KERNEL REGRESSION ESTIMATION FOR INCOMPLETE DATA

A thesis submitted in partial fulfillment of the requirements
For the degree of Master of Science
in Mathematics

by

Timothy Reese

May 2015
The thesis of Timothy Reese is approved:

______________________________  ______________________
Dr. Mark Schilling               Date

______________________________  ______________________
Dr. Ann E. Watkins               Date

______________________________  ______________________
Dr. Majid Mojirsheibani, Chair   Date

California State University, Northridge
Dedication

This thesis is dedicated in memory of my grandmother Patrica, who first inspired me to pursue higher education. I also dedicate this thesis to the friendship and memory of my good friend Mathew Willerford, his passing in 2014 was difficult. Mathew was caring and always put family and friends before himself. He left the world too early.
Acknowledgments

I would first like to express my sincerest gratitude to Dr. Majid Mojirsheibani. The guidance and support provided by Dr. Mojirsheibani has been crucial to the completion of this thesis. The patience Dr. Mojirsheibani has demonstrated throughout this process is greatly appreciated. Dr. Mojirsheibani has shown by example, the necessary traits required to produce meaningful work in the field of statistics. I hope that during this time I have adopted a portion of these traits. I would also like to thank Dr. Mark Schilling for introducing me to the beautiful world of statistics. Without such a great introduction to this field, I may have instead pursued a masters degree in computer science. Furthermore, I would like to thank my committee members Dr. Ann E. Watkins and Dr. Mark Schilling for their time and insightful comments. Additionally, I would like to extend thanks to many of the people in the math department who have provided support in one form or another. Finally, I am always grateful to my family for being supportive throughout my life, so I extend thanks to my parents Ron and Linda, and sister Katrina.
Table of Contents

Signature page iii

Dedication iii

Acknowledgments iv

Abstract vii

Introduction 1

1 Introduction to regression 2
  1.1 What is regression? 2
  1.2 Defining the relationship between response and predictors 3
  1.3 Parametric regression 4
    1.3.1 Least-squares 5
    1.3.2 Supremum norm covers and total boundedness 7
    1.3.3 Empirical $L_p$ covers 9
  1.4 Nonparametric regression 12
    1.4.1 Partitioning Regression Estimator 12
    1.4.2 k-Nearest Neighbors (k-NN) 14
    1.4.3 Kernel Regression 15
  1.5 Comparison between parametric least-squares and kernel regression estimators 17

2 Missing data 21
  2.1 Missing patterns 21
  2.2 Selection probability 23
  2.3 Regression in the presence of missing values 24

3 Kernel regression estimation with incomplete covariates 27
  3.1 Main Results 28
    3.1.1 A kernel-based estimator of the selection probability 32
    3.1.2 A least-squares based estimator of the selection probability 43
  3.2 An application to classification with missing covariates 53
  3.3 Kernel regression simulation with missing covariates 55
ABSTRACT

KERNEL REGRESSION ESTIMATION FOR INCOMPLETE DATA

By

Timothy Reese

Master of Science in Mathematics

The problem of estimating the regression function from incomplete data is explored. The case in which the data is incomplete in the sense that a subset of the covariate vector is missing at random is the concern of chapter 3. Chapter 4 analyzes the case when the response variable is missing at random. In each of the aforementioned chapters, the problem of incomplete data is handled by constructing kernel regression estimators that utilize Horvitz-Thompson-type inverse weights, where the selection probabilities are estimated utilizing both kernel regression and least-squares methods. Exponential upper bounds are derived on the performance of the $L_p$ norms of the proposed estimators, which can then be used to establish various strong convergence results. We conclude chapters 3 and 4 with an application to nonparametric classification in which kernel classifiers are constructed from incomplete data.
### Introduction

Statisticians working in any field are often interested in the relationship between a response variable $Y$ and a vector of covariates $Z = (Z_1, \ldots, Z_d)$. This relationship can best be described by the regression function $m(z) = E[Y|Z = z]$. The regression function is estimated utilizing data which has been collected on the variables of interest. In practice the data is often incomplete, having missing values among the covariates or the response variables. Values may be unavailable for a variety of reasons, for example, corruption may occur in the data collection or storage process, a responding unit may fail to answer a portion of the survey questions, or perhaps the measurement of a variable is too expensive to collect for the entire sample. Regardless of the reason, missing values are problematic, standard regression methods are no longer available and if the statistician performs a complete case analysis, the results may be inefficient at best, and at worst highly biased.

Two common alternatives to the complete case analysis have been established in the literature. The most commonly employed method is that of imputation. Imputation methods attempt to replace the missing values by some estimate. The second approach utilizes a Horvitz-Thompson-type inverse scaling method (Horvitz and Thompson (1952)) which works by weighting the complete cases by the inverse of the missing probabilities. This thesis approaches the problem of estimating the regression function by way of the nonparametric method of kernel regression. The problem of incomplete data is handled by employing the Horvitz-Thompson-type inverse weighting technique.

We start with a review of the problem of regression for the case in which the data is complete. Both parametric and nonparametric regression estimators are investigated. Convergence results for each estimator to the regression function are explored. Chapter 2 begins with an analysis of the patterns in which data can become unavailable. Next, the selection probabilities are clearly defined for a multitude of cases in which the data may be missing. We shall conclude the chapter with a discussion on the methods for estimating the regression function in the presence of missing data i.e., complete case analysis, imputation, and Horvitz-Thompson-type inverse weighting.

In chapters 3 and 4, kernel regression estimators are proposed for handling the problem of missing data. Chapter 3 explores the case when a subset of the covariate vector may be missing. Chapter 4 develops similar results to that of chapter 3 for the case of missing response. In both chapters exponential upper bounds are established on the performance of the $L_p$ norms for each estimator. These bounds are used to establish almost sure convergence results. Each chapter concludes with an application to the problem of nonparametric classification in the presence of incomplete data.
1 Introduction to regression

1.1 What is regression?

Regression is the study of the relationship between some dependent variable \( Y \) called the response variable with a set of independent variables \( Z = (Z_1, \cdots, Z_d) \) called the predictors or covariates. The response variable is assumed to have some functional relationship to the predictor variables i.e., if we observe \( Z = z \) then \( Y = f(z) + \epsilon \), where \( \epsilon \) is some random error term. This relationship is studied in part by the regression function \( m(z) = E[Y|Z = z] \). The importance of the regression function lies in the fact that \( m(\cdot) \) is ”close” to \( Y \). Closeness here is measured by the \( L_2 \) risk or more commonly referred to as the mean squared error. The mean squared error is given by

\[
E |Y - f(Z)|^2. \tag{1.1}
\]

It is easily shown that \( m(\cdot) \) minimizes (1.1), i.e.,

\[
m(z) = \arg\min_{f: \mathbb{R}^d \rightarrow \mathbb{R}} E |Y - f(Z)|^2.
\]

Let \( f \) be any measurable function from \( \mathbb{R}^d \) into \( \mathbb{R} \), then

\[
E |Y - f(Z)|^2 = E |Y - m(Z) + m(Z) - f(Z)|^2 = E |Y - m(Z)|^2 + E |m(Z) - f(Z)|^2 + 2E [(Y - m(Z))(m(Z) - f(Z))] = E |Y - m(Z)|^2 + E |m(Z) - f(Z)|^2. \tag{1.2}
\]

The last line follows by observing that

\[
E [(Y - m(Z))(m(Z) - f(Z))] \overset{\text{a.s.}}{=} E [E [(Y - m(Z))(m(Z) - f(Z))|Z]] = E [(m(Z) - f(Z)) E [(Y - m(Z))|Z]] = E [(m(Z) - f(Z))(m(Z) - m(Z))] = 0.
\]

Therefore, since (1.2) is greater than or equal to zero, we have that (1.2) is minimized when \( f \) is chosen to be the regression function.

The regression function is important to many different fields of study. The following are a few examples of areas where regression techniques can be applied.
1. In gene expression profiling, information from thousands of genes can be used to predict recurrence rates for breast cancer.

2. Regression techniques can be used to predict an inmate’s propensity for violence. One such method is the Risk Assessment Scale for Prison (RASP-Potosi). RASP-Potosi attempts to measure an inmate’s propensity for violence from a vector of covariates which include age, length of sentence, education, prior prison terms, prior probation sentences, conviction for a property offense, and years served (Davis 2013).

3. The success of a business often relies on advertising strategies. Regression techniques may be applied to acquire the necessary information to choose the right advertising strategy. For example, one could model the sales level in relation to the amount of money spent on advertising in different mediums, such as TV, newspaper, billboards, and internet.

In each of these examples the researcher observes some data and then utilizes the data to construct an estimator of the regression function. The regression estimate can then be used to make inferences.

1.2 Defining the relationship between response and predictors

Let \((Z, Y)\) be a \(\mathbb{R}^d \times \mathbb{R}\)-valued random vector, with cumulative distribution function (CDF) \(F(Z, Y) = P(Z \leq z, Y \leq y)\). The random vector \(Z = (Z_1, \ldots, Z_d)\) is the covariate vector and \(Y\) is the response variable of interest. One typically wishes to examine the relationship between these variables by way of the regression function

\[
m(z) = E[Y | Z = z].
\] (1.3)

For example, the covariate vector \(Z\) could consist of the following predictor variables: monthly income, amount of debt, previous political contributions, and asset value, the response variable of interest \(Y\) could be political contribution. The relationship of the potential contribution to the predictors could then be modeled as some function of the covariates with an additional random error term i.e., \(Y = f(Z) + \epsilon\). The mean response of this relationship is portrayed by the regression function and could be used to determine the amount of political pandering a politician should expend on a particular person or group. However, in practice the regression function is unavailable, since the distribution of \((Z, Y)\) is typically unknown. Instead one attempts to estimate the regression function by utilizing some observed data \(D_n = \{(Z_1, Y_1), \ldots, (Z_n, Y_n)\}\), in which \((Z_i, Y_i)’s\) are independent and identically distributed (i.i.d) random vectors with common distribution \(F(Z, Y)\).
To determine the utility of an estimator it is important to show that the estimator converges to the true regression function. The usefulness of the estimator can be measured by the $L_p$ statistic. The $L_p$ statistic is a commonly used measure of global accuracy for the regression estimator denoted as $m_n(\cdot)$, which is defined as follows

$$I_{n}(p) = \int |m_n(z) - m(z)|^p \mu(dz), \quad 1 \leq p < \infty,$$

where $\mu$ is the probability measure of $Z$. $I_{n}(2)$ is the mean squared error, this is the most commonly used measure of accuracy for regression estimators. The primary concern is to show that the $L_p$ statistic converges to 0 in probability or almost surely (see Appendix A) as $n \to \infty$, for some fixed value of $p$. The convergence of the $L_p$ statistic to zero demonstrates the consistency of $m_n(\cdot)$ as an estimator of $m(\cdot)$.

There are two main approaches to estimation of the regression function. The first method is a parametric-based estimator, which assumes some partial knowledge of the underlying functional relationship between the response and vector of covariates. Once the model assumptions have been made, one proceeds to fit the data under some error minimization criterion. The second methodology utilizes local averaging techniques. These techniques function without any reliance on restrictive model assumptions. In the following subsections we briefly discuss both approaches, paying particular attention to the method of least-squares, and the method of kernel regression.

### 1.3 Parametric regression

When there is prior knowledge about the functional relationship between the covariates and the response variable, it is important to use this information in the estimation of the regression function. In parametric regression one makes assumptions about the model $Y = f(Z) + \epsilon$, specifically the function $f$ is assumed to belong to some class of functions $\mathcal{F}$. The observed data is then used to select the best function among the class of functions $\mathcal{F}$. This function is chosen such that the residual error is small for the given data under some chosen minimization criterion. The popular least-squares criterion may be used to minimize the residual error over the given data.
1.3.1 Least-squares

The method of least-squares selects the function from the assumed class of functions that minimizes the sum of the squared errors, and is defined as follows

$$\hat{m}_{LS} = \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{n} (Y_i - f(Z_i))^2.$$  \hspace{1cm} (1.5)

The following is an example of the simplest case of least-squares regression. In this case we assume that the functional relationship between the response variable $Y$ with that of the single predictor $Z$ is linear.

**Example 1** (Simple Linear Model) Let $Z$ be a single predictor random variable and let $Y$ be the corresponding response. Assume the model $Y = \beta_0 + \beta_1 Z + \epsilon$, i.e., assume that the class of functions $\mathcal{F}$ in (1.5) is the class of linear functions of one predictor. Let the sample data be $D_n = \{(Z_1, Y_1), \ldots, (Z_n, Y_n)\}$, then the least-squares method uses simple calculus minimization techniques to yield

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{Z} \quad \hat{\beta}_1 = \frac{\frac{1}{n} \sum_{i=1}^{n} Z_i Y_i - \bar{Y} \bar{Z}}{\frac{1}{n} \sum_{i=1}^{n} Z_i^2 - \bar{Z}^2},$$
where \( \overline{Z} = \frac{1}{n} \sum_{i=1}^{n} Z_i \), and \( \overline{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i \) denote the averages of \( Z_i \) and \( Y_i \). These results yield the best fit line
\[
\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 z.
\]

In figure 1.1 two cases were displayed, case (a) in which the best fit line is a good fit and case (b) where it is a poor fit. We can see based on the figures that model assumptions are crucial to a proper parametric fit.

The most common way of analyzing the performance of the least-squares estimator in (1.5) as an estimator of the true regression function is by way of the \( L^2 \) risk and as such consider the following result, which establishes an upper bound on the \( L^2 \) risk of the least-squares estimator.

**Lemma 1.1 (Györfi et al. (2002))** Let \( \mathcal{F}_n = \mathcal{F}_n(\mathbb{D}_n) \) be a class of functions \( f : \mathbb{R}^d \to \mathbb{R} \) depending on the data \( \mathbb{D}_n = \{(Z_1, Y_1), \ldots, (Z_n, Y_n)\} \). If \( \hat{m}_{LS} \) is as in (1.5) then
\[
\int |\hat{m}_{LS}(z) - m(z)|^2 \mu(dz) \leq 2 \sup_{f \in \mathcal{F}_n} \left| \frac{1}{n} \sum_{i=1}^{n} |Y_i - f(Z_i)|^2 - E |Y - f(Z)|^2 \right| + \inf_{f \in \mathcal{F}_n} \int |m(z) - f(z)|^2 \mu(dz).
\]

**Remark.** For the proof and an explanation why the second term converges to 0, see Lemma 10.1 Györfi et al. (2002, p. 161).

Therefore to establish the convergence of the \( L^2 \) risk in lemma 1.1 to 0, it remains to show that the empirical \( L^2 \) risk
\[
\frac{1}{n} \sum_{i=1}^{n} (Y_i - f(Z_i))^2,
\]
converges uniformly over \( \mathcal{F} \), to the actual \( L^2 \) risk
\[
E |Y - f(Z)|^2,
\]
i.e., we need to establish
\[
\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(Z_i))^2 - E \left[ |Y - f(Z)|^2 \right] \right| \overset{a.s.}{\to} 0. \tag{1.6}
\]
Figure 1.2: This figure is an adaption of Figure 9.1 of Györfi et al. (2002). The dashed lines represent a $\pm \epsilon$ translation of $\tilde{f}$ in red. The figure portrays the supremum norm distance between the function $f$ and one of the class members $\tilde{f}$, being contained within $\epsilon$.

To achieve the uniform convergence in (1.6), we must formally define the idea of totally bounded classes of functions.

1.3.2 Supremum norm covers and total boundedness

As stated before parametric regression methods assume partial knowledge of the functional relationship between the response variable $Y$, to the vector of covariates $Z$. In other words, the regression function is assumed to belong to some class of functions $F$. To establish the convergence in (1.6), we assume that our class of functions $F$ is totally bounded with respect to some norm. First, consider the case in which our class of functions is totally bounded with respect to the supremum norm, as defined below.

**Definition 1.1 (Totally Bounded with Respect to the supremum Norm)** We say that a class of functions $F : \mathbb{R}^d \rightarrow \mathbb{R}$ is totally bounded with respect to the supremum norm if for every $\epsilon > 0$, there is a subclass of functions $F_\epsilon = \{f_1, \cdots, f_{N(\epsilon)}\}$ such that for every $f \in F$ there exists a $\tilde{f} \in F_\epsilon$, satisfying

$$\|f - \tilde{f}\|_\infty := \sup_{z \in \mathbb{R}^d} |f(z) - \tilde{f}(z)| < \epsilon.$$ 

We say that the class $F_\epsilon$ is an $\epsilon$-cover of $F$. The cardinality of the smallest $\epsilon$-cover is denoted by $N_\infty(\epsilon, F)$. 

7
To establish the convergence in (1.6), let $|Y| \leq M < \infty$ and suppose $f$ belongs to some known class of functions $\mathcal{F}$. Put $\mathcal{G}$ to be the class of functions described as

$$\mathcal{G} = \{g_f(z, y) = (y - f(z))^2 | f \in \mathcal{F}, g_f : \mathbb{R}^d \to [0, B], \text{ with } B = 4M^2\},$$

then we have the following theorem

**Theorem 1.1 (see Györfi et al. (2002, Lemma 9.1))** Let $\mathcal{G} : \mathbb{R}^d \to [0, M]$ be a class of functions bounded with respect to the supremum norm, then for every $\epsilon > 0$ and $n$ large enough

$$P\left\{ \sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} g(Z_i, Y_i) - E[g(Z, Y)] > \epsilon \right\} \leq 2N_{\infty}(\epsilon/3, \mathcal{G})e^{-\frac{2n\epsilon^2}{9B^2}}. \quad (1.7)$$

**Remark.** Let $g, \tilde{g} \in \mathcal{G}$ and observe that

$$\|g - \tilde{g}\|_{\infty} = \sup_{Z, Y} |g(Z, Y) - \tilde{g}(Z, Y)|$$

$$= \sup_{Z, Y} \left| (Y - f(Z))^2 - (Y - \tilde{f}(Z))^2 \right|$$

$$\leq \sup_{Z, Y} \left| (Y - f(Z)) - (Y - \tilde{f}(Z)) \right| \times \left| (Y - f(Z)) + (Y - \tilde{f}(Z)) \right|$$

$$\leq 4M \sup_{Z \in \mathbb{R}^d} \left| f(Z) - \tilde{f}(Z) \right|,$$

it then follows that if $\mathcal{F}_{\epsilon/4M} = \{f_1, \cdots, f_n\}$ is a minimal $\epsilon/4M$-cover of $\mathcal{F}$ with respect to the supremum norm, the class of functions $\mathcal{G}_\epsilon = \{g_{f_1}, \cdots, g_{f_n}\}$ will be an $\epsilon$-cover of $\mathcal{G}$. Therefore, the covering numbers for $\mathcal{F}$ and $\mathcal{G}$ satisfy $N_{\infty}(\epsilon, \mathcal{G}) \leq N_{\infty}(\epsilon/4M, \mathcal{F})$, hence (1.7) is also bounded by $2N_{\infty}(\epsilon/12M, \mathcal{F})e^{-\frac{2n\epsilon^2}{9B^2}}$. As a result, applying the Borel-Cantelli lemma yields the convergence in (1.6).

Many important classes of functions are in fact totally bounded with respect to the supremum norm; the following are some examples:

**Example 2** Consider the exponential class of functions

$$\mathcal{F} = \left\{ f_\theta(x) = e^{-\theta|x|}, \theta \in [0, 1], |x| < 1 \right\}.$$
This class is totally bounded: for every $\epsilon > 0$, we have $N_\infty(\epsilon, \mathcal{F}) \leq 2 + [1/(2\epsilon)]$. We note that this is polynomial in $1/\epsilon$. To see this, observe that

$$\sup_{|x| \leq 1} \left| e^{-\theta |x|} - e^{-\bar{\theta} |x|} \right| \leq |\theta - \bar{\theta}|,$$

where the above follows since $e^{-\theta |x|}$ is continuously differentiable and $|x| \leq 1$. This means the subclass of functions needed to cover $\mathcal{F}$ depends on the parameter $\theta$. If we let $\theta_i = 2\epsilon \cdot i$, then taking the set $\Theta = \{\theta_0, \ldots, \theta_i, \ldots, \theta_{\lfloor 1/2 \epsilon \rfloor}, 1\}$, we have for any $\theta$ there is a $\bar{\theta} \in \Theta$ such that $|\theta - \bar{\theta}| \leq \epsilon$. Thus, $\mathcal{F}$ is totally bounded with respect to the supremum norm. Similar bounds hold for the general case when $f(x) = \exp\{-\sum_{i=1}^d \theta_i x_i^2\}$ provided that $\theta_i \in [0, 1]$ and $|x_i| < 1$, $i = 1, \ldots, d$. (See Devroye et al. (1996, Ch. 28)).

**Example 3 (Differentiable functions)** Let $k_1, \ldots, k_s \geq 0$ be non-negative integers and put $k = k_1 + \cdots + k_s$. Also, for any $\psi : \mathbb{R}^s \to \mathbb{R}$, define $D^{(k)}(\psi)(u) = \partial^{k_1} \psi(u)/\partial u_1^{k_1} \cdots \partial u_s^{k_s}$.

Consider the class of functions with bounded partial derivatives of order $r$:

$$\Psi = \left\{ \psi : [0, 1]^d \to \mathbb{R} \mid \sum_{k \leq r} \sup_u |D^{(k)}(\psi)(u)| \leq C < \infty \right\}.$$

Then, for every $\epsilon > 0$, $\log N_\infty(\epsilon, \Psi) \leq M \epsilon^{-\alpha}$, where $\alpha = d/r$ and $M \equiv M(d, r)$; this is due to Kolmogorov and Tikhomirov (1959).

**Example 4 (Lipschitz bounded convex functions)** Let $\mathcal{F}$ be the class of all convex functions $f : C \to [0, 1]$, where $C \subset \mathbb{R}^d$ is compact and convex. If $f$ satisfies $|f(z_1) - f(z_2)| \leq L \|z_1 - z_2\|$ for all $z_1, z_2 \in C$, where $L > 0$ is finite, then $\log N(\epsilon, \mathcal{F}) \leq M \epsilon^{-d/2}$ for all $\epsilon > 0$, where the constant $M$ depends on $d$ and $L$ only; see van der Vaart and Wellner (1996).

### 1.3.3 Empirical $L_p$ covers

Restricting a class of functions to be bounded with respect to the supremum norm may be too restrictive, since the covers required can be exceptionally large. If the supremum norm covers are too large, then we may not obtain convergence in (1.7). An alternative, is to use smaller $L_p$ norm covers.

**Definition 1.2 (Totally Bounded with Respect to the $L_p$ Norm)** We say a class of functions $\mathcal{F} : \mathbb{R}^d \to \mathbb{R}$ is totally bounded with respect to the $L_p$ norm if for every $\epsilon > 0$, there
is a subclass \( \mathcal{F}_\epsilon = \{ f_1, \cdots, f_{N(\epsilon)} \} \) such that for every \( f \in \mathcal{F} \) there exists a \( \tilde{f} \in \mathcal{F}_\epsilon \) such that

\[
\| f - \tilde{f} \|_p = \left\{ \int \left| f(z) - \tilde{f}(z) \right|^p \mu(dz) \right\}^{\frac{1}{p}} < \epsilon .
\]

(1.8)

The cardinality of the smallest \( \epsilon \)-cover is denoted by \( N_p(\epsilon, \mathcal{F}) \).

However, these \( L_p \) covers depend on the unknown distribution of the random variable \( Z \). An alternative is to construct data dependent covers. These covers are obtained by replacing the probability measure \( \mu \) in (1.8) with the empirical measure and they are defined as follows.

**Definition 1.3 (Totally Bounded with Respect to the Empirical \( L_p \) Norm)** We say a class of functions \( \mathcal{F} : \mathbb{R}^d \rightarrow \mathbb{R} \) is totally bounded with respect to the empirical \( L_p \) norm if for every \( \epsilon > 0 \), there is a subclass \( \mathcal{F}_\epsilon = \{ f_1, \cdots, f_{N(\epsilon)} \} \) such that for every \( f \in \mathcal{F} \) there exists a \( \tilde{f} \in \mathcal{F}_\epsilon \), for fixed points \( \{(z_1, y_1), \cdots, (z_n, y_n)\} \) such that

\[
\left\{ \frac{1}{n} \sum_{i=1}^{n} \left| f(z_i) - \tilde{f}(z_i) \right|^p \right\}^{\frac{1}{p}} < \epsilon .
\]

(1.9)

The cardinality of the smallest \( \epsilon \)-cover for the empirical \( L_p \) norm is given by \( N_p(\epsilon, \mathcal{F}, \mathbb{D}_n) \).

Observe that the covering number \( N_p(\epsilon, \mathcal{F}, \mathbb{D}_n) \) relies on the data \( \mathbb{D}_n \), hence the covering number is a random variable. As a result, we will be concerned with the expected value of \( N_p(\epsilon, \mathcal{F}, \mathbb{D}_n) \). Classes of functions exist that are not totally bounded with respect to the supremum norm, but are in fact bounded with respect to the empirical \( L_p \) norm. To appreciate this, consider the following example.

**Example 5 (Non-Lipschitz bounded convex functions (Guntuboyina and Sen 2013))** Let \( \mathcal{F} \) be the class of all convex functions \( f : \mathcal{C} \rightarrow [0, 1] \), where \( \mathcal{C} \subset \mathbb{R}^d \) is compact and convex. Consider the functions given by

\[
f_j(t) := \max(0, 1 - 2^j t), \quad t \in [0, 1], \quad j > 1,
\]
the functions \( f_j(t) \) belong to \( \mathcal{F} \) but are not Lipschitz, observe that these functions are not totally bounded with respect to the supremum norm

\[
\| f_j - f_k \|_{\infty} \geq | f_j(2^{-k}) - f_k(2^{-k}) | \geq 1 - 2^{i-k} \geq 1/2, \quad \text{for all } j < k.
\]

However, it had been shown (see Guntuboyina and Sen 2013) that the class of all convex functions can be totally bounded with respect to the \( L_p \) norm, even if the functions are not Lipschitz. They show that

\[
\log N(\epsilon, \mathcal{F}, L_p) \leq M \epsilon^{-d/2},
\]

where \( M \) is a positive constant that is independent of \( \epsilon \).

To derive the convergence result of (1.6), first suppose that \( |Y| \leq M < \infty \) and let \( f \) belong to a given (known) class of functions \( \mathcal{F} \). Consider the class of functions \( \mathcal{G} \) defined as

\[
\mathcal{G} = \{ g_{f} \mid f \in \mathcal{F}, g_{f} : \mathbb{R}^d \to [0, B], \quad B = 4M^2 \},
\]

we have the following result from the empirical process theory.

**Theorem 1.2 (see Pollard (1984))** Let \( \mathcal{G} : \mathbb{R}^d \to [0, B] \) be a class of functions bounded with respect to the \( L_1 \) empirical norm, then for every \( \epsilon > 0 \) and \( n \) large enough

\[
P \left\{ \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} g(Z_i, Y_i) - E[g(Z, Y)] \right| > \epsilon \right\} \leq 8 \epsilon \log \left[ \frac{\epsilon}{8} \mathcal{N}(\epsilon, \mathcal{G}, \mathbb{D}_n) \right] e^{-2n\epsilon^2 / 128B^2}.
\]

**Remark.** To achieve the convergence in (1.6), first let \( g, g^\dagger \in \mathcal{G} \) and observe

\[
\frac{1}{n} \sum_{i=1}^{n} \left| g(Z_i, Y_i) - g^\dagger(Z_i, Y_i) \right|
\]

\[
= \frac{1}{n} \sum_{i=1}^{n} \left| (Y_i - f(Z_i))^2 - (Y_i - f^\dagger(Z_i))^2 \right|
\]

\[
\leq \frac{1}{n} \sum_{i=1}^{n} \left| (Y_i - f(Z_i)) - (Y_i - f^\dagger(Z_i)) \right| \times \left| (Y_i - f(Z_i)) + (Y_i - f^\dagger(Z_i)) \right|
\]

\[
\leq 4M \frac{1}{n} \sum_{i=1}^{n} \left| f(Z_i) - f^\dagger(Z_i) \right|.
\]

Therefore if \( \mathcal{F}_{\epsilon/4M} = \{ f_1, \cdots, f_N \} \) is a minimal \( \epsilon/4M \)-cover of \( \mathcal{F} \) with respect to the \( L_1 \) empirical norm, then the class of function \( \mathcal{G}_\epsilon = \{ g_{f_1}, \cdots, g_{f_N} \} \) will be an \( \epsilon \)-cover of \( \mathcal{G} \).
Therefore, (1.10) can be bounded with respect to the covering number for \( F \) i.e.,
\[
8E[N_1(\epsilon/8, G, \mathbb{D}_n)] e^{-\frac{2n\epsilon^2}{128d^2}} \leq 8E[N_1(\epsilon/32M, F, \mathbb{D}_n)] e^{-\frac{2n\epsilon^2}{128d^2}}.
\]

Thus, Theorem 1.2 combined with an application of the Borel Cantelli lemma, yield the almost sure convergence in (1.6), provided that
\[
\log(E[N_1(\epsilon/32M, F, \mathbb{D}_n)])/n \to 0, \quad \text{as } n \to \infty.
\]

1.4 Nonparametric regression

When there is little or no knowledge about the underlying relationships between the co-
variates and response variable then parametric methods will be problematic, since any model assumptions need to be accurate to produce reasonable estimates of the regression function. An alternative to the parametric methodology is nonparametric methods. Nonparametric methods perform without any assumptions on the functional relationship. Nonparametric methods use a type of local averaging of the response variables \( Y_i \). More specifically, the \( Y_i \)'s are weighted according to how "close" each corresponding \( Z_i \) is to the \( z \) of interest and then these weighted \( Y_i \)'s are added to produce the predicted response. The corresponding nonparametric regression estimator is as follows
\[
m_n(z) = \sum_{i=1}^{n} W_{n,i}(z)Y_i, \quad (1.11)
\]

where the weights \( W_{n,i} \) are assigned a real number value determined by the "closeness" of \( z \) to \( Z_i \). The way in which these weights \( W_{n,i} \) are chosen determines the type of non-
parametric regression estimator to be used. In this section we give an overview of three standard nonparametric based regression estimators.

1.4.1 Partitioning Regression Estimator

The first nonparametric based estimator to be discussed is the partitioning regression es-
timator. The partitioning estimator or regessogram as referred to by Tukey in 1961. The partitioning regression estimator starts by partitioning the covariate space \( \mathbb{R}^d \) into a collection of nonoverlapping cells denoted as \( \{A_{n,1}, A_{n,2}, \cdots\} \). The regression estimate of \( z \) is then determined by averaging the \( Y_i \)'s corresponding to the \( Z_i \)'s in the cell which \( z \) resides. For, example in figure 1.3 the \( y \)'s corresponding to the points in the circle would be averaged, since they are the only ones in the cell containing the new observation \( x \).
Formally, let $A_n[z]$ denote the cell containing $z$, then the regression estimate is

\[ m_n(z) = \frac{\sum_{i=1}^{n} Y_i I\{Z_i \in A_n[z]\}}{\sum_{i=1}^{n} I\{Z_i \in A_n[z]\}}, \quad (1.12) \]

where $I\{\cdot\}$ denotes the indicator function. The indicator function is one if $Z_i$ belongs to the region containing $z$ and 0 otherwise.

In this case the weight function of (1.11) is chosen to be $I\{Z_i \in A_n[z]\} \div \sum_{i=1}^{n} I\{Z_i \in A_n[z]\}$. A computationally efficient way to choose the partitions is to use cubic and rectangular partitions which partition the space into cells of equal dimension. However, this method can result in some cells containing a large fraction of the data, while other cells contain only a couple points. An alternative to equally spaced cells is to use statistically equivalent blocks. The blocks are constructed to satisfy the criterion that each cell should contain an equal number of points. The almost sure convergence of (1.12) to the regression function is shown below. For a proof, refer to Györfi et al (2002).

**Theorem 1.3** Let $m_n(z)$ be defined as in (1.12). If for each sphere $S$ centered at the origin

\[ \lim_{n \to \infty} \max_{j: A_n \cap S \neq \emptyset} D(A_{n,j}) = 0, \]

where...
where $D$ denotes the diameter of the cell, and

$$\lim_{n \to \infty} \frac{|\{j : A_{n,j} \cap S \neq \emptyset\}|}{n} = 0.$$ 

then

$$E[|m_n(Z) - m(Z)| \mid \mathbb{D}_n] \xrightarrow{a.s.} 0 \text{ as } n \to \infty,$$

provided that $|Y| \xrightarrow{a.s.} M < \infty$.

### 1.4.2 k-Nearest Neighbors (k-NN)

An alternative to the partitioning method or regressogram is the $k$-nearest neighbors ($k$-NN) estimator. Instead of partitioning the space of the covariates, the $k$-NN estimator sorts the data based on how “far” each $Z_i$ is from $z$ and then proceeds to average the $Y_i$’s of the $k$ closest. Let the data used for estimating the regression function be $\mathbb{D}_n = \{(Z_1, Y_1), \ldots, (Z_n, Y_n)\}$ which is an i.i.d. sample from the distribution of $(Z, Y)$. Then denote $\mathbb{D}'_n = \{(Z_{(1)}, Y_{(1)}), \ldots, (Z_{(k)}, Y_{(k)}), \ldots, (Z_{(n)}, Y_{(n)})\}$ as the reordered data set based on how far each $Z_i$ is from $z$, here distance is measured using some distance function (for example Euclidean distance). Then, the corresponding $k$-NN estimator is defined as

$$m_n(z) = \sum_{i=1}^{k} \frac{1}{k} Y_{(i)}.$$  \hspace{1cm} (1.13)

The expression in (1.13) is simply the average of the $Y_i$’s corresponding to the $k$ closest $Z_i$’s to $z$, in this case the weight function of (1.11) is chosen to be

$$W_{n,i}(z) = \begin{cases} \frac{1}{k} & \text{for the } k \text{ closest } Z_i \text{ to } z \\ 0 & \text{otherwise} \end{cases}.$$ 

The observant reader may recognize that issues may arise when there are $Z_i$’s of equal distance from $z$. If one is willing to assume that these events happen with probability 0 (this is true for the case when $Z$ has a density) then it can be shown that $m_n$ in (1.13) converges to the regression function almost surely in the $L_1$ sense, provided that $Y$ is bounded (Devroye and Györfi (1985)). If the distribution of $Z$ is not absolutely continuous, then tie breaking methods must be employed to derive the almost sure convergence of the estimator. There have been a variety of methods proposed for tie breaking. Stone (1977) proposed a pseudo $k$-NN method which during the event of a tie would average the responses for the values of the tie, Stone showed the weak convergence in probability of this pseudo $k$-NN estimator.
An alternative to Stone’s method is the method of randomization in which a new uniform random variable $U$ independent of $Z$ is introduced producing the random vector $(Z, U)$. We then artificially grow the data producing

$$\mathbb{D}_n' = \{(Z_1, U_1, Y_1), \ldots, (Z_n, U_n, Y_n)\},$$

where $U_i$’s are i.i.d. Uniform(0,1) random variables. Next, reorder the data based on the distance the new observation $(z, u)$ is from $(Z_i, U_i)$, for each $i = 1, \ldots, n$, i.e., if no tie exists for the distance between $Z_i$ and $z$ then reorder normally, otherwise if a tie exists between say $Z_i$ and $Z_j$ then compare the distance between $U_i$ to $u$ and $U_j$ to $u$ to break the tie. This method results in the reordered data set

$$\mathbb{D}_n'' = \{(Z_{(1)}, Y_{(1)}), \ldots, (Z_{(n)}, Y_{(n)})\}.$$

The estimator now averages the first $k$ $Y_i$’s in the reordered data set to produce the estimate for the response. The almost sure convergence of the $k$-NN estimator under the randomization tie breaking rule is provided by the following result.

**Theorem 1.4** *(Devroye et al. (1994)).* Let $m_n(z)$ be the $k$-NN regression estimator in (1.13). If $|Y| \leq M < \infty$, if the distribution of $(Z, Y)$ is not absolutely continuous apply the randomization rule, then for every $\epsilon > 0$, and $n$ large enough

$$P\left(\int |m_n(z) - m(z)|\mu(dz) > \epsilon\right) \leq e^{-an},$$

where $a$ is a positive constant that does not depend on $n$.

**Remark.** Furthermore, we obtain $\int |m_n(z) - m(z)|\mu(dz) \overset{\text{a.s.}}{\to} 0$, by the Borel-Cantelli lemma, provided that $k \to \infty$ and $k/n \to 0$ as $n \to \infty$. Therefore, we have that $m_n(z)$ converges almost surely to $m(z)$ in the $L_1$ sense. Further results hold, specifically the above theorem can be extended to the general $L_p$ statistics by the fact that $Y$ is bounded *(Devroye et al. (1994)).*

### 1.4.3 Kernel Regression

Another local averaging method is the Nadaraya-Watson kernel regression estimator *(Nadaraya (1964), Watson (1964)), which assigns weights that decrease the "farther" $Z_i$ is from $z,
and is defined as follows

\[ m_n(z) = \frac{\sum_{i=1}^{n} Y_i \mathcal{K}\left(\frac{z-Z_i}{h_n}\right)}{\sum_{i=1}^{n} \mathcal{K}\left(\frac{z-Z_i}{h_n}\right)}. \]  \hspace{1cm} (1.14)

The kernel function \( \mathcal{K} : \mathbb{R}^d \rightarrow \mathbb{R} \) is a smooth function of the distance between \( z \) and \( Z_i \). The choice of kernel is at the discretion of the statistician but is typically chosen to be a probability density function. Three common choices of kernels for the case of a univariate response are displayed in Figure (1.4). The term \( h_n > 0 \) in (1.14), called the smoothing parameter of the kernel, typically ranges between 0 and 1. The smoothing parameter controls the smoothness of the kernel regression estimate. The smoothing parameter depends on the size of the sample \( n \), as \( n \) gets larger \( h_n \) gets smaller but at a rate that does not decrease too fast. One typically estimates the smoothing parameter from the data at hand.

Various modes of convergence of \( I_n(p) \) defined in (1.4) to zero have been established in the literature; see, for example, Devroye and Wagner (1980), and Spiegelman and Sacks (1980), whose results yield \( I_n(1) \overset{P}{\longrightarrow} 0 \), as \( n \to \infty \). The quantity \( I_n(1) \) also plays an important role in statistical classification; see for example Devroye et al (1996, Sec. 6.2), Devroye and Wagner (1980), and Krzyżak and Pawlak (1984). For the almost sure convergence of \( I_n(1) \) to 0 (i.e., \( I_n(1) \overset{a.s.}{\longrightarrow} 0 \) see, for example, Devroye (1981), and Devroye and Krzyżak (1989). In this last cited work by Devroye and Krzyżak, they develop a number of equivalent results under the assumption that \( |Y| \leq M < \infty \) and the kernel is regular.
**Definition** A nonnegative kernel $K$ is said to be regular if there are positive constants $b > 0$ and $r > 0$ for which $K(z) \geq b I\{z \in S_{0,r}\}$ and \(\int \sup_{y \in \mathbb{R}^d} K(y) dz < \infty\), where $S_{0,r}$ is the ball of radius $r$ centered at the origin.

Many of the kernels used in practice satisfy the above definition. A widely used regular kernel is the Gaussian kernel

\[ K(z) = (2\pi)^{-\frac{d}{2}} e^{-\frac{1}{2}z^Tz}. \]

The results of Devroye and Krzyżak (1989) are of particular importance for later chapters of this thesis. In chapters 3 and 4 we shall produce similar results to the following theorem, for the case of incomplete data.

**Theorem 1.5** (Devroye and Krzyżak (1989)). Assume that $K$ is a regular kernel. Let $m_n(z)$ be the kernel regression estimate in (1.14), and let $0/0$ be defined as $0$. Then for every distribution of $(Z, Y)$, with $|Y| \leq M < \infty$ and for every $\epsilon > 0$ and $n$ large enough we have

\[ P\left(\int |m_n(z) - m(z)| \mu(dz) > \epsilon\right) \leq e^{-cn}, \]

where $c = \min \left\{ \frac{\epsilon^2}{128M^2(1 + c_1)}, \frac{\epsilon}{32M(1 + c_1)} \right\}$.

**Remark.** This result provides $\int |m_n(z) - m(z)| \mu(dz) \xrightarrow{a.s.} 0$, by the Borel-Cantelli lemma, provided that $\lim_{n \to \infty} h_n = 0, \lim_{n \to \infty} nh^d = \infty$. Hence, the almost sure convergence of $m_n(z)$ to $m(z)$ is established. For an in-depth treatment of kernel regression and kernel density estimation see the Monograph by Wand and Jones (1995).

**1.5 Comparison between parametric least-squares and kernel regression estimators**

In this section a simulation is presented to compare kernel regression estimators with parametric least-squares regression estimators. A sample $D_n = \{(X_1, Y_1), \ldots, (X_{75}, Y_{75})\}$ is constructed from $Y_i = e^{-2|X_i| + 5} + \epsilon_i$, where $X_i$ and $\epsilon_i, i = 1, \ldots, n$ are standard normal random variables. The kernel regression estimate is fitted using a Gaussian kernel for fixed smoothing parameters $h = 0.5, 0.3, 0.2, 0.1$. Four choices of parametric models are chosen to demonstrate the importance of correct model specification. Each regression estimator is tested on an independent sample of 1000 points, the results are plotted in Figure 1.5 and the corresponding errors of each regression fit are provided in Table 1.1. The errors in the table are constructed by comparing the fitted value to the actual $Y_i$ for
Table 1.1: Empirical $L_1$, $L_2$, and Max Errors

\[ i = 1, \ldots, 1000. \]  The error is measured using the following methods. The standard $L_1$
error (or least absolute deviations) is given by

\[ \frac{1}{n} \sum_{i=1}^{n} \left| Y_i - \hat{Y}_i \right|, \]  

(1.15)

the empirical \( L_2 \) risk

\[ \frac{1}{n} \sum_{i=1}^{n} \left( Y_i - \hat{Y}_i \right)^2, \]  

(1.16)

and the max error is

\[ \max_{1 \leq i \leq 1000} \left| Y_i - \hat{Y}_i \right|. \]  

(1.17)

In Figure 1.5 (a), we see a poor fit for the case of both estimators. The parametric model of \( Y = \beta_0 + \beta_1 x + \beta_2 x^2 \epsilon \) is a poor choice considering the actual model is nonlinear. For the kernel regression estimate the smoothing parameter of \( h = 0.5 \) is too smooth for a sample size of \( n = 75 \). In Figure 1.5 (b), improvement is seen from the kernel regression estimator, the smoothing parameter choice of \( h = 0.3 \) is closer to the correct choice. The choice of the parametric fourth degree polynomial is better as we would expect, but inadequate, as seen in the image and by the error rates in the provided table. In figure 1.5 (c), there is improvement by both methods. Considering that the parametric fit \( Y = \beta_0 + \beta_1 e^{-|x|} \) is a linear model, the model performs surprisingly well in trying to approximate the nonlinear model. However, the linear model appears to have some trouble approximating the curvature of the relationship. Finally, in figure 1.5 (d) we see that the parametric fit performs great. The correct nonlinear model was specified as \( Y = e^{\beta_1 |x| + \beta_2} \) and as such resulted in an excellent approximation, which can be seen in the errors and by visual inspection. The kernel regression method with smoothing parameter \( h = .1 \) visually appears unstable, but the error rate dropped.

This simulation demonstrates the heavy dependence of parametric regression on the correct specification of the model. When a model is incorrectly specified, the error will be large. Visual inspection of the plots allows for easy identification of model inconsistencies. However, in practice there is no such access to plots, due to the fact that there is usually a large number of predictors. The statistician must be confident in the validity of any model assumptions or risk producing incorrect inferences. If the statistician is unable to rely on model assumptions, it is prudent to use nonparametric methods. Nonparametric methods
are not hindered by correct model specifications. However, the kernel regression method relies on the correct choice of the smoothing parameter $h$. If the smoothing parameter is too large (relative to the sample size) then the estimator will be too smooth, while too small of an $h$ leads to over-fitting. The dependence of the kernel regression method on the smoothing parameter can be seen in the simulation. Additional difficulties arise for both estimators when the data contains missing values, this is the focus of the next chapter.
2 Missing data

The analysis of the regression function was of primary interest in the previous chapter. Both parametric and nonparametric estimators were explored. In particular, the nonparametric kernel regression estimator is of special importance to chapters 3 and 4 of the thesis. Each of the estimators used data \( \mathcal{D}_n = \{(Z_1, Y_1), \cdots, (Z_n, Y_n)\} \) to construct an estimate of the regression function. In practice, statistical researchers encounter data with missing values. In particular, a subset of the covariate vector may be missing, due, for instance to failure of a survey participant to respond to questions deemed too personal. For instance a poll of subjects may include questions on income, sexual preference, and prior drug use. Some of those polled may deem these particular questions too personal to answer. The response variable can also be missing, a situation common in experimental design, where the covariates are controlled by the researcher. The presence of missing data poses additional problems for the statistician to overcome. The standard regression estimators are no longer available to the statistician. New methods need to be considered, and knowledge of the behavior of the missing data is required to produce reliable regression estimators.

This chapter is concerned with providing a brief presentation of the problem of missing data, see Little and Rubin (2002) for an in-depth accounting of the problem of missing data. First, the different patterns in which data may be unavailable are described. Second, the uncertainty of observing or not observing the \( i \)th observation is modeled as a Bernoulli random variable. Different assumptions on the conditional dependencies of the probability associated with this random variable called the selection probability or missing probability mechanism are presented. Lastly, three common procedures to produce regression estimators in the presence of incomplete data are examined.

2.1 Missing patterns

There exist a variety of patterns for which data may be unobservable. We shall briefly discuss five of these patterns. See Little and Rubin (2002) for a more complete treatment of missing data patterns. Figure 2.1 gives a visual representation of these five patterns. In Figure 2.1 any of the variables \( Y_1, \cdots, Y_5 \) can represent either the covariates or the response variable.

The first missing pattern is the univariate nonresponse pattern. This pattern represents the case in which several variables are completely observable, while one variable may have missing values. The variable with missing values can be either part of the vector of covariates or the response variable. However, in regression analysis this pattern is
more representative of the case of missing response. This missing pattern is common in experimental design scenarios, where the covariate vector is controlled by the researcher. Little and Rubin (2002) provide an example in which the yield of a particular crop under different input values of the covariates is of interest. The crop yield may be incomplete due to the failure of the seed to germinate.

The multivariate two pattern problem is the missing pattern in which there are two distinct subsets of vectors. The first subset is always available and the second subset has missing values. The subset with missing values will have the same observation missing for each variable. This is common in survey sampling, in which individuals in the sample fail to report on some of the predictor variables. Another scenario that results in this pattern is when extreme cost of measurement exist for some of the variables. For example, consider a car manufacturer who wishes to measure the safety and durability of their vehicles. Some of the measurements can be cheap, like stopping time and tire durability. However, measurements under accident scenarios can result in the complete destruction of the vehicle. These measurements are expensive and it would not be in the interest of the company to perform repeated measurements.
When there is a monotonic decrease in the availability of observations, we denote this pattern as the monotone pattern. This pattern is common in longitudinal studies, in which measurements are conducted over a long period of time. The first measurement is always complete, since the availability at the start of the experiment determines the initial sample size. When studies progress over a long period of time, unforeseeable circumstances lead to a decrease in the responding units.

The file matching pattern is a pattern in which two subsets of the variables are never observed at the same time. This behavior can be observed in figure 2.1 (e), in this case $Y_2$ and $Y_3$ are never observable together. This pattern can result in unique problems, these problems result from the inability to estimate the relationship between the variables which are never observed together. For example, in quantum mechanics the Heisenberg’s uncertainty principle states that a particle’s position and momentum cannot both be simultaneously known, this behavior could lead to the file matching pattern. This specialized example portrays a situation in which the pattern can occur as a direct result of the experimental procedure being performed. The merging of two large government databases is a more common situation that may result in this type of pattern. The reason the merging may result in this pattern is that the databases being merged may share some common variables, but each database has a set of variables that are specific to the department responsible for that particular database. For more on the file matching pattern and other missing patterns one may refer to Little and Rubin (2002).

To further understand the behavior of missing data, methods need to be employed to deal with the uncertainty of a vector being observable. This uncertainty is explored in the next section. Probability models are developed to represent the probability that a vector is observed.

2.2 Selection probability

When dealing with incomplete data, the uncertainty of observing a variable must be modeled with some probability mechanism. The underlying missing probability mechanism, called the selection probability, represents the probability that a variable will be available. To simplify the presentation of the selection probability, consider the following situation. Let $(Z, Y)$ be an $\mathbb{R}^{d+s} \times \mathbb{R}$ random vector, with $Z = (X, V)$, $X \in \mathbb{R}^d$, $V \in \mathbb{R}^s$, and $d, s \geq 1$. Suppose the subset $V$ of the covariate vector $Z$ has the possibility of being unobservable. Introduce the new random variable $\Delta$, where $\Delta = 1$ if $V$ is observable and $\Delta = 0$ otherwise. The selection probability can then be defined as follows,
Definition 2.1 (Missing Probability Mechanism)

\[ P\{\Delta = 1|Z = z, Y = y\} = P\{\Delta = 1|X = X, V = v, Y = y\}. \]  \hspace{1cm} (2.1)

The conditional probability in (2.1) defines the probability that \( \Delta = 1 \) (\( V \) is observable) given \((X, V, Y)\). When the conditional probability depends on all variables including the possibly missing \( V \) it is called missing not at random (MNAR). When the right hand side of (2.1) equals a constant our missing probability mechanism is called Missing Completely at Random (MCAR), i.e.,

Definition 2.2 (Missing Completely at Random (MCAR))

\[ P\{\Delta = 1|Z = z, Y = y\} = P\{\Delta = 1\} = p. \]  \hspace{1cm} (2.2)

The MCAR assumption may be reasonable if the data is missing due to simple corruption in the data gathering or storage of the sample. This situation is denoted as ignorable, since proceeding if there is no missing data will yield consistent and unbiased estimates. However, this situation is unlikely and if assumed will often lead to biased estimates. Instead, a more common assumption is that of missing at random (MAR). The Missing at Random assumption is a compromise between the overly restrictive MNAR and highly unlikely MCAR assumptions. We define the MAR mechanism as follows

Definition 2.3 (Missing at Random (MAR))

\[ P\{\Delta = 1|Z = z, Y = y\} = P\{\Delta = 1|X = x, Y = y\}. \]  \hspace{1cm} (2.3)

Here, (2.3) assumes that the probability that \( V \) is missing depends only on the always observable vector \((X, Y)\).

2.3 Regression in the presence of missing values

Recall, that our main goal has been to estimate the regression function \( m(z) = E[Y|Z = z] \), when the data \( \mathbb{D}_n = \{(Z_1, Y_1), \ldots, (Z_n, Y_n)\} \) may contain missing values. A variety of techniques have been developed to deal with the problem of missing values among the data. Here we briefly introduce three common methods of estimating the regression function when exposed to incomplete data and analyze the relative merits of each method.

Complete case analysis
The first and most obvious method is to use only the cases with no missing values to estimate the regression function. This method is called complete case analysis. In complete case analysis any observation that has a missing value is ignored in the process of estimating the regression function. In complete case analysis the data used in estimating the regression function is no longer \( \mathbb{D}_n = \{ (Z_1, Y_1), \ldots, (Z_n, Y_n) \} \), instead the effective data is \( \mathbb{D}'_n = \{ (Z_1, Y_1), \ldots, (Z_m, Y_m) \} \). In this case the sample size is reduced to \( m = \sum_{i=1}^{n} \Delta_i \), with \( \Delta_i = 1 \) if there is no missing values among the \( i^{th} \) observation and \( \Delta_i = 0 \) otherwise. This method is only recommended if the number of missing cases is small relative to the sample size. If the number of missing cases is large, using complete case analysis will result in an increase in the variance of the estimator, which may lead to a drastic loss of efficiency. Another issue of complete case methods is that the reduced sample size \( m = \sum_{i=1}^{n} \Delta_i \) is a random variable. This randomness leads to difficulties. The majority of regression estimation methods assume that the sample size is a fixed quantity. Furthermore, if the selection probability is not MCAR the regression estimator is likely to be biased.

**Imputation methods**

The reduction in sample size in complete case analysis is an undesirable property. Imputation methods attempt to reconstruct the sample by imputing or "substituting" the missing values. To fill in a missing value in a case, an estimate of this value is constructed from the available data, this estimate is then inserted in place of the missing value. To be more precise, suppose we have data \( \mathbb{D}_n = \{ (Z_1, Y_1), \ldots, (Z_n, Y_n) \} \) and that a subvector \( V_i \) of \( Z_i = (X_i, V_i) \), could be missing. The imputation method would then produce the new sample \( \mathbb{D}'_n = \{ (X_1, \hat{V}_1, Y_1), \ldots, (X_n, \hat{V}_n, Y_n) \} \), where \( \hat{V}_i \) is either \( V_i \) if \( V_i \) is observable, or an estimate, if \( V_i \) was missing. Two common methods used in the area of survey sampling are hot deck and cold deck imputation. In hot deck imputation, the value of a similar available unit in the sample is used in place of the missing value. Cold deck imputation uses a similar unit from an alternative sample (perhaps a previous realization) in place of the missing value of the current sample. Another obvious possibility is to replace the missing values with a mean from the observed quantities of the same variable, this is referred to as mean imputation. Mean imputation is not recommended, this method can result in inaccurate summary statistics of the sample, in particular it underestimates the variance and tends to shift the sample correlation between variables toward zero. Another common method is to impute the missing values by performing a regression of the other available variables on the observed variables corresponding to the missing value. This regression estimate is then used to predict the missing value. The method of regression imputation conditions on the observed variables which helps reduce bias due to nonresponse.
The above methods of imputation are referred to as single imputation, since one single value is estimated to replace the missing value. Single imputation methods have the disadvantage of treating the imputed value as a known quantity, thus removing the uncertainty associated with the missing value. An alternative to single imputation is to use multiple imputation. Multiple imputation replaces the missing values with several estimates resulting in multiple complete data sets. Inferences on each data set can be combined to achieve one overall inference that properly portrays the uncertainty associated with the presence of missing values.

**Horvitz-Thompson Inverse Weights**

When the data is not MCAR the complete case analysis can result in biased estimates of the regression function. The complete case regression model can be bias-corrected by using Horvitz-Thompson-type inverse weights (Horvitz and Thompson (1952)). This method works by weighting the complete cases by the inverse of the missing data probability mechanism. As a result of weighting by the inverse missing probabilities values that were observed but had a high probability of being unavailable will be given more weight, thus correcting for the bias of the complete case regression estimate. However, one typically does not have knowledge of the missing probability mechanism and as a result the method must use some estimator of these probabilities. This method has previously been used by some authors to estimate the mean of a distribution with missing data; see, for example, Hirano et al. (2003). Robins, Rotnitzky and Zhao (1994) developed a class of locally and globally adaptive semi-parametric estimators using inverse probability weighted estimating equations for the case of covariates missing at random. For a comparison between imputation methods and Horvitz-Thompson-type inverse weights see Vansteelandt (2010). In the following chapters we utilize the method of Horvitz-Thompson-type inverse weights to construct kernel regression estimators for incomplete data.

26
3 Kernel regression estimation with incomplete covariates

In the previous chapter we introduced the problem of regression estimation when the data used in the estimation is incomplete. In this chapter we study nonparametric regression, for the specific case of missing covariates. Until recently, there has been little work in the area of nonparametric regression with missing covariates. Mojirsheibani (2007) considered a multiple imputation method to estimate the regression function when part of the covariate vector could be missing at random, this is similar to the work of Cheng and Chu (1996) in that the regression estimate imputes a specific function of a missing value (and not the missing value itself). Efromovich (2011) constructed an adaptive orthogonal series estimator and derived the minimax rate of the MISE when the regression function belongs to a k-fold Sobolev class. Wand et al. (2012) developed variational Bayes algorithms for penalized spline nonparametric regression when the single covariate variable is susceptible to missingness.

Our contribution will be to construct kernel regression estimators from data in which a subset of the covariate vector may be missing at random. We propose kernel regression estimators that use Horvitz-Thompson-type inverse weights, in which the selection probabilities are estimated utilizing kernel regression or least-squares methods. Exponential upper bounds are derived for the $L_p$ norm of each estimator, these results yield almost sure convergence of the $L_p$ statistic to 0 by an application of the Borel-Cantelli lemma. As an immediate application of these results, the problem of nonparametric classification with missing covariates will be studied. Now, let us formally define the problem.

Let $(Z, Y)$ be an $\mathbb{R}^d \times \mathbb{R}$-valued random vector and let

$$m(Z) = E[Y | Z = z]$$

be the regression function. Given the data $\mathbb{D}_n = \{(Z_1, Y_1), \ldots, (Z_n, Y_n)\}$, where the $(Z_i, Y_i)$’s are independently and identically distributed (iid) random vectors with the same distribution as that of $(Z, Y)$, the regression function $m(z)$ can be estimated by the Nadaraya-Watson kernel estimate (Nadaraya (1964), Watson (1964))

$$m_n(z) = \frac{\sum_{i=1}^{n} Y_i K\left(\frac{z - Z_i}{h_n}\right)}{\sum_{i=1}^{n} K\left(\frac{z - Z_i}{h_n}\right)},$$

(3.1)

which was defined in (1.14). Here $0 < h \equiv h(n)$ is a smoothing parameter depending on $n$. 

27
and the kernel $K$ is an absolutely integrable function. The case in which the data was completely observable was already explored in Chapter 1, where the almost sure convergence of $m_n(\cdot)$ to $m(\cdot)$ was established. The result of Devroye and Krzyżak (1989) in Theorem 1.5 will be of particular importance to the results of this chapter. We are interested in extending these results to the case where a subset of the covariate vector $Z_i$ may be unavailable from $D_n$. Clearly, the estimator (3.1) is no longer available. In the following sections we present our proposed estimators.

3.1 Main Results

Let $(Z, Y)$ be an $\mathbb{R}^{d+s} \times \mathbb{R}$-valued random vector and $Z = (X', V')'$, with $X \in \mathbb{R}^d$, $V \in \mathbb{R}^s$, and $d, s \geq 1$. Our goal is to estimate the regression function $m(z) = E[Y|Z = z]$ when $V$ may be missing at random. Under the missing at random assumption the selection probability is defined as follows.

$$P(\Delta = 1|Z = z, Y = y) = P(\Delta = 1|X = x, Y = y) =: \eta^*(x, y). \quad (3.2)$$

We first consider the unrealistic case in which the selection probability $\eta^*$ is known. In this case we estimate the regression function $m(z)$ by constructing a modified kernel regression estimator from the i.i.d. sample $D_n = \{(Z_1, Y_1, \Delta_1), \cdots, (Z_n, Y_n, \Delta_n)\} = \{(X_1, V_1, Y_1, \Delta_1), \cdots, (X_n, V_n, Y_n, \Delta_n)\}$, with $\Delta_i = 1$ if $V_i$ is observable and 0 otherwise, and is defined as follows

$$\hat{m}_{\eta^*}(z) = \frac{\sum_{i=1}^n \Delta_i Y_i \mathcal{K}(z - Z_i/h_n)}{\sum_{i=1}^n \Delta_i \mathcal{K}(z - Z_i/h_n)}. \quad (3.3)$$

Observe that the estimator in (3.3) works by weighting the complete cases by the inverse of the selection probabilities; this is very much in the spirit of the classical Horvitz-Thompson estimator (Horvitz and Thompson (1952)). As for the usefulness of $\hat{m}_{\eta^*}(z)$ as an estimator of $m(z)$, observe that (3.3) can be rewritten as follows

$$\hat{m}_{\eta^*}(z) = \frac{\sum_{i=1}^n \frac{\Delta_i Y_i}{\eta^*(X_i, Y_i)} \mathcal{K}(z - Z_i/h_n)}{\sum_{i=1}^n \frac{\Delta_i}{\eta^*(X_i, Y_i)} \mathcal{K}(z - Z_i/h_n)} / \frac{\sum_{i=1}^n \mathcal{K}(z - Z_i/h_n)}{\sum_{i=1}^n \mathcal{K}(z - Z_i/h_n)},$$

which is a ratio of kernel estimators for the following ratio of two conditional expectations.
\[
\begin{align*}
E \left[ \frac{\Delta Y_{\eta^*}(X,Y)}{\eta^*(X,Y)} \right] | Z & \equiv E \left[ \frac{Y}{\eta^*(X,Y)} E \left[ \Delta | Z, Y \right] | Z \right] = E \left[ \frac{\Delta | Z, Y}{\eta^*(X,Y)} | Z \right] = E \left[ \frac{Y}{1} | Z \right] = m(Z).
\end{align*}
\]

Therefore when the missing probability mechanism \( \eta^* \) is known, (3.3) can be seen as a kernel regression estimator of the regression function \( E[Y|Z] = m(Z) \). To explore the accuracy of our estimator in (3.3) we examine the convergence of the \( L_p \) statistic to 0. We shall assume in what follows that the selected kernel \( K \) is regular as defined in (1.4.3), which we recall is

**Definition** A nonnegative kernel \( K \) is said to be regular if there are positive constants \( b > 0 \) and \( r > 0 \) for which \( K(z) \geq b I\{z \in S_{0,r}\} \) and \( \int \sup_{y \in \mathbb{R} + S_{0,r}} K(y) dz < \infty \), where \( S_{0,r} \) is the ball of radius \( r \) centered at the origin.

Many of the kernels used in practice are regular kernels, in fact the popular Gaussian kernel is a regular kernel. The regular kernel is the type of kernel used by Devroye and Krzyżak (1989). For more on regular kernels the reader is referred to Györfi et al (2002).

Before proceeding we need to state the following condition

**Condition A1.** \( \inf_{x \in \mathbb{R}^d, y \in \mathbb{R}} P(\Delta = 1 | X = x, Y = y) := \eta_0 > 0 \), where \( \eta_0 \) can be arbitrarily small.

Condition A1 is necessary so that \( V \) is observed with a nonzero probability. This is a rather standard assumption in the literature on missing data; see, for example, Wang and Qin (2010) and Cheng and Chu (1996). The following theorem employs the result of the main theorem of Devroye and Krzyżak (1989).

**Theorem 3.1** Let \( \hat{m}_{\eta^*}(z) \) be the kernel regression estimator defined in (3.3), where \( K \) is a regular kernel, and the condition A1 holds. If \( |Y| \leq M < \infty \) and \( h_n \to 0, nh_n^{d+s} \to \infty \) as \( n \to \infty \), then for every \( \epsilon > 0 \) and \( n \) large enough

\[
P \left\{ \int |\hat{m}_{\eta^*}(z) - m(z)|^p \mu(dz) > \epsilon \right\} \leq 8e^{-na_1}
\]

where \( a_1 \equiv a_1(\epsilon) = \min\left( \epsilon^2 \eta_0^{2p}/[2^{2p+7}M^{2p}(1 + \eta_0)^{p-2}(1 + c_1)], \epsilon \eta_0^{p}/[2^{p+5}M^p(1 + \eta_0)^{p-1}(1 + c_1)] \right) \), with \( c_1 \) the positive constant of Lemma B.1.
Remark. Theorem 3.1 in conjunction with the Borel-Cantelli lemma yields $E[|\hat{m}_{\eta^*}(Z) - m(Z)|^p | \mathbb{D}_n] \overset{a.s.}{\to} 0$, as $n \to \infty$.

PROOF OF THEOREM 3.1

Let $\tilde{m}_{\eta^*}$ be the kernel regression estimator of $m(z)$ and $\bar{m}_{\eta^*}$ be the kernel regression estimator of $Y \overset{a.s.}{=} 1$, defined as

$$\tilde{m}_{\eta^*}(z) = \frac{\sum_{i=1}^{n} \Delta Y_i \eta^*(X_i, Y_i) K\left(\frac{z-Z_i}{h_n}\right)}{\sum_{i=1}^{n} K\left(\frac{z-Z_i}{h_n}\right)} \quad \bar{m}_{\eta^*}(z) = \frac{\sum_{i=1}^{n} \Delta \eta^*(X_i, Y_i) K\left(\frac{z-Z_i}{h_n}\right)}{\sum_{i=1}^{n} K\left(\frac{z-Z_i}{h_n}\right)}.$$  (3.4)

Observing that $|\tilde{m}_{\eta^*}(z)/\bar{m}_{\eta^*}(z)| \leq M$, we have the following

$$|\tilde{m}_{\eta^*}(z) - m(z)| \leq \left| \frac{\tilde{m}_{\eta^*}(z)}{\bar{m}_{\eta^*}(z)} - \frac{m(z)}{1} \right| \leq M \left| \bar{m}_{\eta^*}(z) - 1 \right| + |\tilde{m}_{\eta^*}(z) - m(z)|.$$  (3.5)
As a result of (3.5), we have that for every $\epsilon > 0$, and $n$ large enough

$$P \left\{ \int |\tilde{m}_{\eta^*}(z) - m(z)|^p \mu(dz) > \epsilon \right\}$$

$$\leq P \left\{ \int |M| \tilde{m}_{\eta^*}(z) - 1| + |\tilde{m}_{\eta^*}(z) - m(z)|^p \mu(dz) > \epsilon \right\}$$

$$\leq P \left\{ \int 2^{p-1} M^p |\tilde{m}_{\eta^*}(z) - 1|^p + 2^p - 1 |\tilde{m}_{\eta^*}(z) - m(z)|^p \mu(dz) > \epsilon \right\}$$

$$= P \left\{ \int 2^{p-1} M^p \left( \frac{1}{\eta_0} + 1 \right)^{p-1} |\tilde{m}_{\eta^*}(z) - 1| \mu(dz) > \frac{\epsilon}{2} \right\}$$

$$\leq 8e^{-na_1}, \text{ by (1.5), in view of the result of Devroye and Krzyżak (1989)}.$$

**The kernel estimator defined in (3.3) requires the knowledge of the missing probability mechanism $\eta^*(X,Y) = E[\Delta|X,Y]$ (an unrealistic case). If $\eta^*$ is unknown, then it has to be replaced by some estimator in (3.3). We consider two different estimators of $\eta^*$, a kernel regression estimator and a least-squares regression type estimator.**

Here, $a_1 \equiv a_1(\epsilon) = \min \left\{ \left( \frac{2p}{\eta_0^2} \right), \left( \frac{2p+7}{M^2(1+\eta_0)^{2p-2}(1+c_1)} \right), \left( \frac{\eta_0^p}{2p+5 M^p (1+\eta_0)^{p-1}(1+c_1)} \right) \right\}$, with $c_1$ the positive constant in Lemma B.1.

\[ \square \]
3.1.1 A kernel-based estimator of the selection probability

We first consider replacing \( \eta^*(X_i, Y_i) \) in (3.3) by one of the two following estimators

\[
\hat{\eta}(X_i, Y_i) = \frac{\sum_{j=1}^n \Delta_j I \{ Y_i = Y_j \} \mathcal{H} \left( \frac{X_i - X_j}{\lambda_n} \right)}{\sum_{j=1}^n I \{ Y_i = Y_j \} \mathcal{H} \left( \frac{X_i - X_j}{\lambda_n} \right)},
\]

(3.6)

\[
\hat{\eta}_c(X_i, Y_i) = \frac{\sum_{j=1}^n \Delta_j \mathcal{H} \left( \frac{U_i - U_j}{\lambda_n} \right)}{\sum_{j=1}^n \mathcal{H} \left( \frac{U_i - U_j}{\lambda_n} \right)}, \quad \text{with } U_i = (Z_i', Y_i)', \quad (3.7)
\]

where \( \mathcal{H} : \mathbb{R}^d \to \mathbb{R}^+ \) in (3.6) and \( \mathcal{H} : \mathbb{R}^{d+1} \to \mathbb{R}^+ \) in (3.7), where \( \mathcal{H} \) is a kernel with smoothing parameter \( \lambda_n \), satisfying \( \lambda_n \to 0 \) as \( n \to \infty \), with the convention that \( 0/0 = 0 \).

The estimator in (3.6) is used if \( Y \) is a discrete random variable, otherwise one estimates the selection probability with (3.7). We derive an exponential bound for the more complicated case of discrete \( Y \) i.e., when the selection probability is estimated by the estimator described in (3.6). The continuous case follows with simpler arguments.

Now, suppose that \( Z \in \mathbb{R}^d \) and \( Y \in \mathcal{Y} \), where \( \mathcal{Y} \) is a countable subset of \( \mathbb{R} \), then our modified kernel-type regression estimator of \( m(z) = E[Y | Z = z] \) is given by

\[
\hat{m}(z) = \frac{\sum_{i=1}^n \Delta_i \mathcal{H} \left( \frac{z - Z_i}{h_n} \right)}{\sum_{i=1}^n \mathcal{H} \left( \frac{z - Z_i}{h_n} \right)}, \quad (3.8)
\]

where \( \hat{\eta} \) is as in (3.6). To study (3.8) we first state some standard regularity conditions.

**Condition A2.** The kernel \( \mathcal{H} \) in (3.6) satisfies \( \int_{\mathbb{R}^d} \mathcal{H}(u) du = 1 \) and \( \int_{\mathbb{R}^d} |u| \mathcal{H}(u) du < \infty \), for \( i = 1, \ldots, d \), where \( u = (u_1, \ldots, u_d)' \). Furthermore, the smoothing parameter \( \lambda_n \) satisfies \( \lambda_n \to 0 \) and \( n\lambda_n^d \to \infty \), as \( n \to \infty \).

**Condition A3.** The random vector \( X \) has a compactly supported probability density function, \( f(x) = \sum_{y \in \mathcal{Y}} f_y(x) P(Y = y) \), which is bounded away from zero on its support, where \( f_y(x) \) is the conditional density of \( X \) given \( Y = y \). Additionally, both \( f \) and its first-order partial derivatives are uniformly bounded on its support.
**Condition A4.** The partial derivatives \( \frac{\partial}{\partial x_i} \eta^*(x, y) \), exist for \( i = 1, \cdots, d \) and are bounded uniformly, in \( x \), on the compact support of \( f \).

Condition A2 is not restrictive since the choice of the kernel \( \mathcal{H} \) is at our discretion. Condition A3 is usually imposed in nonparametric regression in order to avoid having unstable estimates (in the tail of the p.d.f of \( X \)). Condition A4 is technical and has already been used in the literature; see, for example, Cheng and Chu (1996).

**Theorem 3.2** Let \( \hat{m}_q(z) \) be defined as in (3.8), where \( \mathcal{K} \) is a regular kernel. Suppose that \( |Y| \leq M < \infty \). If \( h_n \to 0 \), \( nh^{d+s} \to \infty \), as \( n \to \infty \), then under conditions A1, A2, A3, and A4, then for every \( \epsilon > 0 \) and every \( p \in [1, \infty) \), and \( n \) large enough,

\[
P \left\{ \int |\hat{m}_q(z) - m(z)|^p \mu(dz) > \epsilon \right\} \leq 8e^{-na_2} + 2e^{-na_3} + 8ne^{-(n-1)\lambda^d_{a_4}} + 8ne^{-(n-1)\lambda^d_{a_5}},
\]

where

\[
a_2 \equiv a_2(\epsilon) = \min^2 \left( \frac{\epsilon^2 \eta_0^{2p}}{2^{2p+7}M^{2p}(1 + \eta_0)^{2p-2}(1 + c_1)}, \frac{\epsilon \eta_0^p}{2^{2p+5}M^{p}(1 + \eta_0)^{p-1}(1 + c_1)} \right), \quad a_3 = \frac{\epsilon^2 \eta_0^{2p}}{2^{4p+8}3^{2p}c_1^2M^{2p}}, \quad a_4 \equiv a_4(\epsilon) = \frac{\epsilon^2 \eta_0^{2p+2}\psi_0^2}{2^{4p+12}3^{2p-2}c_1^2M^{2p}||\mathcal{K}||_\infty(||f||_\infty + \psi_0/12)}, \quad a_5 = \frac{\eta_0^2\psi_0^2}{2^7||\mathcal{K}||_\infty(||f||_\infty + \psi_0/12)},
\]

with \( \psi_0 \) given by \( 0 < \psi_0 := f_0 \cdot \inf_{y \in Y} P(Y = y) \) and \( c_1 \) the positive constant in Lemma B.1.

**Remark.** In view of Theorem 3.2 and the Borel-Cantelli lemma, if \( \log n/(n\lambda(d)) \to 0 \) as \( n \to \infty \), then \( E[|\hat{m}_q(z) - m(z)|^p] \stackrel{D}{\to} 0 \).

**PROOF OF THEOREM 3.2**

To prove Theorem 3.2, we first need to define a couple terms, let

\[
\hat{m}_q(z) = \frac{\sum_{i=1}^n \hat{\Delta}_Y \eta_i(X_i, Y_i) \mathcal{K} \left( \frac{z - Z_i}{h_n} \right)}{\sum_{i=1}^n \mathcal{K} \left( \frac{Z_i}{h_n} \right)}, \quad \text{and} \quad \bar{m}_q(z) = \frac{\sum_{i=1}^n \Delta_z \eta_i(X_i, Y_i) \mathcal{K} \left( \frac{Z_i}{h_n} \right)}{\sum_{i=1}^n \mathcal{K} \left( \frac{Z_i}{h_n} \right)}.
\]
Therefore, with \( \hat{m}_\eta(z) \) and \( \overline{m}_\eta(z) \) as in (3.9), and \( \hat{m}_{\eta^*}(z) \) and \( \overline{m}_{\eta^*}(z) \) as in (3.4), we have for every \( \epsilon > 0 \)

\[
P \left\{ \int |\hat{m}_\eta(z) - m(z)|^p \mu(dz) > \epsilon \right\}
\leq P \left\{ 2^{p-1} M^p \int \overline{m}_\eta(z) - 1|^p \mu(dz) > \frac{\epsilon}{2} \right\}
+ P \left\{ 2^{p-1} \int |\hat{m}_\eta(z) - m(z)|^p \mu(dz) > \frac{\epsilon}{2} \right\}
\leq P \left\{ 2^{2p-2} M^p \int |\overline{m}_\eta(z) - \overline{m}_{\eta^*}(z)|^p \mu(dz) > \frac{\epsilon}{4} \right\}
+ P \left\{ 2^{2p-2} M^p \int |\hat{m}_\eta(z) - \hat{m}_{\eta^*}(z)|^p \mu(dz) > \frac{\epsilon}{4} \right\}
+ P \left\{ 2^{2p-2} \int |\hat{m}_{\eta^*}(z) - m(z)|^p \mu(dz) > \frac{\epsilon}{4} \right\}
\leq P \left\{ 2^{2p-2} M^p \left| 1 + \frac{1}{\eta_0} \right|^{p-1} \int |\overline{m}_\eta(z) - \overline{m}_{\eta^*}(z)| \mu(dz) > \frac{\epsilon}{4} \right\}
+ P \left\{ 2^{2p-2} M^p \left| 1 + \frac{1}{\eta_0} \right|^{p-1} \int |\hat{m}_\eta(z) - 1| \mu(dz) > \frac{\epsilon}{4} \right\}
+ P \left\{ 2^{2p-2} \left| \frac{M}{\eta_0} \right|^{p-1} \int |\hat{m}_{\eta^*}(z) - \hat{m}_{\eta^*}(z)| \mu(dz) > \frac{\epsilon}{4} \right\}
+ P \left\{ 2^{2p-2} \left| \frac{M}{\eta_0} + M \right|^{p-1} \int |\hat{m}_{\eta^*}(z) - m(z)| \mu(dz) > \frac{\epsilon}{4} \right\}
:= P_{n,1} + P_{n,2} + P_{n,3} + P_{n,4}, \quad \text{(say)}.
\]
In view of Theorem 3.1, it follows that

\[ P_{n,2} + P_{n,4} \leq 8e^{-na_2}, \quad (3.10) \]

where

\[ a_2 \equiv a_2(\epsilon) = \min \left( \frac{\epsilon^2\eta_0^{2p}}{24p+7M2p(1+\eta_0)^{2p-2}(1+c_1)}, \frac{\epsilon\eta_0^p}{22p+5Mp(1+\eta_0)^{p-1}(1+c_1)} \right). \]

To deal with the term \( P_{n,1} \), first observe that

\[
\begin{align*}
|\bar{m}_0(z) - \bar{m}_{\eta^*(z)}| &= \left| \sum_{i=1}^n \frac{\Delta_i}{\eta(X_i,Y_i)} \mathcal{K} \left( \frac{z-Z_i}{h_n} \right) - \sum_{i=1}^n \frac{\Delta_i}{\eta^*(X_i,Y_i)} \mathcal{K} \left( \frac{z-Z_i}{h_n} \right) \right| \\
&= \left| \sum_{i=1}^n \Delta_i \left( \frac{1}{\eta(X_i,Y_i)} - \frac{1}{\eta^*(X_i,Y_i)} \right) \mathcal{K} \left( \frac{z-Z_i}{h_n} \right) \right| \\
&\leq \left\{ \sum_{i=1}^n \Delta_i \left( \frac{1}{\eta(X_i,Y_i)} - \frac{1}{\eta^*(X_i,Y_i)} \right) \mathcal{K} \left( \frac{z-Z_i}{h_n} \right) \right\} \\
&\quad + \left\{ \sum_{i=1}^n \Delta_i \left( \frac{1}{\eta(X_i,Y_i)} - \frac{1}{\eta^*(X_i,Y_i)} \right) \mathcal{K} \left( \frac{z-Z_i}{h_n} \right) \right\} \\
&\quad \times \left| \sum_{i=1}^n \frac{1}{\mathcal{K} \left( \frac{z-Z_i}{h_n} \right)} - \frac{1}{nE \left[ \mathcal{K} \left( \frac{z-Z}{h_n} \right) \right]} \right| \\
&:= S_{n,1}(z) + S_{n,2}(z), \quad (say).
\]
However, we can bound $\int S_{n,1}(z)\mu(dz)$ as follows

$$
\int S_{n,1}(z)\mu(dz) \leq \int \frac{\sum_{i=1}^{n} \left| \frac{1}{\eta^*(X_i,Y_i)} - \frac{1}{\hat{\eta}(X_i,Y_i)} \right|}{nE\left[ \mathcal{K}\left( \frac{z-Z}{h_n} \right) \right]} \mu(dz)
$$

$$
\leq \left( \sup_w \int \frac{\mathcal{K}\left( \frac{z-w}{h_n} \right)}{E\left[ \mathcal{K}\left( \frac{z-Z}{h_n} \right) \right]} \mu(dz) \right) \left( \frac{1}{n} \sum_{i=1}^{n} \left| \frac{1}{\eta^*(X_i,Y_i)} - \frac{1}{\hat{\eta}(X_i,Y_i)} \right| \right)
$$

$$
\leq c_1 \left( \frac{1}{n} \sum_{i=1}^{n} \left| \frac{1}{\eta^*(X_i,Y_i)} - \frac{1}{\hat{\eta}(X_i,Y_i)} \right| \right), \quad (3.11)
$$

where the last line follows from Lemma B.1 in the appendix, and the fact that $|\Delta_i| \leq 1, \forall i = 1, \ldots, n$. Here $c_1$ is the positive constant of Lemma B.1. To bound $\int S_{n,2}(z)\mu(dz)$, we have

$$
\int S_{n,2}(z)\mu(dz) \leq \max_{1 \leq i \leq n} \left| \frac{1}{\eta^*(X_i,Y_i)} - \frac{1}{\hat{\eta}(X_i,Y_i)} \right|
$$

$$
\times \int \left| \frac{\sum_{i=1}^{n} \mathcal{K}\left( \frac{z-Z}{h_n} \right)}{nE\left[ \mathcal{K}\left( \frac{z-Z}{h_n} \right) \right]} - 1 \right| \mu(dz). \quad (3.12)
$$

36
Therefore, we have

\[
P_{n,1} \leq P \left\{ 2^{2p-2}M^p \left[ \frac{1}{\eta_0} + \frac{1}{\Lambda_{i=1}^n \eta(X_i, Y_i)} \right]^{p-1} \int \left( S_{n,1}(z) + S_{n,2}(z) \right) \mu(dz) > \frac{\epsilon}{4} \right\}
\]

\[
\leq P \left\{ 2^{2p-2}M^p \left[ \frac{1}{\eta_0} + \frac{1}{\Lambda_{i=1}^n \eta(X_i, Y_i)} \right]^{p-1} \left\{ \frac{c_1}{n} \sum_{i=1}^n \left| \eta^*(X_i, Y_i) - \eta(X_i, Y_i) \right| \eta(X_i, Y_i) \right\} \right\}
\]

\[
\times \int \left| \sum_{i=1}^n \mathcal{K} \left( \frac{z_i - Z_{ih_n}}{h_n} \right) \right| - 1 \mu(dz) > \frac{\epsilon}{8} \cap \bigcap_{i=1}^n \left[ \eta(X_i, Y_i) \geq \frac{\eta_0}{2} \right] \right\}
\]

\[
+ P \left\{ \frac{1}{n} \sum_{i=1}^n \left| \eta(X_i, Y_i) - \eta^*(X_i, Y_i) \right| > \frac{\epsilon \eta_0^{p+1}}{2^{2p+2}3^{p-1}M^p c_1} \right\}
\]

\[
+ P \left\{ \int \left| \sum_{i=1}^n \mathcal{K} \left( \frac{z_i - Z_{ih_n}}{h_n} \right) \right| - 1 \mu(dz) > \frac{\epsilon \eta_0^p}{2^{2p+1}3^p M^p} \right\}
\]

\[
+ P \left\{ \bigcup_{i=1}^n \eta(X_i, Y_i) < \frac{\eta_0}{2} \right\}.
\]

(3.13)
The bound for the term $P_{n,3}$ follows in a similar fashion as in $P_{n,1}$, we first note that

\[
|\hat{m}_\eta(z) - \bar{m}_{\eta^*}(z)| = \left| \sum_{i=1}^{n} \Delta_i Y_i \left( \frac{1}{\hat{\eta}(X_i, Y_i)} - \frac{1}{\eta^*(X_i, Y_i)} \right) \mathcal{K} \left( \frac{z - Z_i}{h_n} \right) \right| \leq \left| \sum_{i=1}^{n} \Delta_i Y_i \left( \frac{1}{\hat{\eta}(X_i, Y_i)} - \frac{1}{\eta^*(X_i, Y_i)} \right) \mathcal{K} \left( \frac{z - Z_i}{h_n} \right) \right| + n E \left[ \mathcal{K} \left( \frac{z - Z}{h_n} \right) \right]
\]

\[
\leq \left| \sum_{i=1}^{n} \Delta_i Y_i \left( \frac{1}{\hat{\eta}(X_i, Y_i)} - \frac{1}{\eta^*(X_i, Y_i)} \right) \mathcal{K} \left( \frac{z - Z_i}{h_n} \right) \right| + \left| \sum_{i=1}^{n} \Delta_i Y_i \left( \frac{1}{\hat{\eta}(X_i, Y_i)} - \frac{1}{\eta^*(X_i, Y_i)} \right) \mathcal{K} \left( \frac{z - Z_i}{h_n} \right) \right| - \left| \sum_{i=1}^{n} \Delta_i Y_i \left( \frac{1}{\hat{\eta}(X_i, Y_i)} - \frac{1}{\eta^*(X_i, Y_i)} \right) \mathcal{K} \left( \frac{z - Z_i}{h_n} \right) \right| \times \left( \frac{1}{\sum_{i=1}^{n} \mathcal{K} \left( \frac{z - Z_i}{h_n} \right)} - \frac{1}{n E \left[ \mathcal{K} \left( \frac{z - Z}{h_n} \right) \right]} \right)
\]

\[
:= S_{n,1}'(z) + S_{n,2}'(z) \quad \text{(say)}.
\]

Similar to that of $\int S_{n,1}(z)\mu(dz)$, the bound for $\int S_{n,1}'(z)\mu(dz)$ is as follows

\[
\int S_{n,1}'(z)\mu(dz) \leq M \cdot c_1 \left( \frac{1}{n} \sum_{i=1}^{n} \left| \frac{1}{\hat{\eta}(X_i, Y_i)} - \frac{1}{\eta^*(X_i, Y_i)} \right| \right).
\]

The above line follows as a result of Lemma B.1 in the appendix, and the fact that $|\Delta_i Y_i| \leq M, \forall i = 1, \ldots, n, (c_1 \text{ as in Lemma B.1})$. For the term $\int S_{n,2}'(z)\mu(dz)$, we have

\[
\int S_{n,2}'(z)\mu(dz) \leq M \cdot \max_{1 \leq i \leq n} \left| \frac{1}{\hat{\eta}(X_i, Y_i)} - \frac{1}{\eta^*(X_i, Y_i)} \right| \int \left| \sum_{i=1}^{n} \mathcal{K} \left( \frac{z - Z_i}{h_n} \right) \right| - \left| \sum_{i=1}^{n} \mathcal{K} \left( \frac{z - Z}{h_n} \right) \right| \mu(dz).
\]
Therefore, $P_{n,3}$ can be bounded as follows

$$P_{n,3} \leq P \left\{ 2^{2p-2} \left| \frac{M}{\eta_0} + \frac{M}{\wedge_{i=1}^{n} \hat{\eta}(X_i, Y_i)} \right|^{p-1} \int (S_{n,1}'(z) + S_{n,2}'(z)) \mu(dz) > \frac{\epsilon}{4} \right\}$$

$$\leq P \left\{ 2^{2p-2} \left| \frac{M}{\eta_0} + \frac{M}{\wedge_{i=1}^{n} \hat{\eta}(X_i, Y_i)} \right|^{p-1} \sum_{i=1}^{n} \frac{c_1}{n} \left| \frac{\eta^*(X_i, Y_i) - \hat{\eta}(X_i, Y_i)}{\eta^*(X_i, Y_i) \hat{\eta}(X_i, Y_i)} \right| > \frac{\epsilon}{8} \right\}$$

$$\cap \bigcap_{i=1}^{n} \left[ \hat{\eta}(X_i, Y_i) \geq \frac{\eta_0}{2} \right]$$

$$+ P \left\{ 2^{2p-2} \left| \frac{M}{\eta_0} + \frac{M}{\wedge_{i=1}^{n} \hat{\eta}(X_i, Y_i)} \right|^{p-1} \max_{1 \leq i \leq n} \left| \frac{1}{\eta^*(X_i, Y_i)} - \frac{1}{\hat{\eta}(X_i, Y_i)} \right| \right. \left. \times \int \left| \sum_{i=1}^{n} \mathcal{K} \left( \frac{z - Z_{h_n}}{h_n} \right) - 1 \right| \mu(dz) \right| > \frac{\epsilon}{8} \bigcap \bigcap_{i=1}^{n} \left[ \hat{\eta}(X_i, Y_i) \geq \frac{\eta_0}{2} \right] \right\}$$

$$+ P \left\{ \bigcup_{i=1}^{n} \hat{\eta}(X_i, Y_i) < \frac{\eta_0}{2} \right\}$$

$$\leq P \left\{ \frac{1}{n} \sum_{i=1}^{n} \left| \hat{\eta}(X_i, Y_i) - \eta^*(X_i, Y_i) \right| > \frac{\epsilon \eta_0^{p+1}}{2^{2p+2}3^{p-1}M^{c_1}} \right\}$$

$$+ P \left\{ \int \left| \sum_{i=1}^{n} \mathcal{K} \left( \frac{z - Z_{h_n}}{h_n} \right) - 1 \right| \mu(dz) > \frac{\epsilon \eta_0^{p}}{2^{2p+1}3^{p}M^{p}} \right\}$$

$$+ P \left\{ \bigcup_{i=1}^{n} \hat{\eta}(X_i, Y_i) < \frac{\eta_0}{2} \right\} .$$

(3.16)
Therefore, in view of (3.13), (3.14), (3.15), (3.16), (3.17), and (3.18), we find
\[
P_{n,1} + P_{n,3} \leq 2 \sum_{i=1}^{n} P \left\{ |\hat{\eta}(X_i, Y_i) - \eta^*(X_i, Y_i)| > \frac{\epsilon \eta_0^{p+1}}{2^{2p+23p-1} M c_1} \right\}
+ 2 \sum_{i=1}^{n} \int \frac{\sum_{i=1}^{n} \mathcal{K} \left( \frac{z-Z_i}{h_n} \right)}{n E \left[ \mathcal{K} \left( \frac{z-Z}{h_n} \right) \right]} - 1 \left| \mu(dz) > \frac{\epsilon \eta_0^p}{2^{2p+13p} M^p} \right\}
+ 2 \sum_{i=1}^{n} P \left\{ \hat{\eta}(X_i, Y_i) < \frac{\eta_0}{2} \right\}
:= T_{n,1} + T_{n,2} + T_{n,3}, \quad \text{(say).} \tag{3.19}
\]

First note that upon taking \(Y_i \overset{\text{a.s.}}{=} 1\) and \(m(z) = 1\) in Lemma B.2 for \(i = 1, \cdots, n\), we find
\[
T_{n,2} \leq 2e^{-na_3}, \tag{3.20}
\]
where \(a_3 \equiv a_3(\epsilon) = (\epsilon^2 \eta_0^{2p})/(2^{4p+8} 3^{2p} c_1^2 M^{2p})\). To deal with the term \(T_{n,1}\), define the following terms
\[
\Phi (X_i, Y_i) = \eta^*(X_i, Y_i) f(X_i) P(Y = Y_i | Y_i)
\]
\[
\hat{\Phi} (X_i, Y_i) = (n-1)^{-1} \lambda_n^{-d} \sum_{j=1 \neq i}^{n} \Delta_j I \{ Y_i = Y_j \} \mathcal{H} \left( \frac{X_i - X_j}{\lambda_n} \right)
\]
\[
\Psi (X_i, Y_i) = f(X_i) P(Y = Y_i | Y_i)
\]
\[
\hat{\Psi} (X_i, Y_i) = (n-1)^{-1} \lambda_n^{-d} \sum_{j=1 \neq i}^{n} I \{ Y_i = Y_j \} \mathcal{H} \left( \frac{X_i - X_j}{\lambda_n} \right).
\]

Notice that
\[
\frac{\hat{\Phi} (X_i, Y_i)}{\hat{\Psi} (X_i, Y_i)} = \tilde{\eta}(X_i, Y_i) \quad \text{and} \quad \frac{\Phi (X_i, Y_i)}{\Psi (X_i, Y_i)} = \eta^*(X_i, Y_i),
\]
as well as that \(|\hat{\Phi} (X_i, Y_i)/\hat{\Psi} (X_i, Y_i)| \leq 1\). Where \(\lambda_n\) and \(\mathcal{H}\) are as in (3.6). Using the above definitions,
observe that

\[
|\hat{\eta}(X_i, Y_i) - \eta^*(X_i, Y_i)| = \left| \frac{\hat{\Phi}(X_i, Y_i)}{\Psi(X_i, Y_i)} - \frac{\Phi(X_i, Y_i)}{\Psi(X_i, Y_i)} \right| \leq \frac{|\hat{\Psi}(X_i, Y_i) - \Psi(X_i, Y_i)|}{\Psi(X_i, Y_i)} + \frac{|\hat{\Phi}(X_i, Y_i) - \Phi(X_i, Y_i)|}{\Psi(X_i, Y_i)}.
\]

Now, let \(0 < f_0 := \inf_{x \in \mathbb{R}^d} f(x)\) (see condition A3), in view of this we have

\[
\Psi(X_i, Y_i) = f(X_i) P(Y = Y_i | Y_i) \geq f_0 \inf_{y \in \mathcal{Y}} P(Y = y) =: \psi_0 > 0. \tag{3.21}
\]

Therefore,

\[
T_{n,1} \leq 2 \sum_{i=1}^{n} P \left\{ \left| \frac{\hat{\Psi}(X_i, Y_i) - \Psi(X_i, Y_i)}{\Psi(X_i, Y_i)} \right| > t \right\} \tag{3.22}
\]

\[+ 2 \sum_{i=1}^{n} P \left\{ \left| \frac{\hat{\Phi}(X_i, Y_i) - \Phi(X_i, Y_i)}{\Psi(X_i, Y_i)} \right| > t \right\}, \tag{3.23}
\]

where \(t = \frac{c_0^p + \psi_0}{2^{p+1} 3^{p-1} M p c_1} > 0\), with \(\psi_0\) as in (3.21). Now, put

\[
\Gamma_j(X_i, Y_i) = \lambda_n^{-d} \left\{ \Delta_j I\{Y_i = Y_j\} \mathcal{H} \left( \frac{X_i - X_j}{\lambda_n} \right) - E \left[ \Delta_j I\{Y_i = Y_j\} \mathcal{H} \left( \frac{X_i - X_j}{\lambda_n} \right) \bigg| X_i, Y_i \right] \right\},
\]
for \( i = 1, \ldots, n, j = 1 \ldots, n,j \neq i, \) and observe that

\[
P \left\{ \left| \hat{\Phi} (X_i, Y_i) - \Phi (X_i, Y_i) \right| > t \right\}
\]

\[
= P \left\{ \left| \hat{\Phi} (X_i, Y_i) \pm E \left[ \hat{\Phi} (X_i, Y_i) \right] X_i, Y_i \right| - \Phi (X_i, Y_i) \right| > t \right\}
\]

\[
\leq P \left\{ \left| \hat{\Phi} (X_i, Y_i) - E \left[ \hat{\Phi} (X_i, Y_i) \right] X_i, Y_i \right| + \frac{t}{2} > t \right\}
\]

(for large enough \( n, \) by Lemma B.3 in the appendix)

\[
= E \left[ P \left\{ \left| \hat{\Phi} (X_i, Y_i) - E \left[ \hat{\Phi} (X_i, Y_i) \right] X_i, Y_i \right| > \frac{t}{2} X_i, Y_i \right\} \right]
\]

\[
= E \left[ P \left\{ (n - 1)^{-1} \sum_{j=1 \neq i}^{n} \Gamma_j (X_i, Y_i) \right| > \frac{t}{2} X_i, Y_i \right\} \right],
\]

But, conditional on \((X_i, Y_i),\) the terms \( \Gamma_j (X_i, Y_i), j = 1, \ldots, n,j \neq i,\) are independent zero-mean random variables, bounded by \(-\lambda_n^{-d} \|H\|_{\infty} + \lambda_n^{-d} \|\mathcal{H}\|_{\infty}.\) Also, observe that \( \text{Var}[\Gamma_j (X_i, Y_i) | X_i, Y_i] = E[\Gamma_j (X_i, Y_i) | X_i, Y_i] \leq \lambda_n^{-d} \|H\|_{\infty} \|f\|_{\infty}.\) Therefore by an application of Bernstein’s Inequality (Bernstein 1946), we have

\[
P \left\{ (n - 1)^{-1} \sum_{j=1 \neq i}^{n} \left| \Gamma_j (X_i, Y_i) \right| > \frac{t}{2} X_i, Y_i \right\}
\]

\[
\leq 2 \exp \left\{ \frac{-\left( n - 1 \right) \left( \frac{t}{2} \right)^2}{2 \left[ \lambda_n^{-d} \|H\|_{\infty} \|f\|_{\infty} + \frac{1}{3} \lambda_n^{-d} \|\mathcal{H}\|_{\infty} \frac{t}{2} \right]} \right\}
\]

\[
\leq 2 \exp \left\{ \frac{-\left( n - 1 \right) \lambda_n^{-d} \eta_0^{2p+1} \psi_0^{2} \epsilon^{2}}{24p+12 \lambda_n^{2} \psi_0^{2} \epsilon^{2} \|H\|_{\infty} \|f\|_{\infty} + \psi_0^{12}} \right\},
\]

for \( n \) large enough. The last line follows by the fact that in bounding \( P \{ |\tilde{\eta}(X_i, Y_i) - \eta^*(X_i, Y_i)| > \frac{\epsilon \eta_0^{2p+1} \psi_0^{2} \epsilon^{2}}{24p+12 \lambda_n^{2} \psi_0^{2} \epsilon^{2} \|H\|_{\infty} \|f\|_{\infty} + \psi_0^{12}} \},\) we only need to consider \( \epsilon \in (0, \frac{24p+12 \lambda_n^{2} \psi_0^{2} \epsilon^{2} \|H\|_{\infty} \|f\|_{\infty} + \psi_0^{12}}{\eta_0^{2p+1} \psi_0^{2}}),\) because \( |\tilde{\eta}(X_i, Y_i) - \eta^*(X_i, Y_i)| \leq 1.\) Similarly, since \( \tilde{\Psi}(X_i, Y_i) \) is a special case of \( \hat{\Phi}(X_i, Y_i),\) with \( \Delta_j = 1 \) for \( j = 1, \ldots, n \) (compare (3.22) with (3.23)), one finds

\[
P \left\{ \left| \hat{\Psi}(X_i, Y_i) - \Psi(X_i, Y_i) \right| > t \right\} \leq 2 e^{\frac{-\left( n - 1 \right) \lambda_n^{-d} \eta_0^{2p+1} \psi_0^{2} \epsilon^{2}}{24p+12 \lambda_n^{2} \psi_0^{2} \epsilon^{2} \|H\|_{\infty} \|f\|_{\infty} + \psi_0^{12}}},
\]
for $n$ large enough. Combining the above two bounds, we obtain

$$T_{n,1} \leq 8ne^{-(n-1)\lambda_n^d a_4},$$

(3.24)

where $a_4 \equiv a_4(\epsilon) = \frac{\epsilon^2 \eta_0^{2p+2} \psi_0^2}{2^{4p+12} \cdot 3^{2p-2} c_f M^{2p} \|H\|_\infty \|f\|_\infty + \psi_0/12}$.

Finally, to deal with the remaining term $T_{n,3}$, notice that

$$P \left\{ \hat{\eta}(X_i, Y_i) < \frac{\eta_0}{2} \right\} = P \left\{ \hat{\eta}(X_i, Y_i) - \eta^*(X_i, Y_i) < \frac{\eta_0}{2} - \eta^*(X_i, Y_i) \right\}$$

$$\leq P \left\{ \hat{\eta}(X_i, Y_i) - \eta^*(X_i, Y_i) < \frac{-\eta_0}{2} \right\}$$

$$\leq P \left\{ |\hat{\eta}(X_i, Y_i) - \eta^*(X_i, Y_i)| \geq \frac{\eta_0}{2} \right\}.$$

Therefore, employing the arguments that lead to the derivation of the bound on $T_{n,1}$, we have

$$T_{n,3} \leq 8ne^{-(n-1)\lambda_n^d a_5},$$

(3.25)

where $a_5 = \frac{\eta_0^2 \psi_0^2}{2^7 \|H\|_\infty \|f\|_\infty + \psi_0/12}$, which does not depend on $\epsilon$ or $n$. Therefore, in view of (3.19), (3.20), (3.24), and (3.25), one finds

$$P_{n,1} + P_{n,3} \leq 2e^{-\frac{n \eta_0^2 \psi_0^2}{2^{4p+12} \cdot 3^{2p-2} c_f M^{2p} \|H\|_\infty \|f\|_\infty + \psi_0/12}} + 8ne^{-\frac{\eta_0^2 \psi_0^2}{2^7 \|H\|_\infty \|f\|_\infty + \psi_0/12}}.$$

This completes the proof of Theorem 3.2.

3.1.2 A least-squares based estimator of the selection probability

Suppose that it is known in advance that the regression function belongs to a given (known) class of functions. In this case the least-squares estimator is an alternative to the kernel estimator of $\eta^*(x, y)$. This works as follows. Let $\eta^*$ belong to a known class of functions $\mathcal{M}$ of the form $\eta : \mathbb{R}^d \times \mathbb{R} \rightarrow [\eta_0, 1]$, where $\eta_0 = \inf_x P(\Delta = 1|X = x, Y = y) > 0$, as
described in assumption A1. The least-squares estimator of the function $\eta^*$ is

$$\hat{\eta}_{ls}(X_i, Y_i) = \operatorname{argmin}_{\eta \in M} \frac{1}{n} \sum_{i=1}^{n} (\Delta_i - \eta(X_i, Y_i))^2. \quad (3.26)$$

Upon replacing $\eta^*(X_i, Y_i)$ in (3.3) with the least-squares estimator in (3.26), we find the revised regression estimator

$$\hat{m}_{\hat{\eta}_{ls}}(z) = \frac{\sum_{i=1}^{n} \frac{\Delta Y_i}{\hat{\eta}_{ls}(X_i, Y_i)} K \left( \frac{z - Z_i}{h_n} \right)}{\sum_{i=1}^{n} \frac{\Delta_i}{\hat{\eta}_{ls}(X_i, Y_i)} K \left( \frac{z - Z_i}{h_n} \right)}. \quad (3.27)$$

To study the performance of $\hat{m}_{\hat{\eta}_{ls}}(z)$, we employ results from the empirical process theory (see, for example, van der Vaart and Wellner (1996) p.83, Pollard (1984), p.25, or Györfi, et al. (2002) p.134). We say $M$ is totally bounded with respect to the $L_1$ empirical norm if for every $\epsilon > 0$, there exists a subclass of functions $M_{\epsilon} = \{M_1, \cdots, M_N\}$ such that for every $\eta \in M$ there exists a $\eta^\dagger \in M_{\epsilon}$ with the property that for the fixed points $(x_1, y_1), \cdots, (x_n, y_n)$ we have $\frac{1}{n} \sum_{i=1}^{n} |\eta(x_i, y_i) - \eta^\dagger(x_i, y_i)| < \epsilon$. The sub class $M_{\epsilon}$ is called an $\epsilon$-cover of $M$. The cardinality of the smallest such cover is called the $\epsilon$-covering number of $M$ and is denoted by $N_1(\epsilon, M, D_n)$.

**Theorem 3.3** Let $\hat{m}_{\hat{\eta}_{ls}}(z)$ be as in (3.27), where $\hat{\eta}_{ls}$ is the least-squares estimator of $\eta^*$ described in (3.26), with $\eta^*$ a member of a known class of functions $M$, where $M$ is totally bounded with respect to the empirical $L_1$ norm. Also, suppose that condition A1 holds and let $K$ in (3.27) be a regular kernel with smoothing parameter $h_n$. Then, provided that $h_n \to 0$, $nh_n^{d+s} \to \infty$, as $n \to \infty$, and $|Y| \leq M < \infty$, we have for every $\epsilon > 0$ and every $p \in [1, \infty)$, and $n$ large enough

$$P \left\{ \int |\hat{m}_{\hat{\eta}_{ls}}(z) - m(z)|^p \mu(dz) > \epsilon \right\} \leq 8e^{-nb_1} + 2e^{-nb_2} + 16E[N_1(b_3, M, D_n)] e^{-nb_4} + 16E[N_1(b_5, M, D_n)] e^{-nb_6}.$$

Here, the constants $b_1, \cdots, b_6$ (note no attempt was made to find the optimal constants)
are defined as

\[ b_1 \equiv b_1(\epsilon) = \min^2 \left( \frac{e^{2} \eta_0^2}{2^{4p+7} M^{2p}(1 + \eta_0)^{2p-2}(1 + c_1)}, \frac{e \eta_0^p}{2^{2p+5} M^{p}(1 + \eta_0)^{p-1}(1 + c_1)} \right), \]

\[ b_2 \equiv b_2(\epsilon) = \frac{e^2 \eta_0^{2p}}{2^{6p+8} M^{2p}c_1^2}, \quad b_3 \equiv b_3(\epsilon) = \frac{e \eta_0^{p+1}}{2^{3p+3} M^p c_1^2}, \quad b_4 \equiv b_4(\epsilon) = \frac{e^2 \eta_0^{2p+2}}{2^{6p+7} M^{2p}c_1^2}, \]

\[ b_5 \equiv b_5(\epsilon) = \frac{e^2 \eta_0^{2p+2}}{2^{6p+4} M^{2p}c_1^2}, \quad b_6 \equiv b_6(\epsilon) = \frac{e^4 \eta_0^{4p+4}}{2^{12p+6} M^{4p}c_1^4}. \]

**Remark.** Combining an application of the Borel-Cantelli lemma with the conclusion of Theorem 3.3, yields \( E[|\hat{m}_{\eta}(z) - m(z)|^p|D_n]| \xrightarrow{a.s} 0 \), provided that

\[ \log(E[|N_1(b_3 \wedge b_5, \mathcal{M}, D_n)|])/n \to 0 \quad \text{as} \quad n \to \infty. \]

**PROOF OF THEOREM 3.3**

First note that, as in the proof of Theorem 2, for every \( \epsilon > 0 \),

\[ \hat{m}_{\eta}(z) = \frac{\sum_{i=1}^{n} \Delta Y_i}{\eta(z_h, \mathcal{M}, \overline{D}_n)} K \left( \frac{z-Z_i}{h_n} \right), \quad \text{and} \quad \overline{m}_{\eta}(z) = \frac{\sum_{i=1}^{n} \frac{\Delta Y_i}{\eta(z_h, \mathcal{M}, \overline{D}_n)}}{\sum_{i=1}^{n} K \left( \frac{z-Z_i}{h_n} \right)}. \]

Then, upon observing that \(|\hat{m}_{\eta}(z)/\overline{m}_{\eta}(z)| \leq M\), one obtains

\[ |\hat{m}_{\eta}(z) - m(z)| = \frac{\hat{m}_{\eta}(z) - m(z)}{\overline{m}_{\eta}(z)} \leq M \frac{|\overline{m}_{\eta}(z) - 1| + |\hat{m}_{\eta}(z) - m(z)|}{\overline{m}_{\eta}(z)} \]
Therefore, with $\tilde{m}_{\eta s} (z)$ and $\overline{m}_{\eta s} (z)$ as in (3.28), and $\tilde{m}_{\eta^*} (z)$ and $\overline{m}_{\eta^*} (z)$ as in (3.4), we have for every $\epsilon > 0$

$$P \left\{ \int |\tilde{m}_{\eta s} (z) - m(z)|^p \mu(dz) > \epsilon \right\}$$

$$\leq P \left\{ 2^{p-1} M^p \int |\overline{m}_{\eta s} (z) - 1|^p \mu(dz) > \frac{\epsilon}{2} \right\}$$

$$+ P \left\{ 2^{p-1} \int |\tilde{m}_{\eta s} (z) - m(z)|^p \mu(dz) > \frac{\epsilon}{2} \right\}$$

$$\leq P \left\{ 2^{2p-2} M^p \int |\overline{m}_{\eta s} (z) - \overline{m}_{\eta^*} (z)|^p \mu(dz) > \frac{\epsilon}{4} \right\}$$

$$+ P \left\{ 2^{2p-2} M^p \int |\tilde{m}_{\eta s} (z) - \tilde{m}_{\eta^*} (z)|^p \mu(dz) > \frac{\epsilon}{4} \right\}$$

$$+ P \left\{ 2^{2p-2} \int |\tilde{m}_{\eta^*} (z) - m(z)|^p \mu(dz) > \frac{\epsilon}{4} \right\}$$

$$\leq P \left\{ 2^{2p-2} M^p \left| \frac{1}{\eta_0} + \frac{1}{\eta_0} \right|^{p-1} \int |\overline{m}_{\eta s} (z) - \overline{m}_{\eta^*} (z)| \mu(dz) > \frac{\epsilon}{4} \right\}$$

$$+ P \left\{ 2^{2p-2} M^p \left| \frac{1}{\eta_0} + 1 \right|^{p-1} \int |\overline{m}_{\eta^*} (z) - 1| \mu(dz) > \frac{\epsilon}{4} \right\}$$

$$+ P \left\{ 2^{2p-2} \left| \frac{M}{\eta_0} + \frac{M}{\eta_0} \right|^{p-1} \int |\tilde{m}_{\eta s} (z) - \tilde{m}_{\eta^*} (z)| \mu(dz) > \frac{\epsilon}{4} \right\}$$

$$+ P \left\{ 2^{2p-2} \left| \frac{M}{\eta_0} + M \right|^{p-1} \int |\tilde{m}_{\eta^*} (z) - m(z)| \mu(dz) > \frac{\epsilon}{4} \right\}$$

$$:= Q_{n,1} + Q_{n,2} + Q_{n,3} + Q_{n,4}, \text{ (say).}$$

As a result of Theorem 3.1, we have

$$Q_{n,2} + Q_{n,4} \leq 8 e^{-n b_1}, \quad (3.29)$$

46
where

\[ b_1 \equiv b_1(\epsilon) = \min\left( \frac{\epsilon^2 \eta_0^{2p}}{2^{4p+7} M^{2p} (1 + \eta_0)^{2p-2} (1 + c_1)}, \frac{\epsilon \eta_0^p}{2^{2p+5} M^p (1 + \eta_0)^{p-1} (1 + c_1)} \right). \]

To bound \( Q_{n,1} \), first we note that

\[
|\overline{m}_{\hat{\eta}_s}(z) - \overline{m}_{\eta^*}(z)| \\
\leq \left| \sum_{i=1}^{n} \Delta_i \left( \frac{1}{\hat{\eta}_s(X_i, Y_i)} - \frac{1}{\eta^*(X_i, Y_i)} \right) K\left( \frac{z-Z_i}{h_n} \right) \right| nE\left[ K\left( \frac{z-Z}{h_n} \right) \right] \\
+ \left| \sum_{i=1}^{n} \Delta_i \left( \frac{1}{\hat{\eta}_s(X_i, Y_i)} - \frac{1}{\eta^*(X_i, Y_i)} \right) K\left( \frac{z-Z_i}{h_n} \right) \right| \\
\times \left( \frac{1}{\sum_{i=1}^{n} K\left( \frac{z-Z_i}{h_n} \right)} - \frac{1}{nE\left[ K\left( \frac{z-Z}{h_n} \right) \right]} \right) \\
:= \pi_{n,1}(z) + \pi_{n,2}(z), \quad \text{(say)}.
\]

But, using the arguments that lead to (3.11), we have

\[
\int \pi_{n,1}(z) \mu(dz) \leq c_1 \left( \frac{1}{n} \sum_{i=1}^{n} \left| \frac{1}{\hat{\eta}_s(X_i, Y_i)} - \frac{1}{\eta^*(X_i, Y_i)} \right| \right).
\] \quad (3.30)

The term \( \int \pi_{n,2}(z) \mu(dz) \) can be bounded similar to that of (3.12),

\[
\int \pi_{n,2}(z) \mu(dz) \leq \max_{1 \leq i \leq n} \left| \frac{1}{\hat{\eta}_s(X_i, Y_i)} - \frac{1}{\eta^*(X_i, Y_i)} \right| \\
\times \int \left| \sum_{i=1}^{n} K\left( \frac{z-Z_i}{h_n} \right) \right| \left( nE\left[ K\left( \frac{z-Z}{h_n} \right) \right] - 1 \right) \mu(dz). \quad (3.31)
\]
Now, put $\epsilon_1 = \epsilon \eta_0^{p-1}/[2^{2p-1}M^p]$, then combining (3.30) and (3.31), yields

$$Q_{n,1} \leq P \left\{ \int (\pi_{n,1}(z) + \pi_{n,2}(z)) \mu(dz) > \epsilon_1 \right\}$$

$$\leq P \left\{ \frac{c_1}{n} \sum_{i=1}^{n} \left| \frac{1}{\eta^*(X_i, Y_i)} - \frac{1}{\hat{\eta}_h(X_i, Y_i)} \right| > \frac{\epsilon_1}{2} \right\}$$

$$+ P \left\{ \max_{1 \leq i \leq n} \left| \frac{1}{\eta^*(X_i, Y_i)} - \frac{1}{\hat{\eta}_h(X_i, Y_i)} \right| \int \left| \frac{1}{nE [K(z - Z_i h_n)]} - 1 \right| \mu(dz) > \frac{\epsilon_1}{2} \right\}$$

$$\leq P \left\{ \frac{1}{n} \sum_{i=1}^{n} \left| \frac{1}{\eta^*(X_i, Y_i)} - \frac{1}{\hat{\eta}_h(X_i, Y_i)} \right| > \frac{\epsilon_1 \eta_0^2}{c_1} \right\}$$

$$+ P \left\{ \int \sum_{i=1}^{n} \frac{1}{nE [K(z - Z_i h_n)]} - 1 \right| \mu(dz) > \frac{\epsilon_1 \eta_0}{4} \right\}$$

(3.32)

(3.33)

To bound $Q_{n,3}$ we use the arguments for the bound on $Q_{n,1}$, first observing that

$$|\tilde{m}_{\eta^*}(z) - \tilde{m}_{\eta^*}(z)|$$

$$\leq \left| \frac{\sum_{i=1}^{n} \Delta_i Y_i \left( \frac{1}{\hat{\eta}_h(X_i, Y_i)} - \frac{1}{\eta^*(X_i, Y_i)} \right) K(z - Z_i h_n)}{\sum_{i=1}^{n} K(z - Z_i h_n)} \right|$$

$$+ \left| \frac{\sum_{i=1}^{n} \Delta_i Y_i \left( \frac{1}{\hat{\eta}_h(X_i, Y_i)} - \frac{1}{\eta^*(X_i, Y_i)} \right) K(z - Z_i h_n)}{\sum_{i=1}^{n} K(z - Z_i h_n)} \right|$$

$$\times \left( \frac{1}{\sum_{i=1}^{n} K(z - Z_i h_n)} - 1 \right)$$

$$:= \pi'_{n,1}(z) + \pi'_{n,2}(z), \text{ (say)}.$$

But, once again the arguments that lead to (3.11), yield

$$\int \pi'_{n,1}(z) \mu(dz) \leq M c_1 \left( \frac{1}{n} \sum_{i=1}^{n} \left| \frac{1}{\hat{\eta}_h(X_i, Y_i)} - \frac{1}{\eta^*(X_i, Y_i)} \right| \right).$$

(3.34)
The bound for the term \( \int \pi'_{n,2}(z) \mu(dz) \) follows as before in (3.12)

\[
\int \pi'_{n,2}(z) \mu(dz) \leq M \max_{1 \leq i \leq n} \left| \frac{1}{\tilde{\eta}_h(X_i, Y_i)} - \frac{1}{\eta^*(X_i, Y_i)} \right| \times \int \frac{\sum_{i=1}^n K \left( \frac{z-Z_i}{h_n} \right)}{nE \left[ K \left( \frac{z-Z}{h_n} \right) \right]} - 1 \mu(dz). \tag{3.35}
\]

Therefore, with \( \epsilon_1 = \epsilon_0^{p-1}/[2^{3p-1}M^p] \), the bound for \( Q_{n,3} \) follows similarly to that of \( Q_{n,1} \)

\[
Q_{n,3} \leq P \left\{ \int (\pi'_{n,1}(z) + \pi'_{n,2}(z)) \mu(dz) > M\epsilon_1 \right\}
\leq P \left\{ \frac{1}{n} \sum_{i=1}^n |\tilde{\eta}_h(X_i, Y_i) - \eta^*(X_i, Y_i)| > \frac{\epsilon_1\eta_0^2}{c_1} \right\} \tag{3.36}
\]

\[
+ P \left\{ \int \left| \frac{\sum_{i=1}^n K \left( \frac{z-Z_i}{h_n} \right)}{nE \left[ K \left( \frac{z-Z}{h_n} \right) \right]} - 1 \right| \mu(dz) > \frac{\epsilon_1\eta_0}{4} \right\}. \tag{3.37}
\]

Thus, by (3.32), (3.33), (3.36), and (3.37), it follows that

\[
Q_{n,1} + Q_{n,3} \leq 2P \left\{ \frac{1}{n} \sum_{i=1}^n |\tilde{\eta}_h(X_i, Y_i) - \eta^*(X_i, Y_i)| > \frac{\epsilon_1\eta_0^2}{c_1} \right\}
\]

\[
+ 2P \left\{ \int \left| \frac{\sum_{i=1}^n K \left( \frac{z-Z_i}{h_n} \right)}{nE \left[ K \left( \frac{z-Z}{h_n} \right) \right]} - 1 \right| \mu(dz) > \frac{\epsilon_1\eta_0}{4} \right\}
:= U_{n,1} + U_{n,2}, \quad (say). \tag{3.38}
\]

Observe that, upon taking \( Y_i \overset{a.s.}{=} 1 \), and \( m(z) = 1 \) in lemma B.2 for \( i = 1, \ldots, n \), one obtains

\[
U_{n,2} \leq 2e^{-nb_2}, \tag{3.39}
\]

where \( b_2 \equiv b_2(\epsilon) = \frac{\epsilon^2_0\eta_0^{2p}}{2^{5p/2+8}M^{2p}c_1} \).
To deal with the term $U_{n,1}$, observe that

$$U_{n,1} \leq 2P \left\{ \left( \frac{1}{n} \sum_{i=1}^{n} |\hat{\eta}_{ls}(X_i, Y_i) - \eta^*(X_i, Y_i)| \right. \right.$$

$$\left. - E \left[ |\hat{\eta}_{ls}(X, Y) - \eta^*(X, Y)| \right| D_n \right] > \frac{\epsilon_1 \eta_0^2}{2c_1} \right\} \quad (3.40)$$

$$+ 2P \left\{ E \left[ |\hat{\eta}_{ls}(X, Y) - \eta^*(X, Y)| \right| D_n \right] > \frac{\epsilon_1 \eta_0^2}{2c_1} \right\} . \quad (3.41)$$

To handle the term (3.40), first put $\epsilon_2 = \frac{\epsilon_1 \eta_0^2}{2c_1} > 0$ and let the class of functions $\mathcal{G}$ be defined as $\mathcal{G} = \{ g_\eta(x, y) = |\eta(x, y) - \eta^*(x, y)| : \eta \in \mathcal{M} \}$. Now, observe that for any $g, g^\dagger \in \mathcal{G}$ one has

$$\frac{1}{n} \sum_{i=1}^{n} \left| g(X_i, Y_i) - g^\dagger(X_i, Y_i) \right| = \frac{1}{n} \sum_{i=1}^{n} \left| \left| \eta(X_i, Y_i) - \eta^*(X_i, Y_i) \right| - \eta^\dagger(X_i, Y_i) - \eta^*(X_i, Y_i) \right|$$

$$\leq \frac{1}{n} \sum_{i=1}^{n} \left| \eta(X_i, Y_i) - \eta^\dagger(X_i, Y_i) \right| .$$

Therefore, it follows that if $\mathcal{M}_{\epsilon_2} = \{ \eta_1, \ldots, \eta_{N_{\epsilon_2}} \}$ is a minimal $\epsilon_2$-cover of $\mathcal{M}$ with respect to the empirical $L_1$ norm, the class of functions $\mathcal{G}_{\epsilon_2} = \{ g_{\eta_1}, \ldots, g_{\eta_{N_{\epsilon_2}}} \}$ will be an $\epsilon_2$-cover of $\mathcal{G}$. Furthermore, the $\epsilon_2$-covering numbers for $\mathcal{G}$ and $\mathcal{M}$ satisfy the following $N_1(\epsilon_2, \mathcal{G}, \mathbb{D}_n) \leq N_1(\epsilon_2, \mathcal{M}, \mathbb{D}_n)$. Upon employing standard results from the empirical process theory (see Pollard (1984), p.25, or Theorem 9.1 Györfi et al (2002, p. 136)), the upper bound for (3.40) follows

$$(3.40) \leq 2P \left\{ \sup_{\eta \in \mathcal{M}} \left( \frac{1}{n} \sum_{i=1}^{n} |\eta(X_i, Y_i) - \eta^*(X_i, Y_i)| \right. \right.$$

$$\left. - E |\eta(X, Y) - \eta^*(X, Y)| \right| D_n \right] > \frac{\epsilon_1 \eta_0^2}{2c_1} \right\}$$

$$\leq 16 E[N_1(b_3, \mathcal{M}, \mathbb{D}_n)] e^{-nb_4}, \quad (3.42)$$

where $b_3 \equiv b_3(\epsilon) = \frac{\epsilon^{p+1} \eta^{p+1}}{2^{p+1} M M_1 c_1}$ and $b_4 \equiv b_4(\epsilon) = \frac{\epsilon^{2p+2} \eta^{2p+2}}{2^{p+1} M M_1 c_1}$. 

50
To deal with (3.41), note that by the Cauchy Schwartz inequality we have

\[
(3.41) \leq 2P\left\{ \sqrt{E\left[ |\hat{\eta}_{\text{LS}}(X,Y) - \eta^*(X,Y)|^2 \right] D_n} > \frac{\epsilon_1\eta_0^2}{2c_1} \right\} \\
\leq 2P\left\{ E\left[ |\hat{\eta}_{\text{LS}}(X,Y) - \eta^*(X,Y)|^2 \right] D_n \right\} > \frac{\epsilon_1^2\eta_0^4}{4c_1^2} \right\}. \tag{3.43}
\]

Since

\[
E\left[ |\hat{\eta}_{\text{LS}}(X,Y) - \Delta|^2 \right] D_n = \inf_{\eta \in M} E\left[ |\eta(X,Y) - \Delta|^2 \right] + E\left[ |\hat{\eta}_{\text{LS}}(X,Y) - \eta^*(X,Y)|^2 \right] D_n
\]

we find

\[
E\left[ |\hat{\eta}_{\text{LS}}(X,Y) - \eta^*(X,Y)|^2 \right] D_n = E\left[ |\hat{\eta}_{\text{LS}}(X,Y) - \Delta|^2 \right] D_n
- \inf_{\eta \in M} E\left[ |\eta(X,Y) - \Delta|^2 \right],
\]
In other words, if $M$ is a minimal $\epsilon_3$-cover of $\mathcal{M}$ with respect to the empirical $L_1$ norm, then $\mathcal{F}_{\epsilon_3} = \{ f_{\eta_1}, \ldots, f_{\eta_{N_{\epsilon_3}}} \}$ is an $\epsilon_3$-cover of $\mathcal{F}$. Additionally one
has that \( \mathcal{N}_1(\epsilon_3, \mathcal{F}, \mathbb{D}_n) \leq \mathcal{N}_1(\epsilon_3/2, \mathcal{M}, \mathbb{D}_n) \), thus under standard results from empirical process theory (see, for example, Pollard (1984), p.25, or Theorem 9.1 Györfi et al (2002, p. 136 ), one has the following bound for (3.43)

\[
(3.43) \leq 2 P \left\{ \sup_{\eta \in \mathcal{M}} \left| \frac{1}{n} \sum_{i=1}^{n} \eta(X_i, Y_i) - \Delta_i \right|^2 - E \left| \eta(X, Y) - \Delta \right|^2 > \frac{\epsilon_3^4}{2^5 c_1^2} \right\} \\
\leq 16 E \left[ \mathcal{N}_1 \left( b_5, \mathcal{M}, (X_i, Y_i)^n_{i=1} \right) \right] e^{-nb_6} \tag{3.45}
\]

where \( b_5 \equiv b_5(\epsilon) = \frac{\epsilon_3^2 \eta_0^{2p+2}}{2^{3p+3} M p C_1} \) and \( b_6 \equiv b_6(\epsilon) = \frac{\epsilon_3^4 \eta_0^{4p+4}}{2^{12p+6} M p C_1} \). Therefore, in view of (3.38), (3.39), (3.42), and (3.45), we have

\[
Q_{n,1} + Q_{n,3} \leq 2 e^{-\frac{n \epsilon_3^{2p+2}}{2^{3p+3} M p C_1}} \]

\[
+ 16 E \left[ \mathcal{N}_1 \left( \frac{\epsilon_3^4 \eta_0^{p+1}}{2^{3p+3} M p C_1}, \mathcal{M}, \mathbb{D}_n \right) \right] e^{-\frac{n \epsilon_3^{2p+2}}{2^{3p+3} M p C_1}} \]

\[
+ 16 E \left[ \mathcal{N}_1 \left( \frac{\epsilon_3^4 \eta_0^{2p+2}}{2^{6p+4} M^2 p C_1^2}, \mathcal{M}, \mathbb{D}_n \right) \right] e^{-\frac{n \epsilon_3^{4p+4}}{2^{12p+6} M p C_1}}.
\]

This completes the proof of Theorem 3.3.

\[\square\]

### 3.2 An application to classification with missing covariates

In this section we consider an application of our previous results to the problem of statistical classification. Let \((Z, Y)\) be an \(\mathbb{R}^{d+s} \times \{1, \cdots, N\}\)-valued random vector, where \(Y\) is the class label and is to be predicted from the explanatory variables \(Z = (Z_1, \cdots, Z_{d+s})\). In classification one searches for a function \(\phi: \mathbb{R}^{d+s} \rightarrow \{1, \cdots, N\}\) such that the miss-classification error probability \(L(\phi) = P\{\phi(Z) \neq Y\}\) is as small as possible. Let

\[
\pi_k(z) = P\{Y = k|Z = z\}, \quad z \in \mathbb{R}^{d+s}, \quad 1 \leq k \leq N,
\]

then the best classifier (i.e., the one that minimizes \(L(\phi)\)) is called the Bayes classifier and is given by (see, for example, Devroye and Györfi (1985 pp. 253-254))

\[
\pi_B(z) = \arg\max_{1 \leq k \leq N} \pi_k(z). \tag{3.46}
\]
We note that the Bayes classifier satisfies
\[
\max_{1 \leq k \leq N} \pi_k(z) = \pi_{\phi_B}(z).
\]

In practice, the Bayes classifier is unavailable, because the distribution of \((Z, Y)\) is usually unknown. Instead, one only has access to a random sample \(D_n = \{(Z_1, Y_1), \ldots, (Z_n, Y_n)\}\) from the distribution \((Z, Y)\). Now, let \(\hat{\phi}_n(Z)\) be any sample-based classifier to predict \(Y\) from the data \(D_n\) and \(Z\). The error of misclassification of this estimator, conditioned on the data, is given by
\[
L_n(\hat{\phi}_n) = P\left\{ \hat{\phi}_n(Z) \neq Y \mid D_n \right\}.
\]

Then, it can be shown (see, for example, Devroye and Györfi (1985, pp. 254)) that
\[
0 \leq L_n(\hat{\phi}_n) - L(\phi_B) \leq \sum_{k=1}^{N} \int \left| \hat{\pi}_k(z) - \pi_k(z) \right| \mu(dz).
\]

Therefore, to show \(L_n(\hat{\phi}_n) - L(\phi_B) \xrightarrow{a.s.} 0\), it is sufficient to show that \(E[|\hat{\pi}_k(Z) - \pi_k(Z)|]\mid D_n \xrightarrow{a.s.} 0\), as \(n \to \infty\), for \(k = 1, \ldots, N\). Here, we consider the problem of classification with missing covariates, i.e., the case where some subsets of the \(Z_i\)'s may be unavailable. For some relevant results along these lines see Mojirsheibani (2012). Define
\[
\hat{\pi}_k(z) = \frac{\sum_{i=1}^{n} \Delta_i I\{y_i = k\} \mathcal{K}\left(\frac{z - Z_i}{h_n}\right)}{\sum_{i=1}^{n} \Delta_i I\{y_i = k\} \mathcal{K}\left(\frac{z - Z_i}{h_n}\right)}, \quad k = 1, \ldots, N, \tag{3.48}
\]

where \(\hat{\eta}\) is any estimator of \(\eta\), and define the classifier
\[
\hat{\phi}_n(z) = \arg\max_{1 \leq k \leq N} \hat{\pi}_k(z). \tag{3.49}
\]

The following result summarizes the performance of \(\hat{\phi}_n\).

**Theorem 3.4** Let \(\hat{\phi}_n\) be the classifier defined in (3.49) and (3.48).

(i) If \(\hat{\eta}\) in (3.48) is chosen to be the kernel estimator \(\hat{\eta}\) defined in (3.6) then, under the conditions of Theorem 3.2, \(\hat{\phi}_n\) is strongly consistent i.e., \(L_n(\hat{\phi}_n) - L(\phi_B) \xrightarrow{a.s.} 0\), as \(n \to \infty\).
(ii) If \( \hat{\eta} \) in (3.48) is chosen to be the least-squares estimator \( \hat{\eta}_{LS} \) defined in (3.26) then, under the conditions of Theorem 3.3, \( \hat{\phi}_n \) is strongly consistent i.e., \( L_n(\hat{\phi}_n) - L(\phi_B) \xrightarrow{a.s.} 0 \), as \( n \to \infty \).

Proof of Theorem 3.4

(i) Let \( \hat{\eta} \) be the kernel estimator \( \hat{\eta} \) defined in (3.6) then by (3.47) one has for every \( \epsilon > 0 \), and \( n \) large enough

\[
P\left\{ L_n(\hat{\phi}_n) - L(\phi_B) > \epsilon \right\} \leq P\left\{ \sum_{i=1}^{N} \int |\hat{\tau}_k(z) - \tau_k(z)| \mu(dz) > \frac{\epsilon}{N} \right\}.
\]

The proof now follows from an application of Theorem 3.2 and the Borel Cantelli lemma.

(ii) The proof of part (ii) is similar and will not be given.

\[ \square \]

3.3 Kernel regression simulation with missing covariates.

In this section we study the performance of the estimators \( \hat{\eta}^*, \hat{\eta}_c, \) and \( \hat{\eta}_{LS} \). The performance of each estimator will be compared with the complete case estimator of chapter 2 (denoted as \( \hat{\eta}_{co} \)). To explore the performance of these estimators, consider the following simulation. A sample of size 100 is taken from the linear model

\[ Y_i = 0.5V_i + 0.9X_i + \epsilon_i, \]

where \( V_i, X_i, \) and \( \epsilon_i \) are independent and identically distributed normal random variables with means \( \mu_1 = 5, \mu_2 = 9, \mu_3 = 0 \) and standard deviations \( \sigma_1 = 1, \sigma_2 = 6, \sigma_3 = 1 \) respectively, for \( i = 1, \cdots, 100 \). In the simulation \( V_i \) is sometimes missing with missing probability mechanism

\[
P(\Delta_i = 1|X_i, Y_i) = \frac{e^{a|X_i|+b|Y_i|}}{1 + e^{a|X_i|+b|Y_i|}}, \tag{3.50}
\]

here \( \Delta_i = 1 \) if \( V_i \) is available and \( \Delta_i = 0 \) otherwise for \( i = 1, \cdots, n \), \( a \) and \( b \) are chosen to produce approximately 25\% or 75\% missingness among the \( V_i \)'s. The least-squares estimator for the selection probability in the regression estimator \( \hat{\eta}_{LS} \) is correctly specified and the coefficients are estimated using the nls method in the R programming language. For the following smoothing parameters \( h = .2, .4, .6, \) and \( .8 \) we perform a series of fifty Monte Carlo runs and test the empirical \( L_1 \) and \( L_2 \) errors (defined in (1.15) and
Table 3.1: Comparison of empirical $L_1$ and $L_2$ errors of the estimators when $V_i$ is missing at random with missing mechanism as in (3.50). The missing mechanism has parameters $a = -2$ and $b = .195$, the choice of parameters results in approximately 25% of the $V_i$’s missing.

<table>
<thead>
<tr>
<th>$h$</th>
<th>$L_1$</th>
<th>$L_2$</th>
<th>$L_1$</th>
<th>$L_2$</th>
<th>$L_1$</th>
<th>$L_2$</th>
<th>$L_1$</th>
<th>$L_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$m_{Co}$</td>
<td>6.439</td>
<td>0.450</td>
<td>66.228</td>
<td>8.432</td>
<td>5.123</td>
<td>0.369</td>
<td>43.144</td>
<td>5.036</td>
</tr>
<tr>
<td>$m_{\eta^*}$</td>
<td>6.426</td>
<td>0.449</td>
<td>66.041</td>
<td>8.395</td>
<td>5.111</td>
<td>0.356</td>
<td>43.144</td>
<td>4.965</td>
</tr>
<tr>
<td>$m_{\eta^c}$</td>
<td>6.180</td>
<td>0.407</td>
<td>61.689</td>
<td>7.604</td>
<td>4.874</td>
<td>0.359</td>
<td>39.808</td>
<td>5.724</td>
</tr>
<tr>
<td>$m_{\eta^cLS}$</td>
<td>6.409</td>
<td>0.471</td>
<td>65.771</td>
<td>8.767</td>
<td>5.113</td>
<td>0.368</td>
<td>42.925</td>
<td>5.734</td>
</tr>
</tbody>
</table>

Table 3.2: Comparison of empirical $L_1$ and $L_2$ errors of the estimators when $V_i$ is missing at random with missing mechanism as in (3.50). The missing mechanism has parameters $a = -1$ and $b = .7$, the choice of parameters results in approximately 75% of the $V_i$’s missing.

<table>
<thead>
<tr>
<th>$h$</th>
<th>$L_1$</th>
<th>$L_2$</th>
<th>$L_1$</th>
<th>$L_2$</th>
<th>$L_1$</th>
<th>$L_2$</th>
<th>$L_1$</th>
<th>$L_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$m_{Co}$</td>
<td>7.491</td>
<td>0.725</td>
<td>85.428</td>
<td>13.452</td>
<td>5.672</td>
<td>0.417</td>
<td>52.704</td>
<td>7.001</td>
</tr>
<tr>
<td>$m_{\eta^*}$</td>
<td>7.399</td>
<td>0.706</td>
<td>83.779</td>
<td>13.071</td>
<td>5.608</td>
<td>0.402</td>
<td>51.453</td>
<td>6.707</td>
</tr>
<tr>
<td>$m_{\eta^c}$</td>
<td>6.915</td>
<td>0.653</td>
<td>75.163</td>
<td>11.928</td>
<td>5.202</td>
<td>0.373</td>
<td>44.349</td>
<td>6.066</td>
</tr>
<tr>
<td>$m_{\eta^cLS}$</td>
<td>7.391</td>
<td>0.701</td>
<td>83.623</td>
<td>12.984</td>
<td>5.594</td>
<td>0.402</td>
<td>51.353</td>
<td>6.767</td>
</tr>
</tbody>
</table>

(1.16)) for each estimator on an independent sample of 500 complete observations. The smoothing parameter for the kernel regression estimate in (3.7) is selected by using the cross-validation method of Racine and Li (2004) in the R package titled “np” (Racine and Hayfield (2008)). This method utilizes a conjugate gradient search pattern and returns a multi-dimensional bandwidth. Conversion of the vector of bandwidth into a single constant is done by taking the geometric mean of the vector. In tables 3.1 and 3.2 one may observe that the kernel regression estimator $\hat{m}_{\eta^c}$ defined in section 3.1.1 out performs each estimator in every situation. A non-intuitive observation is that the estimator that assumes knowledge of the missing probability mechanism i.e., $\hat{m}_{\eta^*}$ performs considerably worse than the estimator $\hat{m}_{\eta^c}$, which used an estimate of the missing probability mechanism. This behavior is observed in the literature, see for example Robins and Rotnitzky (1995). In almost all cases the complete case estimator performs worse than each of our estimators.
4 Kernel regression estimation with incomplete response

This chapter explores constructing kernel regression estimators in which the response variable may be missing at random. The results in this chapter mirror that of the case of missing covariates in Chapter 3. We derive exponential upper bounds for the $L_p$ norms of our proposed estimators, yielding almost sure convergence of the $L_p$ statistic to 0. At the conclusion of this chapter we once again apply the results to nonparametric classification. We construct kernel classifiers from partially labeled data.

4.1 Main results

We propose kernel estimators of an unknown regression function $m(z) = E[Y|Z = z]$ in the presence of missing response. Our estimator weighs the complete cases, i.e., the observed $Y_i$’s, by the inverse of the missing data probabilities. First consider the simple, but unrealistic, case where the missing probability mechanism (also called the selection probability)

$$
\pi^*(z) = P\{\Delta = 1|Z = z\} = E[\Delta|Z = z]
$$

(4.1)
is completely known, where $\Delta = 0$ if $Y$ is missing (and $\Delta = 1$, otherwise), and define the revised kernel estimator

$$
\hat{m}_{\pi^*}(z) = \frac{\sum_{i=1}^{n} \Delta_i Y_i}{\pi^*(Z_i)} \frac{K(\frac{z - Z_i}{h})}{\sum_{i=1}^{n} K(\frac{z - Z_i}{h})}.
$$

(4.2)

As for the usefulness of $\hat{m}_{\pi^*}(z)$ as an estimator of $m(z)$, first observe that (4.2) can be viewed as the kernel estimator of $E[\frac{\Delta Y}{\pi^*(Z)}|Z = z]$. Furthermore, when the MAR assumption (2.3) holds, one finds

$$
E\left[\frac{\Delta Y}{\pi^*(Z)}\bigg|Z\right] \overset{\text{a.s.}}{=} E\left[E\left[\frac{\Delta Y}{\pi^*(Z)}|Z, Y\right]|Z\right] = E\left[\frac{Y}{\pi^*(Z)} E[\Delta|Z, Y]|Z\right] = E\left[\frac{Y}{\pi^*(Z)} \pi^*(Z)|Z\right] = E[Y|Z] := m(Z).
$$
In other words, (4.2) can be viewed as a kernel regression estimator of the regression function \( E[Y|Z = z] \) whenever the function \( \pi^* \) in (4.1) is completely known. How good is the estimator \( \hat{m}_{\pi^*}(z) \)? To answer this question, and in what follows, we shall assume that the selected kernel \( K \) is regular: as described earlier in (1.4.3) and we state the following condition.

**Condition B1.** \( \inf_z P\{\Delta = 1|Z = z\} := \pi_0 > 0 \), where \( \pi_0 \) can be arbitrarily small.

Condition B1 guarantees, in a sense, that \( Y \) will be observed with a nonzero probability when \( Z = z \), for all \( z \). The following result is an immediate consequence of the main theorem of Devroye and Krzyżak (1989).

**Theorem 4.1** Let \( \hat{m}_{\pi^*}(z) \) be the kernel estimator defined in (4.2), where \( K \) is a regular kernel, and suppose that condition B1 holds. If \( h \to 0 \) and \( nh^d \to \infty \), as \( n \to \infty \), then for any distribution of \((Z, Y)\) satisfying \( |Y| \leq M < \infty \), and for every \( \epsilon > 0 \) and \( n \) large enough

\[
P\left\{ \int |\hat{m}_{\pi^*}(z) - m(z)| \mu(dz) > \epsilon \right\} \leq e^{-an},
\]

where \( a \equiv a(\epsilon) = \min \left( \frac{\pi_0^2 \epsilon^2}{128M^2(1+c_1)} : \frac{\pi_0^3 \epsilon}{32M(1+c_1)} \right) \). Here \( c_1 \) is a constant that depends on the kernel \( K \) only.

**Remarks.** The above result continues to hold for \( \int |\hat{m}_{\pi^*}(z) - m(z)|^p \mu(dz) \), for all \( 1 \leq p < \infty \), with \( a(\epsilon) \) replaced by \( a(\tilde{\epsilon}) \) where \( \tilde{\epsilon} = \epsilon/(M(\pi_0^{-1} + 1))^{p-1} \). To appreciate this, simply note that for all \( z \in \mathbb{R}^d \) and \( p \in [1, \infty) \)

\[
|\hat{m}_{\pi^*}(z) - m(z)|^p \leq (|\hat{m}_{\pi^*}(z)| + |m(z)|)^{p-1} |\hat{m}_{\pi^*}(z) - m(z)|.
\]

Therefore, under the conditions of Theorem 4.1, and in view of the Borel-Cantelli lemma, one finds \( E[|\hat{m}_{\pi^*}(Z) - m(Z)|^p |\mathcal{D}_n]] \xrightarrow{a.s.} 0 \), as \( n \to \infty \).

Of course the estimator \( \hat{m}_{\pi^*} \) defined by (4.2) is useful only if the missing probability mechanism \( \pi^*(z) = P\{\Delta = 1|Z = z\} \) is known. If this is not the case, \( \pi^*(z) \) must be replaced by some estimator in (4.2). In what follows, we consider kernel as well as least-squares
regression type estimators of $\pi^*(z)$.

### 4.1.1 A kernel-based estimator of the selection probability

Our first estimator of $\pi^*(Z_i) = E(\Delta_i|Z_i)$ is defined by

$$\hat{\pi}(Z_i) = \frac{\sum_{j=1,\neq i}^n \Delta_j H \left( \frac{Z_i - Z_j}{a_n} \right)}{\sum_{j=1,\neq i}^n H \left( \frac{Z_i - Z_j}{a_n} \right)}$$

(4.3)

with the convention $0/0 = 0$, where $H : \mathbb{R}^d \to \mathbb{R}_+$ is the kernel used with the smoothing parameter $a_n \to 0$, as $n \to \infty$. The kernel regression estimator in (4.3) is a weighted average of the $\Delta_j$'s, this provides an estimate for the probability that the $i$th observation of the response will be unavailable. Our first proposed kernel-type regression estimator of $m(z) = E[Y|Z = z]$ is given by

$$\hat{m}_r(z) = \frac{\sum_{i=1}^n \Delta_i Y_i \mathcal{K} \left( \frac{z - Z_i}{h} \right)}{\sum_{i=1}^n \mathcal{K} \left( \frac{z - Z_i}{h} \right)}$$

(4.4)

To investigate the performance of the regression estimator (4.4) we first state the following additional conditions.

**Condition B2.** The kernel $H$ in (4.3) satisfies $\int_{\mathbb{R}^d} H(\mathbf{u}) d\mathbf{u} = 1$ and $\int_{\mathbb{R}^d} |u_i| H(\mathbf{u}) d\mathbf{u} < \infty$, for $i = 1, \cdots, d$, where $\mathbf{u} = (u_1, \cdots, u_d)'$. Furthermore, the smoothing parameter $a_n$ satisfies $a_n \to 0$ and $n a_n^d \to \infty$, as $n \to \infty$.

**Condition B3.** The random vector $Z$ has a compactly supported probability density function, $f(z)$, where $f$ is bounded away from zero on its compact support. Additionally, both $f$ and its first-order partial derivatives are uniformly bounded.

**Condition B4.** The partial derivatives $\frac{\partial}{\partial z_i} E[\Delta|Z = z]$, exist for $i = 1, \cdots, d$ and are bounded uniformly, in $z$, on the compact support of $f$. 

59
Condition B2 is not restrictive since the choice of the kernel \( H \) is at our discretion. Condition B3 is usually imposed in nonparametric regression in order to avoid having unstable estimates (in the tail of the p.d.f of \( Z \)). Condition B4 is technical and has already been used in the literature; see, for example, Cheng and Chu (1996).

**Theorem 4.2** Let \( \hat{m}_H(z) \) be as in (4.4), where \( K \) is a regular kernel. Suppose that \( |Y| \leq M < \infty \). If \( h \to 0 \) and \( nh^d \to \infty \), as \( n \to \infty \), then under conditions B1, B2, B3, and B4, for every \( \epsilon > 0 \) and every \( p \in [1, \infty) \), and \( n \) large enough,

\[
P \left\{ \int |\hat{m}_H(z) - m(z)|^p \mu(dz) > \epsilon \right\} \leq 4e^{-bn + 4ne^{-C_1n\alpha_0^d}e^{2} + e^{-C_2n\epsilon^2} + 4ne^{-C_3n\alpha_0^d}},
\]

where \( b \equiv b(\epsilon) = \min \left\{ \frac{\pi_0^2e^{2}}{2^pM^2(1+c_1)(2(M\pi_0^{-1} + M))^{2p-2}} \cdot \frac{\pi_0^e}{2^pM(1+c_1)(2(M\pi_0^{-1} + M))^{p-1}} \right\} \), and

\[
C_1 = \frac{\pi_0^4f_0^2}{2^{2p}(3M\pi_0^{-1})^{2p-2}(32M^3c_1)^2\|H\|_\infty(2\|f\|_\infty + f_0/12)},
\]
\[
C_2 = \frac{\pi_0^2}{[2(3M\pi_0^{-1})^{2p-2}(64M^2c_1)^2]},
\]
\[
C_3 = \frac{\pi_0^2f_0^2}{256\|H\|_\infty(2\|f\|_\infty + \pi_0f_0/24)}.
\]

**Remark.** We note that as an immediate consequence of Theorem 4.2, if \( \log n/(na_0^d) \to 0 \), as \( n \to \infty \), then in view of the Borel-Cantelli lemma, one has \( E [ |\hat{m}_H(Z) - m(Z)|^p ] \overset{a.s.}{\to} 0 \).

**PROOF OF THEOREM 4.2**
First note that
\[
P \left\{ \int |\hat{m}_\pi(z) - m(z)|^p \mu(dz) > \epsilon \right\} \leq P \left\{ \int 2^{p-1} |\hat{m}_\pi(z) - \hat{m}_{\pi^*}(z)|^p \mu(dz) > \frac{\epsilon}{2} \right\} \\
+ P \left\{ \int 2^{p-1} |\hat{m}_{\pi^*}(z) - m(z)|^p \mu(dz) > \frac{\epsilon}{2} \right\} \\
:= P_{n,1} + P_{n,2}, \text{(say).} \quad (4.5)
\]

But, by Theorem 4.1 (and the remarks that follow Theorem 4.1),
\[
P_{n,2} \leq P \left\{ \int [2(M\pi^{-1} + M)]^{p-1} |\hat{m}_{\pi^*}(z) - m(z)| > \frac{\epsilon}{2} \right\} \leq 4e^{-bn},
\]
for \(n\) large enough, with
\[
b \equiv b(\epsilon) = \min^2 \left\{ \frac{\pi_0^2 \epsilon^2}{2^p M^2 (1 + c_1)[2(M\pi^{-1} + M)]^{2p-2}} , \frac{\pi_0 \epsilon}{2^6 M (1 + c_1)[2(M\pi^{-1} + M)]^{p-1}} \right\}.
\]

To deal with the term \(P_{n,1}\), first note that
\[
|\hat{m}_\pi(z) - \hat{m}_{\pi^*}(z)| = \frac{\sum_{i=1}^{n} \left( \frac{1}{\pi(Z_i)} - \frac{1}{\pi^*(Z_i)} \right) \Delta_i Y_i \mathcal{K} \left( \frac{z - Z_i}{h} \right)}{\sum_{i=1}^{n} \mathcal{K} \left( \frac{z - Z_i}{h} \right)} \\
\leq \frac{\sum_{i=1}^{n} \left( \frac{1}{\pi(Z_i)} - \frac{1}{\pi^*(Z_i)} \right) \Delta_i Y_i \mathcal{K} \left( \frac{z - Z_i}{h} \right)}{nE [\mathcal{K} \left( \frac{z - Z_i}{h} \right)]} \\
+ \left( \sum_{i=1}^{n} \left( \frac{1}{\pi(Z_i)} - \frac{1}{\pi^*(Z_i)} \right) \Delta_i Y_i \mathcal{K} \left( \frac{z - Z_i}{h} \right) \right) \\
\times \left( \frac{1}{nE [\mathcal{K} \left( \frac{z - Z_i}{h} \right)]} - \frac{1}{\sum_{i=1}^{n} \mathcal{K} \left( \frac{z - Z_i}{h} \right)} \right) \\
=: |U_{n,1}(z)| + |U_{n,2}(z)|, \text{(say).} \quad (4.6)
\]
Therefore, by the definition of \( \hat{m}_\pi(z) \) in (4.4) and the fact that \( |Y| \leq M \),

\[
|\hat{m}_\pi(z) - \hat{m}_{\pi^*}(z)|^p \leq \left( \frac{M}{\Lambda_{i=1}^n \hat{\pi}(Z_i)} + \frac{M}{\pi_0} \right)^{p-1} \left( |U_{n,1}(z)| + |U_{n,2}(z)| \right). \tag{4.7}
\]

But

\[
\int |U_{n,1}(z)| \mu(dz) \leq \int \sum_{i=1}^n \frac{1}{\Lambda_i(Z_i)} \frac{1}{\hat{\pi}'(Z_i)} |\Delta_i Y_i| \frac{K(z - Z_i)}{E[K(z - Z_i)]} \mu(dz)
\]

\[
= \frac{1}{n} \sum_{i=1}^n |\Delta_i Y_i| \left| \frac{1}{\hat{\pi}'(Z_i)} - \frac{1}{\pi^*(Z_i)} \right| \int \frac{K(z - Z_i)}{E[K(z - Z_i)]} \mu(dz)
\]

\[
\leq \sup_u \int \frac{K(z - u)}{E[K(z - Z_i)]} \mu(dz) \times \frac{1}{n} \sum_{i=1}^n |\Delta_i Y_i| \left| \frac{1}{\hat{\pi}'(Z_i)} - \frac{1}{\pi^*(Z_i)} \right|
\]

\[
\leq \frac{M c_1}{n} \sum_{i=1}^n \left| \frac{1}{\hat{\pi}'(Z_i)} - \frac{1}{\pi^*(Z_i)} \right|, \tag{4.8}
\]

where the last line follows from Lemma B.1 and the fact that \( |\Delta_i Y_i| \leq M \), for all \( i = 1, \cdots, n \); here \( c_1 \) is a finite positive constant depending on the kernel \( K \) only. Now, for
every $\epsilon > 0$,

\[
P\left\{ \int 2^{p-1} \left( \frac{M}{\pi_0} + \frac{M}{\pi_0} \right)^{p-1} \left| U_{n,1}(z) \right| \mu(dz) > \epsilon \right\} \\
\leq P\left\{ 2^{p-1} \left( \frac{M}{\pi_0} + \frac{M}{\pi_0} \right)^{p-1} \times \frac{1}{n} \sum_{i=1}^{n} \left| \frac{1}{\pi_i(Z_i)} - \frac{1}{\pi^*(Z_i)} \right| > \frac{\epsilon}{4M c_1} \right\} \\
\leq P\left\{ 2^{p-1} \left( \frac{M}{\pi_0} + \frac{M}{\pi_0} \right)^{p-1} \times \frac{1}{n} \sum_{i=1}^{n} \left| \frac{\pi(Z_i) - \pi^*(Z_i)}{\pi(Z_i)\pi^*(Z_i)} \right| > \frac{\epsilon}{4M c_1} \right\} \\
\cap \bigcap_{i=1}^{n} \left[ \frac{\pi(Z_i)}{2} \right] \right\} \\
+ P\left\{ \bigcup_{i=1}^{n} \left[ \frac{\pi(Z_i)}{2} \right] \right\} \\
\leq \sum_{i=1}^{n} P\left\{ 2^{p-1} \left( \frac{3M}{\pi_0} \right)^{p-1} \left( \frac{2}{\pi_0} \right) \left| \frac{\pi(Z_i) - \pi^*(Z_i)}{\pi(Z_i)\pi^*(Z_i)} \right| > \frac{\epsilon}{4M c_1} \right\} \\
\leq \sum_{i=1}^{n} P\left\{ \frac{\pi(Z_i)}{2} \right\} \right\}.
\]

As for the second term in (4.6), observe that since $|\Delta_i Y_i| \leq M$, for all $i = 1, \ldots, n$, one finds

\[
\left| U_{n,2}(z) \right| \leq M \max_{1 \leq i \leq n} \left| \frac{1}{\pi(Z_i)} - \frac{1}{\pi(Z_i)} \right| \left| \frac{\sum_{i=1}^{n} \mathcal{K}\left( \frac{z-Z_i}{h} \right)}{nE[\mathcal{K}\left( \frac{z-Z_i}{h} \right)]} - 1 \right|.
\]
Now, for every $\epsilon > 0$,\[
\begin{align*}
P \left\{ \int_{2^{p-1}}^{2^p} \left( \frac{M}{\pi_0} \right)^{p-1} \left| U_{n,2}(z) \right| \mu(dz) > \frac{\epsilon}{4} \right\} \\
\leq P \left\{ \int_{2^{p-1}}^{2^p} \left( \frac{M}{\pi_0} \right)^{p-1} \left| U_{n,2}(z) \right| \mu(dz) > \frac{\epsilon}{4} \right\} \cap \bigcap_{i=1}^{n} \left\{ \hat{\pi}(Z_i) \geq \frac{\pi_0}{2} \right\} \\
+ P \left\{ \bigcup_{i=1}^{n} \left[ \hat{\pi}(Z_i) < \frac{\pi_0}{2} \right] \right\} \\
\leq P \left\{ \int_{2^{p-1}}^{2^p} \left( \frac{3M}{\pi_0} \right)^{p-1} \left| \sum_{i=1}^{n} K \left( \frac{z - Z_i}{h} \right) \right| \left( -1 \right) \mu(dz) > \frac{\epsilon}{4} \right\} \\
+ \sum_{i=1}^{n} P \left\{ \hat{\pi}(Z_i) < \frac{\pi_0}{2} \right\}.
\end{align*}
\]

Therefore, in view of (4.7), the term $P_{n,1}$ in (4.5) can be upper-bounded as follows:

\[
P_{n,1} \leq \sum_{i=1}^{n} \left\{ \left| \hat{\pi}(Z_i) - \pi^*(Z_i) \right| > \frac{\pi_0^2 \epsilon}{8Mc_1[2(3M\pi_0^{-1})]^{p-1}} \right\} \\
+ P \left\{ \int \left| \sum_{i=1}^{n} K \left( \frac{z - Z_i}{h} \right) \right| \left( -1 \right) \mu(dz) > \frac{\pi_0^2 \epsilon}{8M[2(3M\pi_0^{-1})]^{p-1}} \right\} \\
+ 2 \sum_{i=1}^{n} P \left\{ \hat{\pi}(Z_i) < \frac{\pi_0}{2} \right\} := \Delta_{n,1} + \Delta_{n,2} + \Delta_{n,3}, \text{ (say)}.
\]

First observe that by Lemma B.2, \[
\Delta_{n,2} \leq \exp \left\{ \frac{-n \pi_0^2 \epsilon^2}{(64M^2c_1)^2[2(3M\pi_0^{-1})]^{2p-2}} \right\},\]

for $n$ large enough; (this is a special case of Lemma B.2 with $Y_i = 1$ for $i = 1, \cdots, n$, and $m(z) = 1$). To bound the term $\Delta_{n,1}$ we start by defining\[
\psi(Z) = \pi^*(Z) \cdot f(Z) \quad (4.9)
\]
\[
\hat{\psi}(Z_i) = \frac{1}{(n-1)a_n^d} \sum_{j=1, j\neq i}^n \Delta_j H \left( \frac{Z_i - Z_j}{a_n} \right) \tag{4.10}
\]

\[
\hat{f}(Z_i) = \frac{1}{(n-1)a_n^d} \sum_{j=1, j\neq i}^n H \left( \frac{Z_i - Z_j}{a_n} \right) \tag{4.11}
\]

where \(a_n\) and the kernel \(H\) are as in (4.3). Since \(\left| \frac{\hat{\psi}(Z_i)}{\hat{f}(Z_i)} \right| \leq 1\), one finds

\[
\left| \hat{\pi}(Z_i) - \pi^*(Z_i) \right| = \left| \frac{\hat{\psi}(Z_i)}{\hat{f}(Z_i)} - \frac{\psi(Z_i)}{f(Z_i)} \right| = \left| -\frac{\hat{\psi}(Z_i)}{\hat{f}(Z_i)} \cdot \frac{\hat{f}(Z_i) - f(Z_i)}{f(Z_i)} + \frac{\hat{\psi}(Z_i) - \psi(Z_i)}{f(Z_i)} \right|
\]

\[
\leq \frac{1}{f(Z_i)} \left| \hat{f}(Z_i) - f(Z_i) \right| + \frac{1}{f(Z_i)} \left| \hat{\psi}(Z_i) - \psi(Z_i) \right|
\]

Therefore,

\[
\Delta_{n,1} \leq \sum_{i=1}^n P \left\{ \frac{1}{f(Z_i)} \left| \hat{\psi}(Z_i) - \psi(Z_i) \right| > \frac{\pi_0^2 \epsilon}{16Mc_1[2(3M\pi_0^{-1})]^{p-1}} \right\} + \sum_{i=1}^n P \left\{ \frac{1}{f(Z_i)} \left| \hat{f}(Z_i) - f(Z_i) \right| > \frac{\pi_0^2 \epsilon}{16Mc_1[2(3M\pi_0^{-1})]^{p-1}} \right\}.
\]
Now, observe that
\[
P \left\{ \frac{\hat{f}(Z_i) - f(Z_i)}{f(Z_i)} > \frac{\pi_0^2 \epsilon}{16MC_1[2(3M\pi_0^{-1})]^{p-1}} \right\}
\]
\[
\leq P \left\{ \left| \hat{\psi}(Z_i) - E \left[ \hat{\psi}(Z_i) \big| Z_i \right] \right| + E \left[ \hat{\psi}(Z_i) \big| Z_i \right] - \psi(Z_i) > \frac{\pi_0^2 f_0 \epsilon}{16MC_1[2(3M\pi_0^{-1})]^{p-1}} \right\}
\]
(\text{where } f_0 = \inf_z f(z) > 0, \text{ by condition B3})
\[
\leq P \left\{ \left| \hat{\psi}(Z_i) - E \left[ \hat{\psi}(Z_i) \big| Z_i \right] \right| + \frac{\pi_0^2 f_0 \epsilon}{32MC_1[2(3M\pi_0^{-1})]^{p-1}} > \frac{\pi_0^2 f_0 \epsilon}{16MC_1[2(3M\pi_0^{-1})]^{p-1}} \right\}
\]
(for \( n \) large enough, by Lemma B.4)
\[
= E \left\{ P \left\{ \left| \hat{\psi}(Z_i) - E \left[ \hat{\psi}(Z_i) \big| Z_i \right] \right| > \frac{\pi_0^2 f_0 \epsilon}{32MC_1[2(3M\pi_0^{-1})]^{p-1}} \right\} \bigg| Z_i \right\}
\]
\[
\leq 2e^{-2(32MC_1)^2\|H\|_\infty^2 \|f\|_\infty + a_n^2 \epsilon_0^2 / (96MC_1[2(3M\pi_0^{-1})]^{p-1}) \|2(3M\pi_0^{-1})\|2^{p-2}}
\]
(by Lemma B.5)
\[
\leq 2e^{-2(32MC_1)^2\|H\|_\infty^2 \|2\|_\infty + f_0/12\|2(3M\pi_0^{-1})\|2^{p-2}}
\]
for \( n \) large enough, where we have also used the fact that in bounding \( P \{ \left| \hat{\pi}(Z_i) - \pi^*(Z_i) \right| > \pi_0^2 \epsilon / (8MC_1[2(3M\pi_0^{-1})]^{p-1}) \} \) one only needs to consider \( 0 < \epsilon < 8MC_1[2(3M\pi_0^{-1})]^{p-1} / \pi_0^2 \) (because \( \left| \hat{\pi}(Z_i) - \pi^*(Z_i) \right| \leq 1 \)). Similarly, since \( \hat{f} \) is a special case of \( \hat{\psi} \), with \( \Delta_j = 1 \) for all \( j \)'s (compare (4.10) and (4.11)), one finds
\[
P \left\{ \frac{\hat{f}(Z_i) - f(Z_i)}{f(Z_i)} > \frac{\pi_0^2 \epsilon}{16MC_1[2(3M\pi_0^{-1})]^{p-1}} \right\}
\]
\[
\leq 2 \exp \left\{ -\left( n - 1 \right) a_n^4 \pi_0^4 f_0^2 \epsilon^2 \right\}
\]
\[
\leq 2(32MC_1)^2\|H\|_\infty^2 \|f\|_\infty + f_0/12\|2(3M\pi_0^{-1})\|2^{p-2}
\]
for \( n \) large enough. Putting all the above together, we find
\[
\Delta_{n,1} \leq 4ne^{-\left( n - 1 \right) a_n^4 \epsilon_0^2 / 10}
\]
where

\[ C_{10} = \pi_0^4 f_0^2 / \left\{ 2(32M c_1)^2 \| H \|_\infty [2\| f \|_\infty + f_0/12] [2(3M \pi_0^{-1})]^{2p-2} \right\} \]

Finally, since \( P\{\hat{\pi}(Z_i) < \pi_0/2\} \leq P\{|\hat{\pi}(Z_i) - \pi^*(Z_i)| > \pi_0/2\} \), the arguments that lead to the derivation of the bound on \( \Delta_{n,1} \) yield

\[ \Delta_{n,3} \leq 2 \sum_{i=1}^{n} P \left\{ |\hat{\pi}(Z_i) - \pi^*(Z_i)| > \frac{\pi_0}{2} \right\} \]
\[ \leq 4n \exp \left\{ \frac{-(n-1)a_n^d \pi_0^2 f_0^2}{128\| H \|_\infty [2\| f \|_\infty + a_n^d \pi_0 f_0/24]} \right\} \]
\[ \leq 4n e^{-(n-1)a_n^d C_{11}}, \text{ (for } n \text{ large enough)}, \]

where \( C_{11} = \pi_0^2 f_0^2 / (128\| H \|_\infty [2\| f \|_\infty + \pi_0 f_0/24]) \) does not depend on \( n \). This completes the proof of Theorem 4.2.

\[ \square \]

4.1.2 Least-squares-based estimator of the selection probability

As in chapter 3 if we have prior knowledge about the behavior of the missing probability mechanism we can use least-squares methods. As a result, our second estimator of \( \pi^*(z) \) is based on a least-squares approach and works as follows. Suppose that \( \pi^* \) belongs to a known class of functions \( \mathcal{P} \) of the form \( \pi : \mathbb{R}^d \to [\pi_0, 1] \), where \( \pi_0 = \inf_z \pi(z) > 0 \) is as before (see assumption B1). Then the least-squares estimator of the function \( \pi^* \) is

\[ \hat{\pi}_{LS} = \arg\min_{\pi \in \mathcal{P}} \frac{1}{n} \sum_{i=1}^{n} (\Delta_i - \pi(Z_i))^2. \] (4.12)

Replacing \( \pi^* \) by \( \hat{\pi}_{LS} \) in (4.2), one obtains

\[ \hat{m}_{\hat{\pi}_{LS}}(z) = \frac{\sum_{i=1}^{n} \Delta_i Y_i \mathcal{K} \left( \frac{z - Z_i}{h} \right)}{\sum_{i=1}^{n} \mathcal{K} \left( \frac{z - Z_i}{h} \right)}. \] (4.13)

To study the performance of the regression function estimator in (4.13), we assume our class of functions is bounded with respect to the supnorm. Recall from chapter 1 that a class of functions \( \mathcal{P} \) is said to be totally bounded with respect to the supnorm if for every
\(\epsilon > 0\), there is a subclass \(\mathcal{P}_\epsilon = \{\pi_1, \ldots, \pi_{N(\epsilon)}\}\) such that for every \(\pi \in \mathcal{P}\), there is a \(\tilde{\pi} \in \mathcal{P}_\epsilon\) satisfying \(\|\pi - \tilde{\pi}\|_\infty < \epsilon\). The class \(\mathcal{P}_\epsilon\) is called an \(\epsilon\)-cover of \(\mathcal{P}\). The cardinality of the smallest such \(\epsilon\)-cover is denoted by \(N_\infty(\epsilon, \mathcal{P})\).

**Theorem 4.3** Let \(\hat{m}_{\tilde{\pi}_{LS}}(z)\) be as in (4.13), where \(\tilde{\pi}_{LS}\) is the least-squares estimator of \(\pi^*\) \(\epsilon\mathcal{P}\), and where \(\mathcal{P}\) is totally bounded with respect to the supnorm. Let \(K\) be a regular kernel in (4.13) and suppose that condition B1 holds. If \(h \to 0\) and \(nh^{d} \to \infty\), as \(n \to \infty\), then for any distribution of \((Z, Y)\) satisfying \(|Y| \leq M < \infty\), and for every \(\epsilon > 0\) and every \(p \in [1, \infty)\), and \(n\) large enough

\[
P\left\{ \int |\hat{m}_{\tilde{\pi}_{LS}}(z) - m(z)|\mu(dz) > \epsilon \right\} \leq 4e^{-b_m} + 2N_\infty(C_4 \epsilon, \mathcal{P}) e^{-C_5 n \epsilon^2} + 2N_\infty(C_6 \epsilon^2, \mathcal{P}) e^{-C_7 n \epsilon^4} + e^{-C_8 n \epsilon^2}
\]

where \(b\) is as in Theorem 4.2, and

\[
\begin{align*}
C_4 &= \frac{\pi_0^2}{32M c_1 (2(M\pi_0^{-1}+M))^{p-1}}, & C_5 &= \frac{\pi_0^2}{128M^2 c_1 (2(M\pi_0^{-1}+M))^{2p-2}} \\
C_6 &= \frac{\pi_0^2}{1024M^2 c_1 (2(M\pi_0^{-1}+M))^{2p-2}}, & C_7 &= \frac{\pi_0^2}{256M^4 c_1 (2(M\pi_0^{-1}+M))^{4p-4}} \\
C_8 &= \frac{\pi_0^2}{1024M^4 c_1 (2(M\pi_0^{-1}+M))^{2p-2}},
\end{align*}
\]

where \(c_1\) is a constant that depends on the kernel \(K\) only.

Theorem 4.3 in conjunction with the Borel-Cantelli lemma yields

\[
E \left[ |\hat{m}_{\tilde{\pi}_{LS}}(Z) - m(Z)|^p \mid \mathbb{D}_n \right] \xrightarrow{a.s.} 0, \quad \text{as } n \to \infty.
\]

**Remark.** The upper bound found in Theorem 4.3 is meaningful when the class \(\mathcal{P}\) is totally bounded with respect to the supnorm. See Section 1.3.2 for examples of classes of functions totally bounded with respect to the supnorm.

**PROOF OF THEOREM 4.3**

First note that for the least-squares estimator \(\tilde{\pi}_{LS}\) we have \(|\hat{m}_{\tilde{\pi}_{LS}}(z)| \leq \frac{M}{\Lambda_{n_{LS}}(Z)} \leq \frac{M}{\pi_0}.
\]
Therefore, employing the arguments used in the proof of Theorem 4.2, we have
\[
P \left\{ \int |\hat{m}_{\pi_{0}}(z) - m(z)|^p \mu(dz) > \epsilon \right\} \leq q_{n,1} + q_{n,2},
\]
where
\[
q_{n,1} = P \left\{ \int [2(M_{\pi_{0}^{-1}} + M)]^{p-1} \left| \hat{m}_{\pi}(z) - m(z) \right| \mu(dz) > \frac{\epsilon}{2} \right\} \leq 4e^{-bn}, \quad \text{(for large } n),
\]
where \(b\) is as in Theorem 4.2, and where
\[
q_{n,2} = P \left\{ \int [2(M_{\pi_{0}^{-1}} + M)]^{p-1} \left| \hat{m}_{\pi_{0}}(z) - \hat{m}_{
\pi}(z) \right| \mu(dz) > \frac{\epsilon}{2} \right\}. \quad (4.14)
\]
To upper-bound \(q_{n,2}\) first note that
\[
|\hat{m}_{\pi_{0}}(z) - \hat{m}_{\pi}(z)| \leq \sum_{i=1}^{n} \left| \frac{1}{\pi_{0}(Z_i)} - \frac{1}{\pi^{*}(Z_i)} \right| |\Delta Y_i| \frac{\mathcal{K}(z - Z_i)}{n E \left[ \mathcal{K}(z - Z_i) \right]}
\]
\[
+ \sum_{i=1}^{n} \left( \frac{1}{\hat{\pi}_{0}(Z_i)} - \frac{1}{\hat{\pi}^{*}(Z_i)} \right) (\Delta Y_i)
\]
\[
\times \left( \frac{\mathcal{K}(z - Z_i)}{n E \left[ \mathcal{K}(z - Z_i) \right]} - \frac{\mathcal{K}(z - Z_i)}{\sum_{i=1}^{n} \mathcal{K}(z - Z_i)} \right)
\]
\[
:= |S_{n,1}(z)| + |S_{n,2}(z)|, \quad \text{(say)}. \quad (4.15)
\]
Furthermore, as in (4.8),
\[
\int |S_{n,1}(z)| \mu(dz) \leq \frac{Mc_1}{n} \sum_{i=1}^{n} \left| \frac{1}{\pi_{0}(Z_i)} - \frac{1}{\pi^{*}(Z_i)} \right|,
\]
which yields, for every $t > 0$,

$$P \left\{ \int |S_{n,1}(z)| \mu(dz) > t \right\}$$

$$\leq P \left\{ \frac{1}{n} \sum_{i=1}^{n} \left| \frac{1}{\pi_{LS}(Z_i)} - \frac{1}{\pi^{*}(Z_i)} \right| > \frac{t}{M_1} \right\}$$

$$= P \left\{ \frac{1}{n} \sum_{i=1}^{n} \left| \frac{1}{\pi_{LS}(Z_i)} \pi^{*}(Z_i) \right| \left| \pi_{LS}(Z_i) - \pi^{*}(Z_i) \right| > \frac{t}{M_1} \right\}$$

$$\leq P \left\{ \frac{1}{n} \sum_{i=1}^{n} \left| \pi_{LS}(Z_i) - \pi^{*}(Z_i) \right| > \frac{\pi_0^2 t}{M_1} \right\}$$

(since $B_1$ yields $\inf_{z} \pi_{LS}(z) \geq \inf_{z} \pi^{*}(z) = \pi_0 > 0$, and the definition of class $\mathcal{P}$)

$$\leq P \left\{ \frac{1}{n} \sum_{i=1}^{n} \left| \pi_{LS}(Z_i) - \pi^{*}(Z_i) \right| - E \left[ \left| \pi_{LS}(Z) - \pi^{*}(Z) \right| \| D_n \right] \right\} > \frac{\pi_0^2 t}{2M_1} \right\} (4.16)$$

$$+ P \left\{ E \left[ \left| \pi_{LS}(Z) - \pi^{*}(Z) \right| \| D_n \right] > \frac{\pi_0^2 t}{2M_1} \right\}. \quad (4.17)$$

To deal with the term (4.16), put $\epsilon' = \pi_0^2 t/8M_1$ and let $\mathcal{P}_{\epsilon'}$ be a minimal $\epsilon'$-cover of $\mathcal{P}$. In other words, for every $\pi \in \mathcal{P}$, there is a $\tilde{\pi} \in \mathcal{P}_{\epsilon'}$ such that $\| \pi - \tilde{\pi} \| \leq \epsilon'$. Let $g(z) = |\pi(z) - \pi^{*}(z)|$ and $\tilde{g}(z) = |\tilde{\pi}(z) - \pi^{*}(z)|$ and note that
\[
\begin{align*}
\frac{1}{n} \sum_{i=1}^{n} \left| \pi(Z_i) - \pi^*(Z_i) \right| - E \left| \pi(Z) - \pi^*(Z) \right| \\
&\leq \frac{1}{n} \sum_{i=1}^{n} \left| \pi(Z_i) - \pi^*(Z_i) \right| - \sum_{i=1}^{n} \left| \tilde{\pi}(Z_i) - \pi^*(Z_i) \right| \\
&\quad + \left| E \tilde{\pi}(Z) - \pi^*(Z) \right| - E \left| \pi(Z) - \pi^*(Z) \right| \\
&\quad + \frac{1}{n} \sum_{i=1}^{n} \left| \tilde{\pi}(Z_i) - \pi^*(Z_i) \right| - E \left| \tilde{\pi}(Z) - \pi^*(Z) \right| \\
&\leq 2 \| g - \tilde{g} \|_{\infty} + \frac{1}{n} \sum_{i=1}^{n} \left| \tilde{\pi}(Z_i) - \pi^*(Z_i) \right| - E \left| \tilde{\pi}(Z) - \pi^*(Z) \right| \\
&\leq \frac{\pi_0^2 t}{4M c_1} + \frac{1}{n} \sum_{i=1}^{n} \left| \tilde{\pi}(Z_i) - \pi^*(Z_i) \right| - E \left| \tilde{\pi}(Z) - \pi^*(Z) \right|,
\end{align*}
\]

where the last line follows from the fact that
\[
|g(z) - \tilde{g}(z)| = ||\pi(z) - \pi^*(z)| - |\tilde{\pi}(z) - \pi^*(z)|| \leq |\pi(z) - \tilde{\pi}(z)| \leq ||\pi - \tilde{\pi}||_{\infty} \leq \frac{\pi_0^2 t}{8M c_1}.
\]

Therefore with \( \epsilon' = \pi_0^2 t/[8M c_1] \),

\[
(4.16) \quad \leq \quad P \left\{ \sup_{\pi \in D} \left| \frac{1}{n} \sum_{i=1}^{n} \left| \pi(Z_i) - \pi^*(Z_i) \right| - E \left| \pi(Z) - \pi^*(Z) \right| > \frac{\pi_0^2 t}{2M c_1} \right\} \\
\leq \quad P \left\{ \sup_{\pi \in D, \epsilon'} \left| \frac{1}{n} \sum_{i=1}^{n} \left| \pi(Z_i) - \pi^*(Z_i) \right| - E \left| \pi(Z) - \pi^*(Z) \right| + \frac{\pi_0^2 t}{4M c_1} > \frac{\pi_0^2 t}{2M c_1} \right\} \\
\leq \quad \mathcal{N} \left( \frac{\pi_0^2 t}{8M c_1}, D \right) \\
\quad \times \sup_{\pi \in D, \epsilon'} P \left\{ \left| \frac{1}{n} \sum_{i=1}^{n} \left| \pi(Z_i) - \pi^*(Z_i) \right| - E \left| \pi(Z) - \pi^*(Z) \right| > \frac{\pi_0^2 t}{4M c_1} \right\} \\
\leq \quad 2 \mathcal{N} \left( \frac{\pi_0^2 t}{8M c_1}, D \right) \cdot \exp \left\{ -\frac{\pi_0^4 t^2 n}{8M^2 c_1} \right\},
\]

(4.18)
where the exponential bound in (4.18) follows from an application of Hoeffding’s (1963) inequality. Next, to bound (4.17), we first note that

\[(4.17) \leq P \left\{ \sqrt{E \left[ \left| \hat{\pi}_{LS}(Z) - \pi^*(Z) \right|^2 \right]} > \frac{\pi_0^2 t}{2Mc_1} \right\} \]

(by Cauchy-Schwartz inequality)

\[= P \left\{ E \left[ \left| \hat{\pi}_{LS}(Z) - \pi^*(Z) \right|^2 \right] > \frac{\pi_0^4 t^2}{4M^2c_1} \right\} \]

\[\leq P \left\{ \sup_{\pi \in \mathcal{P}} \left| \nu_n(\pi) - E[\pi(Z) - \Delta]^2 \right| > \frac{\pi_0^4 t^2}{8M^2c_1} \right\} \quad (4.19)\]

(by Lemma B.5, where \(\nu_n(\pi) = n^{-1} \sum_{i=1}^n (\pi(Z_i) - \Delta_i)^2\)).

Now, put

\[\epsilon'' = \frac{\pi_0^4 t^2}{64M^2c_1}\]

and let \(\mathcal{P}_{\epsilon'}\) be a minimal \(\epsilon''\)-cover of \(\mathcal{P}\) with respect to the supnorm. Therefore, given \(\pi \in \mathcal{P}\), there is a \(\tilde{\pi} \in \mathcal{P}_{\epsilon'}\) such that \(|\pi - \tilde{\pi}|_\infty < \epsilon''\). Since

\[|\pi(Z) - \Delta|^2 - |\tilde{\pi}(Z) - \Delta|^2 \leq |(\pi(Z) - \Delta) + (\tilde{\pi}(Z) - \Delta)| \times |(\pi(Z) - \Delta) - (\tilde{\pi}(Z) - \Delta)| \]

\[\leq 2|\pi(Z) - \tilde{\pi}(Z)| \leq 2||\pi - \tilde{\pi}||_\infty \leq 2\epsilon'',\]

one finds

\[|\nu_n(\pi) - E[\pi(Z) - \Delta]^2| \leq \frac{1}{n} \left| \sum_{i=1}^n |\pi(Z_i) - \Delta_i|^2 - \sum_{i=1}^n |\tilde{\pi}(Z_i) - \Delta_i|^2 \right| + |E[\pi(Z) - \Delta]^2 - E[\tilde{\pi}(Z) - \Delta]^2| \]

\[+ \frac{1}{n} \sum_{i=1}^n |\tilde{\pi}(Z_i) - \Delta_i|^2 - E|\tilde{\pi}(Z) - \Delta|^2| \]

\[\leq 4\epsilon'' + \frac{1}{n} \sum_{i=1}^n |\tilde{\pi}(Z_i) - \Delta_i|^2 - E|\tilde{\pi}(Z) - \Delta|^2| .\]
Therefore, with \( \epsilon'' = \pi_0 t^2 / (64M^2c_1) \),

\[
\begin{align*}
(4.19) & \leq P \left\{ \sup_{\pi \in \mathcal{P}_{\epsilon''}} \frac{1}{n} \sum_{i=1}^{n} \left| \tilde{\pi}(Z_i) - \Delta_i \right|^2 - E \left| \tilde{\pi}(Z) - \Delta \right|^2 > \frac{\pi_0 t^2}{16M^2c_1} \right\} \\
& \leq \mathcal{N}_\infty \left( \frac{\pi_0 t^2}{64M^2c_1}, \mathcal{P} \right) \\
& \quad \times \sup_{\pi \in \mathcal{P}_{\epsilon''}} P \left\{ \left| \frac{1}{n} \sum_{i=1}^{n} \tilde{\pi}(Z_i) - \Delta_i \right|^2 - E \left| \tilde{\pi}(Z) - \Delta \right|^2 > \frac{\pi_0 t^2}{16M^2c_1} \right\} \\
& \leq 2\mathcal{N}_\infty \left( \frac{\pi_0 t^2}{64M^2c_1}, \mathcal{P} \right) \cdot \exp \left\{ \frac{-\pi_0 t^2}{128M^4c_1} \right\}, \hspace{1cm} (4.20)
\end{align*}
\]

where the last line follows by an application of Hoeffding’s inequality.

To deal with the term \( |S_{n,2}(z)| \) in (4.15), first note that

\[
|S_{n,2}(z)| \leq M \cdot \max_{1 \leq i \leq n} \left| \frac{1}{\pi_{1s}(Z_i)} - \frac{1}{\pi^*(Z_i)} \right| \left| \frac{\sum_{i=1}^{n} K \left( \frac{z - Z_i}{h} \right)}{n E \left[ K \left( \frac{z - Z}{h} \right) \right]} - 1 \right| \\
\leq M \frac{\sum_{i=1}^{n} K \left( \frac{z - Z_i}{h} \right)}{n E \left[ K \left( \frac{z - Z}{h} \right) \right]} - 1.
\]

Now, in view of Lemma 3 with \( Y = 1 \) (and thus \( m(z) = 1 \), for all \( z \)), we find that for every constant \( t > 0 \),

\[
P \left\{ \int |S_{n,2}(z)| \mu(dz) > t \right\} \leq P \left\{ \int |m_n^*(z) - 1| \mu(dz) > \frac{t\pi_0}{M} \right\} \\
\leq \exp \left\{ \frac{-n\pi_0^2t^2}{64M^4c_1} \right\}, \hspace{1cm} (4.21)
\]

for \( n \) large enough, where \( m_n^*(z) = \sum_{i=1}^{n} K \left( \frac{z - Z_i}{h} \right) / (n E[K(z - Z/h)]) \) is a special case of \( m_n^*(z) \) of Lemma B.2 with \( Y, Y_1, \cdots, Y_n = 1 \), (i.e., here \( m_n^*(z) \) is the kernel regression estimator of \( m(z) = 1 \)).
Therefore, in view of (4.14), (4.15), (4.18), (4.20), and (4.21), we have

\[
q_{n,2} \leq \left\{ \int [2(M \pi_0^{-1} + M)]^{p-1} \left| S_{n,1}(z) \right| \mu(dz) > \frac{\epsilon}{4} \right\} \\
+ \left\{ \int [2(M \pi_0^{-1} + M)]^{p-1} \left| S_{n,2}(z) \right| \mu(dz) > \frac{\epsilon}{4} \right\} \\
\leq 2 \mathcal{N}_\infty \left( \frac{2^2 \epsilon}{32 M c_1 [2(M \pi_0^{-1} + M)]^{p-1}}, \mathcal{P} \right) e^{-\frac{2^{15} M^4 c_1 [2(M \pi_0^{-1} + M)]^{2p-4}}{128 M^2 c_1 [2(M \pi_0^{-1} + M)]^{2p-2}}} \\
+ 2 \mathcal{N}_\infty \left( \frac{2^4 \epsilon^2}{1024 M^2 c_1 [2(M \pi_0^{-1} + M)]^{2p-2}}, \mathcal{P} \right) e^{-\frac{2^{15} M^4 c_1 [2(M \pi_0^{-1} + M)]^{2p-4}}{128 M^2 c_1 [2(M \pi_0^{-1} + M)]^{2p-2}}} \\
+ e^{-\frac{2^{15} M^4 c_1 [2(M \pi_0^{-1} + M)]^{2p-4}}{128 M^2 c_1 [2(M \pi_0^{-1} + M)]^{2p-2}}}.
\]

This completes the proof of Theorem 4.3.

\[\square\]

4.2 Applications to classification with partially labeled data

Here we consider an application of the results developed in the previous sections to the problem of statistical classification. Let \((Z, Y)\) be an \(R^d \times \{1, \cdots, M\}\)-valued random pair. In classification one has to predict the integer-valued label \(Y\) based on the covariate vector \(Z\). More formally, one seeks to find a function (a classifier) \(\phi : \mathbb{R}^d \rightarrow \{1, \cdots, M\}\) for which the probability of missclassification error (incorrect prediction), i.e., \(P\{\phi(Z) \neq Y\}\) is as small as possible. Let

\[
p_k(z) = P \{ Y = k \mid Z = z \}, \quad z \in \mathbb{R}^d, \quad 1 \leq k \leq M.
\]

It is straightforward to show that the best classifier (i.e., the one with the lowest probability of error) is given by

\[
\phi_B(z) = \arg\max_{1 \leq k \leq M} p_k(z), \quad (4.22)
\]

i.e, the best classifier \(\phi_B\) satisfies

\[
\max_{1 \leq k \leq M} p_k(z) = p_{\phi_B(z)}(z);
\]

see, for example, Devroye and Györfi (1985, pp. 253-254). Since \(\phi_B\) is almost always un-
known, one uses the data to construct estimates of $\phi_B$. Let $D_n = \{(Z_1, Y_1), \cdots, (Z_n, Y_n)\}$ be a random sample (the data) from the distribution of $(Z, Y)$, where each $(Z_i, Y_i)$ is fully observable. Let $\hat{\phi}_n$ be any sample-based classifier. In other words, $\hat{\phi}_n(Z)$ is the predicted value of $Y$, based on $D_n$ and $Z$. Let

$$L_n(\hat{\phi}_n) = P\left\{\hat{\phi}_n(Z) \neq Y \mid D_n\right\},$$

be the conditional probability of error of the sample-based classifier $\hat{\phi}_n$. Then $\hat{\phi}_n$ is said to be strongly consistent if $L_n(\hat{\phi}_n) \xrightarrow{a.s.} L(\phi_B) := P\{\phi_B(Z) \neq Y\}$, as $n \to \infty$.

For $k = 1, \cdots, M$, let $\hat{p}_k(z)$ be any sample-based estimators of $p_k(z) = P\{Y = k \mid Z = z\}$ and define the classification rule $\hat{\phi}_n$ by

$$\hat{\phi}_n(z) = \arg\max_{1 \leq k \leq M} \hat{p}_k(z).$$

In other words, $\hat{\phi}_n$ satisfies

$$\max_{1 \leq k \leq M} \hat{p}_k(z) = \hat{p}_{\hat{\phi}_n(z)}(z).$$

Then, it can be shown (see, for example, Devroye and Györfi (1985, pp. 254)) that

$$0 \leq L_n(\hat{\phi}_n) - L(\phi_B) \leq \sum_{k=1}^{M} \int |\hat{p}_k(z) - p_k(z)| \mu(dz).$$ (4.23)

Therefore, to show $L_n(\hat{\phi}_n) - L(\phi_B) \xrightarrow{a.s.} 0$, it is sufficient to show that $E[|\hat{p}_k(Z) - p_k(Z)| \mid D_n] \xrightarrow{a.s.} 0$, as $n \to \infty$, for $k = 1, \cdots, M$. Now, consider the case of a partially labeled data, where some of the $Y_i$’s (the labels) may be unavailable in $D_n$. Let

$$\hat{p}_k(z) = \frac{\sum_{i=1}^{n} \Delta_i I\{Y_i = k\} K(z - z_i/h)}{\sum_{i=1}^{n} K(z - z_i/h)}, \quad k = 1, \cdots, M,$$ (4.24)

where $\tilde{\pi}$ is taken to be either $\hat{\pi}$ in (4.3) or $\hat{\pi}_{LS}$ as in (4.12), and where $\Delta_i = 1$ if $Y_i$ is available ($\Delta_i = 0$ if $Y_i$ is missing). With $\hat{p}_k$ as in (4.24), define the classifier

$$\hat{\phi}_n(z) = \arg\max_{1 \leq k \leq M} \hat{p}_k(z).$$ (4.25)
How good is \( \hat{\phi}_n \) in (4.25) as an estimator of the optimal classifier \( \phi_B \) in (4.22)? The answer is given by the following result.

**Theorem 4.4** Let \( \hat{\phi}_n \) be the classifier defined via (4.25) and (4.24).

(i) If \( \hat{\pi} \) in (4.24) is chosen to be the kernel estimator \( \hat{\pi} \) defined in (4.3) then, under the conditions of Theorem 4.2, \( \hat{\phi}_n \) is strongly consistent i.e., \( L_n(\hat{\phi}_n) - L(\phi_B) \xrightarrow{a.s.} 0 \).

(ii) If \( \hat{\pi} \) in (4.24) is chosen to be the least-squares estimator \( \hat{\pi}_{LS} \) defined in (4.12) then, under the conditions of Theorem 4.3, \( \hat{\phi}_n \) is strongly consistent.

**PROOF OF THEOREM 4.4**

(i) Let \( \hat{\pi} \) be the kernel regression estimator \( \hat{\pi} \) defined in (4.3) and observe that by (4.23), for every \( \epsilon > 0 \),

\[
P \left\{ L_n(\hat{\phi}_n) - L(\phi_B) > \epsilon \right\} \leq \sum_{k=1}^{M} P \left\{ \int |\hat{p}_k(z) - p_k(z)| \mu(dz) > \frac{\epsilon}{M} \right\}.
\]

The proof now follows from an application of Theorem 4.2 in conjunction with the Borel-Cantelli lemma.

(ii) The proof is similar to that of part (i).
Conclusion-Future work

In chapter 3, kernel regression estimators were constructed for the situation in which a subset of the covariate vector may be missing at random. In this case we are implicitly assuming that the missing pattern follows the multivariate two pattern problem discussed in Chapter 2. In practice the data can consist of multiple missing patterns, not just multivariate two pattern. To handle the case of multiple missing patterns, modifications to the estimators of chapter 3 are necessary.

The primary focus of the thesis has been on kernel regression estimators in the presence of incomplete data. The partition regression estimator and $k$-NN estimator of chapter 1 could potentially yield similar results as that of the kernel regression estimators constructed in chapters 3 and 4. These estimators in chapters 3 and 4 also dealt with the problem of missing data by using inverse weighting techniques. One weakness of this method is that it is still a function of only the complete cases, resulting in a smaller random sample size. Constructing kernel regression estimators that combine the methods of imputation and inverse weights may result in estimators that perform more efficiently.
References


A Probability Review

This section provides a brief review of some of the important probabilistic definitions used in establishing convergence of the regression estimators.

**Definition A.1 (Convergence in Probability)** Let $X_n$ be a sequence of random variables. We say that $X_n$ converges in probability to $X$ if \( \forall \epsilon > 0, \)

\[
P\{|X_n - X| > \epsilon\} \to 0 \quad \text{as } n \to \infty.
\]

The notation $X_n \xrightarrow{p} X$ denotes “$X_n$ converges to $X$ in probability”.

A stronger mode of convergence is almost sure convergence, this is the mode of convergence used in all our results.

**Definition A.2 (Almost Sure Convergence)** The random variable $X_n$ is said to converge almost surely (with probability 1) to $X$ if

\[
P\left\{ \lim_{n \to \infty} X_n = X \right\} = 1.
\]

The notation $X_n \xrightarrow{a.s.} X$ denotes “$X_n$ converges to $X$ almost surely”.

This definition is not commonly used to establish almost sure convergence. We provide an alternative definition, which can be more easily used to obtain the almost sure convergence of an estimator, through way of the Borel-Cantelli Lemma.

**Lemma A.1** Let $\{A_n : n \geq 1\}$ be a sequence of events in a probability space. If

\[
\sum_{n=1}^{\infty} P(A_n) < \infty,
\]

then

\[
P\left( \bigcap_{k=1}^{\infty} \bigcup_{n=k}^{\infty} A_n \right) = 0
\]

i.e., the probability that $A_n$'s occurs infinitely often is 0.
The following inequalities were used to establish exponential upper bounds for our regression estimators in chapters 3 and 4. See Györfi et al. (2002 p. 594).

**Theorem A.1 (Hoeffding (1963))** Let \(X_1, \ldots, X_n\) be independent random variables such that \(|X_i| \leq M\) with probability 1. Let \(\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i\) denote their average, then for any \(\epsilon > 0\),

\[
P\{\left|\overline{X} - E[\overline{X}]\right| > \epsilon\} \leq 2 \exp\left\{-\frac{2n\epsilon^2}{M^2}\right\}.
\]

**Theorem A.2 (Bernstein (1946))** Let \(X_1, \ldots, X_n\) be independent random variables with \(|X_i| \leq M\) almost surely, for all \(i = 1, \ldots, n\) and \(E[X_i^2] < \infty\). Put \(\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} \text{var}X_i\), then for any \(\epsilon > 0\),

\[
P\left\{\left|\frac{1}{n} \sum_{i=1}^{n} (X_i - E[X_i])\right| > \epsilon\right\} \leq 2 \exp\left\{-\frac{n\epsilon^2}{2\sigma^2 + 2M\epsilon/3}\right\}.
\]
B  Technical Lemmas

A number of important lemmas were used to establish the results of chapters 3 and 4. The lemmas have been provided in this section.

Lemma B.1  Let $K$ be a regular kernel, and let $\mu$ be any probability measure on the Borel sets of $\mathbb{R}^{d+s}$. Then, there exists a finite positive constant $c_1$, only depending on the kernel $K$, such that for all $h_n > 0$

$$
\sup_{u \in \mathbb{R}^{d+s}} \int \frac{K\left(\frac{z-u}{h_n}\right)}{E[K\left(\frac{z-Z}{h_n}\right)]} \mu(dz) \leq c_1.
$$

PROOF OF LEMMA B.1
This lemma and its proof is given in Devroye and Krzyżak (1989; Lemma 1). Also, see Devroye and Wagner (1980) as well as Spiegelman and Sacks (1980).

Lemma B.2  Let $K$ be a regular kernel, and put $m^*_n(z) = \sum_{i=1}^n Y_i K(\frac{z-Z_i}{h_n})/(nE[K(\frac{z-Z}{h_n})])$, where $(Z_i, Y_i)$'s are i.i.d. $\mathbb{R}^{d+s} \times \mathbb{R}$-valued random vectors. Then, if $|Y| \leq M < \infty$, we have for every $\epsilon > 0$ and $n$ large enough,

$$
P\left\{ \int |m^*_n(z) - m(z)| \mu(dz) > \epsilon \right\} \leq \exp\left\{ \frac{-n\epsilon^2}{64M^2c_1^2} \right\}.
$$

Here, $\mu$ is the probability measure of $Z$, $m(z) = E[Y|Z = z]$, and $c_1$ is as in Lemma B.1.

PROOF OF LEMMA B.2
For a proof of this result see, for example, Györfi et al. (2002; Lemma 23.9).
Lemma B.3 Let $\phi(X_i, Y_i) = \eta^\star(X_i, Y_i) f(X_i) P(Y = Y_i|Y_i)$ where $\eta^\star(X_i, Y_i)$ is as in (3.2), and put $\hat{\phi}(X_i, Y_i) = \lambda_n^{-d} n^{-1} \sum_{j=1}^n \Delta_j I\{Y_j = Y_i\} \mathcal{H}((X_i - X_j)/\lambda_n)$ where $\lambda_n$ and the kernel $\mathcal{H}$ are as in (3.6). Suppose that conditions B2, B3, and B4 hold.

Then,

$$|E\left[\hat{\phi}(X_i, Y_i)|X_i, Y_i\right] - \phi(X_i, Y_i)| \overset{a.s.}{\leq} c_2 \lambda_n$$

where $c_2$ is a positive constant not depending on $n$.

The proof is similar to that of Lemma 3 of Mojirsheibani (2012) and goes as follows.
PROOF OF LEMMA B.3

First note that

\[ E \left[ \hat{\phi}(X_i, Y_i) \right] - \phi(X_i, Y_i) \]

\[ = E \left[ \lambda_n I \{ Y_j = Y_i \} \mathcal{H} \left( \frac{X_i - X_j}{\lambda_n} \right) \right] - \phi(X_i, Y_i) \]

\[ \approx \lambda_n E \left[ I \{ Y_1 = Y_i \} \mathcal{H} \left( \frac{X_i - X_1}{\lambda_n} \right) \right] - \phi(X_i, Y_i) \]

\[ = \lambda_n E \left[ I \{ Y_1 = Y_i \} \mathcal{H} \left( \frac{X_i - X_1}{\lambda_n} \right) \eta^*(X_1, Y_1) \right] \]

\[ + \lambda_n E \left[ \eta^*(X_i, Y_i) I \{ Y_1 = Y_i \} \mathcal{H} \left( \frac{X_i - X_1}{\lambda_n} \right) \right] \]

\[ = \left( E \left[ \lambda_n \left( \eta^*(X_1, Y_1) \right) \right] - \eta^*(X_i, Y_i) f(X_i) P(Y = Y_i | Y_i) \right) \]

\[ : = \Delta_{n,i}(1) + \Delta_{n,i}(2) \]

Utilizing a one term Taylor expansion, we can bound \( \Delta_{n,i,j}(1) \) as follows

\[ |\Delta_{n,i,j}(1)| \leq \lambda_n E \left[ \sum_{k=1}^{d} \left| \frac{\partial \eta^*(X_i, Y_i)}{\partial X_k} \right| X_{i,k} - X_{1,k} \mathcal{H} \left( \frac{X_i - X_1}{\lambda_n} \right) \right], \]
where $X_{1,k}$ and $X_{i,k}$ are the $k^{th}$ components of $X_1$ and $X_i$, respectively, and $X^\dagger$ is on the interior of the line segment joining $X_1$ and $X_i$. Therefore,

$$|\Delta_{n,i}(1)| \leq \alpha_1 E \left[ \sum_{k=1}^{d} |X_{i,k} - X_{1,k}| \lambda_n^{-d} \mathcal{H} \left( \frac{X_i - X_1}{\lambda_n} \right) \right]$$

where $\alpha_1 = \max_{1 \leq k \leq d} \sup_{v \in \mathbb{R}^d, y \in Y} \left| \frac{\partial \eta^*(x, y)}{\partial x_k} \right|_{x=v}$

$$= \alpha_1 \sum_{k=1}^{d} \int_{\mathbb{R}^d} |X_{i,k} - x_k| \lambda_n^{-d} \mathcal{H} \left( \frac{X_i - x}{\lambda_n} \right) f(x) dx$$

$$\leq \alpha_1 \|f\|_{\infty} \sum_{k=1}^{d} \int_{\mathbb{R}^d} |u_k| \lambda_n \mathcal{H}(u) du \quad \text{(by condition A3)}$$

$$\leq \alpha \lambda_n \quad \text{(by condition A2)}.$$  

To bound the term $\Delta_{n,i,j}(2)$, first note that

$$\Delta_{n,i}(2) = \eta^* \left( X_i, Y_i \right) E \left[ \lambda_n^{-d} I \{Y_1 = Y_i\} \mathcal{H} \left( \frac{X_i - X_1}{\lambda_n} \right) - f(X_i) P(Y = Y_i | Y_i) \right]$$

Now, let $f_y(x)$ be the conditional pdf of $X$, given $Y = y$, and observe that

$$E \left[ \lambda_n^{-d} I \{Y_i = Y_i\} \mathcal{H} \left( \frac{X_i - X_1}{\lambda_n} \right) \right] = \sum_{y \in Y} P(Y = y) I \{Y_i = y\} \int_{\mathbb{R}^d} \lambda_n^{-d} \mathcal{H} \left( \frac{X_i - x}{\lambda_n} \right) f_y(x) dx$$

Therefore, utilizing condition A2 and the fact that

$$P(Y = Y_i | Y_i) = \sum_{y \in Y} P(Y = y) I \{Y_i = y\},$$

one finds

$$86$$
\[ \Delta_{n,i}(2) = \eta^*(X_i, Y_i) \sum_{y \in \mathcal{Y}} P(Y = y) \mathbb{I}\{Y_i = y\} \]
\[ \times \left[ \int_{\mathbb{R}^d} \lambda_n^{-d} \mathcal{H} \left( \frac{X_i - x}{\lambda_n} \right) (f_y(x) - f_y(X_i)) \, dx \right] \]

Performing a one term Taylor expansion and using the fact that \(|\eta^*(X_i, Y_i)| \leq 1\) and \(I\{Y_i = y\} \leq 1\) for every \(y \in \mathcal{Y}\), yields

\[ |\Delta_{n,i}(2)| \leq \sum_{y \in \mathcal{Y}} P(Y = y) \left| \int_{\mathbb{R}^d} \mathcal{H}(u) (f_y(X_i - \lambda_n u) - f_y(X_i)) \, du \right| \]
\[ \leq \lambda_n \sum_{k=1}^{d} \int_{\mathbb{R}^d} |u_k| \mathcal{H}(u) \, du \sum_{y \in \mathcal{Y}} P(Y = y) \sup_{v \in \mathbb{R}^d} \left| \frac{\partial f_y(x)}{\partial x_k} \right|_{x = v} \]
\[ \leq d\lambda_n \max_{1 \leq k \leq d} \sup_{v \in \mathbb{R}^d} \left| \frac{\partial f(x)}{\partial x_k} \right|_{x = v} \int_{\mathbb{R}^d} |u_k| \mathcal{H}(u) \, du \]
\[ \leq \beta \lambda_n \quad \text{(by condition A2, A3 and A4).} \]

This completes the proof of Lemma B.3

The following lemma is similar to that of Lemma B.3, with slight modifications to deal with the simpler case of missing response instead of missing covariates. Used in the proof of Theorem 4.2.

**Lemma B.4** Let \(\psi(Z) = \pi^*(Z) f(Z)\), where \(f\) is the probability density of \(Z\), and \(\pi^*(z) = P\{\Delta = 1|Z = z\}\). Put \(\hat{\psi}(Z) = n^{-1} a_n^{-d} \sum_{j=1}^{n} \Delta_j H((Z_j - Z) / a_n)\), where \(a_n\) and the kernel \(H\) are as in (4.3). Suppose that conditions B2, B3, and B4 hold. Then

\[ \mathbb{E}\left[ \hat{\psi}(Z) | Z \right] - \psi(Z) \xrightarrow{a.s.} \kappa a_n, \]

where \(\kappa\) is a positive constant not depending on \(n\).
PROOF OF LEMMA B.4
The proof is similar to that of Lemma 4.2 of Mojirsheibani (2007) and goes as follows.

\[
E \left[ \hat{\psi}(Z) \mid Z \right] - \psi(Z) = a_n^{-d} E \left[ \Delta_1 H((Z_1 - Z)/a_n) \mid Z \right] - \pi^*(Z)f(Z)
\]

\[
\overset{a.s.}{=} a_n^{-d} E \left[ H((Z_1 - Z)/a_n) \mid \Delta_1, Z, Z_1 \right] Z - \pi^*(Z)f(Z)
\] (B.1)

But \( E[\Delta_1 \mid Z, Z_1] = E[\Delta_1 \mid Z_1] = \pi^*(Z_1) \) because \( Z \) is independent of \( \Delta_1 \) and \( Z_1 \). Therefore

\[
(B.1) = a_n^{-d} E \left[ \left( \pi^*(Z_1) - \pi^*(Z) \right) H((Z_1 - Z)/a_n) \mid Z \right] + E \left[ \pi^*(Z) \left( a_n^{-d} H((Z_1 - Z)/a_n) - f(Z) \right) \mid Z \right]
\]

\[
:= T_{n,1}(Z) + T_{n,2}(Z), \quad \text{(say)}.
\]

An one-term Taylor expansion gives

\[
T_{n,1}(Z) = a_n^{-d} E \left[ \left( \sum_{i=1}^{d} \frac{\partial \pi^*(Z)}{\partial Z_i} (Z_{1,i} - Z_i) \right) H((Z_1 - Z)/a_n) \mid Z \right],
\]

where \( Z_i \) and \( Z_{1,i} \) are the \( i^{th} \) components of \( Z \) and \( Z_1 \), respectively, and \( \mathbf{Z}^\dagger \) is a point on interior of the line segment joining \( \mathbf{Z} \) and \( \mathbf{Z}_1 \). Therefore,

\[
|T_{n,1}(Z)| \leq C_{14} \sum_{i=1}^{d} E \left[ |Z_{1,i} - Z_i| a_n^{-d} H((Z_1 - Z)/a_n) \mid Z \right], \quad \text{(by condition B4)}
\]

(where \( C_{14} = \max_{1 \leq i \leq d} \sup_{\mathbf{z}} |\partial \pi^*(\mathbf{z})/\partial z_i| \))

\[
= C_{14} \sum_{i=1}^{d} \int |z_i - Z_i| a_n^{-d} H((z - Z)/a_n) f(z) dz
\]

\[
\leq C_{14} \| f \|_{\infty} \sum_{i=1}^{d} \int a_n |v_i| H(v) dv = |C_{15}| \cdot a_n, \quad \text{(by condition B2 and B4)},
\]

88
where we have made the substitution \( v_i = z_i - Z_i, i = 1, \ldots, d \). Next, to bound \( T_{n,2}(Z) \), first note that

\[
T_{n,2}(Z) = \pi^*(Z) \int a_n^{-d}H((z - Z)/a_n)[f(z) - f(Z)]dz
= \pi^*(Z) \int [f(Z + a_n y) - f(Z)]H(y)dy.
\]

Finally, a one-term Taylor expansion gives

\[
|T_{n,2}(Z)| \leq a_n \cdot d\|f'\|\infty \sum_{i=1}^{d} \int |z_i|H(z)dz, \quad \text{(by condition B2)}.
\]

This completes the proof of Lemma B.4.

The following lemma is used in the proof of Theorem 4.2.

**Lemma B.5** Let \( \hat{\psi} \) be as in Lemma B.4. Then, for every \( t > 0 \), one has

\[
P \left\{ \left| \hat{\psi}(Z) - E \left[ \hat{\psi}(Z) \big| Z \right] \right| > t \bigg| Z = z \right\} \leq 2 \exp \left\{ \frac{-nt^2}{2\|H\|\infty \left[ 2\|f\|\infty + 2a_n^d t/3 \right]} \right\}.
\]

**PROOF OF LEMMA B.5**

Let

\[
\Gamma_j(Z) = a_n^{-d} \left[ \Delta_j H \left( \frac{Z - Z_j}{a_n} \right) - E \left( \Delta_j H \left( \frac{Z - Z_j}{a_n} \right) \bigg| Z \right) \right]
\]

and note that, conditional on \( Z \), the terms \( \Gamma_j(Z) \) are independent, zero-mean random variables, bounded by \(-a_n^{-d}\|H\|\infty\) and \(+a_n^{-d}\|H\|\infty\). Also, \( \text{var}(\Gamma_j(Z)|Z) = E[\Gamma_j^2(Z)|Z] \leq 2a_n^{-d}\|H\|\infty\|f\|\infty \). Therefore, by Bernstein’s (1946) inequality,

\[
P \left\{ \left| \hat{\psi}(Z) - E \left[ \hat{\psi}(Z) \big| Z \right] \right| > t \bigg| Z = z \right\} = P \left\{ \frac{1}{n} \sum_{j=1}^{n} \Gamma_j(Z) > t \bigg| Z = z \right\}
\]

\[
\leq 2e^{-t^2/2[2a_n^{-d}\|H\|\infty\|f\|\infty + (a_n^{-d}\|H\|\infty^2)/3]},
\]

89
which does not depend on \( z \).

\[ \square \]

**Lemma B.6** Let \( \hat{\pi}_{ls} \) be as in (4.12) and suppose that \( \pi^* \in \mathcal{P} \), where \( \pi^*(z) = E[\Delta|Z = z] \). Then

\[
E \left[ |\hat{\pi}_{ls}(Z) - \pi^*(Z)| \right| D_n \right] \leq 2 \sup_{\pi \in \mathcal{P}} |\nu_n(\pi) - E|\pi(Z) - \Delta|_2|_D_n \right.,
\]

where

\[
\nu_n(\pi) = \frac{1}{n} \sum_{i=1}^{n} |\pi(Z_i) - \Delta_i|^2.
\]

**Proof of Lemma B.6**

Using the elementary decomposition

\[
E \left[ |\hat{\pi}_{ls}(Z) - \pi^*(Z)|^2 \left| D_n \right. \right] = E \left[ |\pi^*(Z) - \Delta|^2 \right] + E \left[ |\hat{\pi}_{ls}(Z) - \pi^*(Z)|^2 \left| D_n \right. \right],
\]

one has

\[
E \left[ |\hat{\pi}_{ls}(Z) - \pi^*(Z)|^2 \right| D_n \right] = \left\{ E \left[ |\hat{\pi}_{ls}(Z) - \Delta|^2 \right| D_n \right] - \inf_{\pi \in \mathcal{P}} E |\pi(Z) - \Delta|^2 \right\}
\]

\[
+ \left\{ \inf_{\pi \in \mathcal{P}} E |\pi(Z) - \Delta|^2 - E |\pi^*(Z) - \Delta|^2 \right\}
\]

\[ =: T_{n,1} + T_{n,2}. \]

But the term \( T_{n,2} = 0 \), because \( \pi^* \in \mathcal{P} \). As for the term \( T_{n,1} \), for each \( \pi \in \mathcal{P} \), let

\[
\nu_n(\pi) = \frac{1}{n} \sum_{i=1}^{n} |\pi(Z_i) - \Delta_i|^2
\]
and observe that

\[ T_{n,1} = \sup_{\pi \in \mathcal{P}} \left\{ E \left[ |\hat{\pi}_{LS}(Z) - \Delta|^{2} \right| D_n \right] - \nu_n(\pi) + \nu_n(\pi) \]

\[ -\nu_n(\hat{\pi}_{LS}) + \nu_n(\hat{\pi}_{LS}) - E |\pi(Z) - \Delta|^{2} \}

\[ \leq \left\{ -\nu_n(\hat{\pi}_{LS}) + E \left[ |\hat{\pi}_{LS}(Z) - \Delta|^{2} \right| D_n \right\} + \sup_{\pi \in \mathcal{P}} |\nu_n(\pi) - E |\pi(Z) - \Delta|^{2}| \]

(because \( \nu_n(\hat{\pi}_{LS}) - \nu_n(\pi) \leq 0, \forall \pi \in \mathcal{P} \))

\[ \leq 2 \sup_{\pi \in \mathcal{P}} |\nu_n(\pi) - E |\pi(Z) - \Delta|^{2}| . \]