Semiparametric Quantile Regression and Applications to Healthcare Data Analysis

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Adam Maidman

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Lan Wang, Adviser

June 2018

Acknowledgements

I am grateful to my advisor Dr. Lan Wang for her support and guidance in making this dissertation possible. She has been patient with me when I struggled and pushed me toward a successful path when I veered off track. I learned a tremendous amount from her and am fortunate to have had the opportunity to work with Lan these past few years.

I would also like to thank my committee members Drs. Ansu Chatterjee, Chih-Lin Chi, and Nate Helwig for taking the time to read over my dissertation and provide comments. My conversations with them have been helpful in shaping and framing my research.

This dissertation was also made possible by the support, help, and advice of my friends in the School of Statistics and my friends outside it as well; thank you Aaron, Ben, Brad, Brittany, Brittany, Dan, Dootika, Haema, James, Jordan, Matt, Mitch, Sakshi, Sam, and Yang. Our discussions were always interesting and time spent together always fun. You all played a large part in making my time in graduate school and Minneapolis a great experience.

I would not be where I am today without the unwavering support of my family. Thank you to my brothers Rich and Jordan, grandparents, and aunts and uncles who have helped me with my life and career choices, especially during the particularly difficult moments. Thank you to my parents for raising me to the point where I was able to graduate with a PhD in statistics. They have always challenged and encouraged me to pursue my interests and explore new ones. They prepared me well for research and a fulfilling life. Your guidance and insight is invaluable.

Abstract

The ubiquity of healthcare data allows for complex analyses of a variety of topics ranging from healthcare cost to cognitive decline in dementia patients. Healthcare datasets are often highly skewed and heteroskedastic posing great challenges for statistical analyses. Quantile regression is an effective tool for analyzing healthcare datasets because, compared with mean regression, quantile regression has weaker assumptions which are more appropriate for complex data. Additionally, quantile regression models conditional quantiles of the response variable providing a more complete picture of the conditional distribution. In this dissertation, we propose three solutions to challenges in healthcare data analysis. All three solutions either directly rely on quantile regression or extend existing methodology and algorithms.

Motivated by the Medical Expenditure Panel Survey containing data from individuals' medical providers and employers across the United States, we propose a new semiparametric procedure for predicting whether a patient will incur high medical expenditure. The common practice is to artificially dichotomize the response. We propose a new semiparametric prediction rule to classify whether a future response occurs at the upper tail of the response distribution. The new method can be considered a semiparametric estimator of the Bayes rule for classification and enjoys some nice features. It incorporates nonlinear covariate effects and can be adapted to construct a prediction interval and hence provides more information about the future response.

Next, we extend semiparametric quantile regression methodology to longitudinal studies with non-ignorable dropout. Dropout occurs when a patient leaves a study prior to its conclusion. Non-ignorable dropout occurs when the probability of dropout depends on the response. Failing to account for non-ignorable dropout can result in biased estimation. To handle dropout, we propose a weighted semiparametric quantile regression estimator where the weights are inversely proportional to the estimated probability remaining in the study. We show that this weighted estimator gives unbiased estimates of linear effects. We illustrate the advantages of the proposed method on a subset of the National Alzheimer's Coordinating Center Uniform Data Set tracking cognitive decline in dementia patients.

Lastly, we turn our attention to the issue of analyzing very large datasets with a large number of covariates and sample size. Penalized quantile regression is often used to simultaneously select variables and estimate effects by fitting models at many values of a tuning parameter. Existing algorithms have focused on improving computation time at one value of a tuning parameter, however obtaining model estimates for all values of the tuning parameter can still be prohibitively time-consuming. Instead of attempting to solve the penalized quantile regression problem for each value of a tuning parameter, we propose a sparsity path algorithm to approximate the solution allowing for fast exploration of candidate models at many different sparsity levels. Simulations show that the true model is always contained in the set of candidate models returned by the proposed sparsity path algorithm.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Researchers have been collecting and analyzing data to provide answers to a wide variety of problems in healthcare for decades. These data come from a variety of sources including clinical studies, large government sponsored panel surveys, and observations of patients during doctor visits. Analysis of these datasets can determine the efficacy of a new treatment for a disease, inform and influence healthcare policy, and predict future outcomes for individuals.

Healthcare data typically consist of observations with a single response and many covariates. A well known approach for analyzing a response conditional on covariates is mean regression. Typically ordinary least squares methods are used to estimate the conditional mean or variance and make inference or predictions about future observations. Ordinary least squares methods typically require strong assumptions on the distribution of the error. It is common to assume that the errors are identically and independently distributed, Other times, like in weighted least squares, the variance of the error for a particular observation is assumed to be proportional to a scalar which must be estimated from the data [Weisberg, 2005]. Inference and prediction can be inaccurate and misleading if these assumptions are violated. Healthcare and healthcare expenditure data, however, are often skewed and heterogeneous [Zhou et al., 2001], violating a key assumption of ordinary least squares.

Additionally, healthcare and healthcare expenditure analysis often requires estimation and inference of other features of the conditional distribution of the response beyond just the mean. For example, infants with extremely low birth weight need special medical attention immediately upon birth. In this case, estimation of the lower quantiles of the conditional birth weight distribution is needed. Another familiar example is the construction of growth charts for children's height and weights. All quantiles of the growth chart conditioned on age, gender, and potentially other covariates are needed to properly understand a child's growth in the context of his or her age and sex.

Quantile regression is a method that allows for estimation and prediction of all quantiles of the distribution of the response conditioned on covariates without making strong distributional assumptions on the response. Additionally, quantile regression allows for heteroscedastic errors. A benefit of allowing for heteroscedasticity means that covariates can effect the response differently at different quantiles. For example, a patient's sex can have a large effect on the median weight, but have a very small effect on weight at a high quantile. Put simply, the median weight of boys and girls can be different, but the heaviest boys can have similar weights as the heaviest girls. Going one step further, it is also possible that some covariates have a large effect for some quantiles and no effect at others [Wang et al., 2012]. These two features of quantile regression make quantile regression an appealing approach for analyzing healthcare and healthcare cost data.

1.1 Applications in healthcare

In the past two decades, researchers have applied quantile regression to a variety of problems in healthcare. In a study tracking neuropsychological performance, Sherwood et al. [2016] used quantile regression to model a patient's decline in cognitive ability over time. To track if a patient is experiencing more cognitive decline than expected, a patient's baseline cognitive ability is first matched to a quantile. As the patient ages, his or her cognitive ability is compared to this same quantile. A patient is said to be experiencing unnatural cognitive decline (a sign of cognitive impairment or dementia) if he or she fails to maintain performance at this quantile. Sherwood et al. [2016] found that neuropsychological performance decreased much faster over time for high performers than for middle or lower performers. Exploratory analysis of the data suggested that the data was heteroscedastic, violating a key assumption of mean regression. Had mean regression been used here, a large number of patients with high cognitive ability would have been mislabeled with cognitive impairment.

Understanding the effect of the economic recession of 2007-2009 on healthcare expenditure can help policy makers better plan for future recessions. Chen et al. [2014] analyzed healthcare expenditure during this time period using quantile regression and found that the recession was associated with reductions in expenditure for the lower quantiles, but that the recession did not effect expenditure for the higher quantiles. This means that the recession decreased spending among patients who were already spending little on healthcare, but did not effect the spending of those with high expenditure. An analysis using mean regression would not have detected the differences in the recession's effect on low and high spenders.

The above analyses assumed that the effects of the covariates on the response are linear. However, relationships between covariates and the response are not always linear. Consider a patient's age. When younger, a patient requires annual checkups and frequently has a health related problem. As the patient ages into adulthood, he or she probably does not need as much medical attention. But once the patient reaches a certain age, medical attention becomes much more necessary and frequent again. To handle these nonlinear relationships, He and Shi [1996] extended quantile regression to allow for the estimation of nonlinear effects. We use the term semiparametric quantile regression to refer to situations when both linear and nonlinear effects are estimated. We will discuss semiparametric quantile regression in more detail in Section 2.2.

Noting that children do not grow linearly over time, Wei et al. [2006b] modeled children's growth charts using semiparametric quantile regression allowing for age to have a nonlinear effect on growth. Sherwood and Wang [2016] extended the semiparametric quantile regression model to the high-dimensional case and estimated different quantiles of birthweight by allowing the mother's age to have a nonlinear effect. The semiparametric quantile regression model is also useful for analyzing longitudinal data. He et al. [2002] proposed a method for estimating the conditional quantiles with longitudinal data using semiparametric quantile regression and analyzed a hormone study.

Another challenging feature of healthcare and healthcare cost data is missing data. Sometimes some covariates are not always observed or sometimes patients drop out of a longitudinal study before completion. We will discuss different kinds of missing data and methods for handling missingness in more detail in Section 2.3. When missing data is ignored, estimates can be biased leading to incorrect conclusions. Sherwood et al. [2013] showed a method for consistently estimating the quantiles of healthcare costs when not all covaraites are always observed when all effects are linear. This method was later extended to the semiparametric quantile regression model when some covariates are not always observed and was used to model the time patients spend in a rehabilitation center [Sherwood, 2016]. In a longitudinal study where some patients dropped out prior to conclusion of the study, Lipsitz et al. [1997] used similar techniques to estimate the conditional quantiles of the CD4 cell count of HIV patients.

As researchers collect more and more data, they are often faced with the challenge of selecting which variables to include in models. Variable selection is a rich area in the statistical literature and there are many methods to help select a model. In the past year alone, two competing algorithms for variable selection have been proposed [Gu et al., 2017, Yu et al., 2017]. However, not all these methods and algorithms perform well when datasets are large [Fan et al., 2014]. Fast algorithms that work well on large datasets are essential for analysis of healthcare data.

1.2 Overview

The outline of this dissertation is as follows. In Chapter 2, we formally introduce and review quantile regression as it pertains to the rest of this dissertation. Motivated by the Medical Expenditure Panel Survey containing data from individuals' medical providers and employers across the United States we propose a new semiparametric procedure for predicting whether a patient will incur high medical expenditure in Chapter 3. In particular, we propose a new semiparametric prediction rule to classify whether a future response occurs at the upper tail of the response distribution.

Next in Chapter 4, we extend seimparametric quantile regression methodology to longitudinal studies with non-ignorable dropout. Dropout occurs when a patient leaves a study prior to its conclusion and non-ignorable dropout occurs when the probability of dropout depends on the response. Failing to account for non-ignorable dropout can result in biased estimation. To handle dropout, we propose a weighted semiparametric quantile regression estimator where the weights are inversely proportional to the estimated probability remaining in the study. We illustrate the advantages of the proposed method on a subset of the National Alzheimer's Coordinating Center Uniform Data Set tracking cognitive decline in dementia patients. Patients in this study are more likely to dropout as their cognitive abilities decline.

Lastly, we turn our attention to the issue of analyzing very large datasets with a large number of covariates and sample size in Chapter 5. Penalized quantile regression is often used to simultaneously select variables and estimate effects by fitting models at many values of a tuning parameter. Existing algorithms have focused on improving computation time at one value of a tuning parameter, however obtaining model estimates for all values of the tuning parameter can still be prohibitively time-consuming. Instead of attempting to solve the penalized quantile regression problem for each value of a tuning parameter, we propose a sparsity path algorithm to approximate the solution at increasing values of the tuning parameter. We conclude this dissertation in Chapter 6 with a discussion of our contributions and future extensions of our work.

Chapter 2

Review: Quantile Regression

In this chapter, we will review quantile regression and basic properties that are relevant to this dissertation. We will begin the review in Section 2.1 with quantile regression when all the covariates have linear effects on the repsonse. We will extend linear quantile regression to semiparametric quantile regression in Section 2.2 which relaxes the linear assumption and allows some covariates to have nonlinear effects on the response. The review of semiparametric quantile regression will focus on partially linear additive quantile regression, a subclass of semiparametric quantile regression. Finally we will review different types of missing data and some methods for handling missingness in the quantile regression framework in Section 2.3. Computational methods will also be discussed.

First, we will introduce some notation. Let (Y, X') be a random variable where $Y \in \mathbb{R}$ and $X \in \mathbb{R}^p$. The conditional distribution function of Y given X is $F_{Y|X}(y) = P(Y \leq y|X)$. For a given $\tau \in (0, 1)$, the τth conditional quantile of Y given X is defined as $Q_{Y|X}(\tau) = \inf\{t : F_{Y|X}(t) \geq \tau\}$. The conditional median corresponds to $Q_{Y|X}(0.5)$. Interpretation of the conditional quantile is straightforward. For example, given the vector of covariates $X = x$ and $\tau = 0.9$, 90% of observations of Y with associated $X = x$ fall below $Q_{Y|X}(0.9)$. A useful property of the quantile function is the invariance property. For any monotone function $h(\cdot)$, for example the logarithm function, $Q_{h(Y)|X}(\tau) = h\left(Q_{Y|X}(\tau)\right)$; the analog for the conditional mean is not always true, i.e., in general $E[\log(Y)|X] \neq \log(E[Y|X])$.

2.1 Linear quantile regression

Given the random sample $(Y_i, X_i')'$, $i = 1, \ldots, n$, the classical linear quantile regression model assumes that $Q_{Y_i|X_i}(\tau) = X_i'\beta(\tau)$ where $\beta(\tau)$ is a vector of unknown coefficients. Alternatively, we can write

$$Y_i = X_i'\beta(\tau) + \varepsilon_i$$

where the errors $\{\varepsilon_i\}_{i=1}^n$ are independent and satisfy the quantile constraint $P(\varepsilon_i < 0|X_i) = \tau$. Because there is no assumption on any parametric distribution for ε_i and no restriction on homogeneity of variance, quantile regression is an attractive model for modeling heteroscedastic and nonnormal data. When ε_i are independent and identically distributed, the coefficient vector $\beta(\tau)$ does not depend on τ except the intercept; while for heteroscedastic data, the coefficient vector usually varies for different values of τ. By studying different choices of τ, we can gain a more complete understanding of the relationship between Y and X.

Koenker and Bassett [1978] proved that the estimator for $\beta(\tau)$ can be obtained by solving the following convex optimization problem

$$\hat{\beta}(\tau) = \underset{\beta \in \mathcal{R}^p}{\arg\min} \sum_{i=1}^{n} \rho_\tau(Y_i - X_i'\beta), \tag{2.1}$$

with loss function $\rho_\tau(u) = u(\tau - I\{u < 0\})$. Figure 2.1 depicts the quantile loss function, which is a weighted L_1 objective function. The optimization problem in (2.1) can be effectively solved by linear programming [Koenker and d'Orey, 1987, 1994, Koenker and Park, 1996]. The `R` package `quantreg` provides functions for

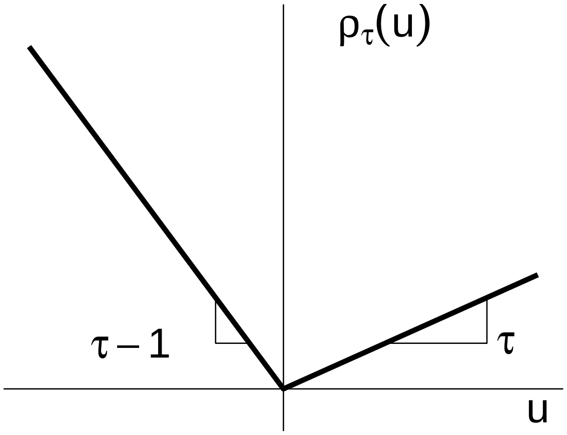

Figure 2.1: Plot of quantile loss function.

obtaining the estimator [Koenker, 2013].

Before stating a key asymptotic property of $\hat{\beta}(\tau)$, we first need two mild conditions.

> (Conditions on the random error) The random error ϵ_i conditioned on the covariates X_i has the distribution function F_i and continuous conditional density function f_i. The f_i are uniformly bounded away from 0 and infinity in a neighborhood of zero and its first derivative f_i' has a uniform upper bound in a neighborhood of zero, for $1 \leq i \leq n$.
>
> (Conditions on the covariates) Let D be a compact subset of $\mathbb{R}^p$ and $X_i \in D$ for $i = 1, \ldots, n$.

Under these two conditions, Theorem 4.1 of Koenker [2005] proves that if $n^{-1} \sum_{i=1}^n X_i X_i' \xrightarrow{p} \Sigma$ and $n^{-1} \sum_{i=1}^n f_i(0) X_i X_i' \xrightarrow{p} \Sigma_1$, then

$$\sqrt{n}\left(\hat{\beta}(\tau) - \beta(\tau)\right) \xrightarrow{d} N\left(0, \tau(1-\tau)\Sigma_1^{-1}\Sigma\Sigma_1^{-1}\right).$$

It is important to note that this result does not require the errors to be identically distributed.

If a distributional assumption can be made on the errors, then mean regression can be used to estimate the quantiles in addition to the mean. To understand this point, consider data generated from the following simple model:

$$Y_i = X_i'\beta + \varepsilon_i, \tag{2.2}$$

where $X_i \in \mathbb{R}$ is an observed covariate, β is the unknown coefficient, and $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$. This model meets the assumptions of ordinary least squares (OLS) and can be rewritten as $Y_i|X_i \sim N(X_i\beta, \sigma^2)$. Let $\hat{\beta}^{OLS}$ denote the OLS estimator of β and $\hat{\sigma}^2$ be the OLS estimate of σ^2. Then the estimate of $E[Y_i|X_i = x_i]$ is $x_i\hat{\beta}^{OLS}$. Because we assumed a normal distribution for $Y_i|X_i$, the estimate of $Q_{Y_i|X_i=x_i}(\tau) = x_i\hat{\beta}^{OLS} + \Phi^{-1}(\tau)\hat{\sigma}$, where $\Phi^{-1}(\tau)$ is the τth quantile of the standard normal distribution. If the errors are not identically distributed, the estimates of the conditional quantiles relying on OLS estimation will be incorrect.

Below we will analyze a food expenditure dataset to demonstrate the importance of directly estimating the conditional quantile function using the estimator in (2.1) and not relying on distributional assumptions of the errors. The food expenditure dataset contains 235 observations from 19th century Belgium working class households about annual income and annual food expenditure in Belgian francs [Koenker and Bassett Jr, 1982]. We will estimate the conditional distribution of food expenditure given annual income. Though not a healthcare dataset, this simple example illustrates the complexity of expenditure data in general and the need to estimate conditional quantiles directly. Figure 2.2 contains estimates of the 0.9 and 0.1 conditional quantiles from directly estimating the quantiles and from using the OLS method discussed above.

We first notice that the two methods result in two different estimates of the conditional quantiles for different incomes. The estimated quantiles from the OLS method

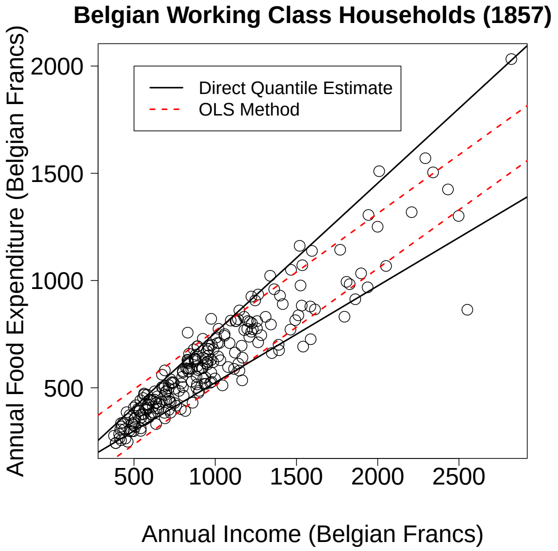

Figure 2.2: Estimated conditional 0.9 (upper lines) and 0.1 (lower lines) quantiles using direct estimates of the quantiles and OLS. The solid lines are the quantile estimates from (2.1) and the dashed lines are the quantile estimates obtained using OLS.

are parallel while the direct estimates are not. Directly estimating the quantiles allows for the effects of covariates to be different at different quantiles. From the definition of a quantile, we can expect that about 90% of the data is below the 0.9 quantile line and about 10% below the 0.1 quantile line. We see that the OLS estimated quantile lines bound the expenditures for low incomes and only contain the middle expenditures for large incomes. However, the directly estimated conditional quantiles maintain the correct proportion of the expenditures below the lines for all incomes. The quantile lines are not parallel because the variance is heteroscedastic.

2.2 Partially linear additive quantile regression

To incorporate nonlinear effects, we make use of the flexible partially linear additive quantile regression model. More specifically, we write $X_i = (V_i', Z_i')' \in \mathbb{R}^{p+q}$, where V_i

denotes a p-vector of covariates with linear effects and $Z_i = (Z_{i1}, \ldots, Z_{iq})'$ denotes a q-vector of covariates with nonlinear effects. The first element of V_i is 1 and corresponds to the intercept. The partially linear additive quantile regression model assumes that

$$Q_{Y_i|X_i}(\tau) = V_i'\beta(\tau) + \sum_{k=1}^{q} g_k(Z_{ik}), \tag{2.3}$$

where $g_k(\cdot)$ is an unknown smooth nonparametric function, $k = 1, \ldots, q$. For identifiability, it is often assumed that $E(g_k(Z_{ik})) = 0$. The semiparametric quantile regression models considered by He and Shi [1996], He et al. [2002], Wang et al. [2009], among others, are useful for incorporating nonlinearity while avoiding the curse of dimensionality.

To approximate the unknown nonparametric components $g_k(\cdot)$, we use a linear combination of basis spline (B-spline) functions. Schumaker [1981] details the construction and many properties of B-splines. Here we provide a description of the construction and relevant results about the approximations of B-spline basis functions. We assume that each covariate Z_{ik} is bounded above and below. We can then, without loss of generality, standardize the Z_{ik} covariates to be in the interval [0,1].

To define the B-spline functions, we first select a dregree r to use for the B-spline functions and the number of internal knots $m_n - 1$ used to divide the support of Z_{ik} into m_n intervals. The number of internal knots selected should grow with the sample size, but in practice a small integer works well. Then we place r knots on the lower and upper bound of the support. Let $t_1 \leq t_2 \leq \ldots \leq t_{2r+m_n-1}$ be the sequence of knots. It is common to choose the internal knots to create m_n equally spaced intervals or to correspond to $m_n - 1$ quantiles of Z_{ik}. This procedure results in a total of $J_n = r + m_n$ basis functions. The formula for the basis functions $b_1^r, \ldots, b_{j_n}^r$

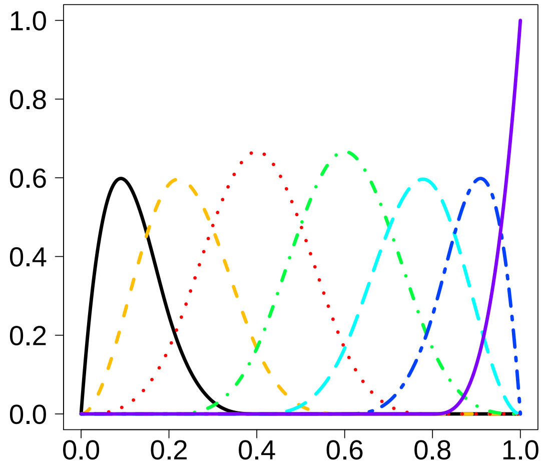

Figure 2.3: Plot of cubic B-splines with $J_n = 8$.

is defined recursively below:

$$
\begin{aligned}
b_1^r(z) &= \begin{cases} 1, & \text{if } t_i \leq z \leq t_{i+1}, \\ 0, & \text{otherwise }, \end{cases} \\
b_i^r(z) &= \frac{z - t_i}{t_{i+r-1} - t_i} b_i^{r-1}(z) + \frac{t_{i+r} - z}{t_{i+r} - t_{i+1}} b_{i+1}^{r-1}(z).
\end{aligned}
$$

Figure 2.3 displays eight cubic B-Splines on a support of $[0, 1]$ with four evenly placed internal knots.

Let $w(z) = (b_1(z), \ldots, b_{k_n+l+1}(z))'$ denote a vector of normalized B-spline basis functions of order $l + 1$ with k_n quasi-uniform internal knots on $[0, 1]$. Then $g_k(Z_{ik})$ can be approximated by $w(Z_{ik})'\xi_k$, where ξ_k are to be estimated from the data, $k = 1, \ldots, q$. The B-spline approximation is known to be flexible and computationally efficient. For simplicity, we use the same number of basis functions for all nonparametric components, but this is not necessary in practice.

To estimate the partially linear additive quantile regression model, we obtain

$$\left\{\hat{\beta}(\tau), \widehat{\xi}_1(\tau), \ldots, \widehat{\xi}_q(\tau)\right\} = \\ \underset{\{\beta,\xi_1,\ldots,\xi_q\}\in\mathbb{R}^{p+(k_n+l+1)q}}{\arg\min} \sum_{i=1}^{n} \rho_\tau \left[Y_i - \left\{V_i'\beta + \sum_{k=1}^{q} w(Z_{ik})'\xi_k\right\}\right]. \tag{2.4}$$

The estimator for the nonparametric function g_k is

$$\hat{g}_k(Z_{ik}) = w(Z_{ik})'\hat{\xi}_k(\tau) - n^{-1}\sum_{i=1}^{n} w(Z_{ik})'\hat{\xi}_k(\tau), \tag{2.5}$$

for $k = 1, \ldots, q$; where the centering is the sample analog of the identifiability condition $E[g_k(Z_{ik})] = 0$. In the sequel, we will omit the dependence on τ in notation for simplicity when the quantile level of interest is clear from the context. The asymptotic theory of the estimators is systematically investigated in Sherwood and Wang [2016]. For consistency, it is required that the number of basis functions $k_n \to \infty$, but in practice usually the choice of a small integer works well.

Many statistical software packages such as `R`, `SAS` and `STATA` can be adapted to obtain estimates of $\hat{\beta}$ and $\hat{g}_k(Z_{ik})$. To estimate the partially linear additive quantile regression model, we recommend using the `plaqr` function inside the `R` package `plaqr` we developed [Maidman, 2016]. Nonlinear effects can be plotted using the `nonlinEffect` and `plot` functions.

2.3 Quantile regression with missing data

Before describing existing methods for estimating conditional quantile function in the presence of missing data, we first need to define different types of missing data. There are three main types of missing data: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). The missing data is usually a subset of the covariates that is not always observed. For example, a variable indicating race is usually an optional question on surveys and thus may not be observed for every patient. A variable is MCAR if the probability of it not being observed does not depend on its value or on the values of any of the other variables (including the response). A variable is MAR if the probability of it not being observed only depends on the values of all or a subset of the always observed variables (including the response). A variable is MNAR if the probability of it not being observed depends on variables which are not always observed (including its own value).

To define the different types of missingness formally, we will first introduce some notation. Using notation defined previously in this chapter, we write $X_i = (s_i', m_i')'$ where s_i is always observed and m_i is the vector of sometimes missing covariates which are not always observed. Let $R_i = 1$ if m_i is fully observed and $R_i = 0$ otherwise. Let $T_i \subseteq (Y_i, X_i')'$ be the vector of always observed variables that effect the probability not fully observing m_i. We can now formally define the missing types of data below:

$$\begin{aligned} \text{(MCAR)} \qquad P(R_i = 1 \mid Y_i, X_i) &= P(R_i = 1), \\ \text{(MAR)} \qquad P(R_i = 1 \mid Y_i, X_i) &= P(R_i = 1 \mid T_i). \end{aligned}$$

For a variable that is MNAR, it is not possible to simplify $P(R_i = 1 \mid Y_i, X_i)$.

When data is MCAR, standard quantile regression techniques can be used on the subset of completely observed data. Though this results in a loss in efficiency because not all the subjects are used for estimation, there is no bias in the estimation. There are no methods for estimating the conditional quantile function when data is MNAR. Assuming the data is MCAR can sometimes be too strong of an assumption, so it is common to assume that missing data is MAR. We will focus our discussion of missing data for the MAR setting.

Two common techniques for handling MAR data are imputation and inverse probability weighting. Imputation is a technique that attempts to fill in the missing data so a "complete" dataset can be used for estimation. Research on imputation for quantile regression has begun only recently (see Wei et al. [2012] and Wei and Yang [2014]). Inverse probability weighting does not rely on estimating the missing data. Instead, the goal is to estimate $P(R_i = 1 \mid T_i)$ and assign weights to each fully observed case inversely to the estimate of $P(R_i = 1 \mid T_i)$. Robins et al. [1994] first used inverse probablity weighting to estimate the conditional mean when some covariates are MAR. Inverse probability weighting was extended to the linear quantile regression case [Sherwood et al., 2013] and later to the partially linear additive quantile regression model case [Sherwood, 2016].

To formally define the inverse probability weighting quantile regression estimator, first define $P(R_i = 1 \mid T_i) = \pi(T_i)$. We can obtain an estimate $\hat{\pi}(T_i)$ of $\pi(T_i)$ using logistic regression. We then can consistently estimate β by solving a weighted version of (2.1):

$$\hat{\beta}(\tau) = \operatorname*{arg\,min}_{\beta \in \mathcal{R}^p} \sum_{i=1}^{n} \frac{R_i}{\hat{\pi}(T_i)} \rho_\tau(Y_i - X_i'\beta). \tag{2.6}$$

Another kind of missing data can occur with longitudinal data. Consider a study on cognitive decline that expects to measure cognitive ability of a patients every year for ten years. Once a patient reaches a certain level of cognitive decline it is common for the patient to drop out of the study and cease returning for all future appointments. The probability of dropout usually depends on the value of the response and possibly other covariates as well. As a result, the dropout cannot be ignored. Lipsitz et al. [1997] originally proposed using inverse probability weighting where the probability of dropout needs to be estimated in order to estimate the conditional quantile function in the longitudinal setting with dropout. These results relied upon a heuristic explanation. A consistent estimator for longitudinal partially linear additive quantile regression is porposed in He et al. [2002]. Yi and He [2009] used weighted estimating equations to estimate the the conditional median in longitudinal studies with dropout. The weights used in their article are the inverse of the estimated probability of dropout. There remains a gap in the literature for estimating the conditional quantiles in a longitudinal model with dropout when not all covariates are linear. Chapter 4 seeks to fill in that gap and apply the method to analyzing cognitive decline and the potential onset of dementia.

Chapter 3

Semiparametric Method for Predicting High-Cost Patients

3.1 Introduction

In this chapter, we propose a new semiparametric prediction procedure using training data from the past one or two years to classify a patient's next-year expenditure into the class of "high-cost" or "not-high-cost". A threshold value c determined by a field expert, typically corresponding to a high quantile of the expenditure distribution, separates the two classes. This problem differs from the traditional classification problem in two important aspects. First, the actual values of the response variable on a continuous scale are available in the training data set, not solely class labels. Second, the two classes are severely imbalanced with high-cost patients in the minority. Ignoring the first issue results in efficiency loss; while ignoring the second issue results in a classification rule with low sensitivity, i.e. low probability of identifying the high-cost patients. An additional difficulty inherent in expenditure data is skewness and heteroscedasticity which pose challenges for statistical analysis [Zhou et al., 2001] and prediction at the tails of the distribution.

A popular approach in the literature for predicting if a new subject will be located in the tails of the response distribution relies on binomial regression using a logistic link function, e.g. Fleishman and Cohen [2010], Meenan et al. [1999], and Hosmer Jr and Lemeshow [2004]. Other link functions such as the complementary log-log function may be used as well. Given the threshold c, the binomial regression approach first artificially discretizes medical expenditure by assigning a value of 1 if the expenditure is greater than c and 0 otherwise. A binomial regression model is then fit to the 0-1 response data and a new patient can be classified as high-cost if his or her predicted probability of being a high-cost patient is more likely than not. By artificially dichotomizing the response, binomial regression results in efficiency loss and it is not clear whether the artificially modified data satisfy modeling assumptions.

Modeling English inpatient healthcare expenditure using the generalized beta distribution of the second kind and the generalized gamma distribution was found to have potential in predicting tail probabilities [Jones et al., 2015]. These methods can suffer from high variability without very large sample sizes. Bertsimas et al. [2008] took algorithmic approaches to predicting future healthcare expenditure using classification trees [Breiman et al., 1984] and clustering algorithms [Kannan et al., 2004]. While clustering algorithms are useful for identifying similar groups of patients, they cannot predict if a future patient belongs to a class defined a priori.

If 10% of all patients are high-cost, the naive classification rule that classifies every patient as not-high-cost has merely a 10% error rate. However, it completely misses the minority class of high-cost patients rendering it unsuitable for many applications [Vickers and Elkin, 2006]. Let r be the ratio of costs of a false positive (a not-high-cost patient predicted to be high-cost) and a false negative (a high-cost patient predicted to be not-high-cost). Simply taking $r = 1$ can result in classification rules with low sensitivity.

We propose a novel procedure that takes into account the missclassification error costs and leads to increased performance of sensitivity and overall classification. Our procedure uses training data to obtain a semiparametric estimation of the $\left(\frac{1}{1+r}\right)$th conditional quantile function. The classification rule amounts to comparing the $\left(\frac{1}{1+r}\right)$th conditional quantile of the response with the given threshold c. We show that this semiparametric procedure consistently estimates the Bayes rule. The new prediction procedure does not require dichotomization of the response and fully uses the information contained in the expenditure data. It does not require parametric distributional assumptions and is possibly more robust. The proposed procedure can be modified to create prediction intervals for future expenditure yielding richer information. In contrast, binomial regression provides little extra information beyond predicting whether the future expenditure is below or above the threshold.

In Section 3.2, we introduce the new semiparametric classification procedure and its connection to binomial regression. We demonstrate the performance of our new estimator with Monte Carlo simulations in Section 3.4. Section 3.5 reports a detailed analysis of MEPS. We conclude with a discussion in Section 3.6. Numerical results in Section 3.4 demonstrate that the new classification procedure is better able to correctly classify new patients, particularly high-cost patients, compared to existing parametric and algorithmic procedures. Proofs of theoretical results are included in Section 3.7.

3.2 New semiparametric prediction procedure

3.2.1 Bayes rule for classification

A patient is considered as high-cost if his or her next-year medical expenditure, denoted by Y, is greater than a predetermined threshold c. We consider a loss function that allows for unequal weighting of a false positive and a false negative. For a new patient with covariates X^*, let $\phi(X^*) \in \{1, -1\}$ be the prediction: -1 for not-high-cost and 1 for high-cost. The loss function of the decision rule $\phi(X^*)$ is

$$L(\phi(X^*)) = \begin{cases} r^{-1}, & \text{if } \phi(X^*) = -1 \text{ and } Y^* > c, \\ 1, & \text{if } \phi(X^*) = 1 \text{ and } Y^* \leq c, \\ 0, & \text{otherwise.} \end{cases}$$

Without loss of generality, the cost of a false positive is 1. Hence, r is the ratio of the cost of a false positive to a false negative. Taking $r = 1$ can result in classification rules with low sensitivity. Smaller ratios that weight the cost of a false negative heavier than that of a false positive (e.g. ratio of 4:1) result in classification rules with higher sensitivity. The ratio can be supplied by field experts or estimated from a pilot study. Similar to the threshold c, the ratio r is driven by the domain of application, not by the data.

The Bayes rule for classification minimizes the expected weighted 0-1 loss function,

$$\begin{aligned} E\left[L(\phi(X^*))\right] &= I\left(\phi(X^*) = 1\right)\left[1 - P(Y^* > c \mid X^*)(1 + r^{-1})\right] \\ &+ \; r^{-1} P(Y^* > c \mid X^*). \end{aligned}$$

It is straightforward to show that the decision rule $\phi(X^*)$ that minimizes $E\left[L(\phi(X^*))\right]$ is given by

$$\phi(X^*) = \begin{cases} 1, & \text{if } \; P(Y^* > c \mid X^*) > \frac{r}{1+r}, \\ -1, & \text{if } \; P(Y^* > c \mid X^*) \leq \frac{r}{1+r}. \end{cases} \tag{3.1}$$

3.2.2 The new prediction method

We want to classify a new patient with known predictors X^* as high-cost if $Y^* > c$. Note that the Bayes rule classifies a new patient with covariates $X = x^*$ as high-cost if

$P\left(Y^* > c \mid x^*\right) > \frac{r}{1+r}$. When $r = 1$ (equally weighted errors), the patient is classified as high-cost if $P\left(Y^* > c \mid x^*\right) > P\left(Y^* \leq c \mid x^*\right)$.

Our new approach can be viewed as a semiparametric method for estimating the Bayes rule without directly estimating the class probability $P\left(Y^* > c \mid x^*\right)$. This is based on the important observation that

$$\operatorname{sign}\left[P(Y^* > c \mid X^*) - \frac{r}{1+r}\right] = \operatorname{sign}\left[Q_{Y^*|X^*}\left(\frac{1}{1+r}\right) - c\right]. \tag{3.2}$$

This equivariance suggests that we can estimate the Bayes rule by obtaining a semiparametric estimate of $Q_{Y^*|X^*}\left(\frac{1}{1+r}\right)$ and comparing our estimate to the given threshold c. The approach is semiparametric in the sense that it does not assume a specific parametric distribution model for Y given X.

The classification rule is constructed from the training data $(Y_i, X_i')'$, $i = 1, \ldots, n$, and the observed vector of predictors $X^* = (V^{*'}, Z^{*'})'$ for the new patient in the following three step algorithm.

1. Fit model (2.3) on the training data and obtain $\hat{\beta}$ and $\hat{g}_k$, $k = 1, \ldots, q$ for $\tau = \frac{1}{1+r}$.

2. For the new patient, we estimate $\widehat{Q}_{Y^*|X^*}\left(\frac{1}{1+r}\right) = V^{*'}\hat{\beta} + \sum_{k=1}^{q} \hat{g}_k(Z_k^*)$.

3. Make the prediction: If $\widehat{Q}_{Y^*|X^*}\left(\frac{1}{1+r}\right) > c$, we classify the new patient as high-cost; otherwise, we classify him or her as not-high-cost.

Useful byproducts of the partially linear additive quantile regression model are prediction intervals. A $(1-\alpha) \times 100\%$ prediction interval for next-year expenditure for a new patient with predictors X^* is $\left(\widehat{Q}_{Y^*|X^*}\left(\alpha/2\right),\ \widehat{Q}_{Y^*|X^*}\left(1-\alpha/2\right)\right)$. Though not necessary for classifying a future patient, the prediction interval provides useful information for the analyst.

Many statistical software packages such as `R`, `SAS` and `STATA` can be adapted for the first step of the prediction procedure. We recommend using the `R` package `plaqr` we developed [Maidman, 2016]. A complete implementation of the prediction procedure using `plaqr` is given in Section 7.1.

3.2.3 Connection to the binomial regression approach

An alternative approach to this prediction problem relies on binomial regression with artificially dichotomized binary response variables. The underlying model with a logistic link function assumes that

$$\log\left(\frac{P\left(Y^* > c \mid X^*\right)}{1 - P\left(Y^* > c \mid X^*\right)}\right) = {X^*}'\alpha$$

and with a complementary log-log link function that

$$\log\left\{-\log\left[1 - P\left(Y^* > c \mid X^*\right)\right]\right\} = {X^*}'\alpha$$

for some unknown parameter α. In practice, α is usually estimated using the likelihood method to yield an estimator of the class probability $P\left(Y^* > c | X^*\right)$.

However, different from the ordinary binary classification problem for which only class labels are observed, in our setting, we also have complete information on the magnitude of the response variable. Our proposed semiparametric procedure fully uses the information in the response variable to make predictions. As binomial regression requires artificially dichotomizing the response variable, loss of information is expected.

3.3 Asymptotic properties

Note that (3.2) is stated for the unknown population conditional quantile function $Q_{Y^*|X^*}$. In order to prove large sample results for the proposed semiparametric procedure, we first state some conditions on the model.

We can write $Y_i = V_i^{'}\beta + \sum_{j=1}^{q} g_j(Z_{ij}) + \epsilon_i$, where the ϵ_i are independent and satisfy the constraint $P(\epsilon_i \leq 0|X_i) = \tau$, where $X_i = (V_i', Z_i')'$ with $V_i = (V_{i1}, \ldots, V_{ip})'$ and $Z_i = (Z_{i1}, \ldots, Z_{iq})'$.

Definition 3.3.1

Let $r \equiv m+v$, where m is a positive integer and $v \in (0, 1]$. Define $\mathcal{H}_r$ as the collection of functions $h(\cdot)$ on $[0, 1]$ whose mth derivative $h^{(m)}(\cdot)$ satisfies the Hölder condition of order v. That is, for any $h(\cdot) \in \mathcal{H}_r$, there exists some positive constant C such that

$$\left|h^{(m)}(z') - h^{(m)}(z)\right| \leq C|z' - z|^v, \quad \forall\ 0 \leq z', z \leq 1.$$

Definition 3.3.2

Given $Z = (Z_1, \ldots, Z_q)'$, the function $g(Z)$ is said to belong to the class of functions $\mathcal{G}$ if it has the representation $g(Z) = \alpha + \sum_{k=1}^{q} g_k(Z_k)$, $\alpha \in \mathcal{R}$, $g_k \in \mathcal{H}_r$ and $E[g_k(Z_k)] = 0$. □

Let $h_j^*(\cdot) = \underset{h_j(\cdot)\in\mathcal{G}}{\arg\inf} \sum_{i=1}^{n} E\left[f_i(0)(x_{ij} - h_j(Z_i))^2\right]$, where $f_i(\cdot)$ is the probability density function of ϵ_i given X_i. Let $m_j(Z) = E\left[x_{ij}|Z_i = Z\right]$, then it can be shown that $h_j^*(\cdot)$ is the weighted projection of $m_j(\cdot)$ into $\mathcal{G}$ under the L_2 norm, where the weights $f_i(0)$ are included to account for the possibly heterogeneous errors. Define $\delta_{ij} \equiv X_{ij} - h_j^*(Z_i)$. Let V be the $n \times p$ matrix of the linear covariates. Let H be the $n \times q$ matrix with the (i, j)th element $H_{ij} = h_j^*(Z_i)$, and write $V = H + \Delta$.

The following conditions are imposed for deriving the properties stated in Section 3.3. These conditions are similar to those in Sherwood and Wang [2016].

(C1) (Conditions on the random error) The random error ϵ_i has the conditional distribution function F_i and continuous conditional density function f_i . The f_i are uniformly bounded away from 0 and infinity in a neighborhood of zero and its first derivative f_i' has a uniform upper bound in a neighborhood of zero, for $1 \leq i \leq n$.

(C2) (Conditions on the covariates) There exist positive constants M_1 and M_2 such that $|V_{ij}| \leq M_1$, $\forall\ 1 \leq i \leq n$, $1 \leq j \leq p$ and $\mathrm{E}[\delta_{ij}^4] \leq M_2$, $\forall\ 1 \leq i \leq n$, $1 \leq j \leq q$. For a matrix X, define $\lambda_{max}(X)$ to be the maximum eigenvalue of X. There exist finite positive constants C_1 and C_2 such that with probability one

$$C_1 \leq \lambda max\left(n^{-1}VV'\right) \leq C_2, \quad C_1 \leq \lambda max\left(n^{-1}\Delta\Delta'\right) \leq C_2.$$

(C3) (Condition on the non-linear functions) For $r = m + v > 1.5$, $g_0 \in \mathcal{G}$.

(C4) (Condition on the B-Spline basis) The dimension of the spline basis k_n satisfies $k_n \approx n^{1/(2r+1)}$. and $n^{-1}k_n^3 = o(1)$.

Sherwood and Wang [2016] showed that in this setting, $\sqrt{n}\left(\hat{\beta} - \beta\right)$ converges in distribution to a Normal random variable and that $n^{-1}\sum_{i=1}^{n}\left(\sum_{k=1}^{q}\widehat{g}_k(z_{ik}) - \sum_{k=1}^{q}g_k(z_{ik})\right)^2 = O_p(n^{-1}k_n)$. In order to prove that our proposed procedure classifies a new patient consistently, we need to strengthen these results. First we will state a key lemma that proves that we are uniformly estimating the nonlinear functions accurately.

Lemma 3.3.3

For a model and function $g_1, \dots, g_q$ satisfying conditions (C1)-(C4),

$$\sup_{z \in [0,1]^q} \left| \sum_{k=1}^{q} \widehat{g}_k(z_k) - \sum_{k=1}^{q} g_k(z_k) \right| = o_p(1).$$

In other words, the difference between the estimated nonlinear functions and the true value of the nonlinear functions goes to zero uniformly as the sample size increases to infinity. This property is essential to ensure that $\widehat{Q}_{Y^*|X^*}(\tau)$ accurately estimates $Q_{Y^*|X^*}(\tau)$, hence the sign function can be predicted correctly with probability approaching one. This leads directly to the theorem proving consistency of the proposed procedure.

Theorem 3.3.4

Under Conditions (C1)-(C4),

$$\text{sign}\left[\widehat{Q}_{Y^*|X^*}(\tau) - c\right] = \text{sign}\left[P\left(Y^* > c|X^*\right) - \frac{r}{1+r}\right] + o_p(1). \qquad (3.3)$$

□

Hence, the proposed semiparametric procedure consistently estimates the Bayes rule. The proofs of Lemma 3.3.3 and Theorem 3.3.4 are included in Section 3.7.

3.4 Monte Carlo studies

3.4.1 Simulation setup

We compare our proposed new method (denoted by PLAQR) with five alternative parametric or semiparametric procedures, linear logistic regression (LLOG), partially linear additive logistic regression (PLALOG), linear complementary log-log regression (LCLOG), partially linear additive complementary log-log regression (PLACLOG), and the proposed prediction algorithm using classical linear quantile regression (LQR) [Koenker and Bassett, 1978], as well as a classification tree (TREE) in Monte Carlo experiments. The classification tree procedure incorporates different choices of r by treating the unequal cost of errors as a priori known class probabilities [Breiman et al., 1984] and is implementable in many software packages. For the binomial regression and classification tree approaches, the continuous response is dichotomized using a predetermined threshold c. The simulation results are based on 10,000 runs.

Mimicking the setting of the real data example in Section 5, we generate the response variable, next-year expenditure Y from the following model

$$Y = \exp\left(3V_1 + 1.5V_2 + 2V_3 + b\left[\sin\left(2\pi Z_1\right) + Z_2^3\right] + \epsilon\right), \tag{3.4}$$

where $V_1 \sim \text{Bernoulli}(0.5)$, $V_2, V_3 \overset{iid}{\sim} N(0,1)$, $Z_1 \sim \text{Uniform}(0,1)$, and
$Z_2 \sim \text{Uniform}(-1,1)$. We consider three different choices for the random error distributions: (1) $\epsilon \sim N(0,1)$, (2) $\epsilon \sim t_3$, and (3) $\epsilon \sim V_2\xi$, where $\xi \sim N(0,1)$. Case (2) corresponds to a heavy-tailed error distribution, and case (3) corresponds to a heteroscedastic error distribution. We consider four different choices of b: 1, 2, 3, and 5, which provide varying magnitudes of nonlinearity. In each simulation scenario, the size of the training and testing data are both 200. A new patient is referred to as high-cost if his or her expenditure exceeds a threshold c. Here, we consider a choice of c corresponding to approximately the marginal 0.9 quantile of Y.

Step (1) of the prediction procedure described in Section 3.2 requires estimating a partially linear additive (or linear for LQR) quantile regression model from the training data. Geraci and Jones [2015] proposes a one-parameter symmetric monotonic transformation of the response to achieve linearity for $\mathbb{R}^+$ valued responses. In each iteration, we estimate the transformation parameter for each value value of τ. Letting $\tilde{Y}$ denote the transformed response, we estimate the quantile function for $\tilde{Y}$.

To simplify computations, transformations for the PLAQR procedure are estimated for the model with three basis functions.

Motivated by Lee et al. [2014], we select the order of the basis functions m used to approximate each nonlinear component in the partially linear additive quantile regression model by minimizing

$$\begin{aligned} \text{BIC}(m) &= \log\left(\sum_{i=1}^{n} \rho_\tau\left(\tilde{Y}_i - \left[V_i'\hat{\beta}^{(m)} + \sum_{k=1}^{q} \pi(Z_{ik})^{(m)'}\xi_k^{(m)}\right]\right)\right) \\ &\quad + (p+1+qm)\frac{\log n}{2n}, \end{aligned}$$

where the superscript (m) denotes estimates from the model with basis functions of order m.

3.4.2 Simulation results

Different procedures are compared by plotting modified decision curves (see Vickers and Elkin [2006]). For each procedure and choice of r, let t_h and f_h denote the number of correctly and incorrectly predicted high-cost patients, respectively, and n denote the size of the prediction set. The decision curves are a plot of the net benefit for each prediction procedure:

$$\text{Net Benefit} \quad = \quad \frac{t_h}{n} - \frac{f_h}{n}r.$$

This measure reflects the simultaneous goals of achieving high sensitivity and high specificity by weighting the number of false positives by the relative cost of an error, r. Higher values indicate better prediction performance. We consider nine choices of r: $1, 9/11, 8/12, 7/13, 6/14, 5/15, 4/16, 3/17,$ and $2/18$ (corresponding to $\tau = .50, .55, .60, \ldots, .90$, respectively), reflecting situations when the cost of misclassifying a high-cost patient is equal to nine times higher than misclassifying a not-high-cost cost patient.

We report the decision curves for all values of b and the three types of errors in Figure 3.1. The decision curves for all values of b follow similar patterns. We summarize the major observations below for the case when $b = 5$.

First, we observe the importance of incorporating the nonlinear covariate effects. The more flexible semiparametric approach to classification outperforms the linear model based approaches and the classification tree, resulting in larger net benefit. Even when the nonlinear effects are milder (b=1 or 2), we observe the semiparametric models outperforming the linear models and classification tree. As the magnitude of nonlinearity increases, the increase in net benefit using the semiparametric approach becomes more evident.

Second, we observe that when the main interest is to predict if a future observation belongs to a small class, it is important to consider different weights for a false positive and a false negative in order to increase sensitivity. The increased sensitivity does not necessarily come at the cost of dramatically reduced specificity. When $r = 2/18$, our proposed new semiparametric procedure achieves a fine balance between sensitivity and specificity, resulting in the largest net benefit.

Finally, the most interesting and important observation is that PLAQR, PLALOG, and PLACLOG all perform similarly with respect to specificity; but PLAQR has higher sensitivity, particularly when $r = 2/18$. The poor performance of TREE can be explained by its low sensitivity for all choices of r, making it unusable in application. To better understand the relative performance of PLAQR versus PLALOG and PLACLOG, we consider a hypothetical situation in which 10,000 patients need to be classified as high-cost or not-high-cost, of which 1,000 are high-cost. When $b = 5$ with heteroscedastic errors and $r = 2/18$, PLAQR has mean sensitivity (SN) 0.981 and mean specificity (SP) 0.958, PLALOG has SN=0.864 and SP=0.964, and

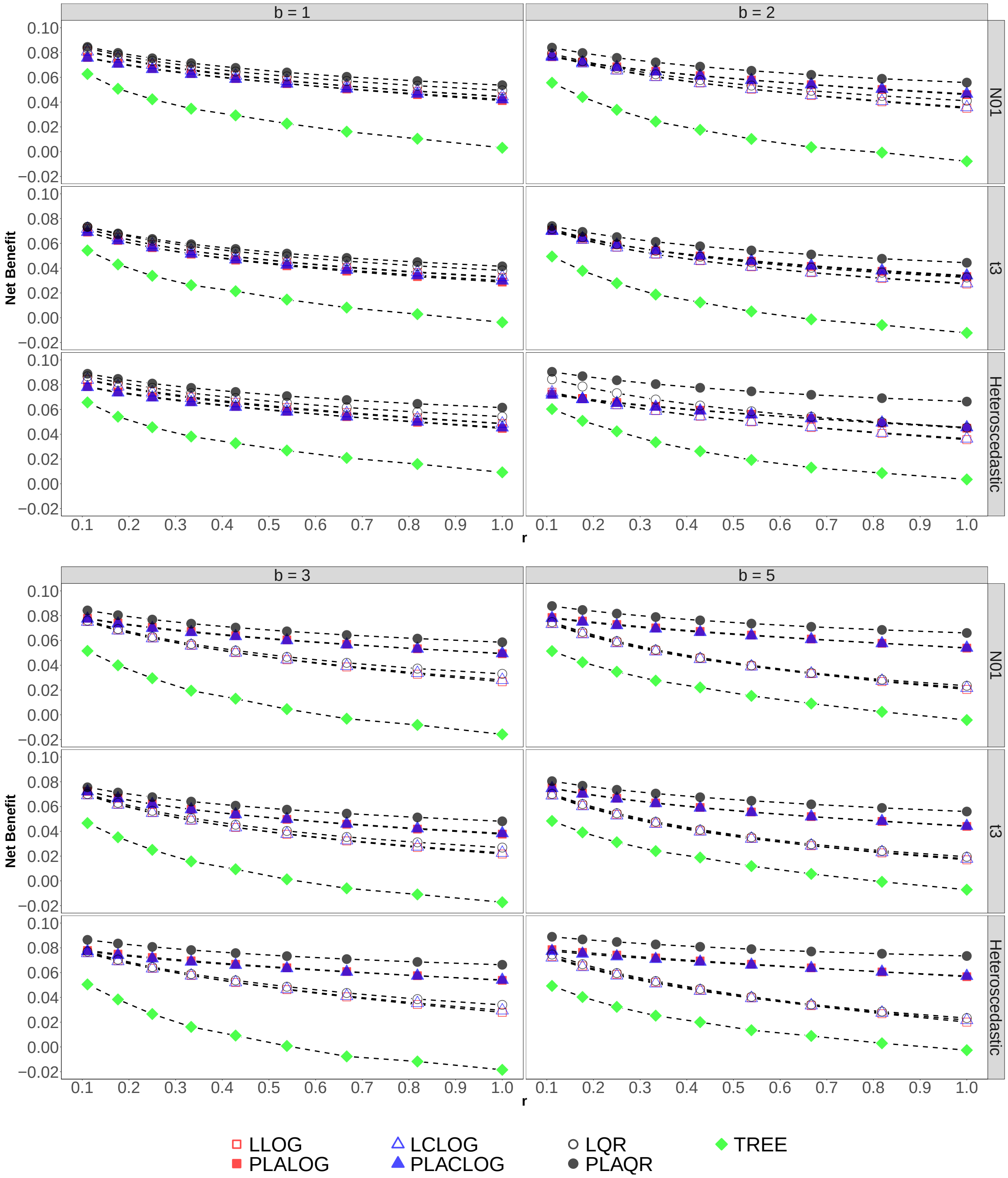

Figure 3.1: Decision curves for the LLOG, PLALOG, LCLOG, PLACLOG, LQR, TREE, and PLAQR procedures for simulations with standard normal errors, t_3 errors, and heteroscedastic errors when $b = 1, 2, 3, 5$.All standard errors are less than 2.7×10^{-4}.

PLACLOG has SN=0.851 and SP=0.966. Translating these results into the above hypothetical setting, PLAQR predicts 19 false negatives and 378 false positives; while PLALOG predicts 136 false negatives and 324 false positives and PLACLOG predicts 149 false negatives and 306 false positives. Hence, applying PLALOG or PLACLOG results in 117 or 130 more high-cost patients falsely predicted as not-high-cost. With normal errors, applying PLALOG or PLACLOG results in 116 or 121 more high-cost patients being misclassified. With t_3 errors, applying PLALOG or PLACLOG results in 71 or 74 more high-cost patients being misclassified.

3.4.3 Sensitivity analysis

In the following we perform a sensitivity analysis to investigate the performance of the proposed semiparametric classification method when the underlying model does not have additive nonlinear effects. In particular, we consider responses generated from the log-linear model

$$Y = \exp\left(3V_1 + 1.5V_2 + 2V_3 + Z_1 + Z_2 + \epsilon\right),$$

and the model with nonadditive effects on the log scale

$$Y = \exp\left(3V_1 + 1.5V_2 + 2V_3 + Z_1 + Z_2 + Z_1Z_2 + \epsilon\right),$$

where the covariates, errors, training and testing data sample sizes, and number of iterations are the same as in Section 3.4.1. The log-linear model is a special case of the partially linear additive assumption while the nonadditive effects model violates it. We compare our proposed method with the correctly specified quantile based procedure (denoted ORACLE_QR). The one-parameter symmetric monotonic transformation is used to estimate transformations [Geraci and Jones, 2015]. Decision

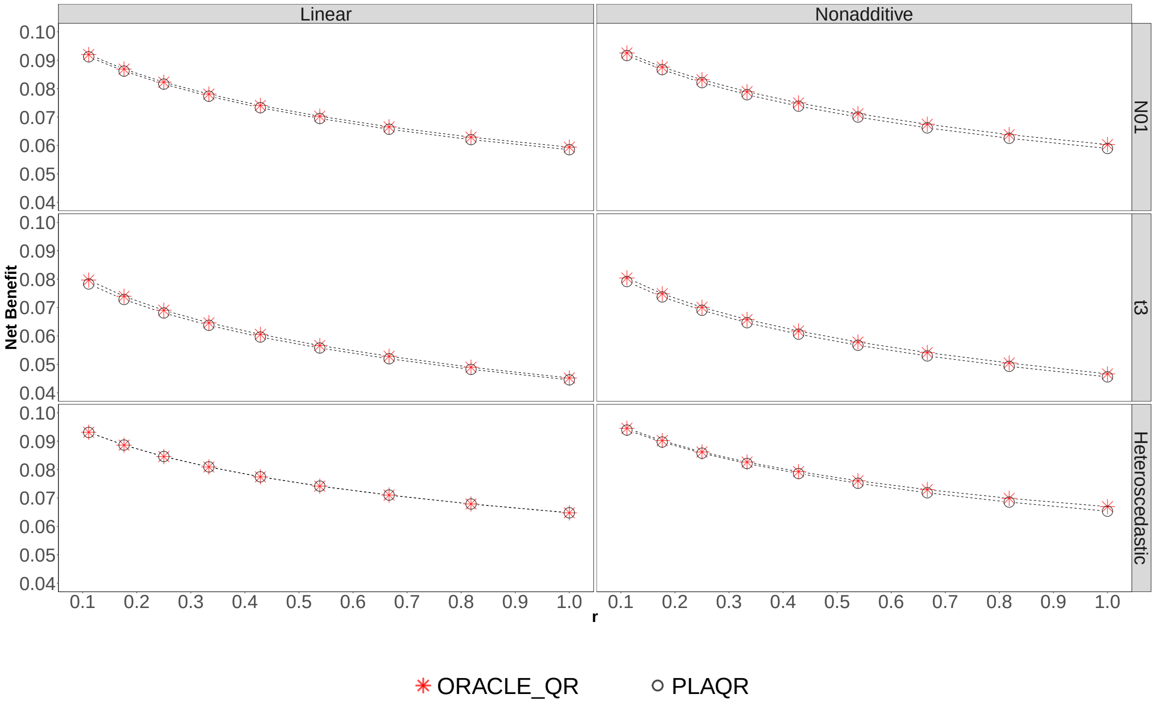

Figure 3.2: Decision curves for the ORACLE_QR and PLAQR procedures for log-linear and log-nonadditive model simulations with standard normal errors, t_3 errors, and heteroscedastic errors.All standard errors are less than 2.3×10^{-4}.

curves are plotted in Figure 3.2.

When all effects are linear on the log scale, PLAQR and ORACLE_QR have almost identical estimated net benefits for all three errors. It is not surprising that PLAQR performs nearly as well as ORACLE_QR in this setting because the class of partially linear additive models contains the class of linear models. Even with nonadditive effects on the log scale, PLAQR only has slightly lower net benefit than ORACLE_QR. The results from this sensitivity analysis suggest that our proposed semiparametric classification procedure works well even when the model does not contain nonlinear effects or has nonadditive effects.

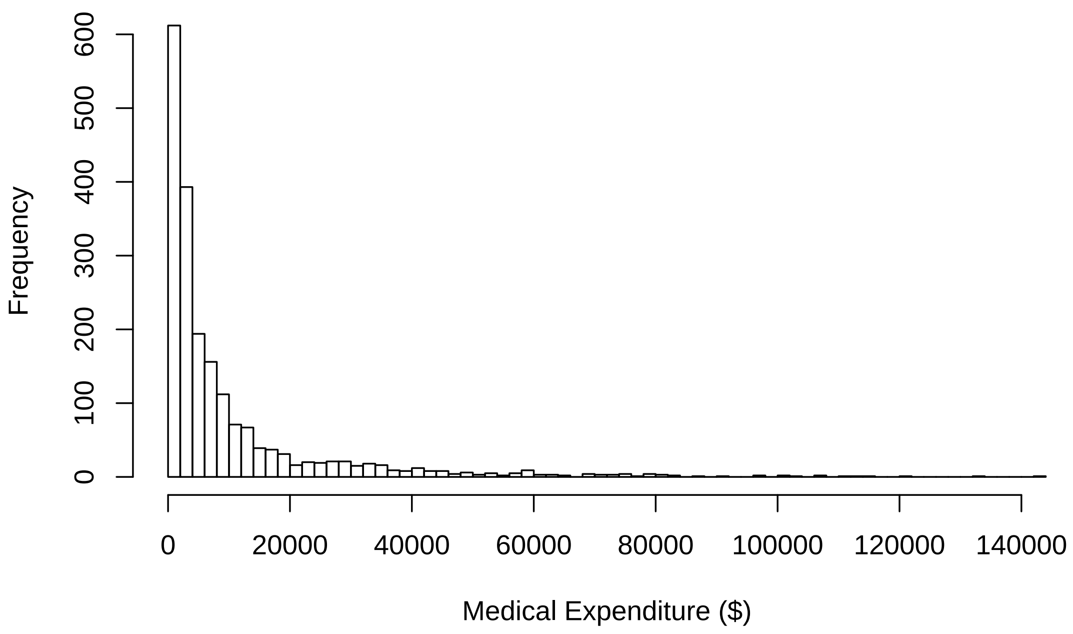

Figure 3.3: A histogram of next-year medical expenditures (second year of Panels 1, 2, and 3).

3.5 Analysis of Medical Expenditure Panel Survey

We now apply the proposed procedure to analyze medical expenditure from MEPS. Each panel consists of data from an individual over a two year span. In our analysis, we consider 1,985 male patients aged 65 or older in Panels 1, 2, and 3 from years 2006-2007 (724 patients), 2007-2008 (568 patients), and 2008-2009 (693 patients), respectively. We use the data from Panels 1 and 2 to predict if patients in Panel 3 will be high-cost in 2009. A threshold of US $28,520 corresponding to the marginal approximate 0.9 quantile of the next-year expenditure in Panel 3 is used to define patients as high-cost or not-high-cost.

Next-year expenditure among the three panels ranges from US $0 to US $314,400 (mean and median are US $10,110 and US $3,900, respectively). Nearly all of next-year expenditures are less than US $150,000 and about 4% of next-year expenditures are US $0. The first and third quantiles are US $1,539 and US $10,586, respectively. A histogram of next-year expenditure excluding the one expenditure greater than US $150,000 (to obtain sufficient resolution on the x-axis) is given in Figure 3.3. Its distribution is highly skewed. We use the following eight predictors observed in the first year of each panel: rgn (region of the country: northeast, midwest, south, west), insr (type of medical insurance: Medicare, private, Medicaid, uninsured), chrnc (number of chronic conditions: 0,1,...,8,9^+), prscrpt (number of prescriptions: 0,1,2,3,4^+), er (number of visits to the emergency room), health (summary score of self-described physical health), age, and rrs (relative risk adjustment score to account for inflation). The relative risk adjustment score, rrs, is a prospective measure of disease burden relying on health condition categories. Studies have shown that individuals with higher relative risk scores go on to use more hospital resources. These variables are important in the medical cost literature for their predictive power [Fleishman and Cohen, 2010].

First, we compare the prediction performance of the seven procedures LLOG, PLALOG, LCLOG, PLACLOG, TREE, LQR, and PLAQR discussed in Section 3.4. For each of the seven procedures, we use the training data to fit the prediction model. To reflect the panel-to-panel changes in the next-year expenditure distribution and the goal of predicting patients with next-year expenditure greater than US $28,520 in Panel 3, we artificially dichotomize the next-year expenditure in Panels 1 and 2 according to their respective marginal approximate 0.9 quantiles (US $29,630 and US $24,000) for the binomial regression and TREE procedures. We assume nonlinear effects for age and rrs based on exploratory data analysis.

Transformations for the quantile regression procedures require strictly positive responses. Because some patients in the training data have US $0 expenditure, we add 1 to each response and apply the recommended one-parameter symmetric transformation [Geraci and Jones, 2015] for each value of τ under consideration.

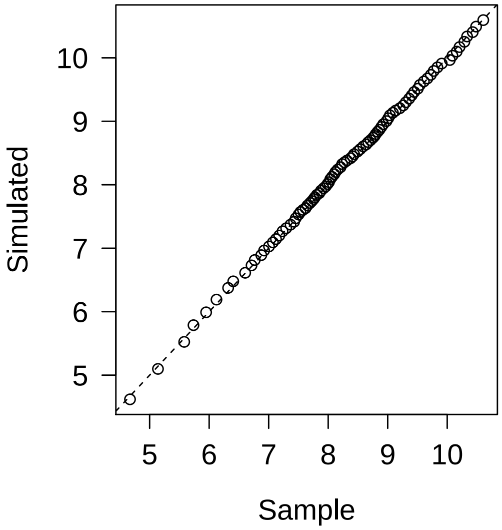

Figure 3.4: Lack-of-fit diagnostic QQ plot for PLAQR.

The 95% bootstrap confidence intervals for the transformation parameters suggest that the transformation $\tilde{y}_i \equiv \log(y_i + 1)$ is appropriate for all quantiles. By the equivariance property of quantile regression, the conditional quantile of y is given by $Q_{Y|X,Z}(\tau) = \exp\left(Q_{\tilde{Y}|X,Z}(\tau)\right) - 1$.

We assess the overall lack-of-fit for the PLAQR model via the simulation based graphical method proposed by Wei et al. [2006a]. More specifically, we generate a random $\tilde{\tau}$ from the Uniform(0,1) distribution and estimate $\widehat{Q}_{Y|X}(\tilde{\tau})$ for a randomly sampled X in the training data. We repeat this process 5000 times to produce 5000 simulated responses from the assumed model and plot the quantiles of the sample responses against the quantiles of the simulated respones in Figure 3.4. The points in the QQ plot fall nearly along the identity line suggesting no lack-of-fit.

We evaluate the performance of all seven procedures for choices of r ranging from 1/9 to 1. When $r = 1$ ($\tau = .5$) none of the seven procedures is able to accurately predict high-cost patients. For smaller choices of r, the prediction procedures achieve a better balance of sensitivity and specificity. When $r = 1/9$ ($\tau = .9$), the procedures

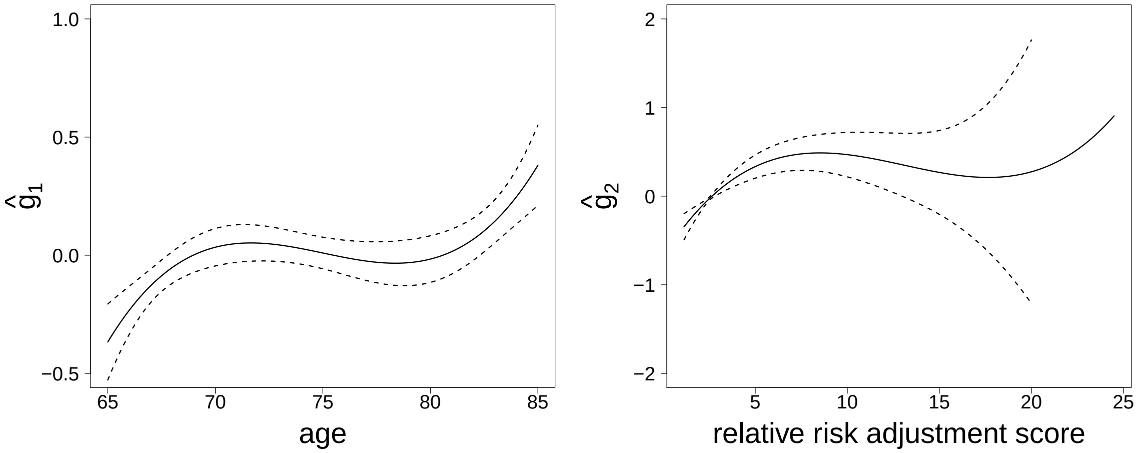

Figure 3.5: Plots of the estimated nonlinear effects.

identify high-cost patients at an acceptable rate without sacrificing much ability to identify not-high-cost patients. Sensitivity from the PLACLOG procedure was 0.671 and from the PLAQR procedure 0.729. PLAQR is able to correctly predict 5.8% more high-cost patients than PLACLOG while maintaining adequate specificity. The specificities of PLACLOG and PLAQR were 0.713 and 0.724, respectively. PLALOG had a sensitivity and specificity of 0.657 and 0.713, respectively. Consistent with findings in Section 3.4.2, the TREE procedure's low sensitivity rendered it inviable as a prediction procedure.

To better understand the practical importance of this increased sensitivity, consider that the subpopulation of males aged 65 and older in the U.S. in 2014 was about 20 million [U.S. Census Bureau, 2016]. If about 10% of patients had high-cost medical expenditure, then PLAQR correctly identifies about 116,000 more high-cost patients than PLACLOG while correctly identifying slightly more not-high-cost patients.

Next, to gain more insight into this data, we further explore the estimated conditional 0.9 quantile of next-year expenditure in Panel 3 using the partially linear additive quantile regression model. The estimated coefficients for the linear effects and the estimated nonlinear functions $\hat{g}_1$ and $\hat{g}_2$ are given in Table 3.1 and Figure 3.5,

Table 3.1: Coefficient estimates of linear effects for MEPS model when $\tau = 0.9$ (90% confidence intervals in parentheses).

Coefficient	Estimate	
(Intercept)	8.088	(7.128, 8.913)
rgn_{MW}	0.163	$(-0.445,\ 0.124)$
rgn_{S}	-0.048	$(-0.337,\ 0.188)$
rgn_{W}	-0.332	$(-0.614,\ -0.093)$
$\text{insr}_{\text{prvt}}$	0.298	$(-0.534,\ 0.727)$
$\text{insr}_{\text{Mdcd}}$	-3.634	$(-13.806,\ -2.519)$
$\text{insr}_{\text{unin}}$	0.683	$(-0.904,\ 2.056)$
chrnc_1	2.198	(1.532, 3.348)
chrnc_2	2.700	(1.681, 3.649)
chrnc_3	2.279	(1.606, 3.296)
chrnc_4	2.582	(1.834, 3.580)
chrnc_5	2.768	(2.060, 3.831)
chrnc_6	3.093	(2.359, 4.160)
chrnc_7	2.856	(2.030, 3.950)
chrnc_8	2.696	(1.993, 3.865)
chrnc_{9+}	2.893	(2.101, 3.994)
prscrpt_1	-0.844	$(-1.664,\ -0.388)$
prscrpt_2	-0.972	$(-1.608,\ -0.481)$
prscrpt_3	-1.076	$(-1.747,\ -0.590)$
prscrpt_{4+}	-0.884	$(-1.567,\ -0.387)$
er	0.079	$(-0.045,\ 0.201)$
health	-0.013	$(-0.022,\ -0.006)$

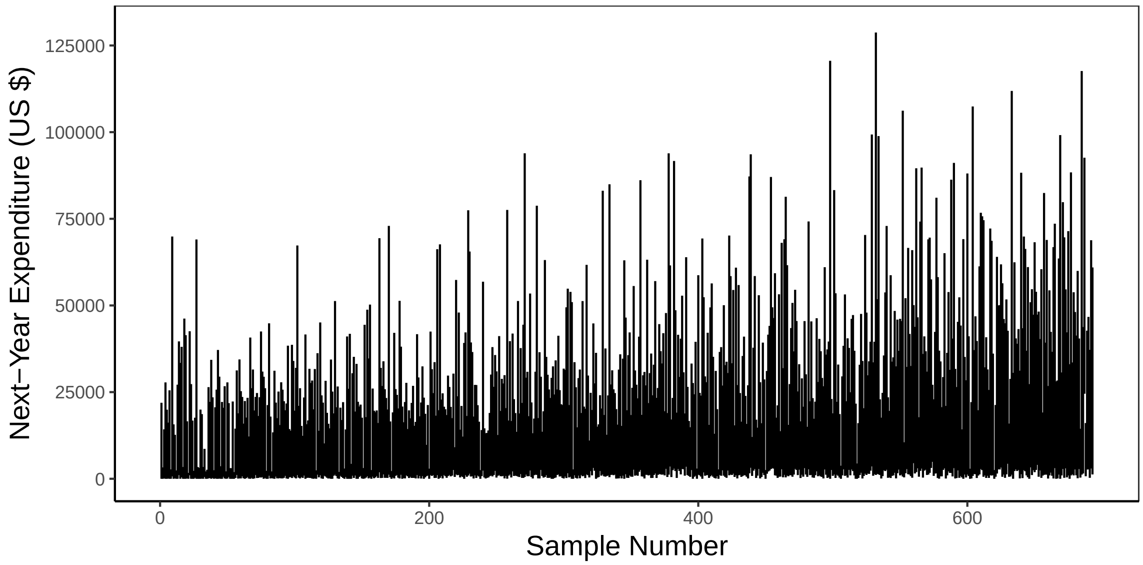

Figure 3.6: 90% prediction intervals for next-year expenditure in Panel 3 of MEPS.

respectively. Dashed lines are one standard deviation above and below the estimated effects. The pointwise standard deviations and confidence intervals are estimated from 999 bootstrapped samples using the wild bootstrap [Feng et al., 2011]. Dashed lines in the plot of $\hat{g}_2$ do not cover the whole range of observed relative risk adjustment score due to sparsity in the large values of the observed relative risk adjustment score causing error estimation to be difficult and untrustworthy.

We conclude our analysis of MEPS by investigating prediction intervals. We computed and plotted 90% prediction intervals for patients' next-year expenditure in Panel 3 in Figure 3.6. About 87% of the prediction intervals cover the true next-year expenditure for each patient. As an example, consider a typical patient with rgn = northeast, insr = Medicare, chrnc = 4, prscrpt = 3, er = 0, age = 75, rrs = 2.5, and health = 55. The 90% prediction interval of this patient's next-year medical expenditure is (US $\$2,550,$ US $\$47,728$) with a predicted 0.9 quantile of US $\$38,155$.

3.6 Discussion

Motivated by a real data application to identify potential future high-cost patients, we propose a new semiparametric procedure to predict whether a new response falls in the tail of the response variable distribution. We prove that the proposed semiparametric procedure is a consistent estimator of the Bayes rule for classification while avoiding estimating the class probability. Empirically, we show that the proposed procedure outperforms popular binomial regression and classification tree based classification procedures. Furthermore, the semiparametric approach incorporates nonlinear covariate effects. As suggested by simulation results, ignoring nonlinear effects may substantially increase the misclassification error rates.

In the real data application, we formulate the problem as a binary prediction problem as the intervention policy (whether to introduce an intervention program) only depends on whether the patient's future expenditure falls in the upper tail of the expenditure distribution. We then consider a decision theory framework to minimize the loss due to misclassification, where the two types of misclassification errors are weighted according to their potential consequences. If we can estimate the effect of the intervention as a percentage of the potential spending, then it is possible to formulate the decision theory framework as in Section 2.3 of Ehm et al. [2016] to take into account the magnitude of gains and losses. This approach will be useful in the future when information about medical expenditure reductions as a result of policy changes is available, for example, from a pilot program. This will be an interesting future research direction.

3.7 Proofs

Proof of Lemma 3.3.3

To facilitate the proof, we will make use of the theoretically centered B-spline basis functions (e.g., Xue and Yang [2006]). More specifically, we consider the B-spline basis functions $b_j(\cdot)$ in Section 2 and let $B_j(z_{ik}) = b_{j+1}(Z_{ik}) - \frac{E[b_{j+1}(Z_{ik})]}{E[b_1(Z_{ik})]} b_1(Z_{ik})$ for $j = 1, \ldots, k_n + l$. Then $E(B_j(Z_{ik})) = 0$. For a given covariate Z_{ik}, let $\mathbf{w}(Z_{ik}) = (B_1(Z_{ik}), ..., B_{k_n+l}(Z_{ik}))'$ be the vector of basis functions, and $\mathbf{W}(Z_i)$ denote the J_n-dimensional vector $\left(k_n^{-1/2}, \mathbf{w}(Z_{i1})', ..., \mathbf{w}(Z_{iq})'\right)'$, where $J_n = q(k_n + l) + 1$.

By the result of Schumaker [1981] (p. 227), there exists a vector $\boldsymbol{\gamma}_0 \in \mathcal{R}^{J_n}$ and a positive constant C_0, such that $\sup_{t\in[0,1]^d} |\sum_{k=1}^q g_k(\mathbf{t}) - \mathbf{W}(\mathbf{t})'\boldsymbol{\gamma}_0| \le C_0 k_n^{-r}$. Let

$$(\hat{\mathbf{c}}_{\mathbf{1}}, \hat{\boldsymbol{\gamma}}) = \underset{(\mathbf{c}_1, \boldsymbol{\gamma})}{\operatorname{argmin}}\, n^{-1} \sum_{i=1}^n \rho_\tau\big(Y_i - V_i'\mathbf{c}_1 - \mathbf{W}(Z_i)'\boldsymbol{\gamma}\big). \tag{3.5}$$

We write $\boldsymbol{\gamma} = (\gamma_0, \boldsymbol{\gamma}_1', \ldots, \boldsymbol{\gamma}_d')'$, where $\gamma_0 \in \mathcal{R}$, $\boldsymbol{\gamma}_j \in \mathcal{R}^{k_n+l}$, $j = 1, \ldots, d$; and we write $\hat{\boldsymbol{\gamma}} = (\hat{\gamma}_0, \hat{\boldsymbol{\gamma}}_1', \ldots, \hat{\boldsymbol{\gamma}}_d')'$ the same fashion. Let $\tilde{g}_j(Z_{ij}) = w(Z_{ij})'\hat{\boldsymbol{\gamma}}_j$ be the estimator of g_j, $j = 1, ..., q$. Let $\tilde{g}(Z_i) = \mathbf{W}(Z_i)'\hat{\boldsymbol{\gamma}} = \tilde{g}_0 + \sum_{j=1}^q \tilde{g}_j(Z_{ij})$; and $\hat{g}(Z_i) = \sum_{j=1}^q \widehat{g}_j(Z_{ij})$. It can be derived that $\hat{g} = \tilde{g}$.

$$\begin{aligned} \sup_z \Big|\sum_{j=1}^q \widehat{g}_j(z) - \sum_{j=1}^q g_j(z)\Big| &= \sup_z \Big|\tilde{g}(z) - \sum_{j=1}^q g_j(z)\Big| \\ &= \sup_z \Big|\mathbf{W}(z)'(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)\Big| \\ &\le \sup_z ||\mathbf{W}(z)|| \cdot ||\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0||. \end{aligned}$$

Let $B_n = \text{diag}(f_1(0), \ldots, f_n(0))$ be the $n \times n$ diagonal matrix; $W = (\mathbf{W}(Z_1), \ldots, \mathbf{W}(Z_n))' \in \mathbb{R}^{n\times J_n}$, and $W_B^2 = W'B_nW \in \mathbb{R}^{J_n\times J_n}$. It follows from

Sherwood and Wang [2016] that $||W_B(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)|| = O_p(\sqrt{k_n})|$. Hence,

$$\begin{aligned} ||\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0|| &= ||W_B^{-1} W_B(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)|| \\ &\leq \sqrt{k_n n^{-1}} O_p(k_n) = O_p(k_n^{3/2} n^{-1/2}). \end{aligned}$$

In our setting, $\sup_z ||\mathbf{W}(z)|| = O_p(1)$.
Thus $\sup_z \left| \sum_{j=1}^q \widehat{g}_j(z) - \sum_{j=1}^q g_j(z) \right| = O_p(k_n^{3/2} n^{-1/2}) = o_p(1)$. Hence 3.3.3 is verified.

□

Proof of Theorem 3.3.4

The proof of 3.3.4 is a direct consequence of 3.3.3. We have

$$\begin{aligned} & \text{sign}[\hat{Q}_{Y^*|X^*}(0.5) - c] \\ = \ & \text{sign}[Q_{Y^*|X^*}(0.5) - c + V_i^{*'}(\hat{\beta} - \beta) + \sum_{j=1}^{q}(\hat{g}_j(Z_{ij}) - g_j(Z_{ij}))] \\ = \ & \text{sign}[Q_{Y^*|X^*}(0.5) - c] + o_p(1) \end{aligned}$$

since $\hat{\beta}$ and $\hat{g}_j$, $j = 1, \dots, q$, are estimated on the training data and are independent of (Y^*, X^*). □

Chapter 4

Longitudinal Quantile Regression with Dropout

4.1 Introduction

Many datasets in healthcare arise from longitudinal studies in which the same subject is measured repeatedly over time. For example, the Uniform Data Set (UDS) maintained by the National Alzheimer's Coordinating Center tracks patients' cognitive decline over a period of ten years. Subsets of the Medical Expenditure Panel Survey track individuals' healthcare expenditures over a period of time. A subset of the Medical Expenditure Panel Survey was analyzed in Chapter 3, but the data included only one previous observation and longitudinal techniques were not needed. Ignoring the longitudinal structure of the data can result in biased estimation. Missing data is another inherent challenge that can yield biased estimates if not handled properly. As discussed in previous chapters, semiparametric quantile regression is a popular tool for analysis. Despite this, the literature lacks a theoretically justified semiparametric quantile regression estimator for the longitudinal setting with missing data.

To properly refer to the longitudinal structure we will introduce some new notation. Following along the lines of the semiparametric model of (2.3), let

$$Y_{ij} \quad = \quad V_{ij}'\beta + \sum_{d=1}^{q} g_d\left(Z_{ijd}\right) + \varepsilon_{ij}, \tag{4.1}$$

where $i = 1, \ldots, n$ denotes the subjects, $j = 1, \ldots, m_i$ denotes the longitudinal structure of the observations, $d = 1, \ldots, q$ denotes the dth nonlinear function, and ε_{ij} is independent of ε_{kj} when $i \neq k$ and satisfies the quantile constraint $P(\varepsilon_{ij} < 0 \mid X_{ij}) = \tau$. He et al. [2002] showed that a slight modification of the estimator in (2.4) to

$$\left\{\hat{\beta}(\tau), \widehat{\xi}_1(\tau), \ldots, \widehat{\xi}_q(\tau)\right\} =$$
$$\underset{\{\beta,\xi_1,\ldots,\xi_q\}\in\mathbb{R}^{p+(k_n+l+1)q}}{\arg\min} \sum_{i=1}^{n}\sum_{j=1}^{m_i} \rho_\tau\left[Y_{ij} - \left\{V_{ij}'\beta + \sum_{k=1}^{q} w(Z_{ijk})'\xi_k\right\}\right]$$

is a consistent estimator of the conditional quantile function even if the dependence structure of the errors is not specified. This modification handles the longitudinal structure, but does not take into account any missingness.

In longitudinal studies, not every patient returns for repeated observations throughout the entire duration of the study. It is very common for a patient to miss a scheduled appointment and then drop out of the study [Hogan et al., 2004]. Of the 5350 patients in the UDS to attend the first appointment for observation, about 75% dropped out after four follow up visits. Only about 10% of patients remained in the study for eight follow up visits, and less than 2% of the original 5360 patients made it to all nine follow up visits.

Often the probability of dropout is not independent of other covariates. Variables like age or even the repsonse can increase or decrease the probability of dropout. To handle the missing data or dropout, we assume that the probability that an observation is missing only depends on data that is always observed. This assumption is called missing at random. Let $R_{ij} = 1$ if (Y_{ij}, X_{ij}') is completely observed and 0 otherwise. Let $t_{ij} \subseteq (Y_{ij}, X'_{ij})'$ and $T_{ij} = (t'_{i1}, \ldots, t'_{ij})'$. Formally, the missing at random assumption is that

$$P(R_{ij} = 1 \mid Y_{i1}, \ldots, Y_{ij}, X_{i1}, \ldots, X_{ij}) = P(R_{ij} = 1 \mid T_{ij}).$$

In the case where the data does not arise from a longitudinal study, Sherwood et al. [2013] proposed using weighted quantile regression to estimate the conditional linear quantile function. The weights used are the inverse of the estimated probability that $R_i = 1$ and are estimated using logistic regression. The intuition behind this idea being that an observation that is unlikely to be observed receives higher weighting to account for the other similar observations that were not completely observed. Inverse probability weighting was extended to the partially linear additive quantile regression model in Sherwood [2016].

In the longitudinal setting, Chen and Zhou [2011] proposed a doubly robust estimator for binary responses in that consistent estimators will be provided if either the missing data model or the missing covariate model is correctly specified. This approach was extended for generalized estimating equations with ordinal data [da Silva et al., 2015].

In the longitudinal setting with dropout, Lipsitz et al. [1997] used the inverse probability weighting method with weights being the inverse of the estimated probability of dropout. Yi and He [2009] took a similar approach in estimating the conditional median. In this chapter, we propose a theoretically justified estimator of the partially linear additive quantile regression model with dropout. Section 4.2 formally defines the dropout model and provides intuition for the proposed estimator. Asymptotic properties are presented in Section 4.3. We demonstrate the performance of our estimator with Monte Carlo sumulations in Section 4.4 and analyze the UDS in Section 4.5. We conclude with a discussion in Section 4.6 and relegate proofs to

Section 4.7.

4.2 Estimation

Inverse probability weighting is a two-step procedure. First, the probability that the ith patient is observed at the jth time point needs to be estimated. These estimated weights are then used to estimate the conditional quantile of the response. We will first discuss estimation of missingness.

4.2.1 Dropout

Let $\pi(T_{ij}) = P(R_{ij} = 1 \mid T_{ij})$ be the probability that the ijth data point is observed. To relax the notation, we write $\pi_{ij0} \equiv \pi(T_{ij})$. The inutition behind inverse probability weighting is that for every observed data point with probability π_{ij0} of being observed, $1/\pi_{ij0}$ data points with the same covariates are expected to be observed if there were no missing data. For example, an observed data point with $\pi_{ij0} = 1/2$ is given the weight of two observations. This accounts for the other observation with similar covariates that is not observed.

Another explanation for not ignoring missingness is to consider the naive estimator which only incorporates the observed data points. For simplicity, we will consider the case with only linear effects

$$\hat{\beta}^N(\tau) \quad = \quad \underset{\beta \in \mathbb{R}^p}{\arg\min} \sum_{i=1}^{n} \sum_{j=1}^{m} R_{ij} \rho_\tau \left[Y_{ij} - V_{ij}' \beta \right] . \tag{4.2}$$

Equation (4.2) implies that $\hat{\beta}^N(\tau)$ approximately solves the estimating equation

$$G^N(\beta) \quad = \quad \sum_{i=1}^{n} \sum_{j=1}^{m} R_{ij} V_{ij} \psi_\tau \left[Y_{ij} - V_{ij}' \beta \right] ,$$

where $\psi_\tau(u) = \tau - I(u < 0)$ is the gradient function of $\rho_\tau(u)$. Assuming covariates are missing at random,

$$\begin{aligned} & \mathrm{E}\left\{\sum_{i=1}^{n}\sum_{j=1}^{m} R_{ij}V_{ij}\psi_\tau\left(Y_{ij} - V'_{ij}\beta\right)\right\} \\ = & \sum_{i=1}^{n}\sum_{j=1}^{m} \mathrm{E}\left\{\mathrm{E}\left[R_{ij}V_{ij}\psi_\tau\left(Y_{ij} - V'_{ij}\beta\right) \mid T_{ij}\right]\right\} \\ = & \sum_{i=1}^{n}\sum_{j=1}^{m} \mathrm{E}\left\{\pi(T_{ij})V_{ij}\psi_\tau\left(Y_{ij} - V'_{ij}\beta\right)\right\}. \end{aligned}$$

Although $\mathrm{E}\left\{V_{ij}\psi_\tau\left(Y_{ij} - V'_{ij}\beta\right) \mid V_{ij}\right\} = 0$, because $\pi(T_{ij})$ is a function of Y_{ij}, generally $\mathrm{E}\left\{\pi(T_{ij})V_{ij}\psi_\tau\left(Y_{ij} - V'_{ij}\beta\right) \mid V_{ij}\right\} \neq 0$. When inverse probability weights are included, the estimating equation becomes

$$G(\beta) = \sum_{i=1}^{n}\sum_{j=1}^{m} \frac{R_{ij}}{\pi_{ij0}} V_{ij}\psi_\tau\left[Y_{ij} - V'_{ij}\beta\right].$$

The expectation of $G(\beta)$ is 0 by the same technique.

In the longitudinal model with dropout, the jth observation in the ith individual can only be observed if all previous observations are observed. We also assume that the first observation for each individual is always observed ($\pi_{i10} = 1$). Formally, for $2 \leq j \leq m$,

$$P\left(R_{ij} = 1 \mid T_{ij}, \prod_{k=1}^{j-1} R_{ik} = 0\right) = 0.$$

Let $\eta_{ij0} = P(R_{ij} = 1 \mid t_{ij}, R_{i1} = \ldots = R_{ij-1} = 1)$. In this section we assume the probability can be modeled using a logistic regression, i.e.,

$$\eta_{ij0} \equiv \eta(t_{ij}, \gamma_{j0}) = \frac{\exp\left(t'_{ij}\gamma_{j0}\right)}{1 + \exp\left(t'_{ij}\gamma_{j0}\right)}.$$

We let $\hat{\eta}_{ij} \equiv \eta(t_{ij}, \hat{\gamma}_j)$ be the estimate of $P(R_{ij} = 1 \mid t_{ij}, R_{i1} = \ldots = R_{ij-1} = 1)$. We then have that $\pi_{ij0} = \prod_{k=1}^{j} \eta_{ij0}$ and we let $\hat{\pi}_{ij} = \prod_{k=1}^{j} \hat{\eta}_{ij}$ be the estimate of $P(R_{ij} = 1|T_{ij})$.

4.2.2 An unbiased estimator

The next step is to incorporate estimates of probabilities of completely observing a case into the estimation of a conditional quantile. Let $N_0 = \sum_{i=1}^{n} \sum_{j=1}^{m} R_{ij}$ be the total number of observations. Following the proposed estimators of He et al. [2002] for the longitudinal model and Sherwood [2016] for the independent model with missing covariates, we define the following estimator

$$\left\{\hat{\beta}^W(\tau), \widehat{\xi}_1^W(\tau), \ldots, \widehat{\xi}_q^W(\tau)\right\} = \\ \underset{\{\beta,\xi_1,\ldots,\xi_q\} \in \mathbb{R}^{p+(k_n+l+1)q}}{\arg\min} \sum_{i=1}^{n} \sum_{i=1}^{n} \frac{R_{ij}}{\hat{\pi}_{ij}} \rho_\tau \left[Y_{ij} - \left\{ V_{ij}'\beta + \sum_{k=1}^{q} w(Z_{ijk})'\xi_k \right\} \right]. \tag{4.3}$$

The estimator for the nonparametric function g_k is

$$\hat{g}_k(Z_{ijk}) = w(Z_{ijk})'\hat{\xi}_k(\tau) - N_0^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} R_{ij} w(Z_{ik})'\hat{\xi}_k(\tau), \tag{4.4}$$

for $k = 1, \ldots, q$. Lipsitz et al. [1997] proposed an inverse probability weighted estimator similar to our proposed estimator in Equation (4.3), but estimated the probability of dropout instead of the probability of observing the ijth data point. The asymptotic results of the next section still hold for this model as well.

4.3 Asymptotic properties

The estimator in Equation (4.3) is unbiased under similar conditions as the procedure proposed in Chapter 3. We will need an additional assumption on the probability of dropout and the estimator of the weights. We state the additional conditions and restate the conditions from Chapter 3 for clarity.

Define the set $\mathcal{H}_r^q = \{\sum_{d=1}^q h_k(z) \mid h_k \in \mathcal{H}_r^q\}$ and

$$h_k^*(\cdot) \quad = \quad \arg\inf_{h_k \in \mathcal{H}_r^q} \sum_{i=1}^n \sum_{j=1}^m \mathrm{E}\left[f_{ij}(0 \mid \mathbf{x}_{ij}, \mathbf{z}_{ij}) \left\{ x_{ijk} - h_k(\mathbf{z}_{ij}) \right\}^2 \right].$$

Let $t_k(z) = \mathrm{E}(x_{ijk} \mid \mathbf{z}_{ij})$, then h_k^* is the weighted projection of $t_k(\cdot)$ into $\mathcal{H}_r^q$ under the L_2 norm, where the weights $f_{ij}(0 \mid \mathbf{x}_{ij}, \mathbf{z}_{ij})$ are included to account for possibly heterogeneous errors. Let x_{ijk} be the element of X at the $(m(i-1)+j)$th row and k column. Define $\delta_{ijk} \equiv x_{ijk} - h_k^*(\mathbf{z}_{ij})$, $\boldsymbol{\delta}_{ij} = (\delta_{ij1}, \ldots, \delta_{ijp})' \in \mathbb{R}^p$, and $\Delta_n = (\delta_{i1}, \ldots, \delta_{nm})' \in \mathbb{R}^{mn \times p}$. Define H as the $mn \times p$ matrix with $(m(i-1)+j, k)$th element $H_{m(i-1)+j,k} = h_k^*(\mathbf{z}_{ij})$. Then $X = H + \Delta_n$. Additionally, define $\psi_\tau(u) = \tau - I(u < 0)$ and $\psi_\tau(\varepsilon_i) = (\psi_\tau(\varepsilon_{i1}), \ldots, \psi_\tau(\varepsilon_{im}))'$.

(C1) (Conditions on the random error) The random error ϵ_{ij} has the conditional distribution function F_{ij} and continuous conditional density function f_{ij} . The f_{ij} are uniformly bounded away from 0 and infinity in a neighborhood of zero and its first derivative f'_{ij} has a uniform upper bound in a neighborhood of zero, for $1 \leq i \leq n$.

(C2) (Conditions on the covariates) There exist positive constants M_1 and M_2 such that $|V_{ijk}| \leq M_1$, $\forall\ 1 \leq i \leq n$, $1 \leq j \leq m$, $1 \leq k \leq p$ and $\mathrm{E}[\delta_{ij}^4] \leq M_2$, $\forall\ 1 \leq i \leq n$, $1 \leq j \leq m$, $1 \leq k \leq p$. For a matrix X, define $\lambda_{max}(X)$ to be the maximum eigenvalue of X. There exist finite positive constants C_1 and C_2

such that with probability one

$$C_1 \leq \lambda max \left(n^{-1}XX'\right) \leq C_2, \quad C_1 \leq \lambda max \left(n^{-1}\Delta\Delta'\right) \leq C_2.$$

(C3) (Condition on the non-linear functions) For $r = m + v > 1.5$, $g_0 \in \mathcal{G}$.

(C4) (Condition on the B-Spline basis) The dimension of the spline basis k_n satisfies $k_n \approx n^{1/(2r+1)}$ and $n^{-1}k_n^3 = o(1)$.

(C5) (Condition on the dropout probability) There exist $0 < \alpha_\ell$ and $\alpha_u < 1$ such that $\alpha_\ell < \pi_{ij0} < \alpha_u$ for all $(i,j) \in \{1,\ldots,n\} \times \{1,\ldots,m\}$ and $\pi_{i10} = 1$.

(C6) (Condition on the estimator of weights) Assume a logistic relationship between η_{ij} and $P(R_{ij} = 1 \mid t_{ij})$ for $j \geq 2$ and $||\partial\eta_{ij}(t_{ij},\gamma)/\partial\gamma||$ and $||\partial^2\eta_{ij}(t_{ij},\gamma)/\partial\gamma\partial\gamma'||$ are bounded in a neighborhood of γ_j.

Define $\psi_\tau(u) = \tau - I(u \leq 0)$ as the gradient of $\rho_\tau(u)$. Let $\mathbf{F}_i$ be a $m \times m$ diagonal matrix with diagonal entries $f_{i1}(0 \mid X_{i1}, Z_{i1}), \ldots, f_{im}(0 \mid X_{im}, Z_{im})$, $\mathbf{k}_i = (\psi_\tau(\varepsilon_{i1}), \psi_\tau(\varepsilon_{i2})R_{i2}/\pi_{i20}, \ldots, \psi_\tau(\varepsilon_{im})R_{im}/\pi_{nm0})$, and

$$\begin{aligned}
\boldsymbol{\Sigma}_1 &= \mathrm{E}\left[\boldsymbol{\delta}_i\mathbf{F}_i\boldsymbol{\delta}_i'\right] \\
\boldsymbol{\Sigma}_2 &= \mathrm{E}\left[\left(\mathbf{k}_i'\boldsymbol{\delta}_i\right)'\mathbf{k}_i'\boldsymbol{\delta}_i\right] \\
\boldsymbol{\Sigma}_{3jk} &= I(k \leq j)\mathrm{E}\left[\frac{\eta_{ij0}\left(1-\eta_{ij0}\right)}{\pi_{ij0}}\psi_\tau(\varepsilon_{ij})\boldsymbol{\delta}_{ij}t_{ik}'\right] \qquad \text{for } k = 2,\ldots,m \\
\boldsymbol{\Sigma}_{3j} &= \left(\boldsymbol{\Sigma}_{3j2},\ldots,\boldsymbol{\Sigma}_{3jm}\right).
\end{aligned}$$

Theorem 4.3.1

Let $\boldsymbol{\Sigma}_m = \boldsymbol{\Sigma}_2 - \sum_{j=2}^m \boldsymbol{\Sigma}_{3j}I(\boldsymbol{\gamma}_0)^{-1}\boldsymbol{\Sigma}_{3j}'$. Under conditions (C1)-(C6),

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \rightarrow N\left(\mathbf{0}, \boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_m\boldsymbol{\Sigma}_1^{-1}\right) \tag{4.5}$$

$$\frac{1}{N_0}\sum_{i=1}^{n}\sum_{j=1}^{m} R_{ij}\left(\hat{g}_k(\mathbf{z}_{ij}) - g_0(\mathbf{z}_{ij})\right)^2 = O_p\left(n^{-2/(2r+1)}\right) \tag{4.6}$$

□

Theorem 4.3.1 proves that using inverse probability weights will yield a consistent estimator of the conditional quantile in the presence of dropout.

4.4 Monte Carlo studies

We perform a simulation study to compare the proposed method for estimating effects with existing ones in the literature. We consider a setting with $n = 300$ individuals. We consider four covariates with linear effects, $X_1, \ldots, X_4$ and two covariates with nonlinear effects, Z_1, Z_2 that do not change across timepoints. X_1 is distributed Uniform$(0, 1)$ and X_2, X_3, X_4 follow the standard normal distribution. Z_1 is distributed Uniform$(0, 1)$ and Z_2 is distribtued Uniform$(-1, 1)$. The jth response for the ith is determined as follows,

$$Y_{ij} \;=\; \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \sin(2\pi Z_{1i}) + Z_{2i}^3 + \epsilon_{ij},$$

where $(\beta_2, \beta_3, \beta_4) = (2, 1, 2)$. Letting $(\xi_{i1}, \ldots, \xi_{im}) \sim N_m(0, \Sigma)$ where the (i, j)th element of Σ is $\rho^{|i-j|}$, we consider two settings for the errors: (1) $\epsilon_{ij} = 2\xi_{ij}$ and (2) $\epsilon_{ij} = 2X_{1ij}\xi_{ij}$ which reflect the case with homogenous errors and heterogeneous errors. With homogenous errors, $\beta_1 = 0$ and with heterogeneous errors, $\beta_1 = \Phi_2^{-1}(\tau)$ where Φ_σ is the CDF of the Normal distribution with mean 0 and variance σ^2. We explore settings with $\rho = .75$, $\rho = .5$, and $\rho = .25$.

The probabilities of returning at time j are generated from the following model,

$$P(R_{ij} = 1 \mid R_{i,j-1} = 1) \;=\; \frac{\exp\left(2 + 2X_{1i} + 2X_{2i} - b_j Y_{i,j-1}\right)}{1 + \exp\left(2 + 2X_{1i} + 2X_{2i} - b_j Y_{i,j-1}\right)},$$

where $b_2 = 1$, $b_3 = 2$, $b_4 = 2$, and $b_5 = 3$. This model implies that an individual with higher responses is more likely to dropout.

We consider two different methods of estimating effects: the proposed inverse probability weighting method (IPW) and the method which ignores the dropout and only uses the complete observations (Naive). The Naive method would yield consistent estimates if the data were missing completely at random. Additionally, we compare results results to the setting in which all data is observed (Oracle). This hypothetical situation is unattainable in practice but is useful for comparing the IPW and Naive methods to the gold standard.

We estimate the $\tau = .5$ and $\tau = .7$ quantiles for the settings when $m = 2$ and $m = 3$ and $n = 300$. We run the simulation for $10,000$ replications. We report the bias of the estimator for each coefficient ($\hat{\beta}_j$), the mean squared error of the estimator for the linear coefficient vector (MSE), and the mean squared error of the estimator of the nonlinear functions (gMSE). Simulation results are contained in Table 4.1 and Table 4.2. Table 4.3 and Table 4.4 contain the ranges of individuals still in the study at each timepoint. All standard errors for the estimates of the bias were less than 0.002.

In both the homogeneous error setting and the heteroscedastic errors setting, the proposed weighted quantile regression estimator has less bias in estimating the linear coefficients than the Naive method for most of the simulations. One surprising observation is that the Naive method almost always has the least bias in estimating β_2. This is likely an artifact of generated data as the Naive method has much more bias than the oracle and proposed estimator in estimating the other linear effects. In the model with heterogeneous errors, the X_{1i} variable causes the heterscedasticity making β_1 a challenging coefficient to estimate. Unsurprisingly, the Oracle method has the least bias, but the IPW method is still acceptable as the Oracle method is unusable in practice.

Table 4.1: Summary of the simulation study with homogenous errors.

Method	τ	m	ρ	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	MSE	gMSE
Oracle	0.5	2	0.75	0.004	0.003	0.006	0.001	0.243	0.102
Naive	0.5	2	0.75	0.057	0.001	0.034	0.079	0.258	0.106
IPW	0.5	2	0.75	0.016	0.004	0.006	0.025	0.309	0.134
Oracle	0.7	2	0.75	0.000	0.003	0.006	0.001	0.267	0.111
Naive	0.7	2	0.75	0.059	0.002	0.033	0.078	0.275	0.111
IPW	0.7	2	0.75	0.020	0.005	0.015	0.044	0.380	0.161
Oracle	0.5	3	0.75	0.001	0.003	0.006	0.002	0.207	0.088
Naive	0.5	3	0.75	0.076	0.020	0.055	0.124	0.245	0.096
IPW	0.5	3	0.75	0.029	0.026	0.032	0.074	0.304	0.130
Oracle	0.7	3	0.75	0.001	0.003	0.005	0.002	0.227	0.095
Naive	0.7	3	0.75	0.078	0.017	0.052	0.120	0.252	0.098
IPW	0.7	3	0.75	0.040	0.025	0.041	0.092	0.363	0.152
Oracle	0.5	2	0.50	0.000	0.005	0.005	0.001	0.212	0.088
Naive	0.5	2	0.50	0.036	0.003	0.020	0.051	0.226	0.094
IPW	0.5	2	0.50	0.008	0.007	0.002	0.018	0.283	0.120
Oracle	0.7	2	0.50	0.000	0.005	0.007	0.000	0.231	0.096
Naive	0.7	2	0.50	0.037	0.003	0.019	0.049	0.243	0.100
IPW	0.7	2	0.50	0.016	0.007	0.009	0.032	0.325	0.139
Oracle	0.5	3	0.50	0.001	0.004	0.005	0.000	0.164	0.070
Naive	0.5	3	0.50	0.045	0.017	0.032	0.077	0.193	0.081
IPW	0.5	3	0.50	0.019	0.021	0.019	0.047	0.268	0.115
Oracle	0.7	3	0.50	0.003	0.005	0.007	0.001	0.179	0.076
Naive	0.7	3	0.50	0.042	0.016	0.029	0.075	0.202	0.084
IPW	0.7	3	0.50	0.024	0.022	0.023	0.062	0.310	0.132
Oracle	0.5	2	0.25	0.005	0.006	0.006	0.001	0.182	0.078
Naive	0.5	2	0.25	0.014	0.004	0.006	0.023	0.200	0.085
IPW	0.5	2	0.25	0.001	0.005	0.001	0.008	0.255	0.109
Oracle	0.7	2	0.25	0.001	0.006	0.004	0.002	0.198	0.086
Naive	0.7	2	0.25	0.019	0.004	0.007	0.022	0.214	0.092
IPW	0.7	2	0.25	0.008	0.005	0.005	0.020	0.278	0.121
Oracle	0.5	3	0.25	0.006	0.006	0.005	0.002	0.129	0.058
Naive	0.5	3	0.25	0.015	0.011	0.011	0.033	0.159	0.070
IPW	0.5	3	0.25	0.000	0.012	0.005	0.020	0.241	0.103
Oracle	0.7	3	0.25	0.003	0.004	0.004	0.002	0.141	0.063
Naive	0.7	3	0.25	0.021	0.009	0.012	0.033	0.168	0.074
IPW	0.7	3	0.25	0.007	0.013	0.010	0.030	0.261	0.115

Table 4.2: Summary of the simulation study with heterogeneous errors.

Method	τ	m	ρ	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	MSE	gMSE
Oracle	0.5	2	0.75	0.001	0.005	0.010	0.005	0.043	0.013
Naive	0.5	2	0.75	0.077	0.001	0.001	0.010	0.050	0.014
IPW	0.5	2	0.75	0.007	0.004	0.009	0.003	0.050	0.015
Oracle	0.7	2	0.75	0.059	0.004	0.010	0.004	0.050	0.014
Naive	0.7	2	0.75	0.149	0.000	0.000	0.012	0.070	0.015
IPW	0.7	2	0.75	0.075	0.004	0.007	0.002	0.064	0.017
Oracle	0.5	3	0.75	0.002	0.005	0.011	0.006	0.036	0.013
Naive	0.5	3	0.75	0.147	0.004	0.005	0.021	0.061	0.015
IPW	0.5	3	0.75	0.054	0.007	0.002	0.005	0.052	0.017
Oracle	0.7	3	0.75	0.054	0.004	0.010	0.004	0.042	0.013
Naive	0.7	3	0.75	0.222	0.002	0.006	0.022	0.091	0.015
IPW	0.7	3	0.75	0.129	0.007	0.002	0.011	0.074	0.019
Oracle	0.5	2	0.50	0.001	0.005	0.011	0.005	0.035	0.011
Naive	0.5	2	0.50	0.052	0.001	0.005	0.005	0.040	0.012
IPW	0.5	2	0.50	0.005	0.004	0.010	0.004	0.043	0.014
Oracle	0.7	2	0.50	0.053	0.005	0.010	0.004	0.041	0.012
Naive	0.7	2	0.50	0.111	0.001	0.004	0.006	0.053	0.013
IPW	0.7	2	0.50	0.068	0.004	0.007	0.001	0.054	0.016
Oracle	0.5	3	0.50	0.003	0.005	0.011	0.006	0.027	0.011
Naive	0.5	3	0.50	0.094	0.003	0.001	0.011	0.040	0.013
IPW	0.5	3	0.50	0.036	0.006	0.005	0.002	0.042	0.015
Oracle	0.7	3	0.50	0.050	0.005	0.010	0.005	0.031	0.012
Naive	0.7	3	0.50	0.151	0.003	0.000	0.012	0.056	0.013
IPW	0.7	3	0.50	0.098	0.006	0.002	0.006	0.056	0.017
Oracle	0.5	2	0.25	0.004	0.005	0.011	0.006	0.031	0.010
Naive	0.5	2	0.25	0.027	0.003	0.008	0.001	0.034	0.011
IPW	0.5	2	0.25	0.005	0.005	0.010	0.005	0.039	0.013
Oracle	0.7	2	0.25	0.054	0.005	0.010	0.005	0.037	0.011
Naive	0.7	2	0.25	0.079	0.003	0.007	0.000	0.043	0.012
IPW	0.7	2	0.25	0.062	0.004	0.008	0.001	0.047	0.014
Oracle	0.5	3	0.25	0.005	0.005	0.011	0.006	0.022	0.010
Naive	0.5	3	0.25	0.044	0.002	0.007	0.002	0.027	0.011
IPW	0.5	3	0.25	0.019	0.005	0.008	0.002	0.036	0.013
Oracle	0.7	3	0.25	0.047	0.005	0.010	0.005	0.026	0.010
Naive	0.7	3	0.25	0.090	0.003	0.006	0.003	0.035	0.012
IPW	0.7	3	0.25	0.073	0.005	0.005	0.002	0.044	0.015

Table 4.3: Ranges of percent of individuals remaining by timepoint (mean in parentheses) with homogenous errors.

ρ	Time 2	Time 3
0.75	71-87 (79)	53-70 (62)
0.50	71-86 (79)	52-69 (60)
0.25	70-87 (79)	50-67 (59)

Table 4.4: Ranges of percent of individuals remaining by timepoint (mean in parentheses) with heterogeneous errors.

ρ	Time 2	Time 3
0.75	77-91 (84)	61-75 (68)
0.50	78-90 (84)	59-76 (67)
0.25	74-91 (84)	60-75 (67)

The IPW method's MSE of the linear coefficients and nonlinear effects is the largest among the three methods. This is attributed to the variability in estimating the inverse probability weights. The simulation study suggests that though the Naive method has larger bias, it has smaller variability because no weights are estimated. The MSE of an estimator is important to consider if the model is being used for prediction.

As the correlation among the errors decreases from 0.75 to 0.25, the bias and MSE decrease. However, as the number of timepoints increases from 2 to 3, the bias and MSE increase for the IPW and Naive methods. This may seem counterintuitive at first, but there are two factors that are likely causing the increase. The IPW method requires estimating the probability of dropout. At timepoint 3, only 50%-75% of patients remain in the study. As patients dropout, estimation of the probability of dropout becomes less acurate. The asymptotic theory shows that more patients will decrease bias and MSE, not more timepoints. The Naive method suffers because it does not handle the dropout properly so more bias is introduced into the estimate.

4.5 Analysis of the Uniform Data Set

In this section we analyze data from the National Alzheimer's Coordinating Center's Uniform Data Set. In particular, we are interested in the effects of covariates on cognitive ability. To measure cognitive ability, we create a composite cognitive (CC) score which is the sum of standardized scores from the Logical Memory IA and IIA tests; Digit Span backwards test; animals, vegetables, and boston naming tests, Digit Symbol, Trail A and B tests, and the mini-mental state exam ([Sano et al., 2017, Nandipati et al., 2012, Cosentino et al., 2010]). The scores for the Trail A and B tests are times until completion, where a shorter time is interpreted as having higher cognitive ability unlike the other tests. To handle this discrepancy, we reverse the scores for the standardized Trail A and B tests before using them to create the CC score.

In the NACC dataset, data was collected on the first visit and patients were asked to return for 9 follow-up visits spaced about one year apart. Though many patients followed-up for one or two years, very few patients remained for all ten visits. As a result, we restrict analysis to the first 5 follow-up visits to ensure that there are enough observations at each timepoint for accurate estimation. To study the covariates' effects on cognitive decline, we define the response variable as the difference in CC scores between the initial visit and each follow-up visit. We consider patients aged 65 or older at the initial visit and those who did not have any missing covariates at the follow-up visits they attended. A patient was removed if he or she increased their CC score from the initial CC score by more than 1. This large of an increase suggests that the score is not reliable. About 7% of patients were removed because of this restriction.

Table 4.5: Dropout in the Uniform Data Set.

Follow-up	1	2	3	4	5
Number of Subjects	3581	3424	3254	2381	1728
Percent of Subjects Remaining	100	96	91	66	48

The percent and number of patients remaining in the study at each follow-up appointment is summarized in Table 4.5. After removing the patients from the dataset who did not meet the inclusion criteria, there were a total of 14,368 observations made across 3,581 unique patients and the 5 follow-up visits.

We analyze the data at the $\tau = 0.9$ conditional quantile which corresponds to cognitive decline among lower performing patients. Table 4.6 contains the estimates of the linear covariates in the UDS for the proposed IPW method and the Naive method. We also compute 90% confidence intervals for the coefficients by resampling the patients to create the bootstrapped dataset. Interestingly at the $\alpha = 0.1$ significance level, the IPW method found that sex has a significant effect on the response while the Naive method did not find significance. More specifically, at the $\tau = 0.9$ conditional quantile, the difference in CC score at baseline was lower in females than males after accounting for the other covariates.

Table 4.6: Linear coefficient estimates with 90% boostrapped confidence intervals (in parentheses) for the $\tau = 0.9$ conditional quantile.

Coefficient	IPW	Naive
(Intercept)	0.2359 (-0.1190, 0.7450)	0.2320 (-0.1119, 0.6854)
sex	-0.0469 (-0.0970, -0.0065)	-0.0420 (-0.0815, 0.0003)
race2Other	0.0764 (0.0274, 0.1334)	0.0790 (0.0329, 0.1305)
hyperten	0.0169 (-0.0221, 0.0632)	0.0260 (-0.0131, 0.0665)
diabetes	0.0739 (0.0203, 0.1403)	0.0590 (0.0095, 0.1193)
stroke	0.2337 (0.1120, 0.3876)	0.2480 (0.1163, 0.4331)
depression	0.1812 (0.1302, 0.2342)	0.1790 (0.1312, 0.2226)
alcohol	-0.1159 (-0.2231, 0.0456)	-0.1060 (-0.2083, 0.0234)
smoke	0.0562 (-0.0513, 0.2019)	0.0480 (-0.0356, 0.1888)
education	-0.0120 (-0.0184, -0.0059)	-0.0130 (-0.0184, -0.0061)

4.6 Discussion

In this chapter, we propose an inversely weighted semiparametric quantile regression estimator for longitudinal data with dropout. Dropout is a pervasive problem in many fields and ignoring the issue often results in biased analyses. We prove that the proposed estimator is consistent and use a bootstrap scheme to create confidence intervals. Additionally, our proposed estimator extends work by He et al. [2002], Lipsitz et al. [1997] in that both nonlinear effects and dropout are present in the model.

An area of research which would be useful to the proposed estimator is producing confidence intervals for the effects. This is very difficult in practice. Bootstrap procedures for quantile regression exist for the nonlongitudinal setting with only linear effects [Feng et al., 2011], but we are not aware of procedures that can handle a semiparametric longitudinal model with dropout.

4.7 Proofs

This notation will be used throughout this section. Let p be the dimension of β and J_n be the number of spline basis functions. Define:

$$
\begin{aligned}
d_n &= p + J_n \in \mathbb{N}, \\
N &= nm \in \mathbb{N}, \\
f_{ij}(0) &= f_{ij}(0 \mid \mathbf{x}_{ij}, \mathbf{z}_{ij}) \in \mathbb{R}, \\
B_N &= \operatorname{diag}\left\{f_{11}(0), \ldots, f_{1m}(0), \ldots, f_{nm}(0)\right\} \in \mathbb{R}^{N \times N}, \\
W &= (\mathbf{W}(\mathbf{z}_{11}), \ldots, \mathbf{W}(\mathbf{z}_{1m}), \ldots, \mathbf{W}(\mathbf{z}_{nm}))' \in \mathbb{R}^{N \times J_n}, \\
P &= W \left(W' B_N W\right)^{-1} W' B_N \in \mathbb{R}^{N \times N}, \\
X^* &= (\mathbf{x}^*_{11}, \ldots, \mathbf{x}^*_{1m}, \ldots, \mathbf{x}^*_{nm})' = (I_N - P)X \in \mathbb{R}^{N \times p}, \\
X^*_i &= \left(\mathbf{x}^*_{i1}, \ldots, \mathbf{x}^*_{im}\right)', \\
W^2_{B_N} &= W' B_N W \in \mathbb{R}^{J_n \times J_n}, \\
\theta_1 &= \sqrt{n}\,(\boldsymbol{\beta} - \boldsymbol{\beta}_0) \in \mathbb{R}^p, \\
\theta_2 &= W_{D_N}\,(\boldsymbol{\xi} - \boldsymbol{\xi}_0) + W^{-1}_{D_N} W' D_N X\,(\boldsymbol{\beta} - \boldsymbol{\beta}_0) \in \mathbb{R}^{J_n}, \\
\theta &= (\theta_1', \theta_2')' \in \mathbb{R}^{d_n}, \\
\tilde{\mathbf{x}}_{ij} &= n^{-1/2} \mathbf{x}^*_{ij} \in \mathbb{R}^p, \\
\tilde{\mathbf{W}}(\mathbf{z}_{ij}) &= W^{-1}_{B_N} \mathbf{W}(\mathbf{z}_{ij}) \in \mathbb{R}^{J_n}, \\
\tilde{\mathbf{s}}_{ij} &= \left(\tilde{\mathbf{x}}_{ij}', \tilde{\mathbf{W}}(\mathbf{z}_{ij})\right)' \in \mathbb{R}^{d_n}, \\
u_{Nij} &= \mathbf{W}(\mathbf{z}_{ij})' \boldsymbol{\xi}_0 - g_0(\mathbf{z}_{ij}) \in \mathbb{R}, \\
Q_{ij}(a_n) &= \rho_\tau\left(\varepsilon_{ij} - a_n\left(\tilde{\mathbf{x}}_{ij}'\theta_1 - \tilde{\mathbf{W}}(\mathbf{z}_{ij})'\theta_2\right) - u_{Nij}\right) \in \mathbb{R}, \\
E_s\left(Q_{ij}\right) &= E\left(Q_{ij} \mid \mathbf{x}_{ij}, \mathbf{z}_{ij}\right) \in \mathbb{R}, \\
D_{ij}(\theta, a_n) &= Q_{ij}(a_n) - Q_{ij}(0) - E_s\left[Q_{ij}(a_n) - Q_{ij}(0)\right] \\
&\quad - a_n\left(\tilde{\mathbf{x}}_{ij}'\theta_1 + \tilde{\mathbf{W}}(\mathbf{z}_{ij})'\theta_2\right) \psi_\tau(\varepsilon_{ij}) \in \mathbb{R}.
\end{aligned}
$$

Note that

$$\begin{aligned}
&\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{R_{ij}}{\hat{\pi}_{ij}}\rho_\tau(Y_{ij}-\mathbf{x}_{ij}'\boldsymbol{\beta}-\mathbf{W}(\mathbf{z}_{ij})'\boldsymbol{\xi}) \\
= &\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{R_{ij}}{\hat{\pi}_{ij}}\rho_\tau(\varepsilon_{ij}-\tilde{\mathbf{x}}_{ij}'\theta_1-\tilde{\mathbf{W}}(\mathbf{z}_{ij})'\theta_2-u_{Nij}),
\end{aligned}$$

and

$$\begin{aligned}
\left(\hat{\theta}_1,\hat{\theta}_2\right) &= \underset{(\theta_1,\theta_2)}{\arg\min}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{R_{ij}}{\hat{\pi}_{ij}}\rho_\tau(\varepsilon_{ij}-\tilde{\mathbf{x}}_{ij}'\theta_1-\tilde{\mathbf{W}}(\mathbf{z}_{ij})'\theta_2-u_{Nij}) \\
&= \underset{(\theta_1,\theta_2)}{\arg\min}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{R_{ij}}{\hat{\pi}_{ij}}\rho_\tau(\varepsilon_{ij}-\tilde{\mathbf{x}}_{ij}'\theta_1-\tilde{\mathbf{W}}(\mathbf{z}_{ij})'\theta_2-u_{Nij})-\rho_\tau(\varepsilon_{ij}-u_{Nij}).
\end{aligned}$$

So we write,

$$\begin{aligned}
\hat{\theta}_1 &= \sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0) \\
\hat{\theta}_2 &= W_{D_N}\left(\hat{\boldsymbol{\xi}}-\boldsymbol{\xi}_0\right)+W_{D_N}^{-1}W'D_NX\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\right).
\end{aligned}$$

Additionally, we consider another form of the check function, $\rho_\tau(u)=\frac{1}{2}|u|+(\tau-\frac{1}{2})u$.

Throughout these proofs, we will let C be a positive constant that may change line to line. For a matrix A, let $||A||$ be the spectral norm and for a vector x, let $||x||$ be the Euclidean distance.

Probability of dropout

Here we will find useful quantities dealing with estimating the probability of dropout. We assume a logistic relationship between $\eta(T_{ij},\gamma_j)$ and $P(R_{ij}=1\mid R_{i,j-1}=1,T_{ij})$,

$$\eta(T_{ij},\gamma_j) = \frac{e^{T_{ij}'\gamma_j}}{1+e^{T_{ij}'\gamma_j}}$$

and note that

$$\nabla\eta(T_{ij},\gamma_j) \quad = \quad T_{ij}e^{T_{ij}'\gamma_j}(1+e^{T_{ij}'\gamma_j})^{-2}.$$

The log-likelihood function is

$$\begin{aligned}
\ell(\boldsymbol{\gamma}) &= \sum_{i=1}^{n}\sum_{j=2}^{m} R_{i,j-1}\left\{R_{ij}\log\left[\eta(T_{ij},\gamma_j)\right] + (1-R_{ij})\log\left[1-\eta(T_{ij},\gamma_j)\right]\right\} \\
&= \sum_{i=1}^{n} R_{i,j-1}\left[R_{ij}T_{ij}'\gamma_j - \log\left(1+e^{T_{ij}'\gamma_j}\right)\right] \\
\nabla_{\gamma_j}\ell(\boldsymbol{\gamma}) &= \sum_{i=1}^{n} R_{i,j-1}\left[R_{ij}-\eta(T_{ij},\gamma_j)\right]T_{ij} \\
\nabla^2_{\gamma_j}\ell(\boldsymbol{\gamma}) &= \sum_{i=1}^{n} -R_{i,j-1}\eta(T_{ij},\gamma_j)\left[1-\eta(T_{ij},\gamma_j)\right]T_{ij}T_{ij}' \\
\nabla_{\gamma_j\gamma_k}\ell\boldsymbol{\gamma}) &= \mathbf{0} \\
I(\gamma_j) &= \mathrm{E}\left[R_{i,j-1}\eta(T_{ij},\gamma_j)\left[1-\eta(T_{ij},\gamma_j)\right]T_{ij}T_{ij}'\right] \\
&= \mathrm{E}\left[\pi_{i,j-1,0}\eta(T_{ij},\gamma_j)\left[1-\eta(T_{ij},\gamma_j)\right]T_{ij}T_{ij}'\right],
\end{aligned}$$

so we have that the $\hat{\gamma}_j$ are independent of one another. Let $I(\boldsymbol{\gamma})$ be the block diagonal matrix with $I(\gamma_j)$, for $j = 2,\ldots,m$, on its diagonal. Let $\nabla_{\boldsymbol{\gamma}}\ell(\boldsymbol{\gamma}) = (\nabla_{\gamma_2}\ell(\boldsymbol{\gamma})',\ldots,\nabla_{\gamma_m}\ell(\boldsymbol{\gamma})')'$. For the ith subject, note that

$$\begin{aligned}
\nabla_{\gamma_j}\ell_i(\boldsymbol{\gamma}) &= R_{i,j-1}\left[R_{ij}-\eta(T_{ij},\gamma_j)\right]T_{ij} \\
\nabla\ell_i(\boldsymbol{\gamma}) &= (\nabla_{\gamma_2}\ell_i(\boldsymbol{\gamma})',\ldots,\nabla_{\gamma_m}\ell_i(\boldsymbol{\gamma})')'
\end{aligned}$$

Lemmas

Lemma 4.7.1

Under conditions (C5)-(C6), then

$$\sup_{ij} |\hat{\pi}_{ij}^{-1} - \pi_{ij}^{-1}| \;\; = \;\; O_p(n^{-1/2}).$$

Proof

Let $f_1(R_{ij};\gamma,T_{ij})$ be the pmf for $R_{ij}|R_{ik}=1$ for $k=1,\ldots,j-1$ and $\ell(\gamma_j)$ be the log-likelihood function for γ_j. Let $f(R_{ij};\gamma,T_{ij})$ be the unconditional pmf for R_{ij}. We also note that $\eta\left(T_{ij},\gamma_j\right)=\exp\left(T_{ij}'\gamma_j\right)/\left[1+\exp\left(T_{ij}'\gamma_j\right)\right]$. Then we have

$$\begin{aligned}
f_1(R_{ij};\gamma,T_{ij}) &= \left\{\exp\left(T_{ij}'\gamma_j\right)/\left[1+\exp\left(T_{ij}'\gamma_j\right)\right]\right\}^{R_{ij}}\left\{1/\left[1+\exp\left(T_{ij}'\gamma_j\right)\right]\right\}^{1-R_{ij}} \\
&\quad \times\prod_{k=1}^{j-1} R_{ik} \\
f(R_{ij};\gamma,T_{ij}) &= \left\{\exp\left(T_{ij}'\gamma_j\right)/\left[1+\exp\left(T_{ij}'\gamma_j\right)\right]\right\}^{R_{ij}}\left\{1/\left[1+\exp\left(T_{ij}'\gamma_j\right)\right]\right\}^{1-R_{ij}} \\
&\quad \times\prod_{k=2}^{j-1} I(R_{ij}=1)\exp\left(T_{ik}'\gamma_k\right)/\left[1+\exp\left(T_{ik}'\gamma_k\right)\right] \\
\ell(\gamma_j) &= R_{ij}T_{ij}'\gamma_j-\log(1+\exp\left(T_{ij}'\gamma_j\right)) \\
&\quad +\sum_{k=2}^{j-1}\log\left\{I(R_{ij}=1)\exp\left(T_{ik}'\gamma_k\right)/\left[1+\exp\left(T_{ik}'\gamma_k\right)\right]\right\} \\
\ell'(\gamma_j) &= T_{ij}\left\{R_{ij}-\exp\left(T_{ij}'\gamma_j\right)/\left[1+\exp\left(T_{ij}'\gamma_j\right)\right]\right\} \\
&= T_{ij}\left(R_{ij}-\eta\left(T_{ij},\gamma_j\right)\right) \\
\ell''(\gamma_j) &= -T_{ij}T_{ij}'\eta\left(T_{ij},\gamma_j\right)\left(1-\eta\left(T_{ij},\gamma_j\right)\right) \\
I(\gamma_j) &= E\left[T_{ij}T_{ij}'\eta\left(T_{ij},\gamma_j\right)\left(1-\eta\left(T_{ij},\gamma_j\right)\right)\right]
\end{aligned}$$

Therefore, for $j = 2, \dots, m$

$$\begin{aligned}
n^{1/2}(\hat{\gamma}_j - \gamma_j) &\overset{d}{\to} N(0, I(\gamma_j)^{-1}) \\
n^{1/2}(T'_{ij}\hat{\gamma}_j - T'_{ij}\gamma_j) &\overset{d}{\to} N\left(0, T'_{ij}I(\gamma_j)^{-1}T_{ij}\right) \\
n^{1/2}\left(\eta\left(T_{ij}, \hat{\gamma}_j\right) - \eta\left(T_{ij}, \gamma_j\right)\right) &\overset{d}{\to} N\left(0, T'_{ij}I(\gamma_j)^{-1}T_{ij}\left[\eta\left(T_{ij}, \gamma_j\right) - \eta\left(T_{ij}, \gamma_j\right)^2\right]^2\right) \\
n^{1/2}\left(\hat{\pi}_{ij} - \pi_{ij}\right) &\overset{d}{\to} N\left(0, \sigma^2_{ij}\prod_{k=2}^{j}\eta\left(T_{ik}, \gamma_k\right)^2\right) \\
n^{1/2}\left(\hat{\pi}^{-1}_{ij} - \pi^{-1}_{ij}\right) &\overset{d}{\to} N\left(0, \sigma^2_{ij}/\pi^2_{ij}\right) \\
\sup_{ij}|\hat{\pi}^{-1}_{ij} - \pi^{-1}_{ij}| &= O_p(n^{-1/2}),
\end{aligned}$$

where $\sigma^2_{ij} = \sum_{k=2}^{j} T'_{ik}I(\gamma_k)^{-1}T_{ik}\left[1 - \eta\left(T_{ik}, \gamma_k\right)\right]^2$. □

Lemma 4.7.2

We have the following properties for the spline basis vector.

1. $E\left(||\mathbf{W}(\mathbf{z}_{ij})||\right) \leq b_1\sqrt{k_n}$ for all $(i, j) \in \{1, \dots, n\} \times \{1, \dots, m\}$ for some positive constant b_1 and large n,

2. There exist some positive constants b_2 and b_3 such that for all n sufficiently large
   $b_2 k_n^{-1} \leq E\left[\lambda_{\min}\left\{\mathbf{W}(\mathbf{z}_{ij})\mathbf{W}(\mathbf{z}_{ij})'\right\}\right]$ and $E\left[\lambda_{\max}\left\{\mathbf{W}(\mathbf{z}_{ij})\mathbf{W}(\mathbf{z}_{ij})'\right\}\right] \leq b_3 k_n^{-1}$,

3. $E(||W_B^{-1}||) \geq b_4 n^{-1/2}$, for some positive constant b_4 and sufficiently large n,

4. $\max_{ij}||\tilde{\mathbf{W}}(\mathbf{z}_{ij})|| = O_p(k_n^{1/2}n^{-1/2})$.

Proof

These results were proven in Lemma 2 from Sherwood and Wang [2016]. They hold for the longitudinal model because the sample size grows with n so the constant can absorb the m additional observations in the ith subject. □

Lemma 4.7.3

For a positive constant L,

$$d_n^{-1} \sup_{||\theta|| \leq L} \left| \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{R_{ij}}{\pi_{ij}} D_{ij}(\theta, d_n^{1/2}) \right| = o_p(1).$$

This is similar to Lemma B.1 from Sherwood and Wang [2016]. □

Proof

We will follow the proof from Sherwood and Wang (2016). Let F_{n1} denote the event $\tilde{s}_{(N)} \leq \alpha_1 \sqrt{d_n/n}$, for some positive constant α_1, where $\tilde{s}_{(N)} = \max_{ij} ||\tilde{s}_{ij}||$. Note that $\max_{ij} ||\tilde{\mathbf{x}}_{ij}|| \leq \alpha_2 \sqrt{p/n}$, for some positive constant α_2. This observation combined with $\max_{i,j} ||\tilde{W}(\mathbf{z}_{ij})|| = O_p\left(\sqrt{k_n/n}\right)$ implies that $P(F_{n1}) \to 1$ as $n \to \infty$. Let F_{n2} denote the event $\max_{ij} |u_{Nij}| \leq \alpha_3 d_n^{-r}$, for some positive constant α_3, then it follows from Schumaker (1981) that $P(F_{n2}) \to 1$.

To prove the lemma, we need to show that $\forall \varepsilon > 0$,

$$P\left(d_n^{-1} \sup_{||\theta|| \leq 1} \left| \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{R_{ij}}{\pi_{ij}} D_{ij}(\theta, L d_n^{1/2}) \right| > \varepsilon,\ F_{n1} \cap F_{n2} \right)$$

Define $\Theta^* \equiv \{\theta \mid ||\theta|| \leq 1, \theta \in \mathbb{R}^{d_n}\}$. We partition Θ as a union of disjoint regions $\Theta_1, \ldots, \Theta_{M_n}$, such that the diameter of each region does not exceed $m_0 = \frac{\varepsilon \alpha_\ell^m}{8 \alpha_1 L m \sqrt{n}}$. This covering can be constructed such that $M_n \leq C\left(\frac{C\sqrt{n}}{\varepsilon}\right)^{d_n+1}$, where C is a positive constant. Let $\theta_1^*, \ldots, \theta_{M_n}^*$ be arbitrary points in $\Theta_1, \ldots, \Theta_{M_n}$, respectively $k = 1, \ldots, M_n$. Then

$$\begin{aligned} & P\left(d_n^{-1} \sup_{||\theta|| \leq 1} \left| \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{R_{ij}}{\pi_{ij}} D_{ij}(\theta, L d_n^{1/2}) \right| > \varepsilon,\ F_{n1} \cap F_{n2} \right) \\ \leq & \sum_{k=1}^{M_n} P\left(d_n^{-1} \sup_{\theta \in \Theta_k} \left| \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{R_{ij}}{\pi_{ij}} D_{ij}(\theta, L d_n^{1/2}) \right| > \varepsilon,\ F_{n1} \cap F_{n2} \right) \end{aligned}$$

$$
\begin{aligned}
\leq\ & \sum_{k=1}^{M_n} P\Bigg(\Big| \sum_{i=1}^{n}\sum_{j=1}^{m} \frac{R_{ij}}{\pi_{ij}} D_{ij}(\theta_k^*, Ld_n^{1/2}) \Big| \\
& + \sup_{\theta\in\Theta_k} \Big| \sum_{i=1}^{n}\sum_{j=1}^{m} \frac{R_{ij}}{\pi_{ij}} \left(D_{ij}(\theta, Ld_n^{1/2}) - D_{ij}(\theta_k^*, Ld_n^{1/2}) \right) \Big| \\
& > d_n\varepsilon,\ F_{n1} \cap F_{n2} \Bigg).
\end{aligned}
$$

We will now show that

$$
\sup_{\theta\in\Theta_k} \Big| d_n^{-1} \sum_{i=1}^{n}\sum_{j=1}^{m} \frac{R_{ij}}{\pi_{ij}} \left(D_{ij}(\theta, Ld_n^{1/2}) - D_{ij}(\theta_k^*, Ld_n^{1/2}) \right) \Big| I(F_{n1} \cap F_{n2}) \quad \leq \quad \varepsilon/2.
$$

Using bounds on $||\tilde{\mathbf{x}}_{ij}||$, $||\tilde{W}(\mathbf{z}_{ij})||$, and π_{ij}^{-1}, we have

$$
\begin{aligned}
& \sup_{\theta\in\Theta_k} \Big| d_n^{-1} \sum_{i=1}^{n}\sum_{j=1}^{m} \frac{R_{ij}}{\pi_{ij}} \left(D_{ij}(\theta, Ld_n^{1/2}) - D_{ij}(\theta_k^*, Ld_n^{1/2}) \right) \Big| I(F_{n1} \cap F_{n2}) \\
=\ & d_n^{-1} \sup_{\theta\in\Theta_k} \Big| \sum_{i=1}^{n}\sum_{j=1}^{m} \frac{R_{ij}}{\pi_{ij}} \frac{1}{2} \Big[\big| \varepsilon_{ij} - \tilde{\mathbf{x}}_{ij}'\theta_1 L\sqrt{d_n} - \tilde{\mathbf{W}}(\mathbf{z}_{ij})'\theta_2 L\sqrt{d_n} - u_{Nij} \big| \\
& \qquad - \big| \varepsilon_{ij} - u_{Nij} \big| \Big] \\
& - \sum_{i=1}^{n}\sum_{j=1}^{m} \frac{R_{ij}}{\pi_{ij}} \frac{1}{2} \mathrm{E}_s \Big[\big| \varepsilon_{ij} - \tilde{\mathbf{x}}_{ij}'\theta_1 L\sqrt{d_n} - \tilde{\mathbf{W}}(\mathbf{z}_{ij})'\theta_2 L\sqrt{d_n} - u_{Nij} \big| - \big| \varepsilon_{ij} - u_{Nij} \big| \Big] \\
& + \sum_{i=1}^{n}\sum_{j=1}^{m} \frac{R_{ij}}{\pi_{ij}} L\sqrt{d_n} \left(\tilde{\mathbf{x}}_{ij}'\theta_1 - \tilde{\mathbf{W}}(\mathbf{z}_{ij})'\theta_2 \right) \psi_\tau(\varepsilon_{ij}) \\
& - \sum_{i=1}^{n}\sum_{j=1}^{m} \frac{R_{ij}}{\pi_{ij}} \frac{1}{2} \Big[\big| \varepsilon_{ij} - \tilde{\mathbf{x}}_{ij}'\theta_{k1}^* L\sqrt{d_n} - \tilde{\mathbf{W}}(\mathbf{z}_{ij})'\theta_{k2}^* L\sqrt{d_n} - u_{Nij} \big| - \big| \varepsilon_{ij} - u_{Nij} \big| \Big] \\
& + \sum_{i=1}^{n}\sum_{j=1}^{m} \frac{R_{ij}}{\pi_{ij}} \frac{1}{2} \mathrm{E}_s \Big[\big| \varepsilon_{ij} - \tilde{\mathbf{x}}_{ij}'\theta_{k1}^* L\sqrt{d_n} - \tilde{\mathbf{W}}(\mathbf{z}_{ij})'\theta_{k2}^* L\sqrt{d_n} - u_{Nij} \big| \\
& \qquad - \big| \varepsilon_{ij} - u_{Nij} \big| \Big] \\
& - \sum_{i=1}^{n}\sum_{j=1}^{m} \frac{R_{ij}}{\pi_{ij}} L\sqrt{d_n} \left(\tilde{\mathbf{x}}_{ij}'\theta_{k1}^* - \tilde{\mathbf{W}}(\mathbf{z}_{ij})'\theta_{k2}^* \right) \psi_\tau(\varepsilon_{ij}) \Big| I(F_{n1} \cap F_{n2})
\end{aligned}
$$

$$
\begin{aligned}
\leq\ & d_n^{-1}\Bigg(\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{R_{ij}}{\pi_{ij}}\frac{1}{2}m_0\sqrt{d_n}\tilde{s}_{(N)}L+\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{R_{ij}}{\pi_{ij}}\frac{1}{2}m_0\sqrt{d_n}\tilde{s}_{(N)}L\\
&+\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{R_{ij}}{\pi_{ij}}m_0\sqrt{d_n}\tilde{s}_{(N)}L\\
&+\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{R_{ij}}{\pi_{ij}}\frac{1}{2}m_0\sqrt{d_n}\tilde{s}_{(N)}L+\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{R_{ij}}{\pi_{ij}}\frac{1}{2}m_0\sqrt{d_n}\tilde{s}_{(N)}L\\
&+\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{R_{ij}}{\pi_{ij}}m_0\sqrt{d_n}\tilde{s}_{(N)}L\Bigg)I(F_{n1}\cap F_{n2})\\
\leq\ & d_n^{-1}nm\alpha_\ell^{-m}\left(4m_0\sqrt{d_n}\tilde{s}_{(N)}L\right)I(F_{n1}\cap F_{n2})\\
\leq\ & d_n^{-1}nm\alpha_\ell^{-m}2m_0\sqrt{d_n}\alpha_1\sqrt{d_n/n}L\\
=\ & \varepsilon/2.
\end{aligned}
$$

We now need to show

$$
\sum_{k=1}^{M_n}P\left(\left|\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{R_{ij}}{\pi_{ij}}D_{ij}(\theta_k^*,Ld_n^{1/2})\right|>d_n\varepsilon/2,\ F_{n1}\cap F_{n2}\right)\to 0.
$$

We have

$$
\begin{aligned}
& \max_{ij}\left|\frac{R_{ij}}{\pi_{ij}}D_{ij}(\theta_k^*,Ld_n^{1/2})\right|I(F_{n1}\cap F_{n2})\\
\leq\ & \max_{ij}\frac{R_{ij}}{\pi_{ij}}\left[\left|\varepsilon_{ij}-\tilde{\mathbf{x}}_{ij}'\theta_{k1}^*L\sqrt{d_n}-\tilde{\mathbf{W}}(\mathbf{z}_{ij})'\theta_{k2}^*L\sqrt{d_n}-u_{Nij}\right|-\left|\varepsilon_{ij}-u_{Nij}\right|\right]\\
&+\ \max_{ij}\frac{R_{ij}}{\pi_{ij}}L\sqrt{d_n}\left(\tilde{\mathbf{x}}_{ij}'\theta_{k1}^*-\tilde{\mathbf{W}}(\mathbf{z}_{ij})'\theta_{k2}^*\right)\psi_\tau(\varepsilon_{ij})\Bigg|I(F_{n1}\cap F_{n2})\\
\leq\ & 2L\alpha_\ell^{-m}\sqrt{d_n}\tilde{s}_{(N)}I(F_{n1}\cap F_{n2})\\
\leq\ & Cd_nn^{-1/2},
\end{aligned}
$$

for a positive constant C.

Define

$$V_{ij}(\theta_k^*, a_n) \;=\; Q_{ij}(a_n) - Q_{ij}(0) + a_n \left(\tilde{\mathbf{x}}_{ij}'\theta_{k1}^* - \tilde{\mathbf{W}}(\mathbf{z}_{ij})'\theta_{k2}^*\right) \psi_\tau(\varepsilon_{ij}).$$

It follows that $D_{ij}(\theta_k^*, a_n) = V_{ij}(\theta_k^*, a_n) - \mathrm{E}_s\left[V_{ij}(\theta_k^*, a_n)\right]$ and $\mathrm{E}_s\left[\frac{R_{ij}}{\pi_{ij}} D_{ij}(\theta_k^*, a_n)\right] = 0$ using iterative expectations. Thus

$$\begin{aligned} & \sum_{i=1}^n \mathrm{Var}\left(\sum_{j=1}^m \frac{R_{ij}}{\pi_{ij}} D_{ij}(\theta_k^*, a_n) I(F_{n1} \cap F_{n2}) \mid \mathbf{x}_i, \mathbf{z}_i\right) \\ \leq\; & \sum_{i=1}^n \mathrm{E}_s\left[\left(\sum_{j=1}^m \frac{R_{ij}}{\pi_{ij}} V_{ij}(\theta_k^*, a_n) I(F_{n1} \cap F_{n2})\right)^2\right] \end{aligned}$$

Using Knight's Identity,

$$\begin{aligned} & V_{ij}(\theta_k^*, L\sqrt{d_n}) \\ =\; & L\sqrt{d_n}\left(\tilde{\mathbf{x}}_{ij}'\theta_{k1}^* - \tilde{\mathbf{W}}(\mathbf{z}_{ij})'\theta_{k2}^*\right)\left[I(\varepsilon_{ij} - u_{Nij} < 0) - I(\varepsilon_{ij} < 0)\right] \\ & + \int_0^{L\sqrt{d_n}\left(\tilde{\mathbf{x}}_{ij}'\theta_{k1}^* - \tilde{\mathbf{W}}(\mathbf{z}_{ij})'\theta_{k2}^*\right)} \left[I(\varepsilon_{ij} - u_{Nij} < s) - I(\varepsilon_{ij} < 0)\right] ds \\ \equiv\; & V_{ij1} + V_{ij2}. \end{aligned}$$

To find the variance, we need to consider all the cross product terms. For $(j, w) = \{1, \ldots, m\} \times \{1, \ldots, m\}$, we have

$$\begin{aligned} & \sum_{i=1}^n \sum_{j,w} \mathrm{E}_s\left[\frac{R_{ij}R_{iw}}{\pi_{ij}\pi_{iw}} V_{ij1}V_{iw1} I(F_{n1} \cap F_{n2})\right] \\ =\; & \sum_{i=1}^n \sum_{j,w} \mathrm{E}_s\Bigg[\frac{R_{ij}R_{iw}}{\pi_{ij}\pi_{iw}} L^2 d_n \left(\tilde{\mathbf{x}}_{ij}'\theta_{k1}^* - \tilde{\mathbf{W}}(\mathbf{z}_{ij})'\theta_{k2}^*\right)\left(\tilde{\mathbf{x}}_{iw}'\theta_{k1}^* - \tilde{\mathbf{W}}(\mathbf{z}_{iw})'\theta_{k2}^*\right) \\ & \times \left|I(\varepsilon_{ij} - u_{Nij} < 0) - I(\varepsilon_{ij} < 0)\right| \left|I(\varepsilon_{iw} - u_{Niw} < 0) - I(\varepsilon_{iw} < 0)\right| \\ & \times I(F_{n1} \cap F_{n2})\Bigg] \end{aligned}$$

$$\begin{aligned}
&\leq \; 2L^2 d_n \sum_{i=1}^{n} \sum_{j,w} \mathrm{E}_s\Bigg[\frac{R_{ij}R_{iw}}{\pi_{ij}\pi_{iw}} \tilde{s}^2_{(N)} \left|I(\varepsilon_{ij} - u_{Nij} < 0) - I(\varepsilon_{ij} < 0)\right| \\
&\qquad \times \left|I(\varepsilon_{iw} - u_{Niw} < 0) - I(\varepsilon_{iw} < 0)\right| I(F_{n1} \cap F_{n2})\Bigg] \\
&\leq \; CL^2 d_n^2 n^{-1} \sum_{i=1}^{n} \sum_{j,w} \mathrm{E}_s\Big[I(0 \leq |\varepsilon_{ij}| \leq |u_{Nij}|) I(0 \leq |\varepsilon_{iw}| \leq |u_{Niw}|) I(F_{n1} \cap F_{n2})\Big] \\
&\leq \; CL^2 d_n^2 n^{-1} \sum_{i=1}^{n} \sum_{j,w} \mathrm{E}_s\Big[I(0 \leq |\varepsilon_{ij}| \leq |u_{Nij}|) I(F_{n1} \cap F_{n2})\Big] \\
&\leq \; CL^2 d_n^2 n^{-1} \sum_{i=1}^{n} \sum_{j,w} \int_{-|u_{Nij}|}^{|u_{Nij}|} f_{ij}(s)\, ds \\
&\leq \; C d_n^2 k_n^{-r}.
\end{aligned}$$

Note that V_{ij2} is always nonnegative and $\max_{ij} \left| \sqrt{d_n} L \left(\tilde{\mathbf{x}}'_{ij}\theta^*_{k1} - \tilde{\mathbf{W}}(\mathbf{z}_{ij})'\theta^*_{k2} \right) \right| \leq \alpha_1 L d_n n^{-1/2}$, then $V_{ij2} \leq C d_n n^{-1/2}$.

$$\begin{aligned}
&\sum_{i=1}^{n} \sum_{j,w} \mathrm{E}_s\left[\frac{R_{ij}R_{iw}}{\pi_{ij}\pi_{iw}} V_{ij2} V_{iw2} I(F_{n1} \cap F_{n2})\right] \\
&= \; C d_n n^{-1/2} \sum_{i=1}^{n} \sum_{j=1}^{m} \mathrm{E}_s\left[V_{ij2} I(F_{n1} \cap F_{n2})\right] \\
&\leq \; C d_n n^{-1/2} \sum_{i=1}^{n} \sum_{j=1}^{m} \int_0^{\sqrt{d_n}L\left(\tilde{\mathbf{x}}'_{ij}\theta^*_{k1} - \tilde{\mathbf{W}}(\mathbf{z}_{ij})'\theta^*_{k2}\right)} \left[F_{ij}(s + u_{Nij}) - F_{ij}(u_{Nij})\right] \\
&\quad \times I(F_{n1} \cap F_{n2})\, ds \\
&\leq \; C d_n n^{-1/2} \sum_{i=1}^{n} \sum_{j=1}^{m} \int_0^{\sqrt{d_n}L\left(\tilde{\mathbf{x}}'_{ij}\theta^*_{k1} - \tilde{\mathbf{W}}(\mathbf{z}_{ij})'\theta^*_{k2}\right)} \left(f_{ij}(0) + o(1)\right)\left(s + O(s^2)\right) ds \\
&\leq \; C d_n^2 n^{-1/2} \Bigg[{\theta^*_{k1}}' \left(\sum_{i=1}^{n} \sum_{j=1}^{m} f_{ij}(0) \tilde{\mathbf{x}}_{ij} \tilde{\mathbf{x}}'_{ij} \right) \theta^*_{k1} \\
&\quad + {\theta^*_{k2}}' \left(\sum_{i=1}^{n} \sum_{j=1}^{m} f_{ij}(0) \tilde{\mathbf{W}}(\mathbf{z}_{ij}) \tilde{\mathbf{W}}(\mathbf{z}_{ij})' \right) \theta^*_{k2} \Bigg] (1 + o(1))
\end{aligned}$$

$$\begin{aligned}
&\leq \; Cd_n^2 n^{-1/2}\left[\theta_{k1}^{*}{}'\left(\sum_{i=1}^{n}\sum_{j=1}^{m} f_{ij}(0)||\tilde{\mathbf{x}}_{ij}||^2\right)\theta_{k1}^{*} + \theta_{k2}^{*}{}'W_{B_N}^{-1}W'B_N W W_{B_N}^{-1}\theta_{k2}^{*}\right] \\
&\quad \times(1+o(1)) \\
&\leq \; Cd_n^2 n^{-1/2}\left[\theta_{k1}^{*}{}'\left(\sum_{i=1}^{n}\sum_{j=1}^{m} n^{-1}\right)\theta_{k1}^{*} + \theta_{k2}^{*}{}'I\theta_{k2}^{*}\right](1+o(1)) \\
&\leq \; Cd_n^2 n^{-1/2}\left[||\theta_{k1}^{*}||^2 + ||\theta_{k2}^{*}||^2\right](1+o(1)) \\
&\leq \; Cd_n^2 n^{-1/2}(1+o(1)).
\end{aligned}$$

We now check the last term.

$$\begin{aligned}
&\sum_{i=1}^{n}\sum_{j,w}\left|\mathrm{E}_s\left[\frac{R_{ij}R_{iw}}{\pi_{ij}\pi_{iw}}V_{ij1}V_{iw2}I(F_{n1}\cap F_{n2})\right]\right| \\
\leq \; & C\sum_{i=1}^{n}\sum_{j,w}\left|\mathrm{E}_s\left[L\sqrt{d_n}\left(\tilde{\mathbf{x}}_{ij}'\theta_{k1}^{*} - \tilde{\mathbf{W}}(\mathbf{z}_{ij})'\theta_{k2}^{*}\right)\left[I(\varepsilon_{ij}-u_{Nij}<0)-I(\varepsilon_{ij}<0)\right]\right.\right. \\
&\left.\left.\times V_{iw2}I(F_{n1}\cap F_{n2})\right]\right| \\
\leq \; & C\sum_{i=1}^{n}\sum_{j,w}\left|\mathrm{E}_s\left[L\sqrt{d_n}\tilde{s}_{(N)}V_{iw2}I(F_{n1}\cap F_{n2})\right]\right| \\
\leq \; & Cd_n n^{-1/2}\sum_{i=1}^{n}\sum_{j,w}\left|\mathrm{E}_s\left[V_{iw2}I(F_{n1}\cap F_{n2})\right]\right| \\
\leq \; & Cd_n^2 n^{-1/2}(1+o(1)).
\end{aligned}$$

Therefore,

$$\sum_{i=1}^{n}\mathrm{Var}\left(\sum_{j=1}^{m}\frac{R_{ij}}{\pi_{ij}}D_{ij}(\theta_k^{*},a_n)I(F_{n1}\cap F_{n2}) \mid \mathbf{x}_{ij},\mathbf{z}_{ij}\right) \;\leq\; Cd_n^2 k_n^{-r}.$$

We now check the maximum value of $\left|\sum_{j=1}^{m}\frac{R_{ij}}{\pi_{ij}}D_{ij}(\theta_k^{*},Ld_n^{1/2})\right|$.

$$V_{ij}(\theta_k^{*},Ld_n^{1/2}) \;\leq\; CLd_n n^{-1/2},$$

therefore

$$\left| \sum_{j=1}^{m} \frac{R_{ij}}{\pi_{ij}} D_{ij}(\theta_k^*, Ld_n^{1/2}) \right| \ \leq \ CLd_n n^{-1/2}.$$

We can now use Bernstein's inequality,

$$\begin{aligned}
& \sum_{k=1}^{M_n} P\left(\left| \left(\sum_{i=1}^{n} \sum_{j=1}^{m} \frac{R_{ij}}{\pi_{ij}} D_{ij}(\theta_k^*, Ld_n^{1/2}) \right) \right| > d_n\varepsilon/2, \ F_{n1} \cap F_{n2} \right) \\
\leq \ & 2\sum_{k=1}^{M_n} \exp\left(\frac{-d_n^2\varepsilon^2/8}{Cd_n^2 k_n^{-r} + C\varepsilon d_n^2 n^{-1/2}} \right) \\
\leq \ & 2\sum_{k=1}^{M_n} \exp\left(-Ck_n^r - Cn^{1/2} \right) \\
\leq \ & 2\sum_{k=1}^{M_n} \exp\left(-Ck_n^r \right) \\
= \ & 2M_n \exp\left(-Ck_n^r \right) \\
\leq \ & C\left(\frac{C\sqrt{n}}{\varepsilon} \right)^{d_n+1} \exp\left(-Ck_n^r \right) \\
= \ & C\exp\left((d_n+1)\log\left(C\sqrt{n}/\varepsilon \right) - Ck_n^r \right) \\
\leq \ & C\exp\left(C(d_n+1)\log(n) - Ck_n^r \right),
\end{aligned}$$

which goes to 0 as $n \to \infty$ because d_n is of the same order as k_n. Thus the proof is complete. □

Proof of second part of Theorem 4.3.1

Proof

This is similar to Lemma 4 of Sherwood and Wang [2016] and Lemma A.1 of Sherwood [2016].

We consider

$$d_n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{R_{ij}}{\hat{\pi}_{ij}}\rho_\tau(\varepsilon_{ij}-d_n^{1/2}\tilde{\mathbf{x}}_{ij}'\theta_1-d_n^{1/2}\tilde{\mathbf{W}}(\mathbf{z}_{ij})'\theta_2-u_{Nij})-\rho_\tau(\varepsilon_{ij}-u_{Nij}).$$

By Knight's Identity,

$$\begin{aligned}
& d_n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{R_{ij}}{\hat{\pi}_{ij}}\rho_\tau(\varepsilon_{ij}-d_n^{1/2}\tilde{\mathbf{x}}_{ij}'\theta_1-d_n^{1/2}\tilde{\mathbf{W}}(\mathbf{z}_{ij})'\theta_2-u_{Nij})-\rho_\tau(\varepsilon_{ij}-u_{Nij}) \\
= \;& d_n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{R_{ij}}{\hat{\pi}_{ij}}\int_{-u_{Nij}}^{-\sqrt{d_n}\left(\tilde{\mathbf{x}}_{ij}'\theta_1+\tilde{\mathbf{W}}(\mathbf{z}_{ij})'\theta_2\right)-u_{Nij}}\psi_\tau(\varepsilon_{ij}+s)\,ds \\
= \;& d_n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{R_{ij}}{\pi_{ij}}\left\{Q_{ij}\left(\sqrt{d_n}\right)-Q_{ij}\left(0\right)\right\} \\
& +d_n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}R_{ij}\left(\frac{1}{\hat{\pi}_{ij}}-\frac{1}{\pi_{ij}}\right)\int_{-u_{Nij}}^{-\sqrt{d_n}\left(\tilde{\mathbf{x}}_{ij}'\theta_1+\tilde{\mathbf{W}}(\mathbf{z}_{ij})'\theta_2\right)-u_{Nij}}\psi_\tau(\varepsilon_{ij}+s)\,ds \\
= \;& K_1+K_2.
\end{aligned}$$

We want to show that $\forall\eta>0$, there exists an $L>0$ such that

$$P\left(\inf_{||\theta||=L}(K_1+K_2)>0\right)\geq 1-\eta.$$

Note that

$$\inf_{||\theta||=L}(K_1+K_2) \;\;\geq\;\; \inf_{||\theta||=L}K_1-\sup_{||\theta||=L}|K_2|$$

so we can show the above result by proving that there exists an η^* such that

$$\begin{aligned}
P\left(\inf_{||\theta||=L}K_1>\eta^*\right) &\geq 1-\eta_1 \\
\text{and}\quad P\left(\sup_{||\theta||=L}|K_2|\leq\eta^*\right) &\geq 1-\eta_2
\end{aligned}$$

for large n and all positive values of η_1 and η_2.

We will first examine K_1.

$$\begin{aligned}
& d_n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{R_{ij}}{\pi_{ij}}\left\{Q_{ij}\left(d_n^{1/2}\right)-Q_{ij}\left(0\right)\right\} \\
= \ & d_n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{R_{ij}}{\pi_{ij}}D_{ij}(\theta,d_n^{1/2}) \\
& +d_n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}E\left[\frac{R_{ij}}{\pi_{ij}}\left(Q_{ij}\left(d_n^{1/2}\right)-Q_{ij}\left(0\right)\right)\mid X,Z\right] \\
& +d_n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{R_{ij}}{\pi_{ij}}\left(\tilde{\mathbf{x}}_{ij}'\theta_1+\tilde{\mathbf{W}}(\mathbf{z}_{ij})'\theta_2\right)\psi_\tau(\varepsilon_{ij}) \\
= \ & G_1+G_2+G_3.
\end{aligned}$$

By Lemma 4.7.3 we have that $\sup_{||\theta||\leq L}|G_1| = o_p(1)$. Next we analyze G_3. We have that

$$\begin{aligned}
& & E\left[\frac{R_{ij}}{\pi_{ij}}\left(\tilde{\mathbf{x}}_{ij}'\theta_1+\tilde{\mathbf{W}}(\mathbf{z}_{ij})'\theta_2\right)\psi_\tau(\varepsilon_{ij})\right] \\
& = & E\left[E\left(\frac{R_{ij}}{\pi_{ij}}\left(\tilde{\mathbf{x}}_{ij}'\theta_1+\tilde{\mathbf{W}}(\mathbf{z}_{ij})'\theta_2\right)\psi_\tau(\varepsilon_{ij})\mid X,Z,Y\right)\right] \\
& = & E\left[\left(\tilde{\mathbf{x}}_{ij}'\theta_1+\tilde{\mathbf{W}}(\mathbf{z}_{ij})'\theta_2\right)\psi_\tau(\varepsilon_{ij})\right] \\
& = & 0 \\
\Longrightarrow \quad E(G_{n3}) & = & 0
\end{aligned}$$

By (C1),

$$\begin{aligned}
\theta_2'\sum_{i=1}^{n}\sum_{j=1}^{m}\tilde{\mathbf{W}}(\mathbf{z}_{ij})\tilde{\mathbf{W}}(\mathbf{z}_{ij})'\theta_2 & \leq & C\theta_2'\sum_{i=1}^{n}\sum_{j=1}^{m}f_{ij}(0)\tilde{\mathbf{W}}(\mathbf{z}_{ij})\tilde{\mathbf{W}}(\mathbf{z}_{ij})'\theta_2 \\
& = & C||\theta_2||^2.
\end{aligned}$$

Thus we have,

$$\begin{aligned} E(G_{n3}^2) &\leq Ck_n^{-1}E\left[n^{-1}\theta_1 X^{*\prime}X^*\theta_1 + ||\theta_2||^2\right] \\ &= O(k_n^{-1/2}||\theta||^2) \\ \Longrightarrow \qquad G_{n3} &= O_p(k_n^{-1/2}||\theta||^2) \end{aligned}$$

We next analyze G_{n2},

$$\begin{aligned} G_{n2} &= k_n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m} E\left[\frac{R_{ij}}{\pi_{ij}}\left(Q_{ij}\left(k_n^{1/2}\right) - Q_{ij}\left(0\right)\right) \mid X, Z\right] \\ &= k_n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m} E\left[\left(Q_{ij}\left(k_n^{1/2}\right) - Q_{ij}\left(0\right)\right) \mid X, Z\right]. \end{aligned}$$

By Knight's Identity twice,

$$\begin{aligned} G_{n2} &= k_n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m} E\Big[-\left\{\sqrt{k_n}\left(\tilde{\mathbf{x}}_{ij}'\theta_1 + \tilde{\mathbf{W}}(\mathbf{z}_{ij})'\theta_2\right) + u_{Nij}\right\}\psi_\tau(\varepsilon_{ij}) \\ &\quad + \int_0^{\sqrt{k_n}\left(\tilde{\mathbf{x}}_{ij}'\theta_1 + \tilde{\mathbf{W}}(\mathbf{z}_{ij})'\theta_2\right)+u_{Nij}} \left[I(\varepsilon_{ij} < s) - I(\varepsilon_{ij} < 0)\right] ds \\ &\quad + u_{Nij}\psi_\tau(\varepsilon_{ij}) - \int_0^{u_{Nij}} \left[I(\varepsilon_{ij} < s) - I(\varepsilon_{ij} < 0)\right] ds \mid X, Z\Big] \\ &= k_n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m} E\Big[\int_{u_{Nij}}^{\sqrt{k_n}\left(\tilde{\mathbf{x}}_{ij}'\theta_1 + \tilde{\mathbf{W}}(\mathbf{z}_{ij})'\theta_2\right)+u_{Nij}} [I(\varepsilon_{ij} < s) \\ &\qquad - I(\varepsilon_{ij} < 0)]\, ds \mid X, Z\Big] \\ &= k_n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m} \int_{u_{Nij}}^{\sqrt{k_n}\left(\tilde{\mathbf{x}}_{ij}'\theta_1 + \tilde{\mathbf{W}}(\mathbf{z}_{ij})'\theta_2\right)+u_{Nij}} E\big[I(\varepsilon_{ij} < s) \\ &\qquad - I(\varepsilon_{ij} < 0) \mid X, Z\big]\, ds \\ &= k_n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m} \int_{u_{Nij}}^{\sqrt{k_n}\left(\tilde{\mathbf{x}}_{ij}'\theta_1 + \tilde{\mathbf{W}}(\mathbf{z}_{ij})'\theta_2\right)+u_{Nij}} \left[\int_0^s f_{ij}(x)\, dx\right] ds \end{aligned}$$

$$\begin{aligned}
&= k_n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}\int_{u_{Nij}}^{\sqrt{k_n}\left(\tilde{\mathbf{x}}'_{ij}\theta_1+\tilde{\mathbf{W}}(\mathbf{z}_{ij})'\theta_2\right)+u_{Nij}}\left[\int_0^s f_{ij}(0)(1+o(1))\,dx\right]ds\\
&= k_n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}\int_{u_{Nij}}^{\sqrt{k_n}\left(\tilde{\mathbf{x}}'_{ij}\theta_1+\tilde{\mathbf{W}}(\mathbf{z}_{ij})'\theta_2\right)+u_{Nij}} f_{ij}(0)s\,ds(1+o(1))\\
&= k_n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m} f_{ij}(0)\Bigg[\frac{1}{2}k_n\left(\tilde{\mathbf{x}}'_{ij}\theta_1+\tilde{\mathbf{W}}(\mathbf{z}_{ij})'\theta_2\right)^2\\
&\quad +u_{Nij}\sqrt{k_n}\left(\tilde{\mathbf{x}}'_{ij}\theta_1+\tilde{\mathbf{W}}(\mathbf{z}_{ij})'\theta_2\right)\Bigg](1+o(1)).
\end{aligned}$$

We consider the cross product term,

$$\begin{aligned}
&k_n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m} f_{ij}(0)k_n\left(\theta'_1\tilde{\mathbf{x}}_{ij}\tilde{\mathbf{W}}(\mathbf{z}_{ij})'\theta_2\right)\\
&= \theta'_1\left(\sum_{i=1}^{n}\sum_{j=1}^{m} f_{ij}(0)\tilde{\mathbf{x}}_{ij}\tilde{\mathbf{W}}(\mathbf{z}_{ij})'\right)\theta_2\\
&= \theta'_1\left(n^{-1/2}X^{*\prime}B_nWW_{B_n}^{-1}\right)\theta_2\\
&= \theta'_1\left(n^{-1/2}X'(I_{nm}-P')B_nWW_{B_n}^{-1}\right)\theta_2\\
&= \theta'_1\left(n^{-1/2}X'(B_nW-B_nW)WW_{B_n}^{-1}\right)\theta_2\\
&= 0.
\end{aligned}$$

Returning to G_{n2} and expanding the quadratic term,

$$\begin{aligned}
G_{n2} &= C\theta'_1\left(n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m} f_{ij}(0)X_i^*X_i^{*\prime}\right)\theta_1\times(1+o(1))\\
&\quad +C\theta'_2\left(\sum_{i=1}^{n}\sum_{j=1}^{m} f_{ij}(0)\tilde{\mathbf{W}}(\mathbf{z}_{ij})\tilde{\mathbf{W}}(\mathbf{z}_{ij})'\right)\theta_2\times(1+o(1))\\
&\quad +k_n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{m} f_{ij}(0)u_{Nij}\left(\tilde{\mathbf{x}}'_{ij}\theta_1+\tilde{\mathbf{W}}(\mathbf{z}_{ij})'\theta_2\right)\\
&= C\theta'_1\left(n^{-1}X^{*\prime}\hat{M}_nBX^*\right)\theta_1\times(1+o(1))+C||\theta_2||\times(1+o(1))
\end{aligned}$$

$$+k_n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{m} f_{ij}(0)u_{Nij}\left(\tilde{\mathbf{x}}_{ij}'\theta_1+\tilde{\mathbf{W}}(\mathbf{z}_{ij})'\theta_2\right).$$

There exists a constant $c > 0$, such that

$$C\theta_1'\left(n^{-1}X^{*\prime}B_nX^*\right)\theta_1\times(1+o(1))+C||\theta_2||\times(1+o(1)) \quad \geq \quad c||\theta||^2$$

with probability one.

Let $U_n=(u_{n11},\ldots,u_{nnm})'$. By Schumaker, we know that $||U_n||=O\left(n^{1/2}k_n^{-r}\right)$. We have

$$\begin{aligned}
k_n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{m} f_{ij}(0)u_{Nij}\tilde{\mathbf{x}}_{ij}'\theta_1 &= k_n^{-1/2}n^{-1/2}\theta_1'X^{*\prime}BU_n\\
&\leq k_n^{-1/2}n^{-1/2}||\theta_1||\cdot||X^*||||BU_n||\\
&= O_p(k_n^{1/2}n^{1/2}k_n^{-r})||\theta||\\
&= O_p(||\theta||).
\end{aligned}$$

Similarly,

$$k_n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{m} f_{ij}(0)u_{Nij}\tilde{\mathbf{W}}(\mathbf{z}_{ij})'\theta_2 \quad = \quad O_p(||\theta||).$$

Thus, for large n, the quadratic term will dominate, and we have that $k_n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}\left(Q_{ij}(k_n^{1/2})-Q_{ij}(0)\right)$ has an asymptotic lower bound of cL^2.

We can then conclude that for any $\eta_1 > 0$

$$P\left(\inf_{||\theta||=L}K_1 > cL^2\right)\geq 1-\eta_1.$$

We now need to show that for any $\eta_2 > 0$,

$$P\left(\sup_{||\theta||=L} K_2 \leq cL^2\right) \geq 1 - \eta_2.$$

We have that

$$\begin{aligned}
|K_2| &= k_n^{-1}\bigg|\sum_{i=1}^{n}\sum_{j=1}^{m} R_{ij}\left(\frac{1}{\hat{\pi}_{ij}} - \frac{1}{\pi_{ij}}\right) \\
&\quad \times \int_{-u_{Nij}}^{-k_n^{1/2}\tilde{\mathbf{x}}_{ij}'\theta_1 - k_n^{1/2}\tilde{\mathbf{W}}(\mathbf{z}_{ij})'\theta_2 - u_{Nij}} \psi_\tau(\varepsilon_{ij} + s)\, ds\bigg| \\
&\leq k_n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m} R_{ij}\bigg|\left(\frac{1}{\hat{\pi}_{ij}} - \frac{1}{\pi_{ij}}\right)\bigg| \\
&\quad \times\bigg|\int_{-u_{Nij}}^{-k_n^{1/2}\tilde{\mathbf{x}}_{ij}'\theta_1 - k_n^{1/2}\tilde{\mathbf{W}}(\mathbf{z}_{ij})'\theta_2 - u_{Nij}} \psi_\tau(\varepsilon_{ij} + s)\, ds\bigg|
\end{aligned}$$

Note that

$$\begin{aligned}
&\max_{ij}\bigg|\int_{-u_{Nij}}^{-k_n^{1/2}\tilde{\mathbf{x}}_{ij}'\theta_1 - k_n^{1/2}\tilde{\mathbf{W}}(\mathbf{z}_{ij})'\theta_2 - u_{Nij}} \psi_\tau(\varepsilon_{ij} + s)\, ds\bigg| \\
\leq\ & |k_n^{1/2}\tilde{\mathbf{x}}_{ij}'\theta_1| + |k_n^{1/2}\tilde{\mathbf{W}}(\mathbf{z}_{ij})'\theta_2| + 2|u_{Nij}|
\end{aligned}$$

We will look at each piece individually.

$$\begin{aligned}
\text{from Schumaker p 227} \quad \max_{ij}|u_{Nij}| &= O(k_n^{-r}) \\
\max_{i,j}|k_n^{1/2}\tilde{\mathbf{W}}(\mathbf{z}_{ij})'\theta_2| &\leq k_n^{1/2}\ ||\theta||\ \max_{i,j}||\tilde{\mathbf{W}}(\mathbf{z}_{ij})|| \\
&= O_p\left(||\theta||\ k_n n^{-1/2}\right) \\
\max_{i,j}|k_n^{1/2}\tilde{\mathbf{x}}_{ij}'\theta_1| &\leq k_n^{1/2}\ ||\theta||\ \max_{i,j}||\tilde{\mathbf{x}}_{ij}|| \\
&= O_p\left(||\theta||\ k_n^{1/2} n^{-1/2}\right).
\end{aligned}$$

Thus

$$\max_{ij}\left|\int_{-u_{Nij}}^{-k_n^{1/2}\tilde{\mathbf{x}}_{ij}'\theta_1-k_n^{1/2}\tilde{\mathbf{W}}(\mathbf{z}_{ij})'\theta_2-u_{Nij}}\psi_\tau(\varepsilon_{ij}+s)\,ds\right| = O_p\left(k_n n^{-1/2}\ ||\theta||\right).$$

We now have

$$\begin{aligned}\sup_{||\theta||=L}|K_2| &\leq k_n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}R_{ij}O_p\left(n^{-1/2}\right)O_p\left(||\theta||k_n n^{-1/2}\right)\\ &= O_p(||\theta||).\end{aligned}$$

Therefore, for any $\eta_2 > 0$,

$$P\left(\sup_{||\theta||=L}K_2 \leq L\right) \geq 1-\eta_2.$$

By convexity implies $||\hat{\theta}|| = O_p(k_n^{1/2})$. Therefore, it follows that

$$||W_B(\hat{\gamma}-\gamma)|| = O_p(k_n^{1/2}).$$

Now

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}R_{ij}f_{ij}(0)\left(\sum_{d=1}^{q}\hat{g}_d(Z_{ij})-\sum_{d=1}^{q}g_d(Z_{ij})\right)^2\\ =&\ n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}f_{ij}(0)\left(\tilde{\mathbf{W}}(\mathbf{z}_{ij})'(\hat{\gamma}-\gamma)-u_{Nij}\right)^2\\ \leq&\ n^{-1}(\hat{\gamma}-\gamma)'W_B^2(\hat{\gamma}-\gamma)+O_p(n^{-2r/(2r+1)})\\ =&\ O_p(n^{-2r/(2r+1)}).\end{aligned}$$

We have that $\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}R_{ij}\left(\sum_{d=1}^{q}\hat{g}_d(Z_{ij})-\sum_{d=1}^{q}g_d(Z_{ij})\right)^2 = O_p(n^{-2r/(2r+1)})$ because $f_{ij}(0)$ has a lower and upper bound.

Now $\hat{\theta}_1 = \sqrt{n}(\hat{\beta} - \beta)$ and it also follows that $||\hat{\theta}_1|| = O_p(k_n^{1/2})$. Thus

$$||\hat{\beta} - \beta|| \;\; = \;\; O_p(n^{-1/2}k_n^{1/2}). \qquad \square$$

We now prove asymptotic normality from the first part of Theorem 4.3.1. First we state some necessary lemmas.

Lemma 4.7.4

$$\begin{aligned} n^{-1/2}X^* &= n^{-1/2}\Delta_n + o_p(1) \\ \text{and} \qquad n^{-1}X^{*\prime}B_nX^* &= \Sigma_1 + o_p(1). \end{aligned}$$

Proof

By definition,

$$\begin{aligned} n^{-1/2}X^* &= n^{-1/2}\left(X - PX\right) \\ &= n^{-1/2}\left(H + \Delta_n - PX\right) \\ &= n^{-1/2}\Delta_n + n^{-1/2}\left(H - PX\right) \end{aligned}$$

Now let $\gamma_k^* \in \mathbb{R}^{J_n}$ be defined as $\gamma_k^* = \arg\min_{\gamma\in\mathbb{R}^{J_n}} \sum_{i=1}^n \sum_{j=1}^m \left(R_{ij}/\hat{\pi}_{ij}\right) f_{ij}(0 \mid \mathbf{x}_{ij}, \mathbf{z}_{ij}) \left\{X_{ijk} - \mathbf{W}(\mathbf{z}_{ij})'\gamma\right\}^2$. Let $\hat{h}_k(\mathbf{z}_{ij}) = \mathbf{W}(\mathbf{z}_{ij})'\gamma_k^*$ and notice that $(PX)_{m(i-1)+j,k} = \hat{h}_k(\mathbf{z}_{ij})$. It follows that

$$\begin{aligned} n^{-1}||H - PX||^2 &= n^{-1}\lambda_{\max}\left\{\left(H - PX\right)'\left(H - PX\right)\right\} \\ &\leq n^{-1}\text{trace}\left\{\left(H - PX\right)'\left(H - PX\right)\right\} \\ &= n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{p}\left(h_k^*(\mathbf{z}_{ij}) - \hat{h}_{ij}(\mathbf{z}_{ij})\right)^2 \end{aligned}$$

$$
\begin{aligned}
&= O_p\left(pn^{-2r/(2r+1)}\right) \\
&= o_p(1).
\end{aligned}
$$

For the second equation, note that

$$
\begin{aligned}
n^{-1}X^{*\prime}B_nX^* &= n^{-1}\left(\Delta_n + o_p(1)\right)' B_n \left(\Delta_n + o_p(1)\right) \\
&= n^{-1}\left[\Delta_n' B_n \Delta_n + \Delta_n' B_n o_p(1) + B_n \Delta_n o_p(1) + B_n o_p(1)\right] \\
&= n^{-1}\Delta_n' B_n \Delta_n + o_p(1) \\
&= \Sigma_1 + o_p(1)
\end{aligned}
$$

□

We define the following variables:

$$
\begin{aligned}
\tilde{R} &= \operatorname{diag}\left(R_{11}\hat{\pi}_{11}^{-1}, \ldots, R_{nm}\hat{\pi}_{nm}^{-1}\right) \\
\psi_\tau(\varepsilon) &\equiv \left(\psi_\tau(\varepsilon_{11}), \ldots, \psi_\tau(\varepsilon_{nm})\right)' \\
\tilde{\theta}_1 &= \sqrt{n}\left(X^{*T}B_nX^*\right)^{-1} X^{*T}\tilde{R}\psi_\tau(\varepsilon) \\
Q_{ij}^*(\theta_1, \tilde{\theta}_1, \theta_2) &= \rho_\tau\left\{\varepsilon_{ij} - \tilde{\mathbf{x}}_{ij}'\theta_1 - \tilde{W}(\mathbf{z}_{ij})'\theta_2 - u_{nij}\right\} \\
&\quad -\rho_\tau\left\{\varepsilon_{ij} - \tilde{\mathbf{x}}_{ij}'\tilde{\theta}_1 - \tilde{W}(\mathbf{z}_{ij})'\theta_2 - u_{nij}\right\}
\end{aligned}
$$

We want to show that $\hat{\theta}_1$ is asymptotically equivalent to $\tilde{\theta}_1$. The following lemmas are similar to Lemmas A2-A5 in Sherwood [2016].

Lemma 4.7.5

Define $\Delta(B)_n = n^{-1}\Delta_n' B_n \Delta_n$. Then,

$$
\sup_{\substack{\|\theta_1-\tilde{\theta}_1\|\le M \\ \|\theta_2\|\le C\sqrt{d_n}}} \Big| \sum_{i=1}^{n}\sum_{j=1}^{m} \frac{R_{ij}}{\hat{\pi}_{ij}} \mathrm{E}_s\left\{Q_{ij}^*(\theta_1, \tilde{\theta}_1, \theta_2)\right\} - \frac{1}{2}\left\{\theta_1'\Delta(B)_n\theta_1 - \tilde{\theta}_1\Delta(B)_n\tilde{\theta}_1\right\} \Big| \quad = \quad o_p(1).
$$

Proof

$$
\begin{aligned}
& \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{R_{ij}}{\hat{\pi}_{ij}}\mathrm{E}_s\left\{Q^*_{ij}(\theta_1,\tilde{\theta}_1,\theta_2)\right\} \\
= & \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{R_{ij}}{\hat{\pi}_{ij}}\mathrm{E}_s\left[\int_{-\left\{\tilde{\mathbf{x}}'_{ij}\tilde{\theta}_1+\tilde{\mathbf{W}}(\mathbf{z}_{ij})'\theta_2+u_{nij}\right\}}^{-\left\{\tilde{\mathbf{x}}'_{ij}\theta_1+\tilde{\mathbf{W}}(\mathbf{z}_{ij})'\theta_2+u_{nij}\right\}}\psi_\tau(\varepsilon_{ij}+s)\,ds\right] \\
= & -\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{R_{ij}}{\hat{\pi}_{ij}}\int_{-\left\{\tilde{\mathbf{x}}'_{ij}\tilde{\theta}_1+\tilde{\mathbf{W}}(\mathbf{z}_{ij})'\theta_2+u_{nij}\right\}}^{-\left\{\tilde{\mathbf{x}}'_{ij}\theta_1+\tilde{\mathbf{W}}(\mathbf{z}_{ij})'\theta_2+u_{nij}\right\}}\left[F_{ij}(-s\mid\mathbf{x}_{ij},\mathbf{z}_{ij})-F_{ij}(0\mid\mathbf{x}_{ij},\mathbf{z}_{ij})\right]ds \\
= & \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{R_{ij}}{\hat{\pi}_{ij}}f_{ij}(0\mid\mathbf{x}_{ij},\mathbf{z}_{ij})(1+o(1))\Big[\left(\tilde{\mathbf{x}}'_{ij}\theta_1+\tilde{\mathbf{W}}(\mathbf{z}_{ij})'\theta_2+u_{nij}\right)^2- \\
& \left(\tilde{\mathbf{x}}'_{ij}\tilde{\theta}_1+\tilde{\mathbf{W}}(\mathbf{z}_{ij})'\theta_2+u_{nij}\right)^2\Big] \\
= & \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{R_{ij}}{\hat{\pi}_{ij}}f_{ij}(0\mid\mathbf{x}_{ij},\mathbf{z}_{ij}) \\
& \times\left[\left(\tilde{\mathbf{x}}'_{ij}\theta_1\right)^2-\left(\tilde{\mathbf{x}}'_{ij}\tilde{\theta}_1\right)^2+2\left(\tilde{\mathbf{W}}(\mathbf{z}_{ij})'\theta_2+u_{nij}\right)\left(\tilde{\mathbf{x}}'_{ij}\theta_1-\tilde{\mathbf{x}}'_{ij}\tilde{\theta}_1\right)\right](1+o(1)) \\
= & \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m}\left(\frac{R_{ij}}{\pi_{ij}}+\frac{R_{ij}}{\hat{\pi}_{ij}}-\frac{R_{ij}}{\pi_{ij}}\right)f_{ij}(0\mid\mathbf{x}_{ij},\mathbf{z}_{ij}) \\
& \left[\left(\tilde{\mathbf{x}}'_{ij}\theta_1\right)^2-\left(\tilde{\mathbf{x}}'_{ij}\tilde{\theta}_1\right)^2+2\left(\tilde{\mathbf{W}}(\mathbf{z}_{ij})'\theta_2+u_{nij}\right)\left(\tilde{\mathbf{x}}'_{ij}\theta_1-\tilde{\mathbf{x}}'_{ij}\tilde{\theta}_1\right)\right](1+o(1)) \\
= & \frac{1}{2}\left(\theta'_1\Delta(B_n)\theta_1-\tilde{\theta}'\Delta(B_n)\tilde{\theta}_1\right)(1+o(1)) \\
& +n^{-1/2}(\theta_1-\tilde{\theta}_1)'\Big)\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{R_{ij}}{\pi_{ij}}f_{ij}(0\mid\mathbf{x}_{ij},\mathbf{z}_{ij})\delta_{ij}u_{nij}(1+o(1)) \\
& +\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m}\left(\frac{R_{ij}}{\hat{\pi}_{ij}}-\frac{R_{ij}}{\pi_{ij}}\right)\left[\left(\tilde{\mathbf{x}}'_{ij}\theta_1\right)^2-\left(\tilde{\mathbf{x}}'_{ij}\tilde{\theta}_1\right)^2\right] \\
& +n^{-1/2}(\theta_1-\tilde{\theta}_1)'\sum_{i=1}^{n}\sum_{j=1}^{m}\left(\frac{R_{ij}}{\hat{\pi}_{ij}}-\frac{R_{ij}}{\pi_{ij}}\right)f_{ij}(0\mid\mathbf{x}_{ij},\mathbf{z}_{ij})\delta_{ij}u_{nij}(1+o(1)).
\end{aligned}
$$

Now we examine each piece individually. First, we have that $\delta_{ij}=\mathbf{x}_{ij}-h^*(\mathbf{z}_{ij})$ and

$\mathrm{E}(\mathbf{x}_{ij}) = 0$ and $\mathrm{E}(h^*(\mathbf{z}_{ij})) = 0$, thus $\mathrm{E}(\delta_{ij}) = 0$ and

$$\mathrm{E}\left[n^{-1/2}(\theta_1 - \tilde{\theta}_1)') \sum_{i=1}^{n}\sum_{j=1}^{m} \frac{R_{ij}}{\hat{\pi}_{ij}} f_{ij}(0 \mid \mathbf{x}_{ij}, \mathbf{z}_{ij}) \delta_{ij} u_{nij}(1 + o(1))\right] = 0.$$

Note that $\max_{ij} |u_{nij}| = O(k_n^{-r})$ so

$$\begin{aligned} \mathrm{Var}\left[n^{-1/2}(\theta_1 - \tilde{\theta}_1)') \sum_{i=1}^{n}\sum_{j=1}^{m} \frac{R_{ij}}{\hat{\pi}_{ij}} f_{ij}(0 \mid \mathbf{x}_{ij}, \mathbf{z}_{ij}) \delta_{ij} u_{nij}(1 + o(1))\right] &= o(1) \\ \Longrightarrow\ n^{-1/2}(\theta_1 - \tilde{\theta}_1)') \sum_{i=1}^{n}\sum_{j=1}^{m} \frac{R_{ij}}{\hat{\pi}_{ij}} f_{ij}(0 \mid \mathbf{x}_{ij}, \mathbf{z}_{ij}) \delta_{ij} u_{nij}(1 + o(1)) &= o_p(1). \end{aligned}$$

The rest of the proof follows Sherwood [2016] (Lemma A2). □

Lemma 4.7.6

Under the conditions of Theorem 4.3.1, then for any given positive constants M and C,

$$\sup_{\substack{||\theta_1 - \tilde{\theta}_1|| \le M \\ ||\theta_2|| \le C\sqrt{d_n}}} \Big| \sum_{i=1}^{n}\sum_{j=1}^{m} \frac{R_{ij}}{\hat{\pi}_{ij}} \left[Q^*_{ij}(\theta_1, \tilde{\theta}_1, \theta_2) - \mathrm{E}_s\left\{Q^*_{ij}(\theta_1, \tilde{\theta}_1, \theta_2)\right\} + \tilde{x}'_{ij}(\theta_1 - \tilde{\theta}_1)\psi_\tau(\varepsilon_{ij})\right] \Big| \\ o_p(1).$$

Proof

The proof follows Sherwood [2016] (Lemma A3). □

Lemma 4.7.7

Under the conditions of Theorem 4.3.1,

$$\hat{\theta}_1 - \tilde{\theta}_1 = o_p(1).$$

Proof

The proof follows Sherwood [2016] (Lemma A4). □

Lemma 4.7.8

Under the conditions of Theorem 4.3.1,

$$\begin{aligned} & n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{R_{ij}}{\tilde{\pi}_{ij}}\delta_{ij}\psi_\tau(\varepsilon_{ij}) \\ = \;& n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{R_{ij}}{\pi_{ij}}\delta_{ij}\psi_\tau(\varepsilon_{ij}) \\ & -n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{R_{ij}-\pi_{ij}}{\pi_{ij}}\mathrm{E}\left[\delta_{ij}\psi_\tau(\varepsilon_{ij}) \mid \mathbf{t}_{ij}\right] + o_p(1) \end{aligned}$$

Proof

The proof follows Sherwood [2016] (Lemma A5). □

Lemma 4.7.9

Let $H_{ij}(\boldsymbol{\gamma}^*) \equiv \nabla^2_{\boldsymbol{\gamma}}\frac{1}{\pi_{ij0}}(\boldsymbol{\gamma}^*)$ be the Hessian matrix evaluated at $\boldsymbol{\gamma}^*$. Under the conditions of Theorem 4.3.1,

$$\begin{aligned} n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{R_{ij}}{\hat{\pi}_{ij}}\delta_{ij}\psi_\tau(\varepsilon_{ij}) &= n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{R_{ij}}{\pi_{ij0}}\delta_{ij}\psi_\tau(\varepsilon_{ij}) \\ &\quad +n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{m}R_{ij}\frac{1}{\pi_{ij0}}\left(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}\right)'\nabla_{\boldsymbol{\gamma}}\frac{1}{\pi_{ij0}}\delta_{ij}\psi_\tau(\varepsilon_{ij}) \\ &\quad +o_p(1). \end{aligned}$$

Proof

$$n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{R_{ij}}{\hat{\pi}_{ij}}\delta_{ij}\psi_\tau(\varepsilon_{ij}) = n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{R_{ij}}{\pi_{ij}}\delta_{ij}\psi_\tau(\varepsilon_{ij})$$

$$+n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{m} R_{ij}\left(\frac{1}{\hat{\pi}_{ij}}-\frac{1}{\pi_{ij}}\right)\delta_{ij}\psi_{\tau}(\varepsilon_{ij}).$$

Note that, $\mathbf{T}_{ij} = (T_{i2}', \ldots, T_{ij}')'$, $\pi(\mathbf{T}_{ij}, \boldsymbol{\gamma}_j) = \prod_{k=2}^{j} e^{T_{ik}'\gamma_k}/(1+e^{T_{ik}'\gamma_k})$, and

$$\begin{aligned}
\nabla_{\gamma_k}\frac{1}{\pi_{ij0}} &= \frac{-1}{\pi_{ij0}} I(k \le j)\eta\left(T_{ik},\gamma_k\right)e^{-T_{ik}'\gamma_k}T_{ik}\\
\nabla_{\gamma_k,\gamma_d}\frac{1}{\pi_{ij0}} &= \frac{1}{\pi_{ij0}} I(k \le j)I(d \le j)\eta\left(T_{ik},\gamma_k\right)\eta\left(T_{id},\gamma_d\right)e^{-T_{ik}'\gamma_k-T_{id}'\gamma_d}T_{ik}T_{id}'\\
\nabla_{\gamma_k}^2\frac{1}{\pi_{ij0}} &= \frac{1}{\pi_{ij0}} I(k \le j)\eta\left(T_{ik},\gamma_k\right)e^{-T_{ik}'\gamma_k}T_{ik}T_{ik}'.
\end{aligned}$$

Let $\hat{\boldsymbol{\gamma}} = (\hat{\gamma}_2', \ldots, \hat{\gamma}_m')'$. By Taylor's Expansion,

$$\frac{1}{\hat{\pi}_{ij}}-\frac{1}{\pi_{ij}} = \left(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}\right)'\nabla\boldsymbol{\gamma} + \frac{1}{2}\left(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}\right)'\nabla_{\boldsymbol{\gamma}}^2\left(\hat{\boldsymbol{\gamma}}_j-\boldsymbol{\gamma}_j\right)+o_p(1).$$

By Taylor's Theorem,

$$\frac{1}{\hat{\pi}_{ij}}-\frac{1}{\pi_{ij}} = \left(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}\right)'\nabla\boldsymbol{\gamma} + \frac{1}{2}\left(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}\right)'H_{ij}(\boldsymbol{\gamma}^*)\left(\hat{\boldsymbol{\gamma}}_j-\boldsymbol{\gamma}_j\right)$$

for some $\boldsymbol{\gamma}_j^* = \boldsymbol{\gamma}_j + t(\hat{\boldsymbol{\gamma}}_j-\boldsymbol{\gamma}_j)$ for $t \in [0,1]$. We now have that

$$\begin{aligned}
& n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{m} R_{ij}\left(\frac{1}{\hat{\pi}_{ij}}-\frac{1}{\pi_{ij}}\right)\delta_{ij}\psi_{\tau}(\varepsilon_{ij})\\
= {} & n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{m} R_{ij}\left(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}\right)'\nabla\boldsymbol{\gamma}\delta_{ij}\psi_{\tau}(\varepsilon_{ij})\\
& +n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{m} R_{ij}\frac{1}{2}\left(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}\right)'H_{ij}(\boldsymbol{\gamma}^*)\left(\hat{\boldsymbol{\gamma}}_j-\boldsymbol{\gamma}_j\right)\delta_{ij}\psi_{\tau}(\varepsilon_{ij})
\end{aligned}$$

We now show that

$n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{m} R_{ij}\frac{1}{2}\left(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}\right)' H_{ij}(\boldsymbol{\gamma}^{*})\left(\hat{\boldsymbol{\gamma}}_j-\boldsymbol{\gamma}_j\right)\delta_{ij}\psi_\tau(\varepsilon_{ij}) = o_p(1).$

$$\begin{aligned}
& n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{m} R_{ij}\frac{1}{2}\left(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}\right)' H_{ij}(\boldsymbol{\gamma}^{*})\left(\hat{\boldsymbol{\gamma}}_j-\boldsymbol{\gamma}_j\right)\delta_{ij}\psi_\tau(\varepsilon_{ij}) \\
= \;& \sqrt{n}\left(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}\right)'\left[\sum_{j=1}^{m}\frac{1}{n\sqrt{n}}\sum_{i=1}^{n} R_{ij}\frac{1}{2}H_{ij}(\boldsymbol{\gamma}^{*})\delta_{ij}\psi_\tau(\varepsilon_{ij})\right]\sqrt{n}\left(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}\right)
\end{aligned}$$

Assuming that $\hat{\boldsymbol{\gamma}}$ for $\boldsymbol{\gamma}$ satisfies the regularity conditions of asymptotic normality of MLEs for exponential family models, then the problem is reduced to showing for any element of the standard basis, e_d, e_k that

$$\sum_{j=1}^{m}\frac{1}{n\sqrt{n}}\sum_{i=1}^{n} R_{ij}\frac{1}{2}e_d' H_{ij}(\boldsymbol{\gamma}^{*})e_k\delta_{ij}\psi_\tau(\varepsilon_{ij}) \quad = \quad o_p(1).$$

We know that $\boldsymbol{\gamma}^{*} \xrightarrow{P} \boldsymbol{\gamma}$ and that $||H(\boldsymbol{\gamma}_0)|| < C < \infty$ for some constant C. Thus $P\left(||H(\boldsymbol{\gamma}^{*})|| < C\right) \to 1$. Thus by the Law of Large Numbers,

$$\sum_{j=1}^{m}\frac{1}{n\sqrt{n}}\sum_{i=1}^{n} R_{ij}\frac{1}{2}e_d' H_{ij}(\boldsymbol{\gamma}^{*})e_k\delta_{ij}\psi_\tau(\varepsilon_{ij}) \quad = \quad o_p(1).$$

We have now shown that

$$\begin{aligned}
& n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{R_{ij}}{\pi(\mathbf{T}_{ij},\hat{\boldsymbol{\gamma}}_j)}\delta_{ij}\psi_\tau(\varepsilon_{ij}) \\
= \;& n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{R_{ij}}{\pi(\mathbf{T}_{ij},\boldsymbol{\gamma}_j)}\delta_{ij}\psi_\tau(\varepsilon_{ij}) \\
& + n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{m} R_{ij}\left[\frac{1}{\pi(\mathbf{T}_{ij},\boldsymbol{\gamma}_j)}\left(\hat{\boldsymbol{\gamma}}_j-\boldsymbol{\gamma}_j\right)' H(\boldsymbol{\gamma}^{*})\mathbf{T}_{ij}\right]\delta_{ij}\psi_\tau(\varepsilon_{ij}) \\
& + o_p(1).
\end{aligned}$$

□

Proof of second part of Theorem 4.3.1

We now continue with proving asymptotic normality of $\hat{\beta}$. Noting that $\hat{\theta}_1 = \sqrt{n}(\hat{\beta} - \beta)$, by Lemma 4.7.7,

$$\begin{aligned}\sqrt{n}\left(\hat{\beta} - \beta\right) &= \sqrt{n}\left(X^* B_n X^{*T}\right)^{-1} X^{*T} \tilde{R} \psi_\tau(\varepsilon) + o_p(1) \\ &= \left(\frac{1}{n} X^* B_n X^{*T}\right)^{-1} \frac{1}{\sqrt{n}} X^{*T} \tilde{R} \psi_\tau(\varepsilon) + o_p(1)\end{aligned}$$

By Lemma 4.7.4,

$$\sqrt{n}\left(\hat{\beta} - \beta\right) = \left\{\Sigma_1 + o_p(1)\right\}^{-1} n^{-1/2} \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{R_{ij}}{\hat{\pi}_{ij}} \delta_{ij} \psi_\tau(\varepsilon_{ij}) \left\{1 + o_p(1)\right\},$$

and by Lemma 4.7.9,

$$\begin{aligned}\sqrt{n}\left(\hat{\beta} - \beta\right) &= \left\{\Sigma_1 + o_p(1)\right\}^{-1} \Big\{ n^{-1/2} \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{R_{ij}}{\pi_{ij0}} \delta_{ij} \psi_\tau(\varepsilon_{ij}) \\ &\quad + n^{-1/2} \sum_{i=1}^{n} \sum_{j=1}^{m} R_{ij} \frac{1}{\pi_{ij0}} \left(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\right)' \nabla_{\boldsymbol{\gamma}} \frac{1}{\pi_{ij0}} \delta_{ij} \psi_\tau(\varepsilon_{ij}) + o_p(1) \Big\} \\ &= \left\{\Sigma_1 + o_p(1)\right\}^{-1} \Bigg\{ n^{-1/2} \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{R_{ij}}{\pi_{ij0}} \delta_{ij} \psi_\tau(\varepsilon_{ij}) \\ &\quad + \left[n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} R_{ij} \frac{1}{\pi_{ij0}} \delta_{ij} \psi_\tau(\varepsilon_{ij}) \nabla_{\boldsymbol{\gamma}} \frac{1}{\pi_{ij0}}' \right] \sqrt{n}\left(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\right) \\ &\quad + o_p(1) \Bigg\}\end{aligned}$$

From MLE theory, we have that

$$\sqrt{n}\left(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\right) = \frac{1}{\sqrt{n}} I(\boldsymbol{\gamma})^{-1} \sum_{i=1}^{n} \nabla \ell_i(\boldsymbol{\gamma}) + o_p(1).$$

Now let $D_{2jk} = \mathrm{E}\left[\frac{I(k\leq j)\eta(T_{ik},\gamma_k)}{\pi_{ij0}}(1-\eta(T_{ik},\gamma_k))\psi_\tau(\varepsilon_{ij})\delta_{ij}T'_{ik}\right]$ and $D_{2j} = (D_{2j2},\ldots,D_{2jm})$. By the WLLN,

$$n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m} R_{ij}\frac{1}{\pi_{ij0}}\delta_{ij}\psi_\tau(\varepsilon_{ij})\nabla_{\boldsymbol{\gamma}}\frac{1}{\pi_{ij0}}' \;=\; \sum_{j=2}^{m} D_{2j} + o_p(1).$$

Putting things back together, we have

$$\begin{aligned}
& n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{m} R_{ij}\frac{1}{\pi_{ij0}}\left(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}\right)'\nabla_{\boldsymbol{\gamma}}\frac{1}{\pi_{ij0}}\delta_{ij}\psi_\tau(\varepsilon_{ij}) \\
={}& \frac{1}{\sqrt{n}}\sum_{j=2}^{m} D_{2j}I(\boldsymbol{\gamma})^{-1}\sum_{i=1}^{n}\nabla\ell_i(\boldsymbol{\gamma}) + o_p(1).
\end{aligned}$$

We now consider the expectation and variance of $\sqrt{n}\left(\hat{\beta}-\beta\right)$. Let $k_i = (\psi_\tau(\varepsilon_{i1}), \psi_\tau(\varepsilon_{i2})R_{i2}/\pi(\mathbf{T}_{i2},\boldsymbol{\gamma}_2),\ldots,\psi_\tau(\varepsilon_{im})R_{im}/\pi(\mathbf{T}_{im},\boldsymbol{\gamma}_m))'$ and K_i be an $m\times m$ diagonal matrix with k_i on the diagonal. Also, note that $\delta_i = (\delta_{i1},\ldots,\delta_{im})'$, $R_{i1}=1$, and $\pi(\mathbf{T}_{i1},\boldsymbol{\gamma}_1)=1$.

$$\begin{aligned}
\mathrm{E}\left[\frac{1}{\sqrt{n}}\sum_{j=2}^{m} D_{2j}I(\boldsymbol{\gamma})^{-1}\sum_{i=1}^{n}\nabla\ell_i(\boldsymbol{\gamma})\right] &= 0 \\
\mathrm{E}\left[n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{R_{ij}}{\pi_{ij0}}\delta_{ij}\psi_\tau(\varepsilon_{ij})\right] &= 0 \\
\mathrm{Var}\left[\frac{1}{\sqrt{n}}\sum_{j=2}^{m} D_{2j}I(\boldsymbol{\gamma})^{-1}\sum_{i=1}^{n}\nabla\ell_i(\boldsymbol{\gamma})\right] &= \sum_{j=1}^{m} D_{2j}I(\boldsymbol{\gamma})^{-1}D'_{2j} \\
\mathrm{Var}\left[n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{R_{ij}}{\pi_{ij0}}\delta_{ij}\psi_\tau(\varepsilon_{ij})\right] &= \mathrm{E}\left[\delta'_i k_i k'\delta_i\right] \\
&\equiv D_3
\end{aligned}$$

For the covariance

$$\begin{aligned}
& \operatorname{Cov}\left[\frac{1}{\sqrt{n}}\sum_{j=2}^{m} D_{2j} I(\boldsymbol{\gamma})^{-1}\sum_{i=1}^{n}\nabla\ell_i(\boldsymbol{\gamma}), n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{R_{ij}}{\pi_{ij0}}\delta_{ij}\psi_\tau(\varepsilon_{ij})\right] \\
= & \operatorname{Cov}\left[\sum_{j=2}^{m} D_{2j} I(\boldsymbol{\gamma})^{-1}\nabla\ell_i(\boldsymbol{\gamma}), \delta_i' k_i\right] \\
= & \operatorname{E}\left[\sum_{j=2}^{m} D_{2j} I(\boldsymbol{\gamma})^{-1}\nabla\ell_i(\boldsymbol{\gamma}) k_i'\delta_i\right] \\
= & \sum_{j=2}^{m} D_{2j} I(\boldsymbol{\gamma})^{-1} D_{2j}'.
\end{aligned}$$

We finally show normality by using the Lindeberg-Feller CLT. Let

$$\begin{aligned}
D_{ni} \;=\; & \Sigma_1^{-1}\Bigg\{ n^{-1/2}\sum_{j=1}^{m}\frac{R_{ij}}{\pi_{ij0}}\delta_{ij}\psi_\tau(\varepsilon_{ij}) \\
& + \left[n^{-1}\sum_{j=1}^{m} R_{ij}\frac{1}{\pi_{ij0}}\delta_{ij}\psi_\tau(\varepsilon_{ij})\nabla\boldsymbol{\gamma}\frac{1}{\pi_{ij0}}'\right]\sqrt{n}\left(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}\right)\Bigg\}
\end{aligned}$$

and we know that

$$\operatorname{Var}(D_{ni}) \;=\; \frac{1}{n}\Sigma_1^{-1}\left(D_3 - \sum_{j=2}^{m} D_{2j} I(\boldsymbol{\gamma})^{-1} D_{2j}'\right)\Sigma_1^{-1}.$$

Let $s_n^2 = n\operatorname{Var}(||D_{ni}||)$ and C be a constant that can change from line to line.

$$\begin{aligned}
& \frac{1}{s_n^2}\sum_{i=1}^{n}\operatorname{E}\left[||D_{ni}||^2 I(||D_{ni}|| > \varepsilon s_n)\right] \\
\leq & \frac{1}{s_n^2}\sum_{i=1}^{n}\operatorname{E}\left[||D_{ni}||^4\right]\operatorname{E}\left[I\left(||D_{ni}|| > \epsilon s_n\right)\right] \\
= & \frac{1}{s_n^2}\sum_{i=1}^{n}\operatorname{E}\left[||D_{ni}||^4\right] P\left(||D_{ni}|| > \epsilon s_n\right)
\end{aligned}$$

$$\begin{aligned}
&\leq \frac{1}{s_n^4 \epsilon^2} \sum_{i=1}^{n} \mathrm{E}\left[||D_{ni}||^4\right] \mathrm{Var}\left(||D_{ni}||\right) \\
&= \frac{s_n^2}{n s_n^4 \epsilon^2} \sum_{i=1}^{n} \mathrm{E}\left[||D_{ni}||^4\right] \\
&= \frac{1}{n s_n^2 \epsilon^2} \sum_{i=1}^{n} \mathrm{E}\left[||D_{ni}||^4\right] \\
&= C\left(n\epsilon\right)^{-2} \sum_{i=1}^{n} \mathrm{E}\left[||D_{ni}||^4\right] \\
&\leq C(n\varepsilon)^{-2} \sum_{i=1}^{n} \mathrm{E}\left[||\delta_i||^4\right] \\
&= o(1)
\end{aligned}$$

Thus we have met the Lindeberg-Feller Condition and shown asymptotic normality.

□

Chapter 5

Sparsity Path Algorithm for Penalized Quantile Regression

5.1 Introduction

In this dissertation, we have discussed methods and theoretical results for semiparametric quantile regression and have ignored the challenges of computing the quantile regression estimators. Additionally, we focused on model estimation and did not worry about model selection. Model selection is an important part of analysis because estimators of smaller models that include only relevant covariates are more accurate and have better prediction performance. As the number of potentially important covariates grows thereby increasing the number of combinations of the covariates, model selection becomes more difficult. In this chapter, we turn our attention to the model selection problem and propose an algorithm for model selection in quantile regression.

Consider a dataset with n samples and p covariates. In practice, all p covariates may not have an effect on the response. A model with only the relevant covariates will have better theoretical estimation and prediction properties than a larger model containing covariates with no effect on the response. When $p < n$, a common approach to model selection is to fit all possible 2^p models and select one model using crossvalidation or an information criterion such as AIC or BIC. When p is too large, fitting all possible models becomes too computationally expensive and infeasible. One remedy to this problem is stepwise regression, in which a model is selected by adding or removing one covariate at a time until modifying the current model returns a worse information criterion. Another approach which also works well even in the high-dimensional setting ($p > n$) is penalized quantile regression which simulataneously selects covariates and estimates effects. Letting $Y_1, \dots, Y_n$ denote the n responses and $X_1, \dots, X_n$ denote the p-dimensional vectors of covariates for the n cases, the effects β are estimated by minimizing,

$$\hat{\beta}(\tau) = \operatorname*{arg\,min}_{\beta \in \mathcal{R}^p} \sum_{i=1}^{n} \rho_\tau(Y_i - X_i'\beta) + \sum_{j=1}^{p} p_\lambda(\beta_j),$$

where ρ_τ is the nonsmooth check loss function and p_λ is a non-negative function that depends a tuning parameter λ. The estimated model increases in sparsity (decreases in size) as the value of λ increases. Common choices for the penalty function are the least absolute shrinkage and selection operator [Tibshirani, 1996], lasso, and the nonconvex smoothly clipped absolute deviation[Fan and Li, 2001], SCAD, and minimax concave [Zhang et al., 2010], MCP, penalties. Theoretical properties of the lasso penalized quantile regression model are investigated in Belloni et al. [2011] and for the SCAD and MCP penalized quantile regression model in Wang et al. [2012] and Sherwood and Wang [2016].

Penalized quantile regression has also been studied when $p < n$ [Zou and Yuan, 2008, Kai et al., 2011, Wu and Liu, 2009]. Much of the theoretical literature for penalized quantile regression is restricted to proving consistency of the estimator and does not consider inference. One exception is a recently proposed wild bootstrap procedure to approximate the sampling distribution of the penalized coefficient estimator [Wang et al.].

Inference is important in practice for testing hypotheses and creating confidence intervals. About four decades ago Cox [1975] argued that it may be advantageous to split data into two groups where the first group is used to generate hypotheses and the second group is used for testing. By splitting the data into independent exploratory and testing sets, we can obtain valid p-values for hypthesis testing. More recent research has also ivestigated inference using data splitting [Meinshausen et al., 2009]. Post-selection inference seems to be a fruitful area of research for making inference on penalized quantile regression estimators. In post-selection inference, one or a set of low-dimensional ($p < n$) models is first selected by penalized regression and then inference is made on the selected model or set of models using classical techniques [Wasserman and Roeder, 2009, Berk et al., 2013, Lee et al., 2016, Tibshirani et al., 2016, Taylor and Tibshirani, 2015].

To compute the unpenalized quantile regression estimator, Koenker and Park [1996] proposed an interior point algorithm. Hunter and Lange [2000] proposed an MM algorithm which majorizes the nonsmooth quantile loss function with a quadratic function. For lasso penalized regression, Li and Zhu [2008] developed an algorithm to compute the estimate at all values of the tuning parameter λ and Peng and Wang [2015] combined an MM algorithm with coordinate descent to quickly compute the estimate for nonconvex penalized quantile regression. Recently, the literature has turned to the alternating direction method of multipliers (ADMM) algorithm for quickly computing the penalized quantile regression estimate, e.g. Gu et al. [2017], Yu et al. [2017], Yu and Lin [2017].

All of these algorithms are designed to exactly compute the penalized estimate for a single value of the tuning parameter. As a solution needs to be computed for a grid of values of the tuning parameter, the computation can be prohibitively time-consuming as n and p become larger. In post-selection inference, the goal is to obtain a few candidate models. Candidate models are determined by which coefficients are nonzero; the exact magnitude is not necessary. For penalized least squares, Hu et al. [2016] modified an ADMM algorithm to quickly obtain a sparsity path, i.e. approximate coefficient estimates at different sparsity levels. Their algorithm requires one update for each value of the tuning parameter. This type of algorithm is very useful for the first stage of post-selection inference where identifying candidate models is more important than estimation. Motivated by their work, we also modify an ADMM algorithm to obtain a sparsity path for penalized quantile regression.

Our proposed algorithm differs from the penalized least squares algorithm substantially. Unlike the squared loss function for least squares, the nonsmooth check loss function drastically increases the complexity of the ADMM algorithm. We alleviate this problem by approximating the check function with a quadratic function, effectively turning the optimization problem into a series of weighted least squares problem with a ridge penalty. We show that our sparsity path returns a good set of candidate models containing the true model and is siginificantly faster than algorithms computing the exact solution.

In Section 5.2, we review the ADMM algorithm and provide intuition for modification. We derive the algorithm in Section 5.3. We demonstrate the effectiveness of our proposed algorithm with Monte Carlo simulations in Section 5.4 and conclude with a discussion in Section 5.5.

5.2 Review of ADMM

First introduced by Glowinski and Marroco [1975], Gabay and Mercier [1975], the ADMM algorithm solves the optimization problem

$$\min_{x,z} f(x) + g(z) \quad \text{s.t. } Ax + Bz = c, \tag{5.1}$$

where $x \in \mathbb{R}^n$, $z \in \mathbb{R}^p$, $A \in \mathbb{R}^{m\times n}$, $B \in \mathbb{R}^{m\times p}$, and $c \in \mathbb{R}^m$. Introducing a strictly positive augmentation parameter γ and Lagrangian multiplier $u \in \mathbb{R}^m$, the augmented Lagrangian function of (5.1) becomes

$$L(x,z,u) \;=\; f(x) + g(z) + u^T(Ax + Bz - c) + \frac{\gamma}{2}||Ax + Bz - c||^2.$$

Boyd et al. [2011] showed the following updates for the ADMM algorithm,

$$\begin{aligned} x^{k+1} &= \arg\min_x L(x, z^k, u^k) \\ z^{k+1} &= \arg\min_z L(x^{k+1}, z, u^k) \\ u^{k+1} &= u^k - \gamma(Ax^{k+1} + Bz^{k+1} - c), \end{aligned}$$

where (x^k, z^k, u^k) denotes the kth iteration of the algorithm for $k \geq 0$. The u variable update is often called the dual update and ensures that the constraint $Ax + Bz = c$ is met.

For penalized quantile regression, it is common to define $r_i = Y_i - X_i^T\beta$ so the ADMM algorithm is

$$L(\beta, r, u) \;=\; \sum_{i=1}^n \rho_\tau(r_i) + \sum_{j=1}^p p_\lambda(\beta_j) + u^T(r + X\beta - Y) + \frac{\gamma}{2}||r + X\beta - Y||^2,$$

resulting in the following updates

$$\begin{aligned} \beta^{k+1} &= \arg\min_\beta \sum_{j=1}^p p_\lambda(\beta_j) + {u^k}^T(r^k + X\beta - Y) + \frac{\gamma}{2}||r^k + X\beta - Y||^2 \qquad (5.2) \\ r^{k+1} &= \arg\min_z \sum_{i=1}^n \rho_\tau(r_i) + {u^k}^T(r + X\beta^{k+1} - Y) + \frac{\gamma}{2}||r + X\beta^{k+1} - Y||^2 \\ u^{k+1} &= u^k - \gamma(r^{k+1} + X\beta^{k+1} - Y). \end{aligned}$$

Recognizing that $\rho_\tau(r_i) = \frac{1}{2}|r_i| + (\tau - \frac{1}{2})r_i$, the r update is easily solved using soft-thresholding. The β update is much more challenging and is similar to solving a lasso penalized least squares problem with "response" $Y_i - r_i^k - u_i^k/\gamma$. Gu et al. [2017] included a proximal term to the objective function in the β update to simplify computation. Splitting the data into smaller subsets can also simplify the updates and allows for computation in parallel [Yu and Lin, 2017, Yu et al., 2017]. These algorithms solve the penalized quantile regression problem for a fixed value of the tuning parameter.

5.3 Sparsity path

Our goal is to develop an algorithm that quickly identifies which variables have effects on the response as the sparsity of the model increases, not necessarily to accurately estimate the effects themselves. Sparse estimates are induced from the ADMM setup in (5.2) because the β update is essentially a lasso penalized least squares problem. We make a simple modification to the ADMM setup to induce sparsity that does not require solving a computationally intensive lasso problem.

Defining z to be a copy of β, we can write the penalized quantile regression problem as

$$\min_{x,z} \sum_{i=1}^{n} \rho_\tau(y_i - x_i^T\beta) + \sum_{j=1}^{p} p_\lambda(z_j) \quad \text{s.t. } \beta - z = 0,$$

with its associated augmented Lagrangian:

$$L(\beta, z, u) \;=\; \sum_{i=1}^{n} \rho_\tau(y_i - x_i^T\beta) + \sum_{j=1}^{p} p_\lambda(z_j) + u^T(\beta - z) + \frac{\gamma}{2}||\beta - z||^2.$$

The updates are

$$\begin{aligned}
\beta^{k+1} &= \arg\min_{\beta} \sum_{i=1}^{n} \rho_\tau(y_i - x_i^T\beta) + u^{k^T}(\beta - z^k) + \frac{\gamma}{2}||\beta - z^k||^2 && (5.3)\\
z^{k+1} &= \arg\min_{z} \sum_{j=1}^{p} p_\lambda(z_j) + u^{k^T}(\beta^{k+1} - z) + \frac{\gamma}{2}||\beta^{k+1} - z||^2 && (5.4)\\
u^{k+1} &= u^k - \gamma(\beta^{k+1} - z^{k+1}).
\end{aligned}$$

In this framework, the z update in (5.4) induces sparsity and is simple to solve for most sparsity inducing penalty functions. The z variable will be sparser for larger values of λ. Candidate models are found by increasing the value of λ after one complete iteration of the algorithm and saving the z update. We continue iterating and increasing λ until the z update yields a completely sparse vector. If the algorithm is initialized with $z^0 = 0$, then the first β update is a ridge penalized quantile regression problem which will yield a dense vector. Therefore, by initializing the algorithm with a dense estimate, we can obtain a set of candidate models ranging from dense to fully sparse.

The augmentation parameter γ is sometimes tuned to decrease the number of iterations until convergence in traditional ADMM algorithms. However, in this setting, the algorithm finishes when z^k is fully sparse. Because the tuning parameter λ controls the sparsity of z, we set $\gamma = 1$ for simplicity.

For example, if the lasso penalty, $p_\lambda(z_j) = \lambda|z_j|$, is used, then the z update simply requires a componentwise soft-threshold operation: $z_j^{k+1} = \text{sign}(\beta_j^{k+1} + u_j^k/\gamma)\max(|\beta_j^{k+1} + u_j^k/\gamma| - \lambda, 0)$. The u update ensures that the nonsparse β update will be shrunk towards the sparse z value.

5.3.1 Updating β

The β update does not have a closed form solution due to to the nonsmoothness of the ρ_τ function. For $k \geq 0$, we approximate the ρ_τ with the same qudratic function that Hunter and Lange [2000] used to majorize an approximation of ρ_τ in their MM algorithm for quantile regression:

$$\rho_\tau^\epsilon(y_i - x_i^T\beta^{k+1} \mid \beta^k) \quad = \quad \frac{1}{4}\left[\frac{(y_i - x_i'\beta)^2}{\epsilon + |y_i - x_i'\beta^k|} + (4\tau - 2)(y_i - x_i'\beta) + c\right],$$

where $\epsilon > 0$ and c is chosen such that $\rho_\tau^\epsilon(y_i - x_i^T\beta^{k+1} \mid \beta^k) = \rho_\tau(y_i - x_i^T\beta) - \frac{\epsilon}{2}\log(\epsilon + |y_i - x_i^T\beta|)$. Defining R_ϵ^k to be a $n \times n$ diagonal matrix with (i, i)th entry $1/(4(\epsilon + |y_i - x_i^T\beta^k|))$, the β update becomes

$$\beta^{k+1} \quad = \quad \arg\min_\beta ||Y - X\beta||_{R_\epsilon^k}^2 - n\left(\tau - \frac{1}{2}\right)\bar{X}^T\beta + u^{k^T}(\beta - z^k) + \frac{\gamma}{2}||\beta - z^k||^2,$$

where for a vector a and matrix B, $||a||_X^2 \equiv a^TXa$. This minimization problem is a weighted least squares problem with a ridge-like penalty whose solution is

$$\beta^{k+1} \quad = \quad \left(X'R_\epsilon^kX + \frac{\gamma}{2}I_p\right)^{-1}\left[X'R_\epsilon^kY + \frac{n}{2}\left(\tau - \frac{1}{2}\right)\bar{X} + \frac{\gamma}{2}\mathbf{z}^k - \frac{1}{2}u^k\right].$$

Computing $\left(X'R_\epsilon^kX + \frac{\gamma}{2}I_p\right)^{-1}$ for each new value of λ is expensive. A recursive algorithm relying on a Searle Identity is useful for quickly approximating the inverse at each iteration. First, define $M_k \equiv \left(X'R_\epsilon^kX + \frac{\gamma}{2}I_p\right)^{-1}$ and $B_k \equiv X^T\left(R_\epsilon^k - R_\epsilon^{k-1}\right)X$. Then the update for the inverse is

$$M_{k+1} \quad \approx \quad M_k - M_kB_kM_k. \tag{5.5}$$

We leave the details of the derivation to Section 5.6.

The inverse can be updated every 5 or 10 steps to decrease computation time.

Empirical evidence suggests that a good set of candidate models can be found by not updating R_ϵ^k at any point. This greatly speeds up computation.

5.4 Monte Carlo studies

In this section, we evaluate the performance of our proposed sparsity path algorithm in a simulation study. This sparsity path algorithm will be useful in practice if it can quickly find a good set of candidate models. All simulations were conducted on an Intel Core i7-4790 processor (single-core, 3.6 GHz).

We consider a simulation setting similar to that in Peng and Wang [2015]. First we generate $(\tilde{X}_1, \ldots, \tilde{X}_p)^T \sim N(0, I_p)$. Then we set $X_1 = \Phi(\tilde{X}_1)$ and $X_k = \tilde{X}_k$ for $k = 2, \ldots, p$ and generate the response from the following heteroscedastic model,

$$Y \;=\; X_6 + X_{12} + X_{15} + X_{20} + 0.7X_1\epsilon,$$

where $\epsilon \sim N(0,1)$. In all simulation settings, $n = 30,000$ and p is set to either 100 or 1000. Three quantiles are considered: $\tau = 0.3$, 0.5, and 0.7. The effect of X_1 is zero when $\tau = .5$.

In each simulation, the sparsity path algorithm is initialized with a lasso penalized quantile regression estimate with a very small value of λ to ensure a dense estimate. After each iteration, the tuning parameter λ is increased by 5% until a fully sparse estimate is reached. The candidate set is chosen to be all models containing between 2 and 15 variables. A final model is chosen from among the candidate set using BIC. In the simulation, we fix $R_\epsilon^k = R_\epsilon^0$. We used the QPADM algorithm [Yu et al., 2017] to obtain the initial dense estimate.

Table 5.1 summarizes the simulation study. We report the mean size of the selected model, size; the percent of times that $X_6, X_{12}, X_{15},$ and X_{20} were all included in the

Table 5.1: Performance of the proposed sparsity path algorithm.

p	τ	Size	P1	P2	AE	M	Time
100	0.3	5.010 (0.010)	100%	100%	0.163 (0.001)	100%	0.538 (0.055)
100	0.5	4.010 (0.010)	100%	1%	0.007 (0.000)	100%	0.587 (0.011)
100	0.7	5.060 (0.028)	100%	100%	0.162 (0.001)	100%	0.493 (0.015)
1000	0.3	5.030 (0.017)	100%	100%	0.163 (0.001)	100%	41.405 (1.425)
1000	0.5	4.000 (0.000)	100%	0%	0.006 (0.000)	100%	41.288 (1.791)
1000	0.7	5.010 (0.010)	100%	100%	0.162 (0.001)	100%	42.434 (1.680)

final model, P1; the percent of times that X_1 was included in the final model, P2; the ℓ_1 estimation error, AE; the percent of times that the correct model was included in the candidate set, M; and the mean time in seconds to approximate the path, Time. Standard errors are in parentheses.

In all simulations, the correct model was always included in the candidate set. Because of this, the final model selection step using BIC was mostly successful in selecting and estimating the true model. It is difficult to compare the computation time of the proposed algorithm as we are not aware of the existence of any other sparsity path algorithms for penalized quantile regression. However, when $p = 100$, the QPADM algorithm required about 10 seconds on average compute the initial estimate and about 115 seconds on average when $p = 1000$. In each simulation, the sparsity path algorithm typically ran for about 250 iterations. Our proposed algorithm was significantly faster than the QPADM algorithm when a fine grid of tuning parameter values are needed and selects a good set of candidate models. The computational burden in the sparsity path algorithm comes from computing the $p \times p$ inverse $\left(X'R_\epsilon^k X + \frac{\gamma}{2} I_p\right)^{-1}$ which requires on the order of p^3 operations. Exact algorithms for penalized quantile regression can be too slow to use in practice when n and p are large.

5.5 Discussion

Motivated by the goal of quickly finding a set of candidate models when the sample size and the number of covariates are large, we proposed a sparsity path algorithm to approximate the model at different sparsity levels. The algorithm differs from traditional ADMM algorithms in that sparsity is imposed on a copy variable and is very easy to compute. The simulation study showed that the sparsity path algorithm is much faster than existing algorithms, which can be prohibitively slow when the dataset is large.

We leave investigation of theoretical properties of the algorithm to future works, but intuit that a good set of candidate models is obtained by squeezing the sparse z update and dense β update together. Ideas used in the sparsity path algorithm may also prove useful in creating an algorithm to quickly estimate the conditional quantile process, that is, estimate the conditional quantile function along a fine grid of $\tau \in (0, 1)$. Algorithms focused on exploring the model space can be useful for very large datasets when traditional algorithms that find exact solutions are too slow.

5.6 Derivation of the matrix inverse approximation

We derive the matrix approximation in (5.5). By the Searle Identity, for any two square matrices A and B,

$$\begin{aligned}(A+B)^{-1} &= A^{-1} - A^{-1}(I + BA^{-1})BA^{-1} \\ &\approx A^{-1} - A^{-1}BA^{-1},\end{aligned}$$

where the approximation follows if B represents a small perturbation such that $I + A^{-1}B \approx I$. Define

$$\begin{aligned} M_k &= \left(X' R_\epsilon^k X + \frac{\gamma}{2} I_p\right)^{-1} \\ B_k &= X' \left(R_\epsilon^k - R_\epsilon^{k-1}\right) X. \end{aligned}$$

We now have a recrusive formula for computing the inverse that is easily updated at each iteration or every few iterations,

$$M_k = M_{k-1} - M_{k-1} B_k M_{k-1}.$$

Chapter 6

Conclusion

In this dissertation, we proposed solutions to statistical challenges in healthcare. Healthcare data is notoriously complex and heterscedastic making analysis difficult. Fortunately, semiparametric quantile regression is an effective tool with mild assumptions that can be used to analyze healthcare data with both linear and nonlinear effects. We applied quantile regression in a classification problem, proposed a consistent estimator for semiparametric quantile regression in a longitudinal study with dropout, and derived an algorithm to quickly find a set of candidate models for large datasets.

The solutions posed all lead to more interesting questions and extensions. The asymptotic covariance matrix for the longitudinal model with dropout (Chapter 4) is very difficult to estimate in practice. A potential solution might be to approximate the distribution of the estimator using the bootstrap. This solution, however, comes with its own set of challenges. First, one of the terms in the covariance matrix is the error of approximating the covariates with linear effects using the covariates with nonlinear effects. We are not aware of any method of estimating this error. Another challenge is handling the correlation among an individual's observations. A nice feature of quantile regression is that the errors need not be homoscedastic to obtain a consistent estimator. By making no distributional assumptions, the difficulty of estimating the correlation increases greatly.

In Chapter 5, we proposed an algorithm to select a model when the dimensions of the data are large and conventional methods do not provide an answer in a reasonable amount of time. We did not investigate the theoretical properties of the selected model. It would be helpful to understand these properties so that we can establish a property similar to that of Theorem 3.3.4. In practice, variable selection is an important part of creating a predictive model. Additionally, the idea of exploiting the update of one variable of the variables in an ADMM algorithm to solve a problem opens up new doors for utilizing the ADMM algorithm in other novel ways.

Healthcare data continues to be collected in new ways with more and more information being gathered each year. Proper analysis can help set effective policies, promote better health outcomes, and understand factors effecting diseases. Methods and algorithms need to modified, adapted, and developed to keep up with the new challenges and questions raised by practioners. This dissertation proposed a few solutions to existing problems and hopefully the methods presented here can be used and adapted to solve future problems as well.

References

Alexandre Belloni, Victor Chernozhukov, et al. ℓ1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics*, 39(1):82–130, 2011.

Richard Berk, Lawrence Brown, Andreas Buja, Kai Zhang, Linda Zhao, et al. Valid post-selection inference. *The Annals of Statistics*, 41(2):802–837, 2013.

Dimitris Bertsimas, Margrét V Bjarnadóttir, Michael A Kane, J Christian Kryder, Rudra Pandey, Santosh Vempala, and Grant Wang. Algorithmic prediction of health-care costs. *Operations Research*, 56(6):1382–1392, 2008.

Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.

Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. *Classification and regression trees.* CRC press, 1984.

Baojiang Chen and Xiao-Hua Zhou. Doubly robust estimates for binary longitudinal data analysis with missing response and missing covariates. *Biometrics*, 67(3): 830–842, 2011.

Jie Chen, Arturo Vargas-Bustamante, Karoline Mortensen, and Stephen B Thomas. Using quantile regression to examine health care expenditures during the great recession. *Health services research*, 49(2):705–730, 2014.

Stephanie A Cosentino, Yaakov Stern, Elisaveta Sokolov, Nikolaos Scarmeas, Jennifer J Manly, Ming Xin Tang, Nicole Schupf, and Richard P Mayeux. Plasma β-amyloid and cognitive decline. *Archives of neurology*, 67(12):1485–1490, 2010.

David R Cox. A note on data-splitting for the evaluation of significance levels. *Biometrika*, 62(2):441–444, 1975.

José Luiz P da Silva, Enrico A Colosimo, and Fábio N Demarqui. Doubly robust-based generalized estimating equations for the analysis of longitudinal ordinal missing data. *arXiv preprint arXiv:1506.04451*, 2015.

Werner Ehm, Tilmann Gneiting, Alexander Jordan, and Fabian Krüger. Of quantiles and expectiles: consistent scoring functions, choquet representations and forecast rankings. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(3):505–562, 2016.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456): 1348–1360, 2001.

Jianqing Fan, Fang Han, and Han Liu. Challenges of big data analysis. *National science review*, 1(2):293–314, 2014.

Xingdong Feng, Xuming He, and Jianhua Hu. Wild bootstrap for quantile regression. *Biometrika*, 98(4):995, 2011.

John A Fleishman and Joel W Cohen. Using information on clinical conditions to predict high-cost patients. *Health services research*, 45(2):532–552, 2010.

Daniel Gabay and Bertrand Mercier. *A dual algorithm for the solution of non linear variational problems via finite element approximation.* Institut de recherche d'informatique et d'automatique, 1975.

Marco Geraci and MC Jones. Improved transformation-based quantile regression. *Canadian Journal of Statistics*, 43(1):118–132, 2015.

Roland Glowinski and A Marroco. Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de dirichlet non linéaires. *Revue française d'automatique, informatique, recherche opérationnelle. Analyse numérique*, 9(R2):41–76, 1975.

Yuwen Gu, Jun Fan, Lingchen Kong, Shiqian Ma, and Hui Zou. Admm for high-dimensional sparse penalized quantile regression. *Technometrics*, (just-accepted), 2017.

Xuming He and Peide Shi. Bivariate tensor-product b-splines in a partly linear model. *Journal of Multivariate Analysis*, 58(2):162–181, 1996.

Xuming He, Zhong-Yi Zhu, and Wing-Kam Fung. Estimation in a semiparametric model for longitudinal data with unspecified dependence structure. *Biometrika*, 89 (3):579–590, 2002.

Joseph W Hogan, Jason Roy, and Christina Korkontzelou. Handling drop-out in longitudinal studies. *Statistics in medicine*, 23(9):1455–1497, 2004.

David W Hosmer Jr and Stanley Lemeshow. *Applied logistic regression*. John Wiley & Sons, 2004.

Yue Hu, Eric C. Chi, and Genevera I. Allen. *ADMM Algorithmic Regularization Paths for Sparse Statistical Machine Learning*, pages 433–459. Springer International Publishing, Cham, 2016. ISBN 978-3-319-41589-5. doi: 10.1007/978-3-319-41589-5_13. URL `https://doi.org/10.1007/978-3-319-41589-5_13`.

David R Hunter and Kenneth Lange. Quantile regression via an mm algorithm. *Journal of Computational and Graphical Statistics*, 9(1):60–77, 2000.

Andrew M Jones, James Lomas, and Nigel Rice. Healthcare cost regressions: going beyond the mean to estimate the full distribution. *Health economics*, 24(9):1192–1212, 2015.

Bo Kai, Runze Li, and Hui Zou. New efficient estimation and variable selection methods for semiparametric varying-coefficient partially linear models. *Annals of statistics*, 39(1):305, 2011.

Ravi Kannan, Santosh Vempala, and Adrian Vetta. On clusterings: Good, bad and spectral. *Journal of the ACM (JACM)*, 51(3):497–515, 2004.

Roger Koenker. *Quantile regression*. Number 38. Cambridge university press, 2005.

Roger Koenker. *quantreg: Quantile Regression.*, 2013. R package version 5.05 (available from `http://CRAN.R-project.org/package=quantreg`).

Roger Koenker and Gilbert Bassett. Regression quantiles. *Econometrica: journal of the Econometric Society*, pages 33–50, 1978.

Roger Koenker and Gilbert Bassett Jr. Robust tests for heteroscedasticity based on regression quantiles. *Econometrica: Journal of the Econometric Society*, pages 43–61, 1982.

Roger Koenker and Vasco d'Orey. Remark as r92: A remark on algorithm as 229: Computing dual regression quantiles and regression rank scores. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 43(2):410–414, 1994.

Roger Koenker and Beum J Park. An interior point algorithm for nonlinear quantile regression. *Journal of Econometrics*, 71(1-2):265–283, 1996.

Roger W Koenker and Vasco d'Orey. Algorithm as 229: Computing regression quantiles. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 36(3): 383–393, 1987.

Eun Ryung Lee, Hohsuk Noh, and Byeong U Park. Model selection via bayesian information criterion for quantile regression models. *Journal of the American Statistical Association*, 109(505):216–229, 2014.

Jason D Lee, Dennis L Sun, Yuekai Sun, Jonathan E Taylor, et al. Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3): 907–927, 2016.

Youjuan Li and Ji Zhu. L 1-norm quantile regression. *Journal of Computational and Graphical Statistics*, 17(1):163–185, 2008.

Stuart R Lipsitz, Garrett M Fitzmaurice, Geert Molenberghs, and Lue Ping Zhao. Quantile regression methods for longitudinal data with drop-outs: application to cd4 cell counts of patients infected with the human immunodeficiency virus. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 46(4):463–476, 1997.

Adam Maidman. *plaqr: Partially Linear Additive Quantile Regression.*, 2016. R package version 1.1 (available from `http://CRAN.R-project.org/package=plaqr`).

Richard T Meenan, Maureen C O'Keeffe-Rosetti, Mark C Hornbrook, Donald J Bachman, Michael J Goodman, Paul A Fishman, and Arnold V Hurtado. The sensitivity and specificity of forecasting high-cost users of medical care. *Medical care*, 37(8): 815–823, 1999.

Nicolai Meinshausen, Lukas Meier, and Peter Bühlmann. P-values for high-dimensional regression. *Journal of the American Statistical Association*, 104(488): 1671–1681, 2009.

Sirisha Nandipati, Xiaodong Luo, Corbett Schimming, Hillel T Grossman, and Mary Sano. Cognition in non-demented diabetic older adults. *Current aging science*, 5 (2):131–135, 2012.

Bo Peng and Lan Wang. An iterative coordinate descent algorithm for high-dimensional nonconvex penalized quantile regression. *Journal of Computational and Graphical Statistics*, 24(3):676–694, 2015.

James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866, 1994.

Mary Sano, Carolyn W Zhu, Hillel Grossman, and Corbett Schimming. Longitudinal cognitive profiles in diabetes: Results from the national alzheimer's coordinating center's uniform data. *Journal of the American Geriatrics Society*, 65(10):2198–2204, 2017.

LL Schumaker. *Spline functions: basic theory*. John Wiley&Sons, New York, NY, USA, 1981.

Ben Sherwood. Variable selection for additive partial linear quantile regression with missing covariates. *Journal of Multivariate Analysis*, 152:206–223, 2016.

Ben Sherwood and Lan Wang. Partially linear additive quantile regression in ultra-high dimension. *The Annals of Statistics*, 44(1):288–317, 2016.

Ben Sherwood, Lan Wang, and Xiao-Hua Zhou. Weighted quantile regression for analyzing health care cost data with missing covariates. *Statistics in medicine*, 32 (28):4967–4979, 2013.

Ben Sherwood, Andrew Xiao-Hua Zhou, Sandra Weintraub, and Lan Wang. Using quantile regression to create baseline norms for neuropsychological tests. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 2:12–18, 2016.

Jonathan Taylor and Robert J Tibshirani. Statistical learning and selective inference. *Proceedings of the National Academy of Sciences*, 112(25):7629–7634, 2015.

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

Ryan J Tibshirani, Jonathan Taylor, Richard Lockhart, and Robert Tibshirani. Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association*, 111(514):600–620, 2016.

U.S. Census Bureau. *U.S. and World Population Clock*, 2016. available from `https://www.census.gov/popclock/`.

Andrew J Vickers and Elena B Elkin. Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making*, 26(6):565–574, 2006.

Huixia Judy Wang, Zhongyi Zhu, and Jianhui Zhou. Quantile regression in partially linear varying coefficient models. *The Annals of Statistics*, pages 3841–3866, 2009.

Lan Wang, Ingrid Van Keilegom, and Adam Maidman. Wild residual bootstrap inference for penalized quantile regression with heteroscedastic errors. *To appear in Biometrika.*

Lan Wang, Yichao Wu, and Runze Li. Quantile regression for analyzing heterogeneity in ultra-high dimension. *Journal of the American Statistical Association*, 107(497): 214–222, 2012.

Larry Wasserman and Kathryn Roeder. High dimensional variable selection. *Annals of statistics*, 37(5A):2178, 2009.

Ying Wei and Yunwen Yang. Quantile regression with covariates missing at random. *Statistica Sinica*, pages 1277–1299, 2014.

Ying Wei, Xuming He, et al. Conditional growth charts. *The Annals of Statistics*, 34 (5):2069–2097, 2006a.

Ying Wei, Anneli Pere, Roger Koenker, and Xuming He. Quantile regression methods for reference growth charts. *Statistics in medicine*, 25(8):1369–1382, 2006b.

Ying Wei, Yanyuan Ma, and Raymond J Carroll. Multiple imputation in quantile regression. *Biometrika*, 99(2):423, 2012.

Sanford Weisberg. *Applied linear regression*, volume 528. John Wiley & Sons, 2005.

Yichao Wu and Yufeng Liu. Variable selection in quantile regression. *Statistica Sinica*, pages 801–817, 2009.

Lan Xue and Lijian Yang. Additive coefficient modeling via polynomial spline. *Statistica Sinica*, 16(4):1423, 2006.

Grace Y Yi and Wenqing He. Median regression models for longitudinal data with dropouts. *Biometrics*, 65(2):618–625, 2009.

Liqun Yu and Nan Lin. Admm for penalized quantile regression in big data. *International Statistical Review*, 85(3):494–518, 2017.

Liqun Yu, Nan Lin, and Lan Wang. A parallel algorithm for large-scale nonconvex penalized quantile regression. *Journal of Computational and Graphical Statistics*, 26(4):935–939, 2017. doi: 10.1080/10618600.2017.1328366. URL `https://doi.org/10.1080/10618600.2017.1328366`.

Cun-Hui Zhang et al. Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, 38(2):894–942, 2010.

Xiao-Hua Zhou, Kevin T Stroupe, and William M Tierney. Regression analysis of health care charges with heteroscedasticity. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 50(3):303–312, 2001.

Hui Zou and Ming Yuan. Composite quantile regression and the oracle model selection theory. *The Annals of Statistics*, pages 1108–1126, 2008.

Chapter 7

Appendix

7.1 Sample Code for using plaqr package

```r
library(plaqr)
set.seed(4)

n = 1000

### Generate the covariates
x1 <- rnorm(n); x2 <- rnorm(n)
z1 <- runif(n); z2 <- runif(n, -1,1)

### Generate the response
y <- exp( x1 + x2 + sin(2*pi*z1) + z2^3 + rnorm(n) )

### Customize the settings for the spline basis functions for z1 and z2
splinesettings <- vector("list", 2)
splinesettings[[2]]$degree <- 4
splinesettings[[2]]$Boundary.knots <- c(-1,1)
```

```r
### Estimate the transformation parameter
trans <- transform_plaqr(y ~ x1 + x2, ~ z1 + z2, tau=.5,
           splinesettings=splinesettings, lambda=seq(0,3,by=.05))
trans$parameter
### Save the transformed response
newy <- trans$Y

### Fit the model
fit <- plaqr(newy ~ x1 + x2, ~ z1 + z2, tau=.5,
           splinesettings=splinesettings)

### Plot the nonlinear effects
plot( nonlinEffect(fit) )

### Make prediction intervals
newdata <- data.frame( x1=c(-1,1), x2=c(0,3),
        z1=c(.2, .6), z2=c(-.5,-.75) )
intervals <- predictInt( fit, newdata=newdata )

### Transform the intervals back to original scale
trans_parameter(intervals, trans$parameter, inverse=TRUE)
```