DISSERTATION

IMPROVED ESTIMATION FOR COMPLEX SURVEYS USING MODERN REGRESSION TECHNIQUES

Submitted by

Kelly McConville

Department of Statistics

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Summer 2011

Doctoral Committee:

Advisor: F. Jay Breidt
Co-Advisor: Thomas C. M. Lee

Jean Opsomer
Myung-Hee Lee
Paul Doherty

ABSTRACT

IMPROVED ESTIMATION FOR COMPLEX SURVEYS USING MODERN REGRESSION TECHNIQUES

In the field of survey statistics, finite population quantities are often estimated based on complex survey data. In this thesis, estimation of the finite population total of a study variable is considered. The study variable is available for the sample and is supplemented by auxiliary information, which is available for every element in the finite population. Following a model-assisted framework, estimators are constructed that exploit the relationship which may exist between the study variable and ancillary data. These estimators have good design properties regardless of model accuracy.

Nonparametric survey regression estimation is applicable in natural resource surveys where the relationship between the auxiliary information and study variable is complex and of an unknown form. Breidt, Claeskens, and Opsomer (2005) proposed a penalized spline survey regression estimator and studied its properties when the number of knots is fixed. To build on their work, the asymptotic properties of the penalized spline regression estimator are considered when the number of knots goes to infinity and the locations of the knots are allowed to change. The estimator is shown to be design consistent and asymptotically design unbiased. In the course of the proof, a result is established on the uniform convergence in probability of the survey-weighted quantile estimators. This result is obtained by deriving a survey-weighted Hoeffding inequality for bounded random variables. A variance estimator is proposed and shown to be design consistent for the asymptotic mean squared error. Simulation results demonstrate the usefulness of the asymptotic approximations.

Also in natural resource surveys, a substantial amount of auxiliary information, typically derived from remotely-sensed imagery and organized in the form of spatial layers in a geographic information system (GIS), is available. Some of this ancillary data may be extraneous and a sparse model would be appropriate. Model selection methods are therefore warranted. The 'least absolute shrinkage and selection operator' (lasso), presented by Tibshirani (1996), conducts model selection and parameter estimation simultaneously by penalizing the sum of the absolute values of the model coefficients. A survey-weighted lasso criterion, which accounts for the sampling design, is derived and a survey-weighted lasso estimator is presented. The root-n design consistency of the estimator and a central limit theorem result are proved. Several variants of the survey-weighted lasso estimator are constructed. In particular, a calibration estimator and a ridge regression approximation estimator are constructed to produce lasso weights that can be applied to several study variables. Simulation studies show the lasso estimators are more efficient than the regression estimator when the true model is sparse. The lasso estimators are used to estimate the proportion of tree canopy cover for a region of Utah. Under a joint design-model framework, the survey-weighted lasso coefficients are shown to be root-N consistent for the parameters of the superpopulation model and a central limit theorem result is found. The methodology is applied to estimate the risk factors for the Zika virus from an epidemiological survey on the island of Yap. A logistic survey-weighted lasso regression model is fit to the data and important covariates are identified.

ACKNOWLEDGMENTS

I would like to thank the Colorado State Department of Statistics faculty and graduate students for shaping me into the statistician, researcher and teacher I am today. Dr. Jay Breidt, my advisor, was quite instrumental in this process. He provided me with generous research support and many times helped me get un-stuck when a problematic proof was blocking my way. I would also like to thank Dr. Thomas Lee, my co-advisor, for allowing me the great pleasure of working with him in Hong Kong and Davis, CA. It is Dr. Jean Opsomer, a member of my doctoral committee, who I must thank for introducing me to the world of survey statistics during STAT605. Also, I am grateful to both Dr. Myung-Hee Lee and Dr. Paul Doherty for being members of my doctoral committee. I appreciated the thought-provoking questions and commentary they both provided. I would like to thank Dr. Duane Boes, my fellow Iowan, for coming out of retirement to teach me STAT730. Duane taught me to not be sloppy in my work because remembering one's indicators can make all the difference.

Additionally, I want to acknowledge the moral and mental support of my friends and family. I can't give Austin enough credit for helping me finish my dissertation. He made sure we celebrated the small victories along the way and stayed confident in my abilities even when my confidence faltered. My step-mother, Beth, was my own personal cheerleader throughout the process and always had a motivating pep-talk ready. I would like to thank Beth and my father for showing more interest in statistics and my work than I would guess they actually had. Also, I need to acknowledge my little brother, Thomas, the mathematician, for always attempting to answer my math questions, no matter how obscure they were. And, when I needed a break from statistics, my friends, Julie, Nick and Sara, all provided me with the laughter and wonderful diversions I needed.

This research was supported in part by the National Science Foundation (SES-0922142) and by the Program for Interdisciplinary Mathematics, Ecology and Statistics, a National Science Foundation IGERT grant (DGE-0221595).

TABLE OF CONTENTS

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Survey statistics

Survey statistics differs from other fields of statistics because of the emphasis placed on inference about a definable, finite populaton at a particular point in time. Here we look at two populations: a region of semi-forested land in Utah in 2010 and the human population on the island of Yap during a Zika outbreak. Much of our discussion centers on the descriptive study of survey samples and in particular on estimating the population total for a study variable, y. For the region of Utah, we are interested in estimating the percent tree canopy cover for the region, which is defined as the percent of forest floor covered by tree crowns when viewed aerially (Toney, Shaw, and Nelson 2008). Tree canopy cover is an important characteristic because it is used directly in the definition of forested land. Survey samples can also be studied analytically to draw inferences about parameters in the hypothetical model which is assumed to have generated the finite population. In this case, the analyst is more interested in understanding the mechanism or system which created the population or in understanding the population over time, of which the current population is just a snapshot. For the human population of Yap, we want to know what covariates are associated with the probability of a person being infected with the Zika virus, a vector borne illness. Therefore, emphasis is placed on better understanding of some underlying model and not on a descriptive value for the finite population.

In this dissertation, we construct estimators which incorporate auxiliary information for both descriptive and analytic inferences. Complex survey data are increasingly augmented by auxiliary information since this ancillary data, such as large-scale photography or other remote sensing information, tends to be less expensive to collect and is often known for each element of the population. For the region of Utah, along with tree canopy cover, we have Landsat satellite bands and geographic information systems layers such as aspect and slope. For the population of Yap, in addition to infection data, we have data from a questionnaire that collected demographic and risk factor information. In each case, we want to use the auxiliary information to inform on the non-sampled study variable elements.

To conduct descriptive inference, we follow the typical framework as given by Särndal, Swensson, and Wretman (1992). For the enumerated finite population $U = \{1, 2, \ldots, N\}$, we want to estimate a function of the study variable y, and we primarily focus on estimating the total of y, $t_y = \sum_{j \in U} y_j$. Since conducting a census is typically too expensive and time consuming, we assume a sample s of size n is collected according to some sampling scheme such as stratified simple random sampling, multi-stage sampling, cluster sampling, etc. Once a sampling scheme is chosen, we can find the sampling design, $p(\cdot)$, where $p(s) = P(S = s)$, S is a random set representing the sample, and s is the realized sample. Since the study variable y_j is collected for $j \in s$, we can estimate the finite population quantity t_y with an estimator based on the sampled values, $\hat{t}_y(s)$. For simplicity of notation, we write this estimator as $\hat{t}_y$, but it is important to note that the estimator is based on the random quantity S. Under design-based inference, the study variable, y, is considered to be a fixed number for each element in the population and the randomness comes from the sample-to-sample variation induced by the sampling design $p(\cdot)$.

To construct estimators and to understand the properties of these estimators, we need to know the probability that any element in the population is included in the sample. Therefore, we define the first-order inclusion probability for element $j \in U$ as $\pi_j = P(j \in s) = \sum_{s: j \in s} p(s)$ and the second order inclusion probability of elements j_1 and $j_2 \in U$ as $\pi_{j_1 j_2} = P(j_1, j_2 \in s) = \sum_{s: j_1, j_2 \in s} p(s)$. Once a sample is obtained, we can find sample membership indicators for each element in the population. To denote sample inclusion for element j, let $I_j = 1$ if $j \in s$ and 0 otherwise. This indicator has the property $E_p I_j = P(j \in s) = \pi_j$ where $E_p(\cdot)$ represents the expectation with respect to the sampling design $p(\cdot)$.

A common estimator for t_y based on the sampled values and their inclusion probabilities, $\{y_j, \pi_j; j \in s\}$, is the Horvitz-Thompson estimator

$$\hat{t}_{y,HT} = \sum_{j \in U} y_j \frac{I_j}{\pi_j} \tag{1.1}$$

(Horvitz and Thompson 1952). The Horvitz-Thompson estimator is called a design-based estimator because it accounts for the sampling design. It is design unbiased, a desirable property for a survey estimator, and the design variance of the Horvitz-Thompson estimator is

$$\mathrm{Var}_p\left(\hat{t}_{y,HT}\right) = \sum\sum_{j,k \in U} \triangle_{jk} \frac{y_j}{\pi_j} \frac{y_k}{\pi_k}$$

where $\triangle_{jk} = \pi_{jk} - \pi_j \pi_k$. Since the Horvitz-Thompson estimator seeks to estimate the total of the population by means of a total over the sample, each value in the sample is inflated by its inverse inclusion probability. This inflation can be loosely understood as the amount of elements in the population that the sample element represents. While the Horvitz-Thompson estimator is both intuitive and easy to calculate, it typically lacks efficiency because it is purely design-based and does not utilize a model. If we assume auxiliary information, which we denote by $\boldsymbol{x}$, is available for all elements in the population, then we can possibly gain efficiency in our estimator by predicting the non-sampled y values with a model that relates the study variable and the auxiliary information.

1.2 Model-assisted estimation

To incorporate the relationship between $\boldsymbol{x}$ and y into the estimation of t_y, we introduce a superpopulation model and consider the finite population values $\{y_j; j \in U\}$ to be realizations of the model. Denoting the superpopulation by ξ, we assume that conditional on $\boldsymbol{x}_j$,

$$y_j = f(\boldsymbol{x}_j) + \epsilon_j \tag{1.2}$$

and the errors, ϵ_j, are independent, identically distributed random variables with mean zero and variance σ^2. We utilize (1.2) under design-based inference by adopting a model-assisted framework where the randomness still stems from the sampling design, $p(s)$, and not the stochastic model (1.2). Under this construction, estimators for t_y are judged based on their design properties, such as design consistency and asymptotic design unbiasedness. In essence, model-assisted estimators should be robust to model misspecification where robustness implies the estimators have good design properties regardless of how accurate the assumed model is. To emphasize this point, the superpopulation model is often referred to as the working model, which implies it is simply an estimation tool and not the foundation for inference. The working model is utilized to increase efficiency of survey estimators. As stated by Hansen, Madow, and Tepping (1983), we seek estimators "that for large enough samples the validity of randomization (design-based) inference does not depend on assumptions concerning the distribution of characteristics in the finite population from which the sample is drawn."

Several model-assisted estimators have been investigated, such as the ratio estimator (Cochran (1977), Ch. 6-7), the calibration estimator (Deville and Särndal 1992), and the generalized regression estimator (Cassel, Särndal, and Wretman 1976). In this dissertation, we study the generalized regression estimator and its properties under different assumed superpopulation models.

1.2.1 Generalized regression estimator

In order to understand the form and properties of the generalized regression estimator, we first introduce the generalized difference estimator. Suppose the mean function of (1.2) can be estimated with a function of the finite population which we denote by $f_U\left(\boldsymbol{x}_j; \boldsymbol{X}_U, \boldsymbol{Y}_U\right)$ where $\boldsymbol{X}_U$ and $\boldsymbol{Y}_U$ are the matrix of covariates and vector of the study variable at the population level, respectively. For ease of notation, write $f_U\left(\boldsymbol{x}_j; \boldsymbol{X}_U, \boldsymbol{Y}_U\right) = f_U\left(\boldsymbol{x}_j\right)$. If the mean function is linear, $f(\boldsymbol{x}_j) = \boldsymbol{x}_j^T \boldsymbol{\beta}$, then the ordinary least squares coefficient estimates, $\boldsymbol{\beta}_U = \left(\boldsymbol{X}_U^T \boldsymbol{X}_U\right)^{-1} \boldsymbol{X}_U^T \boldsymbol{Y}_U$, are appropriate estimates of the superpopulation coefficients, $\boldsymbol{\beta}$, and therefore $f_U\left(\boldsymbol{x}_j\right) = \boldsymbol{x}_j^T \boldsymbol{\beta}_U$ estimates $f(\boldsymbol{x}_j)$. Once we have an estimate of $f(\boldsymbol{x}_j)$, we can construct the generalized difference estimator

$$\hat{t}_{y,diff} = \sum_{j \in s} \frac{y_j - f_U(\boldsymbol{x}_j)}{\pi_j} + \sum_{j \in U} f_U(\boldsymbol{x}_j) \tag{1.3}$$

(Särndal, Swensson, and Wretman 1992). Noting that the finite population quantity $f_U(\boldsymbol{x}_j)$ is not random because it is based on census data, we can easily see the difference estimator is design unbiased. Further, we can find the design variance of the generalized difference estimator

$$\text{Var}_p\left(\hat{t}_{y,diff}\right) = \sum\sum_{j,k \in U} \triangle_{jk} \frac{(y_j - f_U(\boldsymbol{x}_j))}{\pi_j} \frac{(y_k - f_U(\boldsymbol{x}_k))}{\pi_k}.$$

As long as $f_U(\boldsymbol{x}_j)$ is a decent approximation for y_j, $j \in U$, the variance of the difference estimator will be smaller than the variance of the Horvitz-Thompson estimator since it is based on residuals $(y_j - f_U(\boldsymbol{x}_j))$ instead of raw values (y_j).

Based on our assumptions, we cannot compute $f_U(\boldsymbol{x}_j)$, and consequently $\hat{t}_{y,diff}$, since they both depend on $\boldsymbol{Y}_U$ and we only have $\boldsymbol{Y}_s$, the vector of study variable values for the sample. Therefore, we must estimate $f_U(\boldsymbol{x}_j)$ with a sample quantity which we denote by $\hat{f}_s(\boldsymbol{x}_j)$ $(= f_s(\boldsymbol{x}_j; \boldsymbol{X}_s, \boldsymbol{Y}_s, \boldsymbol{\Pi}_s))$. Here, $\boldsymbol{X}_s$ and $\boldsymbol{Y}_s$ are the matrix of covariates and vector of the study variable at the sample level and $\boldsymbol{\Pi}_s$ is a diagonal matrix of the inclusion probabilities for the sampled values. A common survey estimator for a finite population quantity that can be written as a function of population totals is the Horvitz-Thompson 'plug-in' estimator where the population totals are each replaced by their Horvitz-Thompson estimator. Returning to the linear model example, we can write the finite population coefficient vector as

$$\begin{aligned} \boldsymbol{\beta}_U &= \left(\boldsymbol{X}_U^T \boldsymbol{X}_U\right)^{-1} \boldsymbol{X}_U^T \boldsymbol{Y}_U \\ &= \left(\sum_{j \in U} \boldsymbol{x}_j \boldsymbol{x}_j^T\right)^{-1} \sum_{j \in U} \boldsymbol{x}_j y_j \end{aligned}$$

and therefore its Horvitz-Thompson 'plug-in' estimator is found by replacing the totals in $\sum_{j \in U} \boldsymbol{x}_j \boldsymbol{x}_j^T$ and $\sum_{j \in U} \boldsymbol{x}_j y_j$ with their corresponding Horvitz-Thompson estimators to obtain

$$\hat{\boldsymbol{\beta}}_s = \left(\sum_{j \in s} \frac{\boldsymbol{x}_j \boldsymbol{x}_j^T}{\pi_j} \right)^{-1} \sum_{j \in s} \frac{\boldsymbol{x}_j y_j}{\pi_j}$$
$$= \left(\boldsymbol{X}_s^T \boldsymbol{\Pi}_s^{-1} \boldsymbol{X}_s \right)^{-1} \boldsymbol{X}_s^T \boldsymbol{\Pi}_s^{-1} \boldsymbol{Y}_s.$$

Replacing the finite population quantity, $f_U(\boldsymbol{x}_j)$, in (1.3) with the sample quantity, $\hat{f}_s(\boldsymbol{x}_j)$, we obtain the generalized regression estimator

$$\hat{t}_y = \sum_{j \in s} \frac{y_j - \hat{f}_s(\boldsymbol{x}_j)}{\pi_j} + \sum_{j \in U} \hat{f}_s(\boldsymbol{x}_j) \tag{1.4}$$

(Cassel, Särndal, and Wretman 1976). The generalized difference estimator is model-assisted because it is design unbiased and has a valid and usually efficient (in comparison to the Horvitz-Thompson estimator) design variance regardless of the assumed superpopulation model. The generalized regression estimator is not exactly design unbiased but, under a few weak assumptions, is both asymptotically design unbiased and design consistent. These design properties rely on the form of the estimator. Suppose $\hat{f}_s(\boldsymbol{x}_j)$ is a 'bad' estimate for y_j in the sense that $\hat{f}_s(\boldsymbol{x}_j)$ tends to be negatively biased for y_j, $j \in U$. This implies the second component of (1.4) will be negatively biased for t_y. In this situation, typically $\hat{f}_s(\boldsymbol{x}_j) \leq y_j$, which means the first term in (1.4) will be positive so that the overall estimator, $\hat{t}_{y,reg}$ is approximately design unbiased. We can make a similar argument if $\hat{f}_s(\boldsymbol{x}_j)$ tends to be positively biased for y_j, $j \in U$. If $\hat{f}_s(\boldsymbol{x}_j)$ tends to be a 'good' (approximately unbiased) estimator for y_j then the first term in (1.4) will be small and again the overall estimator will be 'good'. Therefore, the first term in (1.4) is referred to as the 'design-bias' adjustment because, using the design weights, it appropriately accounts for a 'bad' model.

An analogous model-based regression estimator is

$$\tilde{t}_y = \sum_{j \in s} y_j + \sum_{j \in U-s} \tilde{f}_s(\boldsymbol{x}_j). \tag{1.5}$$

In this thesis, we assume $\tilde{f}_s(\boldsymbol{x}_j)$ does not directly account for the sampling design because typically in a model-based framework the inclusion probabilities are considered unnecessary information (Hansen, Madow, and Tepping 1983). Continuing the linear model example, a sample model-based estimate for the finite population coefficient vector is the ordinary least squares estimator

$$\tilde{\boldsymbol{\beta}}_s = \left(\sum_{j \in s} \boldsymbol{x}_j \boldsymbol{x}_j^T \right)^{-1} \sum_{j \in s} \boldsymbol{x}_j y_j \tag{1.6}$$

$$= \left(\boldsymbol{X}_s^T \boldsymbol{X}_s \right)^{-1} \boldsymbol{X}_s^T \boldsymbol{Y}_s. \tag{1.7}$$

The estimator (1.5) fails to be model-assisted because if $\tilde{f}_s(\boldsymbol{x}_j)$ is a design biased estimate for y_j then (1.5) is also design biased. Inference on (1.5) relies on the accuracy of the assumed superpopulation model. The strengths and weaknesses of model-based versus model-assisted or design-based estimators along with the corresponding paradigms of inference have been extensively studied (Hansen, Madow, and Tepping (1983); Särndal, Swensson, and Wretman (1992); Smith (1994); and Gregoire (1998)). In this thesis, we primarily study model-assisted estimators under design-based inference because we want to describe a particular finite population without relying on the superpopulation model for accuracy of inference. However, in chapter 3, section 3.9, we discuss analytic inference, which necessitates a model and therefore we employ a joint design-model framework for that scenario.

For the generalized regression estimator, various parametric models have been assumed for $f(\boldsymbol{x}_j)$ and their properties are summarized in Särndal, Swensson, and Wretman (1992). We consider (1.4) under two possible superpopulation models (1.2): a nonparametric model and a linear model where the number of potential covariates is quite large but the true model is sparse.

1.2.2 Nonparametric regression estimators

Since the gain in design efficiency for the generalized regression estimator does rely on the accuracy of the working model, nonparametric models, which are more flexible and can account for more complex model structures, have been proposed to estimate f in (1.4). In such cases, one only needs to assume the mean function is a smooth function in x. Breidt and Opsomer (2000) employed local polynomial regression to estimate f. At the population level, they fit the local polynomial regression mean function for $f_U(\boldsymbol{x}_j)$ in (1.3) and then estimated $f_U(\boldsymbol{x}_j)$ with a survey-weighted local polynomial regression mean function $\hat{f}_s(\boldsymbol{x}_j)$ to produce the local polynomial regression estimator, a nonparametric version of (1.4). They showed design consistency and asymptotic design unbiasedness of the estimator along with proving asymptotic equivalence of the design mean squared error of the estimator and the design variance of the generalized difference estimator. Additionally, they derived a variance estimator for the design mean squared error and showed it was both design consistent and asymptotically design unbiased for the design mean squared error. When the true superpopulation is non-linear, the local polynomial regression estimator out-performed its parametric counterparts.

Breidt, Claeskens, and Opsomer (2005) proposed the penalized spline regression estimator where penalized splines estimate f in (1.4). As is common in the penalized spline literature, they assumed the number and location of the knots to be fixed when studying the asymptotic properties of the penalized spline regression estimator. In chapter 2, we consider the penalized spline regression estimator of Breidt, Claeskens, and Opsomer (2005) and look at its asymptotic properties when the locations of the knots are allowed to change and the number of knots goes to infinity.

1.2.3 Survey estimation and model selection

In the survey setting, there is often a large number of auxiliary variables available. For example, in natural resource inventories conducted by the United States Forest Service, the auxiliary variables consist of multiple layers of processed remote sensing data. Because these layers are frequently correlated and potentially do not have a significant relationship with the variable of interest, model selection is appropriate to remove extraneous variables. The 'least absolute shrinkage and selection operator' (lasso) method proposed by Tibshirani (1996) simultaneously performs model selection and coefficient estimation by shrinking unnecessary coefficients to zero. In a non-survey context, the lasso estimator outperforms the ordinary least squares estimator when the true model is sparse. In chapter 3, we estimate f in (1.4) with a survey-weighted lasso regression model and construct a survey-weighted lasso regression estimator.

Chapter 2

Penalized spline regression estimator

2.1 Introduction

In this chapter, we explore the asymptotic behavior of (1.4) when $f(\boldsymbol{x}_j)$ is modeled with piece-wise penalized splines (p-splines) with a first-order difference penalty. We allow the number of knots to increase and the location of the knots to change as N increases. In section 2.1.1, we apply the methods of Li and Ruppert (2008) to derive the explicit form of the finite population p-spline coefficients and then construct Horvitz-Thompson 'plug-in' estimates of those coefficients. We also prove the asymptotic equivalence of the proposed estimator to the one derived by Breidt, Claeskens, and Opsomer (2005). In section 2.2.2 we show the asymptotic design mean squared error equals the design variance of the difference estimator, in section 2.2.3 we prove the design mean squared consistency of the estimator and in section 2.2.4 we prove the consistency of the variance estimator for the asymptotic design mean squared error. In section 2.3 we discuss alternate estimators and in section 2.4 we present simulation results.

2.1.1 Derivation of the estimator

Assume x_j is univariate and the superpopulation model is (1.2). Also, assume an appropriate estimate of the mean function is

$$f_U(x_j) = \tilde{\boldsymbol{I}}_j^T \boldsymbol{\beta}_U \tag{2.1}$$

where $\tilde{\boldsymbol{I}}_j = (\tilde{I}_{1j}, \tilde{I}_{2j}, \ldots, \tilde{I}_{Kj})^T$ and $\tilde{I}_{ij} = I\{\kappa_{Ui-1} \leq x_j < \kappa_{Ui}\}$ with knots $\{\kappa_{Ui}\}_{i=0}^K$. The number of cells is denoted by K and let $C_U = K^{-1}N$ where we assume C_U is an integer for simplicity. To ensure the x's are placed evenly between knots, every C_U-th x is a knot. The finite population coefficient vector, $\boldsymbol{\beta}_U$, minimizes

$$\sum_{j \in U} \left\{ y_j - \tilde{\boldsymbol{I}}_j^T \boldsymbol{\beta} \right\}^2 + \lambda \sum_{i=2}^{K} (\beta_i - \beta_{i-1})^2$$

where λ, the smoothness parameter, is a fixed, positive number. The p-spline solution for $\boldsymbol{\beta}_U$ can be written in 'ridge regression' format

$$\left(\boldsymbol{X}_U^T \boldsymbol{X}_U + \lambda \boldsymbol{D}^T \boldsymbol{D} \right) \boldsymbol{\beta}_U \quad = \quad \boldsymbol{X}_U^T \boldsymbol{y}_U \tag{2.2}$$

where $\boldsymbol{X}_U = \left[\tilde{\boldsymbol{I}}_1^T, \tilde{\boldsymbol{I}}_2^T, \ldots, \tilde{\boldsymbol{I}}_N^T \right]^T$, $\boldsymbol{y}_U = (y_1, y_2, \ldots, y_N)^T$, and the differencing matrix, $\boldsymbol{D}$, satisfies

$$\boldsymbol{D}\boldsymbol{\beta}_U = \begin{pmatrix} \beta_{U2} - \beta_{U1} \\ \beta_{U3} - \beta_{U2} \\ \vdots \\ \beta_{UK} - \beta_{UK-1} \end{pmatrix}.$$

Dividing both sides of (2.2) by $C_U + 2\lambda$ results in $\boldsymbol{\Omega}_U = (C_U + 2\lambda)^{-1} \left(\boldsymbol{X}_U^T \boldsymbol{X}_U + \lambda \boldsymbol{D}^T \boldsymbol{D} \right)$ with elements

$$\Omega_{U(1,1)} = \Omega_{U(K,K)} = \theta_U = (C_U + 2\lambda)^{-1} (C_U + \lambda), \tag{2.3}$$

for $1 < i < K$, $\Omega_{U(i,i)} = 1$, for $|i - j| = 1$

$$\Omega_{U(i,j)} = \eta_U = -\left(C_U + 2\lambda\right)^{-1}\lambda, \tag{2.4}$$

and for $|i - j| > 1, \Omega_{U(i,j)} = 0$. Following the methods of Li and Ruppert (2008), we exploit the tri-diagonal, banded structure found in all but the first and last columns of $\boldsymbol{\Omega}_U$. This banded structure allows us to find vectors of the form

$$T_t(\rho_U) = (\rho_U^{t-1}, \rho_U^{t-2}, \ldots, \rho_U, 1, \rho_U, \ldots, \rho_U^{K-t})$$

which are orthogonal to all columns of $\boldsymbol{\Omega}_U$ except the first, last, and t-th. Each element in the vector $T_t(\rho_U)$ contains a power of

$$\rho_U \quad = \quad \frac{C_U + 2\lambda - (C_U^2 + 4\lambda C_U)^{1/2}}{2\lambda},$$

and ρ_U, a function of the smoothing parameter, knot size, and population size, is between zero and one. For simplicity of notation, we suppress the dependence on U in θ, η, and ρ. Utilizing the vectors $\boldsymbol{T}_t(\rho)$, we can explicitly solve for the elements of $\boldsymbol{\beta}_U$ without inverting $\boldsymbol{\Omega}_U$. Since $\boldsymbol{T}_1(\rho)$ and $\boldsymbol{T}_K(\rho)$ are orthogonal to all but the first and last columns of $\boldsymbol{\Omega}_U$,

$$\boldsymbol{T}_1(\rho)^T\boldsymbol{\Omega}_U\boldsymbol{\beta}_U = \boldsymbol{T}_1(\rho)^T\left(C_U + 2\lambda\right)^{-1}\boldsymbol{X}_U^T\boldsymbol{y}_U \text{ and } \boldsymbol{T}_K(\rho)^T\boldsymbol{\Omega}_U\boldsymbol{\beta}_U = \boldsymbol{T}_K(\rho)^T\left(C_U + 2\lambda\right)^{-1}\boldsymbol{X}_U^T\boldsymbol{y}_U$$

yield the first and last finite population coefficients

$$\beta_{U1} = \frac{(\theta + \eta\rho)\sum_{i=1}^{K}\rho^{i-1}C_U^{-1}\sum_{j\in U} y_j\tilde{I}_{ij} - \rho^{K-2}\left(\eta + \theta\rho\right)\sum_{i=1}^{K}\rho^{K-i}C_U^{-1}\sum_{j\in U} y_j\tilde{I}_{ij}}{\left\{(\theta + \eta\rho)^2 - \rho^{2(K-2)}\left(\eta + \theta\rho\right)^2\right\}\left(1 + 2\lambda C_U^{-1}\right)} \tag{2.5}$$

and

$$\beta_{UK} = \frac{(\theta+\eta\rho)\sum_{i=1}^{K}\rho^{K-i}C_U^{-1}\sum_{j\in U}y_j\tilde{I}_{ij} - \rho^{K-2}(\eta+\theta\rho)\sum_{i=1}^{K}\rho^{i-1}C_U^{-1}\sum_{j\in U}y_j\tilde{I}_{ij}}{\left\{(\theta+\eta\rho)^2 - \rho^{2(K-2)}(\eta+\theta\rho)^2\right\}\left(1+2\lambda C_U^{-1}\right)}. \tag{2.6}$$

To find the interior coefficients, where $1 < t < K$, substitute (2.5) and (2.6) into

$$\boldsymbol{T}_t(\rho)^T\boldsymbol{\Omega}_U\boldsymbol{\beta}_U = \boldsymbol{T}_t(\rho)^T\left(C_U+2\lambda\right)^{-1}\boldsymbol{X}_U^T\boldsymbol{y}_U$$

to obtain

$$\beta_{Ut} = \frac{\sum_{i=1}^{K}\rho^{|t-i|}C_U^{-1}\sum_{j\in U}y_j\tilde{I}_{ij}}{(1+2\rho\eta)\left(1+\lambda C_U^{-1}\right)} - \frac{\rho^{t-2}(\rho\theta+\eta)\beta_{U1} + \rho^{K-t-1}(\rho\theta+\eta)\beta_{UK}}{(1+2\rho\eta)}.$$

If the coefficient vector, $\boldsymbol{\beta}_U$, is known, then we can compute the p-spline fit in (2.1) and can construct

$$t_{y,diff} = \sum_{j\in s}\frac{y_j - f_U(x_j)}{\pi_j} + \sum_{j\in U}f_U(x_j), \tag{2.7}$$

the generalized difference estimator (Särndal, Swensson, and Wretman 1992). We can also compute the design variance of the generalized difference estimator

$$\mathrm{Var}_p\left(t_{y,diff}\right) = \sum\sum_{j,l\in U}\triangle_{jl}\frac{\left(y_j - \tilde{\boldsymbol{I}}_j^T\boldsymbol{\beta}_U\right)}{\pi_j}\frac{\left(y_l - \tilde{\boldsymbol{I}}_l^T\boldsymbol{\beta}_U\right)}{\pi_l} \tag{2.8}$$

where $\triangle_{jl} = \pi_{jl} - \pi_j\pi_l$. Assuming the linear combination of penalized piece-wise B-splines is a good approximation for the true model, the difference estimator will be more design efficient than the Horvitz-Thompson estimator.

Since the study variable y is collected for the sample, not the population, we must estimate (2.1), or more specifically, the finite population coefficients, $\boldsymbol{\beta}_U$. For each coefficient,

β_{Ui}, $i = 1, 2, \ldots, K$, we propose the following Horvitz-Thompson 'plug-in' estimators

$$\hat{\beta}_{s1} = \frac{(\theta+\eta\rho)\sum_{i=1}^{K}\rho^{i-1}C_U^{-1}\sum_{j\in U}y_j\tilde{I}_{ij}\frac{I_j}{\pi_j} - \rho^{K-2}(\eta+\theta\rho)\sum_{i=1}^{K}\rho^{K-i}C_U^{-1}\sum_{j\in U}y_j\tilde{I}_{ij}\frac{I_j}{\pi_j}}{\left\{(\theta+\eta\rho)^2 - \rho^{2(K-2)}(\eta+\theta\rho)^2\right\}\left(1+2\lambda C_U^{-1}\right)},$$

$$\hat{\beta}_{sK} = \frac{(\theta+\eta\rho)\sum_{i=1}^{K}\rho^{K-i}C_U^{-1}\sum_{j\in U}y_j\tilde{I}_{ij}\frac{I_j}{\pi_j} - \rho^{K-2}(\eta+\theta\rho)\sum_{i=1}^{K}\rho^{i-1}C_U^{-1}\sum_{j\in U}y_j\tilde{I}_{ij}\frac{I_j}{\pi_j}}{\left\{(\theta+\eta\rho)^2 - \rho^{2(K-2)}(\eta+\theta\rho)^2\right\}\left(1+\lambda C_U^{-1}\right)},$$

and for $1 < t < K$,

$$\hat{\beta}_{st} = \frac{\sum_{i=1}^{K}\rho^{|t-i|}C_U^{-1}\sum_{j\in U}y_j\tilde{I}_{ij}\frac{I_j}{\pi_j}}{(1+2\rho\eta)\left(1+\lambda C_U^{-1}\right)} - \frac{\left[\rho^{t-2}(\rho\theta+\eta)\hat{\beta}_{s1} + \rho^{K-t-1}(\rho\theta+\eta)\hat{\beta}_{sK}\right]}{(1+2\rho\eta)}.$$

Each Horvitz-Thompson 'plug-in' estimator, $\hat{\beta}_{Ui}$, is design unbiased for the corresponding finite population coefficient β_{Ui}. Incorporating the estimated mean function $\hat{f}_s(x_j) = \tilde{\boldsymbol{I}}_j^T\hat{\boldsymbol{\beta}}_U$ into (1.4) produces

$$\hat{t}_y = \sum_{j\in s}\frac{y_j - \tilde{\boldsymbol{I}}_j^T\hat{\boldsymbol{\beta}}_s}{\pi_j} + \sum_{j\in U}\tilde{\boldsymbol{I}}_j^T\hat{\boldsymbol{\beta}}_s, \tag{2.9}$$

the penalized spline regression estimator. Since in practice the survey weights are often applied to several study variables, it is useful to write $\hat{t}_y$ as a weighted linear combination of the sampled study variables

$$\begin{aligned}\hat{t}_y &= \sum_{j\in s}\left\{\frac{1}{\pi_j} + \frac{1}{C_U+2\lambda}\left[\sum_{j\in U}\tilde{\boldsymbol{I}}_j\left(1-\frac{I_j}{\pi_j}\right)\right]^T\boldsymbol{\Omega}_U^{-1}\frac{\tilde{\boldsymbol{I}}_j}{\pi_j}\right\}y_j \\ &= \sum_{j\in s}w_jy_j. \end{aligned}\tag{2.10}$$

Since the weights are constructed independent of the study variable y, they can be applied to other study variables.

The spline fit in Breidt, Claeskens, and Opsomer (2005) employs a truncated polynomial basis but could be equivalently represented using the B-spline basis presented here. However, it is important to point out that the matrix form of the estimated B-spline coefficients would be

$$\hat{\boldsymbol{\beta}}_s^* = \left(\boldsymbol{X}_s^T \boldsymbol{\Pi}_s^{-1} \boldsymbol{X}_s + \lambda \boldsymbol{D}^T \boldsymbol{D}\right)^{-1} \boldsymbol{X}_s^T \boldsymbol{\Pi}_s^{-1} \boldsymbol{y}_s \tag{2.11}$$

where $\boldsymbol{\Pi}_s$ is a diagonal matrix of the inclusion probabilities for the sample. These estimates are not equal to the proposed B-spline coefficients

$$\hat{\boldsymbol{\beta}}_s = \left(\boldsymbol{X}_U^T \boldsymbol{X}_U + \lambda \boldsymbol{D}^T \boldsymbol{D}\right)^{-1} \boldsymbol{X}_s^T \boldsymbol{\Pi}_s^{-1} \boldsymbol{y}_s. \tag{2.12}$$

The method for constructing the explicit solutions for $\hat{\boldsymbol{\beta}}_s$ relies on the tri-diagonal, banded structure of $\boldsymbol{\Omega}_U$. The matrix $\boldsymbol{\Omega}_s = (C_U + 2\lambda)^{-1} \left(\boldsymbol{X}_s^T \boldsymbol{\Pi}_s^{-1} \boldsymbol{X}_s + \lambda \boldsymbol{D}^T \boldsymbol{D}\right)$ is also tri-diagonal but no longer banded since the elements on the diagonal need not be equal. Therefore, to find (2.11), the K by K matrix $\boldsymbol{\Omega}_s$ must be inverted, a calculation that becomes troublesome as K increases. To avert this issue, we prove the asymptotic results for the estimator based on (2.12) and show the asymptotic equivalence of $N^{-1}\hat{t}_y$ and

$$N^{-1}\hat{t}_y^* = \frac{1}{N} \sum_{j \in s} \frac{y_j - \tilde{\boldsymbol{I}}_j^T \hat{\boldsymbol{\beta}}_s^*}{\pi_j} + \frac{1}{N} \sum_{j \in U} \tilde{\boldsymbol{I}}_j^T \hat{\boldsymbol{\beta}}_s^* \tag{2.13}$$

in Lemma 2.9.

In this thesis, piece-wise constant penalized splines with a first-order difference penalty are considered. The methods of Li and Ruppert (2008) and those discussed above also can be used to construct finite populations coefficients and their corresponding Horvitz-Thompson 'plug-in' estimators for higher order B-splines and higher order difference penalties.

2.2 Main results

2.2.1 Assumptions

To study the asymptotic behavior of the penalized spline regression estimator, we employ the classical survey asymptotic framework in which nested populations, $U_1 \subset U_2 \subset \ldots \subset U_\zeta \subset \ldots$, are subscripted by an increasing sequence $\{\zeta\}$. For each U_ζ, the sample is selected according to the sampling design $p_\zeta(\cdot)$. Let $\{N_\zeta\}$, $\{n_\zeta\}$, and $\{K_\zeta\}$ be sequences of positive integers with $N_\zeta, n_\zeta, K_\zeta \to \infty$, as $\zeta \to \infty$. Henceforth, we suppress ζ for simplicity of notation but will use N as the asymptotic index when necessary. We write the finite population penalized spline coefficient vector as $\boldsymbol{\beta}_N$ and the sample penalized spline coefficient vector as $\hat{\boldsymbol{\beta}}_N$ to emphasis the dependence on N.

Assumptions for the asymptotic design mean squared error and for the design mean squared consistency:

A1. Let NK^{-1} be an integer for all N.

A2. Assume that for all $i = 1, \ldots, K$, $KN^{-1} \sum_{j \in U_N} y_j^2 \tilde{I}_{ij} \leq M$.

A3. For all N, $\min_{j \in U_N} \pi_j = \pi_{N*} > 0$ and $\min_{(i,j) \in U_N} \pi_{ij} = \pi_{N**} > 0$.

A4. There exists $\tau \geq 0$ such that $\max_{j \in U_N} \sum_{l \in U_N : l \neq j} \triangle_{jl}^2 = O(N^{-2\tau})$ and $\left(\pi_{N*}^2 N^{1/2+\tau}\right)^{-1} n = O(1)$.

A5. Assume $0 < \liminf_{N \to \infty} N\pi_{N*}n^{-1}$ and $\limsup_{N \to \infty} N\pi_{N*}n^{-1} < \infty$.

A6. Let $K^2N^2n^{-3} = o(1)$.

A7. Assumptions on the higher order inclusion probabilities: Let $D_{t,N}$ denote the set of all distinct t-tuples $(j_1, j_2, \ldots, j_t) \in U_N$.

(i) (4 distinct elements) Assume

$$\lim_{N \to \infty} N^2 \max_{(j_1,j_2,j_3,j_4) \in D_{4,N}} \left|E_p\left[(I_{j_1} - \pi_{j_1})(I_{j_2} - \pi_{j_2})(I_{j_3} - \pi_{j_3})(I_{j_4} - \pi_{j_4})\right]\right| < \infty.$$

(ii) (3 distinct elements) Assume

$$\limsup_{N \to \infty} N \max_{(j_1,j_2,j_3) \in D_{3,N}} \left|E_p\left[(I_{j_1} - \pi_{j_1})^2(I_{j_2} - \pi_{j_2})(I_{j_3} - \pi_{j_3})\right]\right| < \infty.$$

(iii) (2 distinct elements) Assume

$$\limsup_{N \to \infty} K \max_{(j_1,j_2) \in D_{2,N}} \left|E_p\left[(I_{j_1} - \pi_{j_1})^3(I_{j_2} - \pi_{j_2})\right]\right| < \infty.$$

(iv) Assume

$$\limsup_{N\to\infty} \quad N \max_{(j_1,j_2,j_3,j_4)\in D_{4,N}} |E_p\left[(I_{j_1}I_{j_2}-\pi_{j_1j_2})(I_{j_3}I_{j_4}-\pi_{j_3j_4})\right]| < \infty.$$

Additional assumptions for the design consistency of the variance estimator:

A8. Assume for all N, $N^{-1}\sum_{j\in U_N} y_j^4 < \infty$.

A9. Assume $n^2(\pi_{N**}N^2)^{-1} = O(1)$ and $nN^{\tau}\pi_{N**}^2 \to \infty$ as $N\to\infty$.

A10. There exists $\xi \geq 0$ such that $\max_{j\in U_N}\sum_{l\in U_N: l\neq j} \triangle_{jl}^4 = O(N^{-2\xi})$ and $n^2\pi_{N**}N^{\xi-3/2} \to \infty$ as $N\to\infty$.

Remark 1. Assumption (A1) ensures the x's are placed evenly between knots. In section 2.3.1, we discuss placement of the x's when the assumption is dropped.

Remark 2. Assumption (A2) bounds the second moment in each cell along with the population second moment. As the number of cells increases, it is important for the second moment in each cell to be bounded uniformly.

Remark 3. While ensuring a measurable, probability sampling design for each N, assumption (A3) allows the first and second-order inclusion probabilities to each go to zero as N goes to infinity. This flexibility, for example, allows the sample size to be of order less than or equal to the order of the population size for simple random sampling without replacement. For the relationship between sample size and number of knots, the assumption (A6) requires $K = o(\sqrt{n})$.

Remark 4. Breidt and Opsomer (2008) have shown the first part of assumption (A4) covers non-trivial dependencies in the sampling design by finding τ for both simple random sampling without replacement and single-stage cluster sampling of equally sized clusters where the clusters are sampled with simple random sampling without replacement. This assumption allows for more potential sampling designs than the usual absolute value assumption on $\triangle_{jl}$ which can be found in (A6) of Breidt and Opsomer (2000).

Remark 5. Assumptions (A8) through (A10), which bound higher order moments and place stricter conditions on the design and model, are utilized in the consistency of the variance estimator.

Remark 6. If $n = cN^\gamma$ where $c > 0$, $K = o(N^{3/2\gamma - 1})$, and $2/3 < \gamma \leq 1$, then all the assumptions hold for simple random sampling without replacement.

2.2.2 Asymptotic mean squared error

In this section we show equivalence of the asymptotic mean squared error of the penalized spline regression estimator and the variance of the difference estimator. This equivalence implies that the dominant source of variability is from the sampling mechanism, not the model fit.

Theorem 2.1. *Under assumptions (A1) – (A7),*

$$E_p\left[\frac{\sqrt{n}}{N}(\hat{t}_y - t_y)\right]^2 = \frac{n}{N^2}\sum\sum_{j,l\in U_N}\triangle_{jl}\frac{(y_j - \tilde{\boldsymbol{I}}_j^T\boldsymbol{\beta}_N)}{\pi_j}\frac{(y_l - \tilde{\boldsymbol{I}}_l^T\boldsymbol{\beta}_N)}{\pi_l} + o(1). \qquad (2.14)$$

Proof. Write

$$\begin{aligned}
& E_p\left[\frac{\sqrt{n}}{N}(\hat{t}_y - t_y)\right]^2 \\
= {} & \frac{n}{N^2}E_p\left[\sum_{j\in U_N}(y_j - \tilde{\boldsymbol{I}}_j^T\hat{\boldsymbol{\beta}}_N)\left(\frac{I_j}{\pi_j} - 1\right)\right]^2 \\
= {} & \frac{n}{N^2}E_p\left[\sum_{j\in U_N}(y_j - \tilde{\boldsymbol{I}}_j^T\boldsymbol{\beta}_N)\left(\frac{I_j}{\pi_j} - 1\right) + \sum_{j\in U_N}\tilde{\boldsymbol{I}}_j^T(\boldsymbol{\beta}_N - \hat{\boldsymbol{\beta}}_N)\left(\frac{I_j}{\pi_j} - 1\right)\right]^2 \\
= {} & \frac{n}{N^2}E_p\left[\sum_{j\in U_N}(y_j - \tilde{\boldsymbol{I}}_j^T\boldsymbol{\beta}_N)\left(\frac{I_j}{\pi_j} - 1\right)\right]^2 + \frac{n}{N^2}E_p\left[\sum_{j\in U_N}\tilde{\boldsymbol{I}}_j^T(\boldsymbol{\beta}_N - \hat{\boldsymbol{\beta}}_N)\left(\frac{I_j}{\pi_j} - 1\right)\right]^2 \\
+ {} & \frac{2n}{N^2}E_p\left[\sum_{j\in U_N}(y_j - \tilde{\boldsymbol{I}}_j^T\boldsymbol{\beta}_N)\left(\frac{I_j}{\pi_j} - 1\right)\right]\left[\sum_{l\in U_N}\tilde{\boldsymbol{I}}_l^T(\boldsymbol{\beta}_N - \hat{\boldsymbol{\beta}}_N)\left(\frac{I_l}{\pi_l} - 1\right)\right].
\end{aligned}$$

The first term equals the variance of the difference estimator and is $O(1)$ by Lemma 2.1 while Lemma 2.2 implies the second term is $o(1)$. The last term is $o(1)$ by the Cauchy-Schwarz Inequality. □

2.2.3 Mean squared error consistency

The results of Theorem 2.1 allow us to look at the order of the design mean squared error of the penalized spline regression estimator directly. From these order statements, we can obtain design mean squared consistency of the penalized spline regression estimator.

Theorem 2.2. *Assume (A1) – (A7). Then $N^{-1}\hat{t}_y$ is design mean squared consistent in the sense that*

$$\lim_{N\to\infty} E_p\left[\frac{\hat{t}_y - t_y}{N}\right]^2 = 0$$

and therefore design consistent in the sense that

$$\lim_{N\to\infty} P\left[\left|\frac{\hat{t}_y - t_y}{N}\right| > \eta\right] = 0$$

for all $\eta > 0$.

Proof. Theorem 2.1 implies

$$E_p\left[\frac{1}{N}(\hat{t}_y - t_y)\right]^2 = \frac{1}{N^2}\sum\sum_{j,l\in U_N}\triangle_{jl}\frac{(y_j - \tilde{\boldsymbol{I}}_j^T\boldsymbol{\beta}_N)}{\pi_j}\frac{(y_l - \tilde{\boldsymbol{I}}_l^T\boldsymbol{\beta}_N)}{\pi_l} + o\left(n^{-1}\right).$$

For the leading term

$$\begin{aligned}
&\frac{1}{N^2}\sum\sum_{j,l\in U_N}\triangle_{jl}\frac{(y_j - \tilde{\boldsymbol{I}}_j^T\boldsymbol{\beta}_N)}{\pi_j}\frac{(y_l - \tilde{\boldsymbol{I}}_l^T\boldsymbol{\beta}_N)}{\pi_l}\\
\le\ & \frac{1}{N\pi_{N*}}\sum_{j\in U_N}\frac{(y_j - \tilde{\boldsymbol{I}}_j^T\boldsymbol{\beta}_N)^2}{N} + \frac{1}{N^{1/2+\tau}\pi_{N*}^2}\left\{N^{2\tau}\max_{j\in U_N}\sum_{l\in U:j\neq l}\triangle_{jl}^2\right\}^{1/2}\left\{\sum_{j\in U_N}\frac{(y_j - \tilde{\boldsymbol{I}}_j^T\boldsymbol{\beta}_N)^2}{N}\right\}\\
=\ & o(1)
\end{aligned}$$

by assumptions (A4) and (A5). □

The mean squared error consistency also implies $\hat{t}_y$ is asymptotically design unbiased in the sense that

$$\lim_{N\to\infty} E_p\left[\frac{\hat{t}_y - t_y}{N}\right] = 0.$$

2.2.4 Consistency of variance estimator

With additional assumptions on the sampling design and model, we now prove the standard variance estimator is consistent for the asymptotic mean squared error. In section 2.4 we explore the performance of the variance estimator via simulation for different combinations of sample size, population size, and number of knots.

Theorem 2.3. *Under assumptions (A1)- (A10),*

$$\lim_{N\to\infty} n\ E_p\left|\widehat{var}\left(\hat{t}_y N^{-1}\right) - AMSE\left(\hat{t}_y N^{-1}\right)\right| = 0$$

where

$$\widehat{var}\left(\hat{t}_y N^{-1}\right) = \frac{1}{N^2}\sum\sum_{i,j\in U_N}(y_i - \tilde{\boldsymbol{I}}_i^T\hat{\boldsymbol{\beta}}_N)(y_j - \tilde{\boldsymbol{I}}_j^T\hat{\boldsymbol{\beta}}_N)\frac{\triangle_{ij}}{\pi_i\pi_j}\frac{I_iI_j}{\pi_{ij}} \tag{2.15}$$

and

$$AMSE\left(\hat{t}_y N^{-1}\right) = \frac{1}{N^2}\sum\sum_{i,j\in U_N}(y_i - \tilde{\boldsymbol{I}}_i^T\boldsymbol{\beta}_N)(y_j - \tilde{\boldsymbol{I}}_j^T\boldsymbol{\beta}_N)\frac{\triangle_{ij}}{\pi_i\pi_j}.$$

Proof. Applying the triangle inequality

$$\begin{aligned}
&nE_p\left|\widehat{var}\left(\hat{t}_y N^{-1}\right) - AMSE\left(\hat{t}_y N^{-1}\right)\right| \\
&\leq E_p\left|\frac{n}{N^2}\sum\sum_{i,j\in U_N}(y_i - \tilde{\boldsymbol{I}}_i^T\boldsymbol{\beta}_N)(y_j - \tilde{\boldsymbol{I}}_j^T\boldsymbol{\beta}_N)\frac{\triangle_{ij}}{\pi_i\pi_j}\left(\frac{I_iI_j}{\pi_{ij}} - 1\right)\right| \\
&+ E_p\left|\frac{n}{N^2}\sum\sum_{i,j\in U_N}\left[(y_i - \tilde{\boldsymbol{I}}_i^T\hat{\boldsymbol{\beta}}_N)(y_j - \tilde{\boldsymbol{I}}_j^T\hat{\boldsymbol{\beta}}_N) - (y_i - \tilde{\boldsymbol{I}}_i^T\boldsymbol{\beta}_N)(y_j - \tilde{\boldsymbol{I}}_j^T\boldsymbol{\beta}_N)\right]\frac{\triangle_{ij}}{\pi_i\pi_j}\frac{I_iI_j}{\pi_{ij}}\right| \\
&= o(1)
\end{aligned}$$

by Lemma 2.3 and Lemma 2.4. □

By Markov's Inequality, Theorem 2.3 implies the variance estimator $\widehat{var}\left(\hat{t}_y N^{-1}\right)$ is both asymptotically design unbiased and design consistent for the asymptotic mean squared error $AMSE\left(\hat{t}_y N^{-1}\right)$.

2.3 Extensions

In this section, we consider the penalized spline regression estimator when there are unequal observations between knots. Additionally, we try to improve the estimator found in (2.9) by constructing an estimator based on sample quantiles and by constructing a Hajek plug-in estimator.

2.3.1 Unequal observations between knots

If we relax assumption (A1), we can define $C_N^* := \lfloor NK^{-1} \rfloor$, $C_{N1}^* := \lfloor (N-(K-2)C_N^*)2^{-1} \rfloor$, and $C_{NK}^* = N - (K-2)C_N^* - C_{N1}^*$ and place C_N^* $x_j's$ between the interior knots, C_1^* $x_j's$ in the first cell and C_K^* $x_j's$ in the last cell. The elements $\Omega_{U(1,1)}$ and $\Omega_{U(K,K)}$ of $\mathbf{\Omega}_U$ are possibly unequal. However, the rest of $\mathbf{\Omega}_U$ remains the same and vectors $\boldsymbol{T}_t(\rho)$ can still be found which are orthogonal to all columns of $\mathbf{\Omega}_U$ except the first, last and t-th columns. Therefore, the methods of Li and Ruppert (2008) still hold for finding $\boldsymbol{\beta}_N$. We must, however, distinguish between $\Omega_{U(1,1)}$ and $\Omega_{U(K,K)}$ when finding the explicit forms for $\boldsymbol{\beta}_N$ and $\hat{\boldsymbol{\beta}}_N$.

2.3.2 Estimator with estimated quantiles

Each estimated coefficient in (2.12) contains terms of the form: $KN^{-1}\sum_{j\in U_N} y_j \tilde{I}_{ij} I_j \pi_j^{-1}$ but it is possible for no x_j $j \in s$ to be between κ_{Ni-1} and κ_{Ni}. To ensure that the estimator has no empty cells, we consider estimated coefficients based on the sample derived knots. The matrix form of this estimator is

$$\hat{\boldsymbol{\beta}}_N^{(2)} = \left(\boldsymbol{X}_U^T \boldsymbol{X}_U + \lambda \boldsymbol{D}^T \boldsymbol{D}\right)^{-1} \widehat{\boldsymbol{X}}_s^T \mathbf{\Pi}_s^{-1} \boldsymbol{y}_s \tag{2.16}$$

where $\widehat{\boldsymbol{X}}_s = \left[\hat{\boldsymbol{I}}_j^T\right]_{j\in s}$, $\hat{\boldsymbol{I}}_j = \left(\hat{\boldsymbol{I}}_{1j}, \hat{\boldsymbol{I}}_{2j}, \ldots, \hat{\boldsymbol{I}}_{Kj}\right)^T$, and $\hat{\boldsymbol{I}}_{ij} = I_{\{}\hat{\kappa}_{Ni-1} \leq x_j < \hat{\kappa}_{Ni}\}$ for estimated knots $\{\hat{\kappa}_{Ni}\}_{i=0}^{K}$. To find the estimated knots, let $p_i = iK^{-1}$ for $i = 0, 1, \ldots, K$. Define the first and last estimated knots as the boundaries of x: $\hat{\kappa}_{N0} = 0$ and $\hat{\kappa}_{NK} = 1$. For $i = 1, 2, \ldots, K-1$, let

$$\hat{\kappa}_{Ni} = \inf\{x : \hat{F}_N(x) \geq p_i\} \tag{2.17}$$

where $\hat{F}_N(x) = \hat{N}^{-1}\sum_{j\in U}\pi_j^{-1}I_jI\{x_j \leq x\}$ and $\hat{N} = \sum_{j\in U}\pi_j^{-1}I_j$. The resulting model-assisted penalized spline survey regression estimator is

$$\hat{t}_y^{(2)} = \sum_{j\in s}\frac{y_j - \hat{\boldsymbol{I}}_j^T\hat{\boldsymbol{\beta}}_N^{(2)}}{\pi_j} + \sum_{j\in U_N}\hat{\boldsymbol{I}}_j^T\hat{\boldsymbol{\beta}}_N^{(2)}. \tag{2.18}$$

To obtain design consistency of $\hat{t}_y^{(2)}$ for t_y, we use the uniform convergence of the sample quantiles $\hat{\kappa}_{Ni}$ for the finite population quantiles

$$\kappa_{Ni} = \inf\{x : F_N(x) \geq p_i\}. \tag{2.19}$$

A proof of the uniform convergence of the sample quantiles is found in Lemma 2.6. The uniform convergence of the sample quantiles requires a probability inequality for bounded, survey-weighted quantities and therefore in Lemma 2.5 we prove a survey-weighted version of Hoeffding's Inequality (Hoeffding 1963). A more general case of the survey-weighted Hoeffding's Inequality is found in Corollary 2.1 and applied in Lemma 2.6. The following three assumptions are used for uniform convergence of the sample quantiles and design consistency of $N^{-1}\hat{t}_y^{(2)}$. Assumption (A12) allows us to ignore the dependence between elements in the sample. Hoeffding (1963) shows assumption (A12) holds for simple random sampling. It can easily be shown that (A12) also holds for stratified simple random sampling.

Additional assumptions for the design consistency of the estimator with sample-based quantiles:

A11. Assume the probability sampling design, $p(\cdot)$, is a fixed size design.

A12. For any bounded function g and constant $h > 0$,

$$E_p \exp\left(h \sum_{j\in U_N} \frac{g(x_j)}{\pi_j} I\{j \in s\}\right) \leq E_{p^*} \exp\left(h \sum_{l=1}^{n} \frac{1}{n} \sum_{j\in U_N} \frac{g(x_j)}{p_j} I\{R_l = j\}\right)$$

where p^* is the sampling design corresponding to sampling with replacement. The random variable, R_l, represents the l-th draw from the finite population and each draw is independent. Therefore, $P(R_l = j) = p_j = n^{-1}\pi_j$, where j is the j-th element in the finite population, U_N.

A13. Let the covariate, x, have compact support on $[a, b]$. The finite population distribution function $F_N(x) = N^{-1}\sum_{j\in U_N} I\{x_j \leq x\}$ converges uniformly in x to $F(x)$, $\lim_{N\to\infty} \sup_{x\in[a,b]} |F_N(x) - F(x)| = 0$, and $F(x)$ is continuous and differentiable. Assume the derivative of $F(x)$, denoted by $f(x)$, is positive on $[a, b]$.

A14. For all N, $K = O(N^{1/4})$.

Theorem 2.4. *Under assumptions (A1) – (A6), (A8), (A11) – (A14), $N^{-1}\hat{t}_y^{(2)}$ is design consistent for $N^{-1}t_y$ in the sense that*

$$\lim_{N\to\infty} P\left[\left|\frac{\hat{t}_y^{(2)} - t_y}{N}\right| > \eta\right] = 0$$

for all $\eta > 0$.

Proof. Write

$$\begin{aligned}
\frac{\hat{t}_y^{(2)} - t_y}{N} =& \frac{1}{N} \sum_{j\in U_N} \left(y_j - \hat{\boldsymbol{I}}_j^T \hat{\boldsymbol{\beta}}_N^{(2)}\right)\left(\frac{I\{j \in s\}}{\pi_j} - 1\right) \\
=& \frac{1}{N} \sum_{j\in U_N} \left(y_j - \tilde{\boldsymbol{I}}_j^T \boldsymbol{\beta}_N\right)\left(\frac{I\{j \in s\}}{\pi_j} - 1\right) + \frac{1}{N} \sum_{j\in U_N} \left(\tilde{\boldsymbol{I}}_j - \hat{\boldsymbol{I}}_j\right)^T \left(\frac{I\{j \in s\}}{\pi_j} - 1\right) \boldsymbol{\beta}_N \\
&+ \frac{1}{N} \sum_{j\in U_N} \left(\hat{\boldsymbol{I}}_j - \tilde{\boldsymbol{I}}_j\right)^T \left(\frac{I\{j \in s\}}{\pi_j} - 1\right)\left(\boldsymbol{\beta}_N - \hat{\boldsymbol{\beta}}_N^{(2)}\right) \\
&+ \frac{1}{N} \sum_{j\in U_N} \tilde{\boldsymbol{I}}_j^T \left(\frac{I\{j \in s\}}{\pi_j} - 1\right)\left(\boldsymbol{\beta}_N - \hat{\boldsymbol{\beta}}_N^{(2)}\right) \\
&:= A_{N1} + A_{N2} + A_{N3} + A_{N4}. \qquad (2.20)
\end{aligned}$$

From Theorem 2.2 we have $A_{N1} = o_p(1)$. For the second term, we can divide it into two parts

$$A_{N2} = \frac{1}{N}\sum_{i=1}^{K}\sum_{j\in U_N}\beta_{Ni}\left(\frac{I\{j\in s\}}{\pi_j}-1\right)\tilde{I}_{ij} - \frac{1}{N}\sum_{i=1}^{K}\sum_{j\in U_N}\beta_{Ni}\left(\frac{I\{j\in s\}}{\pi_j}-1\right)\hat{I}_{ij}$$
$$:= A_{N21} + A_{N22}.$$

Both $|A_{N21}|$ and $|A_{N22}|$ are bounded by

$$\begin{aligned}\left|\frac{1}{N}\sum_{i=1}^{K}\sum_{j\in U_N}\beta_{Ni}\left(\frac{I\{j\in s\}}{\pi_j}-1\right)\right| \leq &\left|\sum_{i=1}^{K}\beta_{Ni}\right|\left|\frac{1}{N}\sum_{j\in U_N}\left(\frac{I\{j\in s\}}{\pi_j}-1\right)\right| \\ =&O(K)O_p(n^{-1/2}) \\ =&o_p(1)\end{aligned}$$

by assumption (A6). In Lemma 2.7, it is shown that $A_{N3} = o_p(1)$. For the last term, we can write in a format similar to A_{N3} and then apply Lemma 2.7,

$$\begin{aligned}|A_{N4}| \leq &\max_i |\beta_{Ni} - \hat{\beta}_{Ni}^{(3)}|\frac{1}{N}\sum_{j\in U_N}\left|\frac{I_j}{\pi_j}-1\right|\sum_{i=1}^{K}\tilde{I}_{ij} \\ = &\max_i |\beta_{Ni} - \hat{\beta}_{Ni}^{(3)}|\frac{1}{N}\sum_{j\in U_N}\left|\frac{I_j}{\pi_j}-1\right| \\ = &o_p(1).\end{aligned}$$

□

2.3.3 Estimator with estimated cells totals

The estimated coefficient vectors essentially boil down to linear combination of the cell means, where the cells are based on the finite population derived knots. However, since the sample is not necessarily divided evenly among the cells, a more accurate cell mean would take the form of the Hajek estimator which contains a estimate of the cell total based on the sample. Therefore, another possible estimator for the population coefficients would be a Hajek plug-in estimator (Hájek 1971). For $1 < t < K$, the estimator for the population coefficient is

$$\hat{\beta}_{Nt}^{(3)} =$$

$$(1+2\rho\eta)^{-1}\left\{\sum_{i=1}^{K}\rho^{|t-i|}(\hat{N}_i+2\lambda)^{-1}\sum_{j\in U}y_j\tilde{I}_{ij}-\left[\rho^{t-2}\left(\rho\theta+\eta\right)\hat{\beta}_{N1}^{(3)}+\rho^{K-t-1}\left(\rho\theta+\eta\right)\hat{\beta}_{NK}^{(3)}\right]\right\}$$

where $\hat{\beta}_{N1}^{(3)}$ and $\hat{\beta}_{NK}^{(3)}$ are the Hajek plug-in estimators for the first and last cells, respectively and $\hat{N}_i = \sum_{j\in s}\pi_j^{-1}\tilde{I}_{ij}$ is the estimated total in the i-th cell. The estimated coefficients $\hat{\boldsymbol{\beta}}_N^{(3)}$ are similar to the estimator in (2.11) since $\boldsymbol{X}_s^T\boldsymbol{\Pi}^{-1}\boldsymbol{X}_s$ is a diagonal matrix of estimated cell totals $\hat{N}_i$. The model-assisted survey regression estimator with Hajek plug-in estimators is

$$t_y^{(3)} = \sum_{j\in s}\frac{y_j-\tilde{\boldsymbol{I}}_j^T\hat{\boldsymbol{\beta}}_N^{(3)}}{\pi_j}+\sum_{j\in U}\tilde{\boldsymbol{I}}_j^T\hat{\boldsymbol{\beta}}_N^{(3)}. \tag{2.21}$$

The asymptotic properties of this estimator are not derived here but in section 2.4 we look at the relationship between (2.21) and $\hat{t}_y^*$ via simulation.

2.4 Simulation

We want to investigate the performance of (2.15) as an estimator of the true variance of the penalized spline regression estimator. Since $\hat{t}_y^*$ is the estimator used in practice, we use the estimated coefficients $\hat{\boldsymbol{\beta}}_N^*$ in the variance estimator. We also consider the performance of the alternate variance estimator presented in Särndal, Swensson, and Wretman (1989) where π_j^{-1} in (2.15) is replaced by the weight, w_j, found in (2.10). Because we approximated the estimator $\hat{t}_y^*$ with the proposed estimators, (2.9), (2.18), and (2.21), it is important to assess the adequacy of these approximations.

The survey design is stratified random sampling with three strata and simple random sampling without replacement in each strata. For the superpopulation model found in (1.2), we fit the mean functions of Breidt and Opsomer (2000):

piece-wise constant:	$f_o(x) = 2I_{\{0\le x\le 0.25\}} + 10I_{\{0.25<x\le 0.6\}} + 5I_{\{0.6<x\le 1\}}$,
linear:	$f_1(x) = 1 + 2(x - 0.5)$,
quadratic:	$f_2(x) = 1 + 2(x - 0.5)^2$,
bump:	$f_3(x) = 1 + 2(x - 0.5) + \exp(-200(x - 0.5)^2)$,
jump:	$f_4(x) = \{1 + 2(x - 0.5)I_{\{x\le 0.65\}}\} + 0.65I_{\{x>0.65\}}$,
cdf:	$f_5(x) = \Phi\left(\frac{1.5-2x}{\sigma}\right)$ where Φ is the standard normal cdf,
exponential:	$f_6(x) = \exp(-8x)$,
cycle1:	$f_7(x) = 2 + \sin(2\pi x)$,
cycle4:	$f_8(x) = 2 + \sin(8\pi x)$

where $x \in [0, 1]$. For stratum one, $x_j \sim \text{Uniform}(0, 0.25)$, for stratum two, $x_j \sim \text{Uniform}(0.25, 0.6)$, and for stratum three, $x_j \sim \text{Uniform}(0.6, 1)$ with stratum population sizes $\lfloor 0.2N\rfloor$, $\lfloor 0.35N\rfloor$, and $N - \lfloor 0.2N\rfloor - \lfloor 0.35N\rfloor$ respectively. We collected equally sized samples from each stratum. The characteristics of interest y_{ij}, are generated by (1.2) with $\epsilon_j \sim N(0, 0.4^2)$ for each mean function $f_i(\cdot)$ except y_{5j}, which are binary realizations of the indicator $y_{5j} = I_{\{y_{1k}\le 1.5\}}$. Since the fitted model consists of linear combinations of piece-wise constant splines, we have various degrees of model misspecification. The smoothing parameter λ is chosen such that the finite population coefficient vector $\boldsymbol{\beta}_N$ has five degrees of freedom. Therefore, for each sample, the estimated coefficients have approximately five degrees of freedom.

We explore the variance estimator over different combinations of n, N, and K while ensuring NK^{-1} is an integer. Of particular interest is the performance of the variance estimator for 'small', 'medium', and 'large' sample sizes. Because the asymptotic results derive what happens when the number of knots goes to infinity as the population and sample size go to infinity, we focus on what happens when n, N, K each grow at rates similar to those discussed in section 2.2.1. For each combination of n, N, and K considered, we generate a population of size N and then sample 10,000 times from the fixed finite population to construct the estimators for each t_{y_i} where $i = 0, \ldots 8$. Therefore, we are able to compute the empirical design bias, empirical design variance, and empirical design mean squared error across the 10,000 samples from the fixed finite population.

To assess the performance of the variance estimator and the alternate variance estimator, denoted by $v\tilde{a}r_p(\hat{t}_y)$, we compute the percent relative design bias of the variance estimator for the design variance

$$\frac{E_p[\widehat{var}_p(\hat{t}_{y_i})] - \mathrm{Var}_p\left(\hat{t}_{y_i}\right)}{\mathrm{Var}_p\left(\hat{t}_{y_i}\right)} \times 100\%$$

for 'small', 'medium', and 'large' sample sizes. In Table 2.1, we consider three cases: $n = 40, N = 600, K = 6$ and $n = 100, N = 2000, K = 8$ and $n = 200, K = 10, N = 5000$. For the smaller sample size, the negative bias is rather significant for both variance estimators though the alternate variance estimator performs slightly better. However, as the sample size increases the negative bias does decrease. Though both variance estimators exhibit negative bias, as we see in Table 2.2, the confidence interval coverage is only slightly too narrow with average rates around 91.5% for the small sample size and around 94.5% for the large sample size.

Table 2.1: Percent relative bias of variance estimator and alternate variance estimator when estimating the empirical variance

	Relative bias of variance estimator			Relative bias of alternate variance estimator		
Mean functions	$n = 40$	$n = 100$	$n = 200$	$n = 40$	$n = 100$	$n = 200$
piece-wise constant	-23.82	-7.02	-5.57	-11.52	-3.17	2.02
linear	-18.78	-7.80	-3.95	-18.47	-6.86	-3.48
quadratic	-19.40	-5.96	-4.13	-19.17	-5.14	-3.69
bump	-20.48	-7.66	-4.32	-19.43	-5.89	-3.51
jump	-15.62	-7.94	-2.97	-15.14	-6.84	-2.46
cdf	-15.93	-8.26	-5.08	-14.52	-7.65	-2.73
exponential	-16.48	-7.76	-1.52	-16.03	-6.88	-1.04
cycle1	-17.25	-5.48	-5.27	-16.43	-4.08	-4.55
cycle4	-18.18	-6.46	-4.67	-15.69	-3.38	-2.79

To assess the effect of model misspecification on the estimators, we compute the percent relative design bias

$$\frac{E_p[\hat{t}_{y_i}] - t_{y_i}}{t_{y_i}} \times 100\%$$

Table 2.2: 95% confidence interval coverage for variance estimator and alternate variance estimator when sample size is small or large

	CI Coverage			CI Coverage using alternate variance estimator		
Mean functions	$n = 40$	$n = 100$	$n = 200$	$n = 40$	$n = 100$	$n = 200$
piece-wise constant	90.51	91.98	93.77	93.22	92.79	94.81
linear	91.19	93.40	94.25	91.16	93.40	94.38
quadratic	90.92	93.92	94.25	90.91	93.98	94.34
bump	90.36	93.70	94.30	90.27	93.95	94.42
jump	91.37	93.47	94.11	91.39	93.77	94.20
cdf	90.35	90.09	94.46	90.64	91.14	95.09
exponential	91.64	93.51	94.70	91.73	93.57	94.75
cycle1	91.53	94.03	94.17	91.79	94.22	94.26
cycle4	90.99	93.82	94.31	91.20	94.24	94.55

for $i = 0, \ldots, 9$. For different combinations of n, N, and K, even small n and averaging across mean functions, the percent relative design bias is less than 2% for all estimators except $\hat{t}_y$. The bias of estimator $\hat{t}_y$, averaging across mean functions, does decrease as sample size decreases with values of -7.30%, -3.08%, and -1.39% for the 'small', 'medium', and 'large' sample sizes respectively. To assess the difference between the estimator used in practice, $\hat{t}_y^*$, and the approximations presented in this thesis, we computed the pairwise correlations. The average correlations across mean functions are given in Table 2.3. The approximations given in section 2.3 are more closely correlated with $\hat{t}_y^*$ but the correlation with $\hat{t}_y$ increases as the sample size increases.

Table 2.3: Average correlation between $\hat{t}_y^*$ and its approximations across the mean functions

	Correlations by sample size		
Approximate estimators	$n = 40$	$n = 100$	$n = 200$
$\hat{t}_y$	0.736	0.865	0.924
$\hat{t}_y^{(2)}$	0.925	0.961	0.979
$\hat{t}_y^{(3)}$	0.952	0.956	0.963

2.5 Appendix A

Lemma 2.1. *Under assumptions (A1) – (A5),*

$$\frac{n}{N^2}E_p\left[\sum_{j\in U_N}(y_j-\tilde{\boldsymbol{I}}_j^T\boldsymbol{\beta}_N)\left(\frac{I_j}{\pi_j}-1\right)\right]^2=O(1).$$

Proof. Following the method of Breidt and Opsomer (2008)

$$\begin{aligned}
&\frac{n}{N^2}E_p\left[\sum_{j\in U_N}(y_j-\tilde{\boldsymbol{I}}_j^T\boldsymbol{\beta}_N)\left(\frac{I_j}{\pi_j}-1\right)\right]^2\\
=&\frac{n}{N^2}\sum_{j\in U_N}(1-\pi_j)\frac{(y_j-\tilde{\boldsymbol{I}}_j^T\boldsymbol{\beta}_N)^2}{\pi_j}+\frac{n}{N^2}\sum\sum_{j\neq l\in U_N}\triangle_{jl}\frac{(y_j-\tilde{\boldsymbol{I}}_j^T\boldsymbol{\beta}_N)}{\pi_j}\frac{(y_l-\tilde{\boldsymbol{I}}_l^T\boldsymbol{\beta}_N)}{\pi_l}\\
\leq&\frac{n}{N\pi_{N*}}\sum_{j\in U_N}\frac{(y_j-\tilde{\boldsymbol{I}}_j^T\boldsymbol{\beta}_N)^2}{N}+\frac{n}{N^2\pi_{N*}^2}\left\{\sum\sum_{j\neq l\in U_N}\triangle_{jl}^2\right\}^{1/2}\left\{\sum\sum_{j\neq l\in U_N}(y_j-\tilde{\boldsymbol{I}}_j^T\boldsymbol{\beta}_N)^2(y_l-\tilde{\boldsymbol{I}}_l^T\boldsymbol{\beta}_N)^2\right\}^{1/2}\\
\leq&\frac{n}{N\pi_{N*}}\sum_{j\in U_N}\frac{(y_j-\tilde{\boldsymbol{I}}_j^T\boldsymbol{\beta}_N)^2}{N}+\frac{n}{N^2\pi_{N*}^2}\left\{N\max_{j\in U_N}\sum_{l\in U_N:j\neq l}\triangle_{jl}^2\right\}^{1/2}\left\{\sum_{j\in U_N}(y_j-\tilde{\boldsymbol{I}}_j^T\boldsymbol{\beta}_N)^2\right\}\\
=&\frac{n}{N\pi_{N*}}\sum_{j\in U_N}\frac{(y_j-\tilde{\boldsymbol{I}}_j^T\boldsymbol{\beta}_N)^2}{N}+\frac{n}{N^{1/2+\tau}\pi_{N*}^2}\left\{N^{2\tau}\max_{j\in U_N}\sum_{l\in U_N:j\neq l}\triangle_{jl}^2\right\}^{1/2}\left\{\sum_{j\in U_N}\frac{(y_j-\tilde{\boldsymbol{I}}_j^T\boldsymbol{\beta}_N)^2}{N}\right\}.
\end{aligned}\tag{2.22}$$

By assumptions (A3) – (A5), (2.22) is bounded as long as $N^{-1}\sum_{j\in U_N}(y_j-\tilde{\boldsymbol{I}}_j^T\boldsymbol{\beta}_N)^2=O(1)$. Assumption (A2) bounds the second moment of y. We must still bound $\max_{i:i=1,2,\ldots,K}\beta_{Ni}^2$. Each coefficient has a similar form, therefore we can consider the first coefficient, β_{N1}^2.

Since $\theta = 1 + o(1)$, $\eta = o(1)$ and $0 < \rho < 1$, the square of (2.5) can be written as

$$\begin{aligned}
\beta_{N1}^2 = &\left[(1+o(1))\sum_{i=1}^{K}\rho^{i-1}\frac{K}{N}\sum_{j\in U_N} y_j\tilde{I}_{ij} - o(1)\sum_{i=1}^{K}\rho^{K-i}\frac{K}{N}\sum_{j\in U_N} y_j\tilde{I}_{ij}\right]^2 \\
=&(1+o(1))\left[\sum_{i=1}^{K}\rho^{i-1}\frac{K}{N}\sum_{j\in U_N} y_j\tilde{I}_{ij}\right]^2 \\
&- o(1)\left\{\left[\sum_{i=1}^{K}\rho^{i-1}\frac{K}{N}\sum_{j\in U_N} y_j\tilde{I}_{ij}\right]\left[\sum_{i=1}^{K}\rho^{K-i}\frac{K}{N}\sum_{j\in U_N} y_j\tilde{I}_{ij}\right] + \left[\sum_{i=1}^{K}\rho^{K-i}\frac{K}{N}\sum_{j\in U_N} y_j\tilde{I}_{ij}\right]^2\right\} \\
\leq&(1+o(1))M^2\left(\frac{1-\rho^K}{1-\rho}\right)^2 - o(1)2M^2\left(\frac{1-\rho^K}{1-\rho}\right)^2 \qquad (2.23) \\
=&O(1)
\end{aligned}$$

taking advantage of the uniform bound in (A2). The other squared coefficients can be bounded uniformly in i by similar methods. Therefore, (2.22) is $O(1)$.

□

Lemma 2.2. *Under assumptions (A1) -(A7),*

$$\frac{n}{N^2}E\left[\sum_{j\in U_N}\tilde{\boldsymbol{I}}_j^T(\boldsymbol{\beta}_N - \hat{\boldsymbol{\beta}}_N)\left(\frac{I_j}{\pi_j}-1\right)\right]^2 = O(1).$$

Proof. Since $\tilde{\boldsymbol{I}}_j^T$ is a vector of indicators, which specify the placement of the j-th observation, write

$$\begin{aligned}
&\frac{n}{N^2}E\left[\sum\sum_{j_1,j_2\in U_N}(\boldsymbol{\beta}_N - \hat{\boldsymbol{\beta}}_N)^T\tilde{\boldsymbol{I}}_{j_1}\tilde{\boldsymbol{I}}_{j_2}^T(\boldsymbol{\beta}_N - \hat{\boldsymbol{\beta}}_N)\left(\frac{I_{j_1}}{\pi_{j_1}}-1\right)\left(\frac{I_{j_2}}{\pi_{j_2}}-1\right)\right] \\
&= \frac{n}{N^2}E\left[\sum_{t_1=1}^{K}\sum_{t_2=1}^{K}(\beta_{Nt_1} - \hat{\beta}_{Nt_1})(\beta_{Nt_2} - \hat{\beta}_{Nt_2})\sum\sum_{j_1,j_2\in U_N}\tilde{I}_{t_1j_1}\tilde{I}_{t_2j_2}\left(\frac{I_{j_1}}{\pi_{j_1}}-1\right)\left(\frac{I_{j_2}}{\pi_{j_2}}-1\right)\right].
\end{aligned} \qquad (2.24)$$

The difference between each coefficient and the corresponding estimator is of the form:

$$\beta_{N1} - \hat{\beta}_{N1} \quad = \quad \sum_{i=1}^{K}\rho^{i-1}D_i + R_{N1} \qquad (2.25)$$

$$\beta_{NK} - \hat{\beta}_{NK} = \sum_{i=1}^{K} \rho^{K-i} D_i + R_{NK} \tag{2.26}$$

$$\beta_{Nt} - \hat{\beta}_{Nt} = \sum_{i=1}^{K} \rho^{|t-i|} D_i - \rho^{t-1}\left(\beta_{N1} - \hat{\beta}_{N1}\right) - \rho^{K-t}\left(\beta_{NK} - \hat{\beta}_{NK}\right) + R_{Nt} \tag{2.27}$$

where $D_i = KN^{-1}\sum_{j\in U_N} y_j \tilde{I}_{ij}(1 - I_j\pi_j^{-1})$ and R_{N1}, R_{NK}, and R_{Nt} are lower order terms. Substituting (2.25), (2.26), and (2.27) into (2.24) while excluding the lower order terms, we find

$$\begin{aligned}
(2.24) \quad = \quad & \frac{n}{N^2}E\Bigg[c_{K1}\sum_{i_1=1}^{K}\sum_{i_2=1}^{K}\rho^{i_1-1}\rho^{i_2-1}D_{i_1}D_{i_2}\sum\sum_{j_1,j_2\in U_N}\tilde{I}_{1j_1}\tilde{I}_{1j_2}\left(\frac{I_{j_1}}{\pi_{j_1}}-1\right)\left(\frac{I_{j_2}}{\pi_{j_2}}-1\right)+ \\
+ \quad & c_{K2}\sum_{i_1=1}^{K}\sum_{i_2=1}^{K}\rho^{K-i_1}\rho^{K-i_2}D_{i_1}D_{i_2}\sum\sum_{j_1,j_2\in U_N}\tilde{I}_{Kj_1}\tilde{I}_{Kj_2}\left(\frac{I_{j_1}}{\pi_{j_1}}-1\right)\left(\frac{I_{j_2}}{\pi_{j_2}}-1\right) \\
+ \quad & c_{K3}\sum_{i_1=1}^{K}\sum_{i_2=1}^{K}\rho^{i_1-1}\rho^{K-i_2}D_{i_1}D_{i_2}\sum\sum_{j_1,j_2\in U_N}\tilde{I}_{1j_1}\tilde{I}_{Kj_2}\left(\frac{I_{j_1}}{\pi_{j_1}}-1\right)\left(\frac{I_{j_2}}{\pi_{j_2}}-1\right) \\
+ \quad & c_{K4}\sum_{t=2}^{K-1}\sum_{i_1=1}^{K}\sum_{i_2=1}^{K}\rho^{i_1-1}\rho^{|t-i_2|}D_{i_1}D_{i_2}\sum\sum_{j_1,j_2\in U_N}\tilde{I}_{1j_1}\tilde{I}_{tj_2}\left(\frac{I_{j_1}}{\pi_{j_1}}-1\right)\left(\frac{I_{j_2}}{\pi_{j_2}}-1\right) \\
+ \quad & c_{K5}\sum_{t=2}^{K-1}\sum_{i_1=1}^{K}\sum_{i_2=1}^{K}\rho^{K-i_1}\rho^{|t-i_2|}D_{i_1}D_{i_2}\sum\sum_{j_1,j_2\in U_N}\tilde{I}_{Kj_1}\tilde{I}_{tj_2}\left(\frac{I_{j_1}}{\pi_{j_1}}-1\right)\left(\frac{I_{j_2}}{\pi_{j_2}}-1\right) \\
+ \quad & c_{K6}\sum_{t_1=2}^{K-1}\sum_{t_2=2}^{K-1}\sum_{i_1=1}^{K}\sum_{i_2=1}^{K}\rho^{|t_1-i_1|}\rho^{|t_2-i_2|}D_{i_1}D_{i_2}\sum\sum_{j_1,j_2\in U_N}\tilde{I}_{t_1j_1}\tilde{I}_{t_2j_2}\left(\frac{I_{j_1}}{\pi_{j_1}}-1\right)\left(\frac{I_{j_2}}{\pi_{j_2}}-1\right)\Bigg] \\
:= \quad & \frac{n}{N^2}E[a_N] + \frac{n}{N^2}E[b_N] + \frac{n}{N^2}E[c_N] + \frac{n}{N^2}E[d_N] + \frac{n}{N^2}E[e_N] + \frac{n}{N^2}E[f_N]
\end{aligned}$$

where $c_{Km} \leq 4$ for $m = 1, 2, \ldots, 6$. Looking more closely at the $nN^{-2}E[f_N]$, which has the largest order of terms, and plugging in D_{i_1} and D_{i_2}, we see

$$\begin{aligned}
\frac{n}{N^2}E[f_N] = & \frac{n}{N^2}E\Bigg[c_{K6}\sum_{t_1=2}^{K-1}\sum_{t_2=2}^{K-1}\sum_{i_1=1}^{K}\sum_{i_2=1}^{K}\rho^{|t_1-i_1|}\rho^{|t_2-i_2|} \\
& \times\frac{K^2}{N^2}\sum\sum\sum\sum_{j_1,j_2,j_3,j_4\in U_N}\frac{y_{j_3}y_{j_4}\tilde{I}_{t_1j_1}\tilde{I}_{t_2j_2}\tilde{I}_{i_1j_3}\tilde{I}_{i_2j_4}}{\pi_{j_1}\pi_{j_2}\pi_{j_3}\pi_{j_4}}\left(I_{j_1}-1\right)\left(I_{j_2}-1\right)\left(I_{j_3}-1\right)\left(I_{j_4}-1\right)\Bigg].
\end{aligned} \tag{2.28}$$

Let $A_{t,N}$ denote the set of all distinct t-tuples from the set $\{1, 2, \ldots, K\}$. Ignoring the second group of summations and their dependence on t_1, t_2, i_1 and i_2, the order of the first part can be bounded

$$\sum_{t_1=2}^{K-1}\sum_{t_2=2}^{K-1}\sum_{i_1=1}^{K}\sum_{i_2=1}^{K}\rho^{|t_1-i_1|}\rho^{|t_2-i_2|} = O(K^2) \tag{2.29}$$

in three cases:

1. $t_1, t_2, i_1, i_2 \in A_{4,N}$.

2. $t_1 = i_1$ and $t_1, t_2, i_2 \in A_{3,N}$ (or $t_2 = i_2$ and $t_1, t_2, i_1 \in A_{3,N}$).

3. $t_1 = i_1$, $t_2 = i_2$ and $t_1, t_2 \in A_{2,N}$.

The order of the first case is found by solving two geometric series,

$$\begin{aligned}
\sum_{\substack{t_1=2 \\ t_1,t_2,i_1,i_2\in A_{4,N}}}^{K-1}\sum_{t_2=2}^{K-1}\sum_{i_1=1}^{K}\sum_{i_2=1}^{K}\rho^{|t_1-i_1|}\rho^{|t_2-i_2|} &= \left[\sum_{\substack{t=2 \\ t\neq i}}^{K-1}\sum_{i=1}^{K}\rho^{|t-i|}\right]^2 \\
&= \left[\sum_{t=2}^{K-1}\rho\left(\sum_{i=0}^{t-2}\rho^i + \sum_{i=0}^{K-t-1}\rho^i\right)\right]^2 \\
&= \rho^2\left[\sum_{t=2}^{K-1}\left(\frac{1-\rho^{t-1}}{1-\rho} + \frac{1-\rho^{K-t}}{1-\rho}\right)\right]^2 \\
&= \frac{\rho^2}{(1-\rho)^2}\left[\sum_{t=2}^{K-1} 2-\rho^{t-1}-\rho^{K-t}\right]^2 \\
&= \frac{\rho^2}{(1-\rho)^2}\left[2(K-2) - 2\rho\sum_{t=0}^{K-3}\rho^t\right]^2 \\
&= \frac{\rho^2}{(1-\rho)^2}\left[2(K-2) - \frac{2\rho(1-\rho^{K-2})}{1-\rho}\right]^2 \\
&= O(K^2).
\end{aligned}$$

The order of the second case is found similarly and the third is trivial. For other combinations of t_1, t_2, i_1 and i_2, (2.29) is $o(K^2)$, therefore, we shall focus on these three cases, which have the highest order. In the first case, the four distinct cells imply that

$(j_1, j_2, j_3, j_4) \in D_{4,N}$,

$$
\begin{aligned}
& \frac{n}{N^2} E\Bigg[c_{K6} \sum_{\substack{t_1=2 \\ (t_1,t_2,i_1,i_2)\in A_{4,N}}}^{K-1} \sum_{t_2=2}^{K-1} \sum_{i_1=1}^{K} \sum_{i_2=1}^{K} \rho^{|t_1-i_1|} \rho^{|t_2-i_2|} \frac{K^2}{N^2} \sum\sum\sum\sum_{(j_1,j_2,j_3,j_4)\in D_{4,N}} \frac{y_{j_3} y_{j_4} \tilde{I}_{t_1 j_1} \tilde{I}_{t_2 j_2} \tilde{I}_{i_1 j_3} \tilde{I}_{i_2 j_4}}{\pi_{j_1}\pi_{j_2}\pi_{j_3}\pi_{j_4}} \\
& \times \left(I_{j_1}-1\right)\left(I_{j_2}-1\right)\left(I_{j_3}-1\right)\left(I_{j_4}-1\right)\Bigg] \\
\leq\; & \frac{n}{N^2} c_{K6} \sum_{\substack{t_1=2 \\ (t_1,t_2,i_1,i_2\in A_{4,N})}}^{K-1} \sum_{t_2=2}^{K-1} \sum_{i_1=1}^{K} \sum_{i_2=1}^{K} \rho^{|t_1-i_1|} \rho^{|t_2-i_2|} \frac{K^2}{N^2} \sum\sum\sum\sum_{(j_1,j_2,j_3,j_4)\in D_{4,N}} \frac{|y_{j_3}||y_{j_4}| \tilde{I}_{t_1 j_1} \tilde{I}_{t_2 j_2} \tilde{I}_{i_1 j_3} \tilde{I}_{i_2 j_4}}{\pi_{j_1}\pi_{j_2}\pi_{j_3}\pi_{j_4}} \\
& \times \left|E\left[\left(I_{j_1}-1\right)\left(I_{j_2}-1\right)\left(I_{j_3}-1\right)\left(I_{j_4}-1\right)\right]\right| \\
\leq\; & \frac{n}{N^2 \pi_{N*}^4} c_{K6} \sum_{\substack{t_1=2 \\ (t_1,t_2,i_1,i_2)\in A_{4,N}}}^{K-1} \sum_{t_2=2}^{K-1} \sum_{i_1=1}^{K} \sum_{i_2=1}^{K} \rho^{|t_1-i_1|} \rho^{|t_2-i_2|} \frac{K}{N} \sum_{j_3} |y_{j_3}| \tilde{I}_{i_1 j_3} \frac{K}{N} \sum_{j_3} |y_{j_4}| \tilde{I}_{i_1 j_4} \sum_{j_1} \tilde{I}_{t_1 j_1} \sum_{j_2} \tilde{I}_{t_2 j_2} \\
& \times \max_{(j_1,j_2,j_3,j_4)\in D_{4,N}} \left|E\left[\left(I_{j_1}-1\right)\left(I_{j_2}-1\right)\left(I_{j_3}-1\right)\left(I_{j_4}-1\right)\right]\right| \\
\leq\; & \frac{n}{N^2 \pi_{N*}^4} c_{K6} \sum_{\substack{t_1=2 \\ (t_1,t_2,i_1,i_2)\in A_{4,N}}}^{K-1} \sum_{t_2=2}^{K-1} \sum_{i_1=1}^{K} \sum_{i_2=1}^{K} \rho^{|t_1-i_1|} \rho^{|t_2-i_2|} M^2 \frac{N^2}{K^2} \\
& \times \max_{(j_1,j_2,j_3,j_4)\in D_{4,N}} \left|E\left[\left(I_{j_1}-1\right)\left(I_{j_2}-1\right)\left(I_{j_3}-1\right)\left(I_{j_4}-1\right)\right]\right| \\
\leq\; & \frac{n}{N^2 \pi_{N*}^4} c_{K6}^* \qquad (2.30)
\end{aligned}
$$

where $c_{K6}^* = O(1)$ by (A7i). Assumption (A6) implies (2.30) goes to zero as $N \to \infty$. For the second case, where we have three distinct cells, either all the elements are distinct or only three elements are distinct (e.g. $j_1 = j_3$). If all elements are distinct, the computations are similar to above. If only three are distinct, assumption (A7ii) ensures the term goes to zero. For case three, where we have two distinct cells, we could have two, three or four distinct elements. For the case where only two elements are distinct, without loss of generality assume $j_1 = j_3$ and $j_2 = j_4$ where $j_1 \neq j_2$. No assumptions on the higher order inclusion probabilities are necessary since

$$\begin{aligned}
& \frac{n}{N^2}\left[c_{K6}\sum_{\substack{t_1=2 \\ (t_1,t_2)\in A_{2,N}}}^{K-1}\sum_{t_2=2}^{K-1}\frac{K^2}{N^2}\sum\sum_{j_1\neq j_2\in U_N}\frac{y_{j_1}y_{j_2}\tilde{I}_{t_1j_1}\tilde{I}_{t_2j_2}}{\pi_{j_1}^2\pi_{j_2}^2}E_p\left(I_{j_1}-1\right)^2\left(I_{j_2}-1\right)^2\right] \\
\leq \quad & \frac{n}{N^2\pi_{N*}^4}(K-2)^2M^2 4 \\
= \quad & O\left(\frac{nK^2}{N^2\pi_{N*}^4}\right) \\
= o(1) & \qquad (2.31)
\end{aligned}$$

by (A6). Following similar arguments, $nN^{-2}E[a_N], nN^{-2}E[b_N], nN^{-2}E[c_N]$, $nN^{-2}E[d_N]$, and $nN^{-2}E[e_N]$ each converge to zero.

$\square$

Lemma 2.3. *Under assumptions (A1) – (A10),*

$$\lim_{N\to\infty}E_p\left|\frac{n}{N^2}\sum\sum_{i,j\in U_N}(y_i-\tilde{\boldsymbol{I}}_i^T\boldsymbol{\beta}_N)(y_j-\tilde{\boldsymbol{I}}_j^T\boldsymbol{\beta}_N)\frac{\triangle_{ij}}{\pi_i\pi_j}\left(\frac{I_iI_j}{\pi_{ij}}-1\right)\right|=0.$$

Proof. Applying the Cauchy-Schwarz Inequality

$$\begin{aligned}
& E_p\left|\frac{n}{N^2}\sum\sum_{i,j\in U_N}(y_i-\tilde{\boldsymbol{I}}_i^T\boldsymbol{\beta}_N)(y_j-\tilde{\boldsymbol{I}}_j^T\boldsymbol{\beta}_N)\frac{\triangle_{ij}}{\pi_i\pi_j}\left(\frac{I_iI_j}{\pi_{ij}}-1\right)\right| \\
& \leq\left\{E_p\left[\frac{n}{N^2}\sum\sum_{i,j\in U_N}(y_i-\tilde{\boldsymbol{I}}_i^T\boldsymbol{\beta}_N)(y_j-\tilde{\boldsymbol{I}}_j^T\boldsymbol{\beta}_N)\frac{\triangle_{ij}}{\pi_i\pi_j}\left(\frac{I_iI_j}{\pi_{ij}}-1\right)\right]^2\right\}^{1/2} \\
& =\left\{E_p\left[\frac{n}{N^2}\sum_{i\in U_N}(y_i-\tilde{\boldsymbol{I}}_i^T\boldsymbol{\beta}_N)^2\frac{(1-\pi_i)}{\pi_i^2}\left(I_i-\pi_i\right)\right.\right. \\
& \left.\left.+\ \frac{n}{N^2}\sum\sum_{i\neq j\in U_N}(y_i-\tilde{\boldsymbol{I}}_i^T\boldsymbol{\beta}_N)(y_j-\tilde{\boldsymbol{I}}_j^T\boldsymbol{\beta}_N)\frac{\triangle_{ij}}{\pi_i\pi_j}\left(\frac{I_iI_j}{\pi_{ij}}-1\right)\right]^2\right\}^{1/2}
\end{aligned}$$

$$= \Bigg\{ E_p \frac{n^2}{N^4} \sum\sum_{i,k \in U_N} (y_i - \tilde{\boldsymbol{I}}_i^T \boldsymbol{\beta}_N)^2 (y_k - \boldsymbol{x}_k^T \boldsymbol{\beta}_N)^2 \frac{(1-\pi_i)(1-\pi_k)}{\pi_i^2 \pi_k^2} (I_i - \pi_i)(I_k - \pi_k)$$

$$+ E_p \frac{2n^2}{N^4} \sum\sum\sum_{i \neq j, k \in U_N} (y_i - \tilde{\boldsymbol{I}}_i^T \boldsymbol{\beta}_N)(y_j - \tilde{\boldsymbol{I}}_j^T \boldsymbol{\beta}_N)(y_k - \boldsymbol{x}_k^T \boldsymbol{\beta}_N)^2 \frac{\triangle_{ij}(1-\pi_k)}{\pi_i \pi_j \pi_k^2 \pi_{ij}} (I_{ij} - \pi_{ij})(I_k - \pi_k)$$

$$+ E_p \frac{n^2}{N^4} \sum\sum\sum\sum_{i \neq j, k \neq l \in U_N} (y_i - \tilde{\boldsymbol{I}}_i^T \boldsymbol{\beta}_N)(y_j - \tilde{\boldsymbol{I}}_j^T \boldsymbol{\beta}_N)(y_k - \boldsymbol{x}_k^T \boldsymbol{\beta}_N)(y_l - \tilde{\boldsymbol{I}}_l^T \boldsymbol{\beta}_N) \frac{\triangle_{ij}}{\pi_i \pi_j \pi_{ij}} \frac{\triangle_{kl}}{\pi_k \pi_l \pi_{kl}}$$

$$\times (I_i I_j - \pi_{ij})(I_k I_l - \pi_{kl}) \Bigg\}^{1/2}$$

$$:= \{a_{N1} + a_{N2} + a_{N3}\}^{1/2}.$$

For the first term,

$$a_{N1} =$$

$$\frac{n^2}{N^4} \sum_{i \in U_N} (y_i - \tilde{\boldsymbol{I}}_i^T \boldsymbol{\beta}_N)^4 \frac{(1-\pi_i)^3}{\pi_i^3} + \frac{n^2}{N^4} \sum\sum_{i \neq k \in U_N} (y_i - \tilde{\boldsymbol{I}}_i^T \boldsymbol{\beta}_N)^2 (y_k - \tilde{\boldsymbol{I}}_k^T \boldsymbol{\beta}_N)^2 \frac{(1-\pi_i)(1-\pi_k)\triangle_{ik}}{\pi_i^2 \pi_k^2}$$

$$\leq \frac{n^2}{N^3 \pi_{N*}^3} \sum_{i \in U_N} N^{-1} (y_i - \tilde{\boldsymbol{I}}_i^T \boldsymbol{\beta}_N)^4 + \frac{n^2}{N^4 \pi_{N*}^4} \left\{ \sum\sum_{i \neq k \in U_N} (y_i - \tilde{\boldsymbol{I}}_i^T \boldsymbol{\beta}_N)^4 (y_k - \tilde{\boldsymbol{I}}_k^T \boldsymbol{\beta}_N)^4 \right\}^{1/2}$$

$$\times \left\{ \sum\sum_{i \neq k \in U_N} \triangle_{ik}^2 \right\}^{1/2}$$

$$= \frac{n^2}{N^3 \pi_{N*}^3} \sum_{i \in U_N} N^{-1} (y_i - \tilde{\boldsymbol{I}}_i^T \boldsymbol{\beta}_N)^4 + \frac{n^2}{N^{5/2+\tau} \pi_{N*}^4} \sum_{i \in U_N} N^{-1} (y_i - \tilde{\boldsymbol{I}}_i^T \boldsymbol{\beta}_N)^4 \left\{ N^{2\tau} \max \sum_{i \neq k \in U_N} \triangle_{ik}^2 \right\}^{1/2}$$

$$= O\left(\frac{1}{n}\right)$$

by assumptions (A4) and (A5) as long as $\sum_{i \in U_N} N^{-1} (y_i - \tilde{\boldsymbol{I}}_i^T \boldsymbol{\beta}_N)^4 = O(1)$. The assumption (A8) and an argument similar to (2.23) for $\max_i \beta_{Ni}^4$ bounds the fourth moment term. The last term expands into components with two, three, or four distinct elements,

$$
\begin{aligned}
a_{N3} &= \frac{n^2}{N^4}\sum\sum_{i\neq j\in U_N}(y_i - \tilde{\boldsymbol{I}}_i^T\boldsymbol{\beta}_N)^2(y_j - \tilde{\boldsymbol{I}}_j^T\boldsymbol{\beta}_N)^2\frac{\triangle_{ij}^2(1-\pi_{ij})}{\pi_i^2\pi_j^2\pi_{ij}} \\
&+ \frac{4n^2}{N^4}\sum\sum\sum_{(i,j,l)\in D_{3,N}}(y_i - \tilde{\boldsymbol{I}}_i^T\boldsymbol{\beta}_N)^2(y_j - \tilde{\boldsymbol{I}}_j^T\boldsymbol{\beta}_N)(y_l - \tilde{\boldsymbol{I}}_l^T\boldsymbol{\beta}_N)\frac{\triangle_{ij}\triangle_{il}}{\pi_i^2\pi_j\pi_l\pi_{ij}\pi_{il}}E\left(I_iI_j-\pi_{ij}\right)\left(I_iI_l-\pi_{il}\right) \\
&+ \frac{n^2}{N^4}\sum\sum\sum\sum_{(i,j,k,l)\in D_{4,N}}(y_i - \tilde{\boldsymbol{I}}_i^T\boldsymbol{\beta}_N)(y_j - \tilde{\boldsymbol{I}}_j^T\boldsymbol{\beta}_N)(y_k - \boldsymbol{x}_k^T\boldsymbol{\beta}_N)(y_l - \tilde{\boldsymbol{I}}_l^T\boldsymbol{\beta}_N)\frac{\triangle_{ij}}{\pi_i\pi_j\pi_{ij}}\frac{\triangle_{kl}}{\pi_k\pi_l\pi_{kl}} \\
&\times E\left(I_iI_j-\pi_{ij}\right)\left(I_kI_l-\pi_{kl}\right) \\
&:= a_{N31} + a_{N32} + a_{N33}.
\end{aligned}
$$

Utilizing the bounded fourth moments and an additional condition on dependencies,

$$
\begin{aligned}
a_{N31} &\leq \frac{n^2}{N^4\pi_{N*}^4\pi_{N**}}\sum\sum_{i\neq j\in U_N}(y_i - \tilde{\boldsymbol{I}}_i^T\boldsymbol{\beta}_N)^2(y_j - \tilde{\boldsymbol{I}}_j^T\boldsymbol{\beta}_N)^2\triangle_{ij}^2 \\
&\leq \frac{n^2}{N^4\pi_{N*}^4\pi_{N**}}\left\{\sum\sum_{i\neq j\in U_N}(y_i - \tilde{\boldsymbol{I}}_i^T\boldsymbol{\beta}_N)^4(y_j - \tilde{\boldsymbol{I}}_j^T\boldsymbol{\beta}_N)^4\right\}^{1/2}\left\{\sum\sum_{i\neq j\in U_N}\triangle_{ij}^4\right\}^{1/2} \\
&\leq \frac{n^2}{N^{5/2+\xi}\pi_{N*}^4\pi_{N**}}\sum_{i\in U_N}\frac{(y_i - \tilde{\boldsymbol{I}}_i^T\boldsymbol{\beta}_N)^4}{N}\left\{N^{2\xi}\max\sum_{i\neq j\in U_N}\triangle_{ij}^4\right\}^{1/2} \\
&O\left(\frac{1}{N^{\xi-3/2}n^2\pi_{N**}}\right) \\
&= o(1)
\end{aligned}
$$

by assumption (A10). Bounding each element of a_{N32} by its absolute value and then applying the Cauchy-Schwarz Inequality,

$$
\begin{aligned}
a_{N32} \leq & \frac{4n^2}{N^4 \pi_{N*}^4 \pi_{N**}^2} \sum\sum\sum_{(i,j,l)\in D_{3,N}} (y_i - \tilde{\boldsymbol{I}}_i^T \boldsymbol{\beta}_N)^2 \left| y_j - \tilde{\boldsymbol{I}}_j^T \boldsymbol{\beta}_N \right| \left| y_l - \tilde{\boldsymbol{I}}_l^T \boldsymbol{\beta}_N \right| |\triangle_{ij}| \\
& \times |\triangle_{il}| \left| E\left(I_i I_j - \pi_{ij}\right)\left(I_i I_l - \pi_{il}\right) \right| \\
\leq & \frac{16n^2}{N^4 \pi_{N*}^4 \pi_{N**}^2} \left\{ \sum\sum\sum_{(i,j,l)\in D_{3,N}} (y_i - \tilde{\boldsymbol{I}}_i^T \boldsymbol{\beta}_N)^4 \left(y_j - \tilde{\boldsymbol{I}}_j^T \boldsymbol{\beta}_N \right)^2 \left(y_l - \tilde{\boldsymbol{I}}_l^T \boldsymbol{\beta}_N \right)^2 \right\}^{1/2} \\
& \times \left\{ \sum\sum\sum_{(i,j,l)\in D_{3,N}} \triangle_{ij}^2 \triangle_{il}^2 \right\}^{1/2} \\
\leq & \frac{16n^2}{N^4 \pi_{N*}^4 \pi_{N**}^2} \left\{ \sum_{i\in U} (y_i - \tilde{\boldsymbol{I}}_i^T \boldsymbol{\beta}_N)^4 \right\} \left\{ N \max_{i\in U} \left[\sum_{j\in U} \triangle_{ij}^2 \right]^2 \right\}^{1/2} \\
\leq & \frac{16n^2}{N^{5/2+2\tau} \pi_{N*}^4 \pi_{N**}^2} \left\{ \sum_{i\in U} \frac{(y_i - \tilde{\boldsymbol{I}}_i^T \boldsymbol{\beta}_N)^4}{N} \right\} N^{2\tau} \max_{i\in U} \sum_{j\in U} \triangle_{ij}^2 \\
= & O\left(\frac{1}{N^\tau n \pi_{N**}^2} \right) \\
= & o(1)
\end{aligned}
$$

by assumption (A9). Employing similar methods and assumption (A7iv),

$$
\begin{aligned}
a_{N33} \leq & \frac{n^2}{N^{3+2\tau} \pi_{N*}^4 \pi_{N**}^2} \left[N \max_{(i,j,k,l)\in D_{4,N}} \left| E_p\left[(I_i I_j - \pi_{ij})(I_k I_l - \pi_{kl}) \right] \right| \right] \\
& \times \sum_{j\in U_N} N^{-1} (y_j - \tilde{\boldsymbol{I}}_j^T \boldsymbol{\beta}_N)^4 \left[N^{2\tau} \max_{i\in U} \sum_{j\in U} \triangle_{ij}^2 \right] \\
= & O\left(\frac{n^2}{N^{3+2\tau} \pi_{N*}^4 \pi_{N**}^2} \right) \\
= & O\left(N^{-1/2} \right)
\end{aligned}
$$

by assumptions (A4), (A5), and (A9). Finally, the term $a_{N2} \to 0$ as $N \to \infty$ by Cauchy-Schwarz Inequality.

□

Lemma 2.4. *Under assumptions (A1) – (A10),*

$$\lim_{N\to\infty} E_p \left| \frac{n}{N^2} \sum\sum_{i,j\in U_N} \left[(y_i - \tilde{\boldsymbol{I}}_i^T \hat{\boldsymbol{\beta}}_N)(y_j - \tilde{\boldsymbol{I}}_j^T \hat{\boldsymbol{\beta}}_N) - (y_i - \tilde{\boldsymbol{I}}_i^T \boldsymbol{\beta}_N)(y_j - \tilde{\boldsymbol{I}}_j^T \boldsymbol{\beta}_N) \right] \frac{\triangle_{ij}}{\pi_i \pi_j} \frac{I_i I_j}{\pi_{ij}} \right| = 0.$$

Proof. Write

$$\begin{aligned}
& E_p \left| \frac{n}{N^2} \sum\sum_{i,j\in U_N} \left[(y_i - \tilde{\boldsymbol{I}}_i^T \hat{\boldsymbol{\beta}}_N)(y_j - \tilde{\boldsymbol{I}}_j^T \hat{\boldsymbol{\beta}}_N) - (y_i - \tilde{\boldsymbol{I}}_i^T \boldsymbol{\beta}_N)(y_j - \tilde{\boldsymbol{I}}_j^T \boldsymbol{\beta}_N) \right] \frac{\triangle_{ij}}{\pi_i \pi_j} \frac{I_i I_j}{\pi_{ij}} \right| \\
&= E_p \left| \frac{n}{N^2} \sum\sum_{i,j\in U_N} \left[2(y_i - \boldsymbol{\beta}_N^T \tilde{\boldsymbol{I}}_i) \tilde{\boldsymbol{I}}_j^T (\boldsymbol{\beta}_N - \hat{\boldsymbol{\beta}}_N) + (\boldsymbol{\beta}_N - \hat{\boldsymbol{\beta}}_N)^T \tilde{\boldsymbol{I}}_i \tilde{\boldsymbol{I}}_j^T (\boldsymbol{\beta}_N - \hat{\boldsymbol{\beta}}_N) \right] \frac{\triangle_{ij}}{\pi_i \pi_j} \frac{I_i I_j}{\pi_{ij}} \right| \\
&\leq E_p \left| \frac{n}{N^2} \sum\sum_{i,j\in U_N} 2(y_i - \boldsymbol{\beta}_N^T \tilde{\boldsymbol{I}}_i) \tilde{\boldsymbol{I}}_j^T (\boldsymbol{\beta}_N - \hat{\boldsymbol{\beta}}_N) \frac{\triangle_{ij}}{\pi_i \pi_j} \frac{I_i I_j}{\pi_{ij}} \right| \\
&+ E_p \left| \frac{n}{N^2} \sum\sum_{i,j\in U_N} (\boldsymbol{\beta}_N - \hat{\boldsymbol{\beta}}_N)^T \tilde{\boldsymbol{I}}_i \tilde{\boldsymbol{I}}_j^T (\boldsymbol{\beta}_N - \hat{\boldsymbol{\beta}}_N) \frac{\triangle_{ij}}{\pi_i \pi_j} \frac{I_i I_j}{\pi_{ij}} \right| \\
&:= b_{N1} + b_{N2}.
\end{aligned}$$

Define the vector $\boldsymbol{g}_N = \left(|\beta_{No} - \hat{\beta}_{No}|, |\beta_{N1} - \hat{\beta}_{N1}|, \ldots, |\beta_{NK} - \hat{\beta}_{NK}| \right)^T$, which is the element-wise absolute value of the vector $\boldsymbol{\beta}_N - \hat{\boldsymbol{\beta}}_N$. Also define the matrix

$$\boldsymbol{H}_N = \frac{n}{N^2} \sum\sum_{i,j\in U_N} \tilde{\boldsymbol{I}}_i \tilde{\boldsymbol{I}}_j^T \frac{|\triangle_{ij}|}{\pi_i \pi_j \pi_{ij}}.$$

Then we can bound the term, b_{N2},

$$\begin{aligned}
b_{N2} &= E_p \left| (\boldsymbol{\beta}_N - \hat{\boldsymbol{\beta}}_N)^T \left[\frac{n}{N^2} \sum\sum_{i,j\in U_N} \tilde{\boldsymbol{I}}_i \tilde{\boldsymbol{I}}_j^T \frac{\triangle_{ij}}{\pi_i \pi_j} \frac{I_i I_j}{\pi_{ij}} \right] (\boldsymbol{\beta}_N - \hat{\boldsymbol{\beta}}_N) \right| \\
&\leq E_p \left(\boldsymbol{g}_N^T \boldsymbol{H}_N \boldsymbol{g}_N \right) \qquad (2.32)
\end{aligned}$$

since $I_i I_j \leq 1$. By a property of quadratic forms,

$$\begin{aligned}
(2.32) &= Tr \left\{ \boldsymbol{H}_N var_p \left(\boldsymbol{g}_N \right) \right\} + \left\{ E_p(\boldsymbol{g}_N) \right\}^T \boldsymbol{H}_N \left\{ E_p(\boldsymbol{g}_N) \right\} \\
&:= b_{N21} + b_{N22}.
\end{aligned}$$

We can bound the variance element-wise by

$$var_p(\boldsymbol{g}_N) = E_p[\boldsymbol{g}_N\boldsymbol{g}_N^T] - [E_p(\boldsymbol{g}_N)][E_p(\boldsymbol{g}_N)]^T \leq 2\boldsymbol{Q}_N \tag{2.33}$$

where $\boldsymbol{Q}_N$ is a $K \times K$ matrix with elements $Q_{Nij} = \left\{E_p(\beta_{Ni} - \hat{\beta}_{Ni})^2 E_p(\beta_{Nj} - \hat{\beta}_{Nj})^2\right\}^{1/2}$. To bound each element of $\boldsymbol{Q}_N$, plug (2.25), (2.26), and (2.27) into $E_p(\beta_{Nt} - \hat{\beta}_{Nt})^2$ for $1 < t < K$,

$$\begin{aligned}
&E_p(\beta_{Nt} - \hat{\beta}_{Nt})^2 \\
=&\Bigg[\sum_{i_1=1}^{K}\sum_{i_2=1}^{K}\rho^{|t-i_1|}\rho^{|t-i_2|} - 2\rho^{t-1}\sum_{i_1=1}^{K}\sum_{i_2=1}^{K}\rho^{i_1-1}\rho^{|t-i_2|} - 2\rho^{K-t}\sum_{i_1=1}^{K}\sum_{i_2=1}^{K}\rho^{K-i_1}\rho^{|t-i_2|} \\
&+ 2\rho^{K-1}\sum_{i_1=1}^{K}\sum_{i_2=1}^{K}\rho^{i_1-1}\rho^{K-i_2} + \rho^{2(t-1)}\sum_{i_1=1}^{K}\sum_{i_2=1}^{K}\rho^{i_1-1}\rho^{i_2-1} \\
&+\rho^{2(K-t)}\sum_{i_1=1}^{K}\sum_{i_2=1}^{K}\rho^{K-i_1}\rho^{K-i_2}\Bigg] \times \frac{K^2}{N^2}\sum\sum_{j_1,j_2\in U_N} y_{j_1}y_{j_2}\frac{\tilde{I}_{i_1j_1}\tilde{I}_{i_2j_2}}{\pi_{j_1}\pi_{j_2}}\triangle_{j_1j_2} + R_{Nt}^* \qquad (2.34) \\
=&\sum_{i=1}^{K} f(\rho,t,i,i)\frac{K^2}{N^2}\sum_{j\in U_N} y_j^2\frac{\tilde{I}_{ij}(1-\pi_j)}{\pi_j} + \sum_{i=1}^{K} f(\rho,t,i,i)\frac{K^2}{N^2}\sum\sum_{j_1\neq j_2\in U_N} y_{j_1}y_{j_2}\frac{\tilde{I}_{ij_1}\tilde{I}_{ij_2}}{\pi_{j_1}\pi_{j_2}}\triangle_{j_1j_2} \\
&+ \sum_{\substack{i_1=1 \\ i_1\neq i_2}}^{K}\sum_{i_2=1}^{K} f(\rho,t,i_1,i_2)\frac{K^2}{N^2}\sum\sum_{j_1\neq j_2\in U_N} y_{j_1}y_{j_2}\frac{\tilde{I}_{i_1j_1}\tilde{I}_{i_2j_2}}{\pi_{j_1}\pi_{j_2}}\triangle_{j_1j_2} + R_{Nt}^* \qquad (2.35)
\end{aligned}$$

where $f(\rho,t,i_1,i_2) = \rho^{|t-i_1|}\rho^{|t-i_2|} - 2\rho^{t-1}\rho^{i_1-1}\rho^{|t-i_2|} - 2\rho^{K-t}\rho^{K-i_1}\rho^{|t-i_2|} + 2\rho^{K-1}\rho^{i_1-1}\rho^{K-i_2} + \rho^{2(t-1)}\rho^{i_1-1}\rho^{i_2-1} + \rho^{2(K-t)}\rho^{K-i_1}\rho^{K-i_2}$.

$$
\begin{aligned}
(2.35) \leq & 20M\frac{K}{N\pi_{N*}} + \sum_{i=1}^{K} f(\rho,t,i,i)\frac{K^2}{N^2\pi_{N*}^2}\sum\sum_{j_1\neq j_2\in U_N} |y_{j_1}||y_{j_2}|\tilde{I}_{ij_1}\tilde{I}_{ij_2}|\triangle_{j_1j_2}| \\
& + \sum_{\substack{i_1=1\\ i_1\neq i_2}}^{K}\sum_{i_2=1}^{K} f(\rho,t,i_1,i_2)\frac{K^2}{N^2\pi_{N*}^2}\sum\sum_{j_1\neq j_2\in U_N} |y_{j_1}||y_{j_2}|\tilde{I}_{i_1j_1}\tilde{I}_{i_2j_2}|\triangle_{j_1j_2}| \\
\leq & 20M\frac{K}{N\pi_{N*}} + \sum_{i=1}^{K} f(\rho,t,i,i)\frac{K^2}{N^2\pi_{N*}^2}\left\{\sum\sum_{j_1\neq j_2\in U_N} y_{j_1}^2 y_{j_2}^2\tilde{I}_{ij_1}\tilde{I}_{ij_2}\right\}^{1/2}\left\{\sum\sum_{j_1\neq j_2\in U_N}\triangle_{j_1j_2}^2\right\}^{1/2} \\
& + \sum_{\substack{i_1=1\\ i_1\neq i_2}}^{K}\sum_{i_2=1}^{K} f(\rho,t,i_1,i_2)\frac{K^2}{N^2\pi_{N*}^2}\left\{\sum\sum_{j_1\neq j_2\in U_N} y_{j_1}^2 y_{j_2}^2\tilde{I}_{i_1j_1}\tilde{I}_{i_2j_2}\right\}^{1/2}\left\{\sum\sum_{j_1\neq j_2\in U_N}\triangle_{j_1j_2}^2\right\}^{1/2} \\
\leq & \frac{20MK}{N\pi_{N*}} + \frac{56MK}{N^{1/2+\tau}\pi_{N*}^2}\left\{N^{2\tau}\max_{j_1\in U_N}\sum_{j_2\in U_N: j_2\neq j_1}\triangle_{j_1j_2}^2\right\}^{1/2}. \qquad (2.36)
\end{aligned}
$$

It can easily be shown that (2.36) also bounds $E(\beta_{Nt} - \hat{\beta}_{Nt})^2$ for $t = 1$ and $t = K$ since these terms appear in (2.34). For the diagonal elements of $\boldsymbol{H}_N$,

$$
\begin{aligned}
h_{Nii} = & \frac{n}{N^2}\sum\sum_{j_1,j_2\in U_N}\tilde{I}_{ij_1}\tilde{I}_{ij_2}\frac{|\triangle_{j_1j_2}|}{\pi_{j_1}\pi_{j_2}\pi_{j_1j_2}} \qquad (2.37) \\
= & \frac{n}{N^2}\sum_{j\in U_N}\tilde{I}_{ij}\frac{(1-\pi_j)}{\pi_j^2} + \frac{n}{N^2}\sum\sum_{j_1\neq j_2\in U_N}\tilde{I}_{ij_1}\tilde{I}_{ij_2}\frac{|\triangle_{j_1j_2}|}{\pi_{j_1}\pi_{j_2}\pi_{j_1j_2}} \\
\leq & \frac{n}{N^2\pi_{N*}^2}\sum_{j\in U_N}\tilde{I}_{ij} + \frac{n}{N^2\pi_{N*}^2\pi_{N**}}\left\{\sum\sum_{j_1\neq j_2\in U_N}\tilde{I}_{ij_1}\tilde{I}_{ij_2}\right\}^{1/2}\left\{\sum\sum_{j_1\neq j_2\in U_N}\triangle_{j_1j_2}^2\right\}^{1/2} \\
\leq & \frac{n}{NK\pi_{N*}^2} + \frac{n}{N^{1/2+\tau}K\pi_{N*}^2\pi_{N**}}\left\{N^{2\tau}\max\sum_{j_1\neq j_2\in U_N}\triangle_{j_1j_2}^2\right\}^{1/2}. \qquad (2.38)
\end{aligned}
$$

For the off-diagonal elements, $h_{Ni_1i_2}$, j_1 and j_2 are distinct and therefore

$$
h_{Ni_1i_2} \leq \frac{n}{N^{1/2+\tau}K\pi_{N*}^2\pi_{N**}}\left\{N^{2\tau}\max\sum_{j_1\neq j_2\in U_N}\triangle_{j_1j_2}^2\right\}^{1/2}.
$$

Finally, applying the bounds obtained in (2.36) and (2.38)

$$
\begin{aligned}
b_{N21} \leq & 2Tr\left(\boldsymbol{H}_{N}\boldsymbol{Q}_{N}\right) \\
\leq & K^2\left[\frac{20MK}{N\pi_{N*}} + \frac{56MK}{N^{1/2+\tau}\pi_{N*}^2}\left\{N^{2\tau}\max_{j_1\in U_N}\sum_{j_2\in U_N:j_2\neq j_1}\triangle_{j_1j_2}^2\right\}^{1/2}\right] \\
& \times\left[\frac{n}{NK\pi_{N*}^2} + \frac{n}{N^{1/2+\tau}K\pi_{N*}^2\pi_{N**}}\left\{N^{2\tau}\max\sum_{j_1\neq j_2\in U_N}\triangle_{j_1j_2}^2\right\}^{1/2}\right] \\
= & o(1)
\end{aligned}
$$

by assumptions (A4) – (A6) and (A9). Both b_{N1} and b_{N22} go to 0 as $N \to \infty$ by similar methods.

□

Lemma 2.5. *For the sample $\{x_j\}_{j\in s}$, assume (A11), (A12) and*

$$0 \leq \frac{x_j}{\pi_j} \leq 1 \quad \text{for } j \in U. \tag{2.39}$$

Denote the Horvitz-Thompson estimator for the population mean as

$$\bar{x}_{HT} = \frac{1}{N}\sum_{j\in U}\frac{x_j I_j}{\pi_j} \tag{2.40}$$

and the finite population mean as

$$\bar{x}_U = \frac{1}{N}\sum_{j\in U} x_j. \tag{2.41}$$

Then for $0 < t < nN^{-1} - \bar{x}_U$,

$$
\begin{aligned}
P\left(\bar{x}_{HT} - \bar{x}_U \geq t\right) \quad &\leq \quad \left\{\left(\frac{\bar{x}_U}{\bar{x}_U + t}\right)^{\bar{x}_U + t}\left(\frac{1-\frac{N}{n}\bar{x}_U}{1-\frac{N}{n}\bar{x}_U - \frac{N}{n}t}\right)^{n/N-\bar{x}_U-t}\right\}^N & (2.42)\\
&\leq \quad \exp\left\{-Nt^2 g(\bar{x}_U)\right\} & (2.43)\\
&\leq \quad \exp\left\{-\frac{2N^2t^2}{n}\right\} & (2.44)
\end{aligned}
$$

where

$$g(\bar{x}_U) = \begin{cases} \left(\frac{n}{N} - 2\bar{x}_U\right)^{-1} log\left(\frac{\frac{n}{N} - \bar{x}_U}{\bar{x}_U}\right) & : 0 < \bar{x}_U < \frac{n}{2N} \\ n\left(2N\bar{x}_U\left[\frac{n}{N} - \bar{x}_U\right]\right)^{-1} & : \frac{n}{2N} \leq \bar{x}_U < \frac{n}{N} \end{cases}.$$

Proof. Following the method of Hoeffding (1963), we can apply the following property of indicator functions

$$\begin{aligned} P\left(\bar{x} - \bar{x}_U \geq t\right) &= E_p\left[I\{\bar{x} - \bar{x}_U \geq t\}\right] \\ &= E_p\left[I\left\{\sum_{j\in U} \frac{x_j I_j}{\pi_j} - N\bar{x}_U - Nt \geq 0\right\}\right] \\ &\leq E_p \exp\left\{h\left(\sum_{j\in U} \frac{x_j I_j}{\pi_j} - N\bar{x}_U - Nt\right)\right\} \end{aligned} \tag{2.45}$$

for $h > 0$. The above relationship holds since $\exp(x) \geq 1$ if $x \geq 0$ and $\exp(x) > 0$ if $x < 0$. Pulling that which is not random out of the expectation, we can now rewrite (2.45) as follows

$$\begin{aligned} (2.45) &= \exp\Big\{-hNt - hN\bar{x}_U\Big\} E_p \exp\left\{h\sum_{j\in U}\frac{x_j I_j}{\pi_j}\right\} \\ &\leq \exp\Big\{-hNt - hN\bar{x}_U\Big\} E_{p^*} \exp\left\{h\sum_{l=1}^{n}\frac{1}{n}\sum_{j\in U}\frac{x_j I\{R_l = j\}}{p_j}\right\} \end{aligned} \tag{2.46}$$

by assumption (A12). Since each draw from the finite population is independent under with replacement sampling, we can take the expectation of each transformed draw individually,

$$\begin{aligned} (2.46) &= \exp\Big\{-hNt - hN\bar{x}_U\Big\} \prod_{l=1}^{n} E_{p^*} \exp\left\{h\left(\frac{1}{n}\sum_{j\in U}\frac{x_j I\{R_l = j\}}{p_j}\right)\right\} \\ &= \exp\Big\{-hNt - hN\bar{x}_U\Big\} \prod_{l=1}^{n} E_{p^*} \exp\left\{hV_l\right\} \end{aligned} \tag{2.47}$$

where we define V_l to be

$$V_l = \frac{1}{n}\sum_{j\in U}\frac{x_j I\{R_l = j\}}{p_j}.$$

Notice $0 \leq V_l \leq 1$ since the vector $[I\{R_l = j\}]_{j \in U}$ contains a single one for some $j^* \in U$ and zero otherwise which implies

$$V_l = \frac{1}{n}\frac{x_j^*}{p_j^*} = \frac{x_j^*}{\pi_j^*} \leq 1$$

by assumption (2.39). Additionally, the mean of V_l under sampling with replacement is given by

$$E_{p^*}V_l = E_{p^*}\left(\frac{1}{n}\sum_{j \in U}\frac{x_j I\{R_l = j\}}{p_j}\right) = \frac{1}{n}\sum_{j \in U} x_j = \frac{N}{n}\bar{x}_U.$$

Applying Lemma 1 in Hoeffding (1963) to the random quantity in (2.47), we obtain

$$\begin{aligned}
(2.47) \quad &= \quad \exp\Big\{-hNt - hN\bar{x}_U\Big\}\prod_{l=1}^{n}\left[1 - \frac{N}{n}\bar{x}_U + \frac{N}{n}\bar{x}_U e^h\right] \\
&\leq \quad \exp\Big\{-hNt - hN\bar{x}_U\Big\}\left[1 - \frac{N}{n}\bar{x}_U + \frac{N}{n}\bar{x}_U e^h\right]^n \\
&:= \quad Q(h, t, \bar{x}_U),
\end{aligned}$$

because the geometric mean is less than or equal to the arithmetic mean. To obtain the first inequality, we minimize the function $Q(h, t, \bar{x}_U)$ with respect to h and find

$$\begin{aligned}
h_o &= \arg\min_h Q(h, t, \bar{x}_U) \\
&= \log\left[\frac{(\bar{x}_U + t)\left(1 - \frac{N}{n}\bar{x}_U\right)}{\bar{x}_U\left(1 - \frac{N}{n}\bar{x}_U - \frac{N}{n}t\right)}\right].
\end{aligned}$$

Since we assumed $0 < t < nN^{-1} - \bar{x}_U, h_o > 0$ and we have

$$\begin{aligned}
P\left(\bar{x} - \bar{x}_U \geq t\right) &\leq Q(h_o, t, \bar{x}_U) \\
&= (2.42).
\end{aligned}$$

We can write $Q(h_o, t, \bar{x}_U)$ in the following way,

$$\begin{aligned} Q(h_o, t, \bar{x}_U) &= \left\{ \left(\frac{\bar{x}_U}{\bar{x}_U + t} \right)^{\bar{x}_U + t} \left(\frac{1 - \frac{N}{n}\bar{x}_U}{1 - \frac{N}{n}\bar{x}_U - \frac{N}{n}t} \right)^{n/N - \bar{x}_U - t} \right\}^N \\ &= \exp\left\{ -Nt^2 G(t, \bar{x}_U) \right\} \end{aligned}$$

where

$$G(t, \bar{x}_U) = \frac{\bar{x}_U + t}{t^2} \log\left(\frac{\bar{x}_U + t}{\bar{x}_U} \right) + \frac{\frac{n}{N} - \bar{x}_U - t}{t^2} \log\left(\frac{1 - \frac{N}{n}\bar{x}_U - \frac{N}{n}t}{1 - \frac{N}{n}\bar{x}_U} \right).$$

If we take the derivative of $G(t, \bar{x}_U)$ with respect to t, we get

$$\begin{aligned} t^2 \frac{\partial}{\partial t} G(t, \bar{x}_U) &= \frac{t^2 - 2t(\bar{x}_U + t)}{t^2} \log\left(\frac{\bar{x}_U + t}{\bar{x}_U} \right) + \frac{t^2 - 2t\left(\frac{n}{N} - \bar{x}_U\right)}{t^2} \log\left(\frac{1 - \frac{N}{n}\bar{x}_U - \frac{N}{n}t}{1 - \frac{N}{n}\bar{x}_U} \right) \\ &= \left(1 - 2\frac{\left(\frac{n}{N} - \bar{x}_U\right)}{t} \right) \log\left(1 - \frac{\frac{N}{n}t}{1 - \frac{N}{n}\bar{x}_U} \right) - \left(1 - 2\frac{\bar{x}_U + t}{t} \right) \log\left(\frac{\bar{x}_U}{\bar{x}_U + t} \right) \\ &= \left(1 - 2\frac{\left(\frac{n}{N} - \bar{x}_U\right)}{t} \right) \log\left(1 - \frac{t}{\frac{n}{N} - \bar{x}_U} \right) - \left(1 - 2\frac{\bar{x}_U + t}{t} \right) \log\left(1 - \frac{t}{\bar{x}_U + t} \right) \\ &= H\left(\frac{t}{\frac{n}{N} - \bar{x}_U} \right) - H\left(\frac{t}{\bar{x}_U + t} \right) \end{aligned}$$

where $H(x) = (1 - 2x^{-1})\log(1 - x)$. Since we assumed $0 < t < nN^{-1} - \bar{x}_U$,

$$0 < \frac{t}{\frac{n}{N} - \bar{x}_U} < 1 \text{ and } 0 < \frac{t}{\bar{x}_U + t} < 1.$$

For $|x| < 1$, we can write out $H(x)$ as two Taylor expansions:

$$\begin{aligned} H(x) &= (1 - \frac{2}{x})\log(1 - x) \\ &= \log(1 - x) - \frac{2}{x}\log(1 - x) \\ &= -\sum_{n=1}^{\infty} \frac{x^n}{n} + \frac{2}{x}\sum_{n=1}^{\infty} \frac{x^n}{n} \\ &= 2 + \sum_{n=1}^{\infty} \left(\frac{2}{n+1} - \frac{1}{n} \right) x^n \\ &= 2 + \left(\frac{2}{3} - \frac{1}{2} \right) x^2 + \left(\frac{2}{4} - \frac{1}{3} \right) x^3 \ldots \end{aligned}$$

Since the coefficients are positive, as x increases, so does $H(x)$. Therefore, $\frac{\partial}{\partial t}G(t, \bar{x}_U) > 0$ if and only if

$$\frac{t}{\frac{n}{N} - \bar{x}_U} > \frac{t}{\bar{x}_U + t}.$$

So if

$$\frac{n}{N} - 2\bar{x}_U > 0$$

then $G(t, \bar{x}_U)$ obtains a minimum at $t = \frac{n}{N} - 2\bar{x}_U$. But if

$$\frac{n}{N} - 2\bar{x}_U \leq 0$$

then $G(t, \bar{x}_U)$ obtains a minimum at $t = 0$. Let $t_o = \arg\min_t G(t, \bar{x}_U)$ and define $g(\bar{x}_U) = G(t_o, \bar{x}_U)$. Since $g(\bar{x}_U) > 0$ for $0 < \bar{x}_U < nN^{-1}$, we can obtain the second inequality where

$$\begin{aligned} Q(h_o, t, \bar{x}_U) &= \exp\left\{-Nt^2 G(t, \bar{x}_U)\right\} \\ &\leq \exp\left\{-Nt^2 g(\bar{x}_U)\right\} \\ &= (2.43). \end{aligned}$$

To obtain the last inequality, we only need to notice that $\min_{\bar{x}_U} g(\bar{x}_U) = 2Nn^{-1}$ and therefore

$$\begin{aligned} \exp\left\{-Nt^2 g(\bar{x}_U)\right\} &\leq \exp\left\{-2N^2 t^2 n^{-1}\right\} \\ &= (2.44). \end{aligned}$$

□

Corollary 2.1. *For the sample $\{x_j^*\}_{j \in s}$, assume (A11), (A12) and*

$$a \leq \frac{x_j^*}{\pi_j} \leq b \quad \textit{for } j \in U. \tag{2.48}$$

Denote the Horvitz-Thompson estimator for the population mean as

$$\bar{x}^*_{HT} = \frac{1}{N} \sum_{j \in U} \frac{x^*_j I_j}{\pi^*_j} \tag{2.49}$$

and the finite population mean as

$$\bar{x}^*_U = \frac{1}{N} \sum_{j \in U} x^*_j. \tag{2.50}$$

Then for $0 < t^* < nbN^{-1} - \bar{x}^*_U$*,*

$$\begin{aligned}
P\left(\bar{x}^*_{HT} - \bar{x}^*_U \geq t^*\right) &\leq \left(\frac{N\bar{x}^*_U - na}{N(\bar{x}^*_U + t^*) - na}\right)^{\frac{N(\bar{x}^*_U + t^*) - na}{(b-a)}} \left(\frac{nb - N\bar{x}^*_U}{nb - N\bar{x}^*_U - Nt^*}\right)^{\frac{nb - N\bar{x}^*_U - Nt^*}{(b-a)}} && (2.51)\\
&\leq \exp\left\{-N(t^*)^2 g(\bar{x}^*_U)\right\} && (2.52)\\
&\leq \exp\left\{-\frac{2N^2(t^*)^2}{n(b-a)^2}\right\} && (2.53)
\end{aligned}$$

where

$$g(\bar{x}^*_U) = \begin{cases} \left(\frac{N}{(b-a)\left(n(b-a) - 2N\bar{x}^*_U\right)}\right) log\left(\frac{nb - N\bar{x}^*_U}{N\bar{x}^*_U - na}\right) & : \frac{na}{N} < \bar{x}^*_U < \frac{n(b+a)}{2N} \\ \frac{nN}{2(N\bar{x}^*_U - na)(nb - N\bar{x}^*_U)} & : \frac{n(b+a)}{2N} \leq \bar{x}^*_U < \frac{na}{N} + (b-a) \end{cases}$$

Proof. If we define

$$x_j = \frac{x^*_j - a\pi_j}{b - a}$$

then by assumption (2.48), we have

$$0 \leq \frac{x_j}{\pi_j} \leq 1$$

Additionally if we define

$$t = \frac{t^*}{b - a}$$

and

$$\bar{x}_U = \frac{N\bar{x}_U^* - na}{N(b-a)}$$

then

$$P\left(\bar{x}^* - \bar{x}_U^* \geq t^*\right) = P\left(\bar{x} - \bar{x}_U \geq t\right)$$

and we can apply the results from Lemma 2.5 to obtain the three inequalities in terms of $\bar{x}_U^*$ and t^*. □

Lemma 2.6. *Under assumptions (A3) – (A5) and (A11) – (A13),*

$$\sup_{i:i=1,2,\ldots,K-1} |\hat{\kappa}_{Ni} - \kappa_{Ni}| \xrightarrow{p} 0$$

where $\hat{\kappa}_{Ni}$ *is defined in (2.17),* κ_{Ni} *in (2.19) and* $0 < p_{Ni} < 1$.

Proof. For population U_N, find the $\hat{\kappa}_N^*$ and κ_N^* such that

$$\max_{i:i=1,2,\ldots,K-1} |\hat{\kappa}_{Ni} - \kappa_{Ni}| = |\hat{\kappa}_N^* - \kappa_N^*| \,.$$

Using the technique in section 2.3.2 of Serfling (1980),

$$P(|\hat{\kappa}_N^* - \kappa_N^*| > \epsilon) \quad = \quad P(\hat{\kappa}_N^* > \epsilon + \kappa_N^*) + P(\hat{\kappa}_N^* < \kappa_N^* - \epsilon).$$

Since $\hat{F}_N(x)$ is a distribution function and using Lemma 1.1.4 in Serfling (1980),

$$
\begin{aligned}
P\left(\hat{\kappa}_N^* > \epsilon + \kappa_N^*\right) =& P\left(p_N^* > \hat{F}_N\left(\epsilon + \kappa_N^*\right)\right) \\
=& P\left(1 - p_N^* < \frac{1}{\hat{N}} \sum_{j \in s} \frac{1}{\pi_j} I\{x_j > \epsilon + \kappa_N^*\}\right) \\
=& P\left(\sum_{j \in s} \frac{1}{\pi_j} I\{x_j > \epsilon + \kappa_N^*\} - \hat{N}\left(1 - p_N^*\right) > 0\right) \\
=& P\left(\frac{1}{N} \sum_{j \in s} \frac{1}{\pi_j} I\{x_j > \epsilon + \kappa_N^*\} + \frac{(N - \hat{N})}{N}\left(1 - p_N^*\right) > \left(1 - p_N^*\right)\right) \\
=& P\left(\frac{1}{N} \sum_{j \in s} \frac{1}{\pi_j} I\{x_j > \epsilon + \kappa_N^*\} + A_N > \left(1 - p_N^*\right); |A_N| > \eta_\epsilon\right) \\
& + P\left(\frac{1}{N} \sum_{j \in s} \frac{1}{\pi_j} I\{x_j > \epsilon + \kappa_N^*\} + A_N > \left(1 - p_N^*\right); |A_N| \leq \eta_\epsilon\right) \qquad (2.54)
\end{aligned}
$$

where $A_N = (N - \hat{N})N^{-1}\left(1 - p_N^*\right)$ and $0 < \eta_\epsilon \leq 2^{-1}\epsilon \min\limits_{x \in [a,b]} f(x)$. Write

$$
(2.54) \quad \leq \quad P\left(|A_N| > \eta_\epsilon\right) + P\left(\sum_{j \in s} \frac{1}{\pi_j} I\{x_j > \epsilon + \kappa_N^*\} + A_N > N\left(1 - p_N^*\right); |A_N| \leq \eta_\epsilon\right).
$$

We know $P\left(|A_N| > \eta_\epsilon\right) \leq P\left(\left|(N - \hat{N})N^{-1}\right| > \eta_\epsilon\right) \to 0$ since $\hat{N}N^{-1}$ is consistent for 1. For the second term

$$
\begin{aligned}
& P\left(\frac{1}{N} \sum_{j \in s} \frac{1}{\pi_j} I\{x_j > \epsilon + \kappa_N^*\} + A_N > \left(1 - p_N^*\right); |A_N| \leq \eta_\epsilon\right) \\
\leq \quad & P\left(\frac{1}{N} \sum_{j \in s} \frac{1}{\pi_j} I\{x_j > \epsilon + \kappa_N^*\} + \eta_\epsilon > \left(1 - p_N^*\right)\right) \\
= \quad & P\left(\frac{1}{N} \sum_{j \in U_N} \frac{I\{j \in s\}}{\pi_j} I\{x_j > \epsilon + \kappa_N^*\} - \frac{1}{N} \sum_{j \in U_N} I\{x_j > \epsilon + \kappa_N^*\} > F_N(\epsilon + \kappa_N^*) - p_N^* - \eta_\epsilon\right)
\end{aligned}
$$

$$
\begin{aligned}
&= P\left(\frac{1}{N}\sum_{j\in U_N}\frac{I\{j\in s\}}{\pi_j}I\{x_j>\epsilon+\kappa_N^*\}-\frac{1}{N}\sum_{j\in U_N}I\{x_j>\epsilon+\kappa_N^*\}>F_N(\epsilon+\kappa_N^*)-p_N^*-\eta_\epsilon\right)\\
&\quad\times I\{F_N(\epsilon+\kappa_N^*)-p_N^*\le\eta_\epsilon\}\\
&\quad+P\left(\frac{1}{N}\sum_{j\in U_N}\frac{I\{j\in s\}}{\pi_j}I\{x_j>\epsilon+\kappa_N^*\}-\frac{1}{N}\sum_{j\in U_N}I\{x_j>\epsilon+\kappa_N^*\}>F_N(\epsilon+\kappa_N^*)-p_N^*-\eta_\epsilon\right)\\
&\quad\times I\{F_N(\epsilon+\kappa_N^*)-p_N^*>\eta_\epsilon\}\\
&\le I\{F_N(\epsilon+\kappa_N^*)-p_N^*\le\eta_\epsilon\}+P\left(\frac{1}{N}\sum_{j\in U_N}\frac{I\{j\in s\}}{\pi_j}I\{x_j>\epsilon+\kappa_N^*\}\right.\\
&\quad\left.-\frac{1}{N}\sum_{j\in U_N}I\{x_j>\epsilon+\kappa_N^*\}>F_N(\epsilon+\kappa_N^*)-p_N^*-\eta_\epsilon\right)I\left\{F_N(\epsilon+\kappa_N^*)-p_N^*>\eta_\epsilon\right\}\\
&:= d_{N1}+d_{N2}.
\end{aligned}
$$

For the first term

$$
\begin{aligned}
d_{N1} &\le I\{F_N(\epsilon+\kappa_N^*)-F_N(\kappa_{N^*})\le\eta_\epsilon\}\\
&= I\{\Big[F_N(\epsilon+\kappa_N^*)-F(\epsilon+\kappa_{N^*})\Big]+\Big[F(\kappa_{N^*})-F_N(\kappa_{N^*})\Big]+\Big[F(\epsilon+\kappa_{N^*})-F(\kappa_{N^*})\Big]\le\eta_\epsilon\}\\
&\le I\{\min_{x\in[a,b]}\Big[F_N(x)-F(x)\Big]+\min_{x\in[a,b]}\Big[F(x)-F_N(x)\Big]+\min_{x\in[a,b]}\Big[\epsilon f(x)\Big]\le\eta_\epsilon\}\\
&\overset{N\to\infty}{\to} I\{\epsilon\min_{x\in[a,b]}\Big[f(x)\Big]\le\eta_\epsilon\}\\
&= 0
\end{aligned}
$$

since $\eta_\epsilon < 2^{-1}\epsilon\min_{x\in[a,b]} f(x)$. Applying Corollary 2.1 to the second term

$$
\begin{aligned}
d_{N2} &\le \exp\left\{-\frac{2N^2(\pi_*)^2}{n}\left(F_N(\epsilon+\kappa_N^*)-p_N^*-\eta_\epsilon\right)^2\right\} I\left\{F_N(\epsilon+\kappa_N^*)-p_N^*>\eta_\epsilon\right\}\\
&\le \exp\left\{-\frac{2N^2(\pi_*)^2}{n}\left(\Big[F_N(\epsilon+\kappa_N^*)-F(\epsilon+\kappa_{N^*})\Big]+\Big[F(\kappa_{N^*})-F_N(\kappa_{N^*})\Big]\right.\right.\\
&\quad\left.\left.+\Big[F(\epsilon+\kappa_{N^*})-F(\kappa_{N^*})\Big]-\eta_\epsilon\right)^2\right\}I\left\{F_N(\epsilon+\kappa_N^*)-p_N^*>\eta_\epsilon\right\}\\
&\le \exp\left\{-\frac{2N^2(\pi_*)^2}{n}\left(\min_{x\in[a,b]}\Big[F_N(x)-F(x)\Big]+\min_{x\in[a,b]}\Big[F(x)-F_N(x)\Big]+\min_{x\in[a,b]}\Big[\epsilon f(x)\Big]-\eta_\epsilon\right)^2\right\}\\
&\to 0
\end{aligned}
$$

as $N \to \infty$. By similar methods

$$P\left(\hat{\kappa}_N^* < \kappa_N^* - \epsilon\right) \to 0$$

as $N \to \infty$. $\square$

Lemma 2.7. *Under assumptions (A1) – (A6), (A8), (A12) – (A14),*

$$\frac{1}{N} \sum_{j \in U_N} \left(\hat{\boldsymbol{I}}_j - \tilde{\boldsymbol{I}}_j\right)^T \left(\frac{I\{j \in s\}}{\pi_j} - 1\right) \left(\boldsymbol{\beta}_N - \hat{\boldsymbol{\beta}}_N^{(3)}\right) = o_p(1).$$

Proof.

$$\begin{aligned}
&P\left(\left|\frac{1}{N} \sum_{j \in U_N} \left(\hat{\boldsymbol{I}}_j - \tilde{\boldsymbol{I}}_j\right)^T \left(\frac{I\{j \in s\}}{\pi_j} - 1\right) \left(\boldsymbol{\beta}_N - \hat{\boldsymbol{\beta}}_N^{(2)}\right)\right| > \epsilon\right) \\
=&P\left(\left|\frac{1}{N} \sum_{i=1}^{K} \sum_{j \in U_N} (\beta_{Ni} - \hat{\beta}_{Ni}^{(2)}) \left(\hat{I}_{ij} - \tilde{I}_{ij}\right) \left(\frac{I\{j \in s\}}{\pi_j} - 1\right)\right| > \epsilon\right) \\
\leq&P\left(\frac{1}{N} \sum_{i=1}^{K} \sum_{j \in U_N} \left|\beta_{Ni} - \hat{\beta}_{Ni}^{(2)}\right| \left|\hat{I}_{ij} - \tilde{I}_{ij}\right| \left|\frac{I\{j \in s\}}{\pi_j} - 1\right| > \epsilon\right) \\
\leq&P\left(\max_i \left|\beta_{Ni} - \hat{\beta}_{Ni}^{(2)}\right| \frac{1}{N} \sum_{j \in U_N} \left|\frac{I\{j \in s\}}{\pi_j} - 1\right| \sum_{i=1}^{K} \left|\hat{I}_{ij} - \tilde{I}_{ij}\right| > \epsilon\right) \\
\leq&P\left(\max_i \left|\beta_{Ni} - \hat{\beta}_{Ni}^{(2)}\right| \left(\frac{2}{N\pi_{N*}} \sum_{j \in U_N} I\{j \in s\} + 2\right) > \epsilon\right) \\
\leq&P\left(\max_i \left|\beta_{Ni} - \hat{\beta}_{Ni}^{(2)}\right| \left(\frac{2n}{N\pi_{N*}} + 2\right) > \epsilon\right).
\end{aligned} \tag{2.55}$$

We can bound $\max_i \left|\beta_{Ni} - \hat{\beta}_{Ni}^{(2)}\right|$ with

$$\max_i \left|\beta_{Ni} - \hat{\beta}_{Ni}^{(2)}\right| \leq c_{N1} D_{N1} + c_{N2} D_{N2}$$

where $c_{N1} = O(1)$, $c_{N2} = O(1)$,

$$D_{N1} = \max_i \left| \frac{K}{N} \sum_{j \in U_N} y_j \tilde{I}_{ij} \left(1 - \frac{I\{j \in s\}}{\pi_j}\right) \right| \text{ and } D_{N2} = \max_i \frac{K}{N} \left| \sum_{j \in U_N} y_j \frac{I\{j \in s\}}{\pi_j} \left(\tilde{I}_{ij} - \hat{I}_{ij}\right) \right|$$

$$\leq \max_i \left[\frac{K}{N} \sum_{j \in U_N} |y_j| \frac{I\{j \in s\}}{\pi_j} \left|\tilde{I}_{ij} - \hat{I}_{ij}\right| \right]$$

$$= \frac{K}{N} \sum_{j \in U_N} |y_j| \frac{I\{j \in s\}}{\pi_j} \left|\tilde{I}_{*j} - \hat{I}_{*j}\right|.$$

In term D_{N2}, define $\tilde{I}_{*j} = I\{\kappa_{N*-1} \leq x_j < \kappa_{N*}\}$ and $\hat{I}_{*j} = I\{\hat{\kappa}_{N*-1} \leq x_j < \hat{\kappa}_{N*}\}$. Also, let $\boldsymbol{d}_{N*} = (d_{N*1}, d_{N*2}) = (\hat{\kappa}_{N*-1} - \kappa_{N*-1}, \hat{\kappa}_{N*} - \kappa_{N*})$ and let $\delta > 0$. Therefore, we have

$$(2.55) \leq P\left(\left[c_{N1} D_{N1} + c_{N2} \frac{K}{N} \sum_{j \in U_N} |y_j| \frac{I\{j \in s\}}{\pi_j} \left|\tilde{I}_{*j} - \hat{I}_{*j}\right|\right] \left(\frac{2n}{N\pi_{N*}} + 2\right) > \epsilon\right)$$

$$\leq P\left(||\boldsymbol{d}_{N*}||_{L1} > \delta\right)$$

$$+ P\left(\left[c_{N1} D_{N1} + c_{N2} \frac{K}{N} \sum_{j \in U_N} |y_j| \frac{I\{j \in s\}}{\pi_j} \left|\tilde{I}_{*j} - \hat{I}_{*j}\right|\right] \left(\frac{2n}{N\pi_{N*}} + 2\right) > \epsilon; ||\boldsymbol{d}_{N*}||_{L1} \leq \delta\right).$$

The first term goes to zero as $N \to \infty$ by Lemma 2.6. For the second term, we have

$$P\left(\left[c_{N1} D_{N1} + c_{N2} \frac{K}{N} \sum_{j \in U_N} |y_j| \frac{I\{j \in s\}}{\pi_j} \left|\tilde{I}_{*j} - \hat{I}_{*j}\right|\right] \left(\frac{2n}{N\pi_{N*}} + 2\right) > \epsilon; ||\boldsymbol{d}_{N*}||_{L1} \leq \delta\right)$$

$$\leq P\left(c_{N1} \left(\frac{2n}{N\pi_{N*}} + 2\right) D_{N1} + \sup_{||\boldsymbol{d}_{N*}|| \leq \delta} c_{N2} \left(\frac{2n}{N\pi_{N*}} + 2\right) \frac{K}{N} \sum_{j \in U_N} |y_j| \frac{I\{j \in s\}}{\pi_j} \left|\tilde{I}_{*j} - \hat{I}_{*j}\right| > \epsilon\right)$$

$$\leq \frac{1}{\epsilon} c_{N1} \left(\frac{2n}{N\pi_{N*}} + 2 \right) E_p D_{N1} \tag{2.56}$$

$$+ \frac{1}{\epsilon} c_{N2} \left(\frac{2n}{N\pi_{N*}} + 2 \right) E_p \left(\frac{K}{N} \sum_{j \in U_N} |y_j| \frac{I_j}{\pi_j} I\{\kappa_{N*-1} - \delta < x_j \leq \kappa_{N*-1}\} I\{x_j \leq \kappa_{N*}\} \right) \tag{2.57}$$

$$+ \frac{1}{\epsilon} c_{N2} \left(\frac{2n}{N\pi_{N*}} + 2 \right) E_p \left(\frac{K}{N} \sum_{j \in U_N} |y_j| \frac{I_j}{\pi_j} I\{\kappa_{N*-1} < x_j \leq \kappa_{N*-1} + \delta\} I\{x_j \leq \kappa_{N*}\} \right) \tag{2.58}$$

$$+ \frac{1}{\epsilon} c_{N2} \left(\frac{2n}{N\pi_{N*}} + 2 \right) E_p \left(\frac{K}{N} \sum_{j \in U_N} |y_j| \frac{I_j}{\pi_j} I\{\kappa_{N*} - \delta < x_j \leq \kappa_{N*}\} I\{x_j > \kappa_{N*-1}\} \right) \tag{2.59}$$

$$+ \frac{1}{\epsilon} c_{N2} \left(\frac{2n}{N\pi_{N*}} + 2 \right) E_p \left(\frac{K}{N} \sum_{j \in U_N} |y_j| \frac{I_j}{\pi_j} I\{\kappa_{N*} < x_j \leq \kappa_{N*} + \delta\} I\{x_j > \kappa_{N*-1}\} \right). \tag{2.60}$$

By assumption (A5), we know $c_{N2}\left(2n(N\pi_{N*})^{-1} + 2\right) = O(1)$. Let $\tilde{I}_{*j} = \underset{\tilde{I}_{ij}: i=1,\ldots,K}{\arg\max} \left| KN^{-1} \sum_{j \in U_N} y_j \tilde{I}_{ij}(1 - I_j \pi_j^{-1}) \right|$ so that we can write the rest of (2.56) as

$$\begin{aligned}
E_p D_{N1} \leq & \left\{ \frac{K^2}{N^2} \sum\sum_{j,l \in U_N} y_j y_l \tilde{I}_{*j} \tilde{I}_{*l} \frac{\triangle_{jl}}{\pi_j \pi_l} \right\}^{1/2} \\
= & \left\{ \frac{K^2}{N^2} \sum_{j \in U_N} y_j^2 \tilde{I}_{*j} \frac{1 - \pi_j}{\pi_j} + \frac{K^2}{N^2} \sum\sum_{j \neq l \in U_N} y_j y_l \tilde{I}_{*j} \tilde{I}_{*l} \frac{\triangle_{jl}}{\pi_j \pi_l} \right\}^{1/2} \\
\leq & \left\{ \frac{K}{N\pi_{N*}} M + \frac{K^2}{N^2 \pi_{N*}^2} \left[\max_i \sum_{j \in U_N} y_j^2 \tilde{I}_{*j} \right] \left[N \max \sum_{j \neq l \in U_N} \triangle_{jl}^2 \right]^{1/2} \right\}^{1/2} \\
\leq & \left\{ \frac{K}{N\pi_{N*}} M + \frac{K}{N^{1/2+\tau} \pi_{N*}^2} M \left[N^{2\tau} \max \sum_{j \neq l \in U_N} \triangle_{jl}^2 \right]^{1/2} \right\}^{1/2} \\
= & O\left(\frac{K^{1/2}}{N^{1/2}} \right) \\
= & o(1)
\end{aligned}$$

by assumption (A6). For last half of (2.57), we can now easily take the expectation and apply assumptions (A8) and (A13) to obtain

$$
\begin{aligned}
&\frac{K}{N}\sum_{j\in U_N}|y_j|I\{\kappa_{N*-1}-\delta<x_j\leq\kappa_{N*-1}\}I\{x_j\leq\kappa_{N*}\}\\
\leq &K\left\{\frac{1}{N}\sum_{j\in U_N}y_j^2\right\}^{1/2}\left\{F_N(\kappa_{N*-1})-F_N(\kappa_{N*-1}-\delta)\right\}^{1/2}\\
\leq &\frac{K}{N^{1/4}}\left\{\frac{1}{N}\sum_{j\in U_N}y_j^4\right\}^{1/4}\left\{2\sup_x|F_N(x)-F(x-\delta)|+\sup_x|F(x)-F(x-\delta)|\right\}^{1/2}\\
= &o\left(\frac{K}{N^{1/4}}\right)\\
= &o(1)
\end{aligned}
$$

by assumption (A14). The terms (2.58) – (2.60) are $o(1)$ by similar logic.

□

The last two lemmas concern the asymptotic equivalence of $N^{-1}\hat{t}_y$ and $N^{-1}\hat{t}_y^*$. Lemma 2.8 provides the rate of convergence of the standardized mean squared error of the estimated cell totals and it is an important result for Lemma 2.9 because the critical difference between $N^{-1}\hat{t}_y$ and $N^{-1}\hat{t}_y^*$ is the estimated cell totals present in Ω_s. We use an additional assumption on the relationship between sample size and knot size along with an assumption concerning the higher order moments of the estimated cell totals:

A15. Assume $K = O(n^{1/4})$.

A16. For $m = 2, 3, \ldots$, assume

$$E_p \max_{i:i=1,2,\ldots,K}\left[\frac{K}{N}\left(\frac{N}{K}-\hat{N}_i\right)\right]^{2m} \leq \frac{c}{n^{m/2}}$$

where $c > 0$.

Lemma 2.8. *Under assumptions (A1)- (A5) and (A15),*

$$E_p \max_{i:i=1,2,\ldots,K}\left[\frac{K}{N}\left(\frac{N}{K}-\hat{N}_i\right)\right]^{2} = O(n^{-3/4}) \tag{2.61}$$

Proof. To bound (2.61), write it as

$$
\begin{aligned}
(2.61) =& E_p \max_{i:i=1,2,\dots,K} \left[\frac{K}{N} \sum_{j \in U_N} \tilde{I}_{ij} \left(1 - \frac{I_j}{\pi_j} \right) \right]^2 \\
=& \max_i \frac{K^2}{N^2} \sum_{j \in U_N} \tilde{I}_{ij} \frac{(1-\pi_j)}{\pi_j} + \max_i \frac{K^2}{N^2} \sum\sum_{j_1 \neq j_2} \tilde{I}_{ij_1} \tilde{I}_{ij_2} \frac{\triangle_{j_1 j_2}}{\pi_{j_1}\pi_{j_2}} \\
\leq& \frac{K^2}{\pi_{N*} N^2} \max_i \sum_{j \in U_N} \tilde{I}_{ij} + \frac{K^2}{\pi_{N*}^2 N^2} \left\{ \max_i \sum\sum_{j_1 \neq j_2} \tilde{I}_{ij_1} \tilde{I}_{ij_2} \right\}^{1/2} \left\{ \sum\sum_{j_1 \neq j_2} \triangle_{j_1 j_2}^2 \right\}^{1/2} \\
\leq& \frac{K}{\pi_{N*} N} + \frac{K^2}{\pi_{N*}^2 N^2} \left\{ \max_i \sum_{j \in U_N} \tilde{I}_{ij} \right\} \left\{ N \max \sum_{j_1 \neq j_2} \triangle_{j_1 j_2}^2 \right\}^{1/2} \\
\leq& \frac{K}{\pi_{N*} N} + \frac{K}{\pi_{N*}^2 N^{1/2+\tau}} \left\{ N^{2\tau} \max \sum_{j_1 \neq j_2} \triangle_{j_1 j_2}^2 \right\}^{1/2} \\
=& \frac{1}{n^{3/4}} \frac{K}{n^{1/4}} \frac{n}{\pi_{N*} N} + \frac{1}{n^{3/4}} \frac{K}{n^{1/4}} \frac{n}{\pi_{N*}^2 N^{1/2+\tau}} \left\{ N^{2\tau} \max \sum_{j_1 \neq j_2} \triangle_{j_1 j_2}^2 \right\}^{1/2} \\
=& O(n^{-3/4})
\end{aligned}
$$

by assumptions (A4), (A5), and (A15). □

Lemma 2.9. *Under assumptions (A1) – (A6), (A15), and (A16),* $N^{-1}\hat{t}_y^* = N^{-1}\hat{t}_y + o_p(1)$.

Proof. Apply Taylor's Theorem to the matrices $\boldsymbol{\Omega}_s^{-1}$ and $\boldsymbol{\Omega}_U^{-1}$ and write the difference between $N^{-1}\hat{t}_y^*$ and $N^{-1}\hat{t}_y$ in three parts:

$$
\begin{aligned}
\frac{\hat{t}_y^* - \hat{t}_y}{N} =& \frac{1}{N} \sum_{j \in U_N} \tilde{\boldsymbol{I}}_j^T \left(1 - \frac{I_j}{\pi_j} \right) \left[\boldsymbol{\Omega}_s^{-1} - \boldsymbol{\Omega}_U^{-1} \right] \frac{1}{NK^{-1} + 2\lambda} \sum_{j \in U_N} y_j \tilde{\boldsymbol{I}}_j \frac{I_j}{\pi_j} \\
=& \frac{1}{N} \sum_{j \in U_N} \tilde{\boldsymbol{I}}_j^T \left(1 - \frac{I_j}{\pi_j} \right) \left[(\boldsymbol{I} + \boldsymbol{A}_s)^{-1} - (\boldsymbol{I} + \boldsymbol{A}_U)^{-1} \right] \frac{1}{NK^{-1} + 2\lambda} \sum_{j \in U_N} y_j \tilde{\boldsymbol{I}}_j \frac{I_j}{\pi_j}
\end{aligned}
$$

$$
\begin{aligned}
=&\frac{1}{N}\sum_{j\in U_N}\tilde{\boldsymbol{I}}_j^T\left(1-\frac{I_j}{\pi_j}\right)\left[\left(\boldsymbol{I}-\boldsymbol{A}_s+\boldsymbol{A}_s^2\ldots\right)-\left(\boldsymbol{I}-\boldsymbol{A}_U+\boldsymbol{A}_U^2\ldots\right)\right]\\
&\times\frac{1}{NK^{-1}+2\lambda}\sum_{j\in U_N}y_j\tilde{\boldsymbol{I}}_j\frac{I_j}{\pi_j}\\
=&\frac{1}{N}\sum_{j\in U_N}\tilde{\boldsymbol{I}}_j^T\left(1-\frac{I_j}{\pi_j}\right)\left[\boldsymbol{A}_U-\boldsymbol{A}_s\right]\frac{1}{NK^{-1}+2\lambda}\sum_{j\in U_N}y_j\tilde{\boldsymbol{I}}_j\frac{I_j}{\pi_j} \qquad (2.62)\\
&+\sum_{m=2}^{\infty}(-1)^m\frac{1}{N}\sum_{j\in U_N}\tilde{\boldsymbol{I}}_j^T\left(1-\frac{I_j}{\pi_j}\right)\boldsymbol{A}_U^m\frac{1}{NK^{-1}+2\lambda}\sum_{j\in U_N}y_j\tilde{\boldsymbol{I}}_j\frac{I_j}{\pi_j} \qquad (2.63)\\
&+\sum_{m=2}^{\infty}(-1)^{m-1}\frac{1}{N}\sum_{j\in U_N}\tilde{\boldsymbol{I}}_j^T\left(1-\frac{I_j}{\pi_j}\right)\boldsymbol{A}_s^m\frac{1}{NK^{-1}+2\lambda}\sum_{j\in U_N}y_j\tilde{\boldsymbol{I}}_j\frac{I_j}{\pi_j} \qquad (2.64)
\end{aligned}
$$

where $A_{U(1,1)} = A_{U(K,K)} = (C_N+\lambda)(C_N+2\lambda)^{-1}-1$, $A_{U(i,i)} = 0$ for $1 < i < K$, $A_{U(i,j)} = -\lambda(C_N+2\lambda)^{-1}$ for $|i-j| = 1$ and $A_{U(i,j)} = 0$ for $|i-j| > 1$. Also, $A_{s(i,j)} = A_{U(i,j)}$ for $|i-j| > 0$ but on the diagonal $A_{s(i,i)} = (\hat{N}_i+\lambda)(C_N+2\lambda)^{-1}-1$ for $i = 1$ and K and $A_{s(i,i)} = (\hat{N}_i+2\lambda)(C_N+2\lambda)^{-1}-1$ for $1 < i < K$. The first term (2.62) can be bounded using the Cauchy-Schwarz Inequality,

$$
\begin{aligned}
&E_p\left|\frac{1}{N}\sum_{j\in U_N}\tilde{\boldsymbol{I}}_j^T\left(1-\frac{I_j}{\pi_j}\right)\left[\boldsymbol{A}_U-\boldsymbol{A}_s\right]\frac{1}{NK^{-1}+2\lambda}\sum_{j\in U_N}y_j\tilde{\boldsymbol{I}}_j\frac{I_j}{\pi_j}\right|\\
=&E_p\left|\frac{1}{K}\sum_{i=1}^{K}\frac{K^2}{N^2}\left(\frac{N}{K}-\hat{N}_i\right)^2\frac{K}{N}\sum_{j\in U_N}y_j\tilde{I}_{ij}\frac{I_j}{\pi_j}\left[\frac{N^2}{K^2}(NK^{-1}+2\lambda)^{-2}\right]\right|\\
\leq&E_p\frac{1}{K}\sum_{i=1}^{K}\frac{K^2}{N^2}\left(\frac{N}{K}-\hat{N}_i\right)^2\frac{K}{N}\sum_{j\in U_N}|y_j|\tilde{I}_{ij}\frac{I_j}{\pi_j}\left[\frac{N^2}{K^2}(NK^{-1}+2\lambda)^{-2}\right]\\
\leq&M\frac{1}{\pi_*}\left[\frac{N^2}{K^2}(NK^{-1}+2\lambda)^{-2}\right]E_p\frac{1}{K}\sum_{i=1}^{K}\frac{K^2}{N^2}\left(\frac{N}{K}-\hat{N}_i\right)^2\\
\leq&M\frac{1}{\pi_*}\left[\frac{N^2}{K^2}(NK^{-1}+2\lambda)^{-2}\right]E_p\max_{i:i=1,2,\ldots,K}\frac{K^2}{N^2}\left(\frac{N}{K}-\hat{N}_i\right)^2\\
=&O\left(\frac{KN}{n^2}\right)\\
=&o(1)
\end{aligned}
$$

by (A6). For the terms in (2.63) and (2.64), we need to bound the element-wise absolute value of $\boldsymbol{A}_U^m$ and $\boldsymbol{A}_s^m$. A crude bound for each is

$$|A_U^m| = \left[\max_{ij} |A_{U(i,j)}|\right]^m K^{m-1}\boldsymbol{J} = \left[\frac{\lambda}{NK^{-1}+2\lambda}\right]^m K^{m-1}\boldsymbol{J}$$

and

$$|A_s^m| = \left[\max_{ij} |A_{s(i,j)}|\right]^m K^{m-1}\boldsymbol{J} = \left[\max\left(\frac{\lambda}{NK^{-1}+2\lambda}, \max_i \left|\frac{K}{N}\left(\frac{N}{K} - \hat{N}_i\right)\right|\right)\right]^m K^{m-1}\boldsymbol{J}$$

where $\boldsymbol{J}$ is a matrix of ones and $||$ represents element-wise absolute value. Then we can bound the expectation of the absolute value of the individual terms of (2.63) as follows:

$$\begin{aligned}
&E_p \left| \frac{1}{N} \sum_{j\in U_N} \tilde{\boldsymbol{I}}_j^T \left(1 - \frac{I_j}{\pi_j}\right) \boldsymbol{A}_U^m \frac{1}{NK^{-1}+2\lambda} \sum_{j\in U_N} y_j \tilde{\boldsymbol{I}}_j \frac{I_j}{\pi_j} \right| \\
\le & E_p \left| \frac{1}{N} \sum_{j\in U_N} \tilde{\boldsymbol{I}}_j^T \left(1 - \frac{I_j}{\pi_j}\right)\right| |\boldsymbol{A}_U^m| \left| \frac{1}{NK^{-1}+2\lambda} \sum_{j\in U_N} y_j \tilde{\boldsymbol{I}}_j \frac{I_j}{\pi_j} \right| \\
\le & E_p \left| \frac{1}{N} \sum_{j\in U_N} \tilde{\boldsymbol{I}}_j^T \left(1 - \frac{I_j}{\pi_j}\right)\right| \left[\frac{\lambda}{NK^{-1}+2\lambda}\right]^m K^{m-1}\boldsymbol{J} \left| \frac{1}{NK^{-1}+2\lambda} \sum_{j\in U_N} y_j \tilde{\boldsymbol{I}}_j \frac{I_j}{\pi_j} \right| \\
= & \left[\frac{\lambda}{NK^{-1}+2\lambda}\right]^m K^{m-1} \left(\frac{NK^{-1}}{NK^{-1}+2\lambda}\right) \\
& \times E_p \sum_{i_1=1}^K \sum_{i_2=1}^K \left| \frac{1}{N} \sum_{j_1\in U_N} \tilde{I}_{i_1 j_1} \left(1 - \frac{I_{j_1}}{\pi_{j_1}}\right)\right| \left| \frac{K}{N} \sum_{j_2\in U_N} y_{j_2} \tilde{I}_{i_2 j_2} \frac{I_{j_2}}{\pi_{j_2}} \right| \\
\le & \left[\frac{\lambda}{NK^{-1}+2\lambda}\right]^m K^{m} \left(\frac{NK^{-1}}{NK^{-1}+2\lambda}\right) E_p \sum_{i_1=1}^K \left| \frac{1}{N} \sum_{j_1\in U_N} \tilde{I}_{i_1 j_1} \left(1 - \frac{I_{j_1}}{\pi_{j_1}}\right)\right| \frac{1}{N} \sum_{j_2\in U_N} |y_{j_2}| \frac{1}{\pi_{j_2}} \\
\le & \left[\frac{\lambda}{NK^{-1}+2\lambda}\right]^m K^{m} \left(\frac{NK^{-1}}{NK^{-1}+2\lambda}\right) M \frac{1}{\pi_{N*}} E_p \frac{1}{K} \sum_{i_1=1}^K \max_i \left| \frac{K}{N} \sum_{j_1\in U_N} \tilde{I}_{i_1 j_1} \left(1 - \frac{I_{j_1}}{\pi_{j_1}}\right)\right| \\
\le & \left[\frac{\lambda}{NK^{-1}+2\lambda}\right]^m K^{m} \left(\frac{NK^{-1}}{NK^{-1}+2\lambda}\right) M \frac{1}{\pi_{N*}} \left\{ E_p \max_i \left[\frac{K}{N} \sum_{j_1\in U_N} \tilde{I}_{i_1 j_1} \left(1 - \frac{I_{j_1}}{\pi_{j_1}}\right)\right]^2 \right\}^{1/2} \\
\le & \left[\frac{\lambda K}{NK^{-1}+2\lambda}\right]^m \left(\frac{NK^{-1}}{NK^{-1}+2\lambda}\right) M \left\{ \frac{K}{\pi_{N*}^3 N} + \frac{K}{\pi_{N*}^4 N^{1/2+\tau}} \left\{ N^{2\tau} \max_{j_1 \neq j_2} \sum \triangle_{j_1 j_2}^2 \right\}^{1/2} \right\}^{1/2} .
\end{aligned} \tag{2.65}$$

Now we can use the individual term bound found in (2.65) to bound the expectation of the absolute value of (2.63):

$$
\begin{aligned}
&E_p\left|\sum_{m=2}^{\infty}(-1)^m\frac{1}{N}\sum_{j\in U_N}\tilde{\boldsymbol{I}}_j^T\left(1-\frac{I_j}{\pi_j}\right)\boldsymbol{A}_U^m\frac{1}{NK^{-1}+2\lambda}\sum_{j\in U_N}y_j\tilde{\boldsymbol{I}}_j\frac{I_j}{\pi_j}\right|\\
&\leq\sum_{m=2}^{\infty}E_p\left|(-1)^m\frac{1}{N}\sum_{j\in U_N}\tilde{\boldsymbol{I}}_j^T\left(1-\frac{I_j}{\pi_j}\right)\boldsymbol{A}_U^m\frac{1}{NK^{-1}+2\lambda}\sum_{j\in U_N}y_j\tilde{\boldsymbol{I}}_j\frac{I_j}{\pi_j}\right|\\
&\leq\sum_{m=2}^{\infty}(2.65)\\
&=\left(\frac{NK^{-1}}{NK^{-1}+2\lambda}\right)M\left\{\frac{K}{\pi_{N*}^3N}+\frac{K}{\pi_{N*}^4N^{1/2+\tau}}\left\{N^{2\tau}\max\sum_{j_1\neq j_2}\triangle_{j_1j_2}^2\right\}^{1/2}\right\}^{1/2}\sum_{m=2}^{\infty}\left[\frac{\lambda K}{NK^{-1}+2\lambda}\right]^m\\
&=o(1)\left[\frac{1}{1-\frac{\lambda K}{NK^{-1}+2\lambda}}-1-\frac{\lambda K}{NK^{-1}+2\lambda}\right]\\
&=o(1).
\end{aligned}
$$

For the terms in (2.64), we assume, without loss of generality, that $\max_{ij}|A_{s(i,j)}|=\max_i\left|\frac{K}{N}\left(\frac{N}{K}-\hat{N}_i\right)\right|$ since the other case is covered by the terms of (2.63) and can find the expectation of the absolute value of a term in (2.64) is bounded by:

$$
\begin{aligned}
&E_p\left|\frac{1}{N}\sum_{j\in U}\tilde{\boldsymbol{I}}_j^T\left(1-\frac{I_j}{\pi_j}\right)\boldsymbol{A}_s^m\frac{1}{NK^{-1}+2\lambda}\sum_{j\in U}y_j\tilde{\boldsymbol{I}}_j\frac{I_j}{\pi_j}\right|\\
\leq&E_p\left|\frac{1}{N}\sum_{j\in U}\tilde{\boldsymbol{I}}_j^T\left(1-\frac{I_j}{\pi_j}\right)\right||\boldsymbol{A}_s^m|\left|\frac{1}{NK^{-1}+2\lambda}\sum_{j\in U}y_j\tilde{\boldsymbol{I}}_j\frac{I_j}{\pi_j}\right|\\
\leq&E_p\left|\frac{1}{N}\sum_{j\in U}\tilde{\boldsymbol{I}}_j^T\left(1-\frac{I_j}{\pi_j}\right)\right|\max_i\left|\frac{K}{N}\left(\frac{N}{K}-\hat{N}_i\right)\right|^mK^{m-1}\boldsymbol{J}\left|\frac{1}{NK^{-1}+2\lambda}\sum_{j\in U}y_j\tilde{\boldsymbol{I}}_j\frac{I_j}{\pi_j}\right|\\
=&K^{m-1}\left(\frac{NK^{-1}}{NK^{-1}+2\lambda}\right)E_p\left[\max_i\left|\frac{K}{N}\left(\frac{N}{K}-\hat{N}_i\right)\right|^m\right.\\
&\left.\sum_{i_1=1}^{K}\sum_{i_2=1}^{K}\left|\frac{1}{N}\sum_{j_1\in U}\tilde{I}_{i_1j_1}\left(1-\frac{I_{j_1}}{\pi_{j_1}}\right)\right|\left|\frac{K}{N}\sum_{j_2\in U}y_{j_2}\tilde{I}_{i_2j_2}\frac{I_{j_2}}{\pi_{j_2}}\right|\right]
\end{aligned}
$$

$$
\begin{aligned}
&\le K^m\left(\frac{NK^{-1}}{NK^{-1}+2\lambda}\right)M\frac{1}{\pi_{N*}}E_p\max_i\left|\frac{K}{N}\left(\frac{N}{K}-\hat{N}_i\right)\right|^m\max_i\left|\frac{K}{N}\left(\frac{N}{K}-\hat{N}_i\right)\right|\\
&\le K^m\left(\frac{NK^{-1}}{NK^{-1}+2\lambda}\right)M\frac{1}{\pi_{N*}}\left\{E_p\max_i\left[\frac{K}{N}\left(\frac{N}{K}-\hat{N}_i\right)\right]^{2m}\right\}^{1/2}\left\{E_p\max_i\left[\frac{K}{N}\left(\frac{N}{K}-\hat{N}_i\right)\right]^{2}\right\}^{1/2}\\
&\le K^m\left(\frac{NK^{-1}}{NK^{-1}+2\lambda}\right)M\left\{\frac{K}{\pi_{N*}^3N}+\frac{K}{\pi_{N*}^4N^{1/2+\tau}}\left\{N^{2\tau}\max_{j_1\neq j_2}\sum\triangle_{j_1j_2}^2\right\}^{1/2}\right\}^{1/2}\left[\frac{c}{n^{m/2}}\right]^{1/2}\\
&=Mc^{1/2}\left(\frac{NK^{-1}}{NK^{-1}+2\lambda}\right)\left\{\frac{K}{\pi_{N*}^3N}+\frac{K}{\pi_{N*}^4N^{1/2+\tau}}\left\{N^{2\tau}\max_{j_1\neq j_2}\sum\triangle_{j_1j_2}^2\right\}^{1/2}\right\}^{1/2}\left[\frac{K}{n^{1/4}}\right]^{m}.
\end{aligned}
\tag{2.66}
$$

As with (2.63), we can bound the expectation of the absolute value of (2.64) using the bound found in (2.66):

$$
\begin{aligned}
&E_p\left|\sum_{m=2}^{\infty}(-1)^m\frac{1}{N}\sum_{j\in U}\tilde{\boldsymbol{I}}_j^T\left(1-\frac{I_j}{\pi_j}\right)\boldsymbol{A}_s^m\frac{1}{NK^{-1}+2\lambda}\sum_{j\in U}y_j\tilde{\boldsymbol{I}}_j\frac{I_j}{\pi_j}\right|\\
&\quad\le\sum_{m=2}^{\infty}(2.66)\\
&=Mc^{1/2}\left(\frac{NK^{-1}}{NK^{-1}+2\lambda}\right)\left\{\frac{K}{\pi_{N*}^3N}+\frac{K}{\pi_{N*}^4N^{1/2+\tau}}\left\{N^{2\tau}\max_{j_1\neq j_2}\sum\triangle_{j_1j_2}^2\right\}^{1/2}\right\}^{1/2}\sum_{m=2}^{\infty}\left[\frac{K}{n^{1/4}}\right]^{m}\\
&=o(1)\left[\frac{1}{1-Kn^{-1/4}}-1-\frac{K}{n^{1/4}}\right]\\
&=o(1).
\end{aligned}
$$

Therefore, we have

$$
E_p\left|\frac{\hat{t}_y^*-\hat{t}_y}{N}\right|=o(1).
$$

□

Chapter 3

Survey-weighted lasso estimator: a model selection and estimation method

3.1 Introduction

In this chapter, we again consider the estimation of the total t_y in the presence of auxiliary information, $\boldsymbol{x}_j$, which is available for each element in the population ($j \in U$). In chapter 2, we only assume the mean function f is smooth in x and can be approximated by a linear combination of penalized B-splines fit on the finite population values. Now, we assume the superpopulation model is parametric and linear where given $\boldsymbol{x}_j$, we have

$$y_j = \boldsymbol{x}_j^T \boldsymbol{\beta} + \epsilon_j. \tag{3.1}$$

Assume the random variables ϵ_j are independent and identically distributed with mean zero and variance σ^2. Additionally, the superpopulation model may be sparse, which means of the p possible covariates, only p_o β's are non-zero where $p_o < p$. Often in survey applications, the number of covariates is large and possibly even greater than the sample size and it is very likely that some covariates do not relate strongly with the study variable. When $p_o < p < n$ but the full model is fit, the design mean squared error of the estimator for t_y may be larger than the design mean squared error of an estimator based on a reduced model

Since our goal is estimation of the finite population quantity t_y we want to select a working model which enables us to more accurately estimate t_y. Therefore, we are interested in constructing a regression estimator whose estimate for f includes model selection. We will explore conditions under which the regression estimator based on model selection outperforms the regression estimator where no model selection occurs.

Frequently, one is interested in estimating several finite population totals and therefore it is advantageous to have one set of regression weights to apply to several study variables. However, since the model selection is based on a regression model for a particular study variable, the corresponding regression estimator weights are also dependent on that study variable. Therefore we are interested in how the model selection regression weights perform when applied to other study variables of interest.

In section 3.1.1 we discuss model selection when the data are independently drawn from an infinite population and specifically look at the lasso method of Tibshirani (1996). We derive a survey-weighted lasso regression estimator when the data are sampled from a finite population, in section 3.1.2. In section 3.2, we present the asymptotic properties of the survey-weighted lasso estimator and then discuss extensions of the estimator in section 3.3. The lasso regression estimator cannot be written as a weighted linear combination of the study variable because the lasso coefficients cannot, in general, be written in closed form. Therefore, lasso regression weights cannot be directly obtained. In section 3.4 we modify the survey-weighted lasso estimator to achieve sampling weights with the caveat that the weights are dependent on the sampled observations. Several model-based lasso estimators are presented in section 3.5 and section 3.6 provides a summary of the estimators discussed in this chapter. We conduct two simulation studies in section 3.7: a study to determine the appropriate criterion for selecting the penalty parameter in the lasso estimators and a study comparing the lasso estimators to other model-assisted and model-based survey estimators. In section 3.8, we estimate the proportion of tree canopy cover for a region in Utah and use the Utah data to conduct a simulation study that compares the performance of the lasso estimators with other model-assisted survey estimators for real data. In section 3.9 we discuss how to use the survey-weighted lasso criterion in a joint design-model framework for analytic inference of survey data.

3.1.1 Background

We first consider model selection for data drawn independently from an infinite population where the regression model is (3.1). Two very widely used discrete methods of model selection are best subsets selection and stepwise selection. The method of best subsets picks a certain number of subset models based on some criterion, such as Mallow's C_p, from the 2^p possible models (Kutner, Nachtsheim, Neter, and Li 2005). A drawback of the best subsets method is that when the number of possible covariates is moderate or large, the method becomes rather computationally infeasible since the number of possible models grows exponentially. Stepwise methods are more computationally efficient than best subsets because instead of considering all possible models, they develop the best model by picking covariates sequentially. For example, the forward stepwise method adds a covariate at each step by selecting the covariate which leads to the largest test statistic (Kutner, Nachtsheim, Neter, and Li 2005). This discrete solution path can lead one to select a model which is locally, but not globally, the best model. The lasso method is a continuous method for model selection that simultaneously performs model selection and parameter estimation by shrinking some coefficients and by forcing other coefficients to be exactly equal to zero (Tibshirani 1996). The lasso method finds coefficients which minimize the sum of squared residuals subject to a constraint on the sum of the absolute value of the coefficients. More specifically, the coefficient estimates for lasso are given by:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\arg\min}\,(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T\,(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) \text{ subject to } \sum_{j=1}^{p} |\beta_j| \leq g \tag{3.2}$$

where the estimate for β_o is not penalized and $g \geq 0$ (Tibshirani 1996). An equivalent solution is given by

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\arg\min}\,(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T\,(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j| \tag{3.3}$$

with $\lambda \geq 0$ since the Lagrangian function of (3.2) is

$$\begin{aligned} L(\boldsymbol{\beta}, \lambda^*) &= (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) - \lambda^* \left(g - \sum_{j=1}^{p} |\beta_j| \right) \\ &= (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda^* \sum_{j=1}^{p} |\beta_j| - \lambda^* g \end{aligned}$$

where $\lambda^* \geq 0$ by the Karush-Kuhn-Tucker conditions. The lasso model selection method is computationally efficient since the solution path is piece-wise linear (Efron, Hastie, Johnstone, and Tibshirani 2004). It selects the global solution since the lasso criterion is convex, which often makes it superior to the best subsets method and the stepwise method. To better understand how the penalty term induces sparsity, we consider the more general penalty term of the bridge estimator, introduced by Frank and Friedman (1993),

$$\lambda \sum_{j=1}^{p} |\beta_j|^\gamma$$

where $\gamma > 0$, $\gamma = 1$ represents the lasso penalty, and $\gamma = 2$ represents the ridge regression penalty. The constrained estimation regions for a regression model with two covariates and for $\gamma = 2, 1,$ and 2^{-1} respectively are displayed in Figure 3.1.1. If the ordinary least squares (OLS) estimator is within or on the constrained estimation region, then the bridge estimator is the OLS estimator. However, if the OLS estimator is outside the region, then the bridge estimator is the point on the constrained estimation region which is touched by the contours

$$(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{(OLS)})^T \boldsymbol{X}^T \boldsymbol{X} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{(OLS)}) \tag{3.4}$$

since the unpenalized criterion can be re-written as a constant term (i.e. a term which does not contain $\boldsymbol{\beta}$) plus (3.4):

$$(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{Y}^T\boldsymbol{Y} - (\hat{\boldsymbol{\beta}}^{(OLS)})^T \boldsymbol{X}^T \boldsymbol{X} \hat{\boldsymbol{\beta}}^{(OLS)} + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{(OLS)})^T \boldsymbol{X}^T \boldsymbol{X} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{(OLS)}).$$

For $\gamma \leq 1$ and when the OLS esimator is outside the constrained estimation region, the contours touch the region on an axis, forcing one coefficient to equal zero, whereas for $\gamma > 1$, this is not the case. Therefore, we can visually see how the lasso and other bridge estimators where $\gamma \leq 1$ produce sparsity in the estimated model. Lasso is more computationally convenient than the bridge estimators with $\gamma < 1$ since the lasso penalty, along with the objective function, are convex.

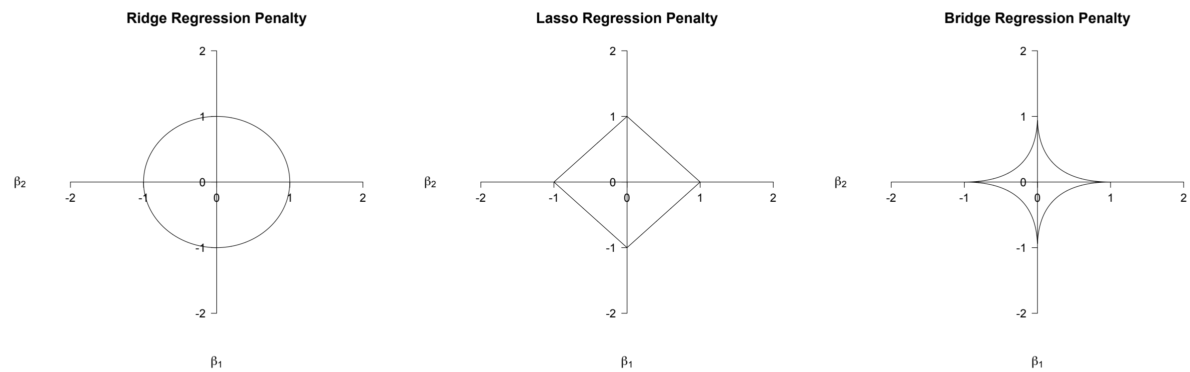

Figure 3.1: Constraint regions for regression model with two covariates

3.1.2 Derivation of survey-weighted lasso and lasso regression estimator

Assume the finite population elements $\{y_j\}_{j \in U}$ are independent realizations from the superpopulation model (3.1), the auxiliary information $\{\boldsymbol{x}_j\}_{j \in U}$ are known and that the sample $\{y_i, \boldsymbol{x}_i\}_{i \in s}$ is obtained according to a measurable sampling design $p(\cdot)$. We can find the first and second order inclusion probabilities as defined in chapter 1, section 1.1. Since the superpopulation model is linear an appropriate regression estimator has the form

$$\hat{t}_{y,lasso} = \sum_{j \in s} \frac{y_j - \boldsymbol{x}_j^T \hat{\boldsymbol{\beta}}_s}{\pi_j} + \sum_{j \in U} \boldsymbol{x}_j^T \hat{\boldsymbol{\beta}}_s. \tag{3.5}$$

To find appropriate sample coefficient estimates, $\hat{\boldsymbol{\beta}}_s$, we should define the finite population quantity, $\boldsymbol{\beta}_U$. Under the model (3.1), an estimate of $\boldsymbol{\beta}$ is the OLS estimator

$$\begin{aligned} \boldsymbol{\beta}_U^{(OLS)} &= \underset{\boldsymbol{\beta}}{\arg\min} \left(\boldsymbol{Y}_U - \boldsymbol{X}_U \boldsymbol{\beta}\right)^T \left(\boldsymbol{Y}_U - \boldsymbol{X}_U \boldsymbol{\beta}\right) \\ &= \left(\boldsymbol{X}_U^T \boldsymbol{X}_U\right)^{-1} \boldsymbol{X}_U^T \boldsymbol{Y}_U \end{aligned} \tag{3.6}$$

where $\boldsymbol{X}_U$ is an $N \times (p+1)$ matrix whose j-th row is the vector $(1, \boldsymbol{x}_j^T)$ and $\boldsymbol{Y}_U = (y_1, y_2, \ldots, y_N)^T$. Although $\boldsymbol{\beta}_U^{(OLS)}$ ignores the sparsity of the model, it is still a good estimate of $\boldsymbol{\beta}$ since it is model unbiased. Also, at the population level, model selection is not as important since N is probably much larger than p and in the descriptive setting, the working model is only a tool used to increase design efficiency. The common survey-weighted estimator of $\boldsymbol{\beta}_U^{(OLS)}$, which under mild assumptions is approximately design unbiased for $\boldsymbol{\beta}_U^{(OLS)}$, is

$$\begin{aligned} \hat{\boldsymbol{\beta}}_s^{(WLS)} &= \underset{\boldsymbol{\beta}}{\arg\min} \left(\boldsymbol{Y}_s - \boldsymbol{X}_s\boldsymbol{\beta}\right)^T \boldsymbol{\Pi}_s^{-1} \left(\boldsymbol{Y}_s - \boldsymbol{X}_s\boldsymbol{\beta}\right) \\ &= \left(\boldsymbol{X}_s^T \boldsymbol{\Pi}_s^{-1} \boldsymbol{X}_s\right)^{-1} \boldsymbol{X}_s^T \boldsymbol{\Pi}_s^{-1} \boldsymbol{Y}_s \end{aligned} \tag{3.7}$$

where $\boldsymbol{X}_s$ is an $n \times (p+1)$ matrix where the j-th row is the vector $(1, \boldsymbol{x}_j^T)$, $\boldsymbol{Y}_s = (y_1, y_2, \ldots, y_n)^T$ and $\boldsymbol{\Pi}_s$ is an $n \times n$ diagonal matrix of the first-order inclusion probabilities for the sampled elements. However, at the sample level, the sparsity of the working model is more important since while N is most likely bigger than p, n may be smaller than p. Also, regardless of whether the model is truly sparse, a reduced model could shrink the overall design variance of the regression estimator, yielding a more efficient estimator. Therefore, we propose estimating $\boldsymbol{\beta}_U^{(OLS)}$ with the following survey-weighted lasso coefficient estimates

$$\hat{\boldsymbol{\beta}}_s^{(L)} = \underset{\boldsymbol{\beta}}{\arg\min} \left(\boldsymbol{Y}_s - \boldsymbol{X}_s\boldsymbol{\beta}\right)^T \boldsymbol{\Pi}_s^{-1} \left(\boldsymbol{Y}_s - \boldsymbol{X}_s\boldsymbol{\beta}\right) \text{ subject to } \sum_{i=1}^{p} |\beta_i| \leq g. \tag{3.8}$$

The survey-weighted lasso coefficient estimates can be found using one of the various algorithms constructed to find (3.3) since we can re-write (3.8) as

$$\hat{\boldsymbol{\beta}}_s^{(L)} = \underset{\boldsymbol{\beta}}{\arg\min} \left(\boldsymbol{Y}_s^* - \boldsymbol{X}_s^*\boldsymbol{\beta}\right)^T \left(\boldsymbol{Y}_s^* - \boldsymbol{X}_s^*\boldsymbol{\beta}\right) \text{ subject to } \sum_{i=1}^{p} |\beta_i| \leq g$$

where $\boldsymbol{Y}_s^* = \boldsymbol{\Pi}_s^{-1/2}\boldsymbol{Y}_s$, $\boldsymbol{X}_s^* = \boldsymbol{\Pi}_s^{-1/2}\boldsymbol{X}_s$ and $\boldsymbol{\Pi}_s^{-1/2} = \operatorname{diag}\left(\pi_k^{-1/2}\right)_{k \in s}$. However, it is important to employ a fitting algorithm which does not require the standardization of the columns of $\boldsymbol{X}_s^*$ since the weighting structure induced by the inverse inclusion probabilities would be lost. In the statistical software package, R (R Development Core Team 2010), we fit the lasso coefficient estimates with the algorithm of Turlach (2005) by using the function `l1ce()` in the package `lasso2` (Lokhorst et al. 2010). Since weight is an argument in the `l1ce()` function, one can either use the original variables $\boldsymbol{Y}_s$ and $\boldsymbol{X}_s$ along with the weight argument or the weighted variables $\boldsymbol{Y}_s^*$ and $\boldsymbol{X}_s^*$ with no weight argument. In both cases, it is important to ensure the intercept term is not weighted.

The estimate for t_y is the regression estimator with (3.8) instead of (3.7) for the coefficient estimates

$$\hat{t}_{y,lasso} = \sum_{j \in s} \frac{y_j - \boldsymbol{x}_j^T \hat{\boldsymbol{\beta}}_s^{(L)}}{\pi_j} + \sum_{j \in U} \boldsymbol{x}_j^T \hat{\boldsymbol{\beta}}_s^{(L)}. \tag{3.9}$$

3.1.3 Selection of the penalty parameter

So far, we have assumed the penalty parameter, g, is a fixed, non-negative number. However, we can also view the penalty parameter as another value to be estimated. Since the goal is estimation of the finite population quantity, t_y, we want to find a criterion for selecting g which leads to a design efficient $\hat{t}_{y,lasso}$. In analytic inference, a useful criterion for selecting g is one which produces coefficient estimates with small mean squared error, and often even more importantly, a useful criterion is one which selects the true, sparse model with a high probability. Although the goals of descriptive and analytic inference differ, we still want to consider some of the criteria used in analytic inference to estimate g in the survey-weighted lasso coefficient estimates since the value of g which yields accurate coefficient estimates should also yield an accurate estimate for the population total. Two common model selection criteria for independently drawn data are the corrected Akaike's Information Criterion (AIC_C) (Hurvich and Tsai 1989) and the Bayesian Information Criterion (BIC) (Schwarz 1978). We want to derive the appropriate AIC_C statistic and the appropriate BIC statistic for survey data.

Assume the true superpopulation model which generated the finite population is (3.1) and denote the true coefficient vector by $\boldsymbol{\beta}_o$ and the true error variance by σ_o^2. Also, assume the errors are normally distributed. Consider working models of the form

$$y_j = \boldsymbol{x}_j^T \boldsymbol{\beta} + \epsilon_j$$

and therefore assume the family of working models contains the true superpopulation model. The Kullback-Leibler information,

$$\triangle(\boldsymbol{\beta}, \sigma^2) = E_\xi \left[-2 \log L(\boldsymbol{\beta}, \sigma^2) \right],$$

measures the accuracy of a particular working model as an approximation to the true model. The Expectation, E_ξ, is taken with respect to the true superpopulation model whereas the likelihood function, $L(\boldsymbol{\beta}, \sigma^2)$, is calculated with respect to the working model. We want to estimate the coefficients, $\boldsymbol{\beta}_o$, with the lasso criterion and stress the dependence on the penalty parameter by writing the estimates as $\boldsymbol{\beta}_U^{(L)}(g)$. To estimate the variance σ_o^2 we use the maximum likelihood estimate

$$\sigma_U^2 = \frac{1}{N} \sum_{j \in U} \left[y_j - \boldsymbol{x}_j^T \boldsymbol{\beta}_U^{(L)}(g) \right]^2.$$

Based on this model fitting procedure, the risk for the working model is

$$E_\xi \triangle \left[\boldsymbol{\beta}_U^{(L)}(g), \sigma_U^2 \right]. \tag{3.10}$$

We want to minimize the risk and therefore we seek the penalty parameter which minimizes (3.10). Hurvich and Tsai (1989) showed

$$E_\xi \triangle \left[\boldsymbol{\beta}_U, \sigma_U^2 \right] \approx E_\xi \left\{ N \log \left[\frac{1}{N} \sum_{j \in U} (y_j - \boldsymbol{x}_j^T \boldsymbol{\beta}_U(\delta))^2 \right] \right\} + \frac{N(\mathrm{df}_U + N)}{N - \mathrm{df}_U - 2} \tag{3.11}$$

when the maximum likelihood procedure is used to estimate both $\boldsymbol{\beta}_o$ and σ_o^2 and df_U is the associated degrees of freedom. Under a few common assumptions (presented in section 3.2), the lasso estimates, $\boldsymbol{\beta}_U^{(L)}(g)$, are consistent for $\boldsymbol{\beta}_o$ and $\boldsymbol{\beta}_U^{(L)}(g) - \boldsymbol{\beta}_o$ approximately follows a multivariate normal distribution (Knight and Fu 2000). Therefore, (3.11) also holds under the lasso procedure. Additionally, Hurvich and Tsai (1989) studied the corrected Akaike's information criterion, which is an approximately unbiased estimate of (3.11). The corrected Akaike's information criterion for the lasso procedure is

$$\mathrm{AIC_{CU}}(g) = N \log\left\{\frac{1}{N}\sum_{j\in U}\left[y_j - \boldsymbol{x}_j^T\boldsymbol{\beta}_U^{(L)}(g)\right]^2\right\} + \frac{N(\hat{\mathrm{df}}_U(g)+N)}{N-\hat{\mathrm{df}}_U(g)-2}. \tag{3.12}$$

Efron (2004) defines the degrees of freedom of any model fitting procedure to be

$$\mathrm{df}_U = \sum_{j\in U}\frac{cov(\hat{y}_j, y_j)}{\sigma^2}.$$

For the maximum likelihood procedure, the degrees of freedom equals the number of coefficients since

$$\mathrm{df}_U = \sum_{j\in U}\frac{\mathrm{cov}(\boldsymbol{x}_j^T(\boldsymbol{X}_U^T\boldsymbol{X}_U)^{-1}\boldsymbol{X}_U^T\boldsymbol{y}, y_j)}{\sigma^2} = \sum_{j\in U}\frac{h_{jj}\sigma^2}{\sigma^2} = p+1$$

where h_{jj} is the j-th diagonal element of the Hat matrix. For the lasso fitting procedure, the fitted values do not have a closed form and therefore, the degrees of freedom cannot be found analytically. Zou, Hastie, and Tibshirani (2007) proved the estimate $\hat{\mathrm{df}}_U(g) = p_L$, where p_L is the number of non-zero lasso coefficients, is an unbiased estimate for the degrees of freedom.

We want to minimize the risk to determine the optimal penalty parameter. Since we cannot compute the risk or its population level estimator, $\mathrm{AIC_{CU}}(g)$, we seek an approximately design unbiased estimator for $\mathrm{AIC_{CU}}(g)$. A sample-based corrected Akaike's information criterion ($\mathrm{AIC_{Cs}}(g)$), which accurately estimates the population-based corrected Akaike's informaton criterion, can be viewed as a reasonable estimator of the risk. A potential sample-based corrected Akaike's information criterion is

$$\mathrm{AIC_{Cs}}(g)^* = N \log\left\{\frac{1}{N}\sum_{j\in s}\frac{1}{\pi_j}\left[y_j - \boldsymbol{x}_j^T\hat{\boldsymbol{\beta}}_s^{(L)}(g)\right]^2\right\} + \frac{N(\hat{\mathrm{df}}_s(g)+N)}{N-\hat{\mathrm{df}}_s(g)-2}. \tag{3.13}$$

The estimated degrees of freedom, $\hat{\mathrm{df}}_s(g)$, equals the number of non-zero entries in $\hat{\boldsymbol{\beta}}_s^{(L)}(g)$.

We can heuristically argue what assumptions are required for the $\text{AIC}_{\text{Cs}}(g)^*$ to be approximately unbiased for the $\text{AIC}_{\text{CU}}(g)$ by looking at the difference between the criteria:

$$\begin{aligned}
&\text{AIC}_{\text{Cs}}(g)^* - \text{AIC}_{\text{CU}}(g) \\
&= N \log \left\{ \frac{N^{-1} \sum_{j \in s} \pi_j^{-1} [y_j - \boldsymbol{x}_j^T \hat{\boldsymbol{\beta}}_s^{(L)}(g)]^2}{N^{-1} \sum_{j \in U} [y_j - \boldsymbol{x}_j^T \boldsymbol{\beta}_U^{(L)}(g)]^2} \right\} + \left[\frac{N(\hat{\text{df}}_s(g) + N)}{N - \hat{\text{df}}_s(g) - 2} - \frac{N(\hat{\text{df}}_U(g) + N)}{N - \hat{\text{df}}_U(g) - 2} \right].
\end{aligned}$$

As long as $\hat{\boldsymbol{\beta}}_s^{(L)}$ is design consistent for $\boldsymbol{\beta}_U^{(L)}$, the ratio in the first term will go to 1 in probability under reasonable assumptions because the numerator is almost a Horvitz-Thompson estimator of the denominator. Also, if the sample-based lasso model is variable selection consistent for the population-based lasso model, the second term will go to 0 in probability.

However, in simulations $\text{AIC}_{\text{Cs}}(g)^*$ performed poorly at selecting the correct model and instead prefered larger models than the true model. It appears to over-penalize the residual term while under-penalizing the degrees of freedom term. Therefore, we prefer a different sample-based corrected Akaike's information criterion:

$$\text{AIC}_{\text{Cs}}(g) = n \log \left\{ \frac{1}{N} \sum_{j \in s} \frac{1}{\pi_j} \left[y_j - \boldsymbol{x}_j^T \hat{\boldsymbol{\beta}}_s^{(L)}(g) \right]^2 \right\} + \frac{n(\hat{\text{df}}_s(g) + n)}{n - \hat{\text{df}}_s(g) - 2}. \tag{3.14}$$

The penalty on model complexity is larger for $\text{AIC}_{\text{Cs}}(g)$ than it is for $\text{AIC}_{\text{Cs}}(g)^*$ and therefore, it selects smaller models. $\text{AIC}_{\text{Cs}}(g)$ performs well in simulations (presented in section 3.7.1). The formal justification of $\text{AIC}_{\text{Cs}}(g)$ as a suitable model selection criteria for survey data should be studied in further detail.

Another common model selection criterion is the Bayesian Information Criterion (BIC) (Schwarz 1978), which is based on the asymptotic Bayes solution for a particular model and fitting procedure. The $\text{BIC}_{\text{U}}(g)$ and $\text{AIC}_{\text{CU}}(g)$ are similar under the model (3.1) with normal errors but have slightly different penalties for model complexity. In particular, $\text{BIC}_{\text{U}}(g)$ tends to favor more parsimonious models than $\text{AIC}_{\text{CU}}(g)$. The finite population $\text{BIC}_{\text{U}}(g)$ criterion for the lasso method is

$$\text{BIC}_{\text{U}}(g) = N \log \left[\frac{1}{N} \sum_{j \in U} (y_j - \boldsymbol{x}_j^T \boldsymbol{\beta}_U(g))^2 \right] + N + \log(N) \left[\text{df}_U(g) + 1 \right].$$

The sample-based estimate for $\mathrm{BIC_U}(g)$ is

$$\mathrm{BIC_s}(g) = n \log \left[\frac{1}{N} \sum_{j \in s} \frac{1}{\pi_j} (y_j - \boldsymbol{x}_j^T \hat{\boldsymbol{\beta}}_s^{(L)}(g))^2 \right] + n + \log(n) \left[\hat{\mathrm{df}}_s(g) + 1 \right]. \tag{3.15}$$

The information statistics attempt to find the optimal model by balancing the bias and variance of the model. However, we are interested in balancing the design-based bias and variance of the estimator of the finite population total by minimizing its design-based mean squared error. Therefore, we also consider the following design-based criterion, proposed by Opsomer and Miller (2005), which accounts for the sampling design and our desired goal of a design efficient estimator,

$$\hat{V}_{CV}(g) = \sum\sum_{i,j \in s} \frac{\triangle_{ij}}{\pi_{ij}} \frac{(y_i - \hat{f}_s(\boldsymbol{x}_i, g)^{(-)})}{\pi_i} \frac{(y_j - \hat{f}_s(\boldsymbol{x}_j, g)^{(-)})}{\pi_j} \tag{3.16}$$

where $\hat{f}_s(\boldsymbol{x}_i, g)^{(-)}$ is the leave-one-out model fit for the i-th observation. Opsomer and Miller (2005) show that (3.16) works well for selecting the bandwidth for the non-parametric model-assisted estimator based on local polynomial regression. In section 3.7.1 we look at how these criteria for selecting the penalty parameter ($\mathrm{AIC_{Cs}}(g)$, $\mathrm{BIC_s}(g)$, and $\hat{V}_{CV}(g)$) compare. We look at both the accuracy of the models and the design efficiency of the resulting estimators. In section 3.5 we present the model-based $\mathrm{AIC_C}$ and the model-based BIC.

3.2 Main results

In section 3.2.1, we list the necessary design assumptions for the survey-weighted lasso regression estimator and in section 3.2.2, we present its asymptotic properties. Theorem 3.1 is a central limit theorem result for the survey-weighted lasso coefficients as estimates of the finite population coefficients and Corollary 3.1 provides the root-N consistency of the survey-weighted lasso coefficients for the finite population coefficients. Throughout this section, the survey-weighted lasso coefficients, $\hat{\boldsymbol{\beta}}_s^{(L)}$, are denoted by $\hat{\boldsymbol{\beta}}_N$ and the finite population coefficients, $\boldsymbol{\beta}_U^{(OLS)}$, are denoted by $\boldsymbol{\beta}_N$ to simplify the notation and to emphasize the dependence on N as we look at the asymptotic properties of these quantities. To find the survey-weighted lasso coefficients, we use the survey-weighted residual sum of squares plus an l_1 penalty on the coefficients:

$$\hat{\boldsymbol{\beta}}_N = \underset{\boldsymbol{\beta}}{\arg\min} \left(\boldsymbol{Y}_s - \boldsymbol{X}_s\boldsymbol{\beta}\right)^T \boldsymbol{\Pi}_s^{-1} \left(\boldsymbol{Y}_s - \boldsymbol{X}_s\boldsymbol{\beta}\right) + \lambda_N \sum_{i=1}^{p} |\beta_i|$$

instead of the equivalent criterion given in (3.8).

Theorem 3.2 is a central limit theorem result for the lasso regression estimator and Corollary 3.2 provides the root-n consistency of the lasso regression estimator for the population total. After showing the usual variance estimator is design consistent in Theorem 3.3, Corollary 3.3 is another central limit theorem result which is immediate from Theorem 3.2 and Theorem 3.3.

In section 3.2.3, we consider the situation where the survey-weighted lasso coefficients are estimates for superpopulation coefficients and derive the properties of the coefficient estimates under a joint design-model framework. In both Theorem 3.4 and Theorem 3.5, we combine standard regression theory with the results of section 3.2.2 to obtain consistency and a central limit theorem result.

3.2.1 Design assumptions

The following assumptions pertain to the sampling design and the asymptotic behavior of both sample and finite population quantities. We follow the asymptotic framework presented in chapter 2, section 2.2.1.

D1. The penalty parameter is allowed to increase as N increases but only at the rate $\lambda_N = o(\sqrt{N})$.

D2. As $N \to \infty$, assume $nN^{-1} \to \pi \in (0, 1)$.

D3. Define the survey-weighted matrix of the covariates

$$\hat{\boldsymbol{C}}_N = \frac{1}{N} \sum_{i \in U_N} \boldsymbol{x}_i \boldsymbol{x}_i^T \frac{I_i}{\pi_i} \tag{3.17}$$

and the corresponding finite population matrix

$$\boldsymbol{C}_N = \frac{1}{N} \sum_{i \in U_N} \boldsymbol{x}_i \boldsymbol{x}_i^T. \tag{3.18}$$

Assume both $\hat{\boldsymbol{C}}_N$ and $\boldsymbol{C}_N$ are positive definite, $\hat{\boldsymbol{C}}_N - \boldsymbol{C}_N = o_p(1)$ elementwise and $\boldsymbol{C}_N - \boldsymbol{C} = o(1)$ elementwise where $\boldsymbol{C}$ is a non-singular matrix. Assume $\boldsymbol{D}_N - \boldsymbol{D} = o(1)$ elementwise for some $\boldsymbol{D} \in \mathbb{R}^{p+1}$ where

$$\boldsymbol{D}_N = \frac{1}{N} \sum_{i \in U_N} \boldsymbol{x}_i y_i.$$

Define $\boldsymbol{\beta}^* = \boldsymbol{C}^{-1}\boldsymbol{D}$.

D4. Define the following $(p+2)(p+1)$ vector of centered, standardized Horvitz-Thompson estimators:

$$\boldsymbol{z}_N = \begin{bmatrix} \frac{\sqrt{n}}{N} \sum_{i \in U_N} \boldsymbol{x}_i y_i \left(\frac{I_i}{\pi_i} - 1 \right) \\ \left\{ \frac{\sqrt{n}}{N} \sum_{i \in U_N} \boldsymbol{x}_i x_{ik} \left(\frac{I_i}{\pi_i} - 1 \right) \right\}_{k=0}^{p} \end{bmatrix} \tag{3.19}$$

where $\boldsymbol{x}_i = (x_{i0}, x_{i1}, \ldots, x_{ip})^T$ and $x_{i0} = 1$. The finite population covariance matrix for $\boldsymbol{z}_N$ is

$$\begin{aligned} \boldsymbol{\Sigma}_N &= \begin{bmatrix} \boldsymbol{\Sigma}_N^{(xyxy)} & \boldsymbol{\Sigma}_N^{(xyxx_o)} & \cdots & \boldsymbol{\Sigma}_N^{(xyxx_p)} \\ \boldsymbol{\Sigma}_N^{(xx_oxy)} & \boldsymbol{\Sigma}_N^{(xx_oxx_o)} & \vdots & \\ \vdots & \vdots & & \\ \boldsymbol{\Sigma}_N^{(xx_pxy)} & \boldsymbol{\Sigma}_N^{(xx_pxx_o)} & \cdots & \boldsymbol{\Sigma}_N^{(xx_pxx_p)} \end{bmatrix} \\ &= \begin{bmatrix} \frac{n}{N^2} \sum\sum_{i,j \in U_N} \frac{\triangle_{ij}}{\pi_i \pi_j} \boldsymbol{x}_i y_i \boldsymbol{x}_j^T y_j & \cdots & \frac{n}{N^2} \sum\sum_{i,j \in U_N} \frac{\triangle_{ij}}{\pi_i \pi_j} \boldsymbol{x}_i y_i \boldsymbol{x}_j^T x_{ip} \\ \vdots & & \vdots \\ \frac{n}{N^2} \sum\sum_{i,j \in U_N} \frac{\triangle_{ij}}{\pi_i \pi_j} \boldsymbol{x}_i x_{ip} \boldsymbol{x}_j^T y_j & \cdots & \frac{n}{N^2} \sum\sum_{i,j \in U_N} \frac{\triangle_{ij}}{\pi_i \pi_j} \boldsymbol{x}_i x_{ip} \boldsymbol{x}_j^T x_{ip} \end{bmatrix}. \end{aligned}$$

The limit of the normalized finite population covariance matrix of $\boldsymbol{z}_N$,

$$\boldsymbol{\Sigma} = \lim_{N\to\infty} \boldsymbol{\Sigma}_N, \tag{3.20}$$

is positive definite.

D5. Assume the following central limit theorem for the normalized, centered, Horvitz-Thompson estimators defined in (3.19):

$$\boldsymbol{z}_N \xrightarrow{D} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}).$$

D6. For the vector $\boldsymbol{z}_N^* = (z_{N1}, z_{Np+2}, z_{Np+3}, \dots, z_{N2p+3})$ which is a subset of the vector $\boldsymbol{z}_N$ defined in (3.19), an estimate of the covariance matrix is

$$\begin{aligned}\hat{\boldsymbol{\Sigma}}_N^* &= \begin{bmatrix} \hat{\Sigma}_N^{(x_oyx_oy)} & \hat{\boldsymbol{\Sigma}}_N^{(x_oyxx_o)} \\ \hat{\boldsymbol{\Sigma}}_N^{(xx_ox_oy)} & \hat{\boldsymbol{\Sigma}}_N^{(xx_oxx_o)} \end{bmatrix} \\ &= \begin{bmatrix} \frac{n}{N^2} \sum\sum_{i,j\in U_N} \frac{\triangle_{ij}}{\pi_i\pi_j} \frac{I_iI_j}{\pi_{ij}} y_iy_j & \frac{n}{N^2} \sum\sum_{i,j\in U_N} \frac{\triangle_{ij}}{\pi_i\pi_j} \frac{I_iI_j}{\pi_{ij}} y_i\boldsymbol{x}_j^T \\ \frac{n}{N^2} \sum\sum_{i,j\in U_N} \frac{\triangle_{ij}}{\pi_i\pi_j} \frac{I_iI_j}{\pi_{ij}} \boldsymbol{x}_iy_j & \frac{n}{N^2} \sum\sum_{i,j\in U_N} \frac{\triangle_{ij}}{\pi_i\pi_j} \frac{I_iI_j}{\pi_{ij}} \boldsymbol{x}_i\boldsymbol{x}_j^T \end{bmatrix}.\end{aligned}$$

Assume $n\left(\hat{\boldsymbol{\Sigma}}_N^* - \boldsymbol{\Sigma}_N^*\right) = o_p(1)$ elementwise where

$$\boldsymbol{\Sigma}_N^* = \begin{bmatrix} \Sigma_N^{(x_oyx_oy)} & \boldsymbol{\Sigma}_N^{(x_oyxx_o)} \\ \boldsymbol{\Sigma}_N^{(xx_ox_oy)} & \boldsymbol{\Sigma}_N^{(xx_oxx_o)} \end{bmatrix}.$$

For ease of notation write,

$$\hat{\boldsymbol{\Sigma}}_N^* = \begin{bmatrix} \hat{\Sigma}_N^{(yy)} & \hat{\boldsymbol{\Sigma}}_N^{(yx)} \\ \hat{\boldsymbol{\Sigma}}_N^{(xy)} & \hat{\boldsymbol{\Sigma}}_N^{(xx)} \end{bmatrix} \text{ and } \boldsymbol{\Sigma}_N^* = \begin{bmatrix} \Sigma_N^{(yy)} & \boldsymbol{\Sigma}_N^{(yx)} \\ \boldsymbol{\Sigma}_N^{(xy)} & \boldsymbol{\Sigma}_N^{(xx)} \end{bmatrix}.$$

Remark 1. The sample size n should be written as n_N since it grows asymptotically. Both the sample size and the population size can be used fairly interchangeably as normalizers since assumption (D2) requires they grow at the same rate.

Remark 2. Assumption (D3) ensures that the finite population parameter $\boldsymbol{\beta}_N$ converges to the vector $\boldsymbol{\beta}^* \in \mathbb{R}^{p+1}$. It does not necessary to converge to $\boldsymbol{\beta}$, the true coefficient vector.

Remark 3. Assumption (D5) allow us to obtain central limit theorem results for the lasso coefficients and regression estimator. Without these assumptions, we would have to restrict our attention to with replacement sampling, simple random sampling or to other special cases with known central limit theorems.

Remark 4. Since $\boldsymbol{x}_i$ contains an intercept term, assumption (D4) covers $\lim_{N\to\infty} \Sigma_N^{(x_o y x_o y)} = \Sigma^{(x_o y x_o y)}$ where

$$\Sigma_N^{(x_o y x_o y)} = \Sigma_N^{(yy)} = \frac{n}{N^2} \sum\sum_{i,j \in U_N} \frac{\triangle_{ij}}{\pi_i \pi_j} y_i y_j$$

and similarly $\lim_{N\to\infty} \boldsymbol{\Sigma}_N^{(x x_o x x_o)} = \boldsymbol{\Sigma}^{(x x_o x x_o)}$ where

$$\boldsymbol{\Sigma}_N^{(x x_o x x_o)} = \boldsymbol{\Sigma}_N^{(xx)} = \frac{n}{N^2} \sum\sum_{i,j \in U_N} \frac{\triangle_{ij}}{\pi_i \pi_j} \boldsymbol{x}_i \boldsymbol{x}_j^T.$$

3.2.2 Design-based asymptotic results

The design based results for both the survey-weighted lasso coefficients and subsequently the survey-weighted lasso regression estimator are given in this section. Under suitable conditions, the asymptotic distribution of the survey-weighted lasso regression estimator is the same as the asymptotic distribution of the full regression estimator.

Theorem 3.1. *Under assumptions (D1) – (D5),*

$$\sqrt{N}\left(\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}_N\right) \xrightarrow{D} \mathcal{N}\left(\mathbf{0}, \pi^{-1} \boldsymbol{C}^{-1} \boldsymbol{V} \boldsymbol{C}^{-1}\right) \tag{3.21}$$

where the matrix $\boldsymbol{V}$ is defined by

$$\boldsymbol{V} = \boldsymbol{\Sigma}^{(xyxy)} - 2\sum_{k=0}^{p} \beta_k^* \boldsymbol{\Sigma}^{(x x_k x y)} + \sum_{k=0}^{p}\sum_{l=0}^{p} \beta_k^* \beta_l^* \boldsymbol{\Sigma}^{(x x_k x x_l)}$$

and the terms in $\boldsymbol{V}$ are components of the limit in (3.20).

Proof. First, define the survey-weighted lasso criterion as

$$\boldsymbol{A}_{N}(\boldsymbol{u}) := (\boldsymbol{Y}_s - \boldsymbol{X}_s\boldsymbol{u})^T \boldsymbol{\Pi}_s^{-1} (\boldsymbol{Y}_s - \boldsymbol{X}_s\boldsymbol{u}) + \lambda_N \sum_{j=1}^{p} |u_j| \tag{3.22}$$

where $\hat{\boldsymbol{\beta}}_N = \arg\min_{\boldsymbol{u}} \boldsymbol{A}_N(\boldsymbol{u})$ and let

$$\boldsymbol{B}_N(\boldsymbol{u}) := \boldsymbol{A}_N\left(\boldsymbol{\beta}_N - \boldsymbol{u}\frac{\sqrt{n}}{N}\right) - \boldsymbol{A}_N(\boldsymbol{\beta}_N). \tag{3.23}$$

Notice the minimum of $\boldsymbol{B}_N(\boldsymbol{u})$ is $Nn^{-1/2}(\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}_N)$. Since $\boldsymbol{\beta}_N$ is the OLS estimator, we have

$$-\frac{2\sqrt{n}}{N}\sum_{i\in U_N} \boldsymbol{x}_i (y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}_N) = \mathbf{0}.$$

Now write

$$\boldsymbol{B}_N(\boldsymbol{u}) = \left(\boldsymbol{Y}_s - \boldsymbol{X}_s\left[\boldsymbol{\beta}_N + \frac{\boldsymbol{u}\sqrt{n}}{N}\right]\right)^T \boldsymbol{\Pi}_s^{-1}\left(\boldsymbol{Y}_s - \boldsymbol{X}_s\left[\boldsymbol{\beta}_N + \frac{\boldsymbol{u}\sqrt{n}}{N}\right]\right) + \lambda_N \sum_{j=1}^{p}\left|\beta_{Nj} + \frac{u_j\sqrt{n}}{N}\right| \tag{3.24}$$

$$\begin{aligned}
&- (\boldsymbol{Y}_s - \boldsymbol{X}_s\boldsymbol{\beta}_N)^T \boldsymbol{\Pi}_s^{-1} (\boldsymbol{Y}_s - \boldsymbol{X}_s\boldsymbol{\beta}_N) - \lambda_N \sum_{j=1}^{p} |\beta_{Nj}| \\
=& -\boldsymbol{u}^T\left[\frac{2\sqrt{n}}{N}\sum_{i\in U_N}\boldsymbol{x}_i(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}_N)\frac{I_i}{\pi_i}\right] + \boldsymbol{u}^T\left[\frac{n}{N^2}\sum_{i\in U_N}\frac{I_i}{\pi_i}\boldsymbol{x}_i\boldsymbol{x}_i^T\right]\boldsymbol{u} \\
&+ \lambda_N\sum_{j=1}^{p}\left|\beta_{Nj} + \frac{u_j\sqrt{n}}{N}\right| - |\beta_{Nj}| \qquad (3.25) \\
=& -\boldsymbol{u}^T\left[\frac{2\sqrt{n}}{N}\sum_{i\in U_N}\boldsymbol{x}_i(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}_N)\left(\frac{I_i}{\pi_i} - 1\right)\right] + \boldsymbol{u}^T\left[\frac{n}{N^2}\sum_{i\in U_N}\frac{I_i}{\pi_i}\boldsymbol{x}_i\boldsymbol{x}_i^T\right]\boldsymbol{u} \\
&+ \lambda_N\sum_{j=1}^{p}\left|\beta_{Nj} + \frac{u_j\sqrt{n}}{N}\right| - |\beta_{Nj}|. \qquad (3.26)
\end{aligned}$$

Assumptions (D3) and (D5) imply

$$\frac{\sqrt{n}}{N}\sum_{i\in U_N}\boldsymbol{x}_i(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}_N)\left(\frac{I_i}{\pi_i} - 1\right) \xrightarrow{D} \mathcal{N}(\mathbf{0}, \boldsymbol{V})$$

since the variance of the centered Horvitz-Thompson estimator can be written

$$
\begin{aligned}
Var_p\left(\frac{\sqrt{n}}{N}\sum_{i\in U_N}\boldsymbol{x}_i(y_i-\boldsymbol{x}_i^T\boldsymbol{\beta}_N)\frac{I_i}{\pi_i}\right) &= \frac{n}{N^2}\sum\sum_{i,j\in U_N}\triangle_{ij}\frac{\boldsymbol{x}_i(y_i-\boldsymbol{x}_i^T\boldsymbol{\beta}_N)}{\pi_i}\frac{\boldsymbol{x}_j^T(y_j-\boldsymbol{x}_j^T\boldsymbol{\beta}_N)}{\pi_j} \\
&= \boldsymbol{\Sigma}_N^{(xyxy)} - 2\sum_{k=0}^{p}\beta_{Nk}\boldsymbol{\Sigma}_N^{(xx_kxy)} + \sum_{k=0}^{p}\sum_{l=0}^{p}\beta_{Nk}\beta_{Nl}\boldsymbol{\Sigma}_N^{(xx_kxx_l)} \\
&:= \boldsymbol{V}_N
\end{aligned}
$$

and by assumptions (D3) and (D4), $\boldsymbol{V} = \lim_{N\to\infty}\boldsymbol{V}_N$. This implies the first component in (3.26) converges in distribution. The second term converges in probability to the quadratic term $\boldsymbol{u}^T\pi C\boldsymbol{u}$ by assumptions (D2) and (D3). Also, we have

$$
\lambda_N\sum_{j=1}^{p}\left|\beta_{Nj}+\frac{u_j\sqrt{n}}{N}\right| - |\beta_{Nj}| = o_p(1) \tag{3.27}
$$

for each $\boldsymbol{u}\in\mathbb{R}^{p+1}$ since

$$
\begin{aligned}
\left|\lambda_N\sum_{j=1}^{p}\left|\beta_{Nj}+\frac{u_j\sqrt{n}}{N}\right| - |\beta_{Nj}|\right| &\le \lambda_N\sum_{j=1}^{p}\left|\left|\beta_{Nj}+\frac{u_j\sqrt{n}}{N}\right| - |\beta_{Nj}|\right| \\
&\le \lambda_N\sum_{j=1}^{p}\left|\frac{u_j\sqrt{n}}{N}\right| \\
&= \frac{\lambda_N}{\sqrt{N}}\frac{\sqrt{n}}{\sqrt{N}}||\boldsymbol{u}||_{L1} \\
&= o_p(1)
\end{aligned}
$$

by assumption (D1). Therefore, we can apply the Corollary in section 1 of Hjort and Pollard (1993) to $\boldsymbol{B}_N(\boldsymbol{u})$ and obtain

$$
Nn^{-1/2}(\hat{\boldsymbol{\beta}}_N-\boldsymbol{\beta}_N)\xrightarrow{D}\mathcal{N}\left(\mathbf{0},\pi^{-2}\boldsymbol{C}^{-1}\boldsymbol{V}\boldsymbol{C}^{-1}\right).
$$

Therefore,

$$
N^{1/2}(\hat{\boldsymbol{\beta}}_N-\boldsymbol{\beta}_N)\xrightarrow{D}\sqrt{\pi}\mathcal{N}\left(\mathbf{0},\pi^{-2}\boldsymbol{C}^{-1}\boldsymbol{V}\boldsymbol{C}^{-1}\right)=\mathcal{N}\left(\mathbf{0},\pi^{-1}\boldsymbol{C}^{-1}\boldsymbol{V}\boldsymbol{C}^{-1}\right).
$$

$\square$

The following Corollary is an immediate consequence of Theorem 3.1.

Corollary 3.1. *Under assumptions (D1) – (D5),* $\hat{\boldsymbol{\beta}}_N$ *is root-N design consistent for* $\boldsymbol{\beta}_N$ *in the sense that* $\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}_N = O_p(N^{-1/2})$.

The next results are the asymptotic design properties of the survey-weighted lasso regression estimator.

Theorem 3.2. *Under assumptions (D1) – (D5),*

$$\left\{var_p(\hat{t}_{y,diff})\right\}^{-1/2}(\hat{t}_y - t_y) \xrightarrow{D} N(0,1).$$

Proof. First, by assumptions (D3) the finite population parameter vector, $\boldsymbol{\beta}_N$, converges to the vector $\boldsymbol{\beta}^* \in \mathbb{R}^{p+1}$ elementwise since

$$\begin{aligned}
\boldsymbol{\beta}_N &= \left(\frac{1}{N}\sum_{i\in U_N} \boldsymbol{x}_i\boldsymbol{x}_i^T\right)^{-1} \frac{1}{N}\sum_{i\in U_N} \boldsymbol{x}_i y_i \\
&\overset{N\to\infty}{\longrightarrow} C^{-1}D \\
&= \boldsymbol{\beta}^*.
\end{aligned}$$

Therefore, $\hat{\boldsymbol{\beta}}_N$ converges in probability to $\boldsymbol{\beta}^*$ since

$$\begin{aligned}
\left|\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}^*\right| &\le \left|\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}_N\right| + \left|\boldsymbol{\beta}_N - \boldsymbol{\beta}^*\right| \\
&= O_p(N^{-1/2}) + o(1) \\
&= o_p(1).
\end{aligned}$$

From assumption (D5), we have

$$\begin{bmatrix} z_{N1} \\ z_{Np+2} \\ z_{Np+3} \\ \vdots \\ z_{N2p+3} \end{bmatrix} = \begin{bmatrix} \frac{\sqrt{n}}{N} \sum_{i \in U_N} y_i \left(\frac{I_i}{\pi_i} - 1 \right) \\ \frac{\sqrt{n}}{N} \sum_{i \in U_N} \boldsymbol{x}_i \left(\frac{I_i}{\pi_i} - 1 \right) \end{bmatrix} \xrightarrow{D} \mathcal{N} \left(\mathbf{0}, \begin{bmatrix} \Sigma^{(x_o y x_o y)} & \boldsymbol{\Sigma}^{(x_o y x x_o)} \\ \boldsymbol{\Sigma}^{(x x_o x_o y)} & \boldsymbol{\Sigma}^{(x x_o x x_o)} \end{bmatrix} \right) \tag{3.28}$$

and for ease of notation, let

$$\begin{bmatrix} \Sigma^{(x_o y x_o y)} & \boldsymbol{\Sigma}^{(x_o y x x_o)} \\ \boldsymbol{\Sigma}^{(x x_o x_o y)} & \boldsymbol{\Sigma}^{(x x_o x x_o)} \end{bmatrix} := \begin{bmatrix} \Sigma^{(yy)} & \boldsymbol{\Sigma}^{(yx)} \\ \boldsymbol{\Sigma}^{(xy)} & \boldsymbol{\Sigma}^{(xx)} \end{bmatrix}.$$

Define the function $g(\cdot, \cdot)$ such that $g(\boldsymbol{a}, \boldsymbol{b}) = (a_1, \boldsymbol{a}_2^T \boldsymbol{b})$. Since $\hat{\boldsymbol{\beta}}_N$ converges in probability to $\boldsymbol{\beta}^*$ we can apply Slutsky's Theorem to (3.28) and obtain

$$\begin{bmatrix} \frac{\sqrt{n}}{N} \sum_{i \in U_N} y_i \left(\frac{I_i}{\pi_i} - 1 \right) \\ \frac{\sqrt{n}}{N} \sum_{i \in U_N} \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_N \left(\frac{I_i}{\pi_i} - 1 \right) \end{bmatrix} \xrightarrow{D} \mathcal{N} \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Sigma^{(yy)} & \boldsymbol{\Sigma}^{(yx)} \boldsymbol{\beta}^* \\ \boldsymbol{\beta}^{*T} \boldsymbol{\Sigma}^{(xy)} & \boldsymbol{\beta}^{*T} \boldsymbol{\Sigma}^{(xx)} \boldsymbol{\beta}^* \end{bmatrix} \right).$$

Now, define the function $h(\cdot, \cdot)$ such that $h(a_1, a_2) = a_1 - a_2$. The Jacobian of $h(a_1, a_2)$ is $\boldsymbol{J}_{h(a_1, a_2)} = (1, -1)$ and

$$\boldsymbol{J}_{h(0,0)} \begin{bmatrix} \Sigma^{(yy)} & \boldsymbol{\Sigma}^{(yx)} \boldsymbol{\beta}^* \\ \boldsymbol{\beta}^{*T} \boldsymbol{\Sigma}^{(xy)} & \boldsymbol{\beta}^{*T} \boldsymbol{\Sigma}^{(xx)} \boldsymbol{\beta}^* \end{bmatrix} \boldsymbol{J}_{h(0,0)}^T = \Sigma^{(yy)} - \boldsymbol{\Sigma}^{(yx)} \boldsymbol{\beta}^* - \boldsymbol{\beta}^{*T} \boldsymbol{\Sigma}^{(xy)} + \boldsymbol{\beta}^{*T} \boldsymbol{\Sigma}^{(xx)} \boldsymbol{\beta}^*.$$

By the Delta Method, we have

$$\frac{\sqrt{n}}{N} \left\{ \Sigma^{(yy)} - \boldsymbol{\Sigma}^{(yx)} \boldsymbol{\beta}^* - \boldsymbol{\beta}^{*T} \boldsymbol{\Sigma}^{(xy)} + \boldsymbol{\beta}^{*T} \boldsymbol{\Sigma}^{(xx)} \boldsymbol{\beta}^* \right\}^{-1/2} (\hat{t}_y - t_y) \xrightarrow{D} N(0, 1).$$

We can write the variance of the difference estimator as

$$var_p(\hat{t}_{y,diff}) = \frac{N^2}{n} \left\{ \Sigma_N^{(yy)} - \boldsymbol{\Sigma}_N^{(yx)} \boldsymbol{\beta}_N - \boldsymbol{\beta}_N^T \boldsymbol{\Sigma}_N^{(xy)} + \boldsymbol{\beta}_N^T \boldsymbol{\Sigma}_N^{(xx)} \boldsymbol{\beta}_N \right\} \tag{3.29}$$

and therefore since $\boldsymbol{\beta}_N$ converges to $\boldsymbol{\beta}^*$ and by (3.20), we have

$$\left\{\Sigma_N^{(yy)} - \boldsymbol{\Sigma}_N^{(yx)}\boldsymbol{\beta}_N - \boldsymbol{\beta}_N^T\boldsymbol{\Sigma}_N^{(xy)} + \boldsymbol{\beta}_N^T\boldsymbol{\Sigma}_N^{(xx)}\boldsymbol{\beta}_N\right\} - \left\{\Sigma^{(yy)} - \boldsymbol{\Sigma}^{(yx)}\boldsymbol{\beta}^* - \boldsymbol{\beta}^{*T}\boldsymbol{\Sigma}^{(xy)} + \boldsymbol{\beta}^{*T}\boldsymbol{\Sigma}^{(xx)}\boldsymbol{\beta}^*\right\} = o(1)$$
$$\Rightarrow \frac{n}{N^2}var_p(\hat{t}_{y,diff}) = \Sigma^{(yy)} - \boldsymbol{\Sigma}^{(yx)}\boldsymbol{\beta}^* - \boldsymbol{\beta}^{*T}\boldsymbol{\Sigma}^{(xy)} + \boldsymbol{\beta}^{*T}\boldsymbol{\Sigma}^{(xx)}\boldsymbol{\beta}^* + o(1).$$

This gives us

$$\begin{aligned}
&\frac{\sqrt{n}}{N}\left\{\Sigma^{(yy)} - \boldsymbol{\Sigma}^{(yx)}\boldsymbol{\beta}^* - \boldsymbol{\beta}^{*T}\boldsymbol{\Sigma}^{(xy)} + \boldsymbol{\beta}^{*T}\boldsymbol{\Sigma}^{(xx)}\boldsymbol{\beta}^*\right\}^{-1/2}(\hat{t}_y - t_y)\\
=&\frac{\sqrt{n}}{N}\left\{\frac{n}{N^2}var_p(\hat{t}_{y,diff})\right\}^{-1/2}(\hat{t}_y - t_y) + o(1)O_p(1)\\
=&\left\{var_p(\hat{t}_{y,diff})\right\}^{-1/2}(\hat{t}_y - t_y) + o_p(1)\\
\xrightarrow{D}&N(0,1).
\end{aligned}$$

□

The design $\sqrt{n}$ consistency of $\hat{t}_{y,lasso}$ for t_y is an immediate consequence of Theorem 3.2

Corollary 3.2. *Under assumptions (D1) – (D5), the estimator $\hat{t}_{y,lasso}$ is design $\sqrt{n}$-consistent for t_y in the sense that $N^{-1}(\hat{t}_{y,lasso} - t_y) = O_p(n^{-1/2})$.*

Theorem 3.3. *Under assumptions (D1) – (D6),*

$$\begin{aligned}
\widehat{V}(\hat{t}_{y,lasso}) &= \sum\sum_{i,j\in s}\frac{\triangle_{ij}}{\pi_{ij}}\frac{\left(y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_N\right)}{\pi_i}\frac{\left(y_j - \boldsymbol{x}_j^T\hat{\boldsymbol{\beta}}_N\right)}{\pi_j}\\
&= var_p(\hat{t}_{y,diff}) + o_p\left(\frac{N^2}{n}\right).
\end{aligned} \tag{3.30}$$

Proof. Similar to the variance of the difference estimator in (3.29), we can write the estimated variance of the survey-weighted lasso regression estimator as

$$\frac{n}{N^2}\widehat{V}(\hat{t}_{y,lasso}) = \left\{\hat{\Sigma}_N^{(yy)} - \hat{\boldsymbol{\Sigma}}_N^{(yx)}\hat{\boldsymbol{\beta}}_N - \hat{\boldsymbol{\beta}}_N^T\hat{\boldsymbol{\Sigma}}_N^{(xy)} + \hat{\boldsymbol{\beta}}_N^T\hat{\boldsymbol{\Sigma}}_N^{(xx)}\hat{\boldsymbol{\beta}}_N\right\}.$$

Therefore, by Theorem 3.1 and assumption (D6), we have

$$\begin{aligned}
&\frac{n}{N^2}\left[\widehat{V}(\hat{t}_{y,lasso}) - var_p(\hat{t}_{y,diff})\right] \\
&= \left\{\hat{\Sigma}_N^{(yy)} - \hat{\boldsymbol{\Sigma}}_N^{(yx)}\hat{\boldsymbol{\beta}}_N - \hat{\boldsymbol{\beta}}_N^T\hat{\boldsymbol{\Sigma}}_N^{(xy)} + \hat{\boldsymbol{\beta}}_N^T\hat{\boldsymbol{\Sigma}}_N^{(xx)}\hat{\boldsymbol{\beta}}_N\right\} \\
&\quad - \left\{\Sigma_N^{(yy)} - \boldsymbol{\Sigma}_N^{(yx)}\boldsymbol{\beta}_N - \boldsymbol{\beta}_N^T\boldsymbol{\Sigma}_N^{(xy)} + \boldsymbol{\beta}_N^T\boldsymbol{\Sigma}_N^{(xx)}\boldsymbol{\beta}_N\right\} \\
&= o_p(1)
\end{aligned}$$

which implies (3.30).

□

Corollary 3.3. *Under assumptions (D1) – (D6),*

$$\left\{\widehat{V}(\hat{t}_{y,lasso})\right\}^{-1/2}(\hat{t}_y - t_y) \xrightarrow{D} N(0,1).$$

Proof. This result is a direct implication of Theorems 3.2 and 3.3 since

$$\begin{aligned}
\left\{\widehat{V}(\hat{t}_{y,lasso})\right\}^{-1/2}(\hat{t}_y - t_y) &= \left\{var_p(\hat{t}_{y,diff})\right\}^{-1/2}(\hat{t}_y - t_y) + o_p\left(\frac{\sqrt{n}}{N}\right)O_p\left(\frac{N}{\sqrt{n}}\right) \\
&= \left\{var_p(\hat{t}_{y,diff})\right\}^{-1/2}(\hat{t}_y - t_y) + o_p(1) \\
&\xrightarrow{D} N(0,1).
\end{aligned}$$

□

3.2.3 Asymptotic results under joint design-model framework

The descriptive study of a finite population includes estimation of a finite population quantity and in our case, estimation of the finite population total of y, t_y. While we utilized the model (3.1) to build the survey-weighted lasso regression estimator, we ignored the variability induced by the model when looking at the properties of the estimator. Ignoring the model error when conducting descriptive inference can often be justified because the error induced by the model is of a smaller order than the design-based error and because we are interested in constructing an estimator that has good design properties regardless of whether or not the model is correctly specified.

The analytic study of survey data sampled from a finite population pertains to the model which generated the population. Typically in statistics, we have data, we assume a model for the data and then we use the data to make statements, often in the form of hypothesis tests and confidence intervals, about the model. In order to make inferences about a model from a complex survey design, we need a framework which accounts for the randomness induced by the model and for the randomness induced by the sampling design. In this situation, the population $\{y_i\}_{i \in U}$ is viewed as independent realizations from the model (3.1) and the finite population coefficients, $\boldsymbol{\beta}_N$, are no longer fixed unknown quantities but instead are estimates (albeit unknown) of the true coefficients $\boldsymbol{\beta}$. The survey-weighted lasso coefficients, $\hat{\boldsymbol{\beta}}_N$, estimate the finite population coefficients $\boldsymbol{\beta}_N$ which in turn estimate the true coefficients $\boldsymbol{\beta}$. Therefore,we can argue that the survey-weighted lasso coefficients $\hat{\boldsymbol{\beta}}_N$, estimate the true coefficient vector $\boldsymbol{\beta}$. The quality of the survey coefficients as estimates of the true coefficients relies on how well the survey coefficients estimate the finite population coefficients and on how well the finite population coefficients estimate the true coefficients.

With the following additional model assumption, we prove the root-N consistency of $\hat{\boldsymbol{\beta}}_N$ for $\boldsymbol{\beta}$ and obtain a joint design-model central limit theorem for $\hat{\boldsymbol{\beta}}_N$. The approximate variance found from the central limit theorem can then be used to make confidence regions or conduct hypothesis tests for $\boldsymbol{\beta}$ while ensuring both the design and model randomness have been taken into account.

M1. The finite population is a realization from the superpopulation model,

$$\xi : y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \epsilon_i$$

where the errors, $\epsilon_1, \ldots, \epsilon_N$, are *iid* random variables with mean 0, variance σ^2, and $E|\epsilon_i^3| = \rho < \infty$. The $\boldsymbol{x}$'s are assumed to be fixed with respect to the model.

Theorem 3.4. *Under assumptions (D1) – (D5), and (M1),* $\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta} = o_p(1)$.

Proof. From assumptions (M1) and (D3), it is well established that the ordinary least squares estimate $\boldsymbol{\beta}_N$ is consistent for $\boldsymbol{\beta}$. So $\boldsymbol{\beta}_N - \boldsymbol{\beta} = o_p(1)$. From Corollary 3.1, we have

$\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}_N = o_p(1)$. Therefore, we have

$$\begin{aligned}\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta} &= \hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}_N + \boldsymbol{\beta}_N - \boldsymbol{\beta} \\ &= o_p(1) + o_p(1) \\ &= o_p(1).\end{aligned}$$

□

Theorem 3.5. *Under (D1) – (D5), and (M1)*

$$\sqrt{N}\left(\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}\right) \xrightarrow{D} \mathcal{N}\left(\mathbf{0}, \left[\pi^{-1}\boldsymbol{C}^{-1}\boldsymbol{V} + \sigma^2\right]\boldsymbol{C}^{-1}\right). \tag{3.31}$$

Proof. From assumptions (M1) and (D3), we have the asymptotic distribution of the finite population ordinary least squares estimator, $\boldsymbol{\beta}_N$ because we can write

$$\sqrt{N}\left(\boldsymbol{\beta}_N - \boldsymbol{\beta}\right) = \left(N^{-1}\boldsymbol{X}_N^T\boldsymbol{X}_N\right)^{-1}\frac{1}{\sqrt{N}}\sum_{i \in U_N}\boldsymbol{x}_i\epsilon_i$$

and by a multivariate version of Theorem 2.7.4 in Lehmann (1999), we have

$$\frac{1}{\sqrt{N}}\sum_{i \in U_N}\boldsymbol{x}_i\epsilon_i \xrightarrow{D} \mathcal{N}\left(\mathbf{0}, \sigma^2 C\right).$$

Then applying Slutsky's Lemma gives us

$$\left(N^{-1}\boldsymbol{X}_N^T\boldsymbol{X}_N\right)^{-1}\frac{1}{\sqrt{N}}\sum_{i \in U_N}\boldsymbol{x}_i\epsilon_i \xrightarrow{D} C^{-1}\mathcal{N}\left(\mathbf{0}, \sigma^2 C\right) = \mathcal{N}\left(\mathbf{0}, \sigma^2 C^{-1}\right).$$

From Theorem 3.1, we have (3.21) conditional on the data. Applying Theorem 1.3.6 in Fuller (2009), we can stack these two asymptotic statements to obtain (3.31).

□

Both Theorem 3.4 and Theorem 3.5 hold for the usual unpenalized, survey-weighted estimator, (3.7). Therefore, the asymptotic distribution of the survey-weighted lasso coefficients is the same as the asymptotic distribution of the unpenalized, survey-weighted coefficients.

3.3 Extensions of the lasso estimator

The lasso method does have a few drawbacks in terms of model selection and parameter estimation. Although the researcher can leave some variables unpenalized by omitting those variables from the penalty term, the lasso method keeps or drops penalized variables on an individual basis. Consider a set of variables, such as the dummy variables of a categorial covariate or the higher order moments of a particular covariate. The lasso method does not keep or drop the set as a group and therefore can produce unsensible models. In section 3.3.1, we discuss an extension which corrects this drawback by introducing a survey-weighted version of the group lasso estimator (Yuan and Lin 2006).

Until now, we have only considered the case where the y variable is continuous. However, there is still a need for model selection when the y is binary or represents counts. Therefore, Park and Hastie (2007) presented a lasso estimator for generalized linear models and an algorithm for fitting the entire regularization path for the estimated coefficients. We look at the survey-weighted version of the lasso estimator for glms in section 3.3.2.

In the criterion for the survey-weighted lasso coefficients, the squared residual terms are weighted by their inverse inclusion probabilities. Both Zou (2006) and Wang and Leng (2008) modified the lasso criterion so that each coefficient in the penalty term is given a different weight. The modified lasso is called adaptive lasso. This work was motivated by the fact that the lasso estimates where the true coefficient is large tend to have negative bias. Additionally, several authors have shown (Zhao and Yu (2006), Zou (2006)) that there are many scenarios where the lasso estimates obtain parameter estimation consistency but do not achieve model selection consistency. In the corrected penalty term, each coefficient is typically weighted by the inverse of a root-n consistent estimator, such as the ordinary least squares or the ridge regression estimator. In this weighting scheme, coefficients which should be large recieve a small penalty whiel coefficients which should be zero receive a large penalty. In section 3.3.3 we discuss the survey-weighted adaptive lasso and how to easily solve for the survey-weighted adaptive lasso estimates using existing algorithms.

Combinations of these lasso extensions are often appropriate and have been explored. For example, Meier, van de Geer, and Bühlmann (2008) studied group lasso for logistic regression and Wang and Leng (2008) presented results for adaptive group lasso. The survey-weighted version of these extensions can also be easily derived but is not investigated further in this thesis.

3.3.1 Survey-weighted group lasso

When variable selection should be done at a group level, not an individual variable level, the lasso estimator is not suitable. Therefore, Yuan and Lin (2006) constructed a group lasso estimator which either includes all or none of the variables in a particular group. The group lasso estimator is invariant to how the dummy variables representing a particular group are coded, a property the lasso estimator lacks. The main difference between the lasso and group lasso criteria is the penalty term in the group lasso is a hybrid between the l_1 and l_2 penalties. If there is one variable in each group, then the penalty reverts to the usual lasso penalty. However, for groups with more than one variable, an l_2 penalty is used on the coefficients in the group. Before presenting the group lasso criterion, we need to define some notation. Let G be the number of groups, p_g the number of factors or variables in group g, and $p = \sum_{g=1}^{G} p_g$ the total number of variables. The coefficient vector is $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T, \ldots, \boldsymbol{\beta}_G^T)^T$ where each $\boldsymbol{\beta}$ is of length p_g and the survey-weighted group lasso criterion is

$$\hat{\boldsymbol{\beta}}_s^{(GL)} = \arg\min_{\boldsymbol{\beta}} (\boldsymbol{Y}_s - \boldsymbol{X}_s\boldsymbol{\beta})^T \boldsymbol{\Pi}_s^{-1} (\boldsymbol{Y}_s - \boldsymbol{X}_s\boldsymbol{\beta}) + \lambda \sum_{g=1}^{G} p_g^{1/2} ||\boldsymbol{\beta}_g||_{l_2}. \tag{3.32}$$

In the penalty term, each group is weighted by the dimension of the group so that the penalty term has the same order as the degrees of freedom (Meier, van de Geer, and Bühlmann 2008). The survey-weighted lasso regression estimator is obtained by replacing the lasso coefficients in (3.9) with the group lasso coefficients.

3.3.2 Survey-weighted lasso for logistic regression

Until now, we have looked at multiple linear regression models where the study variable, y, is continuous. The lasso criterion and its variants have included the residual sums of squares subject to an l_1 penalty term for the coefficient vector. Park and Hastie (2007) extend the lasso method for independent data by constructing a criterion for generalized linear models (glms). Instead of the residual sum of squares, they minimize the negative log likelihood subject to an l_1 penalty. Meier, van de Geer, and Bühlmann (2008) use the same criterion for logistic regression but with the group level penalty on the coefficients, which was discussed in section 3.3.1. Here we propose a survey version of the lasso for logistic regression by minimizing a survey-weighted negative log likelihood subject to the l_1 penalty on the coefficient vector. These results can be extended to group survey-weighted lasso for logistic regression using the penalty term in section 3.3.1.

When the study variable, y, is continuous, the total is a common finite population quantity of interest. However, if we assume the study variable represents a binary variable such as gender or presence/absence of forest, then a common finite population quantity of interest is the population proportion of y, $P_y = N^{-1} \sum_{j \in U} y_j$. Lehtonen and Veijanen (1998) derived a model assisted logistic regression estimator for the population proportion. We build on those results to construct a model-assisted logistic lasso regression estimator.

Consider the superpopulation model, ξ, where the finite population of the study variable, $\{y_j\}_{j \in U}$, are independently distributed realizations from a Bernoulli random variable Y whose distribution is an exponential family. Through the logit function we can model the $E_\xi Y = P_\xi(Y = 1|\boldsymbol{x})$

$$logit(P_\xi(Y = 1|\boldsymbol{x})) = \log\left(\frac{P_\xi(Y = 1|\boldsymbol{x})}{1 - P_\xi(Y = 1|\boldsymbol{x})}\right) = \boldsymbol{x}^T\boldsymbol{\beta} \tag{3.33}$$

where $\boldsymbol{x}^T = (1, x_1, \ldots, x_p)$ are fixed and assumed known for each $j \in U$. The expectation, $E_\xi(\cdot)$, is with respect to the model. Denote the number of non-zero coefficients by p_o and assume the true model is sparse ($p_o < (p+1)$). In order to find the regression estimator, we need to estimate the $P_\xi(Y_j = 1|\boldsymbol{x}_j)[= f(\boldsymbol{x}_j)]$ with the following finite population quantity

$$f_U(\boldsymbol{x}_j) = \frac{\exp(\boldsymbol{x}_j^T \boldsymbol{\beta}_U)}{1 + \exp(\boldsymbol{x}_j^T \boldsymbol{\beta}_U)} \tag{3.34}$$

where the finite population coefficient vector minimizes the negative log-likelihood of the superpopulation model:

$$\begin{aligned}
\boldsymbol{\beta}_U &= \arg\min_{\boldsymbol{\beta}} [-l(\boldsymbol{\beta})] \\
&= \arg\min_{\boldsymbol{\beta}} \left[-\log \prod_{j \in U} P_\xi(Y_j = 1|\boldsymbol{x}_j)^{y_j} \left[1 - P_\xi(Y_j = 1|\boldsymbol{x}_j)\right]^{1-y_j} \right] \\
&= \arg\min_{\boldsymbol{\beta}} \left[-\sum_{j \in U} y_j \log\left[P_\xi(Y_j = 1|\boldsymbol{x}_j)\right] + (1 - y_j) \log\left[1 - P_\xi(Y = 1|\boldsymbol{x})\right] \right] \\
&= \arg\min_{\boldsymbol{\beta}} \left[-\sum_{j \in U} y_j \operatorname{logit}\left[P_\xi(Y_j = 1|\boldsymbol{x}_j)\right] + \log\left[1 - P_\xi(Y = 1|\boldsymbol{x})\right] \right] \\
&= \arg\min_{\boldsymbol{\beta}} \left[-\sum_{j \in U} y_j \boldsymbol{x}_j^T \boldsymbol{\beta} - \log\left[1 + \exp(\boldsymbol{x}_j^T \boldsymbol{\beta})\right] \right].
\end{aligned} \tag{3.35}$$

Since we only have the sampled values for the study variable, we cannot solve (3.35) and must estimate its solution with quantities based on the sample. An unpenalized estimator for the finite population coefficient vector, $\boldsymbol{\beta}_U$, is found by minimizing the survey-weighted negative log-likelihood

$$\hat{\boldsymbol{\beta}}_s = \arg\min_{\boldsymbol{\beta}} \left[-\sum_{j \in s} \frac{1}{\pi_j} \left\{ y_j \boldsymbol{x}_j^T \boldsymbol{\beta} - \log\left[1 + \exp(\boldsymbol{x}_j^T \boldsymbol{\beta})\right] \right\} \right]. \tag{3.36}$$

Since we assumed the true model is sparse and possibly $p > n$, we prefer an estimator which performs both model selection and parameter estimation. The survey-weighted lasso coefficient estimator for logistic regression is found by minimizing the survey-weighted negative log-likelihood subject to an l_1 penalty on the coefficient vector

$$\hat{\boldsymbol{\beta}}_s^{(L)} = \arg\min_{\boldsymbol{\beta}} \left[-\sum_{j \in s} \frac{1}{\pi_j} \left\{ y_j \boldsymbol{x}_j^T \boldsymbol{\beta} - \log\left[1 + \exp(\boldsymbol{x}_j^T \boldsymbol{\beta})\right] \right\} \right] \text{ subject to } \sum_{j=1}^{p} |\beta_j| \leq g$$

where $\hat{\beta}_{so}^{(L)}$ is unpenalized and $g \geq 0$ is the penalty parameter. The finite population mean function is estimated by the sample mean function

$$\hat{f}_s(\boldsymbol{x}_j) = \frac{\exp(\boldsymbol{x}_j^T \hat{\boldsymbol{\beta}}_s^{(L)})}{1 + \exp(\boldsymbol{x}_j^T \hat{\boldsymbol{\beta}}_s^{(L)})} \tag{3.37}$$

and the survey-weighted lasso regression estimator for the population proportion is

$$\hat{P}_{y,reg} = \frac{1}{N} \left[\sum_{j \in s} \frac{y_j - \hat{f}_s(\boldsymbol{x}_j)}{\pi_j} + \sum_{j \in U} \hat{f}_s(\boldsymbol{x}_j) \right]. \tag{3.38}$$

If we use the unpenalized sample coefficient vector (3.36) in the sample mean function (3.37) instead of the survey-weighted coefficient vector, then we obtain the logistic generalized regression estimator of Lehtonen and Veijanen (1998). Note, the survey-weighted coefficient vector, $\hat{\boldsymbol{\beta}}_s^{(L)}$ and sample mean function $\hat{f}_s(\boldsymbol{x}_j)$ are both functions of the penalty parameter, g. To find an appropriate value for g, one can minimize the survey-weighted $\mathrm{AIC_C}$:

$$\mathrm{AIC_{Cs}}(g) = -2 \sum_{j \in s} \frac{1}{\pi_j} \left\{ y_j \boldsymbol{x}_j^T \hat{\boldsymbol{\beta}}_s^{(L)} - \log \left[1 + \exp(\boldsymbol{x}_j^T \hat{\boldsymbol{\beta}}_s^{(L)}) \right] \right\} + \frac{2N(\mathrm{df}+1)}{N - \mathrm{df} - 1}$$

or the survey-weighted BIC:

$$\mathrm{BIC_s}(g) = -2 \sum_{j \in s} \frac{1}{\pi_j} \left\{ y_j \boldsymbol{x}_j^T \hat{\boldsymbol{\beta}}_s^{(L)} - \log \left[1 + \exp(\boldsymbol{x}_j^T \hat{\boldsymbol{\beta}}_s^{(L)}) \right] \right\} + \mathrm{df} \log(N).$$

The methods of Park and Hastie (2007) can also be used to find the survey-weighted lasso criterion for other variables whose distribution is an exponential family.

3.3.3 Survey-weighted adaptive lasso

Two of the drawbacks of the lasso estimator are the over-penalization of ‘large’ coefficients and the situations where the lasso estimates do not obtain model selection consistency. More specifically, the lasso often does not achieve the oracle properties which requires a method to have both model selection consistency and asymptotically the optimal estimation rate. The optimal estimation rate is defined as the estimation rate if we knew the true model ahead of time (Zou 2006). Model selection consistency is not achieved when the unnecessary covariates are highly correlated with the necessary covariates because in this scenario, the lasso criterion has trouble discerning which coefficients are non-zero (Zhao and Yu 2006). Zou (2006) proposed a solution to these issues, the adaptive lasso. In the adaptive lasso criterion function the coefficients in the l_1 penalty are weighted by the inverse of a root-n consistent estimator. The oracle properties are achieved by the adaptive lasso because less weight is placed on 'large' coefficients and more weight is placed on 'small' coefficients, making model selection easier. Also, since the 'large' coefficients are given a smaller penalty, they have less negative bias.

Returning to the estimation of the finite population total, t_y, where y is a continuous variable, we can derive a survey-weighted adaptive lasso regression estimator

$$\hat{t}_{y,alasso} = \sum_{j \in s} \frac{y_j - \boldsymbol{x}_j^T \hat{\boldsymbol{\beta}}_N^{(AL)}}{\pi_j} + \sum_{j \in U_N} \boldsymbol{x}_j^T \hat{\boldsymbol{\beta}}_N^{(AL)} \tag{3.39}$$

where the estimated coefficient vector based on the sample is

$$\hat{\boldsymbol{\beta}}_N^{(AL)} = \arg\min_{\boldsymbol{\beta}} \left(\boldsymbol{Y}_s - \boldsymbol{X}_s \boldsymbol{\beta}\right)^T \boldsymbol{\Pi}_s^{-1} \left(\boldsymbol{Y}_s - \boldsymbol{X}_s \boldsymbol{\beta}\right) \text{subject to } \sum_{i=1}^{p} \frac{|\beta_i|}{|\hat{\beta}_i^{(WLS)}|} \leq g \tag{3.40}$$

and the equation for $\hat{\boldsymbol{\beta}}^{(WLS)}$ is found in (3.7). To compute the survey-weighted adaptive lasso coefficient values, we first need to transform the criterion in (3.40) to look like the criterion in (3.2):

$$\begin{aligned}
&(\boldsymbol{Y}_s - \boldsymbol{X}_s \boldsymbol{\beta})^T \boldsymbol{\Pi}_s^{-1} \left(\boldsymbol{Y}_s - \boldsymbol{X}_s \boldsymbol{\beta}\right) \text{subject to } \sum_{i=1}^{p} \frac{|\beta_i|}{|\hat{\beta}_i^{(WLS)}|} \leq g \\
&\Rightarrow \left(\boldsymbol{\Pi}_s^{-1/2} \boldsymbol{Y}_s - \boldsymbol{\Pi}_s^{-1/2} \boldsymbol{X}_s \boldsymbol{\beta}\right)^T \left(\boldsymbol{\Pi}_s^{-1/2} \boldsymbol{Y}_s - \boldsymbol{\Pi}_s^{-1/2} \boldsymbol{X}_s \boldsymbol{\beta}\right) \text{subject to } \sum_{i=1}^{p} \frac{|\beta_i|}{|\hat{\beta}_i^{(WLS)}|} \leq g \\
&\Rightarrow \left(\boldsymbol{\Pi}_s^{-1/2} \boldsymbol{Y}_s - \boldsymbol{\Pi}_s^{-1/2} \boldsymbol{X}_s \boldsymbol{V}^{-1} \boldsymbol{V} \boldsymbol{\beta}\right)^T \left(\Pi_s^{-1/2} \boldsymbol{Y}_s - \Pi_s^{-1/2} \boldsymbol{X}_s \boldsymbol{V}^{-1} \boldsymbol{V} \boldsymbol{\beta}\right) \text{subject to } \sum_{i=1}^{p} \frac{|\beta_i|}{|\hat{\beta}_i^{(WLS)}|} \leq g \\
&\Rightarrow \left(\boldsymbol{Y}_s^* - \boldsymbol{X}_s^* \boldsymbol{\beta}^*\right)^T \left(\boldsymbol{Y}_s^* - \boldsymbol{X}_s^* \boldsymbol{\beta}^*\right) \text{subject to } \sum_{i=1}^{p} |\beta_i^*| \leq g
\end{aligned} \tag{3.41}$$

where $\boldsymbol{V}$ is the $(p+1) \times (p+1)$ diagonal matrix of the penalty vector

$(1, |\hat{\beta}_1^{(WLS)}|^{-1}, \ldots, |\hat{\beta}_p^{(WLS)}|^{-1})$. Using the function `l1ce()` in R (R Development Core Team 2010), we can fit the lasso criterion in (3.41) using the transformed covariate matrix, $\boldsymbol{X}_s^* = \boldsymbol{\Pi}_s^{-1/2} \boldsymbol{X}_s \boldsymbol{V}^{-1}$ and the transformed study variable vector $\boldsymbol{Y}_s^* = \boldsymbol{\Pi}_s^{-1/2} \boldsymbol{Y}_s$ to obtain $\hat{\boldsymbol{\beta}}_N^{(AL)*}$. The survey-weighted adaptive lasso coefficient values are found by back transforming:

$$\hat{\boldsymbol{\beta}}_N^{(AL)} = \boldsymbol{V}^{-1} \hat{\boldsymbol{\beta}}_N^{(AL)*}.$$

In section 3.7, we compare the survey-weighted lasso regression estimator and the survey-weighted adaptive lasso estimator via simulation.

3.4 Calibration estimators

In practice, it is often the case that several, possibly hundreds or even thousands, of finite population quantities need to be estimated from the same survey data. Therefore, it is desirable to estimate the finite population totals with weighted linear combinations of the sampled study variables:

$$\hat{t}_y = \sum_{j \in s} w_j y_j \tag{3.42}$$

where the weights, $\{w_j\}_{j \in s}$, are dependent on the sample but independent of the study variable. Since the weights are independent of the study variable, they can be applied to many variables of interest. For example, we can write the regression estimator as a linear combination of the sampled study variable:

$$\hat{t}_{y,reg} = \sum_{j \in s} \left[1 + (\boldsymbol{t}_x - \hat{\boldsymbol{t}}_{x,HT})^T \left(\sum_{j \in s} \frac{\boldsymbol{x}_j \boldsymbol{x}_j^T}{\pi_j} \right)^{-1} \boldsymbol{x}_j \right] \frac{1}{\pi_j} y_j \tag{3.43}$$

where $\boldsymbol{t}_x$ is the population total vector of the covariates and $\hat{\boldsymbol{t}}_{x,HT}$ is the corresponding Horvitz-Thompson estimator vector of the covariate totals (Särndal, Swensson, and Wretman 1992). Although the same regression model is not appropriate for each population total of interest, it is much less time consuming to compute one set of weights and as long as the study variables relate even weakly with the covariates, the weights produce a more efficient estimator than the Horvitz-Thompson weights. It is important to rely on model-assisted estimators, not model-based estimators, when the accuracy of the model is not checked for each study variable.

Since the lasso method does not produce an estimator which is linear in y, the lasso regression estimator cannot be written as a linear combination of the y values in the sample. To obtain weights, we employ the method used by Opsomer et al. (2007) and Montanari and Ranalli (2005). We construct a calibration estimator, which can be written as a weighted sum of the sampled study variable as in (3.42) with the caveat that the weights do depend on the sampled study variable, y.

The lasso calibration estimator is found by regressing the study variable, y, on the sample mean function, $\hat{f}_s(\boldsymbol{x}_j)$, over the sample (and without an intercept term). Because the calibration step is a linear regression model, the lasso calibration estimator can be written in the same form as (3.43) where $\boldsymbol{x}_j$ is replaced by $\hat{f}_s(\boldsymbol{x}_j) = \boldsymbol{x}_y^T \hat{\boldsymbol{\beta}}_s^{(L)}$:

$$\hat{t}_{y,cal} = \sum_{j\in s} \left[1 + \left(\sum_{j\in U} \hat{f}_s(\boldsymbol{x}_j) - \sum_{j\in s} \frac{\hat{f}_s(\boldsymbol{x}_j)}{\pi_j} \right) \left(\sum_{j\in s} \frac{\hat{f}_s(\boldsymbol{x}_j)^2}{\pi_j} \right)^{-1} \hat{f}_s(\boldsymbol{x}_j) \right] \frac{1}{\pi_j} y_j. \tag{3.44}$$

Since $\hat{f}_s(\boldsymbol{x}_j)$ is dependent on $\{\boldsymbol{x}_j, y_j\}_{j\in s}$, the weights in the lasso calibration estimator are dependent on the study variable, y. This dependence implies that the utility of applying these weights to other study variables depends on how correlated the variables are with y. The estimator is called the calibration estimator because it has the property that if we set the regressor as the response variable, the resulting estimator will equal the population total of the regressor. Therefore, the estimator agrees with or is calibrated on the regressor. If we let $y_j = \hat{f}_s(\boldsymbol{x}_j)$ in the lasso calibration estimator, we do indeed find

$$\begin{aligned} \hat{t}_{\hat{f}_s(x),cal} &= \sum_{j\in s} \frac{\hat{f}_s(\boldsymbol{x}_j)}{\pi_j} + \left(\sum_{j\in U} \hat{f}_s(\boldsymbol{x}_j) - \sum_{j\in s} \frac{\hat{f}_s(\boldsymbol{x}_j)}{\pi_j} \right) \left(\sum_{j\in s} \frac{\hat{f}_s(\boldsymbol{x}_j)^2}{\pi_j} \right)^{-1} \sum_{j\in s} \frac{\hat{f}_s(\boldsymbol{x}_j)^2}{\pi_j} \\ &= \sum_{j\in U} \hat{f}_s(\boldsymbol{x}_j). \end{aligned}$$

In section 3.7, we compare the lasso calibration estimator with various other finite population total estimators and consider a lasso adaptive calibration estimator where the sample mean function in (3.44) is replaced with

$$\hat{f}_s(\boldsymbol{x}_j) = \boldsymbol{x}_j^T \hat{\boldsymbol{\beta}}_s^{(AL)}. \tag{3.45}$$

3.4.1 Ridge regression approximation

Although the lasso coefficients do not have a closed form solution, Tibshirani (1996) approximated the coefficient estimates with a ridge regression format to derive the standard error. We wish to use this approximate solution as another way to construct weights for an estimator of the form (3.42). In order to utilize ridge regression, we must write the penalty term as $\sum_{i=1}^{p} \beta_i^2 |\beta_i|^{-1}$. This allows us to obtain the following approximate ridge regression coefficient estimates:

$$\hat{\boldsymbol{\beta}}_s^{(ridge)} = \left(\boldsymbol{X}_s^T \boldsymbol{\Pi}_s^{-1} \boldsymbol{X}_s + \mu \boldsymbol{Q}^-\right)^{-1} \boldsymbol{X}_s^T \boldsymbol{\Pi}_s^{-1} \boldsymbol{Y}_s$$

where $\boldsymbol{Q}$ is the diagonal matrix of the vector $(0, |\hat{\beta}_{s1}^{(L)}|, \ldots, |\hat{\beta}_{sp}^{(L)}|)$ and $\boldsymbol{Q}^-$ is the generalized inverse of $\boldsymbol{Q}$. The penalty parameter μ is chosen so that $\sum_{j \in s} |\hat{\beta}_{sj}^{(ridge)}| = g$ where g is defined in (3.8). The survey-weighted lasso ridge regression estimator is

$$\hat{t}_{y,ridge} = \sum_{j \in s} \left[1 + (\boldsymbol{t}_x - \hat{\boldsymbol{t}}_{x,HT})^T \left(\sum_{j \in s} \frac{\boldsymbol{x}_j \boldsymbol{x}_j^T}{\pi_j} + \mu \boldsymbol{Q}^- \right)^{-1} \boldsymbol{x}_j \right] \frac{1}{\pi_j} y_j. \tag{3.46}$$

It is important to again recognize that the weights in (3.46) are dependent on the study variable, y, because the weights are a function of the lasso coefficients, $\hat{\boldsymbol{\beta}}_s^{(L)}$. In section 3.7, we compare the survey-weighted lasso ridge regression estimator to other finite population total estimators. For both the calibration estimators and the ridge regression estimators, we are interested in how the constructed weights compare to the weights of the regression estimator and the Horvitz-Thompson estimator. A survey-weighted adaptive lasso ridge regression estimator is not considered in the simulation because the adaptive lasso ridge coefficients were a fairly unstable approximation of the adaptive lasso coefficients.

3.5 Model-based estimators

Thus far, the estimators discussed in this chapter are model-assisted estimators because they utilize a model but maintain good design properties regardless of the accuracy of the model. Another class of estimators are model-based estimators which tend to be more efficient than the model-assisted estimators if the assumed model is correct and the sampling design is non-informative in the sense that the sample model is the same as the superpopulation model. We wish to present the model-based counterparts to the model-assisted estimators constructed so that we can draw comparisons between the two classes of estimators via simulations.

First, assume the superpopulation model given in (3.1). The model-based generalized regression estimator equals the total of the sampled study variable plus the total of the predicted, non-sampled values [formula given in (1.5)]. For the model-based regression estimator, the model-based lasso regression estimator and the model-based adaptive lasso regression estimator, the fitted values are given by

$$\tilde{f}_s(\boldsymbol{x}_j) = \boldsymbol{x}_j^T \tilde{\boldsymbol{\beta}}_s \tag{3.47}$$

where $\tilde{\boldsymbol{\beta}}_s$ is defined in (1.6) for the regression estimator,

$$\tilde{\boldsymbol{\beta}}_s^{(L)} = \arg\min_{\boldsymbol{\beta}} \left(\boldsymbol{Y}_s - \boldsymbol{X}_s\boldsymbol{\beta}\right)^T \left(\boldsymbol{Y}_s - \boldsymbol{X}_s\boldsymbol{\beta}\right) \text{ subject to } \sum_{j=1}^{p} |\beta_j| \leq g_{MB} \tag{3.48}$$

for the lasso estimator, and

$$\tilde{\boldsymbol{\beta}}_s^{(AL)} = \arg\min_{\boldsymbol{\beta}} \left(\boldsymbol{Y}_s - \boldsymbol{X}_s\boldsymbol{\beta}\right)^T \left(\boldsymbol{Y}_s - \boldsymbol{X}_s\boldsymbol{\beta}\right) \text{ subject to } \sum_{j=1}^{p} \frac{|\beta_j|}{|\tilde{\beta}_j|} \leq g_{MB} \tag{3.49}$$

for the adaptive lasso estimator. The weights in the penalty term of the adaptive lasso estimator are the inverse elements of the ordinary least squares coefficient vector, $\tilde{\boldsymbol{\beta}}_s =$

$(\boldsymbol{X}_s^T\boldsymbol{X}_s)^{-1}\boldsymbol{X}_s^T\boldsymbol{Y}_s$. Therefore, the model-based regression estimator is

$$\tilde{t}_{y,REG} = \sum_{j\in s} y_j + \sum_{j\in U-s} \boldsymbol{x}_j^T\tilde{\boldsymbol{\beta}}_s, \tag{3.50}$$

the model-based lasso regression estimator is

$$\tilde{t}_{y,lasso} = \sum_{j\in s} y_j + \sum_{j\in U-s} \boldsymbol{x}_j^T\tilde{\boldsymbol{\beta}}_s^{(L)}, \tag{3.51}$$

and the model-based adaptive lasso regression estimator is

$$\tilde{t}_{y,alasso} = \sum_{j\in s} y_j + \sum_{j\in U-s} \boldsymbol{x}_j^T\tilde{\boldsymbol{\beta}}_s^{(AL)}. \tag{3.52}$$

To find an appropriate value for g_{MB} in (3.48) and (3.49), we propose using model-based versions of the information criterion presented in section 3.1.3. The model-based $\mathrm{AIC_C}$ is

$$\widetilde{AIC}_{Cs}(g) = n\log\left[\frac{1}{n}\sum_{j\in s}(y_j - \boldsymbol{x}_j^T\tilde{\boldsymbol{\beta}}_s^{(L)}(g))^2\right] + \frac{n(\tilde{\mathrm{df}}_s(g)+n)}{n-\tilde{\mathrm{df}}_s(g)-2} \tag{3.53}$$

where $\tilde{\mathrm{df}}_s(g)$ is the number of non-zero values in $\tilde{\boldsymbol{\beta}}_s^{(L)}(g)$. Simiarly, the model-based BIC is

$$\widetilde{BIC}_s(g) = n\log\left[\frac{1}{n}\sum_{j\in s}(y_j - \boldsymbol{x}_j^T\tilde{\boldsymbol{\beta}}_s^{(L)}(g))^2\right] + n + \log(n)\left[\tilde{\mathrm{df}}_s(g)+1\right]. \tag{3.54}$$

As discussed in section 3.4, it is often more convenient to write the estimator as the sum of a linear combination of the sampled study variable. The model-based regression estimator can be re-written as a weighted sum but the model-based lasso estimators suffer the same drawback as their model-assisted counterparts since the coefficient estimates cannot be written in closed form. Therefore, we also want to approximate the model-based lasso estimators with a calibration estimator and a ridge regression estimator. Both the model-based lasso and adaptive lasso calibration estimators have the form:

$$\tilde{t}_{y,cal} = \sum_{j \in s} \left[1 + \left(\sum_{j \in U-s} \tilde{f}_s(\boldsymbol{x}_j) \right) \left(\sum_{j \in s} \tilde{f}_s(\boldsymbol{x}_j)^2 \right)^{-1} \tilde{f}_s(\boldsymbol{x}_j) \right] y_j \tag{3.55}$$

where the sample mean function is given in (3.47) and the sample coefficient estimates are (3.48) and (3.49), respectively. The model-based calibration estimators are calibrated on the total of the fitted values. The model-based lasso ridge regression estimator is given by:

$$\tilde{t}_{y,ridge} = \sum_{j \in s} \left[1 + \left(\sum_{j \in U-s} \boldsymbol{x}_j^T \right) \left(\sum_{j \in s} \boldsymbol{x}_j \boldsymbol{x}_j^T + \mu \boldsymbol{Q}^- \right)^{-1} \boldsymbol{x}_j \right] y_j \tag{3.56}$$

where $\boldsymbol{Q}$ is the diagonal matrix of the vector $(0, |\tilde{\beta}_{N1}^{(L)}|, \ldots, |\tilde{\beta}_{Np}^{(L)}|)$ and $\boldsymbol{Q}^-$ is the generalized inverse of $\boldsymbol{Q}$.

3.6 Summary of estimators

We have presented several potential estimators for the total of a continuous study variable. When no auxiliary information is present, the Horvitz-Thompson estimator is a good design unbiased estimator for the total. When auxiliary information is present, there are several model-assisted and model-based estimators to consider. The regression estimator utilizes all the potential covariates. If some of the covariates are possibly extraneous, then the lasso or adaptive lasso estimator may be appropriate since they perform model selection. When both model selection and a list of weights are needed, then the calibration or ridge regression approximation are desirable. Additionally, if model selection should occur at a group level instead of on individual variables, a group lasso regression model is appropriate. When the study variable is binary and the finite population quantity of interest is the population proportion, then the lasso regression estimator for logistic regression is appropriate.

3.7 Simulation

We are interested in comparing the lasso regression estimator and some of its variants described in this chapter to other model-assisted and model-based estimators. In particular, we want to compare the lasso estimators to the regression estimator at two extremes: the full regression estimator which includes all covariates and the oracle regression estimator, which includes the true subset of covariates.

3.7.1 Picking the model selection criterion

We first want to evaluate different methods for selecting the penalty parameter in the lasso method. For the model-assisted lasso regression estimator, we denote the penalty parameter by g_{MA} and for the model-based lasso regression estimator, we denote the penalty parameter by g_{MB}. The model-assisted lasso regression estimator is given in (3.9) and the model-based lasso regression estimator is given by (3.51).

The linear superpopulation model of (3.1) with variance $\sigma^2 = 0.16$ is used to generate the finite population. Two mean functions, both from You (2009), are considered: a sparse, first-order mean function:

$$f_1(\boldsymbol{x}) = \boldsymbol{x}^T(1, 0, 1, 0, 1.5, 0, 0, 0, 1) \tag{3.57}$$

where $\boldsymbol{x}^T = (1, x_1, x_2, \ldots, x_8)$ and the covariates are generated to have a correlated uniform distribution and a sparse, second-order mean function

$$f_2(\boldsymbol{x}) = \boldsymbol{x}^T(1.5, 0, -4, 0, 0, 8, 0) \tag{3.58}$$

where $\boldsymbol{x}^T = (1, x_1, x_2, x_3, x_1^2, x_2^2, x_3^2)$ and the covariates are generated from uncorrelated uniform random variables. To generate the correlated covariates of (3.57), we first draw $\boldsymbol{x}^*$ from a multivariate normal with mean $\mathbf{0}$, $var(x_i^*) = 1$ and $cov(x_i^*, x_j^*) = \rho$ for $i \neq j$. The covariates are found by applying the normal cumulative distribution function to the $\boldsymbol{x}^*$ values: $\boldsymbol{x} = \Phi(\boldsymbol{x}^*)$. This construction gives the covariates a correlated uniform distribution and the strength of the correlation depends on the value of ρ. Since we are interested in how the correlation of the covariates affects the accuracy of the model selection criteria we let $\rho = 0, 0.2, 0.5, 0.98$.

The working model used for both mean functions is the first-order model of all the possible covariates. For the data generated by (3.57) the true model is a subset of the working model, whereas for the data generated by (3.58) the true model is not a subset of the working model. This model misspecification allows us to judge how the various criteria behave when the true model is not present.

Once a single population of size $N = 1000$ is generated from one of the superpopulation models, 100 samples are selected by stratified simple random sampling. Because informative sampling is pervasive in practice, we construct the strata such that the inclusion probabilities are correlated with the model errors. Following the method of You (2009), realizations, z_j are generated for each $j \in U$, from a random variable, $Z(r)$ where

$$Z(r) = \begin{cases} Z^* & \text{where } Z^* \sim \mathbb{N}(0,1) & \text{for } r = 0 \\ Z^* + \epsilon & \text{where } Z^* \sim \mathbb{N}\left(0, \frac{1-r}{r}\right) & \text{for } 0 < r < 1 \\ \epsilon & & \text{for } r = 1 \end{cases}$$

and ϵ is the model error defined in (3.1). The finite population data, $\{y_j, \boldsymbol{x}_j, z_j\}_{j \in U}$, are sorted by z_j so that the 250 smallest z values are in stratum one and the next 250 smallest z values are in stratum two and so forth. Within each stratum, simple random samples are collected with sample sizes $n_1 = 15, n_2 = 20, n_3 = 30$, and $n_4 = 35$. The sampling design is unequal probability sampling because the strata are equally sized but the sample sizes within strata differ. The sampling method is considered informative since the model errors and inclusion probabilities are correlated when $r > 0$. As r increases, the sampling method becomes more informative and we look at its effects when $r = 0, 0.25, 0.75, 1$.

For the model-assisted lasso regression estimator, we consider the following three criteria for selecting g_{MA}: survey-weighted AIC_C, survey-weighted BIC, and the design cross validation criterion of Opsomer and Miller (2005), given in (3.14), (3.15), and (3.16) respectively. For the model-based lasso regression estimator we consider the following two criteria for selecting g_{MB}: AIC_C and BIC, given in (3.53) and (3.54) respectively. For each method, the penalty parameter is selected by minimizing the criterion function. We compare the selected penalty parameters to the optimal penalty parameter which is found by minimizing the approximate design mean squared error of each estimator:

$$g_{MA,opt} = \arg\min_{g} \left[\text{MSE}_p(\hat{t}_{y,lasso}(g))\right] \text{ and } g_{MB,opt} = \arg\min_{g} \left[\text{MSE}_p(\tilde{t}_{y,lasso}(g))\right].$$

Both $g_{MA,opt}$ and $g_{MB,opt}$ are found by repeatedly sampling from the single, finite population according to the sampling design. For each sample, both the model-assisted and model-based estimators are constructed on a grid of g values so that the approximate design mean squared error is found as a function of g. Between g values, the approximate design mean squared error is found by spline interpolation. The optimal penalty term is where the approximate design mean squared error attains its minimum value.

Table 3.1 shows the optimal penalty parameter for the model-assisted estimator is fairly stable for f_1. As the sampling becomes more informative or as the correlation among the covariates increases, the optimal value stays around 3.5, which is the sum of the absolute value of the true coefficients (excluding the intercept term since it is not included in the penalty). The optimal penalty parameters for the model-based estimator, shown in Table 3.2, are less consistent. The model-based estimator does not take the sampling design into account and therefore it is naturally more affected by changes in sampling informativeness than by changes in the correlation structure of the covariates. Both the model-assisted and model-based estimator have trouble discerning an optimal penalty parameter when the model is misspecified, as is evident by the last column of Tables 3.1 and 3.2.

Table 3.1: Optimal penalty parameter, $g_{MA,opt}$, for the model-assisted lasso estimator

	Models				
$\boldsymbol{r}$	f_1: no correlation	f_1: mild correlation	f_1: moderate correlation	f_1: strong correlation	f_2
0	3.317	3.414	3.468	3.420	3.735
0.25	3.478	3.578	3.594	3.671	4.072
0.75	3.559	3.497	3.538	3.538	3.912
1.00	3.281	3.325	3.452	3.528	4.450

Table 3.2: Optimal penalty parameter, $g_{MB,opt}$, for the model-based lasso estimator

	Models				
$\boldsymbol{r}$	f_1: no correlation	f_1: mild correlation	f_1: moderate correlation	f_1: strong correlation	f_2
0	3.404	3.518	3.575	3.557	3.837
0.25	2.579	2.776	3.024	3.384	3.903
0.75	2.787	2.942	3.063	3.063	5.557
1.00	1.491	3.486	3.575	3.679	5.557

Tables 3.3 and 3.4 display the proportion of times a covariate was present in the model across the replicate samples. For f_1, the model has three non-zero coefficients but both the model-assisted and model-based estimators tend to include more than three covariates in the model. The model-based estimator builds less greedy models but usually with more than three covariates. Both methods do an excellent job of picking the correct covariates as long as the correlation among the covariates is not too strong. The same conclusions are true for f_2: although the model is misspecified, both methods pick the true covariate every time but also tend to select false signals.

Table 3.3: Average occurrence of coefficients for model-assisted estimator based on $g_{MA,opt}$

Models	r	Average Occurrence of Coefficients for $g_{MA,opt}$	True Occurrence	Average Number of Coefficients for $g_{MA,opt}$
f_1	0	$(0.69, 1.00, 0.58, 1.00, 0.73, 0.65, 0.58, 1.00)$	$(0, 1, 0, 1, 0, 0, 0, 1)$	6.23
no	0.25	$(0.69, 1.00, 0.64, 1.00, 0.71, 0.63, 0.65, 1.00)$		6.32
correlation	0.75	$(0.77, 1.00, 0.69, 1.00, 0.71, 0.71, 0.71, 1.00)$		6.59
	1.00	$(0.62, 1.00, 0.51, 1.00, 0.52, 0.54, 0.55, 1.00)$		5.74
f_1	0	$(0.67, 1.00, 0.56, 1.00, 0.71, 0.64, 0.51, 1.00)$	$(0, 1, 0, 1, 0, 0, 0, 1)$	6.09
mild	0.25	$(0.73, 1.00, 0.67, 1.00, 0.75, 0.66, 0.70, 1.00)$		6.52
correlation	0.75	$(0.75, 1.00, 0.58, 1.00, 0.59, 0.65, 0.60, 1.00)$		6.17
	1.00	$(0.60, 1.00, 0.45, 1.00, 0.49, 0.56, 0.41, 1.00)$		5.51
f_1	0	$(0.65, 1.00, 0.61, 1.00, 0.69, 0.63, 0.55, 1.00)$	$(0, 1, 0, 1, 0, 0, 0, 1)$	6.13
moderate	0.25	$(0.72, 1.00, 0.66, 1.00, 0.72, 0.66, 0.69, 1.00)$		6.46
correlation	0.75	$(0.75, 1.00, 0.59, 1.00, 0.64, 0.64, 0.65, 1.00)$		6.27
	1.00	$(0.61, 1.00, 0.55, 1.00, 0.55, 0.60, 0.53, 1.00)$		5.84
f_1	0	$(0.41, 0.73, 0.47, 0.83, 0.54, 0.58, 0.33, 0.65)$	$(0, 1, 0, 1, 0, 0, 0, 1)$	4.54
strong	0.25	$(0.57, 0.78, 0.58, 0.86, 0.66, 0.61, 0.60, 0.75)$		5.42
correlation	0.75	$(0.75, 1.00, 0.59, 1.00, 0.64, 0.64, 0.65, 1.00)$		6.27
	1.00	$(0.49, 0.75, 0.45, 0.83, 0.51, 0.68, 0.49, 0.68)$		4.88
f_2	0	$(0.31, 1.00, 0.33)$	$(0, 1, 0)$	1.64
	0.25	$(0.64, 1.00, 0.67)$		2.31
	0.75	$(0.43, 1.00, 0.49)$		1.92
	1.00	$(0.96, 1.00, 0.89)$		2.85

Table 3.4: Average occurrence of coefficients for model-based estimator based on $g_{MB,opt}$

Models	$\boldsymbol{r}$	Average Occurrence of Coefficients for $g_{MB,opt}$	True Occurrence	Average Number of Coefficients for $g_{MA,opt}$
f_1	0	$(0.79, 1.00, 0.71, 1.00, 0.78, 0.75, 0.72, 1.00)$	$(0, 1, 0, 1, 0, 0, 0, 1)$	6.75
no correlation	0.25	$(0.10, 1.00, 0.08, 1.00, 0.10, 0.07, 0.09, 1.00)$		3.43
	0.75	$(0.25, 1.00, 0.11, 1.00, 0.23, 0.23, 0.17, 1.00)$		3.99
	1.00	$(0.00, 1.00, 0.00, 1.00, 0.01, 0.00, 0.00, 0.99)$		3
f_1	0	$(0.69, 1.00, 0.65, 1.00, 0.72, 0.69, 0.67, 1.00)$	$(0, 1, 0, 1, 0, 0, 0, 1)$	6.42
mild	0.25	$(0.21, 1.00, 0.13, 1.00, 0.20, 0.14, 0.18, 1.00)$		3.86
correlation	0.75	$(0.32, 1.00, 0.16, 1.00, 0.28, 0.24, 0.31, 1.00)$		4.31
	1.00	$(0.71, 1.00, 0.67, 1.00, 0.59, 0.63, 0.55, 1.00)$		6.15
f_1	0	$(0.74, 1.00, 0.70, 1.00, 0.73, 0.71, 0.68, 1.00)$	$(0, 1, 0, 1, 0, 0, 0, 1)$	6.56
moderate	0.25	$(0.37, 1.00, 0.22, 1.00, 0.33, 0.29, 0.34, 1.00)$		4.55
correlation	0.75	$(0.37, 1.00, 0.17, 1.00, 0.34, 0.31, 0.42, 1.00)$		4.61
	1.00	$(0.73, 1.00, 0.70, 1.00, 0.66, 0.66, 0.65, 1.00)$		6.40
f_1	0	$(0.50, 0.75, 0.59, 0.86, 0.60, 0.59, 0.51, 0.69)$	$(0, 1, 0, 1, 0, 0, 0, 1)$	5.09
strong	0.25	$(0.38, 0.77, 0.39, 0.86, 0.53, 0.51, 0.46, 0.73)$		4.63
correlation	0.75	$(0.75, 1.00, 0.59, 1.00, 0.64, 0.64, 0.65, 1.00)$		4.61
	1.00	$(0.63, 0.77, 0.66, 0.91, 0.62, 0.68, 0.60, 0.77)$		5.64
f_2	0	0.391.000.40	$(0, 1, 0)$	1.79
	0.25	(0.441.000.45)		1.89
	0.75	$(1, 1, 1)$		3
	1.00	$(1, 1, 1)$		3

In order to compare the selection criteria, we find the penalty parameters which minimize the five criteria and then construct the corresponding five estimators for each of the replicate samples. In Figure 3.2, we have the distribution of the penalty parameters across repetitions for each criterion. The horizontal lines respresent to the optimal penalty parameter. Plots are based on data generated from (3.57) with varying degrees of sampling informativeness and varying levels of correlation among the covariates. In the first plot, where the sample is not informative and the covariates are uncorrelated, the methods perform fairly well at selecting a penalty parameter close, on average, to the optimal parameter. However, it is true across all situations that the BIC criterion tends to pick a smaller penalty parameter, which results in a smaller model than the model fit by the optimal parameter. When the sampling becomes informative, as displayed in the plots on the right-hand side, the model-based estimator tends to select a penalty parameter which is larger than the optimal penalty parameter. Since the penalty term equals 3.5 for the true values, we might expect the optimal penalty parameter to be around 3.5. The optimal model-based penalty parameter is smaller than 3.5 and we conjecture this occurs because the model-based methods break down as the sampling becomes more informative. The model-assisted penalty parameters seem to be unaffected by the informativeness with the weighted BIC tending toward a smaller value regardless of whether informative sampling is present.

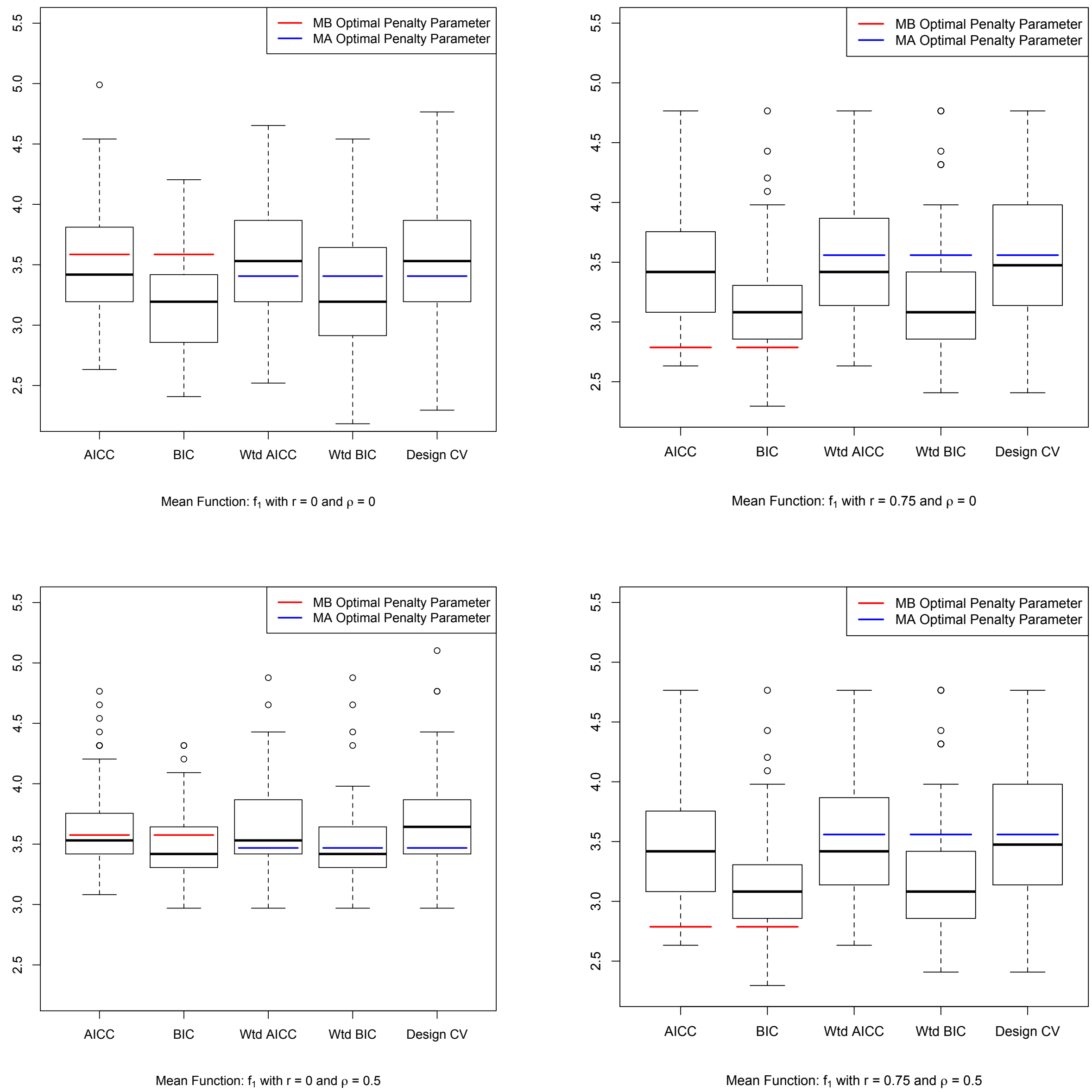

Figure 3.2: Boxplots of penalty parameters selected for each criterion

Tables 3.5 and 3.6 display the design mean squared error ratios where the mean squared error based on the optimal penalty parameter is in the denominator and the mean squared error based on the selection criterion is in the numerator:

$$\frac{\mathrm{MSE}_p(\hat{t}_y(g_{crit}))}{\mathrm{MSE}_p(\hat{t}_y(g_{opt}))}.$$

The mean squared error ratios for the model-assisted estimator are close to one, regardless of the method used to select the penalty parameter. As the sampling becomes more informative, the ratio tends to increase for the $\mathrm{AIC}_{\mathrm{Cs}}$ and the $\mathrm{BIC}_{\mathrm{s}}$ methods whereas, the ratio remains fairly constant for the Design CV method. Therefore, the Design CV method seems to be slightly better at handling the effects of informative sampling. Since the improvement is slight and all three methods appear to be adequate, we use the the $\mathrm{AIC}_{\mathrm{Cs}}$ to select the penalty parameter in section 3.7.2 when comparing the model-assisted lasso regression estimator to other estimators. The the $\mathrm{AIC}_{\mathrm{Cs}}$ method is less computationally intensive than the Design CV method.

Similarly, the ratios are also close to one for the model-based estimator. In the model misspecification case, shown in the last four rows of Table 3.6, the mean squared errors based on the $\mathrm{AIC}_{\mathrm{Cs,\ MB}}(g)$ and the $\mathrm{BIC}_{\mathrm{s,\ MB}}(g)$ is slightly less than the mean squared error based on the optimal penalty parameter. We conjecture that the criteria perform slightly better because the optimal penalty parameter is not optimal for each sample but is optimal overall, since it results in the minimum design mean squared error. The selection criteria, on the other hand, pick a different 'best' penalty parameter for each sample and therefore have the ability to achieve optimality for each particular sample. Again both methods yield similar results but as shown in Figure 3.2 the $\mathrm{AIC}_{\mathrm{Cs,\ MB}}(g)$ picks a penalty term closer to the optimal penalty term. We use $\mathrm{AIC}_{\mathrm{Cs,MB}}(g)$ to select the penalty parameter in section 3.7.2 when comparing the model-based lasso regression estimator to other estimators.

Table 3.5: Ratio of MSE based on each criterion and MSE based on the optimal penalty parameter for the model-assisted estimator

		MSE ratios		
Models	$\boldsymbol{r}$	Weighted $\mathrm{AIC_C}$	Weighted BIC	Design CV
f_1	0	1.028	1.046	1.034
no	0.25	1.029	1.055	1.024
correlation	0.75	1.053	1.100	1.054
	1.00	1.141	1.085	1.040
f_1	0	1.038	1.031	1.059
mild	0.25	1.020	1.035	1.028
correlation	0.75	1.061	1.049	1.055
	1.00	1.249	1.120	1.058
f_1	0	1.045	1.034	1.051
moderate	0.25	1.020	1.022	1.018
correlation	0.75	1.048	1.021	1.051
	1.00	1.244	1.237	1.054
f_1	0	1.030	1.032	1.033
strong	0.25	1.016	1.030	1.020
correlation	0.75	1.048	1.021	1.051
	1.00	1.148	1.144	1.069
f_2	0	1.071	1.061	1.067
	0.25	1.007	1.010	1.000
	0.75	1.035	1.054	1.059
	1.00	1.024	1.056	1.054

Table 3.6: Ratio of MSE based on each criterion and MSE based on the optimal penalty parameter for the model-based estimator

		MSE ratios	
Models	$\boldsymbol{r}$	$\mathrm{AIC_C}$	BIC
f_1	0	1.040	1.052
no	0.25	1.054	1.031
correlation	0.75	1.050	1.040
	1.00	1.036	1.025
f_1	0	1.044	1.048
mild	0.25	1.068	1.044
correlation	0.75	1.063	1.070
	1.00	1.006	1.002
f_1	0	1.044	1.033
moderate	0.25	1.054	1.041
correlation	0.75	1.069	1.052
	1.00	1.009	1.007
f_1	0	1.029	1.025
strong	0.25	1.017	1.017
correlation	0.75	1.069	1.052
	1.00	1.013	1.024
f_2	0	1.065	1.043
	0.25	0.9859	0.9724
	0.75	0.9985	0.9817
	1.00	0.9997	1.0086

3.7.2 Comparing estimators

We wish to compare the model-assisted lasso regression estimator and its variants to other survey estimators when the superpopulation model is sparse. In particular, we want to compare each model-assisted lasso estimator to its corresponding model-based estimator and to compare the model-assisted lasso estimators to other model-assisted or design-based estimators. The model-assisted oracle regression estimator is the usual regression estimator but is fit with only the true subset of covariates. This ideal estimator serves as the benchmark to which each estimator is compared. The following model-assisted and design-based estimators are considered:

LASSO.MA	lasso regression estimator	(3.9)
ALASSO.MA	adaptive lasso regression estimator	(3.39)
CLASSO.MA	lasso calibration estimator	(3.44)
CALASSO.MA	adaptive lasso calibration estimator	(3.44) with (3.45)
RLASSO.MA	lasso ridge regression estimator	(3.46)
REG.MA	regression estimator	(3.43)
ORACLE.MA	oracle regression estimator	(3.43)
HT	Horvitz-Thompson estimator	(1.1)

Since the true model is sparse, the working model contains extraneous covariates. For the REG.MA, the model fit employs all of the working model covariates, as does the RLASSO.MA though with a penalty on some of the covariates. For the LASSO.MA and the rest of its variants, the model fit utilizes some of the working model covariates and the ORACLE.MA fits utilize only the covariates found in the superpopulation model. The ORACLE.MA cannot be found in practice but in simulation serves as a measure of how well the estimators are performing.

The following model-based estimators are considered:

LASSO.MB	lasso regression estimator	(3.51)
ALASSO.MB	adaptive lasso regression estimator	(3.52)
CLASSO.MB	lasso calibration estimator	(3.55 with (3.48)
CALASSO.MB	adaptive lasso calibration estimator	(3.55) with (3.49)
RLASSO.MB	lasso ridge regression estimator	(3.56)
REG.MB	regression estimator	(1.5)
ORACLE.MB	oracle regression estimator	(1.5)

3.7.2.1 Set-up

The following superpopulation model is used to generate the finite population study variable, $\{y_j\}_{j \in U}$:

$$\begin{aligned} \xi : y_j &= \boldsymbol{x}_j^T \boldsymbol{\beta} + \epsilon_j \\ &= (1, x_{1,j}, \ldots, x_{40,j})(1, \mathbf{0}_{10}^T, (1.5)\mathbf{1}_{10}^T, \mathbf{0}_{10}^T, (3)\mathbf{1}_{10}^T)^T + \epsilon_j \end{aligned} \tag{3.59}$$

where the errors have mean 0 and variance, $\sigma^2 = 1$. The covariates are generated by the same process described in section 3.7.1 where $\boldsymbol{x}_j^*$ follows a multivariate normal distribution and $cov(x_{kj}^*, x_{lj})^* = 0.2^{|k-l|}$. The signal-to-noise ratio, defined as $(var(\boldsymbol{X}_U \boldsymbol{\beta})(\sigma^2)^{-1})^{1/2}$, is 3.66. Similar to Example four in section 7.5 of Tibshirani (1996), the superpopulation model is sparse with only 20 of the 40 covariates in the working model relating to the study variable. The oracle regression model is built utilizing only the covariates in the true model: $\{(x_{11,j}, x_{12,j}, \ldots, x_{20,j}, x_{31,j}, x_{32,j}, \ldots, x_{40,j})\}_{j \in U}$.

The population, of size $N = 10,000$, is divided into four equally sized strata by the methods discussed in section 3.7.1 with $r = 0.75$. From the fixed population, $M = 1000$ simple random samples of size $n_h = (15, 20, 30, 35)$ are taken from each strata, respectively, with a total sample size of $n = 100$ for each sample. The sampling is informative since the inclusion probabilities are correlated with the model errors.

The models are utilized to estimate the total of the particular study variable y but often in practice, there are several study variables of interest. Therefore, we also consider how the weights constructed for the lasso estimators perform at estimating totals for other study variables, which have a varying degree of similarity to the study variable, y. The five additional study variables considered are generated by the superpopulation models presented in Table 3.7. The errors in the five superpopulation models which generate the other study variables, ϵ_{ij} are *iid* $\mathbb{N}(0, \sigma^2)$ for $i = 1, 2, \ldots, 5$ and $j = 1, 2, \ldots, N$.

Table 3.7: Superpopulation models for the other study variables and their relationship to the superpopulation model for y

Model	**Relationship with y**
$z_{1j} = (1, x_{1,j}, \ldots, x_{40,j}) \begin{pmatrix} 1 \\ \mathbf{0}_{30} \\ (3)\mathbf{1}_{10} \end{pmatrix} + \epsilon_{1j}$	True model covariates are a subset of the true model covariates for y and the covariates relate similarly to z_1 as they do with y
$z_{2j} = (1, x_{1,j}, \ldots, x_{40,j}) \begin{pmatrix} 1 \\ (3)\mathbf{1}_{10} \\ \mathbf{0}_{30} \end{pmatrix} + \epsilon_{2j}$	True model covariates are a subset of the covariates which are not in the true model of y
$z_{3j} = (1, x_{1,j}, \ldots, x_{40,j}) \begin{pmatrix} 1 \\ \mathbf{1}_{10} \\ \mathbf{0}_{20} \\ (3)\mathbf{1}_{10} \end{pmatrix} + \epsilon_{3j}$	True model covariates include covariates in the true model for y and covariates not in the true model for y
$z_{4j} = (1, x_{1,j}, \ldots, x_{40,j}) \begin{pmatrix} 1 \\ \mathbf{0}_{10} \\ (3)\mathbf{1}_{10} \\ \mathbf{0}_{10} \\ (0.5)\mathbf{1}_{10} \end{pmatrix} + \epsilon_{4j}$	True model covariates are the true model covariates for y but the covariates relate differently with z_4 than with y
$z_{5j} = \epsilon_{5j}$	Noise; no similarity to y

3.7.2.2 Design bias and design mean squared error

Since one, fixed population is generated, we can compute design quantities, such as the design bias and design mean squared error, by averaging across the replicate samples. The design mean is estimated by

$$E_p(\hat{t}) \approx \frac{1}{M} \sum_{m=1}^{M} \hat{t}_m$$

and the design mean squared error is estimated by

$$\text{MSE}_p(\hat{t}) \approx \frac{1}{M} \sum_{m=1}^{M} (\hat{t}_m - t)^2.$$

Table 3.8 displays the percent relative design bias of the estimators,

$$\frac{E_p(\hat{t}) - t}{t} \times 100\%,$$

which measures how biased the estimators are under the sampling design. The model-assisted estimators have a percent relative design bias which is less than 0.5% whereas the model-based estimators all have a design bias which is greater than 1%. To assess the efficiency of the estimators, the ratios of the design mean squared error of the competing estimators to the design mean squared error of the ORACLE.MA are also given in Table 3.8. While the difference in design bias between the model-assisted and model-based estimators does not seem significant, it is compounded in the design mean squared errors. Although the estimator with the lowest design mean squared error is the idealized ORACLE, which assumes the true model is known, the model-assisted lasso estimators are almost as efficient as the ORACLE.MA with ratio values around 1.66 and are much more efficient than the full regression model estimator, the REG.MA. The model-assisted estimators are much more efficient than both the purely design-based estimator, which uses no model, and the model-based estimators, which do not account for the informative sampling. Since the oracle estimator can never be computed in practice, it is fair to say the model-assisted lasso estimators tend to be the most design efficient when the true model is sparse, the working model contains the true model, and the sampling is informative.

Table 3.8: Percent relative design bias and ratio of design MSE for each estimator to design MSE of model-assisted oracle estimator

Estimators		Percent Relative Design Bias	Design MSE Ratios
Model-Assisted:	LASSO.MA	0.37%	1.69
	ALASSO.MA	0.37%	1.63
	CLASSO.MA	0.37%	1.69
	CALASSO.MA	0.37%	1.63
	RLASSO.MA	0.38%	1.95
	REG.MA	0.49%	2.38
	ORACLE.MA	0.25%	1.00
	HT	−0.015%	13.86
Model-Based:	LASSO.MB	1.07%	6.79
	ALASSO.MB	1.07%	6.70
	CLASSO.MB	1.20%	8.31
	CALASSO.MB	1.12%	7.31
	RLASSO.MB	1.07%	7.01
	REG.MB	1.08%	7.05
	ORACLE.MB	1.07%	6.52

3.7.2.3 Survey-weighted model fits

To compare the accuracy of the LASSO.MA, ALASSO.MA and REG.MA fits, we study the selected penalty parameters, the variable selection accuracy, and coefficient estimation accuracy. The average penalty parameter selected by the survey-weighted $\mathrm{AIC_C}$ is 43.84. Plugging the true coefficients of the superpopulation model (3.59) into the penalty of the survey-weighted lasso (excluding the intercept since it is not penalized), the penalty term equals 45. Since the average penalty parameter for the survey-weighted $\mathrm{AIC_C}$ is less than 45, the method slightly over-penalizes the coefficients, on average. For the survey-weighted adaptive lasso, the average penalty parameter is 21.31. The penalty term for adaptive lasso should be roughly equal to 20, the number of non-zero coefficients in (3.59), and therefore the survey-weighted $\mathrm{AIC_C}$ is slightly under-penalizing, on average. Table 3.9 contains the average coefficient values for the LASSO.MA, ALASSO.MA and REG.MA when the coefficient is included in the model. The REG.MA coefficients are approximately unbiased whereas the LASSO.MA coefficients tend to be negatively biased for the non-zero coefficients. The ALASSO.MA coefficients, which were derived to correct for the bias, do have less negative bias than the LASSO.MA coefficients. The intercept, which is not penalized, has positive bias for the ALASSO.MA and even more so for the LASSO. We conjecture the positive bias in the intercept coefficient is attempting to counteract the negative bias in the penalized coefficients so that the overall fit is less biased.

Table 3.9: Average coefficient value for the survey-weighted lasso, survey-weighted adaptive lasso, and the survey-weighted regression estimators when the covariate is included in the model

	Average Coefficient Value										
	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
LASSO.MA	3.02	0.03	0.03	−0.00	0.08	−0.02	−0.03	0.06	0.04	0.08	0.07
ALASSO.MA	1.88	0.03	0.07	0.03	0.16	−0.05	−0.03	0.09	0.04	0.10	0.05
REG.MA	0.99	−0.00	0.03	0.00	0.08	−0.03	−0.02	0.04	0.03	0.05	0.00
		β_{11}	β_{12}	β_{13}	β_{14}	β_{15}	β_{16}	β_{17}	β_{18}	β_{19}	β_{20}
LASSO.MA		1.28	1.32	1.34	1.30	1.32	1.27	1.30	1.32	1.30	1.32
ALASSO.MA		1.41	1.43	1.45	1.41	1.42	1.38	1.41	1.44	1.41	1.46
REG.MA		1.54	1.51	1.55	1.48	1.51	1.48	1.49	1.51	1.49	1.55
		β_{21}	β_{22}	β_{23}	β_{24}	β_{25}	β_{26}	β_{27}	β_{28}	β_{29}	β_{30}
LASSO.MA		0.07	0.02	−0.07	−0.09	0.04	−0.02	0.06	0.06	−0.06	0.09
ALASSO.MA		−0.01	0.01	−0.12	−0.14	0.06	−0.03	0.05	0.12	−0.04	0.09
REG.MA		−0.01	0.01	−0.05	−0.08	0.02	−0.03	0.04	0.04	−0.03	0.01
		β_{31}	β_{32}	β_{33}	β_{34}	β_{35}	β_{36}	β_{37}	β_{38}	β_{39}	β_{40}
LASSO.MA		2.67	2.80	2.82	2.87	2.79	2.83	2.82	2.76	2.86	2.74
ALASSO.MA		2.85	2.93	2.98	3.02	2.92	2.97	2.97	2.90	3.00	2.92
REG.MA		2.91	2.98	3.04	3.08	2.98	3.03	3.02	2.95	3.05	3.00

Although the REG.MA appears to be superior since it has better coefficient estimation accuracy, it estimates all the coefficients to be non-zero, even though half of the coefficients should be exactly zero. The lasso methods are advantageous because they perform model selection and therefore estimate some coefficients to be exactly zero. Table 3.10 displays the average occurrence of the 40 covariates in the LASSO.MA and ALASSO.MA models. Since coefficients which are estimated to be 'small' values in the regression model receive a large penalty in the ALASSO.MA penalty term, the ALASSO.MA method is more efficient at driving extraneous coefficients to zero. Table 3.10 shows that a particular extraneous variable is contained in the LASSO.MA fit about 40% of the time while it is only contained in the ALASSO.MA about 21% of the time. Both methods are very good at keeping the true covariates in the model with 100% accuracy for the covariates with 'large' coefficients in the true model and nearly 100% accuracy for the covariates with smaller coefficients in the true model.

Table 3.10: Average occurrence of covariates in the lasso and adaptive lasso fits

	Average Occurrence of Covariates									
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
LASSO.MA	0.41	0.39	0.43	0.40	0.43	0.46	0.43	0.43	0.44	0.42
ALASSO.MA	0.20	0.22	0.21	0.19	0.21	0.23	0.23	0.22	0.23	0.21
	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	X_{16}	X_{17}	X_{18}	X_{19}	X_{20}
LASSO.MA	0.99	0.99	0.99	0.99	1.00	0.99	0.99	1.00	0.99	0.99
ALASSO.MA	0.98	0.97	0.98	0.97	0.98	0.97	0.97	0.97	0.97	0.97
	X_{21}	X_{22}	X_{23}	X_{24}	X_{25}	X_{26}	X_{27}	X_{28}	X_{29}	X_{30}
LASSO.MA	0.43	0.42	0.44	0.40	0.44	0.42	0.41	0.41	0.40	0.45
ALASSO.MA	0.22	0.21	0.22	0.22	0.23	0.22	0.21	0.22	0.20	0.22
	X_{31}	X_{32}	X_{33}	X_{34}	X_{35}	X_{36}	X_{37}	X_{38}	X_{39}	X_{40}
LASSO.MA	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
ALASSO.MA	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

3.7.2.4 Survey weights

As discussed in section 3.4, a single set of weights is often applied to several study variables with estimators taking the form of a linear combination of the sampled study variable (3.42). The j-th weight, w_j, roughly can be interpreted as the number of similar elements in the population that the j-th element in the sample represents. Large differences in value between weights is undesirable because it implies that some elements are much more influential on the estimate than other elements. Positive weights are also preferred because a negative weight no longer carries the described interpretation. All of the model-assisted estimators which can be written as (3.42) have weights of the form $\pi_j^{-1} + w_j^*$, where the first component is the Horvitz-Thompson weight and the second component is the model adjustment. Figure 3.3 displays the relationship between the weights of the Horvitz-Thompson estimator and the regression and calibration weights across the replicate samples. The darker the hexagon, the more concentrated the points are. Since the weights of the Horvitz-Thompson estimator only take on four different values: $\{71.43, 83.33, 125, 166.67\}$, there are four lines on which the points lie. The weights of the calibration estimator and the adaptive calibration estimator vary much less in their relation to the Horvitz-Thompson weights than the regression and oracle regression weights. We believe the variability relates to the number of covariates on which the estimator is calibrated since the calibration estimator is only calibrated on the model fits whereas the regression estimator is calibrated on each of the 40 covariates and the oracle on each of the 20 true covariates. In Figure 3.3, the red line represents the least squares line for the model-assisted weights regressed on the Horvitz-Thompson weights. The blue line is where the points would fall if the model-assisted weights actually equalled the model-assisted weights. These lines reinforce the conclusion that the calibration weights are very similar to the Horvitz-Thompson weights.

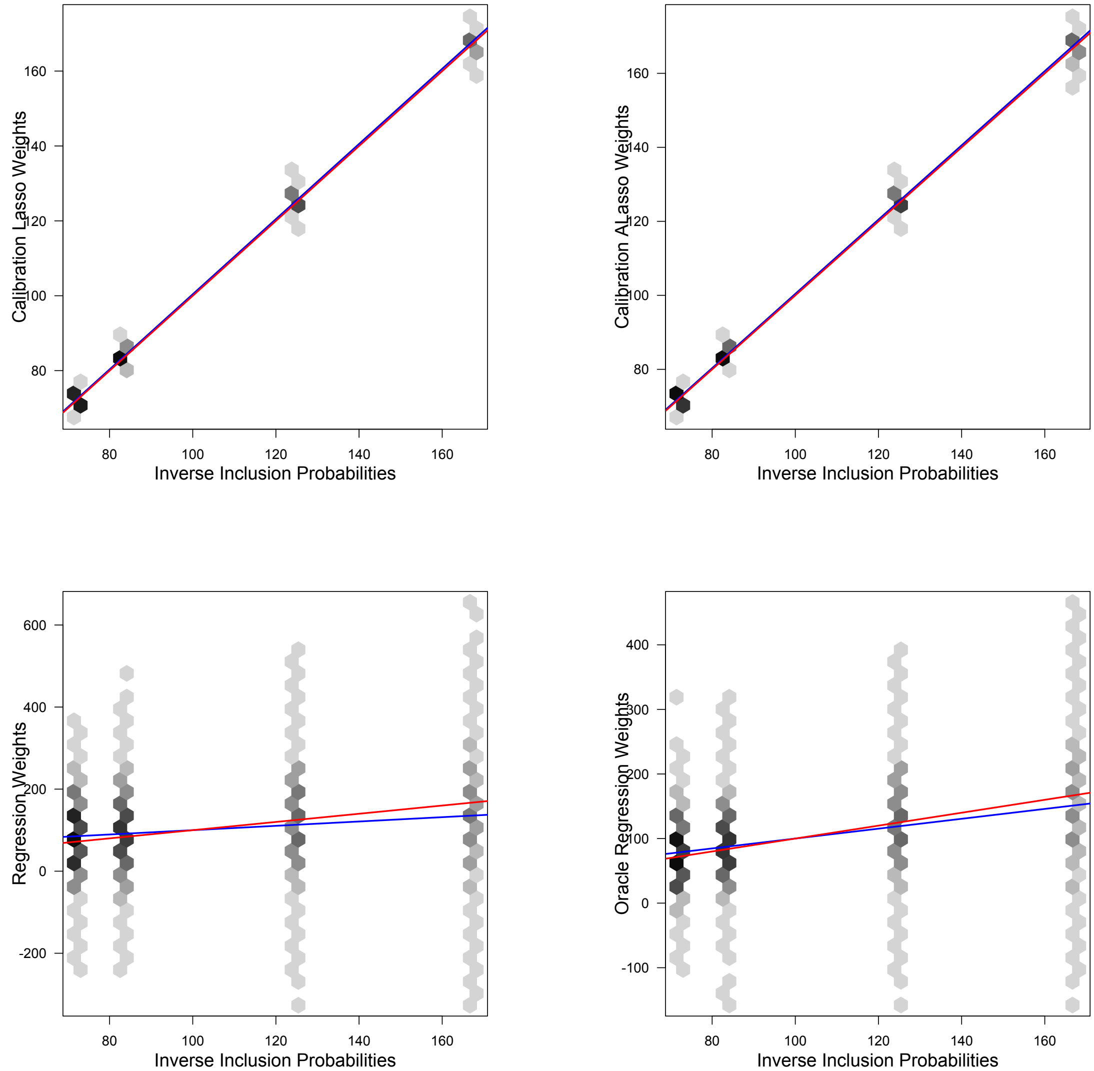

Figure 3.3: Comparing the inverse inclusion probabilities to the regression and calibration weights

To better understand how the weights vary within a sample, we compute the mean variance of the sample weights:

$$\overline{var}(\boldsymbol{w}) = \frac{1}{M}\sum_{m=1}^{M} var(\boldsymbol{w}_m) = \frac{1}{M}\sum_{m=1}^{M}\frac{1}{n-1}\sum_{j\in s}(w_{mj} - \bar{w}_m)^2$$

where $\bar{w}_m = n^{-1}\sum_{j\in s} w_{mj}$. We are also interested in how much the weight for element $j \in U$ varies from sample to sample when element j is in the sample and therefore compute the mean variance of the weight for sampled elements:

$$\begin{aligned}\overline{var}(w_j|j \in s) &= \frac{1}{N}\sum_{j\in U} var(w_j|j \in s)\\ &= \frac{1}{N}\sum_{j\in U}\frac{1}{M^*-1}\sum_{m=1}^{M}(w_{jm}-\bar{w}_j)^2 I_{\{j\in s_m\}}\end{aligned}$$

where $\bar{w}_j = M^{*-1}\sum_{m=1}^{M} w_{jm}I_{\{j\in s_m\}}$ and $M^* = \sum_{m=1}^{M} I_{\{j\in s_m\}}$. Table 3.11 displays both of these variance statistics for the weights within and across samples. The variance of the HT weights within a particular sample is 1172.44 for all repetitions since the sampling design is fixed size sampling from each stratum. Also since the HT weight of a sampled element is constant under this sampling design, the variance across samples for a particular weight given the element is sampled equals zero. Shown in Figure 3.3, the variance measures are only slightly higher for the calibration estimators than for the HT, while the REG.MA weights have the highest variability. As intuition would suggest, since the REG.MA weights are calibrated on twice as many covariates as the ORACLE.MA weights, the variance statistics for the REG.MA weights are about twice the variance statistics of the ORACLE.MA weights. The variability in the RLASSO.MA weights is between the two regression estimators because while the RLASSO.MA is calibrated on 40 covariates, some of the coefficients in the fit are penalized to be nearly zero. To measure the rate of negative weights, the average percentage of negative weights is found. On average, 11.69% of the REG.MA weights are negative, 2.82% of the ORACLE.MA weights are negative, and 5.11% of the RLASSO.MA weights are negative. The calibration estimators produced no negative weights.

Although the small variability in the weights of the calibration estimators is desirable, the weights still depend on the study variable, y, as do the weights of RLASSO.MA. On the other hand, the REG.MA, ORACLE.MA, and HT weights are independent of y and only depend on the sample, s. Therefore, it is important to assess how well the y-dependent weights perform, in comparison to the y-independent weights, when applied to other study variables of interest. Table 3.12 displays the ratio of the MSE of the model-assisted estimators to the MSE of the HT. We use the Horvitz-Thompson estimator as the benchmark in

Table 3.11: Average variances for weights within and across samples for the model-assisted and design based estimators

	Weight Variances	
Estimators	$\overline{var}(\boldsymbol{w})$	$\overline{var}(w_j \mid j \in s)$
CLASSO.MA	1181.07	3.09
CALASSO.MA	1181.09	3.42
RLASSO.MA	4526.23	3644.04
REG.MA	7939.24	5708.55
ORACLE.MA	3655.56	2965.21
HT	1172.44	0.00

this case because its performance should be consistent for different relationships between z_i and y. The ORACLE.MA is only oracle for the study variable y, not necessarily for the study variables z_i. When the true model for z_i contains the same covariates as the true model for y, as is true for z_1 and z_4, the ORACLE.MA is superior whereas when the true model for z_i does not contain any of the same covariates as the true model for y, as is true for z_2, the ORACLE.MA is less efficient than the HT. The calibration estimators are better than the HT when z_i is correlated with y, as is true for z_1, z_3 and z_4, but they are not as efficient as the other model-assisted estimators since much of the information in the individual covariates is lost. Similar to the ORACLE.MA, the calibration estimators perform poorly when the true model for y contains different covariates than the true model for z_i. Since the weights of the calibration estimators are very similar to the weights of the HT, when the study variable is completely random, as in z_5, the calibration estimators perform similarly to the Horvitz-Thompson whereas the others perform worse. Since the RLASSO.MA contains all 40 covariates, the RLASSO.MA is almost as efficient as the REG.MA for the various study variables but has the advantages of less variability in the weights and fewer negative weights.

Table 3.12: Ratios of the design mean squared error of model-assisted estimators to the design mean squared error of the Horvitz-Thompson estimator

	Design MSE Ratios				
Study Variable	z_1	z_2	z_3	z_4	z_5
CLASSO.MA	0.34	1.42	0.38	0.86	1.01
CALASSO.MA	0.33	1.45	0.38	0.88	1.01
RLASSO.MA	0.14	0.21	0.12	0.19	1.29
REG.MA	0.14	0.12	0.12	0.14	1.56
ORACLE.MA	0.11	1.17	0.20	0.10	1.19

3.7.2.5 Design properties as informative sampling, covariate correlation and model error variance are varied

Table 3.13 – Table 3.15 present simulation results when the level of informativeness of the sample (r) is varied, when the correlation among the covariates is varied (ρ), and when the model error variance (σ^2) is varied. Table 3.13, which displays varying r, tells the usual story: when the sampling design is not informative, the model-based estimator is *slightly* more efficient than its corresponding model-assisted estimator. But, as the sampling design becomes informative, the model-assisted estimator quickly becomes more efficient than its model-based counterpart. The efficiency across the model-assisted estimators remains fairly consistent as the sampling becomes more informative with the ALASSO.MA and CALASSO.MA performing the best (after the ORACLE.MA). As the correlation in the covariates increases, the difference in efficiency between estimators shrinks, as shown in Table 3.14. The changes in model errors does not seem to change the differences in efficiency between estimators, as shown in Table 3.15.

Table 3.13: Percent relative design biases and ratios of the design mean squared error of the estimators to the design mean squared error of the model-assisted oracle estimator for varying degrees of informative sampling.

		Percent Relative Design Bias			**Design MSE Ratios**		
Estimators		$r = 0$	$r = 0.25$	$r = 1$	$r = 0$	$r = 0.25$	$r = 1$
Model-Assisted:	LASSO.MA	−0.005	0.20	0.44	1.21	1.32	1.92
	ALASSO.MA	−0.010	0.20	0.44	1.16	1.24	1.86
	CLASSO.MA	−0.005	0.20	0.44	1.21	1.32	1.92
	CALASSO.MA	−0.010	0.20	0.44	1.16	1.24	1.86
	RLASSO.MA	0.001	0.20	0.45	1.42	1.54	2.19
	REG.MA	−0.006	0.27	0.56	1.29	1.55	2.73
	ORACLE.MA	−0.009	0.13	0.29	1.00	1.00	1.00
	HT	0.077	−0.05	0.01	11.58	12.61	15.68
Model-Based:	LASSO.MB	−0.015	0.60	1.26	1.13	2.58	9.54
	ALASSO.MB	−0.017	0.61	1.25	1.06	2.55	9.41
	CLASSO.MB	0.121	0.73	1.38	1.18	3.34	11.44
	CALASSO.MB	0.040	0.67	1.30	1.06	2.86	10.15
	RLASSO.MB	−0.019	0.60	1.26	1.32	2.78	9.93
	REG.MB	−0.006	0.61	1.24	1.22	2.74	9.48
	ORACLE.MB	−0.010	0.61	1.24	0.91	2.41	9.08

Table 3.14: Percent relative design biases and ratios of the design mean squared error of the estimators to the design mean squared error of the model-assisted oracle estimator for varying degrees of correlation among the covariates

		Percent Relative Design Bias			**Design MSE Ratios**		
		Correlation Among the Covariates					
Estimators		none	moderate	strong	none	moderate	strong
Model-Assisted:	LASSO.MA	0.387	0.24	0.15	1.79	1.01	1.00
	ALASSO.MA	0.365	0.27	0.19	1.62	1.01	1.01
	CLASSO.MA	0.387	0.24	0.15	1.79	1.01	1.00
	CALASSO.MA	0.366	0.27	0.19	1.63	1.01	1.01
	RLASSO.MA	0.389	0.25	0.17	1.90	1.02	1.00
	REG.MA	0.460	0.36	0.36	2.25	1.02	1.02
	ORACLE.MA	0.247	0.15	0.14	1.00	1.00	1.00
	HT	0.024	−0.07	0.22	11.03	1.45	3.93
Model-Based:	LASSO.MB	1.044	0.94	0.96	6.86	1.10	1.10
	ALASSO.MB	1.042	0.94	0.95	6.78	1.10	1.10
	CLASSO.MB	1.148	1.09	1.06	8.14	1.13	1.12
	CALASSO.MB	1.088	1.00	0.98	7.33	1.11	1.11
	RLASSO.MB	1.049	0.96	0.96	7.03	1.12	1.11
	REG.MB	1.029	0.94	0.96	6.84	1.10	1.11
	ORACLE.MB	1.042	0.95	0.95	6.61	1.10	1.10

Table 3.15: Percent relative design biases and ratios of the design mean squared error of the estimators to the design mean squared error of the model-assisted oracle estimator as the model variance changes

		Percent Relative Design Bias		**Design MSE Ratios**	
Estimators		$\sigma^2 = 0.5$	$\sigma^2 = 5$	$\sigma^2 = 0.5$	$\sigma^2 = 5$
	LASSO.MA	0.261	0.75	1.67	1.48
	ALASSO.MA	0.241	0.83	1.45	1.65
	CLASSO.MA	0.261	0.75	1.67	1.48
Model-Assisted:	CALASSO.MA	0.241	0.83	1.45	1.65
	RLASSO.MA	0.270	0.73	1.96	1.53
	REG.MA	0.326	1.03	2.26	2.26
	ORACLE.MA	0.176	0.56	1.00	1.00
	HT	0.029	0.03	30.23	3.41
	LASSO.MB	0.741	2.35	6.90	6.90
	ALASSO.MB	0.734	2.33	6.69	6.84
	CLASSO.MB	0.829	2.65	8.46	8.54
Model-Based:	CALASSO.MB	0.765	2.50	7.20	7.76
	RLASSO.MB	0.805	2.39	23.18	9.11
	REG.MB	0.729	2.31	6.88	6.88
	ORACLE.MB	0.738	2.33	6.64	6.64

3.7.2.6 Summary of estimator comparisons

Assuming some of the covariates are extraneous, the sampling is informative, and estimating t_y with precision is more important than estimating t_{z_i} with precision, the calibration estimators or LASSO.MA and ALASSO.MA are the best estimators since they have the smallest design mean squared error (after the fictitious ORACLE.MA) and the design bias of the estimators is negligible. When the same weights need to be applied to several study variables, the calibration estimators are better than LASSO.MA or ALASSO.MA since they produce weights. The calibration estimators are more precise than the HT weights when estimating t_{z_i} as long as z_i is correlated with y. In the case where precision in the estimation of t_{z_i} is required and z_i may not be correlated with y, the RLASSO.MA is a good estimator since its design MSE is competitive with the design MSE of the REG.MA for z_i and smaller for y.

3.8 Applications: United States Forest Inventory and Analysis Program

For a region of Utah, we wish to estimate the proportion of tree canopy cover by modeling the relationship between photo-interpreted data and auxiliary topographic and satellite data. Canopy cover, which is an aerial measure of the amount of ground covered by tree crowns (Toney, Shaw, and Nelson 2008), is an important variable because it is used to define forested lands. We want to compare the performance of the model-assisted estimators presented in this thesis and the Horvitz-Thompson estimator as estimators of canopy cover.

The photo-interpreted data arise from a pilot study of the Forest Inventory and Analysis Program (FIA) in collaboration with the National Land Cover Database (NLCD). To collect the data, a high intensity grid was placed on the region of interest and at each grid point, which represents a 90 by 90 meter plot of land, 105 photo points were placed (Frescino 2010). At each photo point, between two and five trained photo interpreters determined the presence or absence of a tree. For each grid point, the proportion of tree canopy cover is the average across photo interpreters of the proportion of photo points where trees are present. Although the high intensity grid is a sample of this region, we will treat these $N =$ 4,151 grid points as the entire population of interest. We can find the finite population percentage of canopy cover, $N^{-1}t_y \times 100\% = 27.41\%$, and can draw replicate samples from the population to compare the estimators to the truth.

Fourteen auxiliary variables are considered:

Variable	**Description**
2001 NLCD canopy cover estimates	Found by the Multi-Resolution Land Characteristics consortium with the goal of creating land cover maps for the United States (Frescino 2010)
Compound topographic index (CTI)	Topographic variable which measures wetness
Digital Elevation Model (DEM)	Model for elevation, slope, aspect and CTI
Slope	Slope in Degrees
Brightness	Tassel cap transformation on Landsat satellite bands, defined by Huang et al. (2002)
Greenness	Tassel cap transformation on Landsat satellite bands, defined by Huang et al. (2002)
Wetness	Tassel cap transformation on Landsat satellite bands, defined by Huang et al. (2002)
Normalized difference vegetation index (NDVI)	Transformation of Landsat satellite bands three and four
Northness of aspect	Cosine transformation of aspect
Eastness of aspect	Sine transformation of aspect

Each of the auxiliary variables are available at a finer resolution than the photo-interpreted data. The auxiliary variables were collected on a 30 by 30 meter grid, and therefore there are nine observations of every covariate for each photo-interpreted observation. To collapse the auxiliary information, the mean and standard deviation is taken of the nine observations. For the fourteen auxiliary variables, the mean for each grid point is used as a covariate and the standard deviation of the 2001 NLCD canopy cover, CTI, DEM, and slope is used as a covariate. There are 18 covariates in the working model and each is standardized. Although we conjecture that most of the covariates representing standard deviations have a weak relationship with tree canopy cover, we include those covariates in the model so that model selection is appropriate.

Stratifying the region by its ten counties to ensure a good spatial distribution of the sample, we collect 1000 replicate simple random samples of size ten from each county. Therefore, our overall sample size is 100. Since the number of grid points differs by county, we have unequal inclusion probabilities.

The actual design bias, the percent relative design bias and ratios of the design mean squared error of the model-assisted and Horvitz-Thompson estimators to the design mean squared error of the full regression estimator are given in Table 3.16. The model-assisted estimators all slightly overestimate the true proportion of tree canopy cover but still have a relative design bias of less than 1%. The lasso estimators have a smaller design mean squared error than REG.MA or HT. Since the lasso estimators are more design efficient than the full regression estimator, it appears that performing model selection is appropriate.

Table 3.16: Mean estimates of the proportion of canopy cover, percent relative design biases, and the ratios of the design mean squared error of the model-assisted and Horvitz-Thompson estimators to the design mean squared error of the full regression estimator

Estimators		**Mean Estimate**	**Percent Relative Design Bias**	**Design MSE Ratios**
Model-Assisted:	LASSO.MA	27.50%	0.32%	0.94
	ALASSO.MA	27.50%	0.33%	0.93
	CLASSO.MA	27.50%	0.35%	0.94
	CALASSO.MA	27.51%	0.35%	0.93
	RLASSO.MA	27.50%	0.32%	0.95
	REG.MA	27.53%	0.43%	1.00
	HT	27.37%	−0.13%	1.94

Table 3.17 contains the proportion of times each covariate was included in the lasso and adaptive lasso models. Both the lasso and adaptive lasso methods selected the mean 2001 NCLD canopy cover and the mean NDVI the most frequently. No covariate was consistently dropped. The adaptive lasso method selected the standard deviation of the CTI the least often, including it in the model 27.3% of the time. But the lasso method selected the standard deviation of the CTI over half of the time (56%). These values exhibit the lasso property discussed earlier: the lasso method is better at keeping true signals than at dropping false signals. The average value of the coefficient for the standard deviation of CTI (shown in Table 3.17) is essentially zero for each method and therefore this covariate

should probably not be included in the model.

Table 3.17: Average occurrence of the covariates in the survey-weighted lasso and adaptive lasso models and the average value of the coefficients when the covariate is included in the model

	Average Occurrence of Covariates		**Average Value of Coefficients**		
Covariates	LASSO.MA	ALASSO.MA	LASSO.MA	ALASSO.MA	REG.MA
Intercept	1.000	1.000	0.275	0.275	0.275
Mean Canopy Cover 2001	0.973	0.880	0.124	0.130	0.107
Std. Dev. of Canopy	0.737	0.445	0.024	0.034	0.019
Mean of CTI	0.779	0.576	−0.041	−0.050	−0.039
Std. Dev. of CTI	0.560	0.273	−0.001	0.000	−0.000
Mean of DEM	0.618	0.386	−0.009	−0.015	−0.014
Std. Dev. of DEM	0.406	0.495	−0.064	−0.072	−0.057
Mean Slope	0.433	0.518	0.070	0.083	0.058
Std. Dev. of Slope	0.647	0.404	0.009	0.011	0.003
Brightness	0.511	0.408	0.029	0.032	0.022
Greenness	0.542	0.731	−0.180	−0.169	−0.161
Wetness	0.670	0.669	0.086	0.096	0.089
NDVI	0.927	0.919	0.161	0.193	0.210
Northness	0.686	0.434	−0.018	−0.026	−0.017
Eastness	0.651	0.301	−0.003	−0.005	−0.002

For a particular sample, the coefficient paths of LASSO.MA and ALASSO.MA are given in Figures 3.4 and 3.5. The optimal set of coefficients is chosen by the survey-weighted AIC_C criterion and is designated by the vertical black line. In both cases, the optimal model includes only a subset of the potential covariates. The path of a particular coefficient, which is the same color and line type for both plots, is fairly similar whether it was found by the lasso method or the adaptive lasso method. However, the optimal set for ALASSO.MA is a subset of the optimal set for LASSO.MA, which reinforces the conclusion that the ALASSO.MA method is better at dropping false signals. The coefficients of this particular sample are given in Table 3.18. The LASSO.MA and ALASSO.MA coefficient values tend to be more similar to one another than to the REG.MA coefficient values.

In this scenario, the model selection is beneficial since the resulting estimators have a smaller design mean squared error. The adaptive lasso, which tends to select smaller models, has the lowest design mean squared error.

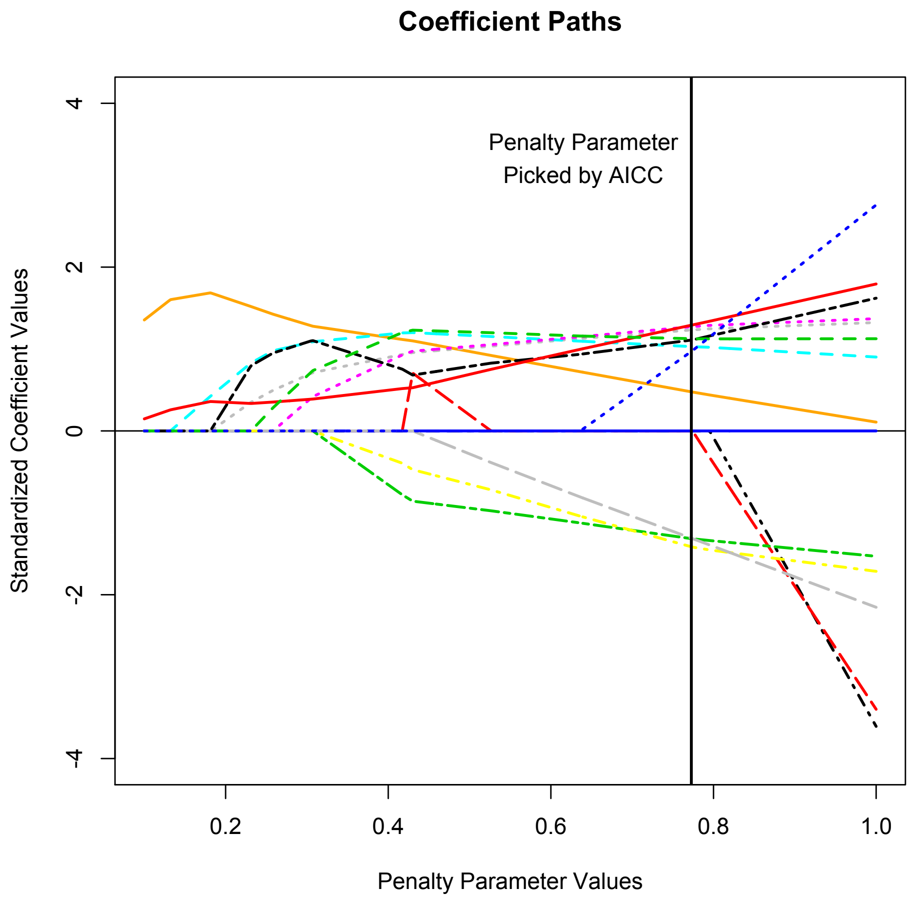

Figure 3.4: Standardized coefficient paths of survey-weighted lasso for US Forest Service data

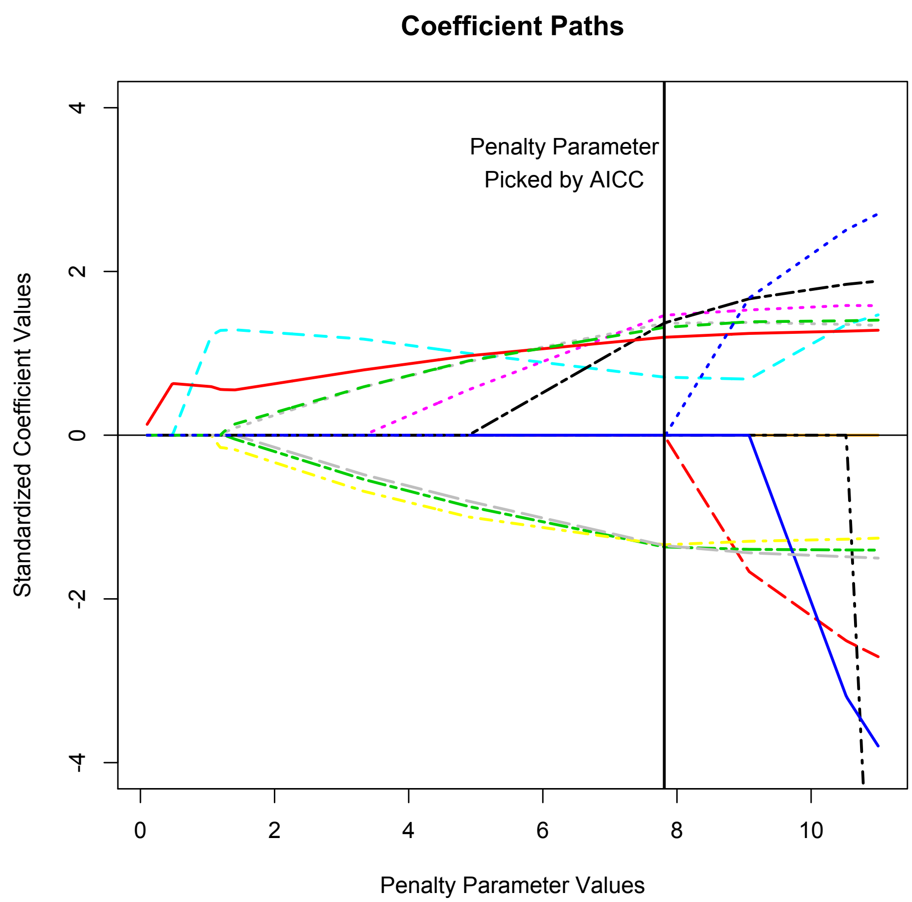

Figure 3.5: Standardized coefficient paths of survey-weighted adaptive lasso for US Forest Service data

Table 3.18: Coefficient estimates for a sample modeling tree canopy cover

	Coefficients		
	LASSO.MA	ALASSO.MA	REG.MA
Intercept	0.252	0.250	0.257
Mean Canopy Cover 2001	0.027		0.002
Std. Dev. of Canopy	0.048	0.053	0.050
Mean of CTI			−0.008
Std. Dev. of CTI			−0.013
Mean of DEM	−0.058	−0.067	−0.069
Std. Dev. of DEM			−0.136
Mean Slope			0.141
Std. Dev. of Slope	0.048	0.051	0.054
Brightness	−0.083	−0.117	−0.104
Greenness	−0.154	−0.241	−0.279
Wetness	0.032	0.035	0.052
NDVI	0.214	0.289	0.320
Northness	0.038	0.034	0.037
Eastness	0.004		0.012

3.9 Analytic inference

In statistics, it is common to use a sample to make inference about a hypothetical model. If the sample is collected from a finite population and if the sampling design is informative, then it is important that the inference accounts for the design. As discussed in section 3.2.3, the survey-weighted lasso coefficient vector $\hat{\boldsymbol{\beta}}_s$ can be viewed as an estimate of the superpopulation coefficient vector $\boldsymbol{\beta}$. In that section, we proved the root-N consistency and a central limit theorem result for $\hat{\boldsymbol{\beta}}_s$ under a joint design-model framework. In the next section, we use the survey-weighted lasso coefficients for logistic regression to estimate the superpopulation coefficients in the model (3.33).

3.9.1 Application: Centers for Disease Control and Prevention

An outbreak of the Zika virus occurred on the island of Yap between April 1 and July 31 of 2007. To better understand the prevalence of the virus and risk factors associated with contracting the virus, a single stage cluster sample of households was conducted.

The virus is believed to be transmitted by infected mosquitos Duffy et al. (2009) so the household survey included demographic information, a blood sample tested for the IgM antibody against Zika, and questions about each household member's activities during the time of the outbreak. We are interested in understanding the relationship between the risk of an infection and the covariates collected in the survey. This information could help us determine which members of the population are at a high risk for contracting the Zika virus.

The finite population on which the data were collected are Yap residents who are three years or older in age. The variable of interest, Y, equals one if an individual has the IgM antibody and showed at least one of the following suspected disease symptoms, as defined by Duffy et al. (2009), during the period of the outbreak: rash, joint pain or red eyes. The covariates we consider are how many days an individual was crabbing during the outbreak period, whether the perimeter of the house was clear of vegetation, whether the house contained any air conditioning units and the age of the individual. We assume the model (3.33) given in section 3.3.2 is appropriate. Therefore, the survey-weighted lasso coefficient vector is

$$\hat{\boldsymbol{\beta}}_s^{(L)} = \underset{\boldsymbol{\beta}}{\arg\min} \left[-\sum_{j \in s} w_j \left\{ y_j \boldsymbol{x}_j^T \boldsymbol{\beta} - \log\left[1 + \exp(\boldsymbol{x}_j^T \boldsymbol{\beta})\right] \right\} + \lambda \sum_{j=1}^{p} |\beta_j| \right] \tag{3.60}$$

where $\lambda \geq 0$ is the penalty parameter. If the response rate were 100%, the j-th weight would be $w_j = \pi_I^{-1} = N_I n_I^{-1}$, the inverse stage I inclusion probability of element j. Here, N_I equals the number of households on the island of Yap ($N_I = 1276$) and n_I is the number of sampled households ($n_I = 200$). For the collected survey, there are two levels of non-response for which the weights must be adjusted. The stage I non-response represents households included in the sample that are not enrolled in the study while the stage II non-response represents the members of an enrolled household who chose not to have their blood tested. Of the 200 households selected for the study, 163 households were enrolled in the survey where at least one household member had their blood tested. Since we have no information on the non-enrolled households, we must assume the non-enrolled houses are missing at random and therefore the adjusted stage I inclusion probability is

$$\pi_I^* = \frac{m_I}{N_I}$$

where $m_I = \sum_{i \in U_I} I\{i \in r_I\} = 163$. The households of Yap are enumerated by the set $U_I = \{1, 2, \ldots, N_I\}$, the sampled households are $s_I \subset U_I$, and the response set of households is $r_I \subseteq s_I$. Of the 808 people in the 163 households enrolled, only 556 allowed their blood to be tested. Because the response rate differed across household and even across gender within households, we cannot assume the stage II non-response is missing at random. Therefore, within a sampled household, we have divided the residents by gender and assume constant response rate within these groups. For the i-th sampled household, the group of females is U_{iF} and the group of males is U_{iM}. Assume the conditional first-order inclusion probability of the j-th person in the i-th sampled house is

$$\pi_{j|i}^* = \begin{cases} m_{iF} N_{iF}^{-1} & \text{for } j \in U_{iF} \\ m_{iM} N_{iM}^{-1} & \text{for } j \in U_{iM} \end{cases}$$

where $N_{ig} = \sum_{j \in U_{ig}} I\{j \in U_{ig}\}$ is the population size of group g in population U_i, $m_{ig} = \sum_{j \in U_{ig}} I\{j \in r_{ig}\}$ is the response size of group g in population U_i, and g is either F or M. Also, assume the individual responses are independent. Therefore, within each household, we are essentially conducting stratified Bernoulli sampling with two strata. There are two cases where the stratification breaks down and we must collapse the two groups into one group: when all members of a household are the same gender or when both genders are present but only members of one gender allowed their blood to be tested. In these cases, the conditional first-order inclusion probability of j-th person in the i-th sampled house is simply

$$\left(m_{iF} + m_{iM}\right)\left(N_{iF} + N_{iM}\right)^{-1}.$$

Adjusted for both levels of non-response, the weight for person j in the i-th sampled house is

$$w_j = (\pi_I^*)^{-1}(\pi_{j|i}^*)^{-1}.$$

To find the survey-weighted lasso coefficient vector, we minimize the criterion given in (3.60) using the R function `optim()` (R Development Core Team 2010) and find the penalty parameter value which minimizes the $\mathrm{AIC_C}$. The standardized coefficient paths are given in Figure 3.6. In this formulation of the criterion, as the penalty parameter increases, the coefficient values decrease and therefore, the penalty parameter axis is flipped. Both the age of the respondent and the number of days spent crabbing during the outbreak period are retained in the model selected by $\mathrm{AIC_C}$ which leads us to believe the true model is sparse and model selection is appropriate. The odds of contracting the Zika virus increase by 3.3% for each additional day spent crabbing when age is held constant. The odds of contracting the Zika virus increase by 5.5% for every ten year increase in age when the number of days spent crabbing is held constant. Therefore, when there is a Zika virus outbreak, those who are older and frequently go crabbing are at a higher risk of infection than those who are younger and rarely go crabbing.

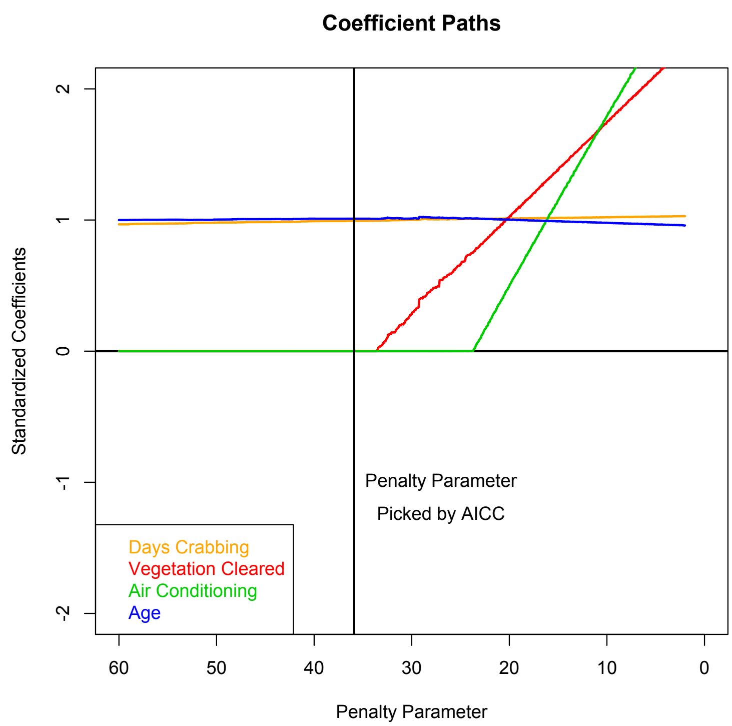

Figure 3.6: Standardized coefficient paths for CDC data

Chapter 4

Discussion and future work

4.1 Summary

In this thesis, we studied two model-assisted estimators for the finite population total: the penalized spline regression estimator and the lasso regression estimator. The penalized spline regression estimator is more efficient than the parametric regression estimator when the superpopulation model is non-linear. When the superpopulation model is linear but sparse, the lasso regression estimator is more efficient than the full regression estimator.

In chapter 2, we derived an asymptotically equivalent approximation of the penalized spline regression estimator and found its asymptotic properties when the number of knots is allowed to increase and the locations of the knots are allowed to change. We also constructed a consistent variance estimator for the asymptotic design mean squared error and demonstrated its accuracy through simulations. We proposed an additional, more accurate approximation to the penalized spline regression estimator, based on sample quantiles. To obtain consistency of the sample quantile based estimator, we showed uniform convergence of the sample quantiles to the finite population quantiles, a result which makes use of a survey-weighted Hoeffding's inequality.

In chapter 3, we considered the need for model selection when the amount of auxiliary information is vast. We proposed a survey-weighted lasso method for fitting the model, which does both model selection and parameter estimation, and used the lasso fits to construct a lasso regression estimator. We derived its asymptotic properties and through simulations, we showed it is more efficient than the regression estimator when the true model is sparse. We also discussed variants of the lasso estimator when the data are grouped, the study variable is binary, or when survey weights are needed. Additionally, we presented an adaptive lasso regression estimator which has less negative bias for large coefficients and has better model selection accuracy than the lasso regression estimator. We measured the proportion of canopy cover for a region of Utah using the lasso estimator along with other model-assisted estimators. In this scenario, the lasso estimators were more efficient than the full regression estimator. Lastly, we discussed how to conduct analytic inference using the survey-weighted lasso coefficients and under a joint design-model framework, we proved the asymptotic properties of the survey-weighted lasso coefficients as estimates for the superpopulation coefficients. We applied the joint design-model framework to estimate the coefficients in a survey-weighted logistic regression model to assess the risk of infection of the Zika virus on the island of Yap.

4.2 Future research

Wu and Sitter (2001) proposed a model calibration estimator for both linear and non-linear superpopulation models and showed the generalized regression estimator is a special case of the model calibration estimator. To extend the results of Wu and Sitter (2001) to different assumed models, Montanari and Ranalli (2005) fit the superpopulation model with neural networks and local polynomials while Opsomer, Breidt, Moisen, and Kauermann (2007) fit the superpopulation model with a generalized additive model. We want to look at the model calibration estimator when fitting the superpopulation model with penalized splines. We wish to derive the asymptotic properties of the penalized spline calibration estimator when the number of knots are allowed to increase.

Support vector machines (SVMs), a popular machine learning technique for classification and regression, are computationally efficient because the procedure only uses a subset of the data to make predictions. The sparse solutions arise from an ϵ-insensitive loss function, where data points inside an 'ϵ-tube' do not contribute to predictions (Bishop 2006). To balance desired accuracy and computational costs, one can bound the fraction of points outside the 'ϵ-tube'. Because many complex surveys consist of very large datasets, we think SVMs would be a cost efficient tool for modeling the regression relationships in the data. We would, therefore, like to look at the properties of a model-assisted SVM estimator and to compare its performance and computational costs to other model-assisted estimators.

In survey statistics, we differentiate between descriptive uses (inferences about quantities from a real, identifiable finite population) and analytic uses (inferences about model parameters from a hypothetical infinite population from which the current finite population is a realization). Analytic inference from survey data may be complicated by informative sampling methods, under which standard methods of analysis (like ordinary least squares estimation for regression models) may lead to biased and inconsistent estimators. Informative sampling can be understood as a sampling method under which the distribution of the sampled data differs from the distribution of the population data.

The problems of informative sampling, however, extend beyond surveys and can be quite common in observational studies. Length-biased sampling, a type of informative sampling where the sampling probabilities are proportional to the size of the variable of interest, is very common in a variety of applications. It is used in wildlife sampling, for example, where the method of capture-recapture selects for longer-living individuals while the method of line transect sampling selects for larger individuals. Under length-biased sampling, the sample mean, a standard estimator of the true mean, is both biased and inconsistent and therefore an inappropriate estimator. Because the usual tools can yield bad results under informative sampling, inferential methods must be adapted for this setting.

When modeling regression relationships, parametric models are often inappropriate if flexibility is needed to cover non-linear functions or if the goal is understanding the underlying trend. In both these cases, scatterplot smoothers can be quite useful. Ruppert, Wand, and Carroll (2003) present scatterplot smoothers in the context of mixed models for non-informative sampling but there are only a few sources in the literature that apply scatterplot smoothers to data derived from informative sampling. Pfeffermann and Sverchkov (1999) develop sample likelihood approaches when the population distribution is parametric while Chambers, Dorfman, and Sverchkov (2003) extend the approach to the case where the population distribution is non-parametric. Wang and Bellhouse (2009) investigated a semiparametric model with both a local polynomial and parametric component. We find this area of research to be quite fascinating and there is still much work to be done to further develop the methodology.

Bibliography

Bishop, C. (2006). *Pattern Recognition and Machine Learning*. New York: Springer.

Breidt, F. and J. Opsomer (2008). Endogenous post-stratification in surveys: Classifying with a sample-fitted model. *Annals of Statistics 36*, 403–427.

Breidt, F. J., G. Claeskens, and J. D. Opsomer (2005). Model-assisted estimation for complex surveys using penalised splines. *Biometrika 92*(4), 831–846.

Breidt, F. J. and J. D. Opsomer (2000). Local polynomial regression estimators in survey sampling. *Annals of Statistics 28*, 1026–1053.

Cassel, C. M., C. E. Särndal, and J. H. Wretman (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika 63*, 615–620.

Chambers, R. L., A. Dorfman, and M. Y. Sverchkov (2003). Nonparametric regression with complex survey data. In R. L. Chambers and C. J. Skinner (Eds.), *Analysis of Survey Data*, pp. 151–174. Chichester, U. K.: John Wiley & Sons.

Cochran, W. G. (1977). *Sampling Techniques* (3rd ed.). New York: John Wiley & Sons.

Deville, J.-C. and C.-E. Särndal (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association 87*, 376–382.

Duffy, M. R., T.-H. Chen, W. T. Hancock, A. M. Powers, J. L. Kool, R. S. Lanciotti, M. Pretrick, M. Marfel, S. Holzbauer, C. Dubray, L. Guillaumot, A. Griggs, M. Bel, A. J. Lambert, J. Laven, O. Kosoy, A. Panella, B. J. Biggerstaff, M. Fischer, and E. B. Hayes (2009). Zika virus outbreak on Yap island, Federated States of Micronesia. *New England Journal of Medicine 360*, 2536–2543.

Efron, B. (2004). The estimation of prediction error: Covariance penalties and cross-validation. *Journal of the American Statistical Association 99*, 619–632.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004). Least angle regression. *Annals of Statistics 32*, 407–499.

Frank, I. E. and J. H. Friedman (1993). A statistical view of some chemometrics regression tools. *Technometrics 35*, 109–135.

Frescino, T. (2010, April). NLCD 2011 tree canopy product. `http://www.fs.fed.us/rm/ogden/about/user_group11.shtml`.

Fuller, W. (2009). *Sampling Statistics*. New Jersey: Wiley.

Gregoire, T. G. (1998). Design-based and model-based inference in survey sampling: Appreciating the difference. *Canadian Journal of Forest Research 28*, 1429–1447.

Hájek, J. (1971). Comment on a paper by D. Basu. In V. P. Godambe and D. A. Sprott (Eds.), *Foundations of Statistical Inference*, pp. 236. Toronto: Holt, Rinehart and Winston.

Hansen, M. H., W. G. Madow, and B. J. Tepping (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys. *Journal of the American Statistical Association 78*, 776–793.

Hjort, N. L. and D. Pollard (1993). Asymptotics for minimisers of convex processes. Unpublished Manuscript. `http://www.stat.yale.edu/~pollard/Papers/`.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association 57*, 13–30.

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

Huang, C., B. Wylie, C. Homer, and G. Zylstra (2002). Derivation of a tasselled cap transformation based on Landsat 7 at-satellite reflectance. *International Journal of Remote Sensing 23*, 1741–1748.

Hurvich, C. M. and C.-L. Tsai (1989). Regression and time series model selection in small samples.

Knight, K. and W. Fu (2000). Asymptotics for lasso-type estimators. *Annals of Statistics 28*, 1356–1378.

Kutner, M. H., C. J. Nachtsheim, J. Neter, and W. Li (2005). *Applied Linear Statistical Models* (5 ed.). Boston: McGraw-Hill Irwin.

Lehmann, E. L. (1999). *Elements of Large-Sample Theory.* New York: Springer.

Lehtonen, R. and A. Veijanen (1998). Logistic generalized regression estimators. *Survey Methodology 24*, 51–55.

Li, Y. and D. Ruppert (2008). On the asymptotics of penalized splines. *Biometrika 95*, 415–436.

Lokhorst, J., B. Venables, B. T. port to R, and tests etc: Martin Maechler (2010). *lasso2: L1 constrained estimation aka 'lasso'.* R package version 1.2-11.

Meier, L., S. van de Geer, and P. Bühlmann (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society, Series B 70*, 53–71.

Montanari, G. E. and M. G. Ranalli (2005). Nonparametric model calibration estimation in survey sampling. *Journal of the American Statistical Association 100*(472), 1429–1442.

Opsomer, J. D., F. J. Breidt, G. G. Moisen, and G. Kauermann (2007). Model-assisted estimation of forest resources with generalized additive models (with discussion). *Journal of the American Statistical Association 102*, 400–416.

Opsomer, J. D. and C. P. Miller (2005). Selecting the amount of smoothing in nonparametric regression estimation for complex surveys. *Journal of Nonparametric Statistics 17*, 593–611.

Park, M. Y. and T. Hastie (2007). L_1–regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society, Series B 69*, 659–677.

Pfeffermann, D. and M. Sverchkov (1999). Parametric and semi-parametric estimation of regression models fitted to survey data. *Sankhyā, Series B 61*, 166–186.

R Development Core Team (2010). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Ruppert, D., M. P. Wand, and R. J. Carroll (2003). *Semiparametric Regression.* Cambridge, UK: Cambridge University Press.

Särndal, C.-E., B. Swensson, and J. Wretman (1989). The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. *Biometrika 76*, 527–537.

Särndal, C.-E., B. Swensson, and J. Wretman (1992). *Model Assisted Survey Sampling.* New York: Springer-Verlag.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics 6*, 461–464.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics.* New York: Wiley.

Smith, T. M. F. (1994). Sample surveys 1975-1990; An age of reconciliation? *International Statistical Review 62*, 5–19.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B 58*, 267–288.

Toney, C., J. D. Shaw, and M. D. Nelson (2008). A stem-map model for predicting tree canopy cover of forest inventory and analysis (FIA) plots. In *McWilliams, Will; Moisen, Gretchen; Czaplewski, Ray, comps. Forest Inventory and Analysis (FIA) Symposium 2008*, Fort Collins, CO. U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station.

Turlach, B. A. (2005). On algorithms for solving least squares problems under an L_1 penalty or an L_1 constraint. In *2004 Proceedings of the American Statistical Association*, Alexandria, VA, pp. 2572–2577. Statistical Computing Section, American Statistical Association.

Wang, H. and C. Leng (2008). A note on adaptive group lasso. *Computational Statistics and Data Analysis 52*, 5277–5286.

Wang, Z. and D. Bellhouse (2009). Semiparametric regression model for complex survey data. *Survey Methodology 35*, 247–260.

Wu, C. and R. R. Sitter (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association 96*, 185–193.

You, L. (2009). *Cross-Validation in Model-Assisted Estimation*. Ph. D. thesis, Iowa State University.

Yuan, M. and Y. Lin (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B 68*, 49–67.

Zhao, P. and B. Yu (2006). On model selection consistency of lasso. *Journal of Machine Learning Research 7*, 2541–2563.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association 101*, 1418–1429.

Zou, H., T. Hastie, and R. Tibshirani (2007). On the degrees of freedom of the lasso. *Annals of Statistics 35*, 2173–2192.