2019

Finite Mixture of Regression Models for Complex Survey Data

Abdelbaset Abdalla
South Dakota State University

FINITE MIXTURE OF REGRESSION MODELS FOR COMPLEX SURVEY DATA

BY

ABDELBASET ABDALLA

A dissertation submitted in partial fulfillment of the requirements for the

Doctor of Philosophy

Major in Computational Science and & Statistics

South Dakota State University

2019

DISSERTATION ACCEPTANCE PAGE

Abdelbaset Abdalla

This dissertation is approved as a creditable and independent investigation by a candidate for the Doctor of Philosophy degree and is acceptable for meeting the dissertation requirements for this degree. Acceptance of this does not imply that the conclusions reached by the candidate are necessarily the conclusions of the major department.

Semhar Michael
Advisor Date

Kurt Cogswell
Department Head Date

Dean, Graduate School Date

To my mother, and the memory of my dear father

ACKNOWLEDGMENTS

This Dissertation would never have been possible without the help and support of my parents and my wife. I would firstly like to express my deepest gratitude to my supervisor, Professor Semhar Michael, for her continuous support from the initial to the final level, for her great patience, motivation, excellent guidance, and immense knowledge. I could not have imagined having a better or friendlier supervisor.

Most importantly, I would also like to thank my committee members Dr. Christopher Saunders, Dr. Gary Hatfield, and Dr. Zhiguang Wang. I have a great deal of respect for each of you, you all have very different careers, and I have learned so much by working with each one of you. I would like to thank the University of Benghazi for scholarship and their financial support.

I am very grateful for Dr. Kurt Cogswel and Mathematics & Statistics Department for their financial support for the first arrival and for giving me this opportunity.

Also, I would like to thank South Dakota State University and the Brookings community for accepting me to complete my study and for all of the lifetime memories that I got during my stay in the United States.

Finally, I would like to thank my parents, sisters, my wife, and my dear kids, for always supporting me and encouraging me with their love and best wishes.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

API . . . . . Academic Performance Index.

BIC . . . . . Bayesian information criterion.

CS . . . . . Complex sample.

EM . . . . . The expectation maximization algorithm.

FMR . . . . . Finite mixture of regression.

IID . . . . . Independent identically distributed.

IQR . . . . . Interquartile range.

MOM . . . . . Method of moments.

MLE . . . . . Maximum likelihood estimation.

MSE . . . . . Mean Squared error.

NCHS . . . . . National Center for Health Statistics.

NHANES . . . . . National Health and Nutrition Examination Survey.

PML . . . . . Pseudo maximum likelihood.

PSU . . . . . Primary sampling unit.

SRS . . . . . Simple random sample.

SSU . . . . . Secondary sampling unit.

NOTATIONS

α_k: The mixing proportion of kth component.

$\boldsymbol{\beta}_k$: The regressions coefficients of kth component.

K: The total number of mixture regression components.

M_i : Number of population units in ith selected cluster

m_i: Number of elements in the sample from the ith PSU.

N: Population size

n: Sample size.

N_h: Number of sampling units (PSU's) in stratum $h, h = 1, ..., H$.

n_h: Number of PSU's sampled from stratum $h, h = 1, ..., H$.

N_c: Number of (PSU's) clusters in the population.

n_c: Number of (PSU's) clusters in the sample.

π: Inclusion probability of a sampling unit.

$\boldsymbol{\Psi}$: The parameter vector.

σ_k^2: The variance of kth component.

$\mathcal{R}$: Percent contribution Index.

U: Finite population

$\mathbf{W}_k^{(t)}$: $n \times n$ diagonal matrix with entries $w_i \times \tau_{ik}^{(t)}$.

$\mathbf{X}$: $n \times (p + 1)$ matrix containing unity for intercept and predictors.

$\mathbf{y}$: $n \times 1$ vector of responses.

ABSTRACT

FINITE MIXTURE OF REGRESSION MODELS FOR COMPLEX SURVEY DATA

ABDELBASET ABDALLA

2019

Over time, survey data has become an essential source of information for modern society. However, to be effective, the structures of survey data require sampling designs that are more complex than simple random sampling. The complex sampling data collected from enormous national surveys via these complex designs ideally include sample weights that allow analysis to take account of complicated population structures. When the target of inference is the parameters of a regression model, it is crucial to know whether these weights should be incorporated into the sampling weight when fitting the model to the survey data. The finite mixture models are one tool for modeling heterogeneity and finding the subgroups in the data. Limited literature is available on modeling survey data via the finite mixture of regression models using a complex survey design.

The principal aim of this dissertation is to develop and evaluate strategies for survey data modeling using a new design-based inference, where sampling weights are integrated into the complete-data log-likelihood function. More specifically, the pseudo maximum likelihood estimator (PML) has been considered, so the expectation-maximization (EM) algorithm was developed accordingly. In order to evaluate this strategy in realistic circumstances, we simulated the performance of the proposed model under numerous scenarios. Comparisons were made using bias-variance components of the mean squared error. Additionally, the Bayesian information criterion was utilized and assessed as a selection tool under the proposed modeling approach. Finally, we applied the proposed approach to original survey datasets to assess its practical usefulness

1 INTRODUCTION

In modern society, survey accumulated data provides significant statistics in ultimately creating a positive philosophy of change. Accurate facts gathered though surveys become instrumental in the process of making decisions. Vital decisions include the implementation of program adjustments to better address the needs of a population, the improvement of community policies and projects, creating priorities when allocating funds in regard to government agencies, and public queries. Data and evidence combined from different surveys facilitates the progress of health issues globally. In this framework, evidence collected from social surveys, one of the most essential data sources, enables understanding of changes in societal social trends, empowering the examinations of change for the benefit of citizens as well as communicating a vision on issues specific to social policy. Equally, health survey data plays a fundamental role in advising policymakers, as well as the public, regarding significant health issues implemented by strategies and procedures. Therefore, survey data contributes the most vital evidence to a focused and successful decision-making processes concerning the implementation of government and global policies. Reliable and unbiased methods of attaining information from a survey commands scrutiny particularly since this information creates the basis for making choices affecting large target populations. Specifically, the necessity to establish dependable survey methods must launch with a small sample in order to consider and infer characteristics of relationships of a vast population. Multifaceted survey datasets contain distinctive structures that require an analytical approach that cannot be achieved using standard techniques. Therefore, necessities for the development of statistical methodologies intensify in order to extract information from data collected from complex survey designs. Statistical sampling techniques and the analysis of complex survey data is detailed in Kish (1965), Cochran (1977), Kalton and Graham (1983), Lohr (2010), and in a more current issue published in the statistical science journal

(Zhang, 2017).

The bulk of all-encompassing surveys use two principal methodologies of statistical inference; namely, design based and model-based inference. These tactics incorporate complexities into survey sampling, such as clustering, stratification, and unequal probabilities of the selection mechanism. In the 1950s, model-based analysis began by Godambe (1955) and Royall (1968). Design-based analysis initiated by Neyman (1934) is used in the framework of survey sampling design to generate inference for population limitations. If we consider the superpopulation model and let $\mathbf{\Psi}_N$ to be a finite population estimator for the model parameter $\mathbf{\Psi}$, which could be computed if the entire population U was observed. If the components of the parameter vector $\mathbf{\Psi}_N$ can be expressed as functions of a finite population parameter, it is again possible to estimate it by a design-based estimator. However, the model-baIf we consider the superpopulation model and let $\mathbf{\Psi}_N$ to be a finite population estimator for the model parameter $\mathbf{\Psi}$, which could be computed if the entire population U was observed, if the components of the parameter vector $\mathbf{\Psi}_N$ can be expressed as functions of a finite population parameter, it is again possible to estimate it by a design-based estimator. However, the model-based is viewing the target population itself as a random realization from the superpopulation model. In this view, the finite population quantity $\mathbf{\Psi}_N$ is viewed as one particular realization of an estimator of the superpopulation model parameter $\mathbf{\Psi}$. Figure 1.1 illustrates the traditional view of the design-based and model-based. In this dissertation, a design-based technique used as an analysis tool, for a given dataset gathered using complex sampling design, will be considered. Principally, the design considers the complication of finite mixture linear regression analysis in order to analyze complex survey data.

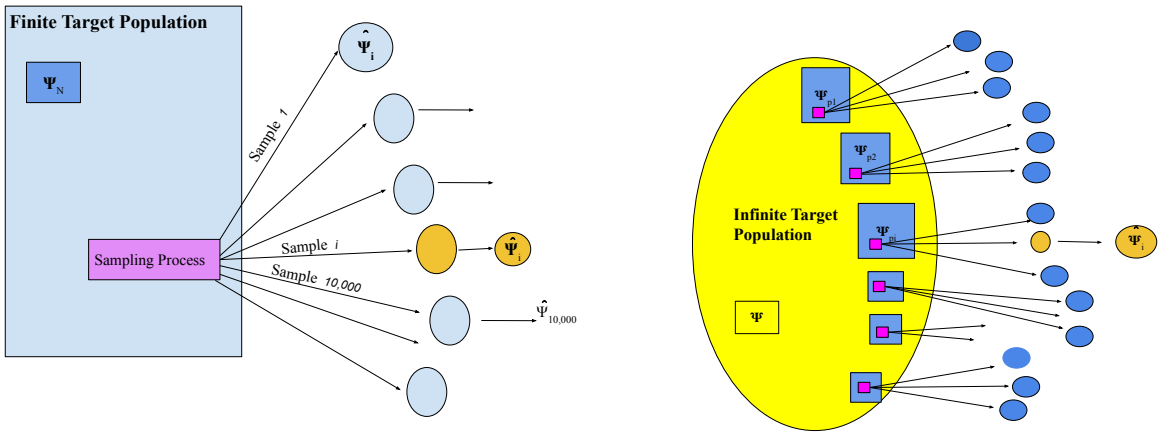


Figure 1.1: Diagrams representing classical design-based inference (on the left), model-based inference for super-population parameters (on the right).

Contemporary statistical applications widely use regression analysis. Essentially, regression analysis demonstrate a path of responses based on its relationship with one or more predictor or explanatory variables. Commonly, applications of linear regression evaluate independent identically distributed (IID) data. However, when investigators perform regression analysis on survey data the assumption is repeatedly inadequate in complex survey sampling designs. Subsequently, linear regression models and estimators typically apply to the inquiry of complex survey data using the PML method first recommended by Binder (1983) following the idea from Skinner et al. (1989). DuMouchel and Duncan (1983) have been discussed the sampling weights in multiple regression analysis for stratified samples. Survey designs through strata, cluster, or a combination of the two, strive to capture the heterogeneity in population in a more economical way. Nonetheless, occasionally subpopulations occur after data collection. One malleable technique for modeling heterogeneity in data uses finite mixture models (McLachlan and Peel, 2000). Finite mixture regression models (Leisch, 2004; Grün and Leisch, 2008) permit simultaneous outcomes of original subpopulations and structuring a regression model for each subpopulation in the data. Correspondingly, this dissertation explores fitting finite mixture linear regression models to sample survey data by including sampling weights to the regression parameter estimators.

1.1 Literature Review of Sampling Designs

Consider a finite population U, comprised of a set of N units termed $1, ..., N$, and a vector of parameters of interest,$\mathbf{\Psi}$, to be assessed. Supposing having used the entire population to estimate $\mathbf{\Psi}$, but regrettably, the population usually exceeds the amount, increasing cost, or too complex to pull together the required statistics from each population division to analyze $\mathbf{\Psi}$. Hence, one obtains a sample of size n from the population, which provides the data with which Ψ can be estimated. Let this estimator be represented by $\widehat{\mathbf{\Psi}}$. The quality of the precision of $\widehat{\mathbf{\Psi}}$ as an estimator of $\mathbf{\Psi}$ depends on, amid other factors, how closely the sample exemplifies the population of interest. An impeccable sample would be similar to Grand view: a scaled-down version of the population, reflecting every distinctive feature of the entire population. Indubitably, such an idyllic sample cannot occur for complex populations. As an alternative, effective sampling ensures that the characteristic of importance in the population, $\mathbf{\Psi}$, can be estimated from the sample by $\widehat{\mathbf{\Psi}}$ and the precision of the estimation can be calculated (Lohr, 2010).

Sampling methods split into two classifications, a probability or a non-probability sampling technique. The methodology behind non-probability techniques, such as convenience or purposive sampling, automatically eliminates specific population units from the sampled population due to techniques that choose sample units via subjective evaluation. In general, this form of sample selection causes the estimate, $\widehat{\mathbf{\Psi}}$, to be biased. Moreover, in the absence of any probability techniques in the selection process, the degree of bias is indefinite. Any effect concluded from non-probability samples subjects itself to an unidentified level of bias (Lohr, 2010). A vital necessity of a probability sampling techniques certify that each possible sample of size n accumulated from the finite population has an identified probability of being selected (Chen et al., 2017). The use of a random mechanism to establish population units selected for the sample reduces the possibility of altering a pre-selected unit for a different unit based on personal judgment. Henceforth, by means of the application of a probability sampling technique, each individual population unit has an assured

chance of appearing in the sample. The possibilities underlying all potential samples of size n gathered using a probability sampling technique allow the establishment of the sampling distribution of $\widehat{\mathbf{\Psi}}$, the estimator of $\mathbf{\Psi}$, making it possible to detect inference using $\hat{\mathbf{\Psi}}$ and likewise defining the quality of the inference by means of the evaluation of standard errors, biases, etc. of the estimators Lohr (2010). Common types of probability sampling include simple random sampling, cluster sampling, stratified sampling, and multistage sampling. Complex sampling, which contrasts with simple random sampling, applies one or more unequal random selection mechanisms. The most popular designs involve using stratified sampling and cluster sampling, or any combination of sampling designs. One might want to consider complex survey design as opposed to simple random sampling as the list of the population may not be available, and even if it is, it might be extremely inefficient to collect data. Besides, any analysis of complex survey data that ignores both sample weights and the sampling design may lead to biased estimation and inaccurate inference. For statistical inference, when studying survey sample data, considering the sampling design is imperative. In chapter 2, reviews of properties of estimates for the principal design mechanisms used in a probability and sample survey design contain stratified, cluster sampling, and complex sampling. The integration of these ideas in section 4.3 demonstrates how they work collaboratively in complex surveys such as the National Health and Nutrition Examination Survey (NHANES).

1.2 Literature Review of Sampling Weights

The primary objective of the sampling theory is to gain insights concerning population parameters of interest. So, insights about those population parameters of interest can be inferred from the sample. Therefore, the importance of employing sampling weights in inference is to adjust for imperfections, for instance, unequal probabilities and population groups that are not adequately embodied in the particular sample. The use of probability

sampling techniques enables the determination of the inclusion probability of a population unit in the acquired sample. Let the inclusion probability of the ith population unit be defined as π_i and let w_i denotes the design weight. The most common definition of sampling weight is as an indicator of the number of population units that are represented by ith sample unit.

The sampling weights in the first stage are assigned to each sample unit to adjust for the unequal selection probabilities. Thus, the sampling weights might not be inverse of inclusion probability. The sampling weights are modified for several reasons. Some customary corrections include nonresponse, misspecification of the sampling frame, and post-stratification. The weights extend the clear-cut idea of design weights by incorporating auxiliary population data. An assortment of adjustments can be executed, and the formation of weights can be complex. Further details regarding weighing in complex surveys can be found in Kish (1992), Gelman et al. (2007), Särndal (2007), Haziza and Lesage (2016); Haziza et al. (2017), and Chen et al. (2017). In this dissertation, it is presumed that the sampling weights are inverse of the inclusion probability of a population unit being selected for the sample,

$$w_i = \frac{1}{\pi_i}, i = 1, ..., n,$$

where n is the size of the selected sample and is interpreted as the number of population units represented by the ith sampled unit. Subsequently $N = \sum_i w_i$, the size of the population from which the sample is selected (Horvitz and Thompson, 1952; Lohr, 2010). Whenever we are dealing with a real-life data application, we either compute the weights associated with each observation based on the sampling design or use the already existing weights available with the data.

1.3 Review of Finite Mixture Regression Models

In the nineteenth century, finite mixture models made their initial recorded appearance in modern statistical literature by Newcomb (1886) who used it in the framework of modeling outliers. In the years following, Pearson (1894) applied a mixture of two univariate Gaussian distributions to analyze a dataset containing ratios of the forehead to body lengths for 1,000 crabs, using the method of moments (MOM) to estimate the parameters in the model. The most prevalent mixture model is the one consisting of Gaussian components (Day, 1969; McLachlan and Basford, 1988; Fraley and Raftery, 2006). We refer to McLachlan and Peel (2000) and Frühwirth-Schnatter (2006) for a complete survey on the history and applications of finite mixture models.

Universally, finite mixture models are used to model data from a heterogeneous population. The power of finite mixture models through model-based clustering is that they allow us to cluster and classify with the assumption that each mixture component represents a set of an observation belonging to one group in the original data (McLachlan and Basford, 1988; Fraley and Raftery, 1998). Various fields of statistical applications such as medicine and biology use mixture distributions for many purposes see, for example,, the review chapter in Schlattmann (2009). All-encompassing dialogue concerning the derivations and applications of finite mixture models are presented in the monographs by McLachlan and Peel (2000) and Frühwirth-Schnatter (2006), and more recent reviews by Melnykov et al. (2015); McNicholas (2016); McLachlan et al. (2019) discusses recent advances and challenges in the topic of finite mixture models and model-based clustering

When a random variable with finite mixture distribution depends on some covariates, it acquires a finite mixture of regression (FMR) model (Khalili and Chen, 2007). The basic idea here is to be able to fit different regression models to portions of data that behave similarly. Quandt and Ramsey (1978) introduced mixtures of linear regression models as a very basic method of switching regression. De Veaux (1989) established an EM approach to fit the two regression situations. Jones and McLachlan (1992) applied combinations of

regressions in data analysis and applied the EM algorithm to suit these models. Applications of FMR models, in many capacities such as market segmentation and social sciences, are studied more carefully in Wedel and Kamakura (2012) and Rabe-Hesketh and Skrondal (2004). The model is implemented in the R software through the FLEXMIX package (Grün and Leisch, 2008).

Fitting regression models to survey data complicates estimating pure population quantities such as totals, means, quantiles and variances. In addition, one of the commonly sought after parameters is the census regression coefficient. This is what would be reached from a regression if the complete population had been sampled. In most detailed demonstration of these and other matters concerning regression, one of the question that surfaces is whether or not sampling weights should play a role when estimating the model parameters. This has been a topic of a debate for many years starting in the seventies. See for example Fuller (1975), Pfeffermann and Smith (1985), Skinner et al. (1989), Pfeffermann (1993) and Lumley and Scott (2017).

This dissertation will explore a finite mixtures of regression models that can be valid as a model when the samples were drawn from complex sample designs. A design-based inference incorporating sampling weight or design weight in the expectation-maximization algorithm will be developed. A presentation will be made of a simulation study and actual datasets, comparing weighted and unweighted models. Furthermore, validation will confirm the effect of incorporating the design weight in log-likelihood function to estimate the finite mixture parameters, using a simulation complex sampling design and a real dataset as well.

1.4 Outline of the Dissertation

The dissertation contains five chapters. In Chapter 2, sampling techniques are discussed in general, and we discuss how sampling weights are calculated and incorporated

to our proposed model procedure. The main focus of this chapter is, therefore, the development of this procedure, together with a discussion of how it can be implemented. This will be followed by the computational strategies used to model data that comes from complex survey designs using finite mixture models. Some theoretical aspects are presented in Chapter 2. Here we discuss the asymptotic behavior and conditions required for the maximum likelihood (ML) estimator. In particular, we introduce a robust estimator of the asymptotic standard error of the pseudo maximum likelihood estimator.

Chapter 3 contains the design of the simulation studies based on stratified sampling, cluster sampling, and complex sampling design. The simulation studies outlined in chapter 3 are very important to investigate the performance of the proposed model. This chapter describes results from a sequence of simulation experiments based on different complex sampling designs. The bias-variance components of the mean squared error will be used to evaluate and compare the proposed model with the alternative model. Some exciting and distinct simulation studies and applications of the proposed model will be presented.

In Chapter 4 we address the topic of how to apply the finite mixture of normal regression models for samples acquired using a complex sampling design, based on real survey datasets. This chapter includes an implementation of the modeling procedure to each complex sampling design through stratified, cluster, and complex sampling data, respectively. Chapter 4 describes results from sequence examples, based on a real-world dataset. One of the most famous national surveys is considered here. Finally, Chapter 5 validates the dissertation with overall remarks and summaries of the conclusions of this research. The chapter concludes with topics identified for further research.

Part of this dissertation can be found in the recent publication "Finite mixture of regression models for a stratified sample" Abdalla and Michael (2019) and can be found in the appendix of the dissertation. A draft manuscript prepared for submission to the Journal of Applied Statistics can be found in the appendix of the dissertation as well.

2 METHODOLOGY

This chapter offers thorough descriptions of the necessary foundation laid regarding finite mixture regression models and displays the proposed methodology used to model data that comes from complex survey designs using a finite mixture of regression models. In the first section, revision of popular probability sampling designs such as stratified sampling, cluster sampling, and complex sampling provides the essential foundation. In the next section, we will introduce definitions and notations related to the finite mixture of regression models. The maximum likelihood estimations of the finite mixture of regression models computed via the EM-algorithm will be described. Furthermore, the Pseudo maximum likelihood (PML) estimations of the parameters of a mixture of regression models are derived under the complex sampling data. This will be followed by discussing the general asymptotic behavior of the ML estimators obtained via the EM-algorithm. We will define the ML estimators as a particular case of M-estimator. Then, we will give a short introduction to the asymptotic concept of ML in general. We also include a section about the asymptotic standard error of ML estimator for mixture models obtained by the EM-algorithm. In the last section, a discussion will be more focused on the asymptotic standard error of the PML estimators of the mixture models when the complex sampling design is assumed.

2.1 Complex Survey Design

The purpose of this section is to revise well-known complex survey designs. This will be followed by a discussion of stratified sampling, cluster sampling, and complex sampling. Finally, the sampling weights will be defined and discussed as an integral part of complex sampling design.

Several statistical analyses assume data being analyzed constitutes a simple random sample (SRS), ensuring that all elements have the same likelihood of being selected in the sample. However, sampling in survey research often works differently. In general, samples are often stratified or clustered by variables of interest. Sampling methods fall into two classifications: (1) non-probability sampling, in which the probability of being selected in the sample is unknown, and (2) probability sampling, in which the probability of being selected is known. The most common types of probability sampling are simple random sampling, cluster sampling, stratified sampling, and multistage sampling. Complex sampling, which contrasts with SRS, applies one or more unequal random selection mechanisms. The most commonly used designs involve applying stratified sampling and cluster sampling, or any combination of sampling designs. For statistical inference, considering the sampling design is imperative when studying survey sample data.

In general, we consider the regression of a dependent variable $\mathbf{y}$ on a vector of independent variables $\boldsymbol{x}$. Then, $(\boldsymbol{x}_i, \mathbf{y}_i)$ denote the row vector of these variables for a unit with label i in the index $U = \{1, ..., N\}$ of a finite population of size N. Without loss of generality, assume a general complex sampling design $p(s)$ from which sample s of size n is drawn without replacement from the population U. The sampling design may involve combinations of sampling schemes. Let δ_i be the indicator variable of the ith unit which is equal to one if $i \in s$ and zero otherwise with restriction $\sum_{i=1}^{N} \delta_i = n$. Suppose that under the sampling design a sampling unit is denoted by $i, (i = 1, ..., n)$, we can define the first-order inclusion probability, π_i, as the probability of ith unit being selected in the sample. The second-order inclusion probability, π_{ij}, is the probability that the two units i, j are selected in the sample. Thus, using the indicator variable, $E(\delta_i) = \pi_i$, and $E(\delta_{ij}) = \pi_{ij}$. The inclusion probability of the ith observation, when we use SRS is defined as, $\pi_i = \frac{n}{N}$. More discussion about the inclusion probability can be found in Horvitz and Thompson (1952), Natarajan et al. (2008) and Lohr (2010).

2.1.1 Stratified Sampling

In this section, we consider modeling data gathered through stratified sampling. A stratified random sample is attained by separating the population elements into non-overlapping groups which are primary sampling units (PSU), called strata. Therefore, the population is the set of strata, $\{U_h\}_{h=1}^{H}$ with sizes $N_1, ..., N_H$ and $\sum_{h=1}^{H} N_h = N$. Then, a simple random sample of size n_h is selected without replacement from each stratum with $\sum_{h=1}^{H} n_h = n$. One property of stratified sampling is that it works best when a heterogeneous population is divided into fairly homogeneous groups. Therefore, strata are to be as homogeneous as possible within, but each stratum as different as much as possible from another with respect to the characteristic being measured. We consider that a finite population contains N units and we split this population into H non-overlapping strata. In this case, we can define the sampling design as

$$p(s) = \begin{cases} \prod_{h=1}^{H} \binom{N_h}{n_h}^{-1} & \text{for all } n_h, h = 1, ..., H \\ 0 & \text{otherwise} \end{cases}.$$

The inclusion probability equals $\pi_i = \frac{n_{h_i}}{N_{h_i}}, i \in U_h$, where h_i is the stratum h from which units i comes (Sugden and Smith, 1984). These first-order inclusion probabilities will play a role when constructing pseudo-likelihood function. Thus, the design weight associated with the ith observation in the hth stratum is

$$w_i = \frac{N_{h_i}}{n_{h_i}},$$

where the sum over all design weights over all the strata equals the population total (Lohr, 2010).

2.1.2 Cluster Sampling

Cluster sampling is a standard sampling design tool for large complex surveys. Cluster sampling is utilized because it is typically more cost effective and more convenient to sample in clusters than in the population at random. Cluster samples are broadly applied in virtually all large surveys executed by governments, commercial businesses or academic institutions, due to enormous cost savings (Scheaffer et al., 2011). A cluster, like a stratum, is defined as a grouping of the members of the population. Considering stratified sampling, for optimal precision, individual elements within each stratum must be as homogeneous as possible, but each stratum must contrast as much a possible from other strata in regard to the characteristic being measured. Clusters bear a superficial resemblance to strata. Both techniques involve the random selection of the sampling units. The selection process, though, is vastly dissimilar in the two methods. In a stratified random sample, observation units within each stratum are selected randomly. In a cluster sample, the clusters, PSU's are randomly chosen from the population of all clusters. Therefore, the elements observed are the SSU's within the clusters. For further specifics, see Horvitz and Thompson (1952).

Cluster sampling breaks up the population into subgroups called clusters. It is then determined which (all or some) of the units in each cluster can be included in the sample. One-stage cluster sampling is when all the units in a sampled cluster are incorporated in the sample. Under this method, the clusters are referred to as primary sampling units (PSU's). Two-stage cluster sampling is when the units in a selected PSU are sub-sampled. Those units are referred to as secondary sampling units (SSU's) (Lohr, 2010). Considering a population of N_c non-overlapping clusters. let M_i denote the number of population units in ith selected cluster (cluster size). Assuming that the number of clusters selected in the sample is n_c and let m_i denote the number of observations to be sampled from each of the chosen clusters. Consider one-stage cluster sampling where clusters are chosen from the population without further sampling from the selected clusters. Thus, in one-stage cluster sampling, $M_i = m_i$. In this case, the inclusion probability for the ith primary sampling

unit equals

$$\pi_i = \frac{n_c}{N_c}, i = 1, ..., n$$

where N_c denotes the number of clusters in the population and n_c is the number of sampled clusters, respectively. Thus, the sampling weights under one-stage cluster sampling is given by

$$w_i = {\pi_i}^{-1} = \frac{N_c}{n_c}.$$

It is important to note that if the secondary sampling units within a cluster are too similar, measuring all the units in the cluster is not beneficial and does not interject any additional information to the sample. Since the variability within a cluster is typically lower than the variability between clusters, it is more valuable to pull more clusters and then procure a random sample of units from each sampled cluster for a given sample size. The final method is the two-stage cluster sampling. In this approach, a sample of clusters is selected at the first stage. Afterward, a sample of units from each sampled cluster is chosen at the second stage (Lohr, 2010). However, with the two-stage sampling cluster the inclusion probability of the jth observation given that the ith cluster has been selected is equal to $\pi_{j|i} = \frac{m_i}{M_i}$. Thus, the overall inclusion probability under two-stage cluster sampling is given by $\pi_{ij} = \pi_i \cdot \pi_{j|i} = \frac{n_c}{N_c} \cdot \frac{m_i}{M_i}$, where $i = 1, \ldots, n_c$ and $j = 1, \ldots, m_i$ (Lohr, 2010). Finally, the sampling weights in this case are given by $w_i = \frac{N_c}{n_c} \cdot \frac{M_i}{m_i}$.

2.1.3 Complex Sampling

The The definition of a complex sample (CS) is a stratified multistage cluster sample. The process of selecting a CS begins by dividing the population into non-overlapping subgroups called strata. Recall the previous exposition of stratified random sampling. The stratification process ensures that all strata in the population are represented in the final sample. Next, each stratum is divided into relevant clusters from which a predetermined number is selected. These first-stage clusters are termed primary sampling units (PSU's).

To facilitate variance estimation, it is vital to ensure that no less than two PSU's are selected per stratum. Each of the selected PSUs is then divided again into smaller clusters. A predetermined number is then chosen from those clusters. These second-stage clusters are called secondary sampling units (SSU's). Notice that the PSU's must be stratified before the SSU's are developed and selected. One continues in this manner until the population units of interest are obtained and thus selected for the sampling. The final stage units are called ultimate sampling units (USU's).

As an example of CS, a stratified two-stage cluster sample design was considered. Assuming that a finite population has been stratified into H strata, then the sample is drawn from each stratum in the population. Assume stratum h was divided into N_h PSU's of which n_h has been sampled, $h = 1, ..., H$ with equal probability. It follows that the selection probability of the ith PSU in the hth stratum, π_{hi}, is given by

$$\pi_{hi} = \frac{n_{hi}}{N_{hi}}.$$

Let the ith sampled PSU be clustered into M_i SSU's of which m_j are sampled with equal probability, $i = 1, ..., n_h$. The selection probability of the jth SSU providing the ith PSU in the hth stratum has been selected, $\pi_{j|i}$, is defined as $\pi_{j|hi} = \frac{m_i}{M_i}$. Lastly, the inclusion probability of the jth SSU in the ith PSU of the hth stratum is calculated as

$$\pi_{ij} = \Big(\frac{n_{hi}}{N_{hi}}\Big)\Big(\frac{m_{hi}}{M_{hi}}\Big), h = 1, ..., H, i = 1, ..., n_h, j = 1, ..., m_{hi}.$$

Consequently, the overall sampling weight is given by

$$w_{ij} = \Big(\frac{N_{hi}}{n_{hi}}\Big)\Big(\frac{M_i}{m_i}\Big),$$

(Lohr, 2010). When conducting the inference about the mixture models under the complex sampling designs, the sampling weights are incorporated in the inference to construct

pseudo-likelihood functions in later sections.

2.2 Gaussian Mixture Models

The density of a one-dimensional random variable can be approximated by a weighted sum of some Gaussian densities

$$g(\boldsymbol{x}_i; \boldsymbol{\Psi}) = \sum_{k=1}^{K} \alpha_k \, \phi(\boldsymbol{x}_i; \mu_k, \sigma_k^2), \tag{2.1}$$

where $\phi(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\sigma}_k^2)$ is a Gaussian density with mean μ_k and variance σ_k^2, and $\alpha_k, k = 1, ..., K$ are the positive mixing proportions that satisfy $\sum_{k=1}^{K} \alpha_k = 1$, then, the entire parameter vector is defined as $\boldsymbol{\Psi} = \{\alpha_1, ..., \alpha_{K-1}, \mu_1, ..., \mu_K, \sigma_1^2, ..., \sigma_K^2\}$. Our goal is to estimate the vector of parameters $\boldsymbol{\Psi}$ which can be conveniently estimated by maximum likelihood via the EM algorithm. Let $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ be a sample of observations from $g(\boldsymbol{x}_i; \boldsymbol{\Psi})$, the log-likelihood function of $\boldsymbol{\Psi}$ is given by

$$\ell(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \alpha_k \, \phi(\boldsymbol{x}_i; \mu_k, \sigma_k^2). \tag{2.2}$$

Now, let Z_{ik} be the indicator variable which takes a value of 1 if the ith observation arises from the kth component and zero otherwise. Then the complete-data log-likelihood function incorporate this indicator random variable and is given by

$$\ell_c(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \sum_{k=1}^{K} I(Z_{ik} = 1) \left\{ \log \alpha_k + \log \phi(\boldsymbol{x}_i; \mu_k, \sigma_k^2) \right\}. \tag{2.3}$$

At the tth iteration of the E-step, we take the conditional expectation of ℓ_c given the previous step parameter estimates $\boldsymbol{\Psi}^{(t-1)}$ and data. These in turn results in the computation of

the posterior probabilities

$$\tau_{ik}^{(t)} = \frac{\alpha_k^{(t-1)}\phi(\boldsymbol{x}_i; \mu_k^{(t-1)}, \sigma; 2(t-1)}{k)} \sum_{k'}^{K} \alpha_{k'}^{(t-1)} \phi(\boldsymbol{x}_i; \mu_{k'}^{(t-1)}, \sigma_{k'}^{2\,(t-1)}), \tag{2.4}$$

for $i = \{1, \ldots, n\}$ and $k \in \{1, \ldots, K\}$. At the M-step of the (t)th iteration, we maximize the conditional expectation of the complete-data log-likelihood function with respect to $\boldsymbol{\Psi}$. This function is commonly known as the Q-function and is given by

$$Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(t)}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik} \left\{ \log \alpha_k^{(t-1)} + \log \phi(\boldsymbol{x}_i; \mu_k, \sigma_k^2) \right\}. \tag{2.5}$$

At the (t)th iteration of the M-step, the Q-function is maximized with respect to $\boldsymbol{\Psi}$. For the Gaussian mixture model the closed form solutions are as follows

$$\alpha_k^{(t)} = \frac{\sum_{i=1}^{n} \tau_{ik}^{(t)}}{\sum_{k=1}^{K} \sum_{i=1}^{n} \tau_{ik}^{(t)}}, \tag{2.6}$$

$$\mu_k^{(t)} = \frac{\sum_{i=1}^{n} \tau_{ik}^{(t)} \boldsymbol{x}_i}{\sum_{i=1}^{n} \tau_{ik}^{(t)}}, \text{ and} \tag{2.7}$$

$$\sigma_k^{2(t)} = \frac{\sum_{i=1}^{n} \tau_{ik}^{(t)} \left(\boldsymbol{x}_i - \mu_k^{(t)}\right)^2}{\sum_{i=1}^{n} \tau_{ik}^{(t)}}. \tag{2.8}$$

Note that the above equations are similar to solutions of the maximum likelihood estimates of the mean and variance of a normal distribution except that they are weighted by the posterior probability from the E-step. The E- and M-steps are iterated until convergence criterion is fulfilled. The criterion used in this paper is the relative difference between consecutive log likelihood values which is given by

$$\frac{\ell(\boldsymbol{\Psi}^{(t)}; \mathbf{x}) - \ell(\boldsymbol{\Psi}^{(t-1)}; \mathbf{x})}{|\ell(\boldsymbol{\Psi}^{(t-1)}; \mathbf{x})|} < 10^{-8},$$

where $\ell(\boldsymbol{\Psi})$ is the log likelihood value evaluated at $\boldsymbol{\Psi}$. We will refer to this modeling approach as the unweighted approach.

2.2.1 Pseudo-Maximum Likelihood Estimation of Gaussian Mixture Models

Assuming a data set of $\{(\boldsymbol{x}_i, w_i); i \in s\}$, where w_i is the sampling weights of n units selected from a finite population of size N under some complex survey design. The models that are frequently used to fit survey data are gathered with complex sampling designs. However, if such a design is considered, then standard maximum likelihood estimators are usually biased. Such a scenario can be avoided using the approximate, or Pseudo-Maximum Likelihood (PML) approach as proposed by Skinner et al. (1989) and described by Pfeffermann (1993), and Chambers and Skinner (2003). We propose a probability weighted estimation procedure for finite mixture models which eliminates the bias estimates that occur when ignoring the sampling design. The reciprocals of the inclusion probabilities, $w_i = \frac{1}{\pi_i}$, at each sampling stage are used to weight the log likelihood function. Then, the Pseudo complete-data log-likelihood function is given by

$$\ell_{pc}(\boldsymbol{\Psi}) = \sum_{i=1}^{n} w_i \sum_{k=1}^{K} I(Z_{ik} = k)[\log \alpha_k + \log \phi(\boldsymbol{x}_i; \mu_k, \sigma_k^2)],$$

and since the sampling weight w_i does not have any effect on the posterior probabilities τ_{ik} the E-step is the same as in the unweighted approach as given in Equation 2.4. The modified Q-function is given by

$$Q_{pw}(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(t)}) = \sum_{i=1}^{n} w_i \sum_{k=1}^{K} \tau_{ik}\{\log \alpha_k - \frac{n}{2}\log(2\pi\sigma_k^2) - \frac{(\boldsymbol{x}_i - \mu_k)^2}{2\sigma_k^2}\}. \tag{2.9}$$

We call the function in Equation 2.9 as the weighted Q-function and is denoted by Q_{pw}. At the (t)th iteration of the M-step, the Q_{pw}-function is maximized with respect to $\boldsymbol{\Psi}$. For the

Gaussian mixture model the closed form solutions are as follows

$$\alpha_k^{(t)} = \frac{\sum_{i=1}^{n} w_i \tau_{ik}^{(t)}}{\sum_{k=1}^{K}\sum_{i=1}^{n} w_i \tau_{ik}^{(t)}}, \tag{2.10}$$

$$\mu_k^{(t)} = \frac{\sum_{i=1}^{n} w_i \tau_{ik}^{(t)} \boldsymbol{x}_i}{\sum_{i=1}^{n} w_i \tau_{ik}^{(t)}}, \tag{2.11}$$

$$\sigma_k^{2(t)} = \frac{\sum_{i=1}^{n} w_i \tau_{ik}^{(t)} \left(\boldsymbol{x}_i - \mu_k^{(t)}\right)^2}{\sum_{i=1}^{n} w_i \tau_{ik}^{(t)}}. \tag{2.12}$$

Note here that the above solutions in Equations 2.10-2.12 are similar to the usual Gaussian mixture M-step solutions given in Equations 2.6-2.8 except they are pre-multiplied by the sampling weights.

2.2.2 Multivariate Gaussian Finite Mixture Models

In the multivariate Gaussian mixture model, the density of a d-dimensional random vector $\mathbf{X}$ is given by

$$g(\mathbf{X}; \boldsymbol{\Psi}) = \sum_{k=1}^{K} \alpha_k \, \phi_k(\mathbf{X}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{2.13}$$

where $\phi_k(\mathbf{X}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ is the kth component Gaussian density with $d \times 1$ mean vector $\boldsymbol{\mu}_k$ and a $d \times d$ covariance matrix, $\boldsymbol{\Sigma}_k$. $\alpha_k, k = 1, ..., K$, are the mixing probabilities that satisfy the constraints: $0 < \alpha_k \leq 1$ and $\sum_{k=1}^{K} \alpha_k = 1$. Following the discussion in Section 2.2, the Q_{pw}-function for the multivariate Gaussian mixture case will be:

$$Q_{pw} = \sum_{i=1}^{n} w_i \left[\sum_{k=1}^{K} \tau_{ik} \log \alpha_k - \frac{p}{2} \sum_{k=1}^{K} \tau_{ik} \log \left(2\pi |\, \boldsymbol{\Sigma}_k| \right) - \frac{1}{2} \sum_{k=1}^{K} \tau_{ik} \left(\mathbf{X} - \boldsymbol{\mu}_k\right)^\top \boldsymbol{\Sigma}_k^{-1} \left(\mathbf{X} - \boldsymbol{\mu}_k\right) \right].$$

Given the Q_{pw}-function, at the (t)th iteration of the M-step for multivariate normal mixture model the closed from of the component means $\boldsymbol{\mu}_k$ and components-covariance matrices $\boldsymbol{\Sigma}_k$ are given by

$$\boldsymbol{\mu}_k^{(t)} = \frac{\sum_{i=1}^{n} w_i \, \tau_{ik}^{(t)} \mathbf{X}}{\sum_{i=1}^{n} w_i \, \tau_{ik}^{(t)}}, \tag{2.14}$$

$$\boldsymbol{\Sigma}_k^{(t)} = \frac{\sum_{i=1}^{n} w_i \, \tau_{ik}^{(t)} \big(\mathbf{X} - \boldsymbol{\mu}_k^{(t)}\big)\big(\mathbf{X} - \boldsymbol{\mu}_k^{(t)}\big)^\top}{\sum_{i=1}^{n} w_i \, \tau_{ik}^{(t)}}. \tag{2.15}$$

The M-step closed form solution for the mixing proportions will be the same as in Equation 2.10.

2.3 Finite Mixture of Gaussian Regression Model

Suppose a random sample $\{(\boldsymbol{x}_i, \mathbf{y}_i),\ i = 1, ..., n\}$ of independent identically distributed (IID) observations is drawn from a finite mixture of normal regression model. In this case, explanatory variables $\boldsymbol{x}_i$ are collected for each observation $\mathbf{y}_i$. Then, the probability distribution function is given by

$$g(\mathbf{y}_i; \boldsymbol{x}_i, \boldsymbol{\Psi}) = \sum_{k=1}^{K} \alpha_k \, \phi(\mathbf{y}_i; \boldsymbol{x}_i \boldsymbol{\beta}_k, \sigma_k^2), \tag{2.16}$$

where K is the total number of mixture regression components, $\phi(\mathbf{y}_i; \boldsymbol{x}_i \boldsymbol{\beta}_k, \sigma_k^2)$ is a Gaussian density function of the kth component with mean $\boldsymbol{x}_i \boldsymbol{\beta}_k$ and variance σ_k^2. The mixing proportions, $\alpha_k, k = 1, ..., K$ have the following restrictions: $0 < \alpha_k \leq 1$ and $\sum_{k=1}^{K} \alpha_k = 1$. Therefore, the parameter vector $\boldsymbol{\Psi} = \{\alpha_1, ..., \alpha_{K-1}, \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_K, \sigma_1^2, ..., \sigma_K^2\}$, where $\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_K, \sigma_1^2, ..., \sigma_K^2$ are the component specific regressions coefficients and variances, respectively. The common goal of statistical inference in this setting is to estimate the parameters of the model. Below we describe two estimation procedures. The first one is the traditional maximum likelihood approach which we will refer as the 'unweighted MLE' and the second one is a pseudo-maximum likelihood approach which we call the 'weighted MLE'. We assume that K is unknown, and regard it as a parameter, when performing model fitting. The matter of how best to select an appropriate K is considered as

part of our model fit and model selection.

2.3.1 Unweighted Maximum Likelihood Approach

In this case, estimation of the parameters is typically performed through the maximum likelihood approach. The log-likelihood function is given by

$$\ell(\mathbf{\Psi}) = \sum_{i=1}^{n} \log \left\{ \sum_{k=1}^{K} \alpha_k \, \phi(\mathbf{y}_i; \boldsymbol{x}_i \boldsymbol{\beta}_k, \sigma_k^2) \right\}. \tag{2.17}$$

Due to the inconvenient form of $\ell(\mathbf{\Psi})$ in Equation 2.17, the expectation maximization algorithm (Dempster et al., 1977), which is based on a complete-data log-likelihood function, is employed. The complete-data setup is given IID samples from $g(y_i; \boldsymbol{x}_i, \mathbf{\Psi})$; we define the latent variable Z_{ik} such that

$$Z_{ik} = \begin{cases} 1 & \text{if the } i\text{th observation} \in k\text{th component} \\ 0 & \text{otherwise} \end{cases}.$$

Then, we can write the complete-data log-likelihood function as

$$\ell_c(\mathbf{\Psi}) = \sum_{i=1}^{n} \sum_{k=1}^{K} I(Z_{ik} = 1) \left\{ \log \alpha_k + \log \phi(\mathbf{y}_i; \boldsymbol{x}_i \boldsymbol{\beta}_k, \sigma_k^2) \right\}. \tag{2.18}$$

The EM-algorithm is an iterative procedure of two steps, the Expectation (E) step, and the Maximization (M) step. At the E-step, we calculate the conditional expectation of the complete-data log-likelihood function given the observed data, $E\left(\ell_c(\mathbf{\Psi}) | \boldsymbol{y}, \boldsymbol{X}\right)$, which simplifies to

$$E\Big(I(Z_{ik} = 1) | \mathbf{y}_i, \boldsymbol{x}_i, \mathbf{\Psi}^{(t-1)}\Big) = Pr(Z_{ik} = 1 | \mathbf{y}_i, \boldsymbol{x}_i, \mathbf{\Psi}^{(t-1)}).$$

This posterior probability will be denoted as τ_{ik}. The expression of τ_{ik} at the (t)th iteration of the E-step is given by

$$\tau_{ik}^{(t)} = \frac{\alpha_k^{(t-1)} \phi\left(y_i; \boldsymbol{x}_i \boldsymbol{\beta}_k^{(t-1)}, \sigma_k^{2(t-1)}\right)}{\sum_{k'=1}^{K} \alpha_{k'}^{(t-1)} \phi\left(y_i; \boldsymbol{x}_i \boldsymbol{\beta}_{k'}^{(t-1)}, \sigma_{k'}^{2\,(t-1)}\right)}.$$

At the M-step of the (t)th iteration, we maximize the conditional expectation of the complete-data log-likelihood function commonly known as the Q-function given by

$$Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(t)}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik}^{(t)} \left\{ \log \alpha_k + \log \phi(\mathbf{y}_i; \boldsymbol{x}_i \boldsymbol{\beta}_k, \sigma_k^2) \right\}. \tag{2.19}$$

The two steps are iterated until a predetermined convergence criterion is met. For a simple linear regression model, $\mathbf{y}_i = \beta_{k0} + \beta_{k1} \boldsymbol{x}_i + \epsilon_{ik}$, where $\mathbf{y}_i$ is the response variable value, $\boldsymbol{x}_i$ denotes a single explanatory variable and $\epsilon_{ik} \sim N(0, \sigma_k^2)$, Equation 2.19 can be written as

$$Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(t)}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik}^{(t)} \left\{ \log \alpha_k - \frac{n}{2} \log(2\pi\sigma_k^2) - \frac{(\mathbf{y}_i - \beta_{k0} - \beta_{k1} \boldsymbol{x}_i)^2}{2\sigma_k^2} \right\}, \tag{2.20}$$

and the closed form solutions for parameters at (t)th iteration of the M-step are given by

$$\alpha_k^{(t)} = \frac{\sum_{i=1}^{n} \tau_{ik}^{(t)}}{\sum_{k=1}^{K} \sum_{i=1}^{n} \tau_{ik}^{(t)}}, \tag{2.21}$$

$$\beta_{k1}^{(t)} = \frac{\sum_{i=1}^{n} \tau_{ik}^{(t)} \sum_{i=1}^{n} \tau_{ik}^{(t)} \boldsymbol{x}_i \mathbf{y}_i - \sum_{i=1}^{n} \tau_{ik}^{(t)} \boldsymbol{x}_i \sum_{i=1}^{n} \tau_{ik}^{(t)} \mathbf{y}_i}{\sum_{i=1}^{n} \tau_{ik}^{(t)} \sum_{i=1}^{n} \tau_{ik}^{(t)} \boldsymbol{x}_i^2 - (\sum_{i=1}^{n} \tau_{ik}^{(t)} \boldsymbol{x}_i)^2}, \tag{2.22}$$

$$\beta_{k0}^{(t)} = \frac{\sum_{i=1}^{n} \tau_{ik}^{(t)} \mathbf{y}_i}{\sum_{i=1}^{n} \tau_{ik}^{(t)}} - \beta_{k1}^{(t)} \frac{\sum_{i=1}^{n} \tau_{ik}^{(t)} \boldsymbol{x}_i}{\sum_{i=1}^{n} \tau_{ik}^{(t)}}, \text{ and} \tag{2.23}$$

$$\sigma_k^{2(t)} = \frac{\sum_{i=1}^{n} \tau_{ik}^{(t)} \left(\mathbf{y}_i - \beta_{k0}^{(t)} - \beta_{k1}^{(t)} \boldsymbol{x}_i\right)^2}{\sum_{i=1}^{n} \tau_{ik}^{(t)}}. \tag{2.24}$$

Note that the Equations 2.22–2.24 are similar to least squares simple linear regression estimates except that they are weighted by the posterior probability from E-step.

2.3.2 Pseudo-Maximum Likelihood Estimation of Mixture Gaussian Regression

We assume the given data set of observations $\{(\boldsymbol{x}_i, \mathbf{y}_i, w_i); i \in s\}$, where w_i is the sampling weights. In this case, we selected a sample of size n units from a finite population of size N under some complex survey design. If such a design is considered, then standard maximum likelihood estimators are usually biased Wedel et al. (1998). Such a scenario can be avoided using the approximate, or pseudo-maximum Likelihood (PML) approach as proposed by Skinner et al. (1989) and described by Pfeffermann (1993) and Chambers and Skinner (2003). We propose a weighted estimation procedure for finite mixture models which minimizes the bias in parameter estimates that occur when the sampling design is not taken into consideration. This is done by incorporating the sampling weights, w_i to the complete data log- pseudo likelihood function. Then the modified Q-function is given by

$$Q_{pw}(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(t)}) = \sum_{i=1}^{n} w_i \sum_{k=1}^{K} \tau_{ik} \{\log \alpha_k - \frac{n}{2} \log(2\pi\sigma_k^2) - \frac{(\mathbf{y}_i - \beta_{k0} - \beta_{k1}\boldsymbol{x}_i)^2}{2\sigma_k^2}\}. \quad (2.25)$$

We refer the function in Equation 2.25 as the weighted Q-function and is denoted by Q_w. At the M-step of the (t)th iteration, the Q_w-function is maximized with respect to $\boldsymbol{\Psi}$. For the simple Gaussian mixture regression model the closed form solutions are as follows

$$\alpha_k^{(t)} = \frac{\sum_{i=1}^{n} w_i \tau_{ik}^{(t)}}{\sum_{k=1}^{K} \sum_{i=1}^{n} w_i \tau_{ik}^{(t)}}, \quad (2.26)$$

$$\beta_{k1}^{(t)} = \frac{\sum_{i=1}^{n} w_i \tau_{ik}^{(t)} \sum_{i=1}^{n} w_i \tau_{ik}^{(t)} \boldsymbol{x}_i \mathbf{y}_i - \sum_{i=1}^{n} w_i \tau_{ik}^{(t)} \boldsymbol{x}_i \sum_{i=1}^{n} w_i \tau_{ik}^{(t)} \mathbf{y}_i}{\sum_{i=1}^{n} w_i \tau_{ik}^{(t)} \sum_{i=1}^{n} w_i \tau_{ik}^{(t)} \boldsymbol{x}_i^2 - \left(\sum_{i=1}^{n} w_i \tau_{ik}^{(t)} \boldsymbol{x}_i\right)^2}, \quad (2.27)$$

$$\beta_{k0}^{(t)} = \frac{\sum_{i=1}^{n} w_i \tau_{ik}^{(t)} \mathbf{y}_i}{\sum_{i=1}^{n} w_i \tau_{ik}^{(t)}} - \beta_{k1}^{(t)} \frac{\sum_{i=1}^{n} w_i \tau_{ik}^{(t)} \boldsymbol{x}_i}{\sum_{i=1}^{n} w_i \tau_{ik}^{(t)}}, \text{ and} \quad (2.28)$$

$$\sigma_k^{2(t)} = \frac{\sum_{i=1}^{n} w_i \tau_{ik}^{(t)} \left(\mathbf{y}_i - \beta_{k0}^{(t)} - \beta_{k1}^{(t)} \boldsymbol{x}_i\right)^2}{\sum_{i=1}^{n} w_i \tau_{ik}^{(t)}}. \tag{2.29}$$

Note here that the update equations in 2.26–2.29 are similar to 2.21–2.24 except the weights are incorporated.

2.3.3 Matrix Approach for the Mixture of Gaussian Multiple Regression Models

We can extend the mixture of simple linear regression model to multiple linear regression model. This can be done using matrix notation as follows

$$\boldsymbol{\beta}_k^{(t)} = \left(\mathbf{X}^\top \mathbf{W}_k^{(t)} \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{W}_k^{(t)} \mathbf{y}, \tag{2.30}$$

where $\mathbf{X}$ is an $n \times (p+1)$ matrix containing unity for intercept and predictors, $\mathbf{W}_k^{(t)}$ is a $n \times n$ diagonal matrix with entries $w_i \times \tau_{ik}^{(t)}$, $\mathbf{y}$ is a $n \times 1$ vector of response variable, and

$$\sigma_k^{2(t)} = \frac{\left\| \mathbf{W}_k^{1/2(t)} \left(\mathbf{y} - \mathbf{X}\, \boldsymbol{\beta}_k^{(t)}\right) \right\|^2}{\mathrm{tr}\left(\mathbf{W}_k^{(t)}\right)}, \tag{2.31}$$

where $\|\mathbf{A}\| = \mathbf{A}^\top \mathbf{A}$ with $\top$ denoting a matrix transpose and $\mathrm{tr}(\mathbf{A})$ means the trace of the matrix $\mathbf{A}$. Equations 2.30 and 2.31 can be used as update equations at at the (t)th iteration of the M-step. The same equation as given in equation 2.26 is used to update mixing proportions.

2.4 Computational Strategies

In this section, we describe some computational strategies that have been used in fitting the proposed model. Initialization is a key step in fitting mixture models to data via the EM algorithm (Baudry and Celeux, 2015). In the simulation study, we considered two strategies

for choosing initial values of parameters. In the first simulation study, we compare the weighted and unweighted models. For this, the true values of parameters were used as the starting values. This will allow for comparing without confounding the issues associated with initialization. In the second simulation study is conducted to assess the validity of Bayesian Information Criterion (BIC) as model selection criterion. For this, we used *Rnd-EM* (Maitra, 2009) to choose initial values. In this initialization method, first random points are selected as seeds and the Euclidean distance is used to assign observations to centers. This is repeated for some fixed number of times. The solution that yields the highest likelihood value is then used for initializing the EM-algorithm. *Rnd-EM* tends to work well if the number of components is not large (Michael and Melnykov, 2016). *Rnd-EM* is used to initialize the EM algorithm for the real data analysis. In the EM algorithm, the E-step and M-step are iterated until a convergence criterion is met. In this dissertation, the algorithm is stopped when the absolute relative change in the likelihood given by

$$\frac{\ell_p(\mathbf{\Psi}^{(t)};\mathbf{y},\mathbf{x}) - \ell_p(\mathbf{\Psi}^{(t-1)};\mathbf{y},\mathbf{x})}{|\ell_p(\mathbf{\Psi}^{t-1};\mathbf{y},\mathbf{x})|} < 10^{-8}.$$

In the real dataset analysis, we used the BIC (Schwarz et al., 1978) to select the optimal number of components. In this dissertation, BIC will be calculated as $BIC(\hat{\mathbf{\Psi}}) = -2\ell_p(\hat{\mathbf{\Psi}}) + M\log n$, where $\ell_p(\hat{\mathbf{\Psi}})$ and M represent the maximized likelihood value for a given K and the number of parameters in the fitted model, respectively. For mixtures of normal regression, $M = (K-1) + K(p+1) + K$, where p represents the number of the predictor variables. The model with lowest BIC value is the best model for a given dataset.

2.5 Identifiability

Identifiability of a given model is one of the major requirements for any model to be meaningful. It is defined for any two parameter vectors $\mathbf{\Psi} \neq \mathbf{\Psi}'$, the respective model $f(\boldsymbol{x};\mathbf{\Psi})$ must be different from $f(\boldsymbol{x};\mathbf{\Psi}')$ for any random vector $\boldsymbol{x}$. The identification issue

for the finite mixture linear model has been and continues to be studied. In general, in the mixture regression model setting, there are two kinds of identification problems that are common. One of them is label switching, and the other is overfitting. The label switching occurs when switching the labels of any two different components does not change the distribution of the response variable at all. Overfitting is a more fundamental lack of identifiability, and it leads to empty components or components with equal parameters. This kind of unidentifiablity can be avoided by restricting the prior mixing ratios to be greater than zero, and the component with specific parameters are different (Leisch, 2004). In this paper, to prevent overfitting, mixing proportions have been restricted to be greater than a particular threshold.

On a similar note, the identifiability of a mixture of regression models depends on the distribution of the response variable. Particularly in this setting, Hennig (2000) pointed out that identifiability issues may arise if there are solely a restricted range of values for covariates and additionally if there is a restricted info per person accessible. Such problems might occur in applications where covariates are generally categorical variables for example race and gender (Grün and Leisch, 2004). As per Hennig (2000), the mixtures of linear regression models with Gaussian random errors are identifiable if the number of components K is smaller than the minimal number of hyperplanes necessary to cover all covariate points. In this dissertation, we mainly focus on continuous response and covariates, but in general, one needs to be cautious of the results obtained.

2.6 Model Comparison

For comparing the weighted and unweighted models, the variance-bias components of MSE are used. The MSE is obtained from the B replications as $MSE(\widehat{\psi_j}) = \frac{1}{B}\sum_{b=1}^{B}(\psi_{jb} - \hat{\psi}_{jb})^2$, where ψ_{jb} and $\hat{\psi}_{jb}$, are the true and estimated parameter, respectively. The variance and bias components are given as $Var(\hat{\psi}_j) = \frac{1}{B}\sum_{b=1}^{B}(\hat{\psi}_{jb} - \bar{\hat{\psi}}_j)^2$ and $Bias(\hat{\psi}_j) =$

$(\bar{\hat{\psi}}_j - \psi_j)^2$, where $\bar{\hat{\psi}}_j = \frac{1}{B}\sum_{b=1}^{B}\hat{\psi}_{jb}$, respectively. Note that, $MSE(\hat{\psi}_j) = Var(\hat{\psi}_j) + Bias^2(\hat{\psi}_j)$. In our setting, $\mathbf{\Psi} = \{\psi_j\}_{j=1}^{M}$, where M is the number of parameters in $\mathbf{\Psi} = \{\alpha_1, ..., \alpha_{K-1}, \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_K, \sigma_1^2, ..., \sigma_K^2\}$, and each element is represented by ψ_j.

In this dissertation, percent contribution, $\mathcal{R}$, is used to compute the relative contribution of a given quantity to a total amount and is calculated as: $\mathcal{R} = \frac{\theta_1}{\theta_1+\theta_2}$, where θ_1 and θ_2 are the two quantities calculated. We will use $\mathcal{R}$ to find out how much percentage contribution take place for two quantities we are trying to compare. Note that, this index will range between 0 and 1 and if both quantities contribute equally to the total amount then R will be equal to 0.5. Values below 0.5 indicate lower percent contribution of θ_1 as compared to θ_2 and values above 0.5 will indicate higher percent contribution of θ_1 to the total $\theta_1 + \theta_2$. In the simulation study, the MSE, and its bias and variance components will be used to compute $\mathcal{R}$. This will be use to compare the performance of the weighted model with the unweighted model. If any of MSE components, bias or the variance of were equal of the compared models; then $\mathcal{R}$ will be equal to 0.5. Also, we have formulated this measurement such that the components of the MSE of the unweighted approach will be on the numerator of the fraction, thus for any of MSE components if $\mathcal{R}$ was less than 0.5 then the performance of the unweighted model will be better than the weighted model. On the other hand, if $\mathcal{R}$ was greater than 0.5, then the performance of the weighted model will be better than the unweighted model.

2.7 Variability Assessment

One common practice in making statistical inference after finding the point estimates of the parameter is to obtain the corresponding standard error of the parameter estimates. The standard errors provide a useful measure of the accuracy of the point estimates being reported. Also, we can use the standard errors when they are available in asymptotic normal theory to obtain the approximate confidence intervals for the parameters of interest or

perform hypothesis tests. This section is concerned with calculation of the standard errors of the estimated parameters obtained when fitting a weighted Gaussian mixture of regression model by maximum likelihood via the EM algorithm. In general statistics theory, the covariance matrix of the MLE, $\widehat{\boldsymbol{\Psi}}$, is determined by using the inverted Fisher Information matrix, $\mathbf{I}_n(\widehat{\boldsymbol{\Psi}})$, where $\mathbf{I}_n(\boldsymbol{\Psi}) = -\partial^2 \log L(\boldsymbol{\Psi})/\partial\boldsymbol{\Psi}\ \partial\boldsymbol{\Psi}^\top$ (Cramer, 1946; Efron and Hinkley, 1978). Computing the direct second partial derivatives of the likelihood function for multivariate mixtures could be challenging. For the $\widehat{\boldsymbol{\Psi}}$ obtained via the EM-algorithm, there are many ways to finding $\mathbf{I}_n$, which are described by (McLachlan and Peel, 2000). One way to proceed is to assume the case of independent and identically distributed observations. In this case, $\mathbf{I}_n$ is approximated using the empirical observed information matrix, $\mathbf{I}_e$ as proposed and termed by Meilijson (1989). This approximation is given by

$$\mathbf{I}_e(\widehat{\boldsymbol{\Psi}}) = \sum_{i=1}^{n} S(y_i, \boldsymbol{x}_i; \widehat{\boldsymbol{\Psi}}) S^\top(y_i, \boldsymbol{x}_i; \widehat{\boldsymbol{\Psi}}),$$

where $S(y_i, \boldsymbol{x}_i; \widehat{\boldsymbol{\Psi}}) = \frac{\partial \log L_i(\widehat{\boldsymbol{\Psi}}|y_i,\boldsymbol{x}_i)}{\partial\boldsymbol{\Psi}} = E(\frac{\partial \log L_{ci}(\widehat{\boldsymbol{\Psi}}|y_i,\boldsymbol{x}_i)}{\partial\boldsymbol{\Psi}})$, with L_i and L_{ci} denoting the likelihood and complete-data likelihood based on a single observation, respectively. Therefore, with this result we can estimate the Fisher information using partial derivatives of the complete-data log-likelihood functions.

The next thing we need to consider is the asymptotic variance of PML estimates. If the sample selection is ignored, the MLE under common regularity conditions are asymptotically normal (See for *e.g.* Holt et al. (1980); White (1982); Pfeffermann (1993); Lohr (2010)). Similarly, the PML estimators of $\boldsymbol{\Psi}$, are shown to be consistent and asymptotically normal. The regularity conditions required for the PML estimator to be strongly consistent had been provided by (White, 1982). We provide these regularity conditions in the Appendix. A robust estimator of the asymptotic variance is in this situation provided by White (1982) and Royall (1986). In misspecified model setting, Royall (1986) had proposed a robust estimator of the Fisher information for a one parameter problem by

replacing I_n by $I_n(\theta)^2 \sum_{i=1}^{n} \frac{\partial \log L_i(\theta)}{\partial \theta}$. In addition, the PML estimator is a consistent estimator if it satisfies the conditions which are given in the Appendix A.3.1. According to (White, 1982) and (Holt et al., 1980) the PML estimators with additional conditions which are given in Appendix A.3.2, are asymptotically normally distributed. Thus, if all conditions A.3.1–A.3.7 are satisfied, we can define the information and covariance matrices as

$$\mathbf{I}(\widehat{\mathbf{\Psi}}_{PML}) = \Big\{\mathbb{E}\Big(\frac{\partial^2 \ \log f(\boldsymbol{y}_i \mathbf{\Psi})}{\partial \mathbf{\Psi} \ \partial \mathbf{\Psi}^\top}\Big)\Big\},$$

and

$$\Sigma(\hat{\mathbf{\Psi}}_{PML}) = \Big\{\mathbb{E}\Big(\frac{\partial \ \log f(\mathbf{y}_i; \mathbf{\Psi})}{\partial \mathbf{\Psi}} \ \times \ \frac{\partial \ \log f(\mathbf{y}_i; \mathbf{\Psi})}{\partial \mathbf{\Psi}^\top}\Big)\Big\}$$

In addition, when the appropriate inverses exist, we can define

$$Var(\hat{\mathbf{\Psi}}_{PML}) = \mathbf{I}(\hat{\mathbf{\Psi}}_{PML})^{-1} \, \Sigma(\hat{\mathbf{\Psi}}_{PML}) \, \mathbf{I}(\hat{\mathbf{\Psi}}_{PML})^{-1}, \tag{2.32}$$

where $\mathbf{I}(\hat{\mathbf{\Psi}})$ is the observed information matrix, $\Sigma(\hat{\mathbf{\Psi}})$ is based on the cross product of the vector of the first derivatives. For more details see White (1982) and (Royall, 1986). In general, the sample size needs to be reasonably large in the finite mixture model analysis for the asymptotic approximation to standard error to be adequate. Finally, we note that, in the classical theory of statistics, the concept of consistency usually refers to the limiting behavior of a sample statistic as the sample size increased. However, here in design-based analysis, the consistency concept requires that the population size will also be allowed to increase (Smith, 1984) and (Pfeffermann, 1993).

3 SIMULATION STUDIES

In the previous chapter, we introduced the proposed methodology for modeling the finite mixture of regression models to samples drawn from complex survey data. In this chapter, we provide simulation results that illustrate the comparative performances of the weighted and unweighted model using the MSE, variance, and bias. As a consequence, in this chapter, we demonstrate a simulation study to assess the utility of BIC as a selection model criterion.

3.1 Simulation of Stratified Sampling Data

In this section, we explore how to fit the finite mixture normal regression model for samples drawn using a stratified sampling design. We introduced some simulation studies and applications of the proposed model. In the first simulation study, the parameter recovery of the weighted and unweighted model has been evaluated. In the second simulation study, more investigation about the parameter recovery capability of the proposed and usual approaches will be present. In the third simulation study, we assess the performance of the weighted model as a classification tool. A way to select the number of components K consisting of computing a convenient model-based selection criterion across a reasonable range of values for the number of components K and then choosing K associated with the best value of the adopted criterion. In this dissertation, the BIC will be adopted for selecting the optimal number of components for a given dataset.

3.1.1 Simulation 1: Parameter Estimation of Stratified Sampling Data

This simulation study was executed to assess the performance of the maximum likelihood estimates obtained via the unweighted and weighted model in various scenarios.

The criteria used for comparison include: Mean Squared Error (MSE), variance, and bias. In this setting, the true values of parameters were used as the starting values. We considered two configurations of the true regression lines: non-overlapping and overlapping which we call Mixture 1 and Mixture 2, respectively. In the first simulation, we generated a finite population composed of $N = 18000$ observations from a two-component mixture of normal regression model. The finite population consists of two stratum, $\{U_h\}_{h=1}^{2}$, with $\{10000, 8000\}$ observations in each stratum. The vector of parameters $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2)$ used to generate the mixture are reported in Table 3.1. Stratified samples of sizes $n_1 = n_2 = \{100, 250, 500, 1000\}$ are drawn from each stratum. Thus, the total sample sizes of $n = 200, 500, 1000, 2000$ are considered. Therefore, for $n = 1000$, we have $n_1 = 500$ from the first stratum and $n_2 = 500$ from second stratum. For example for Mixture 1, with in each stratum, we use $\alpha_1 = 0.34$ and $\alpha_2 = 0.66$ to determine how many observations will belong to component one and component two, respectively. Figure 3.1 shows sample of size $n = 1000$ observations from the considered models Mixture 1 and Mixture 2. The above setup is repeated for $B = 1000$ replications.

Table 3.1: True parameter values for Mixture 1 and Mixture 2.

$\boldsymbol{\psi}$	α_1	α_2	β_{10}	β_{20}	β_{11}	β_{21}	σ_1^2	σ_2^2
Mixture 1	0.34	0.66	-3	3	1	-2	0.1	0.1
Mixture 2	0.34	0.66	-3	-2	1	-2	0.1	0.1

For each replication, the weighted and unweighted models are fitted and parameter estimates are obtained. The true parameter values are compared with the estimated values using the MSE and its components as given in Section 2.6. Since two different methods have been used to fit the model, it is necessary to evaluate their parameter recovery and to check whether accurate the variability of estimates is yielded. Parameter recovery concerns whether the weighted or unweighted models can recover the generating parameters accurately. If the empirical mean of the estimates across replications is statistically meaningfully different from the generating parameter, the estimator is thought to be bi-

ased. There is also a concern regarding the variability of the estimates across replications. If the variability is practically minor, then a slightly biased estimation is negligible. Table 3.2 provides the MSE and its bias and variance components for varying sample sizes when Mixture1 is considered. The bold values show where the minimum is achieved when comparing the weighted and unweighted models. Looking at the table, the estimates obtained by the weighted model have a smaller bias compared to the estimates obtained by the unweighted model in 21 out of 28 cases. Thus, the estimates obtained by the weighted approach have a smaller bias compared with estimates obtained via the unweighted approach. The weighted model estimates have relatively high variability compared to estimates obtained via the unweighted model in 14 out of 28 cases. The variances of the estimates for both models decrease by increasing the size of a sample.

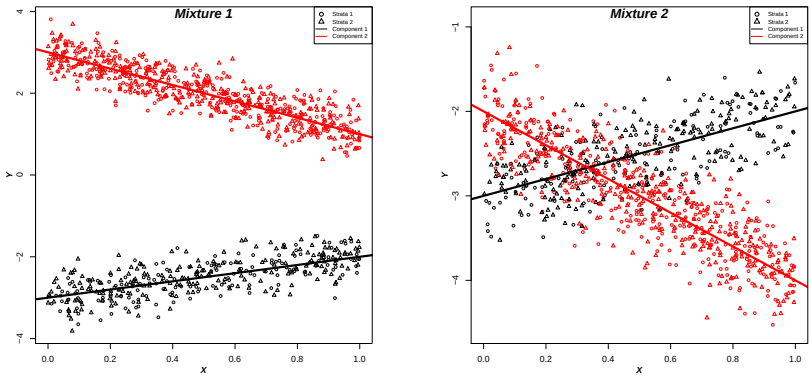


Figure 3.1: Scatter plots of a sample of size $n = 1000$ units. Colors show the two components and plotting characters represent strata. Left plot represents Mixture 1 - non-overlapping components and right plot represents Mixture 2 - overlapping components.

Table 3.2: Mean squared error, bias, and variance of estimated parameters, based on 1000 replications for different sample sizes of the two-component when the Mixture 1 configuration was considered under stratified sampling design. The values reported are $\times 10^{-2}$.

		$\hat{\boldsymbol{\psi}}$	$\hat{\alpha}_1$	$\hat{\beta}_{10}$	$\hat{\beta}_{20}$	$\hat{\beta}_{11}$	$\hat{\beta}_{21}$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$
$n = 200$	MSE	Weighted	**0.1186**	0.5703	**0.2884**	1.7289	0.8935	0.0279	0.0172
		Unweighted	0.1228	**0.5636**	0.2963	**1.7133**	**0.8766**	**0.0274**	**0.0167**
	Bias^2	Weighted	**0.0001**	0.0081	**0.0019**	**0.0272**	**0.0038**	0.0019	0.0010
		Unweighted	0.0034	**0.0076**	0.0027	0.0282	0.0058	**0.0018**	**0.0009**
	Var	Weighted	**0.1185**	0.5622	0.2942	1.7007	0.8886	0.0260	0.0162
		Unweighted	0.1194	**0.5560**	**0.2858**	**1.6861**	**0.8708**	**0.0256**	**0.0158**
$n = 500$	MSE	Weighted	**0.0446**	0.2380	0.1265	0.7559	0.3529	0.0109	0.0061
		Unweighted	0.0492	**0.2332**	**0.1225**	**0.7413**	**0.3478**	**0.0107**	**0.0060**
	Bias^2	Weighted	**0.0001**	**0.0075**	0.0021	0.0427	0.0045	0.0008	**0.0004**
		Unweighted	0.0046	0.0089	**0.0015**	**0.0392**	**0.0032**	**0.0007**	0.0005
	Var	Weighted	**0.0445**	0.2303	0.1234	0.7132	0.3484	0.0101	0.0057
		Unweighted	0.0447	**0.2243**	**0.1211**	**0.7021**	**0.3446**	**0.0100**	**0.0055**
$n = 1000$	MSE	Weighted	**0.0227**	**0.1058**	**0.0567**	**0.3368**	**0.1676**	0.0057	**0.0029**
		Unweighted	0.0249	0.1093	0.0592	0.3451	0.1739	**0.0056**	0.0031
	Bias^2	Weighted	**0.0001**	**0.0067**	**0.0024**	**0.0270**	**0.0025**	**0.0005**	**0.0003**
		Unweighted	0.0024	0.0079	0.0034	0.0295	0.0040	0.0006	0.0004
	Var	Weighted	0.0226	**0.0991**	**0.0543**	**0.3098**	**0.1652**	0.0052	**0.0026**
		Unweighted	**0.0225**	0.1014	0.0559	0.3156	0.1700	**0.0050**	0.0027
$n = 2000$	MSE	Weighted	**0.0096**	**0.0566**	**0.0268**	**0.1842**	**0.0819**	**0.0026**	**0.0014**
		Unweighted	0.0124	0.0586	0.0284	0.1875	0.0855	0.0028	0.0016
	Bias^2	Weighted	**0.0001**	**0.0045**	**0.0029**	**0.0253**	**0.0054**	**0.0004**	**0.0002**
		Unweighted	0.0028	0.0055	0.0041	0.0275	0.0079	0.0005	0.0003
	Var	Weighted	**0.0095**	**0.0522**	**0.0239**	**0.1590**	**0.0766**	**0.0022**	**0.0012**
		Unweighted	0.0096	0.0532	0.0243	0.1602	0.0777	0.0023	0.0013

Table 3.3: Mean squared error, bias, and variance of estimated parameters, based on 1000 replications for different sample sizes of the two-component when the Mixture 2 configuration was considered under stratified sampling design. The values reported are $\times 10^{-2}$.

		$\hat{\boldsymbol{\psi}}$	$\hat{\alpha}_1$	$\hat{\beta}_{10}$	$\hat{\beta}_{20}$	$\hat{\beta}_{11}$	$\hat{\beta}_{21}$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$
$n = 200$	MSE	Weighted	0.1939	**0.3531**	0.2142	**1.2784**	0.7299	**0.0164**	**0.0108**
		Unweighted	**0.1797**	0.3604	**0.2018**	1.2886	**0.7016**	0.0444	0.0221
	Bias^2	Weighted	**0.0009**	**0.0040**	**0.0045**	0.0008	**0.0051**	**0.0001**	**0.0001**
		Unweighted	0.0082	0.0062	0.0046	**0.0001**	0.0058	0.0002	0.0003
	Var	Weighted	0.1930	**0.3491**	0.2097	**1.2776**	0.7248	**0.0163**	**0.0107**
		Unweighted	**0.1715**	0.3542	**0.1972**	1.2885	**0.6956**	0.0442	0.0218
$n = 500$	MSE	Weighted	**0.0646**	0.1354	0.1011	0.5221	0.3135	**0.0078**	**0.0044**
		Unweighted	0.0654	**0.1284**	**0.0986**	**0.4777**	**0.3038**	0.0193	0.0082
	Bias^2	Weighted	**0.0004**	**0.0001**	**0.0035**	0.0012	**0.0018**	**0.0001**	**0.0002**
		Unweighted	0.0017	0.0004	0.0039	0.0001	0.0027	0.0003	0.0003
	Var	Weighted	0.0643	0.1353	0.0976	0.5209	0.3120	**0.0077**	**0.0042**
		Unweighted	**0.0638**	**0.1282**	**0.0949**	**0.4776**	**0.3013**	0.0190	0.0079
$n = 1000$	MSE	Weighted	**0.0214**	0.0692	0.0490	0.2593	0.1533	**0.0036**	**0.0022**
		Unweighted	0.0283	**0.0668**	**0.0472**	**0.2435**	**0.1435**	0.0071	0.0034
	Bias^2	Weighted	**0.0001**	**0.0025**	**0.0038**	**0.0016**	**0.0046**	**0.0001**	**0.0001**
		Unweighted	0.0022	0.0032	0.0039	0.0017	0.0048	0.0002	0.0002
	Var	Weighted	**0.0213**	0.0668	0.0452	0.2577	0.1488	**0.0035**	**0.0021**
		Unweighted	0.0262	**0.0636**	**0.0433**	**0.2418**	**0.1388**	0.0069	0.0032
$n = 2000$	MSE	Weighted	**0.0104**	0.0304	**0.0236**	0.1114	**0.0683**	**0.0019**	**0.0011**
		Unweighted	0.0150	**0.0299**	0.0249	**0.1039**	0.0733	0.0035	0.0017
	Bias^2	Weighted	**0.0002**	**0.0011**	**0.0042**	**0.0005**	**0.0047**	**0.0002**	**0.0001**
		Unweighted	0.0013	0.0018	0.0045	0.0006	0.0054	0.0003	0.0002
	Var	Weighted	**0.0102**	0.0293	**0.0194**	0.1109	**0.0637**	**0.0017**	**0.0010**
		Unweighted	0.0136	**0.0280**	0.0204	**0.1034**	0.0679	0.0032	0.0015

Table 3.3 provides the MSE, the bias, and the variance of the estimated parameters when Mixture 2 is considered. The estimates obtained by the weighted model have lower

bias compared to the estimates obtained by the unweighted model in 26 out of 28 cases, which leads to the conclusion that the estimates obtained by the weighted approach have small bias compared with estimates obtained via the unweighted approach. The weighted model estimates have high variability compared to the unweighted model estimates in only 14 out of 28 cases. However, the variances of the estimates for both models are declined by increasing the size of a sample. Therefore, the estimates obtained via the weighted model for Mixture 2 have a lower bias in about 93% of cases compared by the unweighted model estimates in the same configuration while this percentage to about just decreased to about 78% of instances when Mixture 1 was considered. Hence, we can infer that the weighted model has better performance to reduce the bias of estimated parameters for complicated circumstances.

3.1.2 Simulation 2: Model Comparison of Stratified Sampling Data

To further investigate the parameter recovery capability of both approaches, we will present a diagnosis concerning the results of the first simulation study. This is done to evaluate the impact of the sample size on the parameter recovery and assess the variability associated with the MSE and its components. Based on the results obtained in the previous study, the weighted model has a lower bias in the majority of cases, yet occasionally, the unweighted model estimates have a lower bias compared by those which are obtained via the weighted model. Therefore, the simulation study has not yet determined the general features of the two approaches definitively. Here we considered the Mixture 1 setup; there is a finite population consisting of $\{10000, 8000\}$ observations in each stratum. The vector of parameters is reported in Table 3.1. Stratified samples were drawn from each stratum at different sample sizes, starting with 50 per strata up to 500 with an increment of 50 observations. Thus, the samples that were selected are $n = \{100, 200, 300, \ldots, 1000\}$. Here we replicated $B = 100$ times for each n. These replicates are then used to calculate MSE values and the corresponding bias and variance components. Then, two hundred replications

of the above set up were completed to obtain 200 values of MSE, bias, and variance values for each sample size and parameter under both the weighted and unweighted models. These 200 replicates are then used to calculate the percent contribution index, $\mathcal{R}$, defined in Section 2.6, by setting θ_1 to be the results from the unweighted model and θ_2 to be the results from weighted model. Therefore, if $\mathcal{R}$ is above 0.5, then the weighted model had contributed less to the total MSE, bias or variance. If $\mathcal{R}$ is less than 0.5 then the unweighted model has contributed less to the total MSE, bias or variance.

The results of this analysis can be found in a multiplot provided in Figure 3.2. The top panel of the the figure represents the bias, the middle represents the variance, and the bottom panel represents the MSE. The seven columns correspond to the seven parameters estimated in this study. With in each plot, the x-axis represents the varying sample sizes and y-axis is the $\mathcal{R}$ index. The median values of the index are represented by the black line and the dashed bars indicate ± 1 interquartile range (IQR) values of the index at each sample size. The dashed horizontal line is at 0.5, indicating a threshold for when the two methods perform equally.

Considering the top panel, the median values of the R-index for bias were above 0.5 in all estimated parameters and sample sizes. In the majority of cases, the ± 1 IQR bar of the bias was above the dashed horizontal line except for few cases (estimation of σ^2) where some IQR lines were slightly below the 0.5 line. In case of the mixing proportion $\hat{\alpha}_1$ we noted that on average more than 80% of the total bias was contributed by the unweighted model. Overall, the effect of sample size on bias and its variability was unclear. For three out of four intercept/slope parameter estimates the variability in bias seems to decrease with sample size. In most cases, varying sample sizes did not have a clear trend on median or IQR of the index.

Regarding to the index for variance component presented in the middle panel of Figure 3.2, in five out of seven of the parameters, the median value of R and the ± 1 IQR bars were below the dashed line. The exceptions to this were for the estimates of σ^2. For

both components, the variance seems to be much higher for the unweighted model than the weighted model. In addition, looking at the IQR, we can see that $\mathcal{R}$ associated with variance is much less has shorter bars than the same index for the bias.

Finally, looking at MSE $\mathcal{R}$ index at the bottom panel of Figure 3.2, as MSE is the sum of the bias squared and variance the results shown are reflective of the above two. On most cases, the median value of $\mathcal{R}$ is below the 0.5 threshold line except only the mixing proportion parameter and the variance parameters. Concerning the variability of $\mathcal{R}$ for MSE, we can see that similar to the variance component it has lower variability compared to the same index for the bias. From the three summaries we can conclude that the even if it is unclear which model performs better in terms of MSE, the bias in parameter estimates obtained by using the weighted model is lower than the unweighted model.

3.1.3 Simulation 3: Model Selection of Stratified Sampling Data

In this simulation study, we assess the performance of BIC as a model selection method when using the weighted model as a classification tool. This is using the relationship between finite mixture models and model-based clustering. In model based clustering, each component is associated with a single cluster. Hence, a K-component mixture can be used to identify K homogeneous classes in heterogeneous data. Therefore, we will vary the number of components K that is used to generate the mixture model and assess if BIC is able to retrieve the true K.

The vector of true parameters $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2)$ used to generate the mixtures are shown in Table 3.4. In this setup, samples were drawn using stratified sampling design from the finite population by selecting simple random samples without replacement of size $n_h = 500$ from each stratum, $h = 1, ..., H$. Figure 3.3 shows the stratified samples which were selected in four experiments. In Mixture 1, we generated a finite population containing two strata with $10000, 8000$ observations in each stratum. In this case, there are two components $(K = 2)$, and the total of $n = 1000$ observations was selected. In Mixture 2, we generated

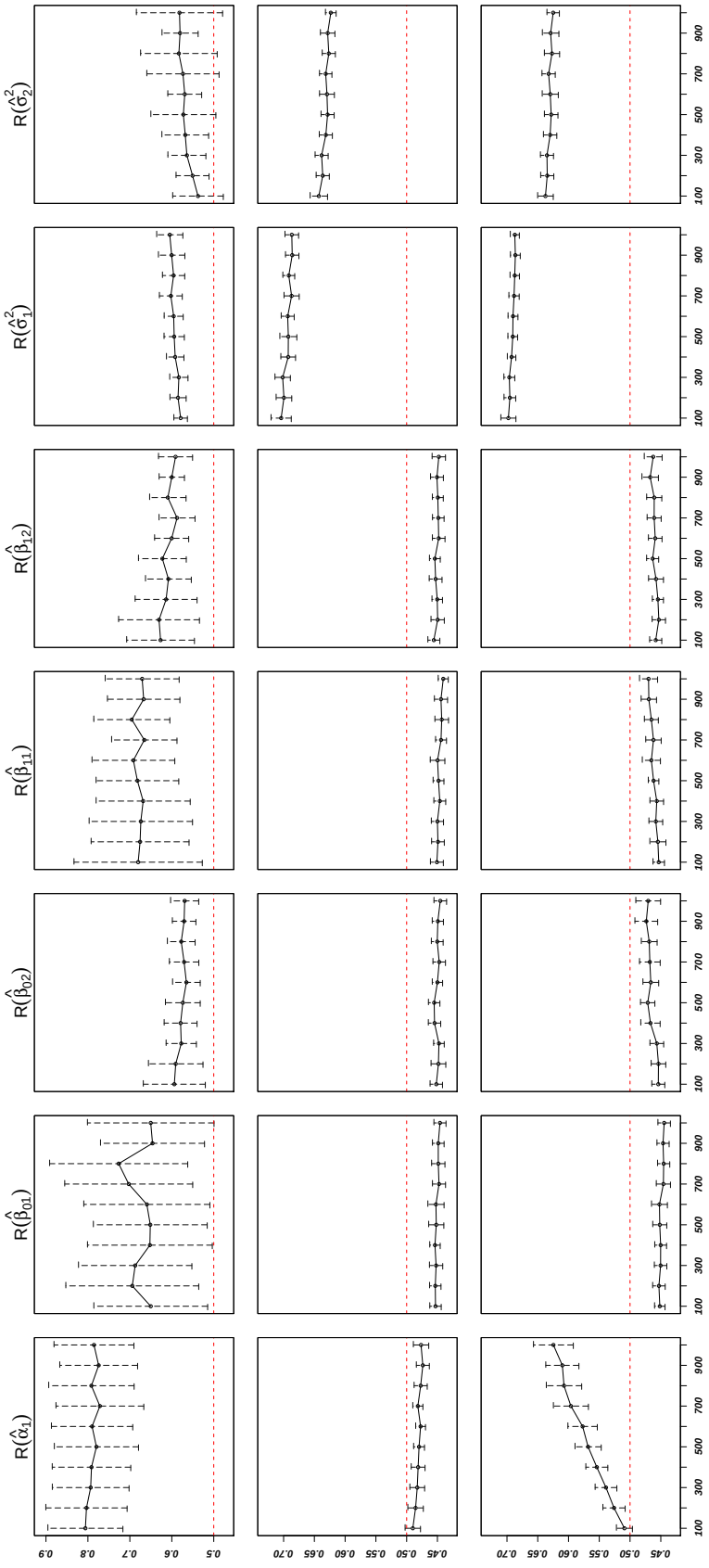


Figure 3.2: Summary of the $\mathcal{R}$ index for bias (top row), variance (middle row), and mean squared error (bottom row) in Simulation study 2. The middle line represents the median and the dashed bars represent the interquartile range (IQR) values at different sample sizes n.

a finite population containing two strata with $10000, 8000$ observations in each stratum. We considered three mixture components $(K = 3)$. Therefore, we have $n = 1000$ observations selected in the total. In Mixture 3, we generated a finite population containing three strata with $1000, 8000, 6000$ observations in each stratum, respectively. The population has four components $(K = 4)$, where the total of $n = 1500$ observations selected with 500 from each stratum. In Mixture 4, we generated a finite population containing two strata with $12000, 8000$ observations, respectively. In this case, the population has five components $(K = 5)$. The total number of observations in the sample was $n = 1000$. After generating data, the weighted model is fitted for different values of K ranging from 1 to 10. The BIC is then calculated for each K. Figure 3.8 shows the results of this experiment including the BIC values for all K and the optimal number of components in the four experiments above. According to the results, BIC was able to choose the optimal number of components under the various circumstance. In all four cases, BIC was the lowest at the true K value.

Table 3.4: True parameter values for Mixtures of linear regression in Simulation 3.

		$\boldsymbol{\psi}$									
H	K	$\boldsymbol{\alpha}_1$	$\boldsymbol{\alpha}_2$	$\boldsymbol{\alpha}_3$	$\boldsymbol{\alpha}_4$	$\boldsymbol{\beta}_{10}$	$\boldsymbol{\beta}_{20}$	$\boldsymbol{\beta}_{30}$	$\boldsymbol{\beta}_{40}$	$\boldsymbol{\beta}_{50}$	$\boldsymbol{\beta}_{11}$
2	2	0.52				-3	3				1
2	3	0.30	0.36			-5	-4	1			1
3	4	0.17	0.32	0.29		-5	-4	1	2		1
2	5	0.24	0.26	0.20	0.15	-5	-4	-1	1	2	1
H	K	$\boldsymbol{\beta}_{21}$	$\boldsymbol{\beta}_{31}$	$\boldsymbol{\beta}_{41}$	$\boldsymbol{\beta}_{51}$	$\boldsymbol{\sigma}_1^2$	$\boldsymbol{\sigma}_2^2$	$\boldsymbol{\sigma}_3^2$	$\boldsymbol{\sigma}_4^2$	$\boldsymbol{\sigma}_5^2$	
2	2	-2				0.1	0.1				
2	3	-2	-3			0.1	0.1	0.5			
3	4	-2	-3	-1		0.1	0.1	0.5	0.5		
2	5	-2	1	-1	-3	0.1	0.1	0.5	0.5	0.4	

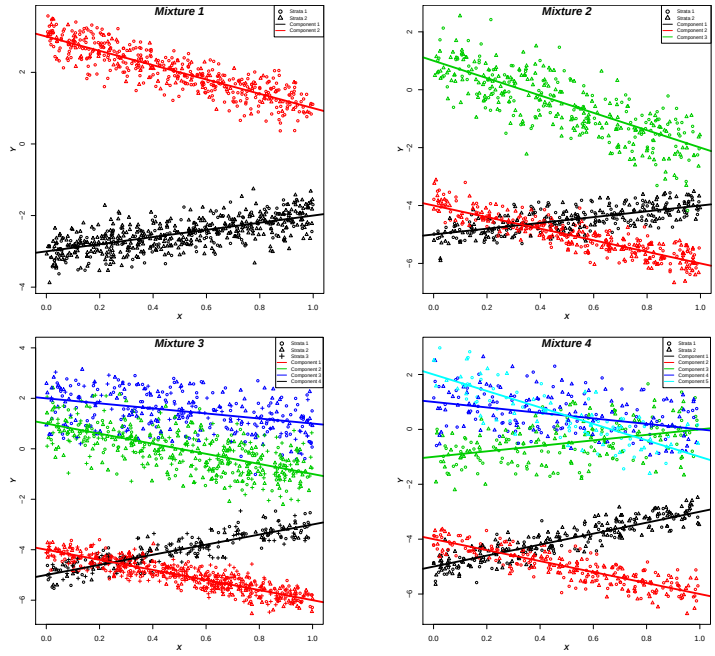


Figure 3.3: Scatter plots of samples which are selected from finite populations for the four experiments in Simulation 3. Colors show the number of components, and plotting characters show the strata.

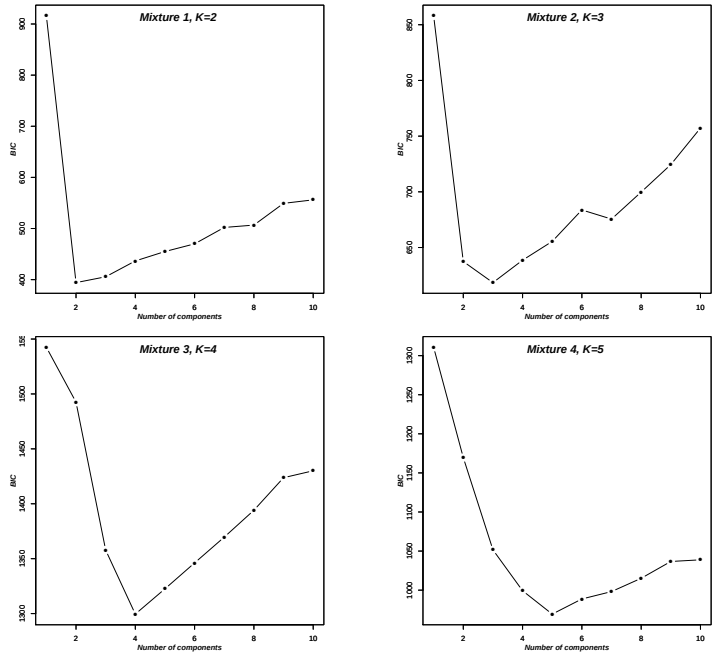


Figure 3.4: BIC values corresponding to the optimal number of components for the four experiments.

3.2 Simulation Studies of Cluster sampling Data

In this section, we explore how to fit the finite mixture normal regression model for samples drawn using a cluster sampling design. We will introduce some simulation studies and applications of the proposed model. In the first simulation study, we conduct a comparison to assess the parameter recovery performance of the proposed model through the usual finite mixture linear regression model. In the second simulation, more examination happens to improve research regarding the capacity of the proposed model with regular methodology available. In the third simulation study, we use the same settings as in the first simulation to assess the performance of the weighted model as a classification tool. We use the utility of the BIC for selecting the optimal number of components for a given dataset.

3.2.1 Simulation 1: Parameter Estimation of Cluster Sampling Data

This simulation study was executed to assess the performance of the maximum likelihood estimates obtained via the unweighted and weighted model in various scenarios. The criteria used for comparison include MSE, variance and bias. In this setting, the true values of parameters were used as the starting values. A finite population composed 23000 observations distributed on $N = 4$ clusters were generated from a two-component mixture of a normal regression model. Two varying scenarios of the true regression line were implemented: non-overlapping and overlapping named Mixture 1 and Mixture 2, respectively. The vector of parameters $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2)$ are reported in Table 3.5. The generated finite population has four clusters of $\{6000, 4000, 8000, 5000\}$ observations in each cluster respectively which are called the primary sampling units PSU's. Two-stage cluster sampling strategy was considered to draw the samples. Thus, a simple random sample of two clusters as PSU's was obtained in the first stage. Then, samples of sizes $m_1 = m_2 = \{100, 250, 500, 1000\}$ are drawn from each sampled cluster. Thus, the to-

tal sample sizes of $n = 200, 500, 1000, 2000$ are considered. For example, when Cluster 1 and Cluster 3 were sampled in the first stage which contained 6000, and 8000 observations, respectively. In this case, in the total sample size $n = 1000$, there are $m_1 = 500$ from Cluster 1 and $m_2 = 500$ from Cluster 3. The mixing proportions $\alpha_1 = 0.37$ and $\alpha_2 = 0.63$ have been used to determine how many observations will belong to component one and component two, respectively. Figure 3.5 shows a sample of size $n = 1000$ observations from the considered models, Mixture 1 and Mixture 2 when the Cluster 1 and Cluster 3 have been sampled. The above setup is repeated to $B = 1000$ replications.

Table 3.5: True parameter values for Mixture 1 and Mixture 2 considering cluster sampling design.

$\boldsymbol{\psi}$	α_1	α_2	β_{10}	β_{20}	β_{11}	β_{21}	σ_1^2	σ_2^2
Mixture 1	0.37	0.63	-3	3	3	1	0.1	0.1
Mixture 2	0.37	0.63	-3	-2	3	-1	0.1	0.1

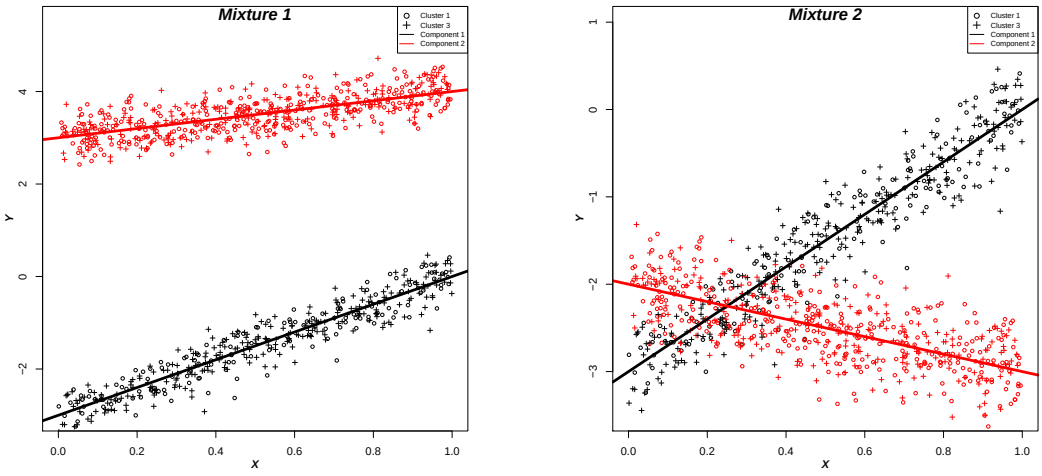


Figure 3.5: Scatter plots of a sample of size $n = 1000$ units. Colors show the two components and plotting characters represent clusters. The left plot represents Mixture 1 - non-overlapping components, and the right plot represent Mixture 2 - overlapping components when cluster sampling design was considered.

Table 3.6 provides the MSE and components of bias and variance for varying sample sizes when Mixture1 is considered. The bold values show where the minimum is achieved when comparing both models. Estimates obtained by the weighted model have a smaller

bias compared to those of an unweighted model in 20 out of 28 cases. Table 3.7 Similarly, provides the MSE, bias and variance components were estimated for Mixture 2. The estimates obtained by the weighted model also have a smaller bias compared to the estimates obtained by the unweighted model in 22 out of 28 cases. Therefore, one can infer that both outcomes using the weighted approach in estimating parameters results in a smaller bias comparatively to an unweighted approach. The weighted model estimates have relatively high variability compared to the unweighted model estimates in 24 out of 28 cases for Mixture 1 and 20 out of 28 cases for Mixture 2. However, the variance of the estimates obtained via the two approaches declined by increasing the size of a sample.

Table 3.6: Mean squared error, bias, and variance of estimated parameters, based on 1000 replications for different sample sizes of the two-component when the Mixture 1 configuration was considered under cluster sampling design. The values reported are $\times 10^{-2}$.

		$\hat{\psi}$	$\hat{\alpha}_1$	$\hat{\beta}_{10}$	$\hat{\beta}_{20}$	$\hat{\beta}_{11}$	$\hat{\beta}_{21}$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$
$n = 200$	MSE	Weighted	0.6486	0.4309	0.3575	1.4556	1.0719	**0.0311**	**0.0194**
		Unweighted	**0.5444**	**0.4092**	**0.3540**	**1.3490**	**1.0262**	0.0313	0.0198
	Bias^2	Weighted	**0.0488**	**0.0090**	**0.0134**	**0.0350**	**0.0211**	**0.0004**	**0.0003**
		Unweighted	0.0770	0.0104	0.0148	0.0362	0.0282	0.0005	0.0004
	Var	Weighted	0.5998	0.4219	0.3441	1.4206	1.0508	**0.0307**	**0.0191**
		Unweighted	**0.4674**	**0.3988**	**0.3392**	**1.3128**	**0.9980**	0.0308	0.0194
$n = 500$	MSE	Weighted	0.4317	0.1580	0.1589	0.6116	0.4425	0.0125	0.0068
		Unweighted	**0.3243**	**0.1542**	**0.1565**	**0.5993**	**0.4292**	**0.0121**	**0.0066**
	Bias^2	Weighted	**0.0001**	0.0016	**0.0001**	0.0320	**0.0003**	**0.0001**	0.0002
		Unweighted	0.0091	**0.0006**	0.0002	**0.0272**	0.0004	0.0002	**0.0001**
	Var	Weighted	0.4316	0.1564	0.1588	0.5796	0.4422	0.0124	0.0066
		Unweighted	**0.3152**	**0.1536**	**0.1563**	**0.5721**	**0.4288**	**0.0119**	**0.0065**
$n = 1000$	MSE	Weighted	0.5891	0.1257	0.0727	0.3366	0.2563	0.0066	0.0036
		Unweighted	**0.4814**	**0.1208**	**0.0693**	**0.3154**	**0.2466**	**0.0061**	**0.0035**
	Bias^2	Weighted	**0.0144**	0.0002	**0.0050**	0.0156	0.0054	**0.0007**	**0.0001**
		Unweighted	0.0306	**0.0001**	0.0058	**0.0134**	**0.0053**	0.0008	0.0002
	Var	Weighted	0.5747	0.1256	0.0677	0.3210	0.2509	0.0059	0.0035
		Unweighted	**0.4508**	**0.1208**	**0.0635**	**0.3020**	**0.2412**	**0.0053**	**0.0034**
$n = 2000$	MSE	Weighted	0.5995	0.0549	0.0304	0.1880	0.1018	0.0044	0.0017
		Unweighted	**0.5175**	**0.0548**	**0.0283**	**0.1849**	**0.0937**	**0.0043**	**0.0015**
	Bias^2	Weighted	**0.0335**	0.0057	**0.0014**	**0.0301**	**0.0019**	**0.0005**	0.0002
		Unweighted	0.0681	**0.0055**	0.0015	0.0324	0.0020	0.0007	**0.0001**
	Var	Weighted	0.5660	**0.0492**	0.0290	0.1579	0.0999	0.0039	0.0015
		Unweighted	**0.4494**	0.0493	**0.0268**	**0.1525**	**0.0917**	**0.0036**	**0.0014**

Table 3.7: Mean squared error, bias, and variance of estimated parameters, based on 1000 replications for different sample sizes of the two-component when the Mixture 2 configuration was considered under cluster sampling design. The values reported are $\times 10^{-2}$.

		$\hat{\psi}$	$\hat{\alpha}_1$	$\hat{\beta}_{10}$	$\hat{\beta}_{20}$	$\hat{\beta}_{11}$	$\hat{\beta}_{21}$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$
$n = 200$	MSE	Weighted	0.6940	0.3359	0.2502	1.5016	0.8560	**0.0250**	**0.0130**
		Unweighted	**0.5882**	**0.3295**	**0.2256**	**1.4147**	**0.7424**	0.0565	0.0201
	Bias^2	Weighted	**0.0044**	**0.0046**	**0.0004**	**0.0009**	**0.0014**	0.0002	**0.0010**
		Unweighted	0.0167	0.0050	0.0006	0.0022	0.0015	**0.0001**	0.0021
	Var	Weighted	0.6896	0.3313	0.2498	1.5007	0.8546	**0.0249**	**0.0120**
		Unweighted	**0.5715**	**0.3245**	**0.2250**	**1.4125**	**0.7409**	0.0565	**0.0180**
$n = 500$	MSE	Weighted	0.6453	0.1603	0.1013	0.5602	0.3429	**0.0106**	**0.0056**
		Unweighted	**0.5794**	**0.1567**	**0.0935**	**0.5390**	**0.3101**	0.0235	0.0088
	Bias^2	Weighted	**0.0712**	0.0239	**0.0002**	**0.0431**	0.0010	**0.0001**	**0.0006**
		Unweighted	0.1115	**0.0225**	0.0003	0.0457	**0.0008**	0.0002	0.0012
	Var	Weighted	0.5741	0.1364	0.1011	0.5171	0.3419	**0.0106**	**0.0050**
		Unweighted	**0.4679**	**0.1342**	**0.0932**	**0.4933**	**0.3093**	0.0233	0.0076
$n = 1000$	MSE	Weighted	0.5551	0.0993	0.0476	0.3150	0.1698	**0.0044**	**0.0029**
		Unweighted	**0.4857**	**0.0952**	**0.0426**	**0.3036**	**0.1585**	0.0100	0.0053
	Bias^2	Weighted	**0.0264**	0.0202	0.0002	**0.0446**	**0.0034**	**0.0001**	**0.0006**
		Unweighted	0.0515	**0.0195**	**0.0001**	0.0450	0.0037	0.0002	0.0013
	Var	Weighted	0.5287	0.0791	0.0476	0.2704	0.1664	**0.0044**	**0.0023**
		Unweighted	**0.4342**	**0.0757**	**0.0426**	**0.2586**	**0.1548**	0.0099	0.0040
$n = 2000$	MSE	Weighted	0.6044	0.0542	0.0201	**0.1674**	0.0681	**0.0018**	**0.0017**
		Unweighted	**0.5240**	**0.0529**	**0.0186**	0.1677	**0.0612**	0.0042	0.0032
	Bias^2	Weighted	**0.0313**	**0.0237**	**0.0001**	**0.0402**	0.0002	**0.0002**	**0.0005**
		Unweighted	0.0510	0.0238	0.0002	0.0454	**0.0001**	0.0004	0.0012
	Var	Weighted	0.5731	0.0305	0.0200	0.1272	0.0679	**0.0017**	**0.0012**
		Unweighted	**0.4730**	**0.0291**	**0.0184**	**0.1223**	**0.0611**	0.0038	0.0020

3.2.2 Simulation 2: Model Comparison of Cluster Sampling Data

Based on previous study results, in most cases, the weighted model had a lower bias, yet intermittently, the unweighted model estimates had a lower bias compared to those found using the weighted model. As a result, the simulation study has yet to establish the numerical evidence of the two approaches conclusively.

Considering the Mixture 1 setup, a finite population containing $\{6000, 4000, 8000, 5000\}$ observations appeared in each individual cluster. The vector of parameters is shown in Table 3.5. In the primary stage, a simple random sample of two clusters was selected as PSU's. At that point, samples of different sizes were extracted from each cluster, beginning with 50 per cluster up to 500 with an increment of 50 observations. Accordingly, the samples that were chosen are $n = \{100, 200, 300, \ldots, 1000\}$. We replicated $B = 100$ times for each n. These replicates were then applied to calculate MSE values and the corresponding bias and variance components. Next, two hundred replications of the above setup were executed to find 200 values of MSE, bias and variance values per sample size and parameter using both the weighted and unweighted models. These 200 replicates were then applied to calculate the percent contribution index, $\mathcal{R}$, defined in Section 2.6, by setting θ_1 to be the results from the unweighted model and θ_2 to be the results from the weighted model. The outcomes of this study can be found in a multiplot presented in Figure 3.6.

Concerning the top panel, the median values of the $\mathcal{R}$-index for bias were above 0.5 in every estimated parameter and sample size. In most cases, the ± 1 IQR bar of the R-index for bias was above the dashed horizontal line with the exception of a few cases of (estimation of β_{10}) where some IQR lines were marginally beneath the 0.5 line. Considering the mixing proportion $\hat{\alpha}_1$, we can noted that on average about 65% of the total bias was contributed by the unweighted model. In general, the effects of sample size on the bias and its variability was not clear. For three out of four intercept/slope parameter estimates, the median value of R appeared to decrease with sample size. Conversely, the median value of R of estimation σ^2 appeared to increase as the sample size increased. In general, varying

sample sizes did not create a well-defined IQR trend of the index.

Regarding the index for the variance component presented in the middle panel of Figure 3.6, in five out of seven parameters, the median value of $\mathcal{R}$ and the ± 1 IQR bars were under the dashed line. The exceptions to this were for the estimates of σ^2. For both components, the variance appeared to be substantially higher for the unweighted model than the weighted model. Furthermore, looking at the IQR, we see that $\mathcal{R}$ associated with variance is far less and has shorter bars than the same index for bias.

Finally, viewing $\mathcal{R}$ index of MSE at the bottom panel of Figure 3.6, as MSE is the sum of the bias squared and variance the results presented are reflective of the above two. On the whole, the median value of $\mathcal{R}$ was below the 0.5 threshold excepting the variance parameters. Regarding the variability of $\mathcal{R}$ for MSE, we see that, similar to the variance component, it had lower variability compared to the same index for the bias. Based on the three summaries it can be concluded that although it is unclear which model works better in terms of MSE, the bias in parameter estimates found using the weighted model is lower than the unweighted model.

3.2.3 Simulation 3: Model Selection of Cluster Sampling Data

In this simulation study, assuming that the finite population involves several nonoverlapping clusters. We assess the performance of the BIC as a model selection tool when considering the weighted approach as a classification method. The finite population composed of four clusters PSU's consisting of $\{6000, 4000, 8000, 5000\}$ observations in each cluster. The vector of true parameters $\boldsymbol{\Psi} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2)$ used to generate the mixtures are shown in Table 3.8. In this setup, samples were drawn using cluster sampling design from the finite population by selecting a simple random sample of two clusters in the first stage. Then, size $m_i = 500$ from each sampled clusters, $i = 1, ..., 4$. Thus, the total of size $n = 1000$ observations were considered. We considered four configurations of the regression lines, which are called: Mixture 1, Mixture 2, Mixture 3, Mixture 4. In Mixture 1,

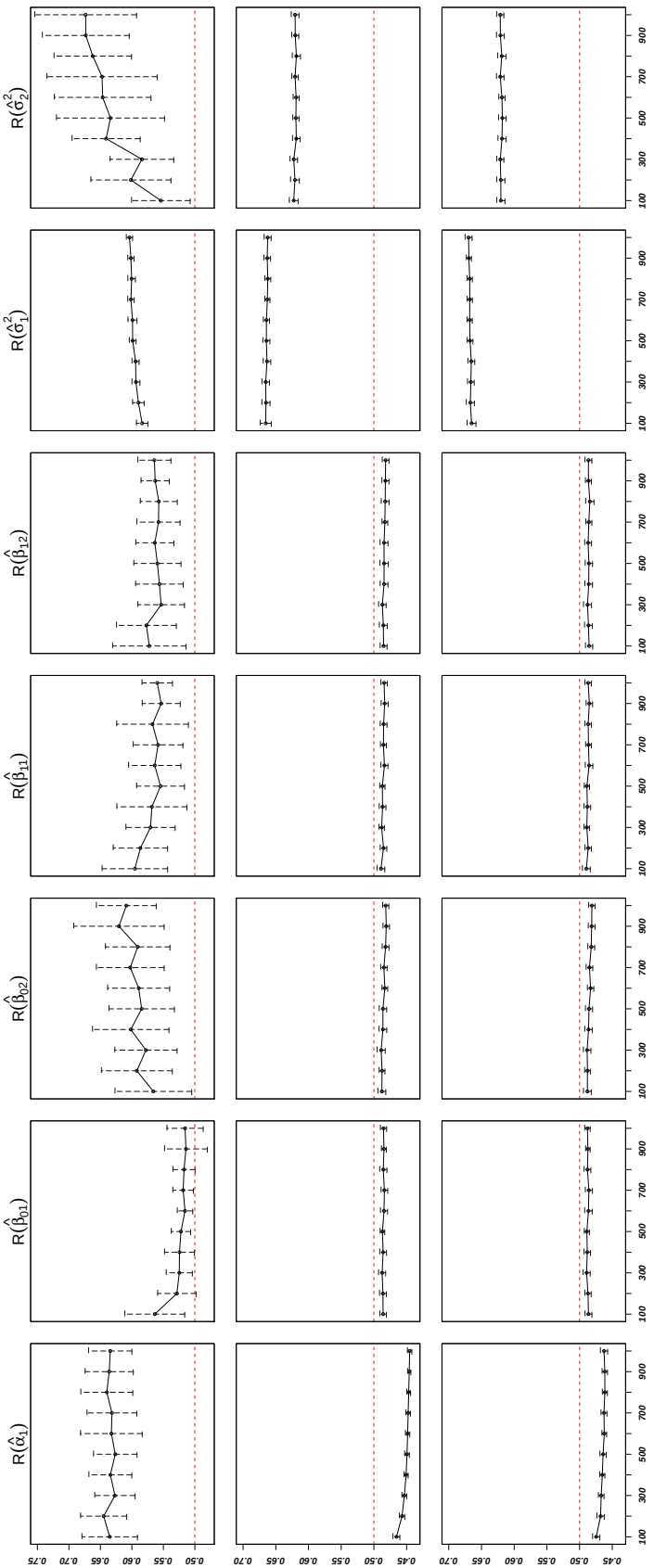


Figure 3.6: Summary of the $\mathcal{R}$ index for bias (top row), variance (middle row), and mean squared error (bottom row) in Simulation study 2. The middle line represents the median and the dashed bars represent the interquartile range (IQR) values at different sample sizes n.

there are two components $(K = 2)$. In Mixture 2, We considered three mixture components $(K = 3)$. In Mixture 3, the population has four components $(K = 4)$. In Mixture 4, in this case, the population has five components $(K = 5)$. Figure 3.7 shows the scatter plots arising from the four scenarios when two-stage cluster sampling considered. Figure 3.8 shows results of BIC and the optimal number of components in the different mixture settings of this simulation. According to the results, BIC was able to choose the optimal number of components under various circumstance. In all four cases, BIC was the lowest at the true K value.

In Mixture 1, there are two components $(K = 2)$, and the total of $n = 1000$ observations were selected. In Mixture 2, We considered three mixture components $(K = 3)$. Therefore, we have $n = 1000$ observations selected in the total. In Mixture 3, the population has four components $(K = 4)$, where a total of $n = 1500$ observations selected with 500 from each stratum. In Mixture 4, in this case, the population has five components $(K = 5)$. The total number of observations in the sample was $n = 1000$.

After generating data, the weighted model is fitted for different values of K ranging from 1 to 10. The BIC is then calculated for each K. Figure 3.8 shows the results of this experiment, including the BIC values for all K and the optimal number of components in the four experiments above. According to the results, BIC was able to choose the optimal number of components under various circumstance. In all four cases, BIC was the lowest at the true K value.

Table 3.8: The true parameters used for simulating Mixtures of linear regression in Simulation 3.

		ψ									
H	K	$\boldsymbol{\alpha}_1$	$\boldsymbol{\alpha}_2$	$\boldsymbol{\alpha}_3$	$\boldsymbol{\alpha}_4$	$\boldsymbol{\beta}_{10}$	$\boldsymbol{\beta}_{20}$	$\boldsymbol{\beta}_{30}$	$\boldsymbol{\beta}_{40}$	$\boldsymbol{\beta}_{50}$	$\boldsymbol{\beta}_{11}$
4	2	0.37				-3	-2				3
4	3	0.39	0.37			-3	-1	1			3
4	4	0.25	0.32	0.27		-5	-1	1	5		1
4	5	0.22	0.25	0.20	0.13	-5	-1	0	1	5	1
H	K	$\boldsymbol{\beta}_{21}$	$\boldsymbol{\beta}_{31}$	$\boldsymbol{\beta}_{41}$	$\boldsymbol{\beta}_{51}$	$\boldsymbol{\sigma}_1^2$	$\boldsymbol{\sigma}_2^2$	$\boldsymbol{\sigma}_3^2$	$\boldsymbol{\sigma}_4^2$	$\boldsymbol{\sigma}_5^2$	
2	2	-1				0.1	0.1				
2	3	1	-1			0.1	0.1	0.1			
3	4	-1	1	-1		0.1	0.1	0.1	0.1		
2	5	-1	1	1	-1	0.1	0.1	0.1	0.1	0.1	

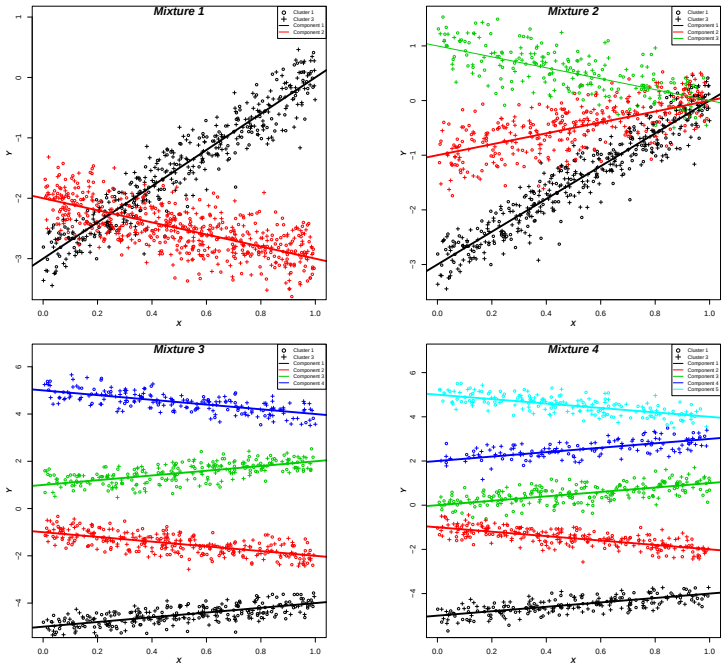


Figure 3.7: Scatter plots of samples selected from finite populations for the four experiments in Simulation 1. Colors show the number of components and plotting characters show the clusters.

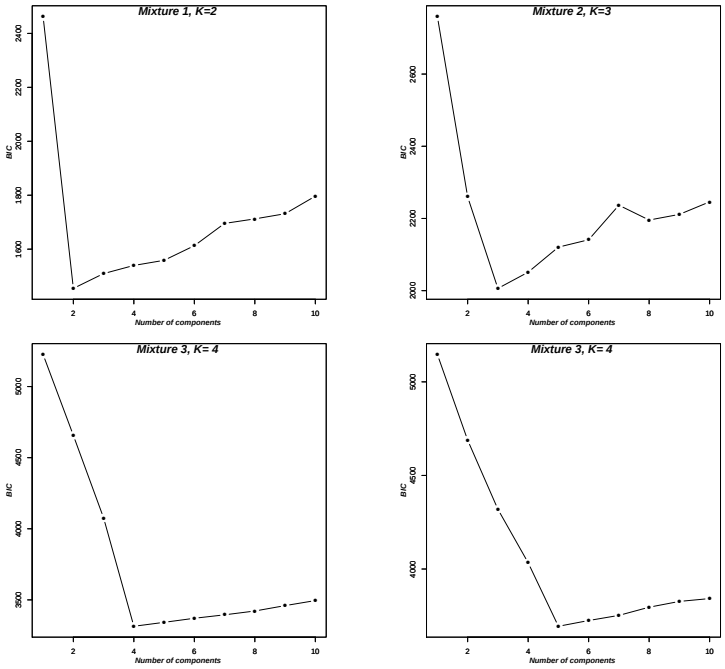


Figure 3.8: BIC values corresponding to the optimal number of components for the four experiments.

3.3 Simulation Studies of Complex Sampling

The aim of this section is to analyze the performance of the finite mixture regression models when the samples are drawn using a complex sampling design. Many simulation studies will be constructed to demonstrate the proficiency of the proposed model to analyze the data drawn from a complex survey design. In the initial simulation, we evaluated the performance of the maximum likelihood estimates acquired using the proposed model and usual finite mixture regression model in different situations using the mean squared error components, including variance and bias. The second simulation study revisits the percent contribution index discussed in the preceding chapters of stratified and cluster sampling to assess the effect of the sample size on the parameter recovery and weigh the variability linked with the MES, its bias, and variance components. The third simulation was executed to gauge the capability of the BIC criterion for selecting the optimal number of components for a given dataset using various settings to the proposed model.

3.3.1 Simulation 1: Parameter Estimation of Complex Sampling Data

For the parameter estimation simulation study, the performance of the suggested model and weighted model in various scenarios was assessed. Assuming that samples were drawn from a complex sampling design under a variety of circumstances, we analyzed the effects of merging the sampling weights in the parameter recovery and in the quality of the bias and standard error estimates. Specifically, a finite population comprised of $N = 22000$ observations was produced from a two-component mixture of a normal regression model. Two instances of the regression line were considered: non-overlapping and overlapping, named Mixture 1 and Mixture 2, respectively. A finite population was stratified into eight strata including $\{4000, 2500, 3500, 1500, 2800, 2400$
$, 3000, 2300\}$ observations in each stratum. The true parameters vector $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2)$ are presented in Table 3.9. A stratified two-stage cluster sample design was considered. The vari-

able 'Strata' has eight strata and was utilized as a stratification variable. The variable 'Cluster 1', which contains nineteen clusters, was utilized as the clustering variable in the first stage. One cluster was drawn from each stratum via simple random sample without replacement as PSU's. The variable 'Cluster 2' was utilized as the clustering variable in the second stage. A simple random sample with sizes $\Big\{\{36, 24, 32, 16, 20, 24, 28, 20\}, \{90, 60, 80, 40, 50, 60, 70, 50\}, \{180, 120, 160, 80, 100, 120, 140, 100)\}, \{\ 360, 240, 320, 160, 200, 240, 280, 200\}\Big\}$ observations was extracted from the eight selected clusters in the first stage. The number of observations sampled in the second stage is proportionate to the size of the original stratum from which the cluster was selected. The total sample sizes of $n = 200, 500, 1000, 2000$ are considered. Thus, for $n = 1000$, we have $\{180, 120, 160, 80, 100, 120, 140, 100\}$ from each selected cluster in the first stage, respectively. For instance, for Mixture 1, within each cluster, $\alpha_1 = 0.46$ and $\alpha_2 = 0.54$ was employed to conclude how many observations would belong to component and component two, respectively. Figure 3.9 Depicts a sample of size $n = 1000$ observations using the stratified two-stage cluster sample design considering models, Mixture 1 and Mixture 2. One thousand replications $B = 1000$ were performed for the arrangement above.

In all the considered simulations, the convergence of the true model was acquired for the weighted and unweighted models. The significance of the bias was virtually negligible in all scenarios. Yet, concerning Mixture 1 or Mixture 2, in 24 out of 28 cases the estimates procured by the weighted model had a smaller bias compared to the estimates procured by the unweighted model as presented in Table 3.10 and Table 3.11. Regarding the mean squared error estimates, no systematic behavior was observed. Relating to the variance of estimates, in all cases, the weighted model estimates produced moderately high variability contrasted to the unweighted model estimates as shown in Table 3.10 while in 19 out of 28 cases as reported in Table 3.11. It is notable that the values roughly decrease with increases in the sample size.

Table 3.9: True parameter values for Mixture 1 and Mixture 2 considering complex sampling design.

$\boldsymbol{\psi}$	α_1	α_2	β_{10}	β_{20}	β_{11}	β_{21}	σ_1^2	σ_2^2
Mixture 1	0.46	0.54	-3	3	3	1	0.1	0.1
Mixture 2	0.46	0.54	-3	-2	3	-1	0.1	0.1

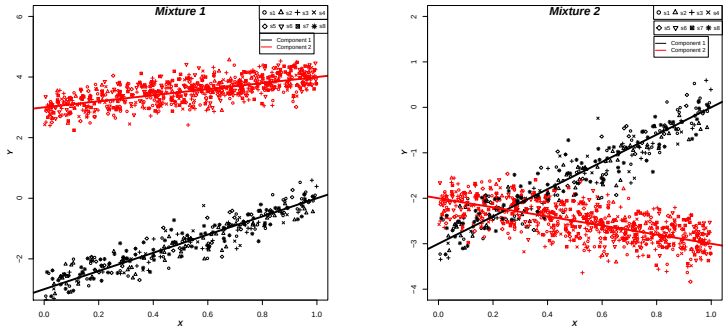


Figure 3.9: Scatter plots of a sample of size $n = 1000$ units. Colors show the components, and plotting characters represent the eight strata as primary sampling units (PSU's) from where the sampled units were drawn. The left plot represents Mixture 1 - non-overlapping components, and the right plot represent Mixture 2 - overlapping components.

Table 3.10: Mean squared error, bias, and variance of estimated parameters, based on 1000 replications for different sample sizes of the two-component when the Mixture 1 configuration was considered under complex survey design. The values reported are $\times 10^{-2}$.

		$\hat{\psi}$	$\hat{\alpha}_1$	$\hat{\beta}_{10}$	$\hat{\beta}_{20}$	$\hat{\beta}_{11}$	$\hat{\beta}_{21}$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$
$n = 200$	MSE	Weighted	**0.1345**	0.0505	0.0418	0.1572	0.1262	0.0029	0.0020
		Unweighted	0.1407	**0.0493**	**0.0391**	**0.1543**	**0.1191**	**0.0027**	**0.0018**
	Bias^2	Weighted	**0.0009**	**0.0001**	**0.0004**	**0.0004**	**0.0004**	0.0004	**0.0001**
		Unweighted	0.0102	0.0005	0.0002	0.0010	0.0012	**0.0003**	0.0002
	Var	Weighted	0.1336	0.0502	0.0418	0.1568	0.1257	0.0025	0.0018
		Unweighted	**0.1305**	**0.0488**	**0.0389**	**0.1533**	**0.1180**	**0.0024**	**0.0017**
$n = 500$	MSE	Weighted	**0.1297**	0.0208	0.0144	0.0613	0.0428	0.0013	0.0008
		Unweighted	**0.1403**	**0.0199**	**0.0135**	**0.0594**	**0.0404**	**0.0012**	**0.0007**
	Bias^2	Weighted	**0.0002**	**0.0003**	**0.0001**	**0.0004**	**0.0001**	**0.0002**	**0.0001**
		Unweighted	0.0132	0.0004	0.0002	0.0008	0.0005	0.0003	0.0002
	Var	Weighted	0.1295	0.0205	0.0144	0.0609	0.0427	0.0011	0.0007
		Unweighted	**0.1271**	**0.0195**	**0.0134**	**0.0586**	**0.0399**	**0.0010**	**0.0007**
$n = 1000$	MSE	Weighted	**0.1410**	0.0101	0.0077	0.0296	0.0236	0.0006	0.0004
		Unweighted	0.1472	**0.0097**	**0.0070**	**0.0292**	**0.0223**	**0.0006**	**0.0003**
	Bias^2	Weighted	**0.0007**	**0.0003**	0.0001	**0.0006**	**0.0001**	0.0003	**0.0001**
		Unweighted	0.0107	0.0003	**0.0002**	0.0012	0.0004	**0.0002**	0.0002
	Var	Weighted	0.1403	0.0098	0.0077	0.0290	0.0235	0.0005	0.0004
		Unweighted	**0.1365**	**0.0093**	**0.0070**	**0.0280**	**0.0219**	**0.0004**	**0.0003**
$n = 2000$	MSE	Weighted	**0.1479**	0.0050	0.0038	0.0143	0.0115	0.0005	0.0005
		Unweighted	0.1514	**0.0047**	**0.0035**	**0.0136**	**0.0113**	**0.0004**	**0.0004**
	Bias^2	Weighted	**0.0022**	**0.0004**	**0.0001**	**0.0006**	**0.0001**	**0.0001**	**0.0001**
		Unweighted	0.0065	0.0005	0.0002	0.0010	0.0004	0.0002	0.0002
	Var	Weighted	0.1457	0.0046	0.0038	0.0138	0.0114	0.0004	0.0004
		Unweighted	**0.1449**	**0.0043**	**0.0035**	**0.0126**	**0.0109**	**0.0002**	**0.0002**

Table 3.11: Mean squared error, bias, and variance of estimated parameters, based on 1000 replications for different sample sizes of the two-component when the Mixture 2 configuration was considered under complex survey design. The values reported are $\times 10^{-2}$.

		$\hat{\psi}$	$\hat{\alpha}_1$	$\hat{\beta}_{10}$	$\hat{\beta}_{20}$	$\hat{\beta}_{11}$	$\hat{\beta}_{21}$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$
$n = 200$	MSE	Weighted	**0.1656**	0.0356	0.0304	0.1358	0.1005	**0.0020**	**0.0016**
		Unweighted	0.2047	**0.0336**	**0.0290**	**0.1274**	**0.0950**	0.0035	0.0025
	Bias^2	Weighted	**0.0221**	**0.0001**	**0.0009**	**0.0011**	**0.0006**	**0.0001**	**0.0001**
		Unweighted	0.0625	0.0002	0.0020	0.0020	0.0026	0.0002	0.0001
	Var	Weighted	0.1435	0.0356	0.0295	0.1348	0.0999	**0.0019**	**0.0016**
		Unweighted	**0.1422**	**0.0336**	**0.0270**	**0.1254**	**0.0924**	0.0035	0.0024
$n = 500$	MSE	Weighted	**0.1344**	0.0129	0.0117	0.0505	0.0395	**0.0007**	**0.0006**
		Unweighted	0.1418	**0.0126**	**0.0103**	**0.0489**	**0.0351**	0.0014	0.0010
	Bias^2	Weighted	**0.0011**	**0.0001**	**0.0002**	**0.0004**	**0.0002**	**0.0001**	0.0002
		Unweighted	0.0109	0.0001	0.0003	0.0010	0.0004	0.0002	**0.0001**
	Var	Weighted	0.1333	0.0129	0.0115	0.0501	0.0393	**0.0007**	**0.0006**
		Unweighted	**0.1309**	**0.0125**	**0.0100**	**0.0479**	**0.0348**	0.0014	0.0010
$n = 1000$	MSE	Weighted	**0.1436**	0.0059	0.0055	0.0222	0.0186	**0.0003**	**0.0003**
		Unweighted	0.146	**0.0056**	**0.0049**	**0.0212**	**0.0169**	0.00070	0.0005
	Bias^2	Weighted	**0.0032**	0.0002	**0.0001**	**0.0001**	**0.0001**	**0.0001**	0.0002
		Unweighted	0.0060	**0.0001**	0.0002	0.0003	0.0003	0.0002	**0.0001**
	Var	Weighted	0.1404	0.0059	0.0053	0.0222	0.0185	**0.0003**	**0.0003**
		Unweighted	**0.1400**	**0.0056**	**0.0047**	**0.0209**	**0.0166**	0.0007	0.0005
$n = 2000$	MSE	Weighted	**0.1303**	0.0029	0.0027	0.0115	0.0097	**0.0003**	**0.0003**
		Unweighted	0.1347	**0.0029**	**0.0026**	**0.0114**	**0.0092**	0.0005	0.0004
	Bias^2	Weighted	**0.0034**	**0.0001**	**0.0001**	**0.0003**	**0.0001**	**0.0001**	0.0002
		Unweighted	0.0045	0.0000	0.0002	0.0007	0.0003	0.0002	**0.0001**
	Var	Weighted	**0.1269**	0.0029	0.0026	0.0112	0.0096	**0.0002**	**0.0001**
		Unweighted	0.1301	**0.0029**	**0.0024**	**0.0107**	**0.0089**	0.0003	0.0003

3.3.2 Simulation 2: Model Comparison of Complex Sampling Data

In all of the studied $1000 \times 4 = 4000$ simulations, the weighted model had a lower bias, yet intermittently, the unweighted model estimates had a lower preference compared to those found via the weighted model. Consequently, the simulation study has yet to establish the general attributes of the two approaches conclusively. Considering the Mixture 1 setup; a finite population was stratified into eight strata including $\{4000, 2500, 3500, 1500, 2800, 2400, 3000, 2300\}$ observations in each stratum. The vector of parameters is shown in Table 3.9. In the first stage, one cluster was drawn from each stratum using a simple random sample without replacement as PSU's. Therefore, eight clusters were sampled in this case. In the second stage, samples of various sizes were taken from each cluster sampled, beginning with $\{18, 12, 16, 8, 10, 12, 14, 10\}$ per cluster up to $\{180, 120, 160, 80, 100, 120, 140, 100\}$ with multiplying the number of observations from each sampled cluster. Accordingly, the samples that were selected are $n = \{100, 200, 300, \ldots, 1000\}$. The MSE values and the corresponding bias and variance components were calculated across 100 replications for each n. Then, 200 replications of the above setup were employed to obtain 200 values of MSE, bias, and variance values per sample size and parameter utilizing both the weighted and unweighted models.

Subsequently, the above setup was then used to elicit further investigate the parameter recovery capability of both approaches. These 200 replicates above were then applied to calculate the percent contribution index, $\mathcal{R}$, defined in Section 2.6, by setting θ_1 to be the results from the unweighted model and θ_2 to be the results from the weighted model. The outcomes of this study can be found in a multiplot presented in Figure 3.10. The top panel of Figure 3.10 shows the R-index for bias. In all of the varying sample sizes and the seven estimated parameters, the median value of $\mathcal{R}$-index was over the threshold value. In general, the ± 1 IQR bars of the $\mathcal{R}$-index for bias were above the dashed horizontal line excepting a few occurrences of (estimation of σ^2) where a few IQR lines were a little lower than the 0.5 line. On the whole, the impact of sample size on the $\mathcal{R}$-index for bias and

its variability was not apparent. For two of intercept/slope parameter estimates, and the estimation σ^2 in the first component, the median value of $\mathcal{R}$ decreased with the size of the sample. All in all, varying sample sizes did not cause a distinct trend on the IQR of the index, ith the exception of estimation σ^2 in the second component, which seemed to decrease with as the sample size increased.

Referring to the index for variance component presented in the middle panel of Figure 3.10, in all sample sizes and the estimated parameters, the median value of $\mathcal{R}$ index, were under the dashed line. In most instances, the ± 1 IQR bars of the $\mathcal{R}$-index for the variance were below the dashed horizontal line with exemptions to this was mixing proportion (estimation of α_1) where 80% of IQR lines were marginally higher than the 0.5 line. Generally, the median values of the $\mathcal{R}$ index for the variance declined with the increase of the sample size.

Finally, examining the $\mathcal{R}$ index of MSE at the bottom panel of Figure 3.10, as MSE is the sum of the bias squared and variance the results depicted are reflective of the above two. For the most part, the median values of $\mathcal{R}$ and the ± 1 IQR bars were beneath the 0.5 thresholds with the exception of the estimated parameter of mixing proportion where the median values of $\mathcal{R}$ were greater than the 0.5 line, along with the majority of occasions of the ± 1 IQR bars which were above the dashed line. Pertaining to the variability of $\mathcal{R}$ for MSE, it is interesting to note that the high index variability compared to the same index for the MSE in the stratified and cluster sample design. We speculate that this might be due to the complex survey design, which was used to select the sample in the simulation study. Based on the three summaries, it can be deduced that while it is unclear which model better performs in terms of MSE, the bias in parameter estimates acquired using the weighted model is lower than the unweighted model.

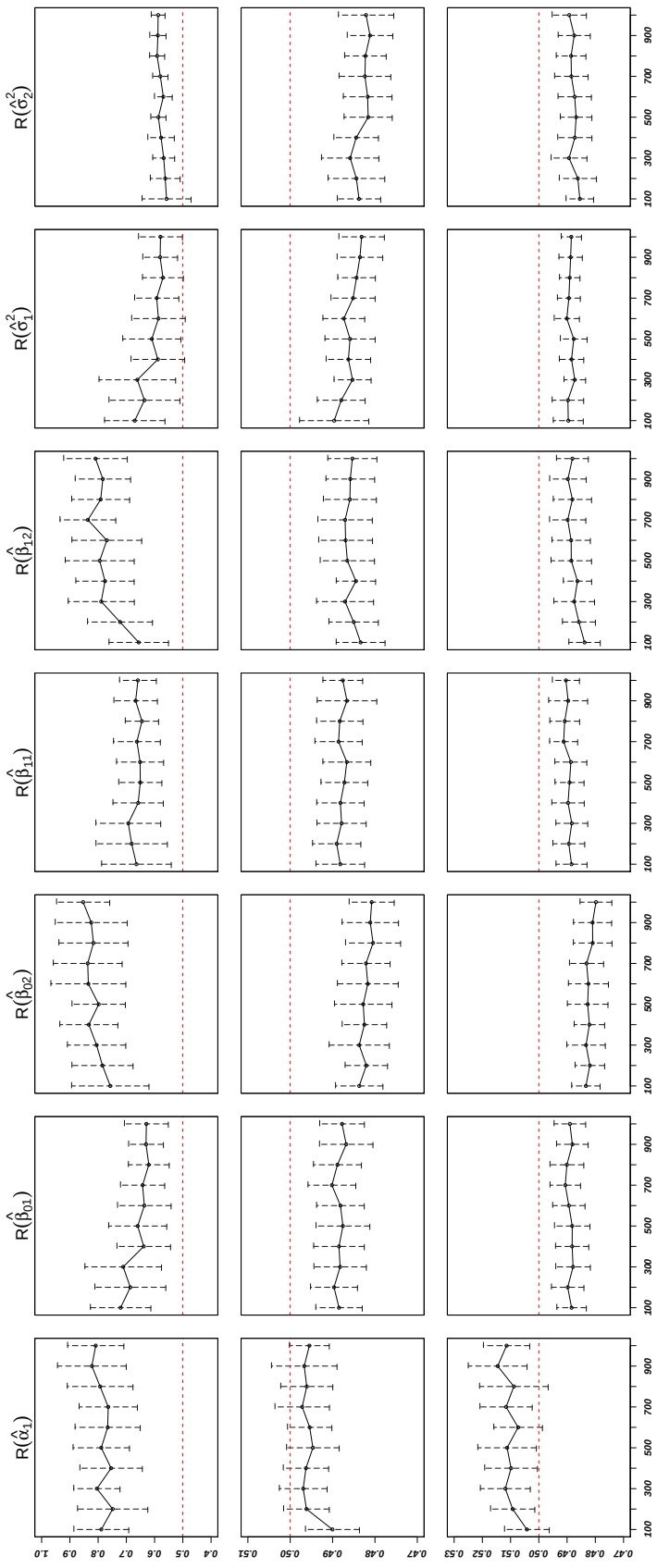


Figure 3.10: Summary of the $\mathcal{R}$ index for bias (top row), variance (middle row), and mean squared error (bottom row) in Simulation study 2. The middle line represents the median and the dashed bars represent the interquartile range (IQR) values at different sample sizes n.

3.3.3 Simulation 3: Model Selection of Complex Sampling Data

Many scenarios of the simulation were implemented by varying the number of components K and the parameters setup. Again, A stratified two-stage cluster sample design was considered. A finite population was stratified into eight strata including $\{4000, 2500, 3500, 1500, 2800, 2400, 3000, 2300\}$ observations in each stratum. The vector of specified value of $\boldsymbol{\Psi} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2)$ for each scenario which used to generate the mixtures are shown in Table 3.12. In the first stage, one cluster was drawn from each stratum using a simple random sample without replacement as PSU's. Therefore, eight clusters were sampled in this case. In the second stage, simple random samples without replacement of sizes 125 observation per cluster. Thus, the total of $n = 125 \times 8 = 1000$ observations was selected. Considering the proposed model, the performance BIC was assessed as a model selection tool. Four configurations of the true regression line were considered which are called: Mixture 1, Mixture 2, Mixture 3, Mixture 4, respectively. In Mixture 1, there are two components $(K = 2)$. In Mixture 2, We considered three mixture components $(K = 3)$. In Mixture 3, the population has four components $(K = 4)$. In Mixture 4, in this case, the population has five components $(K = 5)$. Figure 3.11 shows the scatter plots arising from the four scenarios when stratified two-stage cluster sample design was considered. After generating data, the weighted model is fitted for different values of K ranging from 1 to 10. The BIC is then calculated for each K. Figure 3.12 shows results of BIC and the optimal number of components in the different mixture settings of this simulation., including the BIC values for all K and the optimal number of components in the four experiments above. According to the results, BIC was able to choose the optimal number of components under various circumstance. In all four cases, BIC was the lowest at the true K value.

Table 3.12: True parameter values for Mixtures of linear regression in Simulation 3.

	ψ									
K	$\boldsymbol{\alpha}_1$	$\boldsymbol{\alpha}_2$	$\boldsymbol{\alpha}_3$	$\boldsymbol{\alpha}_4$	$\boldsymbol{\beta}_{10}$	$\boldsymbol{\beta}_{20}$	$\boldsymbol{\beta}_{30}$	$\boldsymbol{\beta}_{40}$	$\boldsymbol{\beta}_{50}$	$\boldsymbol{\beta}_{11}$
2	0.46				-3	3				3
3	0.34	0.42			-4	-1	4			3
4	0.21	0.28	0.22		-5	-1	1	5		1
5	0.20	0.24	0.21	0.21	-10	-4	0	4	10	3
K	$\boldsymbol{\beta}_{21}$	$\boldsymbol{\beta}_{31}$	$\boldsymbol{\beta}_{41}$	$\boldsymbol{\beta}_{51}$	$\boldsymbol{\sigma}_1^2$	$\boldsymbol{\sigma}_2^2$	$\boldsymbol{\sigma}_3^2$	$\boldsymbol{\sigma}_4^2$	$\boldsymbol{\sigma}_5^2$	
2	1				0.1	0.1				
3	1	-1			0.1	0.1	0.5			
4	-1	1	-1		0.1	0.1	0.3	0.2		
5	-3	1	1	-3	0.2	0.1	0.4	0.1	0.3	

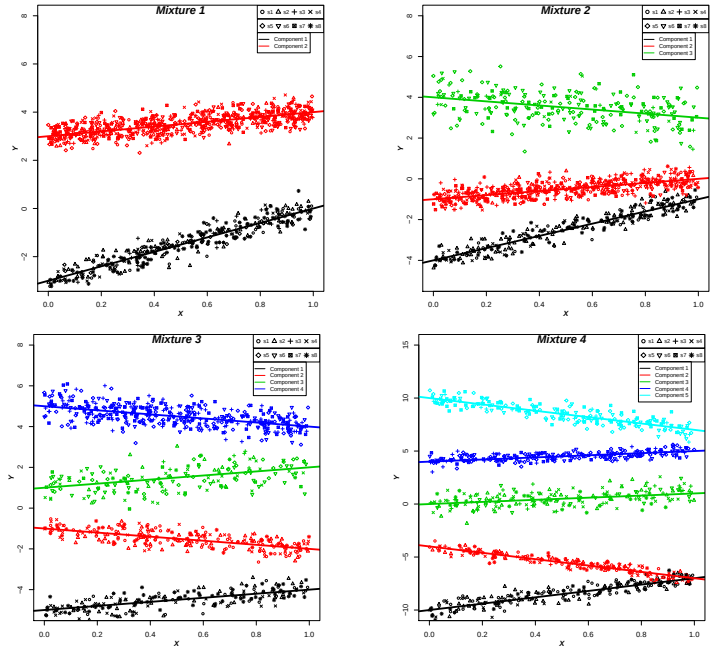


Figure 3.11: Scatter plots of samples selected from finite populations for the four experiments in Simulation 3. Colors show the number of components, and plotting characters represent the eight strata as primary sampling units from where the sampled units were drawn.

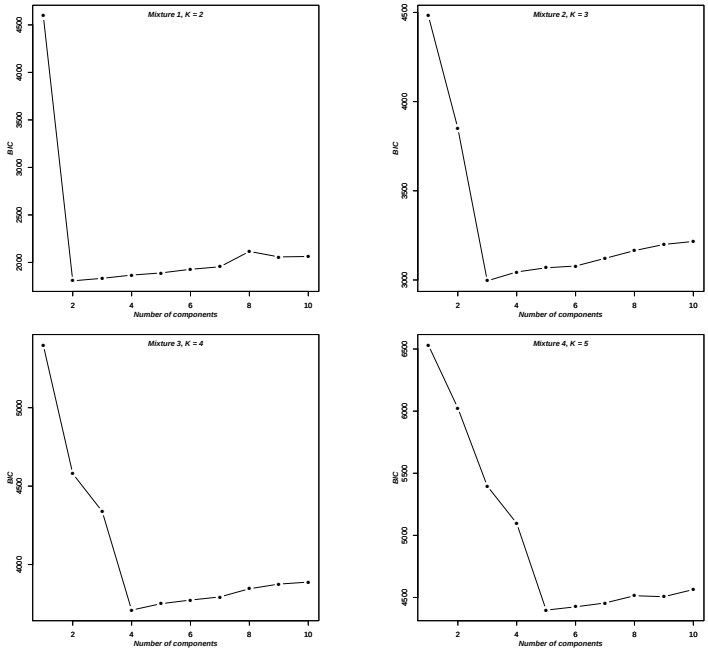


Figure 3.12: BIC values corresponding to the optimal number of components for the four experiments.

4 APPLICATIONS

Chapter 3 considered the applications of the new methodology and techniques to samples drawn from finite populations simulated using a mixture of regression models. Making use of samples from the simulated finite population made it possible to evaluate the proposed model. However, there is a necessity that exists to apply the methodologies of the proposed model on real data and outcomes made to evaluate their performance further. In this chapter, the proposed modeling is implemented for some real datasets. The datasets identified for this purpose are the Academic Performance Index (API) for the students' schools in California and the public data from The National Health and Nutrition Examination Surveys (NHANES), which have been conducted by the US National Center for Health Statistics (NCHS).

4.1 Application to Data from Stratified Sampling Design

This section is dedicated to an application of the proposed model to a real dataset. Our study focuses on the Academic Performance Index (API) dataset. The API is a measurement of academic performance and progress of individual schools in California, United States. This dataset is also available for use in an R package survey (Lumley, 2004). The dataset contains 6,194 observations on 37 variables which provide information for all schools in California with at least 100 students.

4.1.1 Example 1: Academic Performance Index

In this study, we used the variable called $stype$, which indicates the types of school (elementary/middle/high school, for stratification to produce more precise sample estimates, the individual strata should be internally homogeneous and different from one another.

Subsequently, we fitted the mixture regression model for the academic performance index in 2000 ($api00$) as the response variable and percent of parents who were high school graduates, hsg, as a predictor. Then, the parameter estimates are determined based on the proposed approach. A sample with size $n = 750$ observations was selected using stratified sampling design. We implemented pseudo-maximum likelihood procedure for the selected sample, to estimate and deduce the features of the hidden inference associated with the relationship between the response and the explanatory variable. We fitted various finite mixture of Gaussian polynomial regressions for this dataset. Table 4.1 reports the BIC corresponding to the different scenarios that were implemented by varying the number of components K and the polynomial degree r of the independent variable hsg for a mixture regression of linear regression models. According to BIC, the best model was found to be with $K = 2$ as the number of components and $r = 2$ for this part of the dataset. Estimates of the regression parameters for a mixture of quadratic Gaussian regressions have been reported on Table 4.2. Figure 4.1 shows the fitted regression model for the $api00$ on hsg. The first component contains 27% of the total observations in the sample. The expected value of API is about 567 when $hsg = 0$. On average, for each one percent increases in, hsg, the API will decrease by $-3.92 + 2(0.055)hsg$, which corresponds to the first derivative of a quadratic model. The API is declining by about 4 scores when the $hsg = 0$. Moreover, the curve of API scores is slowly decaying by increasing in hsg until the hsg is approximately 36 which represents 87% of observations of hsg and then it is eventually growing when the hsg increases, and that reflect the various behavior of the API of the students in this component. In the second components, there are approximately 73% of the observation in the sample. The expected value of API is about 894 when $hsg = 0$. When the percent parents who graduated from high school, hsg increased by one unit, the expected API changes by $-11.80 + 2(0.14)hsg$, which corresponds to the first derivative of a quadratic model. The API is decreasing by about 12 scores when the $hsg = 0$. Furthermore, the curve of API scores is decreasing in a lower rate as the hsg increases until the hsg is approximately 42

which represents 93% in total of observations of hsg, and it shows a slight increase as the hsg increases.

Table 4.1: BIC values for combination of number of components K and degree of the polynomial r in Example 1. Bold font represents the lowest BIC obtained indicating the best fit.

		r		
		1	2	3
K	1	9259.619	9088.472	9168.37
	2	9231.609	**9073.795**	9149
	3	9282.754	9119.125	9195.36
	4	9311.306	9160.060	9209.54

Table 4.2: Estimated parameters for the mixture regression model for the data in Example 1.

$\hat{\psi}$								
$\hat{\alpha}_1$	$\hat{\beta}_{10}$	$\hat{\beta}_{20}$	$\hat{\beta}_{11}$	$\hat{\beta}_{21}$	$\hat{\beta}_{12}$	$\hat{\beta}_{22}$	$\hat{\sigma}_1$	$\hat{\sigma}_2$
0.27	566.56	894.40	-3.92	-11.80	0.055	0.14	58.52	66.84

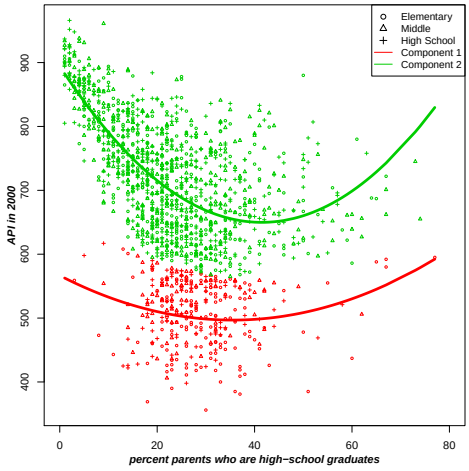


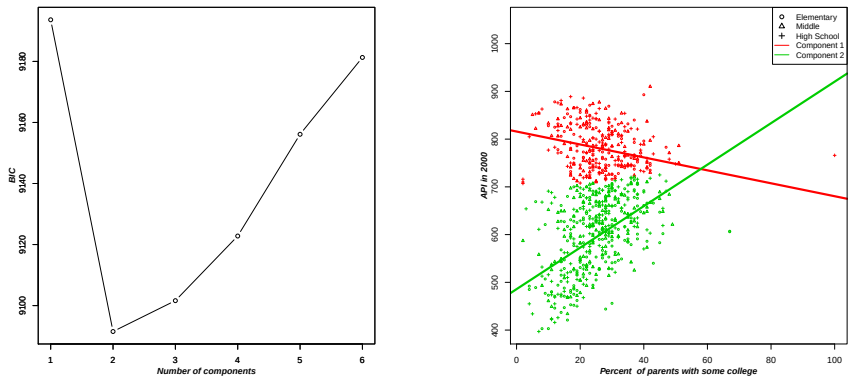
Figure 4.1: The plot shows the best-fitted mixture regression model with a 2-component quadratic Gaussian regressions model to regress the academic performance index in 2000 for the students on percent parents who are high-school graduates

4.1.2 Example 2: Academic Performance Index

We fitted the proposed model for the API in year 2000 $api00$ as a response variable with the percent of parents with some college, $some.col$ as the independent variable. For the mixture regression model of regress the API in 2000 for students on percent of parents with some college, $some.col$. Various finite mixture of polynomial regression models has been fitted. Among these models, we sought a model with a small BIC. The best model was found to be a linear regression with $K = 2$ for this dataset with the smallest BIC value. The BIC corresponding to each of the linear regression fits are presented in Figure 4.2(a). The resulting mixture is given in Figure 4.2(b). The corresponding parameter estimates are provided in Table 4.3. It can be seen that for the first component (red), which consisted of 37% of the observations in the sample, and the average of API was about 816. The API decreases 1 unit for each unit increase in the percent of parents with some college. In the second component, there was approximately 63% of the the sample, and the component had students with an average API of approximately 486 where $some.col = 0$. Their API score increased by 4 units for each unit increase in the percent of parents with some college. In component 1, the conclusion is that on average a student whose parents have lower percent with some college tends to report a higher API score. On the other hand, for component 2 on average a school which has have high percent of parents with some college tended to report a higher API score. The association between API and $some.col$ in first component is not very intuitive and may be indicative of other confounded variable that is not captured.

Table 4.3: Parameters estimated for the mixture regression model with the response the academic performance index in 2000 for the students and the percent of percent parents with some college as explanatory variable.

$\hat{\boldsymbol{\psi}}$						
$\hat{\alpha}_1$	$\hat{\beta}_{10}$	$\hat{\beta}_{20}$	$\hat{\beta}_{11}$	$\hat{\beta}_{21}$	$\hat{\sigma}_1$	$\hat{\sigma}_2$
0.37	815.90	485.80	-1.36	4.35	49.83	72.28



(a) (b)

Figure 4.2: Plots show (a) BIC versus to the number of components, for the mixture regression model of regress the API in 2000 for students on percent of parents with some college, (b) the fitted mixture regression model with a 2-component for the same dataset.

4.2 Application to Data from Cluster Sampling Design

In the previous sections, the proposed model was fitted to stratified sampling selected from real survey data. This section explores an application of the proposed model to cluster sampling data selected from real survey data. We used the public data from the National Health and Nutrition Examination Surveys (NHANES), which were conducted by the U.S. National Center for Health Statistics (NCHS). They are designed to provide national data on health, disease, and dietary and clinical risk factors gained from clinical examinations as well as detailed interviews (Centers for Disease and Prevention, 2013-2016). In this work, we consider the 2013-2014 and 2015-2016 waves of NHANES participants in the clinical exams and dietary questionnaires. The subjects in NHANES who had complete data on a selected set of variables were treated as a finite population (Li and Valliant, 2015).

The data used in this research comes from the National Health and Nutrition Examination Survey (NHANES) program office. Information regarding NHANES can be found on the CDC website. NHANES data can be used as an analysis program designed to assess the health of adults and children within the United States. For many years, the results from the NHANES data have been used as essential indicators for many studies and research

institutes. The survey is one in every of the sole to mix each survey queries and physical examinations.

The R function for reading data in these formats in the R is haven package (Wickham and Miller, 2019). This package is part of the R distribution but is not automatically loaded into memory when R starts. To load this package from the package library, we need to type library(haven). When the package is loaded, all its functions and help pages become available. The functions reade_xport(), will read SAS XPORT files. This function takes a file name as the first argument. In this work, the 2013-2014 and 2015-2016 waves of NHANES data were imported as SAS XPORT files, and we prepared them for our data analysis.

4.2.1 Example 1: Mixture of linear regression models for Systolic Blood pressure

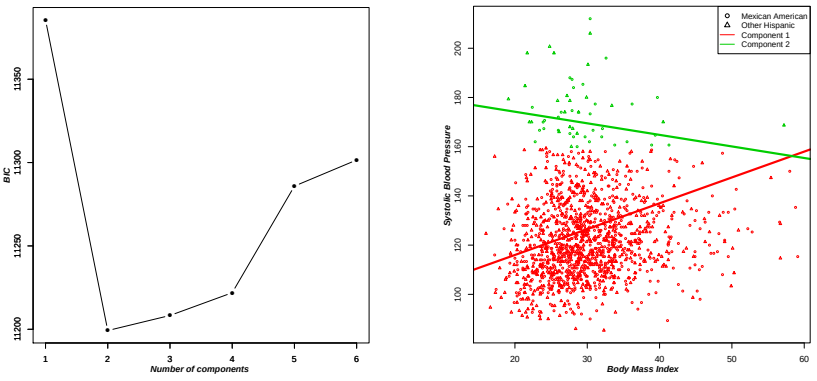
After preparing the dataset, a population of 5053 individuals was readied for analysis. The one-stage cluster sampling design was considered in this example, with race/ ethnicity as a cluster design variable. A random cluster sample of two clusters of 1314 observations was selected. The selected individuals in the sample belong to a Mexican-American cluster and a Hispanic cluster. In this example, we consider the mixture of linear regression models from the NHANES dataset for systolic blood pressure (SBP) as the response variable and body mass index (BMI) as a predictor variable. Numerous mixture linear regression models were fitted. The best model was found to be a linear regression with two components, $K = 2$ with the lowest BIC value.

The BIC corresponding to each of the linear regression fits are presented in Figure 4.3(a). Figure 4.3(b) illustrates a 2-component linear regression model to fit SBP on BMI as a predictor. The corresponding estimated parameters are provided in Table 4.4. It can be seen that, for the first component (red) which consisted of 80% of the observations in the sample, the average systolic blood pressure was about 95 mm Hg. The results indicate that

for each 1 unit increase of BMI, on average, SBP increases by 1.05 mm Hg. In the second component (green), with 20% of the individuals, the systolic blood pressure average of the individuals was approximately 183.6 mm Hg where the body mass index was equal to 0. Their systolic blood pressure decreased by 0.47 mm Hg for each unit increase in BMI level. In component 1, the conclusion is that, on average, an individual who is overweight tends to report a high systolic blood pressure. On the other hand, for component 2, on average, the individuals tend to report a low high systolic blood pressure by increasing their weights. The association between SBP on BMI in this component is not very intuitive and may be indicative of another confounding variable that is not captured.

Table 4.4: Estimated parameters for the mixture regression model with the response variable systolic blood pressure and the body mass index as explanatory variable.

$\hat{\boldsymbol{\psi}}$						
$\hat{\alpha}_1$	$\hat{\beta}_{10}$	$\hat{\beta}_{20}$	$\hat{\beta}_{11}$	$\hat{\beta}_{21}$	$\hat{\sigma}_1$	$\hat{\sigma}_2$
0.80	95	183.6	1.05	-0.47	14.71	26.72

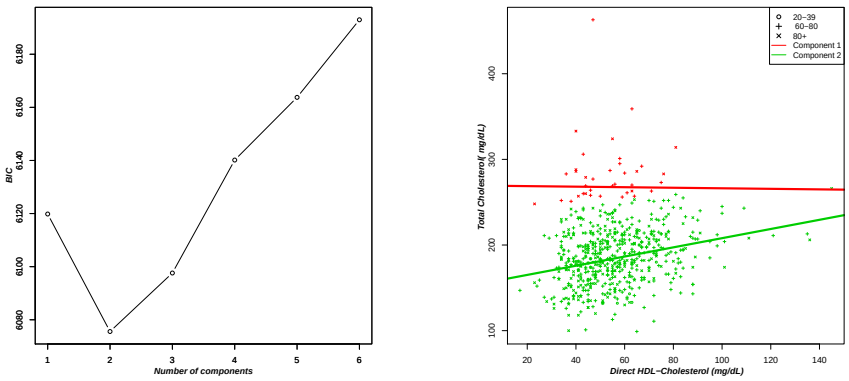


(a) (b)

Figure 4.3: Plots show (a) BIC versus to the number of components, for the mixture regression model of regress systolic blood pressure on the body mass index level (on the left), (b) the fitted mixture regression model with a 2-component for the same dataset (on the right).

4.2.2 Example 2: Mixture of Linear Regression Models for Total Cholesterol

A finite population of 4988 individuals was prepared for analysis. A two-stage cluster sampling design was considered in this example. We considered race/ethnicity and age as the design variables. In the first stage, race/ethnicity was used to divide the population into overlapping clusters. Next, a cluster sample of three clusters was chosen: Mexican-American, Non-Hispanic Black, and Non-Hispanic Asian. In the second stage, the age variable was used as a design variable. Thus, simple random samples of one age category were selected from the PSU's which were sampled in the first stage. Then, we fitted numerous mixture regression models using total cholesterol (TCHOL) as the response variable, with HDL-cholesterol (HDL) as a predictor variable, to estimate the hidden inference between the two variables. The mixture linear regression models were then fitted. Regarding the fitted model, BIC suggests the two-component $K = 2$ of the finite mixture of the regression model with the lowest BIC value. The BIC corresponding to the different finite mixture linear regressions were fitted and are presented in Figure 4.4(a). Figure 4.4(b) illustrates a 2-component linear regression model to fit TCHOL with HDL as a predictor. The corresponding estimated parameters are provided in Table 4.5. Note that for the first component (red), which consisted of 9% of the observations in the sample, the average total cholesterol level was 270 mg/dL. The total cholesterol level on average decreases 0.03 mg/dL for each unit increase in the HDL level. In the second component (green), with approximately 91% of the individuals, the average total cholesterol level of the individuals was approximately 154 mg/dL where the HDL level was equal to 0. Their total cholesterol level, on average, increased by 0.53 mg/dl for each unit increase in the HDL level. Therefore, in component 1, a person who has a low HDL level tends to report a higher level of total cholesterol. On the other hand, in component 2, there is no apparent effect on the HDL level on the TCHOL and may be indicative of another confounding variable that is not captured.



(a) (b)

Figure 4.4: Plots show (a) BIC versus to the number of components, for the mixture regression model of regress the total cholesterol on the direct HDL-cholesterol (on the left), (b) the best fitted mixture regression model with a 2-component of the same dataset (on the right).

Table 4.5: Estimated parameters for the mixture regression model when regress the total cholesterol on the direct HDL-cholesterol.

$\hat{\boldsymbol{\psi}}$						
$\hat{\alpha}_1$	$\hat{\beta}_{10}$	$\hat{\beta}_{20}$	$\hat{\beta}_{11}$	$\hat{\beta}_{21}$	$\hat{\sigma}_1$	$\hat{\sigma}_2$
0.09	269.4	154.4	-0.03	0.53	42.8	29.60

4.3 Application to Data from Complex Survey Design

In the previous application sections, we assessed the utility of the proposed model to real datasets collected through a stratified and complex survey. In this section, we aim to assess how well the proposed method is able to retrieve an underying subgroups. In this case, the public data from the NHANES 2013-2014 and 2015–2016 were used. The NHANES subjects with complete data on a selected set of variables were treated as a finite population.

4.3.1 NHANES Dataset

Our analysis included people who are 18 years of age or older within the NHANES 2013-2016 population. The whole population consists of 2772 observations. After the dataset was prepared and cleaned, the total was $N = 2402$ individuals. Suppose now that one is interested in regressing a response variable $\mathbf{y}$ against a given set of independent variables. For example, Harlan et al. (1985) has fitted regression models to NHANES data with systolic blood pressure as a dependent variable (SBP) and body mass index (BMI), age, and blood lead levels (BLL) as independent variables. Additionally, this example was used by (Li and Valliant, 2015) in a linear regression analysis for data from NHANES. In this section, we fitted the finite mixture of multiple regression models for NHANES data 2013-2016 with the response variable SBP on BMI, age, and BLL as predictor variables.

The dataset consists of two two-year waves of the new (continuous) NHANES data. Thus, it is necessary to download the data on demographics (age, sex, education level, and ethnicity), anthropocentric information (height, weight, and body mass index (BMI)), and blood pressure) for both NHANES 2013-2014 and NHANES 2015-2016, then extract the appropriate variables and merge the datasets. It was also necessary to compute the average of the multiple blood pressure measurements that are provided in the data. The sampling weights also need to be adjusted for the combined data. Since each wave of analysis is weighted to correspond to the full United States population, the combined data represents two copies of the population. Therefore, a new sampling weight variable was created by halving the original weight that is recommended for analysis of complex survey datasets such as NHANES data (Lumley, 2011). Moreover, the weights are created in NHANES to account for the complex survey design, survey non-response, and post-stratification. Let w_{NH} denote the sampling weights included in NHANES data. When a sample is weighted in NHANES, it is representative of the U.S. Census civilian non-institutionalized population. A sample weight is assigned to each sample person, which denotes the number of people in the population represented by that sample person. Throughout this section, we

will assume that a sampling weight of any observation is assigned a weight that is equivalent to the reciprocal of its probability of selection.

4.3.2 Approach

The question that surfaces is whether the proposed method will be able to retrieve the underlying subgroups in a population that heterogeneity was not accounted for by the sampling design. In most clustering methodology development, it is common to use classification dataset to assess performance of a method. However, to our knowledge there is no population level classification dataset. To overcome this problem, we used treated the NHANSE data as population data and determined underlying subpopulation using our method. Then samples are taken from the data using complex survey design and the model is fitted. Then we assessed how well the underlying groups are recovered by looking at co-occurrence of observations in the sample as compared to the population. The steps are given in more detail below.

In step 1, numerous finite mixtures of multiple regression models were fitted to the whole dataset. The sampling weights included in NHANES, w_{NH}, were used in this approach. The best model was found to be a multiple regression model with three components, $K = 3$, with the lowest BIC value. The BIC corresponding to each mixture of the linear regression fits is presented in Figure 4.5. Figure 4.6 displays the plots modeling systolic blood pressure versus the three auxiliary variables using the finite mixture of multiple regression models. The classification solution, based on $K = 3$, was used to construct a $N \times N$ co-occurrence matrix.

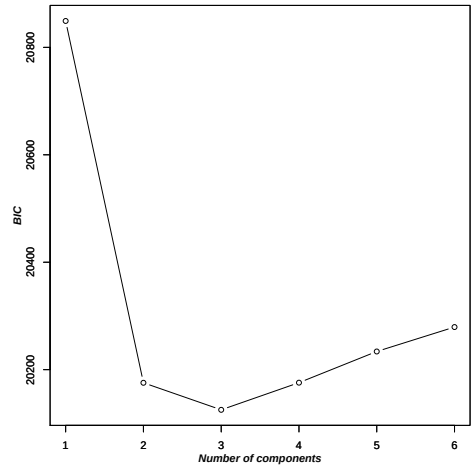


Figure 4.5: The plot shows BIC versus the number of components, for a mixture of multiple regression models of regress the systolic blood pressure on the body mass index, age, and blood lead levels.

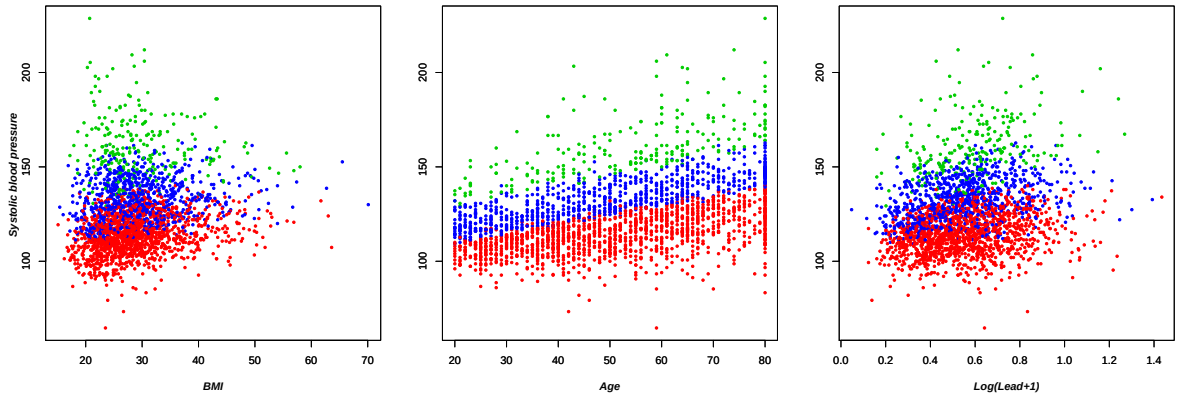


Figure 4.6: plots show the best-fitted mixture of multiple regression models with a 3-component to regress the systolic blood pressure versus three auxiliary variables for NHANES data.

Data: Matrix of finite population X; Design Variables v_1, v_2, v_3

Result: Complex Sample

Step 1;

Use v_1 to divide X into $h = \{1, ..., H\}$ strata;

Step 2;

for *each* $h \in H$ **do**

  Use v_2 to cluster each h into N_h PSU's;

  Select SRS of n_h PSU's form N_h PSU's;

  Use v_3 to cluster each i sampled PSU into M_{hi} SSU's;

  for *each* $j \in M_{hi}$ **do**

    Select SRS of m_{hi} SSU's form M_{hi} SSU's;

  end

end

Algorithm 1: The algorithm presents a stratified two-stage cluster design which uses to selecting a complex sample from a finite population.

In step 2, a sample was drawn from the NHANES 2013-2016 finite population using a stratified two-stage sample design, as described in algorithm 1. Since we re-sampled from the NHANES dataset, new sample weights needed to be constructed. Let w_{ij} be the sampling weights computed by using the stratified two-stage sample design and w_{NH} be the weight associated with the NHANES data, thus, two kinds of sampling weights were used in this step. Lohr (2010) and Thomas et al. (2006) have suggested combining these weights into one weight by multiplying them. Then, the overall sampling weight for an observation unit is $w_i = w_{ij} \times w_{NH}$. The new weights were used in the inference of the proposed model. The sampling was done $B = 200$ times with samples sizes $n_b, b = 1, \ldots, B$. The sample sizes varied depending on the sampling design. The steps to sampling from the NHANSE data are described in the pseudo-code given in Algorithm 1. In our specific application, the gender, the education level, and the race were used as v_1, v_2, and v_3, respectively. After

we selected the sample, the finite mixture of multiple regression model was fitted to the selected sample to find underlying subgroups. This was then used to obtain co-occurrence of pairs of observations forming $n_b \times n_b$.

The previous setup was repeated two hundred times. Contrary to the findings in step 1, which suggested that 3 components may be the best solution for the given dataset, step 2 demonstrated two different solutions a 2-component and a 3-component of the finite mixture of multiple regression models. Overall, in the majority of cases, the best solution is a 3-component of a finite mixture of multiple regression models. The best model was found to be a linear regression with a 3-component, $K = 3$, in approximately 72% of drawn samples, and a 2-component solution, $K = 2$, in 28% of samples. Thus, a set of two hundred $N \times N$ co-occurrence matrices were constructed via the two classification solutions, $K = 2$, and $K = 3$. Therefore, two hundred $n_b \times n_B$ co-occurrence matrices are then combined to $N \times N$ weighed co-occurrence matrix by computing the proportion of times two observations are in the same group given that they were in the sample.

The co-occurrence matrices obtained in both steps were used as distance matrices, so then hierarchical clustering was considered to determine the clusters in the data. After the clusters were obtained, we found the classification solutions between the two steps at $K = 2, ..., 10$. Algorithm 2 describes the strategies used to obtain the co-occurrence matrices and to calculate the proportion of the classification solution agreement between the two steps for different k. The following describes the summary of our approach

- **Step 1**: Full data

  1.1 Find best model fit using BIC. For the considered dataset, a 3-component mixture of multiple regression model was the optimal model according to BIC.

  1.2 Obtain the classification solution, based on $K = 3$ solution to construct was used to construct a $N \times N$ co-occurrence matrix, $A_1 = \{a_{ij}\}_{i,j=1}^{N}$. This matrix is binary with $a_{ij} = 1$ indicating that two observations were in the same group

and $a_{ij} = 0$ that they were not classified together.

- **Step 2**: Sample data

  2.1 Use complex design to get a sample of size n_b from full data of size N. Fit the proposed model for various values of K and find the best model using BIC. For the best model find classification solution and co-occurrence matrix of size $n_b \times n_b$. Repeat this for $b = 1, \ldots, 200$.

  2.2 The 200 $n_b \times n_b$ co-occurrence matrices have been merged to obtain $A_2 = \{b_{ij}\}_{i,j=1}^{N}$, $N \times N$ co-occurrence matrix by finding the proportion, b_{ij}, computed by dividing the number of times observations y_i and y_j are in the same group by the number of times both were in a sample.

- **Step 3:** Comparison

  3.1 Perform hierarchical clustering using $\mathcal{J}_N - A_1$ as dissimilarity matrix, where $\mathcal{J}_N$ denoting an all-ones $N \times N$ matrix. Cut the tree at different values of K to obtain classification solution $\mathcal{C}_{1K}$. Similarly, use $\mathcal{J}_N - A_2$ to perform hierarchical clustering and obtain $\mathcal{C}_{2K}$.

  3.2 Compute classification solutions agreement between the two solutions $\mathcal{C}_{1K}$ and $\mathcal{C}_{2K}$ at different $K = 2, ..., 10$. The pseudo-code with more formal notation is provided in Algorithm 2.

Data: p-dimensional matrix of covariates $\mathbf{X}_{N\times p}$ and a response vector $\mathbf{Y}_{N\times 1}$

Result: Proportion of class agreement

Step 1: Given $\mathbf{Y}_{N\times 1}$ and $\mathbf{X}_{N\times p}$;

for *each* $k \in K$ **do**

  Fit the proposed model $\mathcal{M}_k$;

  obtain $BIC_{\mathcal{M}_k}$;

end

Let k' = $\text{argmin}_k \text{BIC}_{\mathcal{M}_k}$ and $\mathcal{C}_{1k'}$ the corresponding classification solution;

Use $\mathcal{C}_{1k'}$ to construct A_1 an $N \times N$ co-occurrence matrix;

Step 2: Given $\mathbf{Y}_{N\times 1}$ and $\mathbf{X}_{N\times p}$;

for $b \in B$ **do**

  Select a sample $\mathbf{y}_{n_b\times 1}$ and $\mathbf{x}_{n_b\times p}$ form the full data using Algorithm 1;

  for *each* $k \in K$ **do**

    Fit the proposed model $\mathcal{M}_{bk}$;

    obtain $BIC_{\mathcal{M}_{bk}}$;

  end

  Let k' = $\text{argmin}_k \text{BIC}_{\mathcal{M}_{bk}}$ and $\mathcal{C}_{bk'}$ the corresponding classification solution ;

  Use $\mathcal{C}_{bk'}$ to construct an $n_b \times n_b$ co-occurrence matrix ;

end

Obtain B co-occurrence matrices with $n_b \times n_b$ and combine them in one

co-occurrence matrix, A_2 of size $N \times N$

Step 3: Comparison ;

$\mathcal{J}_N - A_1$ and $\mathcal{J}_N - A_2$ as dissimilarity matrix and perform hierarchical clustering;

for $K \in$*{ 1,..., 10}* **do**

  Cut the tree K and obtain classification solutions $\mathcal{C}_{1K}$ and $\mathcal{C}_{2K}$;

  Compute classification solution agreement between $\mathcal{C}_{1k}$ and $\mathcal{C}_{1k}$;

end

Algorithm 2: The algorithm displays the steps to find the classification solution agreement between the two of approaches.

From the short review to the classification solution agreement between the two steps, key findings emerged. When the proposed model was applied to the whole available dataset, the best solution was a 3-component of mixture distributions. Conversely, both a 2-component and 3-component were the best solutions for the sample-based approach. Considering Figure 4.7, it is interesting to note that the classification agreement solution between the co-occurrence matrices was minimal, considering $K = 2$ as a solution for the dataset. The curve of agreement classification proportion suddenly increased to approximately 93% when $K = 3$ was considered. In other words, 93% of the time, the correct number of components for the best solution to the data was $K = 3$. That was not surprising, because the best solution for the whole dataset was $K = 3$. Then, the curve of proportion agreement slowly decayed after $K = 3$, until it reached $K = 4$. Lastly, the agreement classification proportion decreased by increasing the number of components. On this basis, we concluded that the proposed model was about to retrieve a concurrence of observations of sub-populations approximately 93% of the time for the correct number of components when the best solution of the data was $K = 3$.

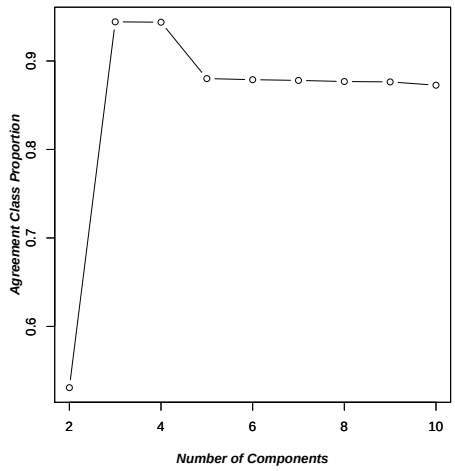


Figure 4.7: The plot shows the proportion of classification solutions agreement between step 1, and step 2 at a different number of components.

5 CONCLUSIONS AND FURTHER RESEARCH

5.1 Summary

The statistical studies based on the regression analysis of complex sampling data have received some attention over time. Notably, these data are widely used by government agencies, which are often concerned with making decisions about the target population. Finite mixture models are widely used for modeling heterogeneity in data. In the classical theory of statistics, the mixture models are estimated under the assumption that observations are drawn from the population using a simple random sampling procedure. However, this optimal assumption is not applicable. The principal aim of this dissertation is to develop and evaluate strategies for modeling data collected via a complex sample survey design using a finite mixture of regression models.

A mixture of regression models is introduced when a sample is gathered from a complex survey design. A new methodology was developed by incorporating sampling weights into the complete-data log-likelihood function, the necessary theoretical groundwork to modeling the mixture of regressions using complex survey data. Parameter estimation was carried out within the EM-algorithm framework and the BIC was used for model selection.

Extensive simulation studies were undertaken in this dissertation. Firstly, a simulation study was used to evaluate the performance of the proposed model under various circumstances. The developed model and the unweighted model were compared using the bias-variance components of the MSE. A simulation study was also conducted to compare different sampling designs, including stratified sampling, cluster sampling, and the complex sampling design based on the results obtained in the first simulation study. The mean squared error for the estimated parameters did not provide evidence significant enough to infer which estimation approach was better. However, the weighted model showed lower bias for the estimated parameters when compared with the unweighted model. Conversely,

the unweighted model had a smaller variance for most of the estimated parameters as compared with the weighted model. Overall, the variability in both models tended to decline as the sample size increased.

During further analysis, we constructed a percent contribution index that indicated how much each model contributed to the total bias, variance, and MSE. Overall, according to the percentage contribution index, $\mathcal{R}$, the proposed model estimates showed lower bias compared with the unweighted model estimates in all complex sampling strategies considered in this dissertation. In the same context, the weighted model estimates demonstrated high variability compared with the estimates obtained via the unweighted model for the majority of the estimated parameters. Regarding the stratified or cluster sampling design, the variability in this index was found to be much higher in bias than either variance or total MSE. However, the variability of $\mathcal{R}$ for MSE, it is interesting to note that the high index variability compared to the same index for the MSE in the stratified and cluster sample design. We speculate that this might be due to the complex survey design. Overall, the effect of sample size on bias and its variability was unclear. The complex sampling designs approach using a stratified, cluster, and complex sampling data were taken into account. The proposed model was then applied to the artificial dataset to assess the utility of the BIC. Here, the BIC was able to select the optimal number of components for a given dataset.

In the real data analysis, the proposed model was applied to real-world datasets. Here, different complex sampling designs such as stratified, cluster, and complex sampling design data were considered. In the real application of stratified sampling, the API scores in 2000 were regressed against the percent of parents who were high-school graduates in California schools in the first example. The optimal regression mixture model was chosen to be the one with the smallest BIC. After several models were fitted, a 2-component quadratic Gaussian regression mixture of regression model performed better than other models, with the BIC being the smallest. In the same context, when the API in 2000 for the students was regressed against percent parents with some college education. After numerous models

were fitted, a 2-component linear mixture regression model had better performance than other models, with I being the smallest.

In the application of cluster sampling, the proposed model has been applied to NHANES data. In the first example, the average of systolic blood pressure as a response variable has been regressed on the body mass index as an independent variable. The optimal regression mixture model was chosen to be the one with the smallest BIC. After several models were fitted, a 2-component mixture of linear regression models performed better than other models, with the BIC being the lowest. In the same context, when Total Cholesterol as a response variable was regressed on the HDL-cholesterol as a predictor variable. After numerous models were fitted, a 2-component mixture of linear regression models performed better than other models, with the BIC being the smallest.

In order to apply the proposed model to complex sampling, the public data from the National Health and Nutrition Examination Surveys 2013-2014 and 2015–2016 NHANES has been used in this analysis. We tried to retrieve the hidden underlying subgroups in the population based on the proposed model. Two different approaches have been assumed. In approach 1, the finite mixture of regression models was fitted to the whole dataset. The best fit model for the NHANES data, which used in this example, was $k = 3$, with the BIC being the smallest. Then, the best solution with $k = 3$ used to construct the approach 1 co-occurrence matrix. In approach 2, the mixture model has been fitted to the sample was selected based on the stratified multistage sampling design. This procedure was repeated multiple times. However, the classification solutions were the 2-component and 3-component mixture of multiple regression models. The classification solutions obtained in this approach were used to construct the sample-based co-occurrence matrix. The hierarchical clustering analysis was considered to find the proportion of the classification agreement index between the two approaches at a different value of K. The proportion of the index was 93% When the best solution was $k = 3$ for both approaches. Thus, there is evidence to suggest the 3-component mixture model solution as the best solution for the

given dataset.

5.2 Further Research

This dissertation represents a starting point in terms of building a framework for the development of strategies for modeling data collected through complex sample survey data using the mixture models. This implies that there are a number of areas for further research that might be addressed in future studies. We will try to develop the statistical inference of the mixture of regression models for complex survey data such as confidence intervals, the test of hypotheses. In addition, exploring a design-based methodology for the mixtures of linear mixed models for a complex sample might prove an essential area for future research.

A APPENDIX

A.1 Asymptotic Properties of the ML Estimators

Consider observations $\mathbf{y} = \{\mathbf{y}_1, ..., \mathbf{y}_n\}$, which are IID from a density function with parameter $\mathbf{\Psi}$ or a function of $\mathbf{\Psi}$. Part of statistical inference deals with estimation of $\mathbf{\Psi}$ and assessment of its variability. One of the most poplar methods to estimate $\mathbf{\Psi}$ to get $\widehat{\mathbf{\Psi}}$ is through M-estimators. Here, we will adapt the notation used in Van der Vaart (2000). M-estimation is done by maximizing the criterion function given by

$$M_n(\mathbf{\Psi}) = \frac{1}{n}\sum_{i=1}^{n} m(\mathbf{y}_i; \mathbf{\Psi}). \tag{A.1}$$

Here $m(\mathbf{y}_i; \mathbf{\Psi})$'s are considered to be known functions from $\mathbf{y}$ to $\mathbf{R}$. In the mathematics literature, the set of maximizing values $\widehat{\mathbf{\Psi}}$ is obtained by setting a (partial) derivative of Equation A.1 equal to zero. Therefore, in multi-parameter setting, say $\mathbf{\Psi} = \{\theta_1, \ldots, \theta_J\}$, with M total number of parameters the vector $\widehat{\mathbf{\Psi}}$ contains solutions to J systems of equations of the form $\sum_{i=1}^{n} \varphi_j(\mathbf{y}_i; \mathbf{\Psi}) = 0$, for $j = 1, \ldots, J$. That is, the M-estimator $\widehat{\mathbf{\Psi}}$ satisfies $\sum_{i=1}^{n} \varphi(\mathbf{y}_i; \widehat{\mathbf{\Psi}}) = 0$.

One of the most famous M-estimators is the Maximum Likelihood (ML) estimator. In this case, we assume that $\mathbf{y}_1, ..., \mathbf{y}_n$ have a common density function $f(\mathbf{y_i}; \mathbf{\Psi})$, then the ML estimator $\widehat{\mathbf{\Psi}}$ of $\mathbf{\Psi}$ maximizes the likelihood, $\prod_{i=1}^{n} f(\mathbf{y_i}; \mathbf{\Psi})$, or alternatively log-likelihood, $\sum_{i=1}^{n} \log f(\mathbf{y_i}; \mathbf{\Psi})$ functions. Hence, the ML estimator is an M-estimator, obtained by putting $m(\mathbf{y_i}; \mathbf{\Psi}) = \log f(\mathbf{y_i}; \mathbf{\Psi})$ in Equation A.1. On the other hand, we can also note that an M-estimator is a generalization of the ML estimator. If the density function is partially differentiable with respect to θ_j for each fixed $\mathbf{y}$, then the ML estimator also solves an $\sum_{i=1}^{n} \varphi_j(\mathbf{y}_i; \Psi) = 0$, for $j = 1, \ldots, J$, and $\varphi_j(\mathbf{y_i}; \mathbf{\Psi})$ equal to the vector of partial derivatives of the form $S_j(\mathbf{\Psi}) = \frac{\partial}{\partial \theta_j} \log f(\mathbf{y_i}; \mathbf{\Psi})$. More comprehensive details and

discussion on the general M estimation can be found in Van der Vaart (2000). Here we will focus on ML estimators.

Generally, when estimating $\mathbf{\Psi}$ using $\widehat{\mathbf{\Psi}}$ some desirable properties of the estimator are assessed. These properties include consistency, which is as $n \to \infty$, we want $\widehat{\mathbf{\Psi}}$ to converge to $\mathbf{\Psi}$ in probability as well as asymptotic normality. Here we will provide details on consistency and asymptotic normality of proposed ML- estimators under common regularity condition.

We are considering $\mathbf{y_1}, ..., \mathbf{y_n}$ to be IID from a pdf $f(\mathbf{y_i}; \mathbf{\Psi})$, and we are interested in estimating $\mathbf{\Psi}_0$. As discussed before, the ML estimator $\widehat{\mathbf{\Psi}}$ is the value of $\mathbf{\Psi}_0$ that maximizes the log-likelihood function say $\ell(\mathbf{\Psi}; \mathbf{y}) = \sum_{i=1}^{n} \log f(\mathbf{y_i}; \mathbf{\Psi})$. Under certain regularity conditions (Royall, 1986), it can be shown $\widehat{\mathbf{\Psi}}$ is consistent and asymptotically normal, with $\sqrt{n}(\widehat{\mathbf{\Psi}} - \mathbf{\Psi}_0) \rightsquigarrow \mathcal{N}(0, \mathcal{I}_n^{-1}(\mathbf{\Psi}_0))$, where $\mathcal{I}_n(\mathbf{\Psi})$ is the Fisher information given by,

$$\mathcal{I}_n(\mathbf{\Psi}) = -\mathbb{E}\left\{S(\mathbf{y}_i; \mathbf{\Psi})\, S^{\top}(\mathbf{y}_i; \mathbf{\Psi})\right\}, \tag{A.2}$$

with $S(\mathbf{\Psi})$ denoting the score function which is the vector of partial derivatives given by,

$$S(\mathbf{\Psi}) = \ell'(\mathbf{\Psi}) = \frac{\partial}{\partial \mathbf{\Psi}} \log f(\mathbf{y_i}; \mathbf{\Psi}). \tag{A.3}$$

Since $\widehat{\mathbf{\Psi}}$ maximizes $\ell(\mathbf{\Psi})$, then in general $S(\widehat{\mathbf{\Psi}}) = \ell'(\widehat{\mathbf{\Psi}}) = 0$. By consistency of ML estimator we have $\widehat{\mathbf{\Psi}} \to \mathbf{\Psi}$ in probability as $n \to \infty$. This allow us to apply a first-order Taylor expansion to the equation $\ell'(\widehat{\mathbf{\Psi}}) = 0$ around $\mathbf{\Psi}_0$ which results in $0 = \ell'(\widehat{\mathbf{\Psi}}) \approx \ell'(\mathbf{\Psi}_0) + (\widehat{\mathbf{\Psi}} - \mathbf{\Psi}_0)\ell''(\mathbf{\Psi}_0)$, then multiplying both sides by n gives

$$\sqrt{n}(\widehat{\mathbf{\Psi}} - \mathbf{\Psi}_0) = -\sqrt{n}\frac{\ell'(\mathbf{\Psi}_0)}{\ell''(\mathbf{\Psi}_0)} = \frac{\frac{\ell'(\mathbf{\Psi}_0)}{\sqrt{n}}}{-\frac{\ell''(\mathbf{\Psi}_0)}{n}} + o_p(1), \text{as } n \to \infty. \tag{A.4}$$

as $n \to \infty$.

Considering the denominator of Equation A.4, by the Weak Law of Large Numbers,

$-\frac{1}{n}\ell''(\mathbf{\Psi}_0) = \frac{1}{n}\sum_{i=1}^{n} -\frac{\partial^2}{\partial \mathbf{\Psi}^2}\big[\log f(\mathbf{y}_i, \mathbf{\Psi}_0)\big]_{\hat{\mathbf{\Psi}}=\mathbf{\Psi}_0}$ converges in probability to

$$-\mathbb{E}\left[S(\mathbf{y}_i; \mathbf{\Psi})S^{\top}(\mathbf{y}_i; \mathbf{\Psi})\right]$$

which is the Fisher information, $\mathcal{I}(\mathbf{\Psi}_0)$. For the numerator in Equation A.4, by the Central Limit Theorem, under familiar regularity conditions,

$$\frac{1}{\sqrt{n}}\ell'(\mathbf{\Psi}_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial}{\partial \mathbf{\Psi}}\big[\log f(\mathbf{y}_i, \mathbf{\Psi})\big]_{\mathbf{\Psi}=\mathbf{\Psi}_0}.$$

This can be written as $\sqrt{n}(1/n\sum_{i=1}^{n} S_i - 0)$, where $S_i = \frac{\partial}{\partial \mathbf{\Psi}}\log f(\mathbf{y_i}; \mathbf{\Psi})$. It is straightforward to show that $E(S_i) = 0$ and by definition $Var(S_i) = \mathcal{I}(\mathbf{\Psi}_0)$. Hence by CLT, $\frac{1}{\sqrt{n}}\ell'(\mathbf{\Psi}_0)$ has a limiting distribution given by $\mathcal{N}(0, \mathcal{I}(\mathbf{\Psi}_0))$. Applying these results, the Continuous Mapping Theorem, and Slutsky's Lemma to A.4, $\sqrt{n}(\widehat{\mathbf{\Psi}} - \mathbf{\Psi}_0) \rightsquigarrow \mathcal{N}(0, \mathcal{I}_n^{-1}(\mathbf{\Psi}_0))$, as desired. It can further be shown that $\sqrt{n}(\widehat{\mathbf{\Psi}} - \mathbf{\Psi}_0) \rightsquigarrow \mathcal{N}(0, \mathcal{I}_n^{-1}(\widehat{\mathbf{\Psi}}))$.

A.2 Properties of ML Estimator for Mixture Models

Now consider a mixture of regression model with an unknown vector of parameters $\mathbf{\Psi}$ with a postulated density function given by

$$g(\mathbf{y}_i, \boldsymbol{x}_i; \mathbf{\Psi}) = \sum_{k=1}^{K}\alpha_k\, f_k(\mathbf{y}_i, \mathbf{x}_i; \boldsymbol{\theta}_\mathbf{k}), \tag{A.5}$$

where $f_k(\boldsymbol{x}_i; \boldsymbol{\theta}_k)$ is a density function of kth component with an unknown parameters $\boldsymbol{\theta}_k$, and α_k is the positive mixing proportion that satisfy $\sum_{k=1}^{K}\alpha_k = 1$, $\mathbf{\Psi}$ denotes the vector for all regression parameters including α_k and θ_k, $\mathbf{y}_i$ is the response for subject i, and $\mathbf{x}$ is the predictors (McLachlan and Peel, 2000). As stated before, the ML estimator $\widehat{\mathbf{\Psi}}$ of $\mathbf{\Psi}$ for the mixture model is provided in regular situations by an appropriate solution of the

log-likelihood equation, $\partial \log \mathcal{L}(\mathbf{\Psi})/\partial \mathbf{\Psi} = 0$, where

$$\log \mathcal{L}(\mathbf{\Psi}) = \sum_{i=1}^{n} \log \left\{ \sum_{k=1}^{K} \alpha_k \, f_k(\mathbf{y}_i, \mathbf{x}_i; \boldsymbol{\theta}_\mathbf{k}) \right\}. \tag{A.6}$$

The asymptotic covariance matrix of the ML estimator $\widehat{\mathbf{\Psi}}$ is the inverse of the expected information matrix $\mathcal{I}(\mathbf{\Psi})$ which is defined in A.2, which can be approximated by $\mathcal{I}_n(\widehat{\mathbf{\Psi}})$. Thus, the asymptotic covariance matrix of $\widehat{\mathbf{\Psi}}$ is $\mathcal{I}_n^{-1}(\widehat{\mathbf{\Psi}})$. Taking a look at Expression A.6, we see that the summation over the K components blocks our log function from being applied to the mixture densities. However, the most agree that the EM-algorithm has been by far the most commonly used approach to fit the mixture distributions. in the next section, we will use the EM to obtain the ML estimators. A more comprehensive description can be found in (McLachlan and Peel, 2000).

In majority of finite mixture modeling development, the most common method of parameter estimation is through M-estimation with a major subclass being Maximum Likelihood. To accomplish this, the famous EM-algorithm is employed. The EM-algorithm attempts to find maximum likelihood estimates for models with latent variables. In mixture modelind framework, the idea is to think of the data as consisting of triples $(\mathbf{x}_i, \mathbf{y}_i, \mathbf{z}_i)$, where $\mathbf{z}_i$ is the unobserved indicator that specifies the mixture component from which the observation $\mathbf{y}_i$. Now, the complete-data log-likelihood for $\mathbf{\Psi}$ is, therefore:

$$\log \mathcal{L}_c(\mathbf{\Psi}) = \sum_{i=1}^{n} \sum_{k=1}^{K} I(Z_i = k) \left\{ \log(\pi_k) + \log(f_k(\mathbf{y}_i, \mathbf{x}_i; \boldsymbol{\theta}_\mathbf{k}) \right\} \tag{A.7}$$

At the E-step of the EM-algorithm, we use the current value of the parameters and compute the posterior probabilities, $\tau_{ik}^{(t)}$, by computing the conditional expectation of the complete-data log-likelihood function given in A.7. At the M-step, we determine the new parameter $\mathbf{\Psi}^{(t)}$ by maximizing $\mathbf{\Psi}^{(t)} = \text{argmax}_{\mathbf{\Psi}} Q(\mathbf{\Psi}, \mathbf{\Psi}^{(t-1)})$. More comprehensive details about the ML estimators of Gaussian mixture models can be found in Section 2.2.

Now, suppose that $\mathcal{L}(\mathbf{\Psi})$ denotes the likelihood function for $\mathbf{\Psi}$ formed from the ob-

served data, and $\mathcal{L}_c(\mathbf{\Psi})$ denotes the complete-data likelihood function for $\mathbf{\Psi}$ formed from the complete-data if it were completely observable. Then the complete-data score function is given by

$$S_c(\mathbf{\Psi}) = \frac{\partial \log \mathcal{L}_c(\mathbf{\Psi})}{\partial \mathbf{\Psi}}, \tag{A.8}$$

and the complete-data based information matrix denoted by $\mathcal{I}_c(\mathbf{\Psi})$ is given as

$$\mathcal{I}_c(\mathbf{\Psi}; \mathbf{y}) = \mathbb{E}\{\mathcal{I}_c(\mathbf{\Psi}; \mathbf{y}_c)|\mathbf{y}\}. \tag{A.9}$$

The expected information matrix corresponding to the complete-data is given by $\mathcal{I}_c(\mathbf{\Psi}) = \mathbb{E}\{I_c(\mathbf{\Psi}; \mathbf{y}_c, \mathbf{x}_c)\}$. The extraction of observed information matrix from complete-data log-likelihood has been shown in Louis (1982). In the paper, they showed that the information matrix $\mathcal{I}(\mathbf{\Psi})$ of the observed data is the negative of the Hessian of the log-likelihood and can be written as

$$\begin{aligned} \mathcal{I}_n(\mathbf{\Psi}; \mathbf{y}) =& \mathcal{I}_c(\mathbf{\Psi}; \mathbf{y}) - \text{cov}\{S_c(\mathbf{y}_c); \mathbf{\Psi}|\mathbf{y}\} \\ =& \mathcal{I}_c(\mathbf{\Psi}; \mathbf{y}) - \mathbb{E}\left\{S_c(\mathbf{y}_c; \mathbf{\Psi})\, S_c^{\top}(\mathbf{y}_c; \mathbf{\Psi})|\mathbf{y}\right\} + S_c(\mathbf{y}_c; \mathbf{\Psi})\, S_c^{\top}(\mathbf{y}_c; \mathbf{\Psi}). \end{aligned} \tag{A.10}$$

In Equation A.10, $S(\mathbf{y}; \mathbf{\Psi})$ and $S_c(\mathbf{y}_c; \mathbf{\Psi})$ denote the observed and complete-data score functions as defined in Equations A.3 and A.8, respectively. In addition, it has been shown that

$$S(\mathbf{y}; \mathbf{\Psi}) = \mathbb{E}\{S_c(\mathbf{y}_c; \mathbf{\Psi})|\mathbf{y}\} \tag{A.11}$$

Since that $S_c(\mathbf{y}_c; \mathbf{\Psi}) = 0$, from A.11, the observed information matrix $I(\widehat{\mathbf{\Psi}})$ can be computed as

$$\mathcal{I}_n(\widehat{\mathbf{\Psi}}; \mathbf{y}) = \mathcal{I}_c(\widehat{\mathbf{\Psi}}; \mathbf{y}) - \left[\mathbb{E}\{S_c(\mathbf{y}_c; \mathbf{\Psi})\, S_c^{\top}(\mathbf{y}_c; \mathbf{\Psi})\}|\mathbf{y}\right]_{\mathbf{\Psi}=\hat{\mathbf{\Psi}}}. \tag{A.12}$$

It can be noted that, the observed information matrix $\mathcal{I}(\widehat{\mathbf{\Psi}})$ given by expression A.12 requires the calculation of the conditional expectation of the complete-data information matrix $\mathcal{I}_c(\widehat{\mathbf{\Psi}})$ given the observed data, $\mathbf{y}, \mathbf{x}$, in addition to the complete-data score func-

tion, $S_c(\mathbf{y}_c; \mathbf{\Psi})$ times its transpose. These may be possible to compute in the simple cases such as mixtures of two univariate Gaussian densities with known common variance. However, for infeasible mixture models, it is infeasible to calculate the information matrix via Equation A.12. Hence, we can consider some practical approaches for approximating the observed information matrix (McLachlan and Peel, 2000).

In case of observed IID data $(\mathbf{y})$, an approximation of the observed information matrix can be obtained without any extra calculation. The log-likelihood can be written as $\log \mathcal{L} = \sum_{i=1}^{n} \log \mathcal{L}_i(\mathbf{\Psi})$, where $\mathcal{L}_i(\mathbf{\Psi}) = f(\mathbf{y}_i; \mathbf{\Psi})$ is the likelihood function of the observation $\mathbf{y}_i$. Then the score function can be written as $S(\mathbf{y}; \mathbf{\Psi}) = \sum_{i=1}^{n} s(\mathbf{y}_i; \mathbf{\Psi})$, where $s(\mathbf{y}_i; \mathbf{\Psi}) = \frac{\partial \log \mathcal{L}_i(\mathbf{\Psi})}{\partial \mathbf{\Psi}}$. In addition, the expectation information matrix $\mathcal{I}_n(\mathbf{\Psi})$ can be written as $\mathcal{I}_n(\mathbf{\Psi}) = n\, \mathcal{I}(\mathbf{\Psi})$, where

$$\mathcal{I}(\mathbf{\Psi}) = \mathbb{E}\left\{ s(\mathbf{y_i}; \mathbf{\Psi})\; s^{\top}(\mathbf{y_i}; \mathbf{\Psi}) \right\}$$

which is $\text{cov}\{s(\mathbf{Y}; \mathbf{\Psi})\}$ is the information contained in one observation. Now, we will define an empirical information matrix corresponding to $\mathcal{I}(\mathbf{\Psi})$ then the empirical information matrix is give by

$$\bar{\mathcal{I}}(\mathbf{\Psi}) = n^{-1} \sum_{i=1}^{n} s(\mathbf{y}_i; \mathbf{\Psi})\; s^{\top}(\mathbf{y}_i; \mathbf{\Psi}) - n^{-2} S(\mathbf{y}; \mathbf{\Psi})\; S^{\top}(\mathbf{y}; \mathbf{\Psi}), \tag{A.13}$$

then let us define $\mathcal{I}_e(\mathbf{\Psi}) = n\, \bar{\mathcal{I}}(\mathbf{\Psi})$ which is equal to

$$\sum_{i=1}^{n} s(\mathbf{y}_i; \mathbf{\Psi})\; s^{\top}(\mathbf{y}_i; \mathbf{\Psi}) - n^{-1} S(\mathbf{y}; \mathbf{\Psi})\; S^{\top}(\mathbf{y}; \mathbf{\Psi}).$$

Finally, by letting $\mathbf{\Psi} = \hat{\mathbf{\Psi}}$, $\mathcal{I}_e(\mathbf{\Psi})$ will be given by

$$\mathcal{I}_e(\widehat{\mathbf{\Psi}}) = \sum_{i=1}^{n} s(\mathbf{y}_i; \widehat{\mathbf{\Psi}})\; s^{\top}(\mathbf{y}_i; \widehat{\mathbf{\Psi}}), \tag{A.14}$$

since $S(\mathbf{y}; \widehat{\mathbf{\Psi}}) = 0$. Therefore, the covariance matrix of the estimated parameter $\widehat{\mathbf{\Psi}}$ is approximated by

$$Var(\widehat{\mathbf{\Psi}}) \approx \mathcal{I}_e^{-1}(\widehat{\mathbf{\Psi}}). \tag{A.15}$$

A.3 The Pseudo-Likelihood Approach

In the classical applications of the finite mixture of models it is assumed that the sample units are drawn from the population via a simple random sample. Hence, the IID assumption is imposed during ML- estimation procedure and variability assessment. When considering models for data collected through complex survey design, such a potentially unrealistic assumption may lead to inconsistent and biased parameter estimates. The more practical solution is to consider a complex sampling design that is based on the pseudo maximum likelihood (PML) estimation approach. The PML approach is now widely used, forming as it does the basis for the methods implemented for the analysis of complex survey data. The basic idea had its origin in Binder (1983), and the development below for the mixture approach is based on Skinner et al. (1989). The PML approach has been applied to several statistical models. However, there is limited work that has been introduced to fit the finite mixture models using the PML approach. Wedel et al. (1998) has made significant contributions to fit the finite mixture models using the PML approach.

In this section, we consider the problem of finding the asymptotic design-based sampling distribution for the parameter estimates obtained through PML approach. These are defined as functions of the data values in the finite population. Suppose that individual pairs $\{(\mathbf{x}_i, \mathbf{y}_i),\ i = 1, ..., N\}$ are generated from a probability distribution with density $g(\mathbf{y}, \mathbf{x}; \mathbf{\Psi})$. Here $\mathbf{\Psi}$ is an unknown parameter denotes the vector for all regression parameters, $\mathbf{\Psi} = \{\alpha_1, ..., \alpha_{K-1}, \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_K, \sigma_1^2, ..., \sigma_K^2\}$, and the aim is to estimate its value from the sample data.

Assume a complex sample was drawn using a complex sampling design. Suppose

that the sampling strategy is such that ith unit in the sample has a probability of being selected of π_i. The conventional estimation approach under simple random sampling is to fit parametric mixture models via maximum likelihood estimation. Assuming the the probability distribution function of the observations is given by (A.5).

The ML estimator of $\mathbf{\Psi}$ maximizes the log-likelihood. Traditionally, the standard formulation of the log-likelihood applies under simple random sampling, in which each unit receives the same weight which is defined in (A.6). The ML estimator solves the likelihood equations

$$\sum_{i=1}^{n} S_i(\mathbf{\Psi}) = \sum_{i=1}^{n} \frac{\partial \ \log f(\mathbf{y}_i; \mathbf{\Psi})}{\partial \mathbf{\Psi}} = 0. \tag{A.16}$$

Often a full ML procedure is intractable since the expression for the likelihood under the complex sampling strategy depends on assumptions about the unknown relationships between the $\mathbf{y}$ and the sample design variables. However, a simple approach is to construct a consistent estimator for $\mathbf{\Psi}$ by solving equations

$$\sum_{i=1}^{n} w_i \ S_i(\mathbf{\Psi}) = \sum_{i=1}^{n} w_i \ \frac{\partial \ \log f(\mathbf{y}_i; \mathbf{\Psi})}{\partial \mathbf{\Psi}} = 0, \tag{A.17}$$

the weights w_i are inverse proportional to the inclusion probabilities π_i. Solving the equation A.17 yields the pseudo- maximum likelihood estimator PML for the vector of parameters $\mathbf{\Psi}$ which is consistent. Inference proceeds with respect to its sampling distribution over repeated samples generated from the population by a particular sampling design (Skinner et al., 1989). We define the pseudo-log-likelihood of the sample as $p\ell_n(\mathbf{\Psi}) \equiv n^{-1} \sum_{i=1}^{n} w_i \log f(\mathbf{y}; \mathbf{\Psi})$, and we define a pseudo-maximum likelihood (PML) estimator as a parameter vector $\widehat{\mathbf{\Psi}}_n$ which solves the problem $\max_{\mathbf{\Psi}} \ p\ell_n(\mathbf{y}; \mathbf{\Psi})$.

A.3.1 Conditions for Consistency

In this appendix, we provide conditions for the the PML estimator to be consistent. According to White (1982) and Royall (1986), the conditions which are required for the

PML estimator to be consistent are as follows

Assumption A.3.1. *(White, 1982, p.2) The independent random* $1 \times M$ *vectors* $\mathbf{y}, i = 1, ..., n$*, have common joint distribution function* G *on* Ω*, a measurable Euclidean space, with measurable Radon-Nikodym density* $g = dG/dv$*.*

Assumption A.3.2. *(White, 1982, p.3) The family of distribution functions* $F(y, \boldsymbol{\psi})$ *has Radon-Nikodym densities* $f(y, \boldsymbol{\psi}) = dF(y, \boldsymbol{\psi})/dy$ *which are measurable in* y *for* $\boldsymbol{\psi}$ *in* $\boldsymbol{\Psi}$*, a compact subset of a p-dimensional Euclidean space, and continuous in* $\boldsymbol{\Psi}$ *for every* $\boldsymbol{\psi}$ *in* $\boldsymbol{\Psi}$*.*

Assumption A.3.3. *(White, 1982, p.3) (a)* $E(\log g(\mathbf{y}_i))$ *exists and* $|\log f(y, \boldsymbol{\Psi})| \leq m(y)$ *for all* Ψ *in* $\boldsymbol{\Psi}$*, where* m *is integrable with respect to* G*; (b)* $I(g : f, \boldsymbol{\Psi})$ *has a unique minimum at* ψ_* *in* $\boldsymbol{\Psi}$*, where* $\boldsymbol{\Psi}_*$ *is the parameter vector which minimizes the* $I(g : f, \boldsymbol{\Psi})$ *is Kullback-Leibler Information Criterion (KLIC), which can be defined as*

$$I(g : f, \boldsymbol{\Psi}) \equiv \mathbb{E}\big(\log \big[\log \ g(\mathbf{y}_i)/f(\mathbf{y}; \boldsymbol{\Psi}) \big] \big)$$

Theorem A.3.4. *(White, 1982, p.4) (Consistency): Given Assumptions A.3.1–A.3.3,* $\hat{\boldsymbol{\Psi}}_n \rightarrow \boldsymbol{\Psi}_*$*, as* $n \rightarrow \infty$ *for almost every sequence* $\{\mathbf{y}_i\}$*; i.e.,* $\widehat{\boldsymbol{\Psi}}_n \xrightarrow{a.s} \boldsymbol{\Psi}_*$*, where* $\widehat{\boldsymbol{\Psi}}_n$ *is a natural estimator for* $\boldsymbol{\Psi}_*$*.*

A.3.2 Conditions for Asymptotic Normality

In this appendix, we provide conditions for the PML estimator to be asymptotically normally distributed. According to White (1982) and Royall (1986), the conditions which are required for the PML estimator to be consistent are as follows

With additional conditions provided in this section, we can show that the PML estimator is asymptotically normally distributed. When the partial derivatives exist, we define the

matrices

$$\mathcal{A}_n(\mathbf{\Psi}) = \Big\{ n^{-1} \sum_{i=1}^{n} \frac{\partial^2 \ \log f(\mathbf{y}_i; \mathbf{\Psi})}{\partial \mathbf{\Psi} \ \partial \mathbf{\Psi}^\top} \Big\},$$

and

$$\mathcal{B}_n(\mathbf{\Psi}) = \Big\{ n^{-1} \sum_{i=1}^{n} \frac{\partial \ \log f(\mathbf{y}_i; \mathbf{\Psi})}{\partial \mathbf{\Psi}} \times \frac{\partial \ \log f(\mathbf{y}_i; \mathbf{\Psi})}{\partial \mathbf{\Psi}^\top} \Big\}$$

If expectations also exist, we define the matrices

$$\mathcal{A}(\mathbf{\Psi}) = \Big\{ \mathbb{E}\Big(\frac{\partial^2 \ \log f(\mathbf{y}_i; \mathbf{\Psi})}{\partial \mathbf{\Psi} \ \partial \mathbf{\Psi}^\top} \Big) \Big\},$$

and

$$\mathcal{B}(\mathbf{\Psi}) = \Big\{ \mathbb{E}\Big(\frac{\partial \ \log f(\mathbf{y}_i; \mathbf{\Psi})}{\partial \mathbf{\Psi}} \times \frac{\partial \ \log f(\mathbf{y}_i; \mathbf{\Psi})}{\partial \mathbf{\Psi}^\top} \Big) \Big\}.$$

When the appropriate inverses exist, also define

$$\begin{aligned} \mathcal{C}_n(\mathbf{\Psi}) =& \mathcal{A}_n(\mathbf{\Psi})^{-1} \ \mathcal{B}_n(\mathbf{\Psi}) \ \mathcal{A}_n(\mathbf{\Psi})^{-1}, \\ \mathcal{C}(\mathbf{\Psi}) =& \mathcal{A}(\mathbf{\Psi})^{-1} \ \mathcal{B}(\mathbf{\Psi}) \ \mathcal{A}(\mathbf{\Psi})^{-1}. \end{aligned}$$

Assumption A.3.5. *(White, 1982, p.5)* $\frac{\partial \ \log f(\mathbf{y}; \mathbf{\Psi})}{\partial \mathbf{\Psi} \ \partial \mathbf{\Psi}^\top}$*, are measurable functions of* $\mathbf{y}$ *for each* ψ *in* Θ *and continuously differentiable functions of* $\mathbf{y}$ $\forall \mathbf{\Psi} \in \Theta$*.*

Assumption A.3.6. *(White, 1982, p.5)* $|\partial \log f(\mathbf{y}, \mathbf{\Psi}) / \partial \psi \ \partial \mathbf{\Psi}^\top|$ *and* $|\partial \ \log f(\mathbf{y}; \mathbf{\Psi}) / \partial \mathbf{\Psi} \times \partial \ \log f(\mathbf{y}; \mathbf{\Psi}) / \partial \mathbf{\Psi}^\top|$ *are dominated by functions integrable with respect to* G *for all* $\mathbf{y}$ *in* Ω *and* $\mathbf{\Psi}$ *in* Θ*.*

Assumption A.3.7. *(White, 1982, p.5) (a)* $\mathbf{\Psi}_*$ *is interior to* θ*; (b)* $\mathcal{B}(\mathbf{\Psi}_*)$ *is nonsingular; (c)* $\mathbf{\Psi}_*$ *is a regular point of* $\mathcal{A}(\mathbf{\Psi})$*.*

Theorem A.3.8. *(White, 1982, p.6) (Asymptotic Normality): Given Assumptions A.3.1–A.3.7*

$$\sqrt{n}(\widehat{\mathbf{\Psi}}_n - \mathbf{\Psi}_*) \rightsquigarrow \mathcal{N}(0, \mathcal{C}(\mathbf{\Psi}_*)).$$

Moreover, $\mathcal{C}(\hat{\mathbf{\Psi}}_n) \xrightarrow{a.s} \mathcal{C}(\mathbf{\Psi}_*)$, *as* $n \to \infty$, *element by element.*

B APPENDIX

Below are some of the products of this dissertation.

Finite Mixture Regression Models for Stratified Sample

Abdelbaset Abdalla and Semhar Michael *

Despite the popularity and importance there is limited work on modeling data which come from complex survey design using finite mixture models. In this work, we explored the use of finite mixture regression models when the samples were drawn using a complex survey design. In particular, we considered modeling data collected based on stratified sampling design. We developed a new design-based inference where we integrated sampling weights in the complete-data log-likelihood function. The expectation-maximization algorithm was developed accordingly. A simulation study was conducted to compare the new methodology with the usual finite mixture of a regression model. The comparison was done using bias-variance components of mean square error. Additionally, a simulation study was conducted to assess the ability of the Bayesian information criterion to select the optimal number of components under the proposed modeling approach. The methodology was implemented on real data with good results.

Keywords: finite mixture regression models, complex survey design, sampling weights, BIC, pseudo-likelihood

This paper has been published in in the Journal of Statistical Computation and Simulation, vol 89, no 14, (2019).

*Abdelbaset Abdalla is a graduate student at South Dakota State University; Semhar Michael is an Assistant Professor of Statistics at South Dakota State University, email: Semhar.Michael@sdstate.edu

1. INTRODUCTION

Data collected from complex surveys are becoming available to researchers in a variety of fields for secondary use. Complex survey sampling design is a probabilistic sampling procedure that differs from simple random sampling. Complex surveys are typically employed on national or multinational levels in studies such as behavioral and social sciences, economics, and public health where simple random sampling is not the most practical option for collecting data. Complex survey data sets have special features which require a distinct analytical view that cannot be performed using standard methods. Hence, there is an increasing need for statistical methodology development to extract information from data collected via complex survey designs. For more details about statistical sampling techniques and analysis of complex survey data see Kish (1965), Cochran (1977), Kalton et al. (1983), Lohr (2010), a more recent issue on several aspects of survey research is published in statistical science journal (Zhang, 2017).

Most large-scale surveys use two main approaches of statistical inference; namely, design based and model-based. These approaches incorporate complexities into survey sampling including clustering, stratification, and unequal probabilities of the selection mechanism. The design based analysis was originated by Neyman (1934) and it is used in the survey sampling design context to make an inference about population parameters. In this work, we consider design based approach as an analysis tool for a given dataset collected using the stratified sampling design. Particularly, we consider the obstacle of regressing a dependent variable against a given set of predictor variables.

Assuming the survey observations are independent identically distributed (IID), regression models can be fitted using conventional methods. However, this assumption is frequently insufficient in complex survey sampling designs. More complex model assumptions are needed to fit the features of the population structure and the complex sampling design. Another way to think of this model is the census regression coefficient model, which

would be obtained from a regression if the entire population had been sampled. When inference is performed, a question that arises is whether the sampling weights should be factored in for parameter estimation. In this paper, we consider fitting finite mixture linear regression models to sample survey data by including sampling weights to the regression parameter estimators. For a detailed presentation of these and other issues regarding regression and survey weights see Pfeffermann (1993) and Lumley and Scott (2017). The parameters of the linear regression models for the complex survey design, in most cases, are derived from the pseudo-maximum likelihood (PML) approach, outlined by Skinner et al. (1989), following ideas of Binder (1983). The effect of a complex design on the identification of the underlying components from the sample using finite mixture models have been studied by (Wedel et al., 1998). They have proposed a pseudo-likelihood approach to obtain consistent estimates of parameters in the population. The paper reported that the estimates of parameters may be severely biased when using the usual maximum likelihood (ML) approach.

Survey designs through strata, cluster, or a combination of the two, aim to capture the heterogeneity in population in a less costly manner. However, sometimes subpopulations may exist after data collection. One flexible tool for modeling heterogeneity in data is through finite mixture models (McLachlan and Peel, 2000). Finite mixture regression models (Leisch, 2004; Grün and Leisch, 2008) allow simultaneously finding underlying subpopulations and building a regression model for each subpopulation in the data. For more details about analyzing a variety of the mixtures of linear regressions see Benaglia et al. (2009). Commonly, the maximum likelihood estimates of model parameters are found assuming that the data are generated through simple random sampling. However, for data that are generated through complex surveys, statistical inference based the usual likelihood approach may not be applicable. Despite the wide use of the regression analysis of survey data and finite mixture regression models, there is limited attention has been given to work

on modeling data collected with complex survey designs by using the mixture regression models. In this paper, we examine the use of regression mixture models on the samples drawn using a complex survey design.

The paper is organized as follows. Section 2 includes some preliminary concepts and the proposed methodology. In Section 3 several simulation studies are presented, and a real data application is shown in Section 4. The paper concludes with discussion and final remarks in Section 5.

2. METHODOLOGY

In this section, some necessary groundwork will be laid concerning finite mixture models and complex survey data that will be used in this paper, then the proposed methodology will be described.

2.1 Complex Survey Design

Several statistical analyses assume that the data being analyzed constitutes a simple random sample (SRS), ensuring that all elements have the same likelihood of being selected in the sample. However, sampling in survey research often works differently. In general, samples are often stratified or clustered by variables of interest. Sampling methods fall into two classifications: (1) non-probability sampling, in which the probability of being selected in the sample is unknown, and (2) probability sampling, in which the probability of being selected is known. The most common types of probability sampling are simple random sampling, cluster sampling, stratified sampling, and multistage sampling. Complex sampling, which contrasts with SRS, applies one or more unequal random selection mechanisms. The most commonly used designs involve applying stratified sampling and cluster sampling, or any combination of sampling designs. For statistical inference, considering the sampling design is imperative when studying survey sample data.

In general, we consider the regression of a dependent variable y on a vector of independent variables x. Then, (x_i, y_i) denote the row vector of these variables for a unit with label i in the index $U = \{1, ..., N\}$ of a finite population of size N. Without loss of generality, assume a general complex sampling design $p(s)$ from which sample s of size n is drawn without replacement from the population U. The sampling design may involve combinations of sampling schemes. Let δ_i be the indicator variable of the ith unit which is equal to one if $i \in s$ and zero otherwise with restriction $\sum_{i=1}^{N} \delta_i = n$. Suppose that under the sampling design a sampling unit is denoted by $i, (i = 1, ..., n)$, we can define the first-order inclusion probability, π_i, as the probability of ith unit being selected in the sample. The second-order inclusion probability, π_{ij}, is the probability that the two units i, j are selected in the sample. Thus, using the indicator variable, $E(\delta_i) = \pi_i$, and $E(\delta_{ij}) = \pi_{ij}$. The inclusion probability of the ith observation, when we use SRS is defined as, $\pi_i = \frac{n}{N}$. More discussion about the inclusion probability can be found in Horvitz and Thompson (1952), Natarajan et al. (2008) and Lohr (2010).

In this paper, we consider modeling data gathered through stratified sampling. A stratified random sample is attained by separating the population elements into non-overlapping groups which are primary sampling units (PSU), called strata. Therefore, the population is the set of strata, $\{U_h\}_{h=1}^{H}$ with sizes $N_1, ..., N_H$ and $\sum_{h=1}^{H} N_h = N$. Then, a simple random sample of size n_h is selected without replacement from each stratum with $\sum_{h=1}^{H} n_h = n$. One property of stratified sampling is that it works best when a heterogeneous population is divided into fairly homogeneous groups. Therefore, strata are to be as homogeneous as possible within, but each stratum as different as much as possible from another with respect to the characteristic being measured. We consider that a finite population U contains N units and we split this population into H non-overlapping strata. In this case, we can

define the sampling design as

$$p(s) = \begin{cases} \prod_{h=1}^{H} \binom{N_h}{n_h}^{-1} & \text{for all } n_h, h = 1, ..., H \\ 0 & \text{otherwise} \end{cases}.$$

The inclusion probability equals $\pi_i = \frac{n_{h_i}}{N_{h_i}}, i \in U_h$, where h_i is the stratum h from which units i comes (Sugden and Smith, 1984). These first-order inclusion probabilities will play a role when constructing pseudo-likelihood function.

2.2 Finite Mixture of Normal Regression

Suppose a random sample $\{(x_i, y_i),\ i = 1, ..., n\}$ of independent identically distributed (iid) observations is drawn from a finite mixture of normal regression model. Then the probability distribution function is given by

$$g(y_i|\boldsymbol{x}_i; \boldsymbol{\Psi}) = \sum_{k=1}^{K} \alpha_k \phi(y_i|\boldsymbol{x}_i\boldsymbol{\beta}_k; \sigma_k^2), \tag{1}$$

where K is the total number of mixture regression components, $\phi(y_i|\boldsymbol{x}_i\boldsymbol{\beta}_k; \sigma_k^2)$ is a Gaussian density function of the kth component with mean $\boldsymbol{x}_i\boldsymbol{\beta}_k$ and variance σ_k^2. The mixing proportions, $\alpha_k, k = 1, ..., K$ have the following restrictions: $0 < \alpha_k \leq 1$ and $\sum_{k=1}^{K} \alpha_k = 1$. Therefore, the parameter vector $\boldsymbol{\Psi}$ contains $\{\alpha_1, ..., \alpha_{K-1}, \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_K, \sigma_1^2, ..., \sigma_K^2\}$, where $\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_K, \sigma_1^2, ..., \sigma_K^2$ are the component specific regressions coefficients and variances, respectively. The common goal of statistical inference in this setting is to estimate the parameters of the model. Below we describe two estimation procedures. The first one is the traditional maximum likelihood approach which we will refer as the 'unweighted MLE' and the second one is a pseudo-maximum likelihood approach which we call the 'weighted MLE'. We assume that K is unknown, and regard it as a parameter, when performing model fitting. The matter of how best to select an appropriate K is considered as part of our model fit and model selection.

2.2.1 Unweighted Maximum Likelihood Approach

In this case, estimation of the parameters is typically performed through the maximum likelihood approach. The log-likelihood function is given by

$$\ell(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \alpha_k \phi(y_i|\boldsymbol{x}_i\boldsymbol{\beta}_k; \sigma_k^2). \tag{2}$$

Due to the inconvenient form of $\ell(\boldsymbol{\Psi})$ in Equation 2, the expectation maximization (EM) algorithm (Dempster et al., 1977), which is based on a complete-data log-likelihood function, is employed. The complete-data setup is given iid samples from $g(y_i|\boldsymbol{x}_i; \boldsymbol{\Psi})$; we define the latent variable Z_{ik} such that

$$Z_{ik} = \begin{cases} 1 & \text{if the } i\text{th observation} \in k\text{th component} \\ 0 & \text{otherwise} \end{cases}.$$

Then, we can write the complete-data log-likelihood function as

$$\ell_c(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \sum_{k=1}^{K} I(Z_{ik} = 1) \left\{ \log \alpha_k + \log \phi(y_i|\boldsymbol{x}_i\boldsymbol{\beta}_k; \sigma_k^2) \right\}. \tag{3}$$

The EM-algorithm is an iterative procedure of two steps, the Expectation (E) step, and the Maximization (M) step. At the E-step, we calculate the conditional expectation of the complete-data log-likelihood function given the observed data, $E\left(\ell_c(\boldsymbol{\Psi})|\boldsymbol{y}; \boldsymbol{X}\right)$, which simplifies to

$$E\Big(I(Z_{ik} = 1)|y_i; \boldsymbol{x}_i; \boldsymbol{\Psi}^{(t-1)}\Big) = Pr(Z_{ik} = 1|y_i, \boldsymbol{x}_i; \boldsymbol{\Psi}^{(t-1)}).$$

This posterior probability will be denoted as τ_{ik}. The expression of τ_{ik} at the (t)th iteration of the E-step is given by

$$\tau_{ik}^{(t)} = \frac{\alpha_k^{(t-1)} \phi\left(y_i|\boldsymbol{x}_i\boldsymbol{\beta}_k^{(t-1)}; \sigma_k^{2(t-1)}\right)}{\sum_{k'=1}^{K} \alpha_{k'}^{(t-1)} \phi\left(y_i|\boldsymbol{x}_i\boldsymbol{\beta}_{k'}^{(t-1)}; \sigma_{k'}^{2\,(t-1)}\right)}.$$

At the M-step of the (t)th iteration, we maximize the conditional expectation of the complete-data log-likelihood function commonly known as the Q-function given by

$$Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(t)}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik}^{(t)} \left\{ \log \alpha_k + \log \phi(y_i|\boldsymbol{x}_i\boldsymbol{\beta}_k; \sigma_k^2) \right\}. \tag{4}$$

The two steps are iterated until a predetermined convergence criterion is met. For a simple linear regression model, $y_i = \beta_{k0} + \beta_{k1}x_i + \epsilon_{ik}$, where y_i is the response variable value, x_i denotes a single explanatory variable and $\epsilon_{ik} \sim N(0, \sigma_k^2)$, Equation 4 can be written as

$$Q(\mathbf{\Psi}; \mathbf{\Psi}^{(t)}) = \sum_{i=1}^{n}\sum_{k=1}^{K} \tau_{ik}^{(t)} \left\{ \log \alpha_k - \frac{n}{2}\log(2\pi\sigma_k^2) - \frac{(y_i - \beta_{k0} - \beta_{k1}x_i)^2}{2\sigma_k^2} \right\}, \quad (5)$$

and the closed form solutions for parameters at (t)th iteration of the M-step are given by

$$\alpha_k^{(t)} = \frac{\sum_{i=1}^{n} \tau_{ik}^{(t)}}{\sum_{k=1}^{K}\sum_{i=1}^{n} \tau_{ik}^{(t)}}, \quad (6)$$

$$\beta_{k1}^{(t)} = \frac{\sum_{i=1}^{n} \tau_{ik}^{(t)} \sum_{i=1}^{n} \tau_{ik}^{(t)} x_i y_i - \sum_{i=1}^{n} \tau_{ik}^{(t)} x_i \sum_{i=1}^{n} \tau_{ik}^{(t)} y_i}{\sum_{i=1}^{n} \tau_{ik}^{(t)} \sum_{i=1}^{n} \tau_{ik}^{(t)} x_i^2 - (\sum_{i=1}^{n} \tau_{ik}^{(t)} x_i)^2}, \quad (7)$$

$$\beta_{k0}^{(t)} = \frac{\sum_{i=1}^{n} \tau_{ik}^{(t)} y_i}{\sum_{i=1}^{n} \tau_{ik}^{(t)}} - \beta_{k1}^{(t)} \frac{\sum_{i=1}^{n} \tau_{ik}^{(t)} x_i}{\sum_{i=1}^{n} \tau_{ik}^{(t)}}, \quad (8)$$

$$\sigma_k^{2(t)} = \frac{\sum_{i=1}^{n} \tau_{ik}^{(t)} \left(y_i - \beta_{k0}^{(t)} - \beta_{k1}^{(t)} x_i\right)^2}{\sum_{i=1}^{n} \tau_{ik}^{(t)}}. \quad (9)$$

Note that the Equations 7–9 are similar to solutions of least squares solutions except that they are weighted by the posterior probability from E-step.

2.2.2 Pseudo-Maximum Likelihood Estimation for Mixture Normal Regression

We assume the given data set of observations $\{(x_i, y_i, w_i); i \in s\}$, where w_i is the sampling weights. In this case, we selected a sample of size n units from a finite population of size N under some complex survey design. The most popular definition of w_i is as an indicator of the number of population units which are represented by ith sample unit. In this paper, w_i will be equal to the reciprocal of the inclusion probability π_i, which is the probability of selecting the ith sample unit under some complex survey sampling design. If such a design is considered, then standard maximum likelihood estimators are usually biased Wedel et al. (1998). Such a scenario can be avoided using the approximate, or pseudo-maximum Likelihood (PML) approach as proposed by Skinner et al. (1989) and described

by Pfeffermann (1993) and Chambers and Skinner (2003). We propose a weighted estimation procedure for finite mixture models which minimizes the bias in parameter estimates that occur when the sampling design is not taken into consideration. This is done by incorporating the sampling weights, w_i to the complete data log-likelihood function. Then the modified Q-function is given by

$$Q_w(\mathbf{\Psi}; \mathbf{\Psi}^{(t)}) = \sum_{i=1}^{n} w_i \sum_{k=1}^{K} \tau_{ik} \{\log \alpha_k - \frac{n}{2} \log(2\pi\sigma_k^2) - \frac{(y_i - \beta_{k0} - \beta_{k1} x_i)^2}{2\sigma_k^2}\}, \tag{10}$$

We refer the function in Equation 10 as the weighted Q-function and is denoted by Q_w. At the M-step of the (t)th iteration, the Q_w-function is maximized with respect to $\mathbf{\Psi}$. For the simple Gaussian mixture regression model the closed form solutions are as follows

$$\alpha_k^{(t)} = \frac{\sum_{i=1}^{n} w_i \tau_{ik}^{(t)}}{\sum_{k=1}^{K} \sum_{i=1}^{n} w_i \tau_{ik}^{(t)}}, \tag{11}$$

$$\beta_{k1}^{(t)} = \frac{\sum_{i=1}^{n} w_i \tau_{ik}^{(t)} \sum_{i=1}^{n} w_i \tau_{ik}^{(t)} x_i y_i - \sum_{i=1}^{n} w_i \tau_{ik}^{(t)} x_i \sum_{i=1}^{n} w_i \tau_{ik}^{(t)} y_i}{\sum_{i=1}^{n} w_i \tau_{ik}^{(t)} \sum_{i=1}^{n} w_i \tau_{ik}^{(t)} x_i^2 - \left(\sum_{i=1}^{n} w_i \tau_{ik}^{(t)} x_i\right)^2}, \tag{12}$$

$$\beta_{k0}^{(t)} = \frac{\sum_{i=1}^{n} w_i \tau_{ik}^{(t)} y_i}{\sum_{i=1}^{n} w_i \tau_{ik}^{(t)}} - \beta_{k1}^{(t)} \frac{\sum_{i=1}^{n} w_i \tau_{ik}^{(t)} x_i}{\sum_{i=1}^{n} w_i \tau_{ik}^{(t)}}, \tag{13}$$

$$\sigma_k^{2(t)} = \frac{\sum_{i=1}^{n} w_i \tau_{ik}^{(t)} \left(y_i - \beta_{k0}^{(t)} - \beta_{k1}^{(t)} x_i\right)^2}{\sum_{i=1}^{n} w_i \tau_{ik}^{(t)}}. \tag{14}$$

Note here that the update equations in 11–14 are similar to 6–9 except the weights are Incorporated.

2.3 Matrix Approach of Mixture Normal Regression

We can extend the mixture of simple linear regression model to multiple linear regression model. This can be done using matrix notation as follows

$$\boldsymbol{\beta}_k^{(t)} = \left(\mathbf{X}^\top \mathbf{W}_k^{(t)} \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{W}_k^{(t)} \mathbf{y}, \tag{15}$$

where $\mathbf{X}$ is an $n \times (p+1)$ matrix containing unity for intercept and predictors, $\mathbf{W}_k^{(t)}$ is a $n \times n$ diagonal matrix with entries $w_i \times \tau_{ik}^{(t)}$, $\mathbf{y}$ is a $n \times 1$ vector of response variable, and

$$\sigma_k^{2(t)} = \frac{\left\| \mathbf{W}_k^{1/2(t)} \left(\mathbf{y} - \mathbf{X}\, \boldsymbol{\beta}_k^{(t)} \right) \right\|^2}{\mathrm{tr}\left(\mathbf{W}_k^{(t)} \right)}, \tag{16}$$

where $\|\mathbf{A}\| = \mathbf{A}^\top \mathbf{A}$ with $\top$ denoting a matrix transpose and $\mathrm{tr}(\mathbf{A})$ means the trace of the matrix $\mathbf{A}$. Equations 15 and 16 can be used as update equations at at (t)th iteration of the M-step. The same equation as given in equation 11 is used to update mixing proportions.

2.4 Computational Strategies

In this section, we describe some computational strategies that have been used in fitting the proposed model. Initialization is a key step in fitting mixture models to data via the EM algorithm (Baudry and Celeux, 2015). In the simulation study, we considered two strategies for choosing initial values of parameters. In the first simulation study, we compare the weighted and unweighted models. For this, the true values of parameters were used as the starting values. This will allow for comparing without confounding the issues associated with initialization. In the second simulation study is conducted to assess the validity of Bayesian Information Criterion (BIC) as model selection criterion. For this, we used *Rnd-EM* (Maitra, 2009) to choose initial values. In this initialization method, first random points are selected as seeds and the Euclidean distance is used to assign observations to centers. This is repeated for some fixed number of times. The solution that yields the highest likelihood value is then used for initializing the EM-algorithm. *Rnd-EM* tends to work well if the number of components is not large (Michael and Melnykov, 2016). *Rnd-EM* is used to initialize the EM-algorithm for the real data analysis. In the EM-algorithm, the E-step and M-step are iterated until a convergence criterion is met. In our paper, the algorithm is stopped when the absolute relative change in the likelihood given by $\frac{\ell(\boldsymbol{\Psi}^{(t)};\mathbf{y},\mathbf{x}) - \ell(\boldsymbol{\Psi}^{t-1};\mathbf{y},\mathbf{x})}{|\ell(\boldsymbol{\Psi}^{t-1};\mathbf{y},\mathbf{x})|}$ is less than 10^{-8}. In the real dataset analysis, we used the BIC (Schwarz et al., 1978) to select

the optimal number of components. In this paper, BIC will be calculated as $BIC(\hat{\boldsymbol{\Psi}}) = -2\ell(\hat{\boldsymbol{\Psi}}) + M \log n$, where $\ell(\hat{\boldsymbol{\Psi}})$ and M represent the maximized likelihood value for a given K and the number of parameters in the fitted model, respectively. For mixtures of normal regression, $M = (K-1) + K(p+1) + K$, where p represents the number of the predictor variables. The model with lowest BIC value is the best model for a given dataset.

2.5 Identifiability

Identifiability of a given model is one of the major requirements for any model to be meaningful. It is defined for any two parameter vectors $\boldsymbol{\Psi} \neq \boldsymbol{\Psi}'$, the respective model $f(\boldsymbol{x}; \boldsymbol{\Psi})$ must be different from $f(\boldsymbol{x}; \boldsymbol{\Psi}')$ for any random vector $\boldsymbol{x}$. The identification issue for the finite mixture linear model has been and continues to be studied. In general, in the mixture regression model setting, there are two kinds of identification problems that are common. One of them is label switching, and the other is overfitting. The label switching which occurs when switching the labels of any two different components does not change the distribution of the response variable at all. Overfitting is a more fundamental lack of identifiability, and it leads to empty components or components with equal parameters. This kind of unidentifiablity can be avoided by restricting the prior mixing ratios to be greater than zero, and the component with specific parameters are different (Leisch, 2004). In this paper, to prevent overfitting, mixing proportions have been restricted to be greater than a particular threshold.

On a similar note, the identifiability of a mixture of regression models depends on the distribution of the response variable. Particularly in this setting, Hennig (2010) pointed out that identifiability issues may arise if there are solely a restricted range of values for covariates and dditionally if there is a restricted info per person accessible. Such problems might occur in applications where covariates are generally categorical variables for example race and gender (Grün and Leisch, 2004). As per Hennig (2010), the mixtures of linear regres-

sion models with Gaussian random errors are identifiable if the number of components K is smaller than the minimal number of hyperplanes necessary to cover all covariate points. In this work, we mainly focus on continuous response and covariates, but in general, one needs to be cautious of the results obtained.

2.6 Model comparison

For comparing the weighted and unweighted models, the variance-bias components of MSE are used. The MSE is obtained from the B replications as $MSE(\hat{\psi}_j) = \frac{1}{B}\sum_{b=1}^{B}(\psi_{jb} - \hat{\psi}_{jb})^2$, where ψ_{jb} and $\hat{\psi}_{jb}$, are the true and estimated parameter, respectively. The variance and bias components are given as $Var(\hat{\psi}_j) = \frac{1}{B}\sum_{b=1}^{B}(\hat{\psi}_{jb} - \bar{\hat{\psi}}_j)^2$ and $Bias(\hat{\psi}_j) = (\bar{\hat{\psi}}_j - \psi_j)^2$, where $\bar{\hat{\psi}}_j = \frac{1}{B}\sum_{b=1}^{B}\hat{\psi}_{jb}$, respectively. Note that, $MSE(\hat{\psi}_j) = Var(\hat{\psi}_j) + Bias^2(\hat{\psi}_j)$. In our setting, $\boldsymbol{\Psi} = \{\psi_j\}_{j=1}^{M}$, where M is the number of parameters in $\boldsymbol{\Psi} = \{\alpha_1, ..., \alpha_{K-1}, \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_K, \sigma_1^2, ..., \sigma_K^2\}$, and each element is represented by ψ_j.

In this paper, percent contribution, $\mathcal{R}$, is used to compute the relative contribution of a given quantity to a total amount and is calculated as: $\mathcal{R} = \frac{\theta_1}{\theta_1+\theta_2}$, where θ_1 and θ_2 are the two quantities calculated. We will use R to find out how much percentage contribution take place for two quantities we are trying to compare. Note that, this index will range between 0 and 1 and if both quantities contribute equally to the total amount then R will be equal to 0.5. Values below 0.5 indicate lower percent contribution of θ_1 as compared to θ_2 and values above 0.5 will indicate higher percent contribution of θ_1 to the total $\theta_1 + \theta_2$. In the simulation study, the MSE, and its bias and variance components will be used to compute R. This will be use to compare the performance of the weighted model with the unweighted model. If any of MSE components, bias or the variance of were equal of the compared models; then R will be equal to 0.5. Also, we have been formulated this measurement such that the components of the MSE of the unweighted approach will be on the numerator of the fraction, thus for any of MSE components if R was less than 0.5 then

the performance of the unweighted model will be better than the weighted model. On the other hand, if R was greater than 0.5, then the performance of the weighted model will be better than the weighted model.

3. SIMULATION STUDIES

3.1 Simulation 1: Parameter Estimation

This simulation study was executed to assess the performance of the maximum likelihood estimates obtained via the unweighted and weighted model in various scenarios. The criteria used for comparison include: Mean Square Error (MSE), variance, and bias. In this setting, the true values of parameters were used as the starting values. We considered two configurations of the true regression lines: non-overlapping and overlapping which we call Mixture 1 and Mixture 2, respectively. In the first simulation, we generated a finite population composed of $N = 18000$ observations from a two-component mixture of normal regression model. The finite population consists of two stratum, $\{U_h\}_{h=1}^{2}$, with $\{10000, 8000\}$ observations in each stratum. The vector of parameters $(\boldsymbol{\tau}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2)$ used to generate the mixture are reported in Table1. Stratified samples of sizes $n_1 = n_2 = \{100, 250, 500, 1000\}$ are drawn from each stratum. Thus, the total sample sizes of $n = 200, 500, 1000, 2000$ are considered. Therefore, for $n = 1000$, we have $n_1 = 500$ from the first stratum and $n_2 = 500$ from second stratum. For example for Mixture 1, with in each stratum, we use $\alpha_1 = 0.34$ and $\alpha_2 = 0.66$ to determine how many observations will belong to component one and component two, respectively. Figure 1 shows sample of size $n = 1000$ observations from the considered models Mixture 1 and Mixture 2. The above setup is repeated for $B = 1000$ replications.

For each replication, the weighted and unweighted models are fitted and parameter estimates are obtained. The true parameter values are compared with the estimated values using the MSE and its components as given in Section 2.6. Since two different methods

Table 1: True parameter values for Mixture 1 and Mixture 2.

$\boldsymbol{\psi}$	α_1	α_2	β_{10}	β_{20}	β_{11}	β_{21}	σ_1^2	σ_2^2
Mixture 1	0.34	0.66	-3	3	1	-2	0.1	0.1
Mixture 2	0.34	0.66	-3	-2	1	-2	0.1	0.1

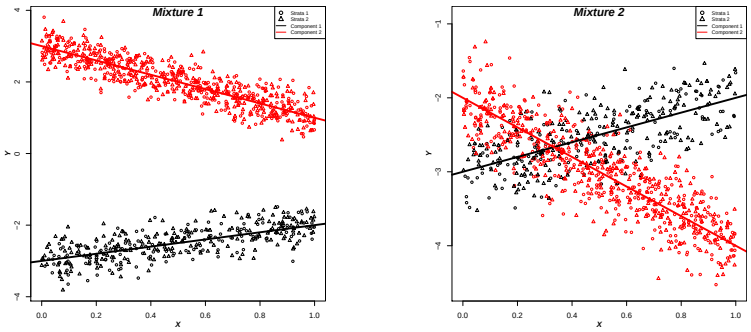


Figure 1: Scatter plots of a sample of size $n = 1000$ units. Colors show the two components and plotting characters represent strata. Left plot represents Mixture 1 - non-overlapping components and right plot represents Mixture 2 - overlapping components.

have been used to fit the model, it is necessary to evaluate their parameter recovery and to check whether accurate the variability of estimates is yielded. Parameter recovery concerns whether the weighted or unweighted models can recover the generating parameters accurately. If the empirical mean of the estimates across replications is statistically meaningfully different from the generating parameter, the estimator is thought to be biased. There is also a concern regarding the variability of the estimates across replications. If the variability is practically minor, then a slightly biased estimation is negligible. Table 2 provides the MSE and its bias and variance components for varying sample sizes when Mixture1 is considered. The bold values show where the minimum is achieved when comparing the weighted and unweighted models. Looking at the table, the estimates obtained by the weighted model have a smaller bias compared to the estimates obtained by the unweighted

Table 2: Mean square error, bias, and variance of estimated parameters, based on 1000 replications for different sample sizes of the two-component when the Mixture 1 configuration was considered.

		$\hat{\psi}$	$\hat{\alpha}_1$	$\hat{\beta}_{10}$	$\hat{\beta}_{20}$	$\hat{\beta}_{11}$	$\hat{\beta}_{21}$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$
$n = 200$	MSE $\times 10^{-2}$	Weighted	**0.1186**	0.5703	**0.2884**	1.7289	0.8935	0.0279	0.0172
		Unweighted	0.1228	**0.5636**	0.2963	**1.7133**	**0.8766**	**0.0274**	**0.0167**
	Bias^2 $\times 10^{-2}$	Weighted	**0.0001**	0.0081	**0.0019**	**0.0272**	**0.0038**	0.0019	0.0010
		Unweighted	0.0034	**0.0076**	0.0027	0.0282	0.0058	**0.0018**	**0.0009**
	Var $\times 10^{-2}$	Weighted	**0.1185**	0.5622	0.2942	1.7007	0.8886	0.0260	0.0162
		Unweighted	0.1194	**0.5560**	**0.2858**	**1.6861**	**0.8708**	**0.0256**	**0.0158**
$n = 500$	MSE $\times 10^{-2}$	Weighted	**0.0446**	0.2380	0.1265	0.7559	0.3529	0.0109	0.0061
		Unweighted	0.0492	**0.2332**	**0.1225**	**0.7413**	**0.3478**	**0.0107**	**0.0060**
	Bias^2 $\times 10^{-2}$	Weighted	**0.0001**	**0.0075**	0.0021	0.0427	0.0045	0.0008	**0.0004**
		Unweighted	0.0046	0.0089	**0.0015**	**0.0392**	**0.0032**	**0.0007**	0.0005
	Var $\times 10^{-2}$	Weighted	**0.0445**	0.2303	0.1234	0.7132	0.3484	0.0101	0.0057
		Unweighted	0.0447	**0.2243**	**0.1211**	**0.7021**	**0.3446**	**0.0100**	**0.0055**
$n = 1000$	MSE $\times 10^{-2}$	Weighted	**0.0227**	**0.1058**	**0.0567**	**0.3368**	**0.1676**	0.0057	**0.0029**
		Unweighted	0.0249	0.1093	0.0592	0.3451	0.1739	**0.0056**	0.0031
	Bias^2 $\times 10^{-2}$	Weighted	**0.0001**	**0.0067**	**0.0024**	**0.0270**	**0.0025**	**0.0005**	**0.0003**
		Unweighted	0.0024	0.0079	0.0034	0.0295	0.0040	0.0006	0.0004
	Var $\times 10^{-2}$	Weighted	0.0226	**0.0991**	**0.0543**	**0.3098**	**0.1652**	0.0052	**0.0026**
		Unweighted	**0.0225**	0.1014	0.0559	0.3156	0.1700	**0.0050**	0.0027
$n = 2000$	MSE $\times 10^{-2}$	Weighted	**0.0096**	**0.0566**	**0.0268**	**0.1842**	**0.0819**	**0.0026**	**0.0014**
		Unweighted	0.0124	0.0586	0.0284	0.1875	0.0855	0.0028	0.0016
	Bias^2 $\times 10^{-2}$	Weighted	**0.0001**	**0.0045**	**0.0029**	**0.0253**	**0.0054**	**0.0004**	**0.0002**
		Unweighted	0.0028	0.0055	0.0041	0.0275	0.0079	0.0005	0.0003
	Var $\times 10^{-2}$	Weighted	**0.0095**	**0.0522**	**0.0239**	**0.1590**	**0.0766**	**0.0022**	**0.0012**
		Unweighted	0.0096	0.0532	0.0243	0.1602	0.0777	0.0023	0.0013

model in 21 out of 28 cases. Thus, the estimates obtained by the weighted approach have a smaller bias compared with estimates obtained via the unweighted approach. The weighted model estimates have relatively high variability compared to estimates obtained via the unweighted model in 14 out of 28 cases. The variances of the estimates for both models decrease by increasing the size of a sample.

Table 3: Mean square error, bias, and variance of estimated parameters, based on 1000 replications for different sample sizes of the two-component when the Mixture 2 configuration was considered.

		$\hat{\psi}$	$\hat{\alpha}_1$	$\hat{\beta}_{10}$	$\hat{\beta}_{20}$	$\hat{\beta}_{11}$	$\hat{\beta}_{21}$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$
$n = 200$	MSE $\times 10^{-2}$	Weighted	0.1939	**0.3531**	0.2142	**1.2784**	0.7299	**0.0164**	**0.0108**
		Unweighted	**0.1797**	0.3604	**0.2018**	1.2886	**0.7016**	0.0444	0.0221
	Bias^2 $\times 10^{-2}$	Weighted	**0.0009**	**0.0040**	**0.0045**	0.0008	**0.0051**	**0.0001**	**0.0001**
		Unweighted	0.0082	0.0062	0.0046	**0.0001**	0.0058	0.0002	0.0003
	Var $\times 10^{-2}$	Weighted	0.1930	**0.3491**	0.2097	**1.2776**	0.7248	**0.0163**	**0.0107**
		Unweighted	**0.1715**	0.3542	**0.1972**	1.2885	**0.6956**	0.0442	0.0218
$n = 500$	MSE $\times 10^{-2}$	Weighted	**0.0646**	0.1354	0.1011	0.5221	0.3135	**0.0078**	**0.0044**
		Unweighted	0.0654	**0.1284**	**0.0986**	**0.4777**	**0.3038**	0.0193	0.0082
	Bias^2 $\times 10^{-2}$	Weighted	**0.0004**	**0.0001**	**0.0035**	0.0012	**0.0018**	**0.0001**	**0.0002**
		Unweighted	0.0017	0.0004	0.0039	0.0001	0.0027	0.0003	0.0003
	Var $\times 10^{-2}$	Weighted	0.0643	0.1353	0.0976	0.5209	0.3120	**0.0077**	**0.0042**
		Unweighted	**0.0638**	**0.1282**	**0.0949**	**0.4776**	**0.3013**	0.0190	0.0079
$n = 1000$	MSE $\times 10^{-2}$	Weighted	**0.0214**	0.0692	0.0490	0.2593	0.1533	**0.0036**	**0.0022**
		Unweighted	0.0283	**0.0668**	**0.0472**	**0.2435**	**0.1435**	0.0071	0.0034
	Bias^2 $\times 10^{-2}$	Weighted	**0.0001**	**0.0025**	**0.0038**	**0.0016**	**0.0046**	**0.0001**	**0.0001**
		Unweighted	0.0022	0.0032	0.0039	0.0017	0.0048	0.0002	0.0002
	Var $\times 10^{-2}$	Weighted	**0.0213**	0.0668	0.0452	0.2577	0.1488	**0.0035**	**0.0021**
		Unweighted	0.0262	**0.0636**	**0.0433**	**0.2418**	**0.1388**	0.0069	0.0032
$n = 2000$	MSE $\times 10^{-2}$	Weighted	**0.0104**	0.0304	**0.0236**	0.1114	**0.0683**	**0.0019**	**0.0011**
		Unweighted	0.0150	**0.0299**	0.0249	**0.1039**	0.0733	0.0035	0.0017
	Bias^2 $\times 10^{-2}$	Weighted	**0.0002**	**0.0011**	**0.0042**	**0.0005**	**0.0047**	**0.0002**	**0.0001**
		Unweighted	0.0013	0.0018	0.0045	0.0006	0.0054	0.0003	0.0002
	Var $\times 10^{-2}$	Weighted	**0.0102**	0.0293	**0.0194**	0.1109	**0.0637**	**0.0017**	**0.0010**
		Unweighted	0.0136	**0.0280**	0.0204	**0.1034**	0.0679	0.0032	0.0015

Table 3 provides the MSE, the bias, and the variance of the estimated parameters when Mixture 2 is considered. The estimates obtained by the weighted model have lower bias compared to the estimates obtained by the unweighted model in 26 out of 28 cases, which leads to the conclusion that the estimates obtained by the weighted approach have small bias compared with estimates obtained via the unweighted approach. The weighted model estimates have high variability compared to the unweighted model estimates in only 14 out of 28 cases. However, the variances of the estimates for both models are declined by increasing the size of a sample. Therefore, the estimates obtained via the weighted model for Mixture 2 have a lower bias in about 93% of cases compared by the unweighted model estimates in the same configuration while this percentage to about just decreased to about 78% of instances when Mixture 1 was considered. Hence, we can infer that the weighted model has better performance to reduce the bias of estimated parameters for complicated circumstances.

3.2 Simulation 2: Model comparison

To further investigate the parameter recovery capability of both approaches, we will present a diagnosis concerning the results of the first simulation study. This is done to evaluate the impact of the sample size on the parameter recovery and assess the variability associated with the MSE and its components. Based on the results obtained in the previous study, the weighted model has a lower bias in the majority of cases, yet occasionally, the unweighted model estimates have a lower bias compared by those which are obtained via the weighted model. Therefore, the simulation study has not yet determined the general features of the two approaches definitively. Here we considered the Mixture 1 setup; there is a finite population consisting of $\{10000, 8000\}$ observations in each stratum. The vector of parameters is reported in Table1. Stratified samples were drawn from each stratum at different sample sizes, starting with 50 per strata up to 500 with an increment of 50 obser-

vations. Thus, the samples that were selected are $n = \{100, 200, 300, \ldots, 1000\}$. Here we replicated $B = 100$ times for each n. These replicates are then used to calculate MSE values and the corresponding bias and variance components. Then, two hundred replications of the above set up were completed to obtain 200 values of MSE, bias, and variance values for each sample size and parameter under both the weighted and unweighted models. These 200 replicates are then used to calculate the percent contribution index, $\mathcal{R}$, defined in Section 2.6, by setting θ_1 to be the results from the unweighted model and θ_2 to be the results from weighted model. Therefore, if $\mathcal{R}$ is above 0.5, then the weighted model had contributed less to the total MSE, bias or variance. If $\mathcal{R}$ is less than 0.5 then the unweighted model has contributed less to the total MSE, bias or variance.

The results of this analysis can be found in a multiplot provided in Figure 2. The top panel of the the figure represents the bias, the middle represents the variance, and the bottom panel represents the MSE. The seven columns correspond to the seven parameters estimated in this study. With in each plot, the x-axis represents the varying sample sizes and y-axis is the $\mathcal{R}$ index. The median values of the index are represented by the black line and the dashed bars indicate ± 1 interquartile range (IQR) values of the index at each sample size. The dashed horizontal line is at 0.5, indicating a threshold for when the two methods perform equally. Considering the top panel, the median values of the R-index for bias were above 0.5 in all estimated parameters and sample sizes. In the majority of cases, the ± 1 IQR bar of the bias was above the dashed horizontal line except for few cases (estimation of σ^2) where some IQR lines were slightly below the 0.5 line. In case of the mixing proportion $\hat{\alpha}_1$ we noted that on average more than 80% of the total bias was contributed by the unweighted model. Overall, the effect of sample size on bias and its variability was unclear. For three out of four intercept/slope parameter estimates the variability in bias seems to decrease with sample size. In most cases, varying sample sizes did not have a clear trend on median or IQR of the index.

Regarding to the index for variance component presented in the middle panel of Figure 2, in four out of six of the parameters, the median value of R and the ± 1 IQR bars were below the dashed line. The exceptions to this were for the estimates of σ^2. For both components, the variance seems to be much higher for the unweighted model than the weighted model. In addition, looking at the IQR, we can see that $\mathcal{R}$ associated with variance is much less has shorter bars than the same index for the bias.

Finally, looking at MSE $\mathcal{R}$ index at the bottom panel of Figure 2, as MSE is the sum of the bias square and variance the results shown are reflective of the above two. On most cases, the median value of $\mathcal{R}$ is below the 0.5 threshold line except only the mixing proportion parameter and the variance parameters. Concerning the variability of $\mathcal{R}$ for MSE, we can see that similar to the variance component it has lower variability compared to the same index for the bias. From the three summaries we can conclude that the even if it is unclear which model performs better in terms of MSE, the bias in parameter estimates obtained by using the weighted model is lower than the unweighted model.

3.3 Simulation 3: Model Selection

In this simulation study, we assess the performance of BIC as a model selection method when using the weighted model as a classification tool. This is using the relationship between finite mixture models and model-based clustering. In model based clustering, each component is associated with a single cluster. Hence, a K-component mixture can be used to identify K homogeneous classes in heterogeneous data. Therefore, we will vary the number of components K that is used to generate the mixture model and assess if BIC is able to retrieve the true K.

The vector of true parameters $\boldsymbol{\psi} = (\boldsymbol{\tau}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2)$ used to generate the mixtures are shown in Table 4. In this setup, samples were drawn using stratified sampling design from the finite population by selecting simple random samples without replacement of size $n_h = 500$

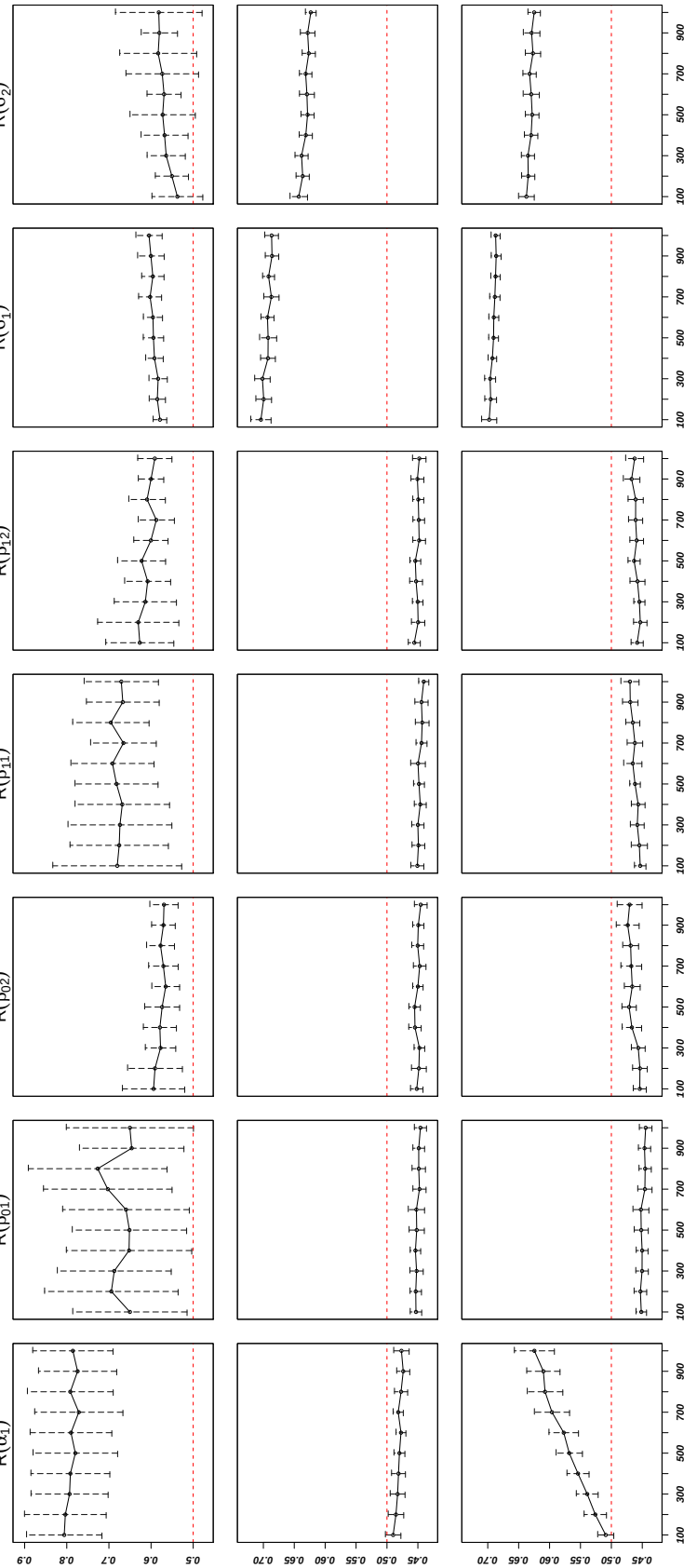


Figure 2: Summary of the $\mathcal{R}$ index for bias (top row), variance (middle row), and mean square error (bottom row) in Simulation study 2. The middle line represents the median and the dashed bars represent the interquartile range (IQR) values at different sample sizes n.

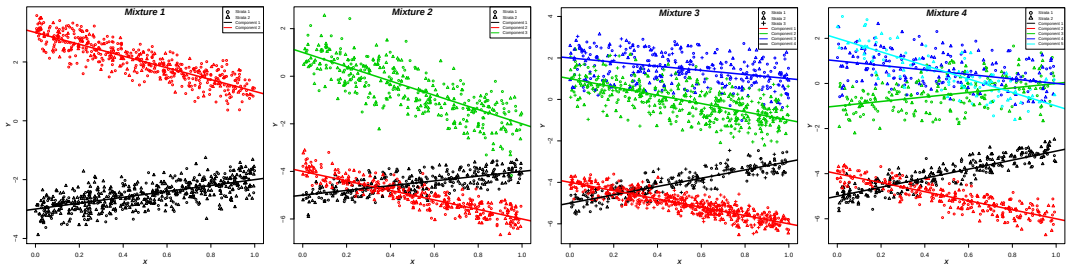


Figure 3: Scatter plots of samples which are selected from finite populations for the four experiments in Simulation 3. Colors show the number of components, and plotting characters show the strata.

from each stratum, $h = 1, ..., H$. Figure 3 shows the stratified samples which were selected in four experiments. In Mixture 1, we generated a finite population containing two strata with $10000, 8000$ observations in each stratum. In this case, there are two components $(K = 2)$, and the total of $n = 1000$ observations was selected. In Mixture 2, we generated a finite population containing two strata with $10000, 8000$ observations in each stratum. We considered three mixture components $(K = 3)$. Therefore, we have $n = 1000$ observations selected in the total. In Mixture 3, we generated a finite population containing three strata with $1000, 8000, 6000$ observations in each stratum, respectively. The population has four components $(K = 4)$, where the total of $n = 1500$ observations selected with 500 from each stratum. In Mixture 4, we generated a finite population containing two strata with $12000, 8000$ observations, respectively. In this case, the population has five components $(K = 5)$. The total number of observations in the sample was $n = 1000$. After generating data, the weighted model is fitted for different values of K ranging from 1 to 10. The BIC is then calculated for each K. Figure 4 shows the results of this experiment including the BIC values for all K and the optimal number of components in the four experiments above. According to the results, BIC was able to choose the optimal number of components under the various circumstance. In all four cases, BIC was the lowest at the true K value.

Table 4: True parameter values for Mixtures of linear regression in Simulation 3.

		ψ									
H	K	$\boldsymbol{\alpha}_1$	$\boldsymbol{\alpha}_2$	$\boldsymbol{\alpha}_3$	$\boldsymbol{\alpha}_4$	$\boldsymbol{\beta}_{10}$	$\boldsymbol{\beta}_{20}$	$\boldsymbol{\beta}_{30}$	$\boldsymbol{\beta}_{40}$	$\boldsymbol{\beta}_{50}$	$\boldsymbol{\beta}_{11}$
2	2	0.52				-3	3				1
2	3	0.30	0.36			-5	-4	1			1
3	4	0.17	0.32	0.29		-5	-4	1	2		1
2	5	0.24	0.26	0.20	0.15	-5	-4	-1	1	2	1
H	K	$\boldsymbol{\beta}_{21}$	$\boldsymbol{\beta}_{31}$	$\boldsymbol{\beta}_{41}$	$\boldsymbol{\beta}_{51}$	$\boldsymbol{\sigma}_1^2$	$\boldsymbol{\sigma}_2^2$	$\boldsymbol{\sigma}_3^2$	$\boldsymbol{\sigma}_4^2$	$\boldsymbol{\sigma}_5^2$	
2	2	-2				0.1	0.1				
2	3	-2	-3			0.1	0.1	0.5			
3	4	-2	-3	-1		0.1	0.1	0.5	0.5		
2	5	-2	1	-1	-3	0.1	0.1	0.5	0.5	0.4	

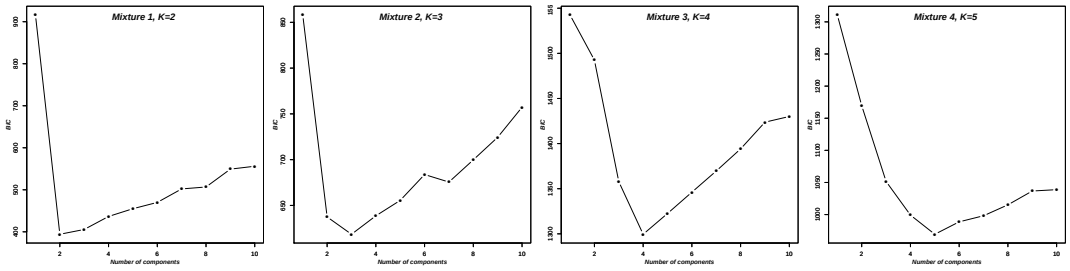


Figure 4: BIC values corresponding to the optimal number of components for the four experiments.

4. APPLICATION

In the previous sections, we established tools to make inference and model assessments for the proposed model. This section is dedicated to an application of the proposed model to a real dataset. Our study focuses on the Academic Performance Index (API) dataset. The API is a measurement of academic performance and progress of individual schools in California, United States. This dataset is also available for use in an R package survey (Lumley, 2004). The dataset contains 6,194 observations on 37 variables which provide information for all schools in California with at least 100 students.

4.1 Example 1

In this study, we used the variable called $stype$, which indicates the types of school (elementary/middle/high school, for stratification to produce more precise sample estimates, the individual strata should be internally homogeneous and different from one another. Subsequently, we fitted the mixture regression model for the academic performance index in 2000 ($api00$) as the response variable and percent of parents who were high school graduates, hsg, as a predictor. Then, the parameter estimates are determined based on the proposed approach. A sample with size $n = 750$ observations was selected using stratified sampling design. We implemented pseudo-maximum likelihood procedure for the selected sample, to estimate and deduce the features of the hidden inference associated with the relationship between the response and the explanatory variable. We fitted various finite mixture of Gaussian polynomial regressions for this dataset. Table 5 reports the BIC corresponding to the different scenarios that were implemented by varying the number of components K and the polynomial degree r of the independent variable hsg for a mixture regression of linear regression models. According to BIC, the best model was found to be with $K = 2$ as the number of components and $r = 2$ for this part of the dataset. Estimates of the regression parameters for a mixture of quadratic Gaussian regressions have been reported on Table 6. Figure 5 shows the fitted regression model for the $api00$ on hsg. The first component contains 27% of the total observations in the sample. The expected value of API is about 567 when $hsg = 0$. On average, for each one percent increases in, hsg, the API will decrease by $-3.92 + 2(0.055)hsg$, which corresponds to the first derivative of a quadratic model. The API is declining by about 4 scores when the $hsg = 0$. Moreover, the curve of API scores is slowly decaying by increasing in hsg until the hsg is approximately 36 which represents 87% of observations of hsg and then it is eventually growing when the hsg increases, and that reflect the various behavior of the API of the students in this component. In the second components, there are approximately 73% of the observation in the

sample. The expected value of API is about 894 when $hsg = 0$. When the percent parents who graduated from high school, hsg increased by one unit, the expected API changes by $-11.80 + 2(0.14)hsg$, which corresponds to the first derivative of a quadratic model. The API is decreasing by about 12 scores when the $hsg = 0$. Furthermore, the curve of API scores is decreasing in a lower rate as the hsg increases until the hsg is approximately 42 which represents 93% in total of observations of hsg, and it shows a slight increase as the hsg increases.

Table 5: BIC values for combination of number of components K and degree of the polynomial r in Example 1.Bold font represents the lowest BIC obtained indicating the best fit.

		r		
		1	2	3
K	1	9259.619	9088.472	9168.37
	2	9231.609	**9073.795**	9149
	3	9282.754	9119.125	9195.36
	4	9311.306	9160.060	9209.54

Table 6: Estimated parameters for the mixture regression model for the data in Example 1.

$\hat{\psi}$								
$\hat{\alpha}_1$	$\hat{\beta}_{10}$	$\hat{\beta}_{20}$	$\hat{\beta}_{11}$	$\hat{\beta}_{21}$	$\hat{\beta}_{12}$	$\hat{\beta}_{22}$	$\hat{\sigma}_1$	$\hat{\sigma}_2$
0.27	566.56	894.40	-3.92	-11.80	0.055	0.14	58.52	66.84

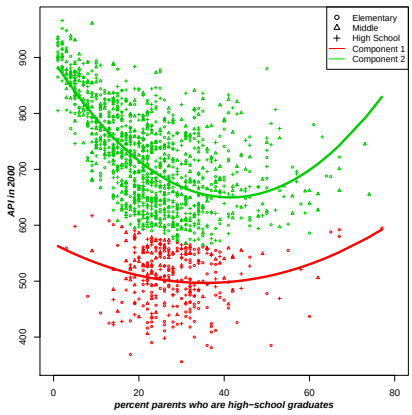
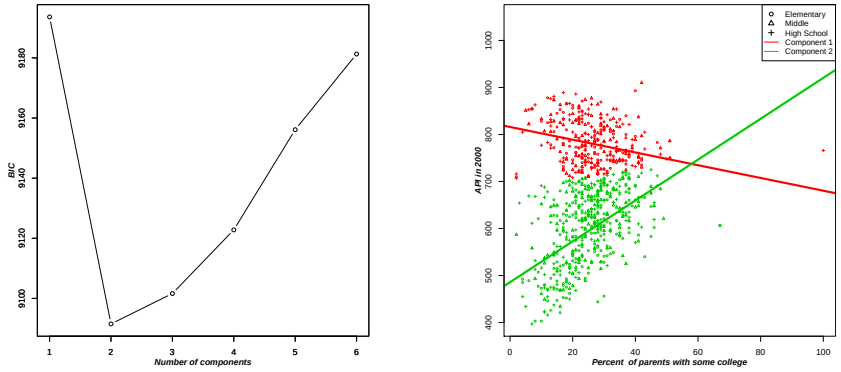


Figure 5: The plot shows the best-fitted mixture regression model with a 2-components quadratic Gaussian regressions model to regress the academic performance index in 2000 for the students on percent parents who are high-school graduates

4.2 Example 2

We fitted the proposed model for the API in year 2000 $api00$ as a response variable with the percent of parents with some college, $some.col$ as the independent variable. For the mixture regression model of regress the API in 2000 for students on percent of parents with some college, $some.col$. Various finite mixture of polynomial regression models has been fitted. Among these models, we sought a model with a small BIC. The best model was found to be a linear regression with $K = 2$ for this dataset with the smallest BIC value. The BIC corresponding to each of the linear regression fits are presented in Figure 6(a). The resulting mixture is given in Figure 6(b). The corresponding parameter estimates are provided in Table 7. It can be seen that for the first component (red), which consisted of 37% of the observations in the sample, and the average of API was about 816. The API decreases 1 unit for each unit increase in the percent of parents with some college. In the second component, there was approximately 63% of the the sample, and the component had students with an average API of approximately 486 where $some.col = 0$. Their API score increased by 4 units for each unit increase in the percent of parents with some college. In

component 1, the conclusion is that on average a student whose parents have lower percent with some college tends to report a higher API score. On the other hand, for component 2 on average a school which has have high percent of parents with some college tended to report a higher API score. The association between API and $some.col$ in first component is not very intuitive and may be indicative of other confounded variable that is not captured.



(a) (b)

Figure 6: Plots show (a) BIC versus to the number of components, for the mixture regression model of regress the API in 2000 for students on percent of parents with some college, (b) the fitted mixture regression model with a 2-components for the same dataset.

Table 7: Parameters estimated for the mixture regression model with the response the academic performance index in 2000 for the students and the percent of percent parents with some college as explanatory variable.

$\hat{\boldsymbol{\psi}}$						
$\hat{\alpha}_1$	$\hat{\beta}_{10}$	$\hat{\beta}_{20}$	$\hat{\beta}_{11}$	$\hat{\beta}_{21}$	$\hat{\sigma}_1$	$\hat{\sigma}_2$
0.37	815.90	485.80	-1.36	4.35	49.83	72.28

5. DISCUSSION AND CONCLUSION

A mixture of regression model was considered when a sample was gathered from a stratified design. A new methodology was developed by incorporating sampling weights to the complete-data log likelihood function, and an EM-algorithm was derived. Several simulation studies were conducted to evaluate the performance of the model under various circumstances. The new weighted model and the unweighted model were compared using the bias-variance components of MSE. In both approaches, based on simulation results, the mean square error for estimated parameters did not provide evidence significant enough to infer which estimation approach was better. However, the weighted model showed lower bias for the estimated parameters compared with the unweighted model. Conversely, the unweighted model had a lower variance for most of the estimated parameters compared with the weighted model. Overall, variability in both models tended to decline as the sample size increased.

To assess this further, we constructed a percent contribution index that shows which shows how much each model contributed to a total bias, variance, and MSE. Overall, according to the relative bias index, the weighted model estimates have lower bias compared with the unweighted model estimates. In the same context, weighted model estimates have high variability compared with estimates which are obtained via the unweighted model for the majority of the parameters to be estimated. The variability in this index was found to be much higher in bias than either variance or total MSE. We assessed the utility of the BIC for selecting the optimal number of components for a given dataset using a simulation study. In all settings, the BIC model resulted in the correct number of components for a given dataset.

In the real data analysis, the API scores in 2000 were regressed against the percent of parents who were high-school graduates in California school of interest in the first example. The optimal regression mixture model was chosen to be the one with the smallest BIC.

After several models were fitted, a 2-component quadratic Gaussian regressions mixture regression model performed better than other models, with the BIC being the smallest. In the same context, when the API in 2000 for the students was regressed against percent parents with some college. After numerous models were fit, a 2-component linear mixture regression model had better performance than other models, with BIC being smallest.

In general, even though the survey detests already have subgroups, we can use finite mixture regression tools, such as the proposed model, to search subpopulations that are not easily amenable using the usual survey analysis tools. In future work, we will strive to extend and develop this work to be appropriated for a mixture of multiple regression models, other survey design techniques, and various types of weight calculations such as non-response rates. Furthermore, we will address the different configurations by giving sufficient scope to analyze the real dataset.

Finite Mixture of Multiple Regression Models for a Complex Sample

Abdelbaset Abdalla and Semhar Michael *

A design-based inference has been developed where sampling weights are integrated into the complete-data log-likelihood function for modeling the mixture model to data collected using complex survey data. A pseudo likelihood approach is proposed and applied to obtain the estimates of the mixture model parameters. A challenging problem that arises in this domain: Is whether our proposed model will be able to retrieve the underlying subgroups in a given population: Is whether our proposed model will be able to retrieve the underlying subgroups in a given population. Two approaches were considered: In the first approach, the mixture of regression models fitted to the available survey data, which was treated as a finite population. The co-occurrence matrix was constructed based on the classification solutions of the best fit model. In the second approach, the mixture model was fitted to the selected sample samples based on a complex design from the survey data. The co-occurrence matrix was constructed based on the classification solutions of the best fit models of multiple samples. The hierarchical clustering used to find clusters in data. Finally, we find out the classification solution agreement between the two methods at different numbers of components. We illustrate the proposed procedure of fitting a mixture model to survey data with an example from NHANES data.

Keywords: mixture models, complex survey design, hierarchical clustering, pseudo likelihood

This is a draft manuscript to be submitted to Journal of Applied Statistics.

*Abdelbaset Abdalla is a graduate student at South Dakota State University; Semhar Michael is an Assistant Professor of Statistics at South Dakota State University, email: Semhar.Michael@sdstate.edu

1. INTRODUCTION

Data collected through surveys are an essential source of information for modern societies. In this context, social survey data are one of the crucial data sources for understanding society and changes in social trends. Besides, health surveys such as the National Health and Nutrition Examination Surveys (NHANES) are vital for ensuring the public health data that inform policymakers as well as members of the community about important health issues for which health policy and procedures need to be implemented. Information from surveys, therefore, represents one of the most important contributions to decision-making processes aimed at effectively implementing international and government policies.

Over the past decades, there are a considerable number of studies describing the use of sampling weights when a model fitting is carried out using complex sample survey data. Many of these focus on the issue of whether sample weights should be used when fitting a model to such data. It should also be pointed out that this is not a new problem. There have been many research papers that aim to answer this question Pfeffermann and Nathan (1981); Nordberg (1989); Pfeffermann (1993); Lohr and Liu (1994). Most of these papers not only point out the solution to the issue but also provide useful guidelines regarding how sample weights should be used in model fitting. The recent issues related with survey data analysis is published in statistical science journal (Zhang, 2017)

Finite mixture models are being used increasingly to model heterogeneous data. Finite mixture models in various practical situations are a powerful device that can be used for clustering by assuming that each mixture of a component represents a subgroup in the data. The most prevalent mixture model is the one consisting of Gaussian components (Day, 1969; McLachlan and Basford, 1988; Fraley and Raftery, 2006). An overview of mixture models is given in McLachlan and Peel (2000) and Frühwirth-Schnatter (2006). The most recent advances and challenges related to mixture models can be found in Melnykov et al. (2015); McNicholas (2016); McLachlan et al. (2019). The issue of estimating the param-

eters of the mixing distribution has a long history and dates back to Pearson (1894), who dealt with a mixture of two components with equal variances by using the method of moments. However, in this work, the expectation-maximization (EM) algorithm (Dempster et al., 1977) will be considered to derive the estimation of the parameters.

The finite mixture models are estimated based on the observations are drawn using the simple random sample design. However, in real applications, this assumption rarely met. However, ignoring the structure of the complex sampling design results in inconsistent and biased estimates Wedel et al. (1998). The pseudo maximum estimation (PML) approach, which is used to deal with finite mixture linear regression models for complex sample designs. However, the inclusion probabilities for each of the sampled units are required when using the PML approach. The development approach in later sections for the mixture model for a complex sample is based on (Skinner et al., 1989). The PML approach has been conducted on several statistical models. Still, as far as we know, there are only limited works to fit the complex sampling data using the mixture of regression models.

The purpose of this article is to try to answer the following question: We will investigate whether the proposed can be retrieved underlying subgroups in a given population if we draw a sample from the data? To our knowledge, there is no previous work using the proposed approach below.

In order to answer this question, an approach of two steps was considered. In the first strategy, the mixture of regression models was fitted to the survey data, which were treated as a target population. The classification solutions are identified based on the fitted model of the data. The co-occurrence matrix has been constructed using these solutions. In the second method, a complex sample is drawn using stratified multistage probability sampling. The mixture of regression models was fitted to the sample and the classification solutions identified based on the best-fitted model of the sample. The previous procedure is repeated multiple times. Then, using these classification solutions, the co-occurrence

matrix was constructed. The hierarchical was considered to find the clusters in the data. Finally, we compute the classification solution agreement between the two approaches at different numbers of components. As a concrete example, we will often refer to public use data from the NHANES data.

2. SAMPLING DESIGN AND MIXTURE OF REGRESSION MODELS

In this section, some necessary groundwork will be laid concerning finite mixture models and complex survey data that will be used in this paper, then the proposed methodology will be described.

2.1 Complex Sampling

In most survey data analyzed in practice are originally collected used non-simple random sample (SRS) designs. These designs, such as stratified and cluster sampling. These designs commonly are combined to obtain so-called complex sampling (CS). CS is a technique employed to ensure that the sample collected represents our target population as closely as possible. In this paper, the sampling weights are calculated as reciprocals of probabilities of selection. The sum of these weights is the population size (Lohr, 2010), denoted here by N. In this paper, we consider a stratified two-stage cluster sample design to draw the sample. Assuming that a finite population has been stratified into H strata, then the sample is drawn from each stratum in the population. Assume stratum h was divided into N_h PSU's of which n_h has been sampled, $h = 1, ..., H$ with equal probability. It follows that the selection probability of the ith PSU in the hth stratum, π_{hi}, is given by $\pi_{hi} = \frac{n_{hi}}{N_{hi}}$.

Let the ith sampled PSU be clustered into M_i SSU's of which m_j are sampled with equal probability, $i = 1, ..., n_h$. The selection probability of the jth SSU providing the ith PSU in the hth stratum has been selected, $\pi_{j|i}$, is defined as $\pi_{j|hi} = \frac{m_i}{M_i}$, suppose that π_{ij}

and which denote in turn the probability of selecting the jth SSU in the ith PSU of the hth stratum, and the sampling weight of the the jth SSU in the ith PSU of the hth stratum. Then $w_{ij} = {\pi_{ij}}^{-1}$. Where

$$\pi_{ij} = \left(\frac{n_{hi}}{N_{hi}}\right)\left(\frac{m_{hi}}{M_{hi}}\right), h = 1, ..., H, i = 1, ..., n_h, j = 1, ..., m_{hi}.$$

Lohr (2010). When conducting the inference about the mixture models under the complex sampling designs, the sampling weights are incorporated in the inference to construct pseudo likelihood functions in later sections.

2.2 General Methods of Estimation for Complex Survey Design

Linear regression models and estimators are usually applies to analyze complex survey data using pseudo maximum likelihood approach (Binder, 1983), (Skinner et al., 1989). Let $\mathbf{y}_1, ..., \mathbf{y}_N$ be the values of the $\mathbf{y}$ in the finite population. These are considered as random variables with the pdf $f(\mathbf{y}_i; \mathbf{\Psi})$ which depends on an unknown parameter vector $\mathbf{\Psi}$. The maximum likelihood estimate (MLE), $\widehat{\mathbf{\Psi}}_{mle}$ of $\mathbf{\Psi}$ is defined as the solution to the equations

$$U(\mathbf{\Psi}) = \sum_{i=1}^{N} \frac{\partial \log f(\mathbf{y}; \mathbf{\Psi})}{\partial \mathbf{\Psi}} = 0. \tag{1}$$

The pseudo maximum likelihood estimator (PMLE), $\widehat{\mathbf{\Psi}}_{pml}$ estimator of $\mathbf{\Psi}$ is defined as the solution of sample estimate of $U(\mathbf{\Psi})$. i.e. $\hat{U}(\mathbf{\Psi}) = 0$. The common estimator of $U(\mathbf{\Psi})$ is the Horvitz-Thompson estimator. Thus, the PMLE of $\mathbf{\Psi}$ is the solution of

$$\sum_{i=1}^{N} w_i \frac{\partial \log f(\mathbf{y}; \mathbf{\Psi})}{\partial \mathbf{\Psi}} = 0. \tag{2}$$

2.3 Finite Mixture of Gaussian Regression Models

Let $\mathbf{y}$ be a response variable of interest and $\mathbf{x} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_p)^\top$ be the vector of p covariates which may have effect on $\mathbf{y}$. We say that $(\mathbf{x}, \mathbf{y})$ follows a finite mixture of Gaussian regression model with the conditional density function of $\mathbf{y}$ given $\mathbf{x}$ has the form

$$f(\mathbf{y}_i; \boldsymbol{x}_i, \boldsymbol{\Psi}) = \sum_{k=1}^{K} \alpha_k \, \phi(\mathbf{y}_i; \boldsymbol{x}_i \boldsymbol{\beta}_k, \sigma_k^2), \tag{3}$$

where K is the total number of mixture regression components, $\phi(\mathbf{y}_i; \boldsymbol{x}_i \boldsymbol{\beta}_k, \sigma_k^2)$ is a Gaussian density function of the kth component with mean $\boldsymbol{x}_i \boldsymbol{\beta}_k$ and variance σ_k^2. The mixing proportions $\alpha_k > 0$ and $\sum_{k=1}^{K} \alpha_k = 1$. The parameter vector $\boldsymbol{\Psi} = (\alpha_1, ..., \alpha_{K-1}, \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_K)$ with $\boldsymbol{\beta}_k = (\boldsymbol{\beta}_{k1}, ..., \boldsymbol{\beta}_{kp})^\top$.

2.4 Pseudo-Maximum Likelihood Estimation Approach

if one considered, a mixture model in (3) the MLE of $\boldsymbol{\Psi}$ maximizes the log-likelihood. The standard formulation of the log-likelihood applies under simple random sampling, in which each unit receives the same weight. The ML estimator solves the likelihood equations in (1).

In this case, estimation of the parameters is typically performed through the maximum likelihood approach. The log-likelihood function is given by

$$\ell(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \log \left\{ \sum_{k=1}^{K} \alpha_k \, \phi(\mathbf{y}_i; \boldsymbol{x}_i \boldsymbol{\beta}_k, \sigma_k^2) \right\}. \tag{4}$$

Due to the inconvenient form of $\ell(\boldsymbol{\Psi})$ in Equation (4), the expectation maximization algorithm (Dempster et al., 1977), which is based on a complete-data log-likelihood function, is employed. The complete-data setup is given IID samples from $f(\mathbf{y}_i; \boldsymbol{x}_i, \boldsymbol{\Psi})$; we define the latent variable Z_{ik} such that

$$Z_{ik} = \begin{cases} 1 & \text{if the } i\text{th observation} \in k\text{th component} \\ 0 & \text{otherwise} \end{cases}.$$

Then, we can write the complete-data log-likelihood function as

$$\ell_c(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \sum_{k=1}^{K} I(Z_{ik} = 1) \left\{ \log \alpha_k + \log \phi(\mathbf{y}_i; \boldsymbol{x}_i \boldsymbol{\beta}_k, \sigma_k^2) \right\}. \tag{5}$$

The EM-algorithm is an iterative procedure of two steps, the Expectation (E) step, and the Maximization (M) step. At the E-step, we calculate the conditional expectation of the complete-data log-likelihood function given the observed data

$$E\Big(I(Z_{ik}=1)|\mathbf{y}_i,\boldsymbol{x}_i,\boldsymbol{\Psi}^{(t-1)}\Big)=Pr(Z_{ik}=1|\mathbf{y}_i,\boldsymbol{x}_i,\boldsymbol{\Psi}^{(t-1)}).$$

This posterior probability will be denoted as τ_{ik}. The expression of τ_{ik} at the (t)th iteration of the E-step is given by

$$\tau_{ik}^{(t)}=\frac{\alpha_k^{(t-1)}\phi\left(\mathbf{y}_i;\boldsymbol{x}_i\boldsymbol{\beta}_k^{(t-1)},\sigma_k^{2(t-1)}\right)}{\sum_{k'=1}^{K}\alpha_{k'}^{(t-1)}\phi\left(\mathbf{y}_i;\boldsymbol{x}_i\boldsymbol{\beta}_{k'}^{(t-1)},\sigma_{k'}^{2\,(t-1)}\right)}.$$

At the M-step of the (t)th iteration, we maximize the conditional expectation of the complete-data log-likelihood function commonly known as the Q-function given by

$$Q(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(t)})=\sum_{i=1}^{n}\sum_{k=1}^{K}\tau_{ik}^{(t)}\left\{\log\alpha_k+\log\phi(y_i;\boldsymbol{x}_i\boldsymbol{\beta}_k,\sigma_k^2)\right\}. \tag{6}$$

Often a full ML procedure is intractable since the expression for the likelihood under the complex sampling strategy depends on assumptions about the unknown relationships between the $\mathbf{y}$ and the sample design variables.

Assume a complex sample $\{(\boldsymbol{x}_i,\mathbf{y}_i,w_i);i\in s\}$, where w_i is the sampling weights. In this case, we selected a sample of size n units from a finite population of size N under some complex survey design. The most popular definition of w_i is as an indicator of the number of population units which are represented by ith sample unit. In this paper, w_i will be equal to the reciprocal of the inclusion probability π_i, which is the probability of selecting the ith sample unit under some complex survey sampling design. If such a design is considered, then standard maximum likelihood estimators are usually biased Wedel et al. (1998). Such a scenario can be avoided using the approximate, or pseudo-maximum Likelihood (PML) approach as proposed by Skinner et al. (1989) and described by Pfeffermann (1993) and Chambers and Skinner (2003). We propose a weighted estimation procedure for finite

mixture models which minimizes the bias in parameter estimates that occur when the sampling design is not taken into consideration. However, a simple approach is to construct a consistent estimator for $\mathbf{\Psi}$ by solving equations in (2).

This is done by incorporating the sampling weights, w_i to the complete data log-likelihood function. Then the modified Q-function is given by

$$Q_{pw}(\mathbf{\Psi}; \mathbf{\Psi}^{(t)}) = \sum_{i=1}^{n} w_i \sum_{k=1}^{K} \tau_{ik} \{\log \alpha_k - \frac{n}{2} \log(2\pi\sigma_k^2) - \frac{(\mathbf{y}_i - \beta_{k0} - \beta_{k1}\boldsymbol{x}_i)^2}{2\sigma_k^2}\}. \quad (7)$$

We refer to the function in Equation (7) as the weighted Q-function and is denoted by Q_{pw}. At the M-step of the (t)th iteration, the Q_{pw}-function is maximized with respect to $\mathbf{\Psi}$. For the mixture of multiple regression model, the closed form solution for mixing proportions at (t)th iteration of the weighted M-step is given by

$$\alpha_k^{(t)} = \frac{\sum_{i=1}^{n} w_i \tau_{ik}^{(t)}}{\sum_{k=1}^{K} \sum_{i=1}^{n} w_i \tau_{ik}^{(t)}}, \quad (8)$$

The closed form solutions for the mixture of multiple regression model parameters at (t)th iteration of the weighted M-step are given by

$$\boldsymbol{\beta}_k^{(t)} = \left(\mathbf{X}^\top \mathbf{W}_k^{(t)} \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{W}_k^{(t)} \mathbf{y}, \quad (9)$$

where $\mathbf{X}$ is an $n \times (p+1)$ matrix containing unity for intercept and predictors, $\mathbf{W}_k^{(t)}$ is a $n \times n$ diagonal matrix with entries $w_i \times \tau_{ik}^{(t)}$, $\mathbf{y}$ is a $n \times 1$ vector of response variable, and

$$\sigma_k^{2(t)} = \left\| \mathbf{W}_k^{1/2(t)} \left(\mathbf{y} - \mathbf{X}\, \boldsymbol{\beta}_k^{(t)}\right) \right\|^2 \times \mathrm{tr}\left(\mathbf{W}_k^{(t)}\right)^{-1}, \quad (10)$$

where $\|\mathbf{A}\| = \mathbf{A}^\top \mathbf{A}$ with $\top$ denoting a matrix transpose and $\mathrm{tr}(\mathbf{A})$ means the trace of the matrix $\mathbf{A}$. Equations (9) and (10) can be used as update equations at at (t)th iteration of the M-step. The same equation as given in equation (8) is used to update mixing proportions.

2.5 Computational Strategies

Rnd-EM (Maitra, 2009) was used to choose initial values. In this initialization method, first random points are selected as seeds, and the Euclidean distance is used to assign observations to centers. This is repeated for some fixed number of times. The solution that yields the highest likelihood value is then used for initializing the EM-algorithm. *Rnd-EM* tends to work well if the number of components is not large (Michael and Melnykov, 2016). In the EM-algorithm, the E-step and M-step are iterated until a convergence criterion is met. In this paper, the algorithm is stopped when the absolute relative change in the likelihood given by

$$\frac{\ell_p(\mathbf{\Psi}^{(t)};\mathbf{y},\mathbf{x}) - \ell_p(\mathbf{\Psi}^{t-1};\mathbf{y},\mathbf{x})}{|\ell_p(\mathbf{\Psi}^{t-1};\mathbf{y},\mathbf{x})|} < \epsilon, \text{ with } \epsilon = 10^{-8},$$

where $\ell_p(\mathbf{\Psi}^{(t)};\mathbf{y}, \ell_p(\mathbf{\Psi}^{(t-1)};\mathbf{y},\mathbf{x})$ are the pseudo log-likelihood values from iterations $t, t-1$, respectively. In the dataset analysis, we used the BIC (Schwarz et al., 1978) to select the optimal number of components. In this paper, BIC will be calculated as $BIC(\hat{\mathbf{\Psi}}) = -2\ell_p(\hat{\mathbf{\Psi}}) + M\log n$, where $\ell_p(\hat{\mathbf{\Psi}})$ and M represent the maximized pseudo likelihood value for a given K and the number of parameters in the fitted model, respectively. For mixtures of normal regression, $M = (K-1) + K(p+1) + K$, where p represents the number of the predictor variables. The model with lowest BIC value is the best model for a given dataset.

Identifiability of a given model is one of the major requirements for any model to be meaningful and still an open question. It is defined for any two parameters vectors $\mathbf{\Psi} \neq \mathbf{\Psi}'$, the respective model $f(\boldsymbol{x};\mathbf{\Psi})$ must be different from $f(\boldsymbol{x};\mathbf{\Psi}')$ for any random vector $\boldsymbol{x}$. Some basic issues are, however, in common with the component of a mixture of regression models to which they belong. To illustrate, it is widely known that finite mixture models are only identifiable up to a permutation of the component labels, which so-called switching labels problem. Nevertheless, this issue only affects the interpretation of the results, but there is no problem with parameter estimation (Leisch, 2004). Overfitting is a more fundamental lack of identifiability, and it leads to empty components or components with equal

parameters. This kind of unidentifiable can be avoided by restricting the prior mixing ratios to be greater than zero, and the component with specific parameters are different (Leisch, 2004). In this paper, to prevent overfitting, mixing proportions have been restricted to be greater than a particular threshold.

3. ANALYSIS OF NHANES DATA

Description of the Data

We will often refer to public use data from the NHANES data conducted by the U.S. National Center for Health Statistics (NCHS). They are designed to provide national data on health, disease, and dietary and clinical risk factors gained from clinical examinations as well as detailed interviews (Centers for Disease and Prevention, 2013-2016). In this work, we consider the 2013-2014 and 2015-2016 waves of NHANES participants in the clinical exams and dietary questionnaires. In this paper, the subjects in NHANES who had complete data on a selected set of variables were treated as a finite population (Li and Valliant, 2015).

The R function for reading data in these formats in the R is haven package (Wickham and Miller, 2019). This package is part of the R distribution but is not automatically loaded into memory when R starts. To load this package from the package library, we need to type library(haven). When the package is loaded, all its functions and help pages become available. The functions reade_xport(), will read SAS XPORT files. This function takes a file name as the first argument. In this work, the 2013-2014 and 2015-2016 waves of NHANES data were imported as SAS XPORT files, and we prepared them for our data analysis. The R function for reading data in these formats in the R is haven package (Wickham and Miller, 2019). This package is part of the R distribution but is not automatically loaded into memory when R starts. To load this package from the package library, we need to type library(haven). When the package is loaded, all its functions and help pages become

available. The functions reade_xport(), will read SAS XPORT files. This function takes a file name as the first argument. In this work, the 2013-2014 and 2015-2016 waves of NHANES data were imported as SAS XPORT files, and we prepared them for our data analysis.

Our analysis included people who are 18 years of age or older within the NHANES 2013-2016 population. The whole population consists of 2772 observations. After the dataset was prepared and cleaned, the total was $N = 2402$ individuals. Suppose now that one is interested in regressing a response variable $\mathbf{y}$ against a given set of independent variables. For example, Harlan et al. (1985) has fitted regression models to NHANES data with systolic blood pressure as a dependent variable (SBP) and body mass index (BMI), age, and blood lead levels (BLL) as independent variables. Additionally, this example was used by (Li and Valliant, 2015) in a linear regression analysis for data from NHANES. In this section, we fitted the finite mixture of multiple regression models for NHANES data 2013-2016 with the response variable SBP on BMI, age, and BLL as predictor variables. The dataset consists of two two-year waves of the new (continuous) NHANES data. Thus, it is necessary to download the data on demographics (age, sex, education level, and ethnicity), anthropocentric information (height, weight, and body mass index (BMI)), and blood pressure) for both NHANES 2013-2014 and NHANES 2015-2016, then extract the appropriate variables and merge the datasets. It was also necessary to compute the average of the multiple blood pressure measurements that are provided in the data. The sampling weights also need to be adjusted for the combined data. Since each wave of analysis is weighted to correspond to the full United States population, the combined data represents two copies of the population. A new sampling weight variable was created by halving the original weight that is recommended for analysis of complex survey datasets such as NHANES data (Lumley, 2011). Moreover, the weights are created in NHANES to account for the complex survey design, survey non-response, and post-stratification. Let w_{NH} denote the

sampling weights included in NHANES data. When a sample is weighted in NHANES, it is representative of the U.S. Census civilian non-institutionalized population. A sample weight is assigned to each sample person, which denotes the number of people in the population represented by that sample person. Throughout this section, we will assume that a sampling weight of any observation is assigned a weight that is equivalent to the reciprocal of its probability of selection.

Statistical Analysis

The question that surfaces is whether the proposed method will be able to retrieve the underlying subgroups in a population that heterogeneity was not accounted for by the sampling design. In most clustering methodology development, it is common to use classification dataset to assess performance of a method. However, to our knowledge there is no population level classification dataset. To overcome this problem, we used treated the NHANSE data as population data and determined underlying subpopulation using our method. Then samples are taken from the data using complex survey design and the model is fitted. Then we assessed how well the underlying groups are recovered by looking at co-occurrence of observations in the sample as compared to the population. The following describes the summary of our approach

- **Step 1**: Full data

  1.1 Find best model fit using BIC. For the considered dataset, a 3-component mixture of multiple regression model was the optimal model according to BIC.

  1.2 Obtain the classification solution, based on $K = 3$ solution to construct was used to construct a $N \times N$ co-occurrence matrix, $A_1 = \{a_{ij}\}_{i,j=1}^{N}$. This matrix is binary with $a_{ij} = 1$ indicating that two observations were in the same group and $a_{ij} = 0$ that they were not classified together.

- **Step 2**: Sample data

  2.1 Use complex design to get a sample of size n_b from full data of size N. Fit the proposed model for various values of K and find the best model using BIC. For the best model find classification solution and co-occurrence matrix of size $n_b \times n_b$. Repeat this for $b = 1, \ldots, 200$.

  2.2 The 200 $n_b \times n_b$ co-occurrence matrices have been merged to obtain $A_2 = \{b_{ij}\}_{i,j=1}^{N}$, $N \times N$ co-occurrence matrix by finding the proportion, b_{ij}, computed by dividing the number of times observations y_i and y_j are in the same group by the number of times both were in a sample.

- **Step 3:** Comparison

  3.1 Perform hierarchical clustering using $\mathcal{J}_N - A_1$ as dissimilarity matrix, where $\mathcal{J}_N$ denoting an all-ones $N \times N$ matrix. Cut the tree at different values of K to obtain classification solution $\mathcal{C}_{1K}$. Similarly, use $\mathcal{J}_N - A_2$ to perform hierarchical clustering and obtain $\mathcal{C}_{2K}$.

  3.2 Compute classification solutions agreement between the two solutions $\mathcal{C}_{1K}$ and $\mathcal{C}_{2K}$ at different $K = 2, ..., 10$. The pseudo-code with more formal notation is provided in Algorithm 2.

Data: Matrix of finite population X; Design Variables v_1, v_2, v_3

Result: Complex Sample

Step 1;

Use v_1 to divide X into $h = \{1, ..., H\}$ strata;

Step 2;

for *each* $h \in H$ **do**

  Use v_2 to cluster each h into N_h PSU's;

  Select SRS of n_h PSU's form N_h PSU's;

  Use v_3 to cluster each i sampled PSU into M_{hi} SSU's;

  for *each* $j \in M_{hi}$ **do**

    Select SRS of m_{hi} SSU's form M_{hi} SSU's;

  end

end

Algorithm 1: The algorithm presents a stratified two-stage cluster design which uses to selecting a complex sample from a finite population.

Data: p-dimensional matrix of covariates $\mathbf{X}_{N\times p}$ and a response vector $\mathbf{Y}_{N\times 1}$

Result: Proportion of class agreement

Step 1: Given $\mathbf{Y}_{N\times 1}$ and $\mathbf{X}_{N\times p}$;

for *each* $k \in K$ **do**

  Fit the proposed model $\mathcal{M}_k$;

  obtain $BIC_{\mathcal{M}_k}$;

end

Let k' = $\text{argmin}_k \text{BIC}_{\mathcal{M}_k}$ and $\mathcal{C}_{1k'}$ the corresponding classification solution;

Use $\mathcal{C}_{1k'}$ to construct A_1 an $N \times N$ co-occurrence matrix;

Step 2: Given $\mathbf{Y}_{N\times 1}$ and $\mathbf{X}_{N\times p}$;

for $b \in B$ **do**

  Select a sample $\mathbf{y}_{n_b\times 1}$ and $\mathbf{x}_{n_b\times p}$ form the full data using Algorithm 1;

  for *each* $k \in K$ **do**

    Fit the proposed model $\mathcal{M}_{bk}$;

    obtain $BIC_{\mathcal{M}_{bk}}$;

  end

  Let k' = $\text{argmin}_k \text{BIC}_{\mathcal{M}_{bk}}$ and $\mathcal{C}_{bk'}$ the corresponding classification solution ;

  Use $\mathcal{C}_{bk'}$ to construct an $n_b \times n_b$ co-occurrence matrix ;

end

Obtain B co-occurrence matrices with $n_b \times n_b$ and combine them in one

co-occurrence matrix, A_2 of size $N \times N$

Step 3: Comparison ;

$\mathcal{J}_N - A_1$ and $\mathcal{J}_N - A_2$ as dissimilarity matrix and perform hierarchical clustering;

for $K \in$*{ 1,. . . , 10}* **do**

  Cut the tree K *and obtain classification solutions* $\mathcal{C}_{1K}$ *and* $\mathcal{C}_{2K}$*;*

  Compute classification solution agreement between $\mathcal{C}_{1k}$ *and* $\mathcal{C}_{1k}$*;*

end

Algorithm 2: The algorithm displays the steps to find the classification solution agreement between the two of approaches.

Results

In step 1, numerous finite mixtures of multiple regression models were fitted to the whole dataset. The sampling weights included in NHANES, w_{NH}, were used in this approach. The best model was found to be a multiple regression model with three components, $K = 3$, with the lowest BIC value. The BIC corresponding to each mixture of the linear regression fits is presented in Figure 1. Figure 2 displays the plots modeling systolic blood pressure versus the three auxiliary variables using the finite mixture of multiple regression models. The classification solution, based on $K = 3$, was used to construct a $N \times N$ co-occurrence matrix.

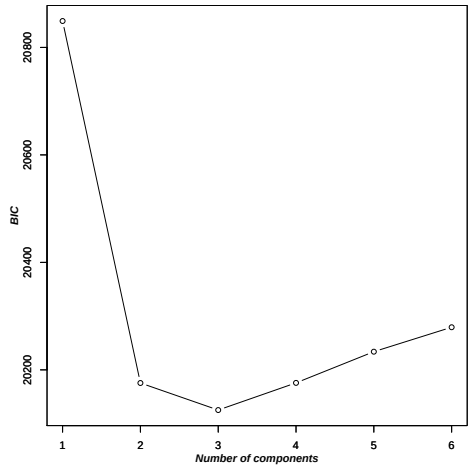


Figure 1: The plot shows BIC versus the number of components, for a mixture of multiple regression models of regress the systolic blood pressure on the body mass index, age, and blood lead levels.

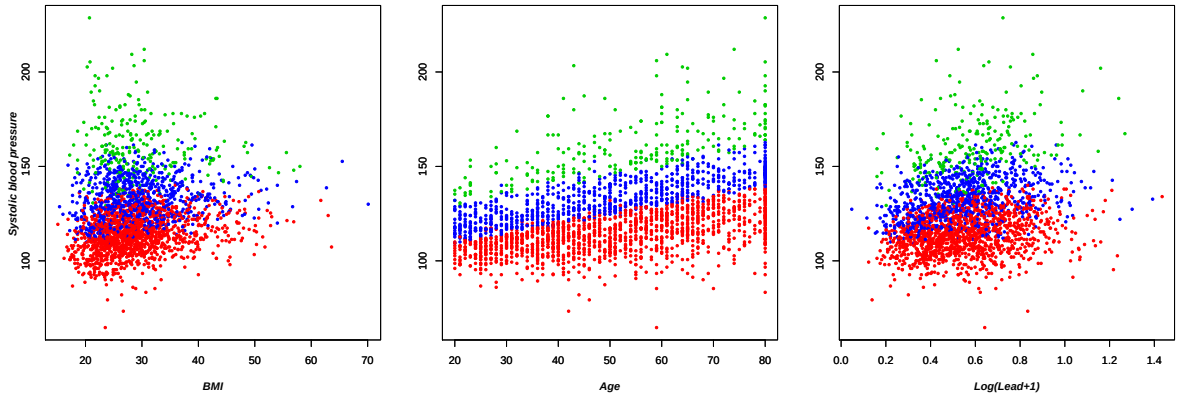


Figure 2: plots show the best-fitted mixture of multiple regression models with a 3-component to regress the systolic blood pressure versus three auxiliary variables for NHANES data.

In step 2, a sample was drawn from the NHANES 2013-2016 finite population using a stratified two-stage sample design, as described in algorithm 1. Since we re-sampled from the NHANES dataset, new sample weights needed to be constructed. Let w_{ij} be the sampling weights computed by using the stratified two-stage sample design and w_{NH} be the weight associated with the NHANES data, thus, two kinds of sampling weights were used in this step. Lohr (2010) and Thomas et al. (2006) have suggested combining these weights into one weight by multiplying them. Then, the overall sampling weight for an observation unit is $w_i = w_{ij} \times w_{NH}$. The new weights were used in the inference of the proposed model. The sampling was done $B = 200$ times with samples sizes $n_b, b = 1, \ldots, B$. The sample sizes varied depending on the sampling design. The steps to sampling from the NHANSE data are described in the pseudo-code given in Algorithm 1. In our specific application, the variables v_1, v_2, and v_3,respectively were used as design variables for each sample stage. After we selected the sample, the finite mixture of multiple regression model was fitted to the selected sample to find underlying subgroups. This was then used to obtain concurrence of pairs of observations forming $n_b \times n_b$.

The previous setup was repeated two hundred times. Contrary to the findings in step

1, which suggested that 3 components may be the best solution for the given dataset, step 2 demonstrated two different solutions a 2-component and a 3-component of the finite mixture of multiple regression models. Overall, in the majority of cases, the best solution is a 3-component of a finite mixture of multiple regression models. The best model was found to be a linear regression with a 3-component, $K = 3$, in approximately 72% of drawn samples, and a 2-component solution, $K = 2$, in 28% of samples. Thus, a set of two hundred $N \times N$ co-occurrence matrices were constructed via the two classification solutions, $K = 2$, and $K = 3$. Therefore, two hundred $n_b \times n_B$ co-occurrence matrices are then combined to $N \times N$ weighed concurrence matrix by computing the proportion of times two observations are in the same group given that they were in the sample.

The co-occurrence matrices obtained in both approaches were used as distance matrices, so then hierarchical clustering was considered to determine the clusters in the data. After the clusters were obtained, we found the classification solutions between the two approaches at $K = 2, ..., 10$. Algorithm 2 describes the strategies used to obtain the co-occurrence matrices and to calculate the proportion of the classification solution agreement between the two approaches for different k.

From the short review to the classification solution agreement between the two steps, key findings emerged. When the proposed model was applied to the whole available dataset, the best solution was a 3-component of mixture distributions. Conversely, both a 2-component and 3-component were the best solutions for the sample-based approach. Considering Figure 3, it is interesting to note that the classification agreement solution between the co-occurrence matrices was minimal, considering $K = 2$ as a solution for the dataset. The curve of agreement classification proportion suddenly increased to approximately 93% when $K = 3$ was considered. In other words, 93% of the time, the correct number of components for the best solution to the data was $K = 3$. That was not surprising, because the best solution for the whole dataset was $K = 3$. Then, the curve of proportion

Bibliography

Abdalla, A. and Michael, S. (2019), "Finite mixture of regression models for a stratified sample," *Journal of Statistical Computation and Simulation*, 89, 2782–2800.

Baudry, J.-P. and Celeux, G. (2015), "EM for mixtures," *Statistics and computing*, 25, 713–726.

Binder, D. A. (1983), "On the variances of asymptotically normal estimators from complex surveys," *International Statistical Review/Revue Internationale de Statistique*, 279–292.

Centers for Disease, C. and Prevention (2013-2016), "National Health and Nutrition Examination Survey," Available at https://www.cdc.gov/nchs/nhanes/.

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

Chen, Q., Elliott, M. R., Haziza, D., Yang, Y., Ghosh, M., Little, R. J., Sedransk, J., Thompson, M., et al. (2017), "Approaches to improving survey-weighted estimates," *Statistical Science*, 32, 227–248.

Cochran, W. G. (1977), *Sampling Techniques (3rd edition)*, New York: John Wiley and Sons, Inc.

Cramer, H. (1946), *Mathematical methods of statistics*, Princeton, New Jersey: Princeton University Press.

Day, N. E. (1969), "Estimating the components of a mixture of normal distributions," *Biometrika*, 56, 463–474.

De Veaux, R. D. (1989), "Mixtures of linear regressions," *Computational Statistics & Data Analysis*, 8, 227–245.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, 39, 1–38.

DuMouchel, W. H. and Duncan, G. J. (1983), "Using sample survey weights in multiple regression analyses of stratified samples," *Journal of the American Statistical Association*, 78, 535–543.

Efron, B. and Hinkley, D. V. (1978), "Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information," *Biometrika*, 65, 457–483.

Fraley, C. and Raftery, A. E. (1998), "How many clusters? Which clustering method? Answers via model-based cluster analysis," *The computer journal*, 41, 578–588.

— (2006), "MCLUST version 3 for R: Normal mixture modeling and model-based clustering," Tech. rep., Citeseer.

Frühwirth-Schnatter, S. (2006), *Finite mixture and Markov switching models*, Springer Science & Business Media.

Fuller, W. A. (1975), "Regression analysis for sample survey," *Sankhya*, 37, 117–132.

Gelman, A. et al. (2007), "Struggles with survey weighting and regression modeling," *Statistical Science*, 22, 153–164.

Godambe, V. P. (1955), "A Unified Theory of Sampling From Finite Populations," *Journal of the Royal Statistical Society. Series B (Methodological)*, 17, 269–278.

Grün, B. and Leisch, F. (2004), *Bootstrapping finite mixture models*, na.

— (2008), *Finite Mixtures of Generalized Linear Regression Models*, Heidelberg: Physica-Verlag HD, pp. 205–230.

Harlan, W. R., Landis, J. R., Schmouder, R. L., Goldstein, N. G., and Harlan, L. C. (1985), "Blood lead and blood pressure: relationship in the adolescent and adult US population," *Jama*, 253, 530–534.

Haziza, D., Beaumont, J.-F., et al. (2017), "Construction of weights in surveys: A review," *Statistical Science*, 32, 206–226.

Haziza, D. and Lesage, É. (2016), "A discussion of weighting procedures for unit nonresponse," .

Hennig, C. (2000), "Identifiablity of models for clusterwise linear regression," *Journal of Classification*, 17, 273–296.

Holt, D., Smith, T., and Winter, P. (1980), "Regression analysis of data from complex surveys," *Journal of the Royal Statistical Society: Series A (General)*, 143, 474–487.

Horvitz, D. G. and Thompson, D. J. (1952), "A generalization of sampling without replacement from a finite universe," *Journal of the American statistical Association*, 47, 663–685.

Jones, P. and McLachlan, G. (1992), "Fitting finite mixture models in a regression context," *Australian Journal of Statistics*, 34, 233–240.

Kalton, G. and Graham, K. (1983), *Introduction to survey sampling*, vol. 35, Sage.

Khalili, A. and Chen, J. (2007), "Variable selection in finite mixture of regression models," *Journal of the american Statistical association*, 102, 1025–1038.

Kish, L. (1965), *Survey sampling*, New York: John Wiley and Sons, Inc.

— (1992), "Weighting for unequal Pi," *Journal of Official Statistics*, 8, 183.

Leisch, F. (2004), "FlexMix: A General Framework for Finite Mixture Models and Latent Class Regression in R," *Journal of Statistical Software*, 11, 1–18.

Li, J. and Valliant, R. (2015), "Linear regression diagnostics in cluster samples," *Journal of Official Statistics*, 31, 61–75.

Lohr, S. (2010), "Sampling: Design and analysis–Second edition. Brooks/Cole Cengage Learning," .

Louis, T. A. (1982), "Finding the observed information matrix when using the EM algorithm," *Journal of the Royal Statistical Society: Series B (Methodological)*, 44, 226–233.

Lumley, T. (2004), "Analysis of Complex Survey Samples," *Journal of Statistical Software*, 9, 1–19, r package verson 2.2.

— (2011), *Complex surveys: a guide to analysis using R*, vol. 565, John Wiley & Sons.

Lumley, T. and Scott, A. (2017), "Fitting regression models to survey data," *Statistical Science*, 32, 265–278.

Maitra, R. (2009), "Initializing Partition-Optimization Algorithms," *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, 6, 144–157.

McLachlan, G. and Peel, D. (2000), *Finite Mixture Models*, New York: John Wiley and Sons, Inc.

McLachlan, G. J. and Basford, K. E. (1988), *Mixture models: Inference and applications to clustering*, vol. 84, M. Dekker New York.

McLachlan, G. J., Lee, S. X., and Rathnayake, S. I. (2019), "Finite mixture models," *Annual review of statistics and its application*, 6, 355–378.

McNicholas, P. D. (2016), *Mixture model-based classification*, Chapman and Hall/CRC.

Meilijson, I. (1989), "A fast improvement to the EM algorithm on its own terms," *Journal of the Royal Statistical Society: Series B (Methodological)*, 51, 127–138.

Melnykov, V., Michael, S., and Melnykov, I. (2015), *Recent Developments in Model-Based Clustering with Applications*, pp. 1–39.

Michael, S. and Melnykov, V. (2016), "An effective strategy for initializing the EM algorithm in finite mixture models," *Advances in Data Analysis and Classification*, 10, 563–583.

Natarajan, S., Lipsitz, S. R., Fitzmaurice, G., Moore, C. G., and Gonin, R. (2008), "Variance estimation in complex survey sampling for generalized linear models," *Journal of the Royal Statistical Society: Series C*, 57, 75–87.

Newcomb, S. (1886), "A generalized theory of the combination of observations so as to obtain the best result," *American Journal of Mathematics*, 8, 343–366.

Neyman, J. (1934), "On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection," *Journal of the Royal Statistical Society*, 97, 558–625.

Pearson, K. (1894), "Contributions to the mathematical theory of evolution," *Philosophical Transactions of the Royal Society of London. A*, 185, 71–110.

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

Pfeffermann, D. and Smith, T. (1985), "Regression Models for Grouped Populations in Cross-Section Surveys, Correspondent Paper," *International Statistical Review/Revue Internationale de Statistique*, 37–59.

Quandt, R. E. and Ramsey, J. B. (1978), "Estimating mixtures of normal distributions and switching regressions," *Journal of the American statistical Association*, 73, 730–738.

Rabe-Hesketh, S. and Skrondal, A. (2004), *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*, Chapman and Hall/CRC.

Royall, R. (1968), "An old approach to finite population sampling theory," *Journal of the American Statistical Association*, 63, 1269–1279.

Royall, R. M. (1986), "Model robust confidence intervals using maximum likelihood estimators," *International Statistical Review/Revue Internationale de Statistique*, 221–226.

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

Scheaffer, R. L., Mendenhall III, W., Ott, R. L., and Gerow, K. G. (2011), *Elementary survey sampling*, Cengage Learning.

Schlattmann, P. (2009), *Medical applications of finite mixture models.*, Springer.

Schwarz, G. et al. (1978), "Estimating the dimension of a model," *The annals of statistics*, 6, 461–464.

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

Smith, T. (1984), "Present position and potential developments: Some personal views: Sample surveys," *Journal of the Royal Statistical Society. Series A (General)*, 208–221.

Sugden, R. A. and Smith, T. M. F. (1984), "Ignorable and Informative Designs in Survey Sampling Inference," *Biometrika*, 71, 495–506.

Thomas, N., Raghunathan, T. E., Schenker, N., Katzoff, M. J., and Johnson, C. L. (2006), "An evaluation of matrix sampling methods using data from the National Health and Nutrition Examination Survey," *Survey Methodology*, 32, 217.

Van der Vaart, A. W. (2000), *Asymptotic statistics*, vol. 3, Cambridge university press.

Wedel, M. and Kamakura, W. A. (2012), *Market segmentation: Conceptual and methodological foundations*, vol. 8, Springer Science & Business Media.

Wedel, M., Ter Hofstede, F., and Steenkamp, J.-B. E. (1998), "Mixture model analysis of complex samples," *Journal of Classification*, 15, 225–244.

White, H. (1982), "Maximum likelihood estimation of misspecified models," *Econometrica: Journal of the Econometric Society*, 1–25.

Wickham, H. and Miller, E. (2019), *haven: Import and Export 'SPSS', 'Stata' and 'SAS' Files*, r package version 2.1.0.

Zhang, C. (ed.) (2017), *Special Section on Complex Surveys*, vol. 32, New York: The Institute of Mathematical Statistics.

CURRICULUM VITAE

https://www.linkedin.com/in/abdelbaset-abdalla-53533315b abdelbaset.abdalla@sdstate.edu
(303) 513-4376

CAREER OBJECTIVE

With a Ph.D. degree in computational science & statistics, I hope to convey my experiences gained to the others, and I strive to develop the teaching, consulting, and my research at a university.

EDUCATION

Ph.D. in Computational Science and Statistics *Fall 2019*
South Dakota State University, Bookings, SD

MS in Statistics *May 2008*
University of Benghazi, Libya GPA: 3.40/4.00

BSc in Statistics *Feb 2001*
University of Benghazi, Libya GPA: 3.13/4.00

TECHNICAL& SKILLS

- **Packages:** R, MINITAB, SPSS, SAS
- **Word processing:** Microsoft, LaTeX

WORK EXPERIENCE

Graduate Teaching Assistance
SDSU Department of Math & Statistics

- Teaching Assistant of Statistical Methods II for the graduate students *Aug - Dec, 2019*
- Teaching Assistance of Statistical Methods I *June - Aug, 2019*
- Teaching Assistance of Statistical Methods II for the graduate students *Aug - May, 2018, 2019*
- Teaching Assistance of Statistical Methods I *Jan - May, 2016, 2017, 2018*
- Teaching Assistance of Statistical Inference I & II for the graduate students *Aug - Dec, 2016, 2017*
- Teaching Assistance of College Algebra *Aug - Dec, 2015, 2018*

Lecturer *Aug 2008 - May 2013*
University of Benghazi, Libya

- Courses Taught: General Statistics, Basic Statistics, Elements of Probability, Statistical Methods, Multivariate Analysis, Distribution Theory, Estimation Theory, Non-Parametric Methods, Advanced Data Analysis.

Graduate Teaching Assistance
University of Benghazi, Libya

- Teaching assistant of two section per semester of Data Analysis *Aug - Jan, 2007, 2008*