DISSERTATION

SURVEY SAMPLING WITH NONPARAMETRIC METHODS: ENDOGENOUS POST-STRATIFICATION AND PENALIZED INSTRUMENTAL VARIABLES

Submitted by

Mark Dahlke

Department of Statistics

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Summer 2012

Doctoral Committee:

Advisor: F. Jay Breidt

Jean Opsomer
Myung-Hee Lee
Ali Pezeshki

ABSTRACT

SURVEY SAMPLING WITH NONPARAMETRIC METHODS:
ENDOGENOUS POST-STRATIFICATION AND
PENALIZED INSTRUMENTAL VARIABLES

Two topics related to the common theme of nonparametric techniques in survey sampling are examined. The first topic explores the estimation of a finite population mean via post-stratification. Post-stratification is used to improve the precision of survey estimators when categorical auxiliary information is available from external sources. In natural resource surveys, such information may be obtained from remote sensing data classified into categories and displayed as maps. These maps may be based on classification models fitted to the sample data. Such "endogenous post-stratification" violates the standard assumptions that observations are classified without error into post-strata, and post-stratum population counts are known. Properties of the endogenous post-stratification estimator (EPSE) are derived for the case of sample-fitted nonparametric models, with particular emphasis on monotone regression models. Asymptotic properties of the nonparametric EPSE are investigated under a superpopulation model framework. Simulation experiments illustrate the practical effects of first fitting a nonparametric model to survey data before post-stratifying.

The second topic explores the use of instrumental variables to estimate regression coefficients. Informative sampling in survey problems occurs when the inclusion probabilities depend on the values of the study variable. In a regression setting under this sampling scheme, ordinary least squares estimators are biased and inconsistent. Given inverse inclusion probabilities as weights for the sample, various consistent estimators can be constructed. In particular, weighted covariates can be used as

instrumental variables, allowing for calculation of a consistent, classical two-stage least squares estimator. The proposed estimator uses a similar two-stage process, but with penalized splines at the first stage. Consistency and asymptotic normality of the new estimator are established. The estimator is asymptotically unbiased, but has a finite-sample bias that is analytically characterized. Selection of an optimal smoothing parameter is shown to reduce the finite-sample variance, in comparison to that of the classical two-stage least squares estimator, offsetting the bias and providing an estimator with a reduced mean square error.

ACKNOWLEDGEMENTS

This research was supported in part by the US National Science Foundation (SES-0922142). For her contribution to Chapter 2, Van Keilegom acknowledges financial support from IAP research network nr. P6/03 of the Belgian government (Belgian Science Policy), and from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC Grant agreement No. 203650.

DEDICATION

for Camryn

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

1.1 Overview

This paper explores the use of nonparametric methods in two distinct survey sampling situations, each involving an atypical reliance on the study variable. First, the study variable is used to form post-strata in an effort to estimate a finite population mean. One chapter is devoted to this topic. Second, the sample inclusion probabilities depend on the study variable and the selected sample is used to estimate regression coefficients. Two chapters are devoted to this topic. The final chapter provides a summary and concluding remarks.

In the remainder of this chapter, we introduce the two main topics and set the stage for more detailed discussions of each. We start with the endogenous post-stratification estimator (EPSE) and follow this with a two-stage regression estimator that uses instrumental variables (IV's) at stage one.

1.2 Endogenous post-stratification

The U.S. Forest Service provides tools to monitor and quantify the current status of the nation's forests. The Forest Inventory and Analysis (FIA) program annually conducts field visits to collect data that are used to determine estimates for a variety of forest attributes (see Frayer and Furnival 1999). Post-stratification (PS) is one

method used to improve the precision of these estimators.

1.2.1 Post-stratification background

In traditional PS (see e.g., Särndal, Swensson, and Wretman 1992, Ch. 7), we collect a sample and classify the observations into two or more post-strata. These classifications are made without error and it is assumed that we have access to the population count for each category. This population count information is typically obtained from a source outside the survey.

Once the observations have been categorized and the population counts are known, we calculate the estimated mean of the study variable by finding a weighted sum of the sample strata means. The weights are the ratios of each stratum population count to the total population count. In practice, it is common to apply these post-strata weights to other study variables of interest.

1.2.2 Endogenous post-stratification estimator (EPSE)

For the FIA, the auxiliary information used to post-stratify takes the form of groundcover categories (e.g., forest, nonforest, etc.) defined for each pixel in a map of a specified region. These groundcover categories are often determined from remotely sensed data. Because the raw satellite imagery is not immediately interpretable for this purpose, classification schemes are developed to allow category prediction for each pixel in the groundcover map. Examples of classification algorithms currently used by the U.S. Forest Service are found in Moisen and Frescino (2002). The FIA uses observed sample data from field visits to aid the development of the classification schemes (see Figure 1). This atypical use of survey data for the PS estimation process leads to the EPSE, which is examined in Breidt and Opsomer (2008). The use of the EPSE leads to the violation of two fundamental assumptions of traditional PS: imperfect classification of sample observations into the post-strata,

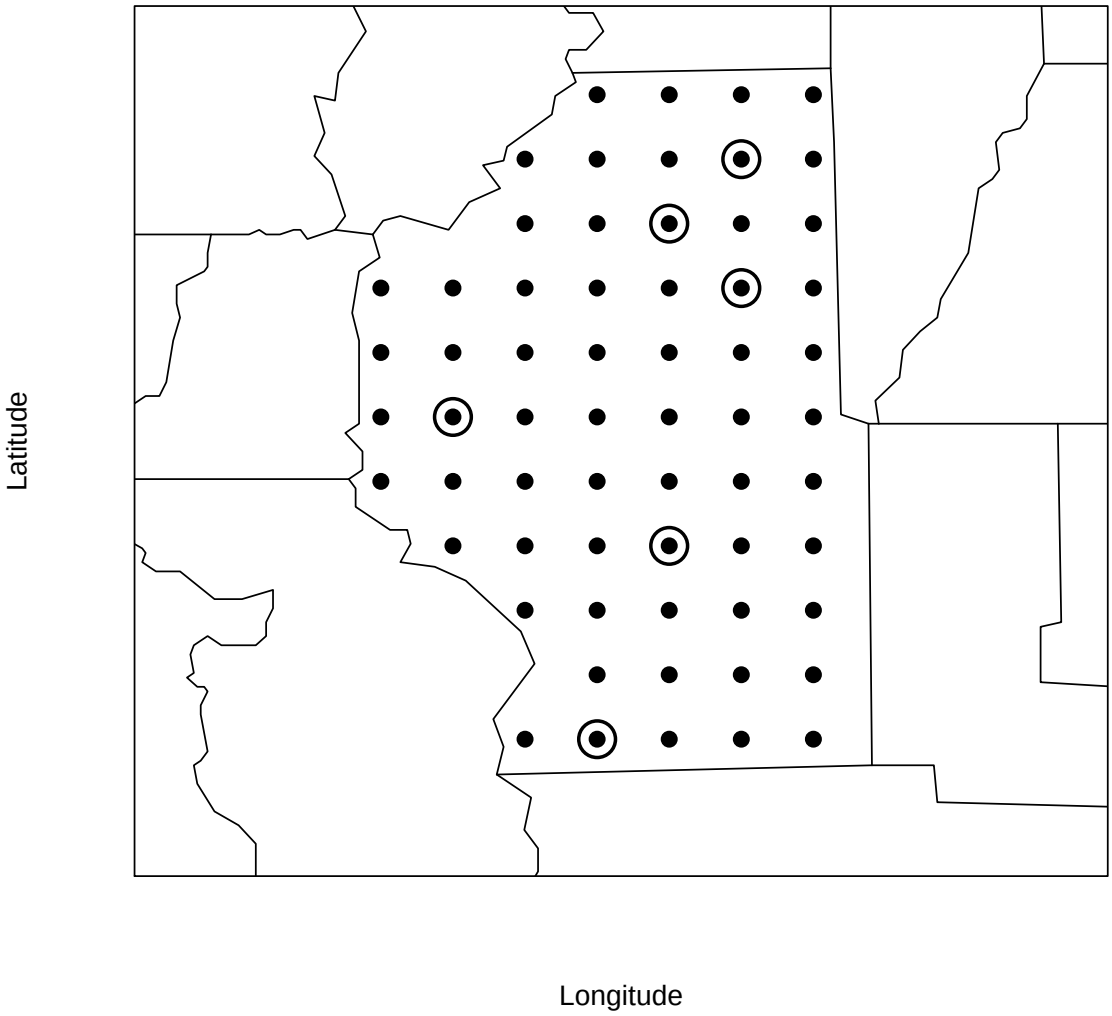

Figure 1: In this simplified example each dot represents a pixel of remotely sensed data (RSD) for a population area of interest. Ground-level sample data (GLSD) is available at each circled location. The relationship between the RSD and the GLSD at the circled locations is used to create the sample-fitted classification scheme that predicts the groundcover classifications at each uncircled dot.

and unknown exact post-stratum population counts. These violations raise concerns about the validity of the FIA process (Scott et al. 2005). The work of Breidt and Opsomer (2008) addresses these concerns by examining the properties of sample-fitted classification schemes based on parametrically specified generalized linear models. The authors demonstrate design consistency of the EPSE under mild conditions, and they demonstrate the consistency and asymptotic normality of the EPSE under a superpopulation model. The EPSE has the same asymptotic variance as the traditional post-stratified estimator with fixed strata.

1.2.3 Contributions of this thesis: Nonparametric EPSE

The nonparametric endogenous post-stratification estimator (NEPSE) of Chapter 2 extends the work of Breidt and Opsomer (2008), specifically focusing on the case where the sample-fitted model is nonparametric and monotone. This chapter is a joint work with F. Jay Breidt, Jean Opsomer, and Ingrid Van Keilegom. A condensed version of this chapter has been accepted for publication in *Statistica Sinica* (Dahlke et al. 2012).

The U.S. Forest Service is interested in this nonparametric extension because it provides justification for the nonparametric methods that are already being used in the FIA context (see Moisen and Frescino 2002 and McRoberts, Nelson, and Wendt 2002).

Asymptotic properties of the NEPSE are investigated under a superpopulation model framework. Consistency and asymptotic normality of the NEPSE are established, showing that the NEPSE has the same asymptotic variance as the traditional post-stratified estimator with fixed strata.

Simulation experiments illustrate the practical effects of first fitting a nonparametric model to the survey data before post-stratifying. We conduct two primary simulations. In the first simulation, the PS variable follows a regression through the

origin. Weights for the estimators are calculated and then applied to seven other study variables with various underlying patterns. In the second simulation, we choose other study variables to serve as the PS variable and investigate the behavior of the NEPSE when monotonicity does not hold. We find that the NEPSE outperforms the estimator with Horvitz-Thompson weights in nearly all the cases of each simulation, with roughly equivalent performance in the other cases. The NEPSE also outperforms the estimator with simple linear regression weights in roughly half of the cases in the first simulation, and nearly two-thirds of the cases in the second simulation. The regression-weights estimator does well when the response variable is linear or near linear. In both simulations, the NEPSE performance is very similar to the traditional PS estimator performance.

1.3 Two-stage regression estimator

1.3.1 Analytic inference

In Chapters 3 and 4, we are interested in the estimation of the coefficients $\boldsymbol{\beta}$ for simple linear regression models. We will actually examine several different estimators, but in this subsection we generically represent one of them using $\hat{\boldsymbol{\beta}}$. We assume a superpopulation model governed by $\boldsymbol{\beta}$ that generates our finite population $\boldsymbol{y}_{U_N}$, and we treat this finite population as a realization of an $N \times 1$ random vector $\boldsymbol{Y}_{U_N}$ (see Chambers and Skinner 2003, Ch. 1). For a specific N, we let $\boldsymbol{\beta}_N$ represent the corresponding finite population (census) parameter. This value also serves as a population-level estimator for $\boldsymbol{\beta}$ and we assume

$$\boldsymbol{\beta}_N - \boldsymbol{\beta} = O_p(N^{-1/2}).$$

We next select a Poisson sample s of size n_N from the finite population. We represent this sample using sample membership indicators I_i, where $I_i = 1$, if $i \in s$ and $I_i = 0$

otherwise, for $i = 1, \ldots, N$.

Once the sample is selected, we let $\hat{\boldsymbol{\beta}}$ represent the sample-level estimator for the census parameter. The asymptotic normality of this estimator yields

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_{_N} = O_p(n_{_N}^{-1/2}),$$

so that

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_{_N}) + (\boldsymbol{\beta}_{_N} - \boldsymbol{\beta}) = O_p(n_{_N}^{-1/2}) + O_p(N^{-1/2}) = O_p(n_{_N}^{-1/2}),$$

since $N \to \infty$ faster than $n_{_N} \to \infty$. Because the orders of the previous two expressions are the same, we focus on the design-based theory and use $\boldsymbol{\beta}$ in place of $\boldsymbol{\beta}_{_N}$ as we examine the asymptotic and finite sample properties of $\hat{\boldsymbol{\beta}}$.

For the consistency of $\hat{\boldsymbol{\beta}}$ (Lehmann and Casella 1998, Ch. 1), we demonstrate that

$$\hat{\boldsymbol{\beta}} \xrightarrow{P} \boldsymbol{\beta},$$

and for the asymptotic normality of $\hat{\boldsymbol{\beta}}$ we use the Lyapunov Central Limit Theorem (Billingsley 1995, Ch. 5) and Slutsky's Theorem (Casella and Berger 2002, Ch. 5) to show

$$\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{V}),$$

where $\boldsymbol{V}$ is the appropriate covariance matrix. To compare the various estimators in our simulations, we find the mean squared error (Casella and Berger 2002, Ch. 7) of each using

$$\text{MSE}(\hat{\boldsymbol{\beta}}) = \text{Var}(\hat{\boldsymbol{\beta}}) + (\text{E}[\hat{\boldsymbol{\beta}}] - \boldsymbol{\beta})^2$$

where $\text{E}[\hat{\boldsymbol{\beta}}]$ and $\text{Var}(\hat{\boldsymbol{\beta}})$ are estimated by the corresponding sample mean and variance.

1.3.2 Informative sampling

The samples we select in Chapters 3 and 4 are based on inclusion probabilities that are related to the study variable. This is known as *informative sampling*. We let $\boldsymbol{x}_{U_N}$ represent the covariates and consider the superpopulation regression model

$$f(\boldsymbol{y}_{U_N}, \boldsymbol{x}_{U_N}) = \prod_{i \in U_N} f(y_i \mid \boldsymbol{x}_i; \boldsymbol{\beta}) \times f(\boldsymbol{x}_{U_N}),$$

where the marginal distribution $f(\boldsymbol{x}_{U_N})$ does not depend on $\boldsymbol{\beta}$. We restrict our attention to the sampled elements (i.e., $I_i = 1$), suppress the subscript i, and consider $f(y \mid x, I = 1; \boldsymbol{\beta})$; that is, the regression of y on x given that it was observed. Selection bias occurs when $f(y \mid x, I = 1; \boldsymbol{\beta}) \neq f(y \mid x; \boldsymbol{\beta})$. Under informative sampling, we have $\pi_i = \pi_i(\boldsymbol{x}_{U_N}, \boldsymbol{y}_{U_N}) = \Pr\left\{I_i = 1 \mid \boldsymbol{x}_{U_N}, \boldsymbol{y}_{U_N}\right\}$. We again suppress the subscript i and find

$$\begin{aligned} f(y \,|\, x, I = 1; \boldsymbol{\beta}) &= \frac{\Pr\{I = 1 \,|\, x, y\}}{\int \Pr\{I = 1 \,|\, x, y\} \, f(y \mid x; \boldsymbol{\beta}) \, dy} f(y \mid x; \boldsymbol{\beta}) \\ &= \frac{\mathrm{E}[\pi(\boldsymbol{x}_{U_N}, \boldsymbol{y}_{U_N}) \,|\, x, y]}{\int \mathrm{E}[\pi(\boldsymbol{x}_{U_N}, \boldsymbol{y}_{U_N}) \,|\, x, y] f(y \mid x; \boldsymbol{\beta}) \, dy} f(y \mid x; \boldsymbol{\beta}) \\ &\neq f(y \mid x; \boldsymbol{\beta}), \end{aligned}$$

since the factor $\mathrm{E}[\pi(\boldsymbol{x}_{U_N}, \boldsymbol{y}_{U_N}) \,|\, x, y]$ depends on y, remains in the integrand, and is not canceled (Pfeffermann and Sverchkov 1999). In this scenario, the sampling process cannot be ignored in our estimation of the regression coefficients.

1.3.3 Classical instrumental variable use

The traditional use of instrumental variables is common in econometrics when there is a suspected correlation between an explanatory variable and the error term (see Wooldridge 2009, Ch. 15). As a simple example, suppose we have the following

model

$$y = \beta_0 + \beta_1 x + \epsilon,$$

and we have reason to believe that x and ϵ are correlated. This correlation causes bias and inconsistency in the ordinary least squares estimator for β_0 and β_1. If we observe an additional variable z that satisfies the assumptions: (1) z is uncorrelated with ϵ, and (2) z is correlated with x; then we call z an *instrumental variable* for x. Given this instrumental variable z, we can calculate a two-stage least squares estimator that will provide consistent estimators for β_0 and β_1. Stage one involves the least squares regression of x on z to obtain fitted values for x, denoted $\hat{x}$. We can think of this stage as "cleaning up" the x's. Stage two is the least squares regression of y on $\hat{x}$ to obtain the consistent estimators for β_0 and β_1. The two-stage estimator we develop in the survey-sampling context is similar to this one, but the need for instrumental variables arises not because x and ϵ are correlated, but rather because the sample inclusion probabilities and the error terms are correlated.

1.3.4 Instrumental variables for analytic inference

We show that the ordinary least squares (OLS) estimator $\hat{\boldsymbol{\beta}}_{ols}$ is biased and inconsistent under informative sampling. As a simple example of the inadequacy of $\hat{\boldsymbol{\beta}}_{ols}$ in an informative sampling context, we look at Figure 2 which is very similar to a graph provided in ten Cate (1986). In this graph we have a finite population (gray circles) of (x, y)-pairs with an obvious linear relationship. A sample (black dots) is drawn from this population based on inclusion probabilities (π_i) that depend on the y-values of the ordered pairs. There are three strata of y-values, each with a fixed inclusion probability. The outer two strata have the same small inclusion probability and the middle stratum has a large inclusion probability. We see in the figure that the OLS estimated regression line does not follow the linear trend of the finite population values. Instead, the slope is too small and the intercept is too large. We

would like to fix this problem.

One solution is to use the probability-weighted least squares estimator $\hat{\boldsymbol{\beta}}_{2sls}$ (see ten Cate 1986, Pfeffermann and Sverchkov 1999, or Fuller 2009, Ch. 6). In Chapter 3, we show that this estimator is consistent under informative sampling. The estimator has the form

$$\hat{\boldsymbol{\beta}}_{2sls} = (\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{y},$$

where $\boldsymbol{X}$ is the matrix of covariates and $\boldsymbol{W}$ is a diagonal matrix with the reciprocal inclusion probabilities (i.e., $1/\pi_i$'s) as weights on the diagonal. We will think of this estimator as the end result of a two-stage least squares process that involves instrumental variables (see Fuller 2009, Ch. 6). The informative sampling issue that plagues $\hat{\boldsymbol{\beta}}_{ols}$ is the correlation between the errors and the inclusion probabilities that arises since both depend on y. The use of appropriate instrumental variables negates this problem. Fuller (2009, Ch. 6) demonstrates that any weighted function of the covariate x can be used as an instrumental variable (IV).

In addition to studying and explaining $\hat{\boldsymbol{\beta}}_{2sls}$'s advantages over $\hat{\boldsymbol{\beta}}_{ols}$, we also seek an estimator that is better than $\hat{\boldsymbol{\beta}}_{2sls}$. Several authors have explored improved estimators for various models and sampling designs and based on various criteria (e.g., Holt, Smith, and Winter 1980, Jewell 1985, and Pfeffermann and Sverchkov 2003). One semi-parametric estimator mentioned in Fuller (2009, Ch. 6) is the estimator developed in Pfeffermann and Sverchkov (1999) having the form

$$\hat{\boldsymbol{\beta}}_{pfsv} = (\boldsymbol{X}^T \boldsymbol{W} \tilde{\boldsymbol{W}}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{W} \tilde{\boldsymbol{W}}^{-1} \boldsymbol{y},$$

where $\tilde{\boldsymbol{W}}$ is obtained by regressing the column vector of weights $\boldsymbol{w}$ on $\boldsymbol{X}$. Generally, the modified weights of $\hat{\boldsymbol{\beta}}_{pfsv}$ make it a more efficient estimator than $\hat{\boldsymbol{\beta}}_{2sls}$. We include comparisons to this estimator in some of our simulations.

Before discussing our new estimator, we again refer to the two-stage least squares

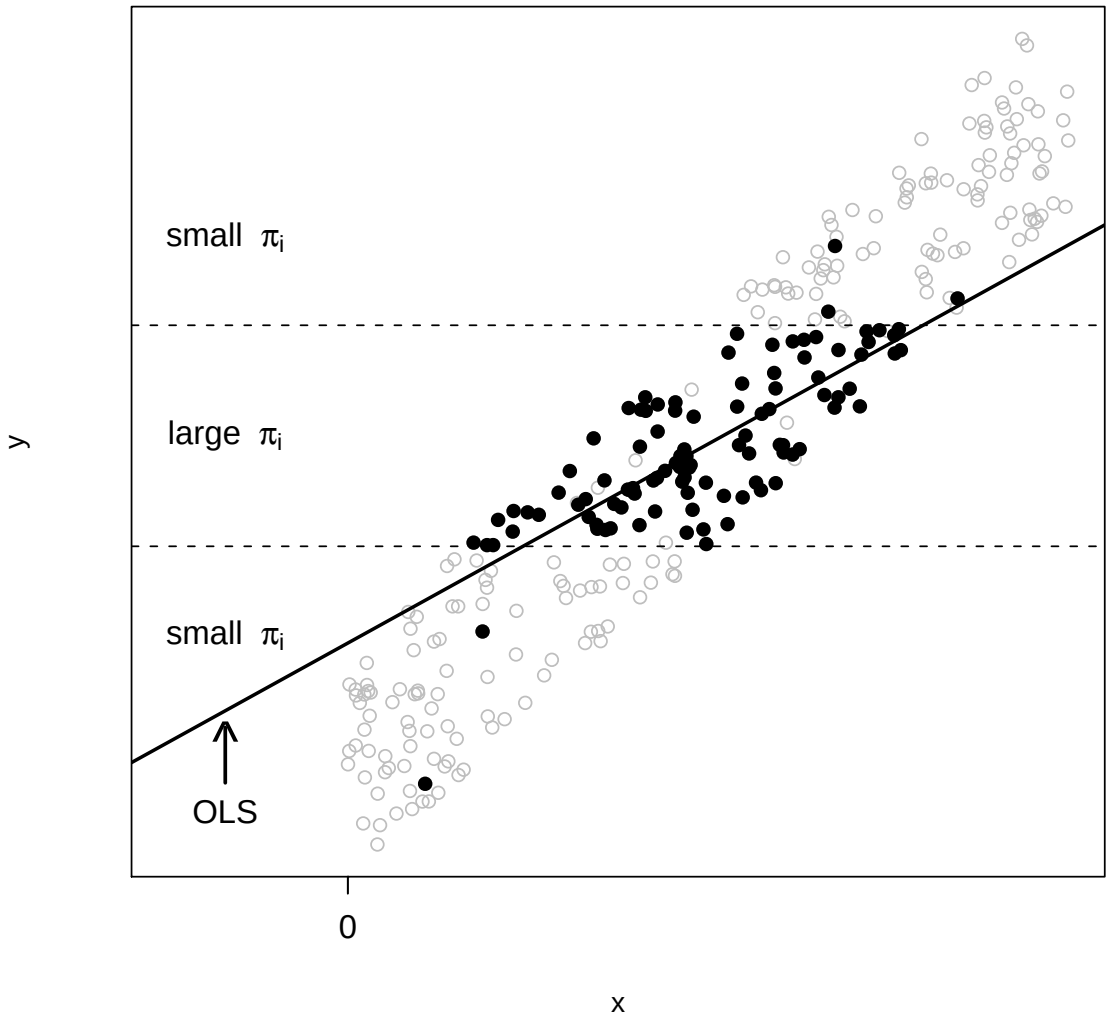

Figure 2: Gray circles are unsampled finite population values and black dots are sampled values. Inclusion probabilities (π_i) depend on the the y-values (informative sampling), larger π_i for the middle y-values and smaller π_i for the extreme y-values. The ordinary least squares (OLS) estimate based on the sample data is provided. Examination of the finite population shows that this line underestimates the actual slope and overestimates the actual intercept, motivating the need for an improved estimator.

process and the use of IV's. Hidden in the final form of $\hat{\boldsymbol{\beta}}_{2sls}$ is the use of IV's at stage one. At stage one we regress $\boldsymbol{X}$ on $\boldsymbol{A}_x$ where $\boldsymbol{A}_x = \boldsymbol{W}\boldsymbol{X}$ is the IV matrix (i.e., weighted $\boldsymbol{X}$). This regression allows us to calculate a fitted covariate matrix $\hat{\boldsymbol{X}}$. At stage two, we regress $\boldsymbol{y}$ on $\hat{\boldsymbol{X}}$ to produce $\hat{\boldsymbol{\beta}}_{2sls}$. A second, similar estimator is formed by increasing the number of IV's used at stage one. We let $\boldsymbol{Z}$ denote a matrix of K functions of x, then $\boldsymbol{A} = \boldsymbol{W}[\boldsymbol{X}\ \boldsymbol{Z}]$ takes the place of $\boldsymbol{A}_x$ as the IV matrix and the two-stage process is repeated to obtain $\hat{\boldsymbol{\beta}}_{2sls}^{K+2}$. The $K+2$ denotes the number of IV's used at stage one when the original model has the form $y = \beta_0 + \beta_1 x + \epsilon$. In Chapter 3, we demonstrate the consistency and asymptotic normality of $\hat{\boldsymbol{\beta}}_{2sls}$ and $\hat{\boldsymbol{\beta}}_{2sls}^{K+2}$ before introducing our new estimator.

1.3.5 Contributions of this thesis: Nonparametric IV's

The proposed estimator, which we call the *penalized spline estimator*, is similar to $\hat{\boldsymbol{\beta}}_{2sls}^{K+2}$, but instead of ordinary least squares at stage one, we use a penalized spline to determine the fitted covariate matrix $\hat{\boldsymbol{X}}$. At stage two, we regress $\boldsymbol{y}$ on $\hat{\boldsymbol{X}}$ to produce $\hat{\boldsymbol{\beta}}_{pspl}$. We demonstrate consistency and asymptotic normality of $\hat{\boldsymbol{\beta}}_{pspl}$, but we also find that a bias term remains for finite samples. We further demonstrate that in certain informative sampling situations, the variance of $\hat{\boldsymbol{\beta}}_{pspl}$ is small enough to offset the finite sample bias and produce an estimator with a smaller MSE than $\hat{\boldsymbol{\beta}}_{2sls}$ and $\hat{\boldsymbol{\beta}}_{2sls}^{K+2}$ (and $\hat{\boldsymbol{\beta}}_{pfsv}$). We provide an informative sampling simulation that verifies the advantage of the penalized spline estimator over the ordinary least squares estimator, the other two-stage least squares estimators, and the Pfeffermann-Sverchkov estimator.

Chapter 4 examines the reasons for the reduction in variance of the regression estimators when additional IV's are used at stage one. Under informative sampling, the development of expressions for the variance of regression estimators is very complicated for even the most basic designs (see e.g., Hausman and Wise 1981 and ten

Cate 1986). We therefore rely on a less analytic approach in this chapter. Through simulations we demonstrate that $\hat{\boldsymbol{\beta}}_{2sls}$ performs well in some informative sampling designs, but in other designs $\hat{\boldsymbol{\beta}}_{2sls}$ performs poorly because some large-weight sample observations have a greater influence on it than on other two-stage estimators that use additional IV's. We quantify the size of the large-weight effect by adopting the concept of Cook's distance from the context of standard regression diagnostics (see e.g., Cook and Weisberg 1982, Ch. 3 or Kutner, Nachtsheim, Neter, and Li 2005, Ch. 10). In the designs where $\hat{\boldsymbol{\beta}}_{2sls}$ performs poorly, the additional IV's increase the flexibility of the stage one "fitting" process and this can reduce the influential effect (i.e. Cook's distance) of the large-weight sampled observations. There is much literature about robust regression in the presence of outliers and there are also many articles about the use of survey weights in regression coefficient estimation, but the literature regarding regression coefficient estimation and the effect of large-weight sample observations under informative sampling is sparse.

CHAPTER 2

NONPARAMETRIC ENDOGENOUS POST-STRATIFICATION ESTIMATION

2.1 Introduction

Post-stratification (Särndal et al. 1992, Ch. 7.6) is the primary method in use today for improving the precision of survey estimators by calibrating the estimates to known population quantities. Calibration is achieved by adjusting the sample weights so that their totals over the strata match the stratum population counts, which is useful to ensure consistency between surveys and other data products released by government agencies. Calibration can facilitate interpretability of the sample weights, because the stratum counts are often highly visible quantities such as the sizes of important subpopulations. Improvement in precision is achieved when stratum membership has predictive power for the survey variables, since post-stratification is a form of model-assisted estimation with regression on categorical covariates. Relative to other calibration methods such as regression estimation or more general model-assisted estimation, post-stratification has the important practical advantages of simplicity and interpretability, often with only a modest loss in efficiency.

In order to post-stratify, categorical auxiliary information is required from sources external to the survey. In surveys of natural resources such as forest inventories, auxiliary information is often obtained from remote sensing data. These data are

typically not directly interpretable, since they are composed of reflectance values at different wavelengths and various indices derived from those values. Models are applied to the remote sensing data to transform them into more useful and interpretable quantities, such as predicted biomass or landcover types. The resulting derived variables are classified into categories, displayed as pixel-based maps and used in post-stratification for surveys. In particular, these are the methods used by the U.S. Forest Service in producing estimators for the Forest Inventory and Analysis (FIA; see Frayer and Furnival 1999). The FIA relies on post-stratification using classification maps derived from satellite imagery and other ancillary information. The assurance of some consistency between the maps derived from remote sensing data and estimates derived from field survey data is regarded as an important practical advantage of the method.

The models used for transformation of remote sensing variables into forestry-relevant variables are built using statistical methods and empirical data. In order to ensure the relevance and accuracy of the post-stratification variables with respect to the survey being post-stratified, the sample data themselves are a very attractive option for the model building. For example, the FIA data represent a source of high quality ground-level information of forest characteristics, so there is a clear desire for being "allowed" to use them in estimating the classification maps used later for post-stratification. However, in traditional survey theory, the post-stratification variables are considered fixed with respect to the population, and the stratum counts are assumed known without error. Using a model fitted on sample data to post-stratify the sample data violates these assumptions, so that existing results on post-stratification do not apply. Breidt and Opsomer (2008) coined the term *endogenous post-stratification estimation* (EPSE) for this scenario, and studied it for the case of a sample-fitted generalized linear model, from which the post-strata are constructed by dividing the range of the model predictions into predetermined

intervals. Under the generalized linear model set-up, Breidt and Opsomer (2008) obtained the design consistency of the endogenous post-stratification estimator for general unequal-probability sampling designs. Model consistency and asymptotic normality of the endogenous post-stratification estimator (EPSE) were also established, showing that EPSE has the same asymptotic variance as the traditional post-stratified estimator with fixed strata. Simulation experiments demonstrated that the practical effect of first fitting a model to the survey data before post-stratifying is small, even for relatively small sample sizes.

The results in Breidt and Opsomer (2008) provided some "weak justification" for using FIA data in estimating classification maps to be used for post-stratification (see Czaplewski 2010). The restriction of those results to parametric models limits their applicability in the FIA context, where the methods being used are often nonparametric in nature (e.g. Moisen and Frescino 2002). As a specific example of this, McRoberts et al. (2002) explored nearest-neighbor methods for creating strata for FIA, which effectively corresponds to using a nonparametric EPSE-like method even though it was not acknowledged as such.

In this chapter, we extend the EPSE methodology to the nonparametric estimation context, and hence strengthen the justification for inferential methods in current use by the U.S. Forest Service in FIA applications. We show here that the superpopulation results obtained for EPSE by Breidt and Opsomer (2008) continue to hold in this nonparametric setting, justifying the use of the nonparametric EPSE, the corresponding normal-theory confidence interval, and the standard variance estimator. We focus on the case where the underlying model is nonparametric but monotone, which is the most practically reasonable scenario in surveys since the model is used to divide the sample into homogeneous classes. Our theoretical results are valid for a general class of nonparametric estimators that includes kernel regression and penalized spline regression.

In the following section we give the definitions of the estimators we propose in this chapter. The asymptotic results are given in Section 2.3. Section 2.4 examines some of the models and estimators satisfying the outlined conditions, and in Section 2.5 we present both a numerical illustration and the results of a small simulation study. Application of the NEPSE methods to U.S. Forest Service data for a region of Utah appears in Section 2.6, followed by a discussion section. The proofs of the asymptotic results are collected in Section 2.8.

2.2 Definition of the estimator

Consider a finite population $U_N = \{1, \ldots, i, \ldots, N\}$. For each $i \in U_N$, an auxiliary vector $\boldsymbol{x}_i$ is observed. A probability sample s of size n is drawn from U_N according to a sampling design $p_N(\cdot)$, where $p_N(s)$ is the probability of drawing the sample s. Assume $\pi_{iN} = \Pr\{i \in s\} = \sum_{s:i\in s} p_N(s) > 0$ for all $i \in U_N$, and define $\pi_{ijN} = \Pr\{i, j \in s\} = \sum_{s:i,j\in s} p_N(s)$ for all $i, j \in U_N$. For compactness of notation we suppress the subscript N and write π_i, π_{ij} in what follows. Various study variables, generically denoted y_i, are observed for $i \in s$.

The targets of estimation are the finite population means of the survey variables, $\bar{y}_N = N^{-1} \sum_{U_N} y_i$. A purely design-based estimator (with all randomness coming exclusively from the selection of s) is provided by the Horvitz-Thompson estimator (HTE)

$$\bar{y}_\pi = \frac{1}{N} \sum_{i\in s} \frac{y_i}{\pi_i}.$$

Post-stratification (PS) and endogenous post-stratification are methods that take advantage of auxiliary information available for the population to improve the efficiency of design-based estimators. Following Breidt and Opsomer (2008), we first introduce some non-standard notation for PS that is useful in our later discussion of endogenous PS. Using the $\{\boldsymbol{x}_i\}_{i\in U_N}$ and a real-valued function $m(\cdot)$, a scalar index

$\{m(\boldsymbol{x}_i)\}_{i\in U_N}$ is constructed and used to partition U_N into H strata according to pre-determined stratum boundaries $-\infty \leq \tau_0 < \tau_1 < \cdots < \tau_{H-1} < \tau_H \leq \infty$. Typically, $m(\cdot)$ is the true relationship between a specific study variable z_i and the auxiliary variable/vector $\boldsymbol{x}_i$. We assume the additive error model

$$z_i = m(\boldsymbol{x}_i) + \sigma(\boldsymbol{x}_i)\epsilon_i, \tag{1}$$

where $\sigma^2(\boldsymbol{x}_i)$ is the unknown variance function, and $\mathrm{E}[\epsilon_i|\boldsymbol{x}_i] = 0, \mathrm{Var}(\epsilon_i|\boldsymbol{x}_i) = 1$. Breidt and Opsomer (2008) considered the particular case in which the index function $m(\cdot)$ is parameterized by a vector, $\boldsymbol{\lambda}$. We write $m_\lambda(\boldsymbol{x}_i)$ in that case.

For exponents $\ell = 0, 1, 2$ and stratum indices $h = 1, \ldots, H$, define

$$A_{Nh\ell}(m) = \frac{1}{N}\sum_{i\in U_N} y_i^\ell I_{\{\tau_{h-1}<m(\boldsymbol{x}_i)\leq\tau_h\}}$$

and

$$A^*_{Nh\ell}(m) = \frac{1}{N}\sum_{i\in U_N} y_i^\ell \frac{I_{\{i\in s\}}}{\pi_i} I_{\{\tau_{h-1}<m(\boldsymbol{x}_i)\leq\tau_h\}}, \tag{2}$$

where $I_{\{C\}} = 1$ if the event C occurs, and zero otherwise. In this notation, stratum h has population stratum proportion $A_{Nh0}(m)$, design-weighted sample post-stratum proportion $A^*_{Nh0}(m)$, and design-weighted sample post-stratum y-mean $A^*_{Nh1}(m)/A^*_{Nh0}(m)$. The traditional design-weighted PS estimator (PSE) for the population mean $\bar{y}_N = N^{-1}\sum_{i\in U_N} y_i$ is then

$$\begin{aligned}\hat{\mu}^*_y(m) &= \sum_{h=1}^{H} A_{Nh0}(m)\frac{A^*_{Nh1}(m)}{A^*_{Nh0}(m)} \\ &= \sum_{i\in s}\left\{\sum_{h=1}^{H} A_{Nh0}(m)\frac{N^{-1}\pi_i^{-1} I_{\{\tau_{h-1}<m(\boldsymbol{x}_i)\leq\tau_h\}}}{A^*_{Nh0}(m)}\right\} y_i = \sum_{i\in s} w^*_{is}(m) y_i, \end{aligned} \tag{3}$$

where the sample-dependent weights $\{w^*_{is}(m)\}_{i\in s}$ do not depend on $\{y_i\}$, and so can be used for any study variable.

For the important special case of equal-probability designs, in which $\pi_i = nN^{-1}$, we write

$$A_{nh\ell}(m) = \frac{1}{n}\sum_{i\in s} y_i^{\ell} I_{\{\tau_{h-1}<m(\boldsymbol{x}_i)\leq\tau_h\}}.$$

In this case, the equal-probability PSE for the population mean $\bar{y}_N$ is

$$\hat{\mu}_y(m) = \sum_{h=1}^{H} A_{Nh0}(m)\frac{A_{nh1}(m)}{A_{nh0}(m)} = \sum_{i\in s} w_{is}(m)y_i, \tag{4}$$

where the weights $\{w_{is}(m)\}_{i\in s}$ are obtained by substituting nN^{-1} for π_i in (3).

In parametric PS, the vector $\boldsymbol{\lambda}$ is known. In parametric endogenous PS, the vector $\boldsymbol{\lambda}$ is not known and needs to be estimated from the sample $\{\boldsymbol{x}_i, z_i : i \in s\}$ using, for example, maximum likelihood estimation or estimating equations. Thus, $m_{\lambda}(\boldsymbol{x}_i)$ is estimated by $m_{\hat{\lambda}}(\boldsymbol{x}_i)$, and the endogenous post-stratification estimator (EPSE) for the population mean $\bar{y}_N$ is then defined as

$$\hat{\mu}_y^*(m_{\hat{\lambda}}) = \sum_{h=1}^{H} A_{Nh0}(m_{\hat{\lambda}})\frac{A^*_{Nh1}(m_{\hat{\lambda}})}{A^*_{Nh0}(m_{\hat{\lambda}})} = \sum_{i\in s} w^*_{is}(m_{\hat{\lambda}})y_i.$$

This parametric EPSE was studied in Breidt and Opsomer (2008). We consider now the case where $m(\cdot)$ is not assumed to follow a specific parametric shape. Again, m is typically the true regression relationship between a specific study variable z_i and an auxiliary variable/vector $\boldsymbol{x}_i$ as in model (1).

The estimator $\hat{\mu}_y^*(m)$ is infeasible, because $m(\cdot)$ is unknown. We can estimate $m(\cdot)$ from the sample $\{(\boldsymbol{x}_i, z_i) : i \in s\}$ by nonparametric regression, and here we explicitly consider both kernel and spline-based methods. However, results should also apply to such other nonparametric and semi-parametric fitting methods as regression trees, neural nets, GAMs, etc. Writing $\hat{m}$ for the nonparametric estimator,

Table 1: Data for example of EPSE calculations for $n = 4$ sample from population with $N = 9$ and $\hat{m}(x)$ computed by ordinary least squares estimation of simple linear regression model.

$x_i, i \in U_N$	-3	-2	-1	-1	0	1	2	3	4
$z_i, i \in s$		-3		-1		1	3		
$\hat{m}(x_i) = 1.4x_i$	-4.2	-2.8	-1.4	-1.4	0	1.4	2.8	4.2	5.6
h	1	1	1	1	1	2	2	2	2

the nonparametric endogenous post-stratified estimator is then defined as

$$\hat{\mu}_y^*(\hat{m}) = \sum_{h=1}^{H} A_{Nh0}(\hat{m}) \frac{A_{Nh1}^*(\hat{m})}{A_{Nh0}^*(\hat{m})}. \tag{5}$$

For the special case of equal-probability designs, in which $\pi_i = nN^{-1}$, the equal-probability NEPSE for the population mean $\bar{y}_N$ is

$$\hat{\mu}_y(\hat{m}) = \sum_{h=1}^{H} A_{Nh0}(\hat{m}) \frac{A_{nh1}(\hat{m})}{A_{nh0}(\hat{m})} = \sum_{i \in s} w_{is}(\hat{m}) y_i. \tag{6}$$

To demonstrate the endogenous post-stratification calculations, we examine an equal-probability sample of size $n = 4$ selected from a finite population of size $N = 9$. Table 1 provides the data. As would be the case in practice, the auxiliary variable x_i is observed for all population elements, while the survey variable z_i is only observed for the sample elements. The HTE is $\bar{z}_\pi = 0$. Given the small sample size, we consider parametric EPSE with $\hat{m}$ obtained as the ordinary least squares fit of the simple linear regression model to the sample data $\{(x_i, z_i) : i \in s\}$, yielding $\hat{m}(x) = 0 + 1.4x$. A single boundary at $\tau_1 = 0.7$ divides the data into two strata based on the $\hat{m}(x_i)$ values. The quantities required to compute the EPSE in (6) are given by

	$A_{Nh0}(\hat{m})$	$A_{nh1}(\hat{m})$	$A_{nh0}(\hat{m})$
$h=1$	$5/9$	$\frac{1}{4}(-3+(-1))=-1$	$2/4$
$h=2$	$4/9$	$\frac{1}{4}(1+3)=1$	$2/4$

and the EPSE is

$$\hat{\mu}_z(\hat{m}) = \frac{5}{9}\frac{(-1)}{2/4} + \frac{4}{9}\frac{1}{2/4} = -\frac{2}{9}.$$

In the next section, we study the theoretical properties of the NEPSE. It is sufficient to consider the following simpler estimators

$$A_{\tau\ell}(\hat{m}) = \frac{1}{N}\sum_{i\in U_N} y_i^{\ell} I_{\{\hat{m}(\boldsymbol{x}_i)\leq\tau\}}$$

and

$$A^*_{\tau\ell}(\hat{m}) = \frac{1}{N}\sum_{i\in U_N} \frac{I_{\{i\in s\}}}{\pi_i} y_i^{\ell} I_{\{\hat{m}(\boldsymbol{x}_i)\leq\tau\}}$$

for a generic boundary value $\tau \in \{\tau_0, \tau_1, \ldots, \tau_H\}$. For equal probability designs we write

$$A_{n\tau\ell}(\hat{m}) = \frac{1}{n}\sum_{i\in s} y_i^{\ell} I_{\{\hat{m}(\boldsymbol{x}_i)\leq\tau\}}.$$

The form of these estimators suggests the use of tools from empirical process theory, which we turn to next.

2.3 Main results

2.3.1 Superpopulation model assumptions

Before we explicitly state the model assumptions for studying the NEPSE, we need the concept of *bracketing number* of empirical process theory (van der Vaart and Wellner 1996). For any $\varepsilon > 0$, any class $\mathcal{G}$ of measurable functions, and any norm $\|\cdot\|_{\mathcal{G}}$ defined on $\mathcal{G}$, $N_{[]}(\varepsilon, \mathcal{G}, \|\cdot\|_{\mathcal{G}})$ is the bracketing number, i.e., the minimal positive integer M for which there exist ε-brackets $\{[l_j, u_j] : \|l_j - u_j\|_{\mathcal{G}} \leq \varepsilon, \|l_j\|_{\mathcal{G}}, \|u_j\|_{\mathcal{G}} < \infty,$

$j = 1, \ldots, M\}$ to cover $\mathcal{G}$ (i.e., for each $g \in \mathcal{G}$, there is a $j = j(g) \in \{1, \ldots, M\}$ such that $l_j \leq g \leq u_j$).

We make the following superpopulation model assumptions.

Assumption 2.3.1. *The covariates $\{\boldsymbol{x}_i\}$ are independent and identically distributed random p-vectors with nondegenerate continuous joint probability density function $f(\boldsymbol{x})$ having compact support. The function $u \to \Pr(m(\boldsymbol{x}) \leq u)$ is Lipschitz continuous of order $0 < \gamma \leq 1$, and*

$$\Pr(m(\boldsymbol{x}) \leq \tau_{h-1}) < \Pr(m(\boldsymbol{x}) \leq \tau_h)$$

for $h = 1, \ldots, H$.

Assumption 2.3.2. *The sample s is selected according to an equal-probability design of fixed size n, with $\pi_i = nN^{-1} \to \pi \in [0, 1]$ as $N \to \infty$.*

Assumption 2.3.3. *The nonparametric estimator $\hat{m}(\cdot)$ satisfies*

$$\sup_{\boldsymbol{x}} |\hat{m}(\boldsymbol{x}) - m(\boldsymbol{x})| = o(1) \ a.s.$$

Assumption 2.3.4. *There exists a space $\mathcal{D}$ of measurable functions that satisfies $m \in \mathcal{D}$, $\Pr(\hat{m} \in \mathcal{D}) \to 1$ as $n \to \infty$, and*

$$\int_0^\infty \sqrt{\log N_{[]}(\lambda, \mathcal{F}, \|\cdot\|_2)}\, d\lambda < \infty,$$

where $\mathcal{F} = \{\boldsymbol{x} \to I_{\{d(\boldsymbol{x}) \leq \tau\}} : d \in \mathcal{D}\}$.

Assumption 2.3.5. *Given $[\boldsymbol{x}_i]_{i \in U_N}$, the study variables $[y_i]_{i \in U_N}$ are conditionally independent of the post-stratification variables $[z_i]_{i \in U_N}$, and $y_i \mid \boldsymbol{x}_i$ are conditionally independent random variables with $E(y_i^{2\ell} \mid \boldsymbol{x}_i) \leq K_1 < \infty$ for $\ell = 0, 1, 2$.*

These assumptions follow those of Section 3.2 in Breidt and Opsomer (2008), generalized to the nonparametric setting. Assumption 2.3.1 gives conditions on the multivariate distribution of covariates $\{\boldsymbol{x}_i\}$ and excludes degenerate situations in which some strata are empty. Assumption 2.3.2 restricts attention to equal probability sampling, and Assumptions 2.3.3 and 2.3.4 specify conditions on the sample fit $\hat{m}(\cdot)$. Finally, Assumption 2.3.5 gives moment conditions and specifies that the survey variables are independent of each other, conditionally on the auxiliary variables used in stratification. In Section 2.4, we discuss specific combinations of nonparametric models and estimators that satisfy them. As noted earlier, we focus on monotone models, because they are of primary interest in applications and because it is easier to establish Assumption 2.3.4. Intuitively, all that is required is that the class of functions is not too large, which is represented by the bracketing number of the class. When the class is too large, the bracketing integral in Assumption 2.3.4 fails to be finite. The class of monotone functions is one example of a well-behaved class, but other classes exist as well, including classes of functions that satisfy certain smoothness conditions. Consider for example the class $\mathcal{D} = C_M^{\alpha}(\mathcal{X})$ of all continuous functions $f : \mathcal{X} \to I\!R$ with $||f||_{\alpha} \leq M$, where

$$||f||_{\alpha} = \max_{k. \leq \underline{\alpha}} \sup_x |D^k f(x)| + \max_{k. = \underline{\alpha}} \sup_{x,y} \frac{|D^k f(x) - D^k f(y)|}{||x - y||^{\alpha - \underline{\alpha}}},$$

$\underline{\alpha}$ is the largest integer strictly smaller than α, $k = (k_1, \ldots, k_d)$, $D^k = \frac{\partial^{k.}}{\partial x_1^{k_1} \ldots \partial x_d^{k_d}}$, and $k. = \sum k_i$. Suppose that the support $\mathcal{X}$ of $\boldsymbol{x}$ is a bounded, convex subset of $I\!R^p$ with nonempty interior. Then it follows from Corollary 2.7.2 in van der Vaart and Wellner (1996) that $\log N_{[]}(\lambda, \mathcal{D}, \|\cdot\|_2) \leq K\lambda^{-p/\alpha}$ for some $0 < K < \infty$, and hence it can be easily seen that Assumption 2.3.4 holds provided $\alpha > p$.

2.3.2 Central limit theorem

For $\ell = 0, 1, 2$, define $\alpha_{\tau\ell}(m) = \mathrm{E}(y_i^\ell I_{\{m(\boldsymbol{x}_i)\leq\tau\}})$. We start with a crucial lemma that shows that $A_{\tau\ell}(\hat{m})$ (which is difficult to handle since it contains the nonparametric estimator $\hat{m}(\boldsymbol{x}_i)$ inside an indicator function) is asymptotically equivalent to $\mathrm{E}(y_i^\ell I_{\{\hat{m}(\boldsymbol{x}_i)\leq\tau\}} \mid \hat{m}) + A_{\tau\ell}(m) - \alpha_{\tau\ell}(m)$.

Lemma 1. *Under Assumptions 2.3.1–2.3.5, for* $\ell = 0, 1, 2$,

$$A_{\tau\ell}(\hat{m}) - E(y_i^\ell I_{\{\hat{m}(\boldsymbol{x}_i)\leq\tau\}} \mid \hat{m}) - A_{\tau\ell}(m) + \alpha_{\tau\ell}(m) = o_p(N^{-1/2}) \tag{7}$$

and

$$A_{n\tau\ell}(\hat{m}) - E(y_i^\ell I_{\{\hat{m}(\boldsymbol{x}_i)\leq\tau\}} \mid \hat{m}) - A_{n\tau\ell}(m) + \alpha_{\tau\ell}(m) = o_p(n^{-1/2}). \tag{8}$$

We are now ready to state the main result of the chapter.

Theorem 1. *Under Assumptions 2.3.1–2.3.5,*

$$\left\{\frac{1}{n}\left(1-\frac{n}{N}\right)\right\}^{-1/2}(\hat{\mu}_y(\hat{m}) - \bar{y}_N) \xrightarrow{d} N(0, V_{ym}),$$

where

$$V_{ym} = \sum_{h=1}^{H} \Pr\{\tau_{h-1} < m(\boldsymbol{x}_i) \leq \tau_h\}\mathrm{Var}(y_i|\tau_{h-1} < m(\boldsymbol{x}_i) \leq \tau_h).$$

The proofs of both results are deferred to Section 2.8.

2.3.3 Variance estimation

For the estimation of the variance V_{ym} we follow Breidt and Opsomer (2008).

Theorem 2. *If*

$$\hat{V}_{y\hat{m}} = \sum_{h=1}^{H} \frac{A_{Nh0}^2(\hat{m})}{A_{nh0}(\hat{m})} \frac{A_{nh2}(\hat{m}) - A_{nh1}^2(\hat{m})/A_{nh0}(\hat{m})}{A_{nh0}(\hat{m}) - n^{-1}}, \tag{9}$$

and Assumptions 2.3.1–2.3.5 hold,

$$\left\{\frac{1}{n}\left(1-\frac{n}{N}\right)\right\}^{-1/2} \hat{V}_{y\hat{m}}^{-1/2}(\hat{\mu}_y(\hat{m}) - \bar{y}_N) \xrightarrow{d} N(0,1).$$

The proof can be found in Section 2.8.

2.4 Applying the results

The results in the previous sections are expressed under quite general conditions on the class $\mathcal{D}$ and the estimator $\hat{m}$. We now give some particular models for the regression function m and some particular estimators $\hat{m}$ for which the conditions are satisfied. The underlying models we consider are at least partly monotone, which is reasonable in this context because the function m is used to split the data into homogeneous cells.

2.4.1 Monotone regression

Let

$$\mathcal{D} = \{d : R_X \to I\!R : d \text{ monotone and } \sup_{x \in R_X} |d(x)| \leq K\}$$

for some $K < \infty$, where R_X is a compact subset of $I\!R$. Suppose for simplicity that the functions in $\mathcal{D}$ are monotone decreasing. Then, the class $\mathcal{F}$ defined in Assumption 2.3.4 is itself a set of one-dimensional bounded and monotone functions, and hence

$$\log N_{[]}(\lambda, \mathcal{F}, \|\cdot\|_2) \leq K_1 \lambda^{-1}$$

for some $K_1 < \infty$, by Theorem 2.7.5 in van der Vaart and Wellner (1996). It follows that the integral in Assumption 2.3.4 is finite.

Let $\hat{m}$ be any estimator of m for which $\sup_{x \in R_X} |\hat{m}(x) - m(x)| = o(1)$ a.s. Then, provided the true regression function m is monotone and bounded, we have $\Pr(\hat{m} \in$

$\mathcal{D}) \to 1$ as $n \to \infty$. The estimator $\hat{m}$ does not need to be monotone itself, a classical local polynomial or spline estimator does the job. Hence, Theorem 1 applies in this case. Moreover, the case of generalized monotone regression functions, obtained by using e.g. a logit transformation, works as well. See Subsection 2.4.4 for more details.

2.4.2 Partially linear monotone regression

Consider now

$$\begin{aligned}\mathcal{D} &= \{R_X \to I\!R : (\boldsymbol{x}_1^T, x_2)^T \to \beta^T \boldsymbol{x}_1 + d(x_2) : \beta \in B \subset I\!R^k \text{ compact},\\ &\quad d \text{ monotone}, \sup_{x_2 \in R_{X_2}} |d(x_2)| \leq K\},\end{aligned}$$

where $R_X = R_{X_1} \times R_{X_2}$ is a compact subset of $I\!R^{k+1}$. Suppose for simplicity that all coordinates of an arbitrary $\boldsymbol{x}_1 \in R_{X_1}$ and $\beta \in B$ are positive. Divide B into $r = O(\lambda^{-2k})$ pairs (β_i^L, β_i^U) $(i = 1, \ldots, r)$ that cover B and are such that $\sum_{l=1}^k (\beta_{il}^U - \beta_{il}^L)^2 \leq \lambda^4$. Similarly, divide R_{X_1} into $s = O(\lambda^{-2k})$ pairs $(\boldsymbol{x}_{1j}^L, \boldsymbol{x}_{1j}^U)$ $(j = 1, \ldots, s)$ that cover R_{X_1} and are such that $\sum_{l=1}^k (x_{1jl}^U - x_{1jl}^L)^2 \leq \lambda^4$. Let $d_1^L \leq d_1^U, \ldots, d_q^L \leq d_q^U$ be the $q = O(\exp(K\lambda^{-1}))$ $\|\cdot\|_\infty$-brackets for the space of bounded and monotone functions (see Theorem 2.7.5 in van der Vaart and Wellner (1996)). Then, for each $\beta \in B$ and d monotone and bounded, there exist i, j and l such that, for all $(\boldsymbol{x}_1, x_2) \in R_X$,

$$\begin{aligned}\ell_{ijl}^L(x_2) &:= I_{\{\beta_i^{UT} \boldsymbol{x}_{1j}^U + d_l^U(x_2) \leq \tau\}}\\ &\leq I_{\{\beta^T \boldsymbol{x}_1 + d(x_2) \leq \tau\}}\\ &\leq I_{\{\beta_i^{LT} \boldsymbol{x}_{1j}^L + d_l^L(x_2) \leq \tau\}} := u_{ijl}^U(x_2).\end{aligned}$$

It is easy to see that the brackets $(\boldsymbol{x}_1, x_2) \to (\ell_{ijl}^L(x_2), u_{ijl}^U(x_2))$ are λ-brackets with respect to the $\|\cdot\|_2$-norm. The number of these brackets is bounded by $\lambda^{-4k} \exp(K\lambda^{-1})$, and hence the integral in Assumption 2.3.4 is finite.

The estimator $\hat{m}$ can, as in the previous example, be chosen as any uniformly consistent estimator of m. Then, $\Pr(\hat{m} \in \mathcal{D}) \to 1$ provided the true regression function m belongs to $\mathcal{D}$. This shows that Theorem 1 also holds for this case.

2.4.3 Single index monotone regression

Our next example concerns a single index model with a monotone link function. Let

$$\mathcal{D} = \{R_X \to I\!R : \boldsymbol{x} \to d(\beta^T \boldsymbol{x}) : \beta \in B \subset I\!R^k \text{ compact}, d \text{ monotone}, \sup_u |d(u)| \leq K\},$$

where R_X is a compact subset of $I\!R^k$. The treatment of this case is similar to that of the partial linear monotone regression model. We omit the details.

2.4.4 Generalized nonparametric monotone regression

The use of generalized linear models in EPSE was initially discussed in Breidt and Opsomer (2008). This approach enjoys the benefit of being able to handle categorical response variables, and has (in many cases) obvious and easily interpretable boundary values. Let the covariate x_i be univariate for ease of presentation, and write

$$\mathrm{E}(z_i|x_i) = \mu(x_i), \mathrm{Var}(z_i|x_i) = \sigma^2(x_i) := V(\mu(x_i)).$$

Consider the case of a known monotone link function $g(\cdot)$, such that $g(\mu(x_i)) = m(x_i)$, following the framework of McCullagh and Nelder (1989). The quasi-likelihood function $Q(\mu(x), z)$ satisfies

$$\frac{\partial}{\partial \mu(x)} Q(\mu(x), z) = \frac{z - \mu(x)}{V(\mu(x))},$$

as in McCullagh and Nelder (1989). The function $m(x)$ can be estimated nonparametrically, as suggested by Green and Silverman (1994) and Fan, Heckman, and

Wand (1995), among others.

Now approximate the function $m(x)$ locally by a pth-degree polynomial $m(x) \approx \beta_0 + \beta_1(x - x_i) + \cdots + \beta_p(x - x_i)^p$, and maximize the weighted quasi-likelihood to estimate the function $m(x)$ at each location x on the support of x_i, as suggested by Fan, Heckman, and Wand (1995),

$$\sum_{i \in s} \frac{1}{\pi_i} Q(g^{-1}(\beta_0 + \beta_1(x - x_i) + \cdots + \beta_p(x - x_i)^p), z_i) K_h(x_i - x), \tag{10}$$

where $K_h(\cdot) = \frac{1}{h} K(\cdot/h)$ and $K(\cdot)$ is a kernel function (for details, see Simonoff 1996 and Silverman 1999).

Let $(\hat{\beta}_{0x}, \hat{\beta}_{1x}, \ldots, \hat{\beta}_{px})$ be the minimizer of (10). Then $\hat{m}(x) = \hat{\beta}_{0x}$, and $\hat{\mathrm{E}}(z|X = x) = g^{-1}(\hat{m}(x)) = g^{-1}(\hat{\beta}_{0x})$. One can retain the boundary values for variable z, $\{\tau_0, \tau_1, \ldots, \tau_H\}$, and define $A^*_{Nh\ell}(\hat{m})$ as in (2):

$$A^*_{Nh\ell}(\hat{m}) = \frac{1}{N} \sum_{i \in U_N} y_i^{\ell} \frac{I_{\{i \in s\}}}{\pi_i} I_{\{\tau_{h-1} < g^{-1}(\hat{m}(\boldsymbol{x}_i)) \leq \tau_h\}}, \tag{11}$$

for $l = 0, 1, 2$. Given (11), a natural estimator for the population mean $\bar{y}_N$ is the same as in (5). The verification of Assumptions 2.3.3 and 2.3.4 is similar to the verification in Subsection 2.4.1, and is therefore omitted.

2.5 Simulations

2.5.1 Numerical example

In Section 2.2, we illustrated the endogenous post-stratification calculations with a linear regression example. To demonstrate the more interesting use of nonparametric regression, we briefly discuss a second small example with penalized splines, as justified in Subsection 2.4.1. Figure 3 shows data for an equal-probability sample of size $n = 25$ selected from a finite population of size $N = 100$. Here, $\hat{m}$ is estimated

using the sample data $\{(x_i, z_i) : i \in s\}$, a penalized spline with 10 knots, and a smoothing parameter which allows approximately five degrees of freedom. A single boundary at $\tau_1 = 0.44$ divides the data into two strata based on the $\hat{m}(x_i)$ values. The "rug" lines at the bottom of the graph indicate the known x_i values for $i \in U_N$. Using the notation of Section 2.2, we have the tabled values

	$A_{Nh0}(\hat{m})$	$A_{nh1}(\hat{m})$	$A_{nh0}(\hat{m})$
$h = 1$	$\frac{1}{100}(30)$	$\frac{1}{25}(0.24)$	$\frac{1}{25}(8)$
$h = 2$	$\frac{1}{100}(70)$	$\frac{1}{25}(24.41)$	$\frac{1}{25}(17)$

,

where 0.24 and 24.41 are the sums of the sample z_i values in each stratum. Based on this, the HTE is $\bar{z}_\pi = 0.99$ and the estimated mean using (6) is

$$\hat{\mu}_z(\hat{m}) = \frac{1}{100}(30)\frac{0.24}{8} + \frac{1}{100}(70)\frac{24.41}{17} = 1.01.$$

2.5.2 Monte Carlo study

The main goal of the simulation was to assess the design efficiency of the NEPSE relative to competing survey estimators. The simulations were performed in a setting that mimics a survey in which characteristics of multiple study variables are estimated using one set of weights. We considered several different sets of weights for estimation of a mean: the Horvitz-Thompson estimator (HTE) weights $\{n^{-1}\}_{i\in s}$, the PSE weights $\{w_{is}(m)\}_{i\in s}$, the NEPSE weights $\{w_{is}(\hat{m})\}_{i\in s}$, and the simple linear regression (REG) weights (e.g. Särndal et al. 1992, equation (6.5.12)). We used $H = 4$ strata with fixed, known boundaries $\boldsymbol{\tau} = (-\infty, 0.5, 1.0, 1.5, \infty)$ for PSE and NEPSE. The HTE did not use auxiliary information; the PSE used auxiliary information with a known model; the REG used auxiliary information with a fitted parametric model, and the NEPSE used auxiliary information with a fitted nonparametric model. Specifically, we used a linear penalized spline with approximate

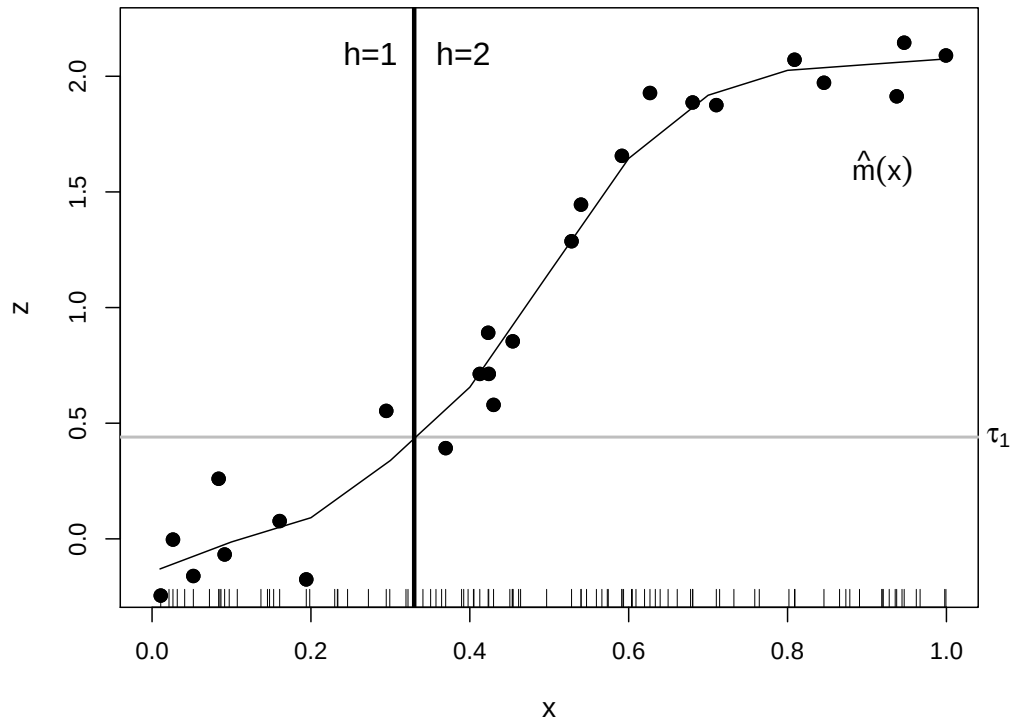

Figure 3: Equal-probability sample of $n = 25$ (x_i, z_i) values from a finite population of size $N = 100$ fitted with a penalized spline, $\hat{m}$, with ten knots and five degrees of freedom. "Rug" lines at the bottom of the graph represent x_i for $i \in U_N$. Boundary value τ_1 determines the strata, $h = 1$ and $h = 2$.

degrees of freedom determined by the smoothing parameter (Ruppert et al. 2003, §3.13).

We generated a population of size $N = 1000$ with eight survey variables of interest. The values $x_1, \ldots, x_N$ were independent and uniformly distributed on $(0, 1)$. The first variable, `ratio`, was generated according to a regression through the origin or ratio model (see e.g. Särndal et al. 1992, p.226), with mean $1 + 2(x - 0.5)$ and with independent normal errors with variance $2\sigma^2 x$. For the next six variables (y_i), we took their mean functions to be

$$2\frac{g_k(x) - \min_{x\in[0,1]} g_k(x)}{\max_{x\in[0,1]} g_k(x) - \min_{x\in[0,1]} g_k(x)}$$

where

$$\begin{aligned}
\texttt{quad}: &\quad g_1(x) = 1 + 2(x - 0.5)^2 \\
\texttt{bump}: &\quad g_2(x) = 1 + 2(x - 0.5) + \exp(-200(x - 0.5)^2) \\
\texttt{jump}: &\quad g_3(x) = \{1 + 2(x - 0.5)\}I_{\{x\le 0.65\}} + 0.65I_{\{x>0.65\}} \\
\texttt{expo}: &\quad g_4(x) = \exp(-8x) \\
\texttt{cycle1}: &\quad g_5(x) = 2 + \sin(2\pi x) \\
\texttt{cycle4}: &\quad g_6(x) = 2 + \sin(8\pi x).
\end{aligned}$$

This means that the minimum was 0 and the maximum was 2 for each of the first seven mean functions. Finally, the eighth survey variable was

$$\texttt{noise}: \quad g_7(x) = 8.$$

Independent normal errors with mean zero and variance equal to σ^2 were then added to each of these mean functions. The variance function for the `ratio` model was chosen so that, averaging over the covariate x, we had $\mathrm{E}[v(x)] = \sigma^2$. Thus, the heteroskedastic `ratio` variable and the remaining seven study variables all had the same variance, averaged over x. See Figure 4 for examples of the population graphs.

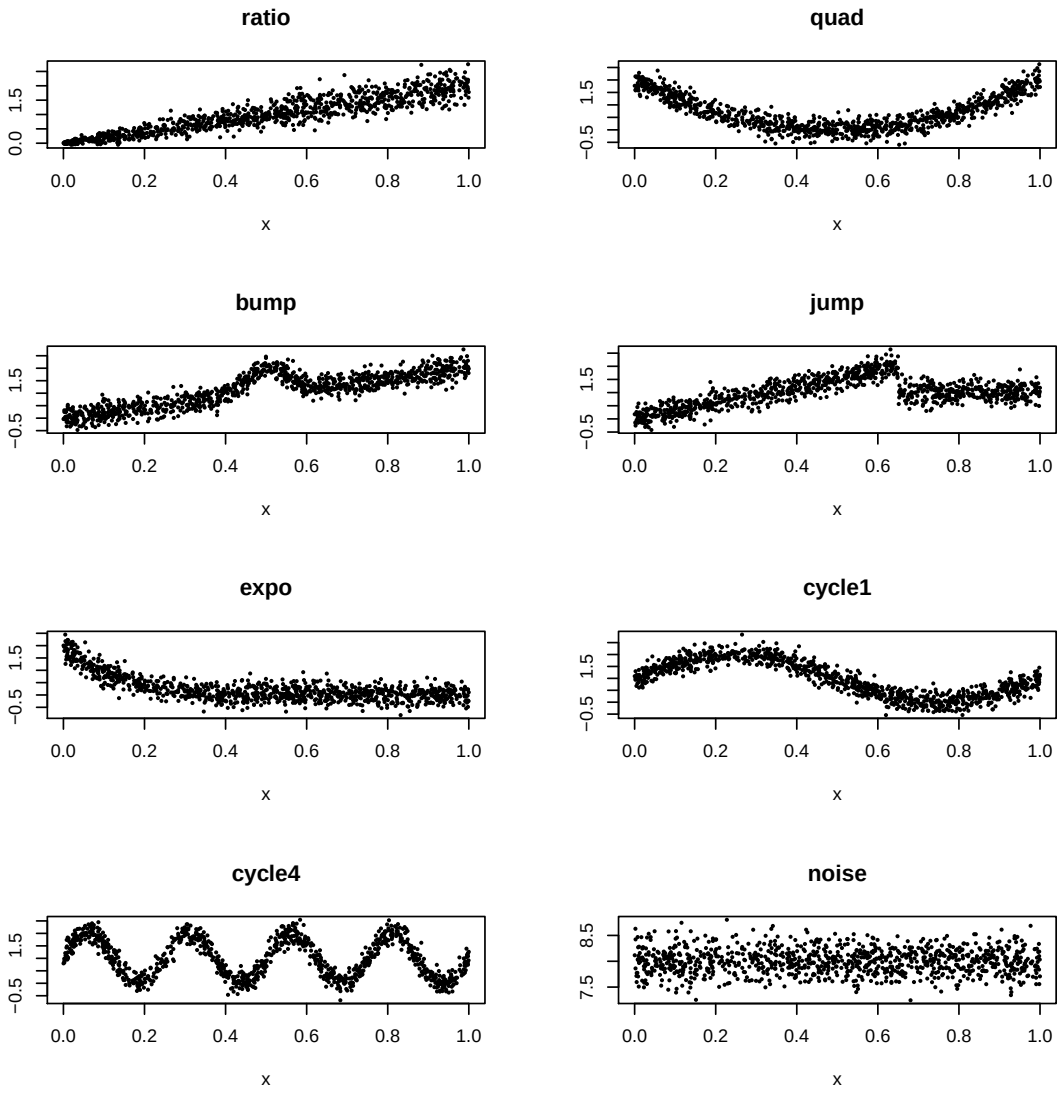

Figure 4: Example population graphs of the eight survey variables of interest

For given values of σ, we fixed the population (that is, simulated N values for each of the eight variables of interest) and drew 1000 replicate samples of size n, each via simple random sampling without replacement from this fixed population. We constructed HTE and REG weights using standard methods. We then computed the ratio of the MSE for each competing estimator to that of the NEPSE.

In the first simulation experiment, we consider in detail the case in which the PS variable follows a regression through the origin or ratio model. We used the `ratio` variable as the PS variable and computed PSE weights with known $m(x) = 1 + 2(x - 0.5)$ and NEPSE weights with (approximately) 2 or 5 degrees of freedom (df) in the smoothing spline. The weights were then applied to the remaining seven study variables. We also varied the noise variance ($\sigma = 0.25$ or $\sigma = 0.5$). With 2 df, the smoothing spline yields the linear (parametric) fit, and thus corresponds to EPSE. Results for this case, presented in Table 2, are qualitatively similar to those in Table 1 of Breidt and Opsomer (2008) (the results are different because the earlier paper fits regression through the origin instead of simple linear regression, and uses different signal-to-noise ratios since the mean functions are not scaled to [0,2]).

NEPSE dominates HTE in every case except `cycle4` (since NEPSE does not have enough df to capture the four cycles and so its estimate of the mean function is oversmoothed and nearly constant) and `noise`, where NEPSE fits an entirely superfluous model. REG beats NEPSE for `ratio`, where REG has the correct working model, and is slightly better for `bump`, which is highly linear over most of its range. REG is also slightly better for `cycle4` and for `noise`. NEPSE performs far better than REG for all of the other variables.

The effect of changing degrees of freedom in NEPSE is negligible in this example, since the true model for the PS variable is in fact linear. The effect of increasing noise variance is quite substantial, bringing the performance of all estimators closer together, as expected. Finally, NEPSE is essentially equivalent to the PSE in terms

Table 2: Ratio of MSE of Horvitz-Thompson (HTE), post-stratification on 4 strata (PSE(4)), and linear regression (REG) estimators to MSE of nonparametric endogenous post-stratification estimator on 4 strata (NEPSE(4)). Numbers greater than one favor NEPSE. Based on `ratio` post-stratification variable in 1000 replications of simple random sampling of size $n = 50$ from a fixed population of size $N = 1000$. Replications in which at least one stratum had fewer than two samples are omitted from the summary: 4 reps at $df \approx 2$, $\sigma = 0.5$ and 33 reps at $df \approx 5$, $\sigma = 0.5$.

		($\sigma = 0.25$)			($\sigma = 0.5$)		
Response		NEPSE(4) versus			NEPSE(4) versus		
Variable	$df \approx$	HTE	PSE(4)	REG	HTE	PSE(4)	REG
`ratio`	2	4.98	1.01	0.74	2.19	1.02	0.91
	5	4.68	0.95	0.69	2.21	1.03	0.91
`quad`	2	2.34	1.03	2.56	1.62	1.05	1.75
	5	2.29	1.01	2.51	1.50	0.97	1.62
`bump`	2	3.22	1.00	0.94	1.88	1.00	0.95
	5	3.26	1.01	0.95	1.90	1.02	0.96
`jump`	2	2.19	1.00	1.80	1.40	0.99	1.26
	5	2.13	0.97	1.76	1.33	0.94	1.20
`expo`	2	1.88	0.99	1.17	1.29	1.01	1.07
	5	1.88	0.99	1.17	1.28	1.01	1.06
`cycle1`	2	3.10	1.04	1.56	1.97	1.03	1.26
	5	3.04	1.02	1.53	1.96	1.02	1.25
`cycle4`	2	0.96	1.00	0.92	0.98	1.02	0.95
	5	0.98	1.02	0.94	1.00	1.05	0.98
`noise`	2	0.93	1.00	0.96	0.92	1.00	0.96
	5	0.92	0.99	0.95	0.93	1.01	0.97

of design efficiency, even for $n = 50$, implying that the effect of basing the PS on a nonparametric regression instead of on stratum classifications and stratum counts known without error from a source external to the survey is negligible for moderate to large sample sizes.

In the second simulation, we fixed $n = 100$, $df \approx 5$, $\sigma = 0.25$ and considered four different PS variables: `ratio`, `quad`, `bump`, and `cycle1`. The latter three allowed us to investigate the behavior of NEPSE when monotonicity did not hold. Table 3

summarizes the design efficiency results as ratios of the MSE of the HTE, PSE(4), or REG over the MSE of the NEPSE(4). Overall, the behavior of the NEPSE is consistent with expectations. Even for the non-monotone functions, NEPSE produces a large improvement in efficiency relative to the HTE for the variable on which the PS is based, and usually for other variables as well. NEPSE is as good or better (i.e. MSE ratio > 0.95) than REG in all but 12 of the 32 cases considered: NEPSE loses out in particular when the true model is linear or nearly so (`bump`). The `noise` variable shows that, when a variable is not related to the stratification variable, the efficiency is near that of the HTE (since the stratification is unnecessary).

We also assessed the coverage of confidence intervals computed using the normal approximation from Theorem 1 and the variance estimator from Theorem 2. Coverage of nominal 95% confidence intervals, $\hat{\mu}_y(\hat{m}) \pm 1.96\{n^{-1}(1 - nN^{-1})\hat{V}_{y\hat{m}}\}^{1/2}$, was consistently in the range of 93% to 96%.

2.6 Application

We illustrate the applicability of the NEPSE approach using pilot study data collected by the U.S. Forest Service in a region of Utah. The field-based data collection methods and variables are similar to those currently in use in the Forest Inventory and Analysis (FIA) program, while the remote sensing variables are among those being considered as post-stratification variables in this context (see e.g. Blackard et al. 2008). FIA is the primary source of information in the United States for assessing status and trends in forested areas, including size, health, growth, mortality, and removals of trees by species. The pilot study is designed to assess the increased use of remote sensing information in the inventory.

The population in this example is a set of $N = 1,707$ 90m$\times$90m plots that are classified as forest and for which extensive remote-sensing data are available. The $n = 250$ sample plots were selected with equal probability from that population, and

Table 3: Ratio of MSE of Horvitz-Thompson (HTE), post-stratification on 4 strata (PSE(4)), and linear regression (REG) estimators to MSE of nonparametric endogenous post-stratification estimator on 4 strata (NEPSE(4)). Numbers greater than one favor NEPSE. Based on four different PS variables in 1000 replications of simple random sampling of size $n = 100$ from a fixed population of size $N = 1000$.

PS Variable	Estimator	ratio	quad	bump	jump	expo	cycle1	cycle4	noise
	HTE	5.17	2.46	3.48	2.12	2.13	3.31	0.99	0.95
ratio	PSE(4)	0.98	1.03	1.02	0.97	1.01	1.02	1.00	1.00
	REG	0.71	2.49	0.97	1.70	1.19	1.64	0.90	0.97
	HTE	0.97	5.47	1.01	1.53	1.31	0.97	0.98	0.96
quad	PSE(4)	1.01	1.00	1.02	1.02	1.04	1.00	1.03	0.99
	REG	0.13	5.53	0.28	1.23	0.73	0.48	0.89	0.98
	HTE	4.07	1.93	4.13	2.02	2.30	2.70	1.13	0.95
bump	PSE(4)	1.27	1.33	0.76	1.07	1.11	0.96	1.05	1.00
	REG	0.56	1.95	1.15	1.62	1.29	1.34	1.03	0.97
	HTE	2.89	1.01	2.53	1.26	1.35	5.68	1.00	0.97
cycle1	PSE(4)	1.01	1.00	1.06	1.04	0.96	0.92	1.03	1.01
	REG	0.40	1.02	0.70	1.01	0.75	2.81	0.91	0.99

a large number of field-based variables were measured on those plots. We considered variables that are representative of the variables typically collected as part of the FIA: basal area of live trees per acre (`BA`), net annual growth of sound live trees (`GROW`), stand age (`STAG`), and a binary forest type code (`FOTP`), chosen here as "Aspen" (code 901). We constructed the NEPSE post-strata using `BA`, since this is a commonly used forestry indicator for the amount of harvestable wood on a plot and is a key FIA variable. From the remote sensing data, we chose as the auxiliary variable the so-called *Greenness index* (`GREEN`). This is a frequently used summary of reflectances at different frequencies with good predictive properties for forestry variables (Crist and Cicone 1984). As in traditional post-stratification, we then applied the resulting NEPSE weights to all of the other survey variables.

As in the simulation study, a linear penalized spline was used in the regression

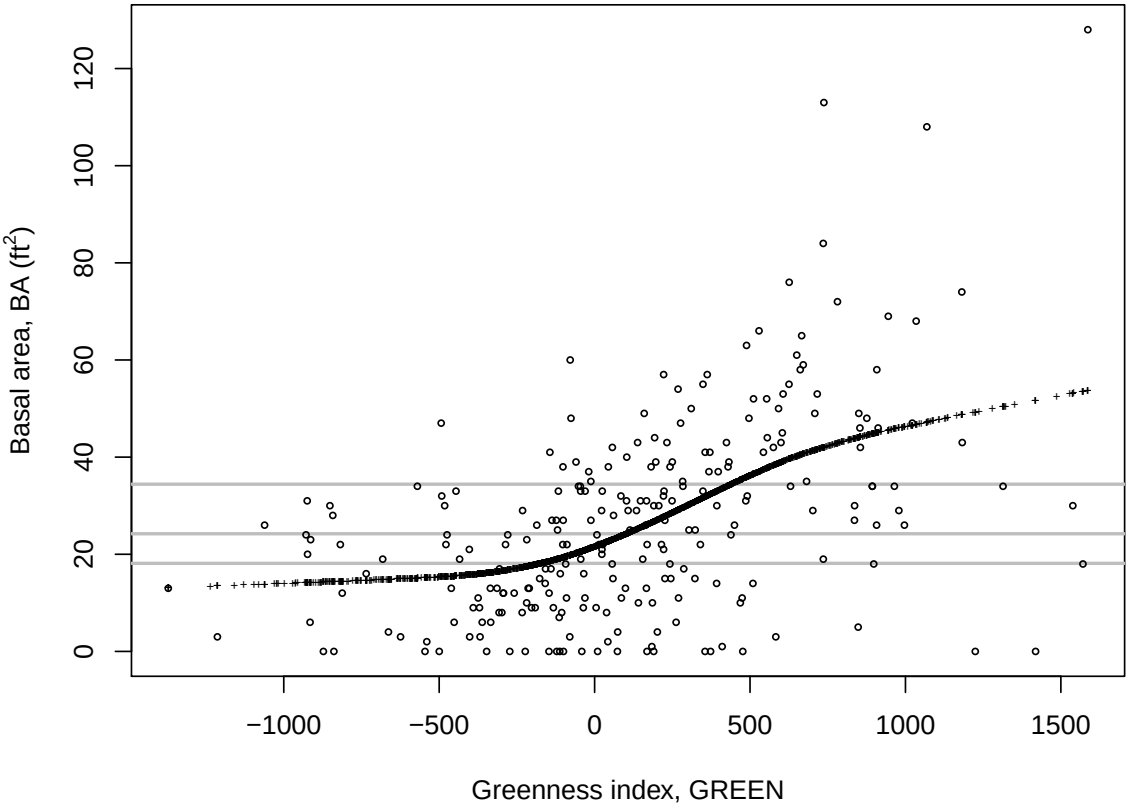

Figure 5: `BA` vs. `GREEN` values ($n = 250$) for a U.S. Forest Service pilot study in Utah. Plus signs (+) indicate penalized spline fitted values, $\{\hat{m}(x_i)\}_{i=1}^{N}$, using four degrees of freedom, where $N = 1,707$. Gray lines are boundaries for the case of four post-strata, based on quartiles of the fitted values.

of `BA` on `GREEN` to form the nonparametric endogenous post-strata. For comparison, the data were analyzed at two levels of degrees of freedom and for four different numbers of strata. The degrees of freedom levels were determined by adjusting the smoothing parameter and the strata were determined by using appropriate quantiles of the $\{\hat{m}(x_i)\}_{i=1}^{N}$ values. For comparison, we also applied the Horvitz-Thompson estimator (HTE) that does not use any auxiliary information.

Figure 5 shows the $n = 250$ `BA` versus `GREEN` values, plotted as open circles, for the Utah pilot study data. Also shown are '+' symbols indicating the penalized spline fitted values, $\{\hat{m}(x_i)\}_{i=1}^{N}$, using four degrees of freedom. The three gray lines indicate

Table 4: Estimates of finite population means for BA, GROW, and STAG, and estimated population proportion of Aspen (FOTP = 901) with estimated standard errors in parentheses. The numbers in parentheses after "NEPSE" indicate the number of strata.

Estimator	BA, ft^2	GROW, ft^3	STAG	FOTP
HTE	26.80 (1.20)	7.92 (1.82)	112.98 (4.87)	0.088 (0.017)
$df = 4$				
NEPSE(2)	26.92 (1.09)	8.02 (1.79)	112.64 (4.63)	0.089 (0.016)
NEPSE(4)	26.30 (0.98)	7.49 (1.67)	114.27 (4.67)	0.080 (0.014)
NEPSE(8)	26.23 (0.95)	6.98 (1.65)	114.12 (4.69)	0.072 (0.013)
NEPSE(16)	26.07 (0.97)	6.98 (1.63)	113.16 (4.77)	0.068 (0.012)
$df = 8$				
NEPSE(2)	26.92 (1.09)	8.02 (1.79)	112.64 (4.63)	0.089 (0.016)
NEPSE(4)	26.30 (0.98)	7.49 (1.67)	114.27 (4.67)	0.080 (0.014)
NEPSE(8)	26.31 (0.99)	7.32 (1.75)	114.31 (4.70)	0.073 (0.013)
NEPSE(16)	26.04 (1.01)	7.12 (1.70)	113.67 (4.75)	0.072 (0.012)

the post-stratum boundaries for the four-stratum case, computed as the quartiles of the fitted values. In this case, the relationship is monotone but nonlinear, so that this application falls under the setting of Subsection 2.4.1. In actual large-scale forestry survey practice, additional auxiliary variables can be expected to be available and more complicated models would undoubtedly be required.

Table 4 shows the estimates and estimated standard deviations for the four forestry variables considered, using NEPSE and HTE. At both *df* levels and for all numbers of strata, the estimated standard error for each variable is smaller for NEPSE than for HTE. The results are reasonably insensitive to the amount of smoothing and the number of post-strata. Averaging across these factors, the HTE has standard error averaging 19% higher than NEPSE for BA, 7% higher for GROW, 4% higher for STAG, and 25% higher for FOTP.

In this particular illustration, the NEPSE-derived post-strata could be interpreted as corresponding to levels of (predicted) tree basal area per acre (e.g. thinly

stocked stratum vs. heavily stocked stratum), facilitating interpretation by forest scientists and other users of FIA data. While a single covariate, `GREEN`, was used here, in actual large-scale forestry survey practice, additional auxiliary variables can be expected to be available and more complicated models would undoubtedly be applied. The interpretation of the strata would remain the same, which is a strong practical advantage of NEPSE. More sophisticated models are also likely to result in increased efficiency, and hence a larger decrease in the estimated standard errors relative to HTE, compared to that seen in Table 4.

2.7 Discussion

In this article, we have obtained the theoretical properties of NEPSE, a new post-stratification-based estimator that uses a sample-fitted nonparametric index to create the post-strata. The finite-sample properties of the estimator are shown in a simulation study, and the applicability of the method is illustrated on a forestry dataset.

There are a number of open issues related to implementation of NEPSE in surveys. Perhaps most importantly, the choice of the number of strata and the selection of the boundaries are of clear interest to practitioners. As noted above, we expect that in many situations these will be dictated by the application. Nevertheless, a data-driven approach that provides guidance in this respect would be desirable, and is currently being investigated.

2.8 Proofs

Proof of Lemma 1. The expression on the left hand side of (7) is

$$N^{-1} \sum_{i \in U_N} \{ y_i^\ell I_{\{\hat{m}(\boldsymbol{x}_i) \le \tau\}} - y_i^\ell I_{\{m(\boldsymbol{x}_i) \le \tau\}} - \mathrm{E}[y_i^\ell I_{\{\hat{m}(\boldsymbol{x}_i) \le \tau\}} \mid \hat{m}] + \mathrm{E}[y_i^\ell I_{\{m(\boldsymbol{x}_i) \le \tau\}}] \}.$$

Let

$$\begin{aligned}\mathcal{H} &= \{(\boldsymbol{x},y) \to y^\ell I_{\{d(\boldsymbol{x})\le\tau\}} - y^\ell I_{\{m(\boldsymbol{x})\le\tau\}} \\ &\quad -\mathrm{E}[y^\ell I_{\{d(\boldsymbol{x})\le\tau\}}] + \mathrm{E}[y^\ell I_{\{m(\boldsymbol{x})\le\tau\}}] : d\in\mathcal{D}\},\end{aligned}$$

where $\mathcal{D}$ is as in Assumption 2.3.4.

In a first step we show that the class $\mathcal{H}$ is Donsker. From Theorem 2.5.6 in van der Vaart and Wellner (1996), it suffices to show that

$$\int_0^\infty \sqrt{\log N_{[]}(\lambda,\mathcal{H},\|\cdot\|_2)}\,d\lambda < \infty. \tag{12}$$

From Assumption 2.3.4 we know that the class

$$\mathcal{F} = \{(\boldsymbol{x},y) \to y^\ell I_{\{d(\boldsymbol{x})\le\tau\}} : d\in\mathcal{D}\}$$

satisfies (12) with $\mathcal{H}$ replaced by $\mathcal{F}$, and hence the same holds for $\mathcal{H}$ itself, since the three other terms in $\mathcal{H}$ do not change its bracketing number.

Let

$$\hat{h}(\boldsymbol{x},y) = y^\ell\left(I_{\{\hat{m}(\boldsymbol{x})\le\tau\}} - I_{\{m(\boldsymbol{x})\le\tau\}}\right) - \mathrm{E}\left[y^\ell\left(I_{\{\hat{m}(\boldsymbol{x})\le\tau\}} - I_{\{m(\boldsymbol{x})\le\tau\}}\right)\Big|\,\hat{m}\right],$$

where $(\boldsymbol{x},y)$ is independent of the fit, $\hat{m}(\cdot)$. Then

$$\begin{aligned}&\mathrm{Var}\left(\hat{h}(\boldsymbol{x},y)\mid\hat{m}\right)\\ &= \mathrm{Var}\left(y^\ell\left(I_{\{\hat{m}(\boldsymbol{x})\le\tau\}} - I_{\{m(\boldsymbol{x})\le\tau\}}\right)\Big|\,\hat{m}\right)\\ &\le \mathrm{E}\left[\left(y^\ell\left(I_{\{\hat{m}(\boldsymbol{x})\le\tau\}} - I_{\{m(\boldsymbol{x})\le\tau\}}\right)\right)^2\Big|\,\hat{m}\right]\\ &= \mathrm{E}\left[y^{2\ell}\left(I_{\{\hat{m}(\boldsymbol{x})\le\tau\}} - I_{\{m(\boldsymbol{x})\le\tau\}}\right)^2\Big|\,\hat{m}\right]\\ &= \mathrm{E}\left[\mathrm{E}\left[y^{2\ell}\left(I_{\{\hat{m}(\boldsymbol{x})\le\tau\}} - I_{\{m(\boldsymbol{x})\le\tau\}}\right)^2\Big|\,\hat{m},\boldsymbol{x}\right]\Big|\,\hat{m}\right]\end{aligned}$$

$$
\begin{aligned}
&= \mathrm{E}\left[\mathrm{E}[y^{2\ell} \mid \hat{m}, \boldsymbol{x}]\left(I_{\{\hat{m}(\boldsymbol{x})\leq\tau\}} - I_{\{m(\boldsymbol{x})\leq\tau\}}\right)^2 \middle| \hat{m}\right] \\
&= \mathrm{E}\left[\mathrm{E}[y^{2\ell} \mid \boldsymbol{x}]\left(I_{\{\hat{m}(\boldsymbol{x})\leq\tau\}} - I_{\{m(\boldsymbol{x})\leq\tau\}}\right)^2 \middle| \hat{m}\right] \\
&\leq K_1 \left\{\Pr(\hat{m}(\boldsymbol{x}) \leq \tau, m(\boldsymbol{x}) > \tau \mid \hat{m}) + \Pr(\hat{m}(\boldsymbol{x}) > \tau, m(\boldsymbol{x}) \leq \tau \mid \hat{m})\right\}, \quad (13)
\end{aligned}
$$

where K_1 is given in Assumption 2.3.5. Let $\epsilon > 0$ be given. By Assumption 2.3.1, $F(u) = \Pr(m(\boldsymbol{x}) \leq u)$ is uniformly continuous, so there exists $\delta > 0$ such that $|u_1 - u_2| \leq \delta$ implies $|F(u_1) - F(u_2)| < \epsilon$. We show that $\Pr(\hat{m}(\boldsymbol{x}) \leq \tau, m(\boldsymbol{x}) > \tau \mid \hat{m}) = o_p(1)$. Consider

$$
\begin{aligned}
&\Pr\Big(\Pr(\hat{m}(\boldsymbol{x}) \leq \tau, m(x) > \tau \mid \hat{m}) > \epsilon\Big) \\
&\leq \Pr\Big(\Pr(\hat{m}(\boldsymbol{x}) \leq \tau, m(x) > \tau \mid \hat{m}) > \epsilon, \sup_{\boldsymbol{x}} |\hat{m}(\boldsymbol{x}) - m(\boldsymbol{x})| \leq \delta\Big) \\
&\quad + \Pr\Big(\sup_{\boldsymbol{x}} |\hat{m}(\boldsymbol{x}) - m(\boldsymbol{x})| > \delta\Big) \\
&\leq \Pr\Big(\Pr(m(\boldsymbol{x}) - \delta \leq \tau, m(\boldsymbol{x}) > \tau \mid \hat{m}) > \epsilon\Big) + o(1) \\
&= \Pr\Big(\Pr(m(\boldsymbol{x}) - \delta \leq \tau, m(\boldsymbol{x}) > \tau) > \epsilon\Big) + o(1) \\
&= I_{\{F(\tau+\delta)-F(\tau)>\epsilon\}} + o(1) = o(1), \quad (14)
\end{aligned}
$$

by choice of δ, where the second inequality follows from Assumption 3.1.3. Similarly,

$$
\Pr(\hat{m}(\boldsymbol{x}) > \tau, m(\boldsymbol{x}) \leq \tau \mid \hat{m}) = o_p(1). \quad (15)
$$

For fixed $\eta > 0, \lambda > 0$ consider

$$
\begin{aligned}
&\Pr\left(N^{1/2}|A_{\tau\ell}(\hat{m}) - \mathrm{E}[y_i^\ell I_{\{\hat{m}(\boldsymbol{x}_i)\leq\tau\}} \mid \hat{m}] - A_{\tau\ell}(m) + \alpha_{\tau\ell}(m)| > \lambda\right) \\
&= \Pr\left(N^{-1/2}\left|\sum_{i\in U_N} \hat{h}(\boldsymbol{x}_i, y_i)\right| > \lambda\right) \\
&\leq \Pr\left(N^{-1/2}\left|\sum_{i\in U_N} \hat{h}(\boldsymbol{x}_i, y_i)\right| > \lambda, \mathrm{Var}(\hat{h}(\boldsymbol{x}, y) \mid \hat{m}) < \eta, \hat{m} \in \mathcal{D}\right)
\end{aligned}
$$

$$
\begin{aligned}
&+ \Pr\left(N^{-1/2} \left| \sum_{i \in U_N} \hat{h}(\boldsymbol{x}_i, y_i) \right| > \lambda, \mathrm{Var}(\hat{h}(\boldsymbol{x}, y) \mid \hat{m}) \geq \eta, \hat{m} \in \mathcal{D} \right) \\
&+ \Pr\left(\hat{m} \notin \mathcal{D} \right) \\
\leq\ & \Pr\left(\sup_{h \in \mathcal{H}, \mathrm{Var}(h) < \eta} N^{-1/2} \left| \sum_{i \in U_N} h(\boldsymbol{x}_i, y_i) \right| > \lambda \right) \\
&+ \Pr\left(\mathrm{Var}(\hat{h}(\boldsymbol{x}, y) \mid \hat{m}) \geq \eta \right) + \Pr\left(\hat{m} \notin \mathcal{D} \right) \\
=\ & d_{1N} + d_{2N} + d_{3N}.
\end{aligned}
$$

As $N \to \infty$, $d_{1N} = o(1)$ as $\eta \downarrow 0$ by Corollary 2.3.12 in van der Vaart and Wellner (1996) and the fact that $\mathcal{H}$ is Donsker. Also, $d_{2N} = o(1)$ by the arguments in (13)–(15), and $d_{3N} = o(1)$ by Assumption 2.3.4. This establishes (7), and similar arguments verify (8). □

Proof of Theorem 1. We start with a statement about equal-probability PSE (i.e., $m(\cdot)$ known) error that is used later for comparison. Since $\bar{y}_N = \sum_{h=1}^{H} A_{Nh1}(\boldsymbol{\gamma})$ for any $\boldsymbol{\gamma}$, we subtract from (4) to obtain

$$
\begin{aligned}
&\hat{\mu}_y(m) - \bar{y}_N \\
&= \sum_{h=1}^{H} A_{Nh0}(m) \frac{A_{nh1}(m)}{A_{nh0}(m)} - \sum_{h=1}^{H} A_{Nh1}(m) \\
&= \sum_{h=1}^{H} \left\{ \frac{A_{nh1}(m)}{A_{nh0}(m)} A_{Nh0}(m) - A_{Nh1}(m) \right\} \\
&= \sum_{h=1}^{H} \left\{ \frac{A_{nh1}(m)}{A_{nh0}(m)} A_{Nh0}(m) - A_{nh1}(m) + A_{nh1}(m) - A_{Nh1}(m) \right\} \\
&= \sum_{h=1}^{H} \left\{ \frac{A_{nh1}(m)}{A_{nh0}(m)} (A_{Nh0}(m) - A_{nh0}(m)) + (A_{nh1}(m) - A_{Nh1}(m)) \right\}. \qquad (16)
\end{aligned}
$$

Next, we have $A_{Nh\ell}(M) = A_{\tau_h \ell}(M) - A_{\tau_{h-1} \ell}(M)$ and $A_{nh\ell}(M) = A_{n\tau_h \ell}(M) -$

$A_{n\tau_{h-1}\ell}(M)$, for $M = \{m, \hat{m}\}$. Let

$$\alpha_{h\ell}(m) = \alpha_{\tau_h\ell}(m) - \alpha_{\tau_{h-1}\ell}(m) = \mathrm{E}[y_i^\ell I_{\{\tau_{h-1} < m(\boldsymbol{x}_i) \leq \tau_h\}}].$$

Then, applying Lemma 1 to two consecutive boundary values, τ_{h-1} and τ_h, the differences of the respective equations are

$$A_{Nh\ell}(\hat{m}) - \mathrm{E}[y_i^\ell I_{\{\tau_{h-1} < \hat{m}(\boldsymbol{x}_i) \leq \tau_h\}} \mid \hat{m}] - A_{Nh\ell}(m) + \alpha_{h\ell}(m) = o_p(N^{-1/2}), \qquad (17)$$

and

$$A_{nh\ell}(\hat{m}) - \mathrm{E}[y_i^\ell I_{\{\tau_{h-1} < \hat{m}(\boldsymbol{x}_i) \leq \tau_h\}} \mid \hat{m}] - A_{nh\ell}(m) + \alpha_{h\ell}(m) = o_p(n^{-1/2}). \qquad (18)$$

Given (17) and (18), the remainder of the proof is very similar to the corresponding proof in Breidt and Opsomer (2008), with adjustments made for the NEPSE context. Define $a_h = A_{Nh0}(m) - A_{nh0}(m)$ and $b_h = A_{Nh1}(m) - A_{nh1}(m)$, and assume, without loss of generality, that in the following calculations the first n values in N constitute the sample. First,

$$\begin{aligned}
&\mathrm{Cov}\,(a_h, a_k) \\
&\quad = \mathrm{Cov}\big(A_{Nh0}(m) - A_{nh0}(m), A_{Nh1}(m) - A_{nh1}(m)\big) \\
&\quad = \mathrm{Cov}\left(\frac{1}{N}\sum_{i=1}^{N} I_{\{\tau_{h-1} < m(\boldsymbol{x}_i) \leq \tau_h\}} - \frac{1}{n}\sum_{i=1}^{n} I_{\{\tau_{h-1} < m(\boldsymbol{x}_i) \leq \tau_h\}},\right. \\
&\qquad\qquad \left.\frac{1}{N}\sum_{i=1}^{N} I_{\{\tau_{k-1} < m(\boldsymbol{x}_i) \leq \tau_k\}} - \frac{1}{n}\sum_{i=1}^{n} I_{\{\tau_{k-1} < m(\boldsymbol{x}_i) \leq \tau_k\}}\right) \\
&\quad = \mathrm{Cov}\left(\left(\frac{1}{N} - \frac{1}{n}\right)\sum_{i=1}^{n} I_{\{\tau_{h-1} < m(\boldsymbol{x}_i) \leq \tau_h\}} + \frac{1}{N}\sum_{i=n+1}^{N} I_{\{\tau_{h-1} < m(\boldsymbol{x}_i) \leq \tau_h\}},\right. \\
&\qquad\qquad \left.\left(\frac{1}{N} - \frac{1}{n}\right)\sum_{i=1}^{n} I_{\{\tau_{k-1} < m(\boldsymbol{x}_i) \leq \tau_k\}} + \frac{1}{N}\sum_{i=n+1}^{N} I_{\{\tau_{k-1} < m(\boldsymbol{x}_i) \leq \tau_k\}}\right)
\end{aligned}$$

$$
\begin{aligned}
&= \operatorname{Cov}\left(\left(\frac{1}{N}-\frac{1}{n}\right)\sum_{i=1}^{n} I_{\{\tau_{h-1}<m(\boldsymbol{x}_i)\le\tau_h\}}, \left(\frac{1}{N}-\frac{1}{n}\right)\sum_{i=1}^{n} I_{\{\tau_{k-1}<m(\boldsymbol{x}_i)\le\tau_k\}}\right) \\
&\quad + \operatorname{Cov}\left(\frac{1}{N}\sum_{i=n+1}^{N} I_{\{\tau_{h-1}<m(\boldsymbol{x}_i)\le\tau_h\}}, \frac{1}{N}\sum_{i=n+1}^{N} I_{\{\tau_{k-1}<m(\boldsymbol{x}_i)\le\tau_k\}}\right). \qquad (19)
\end{aligned}
$$

We examine the two terms in (19) separately. For the first term,

$$
\begin{aligned}
&\operatorname{Cov}\left(\left(\frac{1}{N}-\frac{1}{n}\right)\sum_{i=1}^{n} I_{\{\tau_{h-1}<m(\boldsymbol{x}_i)\le\tau_h\}}, \left(\frac{1}{N}-\frac{1}{n}\right)\sum_{i=1}^{n} I_{\{\tau_{k-1}<m(\boldsymbol{x}_i)\le\tau_k\}}\right) \\
&= \operatorname{E}\left[\left(\frac{1}{N}-\frac{1}{n}\right)^2 \sum_{i=1}^{n}\sum_{j=1}^{n} I_{\{\tau_{h-1}<m(\boldsymbol{x}_i)\le\tau_h\}} I_{\{\tau_{k-1}<m(\boldsymbol{x}_i)\le\tau_k\}}\right] \\
&\quad - \operatorname{E}\left[\left(\frac{1}{N}-\frac{1}{n}\right)\sum_{i=1}^{n} I_{\{\tau_{h-1}<m(\boldsymbol{x}_i)\le\tau_h\}}\right] \operatorname{E}\left[\left(\frac{1}{N}-\frac{1}{n}\right)\sum_{i=1}^{n} I_{\{\tau_{k-1}<m(\boldsymbol{x}_i)\le\tau_k\}}\right] \\
&= \left(\frac{1}{N}-\frac{1}{n}\right)^2 \operatorname{E}\left[\sum_{i=1}^{n} I_{\{\tau_{h-1}<m(\boldsymbol{x}_i)\le\tau_h\}} I_{\{\tau_{k-1}<m(\boldsymbol{x}_i)\le\tau_k\}}\right. \\
&\qquad\qquad \left. + \sum_{i\neq j}\sum I_{\{\tau_{h-1}<m(\boldsymbol{x}_i)\le\tau_h\}} I_{\{\tau_{k-1}<m(\boldsymbol{x}_j)\le\tau_k\}}\right] \\
&\quad - \left(\frac{1}{N}-\frac{1}{n}\right)^2 n^2 \operatorname{E}\left[I_{\{\tau_{h-1}<m(\boldsymbol{x}_i)\le\tau_h\}}\right] \operatorname{E}\left[I_{\{\tau_{k-1}<m(\boldsymbol{x}_i)\le\tau_k\}}\right] \\
&= \left(\frac{1}{N}-\frac{1}{n}\right)^2 \left\{n\alpha_{h0}(m) I_{\{h=k\}} + n(n-1)\alpha_{h0}(m)\alpha_{k0}(m)\right\} \\
&\quad - \left(\frac{1}{N}-\frac{1}{n}\right)^2 n^2 \alpha_{h0}(m)\alpha_{k0}(m) \\
&= \left(\frac{1}{N}-\frac{1}{n}\right)^2 \left\{n\alpha_{h0}(m) I_{\{h=k\}} + \left(n^2-n-n^2\right)\alpha_{h0}(m)\alpha_{k0}(m)\right\} \\
&= \left(\frac{1}{N}-\frac{1}{n}\right)^2 n\left\{\alpha_{h0}(m) I_{\{h=k\}} - \alpha_{h0}(m)\alpha_{k0}(m)\right\}, \qquad (20)
\end{aligned}
$$

and for the second term,

$$
\begin{aligned}
&\operatorname{Cov}\left(\frac{1}{N}\sum_{i=n+1}^{N} I_{\{\tau_{h-1}<m(\boldsymbol{x}_i)\le\tau_h\}}, \frac{1}{N}\sum_{i=n+1}^{N} I_{\{\tau_{k-1}<m(\boldsymbol{x}_i)\le\tau_k\}}\right) \\
&= \left(\frac{1}{N}\right)^2 \left\{(N-n)\alpha_{h0}(m) I_{\{h=k\}} + (N-n)(N-n-1)\alpha_{h0}(m)\alpha_{k0}(m)\right\}
\end{aligned}
$$

$$
\begin{aligned}
&- \left(\frac{1}{N}\right)^2 (N-n)^2 \alpha_{h0}(m)\alpha_{k0}(m) \\
= &\left(\frac{1}{N}\right)^2 \left\{(N-n)\alpha_{h0}(m) I_{\{h=k\}}\right. \\
&\qquad \left. + \left((N-n)^2 - (N-n) - (N-n)^2\right)\alpha_{h0}(m)\alpha_{k0}(m)\right\} \\
= &\left(\frac{1}{N}\right)^2 (N-n)\left\{\alpha_{h0}(m) I_{\{h=k\}} - \alpha_{h0}(m)\alpha_{k0}(m)\right\}. \qquad (21)
\end{aligned}
$$

Inserting (20) and (21) into (19) yields

$$
\begin{aligned}
&\mathrm{Cov}\left(a_h, a_k\right) \\
= &\left(\frac{1}{N} - \frac{1}{n}\right)^2 n \left\{\alpha_{h0}(m) I_{\{h=k\}} - \alpha_{h0}(m)\alpha_{k0}(m)\right\} \\
&+ \left(\frac{1}{N}\right)^2 (N-n)\left\{\alpha_{h0}(m) I_{\{h=k\}} - \alpha_{h0}(m)\alpha_{k0}(m)\right\} \\
= &\left[\left(\frac{1}{N} - \frac{1}{n}\right)^2 n + \left(\frac{1}{N}\right)^2 (N-n)\right]\left\{\alpha_{h0}(m) I_{\{h=k\}} - \alpha_{h0}(m)\alpha_{k0}(m)\right\} \\
= &\left[\left(\frac{1}{N} - \frac{1}{n}\right)\left(\frac{n}{N} - 1\right) + \frac{1}{N}\left(1 - \frac{n}{N}\right)\right]\left\{\alpha_{h0}(m) I_{\{h=k\}} - \alpha_{h0}(m)\alpha_{k0}(m)\right\} \\
= &\left[\left(\frac{1}{n} - \frac{1}{N}\right)\left(1 - \frac{n}{N}\right) + \frac{1}{N}\left(1 - \frac{n}{N}\right)\right]\left\{\alpha_{h0}(m) I_{\{h=k\}} - \alpha_{h0}(m)\alpha_{k0}(m)\right\} \\
= &\left[\left(1 - \frac{n}{N}\right)\left(\frac{1}{n} - \frac{1}{N} + \frac{1}{N}\right)\right]\left\{\alpha_{h0}(m) I_{\{h=k\}} - \alpha_{h0}(m)\alpha_{k0}(m)\right\} \\
= &\frac{1}{n}\left(1 - \frac{n}{N}\right)\left\{\alpha_{h0}(m) I_{\{h=k\}} - \alpha_{h0}(m)\alpha_{k0}(m)\right\}. \qquad (22)
\end{aligned}
$$

Similar calculations show that

$$
\mathrm{Cov}\left(a_h, b_k\right) = \frac{1}{n}\left(1 - \frac{n}{N}\right)\left\{\alpha_{h1}(m) I_{\{h=k\}} - \alpha_{h0}(m)\alpha_{k1}(m)\right\}. \qquad (23)
$$

Because $\mathrm{Var}(a_h) = \mathrm{Cov}(a_h, a_h) = O(n^{-1})$, it follows that $a_h = O_p(n^{-1/2})$. Similar steps show that $b_h = O_p(n^{-1/2})$.

We next examine

$$
\begin{aligned}
&\mathrm{E}\left[\left\{\mathrm{E}[y_i^\ell I_{\{\tau_{h-1}<\hat{m}(\boldsymbol{x}_i)\le\tau_h\}} \mid \hat{m}] - \alpha_{h\ell}(m)\right\}^2\right] \\
&= \mathrm{E}\left[\left\{\mathrm{E}[y_i^\ell I_{\{\tau_{h-1}<\hat{m}(\boldsymbol{x}_i)\le\tau_h\}} \mid \hat{m}] - \mathrm{E}[y_i^\ell I_{\{\tau_{h-1}<m(\boldsymbol{x}_i)\le\tau_h\}}]\right\}^2\right] \\
&= \mathrm{E}\left[\left\{\mathrm{E}[y_i^\ell \left(I_{\{\tau_{h-1}<\hat{m}(\boldsymbol{x}_i)\le\tau_h\}} - I_{\{\tau_{h-1}<m(\boldsymbol{x}_i)\le\tau_h\}}\right) \mid \hat{m}]\right\}^2\right] \\
&\le \mathrm{E}\left[\mathrm{E}\left[y_i^{2\ell}\left(I_{\{\tau_{h-1}<\hat{m}(\boldsymbol{x}_i)\le\tau_h\}} - I_{\{\tau_{h-1}<m(\boldsymbol{x}_i)\le\tau_h\}}\right)^2 \middle| \hat{m}\right]\right] \\
&= \mathrm{E}\left[\mathrm{E}\left[\mathrm{E}\left[y_i^{2\ell}\left(I_{\{\tau_{h-1}<\hat{m}(\boldsymbol{x}_i)\le\tau_h\}} - I_{\{\tau_{h-1}<m(\boldsymbol{x}_i)\le\tau_h\}}\right)^2 \middle| \hat{m}, \boldsymbol{x}_i\right] \middle| \hat{m}\right]\right] \\
&= \mathrm{E}\left[\mathrm{E}\left[\mathrm{E}[y_i^{2\ell} \mid \hat{m}, \boldsymbol{x}_i]\left(I_{\{\tau_{h-1}<\hat{m}(\boldsymbol{x}_i)\le\tau_h\}} - I_{\{\tau_{h-1}<m(\boldsymbol{x}_i)\le\tau_h\}}\right)^2 \middle| \hat{m}\right]\right] \\
&= \mathrm{E}\left[\mathrm{E}\left[\mathrm{E}[y_i^{2\ell} \mid \boldsymbol{x}_i]\left(I_{\{\tau_{h-1}<\hat{m}(\boldsymbol{x}_i)\le\tau_h\}} - I_{\{\tau_{h-1}<m(\boldsymbol{x}_i)\le\tau_h\}}\right)^2 \middle| \hat{m}\right]\right] \\
&\le \mathrm{E}\left[K_1\mathrm{E}\left[\left(I_{\{\tau_{h-1}<\hat{m}(\boldsymbol{x}_i)\le\tau_h\}} - I_{\{\tau_{h-1}<m(\boldsymbol{x}_i)\le\tau_h\}}\right)^2 \middle| \hat{m}\right]\right] \\
&\le \mathrm{E}\Big[K_1\Big\{\Pr(\tau_{h-1} < \hat{m}(\boldsymbol{x}_i) \le \tau_h, m(\boldsymbol{x}_i) > \tau_h \mid \hat{m}) \\
&\quad + \Pr(\tau_{h-1} < \hat{m}(\boldsymbol{x}_i) \le \tau_h, m(\boldsymbol{x}_i) \le \tau_{h-1} \mid \hat{m}) \\
&\quad + \Pr(\hat{m}(\boldsymbol{x}_i) > \tau_h, \tau_{h-1} < m(\boldsymbol{x}_i) \le \tau_h \mid \hat{m}) \\
&\quad + \Pr(\hat{m}(\boldsymbol{x}_i) \le \tau_{h-1}, \tau_{h-1} < m(\boldsymbol{x}_i) \le \tau_h \mid \hat{m})\Big\}\Big]. \qquad (24)
\end{aligned}
$$

We want to show that (24) converges to 0 as $n \to \infty$. For a given $\epsilon > 0$,

$$
\begin{aligned}
&\Pr\Big(\Pr(\tau_{h-1} < \hat{m}(\boldsymbol{x}_i) \le \tau_h, m(\boldsymbol{x}_i) > \tau_h \mid \hat{m}) > \epsilon\Big) \\
&\quad \le \Pr\Big(\Pr(\hat{m}(\boldsymbol{x}_i) \le \tau_h, m(\boldsymbol{x}_i) > \tau_h \mid \hat{m}) > \epsilon\Big) = o(1),
\end{aligned}
$$

by (14). Similar reasoning shows that each of the terms inside the expectation in (24) is $o_p(1)$. By uniform integrability, (24) is $o(1)$. Thus, $\mathrm{E}[y_i^\ell I_{\{\tau_{h-1}<\hat{m}(\boldsymbol{x}_i)\le\tau_h\}} \mid \hat{m}]$ converges to $\alpha_{h\ell}(m)$ in mean square, and hence in probability.

We also have

$$
A_{Nh\ell}(m) - \alpha_{h\ell}(m) = O_p\left(N^{-1/2}\right) \text{ and } A_{nh\ell}(m) - \alpha_{h\ell}(m) = O_p\left(n^{-1/2}\right)
$$

by the Central Limit Theorem, and additionally, $A_{nhl}(m)$ and $A_{Nhl}(m)$ are $O_p(1)$ by the Weak Law of Large Numbers.

Of interest to us is the asymptotic distribution of the NEPSE error,

$$\begin{aligned}\hat{\mu}_y(\hat{m}) - \bar{y}_N &= \sum_{h=1}^{H} A_{Nh0}(\hat{m})\frac{A_{nh1}(\hat{m})}{A_{nh0}(\hat{m})} - \sum_{h=1}^{H} A_{Nh1}(\hat{m}) \\ &= \sum_{h=1}^{H}\left\{\frac{A_{Nh0}(\hat{m})A_{nh1}(\hat{m}) - A_{nh0}(\hat{m})A_{Nh1}(\hat{m})}{A_{nh0}(\hat{m})}\right\}. \qquad (25)\end{aligned}$$

We begin by separately examining the numerator and denominator of the summand in (25), making use of the above order results throughout.

For the numerator, let $D_\ell = \mathrm{E}[y_i^\ell I_{\{\tau_{h-1}<\hat{m}(\boldsymbol{x}_i)\leq\tau_h\}} \mid \hat{m}] - \alpha_{h\ell}(m)$ and recall that $D_\ell = o_p(1)$. Substitute (17) and (18) to obtain

$$\begin{aligned}&A_{Nh0}(\hat{m})A_{nh1}(\hat{m}) - A_{nh0}(\hat{m})A_{Nh1}(\hat{m}) \\ &= \left(A_{Nh0}(m) + D_0 + o_p(N^{-1/2})\right)\left(A_{nh1}(m) + D_1 + o_p(n^{-1/2})\right) \\ &\quad - \left(A_{nh0}(m) + D_0 + o_p(n^{-1/2})\right)\left(A_{Nh1}(m) + D_1 + o_p(N^{-1/2})\right) \\ &= \big(A_{Nh0}(m)A_{nh1}(m) + A_{Nh0}(m)D_1 + A_{Nh0}(m)o_p(n^{-1/2}) + D_0A_{nh1}(m) + D_0D_1 \\ &\quad + D_0o_p(n^{-1/2}) + o_p(N^{-1/2})A_{nh1}(m) + o_p(N^{-1/2})D_1 + o_p(N^{-1/2})o_p(n^{-1/2})\big) \\ &\quad - \big(A_{nh0}(m)A_{Nh1}(m) + A_{nh0}(m)D_1 + A_{nh0}(m)o_p(N^{-1/2}) \\ &\quad + D_0A_{Nh1}(m) + D_0D_1 + D_0o_p(N^{-1/2}) + o_p(n^{-1/2})A_{Nh1}(m) \\ &\quad + o_p(n^{-1/2})D_1 + o_p(n^{-1/2})o_p(N^{-1/2})\big) \\ &= A_{Nh0}(m)A_{nh1}(m) - A_{nh0}(m)A_{Nh1}(m) + \left(A_{Nh0}(m) - A_{nh0}(m)\right)D_1 \\ &\quad + \left(A_{nh1}(m) - A_{Nh1}(m)\right)D_0 + o_p(n^{-1/2}) \\ &= A_{Nh0}(m)A_{nh1}(m) - A_{nh0}(m)A_{Nh1}(m) + O_p(n^{-1/2})o_p(1) \\ &\quad + O_p(n^{-1/2})o_p(1) + o_p(n^{-1/2}) \\ &= A_{Nh0}(m)A_{nh1}(m) - A_{nh0}(m)A_{Nh1}(m) + o_p(n^{-1/2})\end{aligned}$$

$$
\begin{aligned}
&= (A_{Nh0}(m) - \alpha_{h0}(m) + \alpha_{h0}(m))\,(A_{nh1}(m) - \alpha_{h1}(m) + \alpha_{h1}(m)) \\
&\quad - (A_{nh0}(m) - \alpha_{h0}(m) + \alpha_{h0}(m))\,(A_{Nh1}(m) - \alpha_{h1}(m) + \alpha_{h1}(m)) + o_p(n^{-1/2}) \\
&= \{(A_{Nh0}(m) - \alpha_{h0}(m))\,(A_{nh1}(m) - \alpha_{h1}(m)) + (A_{Nh0}(m) - \alpha_{h0}(m))\,\alpha_{h1}(m) \\
&\quad + \alpha_{h0}(m)\,(A_{nh1}(m) - \alpha_{h1}(m)) + \alpha_{h0}(m)\alpha_{h1}(m)\} \\
&\quad - \{(A_{nh0}(m) - \alpha_{h0}(m))\,(A_{Nh1}(m) - \alpha_{h1}(m)) + (A_{nh0}(m) - \alpha_{h0}(m))\,\alpha_{h1}(m) \\
&\quad + \alpha_{h0}(m)\,(A_{Nh1}(m) - \alpha_{h1}(m)) + \alpha_{h0}(m)\alpha_{h1}(m)\} + o_p(n^{-1/2}) \\
&= \Big\{O_p(N^{-1/2})O_p(n^{-1/2}) + A_{Nh0}(m)\alpha_{h1}(m) - \alpha_{h0}(m)\alpha_{h1}(m) \\
&\quad + \alpha_{h0}(m)A_{nh1}(m) - \alpha_{h0}(m)\alpha_{h1}(m) + \alpha_{h0}(m)\alpha_{h1}(m)\Big\} \\
&\quad - \Big\{O_p(n^{-1/2})O_p(N^{-1/2}) + A_{nh0}(m)\alpha_{h1}(m) - \alpha_{h0}(m)\alpha_{h1}(m) \\
&\quad + \alpha_{h0}(m)A_{Nh1}(m) - \alpha_{h0}(m)\alpha_{h1}(m) + \alpha_{h0}(m)\alpha_{h1}(m)\Big\} + o_p(n^{-1/2}) \\
&= \alpha_{h1}(m)\,(A_{Nh0}(m) - A_{nh0}(m)) + \alpha_{h0}(m)\,(A_{nh1}(m) - A_{Nh1}(m)) + o_p(n^{-1/2}) \\
&= \alpha_{h1}(m)a_h + \alpha_{h0}(m)b_h + o_p(n^{-1/2}). \qquad (26)
\end{aligned}
$$

And for the denominator, we use (18) to show

$$
\begin{aligned}
A_{nh0}(\hat{m}) &= \mathrm{E}[y_i^{\ell} I_{\{\tau_{h-1} < \hat{m}(\boldsymbol{x}_i) \le \tau_h\}} \mid \hat{m}] + A_{nh0}(m) - \alpha_{h0}(m) + o_p(n^{-1/2}) \\
&= \alpha_{h0}(m) + \big(\mathrm{E}[I_{\{\tau_{h-1} < \hat{m}(\boldsymbol{x}_i) \le \tau_h\}} \mid \hat{m}] - \alpha_{h0}(m)\big) \\
&\quad + (A_{nh0}(m) - \alpha_{h0}(m)) + o_p(n^{-1/2}) \\
&= \alpha_{h0}(m) + o_p(1) + O_p(n^{-1/2}) + o_p(n^{-1/2}) \\
&= \alpha_{h0}(m) + o_p(1). \qquad (27)
\end{aligned}
$$

Since $\alpha_{h0}(m) > 0$ by Assumption 2.3.1, then by a continuous mapping we have

$$
\frac{1}{A_{nh0}(\hat{m})} = \frac{1}{\alpha_{h0}(m)} + o_p(1). \qquad (28)
$$

Substitute (26) and (28) into (25) to rewrite the NEPSE error as

$$\begin{aligned}
\hat{\mu}_y(\hat{m}) - \bar{y}_N &= \sum_{h=1}^{H} \left\{ (\alpha_{h1}(m)a_h + \alpha_{h0}(m)b_h + o_p(n^{-1/2})) \left(\frac{1}{\alpha_{h0}(m)} + o_p(1) \right) \right\} \\
&= \sum_{h=1}^{H} \left\{ \frac{\alpha_{h1}(m)}{\alpha_{h0}(m)} a_h + \alpha_{h1}(m)a_h o_p(1) + \frac{\alpha_{h0}(m)}{\alpha_{h0}(m)} b_h + \alpha_{h0}(m)b_h o_p(1) \right. \\
&\qquad \left. + o_p(n^{-1/2}) \frac{1}{\alpha_{h0}(m)} + o_p(n^{-1/2})o_p(1) \right\} \\
&= \sum_{h=1}^{H} \left\{ \frac{\alpha_{h1}(m)}{\alpha_{h0}(m)} a_h + \alpha_{h1}(m) O_p(n^{-1/2}) o_p(1) \right. \\
&\qquad \left. + b_h + \alpha_{h0}(m) O_p(n^{-1/2}) o_p(1) + o_p(n^{-1/2}) \right\} \\
&= \sum_{h=1}^{H} \left\{ \frac{\alpha_{h1}(m)}{\alpha_{h0}(m)} a_h + b_h \right\} + o_p(n^{-1/2}). \qquad (29)
\end{aligned}$$

By comparison with (16), the asymptotic distribution for the NEPSE error is the same as that obtained when $m(\cdot)$ is known.

To derive the asymptotic distribution of the NEPSE error we apply the Central Limit Theorem to (29) and note that the limiting distribution is normal with mean zero. We use the earlier covariance computations and the fact that $\sum_{h=1}^{H} b_h = \bar{y}_N - \bar{y}_\pi$ to show that the variance of (29) is approximated by

$$\begin{aligned}
&\mathrm{Var}\left(\hat{\mu}_y(\hat{m}) - \bar{y}_N\right) \\
&\simeq \mathrm{Var}\left(\sum_{h=1}^{H} \left\{ \frac{\alpha_{h1}(m)}{\alpha_{h0}(m)} a_h + b_h \right\} \right) \\
&= \mathrm{Var}\left(\sum_{h=1}^{H} \frac{\alpha_{h1}(m)}{\alpha_{h0}(m)} a_h + \sum_{h=1}^{H} b_h \right) \\
&= \mathrm{Var}\left(\sum_{h=1}^{H} \frac{\alpha_{h1}(m)}{\alpha_{h0}(m)} a_h \right) + \mathrm{Var}\left(\sum_{h=1}^{H} b_h \right) + 2\ \mathrm{Cov}\left(\sum_{h=1}^{H} \frac{\alpha_{h1}(m)}{\alpha_{h0}(m)} a_h, -\sum_{h=1}^{H} b_h \right) \\
&= \left\{ \sum_{h=1}^{H} \left(\frac{\alpha_{h1}(m)}{\alpha_{h0}(m)} \right)^2 \mathrm{Var}(a_h) + 2 \sum_{1 \le h < k \le H} \sum \frac{\alpha_{h1}(m)}{\alpha_{h0}(m)} \frac{\alpha_{k1}(m)}{\alpha_{k0}(m)} \mathrm{Cov}(a_h, a_k) \right\}
\end{aligned}$$

$$
\begin{aligned}
&+ \operatorname{Var}\left(\bar{y}_N - \bar{y}_\pi\right) - 2\sum_{h=1}^{H}\sum_{k=1}^{H}\frac{\alpha_{h1}(m)}{\alpha_{h0}(m)}\operatorname{Cov}(a_h, b_k) \\
= \;& \sum_{h=1}^{H}\left(\frac{\alpha_{h1}(m)}{\alpha_{h0}(m)}\right)^2 \frac{1}{n}\left(1-\frac{n}{N}\right)\left\{\alpha_{h0}(m) - \alpha_{h0}^2(m)\right\} \\
&+ 2\sum\sum_{1\le h<k\le H}\frac{\alpha_{h1}(m)}{\alpha_{h0}(m)}\frac{\alpha_{k1}(m)}{\alpha_{k0}(m)}\frac{1}{n}\left(1-\frac{n}{N}\right)\left\{0 - \alpha_{h0}(m)\alpha_{k0}(m)\right\} \\
&- 2\sum_{h=1}^{H}\sum_{k=1}^{H}\frac{\alpha_{h1}(m)}{\alpha_{h0}(m)}\frac{1}{n}\left(1-\frac{n}{N}\right)\left\{\alpha_{h1}(m)I_{\{h=k\}} - \alpha_{h0}(m)\alpha_{k1}(m)\right\} \\
&+ \operatorname{Var}\left(\bar{y}_N - \bar{y}_\pi\right) \\
= \;& \frac{1}{n}\left(1-\frac{n}{N}\right)\left\{\sum_{h=1}^{H}\frac{\alpha_{h1}^2(m)}{\alpha_{h0}(m)} - \sum_{h=1}^{H}\alpha_{h1}^2(m)\right\} \\
&- 2\,\frac{1}{n}\left(1-\frac{n}{N}\right)\sum\sum_{1\le h<k\le H}\alpha_{h1}(m)\alpha_{k1}(m) \\
&- 2\,\frac{1}{n}\left(1-\frac{n}{N}\right)\left\{\sum_{h=1}^{H}\sum_{k=1}^{H}\frac{\alpha_{h1}^2(m)}{\alpha_{h0}(m)}I_{\{h=k\}} - \sum_{h=1}^{H}\sum_{k=1}^{H}\alpha_{h1}(m)\alpha_{k1}(m)\right\} \\
&+ \operatorname{Var}\left(\bar{y}_N - \bar{y}_\pi\right) \\
= \;& \frac{1}{n}\left(1-\frac{n}{N}\right)\left\{\sum_{h=1}^{H}\frac{\alpha_{h1}^2(m)}{\alpha_{h0}(m)} - \sum_{h=1}^{H}\alpha_{h1}^2(m) - 2\sum\sum_{1\le h<k\le H}\alpha_{h1}(m)\alpha_{k1}(m)\right\} \\
&- 2\,\frac{1}{n}\left(1-\frac{n}{N}\right)\left\{\sum_{h=1}^{H}\frac{\alpha_{h1}^2(m)}{\alpha_{h0}(m)} - \left(\sum_{h=1}^{H}\alpha_{h1}^2(m) + 2\sum\sum_{1\le h<k\le H}\alpha_{h1}(m)\alpha_{k1}(m)\right)\right\} \\
&+ \operatorname{Var}\left(\bar{y}_N - \bar{y}_\pi\right) \\
= \;& \frac{1}{n}\left(1-\frac{n}{N}\right)\left\{\sum_{h=1}^{H}\frac{\alpha_{h1}^2(m)}{\alpha_{h0}(m)} - \sum_{h=1}^{H}\alpha_{h1}^2(m) - 2\sum\sum_{1\le h<k\le H}\alpha_{h1}(m)\alpha_{k1}(m)\right. \\
&\left. - 2\sum_{h=1}^{H}\frac{\alpha_{h1}^2(m)}{\alpha_{h0}(m)} + 2\sum_{h=1}^{H}\alpha_{h1}^2(m) + 4\sum\sum_{1\le h<k\le H}\alpha_{h1}(m)\alpha_{k1}(m)\right\} + \operatorname{Var}\left(\bar{y}_N - \bar{y}_\pi\right) \\
= \;& \frac{1}{n}\left(1-\frac{n}{N}\right)\left\{-\sum_{h=1}^{H}\frac{\alpha_{h1}^2(m)}{\alpha_{h0}(m)} + \left(\sum_{h=1}^{H}\alpha_{h1}^2(m) + 2\sum\sum_{1\le h<k\le H}\alpha_{h1}(m)\alpha_{k1}(m)\right)\right\} \\
&+ \operatorname{Var}\left(\bar{y}_N - \bar{y}_\pi\right) \\
= \;& \frac{1}{n}\left(1-\frac{n}{N}\right)\left\{-\sum_{h=1}^{H}\frac{\alpha_{h1}^2(m)}{\alpha_{h0}(m)} + \left(\sum_{h=1}^{H}\alpha_{h1}(m)\right)^2\right\} + \operatorname{Var}\left(\bar{y}_N - \bar{y}_\pi\right). \qquad (30)
\end{aligned}
$$

Under equal-probability design,

$$\mathrm{Var}\,(\bar{y}_N - \bar{y}_\pi) = \frac{1}{n}\left(1 - \frac{n}{N}\right)\mathrm{Var}(y_i), \tag{31}$$

and

$$\begin{aligned}\sum_{h=1}^{H} \alpha_{h1}(m) &= \sum_{h=1}^{H} \mathrm{E}[y_i I_{\{\tau_{h-1} < m(\boldsymbol{x}_i) \le \tau_h\}}] \\ &= \mathrm{E}[y_i]. \end{aligned} \tag{32}$$

Also, by definition of expectation given an event,

$$\frac{\alpha_{h1}(m)}{\alpha_{h0}(m)} = \mathrm{E}[y_i \,|\, \tau_{h-1} < m(\boldsymbol{x}_i) \le \tau_h],$$

and

$$\begin{aligned}\mathrm{E}[y_i^2] &= \sum_{h=1}^{H} \alpha_{h0}(m)\{\mathrm{Var}(y_i \mid \tau_{h-1} < m(\boldsymbol{x}_i) \le \tau_h) \\ &\qquad +(\mathrm{E}[y_i \mid \tau_{h-1} < m(\boldsymbol{x}_i) \le \tau_h])^2\}.\end{aligned}$$

Substituting (31) and (32) into (30), and making use of the previous two equations, we get

$$\begin{aligned}&\mathrm{Var}\,(\hat{\mu}_y(\hat{m}) - \bar{y}_N) \\ &\simeq \frac{1}{n}\left(1 - \frac{n}{N}\right)\left\{-\sum_{h=1}^{H}\frac{\alpha_{h1}^2(m)}{\alpha_{h0}(m)} + (\mathrm{E}[y_i])^2 + \mathrm{Var}(y_i)\right\} \\ &= \frac{1}{n}\left(1 - \frac{n}{N}\right)\left\{-\sum_{h=1}^{H}\frac{\alpha_{h1}^2(m)}{\alpha_{h0}(m)} + \mathrm{E}[y_i^2]\right\} \\ &= \frac{1}{n}\left(1 - \frac{n}{N}\right)\left\{-\sum_{h=1}^{H}\frac{\alpha_{h1}^2(m)}{\alpha_{h0}(m)} + \sum_{h=1}^{H}\alpha_{h0}(m)\Big\{\mathrm{Var}(y_i \mid \tau_{h-1} < m(\boldsymbol{x}_i) \le \tau_h)\right. \\ &\qquad \left.+ (\mathrm{E}[y_i \mid \tau_{h-1} < m(\boldsymbol{x}_i) \le \tau_h])^2\Big\}\right\}\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{n}\left(1-\frac{n}{N}\right)\left\{-\sum_{h=1}^{H}\frac{\alpha_{h1}^2(m)}{\alpha_{h0}(m)}+\sum_{h=1}^{H}\alpha_{h0}(m)\Big\{\mathrm{Var}(y_i \mid \tau_{h-1}<m(\boldsymbol{x}_i)\le\tau_h)\right.\\
&\qquad\qquad\left.+\left(\frac{\alpha_{h1}(m)}{\alpha_{h0}(m)}\right)^2\Big\}\right\}\\
&= \frac{1}{n}\left(1-\frac{n}{N}\right)\left\{\sum_{h=1}^{H}\alpha_{h0}(m)\mathrm{Var}(y_i \mid \tau_{h-1}<m(\boldsymbol{x}_i)\le\tau_h)\right\}\\
&= \frac{1}{n}\left(1-\frac{n}{N}\right)\left\{\sum_{h=1}^{H}\mathrm{E}[I_{\{\tau_{h-1}<m(\boldsymbol{x}_i)\le\tau_h\}}]\mathrm{Var}(y_i \mid \tau_{h-1}<m(\boldsymbol{x}_i)\le\tau_h)\right\}\\
&= \frac{1}{n}\left(1-\frac{n}{N}\right)\left\{\sum_{h=1}^{H}\Pr\{\tau_{h-1}<m(\boldsymbol{x}_i)\le\tau_h\}\mathrm{Var}(y_i \mid \tau_{h-1}<m(\boldsymbol{x}_i)\le\tau_h)\right\},
\end{aligned}
$$

and the result is proved. □

Proof of Theorem 2. This proof follows the basic structure of the corresponding proof in Breidt and Opsomer (2008) with modifications for the NEPSE setting. Applying arguments similar to those used in (27) to (17) and (18), we find that $A_{Nh\ell}(\hat{m}) \xrightarrow{P} \alpha_{h\ell}(m)$ and $A_{nh\ell}(\hat{m}) \xrightarrow{P} \alpha_{h\ell}(m)$ for $\ell = 0, 1, 2$. So, for the expression given for $\hat{V}_{y\hat{m}}$ in (9), we have

$$
\begin{aligned}
\hat{V}_{y\hat{m}} &= \sum_{h=1}^{H}\frac{A_{Nh0}^2(\hat{m})}{A_{nh0}(\hat{m})}\frac{A_{nh2}(\hat{m})-A_{nh1}^2(\hat{m})/A_{nh0}(\hat{m})}{A_{nh0}(\hat{m})-n^{-1}}\\
&\xrightarrow{P} \sum_{h=1}^{H}\frac{\alpha_{h0}^2(m)}{\alpha_{h0}(m)}\frac{\alpha_{h2}(m))-\alpha_{h1}^2(m)/\alpha_{h0}(m)}{\alpha_{h0}(m)}\\
&= \sum_{h=1}^{H}\alpha_{h0}(m)\left\{\frac{\alpha_{h2}(m)}{\alpha_{h0}(m)}-\left(\frac{\alpha_{h1}(m)}{\alpha_{h0}(m)}\right)^2\right\}\\
&= \sum_{h=1}^{H}\alpha_{h0}(m)\left\{\mathrm{E}[y_i^2 \mid \tau_{h-1}<m(\boldsymbol{x}_i)\le\tau_h]-(\mathrm{E}[y_i \mid \tau_{h-1}<m(\boldsymbol{x}_i)\le\tau_h])^2\right\}\\
&= \sum_{h=1}^{H}\alpha_{h0}(m)\mathrm{Var}(y_i \mid \tau_{h-1}<m(\boldsymbol{x}_i)\le\tau_h)\\
&= \sum_{h=1}^{H}\Pr\{\tau_{h-1}<m(\boldsymbol{x}_i)\le\tau_h\}\mathrm{Var}(y_i \mid \tau_{h-1}<m(\boldsymbol{x}_i)\le\tau_h)\\
&= V_{ym},
\end{aligned}
$$

which implies

$$\hat{V}_{y\hat{m}}^{-1/2} \xrightarrow{P} V_{ym}^{-1/2},$$

so

$$\frac{\hat{V}_{y\hat{m}}^{-1/2}}{V_{ym}^{-1/2}} \xrightarrow{P} 1. \tag{33}$$

By Theorem 1,

$$\left\{\frac{1}{n}\left(1-\frac{n}{N}\right)\right\}^{-1/2} V_{ym}^{-1/2}(\hat{\mu}_y(\hat{m}) - \bar{y}_N) \xrightarrow{d} N(0,1). \tag{34}$$

Applying Slutsky's Theorem to (33) and (34), we get

$$\begin{aligned}
&\frac{\hat{V}_{y\hat{m}}^{-1/2}}{V_{ym}^{-1/2}}\left\{\frac{1}{n}\left(1-\frac{n}{N}\right)\right\}^{-1/2} V_{ym}^{-1/2}(\hat{\mu}_y(\hat{m}) - \bar{y}_N) \\
&= \left\{\frac{1}{n}\left(1-\frac{n}{N}\right)\right\}^{-1/2} \hat{V}_{y\hat{m}}^{-1/2}(\hat{\mu}_y(\hat{m}) - \bar{y}_N) \xrightarrow{d} N(0,1),
\end{aligned}$$

and the result is proved. □

CHAPTER 3

INSTRUMENTAL VARIABLES AND PENALIZED SPLINES: ESTIMATING REGRESSION COEFFICIENTS UNDER INFORMATIVE SAMPLING

3.1 Introduction

3.1.1 Background

In survey problems *informative sampling* occurs when the inclusion probabilities depend on the values of the study variable. This chapter considers the estimation of regression coefficients under informative sampling. We show that ordinary least squares estimators are inconsistent in this sampling scheme, but that consistent estimators can be calculated using a two-stage least squares approach as outlined in Fuller (2009, Ch. 6).

The two-stage least squares process first requires the selection of an appropriate instrumental variable (IV). Stage one is a regression of the auxiliary variable x on the IV to obtain a "fitted" auxiliary variable $\hat{x}$. The study variable y is regressed on $\hat{x}$ at stage two to obtain the consistent estimator of $\boldsymbol{\beta}$.

The proposed estimator is a variation of the two-stage least squares process. In our approach, penalized splines (see Ruppert, Wand, and Carroll 2003, Ch. 3) and IV's are used to determine the "fitted" auxiliary variable $\hat{x}$ at the first stage, and the second stage is an ordinary least-squares regression of y on $\hat{x}$. This estimator is

referred to as the *penalized spline estimator*.

3.1.2 Notation

Consider a finite population $U_N = \{1, \ldots, i, \ldots, N\}$. For each $i \in U_N$, assume a nonrandom auxiliary scalar variable x_i is known. Model the finite population of y_i's, conditioned on the auxiliary variable x_i, as a realization from an infinite superpopulation, ξ, with

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

where ϵ_i are independent random variables with mean zero and variance σ^2_ϵ.

An unequal probability sample s is drawn from U_N according to a Poisson sampling design. Let π_i be the positive inclusion probability for the ith element of U_N, where π_i is a function of x_i and y_i, $\pi_i = \pi(x_i, y_i)$, so that the sampling is *informative*. Let $I_i = 1$ if $i \in s$ and $I_i = 0$ otherwise, where the I_i are independent and distributed as

$$\Pr\{I_i = 1\} = \pi_i, \text{ and } \Pr\{I_i = 0\} = 1 - \pi_i,$$

for $i \in U_N$. Note that $\mathrm{E}[I_i|y_i] = \pi_i$, the expectation with respect to the sampling design (i.e., averaging over all possible samples from the finite population). Let n_N be the size of s, define $w_i = 1/\pi_i$, and assume w_i is known for $i \in s$. The goal is to estimate β_0 and β_1 given the finite population x_i's and the sample y_i's and w_i's.

Select K knots at locations $\kappa_1, \ldots, \kappa_K$ and define the truncated line function $(x-c)_+$ as zero when $x < c$ and $(x-c)$ when $x \geq c$. (Appropriate selection of knots is discussed later.) Define the column vectors

$$\boldsymbol{x}_i = \begin{bmatrix} 1 \\ x_i \end{bmatrix}, \text{ and } \boldsymbol{z}_i = \begin{bmatrix} (x_i - \kappa_1)_+ \\ \vdots \\ (x_i - \kappa_K)_+ \end{bmatrix},$$

for $i \in U_N$ so that

$$\boldsymbol{b}_i = \begin{bmatrix} \boldsymbol{x}_i \\ \boldsymbol{z}_i \end{bmatrix}$$

is a $(K+2)$-vector for each i in the finite population. Because the x_i are nonrandom and known for $i \in U_N$, the vectors $\boldsymbol{x}_i$, $\boldsymbol{z}_i$, and $\boldsymbol{b}_i$ are also nonrandom and known for $i \in U_N$.

In the special case where the entire finite population is considered, define the subscripted $N \times 2$ matrix

$$\boldsymbol{X}_{U_N} = \begin{bmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_N^T \end{bmatrix} = \left[\boldsymbol{x}_i^T\right]_{i \in U_N},$$

but in the more common sample setting, define the $n_N \times 2$ matrix without subscript as

$$\boldsymbol{X} = \left[\boldsymbol{x}_i^T\right]_{i \in s}.$$

Similarly, for the sample, define the $n_N \times K$ matrix

$$\boldsymbol{Z} = \left[\boldsymbol{z}_i^T\right]_{i \in s}.$$

and also define the $n_N \times (K+2)$ matrix

$$\boldsymbol{B} = \left[\boldsymbol{b}_i^T\right]_{i \in s}.$$

Define the $n_N \times n_N$ matrix

$$\boldsymbol{W} = \text{diag}\{w_i\}_{i \in s},$$

and the $n_N \times (K+2)$ matrix

$$\boldsymbol{A} = \boldsymbol{W}\boldsymbol{B} = \left[w_i \boldsymbol{b}_i^T\right]_{i \in s} = \left[\boldsymbol{a}_i^T\right]_{i \in s},$$

where $\boldsymbol{a}_i$ equals $w_i\boldsymbol{b}_i$, a $(K+2)$-vector, for $i \in s$. Use the notation

$$\boldsymbol{A}_x = \boldsymbol{W}\boldsymbol{X} = \left[w_i \boldsymbol{x}_i^T\right]_{i\in s}$$

for the $n_{_N} \times 2$ submatrix of $\boldsymbol{A}$.

Let

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix},$$

and let $\boldsymbol{y} = [y_i]_{i\in s}$ and $\boldsymbol{\epsilon} = [\epsilon_i]_{i\in s}$ be the $n_{_N}$-vectors of y_i's obtained in the sample and the corresponding errors, respectively, so that

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

3.1.3 Chapter outline

In the following section we demonstrate the bias and inconsistency of the ordinary least squares estimator under informative sampling. Sections 3.3 and 3.4 examine the classical two-stage least squares estimators with 2 and $K+2$ IV's, respectively. Section 3.5 introduces and evaluates the new penalized spline estimator. Section 3.6 explores the estimation of the optimal smoothing parameter used by the penalized spline estimator and also contains a simulation study.

3.2 Ordinary least squares estimator, $\hat{\boldsymbol{\beta}}_{ols}$

In this section we demonstrate the bias and inconsistency of the ordinary least squares estimator under informative sampling, indicating the need for an improved estimator.

3.2.1 Assumptions

Let $(\boldsymbol{C})_{jk}$ indicate the element in the jth row and kth column of matrix $\boldsymbol{C}$, and make the following assumptions,

Assumption 3.2.1. $\lim_{N\to\infty} N^{-1} \sum_{i\in U_N} E[\pi_i]\boldsymbol{x}_i\boldsymbol{x}_i^T = \boldsymbol{P}_1$ *is a positive definite* 2×2 *matrix.*

Assumption 3.2.2. $\lim_{N\to\infty} N^{-1} \sum_{i\in U_N} (E[\pi_i] - (E[\pi_i])^2)(\boldsymbol{x}_i\boldsymbol{x}_i^T)^2_{jk} = (\boldsymbol{P}_2)_{jk}$ *for all (j,k) where* $\boldsymbol{P}_2$ *is a* 2×2 *matrix with finite elements.*

Assumption 3.2.3. $\lim_{N\to\infty} N^{-1} \sum_{i\in U_N} \boldsymbol{x}_i Cov(\epsilon_i, \pi_i) = \boldsymbol{p}_3$ *is a* 2×1 *column vector with finite elements.*

Assumption 3.2.4. $\lim_{N\to\infty} N^{-1} \sum_{i\in U_N} \Big(E[\epsilon_i^2\pi_i] - (Cov(\epsilon_i, \pi_i))^2\Big)\boldsymbol{x}_i\boldsymbol{x}_i^T = \boldsymbol{P}_4$ *is a* 2×2 *matrix with finite elements.*

Assumption 3.2.5. $\{\epsilon_i I_i\}_{i\in U_N}$ *are independent.*

3.2.2 Bias and consistency results for $\hat{\boldsymbol{\beta}}_{ols}$

Let $\hat{\boldsymbol{\beta}}_{ols}$ denote the ordinary least squares estimator for $\boldsymbol{\beta}$. We have

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{ols} - \boldsymbol{\beta} &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{\beta} \\
&= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) - \boldsymbol{\beta} \\
&= \boldsymbol{\beta} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\epsilon} - \boldsymbol{\beta} \\
&= \left(\sum_{i\in s} \boldsymbol{x}_i\boldsymbol{x}_i^T\right)^{-1}\left(\sum_{i\in s} \boldsymbol{x}_i\epsilon_i\right) \\
&= \left(\sum_{i\in U_N} \boldsymbol{x}_i\boldsymbol{x}_i^T I_i\right)^{-1}\left(\sum_{i\in U_N} \boldsymbol{x}_i\epsilon_i I_i\right) \qquad (35) \\
&= \left(\frac{1}{N}\sum_{i\in U_N} \boldsymbol{x}_i\boldsymbol{x}_i^T I_i\right)^{-1}\left(\frac{1}{N}\sum_{i\in U_N} \boldsymbol{x}_i\epsilon_i I_i\right). \qquad (36)
\end{aligned}$$

Since the x_i's are fixed, the only randomness is in the sample membership indicators (I_i's) and the errors (ϵ_i's). The expected value of (35) is not a zero vector because

$$\begin{aligned}
\mathrm{E}\left[\sum_{i\in U_N} \boldsymbol{x}_i \epsilon_i I_i\right] &= \sum_{i\in U_N} \boldsymbol{x}_i \mathrm{E}\left[\epsilon_i I_i\right] \\
&= \sum_{i\in U_N} \boldsymbol{x}_i \mathrm{E}\Big[\mathrm{E}\left[\epsilon_i I_i | y_i\right]\Big] \\
&= \sum_{i\in U_N} \boldsymbol{x}_i \mathrm{E}\Big[\epsilon_i \mathrm{E}\left[I_i | y_i\right]\Big] \\
&= \sum_{i\in U_N} \boldsymbol{x}_i \mathrm{E}\left[\epsilon_i \pi_i\right] \\
&= \sum_{i\in U_N} \begin{bmatrix} 1 \\ x_i \end{bmatrix} \mathrm{Cov}(\epsilon_i, \pi_i) \neq \begin{bmatrix} 0 \\ 0 \end{bmatrix},
\end{aligned}$$

since π_i is a function of y_i. As a result, $\hat{\boldsymbol{\beta}}_{ols}$ is biased.

Next, consider the consistency of $\hat{\boldsymbol{\beta}}_{ols}$ by separately examining the two factors of (36). For the first factor,

$$\begin{aligned}
\mathrm{E}[\boldsymbol{x}_i \boldsymbol{x}_i^T I_i] &= \mathrm{E}[\mathrm{E}[I_i|y_i]]\boldsymbol{x}_i \boldsymbol{x}_i^T \\
&= \mathrm{E}[\pi_i]\boldsymbol{x}_i \boldsymbol{x}_i^T,
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{Var}(I_i) &= \mathrm{E}[\mathrm{Var}(I_i|y_i)] + \mathrm{Var}(\mathrm{E}[I_i|y_i]) \\
&= \mathrm{E}[\pi_i(1-\pi_i)] + \mathrm{Var}(\pi_i) \\
&= \mathrm{E}[\pi_i] - \mathrm{E}[\pi_i^2] + \{\mathrm{E}[\pi_i^2] - (\mathrm{E}[\pi_i])^2\} \\
&= \mathrm{E}[\pi_i] - (\mathrm{E}[\pi_i])^2,
\end{aligned}$$

so that

$$\begin{aligned}\lim_{N\to\infty} \mathrm{E}\left[\frac{1}{N}\sum_{i\in U_N}(\boldsymbol{x}_i\boldsymbol{x}_i^T)_{jk}I_i\right] &= \lim_{N\to\infty}\frac{1}{N}\sum_{i\in U_N}\mathrm{E}\left[(\boldsymbol{x}_i\boldsymbol{x}_i^T)_{jk}I_i\right] \\ &= \lim_{N\to\infty}\frac{1}{N}\sum_{i\in U_N}\mathrm{E}[\pi_i](\boldsymbol{x}_i\boldsymbol{x}_i^T)_{jk} \\ &= (\boldsymbol{P}_1)_{jk}\end{aligned}$$

by Assumption 3.2.1, and since the I_i are independent,

$$\begin{aligned}\mathrm{Var}\left(\frac{1}{N}\sum_{i\in U_N}(\boldsymbol{x}_i\boldsymbol{x}_i^T)_{jk}I_i\right) &= \frac{1}{N^2}\mathrm{Var}\left(\sum_{i\in U_N}(\boldsymbol{x}_i\boldsymbol{x}_i^T)_{jk}I_i\right) \\ &= \frac{1}{N^2}\sum_{i\in U_N}(\boldsymbol{x}_i\boldsymbol{x}_i^T)_{jk}^2\mathrm{Var}(I_i) \\ &= \frac{1}{N}\left(\frac{1}{N}\sum_{i\in U_N}(\mathrm{E}[\pi_i]-(\mathrm{E}[\pi_i])^2)(\boldsymbol{x}_i\boldsymbol{x}_i^T)_{jk}^2\right).\end{aligned}$$

Then, as $N\to\infty$,

$$\begin{aligned}&\mathrm{E}\left[\left(\frac{1}{N}\sum_{i\in U_N}(\boldsymbol{x}_i\boldsymbol{x}_i^T)_{jk}I_i-(\boldsymbol{P}_1)_{jk}\right)^2\right] \\ &= \mathrm{Var}\left(\frac{1}{N}\sum_{i\in U_N}(\boldsymbol{x}_i\boldsymbol{x}_i^T)_{jk}I_i-(\boldsymbol{P}_1)_{jk}\right)+\left(\mathrm{E}\left[\frac{1}{N}\sum_{i\in U_N}(\boldsymbol{x}_i\boldsymbol{x}_i^T)_{jk}I_i-(\boldsymbol{P}_1)_{jk}\right]\right)^2 \\ &= \mathrm{Var}\left(\frac{1}{N}\sum_{i\in U_N}(\boldsymbol{x}_i\boldsymbol{x}_i^T)_{jk}I_i\right)+\left(\frac{1}{N}\sum_{i\in U_N}\mathrm{E}\left[(\boldsymbol{x}_i\boldsymbol{x}_i^T)_{jk}I_i\right]-(\boldsymbol{P}_1)_{jk}\right)^2 \\ &= \frac{1}{N}\left(\frac{1}{N}\sum_{i\in U_N}(\mathrm{E}[\pi_i]-(\mathrm{E}[\pi_i])^2)(\boldsymbol{x}_i\boldsymbol{x}_i^T)_{jk}^2\right)+\left(\frac{1}{N}\sum_{i\in U_N}\mathrm{E}[\pi_i](\boldsymbol{x}_i\boldsymbol{x}_i^T)_{jk}-(\boldsymbol{P}_1)_{jk}\right)^2 \\ &= \frac{1}{N}\Big((\boldsymbol{P}_2)_{jk}+o(1)\Big)+\Big((\boldsymbol{P}_1)_{jk}+o(1)-(\boldsymbol{P}_1)_{jk}\Big)^2 \\ &= o(1)\end{aligned}$$

by Assumptions 3.2.1 and 3.2.2, verifying

$$\frac{1}{N}\sum_{i\in U_N}(\boldsymbol{x}_i\boldsymbol{x}_i^T)_{jk}I_i \xrightarrow{m.s.} (\boldsymbol{P}_1)_{jk},$$

for each element, and thus

$$\frac{1}{N}\sum_{i\in U_N}(\boldsymbol{x}_i\boldsymbol{x}_i^T)I_i \xrightarrow{P} \boldsymbol{P}_1. \tag{37}$$

For the second factor of (36),

$$\begin{aligned}\lim_{N\to\infty}\mathrm{E}\left[\frac{1}{N}\sum_{i\in U_N}\boldsymbol{x}_i\epsilon_i I_i\right] &= \lim_{N\to\infty}\frac{1}{N}\sum_{i\in U_N}\boldsymbol{x}_i\mathrm{E}[\epsilon_i I_i]\\ &= \lim_{N\to\infty}\frac{1}{N}\sum_{i\in U_N}\boldsymbol{x}_i\mathrm{Cov}(\epsilon_i,\pi_i)\\ &= \boldsymbol{p}_3\end{aligned}$$

by Assumption 3.2.3, and also

$$\begin{aligned}\mathrm{Var}(\epsilon_i I_i) &= \mathrm{E}[\mathrm{Var}(\epsilon_i I_i|y_i)] + \mathrm{Var}(\mathrm{E}[\epsilon_i I_i|y_i])\\ &= \mathrm{E}[\epsilon_i^2\mathrm{Var}(I_i|y_i)] + \mathrm{Var}(\epsilon_i\pi_i)\\ &= \mathrm{E}[\epsilon_i^2\pi_i(1-\pi_i)] + \{\mathrm{E}[(\epsilon_i\pi_i)^2] - (\mathrm{E}[\epsilon_i\pi_i])^2\}\\ &= \mathrm{E}[\epsilon_i^2\pi_i] - \mathrm{E}[\epsilon_i^2\pi_i^2] + \mathrm{E}[\epsilon_i^2\pi_i^2] - (\mathrm{Cov}(\epsilon_i\pi_i))^2\\ &= \mathrm{E}[\epsilon_i^2\pi_i] - (\mathrm{Cov}(\epsilon_i\pi_i))^2. \end{aligned}\tag{38}$$

Next, to allow work with scalars, we multiply by an arbitrary $\boldsymbol{t}\in\mathbb{R}^2$. Later we recover the desired vector results via the Cramér-Wold device. We also use Assumption 3.2.5 for the variance of the sum in the following equation. As $N\to\infty$,

$$\mathrm{E}\left[\left(\boldsymbol{t}^T\frac{1}{N}\sum_{i\in U_N}\boldsymbol{x}_i\epsilon_i I_i - \boldsymbol{t}^T\boldsymbol{p}_3\right)^2\right]$$

$$
\begin{aligned}
&= \mathrm{Var}\left(\boldsymbol{t}^T \frac{1}{N} \sum_{i \in U_N} \boldsymbol{x}_i \epsilon_i I_i - \boldsymbol{t}^T \boldsymbol{p}_3\right) + \left(\mathrm{E}\left[\boldsymbol{t}^T \frac{1}{N} \sum_{i \in U_N} \boldsymbol{x}_i \epsilon_i I_i - \boldsymbol{t}^T \boldsymbol{p}_3\right]\right)^2 \\
&= \frac{1}{N^2} \boldsymbol{t}^T \left[\sum_{i \in U_N} \boldsymbol{x}_i \mathrm{Var}(\epsilon_i I_i) \boldsymbol{x}_i^T\right] \boldsymbol{t} + \left(\boldsymbol{t}^T \frac{1}{N} \sum_{i \in U_N} \boldsymbol{x}_i \mathrm{E}\left[\epsilon_i I_i\right] - \boldsymbol{t}^T \boldsymbol{p}_3\right)^2 \\
&= \frac{1}{N^2} \boldsymbol{t}^T \left[\sum_{i \in U_N} \mathrm{Var}(\epsilon_i I_i) \boldsymbol{x}_i \boldsymbol{x}_i^T\right] \boldsymbol{t} + \left(\boldsymbol{t}^T \frac{1}{N} \sum_{i \in U_N} \boldsymbol{x}_i \mathrm{Cov}(\epsilon_i, \pi_i) - \boldsymbol{t}^T \boldsymbol{p}_3\right)^2 \\
&= \frac{1}{N} \boldsymbol{t}^T \left[\frac{1}{N} \sum_{i \in U_N} \Big(\mathrm{E}[\epsilon_i^2 \pi_i] - (\mathrm{Cov}(\epsilon_i, \pi_i))^2\Big) \boldsymbol{x}_i \boldsymbol{x}_i^T\right] \boldsymbol{t} + \left(\boldsymbol{t}^T \boldsymbol{p}_3 + o(1) - \boldsymbol{t}^T \boldsymbol{p}_3\right)^2 \\
&= \frac{1}{N} \boldsymbol{t}^T \Big(\boldsymbol{P}_4 + o(1)\Big) \boldsymbol{t} + o(1) \\
&= o(1)
\end{aligned}
$$

by Assumptions 3.2.3 and 3.2.4. This verifies

$$
\boldsymbol{t}^T \frac{1}{N} \sum_{i \in U_N} \boldsymbol{x}_i \epsilon_i I_i \xrightarrow{m.s.} \boldsymbol{t}^T \boldsymbol{p}_3,
$$

which implies

$$
\boldsymbol{t}^T \frac{1}{N} \sum_{i \in U_N} \boldsymbol{x}_i \epsilon_i I_i \xrightarrow{d} \boldsymbol{t}^T \boldsymbol{p}_3,
$$

and by the Cramér-Wold device,

$$
\frac{1}{N} \sum_{i \in U_N} \boldsymbol{x}_i \epsilon_i I_i \xrightarrow{d} \boldsymbol{p}_3.
$$

Because the convergence is to a constant,

$$
\frac{1}{N} \sum_{i \in U_N} \boldsymbol{x}_i \epsilon_i I_i \xrightarrow{P} \boldsymbol{p}_3.
$$

By this result and (37),

$$
(\hat{\boldsymbol{\beta}}_{ols} - \boldsymbol{\beta}) \xrightarrow{P} \boldsymbol{P}_1^{-1} \boldsymbol{p}_3 \neq \begin{bmatrix} 0 \\ 0 \end{bmatrix},
$$

showing that $\hat{\boldsymbol{\beta}}_{ols}$ is not consistent.

3.3 Two-stage least squares estimator, $\hat{\boldsymbol{\beta}}_{2sls}$ (2 IV's)

As indicated in Fuller (2009, Ch. 6), w_i multiplied by any function of x_i is a potential IV under the assumption that $\mathrm{E}[\epsilon_i|x_i] = 0$ for all i in the superpopulation. Thus, we may select $\boldsymbol{A}_x = \boldsymbol{W}\boldsymbol{X}$ as the $n_N \times 2$ IV matrix for a two-stage least squares estimator for $\boldsymbol{\beta}$. We have $\boldsymbol{W}^T = \boldsymbol{W}$ and $\boldsymbol{A}_x^T\boldsymbol{X} = \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}$ is a 2×2 invertible matrix. After stage one, the matrix of "fitted" vectors is

$$\hat{\boldsymbol{X}}_x = \boldsymbol{A}_x\Big[(\boldsymbol{A}_x^T\boldsymbol{A}_x)^{-1}\boldsymbol{A}_x^T\boldsymbol{X}\Big], \tag{39}$$

and after stage two,

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{2sls} &= (\hat{\boldsymbol{X}}_x^T\hat{\boldsymbol{X}}_x)^{-1}\hat{\boldsymbol{X}}_x^T\boldsymbol{y} \\
&= (\boldsymbol{X}^T\boldsymbol{A}_x(\boldsymbol{A}_x^T\boldsymbol{A}_x)^{-1}\boldsymbol{A}_x^T\boldsymbol{A}_x(\boldsymbol{A}_x^T\boldsymbol{A}_x)^{-1}\boldsymbol{A}_x^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{A}_x(\boldsymbol{A}_x^T\boldsymbol{A}_x)^{-1}\boldsymbol{A}_x^T\boldsymbol{y} \\
&= (\boldsymbol{X}^T\boldsymbol{A}_x(\boldsymbol{A}_x^T\boldsymbol{A}_x)^{-1}\boldsymbol{A}_x^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{A}_x(\boldsymbol{A}_x^T\boldsymbol{A}_x)^{-1}\boldsymbol{A}_x^T\boldsymbol{y} \\
&= (\boldsymbol{A}_x^T\boldsymbol{X})^{-1}(\boldsymbol{A}_x^T\boldsymbol{A}_x)(\boldsymbol{X}^T\boldsymbol{A}_x)^{-1}\boldsymbol{X}^T\boldsymbol{A}_x(\boldsymbol{A}_x^T\boldsymbol{A}_x)^{-1}\boldsymbol{A}_x^T\boldsymbol{y} \\
&= (\boldsymbol{A}_x^T\boldsymbol{X})^{-1}(\boldsymbol{A}_x^T\boldsymbol{A}_x)(\boldsymbol{A}_x^T\boldsymbol{A}_x)^{-1}\boldsymbol{A}_x^T\boldsymbol{y} \\
&= (\boldsymbol{A}_x^T\boldsymbol{X})^{-1}\boldsymbol{A}_x^T\boldsymbol{y} \\
&= (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{y}.
\end{aligned} \tag{40}$$

This is also called the probability weighted least squares estimator (see Pfeffermann and Sverchkov 1999).

3.3.1 Assumptions

Make the following assumptions,

Assumption 3.3.1. $\lim_{N\to\infty} N^{-1} \sum_{i\in U_N} \boldsymbol{x}_i\boldsymbol{x}_i^T = \boldsymbol{P}_5$ *is a positive definite* 2×2 *matrix.*

Assumption 3.3.2. $\lim_{N\to\infty} N^{-1} \sum_{i\in U_N} (E[w_i]-1)(\boldsymbol{x}_i\boldsymbol{x}_i^T)^2_{jk} = (\boldsymbol{P}_6)_{jk}$ *for all (j,k) where* $\boldsymbol{P}_6$ *is a* 2×2 *matrix with finite elements.*

Assumption 3.3.3. $\lim_{N\to\infty} N^{-1} \sum_{i\in U_N} E[w_i\epsilon_i^2]\boldsymbol{x}_i\boldsymbol{x}_i^T = \boldsymbol{P}_7$ *is a* 2×2 *matrix with finite elements.*

Assumption 3.3.4. $\{w_iI_i\}_{i\in U_N}$ *are independent.*

Assumption 3.3.5. $\{w_i\epsilon_iI_i\}_{i\in U_N}$ *are independent.*

Assumption 3.3.6. *(Lyapunov condition)* $\lim_{N\to\infty} \frac{1}{V_{Nx}^{1+\delta/2}} \sum_{i\in U_N} E\left[\left|\boldsymbol{t}^T w_i\boldsymbol{x}_i\epsilon_iI_i - 0\right|^{2+\delta}\right] = 0$, *for some* $\delta>0$, *where* $\boldsymbol{t}\in\mathbb{R}^2$ *and* $V_{Nx} = \boldsymbol{t}^T\left(\sum_{i\in U_N} E[w_i\epsilon_i^2]\boldsymbol{x}_i\boldsymbol{x}_i^T\right)\boldsymbol{t}$.

3.3.2 Consistency results for $\hat{\boldsymbol{\beta}}_{2sls}$

Using (40) we can write

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{2sls} - \boldsymbol{\beta} &= (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{y} - \boldsymbol{\beta} \\
&= (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\epsilon}) - \boldsymbol{\beta} \\
&= \boldsymbol{\beta} + (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{\epsilon} - \boldsymbol{\beta} \\
&= \left(\frac{1}{N}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\right)^{-1}\left(\frac{1}{N}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{\epsilon}\right). \qquad (41)
\end{aligned}$$

We examine each factor as $N\to\infty$.

For the first factor of (41), start with

$$\begin{aligned}
\mathrm{E}[w_i\boldsymbol{x}_i\boldsymbol{x}_i^TI_i] &= \mathrm{E}[\mathrm{E}[w_iI_i|y_i]]\boldsymbol{x}_i\boldsymbol{x}_i^T \\
&= \mathrm{E}[w_i\mathrm{E}[I_i|y_i]]\boldsymbol{x}_i\boldsymbol{x}_i^T \\
&= \mathrm{E}[w_i\pi_i]\boldsymbol{x}_i\boldsymbol{x}_i^T \\
&= \boldsymbol{x}_i\boldsymbol{x}_i^T,
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{Var}(w_i I_i) &= \mathrm{E}[\mathrm{Var}(w_i I_i|y_i)] + \mathrm{Var}(\mathrm{E}[w_i I_i|y_i]) \\
&= \mathrm{E}[w_i^2 \mathrm{Var}(I_i|y_i)] + \mathrm{Var}(w_i \mathrm{E}[I_i|y_i]) \\
&= \mathrm{E}[w_i^2 \pi_i(1-\pi_i)] + \mathrm{Var}(w_i \pi_i) \\
&= \mathrm{E}[w_i - 1] + \mathrm{Var}(1) \\
&= \mathrm{E}[w_i] - 1. \qquad (42)
\end{aligned}$$

Next,

$$\begin{aligned}
\lim_{N\to\infty} \mathrm{E}\left[\frac{1}{N}(\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X})_{jk}\right] &= \lim_{N\to\infty} \frac{1}{N} \sum_{i \in U_N} \mathrm{E}[w_i (\boldsymbol{x}_i \boldsymbol{x}_i^T)_{jk} I_i] \\
&= \lim_{N\to\infty} \frac{1}{N} \sum_{i \in U_N} (\boldsymbol{x}_i \boldsymbol{x}_i^T)_{jk} \\
&= (\boldsymbol{P}_5)_{jk}
\end{aligned}$$

by Assumption 3.3.1, and by Assumption 3.3.4,

$$\begin{aligned}
\mathrm{Var}\left(\frac{1}{N}(\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X})_{jk}\right) &= \frac{1}{N^2} \mathrm{Var}\left(\sum_{i \in U_N} w_i (\boldsymbol{x}_i \boldsymbol{x}_i^T)_{jk} I_i\right) \\
&= \frac{1}{N^2} \sum_{i \in U_N} (\boldsymbol{x}_i \boldsymbol{x}_i^T)_{jk}^2 \mathrm{Var}(w_i I_i) \\
&= \frac{1}{N^2} \sum_{i \in U_N} (\mathrm{E}[w_i] - 1)(\boldsymbol{x}_i \boldsymbol{x}_i^T)_{jk}^2 \\
&= \frac{1}{N} \left(\frac{1}{N} \sum_{i \in U_N} (\mathrm{E}[w_i] - 1)(\boldsymbol{x}_i \boldsymbol{x}_i^T)_{jk}^2\right).
\end{aligned}$$

Then, as $N \to \infty$,

$$\mathrm{E}\left[\left(\frac{1}{N}(\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X})_{jk} - (\boldsymbol{P}_5)_{jk}\right)^2\right]$$

$$
\begin{aligned}
&= \mathrm{Var}\left(\frac{1}{N}(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})_{jk} - (\boldsymbol{P}_5)_{jk}\right) + \left(\mathrm{E}\left[\frac{1}{N}(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})_{jk} - (\boldsymbol{P}_5)_{jk}\right]\right)^2 \\
&= \mathrm{Var}\left(\frac{1}{N}(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})_{jk}\right) + \left(\frac{1}{N}\sum_{i\in U_N}\mathrm{E}[w_i(\boldsymbol{x}_i\boldsymbol{x}_i^T)_{jk}I_i] - (\boldsymbol{P}_5)_{jk}\right)^2 \\
&= \frac{1}{N}\left(\frac{1}{N}\sum_{i\in U_N}(\mathrm{E}[w_i]-1)(\boldsymbol{x}_i\boldsymbol{x}_i^T)_{jk}^2\right) + \left(\frac{1}{N}\sum_{i\in U_N}(\boldsymbol{x}_i\boldsymbol{x}_i^T)_{jk} - (\boldsymbol{P}_5)_{jk}\right)^2 \\
&= \frac{1}{N}\Big((\boldsymbol{P}_6)_{jk} + o(1)\Big) + \Big((\boldsymbol{P}_5)_{jk} + o(1) - (\boldsymbol{P}_5)_{jk}\Big)^2 \\
&= o(1)
\end{aligned}
$$

by Assumptions 3.3.1 and 3.3.2, verifying

$$\frac{1}{N}(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})_{jk} \xrightarrow{m.s.} (\boldsymbol{P}_5)_{jk},$$

for each element, and thus

$$\frac{1}{N}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X} \xrightarrow{P} \boldsymbol{P}_5. \tag{43}$$

For the second factor of (41),

$$
\begin{aligned}
\mathrm{E}[w_i\boldsymbol{x}_i\epsilon_i I_i] &= \mathrm{E}[\mathrm{E}[w_i\epsilon_i I_i|y_i]]\boldsymbol{x}_i \\
&= \mathrm{E}[w_i\epsilon_i\mathrm{E}[I_i|y_i]]\boldsymbol{x}_i \\
&= \mathrm{E}[w_i\epsilon_i\pi_i]\boldsymbol{x}_i \\
&= \mathrm{E}[\epsilon_i]\boldsymbol{x}_i \\
&= \mathbf{0}, \qquad (44)
\end{aligned}
$$

and

$$
\begin{aligned}
\mathrm{Var}(w_i\epsilon_i I_i) &= \mathrm{E}[\mathrm{Var}(w_i\epsilon_i I_i|y_i)] + \mathrm{Var}(\mathrm{E}[w_i\epsilon_i I_i|y_i]) \\
&= \mathrm{E}[(w_i\epsilon_i)\mathrm{Var}(I_i|y_i)(\epsilon_i w_i)] + \mathrm{Var}(\epsilon_i) \\
&= \mathrm{E}[w_i^2\pi_i(1-\pi_i)\epsilon_i^2] + \mathrm{Var}(\epsilon_i)
\end{aligned}
$$

$$
\begin{aligned}
&= \mathrm{E}[(w_i - 1)\epsilon_i^2] + \sigma_\epsilon^2 \\
&= \mathrm{E}[w_i\epsilon_i^2] - \mathrm{E}[\epsilon_i^2] + \sigma_\epsilon^2 \\
&= \mathrm{E}[w_i\epsilon_i^2] - \sigma_\epsilon^2 + \sigma_\epsilon^2 \\
&= \mathrm{E}[w_i\epsilon_i^2]. \qquad (45)
\end{aligned}
$$

Let $\boldsymbol{t} \in \mathbb{R}^2$ and let $N \to \infty$ to find

$$
\begin{aligned}
&\mathrm{E}\left[\left(\boldsymbol{t}^T\frac{1}{N}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{\epsilon} - 0\right)^2\right] \\
&= \mathrm{Var}\left(\boldsymbol{t}^T\frac{1}{N}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{\epsilon}\right) + \left(\mathrm{E}\left[\boldsymbol{t}^T\frac{1}{N}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{\epsilon}\right]\right)^2 \\
&= \frac{1}{N^2}\boldsymbol{t}^T\mathrm{Var}\left(\sum_{i\in U_N} w_i\boldsymbol{x}_i\epsilon_i I_i\right)\boldsymbol{t} + \left(\frac{1}{N}\sum_{i\in U_N}\boldsymbol{t}^T\mathrm{E}[w_i\boldsymbol{x}_i\epsilon_i I_i]\right)^2 \\
&= \frac{1}{N^2}\boldsymbol{t}^T\left(\sum_{i\in U_N}\boldsymbol{x}_i\mathrm{Var}(w_i\epsilon_i^2)\boldsymbol{x}_i^T\right)\boldsymbol{t} + \left(\frac{1}{N}\sum_{i\in U_N}\boldsymbol{t}^T\boldsymbol{0}\right)^2 \\
&= \frac{1}{N}\boldsymbol{t}^T\left(\frac{1}{N}\sum_{i\in U_N}\mathrm{E}[w_i\epsilon_i^2]\boldsymbol{x}_i\boldsymbol{x}_i^T\right)\boldsymbol{t} + 0 \\
&= \frac{1}{N}\boldsymbol{t}^T\Big(\boldsymbol{P}_7 + o(1)\Big)\boldsymbol{t} \\
&= o(1)
\end{aligned}
$$

by Assumption 3.3.3, and using Assumption 3.3.5 for the variance of the sum. This result verifies

$$\boldsymbol{t}^T\frac{1}{N}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{\epsilon} \xrightarrow{m.s.} 0.$$

This implies

$$\boldsymbol{t}^T\frac{1}{N}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{\epsilon} \xrightarrow{d} \boldsymbol{t}^T\boldsymbol{0},$$

and by the Cramér-Wold device,

$$\frac{1}{N}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{\epsilon} \xrightarrow{d} \boldsymbol{0}.$$

Because the convergence is to a constant,

$$\frac{1}{N}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{\epsilon} \xrightarrow{P} \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \tag{46}$$

Equations (41), (43), and (46) together imply the consistency of the two-stage least squares estimator,

$$(\hat{\boldsymbol{\beta}}_{2sls} - \boldsymbol{\beta}) \xrightarrow{P} \boldsymbol{P}_5^{-1} \begin{bmatrix} 0 \\ 0 \end{bmatrix} = \mathbf{0}.$$

3.3.3 Central limit theorem results for $\hat{\boldsymbol{\beta}}_{2sls}$

To examine the asymptotic distribution of the estimator, start by writing

$$\begin{aligned} \sqrt{N}\left(\hat{\boldsymbol{\beta}}_{2sls} - \boldsymbol{\beta}\right) &= \sqrt{N}\Big(\boldsymbol{\beta} + (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{\epsilon} - \boldsymbol{\beta}\Big) \\ &= \left(\frac{1}{N}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\right)^{-1}\left(\frac{1}{\sqrt{N}}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{\epsilon}\right). \end{aligned} \tag{47}$$

As $N \to \infty$, by (43),

$$\left(\frac{1}{N}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\right)^{-1} \xrightarrow{P} (\boldsymbol{P}_5)^{-1}. \tag{48}$$

For the second factor of (47), select $\boldsymbol{t} \in \mathbb{R}^2$ to get the scalars

$$\mathrm{E}[\boldsymbol{t}^T w_i \boldsymbol{x}_i \epsilon_i I_i] = \boldsymbol{t}^T \mathrm{E}[w_i \boldsymbol{x}_i \epsilon_i I_i] = \boldsymbol{t}^T \mathbf{0} = 0,$$

using (44), and

$$\mathrm{Var}(\boldsymbol{t}^T w_i \boldsymbol{x}_i \epsilon_i I_i) = \boldsymbol{t}^T \mathrm{Var}(w_i \boldsymbol{x}_i \epsilon_i I_i)\boldsymbol{t} = \boldsymbol{t}^T (\mathrm{E}[w_i \epsilon_i^2]\boldsymbol{x}_i \boldsymbol{x}_i^T)\boldsymbol{t},$$

using (45).

Next, examine

$$\begin{aligned}
\boldsymbol{t}^T \frac{1}{\sqrt{N}} \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{\epsilon} &= \boldsymbol{t}^T \frac{1}{\sqrt{N}} \sum_{i \in U_N} w_i \boldsymbol{x}_i \epsilon_i I_i \\
&= \frac{1}{\sqrt{N}} \sum_{i \in U_N} \boldsymbol{t}^T w_i \boldsymbol{x}_i \epsilon_i I_i \\
&= \frac{1}{\sqrt{N}} V_{Nx}^{1/2} V_{Nx}^{-1/2} \sum_{i \in U_N} \left(\boldsymbol{t}^T w_i \boldsymbol{x}_i \epsilon_i I_i - 0 \right),
\end{aligned}$$

with V_{Nx} as defined in Assumption 3.3.6. We have

$$\begin{aligned}
\frac{1}{\sqrt{N}} V_{Nx}^{1/2} &= \left(\frac{1}{N} V_{Nx} \right)^{1/2} \\
&= \left(\boldsymbol{t}^T \left(\frac{1}{N} \sum_{i \in U_N} \mathrm{E}[w_i \epsilon_i^2] \boldsymbol{x}_i \boldsymbol{x}_i^T \right) \boldsymbol{t} \right)^{1/2} \\
&\to \left(\boldsymbol{t}^T \boldsymbol{P}_7 \boldsymbol{t} \right)^{1/2}
\end{aligned}$$

by Assumption 3.3.3, and by Assumption 3.3.6 and the Lyapunov Central Limit Theorem (Billingsley 1995, Ch. 5),

$$V_{Nx}^{-1/2} \sum_{i \in U_N} \left(\boldsymbol{t}^T w_i \boldsymbol{x}_i \epsilon_i I_i - 0 \right) \xrightarrow{d} N(0,1).$$

Applying Slutsky's Theorem,

$$\begin{aligned}
\boldsymbol{t}^T \frac{1}{\sqrt{N}} \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{\epsilon} &\xrightarrow{d} \left(\boldsymbol{t}^T \boldsymbol{P}_7 \boldsymbol{t} \right)^{1/2} N(0,1) \\
&= N(0, \boldsymbol{t}^T \boldsymbol{P}_7 \boldsymbol{t}) \\
&= N(\boldsymbol{t}^T \boldsymbol{0}, \boldsymbol{t}^T \boldsymbol{P}_7 \boldsymbol{t}) \\
&= \boldsymbol{t}^T N(\boldsymbol{0}, \boldsymbol{P}_7),
\end{aligned}$$

so by the Cramér-Wold device,

$$\frac{1}{\sqrt{N}}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{\epsilon} \xrightarrow{d} N(\mathbf{0}, \boldsymbol{P}_7). \tag{49}$$

Another application of Slutsky's Theorem using (47), (48), and (49) yields

$$\begin{aligned}\sqrt{N}\left(\hat{\boldsymbol{\beta}}_{2sls} - \boldsymbol{\beta}\right) &\xrightarrow{d} \boldsymbol{P}_5^{-1}N(\mathbf{0}, \boldsymbol{P}_7) \\ &= N\left(\begin{bmatrix}0\\0\end{bmatrix}, \boldsymbol{P}_5^{-1}\boldsymbol{P}_7\boldsymbol{P}_5^{-1}\right),\end{aligned} \tag{50}$$

so that

$$\hat{\boldsymbol{\beta}}_{2sls} \text{ is } AN\left(\boldsymbol{\beta}, \frac{1}{N}\boldsymbol{P}_5^{-1}\boldsymbol{P}_7\boldsymbol{P}_5^{-1}\right).$$

3.4 Two-stage least squares estimator, $\hat{\boldsymbol{\beta}}_{2sls}^{K+2}$ ($K+2$ IV's)

In this section we select $\boldsymbol{A} = \boldsymbol{W}\boldsymbol{B}$ as the $n_N \times (K+2)$ IV matrix, utilizing K new IV's in addition to the two IV's used in the previous section. The results we obtain are later useful as a reference for comparisons. Specifically, the central limit theorem results here are shown to be one limiting case of the results for the penalized spline estimator presented in Section 3.5.

For the current IV matrix, define the $(K+2) \times 2$ matrix of stage one coefficients as

$$\hat{\boldsymbol{\Gamma}}_0 = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{X},$$

so that after the first stage, the $n_N \times 2$ matrix of "fitted" vectors is

$$\hat{\boldsymbol{X}}_0 = \boldsymbol{A}\hat{\boldsymbol{\Gamma}}_0,$$

and after the second stage,

$$\hat{\boldsymbol{\beta}}_{2sls}^{K+2} = (\hat{\boldsymbol{X}}_0^T \hat{\boldsymbol{X}}_0)^{-1} \hat{\boldsymbol{X}}_0^T \boldsymbol{y}. \tag{51}$$

3.4.1 Assumptions

As a reference aid, limiting vectors and matrices related strictly to the $\boldsymbol{x}_i$'s are denoted with $\boldsymbol{p}$'s and $\boldsymbol{P}$'s, and limiting matrices related to the $\boldsymbol{b}_i$'s are denoted with $\boldsymbol{M}$'s. When a matrix $\boldsymbol{P}$ is a submatrix of a matrix $\boldsymbol{M}$, the relationship is clearly noted in the text. We make the following additional assumptions,

Assumption 3.4.1. $\lim_{N\to\infty} N^{-1} \sum_{i\in U_N} E[w_i]\boldsymbol{b}_i\boldsymbol{b}_i^T = \boldsymbol{M}_1$ *is a positive definite* $(K+2)\times(K+2)$ *matrix.*

Assumption 3.4.2. $\lim_{N\to\infty} N^{-1} \sum_{i\in U_N} \boldsymbol{b}_i\boldsymbol{x}_i^T = \boldsymbol{M}_2$ *is a* $(K+2)\times 2$ *matrix with finite elements.*

Assumption 3.4.3. $\{w_i^2 I_i\}_{i\in U_N}$ *are independent.*

Assumption 3.4.4. $\lim_{N\to\infty} N^{-1} \sum_{i\in U_N} (E[w_i^3] - (E[w_i])^2)(\boldsymbol{b}_i\boldsymbol{b}_i^T)_{jk}^2 = (\boldsymbol{M}_3)_{jk}$ *for all (j,k) where* $\boldsymbol{M}_3$ *is a* $(K+2)\times(K+2)$ *matrix with finite elements.*

Assumption 3.4.5. $\lim_{N\to\infty} N^{-1} \sum_{i\in U_N} (E[w_i] - 1)(\boldsymbol{b}_i\boldsymbol{x}_i^T)_{jk}^2 = (\boldsymbol{M}_4)_{jk}$ *for all (j,k) where* $\boldsymbol{M}_4$ *is a* $(K+2)\times 2$ *matrix with finite elements.*

Assumption 3.4.6. $\lim_{N\to\infty} N^{-1} \sum_{i\in U_N} E[w_i\epsilon_i^2]\boldsymbol{b}_i\boldsymbol{b}_i^T = \boldsymbol{M}_5$ *is a positive definite* $(K+2)\times(K+2)$ *matrix.*

Assumption 3.4.7. *(Lyapunov condition)* $\lim_{N\to\infty} \frac{1}{V_N^{1+\delta/2}} \sum_{i\in U_N} E\left[\left|\boldsymbol{t}^T w_i \boldsymbol{b}_i \epsilon_i I_i - 0\right|^{2+\delta}\right] = 0$, *for some* $\delta > 0$, *where* $\boldsymbol{t} \in \mathbb{R}^2$ *and* $V_N = \boldsymbol{t}^T \left(\sum_{i\in U_N} E[w_i\epsilon_i^2]\boldsymbol{b}_i\boldsymbol{b}_i^T \right) \boldsymbol{t}$.

3.4.2 Consistency results for $\hat{\boldsymbol{\beta}}_{2sls}^{K+2}$

Using the result

$$\begin{aligned}
\hat{\boldsymbol{X}}_0^T \hat{\boldsymbol{X}}_0 &= \hat{\boldsymbol{\Gamma}}_0^T \boldsymbol{A}^T \boldsymbol{A} \hat{\boldsymbol{\Gamma}}_0 \\
&= \hat{\boldsymbol{\Gamma}}_0^T \boldsymbol{A}^T \boldsymbol{A} (\boldsymbol{A}^T \boldsymbol{A})^{-1} \boldsymbol{A}^T \boldsymbol{X} \\
&= \hat{\boldsymbol{\Gamma}}_0^T \boldsymbol{A}^T \boldsymbol{X} \\
&= \hat{\boldsymbol{X}}_0^T \boldsymbol{X},
\end{aligned}$$

together with (51), allows us to write

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{2sls}^{K+2} - \boldsymbol{\beta} &= (\hat{\boldsymbol{X}}_0^T \hat{\boldsymbol{X}}_0)^{-1} \hat{\boldsymbol{X}}_0^T \boldsymbol{y} - \boldsymbol{\beta} \\
&= (\hat{\boldsymbol{X}}_0^T \hat{\boldsymbol{X}}_0)^{-1} \hat{\boldsymbol{X}}_0^T (\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) - \boldsymbol{\beta} \\
&= (\hat{\boldsymbol{X}}_0^T \hat{\boldsymbol{X}}_0)^{-1} \hat{\boldsymbol{X}}_0^T \boldsymbol{X}\boldsymbol{\beta} + (\hat{\boldsymbol{X}}_0^T \hat{\boldsymbol{X}}_0)^{-1} \hat{\boldsymbol{X}}_0^T \boldsymbol{\epsilon} - \boldsymbol{\beta} \\
&= \boldsymbol{\beta} + (\hat{\boldsymbol{X}}_0^T \hat{\boldsymbol{X}}_0)^{-1} \hat{\boldsymbol{X}}_0^T \boldsymbol{\epsilon} - \boldsymbol{\beta} \\
&= \left(\frac{1}{N} \hat{\boldsymbol{X}}_0^T \hat{\boldsymbol{X}}_0 \right)^{-1} \left(\frac{1}{N} \hat{\boldsymbol{X}}_0^T \boldsymbol{\epsilon} \right). \qquad (52)
\end{aligned}$$

The following work shows that this converges to a zero vector.

We begin by defining $\boldsymbol{\Gamma}_0 = \boldsymbol{M}_1^{-1} \boldsymbol{M}_2$ and show that $\hat{\boldsymbol{\Gamma}}_0 \xrightarrow{P} \boldsymbol{\Gamma}_0$. Start by considering

$$\hat{\boldsymbol{\Gamma}}_0 = (\boldsymbol{A}^T \boldsymbol{A})^{-1} \boldsymbol{A}^T \boldsymbol{X} = \left(\frac{1}{N} \boldsymbol{A}^T \boldsymbol{A} \right)^{-1} \left(\frac{1}{N} \boldsymbol{A}^T \boldsymbol{X} \right).$$

We have

$$\begin{aligned}
\mathrm{E}[w_i^2 \boldsymbol{b}_i \boldsymbol{b}_i^T I_i] &= \mathrm{E}[\mathrm{E}[w_i^2 I_i | y_i]] \boldsymbol{b}_i \boldsymbol{b}_i^T \\
&= \mathrm{E}[w_i^2 \mathrm{E}[I_i | y_i]] \boldsymbol{b}_i \boldsymbol{b}_i^T \\
&= \mathrm{E}[w_i^2 \pi_i] \boldsymbol{b}_i \boldsymbol{b}_i^T \\
&= \mathrm{E}[w_i] \boldsymbol{b}_i \boldsymbol{b}_i^T,
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{Var}(w_i^2 I_i) &= \mathrm{E}[\mathrm{Var}(w_i^2 I_i|y_i)] + \mathrm{Var}(\mathrm{E}[w_i^2 I_i|y_i]) \\
&= \mathrm{E}[w_i^4 \mathrm{Var}(I_i|y_i)] + \mathrm{Var}(w_i^2 \mathrm{E}[I_i|y_i]) \\
&= \mathrm{E}[w_i^4 \pi_i(1-\pi_i)] + \mathrm{Var}(w_i^2 \pi_i) \\
&= \mathrm{E}[w_i^3 - w_i^2] + \mathrm{Var}(w_i) \\
&= \mathrm{E}[w_i^3] - \mathrm{E}[w_i^2] + \mathrm{E}[w_i^2] - (\mathrm{E}[w_i])^2 \\
&= \mathrm{E}[w_i^3] - (\mathrm{E}[w_i])^2.
\end{aligned}$$

Next, show that an individual element of $N^{-1}\boldsymbol{A}^T\boldsymbol{A}$ converges in probability to the corresponding element of $\boldsymbol{M}_1$. Using Assumption 3.4.1,

$$\begin{aligned}
\lim_{N\to\infty} \mathrm{E}\left[\frac{1}{N}(\boldsymbol{A}^T\boldsymbol{A})_{jk}\right] &= \lim_{N\to\infty} \frac{1}{N} \sum_{i\in U_N} \mathrm{E}[w_i^2 (\boldsymbol{b}_i \boldsymbol{b}_i^T)_{jk} I_i] \\
&= \lim_{N\to\infty} \frac{1}{N} \sum_{i\in U_N} \mathrm{E}[w_i] (\boldsymbol{b}_i \boldsymbol{b}_i^T)_{jk} \\
&= (\boldsymbol{M}_1)_{jk},
\end{aligned}$$

and by Assumption 3.4.3,

$$\begin{aligned}
\mathrm{Var}\left(\frac{1}{N}(\boldsymbol{A}^T\boldsymbol{A})_{jk}\right) &= \frac{1}{N^2}\mathrm{Var}\left(\sum_{i\in U_N} w_i^2 (\boldsymbol{b}_i \boldsymbol{b}_i^T)_{jk} I_i\right) \\
&= \frac{1}{N^2} \sum_{i\in U_N} (\boldsymbol{b}_i \boldsymbol{b}_i^T)_{jk}^2 \mathrm{Var}(w_i^2 I_i) \\
&= \frac{1}{N^2} \sum_{i\in U_N} \left(\mathrm{E}[w_i^3] - (\mathrm{E}[w_i])^2\right) (\boldsymbol{b}_i \boldsymbol{b}_i^T)_{jk}^2 \\
&= \frac{1}{N}\left(\frac{1}{N} \sum_{i\in U_N} \left(\mathrm{E}[w_i^3] - (\mathrm{E}[w_i])^2\right) (\boldsymbol{b}_i \boldsymbol{b}_i^T)_{jk}^2\right).
\end{aligned}$$

Then, as $N \to \infty$,

$$\begin{aligned}
&\mathrm{E}\left[\left(\frac{1}{N}(\boldsymbol{A}^T\boldsymbol{A})_{jk} - (\boldsymbol{M}_1)_{jk}\right)^2\right] \\
&= \mathrm{Var}\left(\frac{1}{N}(\boldsymbol{A}^T\boldsymbol{A})_{jk} - (\boldsymbol{M}_1)_{jk}\right) + \left(\mathrm{E}\left[\frac{1}{N}(\boldsymbol{A}^T\boldsymbol{A})_{jk} - (\boldsymbol{M}_1)_{jk}\right]\right)^2 \\
&= \mathrm{Var}\left(\frac{1}{N}(\boldsymbol{A}^T\boldsymbol{A})_{jk}\right) + \left(\frac{1}{N}\sum_{i\in U_N}\mathrm{E}[w_i^2(\boldsymbol{b}_i\boldsymbol{b}_i^T)_{jk}I_i] - (\boldsymbol{M}_1)_{jk}\right)^2 \\
&= \frac{1}{N}\left(\frac{1}{N}\sum_{i\in U_N}\left(\mathrm{E}[w_i^3] - (\mathrm{E}[w_i])^2\right)(\boldsymbol{b}_i\boldsymbol{b}_i^T)_{jk}^2\right) \\
&\quad + \left(\frac{1}{N}\sum_{i\in U_N}\mathrm{E}[w_i](\boldsymbol{b}_i\boldsymbol{b}_i^T)_{jk} - (\boldsymbol{M}_1)_{jk}\right)^2 \\
&= \frac{1}{N}\Big((\boldsymbol{M}_3)_{jk} + o(1)\Big) + \Big((\boldsymbol{M}_1)_{jk} + o(1) - (\boldsymbol{M}_1)_{jk}\Big)^2 \\
&= o(1)
\end{aligned}$$

by Assumptions 3.4.1 and 3.4.4, verifying

$$\frac{1}{N}(\boldsymbol{A}^T\boldsymbol{A})_{jk} \xrightarrow{m.s.} (\boldsymbol{M}_1)_{jk},$$

for each element, and thus

$$\frac{1}{N}\boldsymbol{A}^T\boldsymbol{A} \xrightarrow{P} \boldsymbol{M}_1. \tag{53}$$

Similarly, for the $N^{-1}\boldsymbol{A}^T\boldsymbol{X}$ factor,

$$\begin{aligned}
\mathrm{E}[w_i\boldsymbol{b}_i\boldsymbol{x}_i^T I_i] &= \mathrm{E}[\mathrm{E}[w_i I_i|y_i]]\boldsymbol{b}_i\boldsymbol{x}_i^T \\
&= \mathrm{E}[w_i\mathrm{E}[I_i|y_i]]\boldsymbol{b}_i\boldsymbol{x}_i^T \\
&= \mathrm{E}[w_i\pi_i]\boldsymbol{b}_i\boldsymbol{x}_i^T \\
&= \boldsymbol{b}_i\boldsymbol{x}_i^T,
\end{aligned}$$

and by (42),

$$\mathrm{Var}(w_i I_i) = \mathrm{E}[w_i] - 1.$$

Next,

$$\begin{aligned}
\lim_{N\to\infty} \mathrm{E}\left[\frac{1}{N}(\boldsymbol{A}^T\boldsymbol{X})_{jk}\right] &= \lim_{N\to\infty} \frac{1}{N}\sum_{i\in U_N} \mathrm{E}[w_i(\boldsymbol{b}_i\boldsymbol{x}_i^T)_{jk} I_i] \\
&= \lim_{N\to\infty} \frac{1}{N}\sum_{i\in U_N} (\boldsymbol{b}_i\boldsymbol{x}_i^T)_{jk} \\
&= (\boldsymbol{M}_2)_{jk}
\end{aligned}$$

by Assumption 3.4.2, and by Assumption 3.3.4,

$$\begin{aligned}
\mathrm{Var}\left(\frac{1}{N}(\boldsymbol{A}^T\boldsymbol{X})_{jk}\right) &= \frac{1}{N^2}\mathrm{Var}\left(\sum_{i\in U_N} w_i(\boldsymbol{b}_i\boldsymbol{x}_i^T)_{jk} I_i\right) \\
&= \frac{1}{N^2}\sum_{i\in U_N} (\boldsymbol{b}_i\boldsymbol{x}_i^T)_{jk}^2 \mathrm{Var}(w_i I_i) \\
&= \frac{1}{N^2}\sum_{i\in U_N} (\mathrm{E}[w_i] - 1)(\boldsymbol{b}_i\boldsymbol{x}_i^T)_{jk}^2 \\
&= \frac{1}{N}\left(\frac{1}{N}\sum_{i\in U_N} (\mathrm{E}[w_i] - 1)(\boldsymbol{b}_i\boldsymbol{x}_i^T)_{jk}^2\right).
\end{aligned}$$

Then, as $N \to \infty$,

$$\begin{aligned}
&\mathrm{E}\left[\left(\frac{1}{N}(\boldsymbol{A}^T\boldsymbol{X})_{jk} - (\boldsymbol{M}_2)_{jk}\right)^2\right] \\
&= \mathrm{Var}\left(\frac{1}{N}(\boldsymbol{A}^T\boldsymbol{X})_{jk} - (\boldsymbol{M}_2)_{jk}\right) + \left(\mathrm{E}\left[\frac{1}{N}(\boldsymbol{A}^T\boldsymbol{X})_{jk} - (\boldsymbol{M}_2)_{jk}\right]\right)^2 \\
&= \mathrm{Var}\left(\frac{1}{N}(\boldsymbol{A}^T\boldsymbol{X})_{jk}\right) + \left(\frac{1}{N}\sum_{i\in U_N} \mathrm{E}[w_i(\boldsymbol{b}_i\boldsymbol{x}_i^T)_{jk} I_i] - (\boldsymbol{M}_2)_{jk}\right)^2 \\
&= \frac{1}{N}\left(\frac{1}{N}\sum_{i\in U_N} (\mathrm{E}[w_i] - 1)(\boldsymbol{b}_i\boldsymbol{x}_i^T)_{jk}^2\right) + \left(\frac{1}{N}\sum_{i\in U_N} (\boldsymbol{b}_i\boldsymbol{x}_i^T)_{jk} - (\boldsymbol{M}_2)_{jk}\right)^2 \\
&= \frac{1}{N}\Big((\boldsymbol{M}_4)_{jk} + o(1)\Big) + \Big((\boldsymbol{M}_2)_{jk} + o(1) - (\boldsymbol{M}_2)_{jk}\Big)^2
\end{aligned}$$

$$= \quad o(1)$$

by Assumptions 3.4.2 and 3.4.5, verifying

$$\frac{1}{N}(\boldsymbol{A}^T\boldsymbol{X})_{jk} \xrightarrow{m.s.} (\boldsymbol{M}_2)_{jk},$$

for each element, and thus

$$\frac{1}{N}\boldsymbol{A}^T\boldsymbol{X} \xrightarrow{P} \boldsymbol{M}_2. \tag{54}$$

Therefore, by (53) and (54), as $N \to \infty$,

$$\hat{\boldsymbol{\Gamma}}_0 = \left(\frac{1}{N}\boldsymbol{A}^T\boldsymbol{A}\right)^{-1}\left(\frac{1}{N}\boldsymbol{A}^T\boldsymbol{X}\right) \xrightarrow{P} \boldsymbol{M}_1^{-1}\boldsymbol{M}_2 = \boldsymbol{\Gamma}_0, \tag{55}$$

and (53) together with (55) yields

$$\frac{1}{N}\hat{\boldsymbol{X}}_0^T\hat{\boldsymbol{X}}_0 = \frac{1}{N}\hat{\boldsymbol{\Gamma}}_0^T\boldsymbol{A}^T\boldsymbol{A}\hat{\boldsymbol{\Gamma}}_0 = \hat{\boldsymbol{\Gamma}}_0^T\left(\frac{1}{N}\boldsymbol{A}^T\boldsymbol{A}\right)\hat{\boldsymbol{\Gamma}}_0 \xrightarrow{P} \boldsymbol{\Gamma}_0^T\boldsymbol{M}_1\boldsymbol{\Gamma}_0. \tag{56}$$

For the second factor of (52),

$$\frac{1}{N}\hat{\boldsymbol{X}}_0^T\boldsymbol{\epsilon} = \frac{1}{N}\hat{\boldsymbol{\Gamma}}_0^T\boldsymbol{A}^T\boldsymbol{\epsilon} = \hat{\boldsymbol{\Gamma}}_0^T\left(\frac{1}{N}\boldsymbol{A}^T\boldsymbol{\epsilon}\right),$$

and since convergence of $\hat{\boldsymbol{\Gamma}}_0$ has been established, it remains to show the convergence of $N^{-1}\boldsymbol{A}^T\boldsymbol{\epsilon}$ to a zero vector. First, consider

$$\begin{aligned}
\mathrm{E}[w_i\boldsymbol{b}_i\epsilon_i I_i] &= \mathrm{E}[\mathrm{E}[w_i\epsilon_i I_i|y_i]]\boldsymbol{b}_i \\
&= \mathrm{E}[w_i\epsilon_i\mathrm{E}[I_i|y_i]]\boldsymbol{b}_i \\
&= \mathrm{E}[w_i\epsilon_i\pi_i]\boldsymbol{b}_i \\
&= \mathrm{E}[\epsilon_i]\boldsymbol{b}_i \\
&= \boldsymbol{0},
\end{aligned} \tag{57}$$

where $\mathbf{0}$ is a $(K+2)$-vector of zeros, and by equation (45),

$$\mathrm{Var}(w_i \epsilon_i I_i) = \mathrm{E}[w_i \epsilon_i^2].$$

Next, let $\boldsymbol{t} \in \mathbb{R}^{K+2}$ and let $N \to \infty$ to find

$$\begin{aligned}
&\mathrm{E}\left[\left(\boldsymbol{t}^T \frac{1}{N} \boldsymbol{A}^T \boldsymbol{\epsilon} - 0\right)^2\right] \\
&= \mathrm{Var}\left(\boldsymbol{t}^T \frac{1}{N} \boldsymbol{A}^T \boldsymbol{\epsilon}\right) + \left(\mathrm{E}\left[\boldsymbol{t}^T \frac{1}{N} \boldsymbol{A}^T \boldsymbol{\epsilon}\right]\right)^2 \\
&= \frac{1}{N^2} \boldsymbol{t}^T \mathrm{Var}\left(\sum_{i \in U_N} w_i \boldsymbol{b}_i \epsilon_i I_i\right) \boldsymbol{t} + \left(\frac{1}{N} \sum_{i \in U_N} \boldsymbol{t}^T \mathrm{E}[w_i \boldsymbol{b}_i \epsilon_i I_i]\right)^2 \\
&= \frac{1}{N^2} \boldsymbol{t}^T \left(\sum_{i \in U_N} \boldsymbol{b}_i \mathrm{Var}(w_i \epsilon_i^2) \boldsymbol{b}_i^T\right) \boldsymbol{t} + \left(\frac{1}{N} \sum_{i \in U_N} \boldsymbol{t}^T \mathbf{0}\right)^2 \\
&= \frac{1}{N} \boldsymbol{t}^T \left(\frac{1}{N} \sum_{i \in U_N} \mathrm{E}[w_i \epsilon_i^2] \boldsymbol{b}_i \boldsymbol{b}_i^T\right) \boldsymbol{t} + 0 \\
&= \frac{1}{N} \boldsymbol{t}^T \Big(\boldsymbol{M}_5 + o(1)\Big) \boldsymbol{t} \\
&= o(1)
\end{aligned}$$

by Assumption 3.4.6, and using Assumption 3.3.5 for the variance of the sum. This result verifies

$$\boldsymbol{t}^T \frac{1}{N} \boldsymbol{A}^T \boldsymbol{\epsilon} \xrightarrow{m.s.} 0.$$

This implies

$$\boldsymbol{t}^T \frac{1}{N} \boldsymbol{A}^T \boldsymbol{\epsilon} \xrightarrow{d} \boldsymbol{t}^T \mathbf{0},$$

and by the Cramér-Wold device,

$$\frac{1}{N} \boldsymbol{A}^T \boldsymbol{\epsilon} \xrightarrow{d} \mathbf{0}.$$

Because the convergence is to a constant,

$$\frac{1}{N}\boldsymbol{A}^T\boldsymbol{\epsilon} \xrightarrow{P} \mathbf{0}, \tag{58}$$

and combining with (55), we get

$$\frac{1}{N}\hat{\boldsymbol{X}}_0^T\boldsymbol{\epsilon} = \hat{\boldsymbol{\Gamma}}_0^T\left(\frac{1}{N}\boldsymbol{A}^T\boldsymbol{\epsilon}\right) \xrightarrow{P} \boldsymbol{\Gamma}_0^T\mathbf{0} = \begin{bmatrix}0\\0\end{bmatrix}. \tag{59}$$

Equations (52), (56), and (59) together imply the consistency of the estimator,

$$(\hat{\boldsymbol{\beta}}_{2sls}^{K+2} - \boldsymbol{\beta}) \xrightarrow{P} (\boldsymbol{\Gamma}_0^T\boldsymbol{M}_1\boldsymbol{\Gamma}_0)^{-1}\begin{bmatrix}0\\0\end{bmatrix} = \mathbf{0}.$$

3.4.3 Central limit theorem results for $\hat{\boldsymbol{\beta}}_{2sls}^{K+2}$

To examine the asymptotic distribution of the estimator, start by writing

$$\begin{aligned}\sqrt{N}\left(\hat{\boldsymbol{\beta}}_{2sls} - \boldsymbol{\beta}\right) &= \sqrt{N}\Big(\boldsymbol{\beta} + (\hat{\boldsymbol{X}}_0^T\hat{\boldsymbol{X}}_0)^{-1}\hat{\boldsymbol{X}}_0^T\boldsymbol{\epsilon} - \boldsymbol{\beta}\Big)\\ &= \left(\frac{1}{N}\hat{\boldsymbol{X}}_0^T\hat{\boldsymbol{X}}_0\right)^{-1}\frac{1}{\sqrt{N}}\hat{\boldsymbol{X}}_0^T\boldsymbol{\epsilon}\\ &= \left(\frac{1}{N}\hat{\boldsymbol{X}}_0^T\hat{\boldsymbol{X}}_0\right)^{-1}\hat{\boldsymbol{\Gamma}}_0^T\left(\frac{1}{\sqrt{N}}\boldsymbol{A}^T\boldsymbol{\epsilon}\right).\end{aligned} \tag{60}$$

As $N \to \infty$, by (56),

$$\left(\frac{1}{N}\hat{\boldsymbol{X}}_0^T\hat{\boldsymbol{X}}_0\right)^{-1} \xrightarrow{P} (\boldsymbol{\Gamma}_0^T\boldsymbol{M}_1\boldsymbol{\Gamma}_0)^{-1},$$

a 2×2 matrix, and by (55),

$$\hat{\boldsymbol{\Gamma}}_0^T \xrightarrow{P} \boldsymbol{\Gamma}_0^T,$$

a $(K+2) \times 2$ matrix.

For the third factor of (60), select $\boldsymbol{t} \in \mathbb{R}^{K+2}$ to get the scalars

$$\mathrm{E}[\boldsymbol{t}^T w_i \boldsymbol{b}_i \epsilon_i I_i] = \boldsymbol{t}^T \mathrm{E}[w_i \boldsymbol{b}_i \epsilon_i I_i] = \boldsymbol{t}^T \mathbf{0} = 0,$$

using (57), and

$$\mathrm{Var}(\boldsymbol{t}^T w_i \boldsymbol{b}_i \epsilon_i I_i) = \boldsymbol{t}^T \mathrm{Var}(w_i \boldsymbol{b}_i \epsilon_i I_i) \boldsymbol{t} = \boldsymbol{t}^T (\mathrm{E}[w_i \epsilon_i^2] \boldsymbol{b}_i \boldsymbol{b}_i^T) \boldsymbol{t},$$

using (45).

Next, examine

$$\begin{aligned} \boldsymbol{t}^T \frac{1}{\sqrt{N}} \boldsymbol{A}^T \boldsymbol{\epsilon} &= \boldsymbol{t}^T \frac{1}{\sqrt{N}} \sum_{i \in U_N} w_i \boldsymbol{b}_i \epsilon_i I_i \\ &= \frac{1}{\sqrt{N}} \sum_{i \in U_N} \boldsymbol{t}^T w_i \boldsymbol{b}_i \epsilon_i I_i \\ &= \frac{1}{\sqrt{N}} V_N^{1/2} V_N^{-1/2} \sum_{i \in U_N} \left(\boldsymbol{t}^T w_i \boldsymbol{b}_i \epsilon_i I_i - 0 \right), \end{aligned}$$

with V_N as defined in Assumption 3.4.7. We have

$$\begin{aligned} \frac{1}{\sqrt{N}} V_N^{1/2} &= \left(\frac{1}{N} V_N \right)^{1/2} \\ &= \left(\boldsymbol{t}^T \left(\frac{1}{N} \sum_{i \in U_N} \mathrm{E}[w_i \epsilon_i^2] \boldsymbol{b}_i \boldsymbol{b}_i^T \right) \boldsymbol{t} \right)^{1/2} \\ &\to \left(\boldsymbol{t}^T \boldsymbol{M}_5 \boldsymbol{t} \right)^{1/2} \end{aligned}$$

by Assumption 3.4.6, and by Assumption 3.4.7 and the Lyapunov Central Limit Theorem,

$$V_N^{-1/2} \sum_{i \in U_N} \left(\boldsymbol{t}^T w_i \boldsymbol{b}_i \epsilon_i I_i - 0 \right) \xrightarrow{d} N(0, 1).$$

Applying Slutsky's Theorem,

$$\begin{aligned} \boldsymbol{t}^T \frac{1}{\sqrt{N}} \boldsymbol{A}^T \boldsymbol{\epsilon} &\xrightarrow{d} \left(\boldsymbol{t}^T \boldsymbol{M}_5 \boldsymbol{t}\right)^{1/2} N(0,1) \\ &= N(0, \boldsymbol{t}^T \boldsymbol{M}_5 \boldsymbol{t}) \\ &= N(\boldsymbol{t}^T \boldsymbol{0}, \boldsymbol{t}^T \boldsymbol{M}_5 \boldsymbol{t}) \\ &= \boldsymbol{t}^T N(\boldsymbol{0}, \boldsymbol{M}_5), \end{aligned}$$

so by the Cramér-Wold device,

$$\frac{1}{\sqrt{N}} \boldsymbol{A}^T \boldsymbol{\epsilon} \xrightarrow{d} N(\boldsymbol{0}, \boldsymbol{M}_5). \tag{61}$$

Another application of Slutsky's Theorem using (55), (56), (60), and (61) yields

$$\begin{aligned} \sqrt{N}\left(\hat{\boldsymbol{\beta}}_{2sls}^{K+2} - \boldsymbol{\beta}\right) &\xrightarrow{d} (\boldsymbol{\Gamma}_0^T \boldsymbol{M}_1 \boldsymbol{\Gamma}_0)^{-1} \boldsymbol{\Gamma}_0^T N(\boldsymbol{0}, \boldsymbol{M}_5) \\ &= N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, (\boldsymbol{\Gamma}_0^T \boldsymbol{M}_1 \boldsymbol{\Gamma}_0)^{-1} \boldsymbol{\Gamma}_0^T \boldsymbol{M}_5 \boldsymbol{\Gamma}_0 (\boldsymbol{\Gamma}_0^T \boldsymbol{M}_1 \boldsymbol{\Gamma}_0)^{-1}\right), \end{aligned}$$

so that

$$\hat{\boldsymbol{\beta}}_{2sls}^{K+2} \text{ is } AN\left(\boldsymbol{\beta}, \frac{1}{N}(\boldsymbol{\Gamma}_0^T \boldsymbol{M}_1 \boldsymbol{\Gamma}_0)^{-1} \boldsymbol{\Gamma}_0^T \boldsymbol{M}_5 \boldsymbol{\Gamma}_0 (\boldsymbol{\Gamma}_0^T \boldsymbol{M}_1 \boldsymbol{\Gamma}_0)^{-1}\right).$$

3.5 Penalized spline estimator, $\hat{\boldsymbol{\beta}}_{pspl}$

For this new estimator, the first stage of the two-stage process uses penalized splines and IV's to determine the "fitted" auxiliary variable vector $\hat{\boldsymbol{x}}$. As in the previous two sections, stage two is an ordinary least squares regression of y on $\hat{\boldsymbol{x}}$. The smoothing parameter for the penalized splines is denoted as λ_N. Later in this section we demonstrate that when $\lambda_N \to \infty$, the penalized spline estimator behaves like the probability weighted least squares estimator, which is also the classical two-stage least squares estimator with two instrumental variables, $\hat{\boldsymbol{\beta}}_{2sls}$. Also, when $\lambda_N = 0$,

the resulting estimator is the classical two-stage least squares estimator with $K+2$ instrumental variables, $\hat{\boldsymbol{\beta}}_{2sls}^{K+2}$. As verified previously, both of these estimators are consistent for $\boldsymbol{\beta}$. When λ_N is allowed to vary between 0 and ∞, the resulting estimator is also shown to be consistent, provided $\lambda_N^2/N \to 0$ as $N \to \infty$. In finite samples, however, a bias term $\boldsymbol{\Delta}_N\boldsymbol{\beta}$ is present. Using simulations, we demonstrate that in certain informative sampling cases the finite-sample variance of the estimator for some values of λ_N between 0 and ∞ is smaller than at the two extremes of λ_N. An examination of the trade-off between reduced variance and the square of the induced bias shows that an "optimal" λ_N can be selected to minimize the mean squared error of the estimator. This optimal λ_N can be estimated from the sample data.

In this section, we evaluate the penalized spline estimator under the typical conditions one would find in the informative sampling survey situation. The IV matrix is $\boldsymbol{A} = \boldsymbol{W}\boldsymbol{B}$ and we define $\boldsymbol{D}$ to be the $(K+2)\times(K+2)$ diagonal matrix

$$\boldsymbol{D} = \begin{bmatrix} \mathbf{0}_{2\times 2} & \mathbf{0}_{2\times K} \\ \mathbf{0}_{K\times 2} & \boldsymbol{I}_{K\times K} \end{bmatrix}.$$

In addition, define $\hat{\boldsymbol{\Gamma}}_\lambda$ as the $(K+2)\times 2$ matrix of stage one coefficients,

$$\hat{\boldsymbol{\Gamma}}_\lambda = (\boldsymbol{A}^T\boldsymbol{A} + \lambda_N^2\boldsymbol{D})^{-1}\boldsymbol{A}^T\boldsymbol{X},$$

so that after the first stage, the $n_N \times 2$ matrix of "fitted" vectors is

$$\hat{\boldsymbol{X}}_\lambda = \boldsymbol{A}\hat{\boldsymbol{\Gamma}}_\lambda,$$

and after the second stage,

$$\hat{\boldsymbol{\beta}}_{pspl} = (\hat{\boldsymbol{X}}_\lambda^T\hat{\boldsymbol{X}}_\lambda)^{-1}\hat{\boldsymbol{X}}_\lambda^T\boldsymbol{y}. \tag{62}$$

3.5.1 Assumptions

Make the following additional assumption,

Assumption 3.5.1. $\lim_{N\to\infty} N^{-1}\lambda_N^2 = 0$.

3.5.2 Consistency results for $\hat{\boldsymbol{\beta}}_{pspl}$

We define the 2×2 matrix

$$\boldsymbol{\Delta}_N = (\hat{\boldsymbol{X}}_\lambda^T \hat{\boldsymbol{X}}_\lambda)^{-1} \hat{\boldsymbol{X}}_\lambda^T (\boldsymbol{X} - \hat{\boldsymbol{X}}_\lambda),$$

and write

$$\begin{aligned}
(\hat{\boldsymbol{X}}_\lambda^T \hat{\boldsymbol{X}}_\lambda)^{-1} \hat{\boldsymbol{X}}_\lambda^T \boldsymbol{X}\boldsymbol{\beta} &= (\hat{\boldsymbol{X}}_\lambda^T \hat{\boldsymbol{X}}_\lambda)^{-1} \hat{\boldsymbol{X}}_\lambda^T (\hat{\boldsymbol{X}}_\lambda + \boldsymbol{X} - \hat{\boldsymbol{X}}_\lambda)\boldsymbol{\beta} \\
&= (\hat{\boldsymbol{X}}_\lambda^T \hat{\boldsymbol{X}}_\lambda)^{-1} \hat{\boldsymbol{X}}_\lambda^T \hat{\boldsymbol{X}}_\lambda \boldsymbol{\beta} + (\hat{\boldsymbol{X}}_\lambda^T \hat{\boldsymbol{X}}_\lambda)^{-1} \hat{\boldsymbol{X}}_\lambda^T (\boldsymbol{X} - \hat{\boldsymbol{X}}_\lambda)\boldsymbol{\beta} \\
&= \boldsymbol{\beta} + \boldsymbol{\Delta}_N \boldsymbol{\beta}.
\end{aligned}$$

Together with (62) this yields

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{pspl} - \boldsymbol{\beta} &= (\hat{\boldsymbol{X}}_\lambda^T \hat{\boldsymbol{X}}_\lambda)^{-1} \hat{\boldsymbol{X}}_\lambda^T \boldsymbol{y} - \boldsymbol{\beta} \\
&= (\hat{\boldsymbol{X}}_\lambda^T \hat{\boldsymbol{X}}_\lambda)^{-1} \hat{\boldsymbol{X}}_\lambda^T \boldsymbol{X}\boldsymbol{\beta} + (\hat{\boldsymbol{X}}_\lambda^T \hat{\boldsymbol{X}}_\lambda)^{-1} \hat{\boldsymbol{X}}_\lambda^T \boldsymbol{\epsilon} - \boldsymbol{\beta} \\
&= \boldsymbol{\beta} + \boldsymbol{\Delta}_N \boldsymbol{\beta} + (\hat{\boldsymbol{X}}_\lambda^T \hat{\boldsymbol{X}}_\lambda)^{-1} \hat{\boldsymbol{X}}_\lambda^T \boldsymbol{\epsilon} - \boldsymbol{\beta} \\
&= \boldsymbol{\Delta}_N \boldsymbol{\beta} + \left(\frac{1}{N} \hat{\boldsymbol{X}}_\lambda^T \hat{\boldsymbol{X}}_\lambda \right)^{-1} \left(\frac{1}{N} \hat{\boldsymbol{X}}_\lambda^T \boldsymbol{\epsilon} \right). \qquad (63)
\end{aligned}$$

We show that this converges to a zero vector.

As $N \to \infty$, by (53), (54), and Assumption 3.5.1,

$$\begin{aligned}
\hat{\boldsymbol{\Gamma}}_\lambda &= (\boldsymbol{A}^T \boldsymbol{A} + \lambda_N^2 \boldsymbol{D})^{-1} \boldsymbol{A}^T \boldsymbol{X} \\
&= \left(\frac{1}{N} (\boldsymbol{A}^T \boldsymbol{A} + \lambda_N^2 \boldsymbol{D}) \right)^{-1} \left(\frac{1}{N} \boldsymbol{A}^T \boldsymbol{X} \right)
\end{aligned}$$

$$
\begin{aligned}
&= \left(\frac{1}{N}\boldsymbol{A}^T\boldsymbol{A} + \frac{1}{N}\lambda_N^2\boldsymbol{D}\right)^{-1}\left(\frac{1}{N}\boldsymbol{A}^T\boldsymbol{X}\right) \\
&\xrightarrow{P} (\boldsymbol{M}_1 + \boldsymbol{0})^{-1}\boldsymbol{M}_2 = \boldsymbol{\Gamma}_0, \qquad (64)
\end{aligned}
$$

which leads to

$$
\frac{1}{N}\hat{\boldsymbol{X}}_\lambda^T\hat{\boldsymbol{X}}_\lambda = \frac{1}{N}\hat{\boldsymbol{\Gamma}}_\lambda^T\boldsymbol{A}^T\boldsymbol{A}\hat{\boldsymbol{\Gamma}}_\lambda = \hat{\boldsymbol{\Gamma}}_\lambda^T\left(\frac{1}{N}\boldsymbol{A}^T\boldsymbol{A}\right)\hat{\boldsymbol{\Gamma}}_\lambda \xrightarrow{P} \boldsymbol{\Gamma}_0^T\boldsymbol{M}_1\boldsymbol{\Gamma}_0, \qquad (65)
$$

and

$$
\frac{1}{N}\hat{\boldsymbol{X}}_\lambda^T\boldsymbol{X} = \frac{1}{N}\hat{\boldsymbol{\Gamma}}_\lambda^T\boldsymbol{A}^T\boldsymbol{X} = \hat{\boldsymbol{\Gamma}}_\lambda^T\left(\frac{1}{N}\boldsymbol{A}^T\boldsymbol{X}\right) \xrightarrow{P} \boldsymbol{\Gamma}_0^T\boldsymbol{M}_2,
$$

so that

$$
\begin{aligned}
\boldsymbol{\Delta}_N &= (\hat{\boldsymbol{X}}_\lambda^T\hat{\boldsymbol{X}}_\lambda)^{-1}\hat{\boldsymbol{X}}_\lambda^T(\boldsymbol{X} - \hat{\boldsymbol{X}}_\lambda) \\
&= \left(\frac{1}{N}\hat{\boldsymbol{X}}_\lambda^T\hat{\boldsymbol{X}}_\lambda\right)^{-1}\frac{1}{N}\hat{\boldsymbol{X}}_\lambda^T(\boldsymbol{X} - \hat{\boldsymbol{X}}_\lambda) \\
&= \left(\frac{1}{N}\hat{\boldsymbol{X}}_\lambda^T\hat{\boldsymbol{X}}_\lambda\right)^{-1}\left(\frac{1}{N}\hat{\boldsymbol{X}}_\lambda^T\boldsymbol{X} - \frac{1}{N}\hat{\boldsymbol{X}}_\lambda^T\hat{\boldsymbol{X}}_\lambda\right) \\
&\xrightarrow{P} (\boldsymbol{\Gamma}_0^T\boldsymbol{M}_1\boldsymbol{\Gamma}_0)^{-1}(\boldsymbol{\Gamma}_0^T\boldsymbol{M}_2 - \boldsymbol{\Gamma}_0^T\boldsymbol{M}_1\boldsymbol{\Gamma}_0) \\
&= (\boldsymbol{\Gamma}_0^T\boldsymbol{M}_1\boldsymbol{\Gamma}_0)^{-1}\boldsymbol{\Gamma}_0^T\boldsymbol{M}_2 - \boldsymbol{I}_{2\times 2} \\
&= (\boldsymbol{M}_2^T\boldsymbol{M}_1^{-1}\boldsymbol{M}_1\boldsymbol{M}_1^{-1}\boldsymbol{M}_2)^{-1}\boldsymbol{M}_2^T\boldsymbol{M}_1^{-1}\boldsymbol{M}_2 - \boldsymbol{I}_{2\times 2} \\
&= (\boldsymbol{M}_2^T\boldsymbol{M}_1^{-1}\boldsymbol{M}_2)^{-1}\boldsymbol{M}_2^T\boldsymbol{M}_1^{-1}\boldsymbol{M}_2 - \boldsymbol{I}_{2\times 2} \\
&= \boldsymbol{0},
\end{aligned}
$$

and

$$
\boldsymbol{\Delta}_N\boldsymbol{\beta} \xrightarrow{P} \boldsymbol{0}. \qquad (66)
$$

Results (58) and (64) show that

$$
\frac{1}{N}\hat{\boldsymbol{X}}_\lambda^T\boldsymbol{\epsilon} = \frac{1}{N}\hat{\boldsymbol{\Gamma}}_\lambda^T\boldsymbol{A}^T\boldsymbol{\epsilon}
$$

$$
\begin{aligned}
&= \hat{\boldsymbol{\Gamma}}_\lambda^T \left(\frac{1}{N} \boldsymbol{A}^T \boldsymbol{\epsilon} \right) \\
&\xrightarrow{P} \boldsymbol{\Gamma}_0^T \mathbf{0} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \qquad (67)
\end{aligned}
$$

Equations (63), (65), (66), and (67) together imply the consistency of the penalized spline estimator,

$$
(\hat{\boldsymbol{\beta}}_{pspl} - \boldsymbol{\beta}) \xrightarrow{P} \left(\mathbf{0} + (\boldsymbol{\Gamma}_0^T \boldsymbol{M}_1 \boldsymbol{\Gamma}_0)^{-1} \begin{bmatrix} 0 \\ 0 \end{bmatrix} \right) = \mathbf{0}.
$$

3.5.3 Central limit theorem results for $\hat{\boldsymbol{\beta}}_{pspl}$

To examine the asymptotic distribution of the estimator, start by writing

$$
\begin{aligned}
&\sqrt{N} \left(\hat{\boldsymbol{\beta}}_{pspl} - \boldsymbol{\beta} - \boldsymbol{\Delta}_N \boldsymbol{\beta} \right) \\
&= \sqrt{N} \left((\hat{\boldsymbol{X}}_\lambda^T \hat{\boldsymbol{X}}_\lambda)^{-1} \hat{\boldsymbol{X}}_\lambda^T \boldsymbol{X} \boldsymbol{\beta} + (\hat{\boldsymbol{X}}_\lambda^T \hat{\boldsymbol{X}}_\lambda)^{-1} \hat{\boldsymbol{X}}_\lambda^T \boldsymbol{\epsilon} - \boldsymbol{\beta} - (\hat{\boldsymbol{X}}_\lambda^T \hat{\boldsymbol{X}}_\lambda)^{-1} \hat{\boldsymbol{X}}_\lambda^T (\boldsymbol{X} - \hat{\boldsymbol{X}}_\lambda) \boldsymbol{\beta} \right) \\
&= \sqrt{N} \left((\hat{\boldsymbol{X}}_\lambda^T \hat{\boldsymbol{X}}_\lambda)^{-1} \hat{\boldsymbol{X}}_\lambda^T \boldsymbol{X} \boldsymbol{\beta} + (\hat{\boldsymbol{X}}_\lambda^T \hat{\boldsymbol{X}}_\lambda)^{-1} \hat{\boldsymbol{X}}_\lambda^T \boldsymbol{\epsilon} - \boldsymbol{\beta} - (\hat{\boldsymbol{X}}_\lambda^T \hat{\boldsymbol{X}}_\lambda)^{-1} \hat{\boldsymbol{X}}_\lambda^T \boldsymbol{X} \boldsymbol{\beta} + \boldsymbol{\beta} \right) \\
&= \sqrt{N} (\hat{\boldsymbol{X}}_\lambda^T \hat{\boldsymbol{X}}_\lambda)^{-1} \hat{\boldsymbol{X}}_\lambda^T \boldsymbol{\epsilon} \\
&= (\hat{\boldsymbol{X}}_\lambda^T \hat{\boldsymbol{X}}_\lambda)^{-1} \sqrt{N} \hat{\boldsymbol{X}}_\lambda^T \boldsymbol{\epsilon} \\
&= \left(\frac{1}{N} \hat{\boldsymbol{X}}_\lambda^T \hat{\boldsymbol{X}}_\lambda \right)^{-1} \frac{1}{\sqrt{N}} \hat{\boldsymbol{X}}_\lambda^T \boldsymbol{\epsilon} \\
&= \left(\frac{1}{N} \hat{\boldsymbol{X}}_\lambda^T \hat{\boldsymbol{X}}_\lambda \right)^{-1} \hat{\boldsymbol{\Gamma}}_\lambda^T \left(\frac{1}{\sqrt{N}} \boldsymbol{A}^T \boldsymbol{\epsilon} \right). \qquad (68)
\end{aligned}
$$

As $N \to \infty$, by (65),

$$
\left(\frac{1}{N} \hat{\boldsymbol{X}}_\lambda^T \hat{\boldsymbol{X}}_\lambda \right)^{-1} \xrightarrow{P} \left(\boldsymbol{\Gamma}_0^T \boldsymbol{M}_1 \boldsymbol{\Gamma}_0 \right)^{-1},
$$

and by (64),

$$
\hat{\boldsymbol{\Gamma}}_\lambda^T \xrightarrow{P} \boldsymbol{\Gamma}_0^T,
$$

and also by (61),

$$\frac{1}{\sqrt{N}}\boldsymbol{A}^T\boldsymbol{\epsilon} \xrightarrow{d} N(\mathbf{0}, \boldsymbol{M}_5).$$

An application of Slutsky's Theorem using (61), (64), (65), and (68) yields

$$\begin{aligned}\sqrt{N}\left(\hat{\boldsymbol{\beta}}_{pspl} - \boldsymbol{\beta} - \boldsymbol{\Delta}_N\boldsymbol{\beta}\right) &\xrightarrow{d} (\boldsymbol{\Gamma}_0^T\boldsymbol{M}_1\boldsymbol{\Gamma}_0)^{-1}\boldsymbol{\Gamma}_0^T N(\mathbf{0}, \boldsymbol{M}_5) \\ &= N\left(\begin{bmatrix}0\\0\end{bmatrix}, (\boldsymbol{\Gamma}_0^T\boldsymbol{M}_1\boldsymbol{\Gamma}_0)^{-1}\boldsymbol{\Gamma}_0^T\boldsymbol{M}_5\boldsymbol{\Gamma}_0(\boldsymbol{\Gamma}_0^T\boldsymbol{M}_1\boldsymbol{\Gamma}_0)^{-1}\right),\end{aligned}$$

so that

$$\hat{\boldsymbol{\beta}}_{pspl} \text{ is } AN\left(\boldsymbol{\beta} + \boldsymbol{\Delta}_N\boldsymbol{\beta}, \frac{1}{N}(\boldsymbol{\Gamma}_0^T\boldsymbol{M}_1\boldsymbol{\Gamma}_0)^{-1}\boldsymbol{\Gamma}_0^T\boldsymbol{M}_5\boldsymbol{\Gamma}_0(\boldsymbol{\Gamma}_0^T\boldsymbol{M}_1\boldsymbol{\Gamma}_0)^{-1}\right). \tag{69}$$

CLT results for $\hat{\boldsymbol{\beta}}_{pspl}$, when $\lambda_N = 0$

In this subsection and the next we examine the penalized spline estimator $\hat{\boldsymbol{\beta}}_{pspl}$ at the two extreme values for λ_N. When $\lambda_N = 0$,

$$\hat{\boldsymbol{\Gamma}}_\lambda = (\boldsymbol{A}^T\boldsymbol{A} + 0^2\boldsymbol{D})^{-1}\boldsymbol{A}^T\boldsymbol{X} = \hat{\boldsymbol{\Gamma}}_0,$$

so that $\hat{\boldsymbol{\beta}}_{pspl}$ is equivalent to $\hat{\boldsymbol{\beta}}_{2sls}^{K+2}$, the consistent estimator discussed in Section 3.4.

CLT results for $\hat{\boldsymbol{\beta}}_{pspl}$, when $\lambda_N \to \infty$

Here we assume $\lambda_N^2 \to \infty$ faster than $N \to \infty$, so that $N^{-1}\lambda_N^2 \to \infty$. Under this condition, we demonstrate that the finite-sample bias term vanishes and $\hat{\boldsymbol{\beta}}_{pspl}$ has the same asymptotic variance as $\hat{\boldsymbol{\beta}}_{2sls}$. We define the 2×2 matrix

$$\hat{\boldsymbol{\Gamma}}_x = (\boldsymbol{A}_x^T\boldsymbol{A}_x)^{-1}\boldsymbol{A}_x^T\boldsymbol{X},$$

so that $\hat{\boldsymbol{X}}_x = \boldsymbol{A}_x\hat{\boldsymbol{\Gamma}}_x$. At stage one of our two-stage process, when $\lambda_N \to \infty$, our

estimate for $\hat{\boldsymbol{X}}_\lambda$ converges to $\hat{\boldsymbol{X}}_x$. To verify this, we write

$$\boldsymbol{A} = \boldsymbol{W}\boldsymbol{B} = \boldsymbol{W}\begin{bmatrix}\boldsymbol{X} & \boldsymbol{Z}\end{bmatrix} = \begin{bmatrix}\boldsymbol{W}\boldsymbol{X} & \boldsymbol{W}\boldsymbol{Z}\end{bmatrix} = \begin{bmatrix}\boldsymbol{A}_x & \boldsymbol{W}\boldsymbol{Z}\end{bmatrix},$$

so that

$$\boldsymbol{A}^T\boldsymbol{A} = \begin{bmatrix}\boldsymbol{A}_x^T\boldsymbol{A}_x & \boldsymbol{A}_x^T\boldsymbol{W}\boldsymbol{Z} \\ \boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{A}_x & \boldsymbol{Z}^T\boldsymbol{W}^2\boldsymbol{Z}\end{bmatrix},$$

and

$$\begin{aligned}
\hat{\boldsymbol{\Gamma}}_\lambda &= (\boldsymbol{A}^T\boldsymbol{A} + \lambda_N^2\boldsymbol{D})^{-1}\boldsymbol{A}^T\boldsymbol{X} \\
&= \begin{bmatrix}\boldsymbol{A}_x^T\boldsymbol{A}_x & \boldsymbol{A}_x^T\boldsymbol{W}\boldsymbol{Z} \\ \boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{A}_x & \boldsymbol{Z}^T\boldsymbol{W}^2\boldsymbol{Z} + \lambda_N^2\boldsymbol{I}_{K\times K}\end{bmatrix}^{-1}\begin{bmatrix}\boldsymbol{A}_x^T\boldsymbol{X} \\ \boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{X}\end{bmatrix} \\
&= \begin{bmatrix}\boldsymbol{G}_{11} & \boldsymbol{G}_{12} \\ \boldsymbol{G}_{21} & \boldsymbol{G}_{22}\end{bmatrix}\begin{bmatrix}\boldsymbol{A}_x^T\boldsymbol{X} \\ \boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{X}\end{bmatrix},
\end{aligned}$$

where

$$\begin{aligned}
\boldsymbol{G}_{11} &= (\boldsymbol{A}_x^T\boldsymbol{A}_x)^{-1} + (\boldsymbol{A}_x^T\boldsymbol{A}_x)^{-1}\boldsymbol{A}_x^T\boldsymbol{W}\boldsymbol{Z}\boldsymbol{G}_{22}\boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{A}_x(\boldsymbol{A}_x^T\boldsymbol{A}_x)^{-1}, \\
\boldsymbol{G}_{12} &= -(\boldsymbol{A}_x^T\boldsymbol{A}_x)^{-1}\boldsymbol{A}_x^T\boldsymbol{W}\boldsymbol{Z}\boldsymbol{G}_{22}, \\
\boldsymbol{G}_{21} &= -\boldsymbol{G}_{22}\boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{A}_x(\boldsymbol{A}_x^T\boldsymbol{A}_x)^{-1}, \text{ and} \\
\boldsymbol{G}_{22} &= \left(\boldsymbol{Z}^T\boldsymbol{W}^2\boldsymbol{Z} + \lambda_N^2\boldsymbol{I}_{K\times K} - \boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{A}_x(\boldsymbol{A}_x^T\boldsymbol{A}_x)^{-1}\boldsymbol{A}_x^T\boldsymbol{W}\boldsymbol{Z}\right)^{-1} \\
&= \tfrac{1}{\lambda_N^2}\left(\tfrac{1}{\lambda_N^2}\boldsymbol{Z}^T\boldsymbol{W}^2\boldsymbol{Z} + \boldsymbol{I}_{K\times K} - \tfrac{1}{\lambda_N^2}\boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{A}_x(\boldsymbol{A}_x^T\boldsymbol{A}_x)^{-1}\boldsymbol{A}_x^T\boldsymbol{W}\boldsymbol{Z}\right)^{-1}.
\end{aligned}$$

As $\lambda_N \to \infty$, $\boldsymbol{G}_{22} \to \boldsymbol{0}_{K\times K}$, implying that $\boldsymbol{G}_{11} \to (\boldsymbol{A}_x^T\boldsymbol{A}_x)^{-1}$, $\boldsymbol{G}_{12} \to \boldsymbol{0}_{2\times K}$, and $\boldsymbol{G}_{21} \to \boldsymbol{0}_{K\times 2}$. We denote the convergent $\hat{\boldsymbol{\Gamma}}_\lambda$ as $\hat{\boldsymbol{\Gamma}}_\infty$ where

$$\begin{aligned}
\hat{\boldsymbol{\Gamma}}_\infty &= \begin{bmatrix}(\boldsymbol{A}_x^T\boldsymbol{A}_x)^{-1} & \boldsymbol{0}_{2\times K} \\ \boldsymbol{0}_{K\times 2} & \boldsymbol{0}_{K\times K}\end{bmatrix}\begin{bmatrix}\boldsymbol{A}_x^T\boldsymbol{X} \\ \boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{X}\end{bmatrix} \\
&= \begin{bmatrix}(\boldsymbol{A}_x^T\boldsymbol{A}_x)^{-1}\boldsymbol{A}_x^T\boldsymbol{X} \\ \boldsymbol{0}_{K\times 2}\end{bmatrix} = \begin{bmatrix}\hat{\boldsymbol{\Gamma}}_x \\ \boldsymbol{0}\end{bmatrix}.
\end{aligned}$$

Because $\hat{\boldsymbol{X}}_\lambda = \boldsymbol{A}\hat{\boldsymbol{\Gamma}}_\lambda$, as $\lambda_N \to \infty$,

$$\hat{\boldsymbol{X}}_\lambda \to \boldsymbol{A}\hat{\boldsymbol{\Gamma}}_\infty = \begin{bmatrix} \boldsymbol{A}_x & \boldsymbol{W}\boldsymbol{Z} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\Gamma}}_x \\ \boldsymbol{0} \end{bmatrix} = \boldsymbol{A}_x\hat{\boldsymbol{\Gamma}}_x = \hat{\boldsymbol{X}}_x.$$

As a consequence of this result, when $\lambda_N \to \infty$ the bias term $\boldsymbol{\Delta}_N\boldsymbol{\beta}$ vanishes because

$$\begin{aligned} \hat{\boldsymbol{X}}_x^T\hat{\boldsymbol{X}}_x &= \hat{\boldsymbol{\Gamma}}_x^T\boldsymbol{A}_x^T\boldsymbol{A}_x\hat{\boldsymbol{\Gamma}}_x \\ &= \hat{\boldsymbol{\Gamma}}_x^T\boldsymbol{A}_x^T\boldsymbol{A}_x(\boldsymbol{A}_x^T\boldsymbol{A}_x)^{-1}\boldsymbol{A}_x^T\boldsymbol{X} \\ &= \hat{\boldsymbol{\Gamma}}_x^T\boldsymbol{A}_x^T\boldsymbol{X} \\ &= \hat{\boldsymbol{X}}_x^T\boldsymbol{X}, \end{aligned}$$

and

$$\begin{aligned} \boldsymbol{\Delta}_N &= (\hat{\boldsymbol{X}}_\lambda^T\hat{\boldsymbol{X}}_\lambda)^{-1}\hat{\boldsymbol{X}}_\lambda^T(\boldsymbol{X} - \hat{\boldsymbol{X}}_\lambda) \\ &= (\hat{\boldsymbol{X}}_\lambda^T\hat{\boldsymbol{X}}_\lambda)^{-1}\hat{\boldsymbol{X}}_\lambda^T\boldsymbol{X} - \boldsymbol{I}_{2\times 2} \\ &\to (\hat{\boldsymbol{X}}_x^T\hat{\boldsymbol{X}}_x)^{-1}\hat{\boldsymbol{X}}_x^T\boldsymbol{X} - \boldsymbol{I}_{2\times 2} \\ &= \boldsymbol{I}_{2\times 2} - \boldsymbol{I}_{2\times 2} \\ &= \boldsymbol{0}. \end{aligned}$$

We turn our focus to the asymptotic variance. We start by defining

$$\boldsymbol{P}_8 = \lim_{N\to\infty} \frac{1}{N} \sum_{i\in U_N} \mathrm{E}[w_i]\boldsymbol{x}_i\boldsymbol{x}_i^T,$$

as the symmetric, positive definite 2×2 submatrix of $\boldsymbol{M}_1$, such that

$$\boldsymbol{M}_1 = \begin{bmatrix} \boldsymbol{P}_8 & \lim_{N\to\infty} \frac{1}{N}\mathrm{E}[w_i]\boldsymbol{x}_i\boldsymbol{z}_i^T \\ \lim_{N\to\infty} \frac{1}{N}\mathrm{E}[w_i]\boldsymbol{z}_i\boldsymbol{x}_i^T & \lim_{N\to\infty} \frac{1}{N}\mathrm{E}[w_i]\boldsymbol{z}_i\boldsymbol{z}_i^T \end{bmatrix}.$$

By result (53), $N^{-1}\boldsymbol{A}^T\boldsymbol{A} \xrightarrow{P} \boldsymbol{M}_1$, and therefore

$$\frac{1}{N}\boldsymbol{A}_x^T\boldsymbol{A}_x \xrightarrow{P} \boldsymbol{P}_8. \tag{70}$$

We next refer to result (43),

$$\frac{1}{N}\boldsymbol{A}_x^T\boldsymbol{X} = \frac{1}{N}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X} \xrightarrow{P} \boldsymbol{P}_5, \tag{71}$$

and combine with (70) to show

$$\hat{\boldsymbol{\Gamma}}_\infty = \begin{bmatrix} \left(\frac{1}{N}\boldsymbol{A}_x^T\boldsymbol{A}_x\right)^{-1}\left(\frac{1}{N}\boldsymbol{A}_x^T\boldsymbol{X}\right) \\ \mathbf{0} \end{bmatrix} \xrightarrow{P} \begin{bmatrix} \boldsymbol{P}_8^{-1}\boldsymbol{P}_5 \\ \mathbf{0} \end{bmatrix} = \boldsymbol{\Gamma}_\infty.$$

In the general $\hat{\boldsymbol{\beta}}_{pspl}$ case, $\hat{\boldsymbol{\Gamma}}_\lambda$ converges to $\boldsymbol{\Gamma}_0$, but here, $\hat{\boldsymbol{\Gamma}}_\lambda$ converges to $\boldsymbol{\Gamma}_\infty$, so for the asymptotic variance term we examine

$$(\boldsymbol{\Gamma}_\infty^T\boldsymbol{M}_1\boldsymbol{\Gamma}_\infty)^{-1}\boldsymbol{\Gamma}_\infty^T\boldsymbol{M}_5\boldsymbol{\Gamma}_\infty(\boldsymbol{\Gamma}_\infty^T\boldsymbol{M}_1\boldsymbol{\Gamma}_\infty)^{-1}.$$

Basic matrix calculations show

$$\boldsymbol{\Gamma}_\infty^T\boldsymbol{M}_1\boldsymbol{\Gamma}_\infty = \boldsymbol{P}_5\boldsymbol{P}_8^{-1}\boldsymbol{P}_5,$$

a positive definite 2×2 matrix, and since

$$\boldsymbol{M}_5 = \begin{bmatrix} \boldsymbol{P}_7 \\ \lim\limits_{N\to\infty} \frac{1}{N} \sum\limits_{i\in U_N} \mathrm{E}[w_i\epsilon_i^2]\boldsymbol{z}_i\boldsymbol{z}_i^T \end{bmatrix},$$

further calculations demonstrate

$$\boldsymbol{\Gamma}_\infty^T\boldsymbol{M}_5\boldsymbol{\Gamma}_\infty = \boldsymbol{P}_5\boldsymbol{P}_8^{-1}\boldsymbol{P}_7\boldsymbol{P}_8^{-1}\boldsymbol{P}_5,$$

also a 2×2 matrix. Combining these results yields

$$(\boldsymbol{\Gamma}_\infty^T \boldsymbol{M}_1 \boldsymbol{\Gamma}_\infty)^{-1} \boldsymbol{\Gamma}_\infty^T \boldsymbol{M}_5 \boldsymbol{\Gamma}_\infty (\boldsymbol{\Gamma}_\infty^T \boldsymbol{M}_1 \boldsymbol{\Gamma}_\infty)^{-1} = \boldsymbol{P}_5^{-1} \boldsymbol{P}_7 \boldsymbol{P}_5^{-1},$$

so that under the condition stated at the beginning of the subsection, $\hat{\boldsymbol{\beta}}_{pspl}$ has the same asymptotic variance as the classical two-stage least squares estimator with two instrumental variables, $\hat{\boldsymbol{\beta}}_{2sls}$.

3.6 Estimating the optimal λ_N

For finite samples we need to determine the λ_N that minimizes the MSE of $\hat{\boldsymbol{\beta}}_{pspl}$. To estimate this optimal λ_N, we refer to (69) and develop estimates for the bias and the asymptotic variance of $\hat{\boldsymbol{\beta}}_{pspl}$, both of which depend on λ_N via $\hat{\boldsymbol{\Gamma}}_\lambda$.

The bias term is

$$\begin{aligned} \boldsymbol{\Delta}_N \boldsymbol{\beta} &= (\hat{\boldsymbol{X}}_\lambda^T \hat{\boldsymbol{X}}_\lambda)^{-1} \hat{\boldsymbol{X}}_\lambda^T (\boldsymbol{X} - \hat{\boldsymbol{X}}_\lambda) \boldsymbol{\beta} \\ &= \left((\hat{\boldsymbol{X}}_\lambda^T \hat{\boldsymbol{X}}_\lambda)^{-1} \hat{\boldsymbol{X}}_\lambda^T \boldsymbol{X} - \boldsymbol{I}_{2\times 2} \right) \boldsymbol{\beta}, \end{aligned}$$

where

$$\hat{\boldsymbol{X}}_\lambda = \boldsymbol{A} \hat{\boldsymbol{\Gamma}}_\lambda = \boldsymbol{A} (\boldsymbol{A}^T \boldsymbol{A} + \lambda_N^2 \boldsymbol{D})^{-1} \boldsymbol{A}^T \boldsymbol{X}.$$

The matrices $\boldsymbol{A}$, $\boldsymbol{D}$, and $\boldsymbol{X}$ are all known for the sampling design described in Subsection 3.1.2, and $\boldsymbol{\beta}$ can be estimated by $\hat{\boldsymbol{\beta}}_{2sls}$, so the estimated 2×1 bias vector is

$$\widehat{\mathbf{bias}} = \left((\hat{\boldsymbol{X}}_\lambda^T \hat{\boldsymbol{X}}_\lambda)^{-1} \hat{\boldsymbol{X}}_\lambda^T \boldsymbol{X} - \boldsymbol{I}_{2\times 2} \right) \hat{\boldsymbol{\beta}}_{2sls}.$$

For the asymptotic variance of the estimator,

$$\frac{1}{N} (\boldsymbol{\Gamma}_0^T \boldsymbol{M}_1 \boldsymbol{\Gamma}_0)^{-1} \boldsymbol{\Gamma}_0^T \boldsymbol{M}_5 \boldsymbol{\Gamma}_0 (\boldsymbol{\Gamma}_0^T \boldsymbol{M}_1 \boldsymbol{\Gamma}_0)^{-1},$$

we estimate $\boldsymbol{\Gamma}_0$ using $\hat{\boldsymbol{\Gamma}}_\lambda$ from the sample, and we use $N^{-1}\hat{\boldsymbol{X}}_\lambda^T\hat{\boldsymbol{X}}_\lambda$ as an estimator for $\boldsymbol{\Gamma}_0^T\boldsymbol{M}_1\boldsymbol{\Gamma}_0$, as suggested by the convergence result of (65). The estimation of $\boldsymbol{M}_5$ is slightly more involved. We begin by calculating the residuals

$$\hat{\boldsymbol{\epsilon}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{2sls},$$

and we let $\boldsymbol{\psi}$ be an n_N-vector with components $\{w_i\hat{\epsilon}_i^2\}_{i\in s}$. Motivated by Assumption 3.4.6, we use

$$\widehat{\boldsymbol{M}}_5 = \frac{1}{N}\sum_{i\in U_N}\widehat{\mathrm{E}}[w_i\epsilon_i^2]\boldsymbol{b}_i\boldsymbol{b}_i^T$$

as the estimator for $\boldsymbol{M}_5$, where $\widehat{\mathrm{E}}[w_i\epsilon_i^2] = \widehat{\mathrm{E}}[w_i\epsilon_i^2|\boldsymbol{x}_i]$ is calculated by regressing $\boldsymbol{\psi}$ on $\hat{\boldsymbol{X}}_x$. We obtain the 2×1 vector

$$\boldsymbol{\gamma} = (\hat{\boldsymbol{X}}_x^T\hat{\boldsymbol{X}}_x)^{-1}\hat{\boldsymbol{X}}_x^T\boldsymbol{\psi},$$

so that the fitted values of $\widehat{\mathrm{E}}[w_i\epsilon_i^2]$ for $i\in U_N$ are given by

$$\widehat{\mathrm{E}}[\boldsymbol{\psi}] = \boldsymbol{X}_{U_N}\boldsymbol{\gamma}.$$

As a result, for finite samples the 2×2 estimated variance matrix is

$$\begin{aligned}\widehat{\boldsymbol{V}} &= \frac{1}{N}\left(\frac{1}{N}\hat{\boldsymbol{X}}_\lambda^T\hat{\boldsymbol{X}}_\lambda\right)^{-1}\hat{\boldsymbol{\Gamma}}_\lambda^T\widehat{\boldsymbol{M}}_5\hat{\boldsymbol{\Gamma}}_\lambda\left(\frac{1}{N}\hat{\boldsymbol{X}}_\lambda^T\hat{\boldsymbol{X}}_\lambda\right)^{-1}\\ &= \left(\hat{\boldsymbol{X}}_\lambda^T\hat{\boldsymbol{X}}_\lambda\right)^{-1}\hat{\boldsymbol{\Gamma}}_\lambda^T\left(\sum_{i\in U_N}\widehat{\mathrm{E}}[w_i\epsilon_i^2]\boldsymbol{b}_i\boldsymbol{b}_i^T\right)\hat{\boldsymbol{\Gamma}}_\lambda\left(\hat{\boldsymbol{X}}_\lambda^T\hat{\boldsymbol{X}}_\lambda\right)^{-1}.\end{aligned}$$

As mentioned above, $\widehat{\mathbf{bias}}$ and $\widehat{\boldsymbol{V}}$ are functions of λ_N, and the optimal λ_N may be selected as the λ_N that minimizes the estimated MSE for either $\hat{\beta}_{pspl}$ parameter,

$$\widehat{\mathrm{MSE}}(\hat{\boldsymbol{\beta}}_{0,pspl}) = (\widehat{\boldsymbol{V}})_{11} + (\widehat{\mathbf{bias}})_1^2,$$

or

$$\widehat{\mathrm{MSE}}(\hat{\boldsymbol{\beta}}_{1,pspl}) = (\widehat{\boldsymbol{V}})_{22} + (\widehat{\mathbf{bias}})_2^2.$$

3.6.1 Simulation study

Motivated by an informative sampling example from Pfeffermann and Sverchkov (1999), we select $N = 3000$ values from a Unif$(0, 1)$ distribution and treat them as a fixed, known population of x's. Also, $K = 35$ knot locations are determined by equal-spacing along the interval $(0, 1)$. Next, 5000 finite populations of y's are selected as

$$y_i = 1 + x_i + \epsilon_i; \;\; \epsilon_i \sim N(0, 1), \;\; i = 1 \ldots 3000,$$

so that $\beta_0 = \beta_1 = 1$. An informative Poisson sample is selected from each of the finite populations using inclusion probabilities given by

$$\pi_i = \frac{z_i}{2.5 \max\{z_i\}_{i \in U_N}},$$

where

$$z_i = 5 + 5y_i + 10x_i + u_i,$$

and $u_i \sim$ Unif$(0, 1)$. The estimators $\hat{\boldsymbol{\beta}}_{ols}$, $\hat{\boldsymbol{\beta}}_{2sls}$, $\hat{\boldsymbol{\beta}}_{2sls}^{K+2}$, and $\hat{\boldsymbol{\beta}}_{pspl}$ are calculated for each sample. We also calculate the semi-parametric estimator presented by Pfeffermann and Sverchkov (1999). This estimator is generally expected to be more efficient than $\hat{\boldsymbol{\beta}}_{2sls}$, and has the form

$$\hat{\boldsymbol{\beta}}_{pfsv} = (\boldsymbol{X}^T \boldsymbol{W} \tilde{\boldsymbol{W}}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{W} \tilde{\boldsymbol{W}}^{-1} \boldsymbol{y},$$

where $\tilde{\boldsymbol{W}}$ is an $n_N \times n_N$ diagonal matrix whose diagonal elements are the estimated w_i's found by regressing the sample weights on $\boldsymbol{X}$.

To compare the five estimators, we look at the ratio of the MSE of the given

Table 5: Ratio of MSE of $\hat{\boldsymbol{\beta}}_{ols}$, $\hat{\boldsymbol{\beta}}_{2sls}$, $\hat{\boldsymbol{\beta}}_{2sls}^{K+2}$, $\hat{\boldsymbol{\beta}}_{pfsv}$ to MSE of $\hat{\boldsymbol{\beta}}_{pspl}$. Numbers greater than one favor $\hat{\boldsymbol{\beta}}_{pspl}$. Based on 5000 replications of informative Poisson sampling from finite populations of size $N = 3000$.

Optimal λ_N selection		$\hat{\boldsymbol{\beta}}_{ols}$	$\hat{\boldsymbol{\beta}}_{2sls}$	$\hat{\boldsymbol{\beta}}_{2sls}^{K+2}$	$\hat{\boldsymbol{\beta}}_{pfsv}$
$\lambda_{N,\text{intercept}}$	Intercept	10.03	1.17	1.32	1.07
	Slope	2.01	1.18	1.07	1.03
$\lambda_{N,\text{slope}}$	Intercept	10.82	1.27	1.42	1.16
	Slope	2.16	1.27	1.15	1.11
$(\lambda_{N,\text{intercept}} + \lambda_{N,\text{slope}})/2$	Intercept	10.32	1.21	1.36	1.10
	Slope	2.06	1.21	1.10	1.06

parameter estimator to the MSE of the corresponding parameter estimator using the penalized splines method. Values larger than one favor $\hat{\boldsymbol{\beta}}_{pspl}$. For each sample, the optimal λ_N for $\hat{\boldsymbol{\beta}}_{pspl}$ is selected three different ways: first, using the λ_N that minimized $\widehat{\text{MSE}}(\hat{\boldsymbol{\beta}}_{0,pspl})$, denoted $\lambda_{N,\text{intercept}}$; second, using the λ_N that minimized $\widehat{\text{MSE}}(\hat{\boldsymbol{\beta}}_{1,pspl})$, denoted $\lambda_{N,\text{slope}}$; and third, using an average of $\lambda_{N,\text{intercept}}$ and $\lambda_{N,\text{slope}}$. The results are provided in Table 5.

For this informative sampling scheme, we find that $\hat{\boldsymbol{\beta}}_{pspl}$ easily outperforms $\hat{\boldsymbol{\beta}}_{ols}$, as expected. It also outperforms both of the classical two-stage least squares estimators and has a smaller advantage over $\hat{\boldsymbol{\beta}}_{pfsv}$. There is little difference in the results based on the three options for the optimal λ_N selection. The results are similar when the same simulation is conducted with different penalized spline basis functions.

CHAPTER 4

INSTRUMENTAL VARIABLE SELECTION UNDER INFORMATIVE SAMPLING

4.1 Introduction

In this chapter we explore the selection of instrumental variables in greater detail, and we examine the reason for their relative usefulness under specific informative sampling schemes. For consistency, we retain the notation of the previous chapter unless otherwise noted. The collection of IV's consisting only of the weighted covariates is denoted $\boldsymbol{A}_x$. This matrix of IV's leads to the weighted least squares estimator $\hat{\beta}_{2sls}$ that is analyzed in Fuller (2009, Ch. 6). Under informative sampling, $\hat{\beta}_{2sls}$ provides a better estimator than $\hat{\beta}_{ols}$. We also look at the set of IV's denoted $\boldsymbol{A}$ and the two-stage estimator to which it leads. This collection of IV's consists of $\boldsymbol{A}_x$ plus the additional weighted functions of the covariates, denoted $\boldsymbol{WZ}$. We are interested in comparing the estimator that uses only $\boldsymbol{A}_x$ at stage one to estimators that use additional IV's, $\boldsymbol{A}$, at stage one.

In the next section, we define a no-intercept model and the two two-stage estimators we use in our analysis. Section 4.3 provides simulation results that contrast the use of $\boldsymbol{A}_x$ and $\boldsymbol{A}$ as IV's at stage one. Finally, Section 4.4 examines influential points as a possible explanation for the advantage of $\boldsymbol{A}$ in certain informative sampling designs.

4.2 Model designation and estimators

For simplicity, in this chapter we adopt the no-intercept model

$$y_i = \beta x_i + \epsilon_i; \quad \epsilon_i \sim N(0,1), \quad i \in U_N,$$

written in vector form as $\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{\epsilon}$. We let $\beta = 1$ for our simulations. When working with $i \in s$, $\boldsymbol{W}$ is the $n_N \times n_N$ diagonal matrix of weights and $\boldsymbol{X}$ is an $n_N \times 1$ vector of x's. It will be convenient for us to maintain the use of capital letters for the vectors $\boldsymbol{X}$ and $\boldsymbol{A}_x = \boldsymbol{W}\boldsymbol{X}$.

The two-stage estimators we examine rely on the calculation of a vector of modified x's at stage one. The choice of IV's determines the form of this vector. When only the weighted covariates $\boldsymbol{A}_x$ are used as IV's, we denote the vector of modified x's as

$$\hat{\boldsymbol{X}}_x = \boldsymbol{H}_x\boldsymbol{X},$$

where $\boldsymbol{H}_x = \boldsymbol{A}_x(\boldsymbol{A}_x^T\boldsymbol{A}_x)^{-1}\boldsymbol{A}_x^T$ is the idempotent hat matrix. The estimator of β is written

$$\begin{aligned}
\hat{\beta}_{2sls} &= (\hat{\boldsymbol{X}}_x^T\hat{\boldsymbol{X}}_x)^{-1}\hat{\boldsymbol{X}}_x^T\boldsymbol{y} \\
&= (\hat{\boldsymbol{X}}_x^T\hat{\boldsymbol{X}}_x)^{-1}\hat{\boldsymbol{X}}_x^T\boldsymbol{X}\beta + (\hat{\boldsymbol{X}}_x^T\hat{\boldsymbol{X}}_x)^{-1}\hat{\boldsymbol{X}}_x^T\boldsymbol{\epsilon} \\
&= \beta + (\hat{\boldsymbol{X}}_x^T\hat{\boldsymbol{X}}_x)^{-1}\hat{\boldsymbol{X}}_x^T\boldsymbol{\epsilon}, \qquad (72)
\end{aligned}$$

since $\hat{\boldsymbol{X}}_x^T\hat{\boldsymbol{X}}_x = \boldsymbol{X}^T\boldsymbol{H}_x\boldsymbol{H}_x\boldsymbol{X} = \boldsymbol{X}^T\boldsymbol{H}_x\boldsymbol{X} = \hat{\boldsymbol{X}}_x^T\boldsymbol{X}$.

For the additional IV's beyond the weighted covariates $\boldsymbol{A}_x$, we use functions of the covariates in the form of truncated lines as defined in Chapter 3. The number of additional IV's is indicated by the number of knots K, and $\boldsymbol{Z}$ represents the $n_N \times K$ matrix of these truncated lines. We have

$$\boldsymbol{A} = \boldsymbol{W}\begin{bmatrix}\boldsymbol{X} & \boldsymbol{Z}\end{bmatrix} = \begin{bmatrix}\boldsymbol{A}_x & \boldsymbol{W}\boldsymbol{Z}\end{bmatrix},$$

and we denote the vector of modified x's as

$$\hat{\boldsymbol{X}} = \boldsymbol{H}\boldsymbol{X},$$

with no subscript, for the hat matrix $\boldsymbol{H} = \boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T$. This estimator of β is written

$$\begin{aligned}\hat{\beta}_{add} &= (\hat{\boldsymbol{X}}^T\hat{\boldsymbol{X}})^{-1}\hat{\boldsymbol{X}}^T\boldsymbol{y} \\ &= (\hat{\boldsymbol{X}}^T\hat{\boldsymbol{X}})^{-1}\hat{\boldsymbol{X}}^T\boldsymbol{X}\beta + (\hat{\boldsymbol{X}}^T\hat{\boldsymbol{X}})^{-1}\hat{\boldsymbol{X}}^T\boldsymbol{\epsilon} \\ &= \beta + (\hat{\boldsymbol{X}}^T\hat{\boldsymbol{X}})^{-1}\hat{\boldsymbol{X}}^T\boldsymbol{\epsilon}, \end{aligned} \tag{73}$$

since $\hat{\boldsymbol{X}}^T\hat{\boldsymbol{X}} = \boldsymbol{X}^T\boldsymbol{H}\boldsymbol{H}\boldsymbol{X} = \boldsymbol{X}^T\boldsymbol{H}\boldsymbol{X} = \hat{\boldsymbol{X}}^T\boldsymbol{X}$.

We use the estimator $\hat{\beta}_{add}$ rather than the penalized spline estimator because the smoother (hat) matrix $\boldsymbol{S}_\lambda = \boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A} + \lambda_N^2\boldsymbol{D})^{-1}\boldsymbol{A}^T$ for $\hat{\beta}_{pspl}$ is not idempotent and this leads to an additional complicating term in our expression for the estimator of β. Not using the penalized spline estimator is of little concern in this chapter, because we are primarily interested in exploring the reason for the potential advantage of additional IV's beyond $\boldsymbol{A}_x$. We are able to adjust the number of additional IV's used for the $\hat{\beta}_{add}$ estimator, which is similar to the way we vary λ_N to control the approximate df, or roughness, of the fit for the penalized spline estimator.

4.3 Two informative sampling simulations

Based on the work of ten Cate (1986), we are motivated to examine the estimation of regression parameters for an informative sampling scheme with fixed inclusion probabilities assigned to predefined strata. Stratum inclusion is determined by the y-

values of the data points. The three-strata example of ten Cate (1986) assigns smaller inclusion probabilities to the two extreme strata of y-values and a larger inclusion probability to the middle stratum. The author also explains why an estimator equivalent to $\hat{\beta}_{2sls}$ has advantages over $\hat{\beta}_{ols}$ under this sampling design. We modify his example and use only two strata, with the strata boundary defined so that smaller y-values are in the stratum with the smaller inclusion probability. We look at two separate simulations whose only difference is the location of the boundary. The first simulation, referred to as *high-low*, places the boundary at $y = 0$ and the inclusion probabilities are defined as

$$\pi_i = 0.01 + (0.09)I_{\{y_i>0\}}, \quad i \in U_N.$$

The second simulation, referred to as *extreme*, places the boundary at $y = -1$ and the inclusion probabilities are defined as

$$\pi_i = 0.01 + (0.09)I_{\{y_i>-1\}}, \quad i \in U_N.$$

Our goals are to verify that $\hat{\beta}_{2sls}$ is better than $\hat{\beta}_{ols}$ in these two-strata sampling designs, and to determine if $\hat{\beta}_{add}$ provides additional improvement.

For both simulations we select $N = 10,000$ values from a Unif(0,1) distribution and treat these as a fixed, known population of x's. Next, 1000 finite populations of y's are selected according to the no-intercept model of the previous section. An informative Poisson sample is selected from each of the finite populations and the estimators $\hat{\beta}_{ols}$, $\hat{\beta}_{2sls}$, and $\hat{\beta}_{add}$ for $K = 1$, 2, and 35 are calculated for each sample. Knot placement is based on equal spacing along the interval (0,1). We also calculate the estimator $\hat{\beta}_{pfsv}$ presented in Pfeffermann and Sverchkov (1999) and discussed in the previous chapter. To compare the six estimators, we calculate the simulation MSE of each and then divide these values by the simulation MSE of $\hat{\beta}_{2sls}$. Ratios

Table 6: Ratio of MSE of $\hat{\beta}_{ols}$, $\hat{\beta}_{2sls}$, $\hat{\beta}_{add}$ (for $K = 1$, 2, and 35), and $\hat{\beta}_{pfsv}$ to MSE of $\hat{\beta}_{2sls}$. Numbers greater than one favor $\hat{\beta}_{2sls}$. Based on 1000 replications of informative Poisson sampling from finite populations of size $N = 10,000$.

Simulation	$\hat{\beta}_{ols}$	$\hat{\beta}_{2sls}$	$\hat{\beta}_{add,K=1}$	$\hat{\beta}_{add,K=2}$	$\hat{\beta}_{add,K=35}$	$\hat{\beta}_{pfsv}$
high-low	18.23	1.00	1.04	1.11	6.77	1.41
extreme	2.72	1.00	0.94	0.92	1.61	1.68

larger than one favor $\hat{\beta}_{2sls}$. The results for both examples are provided in Table 6.

We notice in both simulations that all two-stage estimators (i.e., those using IV's) perform better than $\hat{\beta}_{ols}$. We also notice that in *high-low*, $\hat{\beta}_{2sls}$ outperforms all other estimators, but in *extreme*, $\hat{\beta}_{2sls}$ is outperformed by two of the $\hat{\beta}_{add}$ estimators. In fact, the *extreme* simulation verifies the results of Chapter 3, and we could use those concepts to find an optimal λ for the $\hat{\beta}_{pspl}$ estimator if desired. The $\hat{\beta}_{pfsv}$ estimator does not perform well in these two simulations when compared to the two-stage estimators, with the exception of $\hat{\beta}_{add}$ for $K = 35$ in *high-low*.

In this chapter, we are mainly interested in why $\hat{\beta}_{2sls}$ is better than all $\hat{\beta}_{add}$ estimators in the *high-low* simulation and not in the *extreme* simulation. By examining graphs (not shown) of individual realizations from each simulation, and by conducting additional unrecorded simulations with other strata boundaries, we are led to believe that the difference in results between *high-low* and *extreme* for the two-stage estimators depends in some (possibly complex) way on the number of large-weight data points in the individual samples. We explore this further in the next section.

4.4 Influential points

We open this section with a third informative sampling simulation that will more clearly demonstrate the advantage of the additional IV's used in finding $\hat{\beta}_{add}$. We call this simulation *advantage*. The basic design is the same as for the two examples

Table 7: Bias, variance, MSE, and ratio of MSE of $\hat{\beta}_{ols}$, $\hat{\beta}_{2sls}$, $\hat{\beta}_{add}$ (for $K = 1$, 2, and 35), and $\hat{\beta}_{pfsv}$ to MSE of $\hat{\beta}_{2sls}$. Ratio numbers greater than one favor $\hat{\beta}_{2sls}$. Based on 1000 replications of informative Poisson sampling from finite populations of size $N = 10,000$ for *advantage* simulation.

	$\hat{\beta}_{ols}$	$\hat{\beta}_{2sls}$	$\hat{\beta}_{add,K=1}$	$\hat{\beta}_{add,K=2}$	$\hat{\beta}_{add,K=35}$	$\hat{\beta}_{pfsv}$
Bias	0.3956	0.0015	0.0041	0.0065	0.0555	0.0009
Variance	0.0060	0.0087	0.0085	0.0084	0.0076	0.0144
MSE	0.1626	0.0087	0.0085	0.0084	0.0107	0.0144
MSE Ratio	18.71	1.00	0.98	0.97	1.26	1.65

of the previous section and the only difference is the inclusion probability which is now given by

$$\pi_i = \frac{y_i + 3}{80}, \quad i \in U_N.$$

This definition for the inclusion probabilities is similar to the earlier examples in the sense that smaller y-values have smaller inclusion probabilities and larger weights. The results of *advantage*, including the bias and variance components of the MSE, are provided in Table 7.

One result from this simulation that is similar to results found in *high-low* and *extreme* is the indication that based on MSE ratios, $\hat{\beta}_{pfsv}$ does not perform well compared to the two-stage estimators. For this reason, and because we are primarily interested in studying the use of IV's for the two-stage estimators, we omit the $\hat{\beta}_{pfsv}$ estimator from future discussion in this chapter.

The MSE ratios for *advantage* also demonstrate that $\hat{\beta}_{add}$ outperforms $\hat{\beta}_{2sls}$ for some values of K. As we increase K, our fitted x-values approach the original x values and the bias increases, indicating that the decreasing variance component is the reason for the MSE improvement. We would like to look at closed-form expressions for the variances of these two simple estimators in order to explain why the variance for $\hat{\beta}_{add}$ is smaller than the variance for $\hat{\beta}_{2sls}$ in some informative

sampling cases and not in others. We begin by writing $\hat{\beta}_{add}$ as

$$\hat{\beta}_{add} = \hat{\beta}_{2sls} + (\text{additional terms}).$$

As we now demonstrate, the use of two stages unfortunately causes great complexity in the expressions we seek. Because $\hat{\boldsymbol{X}}_x$, $\hat{\boldsymbol{X}}$, and $\boldsymbol{y}$ are all vectors, we can write (72) and (73) as

$$\hat{\beta}_{2sls} = \beta + \frac{\hat{\boldsymbol{X}}_x^T \boldsymbol{\epsilon}}{\hat{\boldsymbol{X}}_x^T \hat{\boldsymbol{X}}_x},$$

and

$$\hat{\beta}_{add} = \beta + \frac{\hat{\boldsymbol{X}}^T \boldsymbol{\epsilon}}{\hat{\boldsymbol{X}}^T \hat{\boldsymbol{X}}},$$

where all numerators and denominators of the fractions are scalars. We use hat matrix notation to derive an expression for $\hat{\beta}_{2sls}$ in terms of sums over the sample. We have

$$\begin{aligned}
\boldsymbol{H}_x &= \boldsymbol{A}_x (\boldsymbol{A}_x^T \boldsymbol{A}_x)^{-1} \boldsymbol{A}_x^T \\
&= \boldsymbol{W}\boldsymbol{X} (\boldsymbol{X}^T \boldsymbol{W}^2 \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{W} \\
&= \boldsymbol{W}\boldsymbol{X} \left(\sum_{i \in s} w_i^2 x_i^2 \right)^{-1} \boldsymbol{X}^T \boldsymbol{W},
\end{aligned}$$

so that

$$\begin{aligned}
\hat{\boldsymbol{X}}_x &= \boldsymbol{W}\boldsymbol{X} \left(\sum_{i \in s} w_i^2 x_i^2 \right)^{-1} \boldsymbol{X}^T \boldsymbol{W}\boldsymbol{X} \\
&= \boldsymbol{W}\boldsymbol{X} \left(\sum_{i \in s} w_i^2 x_i^2 \right)^{-1} \left(\sum_{i \in s} w_i x_i^2 \right) \\
&= c_1 \boldsymbol{W}\boldsymbol{X},
\end{aligned}$$

where c_1 is the scalar

$$c_1 = \frac{\sum_{i \in s} w_i x_i^2}{\sum_{i \in s} w_i^2 x_i^2}.$$

We have

$$\hat{\boldsymbol{X}}_x^T \boldsymbol{\epsilon} = c_1 \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{\epsilon} = c_1 \left(\sum_{i \in s} w_i x_i \epsilon_i \right),$$

and

$$\hat{\boldsymbol{X}}_x^T \hat{\boldsymbol{X}}_x = c_1^2 \boldsymbol{X}^T \boldsymbol{W}^2 \boldsymbol{X} = c_1^2 \left(\sum_{i \in s} w_i^2 x_i^2 \right),$$

so that by (72)

$$\begin{aligned} \hat{\beta}_{2sls} &= \beta + (\hat{\boldsymbol{X}}_x^T \hat{\boldsymbol{X}}_x)^{-1} \hat{\boldsymbol{X}}_x^T \boldsymbol{\epsilon} \\ &= \beta + \frac{c_1 \sum_{i \in s} w_i x_i \epsilon_i}{c_1^2 \sum_{i \in s} w_i^2 x_i^2} \\ &= \beta + \frac{\sum_{i \in s} w_i x_i \epsilon_i}{\sum_{i \in s} w_i x_i^2}. \end{aligned}$$

We follow a similar process to write $\hat{\beta}_{add}$ in a form that contains $\hat{\beta}_{2sls}$. We start with the hat matrix

$$\begin{aligned} \boldsymbol{H} &= \boldsymbol{A}(\boldsymbol{A}^T \boldsymbol{A})^{-1} \boldsymbol{A}^T \\ &= \begin{bmatrix} \boldsymbol{A}_x & \boldsymbol{W}\boldsymbol{Z} \end{bmatrix} \left(\begin{bmatrix} \boldsymbol{A}_x^T \\ \boldsymbol{Z}^T \boldsymbol{W} \end{bmatrix} \begin{bmatrix} \boldsymbol{A}_x & \boldsymbol{W}\boldsymbol{Z} \end{bmatrix} \right)^{-1} \begin{bmatrix} \boldsymbol{A}_x^T \\ \boldsymbol{Z}^T \boldsymbol{W} \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{A}_x & \boldsymbol{W}\boldsymbol{Z} \end{bmatrix} \begin{bmatrix} \boldsymbol{A}_x^T \boldsymbol{A}_x & \boldsymbol{A}_x^T \boldsymbol{W}\boldsymbol{Z} \\ \boldsymbol{Z}^T \boldsymbol{W} \boldsymbol{A}_x & \boldsymbol{Z}^T \boldsymbol{W}^2 \boldsymbol{Z} \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{A}_x^T \\ \boldsymbol{Z}^T \boldsymbol{W} \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{A}_x & \boldsymbol{W}\boldsymbol{Z} \end{bmatrix} \begin{bmatrix} \boldsymbol{E}_{11} & \boldsymbol{E}_{12} \\ \boldsymbol{E}_{21} & \boldsymbol{E}_{22} \end{bmatrix} \begin{bmatrix} \boldsymbol{A}_x^T \\ \boldsymbol{Z}^T \boldsymbol{W} \end{bmatrix} \\ &= \boldsymbol{A}_x \boldsymbol{E}_{11} \boldsymbol{A}_x^T + \boldsymbol{A}_x \boldsymbol{E}_{12} \boldsymbol{Z}^T \boldsymbol{W} + \boldsymbol{W}\boldsymbol{Z} \boldsymbol{E}_{21} \boldsymbol{A}_x^T + \boldsymbol{W}\boldsymbol{Z} \boldsymbol{E}_{21} \boldsymbol{Z}^T \boldsymbol{W}, \end{aligned}$$

where

$$\begin{aligned}
\boldsymbol{E}_{11} &= (\boldsymbol{A}_x^T\boldsymbol{A}_x)^{-1} + (\boldsymbol{A}_x^T\boldsymbol{A}_x)^{-1}\boldsymbol{A}_x^T\boldsymbol{W}\boldsymbol{Z}\boldsymbol{E}_{22}\boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{A}_x(\boldsymbol{A}_x^T\boldsymbol{A}_x)^{-1},\\
\boldsymbol{E}_{12} &= -(\boldsymbol{A}_x^T\boldsymbol{A}_x)^{-1}\boldsymbol{A}_x^T\boldsymbol{W}\boldsymbol{Z}\boldsymbol{E}_{22},\\
\boldsymbol{E}_{21} &= -\boldsymbol{E}_{22}\boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{A}_x(\boldsymbol{A}_x^T\boldsymbol{A}_x)^{-1}, \text{ and}\\
\boldsymbol{E}_{22} &= \left(\boldsymbol{Z}^T\boldsymbol{W}^2\boldsymbol{Z} - \boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{A}_x(\boldsymbol{A}_x^T\boldsymbol{A}_x)^{-1}\boldsymbol{A}_x^T\boldsymbol{W}\boldsymbol{Z}\right)^{-1}\\
&= \left(\boldsymbol{Z}^T\boldsymbol{W}^2\boldsymbol{Z} - \boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{H}_x\boldsymbol{W}\boldsymbol{Z}\right)^{-1},
\end{aligned}$$

so that

$$\begin{aligned}
\boldsymbol{H} &= \boldsymbol{H}_x + \boldsymbol{H}_x\boldsymbol{W}\boldsymbol{Z}\boldsymbol{E}_{22}\boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{H}_x\\
&\quad - \boldsymbol{H}_x\boldsymbol{W}\boldsymbol{Z}\boldsymbol{E}_{22}\boldsymbol{Z}^T\boldsymbol{W} - \boldsymbol{W}\boldsymbol{Z}\boldsymbol{E}_{22}\boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{H}_x + \boldsymbol{W}\boldsymbol{Z}\boldsymbol{E}_{22}\boldsymbol{Z}^T\boldsymbol{W}\\
&= \boldsymbol{H}_x + \boldsymbol{H}_x\boldsymbol{W}\boldsymbol{Z}\boldsymbol{E}_{22}\boldsymbol{Z}^T\boldsymbol{W}\left(\boldsymbol{H}_x - \boldsymbol{I}\right) + \boldsymbol{W}\boldsymbol{Z}\boldsymbol{E}_{22}\boldsymbol{Z}^T\boldsymbol{W}\left(\boldsymbol{I} - \boldsymbol{H}_x\right)\\
&= \boldsymbol{H}_x - \boldsymbol{H}_x\boldsymbol{W}\boldsymbol{Z}\boldsymbol{E}_{22}\boldsymbol{Z}^T\boldsymbol{W}\left(\boldsymbol{I} - \boldsymbol{H}_x\right) + \boldsymbol{W}\boldsymbol{Z}\boldsymbol{E}_{22}\boldsymbol{Z}^T\boldsymbol{W}\left(\boldsymbol{I} - \boldsymbol{H}_x\right)\\
&= \boldsymbol{H}_x + \left(\boldsymbol{I} - \boldsymbol{H}_x\right)\boldsymbol{W}\boldsymbol{Z}\boldsymbol{E}_{22}\boldsymbol{Z}^T\boldsymbol{W}\left(\boldsymbol{I} - \boldsymbol{H}_x\right).
\end{aligned}$$

From this we have

$$\begin{aligned}
\hat{\boldsymbol{X}} &= \boldsymbol{H}\boldsymbol{X}\\
&= \boldsymbol{H}_x\boldsymbol{X} + \left(\boldsymbol{I} - \boldsymbol{H}_x\right)\boldsymbol{W}\boldsymbol{Z}\boldsymbol{E}_{22}\boldsymbol{Z}^T\boldsymbol{W}\left(\boldsymbol{X} - \boldsymbol{H}_x\boldsymbol{X}\right)\\
&= \hat{\boldsymbol{X}}_x + \left(\boldsymbol{I} - \boldsymbol{H}_x\right)\boldsymbol{W}\boldsymbol{Z}\boldsymbol{E}_{22}\boldsymbol{Z}^T\boldsymbol{W}\left(\boldsymbol{X} - \hat{\boldsymbol{X}}_x\right)\\
&= \hat{\boldsymbol{X}}_x + \boldsymbol{g},
\end{aligned}$$

where $\boldsymbol{g}$ represents the second term of the expression that depends on $\boldsymbol{W}$, $\boldsymbol{X}$, and $\boldsymbol{Z}$. Substituting into (73), we have

$$\hat{\beta}_{add} = \beta + (\hat{\boldsymbol{X}}^T\hat{\boldsymbol{X}})^{-1}\hat{\boldsymbol{X}}^T\boldsymbol{\epsilon}$$

$$
\begin{aligned}
&= \beta + \frac{(\hat{\boldsymbol{X}}_x + \boldsymbol{g})^T \boldsymbol{\epsilon}}{(\hat{\boldsymbol{X}}_x + \boldsymbol{g})^T (\hat{\boldsymbol{X}}_x + \boldsymbol{g})} \\
&= \beta + \frac{\hat{\boldsymbol{X}}_x^T \boldsymbol{\epsilon} + \boldsymbol{g}^T \boldsymbol{\epsilon}}{\hat{\boldsymbol{X}}_x^T \hat{\boldsymbol{X}}_x + \hat{\boldsymbol{X}}_x^T \boldsymbol{g} + \boldsymbol{g}^T \hat{\boldsymbol{X}}_x + \boldsymbol{g}^T \boldsymbol{g}} \\
&= \beta + \frac{\hat{\boldsymbol{X}}_x^T \boldsymbol{\epsilon} + c_2}{\hat{\boldsymbol{X}}_x^T \hat{\boldsymbol{X}}_x + c_3} \\
&= \beta + \left(\hat{\boldsymbol{X}}_x^T \boldsymbol{\epsilon} + c_2 \right) \left(\frac{1}{\hat{\boldsymbol{X}}_x^T \hat{\boldsymbol{X}}_x} - \frac{c_3}{\hat{\boldsymbol{X}}_x^T \hat{\boldsymbol{X}}_x \left(\hat{\boldsymbol{X}}_x^T \hat{\boldsymbol{X}}_x + c_3 \right)} \right) \\
&= \beta + \frac{\hat{\boldsymbol{X}}_x^T \boldsymbol{\epsilon}}{\hat{\boldsymbol{X}}_x^T \hat{\boldsymbol{X}}_x} + c_4 \\
&= \hat{\beta}_{2sls} + c_4,
\end{aligned}
$$

where $c_2 = \boldsymbol{g}^T \boldsymbol{y}$, $c_3 = \hat{\boldsymbol{X}}_x^T \boldsymbol{g} + \boldsymbol{g}^T \hat{\boldsymbol{X}}_x + \boldsymbol{g}^T \boldsymbol{g}$, and c_4 is a complicated function of $\boldsymbol{W}$, $\boldsymbol{X}$, $\boldsymbol{Z}$, and $\boldsymbol{y}$. Although we now have expressions relating $\boldsymbol{H}$ and $\boldsymbol{H}_x$, $\hat{\boldsymbol{X}}$ and $\hat{\boldsymbol{X}}_x$, and $\hat{\beta}_{add}$ and $\hat{\beta}_{2sls}$, when we attempt to write these results in terms of sums over the sample, even for the case of one additional IV beyond $\boldsymbol{A}_x$, the expressions are very complex.

Thus, to answer our question about the reduction in variance for the $\hat{\beta}_{add}$ estimator, we take a different approach. We begin by looking at the four graphs of Figure 6 that result from one realization of the *advantage* simulation. In this and similar figures for the remainder of the chapter, we provide graphs of $\hat{\beta}_{ols}$, $\hat{\beta}_{2sls}$, and $\hat{\beta}_{add}$ for $K = 1$ and 2. In each graph of Figure 6, the solid line is $y = x$ and the dashed line is the estimated line for the given $\hat{\beta}$. The plotted points are y vs. x for $\hat{\beta}_{ols}$ and y vs. fitted x for each two-stage estimator. The three two-stage estimators are better predictors than $\hat{\beta}_{ols}$ in this single realization. When we look at similar graphs from several different simulated individual realizations we find similar results most of the time. However, we occasionally find a result that looks like Figure 7. In

Table 8: Bias of $\hat{\beta}_{ols}$, $\hat{\beta}_{2sls}$, and $\hat{\beta}_{add}$ (for $K = 1$, 2, and 35). *Slider* points located at $y = -2.9$ and $x = 0$, 0.2, 0.4, 0.6, 0.8, and 1.0. Based on 1000 replications of informative Poisson sampling from finite populations of size $N = 10,000$ for modified *advantage* simulation.

Slider location	$\hat{\beta}_{ols}$	$\hat{\beta}_{2sls}$	$\hat{\beta}_{add,K=1}$	$\hat{\beta}_{add,K=2}$	$\hat{\beta}_{add,K=35}$
no *slider*	0.40	0.00	0.00	0.01	0.06
$x = 0$	0.40	0.00	0.00	0.01	0.06
$x = 0.2$	0.39	-0.15	-0.06	-0.02	0.05
$x = 0.4$	0.39	-0.30	-0.06	-0.04	0.04
$x = 0.6$	0.38	-0.46	-0.16	-0.07	0.03
$x = 0.8$	0.38	-0.61	-0.42	-0.36	0.03
$x = 1.0$	0.37	-0.75	-0.13	-0.08	0.02

this single realization, $\hat{\beta}_{2sls}$ is not as accurate as in Figure 6. In the $\hat{\beta}_{2sls}$ graph of Figure 7, we also notice the presence of a single point separate from the others in the far lower-right corner. We refer to this circled point as a *slider* and we see that it corresponds to a small y-value in the original sample of (x, y) pairs, and thus has a small inclusion probability and a large weight. Because the two $\hat{\beta}_{add}$ estimators seem less effected by the *slider* than $\hat{\beta}_{2sls}$ in Figure 7, we suspect that the occasional selection of a *slider*-like point in the samples of the *advantage* simulation causes the increase in variation of $\hat{\beta}_{2sls}$.

For our first informal test of this assumption we conduct four separate simulations similar to the *advantage* simulation, except that in each realization of each modified simulation we add one additional pre-selected *slider* point to the sample. This additional "sample point" is always at $y = -2.9$ so that its "inclusion probability" is $\pi = 0.1/80$. We allow the x-value to change from one simulation to the next, but the x-value remains fixed for all finite sample realizations during a given simulation. In these simulations we are only interested in the bias of each estimator. The results from the simulations at x-values of 0, 0.2, 0.4, 0.6, 0.8, and 1.0 are provided in Table 8.

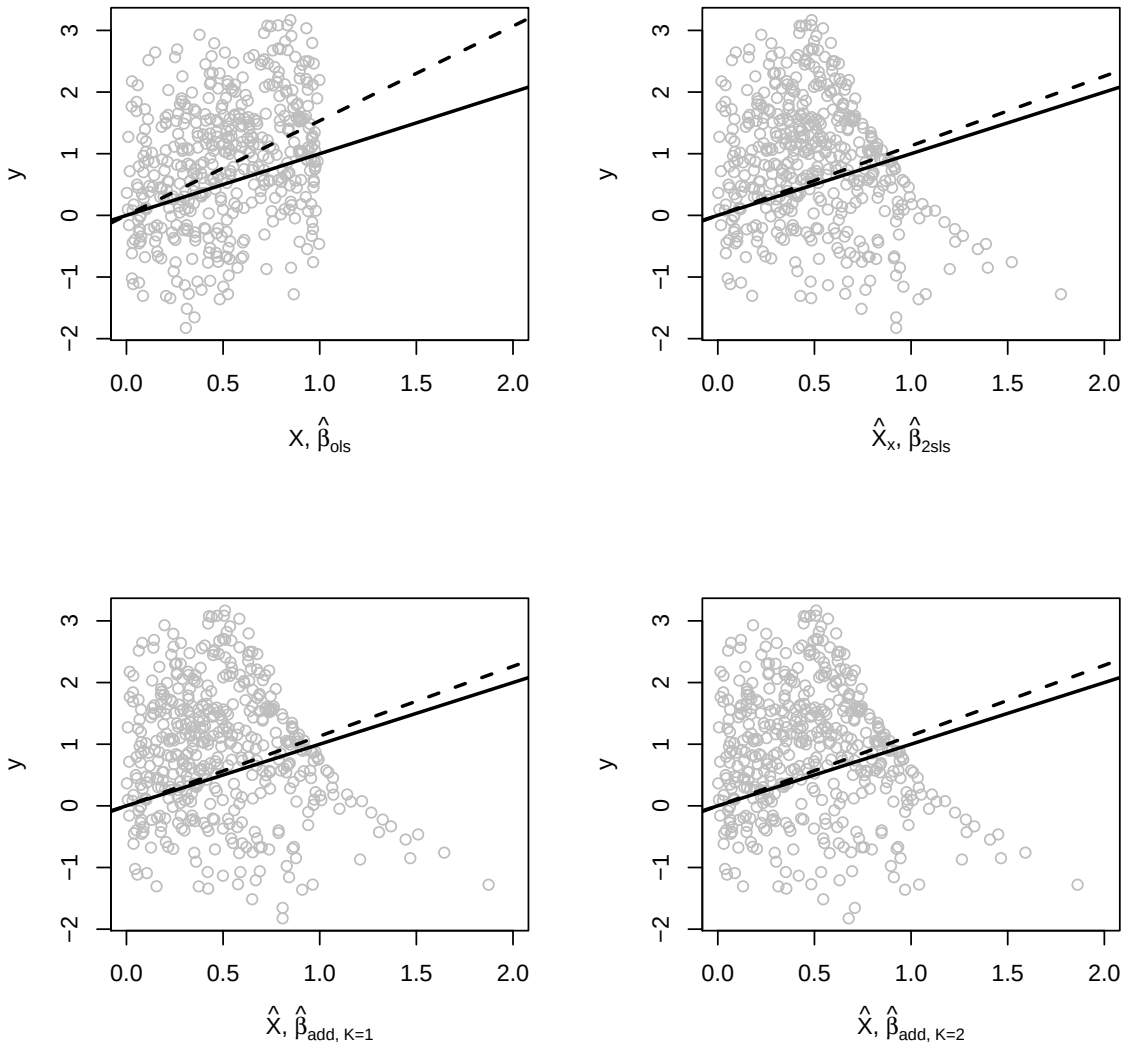

Figure 6: Graphs from one realization of the *advantage* simulation. Plotted points ($n_N = 459$) are y vs. x for $\hat{\beta}_{ols}$ and y vs. fitted x for each two-stage estimator. The solid line is $y = x$ (i.e., $\beta = 1$) and the dashed line is the estimated line for the given $\hat{\beta}$.

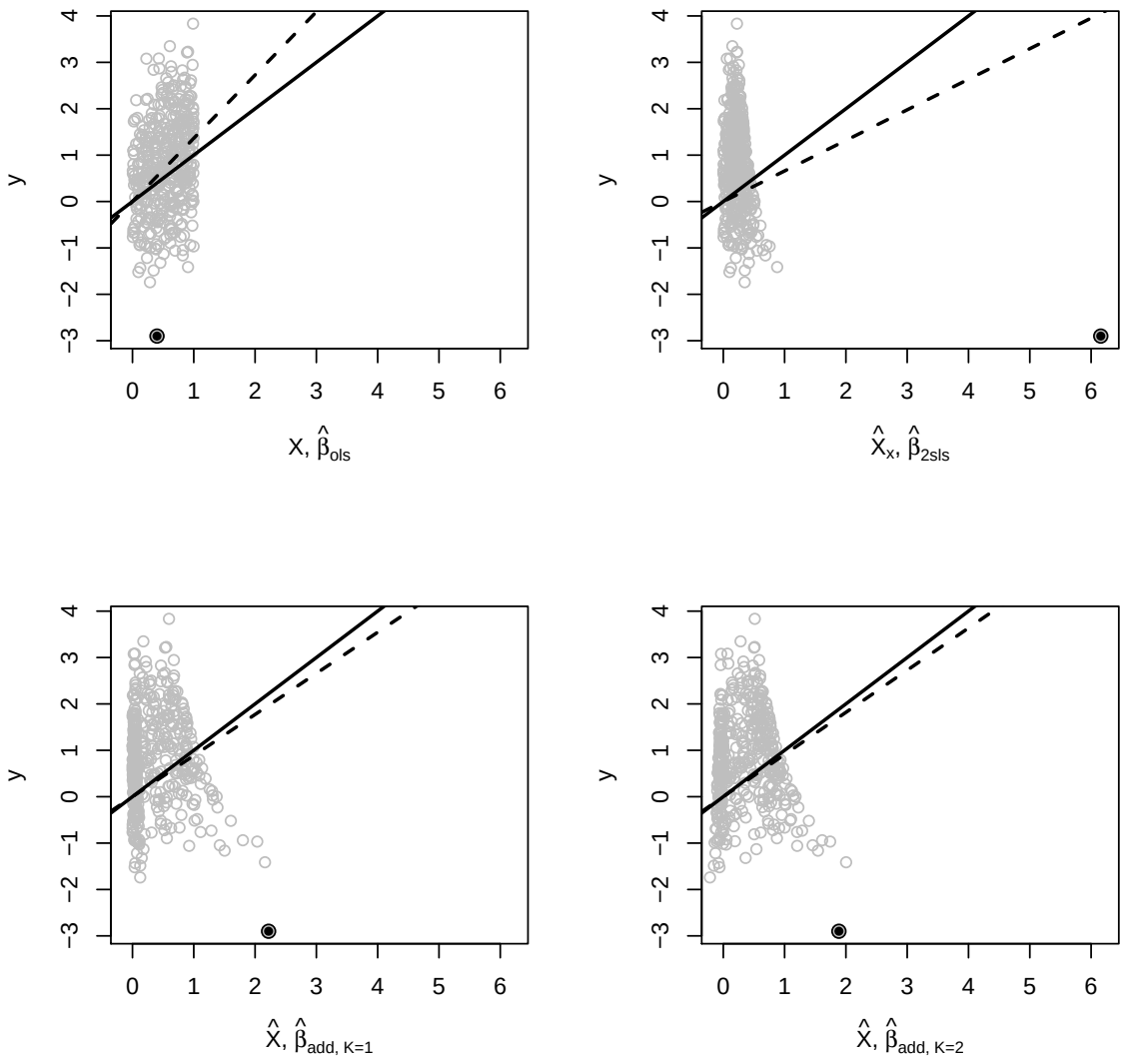

Figure 7: Graphs from one realization of the modified *advantage* simulation ($n_N = 425$). The circled point, the *slider*, has a small inclusion probability and a large weight. The *slider* has the greatest influence on $\hat{\beta}_{2sls}$.

Using the "no *slider*" values for comparison, we find results that verify what we saw in the single example of Figure 7. The presence of a *slider* has the most obvious effect on $\hat{\beta}_{2sls}$. The tabled values imply that in our original *advantage* simulation whenever a truly random *slider* is selected into the sample, it is likely to have the greatest effect on $\hat{\beta}_{2sls}$, a much smaller effect on the $\hat{\beta}_{add}$ estimators for $K = 1$ and 2, and very little effect on $\hat{\beta}_{ols}$ and $\hat{\beta}_{add}$ for $K = 35$. In the *advantage* simulation, we might expect to see an occasional *slider*-like point near the smaller values of x. At $x = 0$, the *slider* has little or no effect on any of the estimators, and at $x = 0.2$ or .04, the bias when even one additional IV is used for $\hat{\beta}_{add}$ (i.e., $K = 1$) is two times less and five times less, respectively, than the bias for $\hat{\beta}_{2sls}$. We do not expect to see *sliders* near the larger x values in the *advantage* simulation, but the bias results for $x = 0.6$, 0.8, or 1.0, while not as convincing as the results for the smaller x-values, still show an advantage for $\hat{\beta}_{add}$ over $\hat{\beta}_{2sls}$.

As a precaution against the possible effects of the varying column spaces of $\boldsymbol{A}$ for the various K values, we also conduct the simulations using principal components regression at stage one. To accomplish this, we set $K = 35$ and calculate the singular value decomposition of the matrix

$$\begin{bmatrix} \boldsymbol{A}_x & (\boldsymbol{I} - \boldsymbol{H}_x)\boldsymbol{W}\boldsymbol{Z} \end{bmatrix}$$

to obtain the matrix $\boldsymbol{U}$ with 36 orthonormal columns. We then select the appropriate number of columns from $\boldsymbol{U}$ to match the number of IV's used in the original simulations. The numerical results are similar to those of the original simulations.

Two additional figures use Cook's distance D_i (see Cook and Weisberg 1982, Ch. 3) as a measure of influence of individual points to explain the difference between biases for $\hat{\beta}_{2sls}$ and $\hat{\beta}_{add}$. The first, Figure 8, shows $ln(D_i)$ for each point in a single realization of the modified *advantage* simulation with the *slider* located at $x = 0.4$. The criteria for designating an observation as influential based on Cook's distance

depends on the problem, but Cook and Weisberg (1982, Ch. 3) suggest $D_i > 1$ (i.e., $ln(D_i) > 0$) as a general rule. In this realization the D_i for the *slider* is very large for $\hat{\beta}_{2sls}$, meaning that it has a large influence on the slope of the estimated regression line. The value of D_i for the *slider* is much reduced, and less than one, for both $\hat{\beta}_{add}$ estimators. To see why D_i for the *slider* is so large only for $\hat{\beta}_{2sls}$, we look at Figure 9. Here we have plots of x vs. weighted x in each graph, and we have added the values of fitted x vs. weighted x using "+" symbols, where the different fitted x's depend on the number of IV's used at stage one. In the $\hat{\beta}_{2sls}$ graph, the inflexibility of the $\boldsymbol{A}_x$ IV's allows the $\hat{x}_x$-value for the *slider* to be very large and thus very influential at the second stage. When additional IV's are added, as in both $\hat{\beta}_{add}$ cases, the flexibility of the fit at the first stage keeps the $\hat{x}$-value for the *slider* from becoming dramatically larger than the other $\hat{x}$-values, and thus it does not have the strong influence at the second stage that we saw for the $\hat{\beta}_{2sls}$ case.

This is a likely explanation for the results of the *advantage* simulation, because for most of the trials $\hat{\beta}_{2sls}$ and both $\hat{\beta}_{add}$ estimators perform similarly, but when the occasional extreme, large-weight data point happens to be included in the sample, $\hat{\beta}_{2sls}$ is effected much more than the $\hat{\beta}_{add}$ estimators. These occasional "bad" estimates lead to an increased variance for $\hat{\beta}_{2sls}$, while not dramatically affecting its bias. Another way to look at this is to run the original *advantage* simulation and record the maximum Cook's distance value for each estimator in each realization of the simulation. Figure 10 shows these maximum D_i values graphed vs. the estimated β's for the corresponding realizations. We see that a handful of realizations for the $\hat{\beta}_{2sls}$ estimator have large Cook's distance values and corresponding low estimates for β, while this is not true of the $\hat{\beta}_{add}$ estimators.

In summary, the empirical evidence of this chapter shows that additional IV's, beyond the weighted covariates, sometimes help at stage one in reducing the magnitude of fitted x-values for large-weight data points. This in turn reduces the

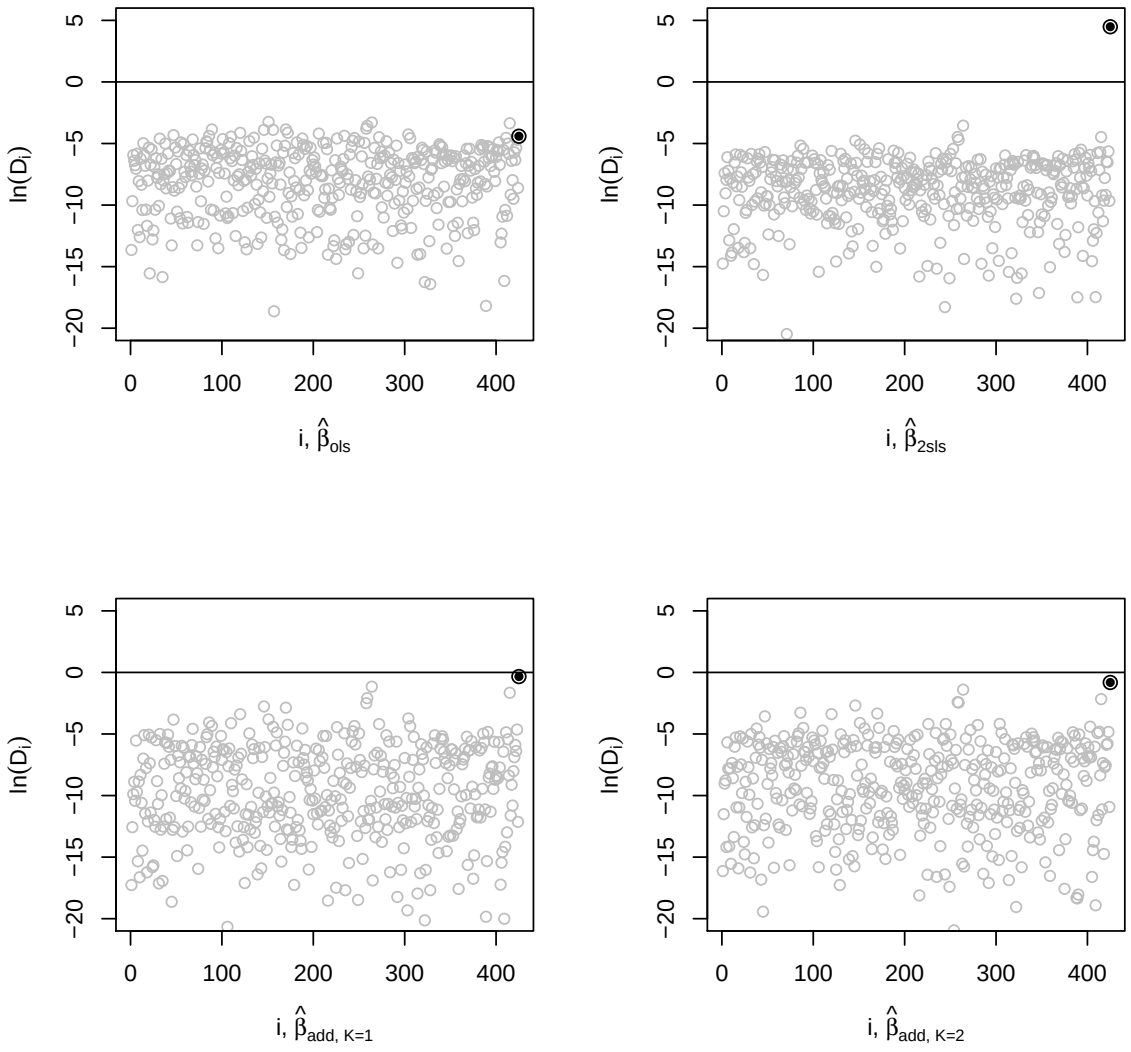

Figure 8: Graphs from one realization of the modified *advantage* simulation (*slider* at $x = 0.4$), with the natural log of Cook's distance $ln(D_i)$ plotted vs. the index for all $n_N = 425$ x's for $\hat{\beta}_{ols}$ or vs. the index for all fitted x's for the two-stage estimators. The horizontal line at $ln(D_i) = 0$ represents the threshold for influential points. In the $\hat{\beta}_{2sls}$ graph the *slider* has an actual $D_i = 88.6$, well above the threshold of $D_i = 1$. In the other three graphs, all D_i values, including the D_i for the *slider*, are below the threshold.

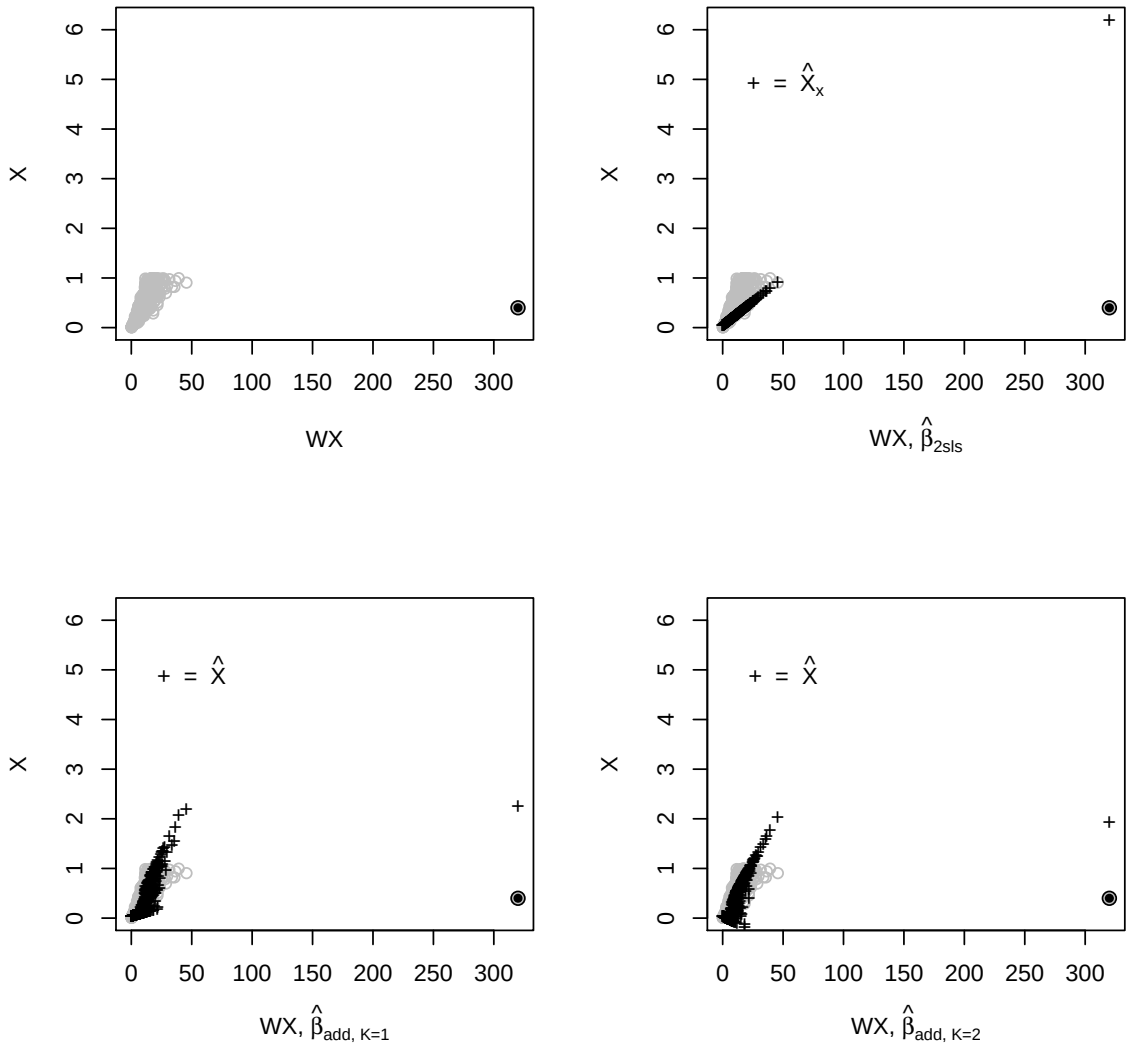

Figure 9: Graphs from one realization of the modified *advantage* simulation (*slider* at $x = 0.4$), with x plotted vs. weighted x ($n_N = 425$). The *slider* has a weighted x equal to 320. For the three two-stage estimators, values of fitted x vs. weighted x are also plotted using plus signs (+). The inflexibility at stage one of $\hat{\beta}_{2sls}$ allows for a large fitted x-value for the *slider*, leading to a large D_i at stage two.

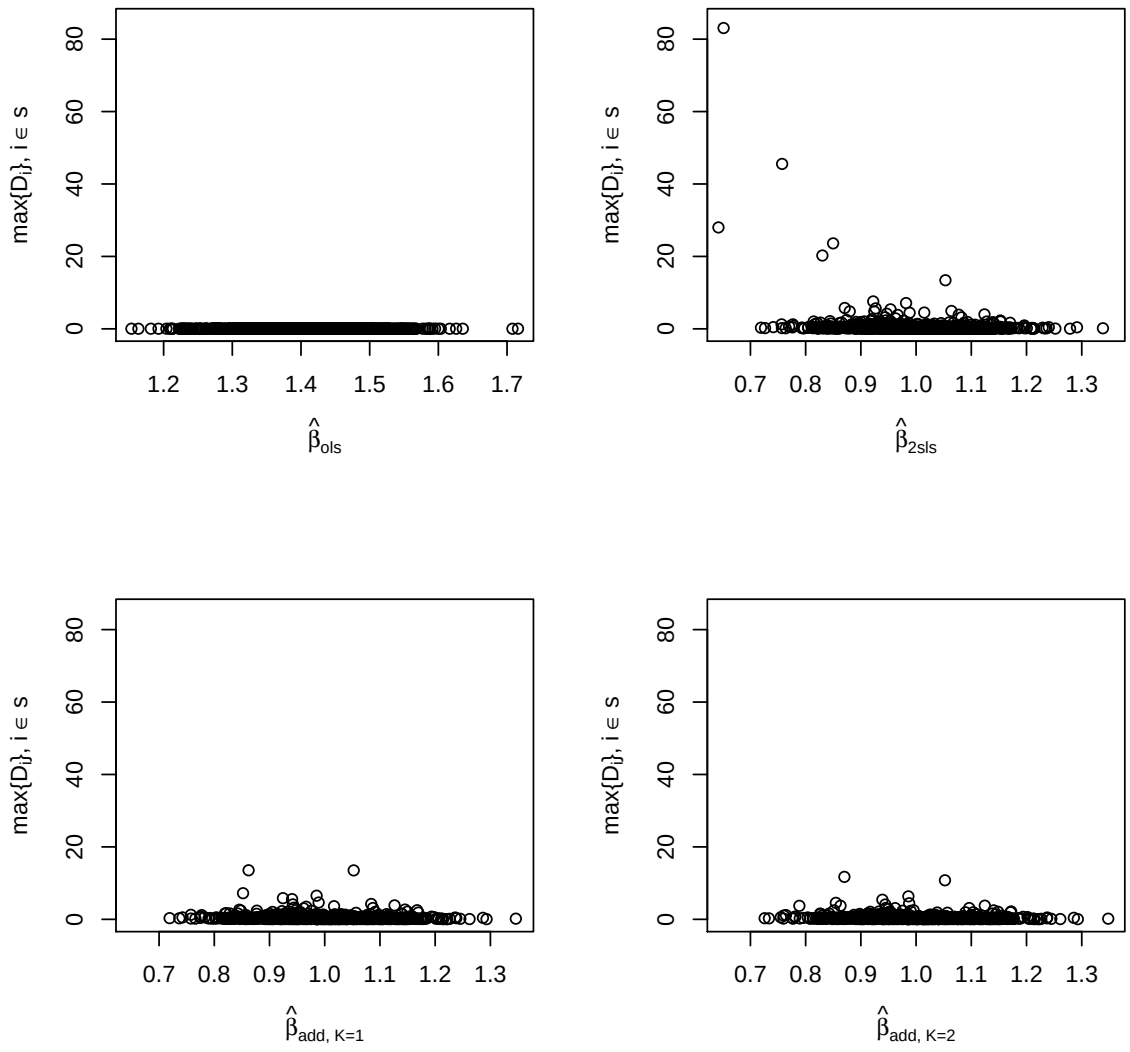

Figure 10: Graphs of maximum Cook's distance vs. estimated β for all 1000 replications of the *advantage* simulation, for all four estimators. The $\hat{\beta}_{2sls}$ graph has large $\max D_i$'s for several of its low-end biased estimates.

excessive influence of those data values at stage two. The complex relationship between weights, x-values, and IV's at stage one makes it difficult to draw specific conclusions, but the evidence suggests that for informative sampling surveys that will possibly include extreme values with large weights, the techniques of Chapter 3 should be considered.

CHAPTER 5

CONCLUSION

We examined two survey sampling problems in this paper. The first was the estimation of a population mean and the second was the estimation of regression coefficients. Both problems involved an atypical reliance on the study variable and both problems utilized nonparametric methods in their solution.

In Chapter 2, the study variable was used in the post-stratification process. We developed the NEPSE and demonstrated its consistency and asymptotic normality under a superpopulation model. We also showed that the NEPSE has the same asymptotic variance as the traditional post-stratified estimator with fixed strata. This work provides justification for the estimation methods currently being used by the U.S. Forest Service. Using simulations, we demonstrated the finite sample properties of the estimator. We found that the NEPSE generally outperformed the Horvitz-Thompson estimator, and performed similarly to the traditional PS estimator. The NEPSE also generally outperformed the estimator with simple linear regression weights, with exceptions in cases where the response variable was linear or near linear. Future work may involve more detailed examination of specific nonparametric methods used in developing sample-fitted classification schemes.

In Chapter 3, the inclusion probabilities were a function of the study variable. We developed an estimator for regression coefficients that uses instrumental variables and a penalized spline at stage one of a two-stage process. Unlike the ordinary least squares estimator, the penalized spline estimator is consistent under informative

sampling. In certain sampling situations, the penalized spline estimator outperforms the traditional two-stage least squares estimator because of a bias/variance trade-off. We also provided a method for estimating the optimal smoothing parameter λ_N for the penalized spline estimator. In our informative sampling simulation, we found that the penalized spline estimator significantly outperformed the ordinary least squares estimator and the other two-stage least squares estimators. The penalized spline estimator also performed slightly better than the Pfeffermann-Sverchkov estimator.

In Chapter 4, we continued our study of the penalized spline estimator by examining the reasons for the decreased variance that often accompanied the use of additional instrumental variables. The analytic expressions were intractable, so we relied mainly on simulations and other tools including Cook's distance which allowed us to quantify the influence of the fitted x-values. In our informative sampling simulations, we found that large-weight observations often had less influence on the estimators that utilized additional instrumental variables because of the increased flexibility of the model at stage one.

Future work with the penalized spline estimator may involve its comparison to other estimators for new data sets and the examination of properties that arise when the technique is applied to more complex regression models.

REFERENCES

Billingsley, P. (1995). *Probability and Measure* (3rd ed.). New York: John Wiley & Sons.

Blackard, J., M. Finco, E. Helmer, G. Holden, M. Hoppus, D. Jacobs, A. Lister, G. G. Moisen, M. Nelson, R. Riemann, B. Ruefenacht, D. Salajanu, D. Weyermann, K. Winterberger, T. Brandies, R. Czaplewski, R. McRoberts, P. Patterson, and R. Tymcio (2008). Mapping U.S. forest biomass using nationwide forest inventory data and moderate resolution information. *Remote Sensing of Environment 112*, 1658–1677.

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

Casella, G. and R. L. Berger (2002). *Statistical Inference* (2nd ed.). India: Cengage Learning.

Chambers, R. L. and C. J. Skinner (Eds.) (2003). *Analysis of Survey Data*. Chichester, U. K.: John Wiley & Sons.

Cook, R. D. and S. Weisberg (1982). *Residuals and Influence in Regression*. Washington, D. C.: Chapman and Hall.

Crist, E. P. and R. C. Cicone (1984). A physically-based transformation of Thematic Mapper data - the TM Tasseled Cap. *IEEE Transactions on Geoscience and Remote Sensing GE-22*, 256–263.

Czaplewski, R. L. (2010). Complex sample survey estimation in static state-space. Gen. Tech. Rep. RMRS-GTR-239, U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station, Fort Collins, CO.

Dahlke, M., F. J. Breidt, J. D. Opsomer, and I. Van Keilegom (2012). Nonparametric endogenous post-stratification estimation. *Statistica Sinica*, (in press), doi:10.5705/ss.2011.272.

Fan, J., N. E. Heckman, and M. P. Wand (1995). Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *Journal of the American Statistical Association 90*(429), 141–150.

Frayer, W. E. and G. M. Furnival (1999). Forest survey sampling designs: A history. *Journal of Forestry 97*, 4–8.

Fuller, W. A. (2009). *Sampling Statistics*. Hoboken, NJ: Wiley & Sons.

Green, P. J. and B. W. Silverman (1994). *Nonparametric Regression and Generalized Linear Models*. tylis. Washington, D. C.: Chapman and Hall.

Hausman, J. A. and D. A. Wise (1981). Stratification on endogenous variables and estimation: the gary income maintenance experiment. In C. F. Manski and D. McFadden (Eds.), *Structural Analysis of Discrete Data with Econometric Applications*, Chapter 10, pp. 365–391. Cambridge: MIT Press.

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

Jewell, N. P. (1985). Least squares regression with data arising from stratified samples of the dependent variable. *Biometrika 72*(1), 11–21.

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

Lehmann, E. L. and G. Casella (1998). *Theory of Point Estimation* (2nd ed.). New York: Springer.

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models* (2 ed.). London:

Chapman and Hall.

McRoberts, R. E., M. D. Nelson, and D. G. Wendt (2002). Stratified estimation of forest area using satellite imagery, inventory data, and the k-nearest neighbors technique. *Remote Sensing of Environment 82*, 457–468.

Moisen, G. G. and T. S. Frescino (2002). Comparing five modelling techniques for predicting forest characteristics. *Ecological Modelling 157*, 209–225.

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

Pfeffermann, D. and M. Y. Sverchkov (2003). Fitting generalized linear models under informative sampling. In R. L. Chambers and C. J. Skinner (Eds.), *Analysis of Survey Data*, Chapter 12, pp. 175–195. England: John Wiley & Sons.

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 (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Scott, C. T., W. A. Bechtold, G. A. Reams, W. D. Smith, J. A. Westfall, M. H. Hansen, and G. G. Moisen (2005). Sample-based estimators used by the forest inventory and analysis national information management system. Gen. Tech. Rep. SRS-80, 53–77, US Department of Agriculture, Forest Service, Southern Research Station, Asheville, NC.

Silverman, B. W. (1999). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall Ltd.

Simonoff, J. S. (1996). *Smoothing Methods in Statistics*. New York: Springer-Verlag.

ten Cate, A. (1986). Regression analysis using survey data with endogenous design. *Survey Methodology 12*(2), 121–138.

van der Vaart, A. W. and J. A. Wellner (1996). *Weak Convergence and Empirical Processes*. Springer-Verlag Inc.

Wooldridge, J. M. (2009). *Introductory Econometrics: A Modern Approach* (4 ed.). Mason, OH: South-Western Cengage Learning.