UNIVERSITY OF SOUTHAMPTON

FACULTY OF SOCIAL AND HUMAN SCIENCES

Social Sciences

Adjusting for Nonresponse in the Analysis and Estimation of Sample Survey Data for Cluster Designs

by

Nuanpan Nangsue

Thesis for the degree of Doctor of Philosophy

June 2014

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF SOCIAL AND HUMAN SCIENCES
Social Sciences

Doctor of Philosophy

ADJUSTING FOR NONRESPONSE IN THE ANALYSIS AND ESTIMATION OF SAMPLE SURVEY DATA FOR CLUSTER DESIGNS

by Nuanpan Nangsue

Nonresponse in sample surveys has been increasing over the years. This thesis covers that issue in two main parts. The first part is concerned with how to use observed data to make inference about regression coefficients in a linear regression model of cluster-level variables when some of the response variable data is missing. A naïve approach estimates the regression coefficients without considering nonresponse. We propose new methods for estimating coefficients which incorporate information on nonresponse at the cluster level. We also extend Heckman estimators to our clustered model. The Workplace Employment Relations Survey (WERS) 2004 data and data from a prepared simulation study are used to compare the new methods with the naïve approach. In the second part the generalized regression estimator (GREG) for two-stage sampling will be considered. We propose new optimum GREG estimators for stratified two-stage sampling and a simulation study is used in order to assess the performance of the new estimators.

Contents

List of Figures

List of Tables

Declaration of Authorship

I, Nuanpan Nangsue , declare that the thesis entitled *Adjusting for Nonresponse in the Analysis and Estimation of Sample Survey Data for Cluster Designs* and the work presented in the thesis are both my own, and have been generated by me as the result of my own original research. I confirm that:

- this work was done wholly or mainly while in candidature for a research degree at this University;
- where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- where I have consulted the published work of others, this is always clearly attributed;
- where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
- none of this work has been published before submission

Signed:..

Date:..

Acknowledgements

I wish to thank my academic supervisor and advisors, Dr. Yves Berger, Prof. Chris Skinner and Prof. Natalie Shlomo for their guidance, direction and support which enabled me to complete my research and write this thesis. Finally, I would like to give special thanks to my friends and family, husband Darren and daughter Naamning and also to the Thai Government whose sponsorship enabled me to undertake and complete my studies at the University of Southampton.

Chapter 1

Introduction

1.1 Cluster Sampling in Sample Surveys

A unifying theme of this thesis will be cluster sampling. This is one of the sampling techniques that statisticians use to select sample survey data for the benefit of reducing costs. In this sampling scheme, the entire population is divided into clusters and a random sample of clusters is selected. We can collect all observations in the selected clusters or sample only some of the elements in the sample set of data (two-stage sampling). This is a useful method that does not require a sampling frame, which is a list of the members of a population, but does need a complete list of clusters of the population instead.

There are usually two different types of uses for survey data: descriptive and analytic. The descriptive use is aimed at making inferences about the whole populations, e.g. to estimate finite population totals and population means. The analytic use is aimed at studying the associations between variables of interest, e.g. regression analysis. Regarding the analytic purpose, the units of analysis may be the clusters or the elements within the clusters or they may include both. For example, if we are interested in finding the association between the level of job satisfaction of employees at workplace level and workplace innovations we would use the workplaces (clusters) as the units of analysis. We will cover both purposes, descriptive and analytic in this thesis. The analytic use will be covered in chapters 3 to 5 and the descriptive use will be considered in chapter 6.

Multilevel modelling will also be considered as it seeks to combine different units of analysis at different levels. Usually multilevel models are applied to hierarchical or clustered structure data that has more than one level. At each level multilevel modelling of data will allow for residual components. At a lower level, the individual level, are the units of analysis that can be nested within aggregate units at a higher level. For example, where the employee is the individual level he might be nested within workplaces which would be considered a higher level. This kind of data usually involves analysis using multilevel modelling.

1.2 Nonresponse in Surveys

Both estimation, for description and analysis, using survey data needs to take account of nonresponse. There are three types of nonresponse in sample surveys: unit nonresponse, item nonresponse and under-coverage. Unit nonresponse occurs when some respondents choose not to, or are unable to, take part in the study. Item nonresponse occurs when some respondents cannot answer or deny answering particular questions i.e., participants do not know how to answer or accidentally skip or refuse to answer some questions. Under-coverage occurs when some members of the population are not represented in the sample (Groves et al. (2009)).

Complex survey designs are used in large scale surveys to collect data and these methods include clustering and stratification designs, where stratification is the sampling scheme where the entire population is divided into homogeneous subgroups called strata before sampling units in each stratum. We normally analyse survey data under the complex survey design in order to do further analysis, e.g. inference to the entire population using the various statistical methodology. Unfortunately, nonresponse often occurs during the process of collecting survey data, which obviously leads to the problem and the problems arising from having missing data.

In surveys with cluster sampling it is possible for nonresponse to occur at either the cluster level or the individual level. For example, in surveys of employees within workplaces nonresponse can occur at a cluster level process or it could be raised at an individual level process. In this thesis we will consider the situation where nonresponse occurs at an individual level only.

Researchers use numerous statistical techniques that have been proposed for analysing survey data from both basic and complex approaches. The basic methods used to analyse survey data are frequency distributions and descriptive statistics. Further than that, complex statistical approaches are applied in survey research such as regression analysis and multilevel modelling. Chambers and Skinner (2003) noted the variation in generalizations of the regression analysis of survey micro data, citing generalised linear modelling, event history analysis and multi-level modelling as three different forms of analysis of the data.

Both sample survey and census always have problems of nonresponse which usually affect results of statistical analysis. In this thesis we consider nonresponse in a sample survey which impact on regression analysis and multilevel modelling for cluster data. Consequently it is essential to look at the reasons for why the data is missing.

Groves et al. (2002) state that nonresponse is of increasing concern due to decreasing response rates in surveys. Response rates are the percentage of people who respond to the survey, with one minus response rate being called the nonresponse rate. The quality of surveys is usually considered from the response rate (e.g. Bethlehem (2009), Groves et al. (2009)). Higher response rates improve the confidence that the survey results are representative of the target population. This is very important in any attempt to deal with nonresponse in surveys with complex sample designs.

The easiest way to handle missing data is to exclude missing data from the data set and continue using the standard statistical techniques to analyse as usual. In this case it is free from missing values. This is known as complete case analysis or listwise or pairwise deletion. However, there are some drawbacks for this method. For example, the estimator may be biased if the missing values are not missing completely at random (MCAR) which will be discussed later in Section 1.4.

For this reason we have to deal with missing data in more appropriate ways. There are two standard methods of compensating for nonresponse: weighting methods are used to deal with unit nonresponse and under-coverage and imputation methods are used to deal with item nonresponse. For the weighting method we can use the inverse of the probability of missingness as a weight to respondents. On the other hand, imputation methods are methods that substitute these missing values with possible estimated values from the recorded responses. Both weighting and imputation methods are used to correct the bias. More details will be shown in Chapter 2.

Missing data also impacts on statistical analysis depending on the missing data pattern and the missing data mechanism. The missing data pattern explains whether values are observed or missing. The missing data mechanism is concerned with the relationship between the reasons why data is missing and the values of variables.

It is helpful to understand the missing data pattern and missing data mechanism in order to determine the suitable approach to use for statistical analysis and may also help in finding a solution for the problems that arise from the missing data. Our analysis involves unit nonresponse which requires an assumption of the missing data mechanism. This Chapter provides general ideas about missing data in surveys in order to understand the nature of nonresponse.

1.3 Survey Estimation

Sample survey data usually has problems of nonresponse that can lead to a biased estimator. Survey weighting is one of the statistical techniques that we often use to correct the bias brought by nonresponse and to improve the efficiency of the estimators for unit nonresponse. There are numerous ways to construct the survey weights. An important way to do this is to use an inverse probability weighting which is a weighting reciprocal to a response probability that has been estimated under a model. (Skinner and Darrigo (2011)).

The Horvitz-Thompson estimator is one of the weighted estimators used for estimating a population total or a population mean. The Horvitz-Thompson estimator is usually used for any probability sampling plan e.g. simple random sampling, stratified sampling and it is also used for accounting for nonresponse. The inverse of the known sampling probability which is drawn from the target population is used for observation weighting.

Auxiliary information can also be used to improve the precision of the estimator, e.g. estimation of the population totals or population mean while the Horvitz-Thompson estimator does not consider the auxiliary information into the estimation steps. The generalized regression (GREG) estimator is one of the methods that uses auxiliary information to improve the precision of the estimations of population totals or population means in survey sampling and it is a special type of calibration estimator which will be discussed in Chapter 6.

Another survey estimation that corrects for sample selection bias is the Heckman model (Heckman (1976), Heckman (1979)). The Heckman model, sometimes called the sample selection model, is a method for estimating regression coefficients that allow for bias selection. The dependent variable in the Heckman model is only observed for a portion of the data and the remaining is unobserved.

1.4 Missing Data Mechanisms

The easiest way for analysing missing data is to ignore unmeasured data and to analyse using the complete set of data. However, this may lead to biased estimators. There are some options to cope with the analysis of incomplete data. In statistical software, listwise and pairwise deletions are the methods for handing missing data by deleting any case with missing data before analysis. To understand these options, we need to understand the reasons why data is missing through missing data mechanism, (refer Little and Rubin (2002) and Groves et al. (2002)). Missing data mechanisms are classified into three categories depending on randomness or non-randomness of missing data. The first one is missing completely at random (MCAR). The data are MCAR if missingness does not depend on observed values and missing values. For example, if we consider the job satisfaction of employees and workplace innovations, the chance to have a missing value on job satisfaction does not depend on the actual value of innovation. The probability of response is equal on all categories of innovations. This is a very strong assumption, and usually does not hold true in practice when using the complete set of data.

The second mechanism which has a less restrictive assumption is called missing at random (MAR). This occurs if the missingnesses is related to the observed data but not on the missing values. Both listwise and pairwise deletion assumes that the missing data mechanism is MCAR. Therefore, if the missing data are MAR, analyzing by these methods may lead to biased estimators. According to the previous example, the chance to have a missing value on job satisfaction may depend on the observed value of innovation but not on the missing value of job satisfaction. In other words within categories of innovation, the probability to respond on job satisfaction is equal but maybe different accross categories of innovation. Analyses of complete cases will bias estimates if MAR. Missing data is called ignorable, if the data are MCAR or MAR

The last mechanism is not missing at random (NMAR). The mechanism is called NMAR, if the missingnesses is related to the missing data itself. Therefore, following the previous example the chance to have a missing value on job satisfaction may depend on the observed value of innovation and also depend on the missing value of job satisfaction. The response probabilities vary both between and within the categories on innovation. When missing data are NMAR, it is very difficult to handle and may lead to severe biased estimates. Missing data are also called non-ignorable, if the data is NMAR.

1.5 Outline of Thesis

The remainder of this thesis is divided into two main parts. As we mentioned earlier, we will cover both descriptive and analytic uses of survey data. The analytic use of survey data will be presented across chapters 3 to 5 and the descriptive use of survey data will be presented in chapter 6 where it is quite different from the other parts of the thesis.

In Chapter 2, we review the literature on approaches to compensate for nonresponse. We also provide an overview of nonresponse in sample surveys. For example, imputation and weighting methods to compensate for nonresponse and also types of nonresponse are described. Finally, the clustered data in survey sampling is discussed.

In Chapter 3, the model of interest is introduced, where assumptions of MAR and NMAR mechanisms are considered. We discuss the estimation of a regression coefficient (β) in a regression model under a naïve approach and introduce new estimators. The bias and variance for both estimators are discussed. We also consider a Heckman estimator. Furthermore, we show that the new methods we propose can be extended to two -stage cluster sampling.

In Chapter 4, we look at the performance of the proposed estimators and Heckman estimators through a simulation study and discuss a model for the simulation study. The simulation results are shown to support the theory. Moreover, we discuss further theory to explain the simulation results.

In Chapter 5, we test our approach on real data. We choose the Workplace Employment Relations Survey (WERS) data where a two-stage cluster sampling design is used where there is also the problem of employee nonresponse in the workplace level. We also discuss the study of Bryson et al. (2009). The proposed analysis is discussed and finally the results of analysis for both the regression model at individual level and at cluster level are presented.

In Chapter 6, we propose the optimum GREG estimators for stratified two-stage cluster sampling. First of all, we review literature that relate to the GREG estimators. Next we discuss the GREG estimators at single stage and two stage respectively. Finally, we look at the performance of the proposed estimators by simulation study.

In the final chapter, Chapter 7, we will consider the full value of our proposed alternative estimator, our developed Heckman estimators and our new GREG estimator leading onto a consideration of possible future works.

Chapter 2

Review of Literature on Analysis of Missing Data

2.1 Introduction

In Chapter 1, we discussed cluster sampling in sample surveys, nonresponse in surveys, survey estimation and described nonresponse mechanisms in sample surveys. In this Chapter, we will review the literature surrounding the methods used for compensating for nonresponse; weighting and imputation. We will describe the literature surrounding survey analysis that deals with missing data in clustered data.

As noted in Chapter 1, missing data is a common problem in survey data. Numerous methods have been developed in order to deal with incomplete survey data; weighting methods are used to compensate for unit nonresponse and for under-coverage while imputation methods are used mainly to compensate for item nonresponse.

2.2 Imputation

Imputation consists of replacing missing data values with plausible estimated values. For single imputation each missing value is replaced with a single estimated value. There are several single imputation techniques used to account for missing data including mean imputation, regression imputation and hot deck imputation.

Mean imputation replaces the missing value by the mean of the existing values of the same field in a data set typically within nonresponse classes. Using the regression imputation method we replace the missing values in a different way with predicted values based on the regression model applied on observed elements (Kalton and Kasprzyk (1982); Little and Rubin (2002); Durrant (2005); Heeringa et al. (2010)).

There are some papers concerning mean and regression imputation. Haziza and Rao (2006) proposed a new approach to contest linear regression imputation for estimating population totals. This new method is assumes an ignorable response mechanism that does not require a model on the variable of interest. They showed that along with the imputed values the estimator of the population totals is also approximately unbiased and show that the variance estimators of the population totals are approximately unbiased following Fay (1991). Their simulation results show that their proposed methods perform better than the naive approach in terms of both bias and mean square error but nevertheless, their simulation study to assess their variance estimators performance is limited with a high regression coefficient of determination appearing in the study model and also, worse still, their research is limited to a single imputation class only.

Hot-deck imputation which replaces missing data with donor information from a respondent who carries similar characteristics as the recipient is another approach for single imputation. A donor can be selected randomly from within an imputation class or from a similarly observed value in a suitable record (Kalton and Kasprzyk (1982); Little and Rubin (2002); Durrant (2005)). Hot-deck imputations are also useful techniques for extrapolating multivariate missing data, as it is common to have incomplete sets of data in many variables where multivariate imputations are needed. In this case, one donor is used for all missing values. Andridge and Little (2010) show different formats of hot deck imputation methods and also provide a detailed review of hot deck properties showing methods to create the donor pool, considering the weights in sample survey for the selection of the donor and also they describe hot deck imputations for multivariate missing data. The variance estimators used to validate hot deck imputation with a detailed description of each are considered and applied to real data. They also gave their views about problems they found with using hot-deck imputation and it might be advantageous for future investigation to address these issues.

Other papers on multivariate imputation are described in Van Buuren et al. (2006). Van Buuren et al. (2006) studies the issue that occurs using the fully conditional specification in the Gibbs sample for imputation for nonresponse in multivariate data. The trouble in theory for this method is that it might not converge under inconsistent conditions. Van Buuren et al. (2006) investigated this problem of non convergence under various missing data mechanisms, e.g. MAR, MCAR for both compatible and incompatible models. The results show that multiple imputation still performs well with incompatible models and it provides an unbiased estimator. A limitation of this research on imputation models is that it only explored MAR models and consequently it might be interesting to apply to other models.

Single imputation has some advantages over multivariate imputation in that it allows the use of standard complete data methods in the analysis and also has the ability to incorporate the data collector's knowledge. However, the major problem of single imputation is that it does not take into account variance structures and multiple imputation (MI) does. Multiple imputation includes more than one set of missing values in a calculation. Rubin (1987) proposed multiple imputations for missing data and since then there has been much work on multiple imputation. Yuan (2000) reviews multiple imputation methods and also develops SAS procedures, PROC MI and PROC MIANALYZE for multivariate missing values. Nonetheless, Reiter et al. (2006) point out that standard software packages do not take into account complex survey designs and therefore this can lead to bias in multiple imputation data. They showed in simulation study that bias can occur in multiple imputation when the researchers ignore complex survey design, e.g. stratified and cluster samples. Two methods which account for the sampling design of the models for imputation are recommended: dummy variables and hierarchical models have been used where both the clustering effects and the stratification effects are accounted for by including random effects and fixed effects respectively. The real data is used to study the difference between imputed data that takes into account design variables while ignoring the design variables. While this study provides valuable results to show bias exists when ignoring complex survey designs in multiple imputation regarding the simulation study and real data, it is also interesting to see how that bias occurs and it should be examined in future research.

It is misleading to calculate the variance of imputed data using standard variance formula as it may lead to underestimation of the variance and gives incorrect precision in measurement. Särndal and Lundström (2005) claim that some statisticians treat the imputed data the same as observed data in order to benefit from the variance estimation. However, this leads to two problems in the estimation of variance which are a biased estimator for the sampling variance and no capacity to account for the additional component of variance due to nonresponse. Previously, Rao (1996) reviewed papers on jackknife variance estimation for a full set of data with no missing values and for a set of data with imputed missing values by a single imputation for item nonresponse under both stratified simple random sampling and stratified multistage sampling. The linearized forms of the jackknife variance estimators are investigated. These can be used by adjusting existing computer software without adding extra code. However, his study is limited to single imputation. Later Berger and Rao (2006) proposed an adjusted jackknife estimator for variance estimation in the case of imputed data under unequal probability sampling without replacement and with non negligible sampling fractions. Three imputation methods are considered; mean, ratio and random imputation. A simulation study is used to compare the proposed method with the naïve jackknife variance estimator. It shows that the proposed estimator performs better than the naïve one in terms of relative bias, relative root mean square error and confidence interval empirical coverage. Their study is limited to single imputation class with uniform response and even though they extended to non-uniform response they still only tested with a single imputation class.

Moreover, Shao and Steel (1999) proposed a new method of variance decomposition for estimating the population totals of the Horvitz-Thomson estimators in two cases using sample surveys where imputation techniques have been utilised upon areas of nonresponse and where there are non-negligible sampling fractions. They showed that their new method can be used to derive variance estimators for both a design-based or a model-assisted approach even where some imputation methods, e.g. deterministic and/or random imputation methods have been used. The variance estimators derived from the new method also hold the properties of being asymptotically unbiased and consistent. They discovered difficulty in using the model-assisted approach over the design-based approach when certain factors are present in the construction of the survey. Later Shao and Wang (2002) examined both bias and variance for the regression imputation method. They also proposed a joint regression imputation method that is an unbiased estimator for marginal totals, second moments and correlation coefficients. A jackknife variance estimation method which takes into account the imputation method is proposed to produce asymptotically unbiased and consistent variance in estimation. The simulation results show that the proposed method performs well in terms of unbiased point estimators and at the estimation of variance.

Brick et al. (2005) compared three methods; the model-assisted, the adjusted jackknife and the multiple imputation methods for variance estimations of the population total under hot deck imputation using a simulation study. They considered both full population and domain estimation under their study where missing variables data is imputed by hot deck imputation under a single-stage stratified sampling design that assumes that the missing at random assumption for hot deck cells of both unbiased estimator and item response are violated. The simulation results show that all variance estimators under hot deck imputation give unbiased estimators for full population but not for domains. Their simulation study however is limited to a single-stage stratified sampling design and also they did not compare the different variance estimations in theory with a mathematical proof.

2.3 Weighting

Weighting methods are used in sample survey data using design weights to scale sample units. It is also used for addressing unit nonresponse in missing data. The weighting adjustment for missing data is a method that allocates a nonresponse weight to respondents based on the inverse of the probability of missingness. This method should decrease nonresponse bias. Although the pros of using weights in survey estimates is to reduce bias from nonresponse, there are also some cons in weighting. It may lead to increased variance in the estimation and complications in analysis due to unrecognised weights in statistical packages which might instead treat them as a frequency weight of the same value (Brick and Kalton (1996)). There are a variety of methods for weighting adjustment for missing data such as inverse probability weighting, adjustment cells weighting and poststratification. Under poststratification and adjustment cell weighting, the weighting classes should be homogenous with respect to the probability to respond in order to reduce nonresponse bias but result in less of an increase in the variance compared to inverse probability weighting .

Adjustment cells weighting is one of the weighting methods that estimates the nonresponse probability from sample data, and also reduces nonresponse bias. There are two methods to create the adjustment cells or subclasses. The first one depends on respondents and non-respondents variables in the survey. The second one depends on external information from a census or larger survey. We assume that respondents and non-respondents are classified into C adjustment cells depending on covariate information. Little and Vartivarian (2003) claim that “the respondents in cell C are weighted by the inverse of the response rate in cell C”. Groves et al. (2002) state that this method is called response propensity stratification which is a method that is effective in reducing nonresponse bias with respect to the background variables.

The poststratification method is a weighting method that needs one or more qualitative auxiliary variables and a population total. It is one of many methods that use the benefit of auxiliary information. This method assigns the different adjustment weights to all units in the same poststratum. It will give unbiased estimators at full response but may be biased if nonresponse occurs (Groves et al. (2002)). Holt and Smith (1979) state that poststratification or stratification after selection is a robust method for estimation, and also unrestricted by assumptions. They point out that if the stratification prior to selection depends on a large number of secondary variables, then post stratification is useful for multi-purpose surveys. However, their discussion is limited to a single stage design.

The difference between inverse probability weights and post stratification weights is stated by Groves et al. (2002). He points out that the difference between these two methods is that post stratification weights come from collected data but that inverse probability weights have already been informed upon survey design. However, inverse probability may cause severe problems. For example, if nonresponse data is often ignored the method will be inefficient. Moreover, the variance estimation problem for estimated weights is covered in explicit formulas only for simple estimators which are not well established for complex survey designs. There is also another approach to calculate the weights e.g. calibration approach which uses the benefit of the auxiliary variables (Särndal and Lundström (2005)).

Some papers use both weighting and imputation to address missing data problems. Brick and Kalton (1996) explore several weighting and imputation techniques and also the advantages and disadvantages of them. Durrant and Skinner (2006) study imputation and weighting methods looking at ways to remove bias that results from measurement errors made estimating of a distribution function, for example fractional imputation, nearest neighbour imputation, predictive mean matching and propensity score weighting. The results show that nearest neighbour performs better in term of bias when compared to other imputation methods.

As we mentioned earlier, nonresponse weighting aims to decrease nonresponse bias. However, it usually causes a problem of increasing the variance. Little and Vartivarian (2005) argue that nonresponse weighting can decrease the variance the same as bias reduction in the case that there is an association between a covariate of a weighting adjustment and the survey outcome. Nevertheless, their study is limited to simple random sampling it is therefore more interesting to see how it will perform in complex survey designs.

2.4 Survey Data Analysis with Missing Data

Little (1982) reviewed many papers related to a modelling approach for handling missing data in survey sampling for both unit and item nonresponse e.g. the methods described in the papers use superpopulation models for a full set of data and also missing data in sample surveys, maximum likelihood methods that do not account for the response mechanism where cases with missing data are removed and analysis continues as though the data was complete. These analyses are called complete-case analysis or listwise deletion. Little and Rubin (2002) pointed out that there are both pros and cons of complete-case analysis. On the positive side, standard statistical techniques can be used with the complete set of data directly. In addition, it allows for the plausibility of comparable univariate statistics. However, complete-case analysis might lead to problems of bias and also loss of efficiency of variance due to throwing away some of the data. There are also some application papers covering problems of bias evident in missing data occurring by complete-case analysis.

Available-case analysis contains every single case of variable of interests. The drawback of this method exists within the pattern of incomplete data that leads to variation of the sample base from one variable to another variable (Little and Rubin (2002)).

Imputed values, as we noted in Section 2.2, are the estimated data values that fill in missing values in order to get a complete set of data for standard statistical analysis, thereby avoiding some of the problems that arise in complete-case and available-case analysis. There are papers on using imputed values for analysis in statistical applications, e.g. regression analysis and multivariate analysis. For example, Little (1992) reviewed regression analysis when independent variables are missing. He compared six methods; complete-case analysis(CC), available-case analysis(AC), least squares on imputed data, maximum likelihood(ML), Bayesian methods and multiple imputation. He pointed out that CC and AC analysis are easy but disadvantaged by the limitations within their method. ML performs well for large samples, but Bayesian methods perform better in small samples. He also discussed software for these methods. This research has limitations in the response mechanism tested using MAR model data but was not tested with other types.

Also, Skinner and Rao (2002) studied the bias found in both standard estimators for the hot-deck method and bias-adjusted estimators. They evaluated the variance for both of them and also jackknife variance estimators for each of them were produced to estimate what missing data might be in a bivariate dataset with imputed data using hot deck imputation. The limitation of their research is that simple random sampling is considered only and not even that for the more complex survey design and also their study considered data with no more than two variables.

Previously, Gelman et al. (1998) proposed a new approach for analysis of a cross-sectional survey with missing data and also a single survey with different questions or different sampling methods applied for each stratum or cluster. Multiple imputations are applied in this new method. However, a hierarchical regression model, allowing for covariate at both individual and survey level, is used for a single survey. Furthermore, the diagnostic checking for the fit of the imputed model under the difference between imputed data and non imputed data is improved.

Alternatively, another approach to deal with missing data is a likelihood method. This method provides unbiased estimators under both MCAR and MAR where complete-case and available case analysis both sometimes have problems, especially when the missing data mechanism is not MCAR. The well known method for the likelihood based approach is called the Expectation Maximization (EM) algorithm and is a method to find maximum likelihood estimates from partial data. (Little and Rubin (2002)).

There are a variety of methods of providing weighting for missing data. Skinner and Coker (1996) extend the methods for dealing with missing covariate values in a linear regression model by using a maximum likelihood estimator under a complex survey design. Also a proposed jackknife variance estimator is applied to estimate the standard errors. Their method and application to real data are restricted in that it only works with a single missing covariate. Moreover, Pfeffermann et al. (1998) presented alternative methods for weighting the estimation in a multilevel model with two-levels. The reciprocals of the selection probabilities at each stage of sampling are used to deal with the bias that occurs during this situation. Also scaled weights and variance estimation have been proposed. The simulation results show that their proposed estimators perform well.

2.5 Clustered Data

As we mention earlier in Chapter 1, this thesis focuses on clustered data where the population is naturally divided into groups (clusters) by similarity of characteristics. If a sample frame of the population is not available, we can draw a random sample of clusters and collect observations from the units in the selected clusters. A single -stage sample design is where observations are collected for all the units in the selected cluster. A two-stage sample design is where observations are collected for only a random sample of units within the selected clusters. In clustered sample designs and a design-based approach for analysis, we can deal with correlations within clusters through the intra-cluster correlation defined approximately as the proportion of the between variance of cluster means to the overall variance. In clustered sample designs, the intra-cluster correlation generally causes variance estimates to increase compared to simple random sample designs. In a model-based approach for analysis, clustered sample designs are usually analysed by treating the clusters as random effects through a multilevel model and the ICC (Intra-class Correlation Coefficient) is the percentage of between-group variability out of the total variance.

In this section we discuss some literature on nonresponse at the cluster level. Haziza and Rao (2001) proposed the use of nested error linear regression models as imputation models in order to take account of the intra-cluster correlation in cluster sampling. The inferential and variance estimation for population total under imputation for missing data on the model-based approach have been proposed following Fay (1991) where Fay (1991) proposed a new method to reverse the sampling and response order which was developed by Shao and Steel (1999). Later Shao (2007) proposed a method to estimate the population mean for nonresponse data under cluster sampling when the missing data is imputed or reweighted. The new estimator is shown to be unbiased under these two situations. Nevertheless, they only considered a single imputation cell in their study.

In the same year, Yuan and Little (2007) suggested new estimators to estimate the population mean when unit nonresponse occurs under the model for two-stage cluster sampling and defined the clusters as random effects. The term, cluster-specific non-ignorable (CSNI) nonresponse has been proposed to represent the type of missing mechanism where the response rate that is varying across cluster data is non-ignorable. They showed that under CSNI, the standard random effect model estimator (RE) for estimating the population mean is biased. Therefore, two adjusted methods have been proposed for bias correction. Both a simulation study and real data are used to compare the performance of the new methods and the naïve estimator. Although their study is limited to two-stage cluster sampling design it can be applied to more complex design such as multistage cluster sampling using hierarchies in multilevel models. Nevertheless, their simulation study is limited to an equal selection probability design. Later, Yuan and Little (2008) proposed an extension of Yuan and Little (2007) but focused on item nonresponse instead of unit nonresponse where missing data depends on covariates and underlying cluster characteristics or depends on covariates and missing outcomes. To see how the new methods perform they considered both a simulation study and applied it to real data. The limitations of this research are that the assumptions of the model are usually violated by multiple imputation and also their assumptions about fully observed covariates are usually wrong in practice.

West (2009) studied various simulation results for the estimator of a population mean for an alternative weighting class adjustment in complex survey designs, e.g. stratified cluster sampling, by developing the previous work of Little and Vartivarian (2003) and Little and Vartivarian (2005). A simulation study is used to investigate five different parameters, e.g. the association between an auxiliary variable for respondents and non-respondents with the variable of interest in sampling survey. The simulation results showed that the weighted response rates in weighted classes are useful in the case of small response rates and in the case that there is an association between the auxiliary variable and the stratum variable for both the variable of interest in the survey and the response propensity in stratified cluster sampling.

Chapter 3

Estimation of Cluster-level Regression Model under Nonresponse within Clusters

3.1 Introduction

In chapter 2, we reviewed the literature which is related to nonresponse in sample surveys especially focusing on nonresponse at the cluster level. In this chapter we discuss nonresponse at the cluster level when some of the response variables data are missing in order to select appropriate regression coefficients in a linear regression model of cluster-level variables and also to extend the Heckman estimators to the clustered model. First, we introduce notation and framework in section 3.2. Second, the model of interest is explained in section 3.3, which focuses on how to use observed data to make inference on coefficients of the model when y data is missing. Next, in section 3.4 we introduce the model for nonresponse, where MAR and NMAR assumptions are also described. Furthermore, we discuss estimation of β for the naïve approach and its bias and variance are shown in section 3.5. The naïve estimator has bias under the NMAR model, we therefore introduce an alternative estimators in section 3.6. Bias and variance for the alternative estimator are also discussed. A Heckman estimator is also developed in section 3.7. Finally, the method we have presented in this chapter is also extended to two-stage cluster sampling.

3.2 Notation and Framework

Let N denote the number of clusters in the population, and m_i the number of elements/units in cluster $i = 1, 2, \ldots, N$.

Let y_{ij} be the value of the study variable y for the j^{th} population element ($j = 1, 2, \ldots, m_i$) for the i^{th} cluster ($i = 1, 2, \ldots, N$).

Let $\bar{y_i}$ be the population mean among all units in cluster i ($i = 1, 2, \ldots, N$).

$$\bar{y_i} = m_i^{-1} \sum_{j=1}^{m_i} y_{ij} \qquad , (i = 1, 2, \ldots, N). \tag{3.1}$$

Let x_i be the cluster-level vector of auxiliary variables in cluster i (no nonresponse error) $x_i = (1, x_{i1}, ..., x_{ik})'$.

Sampling

A simple random sample of n clusters is selected, and all m_i elements($i = 1, 2, \ldots, N$) are sampled in each sampled cluster.

Nonresponse

Let

$$R_{ij} = \begin{cases} 1 & \text{if } y_{ij} \text{ is observed} \\ 0 & \text{if } y_{ij} \text{ is missing} \end{cases}$$

$$\bar{y}_{ri} = \frac{\sum_{j=1}^{m_i} R_{ij} y_{ij}}{\sum_{j=1}^{m_i} R_{ij}} \qquad , (i = 1, 2, \ldots, N) \tag{3.2}$$

$$\sum_{j=1}^{m_i} R_{ij} = r_i \qquad , (i = 1, 2, \ldots, N) \tag{3.3}$$

where r_i is number of respondents in cluster i.

We are now going to make the assumption for number of respondents in cluster i in order to look at new methods for analysis which incorporate information on nonresponse in the model.

Assumption:

$$r_i \geq 1 \qquad ,(i = 1, 2, \ldots, N) \tag{3.4}$$

Data

The observed data count of mean among respondents in cluster i is denoted as $\bar{y}_{ri}, (i = 1, 2, \ldots, N)$.
The response rate in cluster i is denoted as p_i, $(i = 1, 2, \ldots, N)$,

$$p_i = \frac{r_i}{m_i}. \tag{3.5}$$

3.3 Model of Interest

The model of interest is a cluster level linear regression model of $\bar{y}_i$ on x_i, given by the equation

$$\bar{y}_i = {x_i}'\beta + \epsilon_i. \tag{3.6}$$

We shall generally assume $E(\bar{y}_i|x_i) = {x_i}'\beta$, so that $E(\epsilon_i|x_i) = 0, var(\epsilon_i|x_i) = \sigma^2$.

We shall also consider an underlying multilevel model as follows.

Multilevel model
y_{ij} given x_i

$$y_{ij} = {x_i}'\beta + \epsilon_{1i} + \epsilon_{2ij}, \tag{3.7}$$

assuming $E(\epsilon_{1i}|x_i) = 0$ and $E(\epsilon_{2ij}|x_i) = 0, var(\epsilon_{1i}|x_i) = \sigma^2_{By}, var(\epsilon_{2ij}|x_i) = \sigma^2_{Wy}$.

If both (3.6) and (3.7) hold then $\epsilon_i = \epsilon_{1i} + \bar{\epsilon}_{2i}$, where $\bar{\epsilon}_{2i} = \sum_{j=1}^{m_i} \epsilon_{2ij}/m_i$.

Problem:
The problem is how to use observed data to make inference about β when some of the y data is missing.

3.4 Model for Nonresponse

In this section, we introduce a model for the response outcome R_{ij}. It is motivated by Heckman (1976), who introduced a variable u_{ij} to control the response mechanism so

that $R_{ij} = 1$ if $u_{ij} > 0$ and $R_{ij} = 0$, otherwise(Cameron and Trivedi (2005)).
We extend the multilevel model in (3.7) to a bivariate joint model for y_{ij} and u_{ij}, given by

$$\begin{bmatrix} y_{ij} \\ u_{ij} \end{bmatrix} \sim N_2 \left[\begin{bmatrix} \mu_{yi} \\ \mu_{ui} \end{bmatrix} \quad \begin{bmatrix} \sigma^2_{Wy} & \sigma_{Wyu} \\ \sigma_{Wuy} & 1 \end{bmatrix} \right], (j = 1, 2, \ldots, m_i), \tag{3.8}$$

where the matrix

$$\begin{bmatrix} \sigma^2_{Wy} & \sigma_{Wyu} \\ \sigma_{Wuy} & 1 \end{bmatrix}$$

is constant across clusters.

The μ_{yi} and μ_{ui} are assumed to be random effects, where

$$\begin{bmatrix} \mu_{yi} \\ \mu_{ui} \end{bmatrix} \sim N_2 \left[\begin{bmatrix} {x_i}'\beta \\ {x_i}'\gamma \end{bmatrix} \quad \begin{bmatrix} \sigma^2_{By} & \sigma_{Byu} \\ \sigma_{Buy} & \sigma^2_{Bu} \end{bmatrix} \right], (j = 1, 2, \ldots, m_i). \tag{3.9}$$

Hence, one can write analogously to (3.7)

$$y_{ij} = {x_i}'\beta + \epsilon_{ij}, \tag{3.10}$$

where $\epsilon_{ij} \sim N(0, \sigma^2_\epsilon), \sigma^2_\epsilon = \sigma^2_{Wy} + \sigma^2_{By}$,
$\epsilon_{ij} = \epsilon_{1i} + \epsilon_{2ij}, \epsilon_{1i} \sim N(0, \sigma^2_{By}), \epsilon_{2ij} \sim N(0, \sigma^2_{Wy})$, and $\epsilon_{1i}, \epsilon_{2ij}$ are independent.

Similarly, one can write

$$u_{ij} = {x_i}'\gamma + \eta_{ij}, \tag{3.11}$$

where $\eta_{ij} \sim N(0, \sigma^2_\eta), \sigma^2_\eta = 1 + \sigma^2_{Bu}$
$\eta_{ij} = \eta_{1i} + \eta_{2ij}, \eta_{1i} \sim N(0, \sigma^2_{Bu}), \eta_{2ij} \sim N(0, 1)$, and η_{1i}, η_{2ij} are independent.

We have

$$\begin{bmatrix} \epsilon_{ij} \\ \eta_{ij} \end{bmatrix} \sim N_2 \left[\begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \begin{bmatrix} \sigma^2_\epsilon & \sigma_{\epsilon\eta} \\ \sigma_{\epsilon\eta} & \sigma^2_\eta \end{bmatrix} \right], \text{where } \sigma_{\epsilon\eta} = \sigma_{Wyu} + \sigma_{Byu}. \tag{3.12}$$

MAR Assumption and NMAR Model

R_{ij} are MAR if they are conditionally independent of y_{ij} given x_i. Since R_{ij} is determined by u_{ij}, nonresponse is MAR if u_{ij} and y_{ij} are conditionally independent given x_i. From the model (3.10) and (3.11), this occurs if ϵ_{ij} and η_{ij} are conditionally independent given x_i. From (3.12) it follows that $\sigma_{\epsilon\eta} = 0$ if nonresponse is MAR.

Let

$$\delta_i = E[(\bar{y}_{ri} - \bar{y}_{nri})|\vec{R_i}], \qquad (i = 1, 2, \ldots, n), \tag{3.13}$$

where $\vec{R_i} = (R_{i1}, ..., R_{im_i})$, $\bar{y}_{ri}$ is defined in (3.2) and $\bar{y}_{nri}$ is the mean of y among non-respondents,

$$\bar{y}_{nri} = \sum_{j=r+1}^{m_i} y_{ij}/(m_i - r_i). \tag{3.14}$$

We have, provided assumption (3.4) holds,

$$\begin{aligned} E(\bar{y}_{ri}|R_{i1}, ...R_{im_i}) &= \frac{E(\sum_{j=1}^{m_i} R_{ij}y_{ij}|R_{i1},....R_{im_i})}{\sum_{j=1}^{m_i} R_{ij}} \\ &= \frac{\sum_{j=1}^{m_i} E(R_{ij}y_{ij}|R_{i1},....R_{im_i})}{\sum_{j=1}^{m_i} R_{ij}} \\ &= \frac{\sum_{j=1}^{m_i} R_{ij}E(y_{ij}|R_{i1},....R_{im_i})}{\sum_{j=1}^{m_i} R_{ij}}. \end{aligned}$$

If nonresponse is MAR then $E(y_{ij}|R_{i1}, ...R_{im_i}) = x_i{}'\beta$ (treating x_i as fixed here).

Hence, under MAR

$$E(\bar{y}_{ri}|R_{i1}, ...R_{im_i}) \;=\; x_i{}'\beta. \tag{3.15}$$

Similarly,

$$\begin{aligned} E(\bar{y}_{nri}|R_{i1}, ...R_{im_i}) &= E(\textstyle\sum_{j=r+1}^{m_i} y_{ij}/(m_i - r_i)|R_{i1}, ...R_{im_i}) \\ &= \frac{\sum_{j=r+1}^{m_i} E(y_{ij}|R_{i1},....R_{im_i})}{m_i - r_i}. \end{aligned} \tag{3.16}$$

If nonresponse is MAR then,

$$\begin{aligned} E(y_{ij}|R_{i1}, ...R_{im_i}) &= E(y_{ij}) \\ &= x_i{}'\beta. \end{aligned} \tag{3.17}$$

Substitute (3.17) into equation (3.16),

$$\begin{aligned} E(\bar{y}_{nri}|R_{i1}, ...R_{im_i}) &= \frac{\sum_{j=r+1}^{m_i} x_i{}'\beta}{m_i - r_i} \\ &= \frac{(m_i - r)x_i{}'\beta}{m_i - r_i} \end{aligned}$$

$$= {x_i}'\beta \qquad \text{provided } r_i < m_i. \tag{3.18}$$

From (3.15) and (3.18), we have

$$E(\bar{y}_{ri} - \bar{y}_{nri}|R_{i1}, ...R_{im}) = 0.$$

Hence, $\delta_i = 0$ for all i under MAR.

If any of the δ_i are non-zero then the model is NMAR.
For simplicity, we consider a particular NMAR model when

$$\delta_i = \delta, \qquad \delta \neq 0. \tag{3.19}$$

3.5 Estimation of β: Naïve Approach

In this section we introduce a naïve estimator of β in section 3.5.1, we discuss its bias focusing on MAR and NMAR model. Then, the variance of the estimator is explained in section 3.5.2.

From equation (3.6), this model can be written in matrix form for sampled clusters as follows.

$$Y = X\beta + \epsilon, \tag{3.20}$$

where

$$Y = \begin{bmatrix} \bar{y}_1 \\ \bar{y}_2 \\ \vdots \\ \bar{y}_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{bmatrix}, x_i = \begin{bmatrix} 1 \\ x_{i1} \\ \vdots \\ x_{ik} \end{bmatrix} \qquad (i = 1, 2, \dots, n)$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, and \qquad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

Naïve approach replaces $\bar{y}_i$ by $\bar{y}_{ri}$ and Y by

$$Y_r = \begin{bmatrix} \bar{y}_{r1} \\ \bar{y}_{r2} \\ \vdots \\ \bar{y}_{rn} \end{bmatrix},$$

then the estimator of β from the naïve approach is

$$\widehat{\beta_n} = (X'X)^{-1}X'Y_r. \tag{3.21}$$

$\widehat{\beta_n}$ is the best linear unbiased estimator(BLUE). Note assumption (3.4) is required to construct $\widehat{\beta}_n$.

3.5.1 Bias of Naïve Estimator

We now discuss the bias of the naïve estimator under both MAR assumption and NMAR model in section 3.4. We begin with the bias under MAR model before focusing on the bias under NMAR model.

The bias of $\widehat{\beta_n}$ is obtained as follows.

$$\begin{aligned} E(\widehat{\beta_n}) &= E[(X'X)^{-1}X'Y_r] \\ &= (X'X)^{-1}X'E[Y_r]. \end{aligned} \tag{3.22}$$

From (3.15), under MAR

$$E(Y_r) = X\beta. \tag{3.23}$$

Substitute (3.23) into equation (3.22),

$$\begin{aligned} E(\widehat{\beta_n}) &= (X'X)^{-1}X'E[Y_r] \\ &= (X'X)^{-1}X'X\beta \\ &= \beta. \end{aligned} \tag{3.24}$$

So naïve estimator is unbiased under MAR.

Now consider bias under the NMAR model (3.19)

Let us express $\bar{y}_i$ in terms of response rate (p_i),

$$\begin{aligned}
\bar{y}_i &= m_i^{-1}\textstyle\sum_{j=1}^{m_i} y_{ij} \qquad ,(i=1,2,\ldots,N) \\
&= m_i^{-1}(\textstyle\sum_{j=1}^{r_i} y_{ij} + \sum_{j=r_i+1}^{m_i} y_{ij}) \\
&= m_i^{-1}(r_i\bar{y}_{ri} + (m_i - r_i)\bar{y}_{nri} \\
&= \tfrac{r_i}{m_i}\bar{y}_{ri} + \tfrac{m_i - r_i}{m_i}\bar{y}_{nri} \\
&= \tfrac{r_i}{m_i}\bar{y}_{ri} + (1 - \tfrac{r_i}{m_i})\bar{y}_{nri}
\end{aligned}$$

$$\bar{y}_i = p_i\bar{y}_{ri} + (1-p_i)\bar{y}_{nri}, \tag{3.25}$$

where $\bar{y}_{ri}$ is defined in (3.2) and $\bar{y}_{nri}$ is defined in (3.14), and p_i is defined in (3.5).

To find $E(\bar{y}_i|\vec{R}_i)$ in (3.25),

$$\begin{aligned}
E(\bar{y}_i|\vec{R}_i) &= E[(p_i\bar{y}_{ri} + (1-p_i)\bar{y}_{nri})|\vec{R}_i] \\
&= p_iE(\bar{y}_{ri}|\vec{R}_i) + (1-p_i)E(\bar{y}_{nri}|\vec{R}_i).
\end{aligned} \tag{3.26}$$

From (3.13), under the NMAR model

$$\begin{aligned}
\delta &= E[(\bar{y}_{ri} - \bar{y}_{nri})|\vec{R}_i] \\
&= E(\bar{y}_{ri}|\vec{R}_i) - E(\bar{y}_{nri}|\vec{R}_i).
\end{aligned}$$

Hence,

$$E(\bar{y}_{nri}|\vec{R}_i) = E(\bar{y}_{ri}|\vec{R}_i) - \delta. \tag{3.27}$$

Substitute (3.27) into equation (3.26),

$$\begin{aligned}
E(\bar{y}_i|\vec{R}_i) &= p_iE(\bar{y}_{ri}|\vec{R}_i) + (1-p_i)[E(\bar{y}_{ri}|\vec{R}_i) - \delta] \\
&= p_iE(\bar{y}_{ri}|\vec{R}_i) + E(\bar{y}_{ri}|\vec{R}_i) - p_iE(\bar{y}_{ri}|\vec{R}_i) - (1-p_i)\delta \\
&= E(\bar{y}_{ri}|\vec{R}_i) - (1-p_i)\delta.
\end{aligned}$$

Therefore,

$$E(\bar{y}_{ri}|\vec{R}_i) \quad = \quad E(\bar{y}_i|\vec{R}_i) + (1-p_i)\delta.$$

Now

$$E[E(\bar{y}_i|\vec{R}_i)] = E(\bar{y}_i) = x_i{}'\beta.$$

Hence

$$\begin{aligned}
E(\bar{y}_{ri}) - x_i{}'\beta &= E[E(\bar{y}_{ri}|\vec{R}_i)] - x_i{}'\beta \\
&= [1 - E(p_i)]\delta.
\end{aligned}$$

So we may view $[1 - E(p_i)]\delta$ as the bias of $\bar{y}_{ri}$ as an estimator of $x_i'\beta$ under the NMAR model and view $(1 - p_i)\delta$ as the approximate bias,

$$E(\bar{y}_{ri}|\vec{R_i}) \approx x_i{}'\beta + (1 - p_i)\delta. \tag{3.28}$$

3.5.2 Variance of Naïve Estimator

In order to look at the efficiency of the estimator, we now consider the variance of the estimator under the model of interest.

From (3.10) we may write,

$$\bar{y}_{ri} = x_i{}'\beta + \epsilon_{1i} + \bar{\epsilon}_{2i}.$$

Thus,

$$\begin{aligned} V(\bar{y}_{ri}) &= V(\epsilon_{1i} + \bar{\epsilon}_{2i}) \\ &= \sigma^2_{By} + \sigma^2_{Wy}/m_i. \end{aligned}$$

For simplicity, suppose $m_i = m$. Then,

$$\begin{aligned} V(Y_r) &= diag[\sigma^2_{By} + \sigma^2_{Wy}/m] \\ &= (\sigma^2_{By} + m^{-1}\sigma^2_{Wy})I. \end{aligned} \tag{3.29}$$

Therefore, the variance of the naïve estimator is as below.

$$\begin{aligned} V(\widehat{\beta}_n) &= V[(X'X)^{-1}X'Y_r] \\ &= ((X'X)^{-1}X'V(Y_r)X(X'X)^{-1} \\ &= (\sigma^2_{By} + m^{-1}\sigma^2_{Wy})(X'X)^{-1} \end{aligned} \tag{3.30}$$

For simplicity, consider k=1, so that

$$\widehat{\beta}_n = \begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_{1n} \end{bmatrix},$$

$$\begin{aligned} (X'X)^{-1} &= \begin{bmatrix} \frac{\sum_{i=1}^n x_{i1}^2}{n\sum_{i=1}^n (x_{i1}-\bar{x_1})^2} & -\frac{\sum_{i=1}^n x_{i1}}{n\sum_{i=1}^n (x_{i1}-\bar{x_1})^2} \\ -\frac{\sum_{i=1}^n x_{i1}}{n\sum_{i=1}^n (x_{i1}-\bar{x_1})^2} & \frac{1}{\sum_{i=1}^n (x_{i1}-\bar{x_1})^2} \end{bmatrix} \\ &= \begin{bmatrix} \frac{\sum_{i=1}^n x_{i1}^2}{nS^2_{x_1}} & -\frac{\sum_{i=1}^n x_{i1}}{nS^2_{x_1}} \\ -\frac{\sum_{i=1}^n x_{i1}}{nS^2_{x_1}} & \frac{1}{S^2_{x_1}} \end{bmatrix}, \end{aligned}$$

Therefore,

$$V(\widehat{\beta}_{1n}) = (\sigma^2_{By} + m^{-1}\sigma^2_{Wy})(S^2_{x_1})^{-1} \tag{3.31}$$

3.6 Alternative Estimator

It was shown in section 3.5.1 that the naïve approach is biased under NMAR assumption. In this section we therefore look at an alternative estimator. Its bias is shown in section 3.6.1 before moving on to variance of the estimator in section 3.6.2.

From (3.28), we can regress $\bar{y}_{ri}$ on x_i and $(1-p_i)$ and obtain valid estimates of coefficients of x_i under assumption (3.4). We can express the estimator of β using ordinary least squares method following equation (3.23), Let

$$W = \begin{bmatrix} 1 & x_{11} & \dots & x_{1k} & (1-p_1) \\ 1 & x_{21} & \dots & x_{2k} & (1-p_2 \\ \vdots & \vdots & \ddots & \vdots & \\ 1 & x_{n1} & \dots & x_{nk} & (1-p_n) \end{bmatrix}, w_i = \begin{bmatrix} 1 \\ x_{i1} \\ \vdots \\ x_{ik} \\ (1-p_i) \end{bmatrix} \qquad (i = 1, 2, \dots, n)$$

and

$$\gamma_a = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \\ \delta \end{bmatrix} = \begin{bmatrix} \beta \\ \delta \end{bmatrix}.$$

The alternative estimator $\widehat{\beta}_a$ of β is derived from

$$\widehat{\gamma_a} = (W'W)^{-1}W'Y_r \tag{3.32}$$

$$\widehat{\gamma_a} = \begin{bmatrix} \widehat{\beta_a} \\ \widehat{\delta_a} \end{bmatrix}. \tag{3.33}$$

3.6.1 Bias of Alternative Estimator

Following the discussion in section 3.5.1, we now consider the bias of the alternative estimator.

The bias of $\widehat{\gamma_a}$ is obtained as follows.

$$\begin{aligned} E(\widehat{\gamma_a}) &= E[(W'W)^{-1}W'Y_r] \\ &= (W'W)^{-1}W'E[Y_r]. \end{aligned} \tag{3.34}$$

To find $E[Y_r]$, consider $E(\bar{y}_{ri}|\vec{R_i})$ from equation (3.28),

$$\begin{aligned} E(\bar{y}_{ri}|\vec{R_i}) &= {x_i}'\beta_a + (1-p_i)\delta \\ &= w_i'\gamma_a. \end{aligned}$$

Hence,

$$E(Y_r) = W\gamma_a. \tag{3.35}$$

Substitute (3.35) into equation (3.34),

$$\begin{aligned} E(\widehat{\gamma_a}) &= (W'W)^{-1}W'E[Y_r] \\ &= (W'W)^{-1}W'W\gamma_a \end{aligned}$$

$$= \gamma \tag{3.36}$$

and so

$$E(\widehat{\beta_a}) = \beta.$$

Therefore, the alternative estimator is unbiased under the NMAR model in (3.19).

3.6.2 Variance of Alternative Estimator

Now, consider the special case $k = 1$ and $m_i = m$.
Suppose x_i is corrected for its mean, so that a model of interest (section 3.3) is

$$\bar{y}_r = \widetilde{\beta_0} + \beta_1\widetilde{x}_{i11} + \delta x_{i2}, \tag{3.37}$$

where $\widetilde{x}_{i1} = x_{i1} - \bar{x}_1, x_{i2} = 1 - p_i$ and $\widetilde{\beta_0} = \beta_0 + \beta_1\bar{x}_1$.

The W-matrix for centred x_{i1} values is denoted by $\widetilde{W}$,

$$\widetilde{W} = \begin{bmatrix} 1 & \widetilde{x}_{11} & (1-p_1) \\ 1 & \widetilde{x}_{21} & (1-p_2) \\ \vdots & \vdots & \vdots \\ 1 & \widetilde{x}_{n1} & (1-p_n) \end{bmatrix}, \qquad \text{and } \widehat{\gamma}_a \text{ in (3.32) become} \qquad \widehat{\gamma}_a = \begin{bmatrix} \widehat{\beta}_{0a} \\ \widehat{\beta}_{1a} \\ \widehat{\delta}_a \end{bmatrix} = (\widetilde{W}'\widetilde{W})^{-1}\widetilde{W}'Y_r.$$

We now consider the variance of alternative estimator under the NMAR model.

$$\begin{aligned} V(\widehat{\gamma}_a) &= V[(\widetilde{W}'\widetilde{W})^{-1}\widetilde{W}'Y_r] \\ &= (\widetilde{W}'\widetilde{W})^{-1}\widetilde{W}'V(Y_r)\widetilde{W}(\widetilde{W}'\widetilde{W})^{-1} \\ &= (\sigma_{Byy} + m^{-1}\sigma_{Wyy})(\widetilde{W}'\widetilde{W})^{-1} \qquad \text{using (3.30).} \end{aligned} \tag{3.38}$$

Consider $(\widetilde{W}'\widetilde{W})$,

$$(\widetilde{W}'\widetilde{W}) = \begin{bmatrix} n & \sum_{i=1}^n \widetilde{x}_{i1} & \sum_{i=1}^n x_{i2} \\ \sum_{i=1}^n \widetilde{x}_{i1} & \sum_{i=1}^n \widetilde{x}_{i1}^2 & \sum_{i=1}^n \widetilde{x}_{i1}x_{i2} \\ \sum_{i=1}^n x_{i2} & \sum_{i=1}^n \widetilde{x}_{i1}x_{i2} & \sum_{i=1}^n x_{i2}^2 \end{bmatrix}.$$

Therefore, if $\widetilde{x}_{i1}$ is centred, $\widetilde{W}'\widetilde{W}$ is shown as below.

$$(\widetilde{W}'\widetilde{W}) = \begin{bmatrix} n & 0 & \sum_{i=1}^n x_{i2} \\ 0 & \sum_{i=1}^n \widetilde{x}_{i1}^2 & \sum_{i=1}^n \widetilde{x}_{i1}x_{i2} \\ \sum_{i=1}^n x_{i2} & \sum_{i=1}^n \widetilde{x}_{i1}x_{i2} & \sum_{i=1}^n x_{i2}^2 \end{bmatrix}.$$

$(\widetilde{W}'\widetilde{W})$ has been partitioned into sub-matrices as follows.

$$(\widetilde{W}'\widetilde{W}) = \begin{bmatrix} \widetilde{W}_{11} & \widetilde{W}_{12} \\ \widetilde{W}_{21} & \widetilde{W}_{22} \end{bmatrix},$$

where

$$\widetilde{W}_{11} = \begin{bmatrix} n & 0 \\ 0 & \sum_{i=1}^n \widetilde{x}_{i1}^2 \end{bmatrix}, \widetilde{W}_{12} = \begin{bmatrix} \sum_{i=1}^n x_{i2} \\ \sum_{i=1}^n \widetilde{x}_{i1}x_{i2} \end{bmatrix},$$

$$\widetilde{W}_{21} = \begin{bmatrix} \sum_{i=1}^n x_{i2} & \sum_{i=1}^n \widetilde{x}_{i1}x_{i2} \end{bmatrix}, \qquad \text{and} \qquad \widetilde{W}_{22} = \begin{bmatrix} \sum_{i=1}^n x_{i2}^2 \end{bmatrix}.$$

We can find $(\widetilde{W}'\widetilde{W})^{-1}$ using partitioned matrices as shown below.

$$(\widetilde{W}'\widetilde{W})^{-1} = \begin{bmatrix} \widetilde{W}_{11.2}^{-1} & -\widetilde{W}_{11.2}^{-1}\widetilde{W}_{12}\widetilde{W}_{22}^{-1} \\ -\widetilde{W}_{22}^{-1}\widetilde{W}_{21}\widetilde{W}_{11.2}^{-1} & \widetilde{W}_{22}^{-1}\widetilde{W}_{21}\widetilde{W}_{11.2}^{-1}\widetilde{W}_{12}\widetilde{W}_{22}^{-1} + \widetilde{W}_{22}^{-1} \end{bmatrix},$$

where $\widetilde{W}_{11.2} = \widetilde{W}_{11} - \widetilde{W}_{12}\widetilde{W}_{22}^{-1}\widetilde{W}_{21}$.
To find $\widetilde{W}_{11.2}$,

$$\begin{aligned}
\widetilde{W}_{11.2} &= \widetilde{W}_{11} - \widetilde{W}_{12}\widetilde{W}_{22}^{-1}\widetilde{W}_{21} \\
&= \begin{bmatrix} n & 0 \\ 0 & \sum_{i=1}^{n} \widetilde{x}_{i1}^{2} \end{bmatrix} - \begin{bmatrix} \sum_{i=1}^{n} x_{i2} \\ \sum_{i=1}^{n} \widetilde{x}_{i1} x_{i2} \end{bmatrix} \left[\frac{1}{\sum_{i=1}^{n} x_{i2}^{2}}\right] \left[\sum_{i=1}^{n} x_{i2} \quad \sum_{i=1}^{n} \widetilde{x}_{i1} x_{i2}\right] \\
&= \begin{bmatrix} n & 0 \\ 0 & \sum_{i=1}^{n} \widetilde{x}_{i1}^{2} \end{bmatrix} - \begin{bmatrix} \frac{(\sum_{i=1}^{n} x_{i2})^2}{\sum_{i=1}^{n} x_{i2}^{2}} & \frac{(\sum_{i=1}^{n} x_{i2})(\sum_{i=1}^{n} \widetilde{x}_{i1} x_{i2})}{\sum_{i=1}^{n} x_{i2}^{2}} \\ \frac{(\sum_{i=1}^{n} x_{i2})(\sum_{i=1}^{n} \widetilde{x}_{i1} x_{i2})}{\sum_{i=1}^{n} x_{i2}^{2}} & \frac{(\sum_{i=1}^{n} \widetilde{x}_{i1} x_{i2})^2}{\sum_{i=1}^{n} x_{i2}^{2}} \end{bmatrix} \\
&= \begin{bmatrix} n - \frac{(\sum_{i=1}^{n} x_{i2})^2}{\sum_{i=1}^{n} x_{i2}^{2}} - & \frac{(\sum_{i=1}^{n} x_{i2})(\sum_{i=1}^{n} \widetilde{x}_{i1} x_{i2})}{\sum_{i=1}^{n} x_{i2}^{2}} \\ -\frac{(\sum_{i=1}^{n} x_{i2})(\sum_{i=1}^{n} \widetilde{x}_{i1} x_{i2})}{\sum_{i=1}^{n} x_{i2}^{2}} & \sum_{i=1}^{n} \widetilde{x}_{i1}^{2} - \frac{(\sum_{i=1}^{n} \widetilde{x}_{i1} x_{i2})^2}{\sum_{i=1}^{n} x_{i2}^{2}} \end{bmatrix}.
\end{aligned}$$

Consider,

$$\begin{aligned}
|\widetilde{W}_{11.2}| &= \left[n - \frac{(\sum_{i=1}^{n} x_{i2})^2}{\sum_{i=1}^{n} x_{i2}^{2}}\right] \left[\sum_{i=1}^{n} \widetilde{x}_{i1}^{2} - \frac{(\sum_{i=1}^{n} \widetilde{x}_{i1} x_{i2})^2}{\sum_{i=1}^{n} x_{i2}^{2}}\right] - \left[(\textstyle\sum_{i=1}^{n} x_{i2})(\sum_{i=1}^{n} \widetilde{x}_{i1} x_{i2})\right]^2 \\
&= \sum_{i=1}^{n} \widetilde{x}_{i1}^{2} \left[n - \frac{(\sum_{i=1}^{n} x_{i2})^2}{\sum_{i=1}^{n} x_{i2}^{2}}\right] \left[-n \frac{(\sum_{i=1}^{n} \widetilde{x}_{i1} x_{i2})^2}{\sum_{i=1}^{n} x_{i2}^{2}}\right].
\end{aligned}$$

Therefore,

$$\widetilde{W}_{11.2}^{-1} = \left[\frac{1}{\sum_{i=1}^{n} \widetilde{x}_{i1}^{2}\left[n - \frac{(\sum_{i=1}^{n} x_{i2})^2}{\sum_{i=1}^{n} x_{i2}^{2}}\right] - n\frac{(\sum_{i=1}^{n} \widetilde{x}_{i1} x_{i2})^2}{\sum_{i=1}^{n} x_{i2}^{2}}}\right] \begin{bmatrix} \sum_{i=1}^{n} \widetilde{x}_{i1}^{2} - \frac{(\sum_{i=1}^{n} \widetilde{x}_{i1} x_{i2})^2}{\sum_{i=1}^{n} x_{i2}^{2}} & \frac{(\sum_{i=1}^{n} x_{i2})(\sum_{i=1}^{n} \widetilde{x}_{i1} x_{i2})}{\sum_{i=1}^{n} x_{i2}^{2}} \\ -\frac{(\sum_{i=1}^{n} x_{i2})(\sum_{i=1}^{n} \widetilde{x}_{i1} x_{i2})}{\sum_{i=1}^{n} x_{i2}^{2}} & n - \frac{(\sum_{i=1}^{n} x_{i2})^2}{\sum_{i=1}^{n} x_{i2}^{2}} \end{bmatrix}.$$

Hence,

$$\begin{aligned}
[\widetilde{W}_{11.2}^{-1}]_{22} = \frac{n - \frac{(\sum_{i=1}^{n} x_{i2})^2}{\sum_{i=1}^{n} x_{i2}^{2}}}{\sum_{i=1}^{n} \widetilde{x}_{i1}^{2}\left[n - \frac{(\sum_{i=1}^{n} x_{i2})^2}{\sum_{i=1}^{n} x_{i2}^{2}}\right] - n\frac{(\sum_{i=1}^{n} \widetilde{x}_{i1} x_{i2})^2}{\sum_{i=1}^{n} x_{i2}^{2}}} &= \frac{1}{\sum_{i=1}^{n} \widetilde{x}_{i1}^{2} - \frac{n(\sum_{i=1}^{n} \widetilde{x}_{i1} x_{i2})^2}{\sum_{i=1}^{n} x_{i2}^{2}\left[n - \frac{(\sum_{i=1}^{n} x_{i2})^2}{\sum_{i=1}^{n} x_{i2}^{2}}\right]}} \\
&= \frac{1}{S_{x_1}^{2} - \frac{n(\sum_{i=1}^{n} \widetilde{x}_{i1} x_{i2})^2}{n\sum_{i=1}^{n} x_{i2}^{2} - (\sum_{i=1}^{n} x_{i2})^2}}.
\end{aligned}$$

From (3.38), as a result,

$$\begin{aligned}
V(\widehat{\beta}_{1a}) &= (\sigma_{By}^{2} + m^{-1}\sigma_{Wy}^{2}) \left[\frac{1}{S_{x_1}^{2} - \frac{n(\sum_{i=1}^{n} \widetilde{x}_{i1} x_{i2})^2}{n\sum_{i=1}^{n} x_{i2}^{2} - (\sum_{i=1}^{n} \widetilde{x}_{i1})^2}}\right] \\
&= (\sigma_{By}^{2} + m^{-1}\sigma_{Wy}^{2})(S_{x_1}^{2})^{-1} \left[\frac{1}{1 - \frac{(\sum_{i=1}^{n} \widetilde{x}_{i1} x_{i2})^2}{\sum_{i=1}^{n} \widetilde{x}_{i1}^{2}\left[\sum_{i=1}^{n} x_{i2}^{2} - \frac{(\sum_{i=1}^{n} x_{i2})^2}{n}\right]}}\right]
\end{aligned}$$

$$= (\sigma_{By}^{2} + m^{-1}\sigma_{Wy}^{2})(S_{x_1}^{2})^{-1} \left[\frac{1}{1 - r_{12}^{2}}\right], \tag{3.39}$$

where r_{12} is the sample correlation between $\widetilde{x}_{i1}$ and $x_{i2} = (1 - p_i)$.

As a result, found by comparing the estimated regression coefficient variance of $(\widehat{\beta}_1)$ from the naïve approach demonstrated in equation (3.31) and the variance of the alternative approach from equation (3.39) we have that.

$$\frac{V(\widehat{\beta}_{1a})}{V(\widehat{\beta}_{1n})} = \frac{1}{1 - r_{12}^2}. \tag{3.40}$$

If $r_{12}^2$is equal to zero, then $V(\widehat{\beta}_{1a})$ is equal to $V(\widehat{\beta}_{1n})$. However, if r_{12}^2 is positive, then $V(\widehat{\beta}_{1a})$ is larger than $V(\widehat{\beta}_{1n})$. Both approaches give the same result if there is no relationship between auxiliary variable $\widetilde{x}_{i1}$ and nonresponse variable $x_{i2} = (1 - p_i)$. On the other hand, the naïve approach performs better than the alternative approach if there is some correlation between those two variables.

3.7 Heckman Estimator

The Heckman model sometimes called the sample selection model was studied by Heckman (1979) in the case of independent observations and allows for biased selection. Heckman's techniques are very popular in econometrics. In this section we extend this approach to our clustered model, beginning with a Heckman two-step estimator in section 3.7.1 which will introduce Heckman two-step estimator related to alternative approach. Next, we will move on to an approximate Heckman two-step estimator using p_i in section 3.7.2. We end with section 3.7.3, the approximate Heckman maximum likelihood estimator.

According to model for nonresponse in section 3.4, consider

$$E(y_{ij}|R_{ij} = 1) = {x_i}'\beta + E(\epsilon_{ij}|R_{ij} = 1)$$

(Note that expectations are assumed to be conditional on x_i).

$$\begin{aligned} E(y_{ij}|R_{ij} = 1) &= {x_i}'\beta + E(\epsilon_{ij}|u_{ij} > 0) \\ &= {x_i}'\beta + E(\epsilon_{ij}|{x_i}'\gamma + \eta_{ij} > 0) \\ &= {x_i}'\beta + E(\epsilon_{ij}|\eta_{ij} > -{x_i}'\gamma). \end{aligned} \tag{3.41}$$

From (3.12) we can write, $\epsilon_{ij} = \sigma_{\epsilon\eta}{\sigma_\eta}^{-2}\eta_{ij} + \xi_{ij}$, where ξ_{ij} is independent of η_{ij}.
So

$$E(y_{ij}|R_{ij} = 1) = {x_i}'\beta + E(\sigma_{\epsilon\eta}{\sigma_\eta}^{-2}\eta_{ij} + \xi_{ij}|\eta_{ij} > -{x_i}'\gamma). \tag{3.42}$$

We transform $\eta_{ij} \sim N(0, \sigma_\eta^2)$ to $Z_{ij} = \sigma_\eta^{-1}\eta_{ij}$ so that $Z_{ij} \sim N(0,1)$ and plug into (3.42),

$$E(y_{ij}|R_{ij}=1) = x_i{}'\beta + \sigma_{\epsilon\eta}\sigma_\eta^{-1}E\left(Z_{ij}|Z_{ij} > -\frac{x_i{}'\gamma}{\sigma_\eta}\right). \tag{3.43}$$

Now use Proposition 16.1 of Cameron and Trivedi (2005) (see Appendix A)

$$\begin{aligned} E(y_{ij}|R_{ij}=1) &= x_i{}'\beta + \sigma_{\epsilon\eta}\sigma_\eta^{-1}\lambda\left(\tfrac{x_i{}'\gamma}{\sigma_\eta}\right) \\ &= x_i{}'\beta + c\lambda\left(\tfrac{x_i{}'\gamma}{\sigma_\eta}\right), \end{aligned} \tag{3.44}$$

where $c = \sigma_{\epsilon\eta}\sigma_\eta^{-1}, \lambda\left(\frac{x_i{}'\gamma}{\sigma_\eta}\right) = \phi\left(\frac{x_i{}'\gamma}{\sigma_\eta}\right)/\Phi\left(\frac{x_i{}'\gamma}{\sigma_\eta}\right), \phi$ is the probability density function of the standard normal distribution and Φ is the cumulative distribution function of this distribution.

3.7.1 Heckman Two-Step Estimator

This approach is based on result (3.44). Write

$$y_{ij} = x_i{}'\beta + c\lambda\left(\frac{x_i{}'\gamma}{\sigma_\eta}\right) + \upsilon_{ij}, \tag{3.45}$$

where υ_{ij} is an error term.

The estimation steps follow Heckman (1979) as below.
Step 1 Find the estimate $\widehat{\sigma}_\eta^{-1}\widehat{\gamma}$ of $\sigma_\eta^{-1}\gamma$ by probit regression of R_{ij} on x_i.
where $\pi_i = Pr(R_{ij}=1) = Pr(u_{ij} > 0) = Pr(\eta_{ij} > -x_i{}'\gamma) = Pr\left(Z_{ij} > -\frac{x_i{}'\gamma}{\sigma_\eta}\right) = \Phi\left(\frac{x_i{}'\gamma}{\sigma_\eta}\right)$.

Step 2 Calculate the estimated inverse Mills ratio,

$$\lambda\left(\frac{x_i{}'\widehat{\gamma}}{\widehat{\sigma}_\eta}\right) = \phi\left(\frac{x_i{}'\widehat{\gamma}}{\widehat{\sigma}_\eta}\right)/\Phi\left(\frac{x_i{}'\widehat{\gamma}}{\widehat{\sigma}_\eta}\right).$$

Step 3 Plug in the estimated inverse Mills ratio into (3.45) and regress y_{ij} on x_i and the estimated inverse Mills ratio to find the estimation regression coefficients of β_0, β_1 and c.

Relation of Heckman Two-Step Estimator to Alternative Approach

Recall NMAR model in (3.19) $E(\bar{y}_{ri}) - E(\bar{y}_{nri}) = \delta$.
From equation (3.44)

$$E(\bar{y}_{ri}) = E(y_{ij}|R_{ij}=1) = {x_i}'\beta + \sigma_{\epsilon\eta}\sigma_\eta^{-1}\lambda\left(\frac{{x_i}'\gamma}{\sigma_\eta}\right)$$

$$= {x_i}'\beta + c\lambda\left(\frac{{x_i}'\gamma}{\sigma_\eta}\right). \tag{3.46}$$

Similarly,

$$\begin{aligned} E(\bar{y}_{nri}) &= E(y_{ij}|R_{ij}=0) \\ &= {x_i}'\beta + E(\epsilon_{ij}|\mu_{ij}<0) \\ &= {x_i}'\beta + E(\epsilon_{ij}|{x_i}'\gamma + \eta_{ij} < 0) \end{aligned}$$

$$= {x_i}'\beta + E(\epsilon_{ij}|\eta_{ij} < -{x_i}'\gamma). \tag{3.47}$$

As before, write

$$\epsilon_{ij} = \sigma_{\epsilon\eta}\sigma_\eta^{-2}\eta_{ij} + \xi_{ij},$$

where ξ_{ij} is independent of η_{ij}.

So

$$E(\bar{y}_{nri}) = {x_i}'\beta + E(\sigma_{\epsilon\eta}\sigma_\eta^{-2}\eta_{ij} + \xi_{ij}|\eta_{ij} < -{x_i}'\gamma). \tag{3.48}$$

We transform $\eta_{ij} \sim N(0,\sigma_\eta^2)$ to $Z_{ij} = \sigma_\eta^{-1}\eta_{ij}$ so that $Z_{ij} \sim N(0,1)$ and plug into (3.48) ,

$$E(\bar{y}_{nri}) = E(y_{ij}|R_{ij}=0) = {x_i}'\beta + \sigma_{\epsilon\eta}\sigma_\eta^{-1}E\left(Z_{ij}|Z_{ij} < -\frac{{x_i}'\gamma}{\sigma_\eta}\right) \tag{3.49}$$

Now use Proposition 16.1 of Cameron and Trivedi (2005) (see Appendix A)

$$E(\bar{y}_{nri}) = {x_i}'\beta - \sigma_{\epsilon\eta}\sigma_\eta^{-1}\left[\frac{\phi(\frac{{x_i}'\gamma}{\sigma_\eta})}{1-\Phi(\frac{{x_i}'\gamma}{\sigma_\eta})}\right].$$

$$= {x_i}'\beta - c\left[\frac{\phi(\frac{{x_i}'\gamma}{\sigma_\eta})}{1-\Phi(\frac{{x_i}'\gamma}{\sigma_\eta})}\right], \tag{3.50}$$

where $c = \sigma_{\epsilon\eta}\sigma_\eta^{-1}$.

Therefore, from (3.46) and (3.50)

$$\begin{aligned} E(\bar{y}_{ri}) - E(\bar{y}_{nri}) &= \left[{x_i}'\beta + c\left[\frac{\phi({x_i}'\gamma/\sigma_\eta)}{\Phi({x_i}'\gamma/\sigma_\eta)}\right]\right] - \left[{x_i}'\beta - c\left[\frac{\phi(\frac{{x_i}'\gamma}{\sigma_\eta})}{1-\Phi(\frac{{x_i}'\gamma}{\sigma_\eta})}\right]\right] \\ &= \frac{c}{\phi({x_i}'\gamma/\sigma_\eta)(1-\Phi({x_i}'\gamma/\sigma_\eta))}[\phi({x_i}'\gamma/\sigma_\eta) - \phi({x_i}'\gamma/\sigma_\eta)\Phi({x_i}'\gamma/\sigma_\eta) + \phi({x_i}'\gamma/\sigma_\eta)\Phi({x_i}'\gamma/\sigma_\eta)] \\ &= \frac{c(\phi({x_i}'\gamma/\sigma_\eta))}{\phi({x_i}'\gamma/\sigma_\eta)(1-\Phi({x_i}'\gamma/\sigma_\eta))} \\ &= \frac{c\phi_i}{p_i(1-p_i)} \end{aligned}$$

$$= \frac{c\phi_i}{p_i(1-p_i)} = \delta_i \qquad \text{defined in (3.13)}, \tag{3.51}$$

where $\phi_i = \phi\left(\frac{x_i'\widehat{\gamma}}{\widehat{\sigma}_\eta}\right), p_i = \Phi\left(\frac{x_i'\widehat{\gamma}}{\widehat{\sigma}_\eta}\right)$.

3.7.2 Approximate Heckman Two-Step Estimator using p_i

Recall from (3.5) that

$$p_i = m_i^{-1}\sum_{j=1}^{m_i} R_{ij}. \tag{3.52}$$

Let

$$\begin{array}{rcl} \Psi_i &=& \frac{x_i'\gamma}{\sigma_\eta} \\ \widehat{\Psi_i} &=& \frac{x_i'\widehat{\gamma}}{\widehat{\sigma_\eta}}. \end{array} \tag{3.53}$$

For large m_i,

$$p_i = E(R_{ij}) = \Phi\left(\frac{x_i'\widehat{\gamma}}{\widehat{\sigma}_\eta}\right) = \Phi(\Psi_i). \tag{3.54}$$

Now set

$$\widehat{\Psi}_i = \Phi^{-1}(p_i). \tag{3.55}$$

For the Heckman two-step approach replace $\lambda\left(\frac{x_i'\widehat{\gamma}}{\widehat{\sigma}_\eta}\right)$ by $\lambda(\widehat{\Psi}_i) = \lambda_i$ where $\widehat{\Psi}_i$ obtained from (3.55).

3.7.3 Approximate Heckman Maximum Likelihood Estimator

Under working assumption that observations are independent, the likelihood for model in section 3.4 is (Note that when the method is not working well it might be because all observations are not independent.)

$$\prod_{i=1}^{n}\prod_{j=1}^{m_i} Pr(u_{ij} \leq 0)^{1-R_{ij}} f(y_{ij}|u_{ij} > 0) \times Pr(u_{ij} > 0)^{R_{ij}}.$$

The log-likelihood is

$$\sum_{i=1}^{n}\sum_{j=1}^{m_i} (1-R_{ij})log[1 - Pr(u_{ij} > 0)] + R_{ij}log[Pr(u_{ij}) > 0] + R_{ij}log[f(y_{ij}|u_{ij} > 0)].$$

Now $f(y_{ij}|u_{ij} > 0) = Pr(u_{ij} > 0|y_{ij})f(y_{ij})/Pr(u_{ij} > 0)$
and $y_{ij} \sim N(x_i{}'\beta, \sigma_\epsilon^2), u_{ij}|y_{ij} \sim N[x_i{}'\gamma + \sigma_{\epsilon\eta}\sigma_\epsilon^{-2}(y_{ij} - x_i{}'\beta), \sigma_\eta^2 - \sigma_{\epsilon\eta}^2\sigma_\epsilon^{-2}]$.

So evaluation of log likelihood requires evaluating $Pr(u_{ij} > 0)$ for all cases (i, j) and evaluating

$$f(y_{ij}) = \frac{1}{\sqrt{2\pi\sigma_\epsilon^2}} exp\left(\frac{-(y_{ij} - x_i{}'\beta)}{2\sigma_\epsilon^2}\right)$$

and $Pr(u_{ij} > 0|y_{ij})$ for all cases with $R_{ij} = 1$.

The maximum likelihood estimator is used to estimate regression coefficients β_0, β_1 and c in (3.45) instead of probit regression and inverse Mills ratio in Heckman two-step estimator.

3.8 Two-Stage Sampling

The method we have presented in this chapter can be extended to two-stage cluster sampling.

Following, as far as possible, the earlier notation in this chapter, now let

N be the number of clusters in the population, and M_i the number of elements in cluster $i = 1, 2, \ldots, N$.

y_{ij} is the value of the study variable y for the j^{th} population element$(j = 1, 2, \ldots, M_i)$ for the i^{th} cluster $(i = 1, 2, \ldots, N)$.

For two-stage sampling, a simple random sample of n clusters is selected, and a sample of m_i elements $(i = 1, 2, \ldots, n)$ in each sampled cluster is also selected.

Let $\bar{Y}_i$ be the population mean among all units in cluster i $(i = 1, 2, \ldots, N)$.

$$\bar{Y}_i = M_i^{-1} \sum_{j=1}^{M_i} y_{ij} \qquad , (i = 1, 2, \ldots, N). \tag{3.56}$$

Let $\bar{y_i}$ be the sample mean among all units in sampled cluster i $(i = 1, 2, \dots, N)$.

$$\bar{y_i} = m_i^{-1} \sum_{j=1}^{m_i} y_{ij} \qquad , (i = 1, 2, \dots, n). \tag{3.57}$$

Let x_i be the cluster-level vector of variables in cluster i (no nonresponse error) $x_i = (1, x_{i1}, ..., x_{ik})'$.

Let

$$\bar{y_i} = \bar{Y_i} + \delta_i, \tag{3.58}$$

where δ_i be sampling error, and

$$\bar{Y_i} = x_i'\beta + \epsilon_i'. \tag{3.59}$$

From (3.58) and (3.59), we have

$$\begin{aligned} \bar{y_i} &= x_i'\beta + (\epsilon_i' + \delta_i) \\ &= x_i'\beta + \epsilon_i, \end{aligned} \tag{3.60}$$

where $\epsilon_i = \epsilon_i' + \delta_i$.

The model (3.60) is the same model that we applied in the one-stage cluster sampling shown in (3.6) so we can still apply MAR assumption and the NMAR model and follow all of the same estimation processes we did with the naïve approach, the alternative approach and Heckman estimators for our two-stage cluster sampling as we used in one-stage cluster sampling .

3.9 Conclusion

In this chapter we discuss nonresponse at the cluster level where the problem is how to use observed data to make inference about β when some of the y data is missing. We firstly introduce notation and framework for this chapter including sampling techniques that we use, define nonresponse, make the assumption for number of respondents in cluster i in order to look at new methods for analysis which incorporate information on nonresponse in the model and also define data that we consider in our study.

Next, the model of interest which are a cluster level linear regression model of $\bar{y}_i$ on x_i and also a multilevel model are explained but moreover, the model for nonresponse is explained, for example a model for the response outcome R_{ij} that is motivated by Heckman (1976) is introduced, where MAR and NMAR assumptions are also described.

Furthermore, we discuss estimation of β for the naïve approach and also its bias and variance. The naïve estimator produces bias under the NMAR model, we therefore introduce an alternative estimator where we can regress $\bar{y}_{ri}$ on x_i and $(1 - p_i)$ and obtain valid estimates of coefficients of x_i. Bias and variance for the alternative estimator are also discussed.

In addition, a Heckman estimator that was studied by Heckman (1979) is developed upon. We extend this approach to our clustered model, beginning with a Heckman two-step estimator we explore the relationship between it and our alternative approach. Next, we move on to an approximate Heckman two-step estimator using p_i where we replace $\lambda\left(\frac{{x_i}'\widehat{\gamma}}{\widehat{\sigma}_\delta}\right)$ in the Heckman two-step approach with $\lambda(\widehat{\Psi}_i) = \lambda_i$ and end with the approximate Heckman Maximum Likelihood Estimator where the maximum likelihood estimator is used to estimate regression coefficients β_0, β_1 and c instead of probit regression and inverse Mills ratio in the Heckman two-step estimator. Finally, we extend the method we have presented in this chapter to two-stage cluster sampling.

Chapter 4

Simulation Study of Estimators of Cluster-level Regression Model

In chapter 3, we considered some estimators of coefficients in a cluster-level regression model under nonresponse. We now consider the performance of these estimators by means of a simulation study. In this chapter we first discuss the models for the simulation study in section 4.1. Secondly, we show the results of the simulation study in section 4.2. Next, in section 4.3 further theory to explain the simulation results is presented. Finally, the conclusion for our findings is given in section 4.4.

4.1 Models for the Simulation Study

In this section we describe models used to generate y_{ij} and R_{ij} for the simulation studies. This section is divided into two parts; models underlying naïve and alternative approaches and models underlying Heckman estimators approach. For the simulation study we are focusing on a one-stage cluster design where all the clusters have equal sizes. For each method we generate a population with $N = 1000$ clusters and with $m = 25$ elements in each cluster showing the number of workplaces found in the real data that we will apply in Chapter 5 following the number of workplaces in the real data that we will apply in Chapter 5 and repeat 10,000 times.

4.1.1 Models underlying Naïve and Alternative Approaches

We consider two different models based on the assumptions of the naïve approach and alternative approach as follows.

MAR Model

We first generate y_{ij} from the multilevel model given in equation 3.7,

$$y_{ij} = x_i{}'\beta + \epsilon_{1i} + \epsilon_{2ij} \tag{4.1}$$

assuming $E(\epsilon_{1i}|x_i) = 0$ and $E(\epsilon_{2ij}|x_i) = 0$.

We consider a MAR response mechanism with response rate ranging from 0.6 to 0.9. The model starts from generating y and then generates $R|y$ in order to get the joint distribution of (R, y).

The simulation steps for MAR model are as follows.

Step 1 In order to generate y_{ij} from equation (4.1), we generate $\epsilon_{1i} \sim N(0,1), x_i \sim N(20,1)$ and $\epsilon_{2ij} \sim N(0, \sigma^2_{\epsilon_{2ij}})$, where

$$\sigma^2_{\epsilon_{2ij}} = \left[\tfrac{1-\rho}{\rho}\right] \sigma^2_{\epsilon_{1i}},$$

$i = 1, \ldots, N, j = 1, \ldots, m, \rho = 0.1, 0.2, 0.4$, and $\beta_0 = 0, \beta_1 = 1$.

Step 2 Select sample of n = 20 clusters.

Step 3 Generate $A_{ij} \sim U(0,1)$. For cluster 1 to 5, if $0 \leq A_{ij} \leq 0.9$ then $R_{ij} = 1$ else $R_{ij} = 0$. For cluster 6 to 10, if $0 \leq A_{ij} \leq 0.8$ then $R_{ij} = 1$ else $R_{ij} = 0$. For cluster 11 to 15, if $0 \leq A_{ij} \leq 0.7$ then $R_{ij} = 1$ else $R_{ij} = 0$ and for cluster 16 to 20, if $0 \leq A_{ij} \leq 0.6$ then $R_{ij} = 1$ else $R_{ij} = 0$.

Step 4 Compute estimators from naïve approach, alternative approach, Heckman two-step estimator and approximate Heckman two-step estimator using p_i.

Step 5 Compare each estimator using MSE.

NMAR Model

Now consider a model which is designed to capture the NMAR model in (3.19).
The model is shown below.

$$\begin{aligned} P(R_{ij}=1) &= \pi_i \\ y_{ij} &= \begin{cases} (x_i{}'\beta+\epsilon_{1i}+\epsilon_{2ij})+(1-p_i)\delta, & \text{if } R_{ij}=1 \\ (x_i{}'\beta+\epsilon_{1i}+\epsilon_{2ij})-p_i\delta, & \text{if } R_{ij}=0 \end{cases}. \end{aligned} \tag{4.2}$$

This model is the opposite of the previous model as it starts from generating R and then generates $y|R$ in order to get (R, y). The details of how the data and nonresponse are generated are shown as follows.

We also consider replacing p_i in (4.2) by π_i in order to see how the alternative approach performs. We will give some discussions about the simulation results in section (4.2).

The simulation study for NMAR model is divided into three parts as below.

a) π_i fixed

Suppose

$$\pi_i = \pi, \qquad \text{constant for } i. \tag{4.3}$$

π_i variable but independent of x_i Set,

$$\pi = 0.8. \tag{4.4}$$

b)

Define,

$$\begin{aligned} logit(\pi_i) &= logit(0.8)+Z_i, \qquad Z_i \sim N(0,0.5), \\ \pi_i &= \frac{exp(log(\frac{0.8}{0.2})+Z_i)}{1+exp(log(\frac{0.8}{0.2})+Z_i)}. \end{aligned} \tag{4.5}$$

c) π_i depends on x_i

Define,

$$\begin{aligned} logit(\pi_i) &= logit(0.8) + Z_i, \qquad Z_i = x_i - 20, \\ \pi_i &= \tfrac{exp(log(\frac{0.8}{0.2})+Z_i)}{1+exp(log(\frac{0.8}{0.2})+Z_i)}. \end{aligned} \tag{4.6}$$

Simulation steps:

Step 1 For case (b), generate $Z_i \sim N(0, 0.5)$, and then calculate π_i from (4.5). For case (c), generate $x_i \sim N(20, 1)$, and then calculate Z_i and π_i from (4.6).

Step 2 Generate $A_{ij} \sim U(0, 1)$. If $0 \leq A_{ij} \leq \pi_i$ then $R_{ij} = 1$ else $R_{ij} = 0$, and then compute $p_i = \bar{R}_i$.

Step 3 Generate y_{ij} from equation (4.2) with $\epsilon_{1i} \sim N(0, 1), x_i \sim N(20, 1)$ and $\epsilon_{2ij} \sim N(0, 9), \beta_0 = 0, \beta_1 = 1$, and $\delta = 1, 2, 4$.

Step 4 Select sample of n = 20 clusters.

Step 5 Compute estimators from naïve approach, alternative approach, Heckman two-step estimator and approximate Heckman two-step estimator using p_i.

Step 6 Compare each estimator using MSE.

4.1.2 Models underlying Heckman Estimators

The simulation study based on the model underling the Heckman estimators is as follows.

We set
$\sigma_{Wyy} = \sigma_{Byy} = 1$ and $\sigma_{Buu} = 0$ (It implies that there is no random effect so it is only x_i). $\sigma_{Byu} = \rho\sigma_{Byy}\sigma_{Buu} = 0$ and $\sigma_{Wyu} = \rho\sigma_{Wyy} = \rho, \sigma^2_\delta = 1 + \sigma_{Buu} = 1$.
Therefore, $\sigma_{\epsilon\delta} = \sigma_{Wyu} + \sigma_{Byu} = \rho$.

Simulation steps:

Step 1 Generate ϵ_{ij} and δ_{ij} from bivariate normal distribution following equation (3.12)

$$\begin{bmatrix} \epsilon_{ij} \\ \delta_{ij} \end{bmatrix} \sim N_2 \begin{bmatrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix} & \begin{bmatrix} \sigma^2_\epsilon & \sigma_{\epsilon\delta} \\ \sigma_{\epsilon\delta} & \sigma^2_\delta \end{bmatrix} \end{bmatrix}, \text{where } \sigma_{\epsilon\delta} = \sigma_{Wyu} + \sigma_{Byu}, \sigma^2_\epsilon = \sigma_{Wyy} + \sigma_{Byy}, \sigma^2_\delta = 1 + \sigma_{Buu}.$$

In order to specify $\sigma_{\epsilon\delta} = \sigma_{Wyu} + \sigma_{Byu}$, we have to calculate $\sigma_{Byu} = \rho\sigma_{Byy}\sigma_{Buu}$ and $\sigma_{Wyu} = \rho\sigma_{Wyy}$ (vary $0 \leq \rho \leq 1$).

Step 2 To find y_{ij}, generate $x_i \sim N(0,1), \beta_0 = 0, \beta_1 = 1$ and plug in the value of ϵ_{ij} from step 1 into equation (3.10).

Step 3 Similarly, to find u_{ij}, generate $x_i \sim N(0,1), \gamma_0 = 0, \gamma_1 = 1$ and plug in the value of δ_{ij} from step 1 into equation (3.11).

Step 4 If $u_{ij} > 0$ then $R_{ij} = 1$ and $R_{ij} = 0$, otherwise. The overall response rate is 50 %.

Step 5 Select sample of n = 20 clusters.

Step 6 Compute estimators from naïve approach, alternative approach, Heckman two-step estimator and approximate Heckman two-step estimator using p_i.

Step 7 Compare each estimators using MSE.

4.2 Simulation Results

The simulation results are divided into two parts; comparing the results for each approach using models underlying naïve and alternative approaches and models underlying Heckman estimators approach.

4.2.1 Simulation Results from Models underlying Naïve and Alternative Approaches

For the models underlying naïve and alternative approaches, we consider the simulation model at the cluster level. The results will be presented in Tables 4.1 to 4.5 following two models in the simulation study; MAR and NMAR model where NMAR model is divided into 3 separate cases; π_i constant, π_i is variable but independent of x_i and where π_i depends on x_i. The results are as follows.

Table 4.1: Mean, variance and mean square error of the simulation results of MAR model for $N = 1000, m = 25$ and $n = 20, \beta_0 = 0, \beta_1 = 1$. The simulation standard error is shown in parenthesis.

			Naïve approach	**Alternative approach**	**Heckman two-step**	**Approximate Heckman two-step using p_i**
$\rho = 0.1$						
	Mean	Beta 0	-0.092(0.06)	-0.103(0.06)	0.116(0.22)	-0.086(0.06)
		Beta 1	1.005(0.003)	1.005(0.003)	0.992(0.01)	1.004(0.003)
	Variance	Beta 0	**35.837**	38.437	502.313	38.604
		Beta 1	**0.089**	0.095	1.730	0.096
	MSE	Beta 0	**35.845**	38.447	502.327	38.612
		Beta 1	**0.089**	0.095	1.730	0.096
$\rho = 0.2$						
	Mean	Beta 0	-0.067(0.05)	-0.081(0.06)	0.135(0.20)	-0.065(0.06)
		Beta 1	1.004(0.003)	1.004(0.003)	0.991(0.003)	1.003(0.003)
	Variance	Beta 0	**29.080**	31.236	409.407	31.666
		Beta 1	**0.073**	0.077	1.417	0.078
	MSE	Beta 0	**29.085**	31.242	409.425	31.671
		Beta 1	**0.073**	0.077	1.417	0.078
$\rho = 0.4$						
	Mean	Beta 0	-0.049(0.05)	-0.064(0.05)	0.150(0.19)	-0.049(0.05)
		Beta 1	1.003(0.002)	1.003(0.002)	0.991(0.01)	1.002(0.01)
	Variance	Beta 0	**25.660**	27.596	362.222	28.170
		Beta 1	**0.064**	0.068	1.260	0.070
	MSE	Beta 0	**25.663**	27.600	362.245	28.173
		Beta 1	**0.064**	0.068	1.260	0.070

Table 4.1 presents the results of the MAR model for naïve approach, alternative approach, Heckman two-step estimator and approximate Heckman two-step estimator using p_i. We see that the naïve approach performs better than the alternative approach for each ρ. The naïve approach gives smaller variance and also a lower mean square error when compared to all estimators. The alternative approach and the approximate Heckman two-step estimator using p_i behave similarly in this situation. The Heckman two-step estimator performs even worse in that it produces both bigger variance and mean square error comparing to other estimators. Therefore, the Heckman two-step estimator is not suitable to use at all in this situation.

According to equation (3.40) in section 3.6.2, the sample correlation between $\widetilde{x}_{i1}$ and $x_{i2} = (1 - p_i)$ for Table 4.1 is equal to 0.24. As a result, the difference between the variance for estimated regression coefficient $(\widehat{\beta}_1)$ from the naive approach from equation (3.31) and alternative approach from equation (3.39) is 1.061 $(V(\widehat{\beta}_{1a})/V(\widehat{\beta}_{1n}) =$

$1/(1 - r_{12}^2) = 1.06$). If we consider, for example, when $\rho = 0.1$ the simulation variance for estimated regression coefficient ($\widehat{\beta_1}$) from the naive approach is 0.089 and from the alternative approach is 0.095. As a result, the difference between the variance of two methods is 1.067 which is close to 1.061 and it is also similar for all ρ. Therefore, theory implies that the difference between the variances across two methods, alternative and naïve, should be 1.061 and examination of that assumption found it to be true with our simulation results being very close to the prescribed outcome.

As expected, the naïve approach performs well in this scenario because there is some relationship between the auxiliary variable $\widetilde{x}_{i1}$ and the nonresponse variable $x_{i2} = (1-p_i)$. However, if there is no relationship between those two variables at all the alternative approach will give the same results when compared to the naïve approach.

Table 4.2: Mean, variance and mean square error of the simulation results of NMAR model with π_i constant for $N = 1000, m = 25, n = 20$ and $\rho = 0.1, \beta_0 = 0, \beta_1 = 1$. The simulation standard error is shown in parenthesis.

			Naïve approach	**Alternative approach**	**Heckman two-step**	**Approximate Heckman two-step using p_i**
$\delta = 1$						
	Mean	Beta 0	0.115(0.06)	-0.094(0.06)	0.057(0.11)	-0.124(0.06)
		Beta 1	1.004(0.003)	1.005(0.003)	1.003(0.006)	1.005(0.003)
		Delta		1.020(0.04)		
	Variance	Beta 0	**34.907**	37.660	110.875	37.880
		Beta 1	**0.087**	0.092	0.392	0.092
		Delta		14.163		
	MSE	Beta 0	**34.921**	37.669	110.879	37.896
		Beta 1	**0.087**	0.092	0.392	0.092
		Delta		14.163		
$\delta = 2$						
	Mean	Beta 0	**0.318(0.06)**	-0.094(0.06)	0.155(0.11)	-0.153(0.06)
		Beta 1	1.004(0.003)	1.005(0.003)	1.005(0.006)	1.005(0.003)
		Delta		2.020(0.04)		
	Variance	Beta 0	**35.469**	37.660	111.900	37.880
		Beta 1	**0.088**	0.092	0.397	0.092
		Delta		14.163		
	MSE	Beta 0	**35.570**	37.669	111.924	37.903
		Beta 1	**0.089**	0.092	0.397	0.092
		Delta		14.163		
$\delta = 4$						
	Mean	Beta 0	**0.723(0.06)**	-0.094(0.06)	0.351(0.11)	-0.211(0.06)
		Beta 1	1.004(0.003)	1.005(0.003)	1.010(0.006)	1.005(0.003)
		Delta		4.020(0.04)		
	Variance	Beta 0	**37.515**	**37.660**	116.388	37.880
		Beta 1	0.094	**0.092**	0.414	**0.092**
		Delta		14.163		
	MSE	Beta 0	38.038	**37.669**	116.511	37.925
		Beta 1	0.094	**0.092**	0.414	**0.092**
		Delta		14.163		

Table 4.2 presents the results of NMAR model with π_i constant for naïve approach, alternative approach, Heckman two-step estimator and approximate Heckman two-step estimator using p_i. We can see that using the alternative approach seems to correct the bias for the estimator of β_0 better than using the naïve approach does but for β_1 they

both perform properly. There is also some evidence of bias of β_0 visible using the Heckman estimators when δ increases. Using the approximate Heckman two-step estimator with p_i yields similar results to the alternative approach. The alternative approach gives minimum variance and mean square error when $\delta = 4$. It seems that the alternative approach performs better in terms of bias and mean square error as δ increases. Variance and mean square error using alternative approach are not affected by departure from MAR.

Table 4.3: Mean, variance and mean square error of the simulation results of NMAR model with variable π_i but independent of $x_i$for $N = 1000, m = 25, n = 20$ and $\rho = 0.1, \beta_0 = 0, \beta_1 = 1$. The simulation standard error is shown in parenthesis.

			Naïve approach	**Alternative approach**	**Heckman two-step**	**Approximate Heckman two-step using p_i**
$\delta = 1$						
	Mean	Beta 0	0.189(0.06)	-0.008(0.06)	0.210(0.25)	-0.027(0.06)
		Beta 1	1.0017(0.003)	1.0004(0.003)	0.993(0.01)	1.0003(0.003)
		Delta		1.0137(0.02)		
	Variance	Beta 0	**34.820**	36.943	651.751	37.057
		Beta 1	**0.087**	0.091	2.077	0.092
		Delta		5.119		
	MSE	Beta 0	**34.856**	36.943	651.796	37.057
		Beta 1	**0.087**	0.091	2.077	0.092
		Delta		5.119		
$\delta = 2$						
	Mean	Beta 0	**0.412(0.06)**	-0.0084(0.06)	0.321(0.26)	-0.050(0.06)
		Beta 1	1.0016(0.003)	1.0004(0.003)	0.992(0.01)	1.0003(0.003)
		Delta		2.014(0.02)		
	Variance	Beta 0	**36.195**	36.943	676.371	37.068
		Beta 1	**0.090**	**0.091**	2.163	0.092
		Delta		5.119		
	MSE	Beta 0	**36.365**	36.943	676.474	37.070
		Beta 1	**0.090**	**0.091**	2.163	0.092
		Delta		5.119		
$\delta = 4$						
	Mean	Beta 0	**0.856(0.06)**	-0.0084(0.06)	0.542(0.28)	-0.095(0.06)
		Beta 1	1.0014(0.003)	1.0004(0.003)	0.988(0.02)	1.0003(0.003)
		Delta		4.014(0.02)		
	Variance	Beta 0	41.702	**36.943**	759.691	37.096
		Beta 1	0.104	**0.091**	2.444	0.092
		Delta		5.119		
	MSE	Beta 0	42.435	**36.943**	759.985	37.105
		Beta 1	0.104	**0.091**	2.444	0.092
		Delta		5.119		

Table 4.3 presents the results of NMAR model with variable π_i independent of x_i for naïve approach, alternative approach, Heckman two-step estimator and approximate Heckman two-step estimator using p_i. We see a similar result here to that found in Table 4.2. Using the alternative approach seems to correct the bias for the estimator of β_0 quite well when compared to the naïve approach but for β_1 they both perform properly. The β_0 for the Heckman two-step estimator is also biased. Both the naïve estimator and Heckman two-step estimator have higher variance and higher mean square error than the alternative approach for $\delta = 4$ and the approximate Heckman two-step estimator using

p_i has the similar result to alternative approach. Similar to Table 4.2 it seems that the alternative approach performs better in terms of bias and mean square error as δ increases.

Table 4.4: Mean, variance and mean square error of the simulation results of NMAR model with variable π_i depending on x_i for $N = 1000, m = 25, n = 20$ and $\rho = 0.1, \beta_0 = 0, \beta_1 = 1, \beta_0 = 0, \beta_1 = 1$. The simulation standard error is shown in parenthesis.

			Naïve approach	**Alternative approach**	**Heckman two-step**	**Approximate Heckman two-step using p_i**
$\delta = 1$						
	Mean	Beta 0	**3.401(0.063)**	**-0.108(0.13)**	-1.161(0.36)	**-0.088(0.13)**
		Beta 1	**0.842(0.003)**	**1.0048(0.006)**	1.052(0.017)	**1.003(0.003)**
		Delta		1.046(0.035)		
	Variance	Beta 0	39.951	168.870	1316.629	171.982
		Beta 1	0.099	0.377	2.809	0.381
		Delta		12.425		
	MSE	Beta 0	**51.519**	168.882	1317.978	171.990
		Beta 1	**0.124**	0.377	2.812	0.381
		Delta		12.427		
$\delta = 2$						
	Mean	Beta 0	**6.769(0.064)**	**-0.108(0.13)**	-1.795(0.37)	**-0.071(0.13)**
		Beta 1	**0.685(0.003)**	**1.005(0.006)**	1.080(0.017)	**1.001(0.003)**
		Delta		2.046(0.035)		
	Variance	Beta 0	40.967	168.870	1334.745	172.154
		Beta 1	0.101	0.377	2.847	0.381
		Delta		12.425		
	MSE	Beta 0	**86.793**	168.882	1337.967	172.159
		Beta 1	**0.201**	0.377	2.854	0.381
		Delta		12.427		
$\delta = 4$						
	Mean	Beta 0	**13.505(0.064)**	**-0.108(0.13)**	-3.062(0.37)	**-0.037(0.13)**
		Beta 1	**0.372(0.003)**	**1.005(0.006)**	1.136(0.017)	**0.998(0.003)**
		Delta		4.046(0.035)		
	Variance	Beta 0	44.724	168.870	1399.477	172.681
		Beta 1	0.111	0.377	2.985	0.382
		Delta		12.425		
	MSE	Beta 0	227.136	**168.882**	1408.855	172.682
		Beta 1	0.505	**0.377**	3.003	0.382
		Delta		12.427		

Table 4.4 presents the results of NMAR model with the variable π_i depending on x_i for naïve approach, alternative approach, Heckman two-step estimator and approximate Heckman two-step estimator using p_i. We see that the naïve approach now has both bias in β_0 and β_1 as δ increases. Using the alternative approach removes the bias but with quite different variance than the naïve approach yields. The approximate Heckman two-step estimator using p_i behaves similarly to the alternative approach in this scenario. The Heckman two-step estimator is a poor estimator giving both large variance and bias. Therefore, the alternative approach perform very well in this situation especially when δ increases as we can see that it gives minimum variance and also mean square error when compare to other methods including naïve approach.

We try to run the simulation for $\delta > 4$, we got similar results in the case of $\delta = 4$ in Tables 4.2 to 4.4.

We can see that under the NMAR mechanism our alternative approach performs better than the naïve approach with a lower bias shown in the results at all times, and a lower variance and mean square error as δ increases. The outcome is not as positive when using our alternative approach under the MAR mechanism where we see increased variance. Overall, our alternative approach is better for dealing with NMAR data than the naïve approach. Moreover, the approximate Heckman two-step estimator using p_i behaves similarly to our alternative approach.

We also repeated the simulation study replacing p_i in (4.2) by π_i. The alternative approach shows that bias correction is worse even though the variance is about the same, for example, the results of NMAR model with variable π_i depends on x_i for N=1000, m = 25, n = 20 and $\rho = 0.1$ when $\delta = 4$. We see bias in the alternative approach ($\widehat{\beta_0} = 10.353, \widehat{\beta_1} = 0.518$, and $\widehat{\delta} = 0.943$), and also higher variance and mean square error when compared with the naïve approach. The reason for the discrepancy is that the assumed model no longer holds.

Moreover, if we repeated the simulation study replacing x_i in (3.11) and (4.2) by z_i for example, the alternative approach and the Heckman approaches perform very poor. We regenerate the simulation in Table 4.4 for $N = 1000, m = 25, n = 20$ and $\rho = 0.1$ as shown in Table 4.5.

Table 4.5: Mean, variance and mean square error of the simulation results of NMAR model with variable π_i depending on x_i by replacing x_i by z_i in (3.11) and (4.2) for $N = 1000, m = 25, n = 20$ and $\rho = 0.1, \beta_0 = 0, \beta_1 = 1$. The simulation standard error is shown in parenthesis.

			Naïve approach	**Alternative approach**	**Heckman two-step**	**Approximate Heckman two-step using p_i**
$\delta = 1$						
	Mean	Beta 0	**11.730(0.071)**	**17.914(0.076)**	**20.824(0.075)**	**17.976(0.076)**
		Beta 1	**0.425(0.004)**	**0.157(0.004)**	**0.037(0.004)**	**0.157(0.004)**
		Delta		-3.489(0.022)		
	Variance	Beta 0	49.995	57.340	56.887	57.605
		Beta 1	0.125	0.135	0.130	0.135
		Delta		4.669		
	MSE	Beta 0	**187.597**	378.278	490.557	380.779
		Beta 1	**0.455**	0.846	1.05 7	0.846
		Delta		24.821		
$\delta = 2$						
	Mean	Beta 0	**13.532(0.068)**	**17.914(0.076)**	**20.782(0.76)**	**17.965(0.076)**
		Beta 1	**0.347(0.003)**	**0.004(0.006)**	**0.038(0.004)**	**0.157(0.006)**
		Delta		-2.489(0.022)		
	Variance	Beta 0	46.392	57.340	57.673	57.587
		Beta 1	0.116	0.135	0.132	0.135
		Delta		4.669		
	MSE	Beta 0	**229.530**	378.278	489.602	380.370
		Beta 1	**0.542**	0.846	1.058	0.846
		Delta		24.821		
$\delta = 4$						
	Mean	Beta 0	**17.137(0.066)**	**17.914(0.076)**	**20.698(0.078)**	**17.943(0.076)**
		Beta 1	**0.190(0.003)**	**0.157(0.004)**	**0.040(0.004)**	**0.156(0.004)**
		Delta		4.046(0.022)		
	Variance	Beta 0	43.042	57.340	60.571	57.580
		Beta 1	0.107	0.135	0.139	0.135
		Delta		4.669		
	MSE	Beta 0	**336.745**	378.278	489.030	379.581
		Beta 1	**0.763**	0.846	1.061	0.847
		Delta		24.821		

Table 4.5 presents the results of NMAR model with the variable π_i depending on x_i replacing x_i with z_i in (3.11) and (4.2) we see bias in all estimators. The alternative approach and Heckman estimators perform very poorly in terms of both bias in β_0 and β_1 as δ increases and also higher variance and mean square error when compared to the naïve approach. Therefore, the alternative approach and the Heckman approaches are not working at all in this scenario even though it works well in the model underlying Heckman estimators that we will show later on in this chapter. We can not find an exact reason why this occurred, it might happened because there is a difference in correlation and covariance between these two variables or it could actually be for other related reasons due to measurement errors.

4.2.2 Simulation Results from Models underlying Heckman Estimators

For the models underlying Heckman estimators approach, we consider the simulation model at the individual. We also consider to replace x_i in (3.11) by z_i in the simulation study. The results are as follows.

Table 4.6 presents the results found using the Heckman estimator with a multi-level model at the individual level for N=1000, n=20, m = 25. We see that the naïve approach, as expected, displays bias when ρ is not equal to zero. The alternative approach and all Heckman estimators in some ways reduce bias but higher variance for small ρ when compare to the naïve approach. However, the variance is not much difference in some cases eg. the Heckman two-step estimator and the Heckman maximum likelihood estimator when ρ increases. Moreover, the Heckman two-step estimator and the Heckman maximum likelihood estimator have smaller mean square error for all cases except $\rho = 0$ and also in some cases with higher ρ have minimum variance than the the naïve approach. For small ρ, $\rho = 2$ the approximate Heckman two-step using p_i and the Heckman maximum likelihood estimator perform well in terms of minimum mean square error comparing to the naïve approach.

Table 4.7 presents the results found using the Heckman estimator with a multi-level model at the individual level for N=1000, n=50, m = 10 we see similar results to those in Table 4.6. The estimators found using the Heckman maximum likelihood estimator seem to have smaller variance and mean square error shown with an increasing ρ ($\rho > 0.5$). The approximate Heckman maximum likelihood estimator seems to correct bias better, has a lower minimum variance and mean square error than the naïve when ρ increases. The approximate Heckman two-step estimator using p_i behaves similarly to the alternative approach in this scenario and both of them perform well in term of minimum mean square error when $\rho = 0.2$ and $\rho = 0.5$ in the comparison to the naïve approach.

Table 4.8 presents the results found using the Heckman estimator with a multi-level model at the individual level for N=1000, n=100, m = 5. The results are similar to those found in Table 4.7 except that all estimators, other than the Heckman maximum likelihood estimator, for all approaches are biased. It seems that for small m these approaches do not perform well in terms of bias except the Heckman maximum likelihood estimator.

Table 4.9 presents the results found using the Heckman estimator with a multi-level model at the individual level for N=1000, n=5, m = 100. We see the similar results when compared to Table 4.6 except the naïve approach which seems to perform well in term of minimum variance and mean square error for $\rho = 0.2$ but it is poor in term of bias. The alternative approach and all Heckman estimators in some ways reduces bias but higher variance when compare to the naïve approach. However, the Heckman two-step estimator and the Heckman maximum likelihood estimator have smaller mean square error when ρ increases.

Moreover, if we repeated the simulation study by not replacing x_i in (3.11) by z_i for example, the alternative approach and the Heckman approaches perform very poor. We regenerate the simulation in Table 4.6 for $N = 1000, n = 20, m = 25$ and intra-cluster correlation = 0.1 as shown in Table 4.10.

Table 4.10 presents the results found using the Heckman estimator with a multi-level model at the individual level using the same x_i in the simulation model for N=1000, n=20, m = 25. We see that the alternative approach and all Heckman estimators perform poor both for bias and minimum variance and mean square error except for the Heckman maximum likelihood that gives smaller mean square error when ρ is high ($\rho = 0.8$).

Table 4.6: Mean, variance and mean square error for the simulation results of Heckman estimator using multilevel model at the individual level for N=1000, n=20, m = 25 and intra-cluster correlation = 0.1, $\beta_0 = 0, \beta_1 = 1$, $c = 0, 0.6, 1.5, 2.4$ for $\rho = 0, 0.2, 0.5$ and 0.8 respectively. The simulation standard error is shown in parenthesis.

	Estimators		**Mean (simulation s.e.)**			**Variance**			**MSE**	
		β_0	β_1	c	β_0	β_1	c	β_0	β_1	c
$\rho = 0$										
	1. Naïve approach	0.009(0.002)	0.973(0.002)		**0.047**	**0.048**		**0.047**	**0.049**	
	2. Alternative approach	0.038(0.004)	0.965(0.002)		0.173	0.061		0.174	0.062	
	3. Heckman two-step	0.023(0.004)	0.968(0.003)	-0.029(0.006)	0.185	0.062	0.337	0.186	0.063	0.338
	4. Approximate Heckman two-step using p_i	0.037(0.004)	0.965(0.002)	-0.054(0.005)	0.163	0.060	0.299	0.164	0.061	0.302
	5. Heckman ML	0.024(0.004)	0.968(0.002)	-0.030(0.006)	0.180	0.061	0.325	0.180	0.062	0.326
$\rho = 0.2$										
	1. Naïve approach	**0.391(0.002)**	**0.897(0.002)**		**0.050**	**0.060**		0.203	0.07061	
	2. Alternative approach	0.001(0.004)	1.007(0.003)		0.175	0.071		0.175	0.07066	
	3. Heckman two-step	-0.039(0.004)	1.019(0.003)	0.686(0.006)	0.189	0.070	0.338	0.191	0.07061	0.345
	4. Approximate Heckman two-step using p_i	0.006(0.004)	1.008(0.003)	0.638(0.005)	0.165	0.070	0.301	**0.165**	**0.0699**	0.303
	5. Heckman ML	-0.032(0.004)	1.017(0.003)	0.675(0.006)	0.183	0.069	0.325	**0.184**	**0.0694**	0.431
$\rho = 0.5$										
	1. Naïve approach	**0.982(0.003)**	**0.693(0.003)**		**0.066**	0.069		1.031	0.163	
	2. Alternative approach	0.053(0.004)	0.963(0.002)		0.172	0.062		0.175	0.063	
	3. Heckman two-step	-0.065(0.004)	0.991(0.002)	1.657(0.006)	0.183	**0.058**	0.317	0.188	**0.058**	0.342
	4. Approximately to Heckman two-step using p_i	0.067(0.004)	0.965(0.002)	1.505(0.005)	0.163	0.061	0.286	0.167	0.062	0.286
	5. Heckman ML	-0.035(0.004)	0.985(0.002)	1.609(0.005)	0.163	**0.057**	0.263	**0.164**	**0.057**	0.275
$\rho = 0.8$										
	1. Naïve approach	**1.557(0.003)**	**0.582(0.003)**		**0.090**	0.092		2.515	0.267	
	2. Alternative approach	0.163(0.004)	0.984(0.002)		0.150	0.052		0.177	0.052	
	3. Heckman two-step	-0.014(0.004)	1.019(0.002)	2.506(0.005)	0.158	0.042	0.252	0.158	0.043	0.264
	4. Approximate Heckman two-step using p_i	0.178(0.004)	0.988(0.002)	2.280(0.005)	0.138	0.050	0.229	0.169	0.050	0.244
	5. Heckman ML	0.009(0.003)	1.023(0.002)	2.471(0.004)	**0.100**	**0.039**	0.131	**0.101**	**0.040**	0.136

Table 4.7: Mean, variance and mean square error for the simulation results of Heckman estimator using multilevel model at the individual level for N=1000, n=50, m = 10 and intra-cluster correlation = 0.1, $\beta_0 = 0, \beta_1 = 1$, $c = 0, 0.6, 1.5, 2.4$ for $\rho = 0, 0.2, 0.5$ and 0.8 respectively. The simulation standard error is shown in parenthesis.

	Estimators	**Mean (simulation s.e.)**			**Variance**			**MSE**		
		β_0	β_1	c	β_0	β_1	c	β_0	β_1	c
$\rho = 0$										
	1. Naïve approach	-0.002(0.002)	0.948(0.002)		**0.038**	**0.041**		**0.039**	**0.044**	
	2. Alternative approach	-0.014(0.003)	0.953(0.002)		0.109	0.049		0.109	0.051	
	3. Heckman two-step	-0.036(0.004)	0.960(0.002)	0.055(0.005)	0.153	0.052	0.294	0.154	0.054	0.297
	4. Approximate Heckman two-step using p_i	-0.016(0.003)	0.954(0.002)	0.024(0.005)	0.107	0.049	0.226	0.107	0.051	0.226
	5. Heckman ML	-0.038(0.004)	0.961(0.002)	0.058(0.005)	0.156	0.053	0.302	0.158	0.054	0.305
$\rho = 0.2$										
	1. Naïve approach	**0.391(0.002)**	**0.932(0.002)**		**0.047**	**0.044**		0.200	**0.049**	
	2. Alternative approach	0.007(0.004)	1.048(0.002)		0.140	0.055		**0.140**	**0.057**	
	3. Heckman two-step	-0.075(0.004)	1.060(0.002)	0.743(0.006)	0.183	0.056	0.330	0.159	0.060	0.351
	4. Approximate Heckman two-step using p_i	0.002(0.004)	1.051(0.002)	0.675(0.005)	0.135	0.054	0.253	**0.135**	**0.057**	0.258
	5. Heckman ML	-0.073(0.004)	1.058(0.002)	0.739(0.006)	0.180	0.055	0.322	0.185	**0.058**	0.341
$\rho = 0.5$										
	1. Naïve approach	**0.967(0.002)**	**0.733(0.002)**		**0.050**	**0.051**		0.986	0.123	
	2. Alternative approach	0.136(0.004)	0.977(0.002)		0.138	0.056		**0.157**	**0.057**	
	3. Heckman two-step	-0.253(0.003)	1.065(0.002)	1.910(0.006)	0.183	0.056	0.304	0.246	0.060	0.473
	4. Approximately to Heckman two-step using p_i	0.136(0.004)	0.982(0.002)	1.419(0.005)	0.133	0.060	0.225	**0.152**	**0.056**	0.232
	5. Heckman ML	-0.186(0.004)	1.048(0.002)	1.810(0.005)	0.137	**0.051**	0.206	0.172	**0.053**	0.302
$\rho = 0.8$										
	1. Naïve approach	**1.541(0.002)**	**0.486(0.002)**		0.055	0.049		2.431	0.313	
	2. Alternative approach	0.270(0.003)	0.854(0.002)		0.122	0.044		0.195	0.065	
	3. Heckman two-step	-0.305(0.004)	0.999(0.002)	2.934(0.005)	0.151	**0.037**	0.245	0.245	**0.037**	0.530
	4. Approximate Heckman two-step using p_i	0.282(0.003)	0.858(0.002)	2.180(0.004)	0.118	0.044	0.200	0.198	0.064	0.249
	5. Heckman ML	-0.158(0.003)	0.965(0.002)	2.718(0.003)	0.081	**0.031**	0.095	**0.106**	**0.032**	0.196

Table 4.8: Mean, variance and mean square error for the simulation results of Heckman estimator using multilevel model at the individual levelfor N=1000, n=100, m = 5 and intra-cluster correlation = 0.1, $\beta_0 = 0, \beta_1 = 1$, $c = 0, 0.6, 1.5, 2.4$ for $\rho = 0, 0.2, 0.5$ and 0.8 respectively. The simulation standard error is shown in parenthesis.

	Estimators	**Mean (simulation s.e.)**			**Variance**			**MSE**		
		β_0	β_1	c	β_0	β_1	c	β_0	β_1	c
$\rho = 0$										
	1. Naïve approach	-0.063(0.002)	0.981(0.002)		**0.035**	**0.032**		0.039	**0.032**	
	2. Alternative approach	-0.123(0.003)	0.998(0.002)		0.097	0.037		0.112	**0.037**	
	3. Heckman two-step	-0.108(0.004)	0.993(0.002)	0.071(0.006)	0.166	0.040	0.303	0.178	0.040	0.308
	4. Approximate Heckman two-step using p_i	-0.118(0.003)	0.997(0.002)	0.101(0.004)	0.095	0.037	0.190	0.109	**0.037**	0.201
	5. Heckman ML	-0.111(0.004)	0.994(0.002)	0.075(0.006)	0.188	0.042	0.354	0.200	0.042	0.360
$\rho = 0.2$										
	1. Naïve approach	0.544(0.002)	0.883(0.004)		**0.035**	0.175		0.331	0.049	
	2. Alternative approach	0.330(0.003)	0.943(0.002)		0.095	**0.039**		0.204	**0.042**	
	3. Heckman two-step	0.144(0.004)	0.987(0.002)	0.640(0.006)	0.171	0.044	0.334	**0.191**	0.044	0.336
	4. Approximate Heckman two-step using p_i	0.332(0.003)	0.943(0.002)	0.395(0.004)	0.094	**0.039**	0.187	0.205	**0.042**	0.230
	5. Heckman ML	**0.143(0.004)**	**0.987(0.002)**	0.642(0.006)	0.171	**0.044**	0.336	**0.191**	0.044	0.338
$\rho = 0.5$										
	1. Naïve approach	1.058(0.002)	0.486(0.002)		**0.034**	**0.032**		1.153	0.296	
	2. Alternative approach	0.395(0.005)	0.685(0.002)		0.084	0.033		0.240	0.133	
	3. Heckman two-step	-0.155(0.004)	0.809(0.002)	1.966(0.005)	0.143	0.035	0.273	0.167	**0.071**	0.490
	4. Approximately to Heckman two-step using p_i	0.390(0.003)	0.689(0.002)	1.247(0.004)	0.083	0.033	0.171	0.236	0.130	0.235
	5. Heckman ML	**-0.170(0.003)**	**0.818(0.002)**	1.989(0.005)	0.117	0.034	0.207	**0.146**	**0.067**	0.446
$\rho = 0.8$										
	1. Naïve approach	1.706(0.002)	0.442(0.002)		**0.035**	0.034		2.947	0.345	
	2. Alternative approach	0.759(0.003)	0.711(0.002)		0.077	0.033		0.653	0.117	
	3. Heckman two-step	-0.243(0.004)	0.969(0.002)	3.243(0.005)	0.141	0.032	0.264	**0.201**	**0.033**	0.975
	4. Approximate Heckman two-step using p_i	0.751(0.003)	0.715(0.002)	1.872(0.004)	0.077	0.033	0.153	**0.201**	**0.033**	0.975
	5. Heckman ML	**0.050(0.003)**	**0.903(0.002)**	2.789(0.003)	0.066	**0.027**	0.087	**0.068**	**0.037**	0.238

Table 4.9: Mean, variance and mean square error for the simulation results of Heckman estimator using multilevel model at the individual levelfor N=1000, n=5, m = 100 and intra-cluster correlation = 0.1, $\beta_0 = 0, \beta_1 = 1$, $c = 0, 0.6, 1.5, 2.4$ for $\rho = 0, 0.2, 0.5$ and 0.8 respectively. The simulation standard error is shown in parenthesis.

	Estimators	**Mean (simulation s.e.)**			**Variance**			**MSE**		
		β_0	β_1	c	β_0	β_1	c	β_0	β_1	c
$\rho = 0$										
	1. Naïve approach	-0.010(0.003)	1.000(0.004)		**0.084**	**0.154**		**0.084**	0.154	
	2. Alternative approach	-0.002(0.011)	1.000(0.006)		1.173	0.362		1.173	0.362	
	3. Heckman two-step	-0.016(0.009)	1.003(0.006)	0.007(0.010)	0.829	0.364	1.041	0.829	0.364	1.041
	4. Approximate Heckman two-step using p_i	-0.006(0.008)	1.000(0.006)	0.001(0.010)	0.722	0.355	0.980	0.723	**0.355**	0.980
	5. Heckman ML	-0.015(0.008)	1.004(0.005)	0.006(0.009)	0.674	0.287	0.828	0.674	0.287	0.828
$\rho = 0.2$										
	1. Naïve approach	**0.440(0.003)**	**0.865(0.004)**		**0.099**	**0.192**		**0.293**	**0.210**	
	2. Alternative approach	-0.014(0.017)	0.982(0.006)		2.723	0.421		2.723	0.421	
	3. Heckman two-step	0.011(0.010)	0.992(0.006)	0.625(0.010)	0.922	0.367	0.997	0.922	0.367	0.997
	4. Approximate Heckman two-step using p_i	0.025(0.009)	0.987(0.006)	0.607(0.009)	0.745	0.381	0.900	0.745	0.381	0.900
	5. Heckman ML	0.025(0.009)	0.983(0.006)	0.606(0.009)	0.763	0.318	0.785	0.764	0.319	0.785
$\rho = 0.5$										
	1. Naïve approach	**1.044(0.005)**	**0.668(0.006)**		**0.217**	0.367		1.306	0.478	
	2. Alternative approach	-0.099(0.015)	0.973(0.006)		2.110	0.352		2.119	0.353	
	3. Heckman two-step	-0.020(0.011)	0.994(0.005)	1.556(0.010)	1.180	**0.284**	1.081	1.180	**0.284**	1.084
	4. Approximately to Heckman two-step using p_i	-0.008(0.008)	0.988(0.005)	1.533(0.009)	0.706	0.291	0.878	**0.706**	0.291	0.879
	5. Heckman ML	0.005(0.009)	0.981(0.005)	1.519(0.008)	0.806	**0.229**	0.674	**0.806**	**0.229**	0.875
$\rho = 0.8$										
	1. Naïve approach	**1.666(0.006)**	**0.493(0.008)**		0.421	0.690		3.199	0.947	
	2. Alternative approach	-0.073(0.012)	0.9521(0.006)		1.474	0.362		1.479	0.365	
	3. Heckman two-step	0.001(0.007)	0.984(0.005)	2.429(0.009)	0.556	0.234	0.743	**0.556**	**0.234**	0.744
	4. Approximate Heckman two-step using p_i	0.051(0.008)	0.970(0.005)	2.367(0.009)	0.709	0.262	0.771	0.712	0.262	0.772
	5. Heckman ML	0.016(0.006)	0.979(0.004)	2.409(0.006)	**0.310**	**0.159**	0.322	**0.310**	**0.159**	0.322

Table 4.10: Mean, variance and mean square error for the simulation results of Heckman estimator using multilevel model at the individual levelfor N=1000, n=20, m = 25 and intra-cluster correlation = 0.1, $\beta_0 = 0, \beta_1 = 1$, $c = 0, 0.6, 1.5, 2.4$ for $\rho = 0, 0.2, 0.5$ and 0.8 respectively. The simulation standard error is shown in parenthesis.

	Estimators	**Mean (simulation s.e.)**			**Variance**			**MSE**		
		β_0	β_1	c	β_0	β_1	c	β_0	β_1	c
$\rho = 0.2$										
	1. Naïve approach	0.525(0.002)	0.672(0.003)		**0.055**	**0.065**		**0.330**	**0.173**	
	2. Alternative approach	0.066(0.011)	0.928(0.007)		1.251	0.462		1.255	0.467	
	3. Heckman two-step	0.085(0.022)	0.904(0.014)	0.534(0.027)	5.043	2.100	7.130	5.051	2.110	7.135
	4. Approximate Heckman two-step using p_i	0.076(0.010)	0.926(0.007)	0.563(0.012)	1.085	0.431	1.538	1.091	0.437	1.539
	5. Heckman ML	0.259(0.013)	0.812(0.009)	0.326(0.006)	1.783	0.782	0.325	1.850	0.817	2.385
$\rho = 0.5$										
	1. Naïve approach	1.310(0.002)	0.228(0.003)		**0.052**	**0.070**		**1.770**	**0.666**	
	2. Alternative approach	1.077(0.011)	0.346(0.007)		1.234	0.506		2.395	0.934	
	3. Heckman two-step	0.118(0.022)	0.900(0.014)	1.435(0.026)	4.854	2.022	6.789	4.868	2.032	6.793
	4. Approximately to Heckman two-step using p_i	0.955(0.011)	0.418(0.007)	0.448(0.013)	1.109	0.486	1.605	2.021	0.825	2.713
	5. Heckman ML	0.444(0.015)	0.711(0.010)	1.038(0.017)	2.232	0.975	2.897	2.429	1.058	3.110
$\rho = 0.8$										
	1. Naïve approach	2.018(0.002)	-0.112(0.003)		**0.039**	**0.083**		4.112	1.320	
	2. Alternative approach	1.537(0.010)	0.139(0.007)		1.077	0.515		3.440	1.257	
	3. Heckman two-step	0.193(0.018)	0.930(0.012)	2.221(0.022)	3.296	1.376	4.682	3.333	1.381	4.714
	4. Approximate Heckman two-step using p_i	1.391(0.010)	0.230(0.007)	0.793(0.012)	0.923	0.465	1.343	2.859	1.057	3.926
	5. Heckman ML	0.326(0.013)	0.854(0.009)	2.049(0.015)	1.672	0.743	2.180	**1.778**	**0.765**	2.303

4.3 Further Theory to Explain the Simulation Results

According to the simulation results from table 4.2 to 4.4, the naïve estimator of β_1 is unbiased if π_i is independent of x_i but biased if π_i depends on x_i. We are now going to explain this finding.

Assumption: π_i is independent of X_i

$$E(p_i) = E(E(p_i|\pi_i)) = E(\pi_i) = \pi. \tag{4.7}$$

As in (3.22),

$$\begin{aligned} E(\hat{\beta}_n) &= E[(X'X)^{-1}X'Y_r] \\ &= (X'X)^{-1}X'E(Y_r). \end{aligned} \tag{4.8}$$

To find $E(Y_r)$, under model (4.2), we write

$$\begin{aligned} \bar{y}_{ri} &= {x_i}'\beta + \epsilon_{1i} + \bar{\epsilon}_{2ij} + (1-p_i)\delta, \\ E(\bar{y}_{ri}) &= E({x_i}'\beta) + E(\epsilon_{1i}) + E(\bar{\epsilon}_{2ij}) + E((1-p_i)\delta) \\ &= {x_i}'\beta + (1-\pi)\delta, \end{aligned} \tag{4.9}$$

assuming $E(\epsilon_{1i}|x_i) = 0$ and $E(\epsilon_{2ij}|x_i) = 0$.

Consequently,

$$E(\bar{y}_{ri}) = {x_i}'\beta^*, \tag{4.10}$$

where

$$x_i = \begin{bmatrix} 1 \\ x_{i1} \end{bmatrix} \qquad (i = 1, 2, \ldots, n) \qquad and \qquad \beta^* = \begin{bmatrix} \beta_0 + (1-\pi)\delta \\ \beta_1 \end{bmatrix}.$$

Hence,

$$E(Y_r) = X'\beta^*. \tag{4.11}$$

Substitute(4.11) into (4.8),

$$\begin{aligned} E(\widehat{\beta}_n) &= (X'X)^{-1}X'E(Y_r) \\ &= (X'X)^{-1}X'X\beta^* \\ &= \beta^*. \end{aligned} \tag{4.12}$$

As a result, $\widehat{\beta}_{0n}$ is a biased estimator if π_i is independent of X_i. The bias can be shown as follows.

$$\begin{aligned} Bias(\widehat{\beta}_0) &= E(\widehat{\beta}_0) - \beta_0 \\ &= \beta_0 + (1-\pi)\delta - \beta_0 \\ &= (1-\pi)\delta, \end{aligned}$$

and $\widehat{\beta}_1$ is unbiased estimator if π_i is independent of X_i, as $E(\widehat{\beta}_1) = \beta_1$.

Assumption, π_i depends on x_i

To show that $E(\widehat{\beta}_{1n}) \neq \beta_1$:
We have,

$$E(p_i|\pi_i, x_i) = \pi_i.$$

Hence

$$\begin{aligned} E(p_i|x_i) &= E(\pi_i|x_i) \\ &= h(x_i), \end{aligned} \tag{4.13}$$

where $h(x_i)$ is an inverse logistic function.

According to (4.9),

$$\begin{aligned} E(\bar{y}_{ri}) &= E({x_i}'\beta) + E(\epsilon_{1i}) + E(\bar{\epsilon}_{2ij}) + E((1-p_i)\delta) \\ &= {x_i}'\beta + \delta - \delta E(p_i|x_i). \end{aligned} \tag{4.14}$$

Substitute (4.13) into (4.14),

$$E(\bar{y}_{ri}) = {x_i}'\beta + \delta(1 - h(x_i)). \tag{4.15}$$

Let

$$\widetilde{x}_i = \begin{bmatrix} 1 \\ x_{i1} \\ 1 - h(x_i), \end{bmatrix}, \qquad \widetilde{X} = \begin{bmatrix} X & U \end{bmatrix}, \qquad U = \begin{bmatrix} 1 - h(x_1) \\ \vdots \\ 1 - h(x_n) \end{bmatrix}, \qquad and \qquad \beta^* = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \delta \end{bmatrix}.$$

Hence,

$$E(Y_r) = \widetilde{X}\beta^*. \tag{4.16}$$

Therefore,

$$\begin{aligned} E(\widehat{\beta}_n) &= (X'X^{-1})X'E(Y_r) \\ &= (X'X^{-1})X'\widetilde{X}\beta^* \\ &= (X'X^{-1})(X'XX'U)\beta^* \\ &= (I(X'X^{-1})X'U)\beta^* \\ &= \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \delta(X'X^{-1})X'U. \end{aligned} \tag{4.17}$$

As a result, β_0 and β_1 are generally biased estimators if π_i depends on X_i.

4.4 Conclusion

Under the models underlying naïve and alternative approaches, the naïve approach performs well in terms of minimum variance and mean square error under the MAR model. However, it is biased for both β_0 and β_1 under NMAR model when p_i depends on x_i and when p_i independent of x_i, and also β_0 is biased under NMAR model with p_i constant.

The alternative approach is successful in removing bias from both β_0 and β_1 under NMAR model when π_i depends on x_i and also corrects the bias that the estimator produces for β_0 quite well when compared to results found using the naïve approach under NMAR model with π_i constant and π_i variable but independent of x_i. However, it does have a higher variance than the naïve approach yields except when δ is increasing under MAR. It does not work at all for the Heckman model.

Under the model underlying Heckman estimator, the Heckman two-step estimator and approximate Heckman maximum likelihood estimator both in some way reduced bias but produced very high variance while the approximate Heckman maximum likelihood estimator seems to have corrected biased well when compare to the one produced by the naïve approach, and also reduces minimum variance and lowers mean square error when ρ increases but they do not work well with a multi-level model.

Chapter 5

Application using WERS Data

In Chapter 4, we undertook a simulation study to compare the performance of the estimators from chapter 3. In this Chapter, we will apply the new methods we devised to real data from the Workplace Employment Relations Survey (WERS) 2004 that will be discussed in section 5.1. In this survey there were 2 levels, a single cluster and a single element, employees within the workplace (there was a particular problem of nonresponse by employees), the clusters are workplaces and the elements are employees. The literature related to the WERS 2004 data will be reviewed in section 5.2 and then we will consider the Bryson et al. (2009) study in section 5.3 as the basis for the main empirical work in this chapter. The proposed analysis and the results of the analyses are discussed in section 5.4 and 5.5 respectively which will be assessed using the regression model at the individual and workplace levels respectively. This section will also discuss both the models that consider survey weight using the generalised regression (GREG) estimation and those that ignore survey weight. Finally, the conclusion for our findings is given in section 5.6.

5.1 The WERS 2004 Survey

The linked employer-employee Workplace Employment Relations Survey (WERS) 2004 data is used in our analysis. According to the information in the Workplace Employment Relation Survey 2004, information and advice service (http://www.wers2004.info/), which offers a recognised account of working life inside most British workplaces. Following the information in WERS 2004 technical report for cross-section and panel surveys, Chaplin et al. (2005) state that the WERS 2004 survey was conducted under a stratified random sample of workplaces and a sample of employees at those workplaces . The strata have been defined by the establishment size and a Standard Industrial Classification 2003 (SIC(2003)) as it was defined in WERS 1998 which leads to the sampling fraction being equal to 0.675 for each strata(Chaplin et al. (2005)). It was collected at workplace level and targeted data from about 2,300 managers, 1,000 employee representatives and 22,500 employees which covers both the private and public sectors (http://www.wers2004.info).

This data does not cover these sectors of the British economy; mining and quarrying, agriculture, hunting and forestry, fishing and private households without employed persons occupying them and extra-territorial bodies (Bryson et al. (2009)). It contains both the Cross-Section Management and Employee Representative data files.

From the Inter-Departmental Business Register, the Workplaces with at least 5 employees were sampled with aim to conduct a face-to-face interview with the most senior person at each workplace looking at industrial relations, employee relations or personnel matters. The data has been conducted from management interview in total of 2,295 workplaces from a sample of 3,587 addresses which yielded a response rate of 64% (http://www.wers2004.info).

The managers in 1,967 (86 %) of the 2,295 workplaces had permission to distribute an eight page self-completion questionnaire for the Survey of Employees in the WERS 2004 Cross-Section to 25 randomly-selected employees in each workplace (or, to every employee in workplaces with from 5 to 24 employees). A further 10% of workplaces did not return any questionnaires. There were 22,451 employees out of 37,000 questionnaires who completed and returned the questionnaire, indicating a fieldwork response rate of 60% (http://www.wers2004.info).

5.2 Analyses of WERS 2004 Data

In this section we describe some published analyses of WERS 2004 data as background to our application. Wood and Fairleigh (2007) used WERS 2004 data to find the change in well-being within all contributing groups forming the British economy utilising Warr's contentment measure and job satisfaction levels to constitute their study. Regression analysis and a psychological model known as the Karasek model are used to analyse levels of well-being, e.g. studying the correlation between well-being and job demands. The results show that highly demanding jobs which do not apply rules to staff produce the most stress and least job satisfaction. Later, Wood (2008) considered the relationship between job characteristics, the employee voice and well-being using the Karasek model. The results show that there is a negative relationship between well-being and job demands while there is a positive relationship between well-being and a sense of control in the work place which is contradictory to his earlier results.

Additionally, some other works have been taken to examine the association between job satisfaction and other variables. Rose (2007) examined the link between occupation and employee job satisfaction using WERS 2004 data. Regression analysis is used to predict job satisfaction taken from studies of the individual-level variables and workplace-level variables in the data. The 81 minor occupation groupings specified in UK Standards Occupational Classification 2000 were considered. The individual-level variables are education attainment, sex, age, pay level and skill type. The workplace-level variables are workplace flexibility, sense of autonomy and a description of the workplace task structure, involvement of employees and workplace structure and practice are described by the managers in a workplace. The results show a forty percent variance in outcomes that cannot be explained using assumptions made before. Consequently, the further investigation will be essentially before finalise the results. Later, Schyns et al. (2009) investigated the effect of supportive leadership climate on job satisfaction. They focused their analysis on three supportive leadership climate variables; supportive leadership climate quality (the sample mean of the scores of supportive leadership climate from employees, taken in survey, for each workplace), supportive leadership climate strength (the sample standard deviation of the distribution of the scores of supportive leadership climate for each workplace) and relative individual psychological supportive leadership climate (the different between each employee's climate score and the mean of the workplace) and then they are tested whether or not related to job satisfaction. The univariate analysis and multilevel model are used to analyse the result. The results show that supportive leadership climate and individual leadership climate are related to job satisfaction but not for supportive leadership climate strength. Nonetheless, the studying data is limited to the employee questionnaire which employees are fully response to all their interest variables which is lead to small data set due to nonresponse problem.

Bryson and Freeman (2008) investigated the effect of economic performance in employee owned business on pay in the UK. Regression analysis is used to explore the relationship between shared capitalist modes of pay (other factors held fixed) and individual pay for results, managerial monitoring, and worker decision-making. They find that the growth of shared capitalism is the same in the UK and the US.

Chatterji and Mumford (2008) studied wages in both the public and private sectors for male employees on full-time contracts. WERS 2004 data is used to study individual worker characteristics data for both public and private sector workplaces. The regression analysis has been applied in the study. It focuses on earning outcomes for men employed to work full time where the mid-point of the interval has been considered to measure weekly wages. The dependent variable is the hourly wage for each employee in each workplace. The explanatory variables are potential experience (years), training (number of training days taken in the previous year), education, vocational qualifications, children, marital status, disabilities, ethnic origin, fixed term employment contract, length of employment, union membership if any and occupation. The results show that the earnings points for public sector workers is 11.7 log wage more than their private sector and also suggests no correlation between public sector pay and private sector pay exists.

Sessions and Theodoropoulos (2009) examined the association between the slope of the wage-tenure profile and the level of monitoring using combined data from the Management and Employee Questionnaires, WERS 1998 and WERS 2004. Interval regression is used to analyze the result at individual level. The dependent variable is the weekly wage earned for each employee at each workplace in term of logarithm. The explanatory variables are the employee tenure, the level of monitoring, other individual regressors (e.g. education, occupation, demographics, training, and first characteristics) along with other variables at each establishment. By using dual cross sections of employee data the analysts are confident that this prediction will be supported.

Antcliff and Saundry (2009) analysed the effect of the introduction of the statutory right to accompaniment at both grievance and disciplinary hearings on three categories; rates of disciplinary sanctions, dismissals and employment tribunal applications. It considers the relationship between employee representation and rates of disciplinary sanctions, dismissal and applications to employment tribunals. Tobit regression models and multivariate regression models are used to examine separate models for these rates using a set of independent variables that measure characteristics of the workforce and workplace including considerations of any employer's legal compliance with the right to accompaniment; a measure of formality within grievance and disciplinary procedures; demographic features(e.g. gender, ethnicity and age of the workforce), workplace size, type of establishment and union density. It shows that probably the grievance and disciplinary processes have been affected by the introduction of the right to accompaniment.

Salis and Williams (2010) examined the association between the face-to-face communications (FTFC) of workplaces that have human resources management (HRM) practices and looked for productivity gains. The response variable is the labour yield measured in thousands of pounds for each employee. The explanatory variables are selected HRM practices with their potential to enhance FTFC which are working in teams, forming problem-solving (PS) groups, meetings between line managers and employees, meetings between senior managers and employees and presentations from employees and managers. Regression analysis is used to explore the data that they also control for the workplace, organisation and market controls variables. Finally, the presence of the HRM practices variables are also controlled. The results show that there is a linear relationship between productivity and FTFC in problem-solving groups, teams and meetings of senior or line managers and employees.

5.3 The Study of Bryson et al. (2009)

We applied our methods to WERS 2004 data following Bryson et al. (2009) who focussed their analysis on private sector workplaces only and examined the association between innovations (management-initiated workplace change) and worker well-being using this set of data which consisted of 13,500 employees in 1,238 workplaces. This research studies well-being measurements in two data sets. The first one is found in analysing employee responses to the following question: “Thinking of the past few weeks how much of the time has your job made you feel each of the following: tense, calm, relaxed, worried, uneasy, content?” A 5-point scale is used to categorise the responses. However, the six anxiety-comment items are combined into single scale following Wood (2008), and the five-point scores are rescaled to scale from -2 to 2. Therefore, the scale varies between -12 and 12. The second one is job satisfaction variable where all eight aspects of job satisfaction are used. Employees are asked to respond to the following questions: “ How satisfied are you with the following aspects of your job?... achievement you get from your work; the scope for using your own initiative; the amount of influence you have over your job; the training you receive; the amount of pay you receive; your job security; the work itself; the amount of involvement you have in decision-making at this workplace?” Responses have been coded into a 5-point Likert scale varying from very satisfied to very dissatisfied. Similar to the first variable, the second one is combined and a 5-point Likert scale is recoded to scale from -2 to 2. Therefore, the scale varies between -16 and 16.

Bryson et al. (2009) considered innovation variables as their independent variables where innovation variables are depending on response from the manager at each workplace according to the question below:

“Over the past two years has management here introduced any of the changes listed on this card? PROBE: Which others? UNTIL 'None':

1) Introduction of performance related pay
2) Introduction or upgrading of computers
3) Introduction or upgrading of other types of new technology
4) Changes in working time arrangements

5) Changes in the organisation of work
6) Changes in work techniques or procedures
7) Introduction of initiatives to involve employees
8) Introduction of technologically new or significantly improved product or service
9) NONE None of these"

They construct three count variables for innovations following the question above, the first one is the summation of all eight innovations (innovations_all); the second one for labour innovations depend on items 4, 5, 6, and 7 which 4 is the maximum value(innovations_work), and the third one for capital innovations depend on items 2, 3, and 8 which 3 is the maximum value (innovations_technology).

In addition, Bryson et al. (2009) considered the unionization variables and other important control variables as independent variables. Unionization variables are controlled for both individual union membership and workplace level union membership. The union membership data at individual level is controlled where 1 represents individual union membership and zero represents none union membership which is obtained directly from employee answers in the employee self-completion questionnaire. In contrast, the workplace level union membership data is obtained from the manager at each workplace where its value is equal to 1 for union coverage and equal to zero for not covered. For other control variables, we again follow Bryson et al. (2009), the individual level control variables are age (9 dummies); academic qualifications (8 dummies); single-digit occupation (9 dummies); and dummies for disability and gender. The workplace level controls are:single-digit industry (11 dummies); log workplace employment size and a quadratic term; and a dummy for low travel-to-work-area unemployment (below 1.2%).

5.3.1 Models

Using the variables discussed in this section, the model of interest following Bryson et al. (2009) is shown as below.

$$W_{ij} = \beta_1 Innovations_j + \beta_2 Union_{ij} + \beta_3 Innovations_j \times Union_{ij} + \beta_x' X_{ij} + \epsilon_{ij}, \quad (5.1)$$

where W_{ij} is well-being(or job satisfaction) for individual i in workplace $j, Innovations_j$ is the number of innovations recommenced in any workplace $j, Union_{ij}$ is a dummy for union coverage, the $X's$ represents the control vector and ϵ_{ij} is a standard normal distributed error term.

Bryson et al. (2009) considered only unweighed regression models as shown above. They applied these models to WERS2004 data and they found that there is a relationship between management innovations and lower employee sense of well-being.

5.4 Proposed Analysis

In our study, we focused on using the regression model of job satisfaction on the innovations and control variables only then used the alternative approach and the naïve approach with the WERS 2004 data. We analyse both the regression model at individual level and at the workplace level. We also consider the models that take into account the complex survey design and the models that ignore the complex survey design. Weighted least square regression and raking are also used in the analysis shown as follows:

5.4.1 The Regression Model at Individual Level

In this section, the regression model of job-satisfaction at individual level is considered. We regress job-satisfaction on three innovation variables (innovations_all, innovations_work and innovations_technology), nonresponse rate and other control variables. We considered seven different estimation methods (Model 1: Naïve approach, Model 2: Alternative approach, Model 3: Heckman two-step estimator, Model 4: Approximate to Heckman two-step estimator using p_i, Model 5: Alternative approach with control variables in the model, Model 6: Heckman two-step estimator with control variables in the model and Model 7: Approximate to Heckman two-step estimator using p_i). These models are shown as follows.

Model 1: Naïve approach

$$y_{ij} = \beta_0 + \beta_1 x_i + \epsilon_{ij}, \tag{5.2}$$

Model 2: Alternative approach

$$y_{ij} = \beta_0 + \beta_1 x_i + \delta(1 - p_i) + \epsilon_{ij}, \tag{5.3}$$

Model 3: Heckman two-step estimator

$$y_{ij} = \beta_0 + \beta_1 x_i + c\lambda\left(\frac{z_i'\gamma}{\sigma_\eta}\right) + \upsilon_{ij}, \tag{5.4}$$

In the probit model we regress R_{ij} on number of employee at each workplace.

Model 4: Approximate to Heckman two-step estimator using p_i

For Heckman two-step approach replace $\lambda\left(\frac{z_i'\widehat{\gamma}}{\widehat{\sigma}_\eta}\right)$ by $\lambda(\widehat{\Psi}_i) = \lambda_i$. Here $\widehat{\Psi}_i$ obtained from $\widehat{\Psi}_i = \Phi^{-1}(p_i)$.

Model 5: Alternative approach including control variables

$$y_{ij} = \beta_0 + \beta_1 x_i + \delta(1 - p_i) + \beta_x' X_{ij} + \epsilon_{ij}, \tag{5.5}$$

Model 6: Heckman two-step estimator including control variables

$$y_{ij} = \beta_0 + \beta_1 x_i + c\lambda\left(\frac{z_i'\gamma}{\sigma_\eta}\right) + \beta_x' X_{ij} + \upsilon_{ij}, \tag{5.6}$$

Model 7: Approximate to Heckman two-step estimator using p_i including control variables

For this model we followed the same steps described for the Approximate Heckman two-step estimator using the variable p_i and including control variables into the model.

where y_{ij} is job satisfaction for employee j in workplace i, x_i is the number of innovations in a workplace i, the $X's$ represent the control vector (for employee level control variables these are: academic qualifications, single-digit occupation, and dummies for disability and gender, the workplace-level control variables are single-digit industry, log workplace employment size and a quadratic term, and a dummy for low travel-to-work-area unemployment), p_i is the employee response rate at each workplace i, ϵ_{ij} is a standard normal distributed error term and υ_{ij} is an error term.

5.4.2 The Regression Model at Workplace Level

We again regress job-satisfaction on innovation variables and nonresponse rate but this time at workplace level which is different from the approach Bryson et al. (2009) took. We considered two different estimation methods (Model 1: Naïve approach, and Model 2: Alternative approach. These models are shown here:

Model 1: Naïve approach

$$\bar{y_i} = \beta_0 + \beta_1 x_i + \epsilon_i, \tag{5.7}$$

Model 2: Alternative approach

$$\bar{y_i} = \beta_0 + \beta_1 x_i + \delta(1 - p_i) + \epsilon_i, \tag{5.8}$$

where $\bar{y_i}$ is the mean job satisfaction of employees in workplace i, x_i is the innovations employees in each workplace i, p_i is the employees response rate i and ϵ_i is a standard normal distributed error term.

5.5 Results of Analysis

The results are divided into two sections; the regression model when not considering a weighted survey and considering a survey weighted for three panels both at individual level and at workplace level as follows:

5.5.1 The Results of the Regression Model at Individual Level

5.5.1.1 Unweighted Estimates of the Regression Model

In chapter 3, we look at how the alternative approach performs when we include a non-response variable $(1 - p_i)$ into the regression model. In order to see how the alternative estimator performs we undertook the simulation study we discussed in chapter 4. In this section we also applied the alternative approach and Heckman estimators to real data at individual level. We regress job-satisfaction on innovations and using the control variables we discussed in the previous section.

Table 5.1 presents unweighted estimates of the regression model of job-satisfaction on innovations_all and control variables. We see that the nonresponse rate variable is significant at 0.05 level for both model 2 alternative approach and model 5 alternative approach including control variables which use the regression model of job-satisfaction on innovations_all although it was not accounted for at all in the Bryson et al. (2009) approach. However, the inverse mills ratio variable from the Heckman two-step estimator is significant at 0.05 level only for models 3 but not for model 6 which includes control variables in the model. Approximate to Heckman two-step estimator using p_i, λ variable is significant at 0.05 level for both model 4 and model 7 which includes control variables in the model. Innovations_all variable is significant at 0.05 level only for models 1 to 4 which use the regression model of job-satisfaction on innovations_all but not including control variables. Nevertheless, including control variables in models 5 and 7 we see that $1 - p_i$ and λ variables are still significant at 0.05 level even though innovations_all became insignificant and also benefit from using more available auxiliary variables to increase the accuracy of the model.

Similar results have been shown in Table 5.2. Table 5.2 presents unweighted estimates of the regression model of job-satisfaction on innovation_work and control variables. We see that for both model 2 alternative approach and model 5 alternative approach including control variables which use the regression model of job-satisfaction on innovations_work the nonresponse rate variable is significant at 0.05 level although it was not accounted for at all in the Bryson et al. (2009) approach. Similarly, approximate to Heckman two-step estimator using p_i, λ variable is significant at 0.05 level for both model 4 and model 7 which includes control variables in the model. Regardless, inverse mills ratio variable from Heckman two-step estimator is significant at 0.05 level only for models 3 but not for model 6 which includes control variables in the model. The only difference for Table 5.2 when compare to Table 5.1 is that innovation_work is significant at 0.05 level for models 1 to 7.

Table 5.3 also gives similar results to Table 5.1. Table 5.3 presents unweighted estimates of the regression model of job-satisfaction on innovation_technology and control variables. We see that the nonresponse rate variable is significant at 0.05 level for both model 2 alternative approach and model 5 alternative approach including control variables which use the regression model of job-satisfaction on innovations_technology although it was not accounted for at all in the Bryson et al. (2009) approach. Nevertheless, the inverse mills ratio variable from the Heckman two-step estimator is significant at 0.05 level only for models 3 but not for model 6 which includes control variables in the model. Approximate to Heckman two-step estimator using p_i, λ variable is significant at 0.05 level for both model 4 and model 7 which including control variables in the model. Innovations_technology variable is significant at 0.05 level only for models 1 to 4 which use the regression model of job-satisfaction on innovations_technology but does not include control variables. However, including control variables in models 5 and 7 we see that $1 - p_i$ and λ variables are still significant at 0.05 level although innovations_technology became insignificant and also benefit from using more available auxiliary variables to increase the accuracy of the model.

If we compare the results in Tables 5.1 - 5.3 with the results of Bryson et al. (2009) we can see that we obtain exactly the same result for model 1 leaving $1 - p_i$. If we include the nonresponse rate variable $(1-p_i)$ into both model 2 and model 3 we still get the same results as Bryson et al. (2009). Table 5.1, for example, innovation_all is not significant in model 3 if we include control variables and $1-p_i$. It is also not significant if we leave $1-p_i$ in the regression model for both robust standard error following Bryson et al. (2009) or non-robust standard error. The robcov function (robust covariance matrix estimates) in R package is used for robust standard error which corrects heteroscedasticity and for correlated responses from cluster samples and is normally bigger than non-robust ones.

Panel A : INNOVATIONS_ALL

Table 5.1: The unweighted estimates of the regression model of job-satisfaction on innovations_all and control variables. The T-statistics is shown in parenthesis.

	Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 5**	**Model 6**	**Model 7**
Intercept	4.547(46.735*)	4.779(39.534*)	-1.001(-1.537)	4.773(39.204*)	10.625(22.936*)	10.242(6.042*)	10.612(22.923*)
innovations_all	-0.160(-7.106*)	-0.162(-7.183*)	-0.124(-5.409*)	-0.162(-7.183*)	-0.027(-1.206)	-0.025(-1.123)	-0.027(-1.205)
1-p_i/ imr/λ		-0.755(-3.234*)	35.015(8.614*)	-0.440(-3.086*)	-0.730(-3.168*)	1.840(0.146)	-0.426(-3.029*)
unionrec					-0.355(-2.926*)	-0.348(-2.872*)	-0.355(-2.932*)
member					-0.699(-5.545*)	-0.690(-5.471*)	-0.697(-5.534*)
male					-0.445(-4.147*)	-0.441(-4.109*)	-0.445(-4.153*)
disability					-1.516(-6.636*)	-1.528(-6.687*)	-1.516(-6.638*)
age1					0.247(0.629)	0.216(0.549)	0.245(0.623)
age2					-0.112(-0.395)	-0.140(-0.496)	-0.112(-0.396)
age3					-0.455(-1.688)	-0.472(-1.751)	-0.456(-1.693)
age4					-0.595(-4.210*)	-0.601(-4.246*)	-0.595(-4.207*)
age6					-0.051(-0.402)	-0.051(-0.400)	-0.051(-0.401)
age7					0.070(0.487)	0.068(0.473)	0.070(0.490)
age8					1.536(5.727*)	1.526(5.690*)	1.535(5.724*)
age9					3.176(5.975*)	3.148(5.922*)	3.174(5.971*)
academic2					-0.220(-1.039)	-0.222(-1.047)	-0.219(-1.035)
academic3					-0.125(-0.658)	-0.117(-0.612)	-0.123(-0.647)
academic4					-0.795(-5.171*)	-0.785(-5.103*)	-0.793(-5.159*)
academic5					-0.989(-4.209*)	-0.982(-4.181*)	-0.988(-4.206*)
academic6					-1.116(-5.526*)	-1.105(-5.469*)	-1.115(-5.519*)
academic7					-1.291(-7.298*)	-1.274(-7.204*)	-1.289(-7.289*)
academic8					-1.474(-5.855*)	-1.472(-5.842*)	-1.474(-5.853*)
occupation2					-1.555(-7.595*)	-1.534(-7.493*)	-1.555(-7.592*)
occupation3					-1.727(-10.014*)	-1.713(-9.921*)	-1.726(-10.009*)
occupation4					-2.295(-13.578*)	-2.280(-13.482*)	-2.294(-13.573*)
occupation5					-2.744(-13.056*)	-2.746(-13.062*)	-2.744(-13.058*)
occupation6					-2.067(-8.176*)	-2.087(-8.250*)	-2.069(-8.185*)
occupation7					-3.065(-14.511*)	-3.075(-14.554*)	-3.066(-14.515*)
occupation8					-3.157(-15.205*)	-3.153(-15.177*)	-3.158(-15.209*)
occupation9					-2.936(-14.718*)	-2.956(-14.823*)	-2.937(-14.726*)
manu					-0.884(-3.949*)	-0.859(-3.833*)	-0.880(-3.932*)
utility					-0.512(-1.466)	-0.501(-1.433)	-0.511(-1.463)
construction					0.784(2.947*)	0.804(3.023*)	0.787(2.959*)
wholeret					-0.105(-0.452)	-0.128(-0.554)	-0.103(-0.444)
hotrest					0.431(1.386)	0.395(1.272)	0.436(1.403)
transcom					-0.673(-2.545*)	-0.700(-2.640*)	-0.673(-2.543*)
finserv					-1.411(-5.738*)	-1.353(-5.514*)	-1.405(-5.716*)
othbus					-0.380(-1.736)	-0.350(-1.599)	-0.376(-1.719)
education					0.784(2.545*)	0.788(2.534*)	0.785(2.547*)
health					1.093(4.481*)	1.128(4.625*)	1.096(4.494*)
lemp					-0.945(-5.736*)	-1.020(-3.637*)	-0.947(-5.746*)
lempsq					-0.070(4.178*)	0.078(2.219*)	0.070(4.185*)
durate1					0.541(3.789*)	0.547(3.824*)	0.541(3.787*)

* Significant at the 5 percent level

Panel B : INNOVATION_WORK

Table 5.2: The unweighted estimates of the regression model of job-satisfaction on innovations_work and control variables. The T-statistics is shown in parenthesis.

	Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 5**	**Model 6**	**Model 7**
Intercept	4.481(54.995*)	4.694(43.799*)	-0.955(-1.485)	4.688(43.360*)	10.647(23.032*)	10.349(6.105*)	10.642(23.019*)
innovations_work	-0.286(-8.115*)	-0.286(-8.114*)	-0.234(-6.563*)	-0.286(-8.118*)	-0.100(-2.842*)	-0.100(-2.821*)	-0.101(-2.845*)
1-p_i/ imr/λ		-0.713(-3.057*)	34.546(8.525*)	-0.415(-2.915*)	-0.726(-3.154*)	1.212(0.096)	-0.424(-3.018*)
unionrec					-0.341(-2.819*)	-0.335(-2.766*)	-0.342(-2.824*)
member					-0.692(-5.494*)	-0.682(-5.417*)	-0.691(-5.483*)
male					-0.449(-4.186*)	-0.446(-4.149*)	-0.449(-4.191*)
disability					-1.520(-6.655*)	-1.532(-6.708*)	-1.520(-6.657*)
age1					0.247(0.630)	0.216(0.549)	0.245(0.624)
age2					-0.109(-0.386)	-0.138(-0.486)	-0.110(-0.388)
age3					-0.458(-1.701)	-0.476(-1.764)	-0.460(-1.705)
age4					-0.591(-4.181*)	-0.596(-4.216*)	-0.591(-4.1783*)
age6					-0.050(-0.392)	-0.050(-0.391)	-0.050(-0.392)
age7					0.072(0.500)	0.070(0.487)	0.072(0.503)
age8					1.532(5.715*)	1.523(5.679*)	1.532(5.713*)
age9					3.160(5.945*)	3.132(5.892*)	3.157(5.941*)
academic2					-0.215(-1.016)	-0.217(-1.022)	-0.214(-1.012)
academic3					-0.119(-0.622)	-0.110(-0.575)	-0.117(-0.612)
academic4					-0.791(-5.143*)	-0.780(-5.074*)	-0.789(-5.131*)
academic5					-0.984(-4.190*)	-0.978(-4.162*)	-0.983(-4.187*)
academic6					-1.106(-5.475*)	-1.094(-5.417*)	-1.104(-5.468*)
academic7					-1.286(-7.270*)	-1.269(-7.176*)	-1.284(-7.262*)
academic8					-1.476(-5.861*)	-1.473(-5.848*)	-1.475(-5.859*)
occupation2					-1.570(-7.666*)	-1.549(-7.565*)	-1.569(-7.663*)
occupation3					-1.734(-10.055*)	-1.720(-9.965*)	-1.733(-10.050*)
occupation4					-2.300(-13.616*)	-2.286(-13.522*)	-2.300(-13.612*)
occupation5					-2.757(-13.119*)	-2.760(-13.127*)	-2.757(-13.120*)
occupation6					-2.076(-8.217*)	-2.097(-8.294*)	-2.078(-8.225*)
occupation7					-3.067(-14.524*)	-3.077(-14.566*)	-3.068(-14.527*)
occupation8					-3.168(-15.262*)	-3.164(-15.236*)	-3.169(-15.266*)
occupation9					-2.946(-14.771*)	-2.966(-14.878*)	-2.947(-14.779*)
manu					-0.901(-4.002*)	-0.875(-3.903*)	-0.897(-4.006*)
utility					-0.508(-1.455)	-0.496(-1.420)	-0.507(-1.452)
construction					0.736(2.761*)	0.755(2.833*)	0.739(2.773*)
wholeret					-0.108(-0.468)	-0.131(-0.567)	-0.107(-0.461)
hotrest					0.431(1.388)	0.396(1.275)	0.437(1.405)
transcom					-0.689(-2.605*)	-0.715(-2.699*)	-0.689(-2.603*)
finserv					-1.418(-5.770*)	-1.360(-5.546*)	-1.413(-5.748*)
othbus					-0.397(-1.815)	-0.367(-1.678)	-0.393(-1.798)
education					0.734(2.377*)	0.734(2.358*)	0.734(2.379*)
health					1.106(4.536*)	1.141(4.679*)	1.109(4.548*)
lemp					-0.925(-5.620*)	-0.988(-3.520*)	-0.927(-5.629*)
lempsq					-0.069(4.131*)	0.076(2.151*)	0.069(4.137*)
durate1					0.539(3.776*)	0.546(3.814*)	0.539(3.774*)

* Significant at the 5 percent level

Panel C : INNOVATION_TECHNOLOGY

Table 5.3: The unweighted estimates of the regression model of job-satisfaction on innovation_technology and control variables. The T-statistics is shown in parenthesis.

	Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 5**	**Model 6**	**Model 7**
Intercept	4.261(45.377*)	4.509(37.534*)	-1.658(-2.575*)	4.501(37.534*)	10.543(22.744*)	10.234(6.037*)	10.537(22.730*)
innovation_technolog	-0.180(-3.926*)	-0.189(-4.123*)	-0.116(-2.506*)	-0.188(-4.110*)	-0.044(0.958)	0.052(1.140)	0.044(0.970)
1-p_i/ imr/λ		-0.775(-3.308*)	4.044(9.291*)	-0.450(-3.151*)	-0.709(-3.074*)	1.273(0.101)	-0.413(-2.935*)
unionrec					-0.343(-2.824*)	-0.335(-2.761*)	-0.343(-2.828*)
member					-0.708(-5.622*)	-0.6992(-5.551*)	-0.707(-5.611*)
male					-0.445(-4.149*)	-0.442(-4.115*)	-0.446(-4.154*)
disab					-1.511(-6.614*)	-1.523(-6.664*)	-1.511(-6.615*)
age1					0.249(0.632)	0.218(0.554)	0.246(0.627)
age2					-0.122(-0.431)	-0.151(-0.532)	-0.123(-0.434)
age3					-0.457(-1.696)	-0.474(-1.758)	-0.458(-1.700)
age4					-0.601(-4.250*)	-0.606(-4.287*)	-0.601(-4.247*)
age6					-0.054(-0.418)	-0.054(-0.419)	-0.054(-0.418)
age7					0.066(0.462)	0.064(0.446)	0.067(0.464)
age8					1.533(5.714*)	1.523(5.676*)	1.532(5.711*)
age9					3.178(5.979*)	3.151(5.926*)	3.176(5.975*)
academic2					-0.223(-1.053)	-0.225(-1.059)	-0.222(-1.049)
academic3					-0.131(-0.687)	-0.122(-0.641)	-0.129(-0.676)
academic4					-0.801(-5.207*)	-0.791(-5.142*)	-0.799(-5.196*)
academic5					-0.989(-4.211*)	-0.983(-4.184*)	-0.989(-4.208*)
academic6					-1.125(-5.571)	-1.114(-5.517*)	-1.124(-5.565*)
academic7					-1.298(-7.336*)	-1.282(-7.247*)	-1.296(-7.328*)
academic8					-1.481(-5.880*)	-1.479(-5.869*)	-1.480(-5.877*)
occupation2					-1.556(-7.600)	-1.537(-7.504*)	-1.556(-7.596*)
occupation3					-1.729(-10.021*)	-1.715(-9.935*)	-1.728(-10.016*)
occupation4					-2.295(-13.580*)	-2.281(-13.489*)	-2.294(-13.575*)
occupation5					-2.734(-13.015*)	-2.737(-13.021*)	-2.735(-13.016*)
occupation6					-2.046(-8.093*)	-2.065(-8.162*)	-2.048(-8.101*)
occupation7					-3.058(-14.477*)	-3.067(-14.515*)	-3.059(-14.480*)
occupation8					-3.146(-15.156*)	-3.142(-15.128*)	-3.147(-15.159*)
occupation9					-2.927(-14.675*)	-2.946(-14.776*)	-2.928(-14.682*)
manu					-0.895(-3.995)	-0.872(-3.887*)	-0.892(-3.979*)
utility					-0.509(-1.459)	-0.498(-1.423)	-0.508(-1.455)
construction					0.795(2.993)	0.813(3.059*)	0.798(3.005*)
wholeret					-0.104(-0.448)	-0.126(-0.545)	-0.102(-0.441)
hotrest					0.420(1.352)	0.385(1.239)	0.425(1.368)
transcom					-0.668(-2.524*)	-0.693(-2.616*)	-0.667(-2.522*)
finserv					-1.418(-5.767*)	-1.362(-5.553*)	-1.413(-5.745*)
othbus					-0.376(-1.721)	-0.348(-1.590)	-0.373(-1.704)
education					0.814(2.646*)	0.816(2.630*)	0.815(2.649*)
health					1.094(4.484*)	1.128(4.627*)	1.097(4.497*)
lemp					-0.972(-5.906*)	-1.036(-3.697*)	-0.974(-5.916*)
lempsq					0.071(4.231*)	0.077(2.201*)	0.071(4.238*)
durate1					0.545(3.817*)	0.552(3.855*)	0.545(3.815*)

* Significant at the 5 percent level

5.5.1.2 Survey Weighting of the Regression Model

We did a generalised regression (GREG) estimation in order to construct survey weight. The GREG estimation requires the use of auxiliary information of population totals to create survey weight and it is used for design-based estimations of population totals in survey sampling which we discuss in more detail in Chapter 6. We consider two variables from management files in WERS data in section 5.1; gender and occupation and use the gender and occupation distributions for both responding and sample employees. Then we calculate the unweighted results following the 'WERS 2004 Cross-Section: Survey of Employees Revisions to survey weighting(2007)' method.

We finally obtain the new weight with coefficients of variation equal to 0.848 to use for survey weight and apply this to the regression model that considers survey weight. We use the R package with function svydesign to take into account survey design and the results are shown in Tables 5.4 to 5.6.

Table 5.4 presents the weighted estimates of the regression model of job-satisfaction which regresses on innovations_all. Model 1 naïve approach is shown the regression model of job-satisfaction which regresses on innovations_all. The alternative approach which is the regression model of job-satisfaction which regresses on innovations_all and the nonresponse rate variable $(1 - p_i)$ is shown in model 2 and the alternative approach which includes nonresponse rate variable and control variables is shown in model 3 respectively . We see that nonresponse rate variable is significant at 0.05 level in model 2 alternative approach but it is not significant in model 3 alternative approach including control variables. Moreover, the innovation_all variable is significant in models 1 and 2 but not for model 3 which includes nonresponse rate and control variables.

In comparison with the results in Table 5.1 the unweighted estimates of the regression model of job-satisfaction on innovation_all and control variables in model 1 naïve approach, model 2 alternative approach and model 5 alternative approach including control variables we can see some differences as follows. The response rate variable from the weighted estimates of the regression model from alternative approach including control variables is not significant which is different than the one in the unweighed estimates in the same model and also some control variables became insignificant, e.g. union coverage (unionrec), a quadratic term of log workplace employment size (lempsq) and a dummy variable for low travel-to-work-area unemployment (durate1). Nevertheless, similar results are shown in models 1 and 2.

Table 5.5 presents the weighted estimates of the regression model of job-satisfaction which regresses on innovations_work. Similar to Table 5.4 model 1 represents the naïve approach that is the regression model of job-satisfaction which regresses on innovations_work. The alternative approach which is the regression model of job-satisfaction which regresses on innovations_all and the nonresponse rate variable $(1 - p_i)$ shown in model 2 and the alternative approach which includes nonresponse rate variable and control variables which was shown in model 3. Our results in Table 5.5 give similar patterns to Table 5.4 where we see that the nonresponse rate variable is significant at 0.05 level in model 2 alternative approach but it is not significant in model 3 alternative approach which includes control variables. Moreover, the innovation_all variable is significant in models 1 and 2 but not for model 3 which includes nonresponse rate and control variables.

Compared to the results in Table 5.2 the unweighted estimates of the regression model of job-satisfaction on innovation_work and control variables in model 1 naïve approach, model 2 alternative approach and model 5 alternative approach including control variables, we can see that the response rate variable from the weighted estimates of the regression model from alternative approach including control variables is not significant which is different than the one in the unweighed estimates in the same model and also found that some control variables became insignificant, e.g. a quadratic term of log workplace employment size (lempsq) and a dummy variable for low travel-to-work-area unemployment (durate1). Furthermore, innovation_work variable from the unweighed model became insignificant but nevertheless, similar results are shown in models 1 and 2.

Table 5.6 presents the weighted estimates of the regression model of job-satisfaction which regresses on innovations_technology. Model 1 naïve approach is shown the regression model of job-satisfaction which regresses on innovations_technology. The alternative approach which is the regression model of job-satisfaction which regresses on innovations_technology and the nonresponse rate variable $(1 - p_i)$ shown in model 2 and the alternative approach which includes nonresponse rate variable and control variables shown in model 3 respectively. We see that nonresponse rate variable is significant at 0.05 level in model 2 alternative approach but again it is not significant in model 3 alternative approach including control variables. On the other hand, the innovation_technology variable is not significant in all models.

In comparison with the results in Table 5.3 the unweighted estimates of the regression model of job-satisfaction on innovation_technology and control variables in model 1 naïve approach, model 2 alternative approach and model 5 alternative approach including control variables we can see some differences as follows. The response rate variable from the weighted estimates of the regression model from alternative approach including control variables is not significant. Also, the innovation_technology variable from the weighted estimates of the models 1 and 2 became insignificant.

Panel A :INNOVATIONS_ALL

Table 5.4: The weighted estimates of the regression model of job-satisfaction regresses on innovations_all and control variables. The T-statistics is shown in parenthesis.

	Model 1	**Model 2**	**Model 3**
Intercept	4.524(23.132*)	4.828(18.410*)	10.393(13.180*)
innovations_all	-0.190(-4.022*)	-0.193(-4.050*)	-0.042(-1.003)
1-p_i		-0.992(-2.175*)	-0.719(-1.689)
unionrec			-0.378(-1.832)
member			-0.844(-4.355*)
male			-0.391(-2.549*)
disability			-1.551(-4.332*)
age1			0.203(0.483)
age2			0.140(0.379)
age3			-0.544(-1.640)
age4			-0.500(-2.535)
age6			0.028(0.165)
age7			0.075(0.376)
age8			1.366(4.271*)
age9			3.479(5.307*)
academic2			-0.081(-0.299)
academic3			-0.081(-0.314)
academic4			-0.669(-3.236*)
academic5			-0.957(-3.252*)
academic6			-0.8726(-3.107)
academic7			-1.167(-4.425*)
academic8			-1.328(-3.396*)
occupation2			-1.793(-6.447)
occupation3			-1.702(-7.120*)
occupation4			-2.180(-9.706*)
occupation5			-2.534(-9.497)
occupation6			-2.059(-6.226*)
occupation7			-2.922(-10.249*)
occupation8			-3.112(-11.185)
occupation9			-2.917(-9.543*)
manu			-1.369(-3.074)
utility			-1.406(-1.890)
construction			0.373(0.8317)
wholeret			-0.656(-1.659)
hotrest			0.208(0.338)
transcom			-1.259(-2.584*)
finserv			-2.209(-4.939*)
othbus			-0.681(-1.678)
education			0.687(1.303)
health			0.621(1.436)
lemp			-0.695(-2.483*)
lempsq			-0.048(1.677)
durate1			0.315(1.306)

* Significant at the 5 percent level

Panel B : INNOVATION_WORK

Table 5.5: The weighted estimates of the regression model of job-satisfaction regresses on innovation_work and control variables. The T-statistics is shown in parenthesis.

	Model 1	**Model 2**	**Model 3**
Intercept	4.509(27.586*)	4.783(21.009*)	10.375(13.377*)
innovation_work	-0.385(-5.212*)	-0.385(-5.187*)	-0.152(-2.442)
1-p_i		-0.928(-2.071*)	-0.725(-1.718)
unionrec			-0.352(-1.725*)
member			-0.836(-4.330*)
male			-0.400(-2.621*)
disability			-1.553(-4.341*)
age1			0.202(0.482)
age2			0.151(0.407)
age3			-0.542(-1.630)
age4			-0.490(-2.492*)
age6			0.027(0.156)
age7			0.081(0.407)
age8			1.366(4.270*)
age9			3.473(5.272*)
academic2			-0.074(-0.270)
academic3			-0.068(-0.266)
academic4			-0.659(-3.201*)
academic5			-0.952(-3.249*)
academic6			-0.856(-3.058*)
academic7			-1.155(-4.392*)
academic8			-1.331(-3.406*)
occupation2			-1.822(-6.584)
occupation3			-1.717(-7.196*)
occupation4			-2.189(-9.761*)
occupation5			-2.555(-9.559)
occupation6			-2.074(-6.277*)
occupation7			-2.921(-10.259*)
occupation8			-3.128(-11.300)
occupation9			-2.939(-9.666*)
manu			-1.389(-3.166)
utility			-1.403(-1.919)
construction			0.322(0.726)
wholeret			-0.654(-1.688)
hotrest			0.221(0.362)
transcom			-1.273(-2.654*)
finserv			-2.214(-5.036*)
othbus			-0.694(-1.740)
education			0.615(1.176)
health			0.660(1.538)
lemp			-0.652(-2.331*)
lempsq			-0.045(1.611)
durate1			0.318(1.325)

* Significant at the 5 percent level

Panel C : INNOVATION_TECHNOLOGY

Table 5.6: The weighted estimates of the regression model of job-satisfaction regresses on innovation_technology and control variables. The T-statistics is shown in parenthesis.

	Model 1	**Model 2**	**Model 3**
Intercept	4.125(22.055*)	4.447(17.379*)	10.280(12.956*)
innovation_technology	-0.172(-1.761)	-0.185(-1.903)	0.048(0.533)
1-p_i		-1.003(-2.191*)	-0.672(-1.585)
unionrec			-0.372(-1.800)
member			-0.854(-4.405*)
male			-0.394(-2.572*)
disability			-1.545(-4.322*)
age1			0.223(0.534)
age2			0.121(0.329)
age3			-0.555(-1.671)
age4			-0.503(-2.558)
age6			0.028(0.164)
age7			0.072(0.362)
age8			1.364(4.257*)
age9			3.479(5.256*)
academic2			-0.085(-0.312)
academic3			-0.089(-0.346)
academic4			-0.677(-3.276*)
academic5			-0.9589(-3.259*)
academic6			-0.890(-3.177)
academic7			-1.177(-4.469*)
academic8			-1.334(-3.421*)
occupation2			-1.783(-6.418)
occupation3			-1.697(-7.077*)
occupation4			-2.177(-9.680*)
occupation5			-2.504(-9.362)
occupation6			-2.013(-6.101*)
occupation7			-2.915(-10.249*)
occupation8			-3.085(-11.033)
occupation9			-2.899(-9.481*)
manu			-1.370(-3.062)
utility			-1.409(-1.883)
construction			0.398(0.889)
wholeret			-0.648(-1.625)
hotrest			0.207(0.339)
transcom			-1.242(-2.514*)
finserv			-2.203(-4.890*)
othbus			-0.673(-1.652)
education			0.733(1.388*)
health			0.626(1.452)
lemp			-0.738(-2.627*)
lempsq			0.049(1.730)
durate1			0.305(1.268)

* Significant at the 5 percent level

5.5.2 The Results of the Regression Model at Workplace Level

Similar to the previous section, the results are divided into two sections; the regression model when not considering a weighted survey and considering a weighted survey for three panels (A:innovations_all, B:innovations_work and C:innovations_technology) as follows:

5.5.2.1 Unweighted Estimates of the Regression Model

Following the theory in chapter 3, we again look at the alternative approach including the nonresponse variable $(1-p_i)$ into the regression model at workplace level. We applied the alternative approach to real data at workplace level beside of simulation study in chapter 4. Table 5.7 and 5.8 present the results of the regression model of job-satisfaction regressed on innovations (innovation_all, innovation_work and innovation_technology) and nonresponse rate (n_r) considering non-weighted and weighted surveys respectively. We see a difference in results when we compare the individual level results with the workplace level results because nonresponse rate is not significant at workplace level but at individual level it is at 0.05 in both unweighted and weighted surveys. Bryson et al. (2009) did not consider the regression model at workplace level.

Table 5.7: The unweighted estimates of the regression model of job-satisfaction regresses on innovations. The T-statistics is shown in parenthesis.

		Model 1	**Model 2**
Panel A : INNOVATIONS_ALL			
	Intercept	4.943(29.818*)	5.106(23.353*)
	innovations_all	-0.213(-5.307*)	-0.216(-5.362*)
	1-p_i		-0.393(-1.142)
Panel B : INNOVATION_WORK			
	Intercept	4.816(34.297*)	4.941(25.293*)
	innovations_work	-0.357(-5.631*)	-0.358(-5.640*)
	1-p_i		-0.314(-0.915)
Panel C : INNOVATION_TECHNOLOGY			
	Intercept	4.602(28.865*)	4.7769(21.839*)
	innovations_technology	-0.249(-3.073*)	-0.258(-3.173*)
	1-p_i		-0.402(-1.157)

* Significant at the 5 percent level

5.5.2.2 Survey Weighting of the Regression Model

Table 5.8: The weighted estimates of the regression model of job-satisfaction regresses on innovations. The T-statistics is shown in parenthesis.

		Model 1	**Model 2**
Panel A : INNOVATIONS_ALL			
	Intercept	4.990(20.743*)	5.374(17.316*)
	innovations_all	-0.213(-3.898*)	-0.219(-3.986*)
	1-p_i		-0.951(-1.788)
Panel B : INNOVATION_WORK			
	Intercept	4.930(25.111*)	5.255(19.319*)
	innovations_work	-0.414(-4.927*)	-0.414(-4.893*)
	1-p_i		-0.849(-1.605)
Panel C : INNOVATION_TECHNOLOGY			
	Intercept	4.635(19.647*)	5.043(16.658*)
	innovations_technology	-0.215(-1.858*)	-0.238(-2.075*)
	1-p_i		-0.962(-1.823)

* Significant at the 5 percent level

5.5.3 Comparing the Differences in the Significance of the Coefficients between the Models that Take into Account the Complex Survey Design and the Models that Ignore the Complex Survey Design

Considering the models that ignore the complex survey design at individual level the nonresponse rate variable is significant at 0.05 level for both the regression model of job-satisfaction with regresses on innovation (innovations_all, innovations_work and innovations_technology) and nonresponse rate included, and the regression model of job-satisfaction regressed on innovation (innovations_all, innovations_work and innovations_technology) and the nonresponse rate and control variables (Similarly, if we consider robust standard error for this model, nonresponse rate is also significant).

However, the nonresponse rate is not significant at workplace level for the models that ignore complex survey design. If we look at control variables we can see that the results do not change very much between unweighted and weighted surveys e.g. academic qualification and occupation variables do not change but some of the results from single-digit industry variables became insignificant e.g. education.

For the models that take into account the complex survey design at individual level nonresponse rate is still significant at 0.05 level but only when the model of job-satisfaction is regressed on innovation (innovations_all, innovations_work and innovations_technology) and on the nonresponse rate, but not for regression models that include control variables.

Nevertheless, if we consider some models of job-satisfaction regressed on innovation (e.g. innovations_all, innovations_work), and on nonresponse rate and control variables the nonresponse rate is significant at 0.10 level.

We can see that the estimation of the intercept and innovation variables does not change much between weighted and unweighted models but in some cases the results are affected e.g. Panel C: innovation_technology.

On the other hand, for the models that take into account the complex survey design at workplace level nonresponse rate is again not significant at 0.05 level although there are some cases like that found in the regression model of job-satisfaction on innovation (innovations_all, and innovations_technology) and nonresponse rate where nonresponse rate is significant at 0.10 level.

We can see that the estimation of the intercept and innovation variables does not change much between the weighted and unweighted models while the regression coefficient of nonresponse term for survey weighted model looks better but still not significant at 0.05 level.

In conclusion, there is a problem with the model at workplace level as nonresponse rate is not significant. Therefore we will do further analysis by applying lowess plot and weighted least square regression(WLS) onto our model. The details are shown in section 5.5.4 as follows:

5.5.4 Weighted Least Square Regression

According to the results for regression models at workplace level we can see that the nonresponse rate variable is not significant and therefore we will do some more analysis by looking at the lowess plot between the residuals from the regression models in section 5.5.1 (unweighted estimates) and response rate (p_i). We can see evidence of unequal variance. Hence, we use the number of employee returned questionnaires (nnumseq) as the weight in the regression models. In this case the nnumseq is inverse to $Var(\epsilon_i)$, because the nnumseq is inverse to $Var(\epsilon_i)$ we define the weights as below:

The weights = nnumseq(number of employees returned)

We consider both normal standard error and robust standard error. The OLS function in R package is used for WLS and also robcov function for robust standard error. The results are shown as follows:

WLS with normal standard error

Table 5.9: The regression model of job-satisfaction regresses on innovations. The T-statistics is shown in parenthesis.

		Model
Panel A :INNOVATIONS_ALL		
	Intercept	4.776(25.861*)
	innovations_all	-0.163(-4.735*)
	1-p_i	-0.761(-2.139*)
Panel B : INNOVATION_WORK		
	Intercept	4.697(28.718*)
	innovations_work	-0.292(-5.426*)
	1-p_i	-0.713(-2.010*)
Panel C : INNOVATION_TECHNOLOGY		
	Intercept	4.494(24.334*)
	innovations_technology	-0.184(-2.612*)
	1-p_i	-0.784(-2.185*)

* Significant at the 5 percent level

Table 5.9 present the results for the regression of job-satisfaction regresses on innovations (innovation_all, innovation_work and innovation_technology). We see nonresponse rate variable is significant at 0.05 level for all cases.

WLS with robust standard error

Table 5.10: The regression model of job-satisfaction regresses on innovations. The T-statistics is shown in parenthesis.

		Model
Panel A : INNOVATIONS_ALL		
	Intercept	4.776(28.994*)
	innovations_all	-0.163(-5.358*)
	1-p_i	-0.761(-1.417)
Panel B :INNOVATION_WORK		
	Intercept	4.697(30.483*)
	innovations_work	-0.292(-6.247*)
	1-p_i	-0.713(-1.325)
Panel C : INNOVATION_TECHNOLOGY		
	Intercept	4.494(26.918*)
	innovations_technology	-0.184(-2.876*)
	1-p_i	-0.784(-1.461)

* Significant at the 5 percent level

Table 5.10 presents the results for the regression of job-satisfaction regressed on innovations (innovation_all, innovation_work and innovation_technology) with robust standard error. Unfortunately the nonresponse rate variable is not significant at 0.05 level for all cases.

5.6 Conclusion

The alternative approach performs well at individual level for both unweighted and weighted surveys on the job-satisfaction regressed on innovation variables and nonresponse rate but if we include control variables into the model the nonresponse rate variable becomes insignificant at 0.05 level for weighted surveys but is still significant for unweighted model.

The alternative approach does not work well at workplace level for both unweighted and weighted surveys. However, the alternative approach performs better after we apply WLS into workplace level models but only for unweighted estimates.

The Heckman two-step estimator also works well because the inverse mills ratio variable is significant at 0.05 level but only for the model of job-satisfaction regressed on innovation variable not for model which includes control variables in the model.

The approximate to Heckman two-step estimator using p_i perform well in both models of the job-satisfaction regressed on innovation variables and also the model including control variables.

If we compare the results of real data with the simulation results from Chapter 4, we can see that the alternative approach only works well at individual level but not at workplace level. There might be some specific reasons why this set of real data does not suit the alternative approach. For example, we assumed the population has normal distribution in the simulation study but real data distribution might not have normal distribution and moreover the variance is not constant and that is why we applied WLS for workplace level. The alternative approach seems to work well after we used WLS with normal standard error.

Chapter 6

GREG estimators for Two-Stage Sampling

6.1 Introduction

This chapter considers a new topic not covered in the previous chapters but with a similar framework, population and sampling set up. The generalised regression estimator (GREG) for two-stage sampling is considered.

In section 6.2, we review the literature related to the GREG estimators. In section 6.3, the customary GREG estimator is considered and in section 6.4, we will propose a new GREG estimator for two stage sampling. Finally, in section 6.5 we will show the simulation results.

6.2 Literature Review

The generalised regression (GREG) estimator is used for design-based estimation of population totals in survey sampling. The GREG estimator uses auxiliary information and is a special type of calibration estimator. In this chapter we introduce some new ways of using this GREG estimator under two-stage sampling. Some early works have been done on GREG and the calibration estimators. Bethlehem and Keller (1987) proposed a weighting method to calculate weights using linear regression models at the person level producing better results than post- stratification other than where there is only one qualitative auxiliary variable in the model. The new weighting method can address the issues occurring in post-stratification specifically where some strata do not contain any members and details about population are not known but still where their theory is limited to simple random sampling.

Alexander (1987) proposed methods to find household weights which are subject to the constraints of consistency with known control counts in data. This data contains many cells with data about persons where the value of the weightings are close to the initial calculated vectors of household weightings. He considered three types of methods called constrained minimum distant methods which include the principal person method used under two-stage cluster sampling. The proposed method has been compared to the principal person method. The conclusion is that any analysis using these methods would best be served by first gathering more information about survey undercoverage in order to decide which method would provide the best results.

Lemaître and Dufour (1987) proposed an integrated procedure that calculates household weights and can be used to estimate person weights as well. This method is based on assumptions made by Bethlehem and Keller (1987). In order to apply this method, they suggested using the household mean instead of the corresponding auxiliary variables at the person level, where the same value will be applied to each person within a household. They also compare the efficacy of these methods by applying them to real data in test scenarios. They chose the Canadian Labour Force Survey in order to see how their estimators perform. The estimators both gave unbiased and similar results. However there might be a chance of negative weightings using this method.

Later, Steel and Clark (2007) considered generalized regression estimations at household level where people within households have equal weightings. The weight are called integrated weights. They also compared the design variance of GREG estimators at the household level with GREG estimators at the person level in terms both theoretically and empirically where they point out that this was not covered *at all* in Alexander (1987) and Lemaître and Dufour (1987) . The optimal estimator for simple cluster sampling, the explanation of the difference in the asymptotic variances and the linear contextual of GREG estimators are all discussed theoretically. The results show that GREG estimators at the household level have smaller variance than GREG estimators at the person level in large samples. We can see the benefit of this research in looking at sampling variance of GREG estimators at the household level instead of the person level only. However, their sampling plan is limited to single stage cluster sampling.

Montanari (1987) proposed a new GREG estimator with design optimality if the population regression coefficient is known. However, the population regression coefficient is usually unknown and it has to be estimated from the sample. This optimal GREG estimator is also very complex to implement for two stage sampling designs because it requires joint-inclusion probabilities. Berger et al. (2003) proposed an optimal GREG estimator based on the Montanari (1987) estimator. Their optimal GREG estimator is not dependent on joint-inclusion probability. Berger et al. (2003) showed that their estimator may be more accurate than the Montanari (1987) estimator and the standard generalised regression estimator. Nevertheless, their simulation is limited to single stage sampling design. Recently, Tan (2013) proposed an optimal regression estimator which is a particular case of the estimator proposed by Berger et al. (2003). Tan (2013) proposed to expand the calibration estimators that have design-efficiency for the case of known population totals or measured auxiliary variables for all units in the population in both sampling techniques; rejective or high-entropy samplings in the presence of missing data in survey samplings. The proposed estimators have a similar property to an optimal regression estimator that has been proposed by many authors including Montanari (1987). Nevertheless, they showed that the new method can solve two problems, one always existent in the efficiency of a linear superpopulation model applying generalized regression and calibration estimation and also it offers an easy way to approximate the optimal regression estimation.

Rao (1994) considered the use of auxiliary information at estimation stage for the estimation of both the population totals and distribution functions by giving a general set-up for making the estimations under probability sampling and model-assisted approaches. Rao proposed alternative model-assisted estimators having conditional repeated sampling inferences for dealing with model misspecification. He also proposed an optimal calibration estimator that is more accurate than the GREG estimator or the basic estimator of population total and that also expresses in calibration form under stratified simple random sampling and stratified multi-stage sampling.

Estevao and Särndal (2006) studied many scenarios in complex survey designs using calibrations such as the estimation of domains in one-phase sampling, estimation for two-phase sampling, and estimation for two-stage sampling. They first reviewed auxiliary information used in one phase survey design. A vector of auxiliary variables with known population totals is used to calculate weights and corresponding calibration estimators and this helps to decrease variance over estimators that do not account for this auxiliary information. They reviewed each step used during their exploration of instrument vector approach and automated linearisation. They also examined calibration estimation for use with two-phase sampling and two-stage cluster sampling where auxiliary information is available at both the cluster and unit level. They also discussed integrated weightings required for combining auxiliary information with two stage data and moreover; they compare their method with the approach of Lemaître and Dufour (1987) and on the issue of equally weighting individuals within selected households they also discuss the effects of residuals on two-stage estimation of both unit and cluster statistics.

A literature review can be found in Särndal (2007). He compares the generalized regression estimation with methods given in them pointing out that it is not the same way of using auxiliary information in the estimation process although it is a special case calibration estimator. He discussed how the methods can be used with both simple and more complex survey design where sampling in two or more phases or stages is used. Discussion was given on how effective approaches might be in situations of complex survey design where the auxiliary information might be available for more than one component e.g. there might be primary sampling unit and/or secondary sampling unit information available for two-stage design. Finally, calibration for nonresponse adjustment and non-sampling error are investigated (see also Skinner (1998), Särndal and Lundström (2005), and Kott (2006) for nonresponse adjustment).

6.3 GREG Estimation for Single Stage Sampling

6.3.1 The Customary GREG Estimator

Consider a finite population $U = \{1, 2, ..., j, ..., N\}$. Let y_j be the value of the study variable y for the jth population unit. The aim is to estimate the unknown population total t_y given by

$$t_y = \sum_{j=1}^{N} y_j. \tag{6.1}$$

The Horvitz-Thompson estimator for t_y is calculated from a sample s of size n drawn from U and is given by

$$\widehat{t}_{y\pi} = \sum_{j=1}^{n} y_j/\pi_j = \sum_{j=1}^{n} \breve{y}_j = \boldsymbol{Y}_s'\boldsymbol{\Sigma}_s\boldsymbol{1_s}, \tag{6.2}$$

where $\boldsymbol{Y}_s = (y_1, y_2..., y_n)', \boldsymbol{\Sigma}_s = diag(d_j)$, d_j is the design weights defined $d_j = 1/\pi_j$, π_j is the inclusion probabilities for the j^{th} element and $\boldsymbol{1_s}$ is a vector of dimension n with all one units.

Suppose we have auxiliary information available. Let $\boldsymbol{x}_j$ be the vector of k auxiliary variables for the jth unit, $\boldsymbol{x}_j = (\boldsymbol{x}_{j1}, ..., \boldsymbol{x}_{jk})'$. We assume vector total $\boldsymbol{t_x}$ is known and given by

$$\boldsymbol{t}_x = \sum_{j=1}^{N} \boldsymbol{x}_j. \tag{6.3}$$

Suppose we want to use the auxiliary information $\boldsymbol{t}_x$ to estimate t_y. The GREG estimator is given by

$$\widehat{Y}_{GREG} = \sum_{j=1}^{n} w_j y_j, \tag{6.4}$$

where $\boldsymbol{w}_j$ denote GREG weight given by

$$w_j = d_j(1 + \lambda' \boldsymbol{x}_j), \tag{6.5}$$

with

$$\lambda' = (\boldsymbol{t}_x - \widehat{t}_{x\pi})' \left(\sum_{j=1}^{n} d_j \boldsymbol{x}_j \boldsymbol{x}_j' \right)^{-1}, \tag{6.6}$$

where $\widehat{\boldsymbol{t}}_{x\pi}$ is the Horvitz-Thompson estimator given by

$$\widehat{\boldsymbol{t}}_{x\pi} = \sum_{j=1}^{n} x_j/\pi_j = \sum_{j=1}^{n} \breve{x}_j = \boldsymbol{X}_s' \boldsymbol{\Sigma}_s \boldsymbol{1}_s, \tag{6.7}$$

where $\boldsymbol{X}_s = (\boldsymbol{x}_1, \boldsymbol{x}_2 ..., \boldsymbol{x}_n)'$.

Alternatively, the GREG estimator in (6.4) can be rewritten in matrix form as follows.

$$\widehat{Y}_{GREG} = \widehat{t}_{y\pi} + (\boldsymbol{t}_x - \widehat{\boldsymbol{t}}_{x\pi})' \widehat{\boldsymbol{\beta}}_{xy}, \tag{6.8}$$

where

$$\widehat{\boldsymbol{\beta}}_{xy} = (\breve{\boldsymbol{X}}_s' \boldsymbol{\Sigma}_s \breve{\boldsymbol{X}}_s)^{-1} \breve{\boldsymbol{X}}_s' \boldsymbol{\Sigma}_s \breve{\boldsymbol{Y}}_s, \tag{6.9}$$

$$\breve{\boldsymbol{X}}_s = (\breve{\boldsymbol{x}}_1, \breve{\boldsymbol{x}}_2 ..., \breve{\boldsymbol{x}}_n)' \qquad \text{with} \qquad \breve{\boldsymbol{x}}_j = \boldsymbol{x}_j/\pi_j, \tag{6.10}$$

$$\breve{\boldsymbol{Y}}_s = (\breve{y}_1, \breve{y}_2 ..., \breve{y}_n)' \qquad \text{with} \qquad \breve{y}_j = y_j/\pi_j, j \in s, \tag{6.11}$$

$\widehat{\boldsymbol{t}}_{y\pi}$, $\boldsymbol{t}_x$ and $\widehat{\boldsymbol{t}}_{x\pi}$ are defined by (6.2), (6.3) and (6.7) respectively.

Note that the weights w_j are calibrated because they are such that

$$\sum_{j=1}^{n} w_j \boldsymbol{x_j} = \sum_{j=1}^{N} \boldsymbol{x_j}. \tag{6.12}$$

6.3.2 The Optimal GREG Estimator (Montanari 1987)

Montanari (1987) considered the following random variable

$$\breve{Y}_M = \widehat{t}_{y\pi} + (\boldsymbol{t}_x - \widehat{\boldsymbol{t}}_{x\pi})'\boldsymbol{\beta}_M, \tag{6.13}$$

where

$$\boldsymbol{\beta}_M = var(\widehat{\boldsymbol{t}}_{x\pi})^{-1}cov(\widehat{\boldsymbol{t}}_{x\pi}, \widehat{\boldsymbol{t}}_{y\pi}) = (\breve{\boldsymbol{X}}'_U\boldsymbol{\Delta}_U\breve{\boldsymbol{X}}_U)^{-1}\breve{\boldsymbol{X}}'_U\boldsymbol{\Delta}_U\breve{\boldsymbol{Y}}_U, \boldsymbol{\beta}_M \tag{6.14}$$

is a population parameter of size $N \times N$ positive matrix, $\breve{\boldsymbol{X}}_U = (\breve{\boldsymbol{x}}_1, \breve{\boldsymbol{x}}_2..., \breve{\boldsymbol{x}}_N)'$ with $\breve{\boldsymbol{x}}_j = \boldsymbol{x}_j/\pi_j$, $\breve{\boldsymbol{Y}}_U = (\breve{y}_1, \breve{y}_2..., \breve{y}_N)'$, $j \in U$ and $\boldsymbol{\Delta}_U$ is the $N \times N$ matrix given by

$$\boldsymbol{\Delta}_U = [\Delta_{ij}], \tag{6.15}$$

where $\Delta_{ij} = \pi_{ij} - \pi_i\pi_j$, and π_{ij} is the joint inclusion probability of units i and j.

Montanari (1987) showed that $\breve{Y}_M$ is optimal because the expectation of $\breve{Y}_M$ equal t_y and the variance of $\breve{Y}_M$ is minimal. Montanari (1987) proposed to predict $\widehat{Y}_M$ by $\widehat{Y}_M$ after the substitution of optimal choice β_M by $\widehat{\beta}_M$ in order to minimize $Var(Y_M)$, where $\widehat{\boldsymbol{\beta}}_M$ is the estimator of $\boldsymbol{\beta}_M$ given by

$$\widehat{\boldsymbol{\beta}}_M = (\breve{\boldsymbol{X}}'_s\breve{\boldsymbol{\Delta}}_s\breve{\boldsymbol{X}}_s)^{-1}\breve{\boldsymbol{X}}'_s\breve{\boldsymbol{\Delta}}_s\breve{\boldsymbol{Y}}_{\boldsymbol{s}}, \tag{6.16}$$

where $\breve{\boldsymbol{\Delta}}_s = [\Delta_{ij}\pi_{ij}^{-1}]$, and $\widehat{t}_{y\pi}, \boldsymbol{t}_x$ and $\widehat{\boldsymbol{t}}_{x\pi}$ are defined by (6.2) (6.3) and (6.7) respectively. The proposed GREG estimator, is given by

$$\widehat{Y}_M = \widehat{t}_{y\pi} + (\boldsymbol{t}_x - \widehat{\boldsymbol{t}}_{x\pi})'\widehat{\boldsymbol{\beta}}_M, \tag{6.17}$$

6.3.3 Optimal GREG Estimator proposed by Berger et al. (2003)

Berger et al. (2003) showed that under single stage design a consistent estimator of $\boldsymbol{\beta}_M$ can be obtained by including the stratification variable into the regression estimator. Berger et al. (2003) adjusted the Montanari (1987)'s estimator by replacing $\boldsymbol{\Delta}_U$ by

$$\widetilde{\boldsymbol{\Delta}}_U = \boldsymbol{C}_U(\boldsymbol{I}_U - \breve{\boldsymbol{Q}}_U(\breve{\boldsymbol{Q}}'_U\boldsymbol{C}_U\breve{\boldsymbol{Q}}_U) - \breve{\boldsymbol{Q}}'_U\boldsymbol{C}_U) \tag{6.18}$$

which is the estimator of $\boldsymbol{\Delta}_U$ under a conditional stratified Poisson sampling (CSPS) design, where $\boldsymbol{C}_{\boldsymbol{U}} = diag(c_j; j \in U)$, with $c_j = \pi_j(1 - \pi_j)$, I_U is an identity matrix of size $N \times N$, and $\breve{\boldsymbol{Q}}_U$ is the matrix of size $N \times H$ of stratification variables that contain $\breve{q}_{hj} = q_{hj}/\pi_j$, $h = 1, 2, \ldots, H$ and $j = 1, 2, \ldots, N$ where $q_{hj} = \pi_j$ if the jth unit belong to stratum h and otherwise $q_{hj} = 0$. Berger et al. (2003) showed that if we replace $\boldsymbol{\Delta}_U$ by $\widetilde{\boldsymbol{\Delta}}_U$ we will get $\boldsymbol{\beta}_M$ which is the vector of the first elements of M auxiliary variables, $(\boldsymbol{\beta}_\Gamma^{opt})$ of the vector

$$\boldsymbol{\beta}_\Gamma^{opt} = (\breve{\boldsymbol{\Gamma}}_U' \boldsymbol{C}_U \breve{\boldsymbol{\Gamma}}_U)^{-} \breve{\boldsymbol{\Gamma}}_U' \boldsymbol{C}_U \breve{\boldsymbol{Y}}_U, \tag{6.19}$$

where $\breve{\boldsymbol{\Gamma}}_U = [\breve{\boldsymbol{X}}_U, \breve{\boldsymbol{Q}}_U]$ represents the partitioned $N \times (k + H)$ matrix.

The estimator of $\boldsymbol{\beta}_\Gamma^{opt}$ is given by

$$\widehat{\boldsymbol{\beta}}_\Gamma^{opt} = (\breve{\boldsymbol{\Gamma}}_s' \breve{\boldsymbol{C}}_s \breve{\boldsymbol{\Gamma}}_s)^{-} \breve{\boldsymbol{\Gamma}}_s' \breve{\boldsymbol{C}}_s \breve{\boldsymbol{Y}}_s, \tag{6.20}$$

where $\breve{\boldsymbol{\Gamma}}_s = [\breve{\boldsymbol{X}}_s, \breve{\boldsymbol{Q}}_s]$ is a $n \times (k + H)$ matrix, $\breve{\boldsymbol{X}}_s$ is the $n \times k$ matrix defined earlier in section 6.3.1, $\breve{\boldsymbol{C}}_s = diag(\breve{c}_j), j \in s$ with $\breve{c}_j = 1 - \pi_j$, with $\breve{\boldsymbol{Q}}_s$ is the matrix of size $n \times H$ of stratification variables that contain $\breve{q}_{hj} = q_{hj}/\pi_j$, $h = 1, 2, \ldots, H$ and $j = 1, 2, \ldots, n$ where $q_{hj} = \pi_j$ if the jth unit belong to stratum h and otherwise $q_{hj} = 0$.

The GREG estimator proposed by Berger et al. (2003) is given by

$$\widehat{Y}_{opt} = \widehat{t}_{y\pi} + (\boldsymbol{t}_{xq} - \widehat{\boldsymbol{t}}_{xq\pi})' \widehat{\boldsymbol{\beta}}_\Gamma^{opt}, \tag{6.21}$$

where $\widehat{\boldsymbol{t}}_{y\pi}$ is defined by (2), $\boldsymbol{t}_{xq} = [\sum_{j=1}^N \boldsymbol{q}_{1j}, ..., \sum_{j=1}^N \boldsymbol{q}_{Hj}, \sum_{j=1}^N \boldsymbol{x}_j]'$ and $\widehat{\boldsymbol{t}}_{xq\pi} = [\sum_{j=1}^n \breve{\boldsymbol{q}}_{1j}, ..., \sum_{j=1}^N \breve{\boldsymbol{q}}_{Hj}, \sum_{j=1}^N \breve{\boldsymbol{x}}_j]'$.

The simulation results of Berger et al. (2003) showed that their proposed estimator performed better than the standard generalised regression estimator, particularly for stratified sampling design, including the Montanari estimator which gave a poorer result than others in terms of higher relative standard error. Nevertheless, in some situations the generalised regression estimator performed better than other estimators, particularly for stratified sampling with small sample size (two units in each stratum) and high correlation between y and x.

6.4 GREG Estimator for two stage sampling

Let N be the number of primary sampling unit (PSU) in the population, and M_i the number of secondary sampling unit (SSU) in PSU i where $i = 1, 2, \ldots, N$. Let y_{ij} be the value of the study variable y for the jth SSU $(j = 1, 2, \ldots, M_i)$ of the ith PSU $(i = 1, 2, \ldots, N)$. For two-stage sampling, a sample of n PSU is selected and a sample of m_i SSU $(i = 1, 2, \ldots, n)$ is selected in each sampled PSU. Let π_i be the inclusion probabilities for the ith PSU for the first-stage sampling, and let $\pi_{j|i}$ be the inclusion probabilities of the jth SSU of PSU i.

In two-stage sampling we may have auxiliary information available at both PSU and SSU level. Let $\boldsymbol{z}_i$ be the PSU vector of p auxiliary variables in PSU i, $\boldsymbol{z}_i = (\boldsymbol{z}_{i1}, ..., \boldsymbol{z}_{ip})'$. Let $\boldsymbol{x}_{ij}$ be the SSU vector of k auxiliary variables in PSU i and SSU j, $\boldsymbol{x}_{ij} = (\boldsymbol{x}_{ij1}, ..., \boldsymbol{x}_{ijk})'$.

We assume that the sample of PSU is stratified, the population of PSU is divided into H strata and within stratum h SSU are grouped into N_h PSU. Let N be the number of PSU in the population, $N = \sum_{h=1}^{H} N_h$. Let n be the number of PSU in the sample, $n = \sum_{h=1}^{H} n_h$. Let M_h be the number of SSU in stratum $h, M_h = \sum_{i=1}^{N_h} M_{hi}$. Let m_h be the number of sampled SSU in stratum $h, m_h = \sum_{i=1}^{n_h} m_{hi}$, and let m be the total over all strata, $m = \sum_{h=1}^{H} m_h$.

For stratified two-stage sampling, a sample of n_h PSU is selected in each stratum h from the total of N_h PSU in stratum h and a subsample of m_{hi} SSU is selected in each sampled PSU (h_i) from the total of M_{hi} SSU in the PSU, where $h = 1, 2, \ldots, H$.

The aim is to estimate the unknown population total t_y given by

$$t_y = \sum_{i=1}^{N} \sum_{j=1}^{M_i} y_{ij}, \tag{6.22}$$

where y_{ij} is the value of the study variable y for the j^{th} SSU $(j = 1, 2, \ldots, M_i)$ of the i^{th} PSU $(i = 1, 2, \ldots, N)$.

The Horvitz-Thompson estimator for t_y is given by

$$\widehat{t}_{y\pi} = \sum_{i=1}^{n} \sum_{j=1}^{m_i} y_{ij}/\pi_{ij} = \sum_{i=1}^{n} \sum_{j=1}^{m_i} \breve{y}_{ij} = \mathbf{Y}'_s \mathbf{\Sigma}_s \mathbf{1}_s, \tag{6.23}$$

where $\pi_{ij} = \pi_i \times \pi_{j|i}, \boldsymbol{Y}_s = (y_{11}, y_{12} ..., y_{nm_n})', \mathbf{\Sigma}_s = diag(d_{ij})$, d_i and d_{ij} are the design weights defined $d_i = 1/\pi_i$ and $d_{j|i} = 1/\pi_{j|i}$, and $\mathbf{1}_s$ is a vector of dimension $(m_1 + ..., +m_n)$ with all one units. The overall design weight for the j-th SSU in the i-th PSU is given by $d_{ij} = d_i d_{j|i}$, $i = 1, 2, \ldots, N, j = 1, 2, \ldots, M_i$.

We assume both the SSU and PSU vector total $\boldsymbol{t}_x = \sum_{i=1}^{N} \sum_{j=1}^{M_i} \boldsymbol{x}_{ij}$ and $\boldsymbol{t}_z = \sum_{i=1}^{N} \boldsymbol{z}_i$ are known. The vector of population totals is given by

$$\boldsymbol{t}_{xz} = \begin{bmatrix} \sum_{i=1}^{N} \boldsymbol{z}_i \\ \sum_{i=1}^{N} \sum_{j=1}^{M_i} \boldsymbol{x}_{ij} \end{bmatrix} = \begin{bmatrix} \boldsymbol{t}_z \\ \boldsymbol{t}_x \end{bmatrix}. \tag{6.24}$$

6.4.1 The Estevao and Särndal (2006) regression estimator

Suppose we have auxiliary information available at PSU and SSU level, let $\boldsymbol{z}_i$ be cluster level variable and let $\boldsymbol{x}_{ij}$ be unit level variable, Estevao and Särndal (2006) suggested to use

$$\boldsymbol{z}_{ij} = \boldsymbol{z}_i / M_i \tag{6.25}$$

to assign to every selected unit in PSU i.

The GREG estimator using just the Z information is

$$\widehat{Y}_{GREG}^{(1)} = \widehat{t}_{y\pi} + (\boldsymbol{t}_z - \widehat{\boldsymbol{t}}_{z\pi})' \widehat{\boldsymbol{\beta}}_{zy}, \tag{6.26}$$

where

$$\widehat{\boldsymbol{\beta}}_{zy} = (\boldsymbol{Z}_s' \boldsymbol{\Sigma}_s \boldsymbol{Z}_s)^{-1} \boldsymbol{Z}_s' \boldsymbol{\Sigma}_s \boldsymbol{Y}_s, \boldsymbol{Z}_s = (\boldsymbol{z}_{11}, \boldsymbol{z}_{12} ..., \boldsymbol{z}_{nm_n})' \tag{6.27}$$

and

$$\widehat{\boldsymbol{t}}_{z\pi} = \sum_{i=1}^{n} \sum_{j=1}^{m_i} z_{ij} / \pi_{ij} = \boldsymbol{Z}_s' \boldsymbol{\Sigma}_s \boldsymbol{1}_s. \tag{6.28}$$

Estevao and Särndal (2006) also suggested using $\boldsymbol{x}_{ij}$ and $\boldsymbol{z}_{ij}$ in GREG estimator. A GREG estimator is

$$\widehat{Y}_{GREG}^{(2)} = \widehat{t}_{y\pi} + (\boldsymbol{t}_{xz} - \widehat{\boldsymbol{t}}_{xz\pi})' \widehat{\boldsymbol{\beta}}_{\boldsymbol{x}zy}, \tag{6.29}$$

where

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{x}zy} = (\boldsymbol{W}_s' \boldsymbol{\Sigma}_s \boldsymbol{W}_s)^{-1} \boldsymbol{W}_s' \boldsymbol{\Sigma}_s \boldsymbol{Y}_s, \tag{6.30}$$

$$\widehat{\boldsymbol{t}}_{xz\pi} = (\widehat{\boldsymbol{t}}_{x\pi}', \widehat{\boldsymbol{t}}_{z\pi}')' = \boldsymbol{W}_s' \boldsymbol{\Sigma}_s \boldsymbol{1}_s \tag{6.31}$$

and

$$\widehat{\boldsymbol{t}}_{x\pi} = \sum_{i=1}^{n} \sum_{j=1}^{m_i} x_{ij} / \pi_{ij} = \boldsymbol{X}_s' \boldsymbol{\Sigma}_s \boldsymbol{1}, \boldsymbol{X}_s = (\boldsymbol{x}_{11}, \boldsymbol{x}_{12} ..., \boldsymbol{x}_{nm_n})' \tag{6.32}$$

and a stacked auxiliary vector is defined as

$$\boldsymbol{W}_s = \begin{bmatrix} \boldsymbol{Z}_s \\ \boldsymbol{X}_s \end{bmatrix}. \tag{6.33}$$

6.4.2 Proposed Alternative GREG Estimators at the two stage sampling

Montanari (1987)'s estimator is expressed in term of variances and covariances in the estimation of $\boldsymbol{\beta}$. The ultimate cluster approach is a common method for variance estimation in complex survey designs which treat the PSU total estimates as response variables. The ultimate cluster approach was proposed by Hansen and Madow (1953). The ultimate cluster approach calculates the variance at PSU level within each stratum. This approach is valid for small sampling fraction.

We propose to adjust the GREG estimator from Estevao and Särndal (2006) by following Berger et al. (2003) approach and using the idea of the ultimate cluster approach to estimate $\boldsymbol{\beta}$ at PSU level. Berger et al. (2003) proposed an alternative regression estimator which is design - optimal. The idea of this alternative GREG estimator is to add the matrix of stratification variable into the GREG estimators. For this new GREG estimator we will incorporate the design variable into the GREG. The details show as follow.

Let $\breve{Q}_U$ be an $N \times H$ matrix that contain $\breve{q}_{hi} = q_{hi}/\pi_{hi}$, where q_{hi} is the stratification variable for stratum h, $(h = 1, 2, \ldots, H)$ and PSU $i, (i = 1, 2, \ldots, N_h)$ which is defined by

$$q_{hi} = \begin{cases} \pi_{hi} & \text{if PSU } i \text{ is in stratum h,} \\ 0 & \text{other;} \end{cases} \tag{6.34}$$

where $\pi_{hi} = n_h M_{hi} / \sum_{i=1}^{N_h} M_{hi}$ is the inclusion probability of PSU i within stratum, for example under probability proportional to size sampling, $h = 1, 2, \ldots, H, i = 1, 2, \ldots, n_h, j = 1, 2, \ldots, m_{hi}$. Therefore,

$$\breve{q}_{hi} = \begin{cases} 1 & \text{if PSU } i \text{ is in stratum h,} \\ 0 & \text{other.} \end{cases} \tag{6.35}$$

We propose to adjust the GREG estimators by adding the stratification variables into the GREG estimators. We consider that we have a stratified two-stage cluster sampling.

The estimator of $\boldsymbol{\beta}_{\Gamma}^{opt}$ is given by a PSU level

$$\widehat{\boldsymbol{\beta}}_{\Gamma}^{opt} = var(\widehat{\boldsymbol{t}}_{xzq\pi})^{-1} cor(\widehat{\boldsymbol{t}}_{xzq\pi}, \widehat{\boldsymbol{t}}_{y\pi}) = (\breve{\boldsymbol{\Gamma}}'_s \breve{\boldsymbol{C}}_s \breve{\boldsymbol{\Gamma}}_s)^{-} \breve{\boldsymbol{\Gamma}}'_s \breve{\boldsymbol{C}}_s \breve{\boldsymbol{Y}}_s, \tag{6.36}$$

where $\breve{\boldsymbol{\Gamma}}_s = [\breve{\boldsymbol{X}}_s, \breve{\boldsymbol{Z}}_s, \breve{\boldsymbol{Q}}_s]$ represent the partitioned $n \times (k+p+H)$ matrix of the estimates of PSU totals, $\breve{\boldsymbol{C}}_s = diag(\breve{c}_{hi}), h = 1, 2, \ldots, H, i = 1, 2, \ldots, n_h$ with $\breve{c}_{hi} = 1 - \pi_{hi}$, where

- $\breve{\boldsymbol{X}}_s = (\breve{\boldsymbol{x}}_{\boldsymbol{hi}})$ represent the $n \times k$ matrix of the estimates of PSU totals of the unit level variable, with $\breve{\boldsymbol{x}}_{hi} = \sum_{j=1}^{m_{hi}} \boldsymbol{x}_{hij}/\pi_{hj|i}$,
- $\breve{\boldsymbol{Z}}_{\boldsymbol{s}} = (\breve{\boldsymbol{z}}_{\boldsymbol{hi}})$ represent the $n \times p$ matrix of the estimates of PSU totals of the cluster level variable, with $\breve{\boldsymbol{z}}_{hi} = \boldsymbol{z}_{hi}$,
- $\breve{\boldsymbol{Q}}_{\boldsymbol{s}} = (\breve{\boldsymbol{q}}_{\boldsymbol{hi}})$ represent the $n \times H$ matrix of the estimates of PSU totals of the cluster level variable, with $\breve{\boldsymbol{q}}_{hi} = \boldsymbol{q}_{hi}$,
- $\breve{\boldsymbol{Y}}_s = (\breve{y}_{hi})$, represent the $n \times 1$ vector of the estimates of PSU totals of the study variable, with $\breve{y}_{hi} = \sum_{j=1}^{m_{hi}} y_{hij}/\pi_{hj|i}$, where $\pi_{hj|i} = m_{hi}/M_{hi}$.

The overall design weight for the j-th SSU in the i-th PSU in stratum h is given by $d_{hij} = d_{hi} d_{hj|i}$, where $h = 1, 2, \ldots, H, i = 1, 2, \ldots, n_h, j = 1, 2, \ldots, m_{hi}$. The quantities d_{hi} and d_{hij} are the design weights defined $d_{hi} = 1/\pi_{hi}$ and $d_{hj|i} = 1/\pi_{hj|i}$, $\pi_{hij} = \pi_{hi} \times \pi_{hj|i}$.

The proposed GREG estimator is

$$\widehat{Y}_{GREG}^{(3)} = \widehat{t}_{y\pi} + (\boldsymbol{t}_{xzq} - \widehat{\boldsymbol{t}}_{xzq\pi})' \widehat{\boldsymbol{\beta}}_{\Gamma}^{opt}, \tag{6.37}$$

where

$$\widehat{t}_{y\pi} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} y_{hij}/\pi_{hij} = \boldsymbol{Y}'_s \boldsymbol{\Sigma}_s \boldsymbol{1}_s, \tag{6.38}$$

$$\boldsymbol{Y}_s = (y_{111}, y_{112}, ..., y_{hn_h m_n}, .., y_{Hn_H m_n})', \tag{6.39}$$

$$\boldsymbol{\Sigma}_s = diag(d_{hij}), d_{hij} = d_{hi} d_{hj|i}, \tag{6.40}$$

$h = 1, 2, \ldots, H, i = 1, 2, \ldots, n_h, j = 1, 2, \ldots, m_{hi}$, where d_{hi} and d_{hij} are the design weights defined $d_{hi} = 1/\pi_{hi}$ and $d_{hj|i} = 1/\pi_{hj|i}$. The vector $\boldsymbol{1}_s$ is dimension $n \times m$ with all one units, $\widehat{\boldsymbol{t}}_{xzq\pi} = (\widehat{\boldsymbol{t}}'_{x\pi}, \widehat{\boldsymbol{t}}'_{z\pi}, \widehat{\boldsymbol{t}}'_{q\pi})'$, $\widehat{\boldsymbol{t}}_{x\pi}, \widehat{\boldsymbol{t}}_{z\pi}$ and $\widehat{\boldsymbol{t}}_{q\pi}$ are Horvitz-Thompson estimator given by

$$\widehat{\boldsymbol{t}}_{x\pi} = \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} x_{hij}/\pi_{hij}, \tag{6.41}$$

$$\widehat{\boldsymbol{t}}_{z\pi} = \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} z_{hij}/\pi_{hij} \tag{6.42}$$

with

$$z_{hij} = z_{hi}/M_{hi} \tag{6.43}$$

and $\widehat{\boldsymbol{t}}_{q\pi} = [\sum_{i=1}^{n_1} q_{1i}/\pi_{1i}, \sum_{i=1}^{n_2} q_{2i}/\pi_{2i}, ..., \sum_{i=1}^{n_H} q_{Hi}/\pi_{Hi}]' = (n_1, n_2..., n_H)'$ respectively.

The vector of population totals $\boldsymbol{t}_{xzq}$ is given by

$$\boldsymbol{t}_{xzq} = \begin{bmatrix} \sum_{i=1}^{N_1} \boldsymbol{q}_{1i} \\ \sum_{i=1}^{N_2} \boldsymbol{q}_{2i} \\ . \\ . \\ \sum_{i=1}^{N_H} \boldsymbol{q}_{Hi} \\ \sum_{h=1}^{H}\sum_{i=1}^{N_h} \boldsymbol{z}_{hi} \\ \sum_{h=1}^{H}\sum_{i=1}^{N_h}\sum_{j=1}^{M_{hi}} \boldsymbol{x}_{hij} \end{bmatrix}, \tag{6.44}$$

$\boldsymbol{t}_q = [\sum_{i=1}^{N_1} q_{1i}, \sum_{i=1}^{N_2} q_{2i}, ..., \sum_{i=1}^{N_H} q_{Hi}]' = [\sum_{i=1}^{N_1} \pi_{1i}, \sum_{i=1}^{N_2} \pi_{2i}, ..., \sum_{i=1}^{N_H} \pi_{Hi}]' = (n_1, n_2..., n_H)'$. Therefore, $\boldsymbol{t}_q = \widehat{\boldsymbol{t}}_{q\pi}$.

6.5 Simulation Study

We consider stratified two-stage cluster sampling with unequal size.

Let $H = 3$ stratum, $N_1 = 1000, N_2 = 3000, N_3 = 2000, N = N_1 + N_2 + N_3$ and $m = 10$ (We also consider bigger sizes of stratum, e.g. $H = 20$ and $H = 60$).

6.5.1 Simulation steps

Step 1 We generate the number of elements in each cluster for each stratum which is $M_{hi}, h = 1, 2, 3, i = 1, 2, ..., N_h$ for $H = 3$ strata N and N_{hi} clusters by using the following function (see Deville (1997), Berger (2005)),

$$M_{hi} = ((W_{hi} - Min)/Max) \times 6 + 10,$$

where $W_{hi} = (i/N_{hi})^{\alpha} + (1/\alpha), \alpha = 4$, Min $= 10$ and Max $= 15$.

The number of elements in each cluster (M_{hi}) will vary between Min and Max and also vary between 10 and 50, holding a smaller variation to begin with and increasing to a larger variation as we continue the calculations by sample.

We consider two type of stratifications. The first one is stratification by cluster size when the PSU of similar sizes are grouped in the same strata; that is, when the strata are homogeneous according to the PSU sizes. The second one is called random stratification when the PSU of different sizes are grouped in the same strata randomly.

Step 2 The value of y_{hij} are generated from the multilevel model shown as below.

$$y_{hij} = x'_{hij}\beta + z'_{hi}\gamma + \epsilon_{1hi} + \epsilon_{2hij}. \tag{6.45}$$

In order to generate y_{hij} we generate $x_{hij} \sim N(20, 1), z_{hi} \sim N(0, 1)$.

$$\epsilon_{1hi} \sim \begin{cases} N(-10, 1) & \text{if PSU } i \text{ is in stratum h=1,} \\ N(0, 1) & \text{if PSU } i \text{ is in stratum h=2,} \\ N(10, 1) & \text{if PSU } i \text{ is in stratum h=3} \end{cases} \tag{6.46}$$

and $\quad \epsilon_{2hij} \sim N(0, \sigma^2_{\epsilon_{2hij}})$, where

$$\sigma^2_{\epsilon_{2hij}} = \left[\frac{1-\rho}{\rho}\right] \sigma^2_{\epsilon_{1hi}},$$

$\rho = 0.1, 0.4$, and $\beta_0 = \gamma_0 = 0$ and vary β_1 and γ_1.

Step 3 We will calculate $\pi_{hi}, z_{hij} = z_{hi}/M_{hi}$ for each cluster i in each stratum h. For unequal size sampling,$\pi_{hij} = \pi_{hi}\pi_{hj|i}$, $\pi_{hi} = n_h M_{hi}/\sum_{i=1}^{N_h} M_{hi}, \pi_{hj|i} = m_{hi}/M_{hi}, m_{hi} = m$.

Step 4 We selected a sample of 5% for each stratum using non-probability sampling.

Step 5 We compute the HT estimator, the naive GREG and the Estevao and Särndal GREG at individual level and the proposed optimal GREG. We compare their relative root mean squared errors and relative biases.

Table 6.1: Sample relative bias and relative root mean squared error in percentage for HT, classical GREG with x, GREG with z, GREG with x and z at individual level with β at individual level and with β at PSU level, optimal GREG with x and q, optimal GREG with z and q, and optimal GREG with x, z and q estimators at individual level with β at PSU level for PPS sampling. The population size $N_1 = 3,000, N_2 = 2,000, N_3 = 1,000, H = 3$. The intra-cluster correlation ρ is equal to 0.1 and 0.4. The number of SSU in each stratum vary between 10 and 50. Sample 5% from each PSU, $m = 10$ and repeat 1,000 times.

				Relative bias							**Relative root mean squared error**						
				HT	**Estevao and Särndal**			**Optimal GREG**			**HT**	**Estevao and Särndal**			**Optimal GREG**		
					X	Z	XZ	XQ	ZQ	XZQ		X	Z	XZ	XQ	ZQ	XZQ
$\rho = 0.1$	**Stratification by cluster size**	r_{yx}	r_{yz}														
		0.2	0.2	0.36	0.35	0.35	0.25	0.35	0.25	**0.24**	0.46	0.44	0.47	0.32	0.45	0.32	**0.30**
		0.2	0.7	1.06	1.05	0.66	0.58	1.06	0.52	**0.51**	1.33	1.32	0.85	0.74	1.33	0.66	**0.65**
		0.7	0.2	0.12	0.08	0.27	0.06	0.08	0.10	**0.05**	0.15	0.11	0.41	0.07	0.11	0.12	**0.06**
		0.7	0.7	0.31	0.30	0.33	0.16	0.30	0.17	**0.14**	0.39	0.37	0.47	0.21	0.37	0.21	**0.18**
	Random stratification	r_{yx}	r_{yz}														
		0.2	0.2	0.36	0.34	0.46	**0.24**	0.34	0.30	0.29	0.45	0.43	0.66	**0.30**	0.43	0.38	0.36
		0.2	0.7	0.99	0.98	0.62	**0.46**	0.99	0.76	0.75	1.23	1.22	0.83	**0.57**	1.23	0.95	0.95
		0.7	0.2	0.12	0.09	0.37	**0.05**	0.09	0.10	0.07	0.15	0.11	0.60	**0.06**	0.11	0.13	0.09
		0.7	0.7	0.28	0.26	0.40	**0.12**	0.26	0.21	0.20	0.34	0.32	0.61	**0.15**	0.32	0.26	0.25
$\rho = 0.4$	**Stratification by cluster size**	r_{yx}	r_{yz}														
		0.2	0.2	0.34	0.33	0.33	0.22	0.33	0.22	**0.20**	0.43	0.42	0.45	0.28	0.42	0.28	**0.26**
		0.2	0.7	1.06	1.05	0.66	0.58	1.06	0.52	**0.51**	1.33	1.32	0.85	0.74	1.33	0.66	**0.65**
		0.7	0.2	0.12	0.08	0.27	0.05	0.08	0.09	**0.05**	0.15	0.10	0.41	0.06	0.10	0.12	**0.06**
		0.7	0.7	0.31	0.30	0.33	0.16	0.30	0.17	**0.14**	0.39	0.37	0.47	0.21	0.37	0.21	**0.18**
	Random stratification	r_{yx}	r_{yz}														
		0.2	0.2	0.33	0.32	0.45	**0.21**	0.32	0.28	0.26	0.42	0.40	0.64	**0.26**	0.40	0.34	0.32
		0.2	0.7	0.99	0.98	0.62	**0.45**	0.98	0.75	0.75	1.23	1.22	0.82	**0.56**	1.22	0.94	0.94
		0.7	0.2	0.12	0.08	0.37	**0.05**	0.08	0.10	0.07	0.15	0.10	0.60	**0.06**	0.10	0.13	0.08
		0.7	0.7	0.28	0.26	0.40	**0.12**	0.26	0.21	0.20	0.34	0.32	0.61	**0.15**	0.32	0.26	0.25

Table 6.1 presents the results of the relative bias and relative root mean squared error. There are 3 strata and the number of SSU in each stratum vary between 10 and 50. For $\rho = 0.1$, under stratification by cluster size, we see that the optimal GREG with x, z and q has a minimum relative bias compared to the other GREG estimators and the Horvitz-Thomson estimator. In this situations the optimal GREG estimator with z variable alone also performs better than the Horvitz-Thomson estimator in all situations and performs better than the Estevao and Särndal GREG estimators. However, the Estevao and Särndal GREG estimator with x and z performs slightly better when there is a large correlation between y and x, $r_{yx} = 0.7$ and a small correlation between y and z, $r_{yz} = 0.2$ and also when there is a large correlation between both y and x and y and z. The optimal GREG estimator with x only gives similar results to those found using the Estevao and Särndal GREG estimator with variable x only and with x and z. There is a higher relative bias when the correlation at the individual level between y and z is equal to 0.7 compare to other situations.

The Estevao and Särndal GREG estimator with x and z performs better than all other Estevao and Särndal GREG estimators. The Estevao and Särndal GREG estimator based upon z variable performs better than the Estevao and Särndal GREG estimator with the x variable only when the correlation between y and z is large ($r_{yz} = 0.7$). Nevertheless, the optimal GREG estimator with z variable only performs as well as or better than all the optimal GREG estimators with x variable only with the same or higher correlation between y and x respectively except when there is a large correlation between y and x, $r_{yx} = 0.7$ and a small correlation between y and z, $r_{yz} = 0.2$. Most of the Estevao and Särndal GREG estimators perform better than the Horvitz-Thomson estimator except when there is a large correlation between y and z. However, all of the optimal GREG estimators perform better or similarly than the Horvitz-Thomson estimator.

Similar patterns are shown in the relative root mean squared error output. We see that the optimal GREG with x, z and q variables has a smaller relative root mean squared error than the Horvitz-Thomson estimator and the Estevao and Särndal GREG estimators. It seems that the optimal GREG is acurate in this scenario.

However, for random stratification we notice that the Estevao and Särndal GREG estimator with x and z variables has both minimum relative bias and relative root mean squared error. The optimal GREG estimator with x, z and q has a similar (lower) mean squared error with the same level of correlation in place between y and z. The optimal GREG estimator with x and q variables give similar results to the Estevao and Särndal GREG estimator with x only. Random stratification is not suited to the optimal GREG variable because there is not much difference between the strata.

We have similar pattern with an intra-cluster correlation $\rho = 0.4$. Under stratification by cluster size we see that, the optimal GREG with variables x, z and q has a minimum relative bias and relative root mean squared error when compared to other GREG estimators and the Horvitz-Thomson estimator. For a lower intra-cluster correlation, the optimal GREG estimator with z variable alone performs better than the Horvitz-Thomson estimator in all situations. It also performs better than the Estevao and Särndal GREG estimators unless $r_{yx} = 0.7$ and $r_{yz} = 0.2$ and when $r_{yx} = 0.7$ and $r_{yz} = 0.7$. The optimal GREG estimator with x only gives similar results to those found using the Estevao and Särndal GREG estimator with variable x only and with x and z.

If we compare the GREG estimators, the Estevao and Särndal GREG estimators with x only and the optimal GREG estimators with x and q respectively, we see that they both give similar results in terms of relative bias and relative root mean squared error. The optimal GREG with z and q performs better than the Estevao and Särndal GREG estimator with z only for all the cases. Moreover, we notice that the optimal GREG with x, z and q variables has a smaller relative bias and relative root mean squared error than the Estevao and Särndal GREG estimator with x and z.

Moreover, for random stratification, we observe similar pattern to the situation where $\rho = 0.1$. We see that the Estevao and Särndal GREG estimator with x and z variables has a both minimum relative bias and relative root mean squared error. Under stratification by cluster, the Estevao and Särndal GREG estimators with x only and the optimal GREG estimators with x and q respectively, both give similar results in term of relative bias and relative root mean squared error. Moreover, the Estevao and Särndal GREG estimator with z only performs better than the optimal GREG with z and q only when $r_{yx} = 0.2$ and $r_{yz} = 0.7$.

The Horvitz-Thomson estimator performs poorly when compared to other GREG estimators in both situations as shown under cluster size random stratification. It seems that there is no difference in the results when intra-cluster correlation increases.

Table 6.2 gives the results of the relative bias and relative root mean squared error. There are 3 strata. We see similar patterns to those shown in Table 6.1 but slightly different outcomes in some situations. For $\rho = 0.1$, under stratification by cluster size, we see that the GREG with variables x, z and q has a minimum relative bias when compared to the other GREG estimators including the Horvitz-Thomson estimator. In this situation the optimal GREG estimator with variable x and z when $r_{yx} = 0.7$ and

Table 6.2: Sample relative bias and relative root mean squared error in percentage for HT, classical GREG with x, GREG with z, GREG with x and z at individual level with β at individual level and with β at PSU level, optimal GREG with x and q, optimal GREG with z and q, and optimal GREG with x, z and q estimators at individual level with β at PSU level for PPS sampling. The population size $N_1 = 3,000, N_2 = 2,000, N_3 = 1,000, H = 3$. The intra-cluster correlation ρ is equal to 0.1 and 0.4. The number of SSU in each stratum vary between 10 and 30. Sample 5% from each stratum, $m = 10$ and repeat 1,000 times.

				Relative bias							**Relative root mean squared error**						
				HT	**Estevao and Särndal**			**Optimal GREG**			**HT**	**Estevao and Särndal**			**Optimal GREG**		
					X	Z	XZ	XQ	ZQ	XZQ		X	Z	XZ	XQ	ZQ	XZQ
$\rho = 0.1$	**Stratification by cluster size**	r_{yx}	r_{yz}														
		0.2	0.2	0.35	0.34	0.36	0.22	0.34	0.22	**0.21**	0.44	0.43	0.51	0.28	0.43	0.28	**0.27**
		0.2	0.7	1.00	1.00	0.51	0.38	1.00	0.37	**0.36**	1.26	1.26	0.69	0.48	1.26	0.46	**0.45**
		0.7	0.2	0.11	0.08	0.31	**0.04**	0.08	0.09	**0.04**	0.14	0.10	0.48	**0.05**	0.10	0.11	**0.05**
		0.7	0.7	0.29	0.28	0.33	0.15	0.28	0.16	**0.13**	0.37	0.35	0.47	0.19	0.35	0.20	**0.17**
	Random stratification	r_{yx}	r_{yz}														
		0.2	0.2	0.34	0.33	0.44	**0.22**	0.33	0.23	**0.22**	0.43	0.41	0.63	**0.28**	0.41	0.29	**0.28**
		0.2	0.7	0.97	0.97	0.51	**0.32**	0.97	0.40	0.40	1.22	1.21	0.69	**0.40**	1.22	0.50	0.50
		0.7	0.2	0.11	0.08	0.36	**0.04**	0.08	0.09	0.05	0.14	0.10	0.57	**0.06**	0.10	0.11	0.06
		0.7	0.7	0.26	0.24	0.37	**0.07**	0.34	0.12	0.10	0.32	0.30	0.58	**0.10**	0.30	0.15	0.12
$\rho = 0.4$	**Stratification by cluster size**	r_{yx}	r_{yz}														
		0.2	0.2	0.33	0.32	0.34	0.18	0.32	0.19	**0.18**	0.42	0.41	0.49	0.23	0.41	0.24	**0.22**
		0.2	0.7	1.00	1.00	0.50	0.37	1.00	0.36	**0.35**	1.26	1.26	0.68	0.47	1.26	0.45	**0.44**
		0.7	0.2	0.11	0.08	0.30	**0.04**	0.08	0.09	**0.04**	0.14	0.10	0.48	**0.05**	0.10	0.11	**0.05**
		0.7	0.7	0.27	0.26	0.33	**0.09**	0.26	0.12	**0.09**	0.34	0.33	0.51	0.12	0.33	0.15	**0.11**
	Random stratification	r_{yx}	r_{yz}														
		0.2	0.2	0.32	0.31	0.42	**0.18**	0.31	0.20	0.19	0.41	0.39	0.62	**0.23**	0.39	0.25	**0.23**
		0.2	0.7	0.97	0.97	0.50	**0.31**	0.97	0.39	0.39	1.22	1.21	0.69	**0.39**	1.22	0.50	0.49
		0.7	0.2	0.11	0.08	0.36	**0.04**	0.08	0.08	**0.04**	0.14	0.10	0.57	**0.05**	0.10	0.11	**0.05**
		0.7	0.7	0.26	0.24	0.37	**0.07**	0.24	0.12	0.10	0.32	0.30	0.58	**0.10**	0.30	0.15	0.12

$r_{yz} = 0.2$ also performs as well as the optimal GREG estimator with variable x, z and q.

The optimal GREG estimator with the z variable alone also performs better than the Estevao and Särndal GREG estimators and the Horvitz-Thomson estimator unless the correlation between y and x is equal to 0.7 and the correlation between y and z is equal to 0.2. The optimal GREG estimator with x only gives similar results to those found using the Estevao and Särndal GREG estimator but there is a small increase in relative bias when the correlation at the individual level between y and x increases.

If we compare the same GREG estimators, the Estevao and Särndal GREG estimators and the optimal GREG estimators with x and z respectively, we notice that the Estevao and Särndal GREG estimator upon z variable only performs better than the Estevao and Särndal GREG estimator with the x variable only when $r_{yx} = 0.2$ and $r_{yz} = 0.7$. Similarly, the optimal GREG estimator with z variable only performs better than the optimal GREG estimator with the x variable only when $r_{yx} = 0.2$ and $r_{yz} = 0.7$.

Compared to the Horvitz-Thomson estimator, the Estevao and Särndal GREG estimator with the variable x only and with variable x and z performs better or at least the same as the Horvitz-Thomson estimator and the Estevao and Särndal GREG estimator upon z variable only. Similar patterns are shown for the optimal GREG estimators, the optimal GREG estimator with variable x only and with variable x and z performs better or at least the same as the Horvitz-Thomson estimator. The optimal GREG estimator upon z variable only performs better than the Horvitz-Thomson estimator when it has a small correlation between y and x and when it has a large correlation between y and z.

When we consider the relative root mean squared error output, we see that the optimal GREG with x, z and q variables has a smaller relative root mean squared error output to the Horvitz-Thomson estimator and the Estevao and Särndal GREG estimators produce except when $r_{yx} = 0.7$ and $r_{yz} = 0.2$. In this case, the Estevao and Särndal GREG with variable x and z also gives the same results. It seems that the optimal GREG is accurate in these scenarios.

On the other hand, for random stratification, we notice that the Estevao and Särndal GREG estimator with x and z variables has a both minimum relative bias and relative root mean squared error but a slightly different output to that of the optimal GREG estimator with x, z and q variables with the same level of correlation. Interestingly, the optimal GREG with variable x, z and q produces the same results as it has shown in the Estevao and Särndal one with small correlation. Including q variable into the optimal

GREG estimator with z only seems to reduce high relative bias for the optimal GREG with z and q. The optimal GREG estimators still give a better result or at least the same results compared to the Horvitz-Thomson estimator.

Similar patterns are shown for intra-cluster correlation $\rho = 0.4$. Under stratification by cluster size, we see that the GREG with x, z and q has a minimum relative bias compared to the other GREG estimators including the Horvitz-Thomson estimator except where the correlation at the individual level variables between y and x is equal to 0.7 where the Estevao and Särndal GREG estimator with variable x and z produces the same result.

The optimal GREG estimator with z variable alone also performs better than the Estevao and Särndal GREG estimators and the Horvitz-Thomson estimator unless the correlation at the individual level variables between y and x is equal to 0.7 and the correlation between y and z is equal to 0.2. In this case, the Estevao and Särndal GREG estimator only upon x variable performs slightly better. When the correlation is 0.7 the Estevao and Särndal GREG estimator with variable x and z performs slightly better. The optimal GREG estimator with x only gives similar results to those found using the Estevao and Särndal GREG estimator.

Moreover, for random stratification, we also see the similar results to the one with $\rho = 0.1$. We see that the Estevao and Särndal GREG estimator with x and z variables has a small relative bias and relative root mean squared error. Surprisingly, the optimal GREG with variable x, z and q (with a large correlation between y and x) also produces the same relative bias and relative root mean sqaured error as it has shown in the Estevao and Särndal with variable x and z. The relative bias decreases using the optimal GREG with z and q. The optimal GREG estimators still give a better result or at least the same results as the Horvitz-Thomson estimator.

Table 6.3 presents the results of the sample relative bias and relative root mean squared error. There are 3 strata and the number of SSU in each stratum varies between 10 and 15. Interestingly we see different results to those found in Table 6.2, for $\rho = 0.1$ under stratification by cluster size. We see that the GREG with x, z and q has a minimum relative bias. In this situation the optimal GREG estimator with variable z and q when $r_{yx} = 0.2$ and $r_{yz} = 0$ also performs as well as the optimal GREG estimator with variable x, z and q.

The optimal GREG estimator with z variable alone also performs better than the Estevao and Särndal GREG estimator with variable x alone and with variable z alone in

Table 6.3: Sample total, variance, relative root mean squared error and relative bias in percentage for HT, classical GREG with x, GREG with z, GREG with x and z at individual level with β at individual level and with β at PSU level, optimal GREG with x and q, optimal GREG with z and q, and optimal GREG with x, z and q estimators at individual level with β at PSU level for PPS sampling. The population size $N_1 = 3,000, N_2 = 2,000, N_3 = 1,000, H = 3$. The intra-cluster correlation ρ is equal to 0.1 and 0.4. The number of SSU in each stratum vary between 10 and 15. Sample 5% from each stratum, $m = 10$ and repeat 1,000 times.

				Relative bias							**Relative root mean squared error**						
				HT	**Estevao and Särndal**			**Optimal GREG**			**HT**	**Estevao and Särndal**			**Optimal GREG**		
					X	Z	XZ	XQ	ZQ	XZQ		X	Z	XZ	XQ	ZQ	XZQ
$\rho = 0.1$	**Stratification by cluster size**	r_{yx}	r_{yz}														
		0.2	0.2	0.34	0.33	0.35	0.24	0.33	0.25	**0.23**	0.43	0.41	0.47	0.30	0.41	0.31	**0.29**
		0.2	0.7	1.00	0.99	0.66	0.58	1.00	**0.52**	**0.52**	1.24	1.24	0.83	0.73	1.24	0.66	**0.65**
		0.7	0.2	0.12	0.08	0.25	0.06	0.08	0.10	**0.05**	0.15	0.10	0.40	**0.07**	0.10	0.12	**0.07**
		0.7	0.7	0.28	0.27	0.31	0.16	0.27	0.16	**0.14**	0.35	0.34	0.46	0.20	0.34	0.20	**0.18**
	Random stratification	r_{yx}	r_{yz}														
		0.2	0.2	0.33	0.32	0.43	0.21	0.33	0.26	**0.25**	0.42	0.41	0.63	0.27	0.41	0.32	**0.31**
		0.2	0.7	0.82	0.82	0.39	**0.14**	0.82	0.16	**0.14**	1.03	1.02	0.60	0.18	1.03	0.20	**0.17**
		0.7	0.2	0.11	0.09	0.35	0.04	0.09	0.08	**0.03**	0.14	0.11	0.57	0.05	0.11	0.10	**0.04**
		0.7	0.7	0.26	0.24	0.35	**0.03**	0.24	0.08	**0.03**	0.32	0.30	0.57	**0.04**	0.31	0.10	**0.04**
$\rho = 0.4$	**Stratification by cluster size**	r_{yx}	r_{yz}														
		0.2	0.2	0.33	0.32	0.37	0.15	0.32	0.16	**0.14**	0.41	0.40	0.54	0.19	0.40	0.21	**0.18**
		0.2	0.7	0.80	0.80	0.36	**0.13**	0.80	0.14	**0.13**	1.01	1.00	0.54	**0.16**	1.00	0.18	**0.16**
		0.7	0.2	0.11	0.08	0.33	**0.03**	0.08	0.08	**0.03**	0.14	0.10	0.52	**0.04**	0.10	0.10	**0.04**
		0.7	0.7	0.25	0.24	0.33	**0.03**	0.24	0.08	**0.03**	0.32	0.31	0.53	**0.04**	0.30	0.10	**0.04**
	Random stratification	r_{yx}	r_{yz}														
		0.2	0.2	0.33	0.32	0.40	0.16	0.32	0.17	**0.15**	0.41	0.60	0.20	0.24	0.40	0.21	**0.18**
		0.2	0.7	0.82	0.81	0.38	0.13	0.82	0.14	**0.12**	1.02	1.02	0.59	**0.16**	1.02	0.18	**0.16**
		0.7	0.2	0.11	0.08	0.35	**0.03**	0.08	0.08	**0.03**	0.14	0.10	0.57	**0.04**	0.10	0.10	**0.03**
		0.7	0.7	0.26	0.24	0.35	**0.03**	0.24	0.08	**0.03**	0.32	0.30	0.57	**0.04**	0.31	0.10	**0.04**

most situations. The Estevao and Särndal GREG estimator with variable x and q performs similarly to the optimal GREG estimator with variable with z and q. Similarly the other optimal GREG estimators with x only gives similar results to the Estevao and Särndal GREG estimator.

If we compare the Estevao and Särndal GREG estimators and the optimal GREG estimators with x and z respectively, we notice that the Estevao and Särndal GREG estimator with x performs better than the Estevao and Särndal GREG estimator with the z variable only in almost all situations, except when $r_{yx} = 0.7$ and $r_{yz} = 0.2$. Nevertheless, the optimal GREG estimator with z variable performs better than the optimal GREG estimator with the x variable only in almost all situations except when $r_{yx} = 0.7$ and $r_{yz} = 0.2$. Compared to the Horvitz-Thomson estimator, the optimal GREG estimators perform better than the Horvitz-Thomson estimator in all situations. However, the Estevao and Särndal GREG estimator with z only performs slightly worse than the Horvitz-Thomson estimator when it has a lower correlation between y and x and between y and z.

Under random stratification, we notice that the optimal GREG estimator with variable x, z and q still performs better than other GREG variables including the Horvitz-Thomson estimator. It is as accurate as the Estevao and Särndal GREG estimator with x and z with large correlation between y and z. In this situation, it produces different results than those shown in Tables 6.1 and 6.2.

For intra-cluster correlation $\rho = 0.4$ with stratification by cluster size, the Estevao and Särndal GREG estimator with x and z works well as it produces at least the same results as the optimal GREG with x, z and q, in terms of the relative bias and the relative root mean squared error.

However, for random stratification, we see similar results to the one with $\rho = 0.1$. In this situation, the optimal GREG with x, z and q performs well in all situations although the Estevao and Särndal GREG estimator with x and z variables produces the same results, except with a small correlation between y and x for the relative bias and with a large correlation between y and x and between y and z, for the relative root mean squared error.

The Horvitz-Thomson estimator performs poorly or at best the same as the optimal GREG estimators. However, it performs better than the Estevao and Särndal GREG estimator with z alone in some situation, such as with a small correlation between y and x and y and z. The small number of SSU in each stratum affects the results in this situation.

Table 6.4 presents the results of the relative bias and relative root mean squared error. The population size is equal to 300 for 20 strata. Considering a large stratum sizes, we still see similar patterns to those results for the smaller stratum, for $\rho = 0.1$, under stratification by cluster size we see that, the GREG with variables x, z and q has a minimum relative bias when compared to the other GREG estimators including the Horvitz-Thomson estimator. Nevertheless, the Estevao and Särndal GREG estimator with x and z (with a large correlation between y and x and a small correlation between y and z) gives the same result as the optimal GREG with x, z and q.

We observe similar patterns to Table 6.1 and Table 6.2 with random stratification. We see that the optimal GREG estimator with x, z and q still performs better than the other GREG estimators including the Horvitz-Thomson estimator in terms of both minimum relative bias and relative root mean squared error. However, we observe slightly different results than those of the optimal GREG estimator with x, z and q variables with the same level of correlation.

When $\rho = 0.4$, with stratification by cluster size, we see that the optimal GREG with x, z and q performs well in terms of both relative bias and relative root mean squared error. The optimal GREG with z and q also produces the same results to those with x variable (with a small correlation between y and x and a large correlation between y and z). Moreover, the Estevao and Särndal GREG estimator with x and z also works well in some situations ($r_{yx} = 0.7$ and $r_{yz} = 0.2$). It produces at least the same results as those of the optimal GREG with x, z and q in terms of the relative bias and the relative root mean squared error.

However, under random stratification, we also see similar results to those with $\rho = 0.1$. We see that the Estevao and Särndal GREG estimator with x and z also performs well in terms of relative bias and relative root mean squared error even though the optimal GREG estimator with x, z and q produces the same results with a small correlation between y and x and between y and x.

The Horvitz-Thomson estimator performs poorly or at best the same as optimal GREG estimators but performs better than the Estevao and Särndal GREG estimator with variable z alone in some situations, such as with a large correlation between y and x and a small correlation between y and z.

Table 6.4: Sample relative bias and relative root mean squared error in percentage for HT, classical GREG with x, GREG with z, GREG with x and z at individual level with β at individual level and with β at PSU level, optimal GREG with x and q, optimal GREG with z and q, and optimal GREG with x, z and q estimators at individual level with β at PSU level for PPS sampling. The population size is equal to 300 for 20 stratums. The intra-cluster correlation ρ is equal to 0.1. The number of SSU in each stratum vary between 10 and 50. Sample 5% from each stratum, $m = 10$ and repeat 1,000 times.

				Relative bias							**Relative root mean squared error**						
				HT	**Estevao and Särndal**			**Optimal GREG**			**HT**	**Estevao and Särndal**			**Optimal GREG**		
					X	Z	XZ	XQ	ZQ	XZQ		X	Z	XZ	XQ	ZQ	XZQ
$\rho = 0.1$	**Stratification by cluster size**	r_{yx}	r_{yz}														
		0.2	0.2	0.35	0.34	0.36	0.22	0.34	0.22	**0.21**	0.44	0.43	0.51	0.28	0.43	0.28	**0.27**
		0.2	0.7	1.00	1.00	0.51	0.38	1.00	0.37	**0.36**	1.26	1.26	0.6 9	0.48	1.26	0.46	**0.45**
		0.7	0.2	0.11	0.08	0.31	**0.04**	0.08	0.09	**0.04**	0.14	0.10	0.48	**0.05**	0.10	0.11	**0.05**
		0.7	0.7	0.29	0.28	0.33	0.15	0.28	0.16	**0.13**	0.37	0.35	0.47	0.19	0.35	0.20	**0.17**
	Random stratification	r_{yx}	r_{yz}														
		0.2	0.2	0.33	0.32	0.43	**0.21**	0.33	0.26	0.25	0.42	0.41	0.63	**0.27**	0.41	0.32	0.31
		0.2	0.7	0.92	0.92	0.59	**0.43**	0.93	0.66	0.66	1.17	1.17	0.79	**0.54**	1.17	0.82	0.82
		0.7	0.2	0.12	0.09	0.37	**0.05**	0.09	0.10	0.07	0.15	0.11	0.60	**0.06**	0.11	0.13	0.09
		0.7	0.7	0.25	0.24	0.38	**0.11**	0.24	0.18	0.17	0.32	0.30	0.59	**0.14**	0.30	0.23	0.21
$\rho = 0.4$	**Stratification by cluster size**	r_{yx}	r_{yz}														
		0.2	0.2	0.32	0.31	0.32	0.21	0.31	0.22	**0.20**	0.40	0.38	0.44	0.27	0.39	0.27	**0.25**
		0.2	0.7	1.00	0.99	0.65	0.58	0.99	0.52	**0.51**	1.24	1.24	0.83	0.72	1.24	**0.65**	**0.65**
		0.7	0.2	0.12	0.08	0.25	**0.05**	0.08	0.10	**0.05**	0.14	0.10	0.40	**0.06**	0.10	0.12	**0.06**
		0.7	0.7	0.28	0.27	0.31	**0.16**	0.27	0.16	**0.14**	0.35	0.34	0.46	0.20	0.34	0.20	**0.18**
	Random stratification	r_{yx}	r_{yz}														
		0.2	0.2	0.31	0.30	0.41	**0.18**	0.30	0.23	0.22	0.39	0.38	0.62	**0.23**	0.38	0.29	**0.28**
		0.2	0.7	0.92	0.92	0.59	**0.42**	0.92	0.66	0.66	1.17	1.16	0.79	**0.53**	1.17	0.82	0.81
		0.7	0.2	0.11	0.08	0.35	**0.04**	0.08	0.09	0.06	0.13	0.10	0.58	**0.05**	0.10	0.11	0.07
		0.7	0.7	0.25	0.24	0.38	**0.11**	0.24	0.18	0.17	0.32	0.30	0.59	**0.14**	0.30	0.23	0.21

Table 6.5: Sample relative bias and relative root mean squared error in percentage for HT, classical GREG with x, GREG with z, GREG with x and z at individual level with β at individual level and with β at PSU level, optimal GREG with x and q, optimal GREG with z and q, and optimal GREG with x, z and q estimators at individual level with β at PSU level for PPS sampling. The population size is equal to 100 for 60 stratums. The intra-cluster correlation ρ is equal to 0.1. The number of SSU in each stratum vary between 10 and 50. Sample 5% from each stratum, $m = 10$ and repeat 1,000 times.

				Relative bias							**Relative root mean squared error**						
				HT	**Estevao and Särndal**			**Optimal GREG**			**HT**	**Estevao and Särndal**			**Optimal GREG**		
					X	Z	XZ	XQ	ZQ	XZQ		X	Z	XZ	XQ	ZQ	XZQ
$\rho = 0.1$	**Stratification by cluster size**	r_{yx}	r_{yz}														
		0.2	0.2	0.28	0.27	0.28	0.18	0.28	0.26	**0.19**	0.35	0.34	0.38	0.23	0.36	0.32	**0.24**
		0.2	0.7	0.94	0.94	0.61	0.55	0.97	0.52	**0.50**	1.19	1.18	0.78	0.69	1.22	0.65	**0.63**
		0.7	0.2	0.11	0.07	0.25	**0.04**	0.07	0.17	**0.04**	0.14	0.09	0.37	0.06	0.09	0.20	**0.05**
		0.7	0.7	0.29	0.28	0.31	0.16	0.29	0.21	**0.15**	0.37	0.35	0.44	0.20	0.36	0.26	**0.18**
	Random stratification	r_{yx}	r_{yz}														
		0.2	0.2	0.27	0.26	0.41	**0.18**	0.28	0.32	0.22	0.34	0.33	0.62	**0.22**	0.35	0.38	0.28
		0.2	0.7	0.88	0.88	0.58	**0.41**	0.93	0.64	0.62	1.10	1.10	0.78	**0.51**	1.16	0.80	0.77
		0.7	0.2	0.09	0.06	0.35	**0.03**	0.07	0.19	0.05	0.12	0.08	0.58	**0.04**	0.08	0.22	0.06
		0.7	0.7	0.25	0.24	0.38	**0.11**	0.26	0.23	0.17	0.32	0.30	0.60	**0.14**	0.32	0.29	0.21
$\rho = 0.4$	**Stratification by cluster size**	r_{yx}	r_{yz}														
		0.2	0.2	0.27	0.25	0.27	**0.17**	0.27	0.25	**0.17**	0.34	0.32	0.36	**0.21**	0.34	0.31	0.22
		0.2	0.7	0.94	0.94	0.61	0.55	0.97	0.52	**0.50**	1.18	1.18	0.78	0.69	1.21	0.65	**0.63**
		0.7	0.2	0.11	0.07	0.24	**0.04**	0.07	0.17	**0.04**	0.14	0.09	0.37	**0.05**	0.09	0.20	**0.05**
		0.7	0.7	0.29	0.28	0.31	0.16	0.29	0.21	**0.15**	0.37	0.35	0.44	0.20	0.36	0.26	**0.18**
	Random stratification	r_{yx}	r_{yz}														
		0.2	0.2	0.26	0.25	0.40	**0.16**	0.27	0.31	0.21	0.33	0.32	0.62	**0.20**	0.34	0.37	0.26
		0.2	0.7	0.89	0.88	0.58	**0.40**	0.93	0.64	0.61	1.10	1.10	0.78	**0.51**	1.16	0.7 9	0.77
		0.7	0.2	0.09	0.06	0.35	**0.03**	0.06	0.19	0.05	0.12	0.08	0.58	**0.04**	0.08	0.22	0.06
		0.7	0.7	0.25	0.24	0.38	**0.11**	0.26	0.23	0.17	0.32	0.30	0.60	**0.14**	0.32	0.29	0.21

Table 6.5 presents the results of the relative bias and relative root mean squared error. The population size is equal to 100 for 60 strata. Considering 60 strata, we see similar patterns to those results for the smaller stratum, for $\rho = 0.1$. Under stratification by cluster size, we see that the GREG with x, z and q has a minimum relative bias when compared to the other GREG estimators including the Horvitz-Thomson estimator. Nevertheless, the Estevao and Särndal GREG estimator with x and z, a large correlation between y and x and a small correlation between y and z also gives the same result to the optimal GREG with variable x, z and q in terms of minimum relative bias.

Similar patterns are observed on Table 6.4 under random stratification. We see that the optimal GREG estimator with x, z and q performs better than the other optimal GREG estimators including the Horvitz-Thomson estimator in terms of relative bias and relative root mean squared error. However, we observe a slightly different result to that of the optimal GREG estimator with x, z and q and with the same level of correlation between y and z.

When intra-cluster correlation $\rho = 0.4$, and under stratification by cluster size, we see that the optimal GREG with x, z and q performs well in terms of relative bias and relative root mean squared error. Moreover, the Estevao and Särndal GREG estimator with x and z works well in some situations, $r_{yz} = 0.2$, we observe the same results as the optimal GREG with x, z and q in terms of the relative bias and the relative root mean squared error.

Under random stratification, we also see similar results to those with $\rho = 0.1$. We see that in this situation the Estevao and Särndal GREG estimator with x and z variables performs well in terms of both relative bias and relative root mean squared error. The Horvitz-Thomson estimator performs poorly or at best the same as the Estevao and Särndal GREG estimators. However, the Horvitz-Thomson estimator performs better than the optimal GREG estimator with variable x alone when $r_{yx} = 0.2$ and $r_{yz} = 0.7$. It seems that the sizes of the strata do not affect the results.

6.5.2 Conditional Bias

We propose to investigate the conditional bias of the estimators considered in the previous section. First of all, we calculate the relative error for the population total for each estimators from the 1,000 samples in the simulation study. Then we ordered them by their total mean and classified them into 20 groups with 50 sampled each following Chambers and Dunstan (1986). Finally, we calculate the mean of their overall bias for each estimators and total mean of variable x and z. We chose to study some cases of the Table 6.2: stratification by cluster size and random cluster stratification, when $r_{yx} = 0.2$ and $r_{yz} = 0.7$ and $r_{yx} = 0.7$ and $r_{yz} = 0.2$. The population size are $N_1 = 3,000, N_2 = 2,000, N_3 = 1,000$ and there are 3 strata. The results are represented in the graphs below.

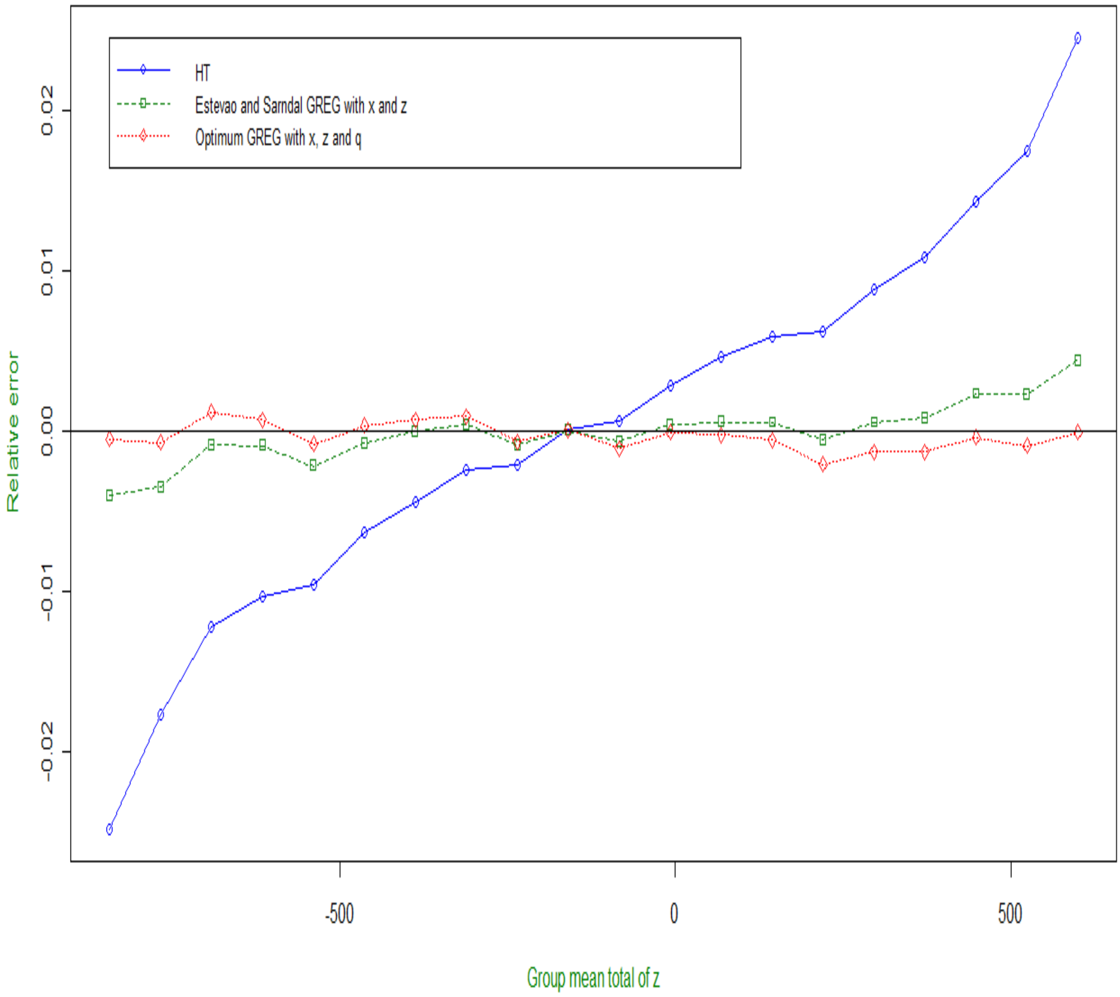

Figure 6.1: Relative bias for HT, Estevao and Särndal with x and z and optimal GREG estimators against the group mean total of z for stratification by cluster size when $r_{yx} = 0.2$ and $r_{yz} = 0.7$.

Figure 6.1 shows that the optimal GREG with x, z and q performs well compared to the Estevao and Särndal GREG and the Horvitz-Thomson GREG estimators. The Horvitz-Thomson GREG estimator performs the worse as it shows a linear trend of z.

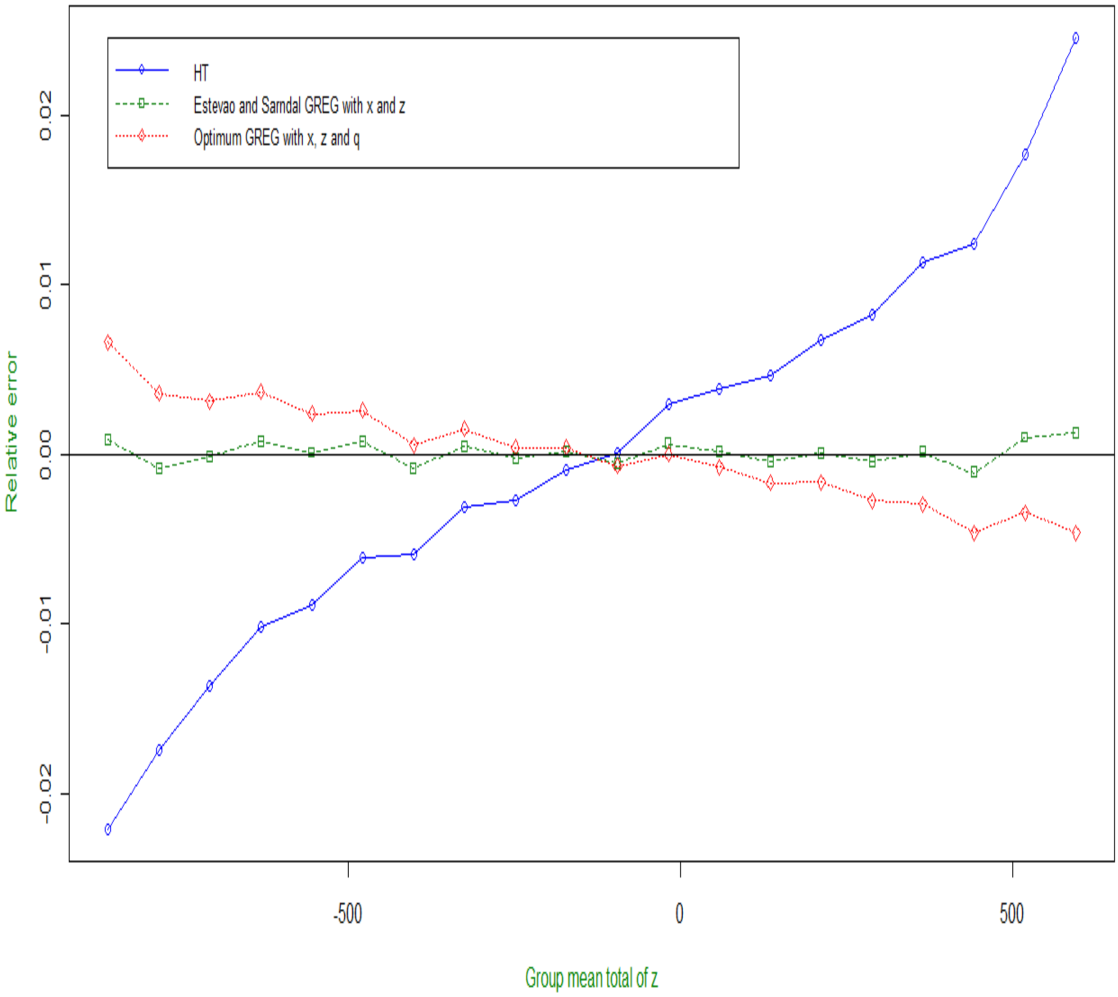

Figure 6.2: Relative bias for HT, Estevao and Särndal with x and z and optimal GREG estimators against the group mean total of z for random stratification when $r_{yx} = 0.2$ and $r_{yz} = 0.7$.

In Figure 6.2 we consider the case of a random stratification. In this situation, we see that the Estevao and Särndal GREG estimator performs better than the optimal GREG with x, z and q and the Horvitz-Thomson GREG estimators. The Horvitz-Thomson GREG estimator shows a linear trend.

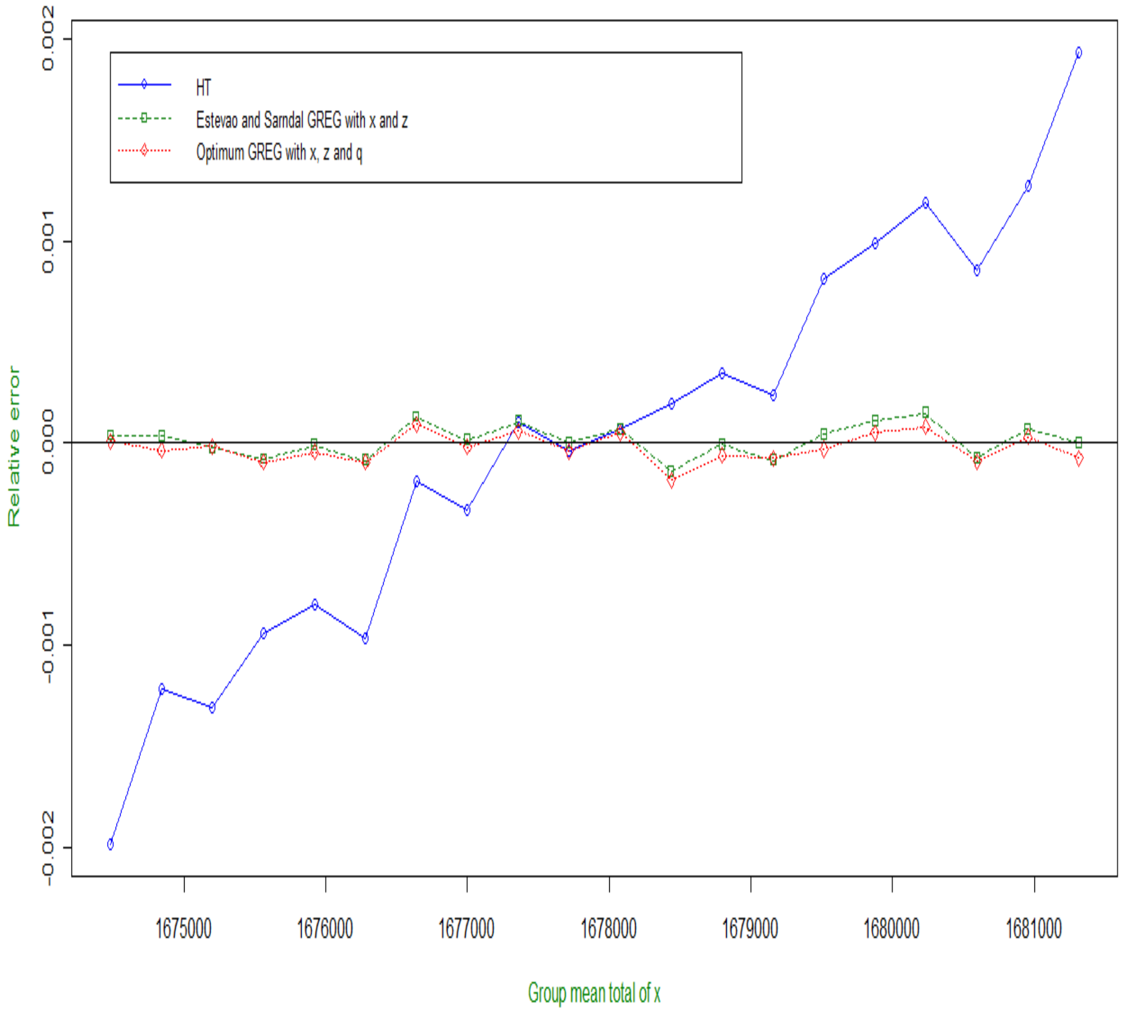

Figure 6.3: Relative bias for HT, Estevao and Särndal with x and z and optimal GREG estimators against the group mean total of x stratification by cluster size when $r_{yx} = 0.7$ and $r_{yz} = 0.2$.

Similar pattern is observed in Figure 6.3. The optimal GREG with x, z and q performs the best. We do not observe much differences when compare to the Estevao and Särndal GREG estimator. There is a linear trend for the Horvitz-Thomson GREG estimator.

Figure 6.4 we consider a random stratification. We observe similar results to Figure 6.2. We see that the Estevao and Särndal GREG estimator is the best. However, we do not observe significant difference compared to the optimal GREG with x, z and q. The Horvitz-Thomson estimator shows a linear trend.

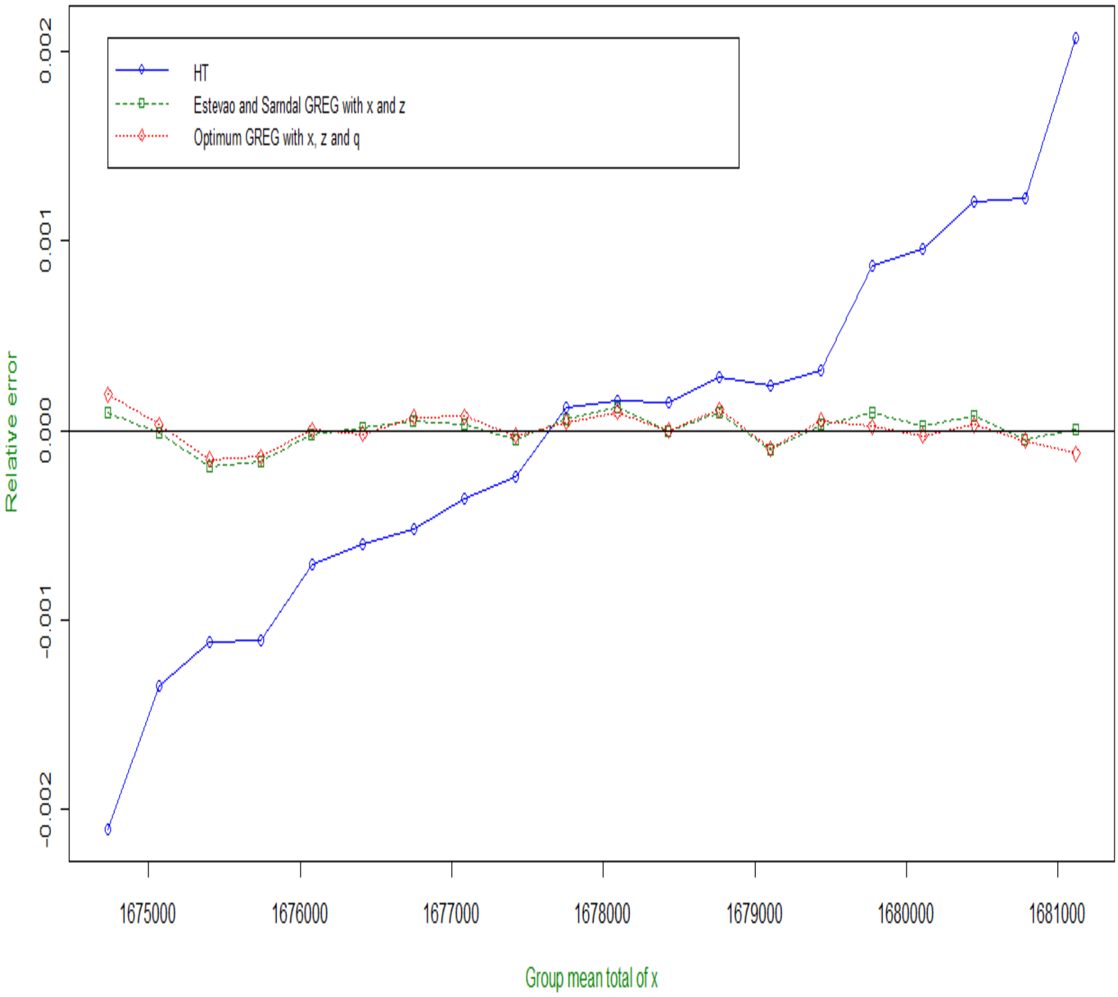

Figure 6.4: Relative bias for HT, Estevao and Särndal with x and z and optimal GREG estimators against the group mean total of x for random stratification when $r_{yx} = 0.7$ and $r_{yz} = 0.2$.

This section's results support the simulation results in Table 6.2. The optimal GREG estimator with variable x, z and q performs well compared to the Estevao and Särndal estimator with x and z. It is also better than the Horvitz-Thomson estimator for stratification by cluster size but not for random stratification.

6.6 Conclusion

The simulation results show that the optimal GREG estimators with variable x, z and q works well under stratification by cluster size in terms of relative bias and relative root mean squared error. The optimal GREG estimator with z and q is accurate compared with the other GREG estimators including the Horvitz-Thomson estimator. The optimal GREG estimator is less accurate for random stratification. The Estevao and Särndal estimator with x and z is more accurate in this situation. The optimal GREG estimator with x, z and q may be as accurate as the Estevao and Särndal estimator. This is the case when there are a small number of SSU in each stratum. We did not observe significant difference with large intra-cluster correlation and stratum size.

We observe different results in Table 6.3, where there are 3 strata and the number of SSU in each stratum varies between 10 and 15. We see that the optimal GREG estimator with x, z and q performs well under both stratifications. Although, the Estevao and Särndal GREG estimator with variable x and z also performs well under stratification by cluster size in this situation as well and some situations with random stratification.

Chapter 7

Conclusions and Discussion

In chapters 1 to 6, we looked at the existence of nonresponse in surveys with cluster designs and formulated new methods of analysis which treats missing data. We proposed new estimators of regression coefficients in a linear regression model with cluster-level variables when some of the data of the response variables are missing. We also proposed an extension to the Heckman estimators under a model for clustered survey data. In chapter 7, we proposed a new generalised regression estimator (GREG) for estimating a population total for two-stage (cluster) sampling. To compare the performance of our estimators with existing estimators, we performed simulation studies. In addition, we applied the new methods to the Workplace Employment Relations Survey (WERS) 2004 data. In the final chapter, we aim to discuss the value of the proposed alternative estimator and the Heckman estimator when analysing survey data under clustered designs and the new GREG estimator for estimation. We also consider possible further avenues of study.

Nonresponse in sample surveys is a common problem which can potentially result in large biases in the analysis and estimation of sample survey data. In the first part of the thesis, we considered how to use observed data in order to estimate appropriate regression coefficients in a linear regression model of cluster-level variables under missing data. There is some literature related to how to deal with nonresponse at the cluster level as described in Chapter 2 but the new method in this thesis explores the idea of adding nonresponse information at cluster level variables into the linear regression model. This was compared to the naïve approach which simply estimates the regression coefficients without considering the nonresponse. Furthermore, we extended the Heckman estimator to take into account nonresponse under the cluster sample designs. This research on the analysis of cluster level survey data which takes into account nonresponse in the regression model is new.

We investigated the efficacy of the new method for analysing clustered survey data for both MAR and NMAR models under the following assumptions: firstly, that we need to have at least one respondent in each cluster and secondly, that the difference between the expectation of the response variable and nonresponse variable must be constant across clusters. The first assumption is reasonable as no cluster can exist in any of our functioning models that has full nonresponse. The second assumption simplifies the model as it provides one constant across clusters but there is scope for future work to consider a more complex model, e.g. by allowing differing values of nonresponse and the expectation of it across clusters.

We developed the theory and showed under simulation that the proposed alternative estimator is unbiased under the NMAR model and is preferable to the naïve approach which is biased. Unfortunately, there was some indication of bias under the MAR model however in practice the nonresponse mechanism occurs more often with NMAR model than with MAR model so we do not consider the small bias to be a significant limitation with our estimator. Therefore, our proposed alternative estimator seems to be more powerful than the existing naïve one.

The models of interest in this thesis for analyzing cluster level survey data is the linear regression model and the multilevel model. We considered these models because they cover the models more commonly used when analysing complex survey designs. The focus of our work on nonresponse is at the cluster level only.

We extended the Heckman estimators to cover the modelling for clustered survey data. The Heckman estimator is well known to economists and deals with selectivity in the modelling. For that reason, we included the Heckman estimator to incorporate the selectivity of nonresponse into our model to see how it performs when analysing cluster data. We also extended the study of the proposed methods to two-stage cluster sampling and it worked in the same way as expected while assuming both MAR and NMAR mechanisms. This points to our methods being useful in even more complicated sample surveys if we follow the same iteration steps for estimation.

The main findings in the simulation results from models underlying the naïve and alternative approaches indicate that the new alternative approach estimator produced unbiased results with the NMAR model but did show higher variance and mean square error in all tests except when the δ factor increased. This is not surprising and is expected for unbiased estimators which tend to give higher variance and higher mean square error than biased estimators. As expected the naïve approach gave an unbiased, minimum variance estimator under the MAR model but is biased under the NMAR model. In the simulation studies based on the Heckman estimators, the first Heckman two-step estimator and the approximate Heckman maximum likelihood estimator performed well showing reduced bias when compared to that produced under the naïve approach. The Heckman maximum likelihood estimator reduces variance and lowers mean square error when ρ increased. The new alternative approach behaves similarly to the Heckman two-step estimator using p_i in both models. We can see that in all simulations that we ran, at least one of our proposed estimators worked well but it would be interesting to investigate in further research using different models with larger variances in parameters to see how our proposed estimators perform in other scenarios.

Surprisingly, we found that when replacing p_i in (4.2) by π_i that quite unexpectedly we achieved a biased result. The theory that we presented stated that under the NMAR model the alternative estimator should be unbiased. The reason for the discrepancy is that the assumed model no longer holds. If m_i is not large then p_i and π_i may be quite different. This may explain the resulting bias which can be viewed as bias caused by measurement error.

Unexplained results appeared in the models underlying the naïve and our alternative approach when we repeated the simulation study replacing x_i in (3.11) and (4.2) by z_i. The alternative approach and the Heckman approaches performed completely differently than when we used x_i in both models (3.11) and (4.2). Both estimators performed poorly as we can see bias in all estimators. However replacing x_i works well in the model underlying Heckman estimators and although we can not find an exact reason why this occurred it might happen because there is a difference in correlation and covariance between these two variables or it could actually be for other related reasons due to measurement errors. Although, there is no clear indication in our theory that there should be a difference in any results caused by using different variables in these two Heckman estimator models it would be useful to investigate and find out why the difference occurred under measurement errors.

We were interested to see how our proposed estimators would perform with real data so we chose the data from the Workplace Employment Relations Survey (WERS) 2004 data where there are 2 levels, a single cluster and a single element which is the employees within the workplace. There was a particularly large incidence of nonresponse by employees in this survey. Although there were difficulties in dealing with a large set of data we applied the proposed methods to the WERS 2004 data following Bryson et al. (2009) who focussed their analysis on private sector workplaces only and examined the effects of innovations (management-initiated workplace changes) had on worker well-being. We run simulations at both the individual and cluster level unlike Bryson et al. (2009) who only focussed his study at the individual level. Our results show that the proposed alternative approach only worked well at individual level so we carried out further analysis to try and identify the problems with this dataset. We found an issue of unequal variance across clusters which is a common occurance in real applications. We successfully used the weighted least square method to solve the problem and got a better result for unweighted estimates. Surprisingly, the alternative approach did not work well at workplace level for both unweighted and weighted surveys where we found results other than those expected. This can be explained because the assumed model no longer holds. For example, we assumed the population has normal distribution in the simulation study but that might not be true in real data. The extended Heckman estimators work well at individual level as well but we did not consider it at the cluster level variable where there tends to be more problems in allowing for maximum likelihood estimation with real data.

In the second area of this thesis the generalized regression estimator (GREG) for two-stage (cluster) sampling is considered. We proposed a new regression estimator based upon the optimal estimator proposed by Berger et al. (2003) which can be used for stratified two-stage sampling designs when the sampling fraction is negligible and the primary sampling units are selected with unequal probabilities. We assume that there are auxiliary variables available for the secondary sampling units and the primary sampling units. We proposed to use an ultimate cluster approach to estimate the regression coefficient of the regression estimator. Estevao and Särndal (2006) proposed a regression estimator for two-stage sampling so we compared the proposed estimator with the Estevao and Särndal (2006) estimator under a self-weighted two-stage sampling design. The simulation results show that the proposed estimator may be more accurate than the Estevao and Särndal (2006) estimator when PSU of similar sizes are grouped in the same strata; that is, when the strata are homogeneous according to the PSU sizes. Note that this is a situation which is not uncommon in practice. If the strata are not related to the PSU sizes, the Estevao and Särndal (2006) estimator is slightly more accurate than our proposed estimator. In this situation, the loss of efficiency of the proposed estimator is minor.

We have developed statistical theory for investigating nonresponse in the analysis and estimation for regression models with clustered data by introducing the alternative estimator which incorporate information on nonresponse at the cluster level and also developed the Heckman estimators for use at cluster level. Moreover, we proposed to investigate different approaches for estimating a population total under two-stage (cluster) sampling. We proposed to adjust the Estevao and Särndal (2006) GREG estimator following studies by Berger et al. (2003). We compared the estimators for clustered data using both simulation study and WERS 2004 data.

Following the theory described in chapters 3 and 6, a number of tasks could be undertaken in future work. First of all, we can apply the proposed estimators to the logistic regression model. Secondly, we can adjust some of the assumptions allowing δ_i to have variability across the cluster even though it could lead to more complications in analysis. We can apply the proposed estimators to different sets of real data and see how they perform. Furthermore, we can investigate how the proposed estimators perform under more complicated survey sampling designs. We can also compare in theory the proposed GREG estimators with other estimators. Finally, we can investigate how the proposed estimators can be extended under imputation and weighting methods.

Appendix A

Proposition 16.1 Cameron and Trivedi (2005)

Preposition 16.1 of Cameron and Trivedi (2005) *Truncated Moments of the Standard Normal.*

Supposed z has a normal distribution with mean is equal to zero and variance is equal to one. The left-truncated moment of z are
$E(z|z > c) = \phi(c)/[1 - \Phi(c)]$ and $E(z|z > -c) = \phi(c)/\Phi(c)$, where ϕ is the probability density function of the standard normal distribution and Φ is the cumulative distribution function of this distribution.

References

Alexander, C. (1987). A class of methods for using person controls in household weighting. *Survey Methodology*, 13:183–198.

Andridge, R. R. and Little, R. J. A. (2010). A review of hot deck imputation for survey non-response. *International Statistical Review*, 78(1):40–64.

Antcliff, V. and Saundry, R. (2009). Accompaniment, workplace representation and disciplinary outcomes in british workplaces - just a formality? *British Journal of Industrial Relations*, 47(1):100–121.

Berger, Y. (2005). Variance estimation with chao's sampling scheme. *Journal of Statistical Planning and Inference*, 127(12):253 – 277.

Berger, Y. G. and Rao, J. N. K. (2006). Adjusted jackknife for imputation under unequal probability sampling without replacement. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):531–547.

Berger, Y. G., Tirari, M. E., and Till, Y. (2003). Towards optimal regression estimation in sample surveys. *Australian & New Zealand Journal of Statistics*, 45(3):319–329.

Bethlehem, J. G. (2009). *Applied survey methods: a statistical perspective*. John Wiley & Sons, Inc.

Bethlehem, J. G. and Keller, W. J. (1987). Linear weighting of sample survey data. *Journal of Official Statistics*, 3(2):141–153.

Brick, J. M., Jones, M. E., Kalton, G., and Valliant, R. (2005). Variance estimation with hot deck imputation: A simulation study of three methods. *Survey Methodology*, 31(2):151–159.

Brick, J. M. and Kalton, G. (1996). Handling missing data in survey research. *Statistical Methods in Medical Research*, 5(3):215–238.

Bryson, A., Dale-Olsen, H., and Barth, E. (2009). How does innovation affect worker well-being? *CEP discussion paper*, (953).

Bryson, A. and Freeman, R. B. (2008). How does shared capitalism affect economic performance in the uk? *CEP discussion paper*, (885).

Cameron, A. and Trivedi, K. P. (2005). *Microeconometrics: Methods and applications.* Cambridge University Press.

Chambers, R. L. and Dunstan, R. (1986). Estimating distribution functions from survey data. *Biometrika*, 73(3):597–604.

Chambers, R. L. and Skinner, C. J. (2003). *Analysis of survey data.* John Wiley & Sons, Ltd.

Chaplin, J., Jane Mangla, S. P., and Airey, C. (2005). *The Workplace Employment Relations Survey (WERS) 2004: Technical report(cross-section and panel surveys).* National Centre for Social Research.

Chatterji, M. and Mumford, K. (2008). Flying high and laying low in the public and private sectors: a comparison of pay differentials for full-time male employees in britain. *Dundee Discussion Papers in Economics.*

Deville, J. (1997). Estimation de la variance du coefficient de gini mesur par sondage. *Actes des Journes de Mthodologie Statistiques, INSEE mthodes[In French].*

Durrant, G. B. (2005). Imputation methods for handling item-nonresponse in the social sciences&58; a methodological review. Working Papers id:1918, eSocialSciences.

Durrant, G. B. and Skinner, C. (2006). Using missing data methods to correct for measurement error in a distribution function. *Survey Methodology*, 32(1):25–36.

Estevao, V. M. and Särndal, C.-E. (2006). Survey estimates by calibration on complex auxiliary information. *International Statistical Review*, 74(2):127–147.

Fay, R. (1991). A design based perspective on missing data variance. *Proceedings of the 1991 Annual Research Conference*, pages 429–440.

Gelman, A., King, G., and Liu, C. (1998). Not asked and not answered: Multiple imputation for multiple surveys. *Journal of the American Statistical Association*, 93(443):pp. 869–874.

Groves, R. M., Dan, John, and Little (2002). *Survey nonresponse.* John Wiley & Sons, Ltd., second edition.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., and Tourangeau, R. (2009). *Survey methodology.* John Wiley & Sons, Ltd., second edition.

Hansen, M.H.and Hurwitz, W. and Madow, W. (1953). *Sample Survey Methods and Theory.* John Wiley & Sons, Inc.

Haziza, D. and Rao, J. (2001). Model-assisted approach to inference for totals in cluster sampling under imputation for missing data, cd-rom. *Proceedings of the Annual Meeting of the American Statistical Association.*

Haziza, D. and Rao, J. (2006). A nonresponse model approach to inference under imputation for missing survey data. *Survey Methodology*, 32(1):53–64.

Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement*, 5(4):475–492.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, 47(1):153–61.

Heeringa, G. S., West, B., and Berglund, P. A. (2010). *Applied survey data analysis*. Chapman and Hall.

Holt, D. and Smith, T. M. F. (1979). Post stratification. *Journal of the Royal Statistical Society. Series A (General)*, 142(1):pp. 33–46.

Kalton, G. and Kasprzyk, D. (1982). Imputing for missing survey responses. *Proceedings of the Section on Survey Research Methods*, pages 22–31.

Kott, P. S. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology*, 32(2):133–142.

Lemaître, G. and Dufour, J. (1987). An integrated method for weighting persons and families. *Survey Methodology*, 13:199–207.

Little, R. J. and Vartivarian, S. (2005). Does weighting for nonresponse increase the variance of survey means? *Survey Methodology*, 31(2):161–168.

Little, R. J. A. (1982). Models for nonresponse in sample surveys. *Journal of the American Statistical Association*, 77(378):pp. 237–250.

Little, R. J. A. (1992). Regression with missing x's: A review. *Journal of the American Statistical Association*, 87(420):pp. 1227–1237.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical analysis with missing data*. John Wiley & Sons, Ltd., second edition.

Little, R. J. A. and Vartivarian, S. (2003). On weighting the rates in non-response weights. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 60(1):pp. 23–40.

Montanari, G. (1987). Post sampling efficient qr-prediction in large sample survey. *International Statistics*, 55(2):191–202.

Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H., and Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 60(1):pp. 23–40.

Rao, J. (1994). Estimating totals and distribution functions using auxiliary information at the estimation stage. *Journal of Official Statistical Review*, 10(2):153–165.

Rao, J. N. K. (1996). On variance estimation with imputed survey data. *Journal of the American Statistical Association*, 91(434):pp. 499–506.

Reiter, J. P., Raghunathan, T. E., and Kinney, S. K. (2006). The importance of modeling the sampling design in multiple imputation for missing data. *Survey Methodology*, 32(2):pp. 143–149.

Rose, M. (2007). Why so fed up and footloose in it? spelling out the associations between occupation and overall job satisfaction shown by wers 2004. *Industrial Relations Journal*, 38(4):356–384.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* New York, Chichester.

Salis, S. and Williams, A. M. (2010). Knowledge sharing through face-to-face communication and labour productivity: Evidence from british workplaces. *British Journal of Industrial Relations*, 48(2):436–459.

Särndal, C.-E. (2007). The calibration approach in survey theory and practice. *Survey Methodology*, 33(2):99–119.

Särndal, C.-E. and Lundström, S. (2005). *Estimation in surveys with nonresponse.* John Wiley & Sons, Ltd., second edition.

Schyns, B., van Veldhoven, M., and Wood, S. (2009). Organizational climate, relative psychological climate and job satisfaction : the example of supportive leadership climate. *Leadership & organization development journal.*, 30(7):649–663.

Sessions, J. and Theodoropoulos, N. (2009). Tenure, wage profiles and monitoring. *Working Paper*, (27/09).

Shao, J. (2007). Handling survey nonresponse in cluster sampling. *Survey Methodology*, 33:pp. 81–85.

Shao, J. and Steel, P. (1999). Variance estimation for survey data with composite imputation and nonnegligible sampling fractions. *Journal of the American Statistical Association*, 94(445):pp. 254–265.

Shao, J. and Wang, H. (2002). Sample correlation coefficients based on survey data under regression imputation. *Journal of the American Statistical Association*, 97(458):pp. 544–552.

Skinner, C. (1998). Calibration weighting and non-sampling errors. *Research in Official Statistics*, 2:33–43.

Skinner, C. J. and Coker, O. (1996). Regression analysis for complex survey data with missing values of a covariate. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 159(2):pp. 265–274.

Skinner, C. J. and Darrigo (2011). Inverse probability weighting for clustered nonresponse. *Biometrika*, 98(4):953–966.

Skinner, C. J. and Rao, J. N. K. (2002). Jackknife variance estimation for multivariate statistics under hot-deck imputation from common donors. *Journal of Statistical Planning and Inference*, 102:pp. 149–167.

Steel, D. G. and Clark, R. G. (2007). Person-level and household-level regression estimation in household surveys. *Survey Methodology*, 33(1).

Tan, Z. (2013). Simple design-efficient calibration estimators for rejective and high-entropy sampling. *Biometrika*, 100(2):399415.

Van Buuren, S., Brand, J. P., Groothuis-Oudshoorn, C. G., and Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, 76(12):1049–1064.

West, B. T. (2009). A simulation study of alternative weighting class adjustments for nonresponse when estimating a population mean from complex sample survey data. *Section on Survey Research Methods JSM*, pages 4920–4933.

Wood, S. (2008). Job characteristics, employee voice and well-being in britain. *Industrial Relations Journal*, 39(2):153–168.

Wood, S. and Fairleigh, E. (2007). Well-being amongst british employees: The evidence from the workplace employee relations survey of 2004. *Background paper for Policy Studies Institute, Employment and Social Policy seminar, 8th March 2007.*

Yuan, Y. and Little, R. J. A. (2007). Model-based estimates of the finite population mean for two-stage cluster samples with unit non-response. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 56(1):79–97.

Yuan, Y. and Little, R. J. A. (2008). Model-based inference for two-stage cluster samples subject to nonignorable item nonresponse. *Journal of Official Statistics*, 24(2):193–211.

Yuan, Y. C. (2000). Multiple imputation for missing data : Concepts and new development. *Proceedings of the twentyfifth annual SAS Users group international conference*, pages 1–11.