left\{\eta_i\left(\mathbf{y}_i, \mathbf{x}_i\right) + d_{it}\right\}\right]^{-1}, \tag{13}$$

where d_{it} is defined earlier, η_i can differ between units but is assumed to be constant over a time window $(0, ..., T)$. Further,

$$E\left(\pi_{it}\right) \approx \frac{e^{\eta_i}}{1 + e^{\eta_i}}\left[1 + \frac{\sigma_d^2\left(1 - e^{\eta_i}\right)}{2\left(1 + e^{\eta_i}\right)^2}\right] = \pi_i(say), \tag{14}$$

where σ_d^2 is the variance of respective d_{it}.

For the estimator of unknown reporting probabilities, suppose for each population unit, define $R_i = \sum_{t=1}^{T} \delta_{it}$ to be the number of past responses over $(1, ..., T)$. Then the estimated reporting probability can be $\hat{\pi}_i = R_i/T$, where $\delta_{it} \sim Bernoulli\,(1, \pi_{it})$, then provided identical marginal distribution of d_{it} for every t,

$$E\left(\hat{\pi}_i|T\right) = E\left\{\frac{1}{T}\sum_{t=1}^{T} E\left(\delta_{it}|\pi_{it}\right)\right\} = \frac{1}{T}\sum_{t=1}^{T} E\left(\pi_{it}|\eta_i\right) = \frac{T\pi_i}{T} = \pi_i. \tag{15}$$

From (14) and (15), we can see that $\hat{\pi}_i$ and π_{it} are approximately same on average. It indicates the candidature of $\hat{\pi}_i$ as an estimator of π_{it} under the model (13).

To use estimating equations approach for analytical purpose, suppose model for response variable is

$$y_{it} = \beta_0 + \beta_1 x_{it} + \epsilon_{it}, \tag{16}$$

where $\epsilon_{it} \sim N(0, \sigma_{it}^2)$ and $\sigma_{it}^2 = \sigma^2 x_{it}^{\alpha}$.

2.2.2 Estimating Equations

Let θ_0 be a finite population parameter defined as the solution to the following estimating equations

$$H_N\left(\theta\right) = N^{-1}\sum_{i=1}^{N} S_i\left(\theta\right); \quad H_N\left(\theta_0\right) = 0, \tag{17}$$

where $S_i()$ is a score function with the y-values considered fixed. The population parameters are defined by the estimating equations (17) and the basis of inference is the model for reporting indicator given in (13). The (unobserved) estimating equations for current $(t = 0)$ reporting units are given by,

$$\tilde{H}_N\left(\theta\right) = N^{-1}\sum_{i=1}^{N} \frac{\delta_{i0}}{\pi_{i0}} S_i\left(\theta\right) = N^{-1}\sum_{i=1}^{N} W_i\left(\theta\right) \quad and \quad \tilde{H}_N(\tilde{\theta}) = 0, \tag{18}$$

where π_{i0} denote the current reporting probability in which '0' stands for current time point of interest. On replacing π_{i0} by $\hat{\pi}_i$ in (18), we have the observed EE,

$$\hat{H}_N\left(\theta\right) = N^{-1}\sum_{i=1}^{N} \frac{\delta_{i0}}{\hat{\pi}_i} S_i\left(\theta\right) \quad and \quad \hat{H}_N(\hat{\theta}) = 0. \tag{19}$$

One might draw analogy between the observed EE and the pseudo-MLE approach; but the score is not necessarily derived from likelihood, and the $\hat{\pi}_i$ is estimated instead of known. When the census estimating equations (17) are the likelihood equations, the estimators obtained by solving (18) with known inclusion probabilities are known in the sampling literature as pseudo mle (pmle). See Binder (1983), Skinner et al. (1989), Pfeffermann (1993), Pfeffermann (1996) and Godambe and Thompson (2009)for discussion with many examples.

Currently under finite population estimating equations approach $\tilde{H}_N(\theta)$ is unbiased estimating equations and it seems difficult to have $\hat{H}_N(\theta)$ asymptotically unbiased. So there is need to obtain bias-adjusted $\hat{H}_N(\theta)$. Two possible bias-adjusted estimating equations are

$$\hat{H}_N^*\left(\theta\right) = N^{-1}\sum_{i=1}^{N} \delta_{i0}\left(\frac{1}{\hat{\pi}_i} - \frac{V\left(\hat{\pi}_i\right)}{\pi_i^3}\right) S_i\left(\theta\right), \tag{20}$$

$$\hat{H}_N^{**} = N^{-1} \sum_{i=1}^{N} \delta_{i0} \left(\frac{\pi_i^2}{\hat{\pi}_i(\pi_i^2 + V(\hat{\pi}_i))} \right) S_i(\theta) \tag{21}$$

2.2.3 Mean Square Error Estimation

For estimation of mean square error of $\hat{\theta}$, by Taylor expansion, we have

$$0 = \hat{H}_N(\hat{\theta}) = \hat{H}_N(\theta_0) + \hat{H}'_N(\theta_0)(\hat{\theta} - \theta_0) + \frac{1}{2}\hat{H}''_N(\theta_0)(\theta^* - \theta_0)^2 \tag{22}$$

where θ^* lies between $\hat{\theta}$ and θ_0. Ignoring the remainder terms

$$(\hat{\theta} - \theta_0) \approx -\left(\hat{H}'_N(\theta_0)\right)^{-1} \hat{H}_N(\theta_0)$$

Post multiplying by $(\hat{\theta} - \theta_0)'$ on both sides, we have

$$(\hat{\theta} - \theta_0)(\hat{\theta} - \theta_0)' \approx \left(\hat{H}'_N(\theta_0)\right)^{-1} \hat{H}_N(\theta_0)\hat{H}_N^T(\theta_0) \left(\hat{H}'_N(\theta_0)\right)^{-T} \tag{23}$$

From (23), mean square error can be obtained provided we know θ_0. Currently we are working on various ways of estimating MSE, which can improve the direct plug-in estimator.

2.3 Simulation Set-up

Simulation study is conducted for both population prediction and estimation of model parameters to illustrate both approaches.

To illustrate the application of the BBL for prediction we need to assume a density for active population and a reporting model to obtain current reporting indicator. Suppose the density for the response variable of active population is normal, i.e. $y_{it} \sim N(\theta, \sigma^2)$.

For analytical use of BBL approach, suppose the density for the response variable of active population is normal, i.e. $y_{it} \sim N(\beta_0 + \beta_1 y_{i(t-1)}, \sigma_{it}^2)$, where $\sigma_{it}^2 = \sigma^2 y_{i(t-1)}^{\alpha}$ and it is also assumed that repose variable at time $t-1$ is available at current time period. The reporting probability model is defined in (3) along with three $d'_{it}s$. The likelihood function (7) is used to obtain the estimators for density of reporting population. Then (8) and (9) are used for prediction of population mean. The likelihood function (10) is used to estimate the model parameters. The parameter $\gamma's$ of reporting models are estimated by solving the calibration constraints $\sum_{i=1}^{r} w_i = N$ and $\sum_{i=1}^{r} w_i \delta_{i1} = \sum_{i=1}^{N} \delta_{i1}$ iteratively, where w_i is the reporting weight of i^{th} unit and δ_{i1} is previous reporting indicator. The estimates of $\gamma's$ are then used in likelihood function before maximising it.

For the illustration of EE approach in finite population prediction and estimation of model parameters we do not need to assume density for response variable, we only require the reporting probability model for current as well as for historic reporting indicators that is actually the basis of inference in finite population prediction approach. The reporting probabilities for current and historic time periods are obtained under the postulated model (13), where η_i is assumed to be $\frac{1}{T}\sum_{t=1}^{T} y_{it}$ and it also assumed that past values are available at time 0. The historic reporting probabilities will be then used to obtain reporting history and ultimately the estimator for current reporting probabilities. The finite population estimating equation

(17) are used to define the population parameters for both prediction and regression model and the reporting estimating equations based on estimated reporting probabilities given in (19), (20) and (21) are used to obtain the estimators for finite population parameters.

The reporting probability models for BBL and EE approached are defined in (3) and (13) respectively. These are our postulated models. We also want to see the effect of different reporting models on the results of both approaches as a sensitivity check. One can imagine that the reporting can depends on previous reporting indicator and it can also be used in reporting probability model. So the following table is provided for different possible reporting probabilities models.

Table 1: Reporting probability models

	Without $\delta_{1,j}$	With $\delta_{1,j}$
	Case-(i)	Case-(iii)
η_i	$logit(\pi_{it,j}) = \gamma_j \eta_i + d_{it,j}$	$logit(\pi_{it,j}) = \gamma_j \eta_i + d_{it,j}$
	$logit(\pi_{i0,j}) = \gamma_{0,j} \eta_i + d_{i0,j}$	$logit(\pi_{i0,j}) = \gamma_{01,j} \eta_i + \gamma_{02,j} \delta_{i1,j} + d_{i0,j}$
	Case-(ii)	Case-(iv)
y_{it}	$logit(\pi_{it,j}) = \gamma_j y_{it} + d_{it,j}$	$logit(\pi_{it,j}) = \gamma_j y_{it} + d_{it,j}$
	$logit(\pi_{i0,j}) = \gamma_{0,j} y_{i0} + d_{i0,j}$	$logit(\pi_{i0,j}) = \gamma_{01,j} y_{i0} + \gamma_{02,j} \delta_{i1,j} + d_{i0,j}$

In above table $j = 1, 2, 3$ is for each d_{it}. The $\pi_{it,j}$ denotes the historic reporting probabilities and $\pi_{i0,j}$, the current reporting probabilities. Similarly $d_{it,j}$ denotes the historic ${d_{it}}'s$ and $d_{i0,j}$, is for current time period. The η_i and $y's$ are multiplied by the coefficient $\gamma's$, suitable values of $\gamma's$ are used to restrict the overall reporting rate around 80%. The function η_i will be replaced with function η for BBL approach that is based on concurrent $y's$ rather historic $y's$.

The reporting models given in above table for case (i) are that assumed by EE; the models given in case (ii) that by BBL. Applying BBL to the current reporting indicators generated under case (i, iii & iv) model entails model misspecification; likewise with EE when the current reporting indicators are generated under case (ii, iii & iv).

For EE approach the reporting probabilities for historic time periods, i.e. π_{it}, are obtained using the models given in above table. These reporting probabilities are used to generate reporting indicators of corresponding time periods, i.e. δ_{it}. The average of these historic reporting indicators is then used to obtain the estimated reporting probability, i.e. $\hat{\pi}_i$, which is an estimator of unknown current reporting probabilities, π_{i0}. Below we present the results for first two cases with $d_{it,1}$.

2.4 Simulation Study

2.4.1 Population Prediction

To simulate the longitudinal data, we considered 11 time periods ($t = 0, 1, 2, ..., 10(= T)$), where $t = 0$ is the current time period, then data for y_{it} are generated using multivariate normal distribution with mean vector $(10.50, 10.55, 10.60, 10.65, 10.70, 10.75, 10.80, 10.85, 10.90, 10.95, 12)$ and variance covariance matrix with same covariances of 10 and variances of 30. The $d_{it,1}$ are generated using multivariate normal distribution with zero mean vector and variance covariance matrix with 0 covariances and same variances of 0.05.

The reporting probabilities π_{it} for historic time periods are obtained under the

first two cases of table 1. These reporting probabilities are used to generate reporting indicators of corresponding time periods. The average of historic reporting indicators is then used to obtain the estimated reporting probability for current time period that will be used as an estimator of unknown reporting probabilities required in EE approach.

The size of population is assumed to be $N = 1000$ and no of simulations are 200. The population data is generated once and kept fixed for all simulations. The parameters are estimated using naive unadjusted approach, BBL approach with likelihood function given in (7) and estimating equations given in (19), (20) and (21). The R function *nlm* is used to obtain the maximum likelihood estimates. The results for all approaches are given in following table.

Table 2: Estimate, bias, empirical SE, RMSE and average of squared error for both cases
($\theta_{FP} = 12.003$, and $\sigma^2_{FP} = 31.06$; $N = 1000$ and $S = 200$)

		Naive	$\hat{H}_N(\theta)$	$\hat{H}^*_N(\theta)$	$\hat{H}^{**}_N(\theta)$	BBL
	Case-(i): $logit(\pi_{it}) = \gamma\eta_i + d_{it}$ and $logit(\pi_{i0}) = \gamma_0\eta_i + d_{i0}$					
$\hat{\theta}$	*Mean Est.*	12.2838	11.9109	11.9737	11.9670	12.7030
	Bias	0.2794	0.0936	0.0308	0.0375	0.6985
	Emp. SE	0.0814	0.0874	0.0857	0.0859	0.0775
	Emp. RMSE	0.2910	0.1281	0.0910	0.0937	0.7044
	$RMSE_{\theta_0}$	0.2813	0.1052	0.0710	0.0729	.....
$\hat{\sigma}^2$	*Mean Est.*	31.0012	31.0200	31.0040	31.0096	30.1394
	Bias	0.0597	0.0409	0.0569	0.0513	0.9261
	Emp. SE	0.5230	0.5729	0.5579	0.5602	0.4792
	Emp. RMSE	0.5264	0.5744	0.5608	0.5626	1.1010
	$RMSE_{\theta_0}$	0.4597	0.4656	0.4582	0.4586	.....
	Case-(ii): $logit(\pi_{it}) = \gamma y_{it} + d_{it}$ and $logit(\pi_{i0}) = \gamma_0 y_{i0} + d_{i0}$					
$\hat{\theta}$	*Mean Est.*	12.6907	12.3915	12.4417	12.4364	12.1394
	Bias	0.6863	0.3870	0.4373	0.4320	0.1350
	Emp. SE	0.0761	0.0835	0.0816	0.0818	0.1142
	Emp. RMSE	0.6905	0.3959	0.4448	0.4397	0.1769
	$RMSE_{\theta_0}$	0.7294	0.4098	0.4625	0.4415	.....
$\hat{\sigma}^2$	*Mean Est.*	29.8130	29.6781	29.7246	29.7149	30.2354
	Bias	1.2479	1.3828	1.3363	1.3460	0.8253
	Emp. SE	0.5200	0.5742	0.5591	0.5614	0.6023
	Emp. RMSE	1.3519	1.4973	1.4485	1.4584	1.0218
	$RMSE_{\theta_0}$	1.9005	1.7200	1.7069	1.7140	.....

The estimates for γ are obtained only for BB approach for both case. For first cases $\hat{\gamma}$=0.1301, $Bais(\hat{\gamma})$=0.0051, $Emp.\ SE(\hat{\gamma})$=0.0072 and $Emp.\ RMSE(\hat{\gamma})$=0.0088. For second case $\hat{\gamma}$=0.12037, $Bais(\hat{\gamma})$=0.00037, $Emp.\ SE(\hat{\gamma})$=0.0086 and $Emp.\ RMSE(\hat{\gamma})$=0.0086.

From above table, the naive approach exhibits bias. While the bias seems small in absolute value, it actually dominates the SE. The EE approach is performing well under the postulated model i.e. case-(i) and BBL is performing well under the postulated model i.e. case-(ii). The bias of EE under case-(ii) is still smaller than the naive it means that it is not that sensitive. But the bias of BBL under case-(i) is larger than the naive it means that it is sensitive.

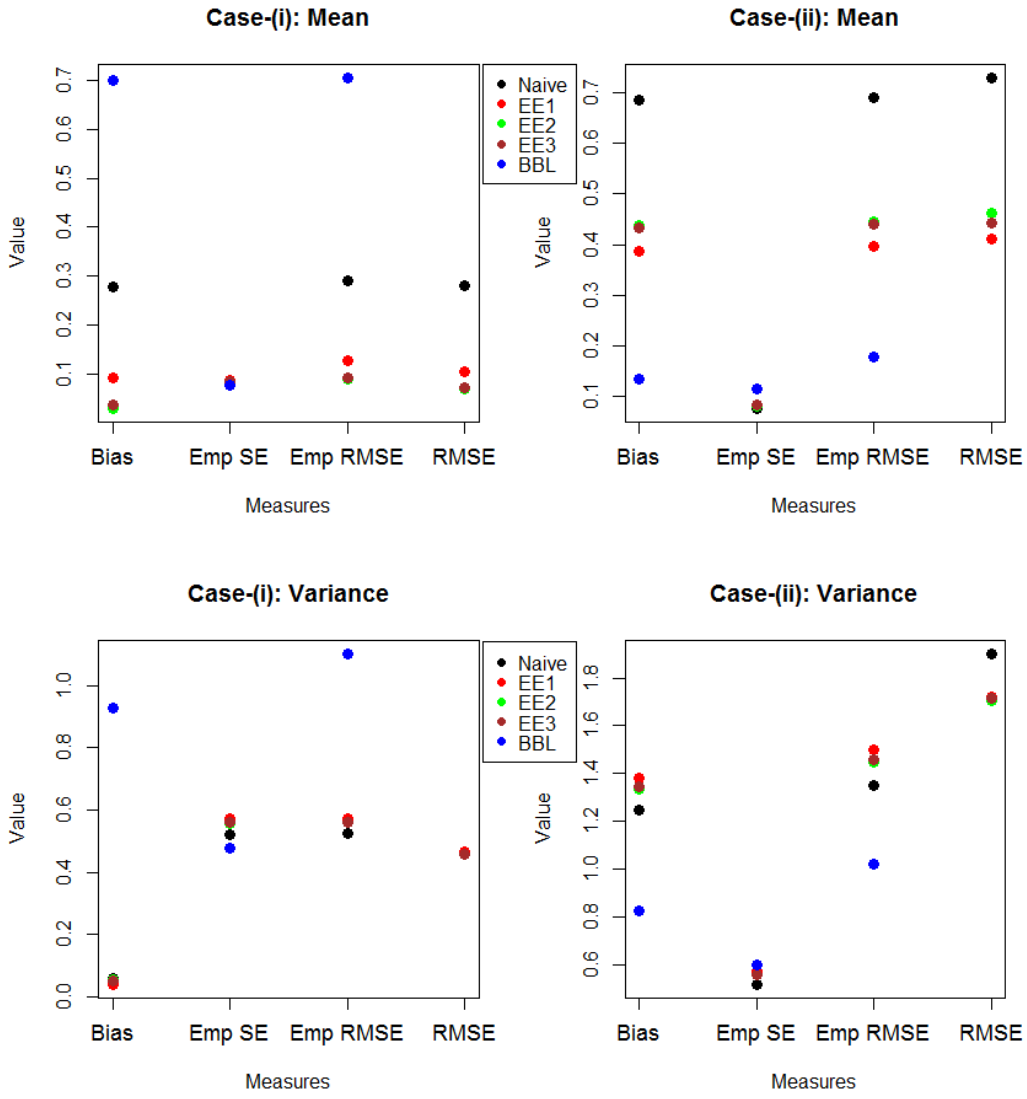


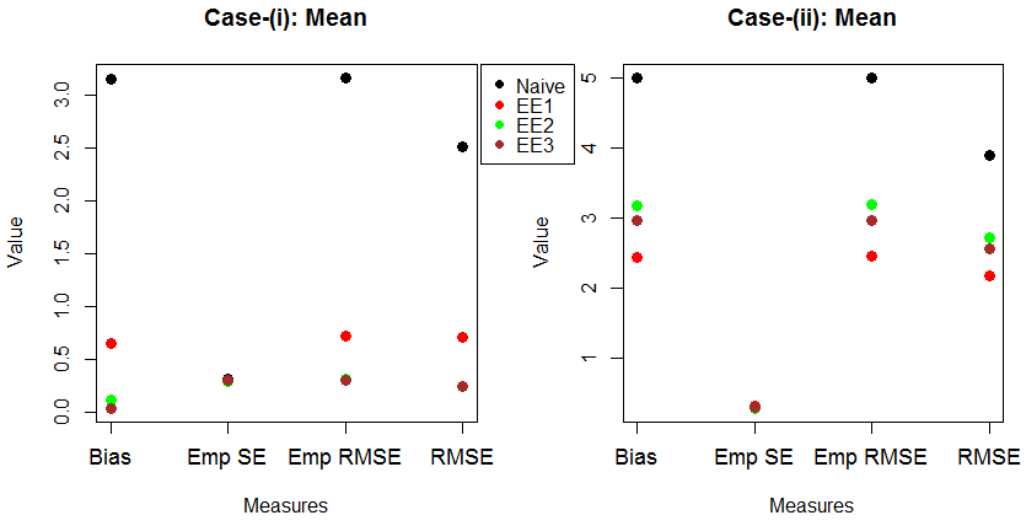
Figure 1: Graphical comparison of different estimation approaches

Next we looked at more skewed data that is common in business surveys. To generate skewed data, the data for y_{it} are simulated using multivariate log-normal distribution with mean vector $(0.50, 0.55, 0.60, 0.65, 0.70, 0.75, 0.80, 0.85, 0.90, 0.95, 1)$ and variance covariance matrix with same covariances of 1 and variances of 3. The remaining setting is same as above case except we could not get the results for BBL approach while using log-normal to generate the data for outcome variable because *nlm* routine was producing wrong estimates. We will try later to get rid of this computational bug. The results for naive and EE approach are given in the following table.

Table 3: Estimate, bias, empirical SE, RMSE and average of squared error for both cases
($\theta_{FP} = 13.06$, and $\sigma^2_{FP} = 1422.3$; $N = 1000$ and $S = 200$)

		Nave	$\hat{H}_N(\theta)$	$\hat{H}^*_N(\theta)$	$\hat{H}^{**}_N(\theta)$
	Case-i: $logit(\pi_{it}) = \gamma\eta_i + d_{it}$ and $logit(\pi_{i0}) = \gamma_0\eta_i + d_{i0}$				
$\hat{\theta}$	*Mean Est.*	16.2059	12.3999	13.1814	13.0147
	Bias	3.1502	0.6559	0.1257	0.0410
	Emp. SE	0.3237	0.3056	0.2972	0.3012
	Emp. RMSE	3.1668	0.7235	0.3227	0.3040
	$RMSE_{\theta_0}$	2.5208	0.7132	0.2511	0.2455
$\hat{\sigma}^2$	*Mean Est.*	1909.9334	1340.9651	1445.8350	1426.3838
	Bias	488.0298	80.9384	23.9315	4.4803
	Emp. SE	38.5870	33.4300	33.4789	33.6360
	Emp. RMSE	489.5529	87.5705	41.1528	33.9331
	$RMSE_{\theta_0}$	283.0270	94.4655	31.7511	27.1976
	Case-ii: $logit(\pi_{it}) = \gamma y_{it} + d_{it}$ and $logit(\pi_{i0}) = \gamma_0 y_{i0} + d_{i0}$				
$\hat{\theta}$	*Mean Est.*	18.0508	15.5006	16.2305	16.0131
	Bias	4.9950	2.4449	3.1748	2.9574
	Emp. SE	0.2919	0.3071	0.2818	0.2907
	Emp. RMSE	5.0036	2.4641	3.1872	2.9717
	$RMSE_{\theta_0}$	3.9017	2.1820	2.7167	2.5562
$\hat{\sigma}^2$	*Mean Est.*	2019.1321	1621.6746	1722.9971	1696.8667
	Bias	597.2286	199.7711	301.0936	274.9632
	Emp. SE	34.7365	34.6352	32.6455	33.4409
	Emp. RMSE	598.2379	202.7513	302.8582	276.9893
	$RMSE_{\theta_0}$	310.7262	150.4409	203.8914	191.3530

From above table, in the case of skewed data EE is performing far better than naive unadjusted reporting population estimates.



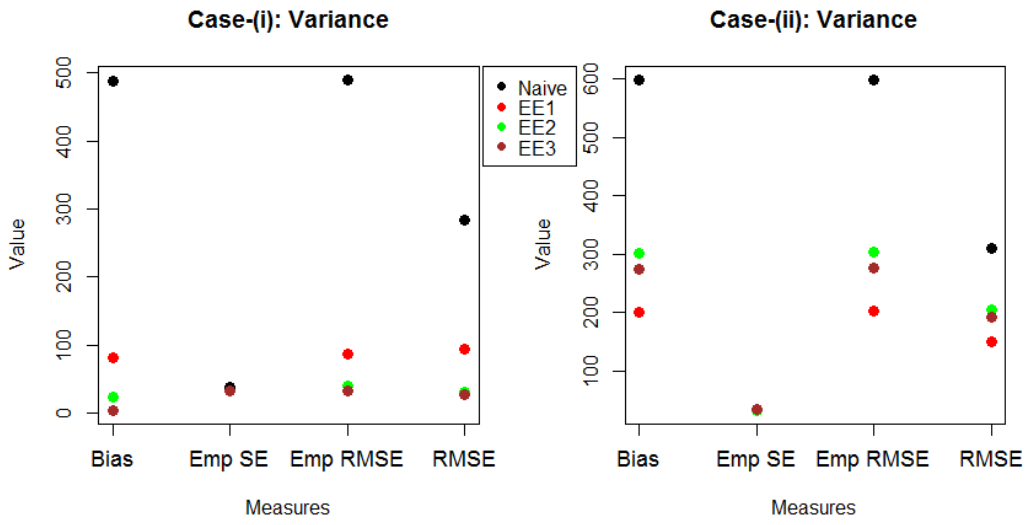


Figure 2: Graphical comparison of different estimation approaches

2.4.2 Estimation of Model Parameters

To illustrate the BBL and estimating equations approach in the case of estimation of model parameters, the data for y_{it} is simulated using the model, $y_{it} = \beta_0 + \beta_1 y_{it-1} + \epsilon_{it}$, where y_{it-1} is the response variable of previous time period and $\epsilon \sim N(0, \sigma_{it}^2)$. It is assumed that there is full reporting in the previous time period for response variable. We considered 11 time periods $(t = 0, 1, 2, ..., 10(= T))$, where $t = 0$ is the current time period. The y_{i11} is generated from log-normal distribution with mean 1 and variance 2. The y_{i11} is used as covariate for y_{i10} and further y_{i10} is used as covariate for y_{i9} and so on y_{i1} is used as covariate for y_{i0}. To generate the values of y_{it} for 11 time periods including the current time period, the $\beta_0 = (1.5, 1.6, 1.7, 1.8, 1.9, 2.0, 2.1, 2.2, 2.2, 2.4, 2.5)$, $\beta_1 = (0.05, 0.15, 0.25, 0.35, 0.45, 0.55, 0.65, 0.75, 0.85, 0.95, 1)$ and $\sigma_{it}^2 = \sigma^2 y_{i(t-1)}^{0.75}$ with $\sigma^2 = 1.25$. The data for $d_{it,1}$ are obtained similar to above population prediction case. The reporting probabilities are generated using the model for Case-(1) and Case-(ii) given in table 1. The following figure is given to assess the informativeness when η_i is used instead of y_{i0}

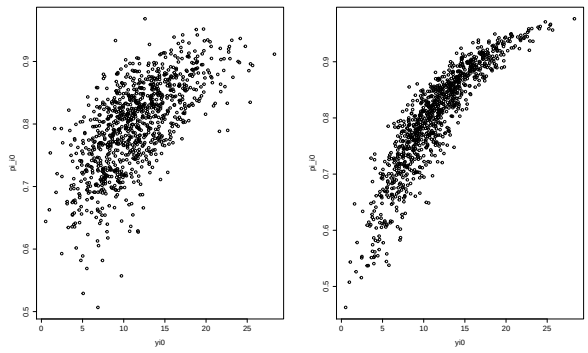


Figure 3: Assessment of Informativeness

From Figure 3, one can see that using η_i instead of y_{i0} gives less pronounced informative reporting pattern.

The population data is generated once and kept fixed for all simulations. The size of population is assumed to be $N = 1000$ and no of simulations are 200. The parameters are estimated using BBL approach with likelihood function given in (10) and estimating equations given in (19), (20) and (21). The results are given below.

Table 4: Estimate, bias, empirical SE & RMSE, and average of squared error for Case-(i)
(Finite Population Parameters: $\beta_{0_{FP}} = 2.573$, $\beta_{1_{FP}} = 0.9773$ and $\sigma^2_{FP} = 1.243$)
(Theoretical Parameters: $\beta_0 = 2.5$, $\beta_1 = 1.0$, $\sigma^2 = 1.25$ and $\gamma_0 = (0.30, 0.1)$)

		Naive	$\hat{H}_N(\theta)$	$\hat{H}^*_N(\theta)$	$\hat{H}^{**}_N(\theta)$	BBL
	Case-(i): $logit(\pi_{it}) = \gamma\eta_i + d_{it}$ and $logit(\pi_{i0}) = \gamma_0\eta_i + d_{i0}$					
$\hat{\beta}_0$	*Mean Est.*	2.6299	2.5001	2.5309	2.5257	2.6266
	Bias	0.1299	0.0725	0.0417	0.0469	0.1394
	Emp. SE	0.1048	0.1111	0.1091	0.1094	0.1329
	Emp. RMSE	0.1669	0.1326	0.1168	0.1190	0.2096
	$RMSE_{\theta_0}$	0.0955	0.1097	0.0964	0.0983	.....
$\hat{\beta}_1$	*Mean Est.*	0.9743	0.9819	0.9797	0.9802	0.9868
	Bias	0.0257	0.0046	0.0024	0.0028	0.0143
	Emp. SE	0.0110	0.0117	0.0115	0.0116	0.0134
	Emp. RMSE	0.0280	0.0126	0.0118	0.0119	0.0215
	$RMSE_{\theta_0}$	0.0092	0.0104	0.0097	0.0098	.....
$\hat{\sigma}^2$	*Mean Est.*	1.2360	1.2473	1.2454	1.2457	1.2184
	Bias	0.0140	0.0040	0.0022	0.0024	0.0325
	Emp. SE	0.0277	0.0302	0.0295	0.0296	0.0256
	Emp. RMSE	0.0311	0.0304	0.0296	0.0297	0.0446
	$RMSE_{\theta_0}$	0.0254	0.0239	0.0234	0.0235	.....

Table 5: Estimate, bias, empirical SE & RMSE, and average of squared error for Case-(ii)
(Finite Population Parameters: $\beta_{0_{FP}} = 2.573$, $\beta_{1_{FP}} = 0.9773$ and $\sigma^2_{FP} = 1.243$)
(Theoretical Parameters: $\beta_0 = 2.5$, $\beta_1 = 1.0$, $\sigma^2 = 1.25$ and $\gamma_0 = (0.30, 0.1)$)

		Naive	$\hat{H}_N(\theta)$	$\hat{H}^*_N(\theta)$	$\hat{H}^{**}_N(\theta)$	BBL
	Case-(ii): $logit(\pi_{it}) = \gamma y_{it} + d_{it}$ and $logit(\pi_{i0}) = \gamma_0 y_{i0} + d_{i0}$					
$\hat{\beta}_0$	*Mean Est.*	2.7833	2.6218	2.6611	2.6527	2.4811
	Bias	0.2833	0.0492	0.0885	0.0801	0.0189
	Emp. SE	0.1124	0.1163	0.1140	0.1145	0.1444
	Emp. RMSE	0.3048	0.1263	0.1443	0.1397	0.1564
	$RMSE_{\theta_0}$	0.2226	0.1026	0.1198	0.1156	.....
$\hat{\beta}_1$	*Mean Est.*	0.9697	0.9832	0.9803	0.9808	0.9868
	Bias	0.0303	0.0059	0.0030	0.0035	0.0135
	Emp. SE	0.0110	0.0112	0.0111	0.0111	0.0190
	Emp. RMSE	0.0322	0.0126	0.0115	0.0116	0.0231
	$RMSE_{\theta_0}$	0.0114	0.0101	0.0093	0.0094	.....
$\hat{\sigma}^2$	*Mean Est.*	1.2198	1.2545	1.2431	1.2461	1.2391
	Bias	0.0302	0.0112	0.0001	0.0028	0.0108
	Emp. SE	0.0288	0.0364	0.0327	0.0338	0.0338
	Emp. RMSE	0.0418	0.0381	0.0327	0.0339	0.0355
	$RMSE_{\theta_0}$	0.0571	0.0327	0.0330	0.0326	.....

The estimates for γ are obtained only for BB approach for both case. For first case $\hat{\gamma}$ = 0.1308, $Bais(\hat{\gamma}) = 0.0048$, $Emp.\ SE(\hat{\gamma}) = 0.0071$ and $Emp.\ RMSE(\hat{\gamma}) = 0.0095$. For second case $\hat{\gamma} = 0.1203$, $Bais(\hat{\gamma}) = 0.0019$, $Emp.\ SE(\hat{\gamma}) = 0.0081$ and $Emp.\ RMSE(\hat{\gamma}) = 0.0084$. For regression analysis, BBL seems slightly sensitive under case-(i) and EE approach seems slightly sensitive under case-(ii). Bias remains the dominate error under the naive approach.

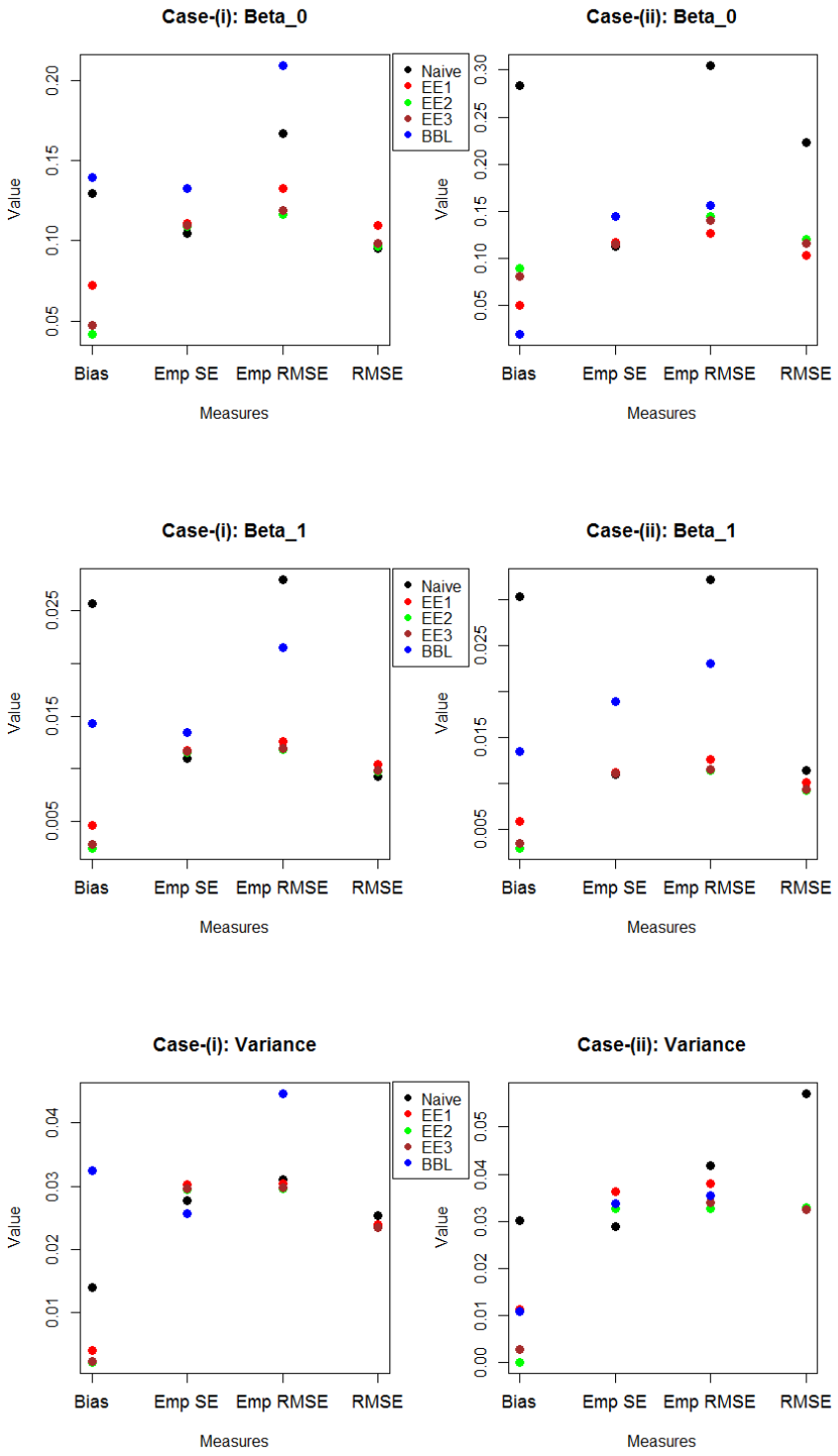


Figure 4: Graphical comparison of different modelling approaches

3. Conclusions

In conclusion, the results of prediction and estimation of model parameters suggest that if sufficient long reporting history exists in longitudinal data that can provide an empirical estimator of reporting probability, one can use EE approach as an alternative to likelihood approaches for prediction as well as for the estimation of model parameters under informative non-response. This approach is simple, computationally easy and free from strict distributional assumptions of the outcome variable. However, when the report probabilities depend on the concurrent value of the target variable, the more sophisticated BBL may be necessary. In practice, one could apply the BBL for every year separately and test the conditions underlying the use of the EE. But applying separate BBL to test the constancy of reporting probability over time is possible in theory but troublesome in practice. One could check this assumption even using EE approach by comparing the estimated reporting probability for different span of consecutive historic time periods depending on the length of history or observing the reporting rate over time. We intend to investigate these possibilities more closely in future.

References

Beaumont, J. (2000). An estimation method for nonignorable nonresponse, *Survey Methodology* **26**: 131–136.

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

Chambers, R. and Skinner, C. (2003). *Analysis of survey data*, John Wiley and Sons, Inc., New York.

Diggle, P. and Kenward, M. (1994). Informative dropout in longitudinal data analysis (with discussion), *Appl. Stat.* **43**: 49–94.

Feder, M. and Pfeffermann, D. (2015). Statistical inference under non-ignorable sampling and non-response. an empirical likelihood approach, *Southampton, GB, University of Southampton* .

Godambe, V. P. (1991a). *Estimating Functions*, Oxford Science Publications, New York.

Godambe, V. and Thompson, M. (2009). *Estimating functions and survey sampling. In Handbook of Statistics 29B; Sample Surveys: Inference and Analysis, (Eds., D. Pfeffermann and C.R. Rao)*, Amsterdam, North Holland.

Greenlees, J. S., Reece, W. S. and Zieschang, K. D. (1982). Imputation of missing values when the probability of response depends on the variable being imputed, *J. Amer. Statist. Assoc.* **77**: 251–261.

Hedlin, D., Fenton, T., McDonald, J. W., Pont, M. and Wang, S. (2006). Estimating the under coverage of a sampling frame due to reporting delays, *Journal of Official Statistics* **22**: 53–70.

Ibrahim, J., Chen, M. and Lipsitz, S. (2001). Missing responses in generalized linear mixed models when the missing data mechanism is nonignorable, *Biometrika* **88**: 551–564.

Ibrahim, J., Lipsitz, S. and Chen, M. (1999). Missing covariates in generalized linear models when the missing data mechanism is non-ignorable, *J. R. Stat. Soc. Ser. B* **61**: 173–190.

Kasprzyk, D., Duncan, G., Kalton, G. and Singh, M. (1989). *Panel Surveys*, John Wiley and Sons, Inc., New York.

Linkletter, C. D. and Sitter, R. R. (2007). Predicting natural gas production in texas: An application of nonparametric reporting lad distribution estimation, *Journal of Official Statistics* **23**: 239–251.

Little, R. (1993). Pattern-mixture models for multivariate incomplete data, *Journal of the American Statistical Association* **88**: 125–134.

Little, R. (1995). Modeling the drop-out mechanism in repeated-measures studies, *Journal of the American Statistical Association* **20**: 1112–1121.

Little, R. and Rubin, D. (2002). *Statistical Analysis With Missing Data*, John Wiley and Sons, Inc., New York.

Molenberghs, G. and Verbeke, G. (2001). A review on linear mixed models for longitudinal data, possibly subject to dropout, *Statist. Model.* **1**: 235–269.

Pfeffermann, D. (1993). The role of sampling weights when modelling survey data, *International Statistical Review* **61**: 317–337.

Pfeffermann, D. (1996). The use of sampling weights for survey data analysis, *Statistical Methods in Medical Research* **5**: 239–261.

Pfeffermann, D. (2011). Modelling of complex survey data: Why model? why is it a problem? how can we approach it?, *Survey Methodology* **37**: 115–136.

Pfeffermann, D. and Landsman, V. (2011). Are private schools better than public schools? appraisal for ireland by methods for observational studies, *The Annals of Applied Statistics* **5**: 1726–1751.

Pfeffermann, D. and Sikov, N. (2011). Imputation and estimation under nonignorable nonresponse in household surveys with missing covariate information, *Journal of Official Statistics* **27**: 181–209.

Pfeffermann, D. and Sverchkov, M. (2009). *Inference under Informative Sampling. In Handbook of Statistics 29B; Sample Surveys: Inference and Analysis, (Eds., D. Pfeffermann and C.R. Rao)*, Amsterdam, North Holland.

Qin, J., Leung, D. and Shao, J. (2002). Estimation with survey data under nonignorable nonresponse or informative sampling, *Journal of the American Statistical Association* **97**: 193–200.

Rubin, D. (1987). *Multiple imputation for nonresponse in surveys*, John Wiley and Sons, Inc., New York.

Sinha, S., Laird, N. and Fitzmaurice, G. (2010). Multivariate logistic regression with incomplete covariate and auxiliary information, *J. Multivariate Anal.* **101**: 2389–2397.

Sinha, S., Troxel, A., Lipsitz, S., Sinha, D., Fitzmaurice, G., Molenberghs, G. and Ibrahim, J. (2011). A bivariate pseudo-likelihood for incomplete longitudinal binary data with nonignorable non-monotone missingness, *Biometrics* **67**: 1119–1126.

Skinner, C., Holt, D. and Smith, T. (1989). *Analysis of complex surveys*, John Wiley and Sons, Inc., New York.

Statistics Denmark (1995). *Statistics on persons in Denmark a register based statistical system*, Eurostat, Luxembourg.

Statistics Finland (2004). *Use of registers and administrative data sources for statistical purposes best practices in Statistics Finland, Handbook 45*, Statistics Finland, Helsinki.

Valliant, R., Dorfman, A. H. and Royall, R. M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*, John Wiley and Sons, Inc.

Verbeke, G. and Molenberghs, G. (2005). Longitudinal and incomplete clinical studies, *Metron* **63**: 143–170.

Wallgren, A. and Wallgren, B. (2007). *Register based statistics Administrative data for statistical purposes*, John Wiley and Sons, Chichester.

Wu, L., Liu, W. and Liu, J. (2009). A longitudinal study of childrens aggressive behaviours based on multivariate mixed models with incomplete data, *Metron* **37**: 435–452.

Xie, H. (2008). A local sensitivity analysis approach to longitudinal non-gaussian data with non-ignorable dropout, *Stat. Med.* **27**: 3155–3177.

Yi, G. and Cook, R. (2002). Marginal methods for incomplete longitudinal data arising in clusters, *Journal of the American Statistical Association* **97**: 1071–1080.

Zhang, L.-C. and Fosen, J. (2012). A modelling approach for uncertainty assessment of register based small area statistics, *Journal of the Indian Society of Agricultural Statistics* **66**: 91–104.

Zhang, L.-C. and Pritchard, A. (2013). Short-term turnover statistics based on vat and monthly business survey data sources, *ENBES workshop*, Nuremberg.