Essays in Semiparametric and Nonparametric Microeconometrics

by

Matias Damian Cattaneo

Licentiate (University of Buenos Aires, Argentina) 2001
M.A. (University of California at Berkeley) 2005

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Economics

in the

GRADUATE DIVISION
of the
UNIVERSITY OF CALIFORNIA, BERKELEY

Committee in charge:
Professor James L. Powell, Chair
Professor David R. Brillinger
Professor Michael Jansson

Spring 2008

www.manaraa.com

UMI Number: 3334300

INFORMATION TO USERS

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleed-through, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

UMI Microform 3334300
Copyright 2008 by ProQuest LLC
All rights reserved. This microform edition is protected against unauthorized copying under Title 17, United States Code.

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Essays in Semiparametric and Nonparametric Microeconometrics

Copyright 2008

by

Matias Damian Cattaneo

www.manaraa.com

Abstract

Essays in Semiparametric and Nonparametric Microeconometrics

by

Matias Damian Cattaneo

Doctor of Philosophy in Economics

University of California, Berkeley

Professor James L. Powell, Chair

A large fraction of the literature on program evaluation focuses on efficient, flexible estimation of treatment effects under the assumption of unconfoundedness. The first two chapters of this dissertation contribute to this literature by studying the efficient estimation of a large class of multi-valued treatment effects as implicitly defined by a collection of possibly over-identified non-smooth moment conditions when treatment assignment is assumed to be ignorable. Chapter 2 proposes two general estimators, one based on an inverse probability weighting scheme and the other based on the efficient influence function of the model, and provides a set of sufficient conditions that ensure root-n consistency, asymptotic normality and efficiency of these estimators. Chapter 3 shows that, under mild assumptions, these conditions are satisfied for the

www.manaraa.com

marginal mean treatment effect and marginal quantile treatment effect, two estimands of particular importance for empirical applications. Previous results for average and quantile treatments effects may be seen as particular cases of the methods proposed in Chapter 2 when treatment is assumed to be dichotomous. Chapter 3 also illustrates the empirical applicability of the results derived in Chapter 2 by studying the effect of maternal smoking intensity during pregnancy on birth weight. The main empirical findings suggest the presence of approximately homogeneous, non-linear treatment effects concentrated on the first 10 cigarettes-per-day smoked during pregnancy.

Finally, Chapter 4 derives the optimal rates of convergence for the Block Regression Estimator, a nonparametric estimator of the regression function that is implicitly used when estimating the Average Treatment Effect by subclassification on the propensity score. This result contributes to both the literature of program evaluation and the literature of nonparametric estimation.

Professor James L. Powell
Dissertation Committee Chair

www.manaraa.com

To my wife, Rocio.

www.manaraa.com

Contents

www.manaraa.com

المنارة للاستشارات

www.manaraa.com

Acknowledgments

I am specially indebted to Guido Imbens, Michael Jansson and Jim Powell for advice and support. I am also deeply grateful to David Brillinger, Richard Crump, Sebastian Galiani, Enrico Moretti, Paul Ruud and Rocio Titiunik for valuable comments and suggestions that helped improve the content of this dissertation. I also thank Douglas Almond, Ken Chay and David Lee for generously providing the data used in the empirical illustration reported in Chapter 3, and seminar participants at Boston University, Brown University, Duke University, Harvard University, University of Michigan, University of Pennsylvania, University of Texas at Austin, and Washington University in St. Louis for comments. The usual disclaimers apply.

www.manaraa.com

Chapter 1

Introduction

Econometric program evaluation has a crucial role in many different fields of study ranging from the social sciences to public health and biostatistics. The main focus of the program evaluation literature is on the estimation of (causal) treatment effects on some outcome of interest. A large fraction of this literature focuses on the (efficient) estimation of different treatment effects under the assumption of unconfoundedness, this is, assuming that any selection into treatment undertaken by the observational units can be removed by conditioning on a sufficiently rich set of observable characteristics. Even though it is usually a strong assumption in applied work, unconfoundedness has received considerable attention in recent years.

The literature of program evaluation concentrates almost exclusively on the special case of binary treatment assignments, despite the fact that in many empirical applications treatments are implicitly or explicitly multi-valued in nature. For example, in

www.manaraa.com

training programs participants receive different hours of training, in conditional cash transfer programs households receive different levels of transfers, and in educational interventions individuals are assigned to different classroom sizes. In cases such as these, a common empirical practice is to collapse the multi-valued treatment status into a binary indicator for eligibility or participation, a procedure that allows for the application of available semiparametric econometric techniques at the expense of a considerable loss of information. Important phenomena such as non-linearities and differential effects across treatment levels cannot be captured by the classical dichotomous treatment literature. This is especially important in a policy-making context where this additional information may provide a better understanding of the policy under consideration.

The first two chapters of this dissertation are concerned with the efficient estimation of a general class of finite multi-valued treatment effects when treatment assignment is assumed to be ignorable. Chapter 2 studies two estimation procedures for a population parameter implicitly defined by a possibly over-identified non-smooth collection of moment restrictions and provides a set of sufficient conditions that guarantees that these estimators be efficient in large samples. This chapter provides general high-level conditions that ensure that these estimators be asymptotically efficient without explicitly imposing any particular nonparametric estimator for the unknown nuisance infinite dimensional components involved in the estimation.

The general results presented in Chapter 2 are then used to analyze several lead-

www.manaraa.com

ing examples of treatment effects of interest both theoretically and in an empirical illustration. In particular, Chapter 3 discusses the efficient semiparametric estimation of marginal mean treatment effects and marginal quantile treatment effects, which provides the basis for the analysis of a rich set of population parameters by allowing not only for comparisons across and within treatment levels, but also for the construction of other quantities of interest. For example, the researcher may easily construct measures of inequality, differential treatment effects, and heterogeneous treatment effects by considering different functions of means and quantiles such as pairwise differences, interquantile ranges and incremental ratios. This chapter also introduces a new nonparametric estimator appropriate for the model under study and shows how the high-level conditions developed in Chapter 2 can be verified using this nonparametric estimator. Finally, Chapter 3 also illustrates the applicability of the theoretical results developed in Chapter 2 by estimating the effect of maternal smoking intensity on birth weight.

The main results of Chapters 2 and 3 are closely related to the program evaluation literature, the missing data literature and the measurement error literature in both econometrics and statistics.[1] Most of these works may be traced back to the seminal papers of Rubin (1974) and Rosenbaum and Rubin (1983), and often focus on the identification and semiparametric (efficient) estimation of different population parameters of interest. In the context of program evaluation and for the particular case

[1]For recent surveys on these topics, usually with a particular emphasis on binary treatment assignments, see Rosenbaum (2002), Frölich (2004), Imbens (2004), Lee (2005), or Tsiatis (2006).

www.manaraa.com

of binary treatments, great effort has been devoted to the efficient estimation of the average treatment effect (ATE) and average treatment effect on the treated (ATT) using either nonparametric regression methods (Hahn (1998), Heckman, Ichimura, and Todd (1998), Imbens, Newey, and Ridder (2006)), matching techniques (Abadie and Imbens (2006)), or procedures based on the nonparametric estimation of the propensity score (Hirano, Imbens, and Ridder (2003)). Recently, Firpo (2007) considered a different population parameter by studying the efficient estimation of quantile treatment effects for dichotomous treatment assignments using a nonparametrically estimated propensity score. In the closely related context of missing data, Robins, Rotnitzky, and Zhao (1994), Robins and Rotnitzky (1995) and Robins, Rotnitzky, and Zhao (1995) develop a general (locally) efficient estimation strategy for models where the missingness indicator is binary involving the parametric estimation of both a regression function and the propensity score. Finally, two recent contributions by Chen, Hong, and Tamer (2005) and Chen, Hong, and Tarozzi (2007) study efficient GMM estimation in the context of measurement error models under a set of assumptions similar to ignorability with a binary missingness indicator.

Considerably less work is available in the literature for the case of multiple treatment assignments. In the context of program evaluation, Imbens (2000) derives a generalization of the propensity score, termed the Generalized Propensity Score (GPS), and shows that the results of Rosenbaum and Rubin (1983) continue to hold when the treatment status is multi-valued. Concerning identification and estimation, Im-

www.manaraa.com

bens (2000) and Lechner (2001) discuss marginal mean treatment effects but do not assess the asymptotic properties of their estimators, while Abadie (2005) studies the large sample properties of an estimator for the marginal mean treatment effect conditional on a treatment level in the context of a difference-in-differences model. In the framework of missing data and under the assumption of missing at random there are further results in terms of limiting distributions and (local) efficiency properties for estimators of the marginal means; for a survey on these results see the recent paper of Bang and Robins (2005) and the references therein. Finally, in the context of missing data but without the assumption of missing at random, Horowitz and Manski (2000) develop sharp bounds for different multi-valued marginal mean treatment effects.

Finally, Chapter 4 of this dissertation derives optimal rates of convergence for the Block Regression Estimator, a nonparametric estimator of an unknown regression function which generalizes a well known estimator in the statistical literature known as Partitioning (see, Kohler, Krzyzak, and Walk (2006) and references therein). It turns out that this nonparametric estimator plays a key role in the construction of a commonly used estimation procedure for the ATE in program evaluation generally referred to as Subclassification on the Propensity Score. Thus, establishing the large sample properties of the Block Regression Estimator is important not only for the statistical literature on nonparametric estimation, but also for the literature on program evaluation.

Chapter 5 outlines some possible extensions and concludes.

www.manaraa.com

Chapter 2

Efficient Semiparametric GMM Estimation with Missing Data at Random

The main contribution of this chapter to the literature of program evaluation is to develop a unified framework for the efficient estimation of a large class of multi-valued treatment effects. This general framework not only includes as particular cases important results from the program evaluation literature when treatment is binary, but also allows for the efficient estimation of other estimands of interest. First, the Efficient Influence Function (EIF) and Semiparametric Efficiency Bound (SPEB) for the general population parameter of interest using the methodology outlined in Newey (1990) and Bickel, Klaasen, Ritov, and Wellner (1993) is computed. Then,

www.manaraa.com

two estimators of multi-valued treatment effects are introduced and motivated as the solution to a general GMM model with missing data at random. To circumvent the fundamental problem of causal inference, these estimators are constructed by forming sample analogues of two (possibly over-identified) moment conditions that depend only on observed data. For the first estimator, the observed moment condition is obtained by an Inverse Probability Weighting (IPW) scheme based on the GPS which may be interpreted as a moment condition exploiting a portion of the EIF. For the second estimator, the observed moment condition is obtained by using the complete form of the EIF and involves both the GPS and another conditional expectation. Because the observed moment conditions include not only the treatment effects of interest but also some infinite dimensional nuisance parameters, both estimators are of the two-step variety. In the first step, the infinite dimensional nuisance parameters are estimated and, in the second step, the corresponding GMM problem is solved. Notice that the model considered here may provide further efficiency gains in the estimation of treatment effects by allowing for over-identification.

The large sample results are derived in two basic stages. In the first stage, developed in this chapter, consistency, asymptotic normality and efficiency of both estimators is established for any given nonparametric estimator of the infinite dimensional nuisance parameters. These results are obtained by imposing a set of mild sufficient conditions concerning the underlying moment identification functions as well as two well-known high-level conditions involving the nonparametric estimators. This strat-

www.manaraa.com

egy provides a better understanding of the set of sufficient conditions required for the general procedure to work and allows for different choices of the nonparametric estimator of the nuisance parameters. The mild conditions imposed for the underlying moment identification functions are easily verified in applications, as shown in the examples discussed in Chapter 3, while the two-high level conditions generally require additional work. Thus, in the second stage, developed in Chapter 3, the nonparametric estimation of the two nuisance parameters for the particular case of series estimation is discussed in detail.

Once an efficient estimation procedure is available, it is discussed how other important population parameters of interest may be efficiently estimated by means of transformations. Intuitively, because semiparametric efficiency is preserved by a standard delta-method argument, other treatment effects that may be written as functions of the general population parameter of interest are also efficiently estimated. For the case of binary treatments, this implies that the results of Hahn (1998), Hirano, Imbens, and Ridder (2003), and Firpo (2007) may be seen as particular cases of the procedure developed here, as discussed in detail in Chapter 3. Furthermore, this general procedure allows for the efficient estimation of restricted treatment effects by means of a simple minimum distance estimator based on the efficiently estimated, unrestricted treatment effects. In addition to enlarging the class of treatment effects covered by the results presented here, these ideas also allow for "optimal" testing of many hypotheses of interest.

www.manaraa.com

2.1 Preliminaries

2.1.1 The Model

Consider the multi-valued treatment effect model, which is the natural extension of the well-known model used in the classical binary treatment effect literature.[1] Assume there exists a finite collection of multiple treatment status (categorical or ordinal) indexed by $t \in \mathcal{T}$, where without loss of generality $\mathcal{T} = \{0, 1, 2, \cdots, J\}$ with $J \in \mathbb{N}$ fixed. The random variables $\{Y(t),\ t \in \mathcal{T}\}$, with $Y(t) \in \mathcal{Y} \subset \mathbb{R}$, denote the collection of potential outcomes under treatment $t \in \mathcal{T}$, while the random variable $T \in \mathcal{T}$ indicates which of the $J+1$ potential outcomes is observed. Thus, the observed outcome is the random variable $Y = \sum_{t \in \mathcal{T}} D_t Y(t)$, where $D_t = \mathbf{1}\{T = t\}$ for all $t \in \mathcal{T}$ and $\mathbf{1}\{\cdot\}$ is the indicator function. It is also assumed that there exists a real-valued random vector $X \in \mathcal{X} \subset \mathbb{R}^{d_x}$, $d_x \in \mathbb{N}$, which is always observed.

The population parameter of interest is the vector $\beta^* = [\beta_0^{*\prime}, \beta_1^{*\prime}, \cdots, \beta_J^{*\prime}]'$, where $\beta_t^* \in \mathcal{B} \subset \mathbb{R}^{d_\beta}$ for all $t \in \mathcal{T}$ and $d_\beta \in \mathbb{N}$. It is assumed that this parameter solves a collection of $J+1$ (possibly over-identified) moment conditions given by

$$\mathbb{E}\left[m\left(Y(t); \beta_t\right)\right] = 0 \text{ if and only if } \beta_t = \beta_t^*,\ \forall t \in \mathcal{T}, \tag{2.1}$$

where the function $m : \mathcal{Y} \times \mathcal{B} \to \mathbb{R}^{d_m}$ is known (possibly non-smooth) with $d_m \geq d_\beta$. This model corresponds to a slightly specialized case of a general GMM model with multi-level missing data. This may be verified by a simple change in notation: let

[1] For a review of the binary treatment effect literature see Imbens (2004), and for a review of the multi-valued treatment effect literature see Frölich (2004).

www.manaraa.com

$Y(t) \in \mathbb{R}^{d_y}$ with $d_y \geq 1$ and (abusing notation) redefine $Y(t) = (Y(t), X)$ for all $t \in \mathcal{T}$. Although all the results in this chapter apply to this more general model without changes, for simplicity the discussion is restricted to the multi-valued treatment effect model.[2]

The maintained assumption in equation (2.1) imposes a conventional high-level identification condition for GMM estimation as defined by the collection of moment conditions. This model allows for a large class of population parameters of interest including those defined by non-smooth moment functions such as quantiles or other robust estimands.

Finally, it is assumed that a random sample of size n from (Y, T, X) is observed, which is denoted by (Y_i, T_i, X_i), $i = 1, 2, \cdots, n$. This leads to a cross-sectional random sample scheme where only the potential outcome corresponding to $T = t$ is observed, which implies that effectively the sample comes from the conditional distribution of $Y(t)$ given $T = t$ rather than from the marginal distribution of $Y(t)$, a fact that will in general induce a bias in the estimation. Notice that in this model the fundamental problem of causal inference is exacerbated: only one of the $J + 1$ potential outcomes for each unit is observed.

[2]Furthermore, observe that all dimensions and moment conditions have been set equal across treatment levels $t \in \mathcal{T}$. This is done only to simplify notation since all the results presented continue to hold in the more general case where different dimensions and/or moment conditions depending on t are considered.

www.manaraa.com

2.1.2 Identification

The identification condition in equation (2.1) covers many cases of interest. However, it has the obvious drawback of being based on unobservable random variables, the potential outcomes, which makes estimation infeasible. To proceed, an additional identification restriction is needed. Following the program evaluation literature, the "selection on observables" assumption is imposed based on the always observed random vector X:

Assumption 1 *For all* $t \in \mathcal{T}$*:*

(a) $Y(t) \perp\!\!\!\perp D_t \mid X$*; and*

(b) $0 < p_{\min} \leq p_t^*(X) \equiv \mathbb{P}[T = t \mid X]$.

In the context of multi-valued treatment effects, Assumption 1 is sometimes referred to as Ignorability while the conditional probabilities $p_t^*(X)$, $t \in \mathcal{T}$, are known as the Generalized Propensity Score. Imbens (2000) and Lechner (2001) provide a detailed discussion of this assumption and discuss the role of the GPS in the estimation of the particular population parameter, which coincides with the first example of possible estimands of interest presented in Chapter 3.

Part (a) of Assumption 1 has been widely used in the program evaluation, missing data and measurement error literatures. This condition, sometimes called Unconfoundedness or Missing at Random, ensures that the distribution of each potential outcome and the treatment level indicator be conditionally independent and con-

www.manaraa.com

sequently provides identification by imposing "random assignment" conditional on observables. Intuitively, this assumption guarantees that, after conditioning on X, the conditional distribution of $Y(t)$ given $T = t$ and the marginal distribution of $Y(t)$ be identical. This assumption turns out to be sufficient for identification of β^* because it leads to

$$\mathbb{E}\left[\mathbb{E}\left[m\left(Y;\beta_t\right) \mid T = t, X\right]\right] = \mathbb{E}\left[m\left(Y(t);\beta_t\right)\right] = 0 \text{ if and only if } \beta_t = \beta_t^*, \forall t \in \mathcal{T}. \tag{2.2}$$

Part (b) of Assumption 1 is important for at least two reasons. First, it is a necessary condition for finiteness of the semiparametric efficiency bound for regular estimators of β^* as discussed in the next section. Second, together with part (a), it provides the opportunity to consider alternative identification conditions based on the observed random variables. For example, it may easily be verified that

$$\mathbb{E}\left[\frac{D_t m\left(Y;\beta_t\right)}{p_t^*\left(X\right)}\right] = \mathbb{E}\left[m\left(Y(t);\beta_t\right)\right] = 0 \text{ if and only if } \beta_t = \beta_t^*, \forall t \in \mathcal{T}, \tag{2.3}$$

and

$$\mathbb{E}\left[\frac{D_t \mathbb{E}\left[m\left(Y;\beta_t\right) \mid X\right]}{p_t^*\left(X\right)}\right] = \mathbb{E}\left[m\left(Y(t);\beta_t\right)\right] = 0 \text{ if and only if } \beta_t = \beta_t^*, \forall t \in \mathcal{T}, \tag{2.4}$$

which leads to two additional observed moment conditions.[3]

Using equations (2.2), (2.3) and (2.4) as a starting point, several estimation procedures and their corresponding efficiency properties have been considered in the

[3]Other identification conditions are also available in the literature based on this idea. For example, see Hahn (1998).

www.manaraa.com

literature for the particular case of binary treatment effects (or binary missingness indicator). Estimators that exploit moment conditions (2.2) or (2.4) are usually known as "imputation" or "projection" estimators because first a conditional expectation function is (nonparametrically) estimated, and then missing outcomes are imputed for all (or some subset of) the observations and averaged out. Recent examples of papers studying this kind of estimators are Hahn (1998) and Imbens, Newey, and Ridder (2006) in the context of program evaluation with binary treatments, and Chen, Hong, and Tamer (2005) and Chen, Hong, and Tarozzi (2007) in the context of nonclassical measurement error. In the framework of missing data, there is a vast literature known as Doubly Robust Estimation that is based on moment conditions such as equation (2.4) and uses parametric specifications of the unknown functions. Bang and Robins (2005) present a comprehensive review on this topic.

Estimators constructed from the moment condition (2.3) lead naturally to an Inverse Probability Weighting (IPW) scheme and have been considered by many authors in different contexts at least since the work of Horvitz and Thompson (1952). Intuitively, this procedure achieves identification by reweighting the observations to make them representative of the population of interest. This idea has been exploited in the program evaluation literature by Imbens (2000), Hirano, Imbens, and Ridder (2003) and Firpo (2007), in the missing data literature by Robins, Rotnitzky, and Zhao (1994) and Robins, Rotnitzky, and Zhao (1995), and in the measurement error literature by Chen, Hong, and Tarozzi (2007), among others. Wooldridge (2007)

www.manaraa.com

provides a very interesting discussion of this estimation strategy.

Assumption 1 leads to an important collection of alternative asymptotically equivalent efficient estimators in the context of program evaluation. In this chapter, two general efficient estimators for the case of multi-valued treatment effects are studied. The first estimator is based on equation (2.3), while the second estimator is based on a different moment condition that may be constructed as a linear combination of equations (2.2), (2.3) and (2.4). These estimators are also asymptotically equivalent to those available in the literature in the special case of binary treatment effects. It remains an important open research question to rank the large class of available semiparametric efficient estimators.

2.1.3 Notation

Before turning to the discussion of efficient estimation in the context of multi-valued treatment effects, it is convenient to introduce some notation that will simplify the presentation. Two important functions are: the $J+1$ vector-valued function representing the GPS, denoted by $p^*(\cdot) = [p_0^*(\cdot), \cdots, p_J^*(\cdot)]'$, and the $(J+1)\,d_m$ vector-valued function of conditional expectations denoted by $e^*(\cdot;\beta) = [e_0^*(\cdot;\beta_0)', \cdots, e_J^*(\cdot;\beta_J)']'$, where $e_t^*(X;\beta_t) = \mathbb{E}[m(Y(t);\beta_t) \mid X]$. It is assumed that $p_t^*(\cdot) \in \mathcal{P}$ and $e_t^*(\cdot;\beta_t) \in \mathcal{E}$ for all $\beta_t \in \mathcal{B}$ and $t \in \mathcal{T}$, where $\mathcal{P}$ and $\mathcal{E}$ represent some space of (smooth) functions. For simplicity, in the remaining parts of Chapter 2 and 3 the arguments of the functions considered are dropped whenever it is clear from the

www.manaraa.com

context.

Let $|\cdot|$ denote the matrix norm given by $|A| = \sqrt{\text{trace}\,(A'A)}$ for any matrix A. As for functions, the sup-norm in all arguments is denoted by $\|\cdot\|_\infty$. In particular, for all $t \in \mathcal{T}$, it is denoted $\|p_t\|_\infty = \sup_{x\in\mathcal{X}} |p_t(x)|$ for some $p_t \in \mathcal{P}$, $\|e_t(\beta_t)\|_\infty = \sup_{x\in\mathcal{X}} |e_t(x;\beta_t)|$ and $\|e_t\|_\infty = \sup_{\beta\in\mathcal{B},x\in\mathcal{X}} |e_t(x;\beta_t)|$ for some $e_t(\beta_t) \in \mathcal{E}$, and similarly for the vector-valued functions p and e. Later in this chapter the class of functions considered will be restricted to enable the nonparametric estimation of these nuisance parameters.

2.2 Semiparametric Efficiency Calculations

This section provides basic semiparametric efficiency calculations essential for the construction of efficient estimators of β^*. Semiparametric efficiency theory has received considerable attention in econometrics at least since the seminal work of Bickel, Klaasen, Ritov, and Wellner (1993) (see also Newey (1990) for an excellent survey). This general theory provides the necessary ingredients for the construction of efficient estimators of finite dimensional parameters in the context of semiparametric models under some mild regularity conditions. First, it provides the analogue concept of the Cramer-Rao Lower Bound for semiparametric models, that is, an efficiency benchmark for regular estimators of the population parameter of interest. Second, and more importantly, it provides a way of constructing efficient estimators using the efficient influence function or the efficient score of the model. In the simplest possible case,

www.manaraa.com

the construction of an efficient estimator starts by deriving the EIF in the statistical model and then verifying that the proposed estimator admits an asymptotic linear representation based on this function. This chapter uses these ideas to verify that the estimators considered are in fact efficient.

Several semiparametric efficiency calculations are available in the literature when some form of Assumption 1 holds. In the context of program evaluation with binary treatments, efficient influence functions and efficiency bounds have been computed by Hahn (1998), Hirano, Imbens, and Ridder (2003) and Firpo (2007) for average and quantile treatment effect parameters using the methodology outlined in Bickel, Klaasen, Ritov, and Wellner (1993). In models of missing data, Robins, Rotnitzky, and Zhao (1994) and Robins and Rotnitzky (1995) develop a general methodology to construct efficient scores and compute the corresponding efficiency bounds when the missingness indicator is binary. In a recent contribution, Chen, Hong, and Tarozzi (2007) provide semiparametric efficiency calculations for GMM models in the context of nonclassical measurement error with one auxiliary sample. The results presented in this section cover all these cases by considering a multi-level missing mechanism in a GMM model. In Chapter 3, it is discussed how the efficiency bounds derived in the program evaluation literature may be recovered from the calculations presented here.

Assumption 2 *For all* $t \in \mathcal{T}$,

(a) $\mathbb{E}[|m(Y(t);\beta_t)|^2] < \infty$ *and* $\mathbb{E}[m(Y(t);\beta_t)]$ *is differentiable in* $\beta_t \in \mathcal{B}$ *at* β_t^*;

www.manaraa.com

and

(b) $\operatorname{rank}(\Gamma_*) = (J+1)\, d_\beta$, *where*

$$\Gamma_* = \begin{bmatrix} \Gamma_0^* & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \Gamma_1^* & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \Gamma_J^* \end{bmatrix},$$

where $\mathbf{0}$ *is a* $(d_m \times d_\beta)$ *matrix of zeros and*

$$\Gamma_t^* = \frac{\partial}{\partial \beta_t'} \mathbb{E}\left[m\left(Y(t); \beta_t\right)\right]\Big|_{\beta_t = \beta_t^*}.$$

The main role of Assumption 2 (together with part (b) of Assumption 1) is to ensure that the bound is finite. The full column rank assumption on the gradient matrix Γ_* guarantees a local identification condition necessary for the semiparametric calculations. A key necessary requirement to provide semiparametric calculations is to establish the pathwise differentiability of the population parameter of interest, which is done in Appendix A under this assumption (and Assumption 1).

The following theorem provides the general form of the EIF and SPEB for the model considered in this chapter.

Theorem 1 *(*EIF AND SPEB*) Let Assumptions 1 and 2 hold. Then the EIF for any regular estimator of* β^* *is given by*

$$\Psi\left(y, t, x; \beta^*, p^*, e^*\left(\beta^*\right)\right) = -\left(\Gamma_*' V_*^{-1} \Gamma_*\right)^{-1} \Gamma_*' V_*^{-1} \psi\left(y, t, x; \beta^*, p^*, e^*\left(\beta^*\right)\right),$$

www.manaraa.com

where $\psi(y,t,x;\beta^*,p^*,e^*(\beta^*)) = m(y,t,x;\beta^*,p^*) - \alpha(t,x;\beta^*,p^*,e^*(\beta^*))$ *and*

$$V_* = \mathbb{V}\left[\psi(Y,T,X;\beta^*,p^*,e^*(\beta^*))\right].$$

Consequently, the SPEB for any regular estimator of β^* *is given by*

$$V^* = \left(\Gamma_*' V_*^{-1} \Gamma_*\right)^{-1}.$$

The results in Theorem 1 may be directly compared to those presented in Newey (1994). This leads to a natural interpretation for the EIF, where the vector-valued function $\alpha(\cdot)$ corresponds to the "adjustment term" in the influence function due to the presence of the unknown nuisance parameter (GPS) when the estimator is constructed from the sample analogue of the moment condition given by equation (2.3). This interpretation is used below to compare the two estimators considered in this chapter.

It is possible to provide additional intuition for the structure of the SPEB after noting that

$$V_* = \mathbb{E}\left[\mathbb{V}\left[m(Y,T,X;\beta^*,p^*) \mid X\right]\right] + \mathbb{E}\left[e^*(X;\beta^*)\, e^*(X;\beta^*)'\right]. \tag{2.5}$$

Using this decomposition, it is seen how the results in Theorem 1 may be interpreted as the multi-level generalization of the SPEB in Theorem 1 of Chen, Hong, and Tarozzi (2007) in the context of measurement error with "verify-in-sample" auxiliary data. Extending the results of Hahn (1998) and Chen, Hong, and Tarozzi (2007) to the context of multi-valued treatments, it is possible to verify that (i) the GPS is

www.manaraa.com

ancillary for the estimation of β^* (i.e., the SPEB does not change whether or not the GPS it is assumed to be known), and (ii) if the distribution of X is known or correctly specified the SPEB is reduced (in particular, if the distribution of X is assumed to be known, then the second term in equation (2.5) drops out). Details for these results are not provided to conserve space.

It is important to note that these calculations have explicitly allowed for the components $\beta_0^*, \cdots, \beta_J^*$ of the population parameter β^* to be different. Under this assumption, the SPEB obtained in Theorem 1 will be in general larger than the one that would have been obtained had $\beta_0^* = \cdots = \beta_J^*$ been imposed. Since the main goal is to estimate efficiently the components of β^* (i.e., treatment effects), the result presented in Theorem 1 seems to be the most appropriate. The SPEB for the "restricted" case may be easily obtained by similar derivations to those presented in Appendix A.

One important simplification in Theorem 1 is achieved in the important case of exact identification:

Corollary 2 *If $d_m = d_\beta$, then Theorem 1 implies that the EIF for any regular estimator of β^* is given by*

$$\Psi\left(y, t, x; \beta^*, p^*, e^*\left(\beta^*\right)\right) = \Gamma_*^{-1} \psi\left(y, t, x; \beta^*, p^*, e^*\left(\beta^*\right)\right).$$

Consequently, the SPEB for any regular estimator of β^* is given by

$$V^* = \Gamma_*^{-1} V_* \Gamma_*'^{-1}.$$

www.manaraa.com

Notice further that in this case, $\Gamma_* = \text{diag}\left(\Gamma_0^*, \cdots, \Gamma_J^*\right)$. The result in Corollary 2 is important because it shows that in the just-identified case the EIF may be constructed by collecting in a single vector the EIF's corresponding to each $\beta_0^*, \cdots, \beta_J^*$. Moreover, using this result, it follows that in the just-identified case it is possible to estimate efficiently β^* by estimating each $\beta_0^*, \cdots, \beta_J^*$ independently. This result is discussed further below.

2.3 Estimation Procedures

In this chapter two estimators for the multi-valued treatment effects are considered. The first estimation procedure uses an IPW approach and is based on equation (2.3), while the second estimation procedure combines the IPW and imputation approaches and is based on the EIF derived in Theorem 1. For simplicity, in the over-identified case the construction does not use continuously updated GMM but rather uses a consistent estimator of the corresponding weighting matrix.[4] In particular, assume that A_n is a $(J+1)\, d_\beta \times (J+1)\, d_m$ (random) matrix such that $A_n = A + o_p(1)$ for some positive semidefinite matrix $W = A'A$.

[4] A generalization to a continuously updated GMM model is straightforward provided the corresponding additional regularity conditions are imposed.

www.manaraa.com

2.3.1 Inverse Probability Weighting Estimator (IPWE)

It is possible to motivate this procedure by a simple sample analog principle. Recall that the goal is to estimate the parameters implicitly defined by the moment conditions $\mathbb{E}\left[m\left(Y(t);\beta_t^*\right)\right]=0$ for all $t\in\mathcal{T}$. Had the random variables $(Y(0),\cdots,Y(J))$ been observed, a natural estimator would simply solve the sample analog counterpart of the $J+1$ moment conditions leading to a standard GMM estimation procedure. Unfortunately, due to the presence of the missingness mechanism, it is impossible to perform such estimation since only Y is observed. Instead, it is possible to use the result in Equation (2.3) to obtain a moment condition based only on observed random variables. This alternative has the drawback that now the feasible moment conditions involve both the finite dimensional parameter of interest, β^*, and an infinite dimensional nuisance parameter (GPS). This reasoning suggests that if a preliminary estimator for the GPS that converges to the true GPS sufficiently fast may be constructed, then it would still be possible to consistently estimate the finite dimensional parameter of interest.

Using these ideas, a simple semiparametric two-step GMM estimation procedure may be considered where the parameter β^* is estimated after a preliminary nonparametric estimator for the GPS has been constructed. In particular, to save notation, define the moment condition

$$M^{IPW}\left(\beta,p\right)=\mathbb{E}\left[m\left(Y,T,X;\beta,p\right)\right],$$

www.manaraa.com

and its sample analogue

$$M_n^{IPW}(\beta, p) = \frac{1}{n}\sum_{i=1}^{n} m(Y_i, T_i, X_i; \beta, p).$$

Formally the IPWE may be described by the following steps. First, construct a nonparametric estimator of the GPS based on the full sample, which is denoted $\hat{p} = [\hat{p}_0, \cdots, \hat{p}_J]'$. Second, the IPWE for β^* is given by

$$\hat{\beta}^{IPW} = \arg\min_{\beta \in \mathcal{B}^{J+1}} \left| A_n M_n^{IPW}(\beta, \hat{p}) \right| + o_p(n^{-1/2}).$$

This estimation procedure has the important advantage of being based only on the nonparametric estimator of the GPS. Note that the infinite dimensional component does not depend on β and therefore it needs to be estimated only once to form the GMM problem, leading to a very simple two-step procedure. On the other hand, this estimation procedure has an important drawback based on its construction. Because it only involves the first part of the EIF derived previously, to ensure its semiparametric efficiency the nonparametric estimator $\hat{p}$ will have to play two roles simultaneously: not only does it have to approximate p^* fast enough, but it also has to do it in such a way that the limiting GMM problem becomes a GMM problem based on the EIF. For example, as pointed out by Hirano, Imbens, and Ridder (2003) in the model of binary treatment effects, the extreme case where $\hat{p} = p^*$ will not lead in general to an efficient estimator because this procedure will be solving the incorrect GMM problem. The necessary requirements on $\hat{p}$ will be made explicit below when studying the large sample properties of this estimator.[5]

[5]The role of the propensity score and how information about it may be efficiently incorporated in

www.manaraa.com

For the just-identified case, the procedure leading to the IPWE is equivalent to solving

$$\hat{\beta}_t^{IPW} = \arg\min_{\beta \in \mathcal{B}} \left| \frac{1}{n} \sum_{i=1}^{n} \frac{D_{t,i} m\left(Y_i; \beta_t\right)}{\hat{p}_t\left(X_i\right)} \right| + o_p(n^{-1/2}),\ \forall t \in \mathcal{T},$$

which leads to a very simple estimator.

2.3.2 Efficient Influence Function Estimator (EIFE)

This estimator is based on the EIF derived in Theorem 1. This procedure can also be motivated by the analogue principle after observing that $\mathbb{E}\left[\psi\left(Y, T, X; \beta, p, e\left(\beta\right)\right)\right] = 0$ if and only if $\beta = \beta^*$, $p = p^*$ and $e = e^*$. In words, the EIF provides another collection of moment conditions that can be exploited to obtain a GMM estimator. Inspection of $\mathbb{E}\left[\psi\left(Y, T, X; \beta, p, e\left(\beta\right)\right)\right]$ shows that its sample analogue corresponds to a linear combination of three sample analogues already discussed in the literature for the special case of binary treatment effects. In particular, this moment condition includes (i) the moment condition leading to an IPW estimator, (ii) the moment condition leading to a nonparametric version of the doubly robust estimator, and (iii) the moment condition leading to an imputation estimator.

To describe the estimator, define the moment condition

$$M^{EIF}\left(\beta, p, e\left(\beta\right)\right) = \mathbb{E}\left[\psi\left(Y, T, X; \beta, p, e\left(\beta\right)\right)\right],$$

semiparametric models have received considerable attention in the literature of program evaluation and related areas of study. See, e.g., Hahn (1998), Heckman, Ichimura, and Todd (1998), Hirano, Imbens, and Ridder (2003), and Chen, Hong, and Tarozzi (2007), among others, for a discussion on this topic.

www.manaraa.com

and its sample analogue

$$M_n^{EIF}\left(\beta, p, e\left(\beta\right)\right) = \frac{1}{n}\sum_{i=1}^{n}\psi\left(Y_i, T_i, X_i; \beta, p, e\left(\beta\right)\right).$$

Formally the EIFE may be described by the following steps. First, construct a nonparametric estimator of the GPS, denoted $\hat{p} = [\hat{p}_0, \cdots, \hat{p}_J]'$, and for each $\beta \in \mathcal{B}$ construct a nonparametric estimator of $e\left(\beta\right)$, denoted $\hat{e}\left(\beta\right) = [\hat{e}_0\left(\beta\right)', \cdots, \hat{e}_J\left(\beta\right)']'$. Second, the EIFE for β^* is given by

$$\hat{\beta}^{EIF} = \arg\min_{\beta\in\mathcal{B}^{J+1}}\left|A_n M_n^{EIF}\left(\beta, \hat{p}, \hat{e}\left(\beta\right)\right)\right| + o_p(n^{-1/2}).$$

This estimator appears to be in general more complicated than the IPWE because it requires the nonparametric estimation of two infinite dimensional parameters, one of which is a function of β itself. On the other hand, it has the attractive feature of being based on the EIF and therefore each nonparametric estimator would only be required to have the intuitive role of approximating well its own population counterpart. For example, it is now possible to consider the extreme case of $\hat{p} = p^*$ and still obtain an efficient estimator, as discussed below.[6]

As in the case of the IPWE, in the just-identified case this procedure is equivalent to solve for all $t \in \mathcal{T}$,

$$\hat{\beta}_t^{EIF} = \arg\min_{\beta_t\in\mathcal{B}}\left|\frac{1}{n}\sum_{i=1}^{n}\frac{D_{t,i}m\left(Y_i;\beta_t\right) - \hat{e}_t\left(X_i;\beta_t\right)\left(D_{t,i} - \hat{p}_t\left(X_i\right)\right)}{\hat{p}_t\left(X_i\right)}\right| + o_p(n^{-1/2}).$$

[6]It is important to note that this is not the only way in which information about the (generalized) propensity score may be incorporated in semiparametric efficient estimators. For two other examples, see the recent work of Chen, Hong, and Tarozzi (2007) in the context of measurement error models.

www.manaraa.com

2.4 Large Sample Properties

This section presents the main large sample results of the chapter in four stages. First, consistency of both the IPWE and EIFE is established under mild regularity conditions. Second, sufficient conditions are provided that guarantee asymptotic normality and efficiency of the IPWE and EIFE for any nonparametric estimators of the infinite dimensional nuisance parameters based on a set of high-level conditions. Third, estimators for the different components of the SPEB derived in Theorem 1 are constructed. Finally, it is discussed how other population parameters of interest may be efficiently estimated and/or optimal inference may be performed based on these general results.

The large sample theory presented here is based on the work of Pakes and Pollard (1989).[7] In the following discussion, terminology and results from the modern theory of empirical processes will be repeatedly employed. For consistency and to simplify the exposition, all references to this literature are based on van der Vaart and Wellner (1996) (see also Andrews (1994) and van der Vaart (1998) for excellent reviews on this topic).

[7] Alternatively, it is possible to apply the general large sample theory of Chen, Linton, and van Keilegom (2003). However, because in the case under study the criterion function is smooth in the infinite dimensional nuisance parameters, the results from Pakes and Pollard (1989) turn out to be sufficient.

www.manaraa.com

2.4.1 Consistency

Consistency of the IPW estimator will follow from two mild conditions imposed on the underlying identifying function $m(\cdot;\beta)$:

Assumption 3 *For all* $t \in \mathcal{T}$,

(a) the class of functions $\{m(\cdot;\beta_t) : \beta_t \in \mathcal{B}\}$ *is Glivenko-Cantelli, and*

(b) $\mathbb{E}[\sup_{\beta_t \in \mathcal{B}} |m(Y(t);\beta_t)|] < \infty$.

Part (a) of Assumption 3 restricts the class of functions that may be considered to implicitly define the population parameter of interest. Functions in this class enjoy an important property: sample averages of these functions are uniform consistent in β for their population mean. Although consistency may be established by other means, requiring an uniform consistency property of the underlying sample moment conditions is standard in the GMM literature. Newey and McFadden (1994) discuss this and other related conditions. A simple set of sufficient conditions for Assumption 3(a) are $\mathcal{B}$ compact, $m(\cdot;\beta_t)$ continuous in β_t, and Assumption 3(b). Although this set of conditions is reasonably weak, it is still stronger than necessary. In fact, to cover interesting nonsmooth cases (such as quantiles) it is necessary to rely on slightly stronger results such as those presented in the empirical process literature. From this literature, many classes of functions are known to be Glivenko-Cantelli and many other classes may be formed by some "permanence" theorem.[8] Part (b) of Assumption

[8]Primitive conditions that ensure a given class of functions to be Glivenko-Cantelli (or Donsker) usually involve some explicit assumption concerning the "size" of the class as measured by some

www.manaraa.com

3 is a usual dominance condition.

Theorem 3 *(*Consistency of IPWE*) Let Assumptions 1 and 3 hold. Assume the following additional condition holds:*

(3.1) $\|\hat{p} - p^*\|_\infty = o_p(1)$.

Then, $\hat{\beta}^{IPW} = \beta^* + o_p(1)$.

The additional condition (3.1) in Theorem 3 is very weak, requiring only that the nonparametric estimator of the GPS is uniformly consistent.

Next, consider the EIFE. For this estimator, assume additionally:

Assumption 4 *For all* $t \in \mathcal{T}$*, the class of functions* $\{e_t^*(\cdot;\beta_t) : \beta_t \in \mathcal{B}\}$ *is Glivenko-Cantelli.*

Assumption 4 captures the ideas implied by Assumption 3(a). In this case, however, this assumption may be easier to verify because the functions $e_t^*(\cdot;\beta_t)$ are conditional expectations and therefore it is natural to assume they are smooth in β_t. Thus, verifying the underlying uniform consistency requirement should be straightforward in this case, possibly after imposing some additional mild regularity conditions.

Theorem 4 *(*Consistency of EIFE*) Let Assumptions 1, 3, and 4 hold. Assume the following additional condition holds:*

version of the entropy numbers. For a recent example in the context of GMM estimation see Ai and Chen (2003).

www.manaraa.com

(4.1) $\|\hat{p} - p^*\|_\infty = o_p(1)$ *and* $\|\hat{e} - e^*\|_\infty = o_p(1)$.

Then, $\hat{\beta}^{EIF} = \beta^* + o_p(1)$.

Since now the full EIF is used to construct the estimator, it is natural to observe that Theorem 4 also requires the nonparametric estimator $\hat{e}$ to be uniformly consistent for e^* in both arguments (the covariates X and the parameter β). This condition is still weak and reasonable for most nonparametric estimators.

2.4.2 Asymptotic Normality and Efficiency

It is now possible to discuss the conditions needed to establish the limiting distribution and efficiency of the two estimators considered in this chapter. First, a set of sufficient conditions for the IPWE is given:

Assumption 5 *For all* $t \in \mathcal{T}$ *and some* $\delta > 0$*:*

(a) $\{m(\cdot; \beta_t) : |\beta_t - \beta_t^*| < \delta\}$ *is a Donsker class;*

(b) $\mathbb{E}[\,|m(Y(t); \beta_t) - m(Y(t); \beta_t^*)|^2] \to 0$ *as* $\beta_t \to \beta_t^*$*;*

(c) there exists a constant $C > 0$ *such that*

$$\mathbb{E}\left[|m(Y(t); \beta_t) - m(Y(t); \beta_t^*)|\right] \leq C\,|\beta_t - \beta_t^*|$$

for all β_t *with* $|\beta_t - \beta_t^*| < \delta$*; and*

(d) $\mathbb{E}[\sup_{|\beta_t - \beta_t^*| < \delta} |m(Y(t); \beta_t)|^2] < \infty$.

www.manaraa.com

Similar to the requirement for consistency, Part (a) of Assumption 5 restricts the class of functions defining the population parameter of interest that may be considered. This assumption is standard from the empirical process literature and ensures that an uniform (in β_t) central limit theorem hold. In turn, this result together with part (b) and part (c) will ensure that a certain stochastic equicontinuity condition apply, which allows to obtain an asymptotic linear representation for the estimator. For most applications, Assumption 5(a) is already established or can be easily verified by some "permanence theorem". Assumptions 5(b) and 5(c) are standard in the literature and may be verified directly, while Assumption 5(d) is a usual dominance condition.

Theorem 5 *(*Asymptotic Linear Representation, IPWE*) Let* $\beta^* \in \text{int}(\mathcal{B}^{J+1})$ $\hat{\beta}^{IPW} = \beta^* + o_p(1)$*, and Assumptions 1, 2, and 5 hold. Assume the following additional conditions hold:*

(5.1) $\|\hat{p} - p^*\|_\infty = o_p\left(n^{-1/4}\right)$.

(5.2) $M_n^{IPW}(\beta^*, \hat{p}) = M_n^{EIF}(\beta^*, p^*, e^*(\beta^*)) + o_p\left(n^{-1/2}\right)$.

Then,

$$\hat{\beta}^{IPW} - \beta^* = -\left(\Gamma_*' W \Gamma_*\right)^{-1} \Gamma_*' W M_n^{EIF}(\beta^*, p^*, e^*(\beta^*)) + o_p(n^{-1/2}).$$

Asymptotic normality of $\hat{\beta}^{IPW}$ follows directly from Theorem 5 while efficiency is easily obtained by an appropriate choice of the limiting weighting matrix W. This

www.manaraa.com

theorem requires two important additional conditions involving the estimator of the GPS. These conditions imply certain restrictions in terms of smoothness for the class of functions $\mathcal{P}$ and $\mathcal{E}$, depending on the nonparametric estimator chosen and the dimension of $\mathcal{X}$.

Condition (5.1) is standard in the literature and imposes a lower bound in the uniform rate of convergence of $\hat{p}$. Condition (5.2) is crucial. This condition involves the sample moment condition (at $\beta = \beta^*$) and the nonparametric estimator, and requires a particular linear expansion based on the EIF to hold. Newey (1994) provides an in-depth general discussion of this particular condition and outlines high-level assumptions that ensure this condition holds. This assumption is very important because it employs the exact form of the EIF to guarantee that the resulting estimator is efficient (provided the weighting matrix is chosen appropriately). If condition (5.2) holds for a function different than $M_n^{EIF}(\beta^*, p^*, e^*(\beta^*))$, then the estimator cannot be efficient. For example, if the GPS is known and $\hat{p} = p^*$ is replaced in $M_n^{IPW}(\beta^*, \hat{p})$ when constructing the estimation procedure, then the resulting estimator will not be efficient as mentioned before. In this sense, Condition (5.2) imposes an upper bound on the uniform rate of convergence of $\hat{p}$. Intuitively, this is due to the fact that $\hat{p}$ plays two roles simultaneously: it estimates nonparametrically p^*, and it also nonparametrically approximates the correction term $\alpha(\cdot; p, e(\beta))$ present in the EIF. Consequently, even if the GPS is known, one may obtain an efficient estimator only if the GPS is nonparametrically estimated.

www.manaraa.com

One way to avoid requiring $\hat{p}$ to play this dual role is to consider the full EIF, which leads to the EIFE. This estimator will be asymptotically normal if the following additional assumption holds:

Assumption 6 *For all $t \in \mathcal{T}$, some $\delta > 0$, and for all $x \in \mathcal{X}$ and all β_t such that $|\beta_t - \beta_t^*| < \delta$:*

(a) $e_t^*(x;\beta_t)$ is continuously differentiable with derivative given by $\partial_{\beta_t} e_t^*(x;\beta_t) \equiv \frac{\partial}{\partial \beta_t} e_t^*(x;\beta_t)$ with $\mathbb{E}[\sup_{|\beta_t - \beta_t^*|<\delta} \left|\partial_{\beta_t} e_t^*(X;\beta_t)\right|] < \infty$; and

(b) there exists $\epsilon > 0$ and a measurable function $b(x)$, with $\mathbb{E}[|b(X)|] < \infty$, such that

$$\left|\partial_{\beta_t} e_t(x;\beta_t) - \partial_{\beta_t} e_t^*(x;\beta_t)\right| \leq b(x) \left\|e_t - e_t^*\right\|_\infty^\epsilon$$

for all functions $e_t(\beta_t) \in \mathcal{E}$ such that $\|e_t - e_t^*\|_\infty < \delta$.

Assumption 6 basically restricts the class of functions $\mathcal{G} = \{e_t : e_t(\beta) \in \mathcal{E}, \|e_t - e_t^*\|_\infty < \delta \text{ and } |\beta_t - \beta_t^*| < \delta\}$, where $e_t^* \in \mathcal{G}$ by construction. Part (a) of this assumption is simple and natural, requiring only mild smoothness conditions of the conditional expectation $e_t(\beta_t)$ in β_t as well as a usual dominance condition. Note that this part of the assumption will imply the smoothness requirement in Assumption 2 whenever integration and differentiation can be interchanged. Part (b) of Assumption 6 further restricts the possible class of functions by requiring that functions that are uniformly close also have their derivatives close. This special technical requirement has also been used by Chen, Hong, and Tamer (2005) and Chen, Hong, and

www.manaraa.com

Tarozzi (2007) in the context of nonclassical measurement error. Assumption 6(b) is imposed because uniform convergence is not enough to ensure uniform convergence of derivatives, a result needed in the proof of the following theorem.

Theorem 6 *(*Asymptotic Linear Representation, EIFE*) Let* $\beta^* \in \text{int}(\mathcal{B}^{J+1})$, $\hat{\beta}^{EIF} = \beta^* + o_p(1)$ *and Assumptions 1, 2, 5 and 6 hold. Assume the following additional conditions hold:*

(6.1) $\|\hat{p} - p^*\|_\infty = o_p\left(n^{-1/4}\right)$.

(6.2) $\sup_{|\beta-\beta^*|<\delta} \|\hat{e}(\beta) - e^*(\beta)\|_\infty = o_p(1)$, *for some* $\delta > 0$.

(6.3) $M_n^{EIF}(\beta^*, \hat{p}, \hat{e}(\beta^*)) = M_n^{EIF}(\beta^*, p^*, e^*(\beta^*)) + o_p(n^{-1/2})$.

Then,

$$\hat{\beta}^{EIF} - \beta^* = -\left(\Gamma_*' W \Gamma_*\right)^{-1} \Gamma_*' W M_n^{EIF}(\beta^*, p^*, e^*(\beta^*)) + o_p\left(n^{-1/2}\right).$$

Asymptotic normality of $\hat{\beta}^{EIF}$ also follows directly from Theorem 6. This time, three additional conditions involving the nonparametric estimators are imposed. Condition (6.1) is the same as Condition (5.1) in Theorem 5. Condition (6.2) further requires uniform consistency of the nonparametric estimator of e^* in both arguments, although in this case no particular rate is required. This result follows from the additional smoothness assumptions imposed in this theorem. Finally, Condition (6.3) is the analogue of Condition (5.2) in Theorem 5, although much easier to verify in general. In this case, additional knowledge about the GPS may be easily incorporated

www.manaraa.com

in the estimation without affecting the asymptotic variance, provided the asymptotic linear representation continues to hold.

Efficiency of the estimators follows directly from Theorems 5 and 6:

Corollary 7 *If $d_m = d_\beta$ (just-identified case) or $W = V_*^{-1}$ (as given in Theorem 1), then the IPWE and EIFE are efficient for β^*.*

This corollary distinguishes two cases. First, if the problem is exactly identified, then the estimators are efficient without further work. Second, if the problem is over-identified, then a consistent estimator of the matrix V_*^{-1} is needed, generating an intermediate step in the construction of the GMM problems for the IPWE and the EIFE. A consistent estimator for V_*^{-1} is easy to construct without further assumptions, as shown next.

2.4.3 Optimal Weighting Matrix and Uncertainty Estimation

The next step in the construction of a feasible estimation procedure is to consider the estimation of V_* and Γ_*, the variance of the EIF and the "sandwich" matrix appearing in the SPEB, respectively. For the over-identified case, estimation of V_* is crucial since the square-root of this matrix is the optimal weighting matrix of both GMM problems.

The natural plug-in estimator of V_* is given by

$$V_n = \frac{1}{n}\sum_{i=1}^{n} \psi(Y_i, T_i, X_i, \hat{\beta}, \hat{p}, \hat{e}(\hat{\beta}))\psi(Y_i, T_i, X_i, \hat{\beta}, \hat{p}, \hat{e}(\hat{\beta}))',$$

www.manaraa.com

for some consistent estimator $\hat{\beta}$ of β^*.

Theorem 8 gives a set of simple sufficient conditions that ensure $\hat{V}_n$ is consistent for V_*.

Theorem 8 (Consistent Estimator of V^*) *Let Assumptions 1, 2, 5, and 6(a) with* $\mathbb{E}[\sup_{|\beta_t-\beta_t^*|<\delta} \left|\partial_{\beta_t} e_t^*(X;\beta_t)\right|^2] < \infty$ *hold. If* $\hat{\beta} = \beta^* + o_p(1)$, $\|\hat{p} - p^*\|_\infty = o_p(1)$ *and* $\sup_{|\beta-\beta^*|<\delta} \|\hat{e}(\beta) - e^*(\beta)\|_\infty = o_p(1)$, *for some* $\delta > 0$, *then* $V_n = V_* + o_p(1)$.

Observe that the conditions imposed in Theorem 8 are the same as those assumed in Theorem 5 plus the mild smoothness and dominance condition on e^*.

Next, consider the estimation of Γ_*. Because this matrix has a very particular structure there are several simple alternative approaches to construct a consistent estimator. For example, it is possible to consider a numerical derivative approach directly applied to the sample analogue (e.g., Pakes and Pollard (1989)) or, in some cases, the estimator may be constructed by taking into consideration the explicit form of the matrix (for an example, see the second estimand presented in Chapter 3). As a third alternative, it is also possible to construct a generic estimator, under the assumptions already imposed, if integration and differentiation can be interchanged. In this case, it is seen that for all $t \in \mathcal{T}$,

$$\Gamma_t^* = \frac{\partial}{\partial \beta_t} \mathbb{E}\left[m\left(Y(t);\beta_t\right)\right]\Big|_{\beta_t=\beta_t^*} = \mathbb{E}\left[\frac{\partial}{\partial \beta_t} e_t\left(X;\beta_t\right)\Big|_{\beta_t=\beta_t^*}\right],$$

which suggests the plug-in estimator given by

$$\hat{\Gamma}_{t,n} = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial \beta_t} \hat{e}_t\left(X;\beta_t\right)\Big|_{\beta_t=\hat{\beta}_t}.$$

www.manaraa.com

Consistency of this plug-in estimator is verified in the following theorem.

Theorem 9 (Consistent Estimator of Γ_*) *Let Assumptions 1, 2, 6 hold. If* $\hat{\beta} = \beta^* + o_p(1)$ *and* $\sup_{|\beta-\beta^*|<\delta} \|\hat{e}(\beta) - e^*(\beta)\|_\infty = o_p(1)$*, for some* $\delta > 0$*, then* $\hat{\Gamma}_{t,n} = \Gamma_t^* + o_p(1)$.

From Theorem 9 it is straightforward to form a consistent estimator of the gradient matrix Γ_*.

2.4.4 Other Population Parameters and Optimal Inference

The results presented in this chapter so far allow for the joint efficient estimation of several multi-valued treatment effects. For instance, using the discussed procedures it is easy to estimate jointly (and efficiently) several marginal quantiles as well as the marginal mean of all potential outcomes, as discussed in the next chapter. However, in many applications the population parameters of interest may be not only the marginal treatment effects but also other quantities involving possibly more than one marginal treatment effect. Fortunately, because differentiable transformations of efficient estimators of Euclidean parameters lead to efficient estimators for the corresponding population parameters, a simple delta-method argument is sufficient to easily recover any collection of treatment effects that may be written as (or approximated by) a differentiable function of the marginal treatment effects.

Using this idea it is possible to efficiently estimate many other treatment effects such as pairwise comparisons (in the spirit of ATE), differences between pairwise com-

www.manaraa.com

parisons, incremental ratios, interquantile ranges, quantile ratios or other measures of differential and heterogeneous treatments effects. Moreover, is possible to also consider the efficient estimation of the effect of different treatments on dispersion as measured by the standard deviation of the potential outcome distribution. These ideas are exploited further in the next chapter when discussing the leading examples and the empirical illustration.

Furthermore, because in some applications incorporating additional information about the treatment effects in a general over-identified model may be challenging, it is possible to consider an alternative approach to the efficient estimation of multiple restricted treatment effects. In particular, suppose that the restrictions of interest can be imposed by writing the marginal treatment effects as a function of the parameters π^*, and denote this function by $\beta(\pi^*)$. Then, it can be verified that under mild regularity conditions an efficient estimator of π^* is given by

$$\hat{\pi} = \arg\min_{\pi} [\hat{\beta} - \beta(\pi)]' \left(\Gamma_n' V_n^{-1} \Gamma_n\right) [\hat{\beta} - \beta(\pi)],$$

where $\hat{\beta}$ is an efficient estimator of β^*, Γ_n is a consistent estimator of Γ_*, and V_n is a consistent estimator of V_*. In this case, it is not hard to verify that

$$\sqrt{n}(\hat{\pi} - \pi^*) \xrightarrow{d} \mathcal{N}\left[0, \left(\partial\beta(\pi^*)' \Gamma_*' V_*^{-1} \Gamma_* \partial\beta(\pi^*)\right)^{-1}\right],$$

where $\partial\beta(\pi^*) = \frac{\partial}{\partial\pi}\beta(\pi)\big|_{\pi=\pi^*}$. From this result, a consistent estimator of the covariance matrix of $\hat{\pi}$ may be constructed using a plug-in approach.

Finally, because testing procedures based on efficient estimators are optimal (pos-

www.manaraa.com

sibly after restricting the class of allowed tests), it is straightforward to perform optimal testing of different hypotheses concerning multi-valued treatment effects. This can be done within and across treatment levels for marginal treatment effects, for treatment effects obtained by means of some (differentiable) transformation of these parameters, and for restricted treatment effects by relying on standard testing strategies.

www.manaraa.com

Chapter 3

Efficient Semiparametric Estimation of Multi-valued Treatment Effects

Chapter 2 developed a general theory for the efficient estimation of multi-valued treatment effects. This chapter specializes these general results to two leading examples: Marginal Mean Treatment Effects (MMTE) and Marginal Quantile Treatment Effects (MQTE). It is shown how the general conditions developed in Chapter 2 may be applied directly to these examples, and how available results in the literature of program evaluation may be seen as particular cases of the findings reported in Chapter 2 when considering these examples.

Further, since the estimation procedures introduced in Chapter 2 involve infinite

www.manaraa.com

dimensional nuisance parameters, a nonparametric estimation procedure for these nuisance parameters appropriate for the problem under study is also discussed. These additional results effectively outline a full data-driven procedure for the efficient estimation of β^*. In particular, in this case results from the nonparametric series (or sieve) estimation literature may be applied directly. However, since the GPS is a conditional probability a new nonparametric estimator, labeled Multinomial Logistic Series Estimator, is proposed which is based on series estimation and captures the specific features of this nuisance parameter. This estimator generalizes the nonparametric estimator for the propensity score introduced by Hirano, Imbens, and Ridder (2003) and may be interpreted as a nonlinear sieve procedure (Chen (2007)) having the key advantage of providing predicted positive probabilities that add up to one. Using these nonparametric estimators, simple primitive conditions that guarantee the efficient estimation of general multi-valued treatment effects are provided.

Finally, to illustrate the results the last portion of this chapter reports a brief empirical study of the effect of maternal smoking intensity on birth weight that extends the analysis of Almond, Chay, and Lee (2005). These authors study the costs of low birth weight using different non-experimental techniques and find an important negative effect of maternal smoking on birth weight defining maternal smoking as a binary treatment. Exploiting the fact that their rich database includes the number of cigarettes-per-day smoked by the mother, the analysis is extended to a multi-valued treatment setup which studies the effect of maternal smoking *intensity*

www.manaraa.com

on birth weight. Somehow surprising, the main findings suggest the presence of a nonlinear negative effect where two thirds of the full impact of smoking on birth weight are due to the first 5 cigarettes, while the remaining third is explained by the next 5 cigarettes with no important effects beyond the tenth cigarette-per-day smoked. Moreover, these effects appear to be additive, shifting parallelly the entire distribution of birth weight along smoking intensity.

3.1 Leading Examples

3.1.1 Marginal Mean Treatment Effects

The first leading example captures the idea of a canonical population parameter of interest in the literature of Biostatistics, Public Health and Medicine, among other fields. This population parameter, sometimes called the Dose-Response Function, reflects the mean response for each treatment level and, in the context of program evaluation, may be seen as an extension of the ATE. The MMTE is denoted by $\mu^* = [\mu_0^*, \mu_1^*, \cdots, \mu_J^*]'$ and solves equation (2.1) with $m\left(Y(t), X; \mu_t\right) = Y(t) - \mu_t$, for all $t \in \mathcal{T}$, which leads to $\mu_t^* = \mathbb{E}\left[Y(t)\right]$.

Observe that in this case identification follows immediately after assuming a finite first moment of the potential outcomes. Next, assume $\mathbb{E}[Y(t)^2] < \infty$ and note that $\Gamma_t^* = 1$ for all $t \in \mathcal{T}$ in this case. Thus, Assumption 2 is satisfied and Theorem 1

www.manaraa.com

implies that the SPEB for the MMTE is given by V^* with typical (i,j)-th element

$$V^*_{[i,j]} = \mathbb{E}\left[\mathbf{1}\left\{i=j\right\}\frac{\sigma_i^2\left(X\right)}{p_i^*\left(X\right)} + \left(\mu_i\left(X\right)-\mu_i^*\right)\left(\mu_j\left(X\right)-\mu_j^*\right)\right],$$

where $\sigma_i^2\left(X\right) = \mathbb{V}\left[Y\left(i\right) \mid X\right]$, $\mu_i\left(X\right) = \mathbb{E}\left[Y\left(i\right) \mid X\right]$, for all $i \in \mathcal{T}$.

In terms of the estimation procedures, it is possible to obtain a closed-form solution for this estimand for both IPWE and EIFE. In particular, IPWE is given by

$$\hat{\mu}_t^{IPW} = \left(\sum_{i=1}^{n}\frac{D_{t,i}}{\hat{p}_t\left(X_i\right)}\right)^{-1}\sum_{i=1}^{n}\frac{D_{t,i}Y_i}{\hat{p}_t\left(X_i\right)},$$

which corresponds to a properly re-weighted average for each $t \in \mathcal{T}$, while the EIFE is given by

$$\hat{\mu}_t^{EIF} = \frac{1}{n}\sum_{i=1}^{n}\frac{D_{t,i}Y_i - \hat{\mu}_t\left(X_i\right)\left(D_{t,i}-\hat{p}_t\left(X_i\right)\right)}{\hat{p}_t\left(X_i\right)},$$

where $\hat{\mu}_t\left(x\right)$ represents some nonparametric estimator of $\mu_t^*\left(x\right)$.

Next, to establish the large sample results first assume that $\mathcal{B}$ is compact and $\mathbb{E}\left[|Y(t)|\right] < \infty$ for all $t \in \mathcal{T}$. Assumption 3 follows directly because the class of functions $\{(\cdot - \mu_t) : \mu_t \in \mathcal{B}\}$ is Glivenko-Cantelli. Therefore, Theorem 3 implies $\hat{\mu}^{IPW} = \mu^* + o_p\left(1\right)$, while the class of functions $\{(\mu_t^*\left(\cdot\right) - \mu_t) : \mu_t \in \mathcal{B}\}$ is also Glivenko-Cantelli and Theorem 4 implies $\hat{\mu}^{EIF} = \mu^* + o_p\left(1\right)$.

Now, the class of functions $\{(\cdot - \mu_t) : |\mu_t - \mu_t^*| < \delta\}$ is Donsker and in this case $\mathbb{E}[|m\left(Y\left(t\right);\mu_t\right) - m\left(Y\left(t\right);\mu_t^*\right)|] = |\mu_t - \mu_t^*|$, giving Assumption 5. Thus, under the conditions of Theorem 5 and Corollary 7 it follows that

$$\sqrt{n}(\hat{\mu}^{IPW} - \mu^*) \xrightarrow{d} \mathcal{N}\left(0, V^*\right),$$

www.manaraa.com

and the estimator $\hat{\mu}^{IPW}$ is efficient. Further, in this case Assumption 6 is trivially satisfied and therefore under the conditions of Theorem 6 and Corollary 7 it is obtained

$$\sqrt{n}(\hat{\mu}^{EIF} - \mu^*) \xrightarrow{d} \mathcal{N}(0, V^*),$$

and the estimator $\hat{\mu}^{EIF}$ is also efficient.

Finally, note that if $\mathcal{T} = \{0, 1\}$ and because the ATE can be written as $\Delta^{ATE} \equiv \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] = v'\mu^*$, where $v = (-1, 1)'$, using Theorem 1 it is easy to verify that

$$V^* = \mathbb{E}\begin{bmatrix} \frac{\sigma_0^2(X)}{p_0(X)} + (\mu_0(X) - \mu_0^*)^2 & (\mu_0(X) - \mu_0^*)(\mu_1(X) - \mu_1^*) \\ (\mu_0(X) - \mu_0^*)(\mu_1(X) - \mu_1^*) & \frac{\sigma_1^2(X)}{p_1(X)} + (\mu_1(X) - \mu_1^*)^2 \end{bmatrix}.$$

Then, either Theorem 5 or Theorem 6 and the transformation $g(z) = v'z$ gives

$$\sqrt{n}\left(\hat{\Delta}^{ATE} - \Delta^{ATE}\right) \xrightarrow{d} \mathcal{N}[0, v'V^*v],$$

where

$$v'V^*v = \mathbb{E}\left[\frac{\sigma_0^2(X)}{p(0, X)} + \frac{\sigma_1^2(X)}{p(1, X)} + \left(\Delta^{ATE}(X) - \Delta^{ATE}\right)^2\right],$$

and $\Delta^{ATE}(X) = \mu_1(X) - \mu_0(X)$. In this case, the asymptotic variance is the SPEB found by Hahn (1998) and the resulting estimator in the case of Theorem 5 is essentially the same as the one considered in Hirano, Imbens, and Ridder (2003) (see also Imbens, Newey, and Ridder (2006) for another similar modification of this estimator).

www.manaraa.com

3.1.2 Marginal Quantile Treatment Effects

Characterizing distributional impacts of a multi-valued treatment is crucial because these effects are closely related to usual inequality and heterogeneity measures. The second leading example captures this idea by looking at the treatment effect at different quantiles of the outcome variable. For some $\tau \in (0,1)$, the MQTE is denoted by $q^*(\tau) = [q_0^*(\tau), q_1^*(\tau), \cdots, q_J^*(\tau)]'$ and it is assumed to solve equation (2.1) with $m(Y(t); q_t(\tau)) = \mathbf{1}\{Y(t) \leq q_t(\tau)\} - \tau$, for all $t \in \mathcal{T}$, which leads to $q_t^*(\tau) \in \inf\{q : F_{Y(t)}(q) \geq \tau\}$, where $F_{Y(t)}$ is the c.d.f. of $Y(t)$. In this case, a simple sufficient condition for identification is that $Y(t)$ be a continuous random variable with density $f_{Y(t)}(q_t^*(\tau)) > 0$.

Using Leibniz's rule $\Gamma_t^* = f_{Y(t)}^*(q_t^*(\tau))$ for $t \in \mathcal{T}$, which was assumed strictly positive. Thus, Assumption 2 is satisfied and Theorem 1 implies that the SPEB for the MQTE is given by V^* with typical (i,j)-th element

$$V_{[i,j]}^* = \mathbb{E}\left[\mathbf{1}\{i=j\}\frac{\sigma_i^2(X;\tau)}{f_{Y(i)}^*(q_i^*(\tau))^2 p_i^*(X)} + \frac{q_i(X;\tau)\,q_j(X;\tau)}{f_{Y(i)}^*(q_i^*(\tau))\,f_{Y(j)}(q_j^*(\tau))}\right],$$

where

$$\begin{aligned}\sigma_i^2(X;\tau) &= \mathbb{V}[\mathbf{1}\{Y(i) \leq q_i^*(\tau)\} \mid X], \\ q_i(X;\tau) &= \mathbb{E}[\mathbf{1}\{Y(i) \leq q_i^*(\tau)\} - \tau \mid X],\end{aligned}$$

for all $i \in \mathcal{T}$.

In terms of the estimation procedures, in this case it is not possible to obtain a closed-form solution to the minimization problem. Thus, the estimator solves for

www.manaraa.com

fixed $\tau \in (0,1)$,

$$\hat{q}_t^{IPW}(\tau) = \arg\min_{q \in \mathcal{B}} \left| \frac{1}{n} \sum_{i=1}^{n} \frac{D_{t,i} \left(\mathbf{1} \{ Y_i \leq q \} - \tau \right)}{\hat{p}_t(X_i)} \right|$$

for all $t \in \mathcal{T}$ in the case of IPWE, while the EIFE is given by, for fixed $\tau \in (0,1)$,

$$\hat{q}_t^{EIF}(\tau) = \arg\min_{q_t \in \mathcal{B}} \left| \frac{1}{n} \sum_{i=1}^{n} \frac{D_{t,i} \left(\mathbf{1} \{ Y_i \leq q_t \} - \tau \right) - \left(\hat{F}_{Y(t)}(q_t | X_i) - \tau \right) \left(D_{t,i} - \hat{p}_t(X_i) \right)}{\hat{p}_t(X_i)} \right|$$

for $t \in \mathcal{T}$, where $e_t^*(X; \beta_t) = F_{Y(t)}^*(q_t(\tau) \mid X) - \tau$ and $\hat{F}_{Y(t)}(y \mid x)$ represents some nonparametric estimator of $F_{Y(t)}^*(y \mid x)$.

Next, to establish the large sample results first note that Assumption 3 follows immediately because the class of functions $\{(\mathbf{1} \{ \cdot \leq q_t \} - \tau) : q_t \in \mathcal{B}\}$ is Glivenko-Cantelli and Theorem 3 gives $\hat{q}^{IPW}(\tau) = q^*(\tau) + o_p(1)$, while if the class of functions $\{F_{Y(t)}^*(q_t \mid \cdot) - \tau : q_t \in \mathcal{B}\}$ is Glivenko-Cantelli, Theorem 4 gives $\hat{q}^{EIF}(\tau) = q^*(\tau) + o_p(1)$. The last requirement may be verified if, for example, $\mathcal{B}$ is compact and $F_{Y(t)}^*(y \mid x)$ is continuous in y for every x.

Now, observe that the class of functions $\{(\mathbf{1} \{ y \leq q_t(\tau) \} - \tau) : |q_t(\tau) - q_t^*(\tau)| < \delta\}$ is Donsker and

$$\begin{aligned} &\mathbb{E}\left[\left| m\left(Y(t); q_t(\tau)\right) - m\left(Y(t); q_t^*(\tau)\right) \right| \right] \\ &= \int \left| \mathbf{1} \{ y \leq q_t(\tau) \} - \mathbf{1} \{ y \leq q_t^*(\tau) \} \right| dF_{Y(t)}(y) \leq C \left| q_t(\tau) - q_t^*(\tau) \right|, \end{aligned}$$

for all $q_t(\tau)$ such that $|q_t(\tau) - q_t^*(\tau)| < \delta$, for some $\delta > 0$, under regularity conditions. It follows from this calculation that Assumption 5 is satisfied in this case and under the conditions of Theorem 5 and Corollary 7 it follows that

$$\sqrt{n}(\hat{q}^{IPW}(\tau) - q^*(\tau)) \xrightarrow{d} \mathcal{N}(0, V^*),$$

www.manaraa.com

and $\hat{q}_t^{IPW}(\tau)$ is efficient. Turning to Assumption 6, part (a) may be easily verified under mild regularity conditions because $e_t^*(X;\beta_t) = F_{Y(t)}^*(q_t(\tau) \mid X) - \tau$, while part (b) requires further restrictions on the class of distribution functions allowed for in this case. Thus, under regularity conditions, it is verified that

$$\sqrt{n}(\hat{q}^{EIF}(\tau) - q^*(\tau)) \xrightarrow{d} \mathcal{N}(0, V^*),$$

with $\hat{q}^{EIF}(\tau)$ efficient.

Finally, and similarly as in the case of ATE, if $\mathcal{T} = \{0,1\}$ and because the QTE may also be written as $\Delta^{QTE} \equiv q_1^*(\tau) - q_0^*(\tau) = v'q^*(\tau)$, where $v = (-1,1)'$, either Theorem 5 or Theorem 6 gives

$$\sqrt{n}\left(\hat{\Delta}^{QTE} - \Delta^{QTE}\right) \xrightarrow{d} \mathcal{N}[0, v'V^*v].$$

In this case, the asymptotic variance coincides with the SPEB derived in Firpo (2007) and the resulting estimator in the case of Theorem 5 corresponds to the Z-estimator version of Firpo's estimator for the QTE.

3.1.3 Other Treatment Effects

Once efficient estimators of marginal treatment effects are available, it is straightforward to derive efficient estimators for other treatment effects of interest whenever these can be written as a (differentiable) function of the marginal treatment effects. This general idea was already discussed in the previous chapter. Using the examples presented above, it is easy to recover efficient estimators of other treatment effects of

www.manaraa.com

interest such as differential treatment effects or quantile ratios, just to mention two possibilities.

Furthermore, the general GMM model considered in Chapter 2 allows for further efficiency gains whenever the treatment effects of interest are over-identified. To fix ideas, suppose the distribution of $Y(t)$ is assumed to be symmetric for location. In this case, mean and median coincide and hence there are (at least) two moment conditions for the same parameter of interest. Thus, the population parameter of interest solves the following (over-identified) moment condition: $m\left(Y(t), X; \vartheta_t\right) = \left(Y(t) - \vartheta_t, \mathbf{1}\left\{Y(t) \leq \vartheta_t\right\} - 1/2\right)$, for all $t \in \mathcal{T}$. Since this moment condition collects the moment conditions of the previous examples, the results for this case will follow from the conditions and results discussed before. On the other hand, it is possible to use this example to illustrate the idea of solving a minimum distance problem based on unrestricted efficient estimators as discussed in Chapter 2. In particular, assume that $\left[\mu^*, q^*\left(.5\right)\right]'$ is efficiently estimated by (say) $\left[\hat{\mu}, \hat{q}\left(.5\right)\right]'$ (for example, using either the IPWE or the EIFE). Then, solving

$$\hat{\pi} = \arg\min_{\pi} \begin{bmatrix} \hat{\mu} - \pi \\ \\ \hat{q}\left(.5\right) - \pi \end{bmatrix}' \left(\Gamma_n' V_n^{-1} \Gamma_n\right) \begin{bmatrix} \hat{\mu} - \pi \\ \\ \hat{q}\left(.5\right) - \pi \end{bmatrix},$$

gives an efficient estimator of the multi-valued treatment effect for location under symmetry. This is, $\hat{\pi}_t$ is an efficient estimator of ϑ_t, for all $t \in \mathcal{T}$.

Using this idea it is possible to incorporate additional restrictions on other estimands of interest, such as different quantiles of the underlying distribution of the

www.manaraa.com

potential outcomes.

3.2 Nonparametric Estimation of Nuisance Parameters

Chapter 2 established consistency, asymptotic normality and efficiency for two estimators of multi-valued treatment effects. These results have been obtained by imposing high-level assumptions concerning the behavior of the nonparametric estimators used for the estimation of the infinite dimensional nuisance parameters rather than by specifying a particular form of such estimators. This section discusses explicitly the nonparametric estimation of p^* and e^* and verifies the additional high-level conditions imposed in Theorems 5 and 6.

Since both p^* and e^* are (possibly high-dimensional) conditional expectations, a nonparametric series estimator seems an appropriate choice. These estimators are attractive because they are computational convenient and can incorporate dimension reduction restrictions easily. This nonparametric estimation procedure has been studied in detail by Newey (1997) and may be interpreted as a linear sieve estimator as discussed in Chen (2007). To briefly describe the estimator, let $g(X) = \mathbb{E}[Z \mid X]$ for some random variable Z and random vector $X \in \mathcal{X}$, and let $\{r_k(x)\}_{k=1}^{\infty}$ be a sequence of known approximating functions with the property that a linear combination of $R_K(x) = (r_1(x), \cdots, r_K(x))'$ can approximate $g(x)$ for $K = 1, 2, \cdots$. An approx-

www.manaraa.com

imating function is formed by $g(X;\gamma_K) = R_K(X)'\gamma_K$ and the series estimator based on an i.i.d. random sample (Z_i, X_i), $i = 1, 2, \cdots, n$, is given by $\hat{g}(X) = g(X;\hat{\gamma}_K)$, with

$$\hat{\gamma}_K = \arg\min_{\gamma_K} \sum_{i=1}^{n} (Z_i - g(X_i;\gamma_K))^2,$$

where, in this case, the closed-form solution is given by

$$\hat{\gamma}_K = \left(\sum_{i=1}^{n} R_K(X_i) R_K(X_i)' \right)^{-} \sum_{i=1}^{n} R_K(X_i) Z_i \tag{3.1}$$

with B^- denoting the generalized inverse of the matrix B.

By choosing the approximating basis appropriately and under suitable conditions on the function $g(\cdot)$ and growth rate of K it is possible to establish the consistency and rate of convergence (in both L_2 and uniform sense) of this nonparametric estimator. Two common choices for an approximating basis are power series and splines, leading to polynomial regression and spline regression, respectively. See Newey (1997) for further details.

This nonparametric estimator may be used directly to estimate the vector valued function e^*. For all $t \in \mathcal{T}$, let $Z(\beta_t) = m(Y;\beta_t)'$ and let $\hat{\gamma}_{t,K}(\beta_t)$ be defined as in equation (3.1) but when only the data for $T = t$ is used. Then, for all $t \in \mathcal{T}$, the series nonparametric estimator of $e_t^*(X;\beta_t)$, $\beta_t \in \mathcal{B}$, is given by $\hat{e}_t(X;\beta_t)' = R_K(X)'\hat{\gamma}_{t,K}(\beta_t)$ where

$$\hat{\gamma}_{t,K}(\beta_t) = \left(\sum_{i=1}^{n} D_{t,i} R_K(X_i) R_K(X_i)' \right)^{-} \sum_{i=1}^{n} D_{t,i} R_K(X_i) m(Y_i;\beta_t)'.$$

Similarly, it is possible to construct a series estimator for p^*. However, the GPS is

www.manaraa.com

not only a conditional expectation but also a conditional probability (i.e., all elements are positive and add up to one), which imposes additional restrictions that cannot be captured by this standard nonparametric estimator. Thus, a nonparametric estimator consistent with these additional requirements is preferred. In particular, this section introduces a generalization of the estimator introduced by Hirano, Imbens, and Ridder (2003) for the special context of binary treatments, labeled Multinomial Logistic Series Estimator (MLSE), which may be interpreted as a non-linear sieve (Chen (2007)) estimation procedure.

Intuitively, since $J+1$ conditional probabilities are nonparametrically estimated it is reasonable to embed them within a multinomial logistic model. Using the notation introduced for series estimation, for all $t \in \mathcal{T}$, let $g\left(X;\gamma_{t,K}\right) = R_K\left(X\right)'\gamma_{t,K}$ be the approximating function and for notational simplicity let $\gamma_K = (\gamma'_{0,K}, \gamma'_{1,K}, \cdots, \gamma'_{J,K})'$. When the coefficients $\gamma_{t,K}$, $t \in \mathcal{T}$, are chosen as in equation (3.1) with $Z = D_t$ the usual series estimator for the components of p^* is obtained. Alternatively, the MLSE chooses simultaneously all the vectors in γ_K by solving the maximum likelihood multinomial logistic problem

$$\hat{\gamma}_K = \arg \max_{\gamma_K | \gamma'_{0,K} = 0_K} \sum_{i=1}^{n} \sum_{t=0}^{J} D_{t,i} \log \left(\frac{\exp\left\{g\left(X_i;\gamma_{t,K}\right)\right\}}{\sum_{j=0}^{J} \exp\left\{g\left(X_i;\gamma_{t,K}\right)\right\}} \right),$$

where 0_K represents a $K \times 1$ vector of zeros used to impose the usual normalization $\gamma_{K,0} = 0_K$ needed to achieve identification in this model. In this case, the nonpara-

www.manaraa.com

metric estimator $\hat{p}(\cdot)$ has typical t-th element given by

$$\hat{p}_t(X) = \frac{\exp\left\{R_K(X)'\hat{\gamma}_{t,K}\right\}}{1+\sum_{j=1}^{J}\exp\left\{R_K(X)'\hat{\gamma}_{t,K}\right\}}.$$

It is straightforward to verify that this nonparametric estimator satisfies the additional restrictions underlying the GPS. The rates of convergence of this non-linear sieve estimator are established in Appendix B.

For simplicity and to reduce the notational burden, this section restricts attention to power series and splines as possible approximation basis and assumes that the same basis is used for all the nonparametric estimators. The following simple assumption is enough to establish the appropriate large sample results for both the linear series estimator and the MLSE.

Assumption 7 *For all $t \in \mathcal{T}$,*

(a) $p_t^*(\cdot)$ and $e_t^*(\cdot, \beta_i^*)$ are s times differentiable with $s/d_x > 2\eta + 2$, where $\eta = 1$ or $\eta = 1/2$ depending on whether power series or splines are used as basis functions, respectively;

(b) X is continuously distributed with density bounded and bounded away from zero on its compact support $\mathcal{X}$; and

(c) for all $t \in \mathcal{T}$ and some $\delta > 0$, $\mathbb{V}[m(Y(t);\beta_t) \mid X = x]$ is uniformly bounded for all $x \in \mathcal{X}$ and all β_t such that $|\beta_t - \beta_t^*| < \delta$.

Part (a) of Assumption 7 provides the exact restrictions needed on the spaces $\mathcal{P}$ and $\mathcal{E}$, describing the minimum smoothness required as a function of the dimension of

www.manaraa.com

X and the choice of basis of approximation. Part (b) of Assumption 7 restricts X to be continuous on a compact support with "well-behaved" density. These assumptions may be relaxed considerably at the expense of some additional notation. For example, it is possible to allow for some components of X to be discretely distributed and to permit $\mathcal{X}$ to be unbounded by restricting the tail-behavior of the density of X (see Chen, Hong, and Tamer (2005) for an example). Part (c) of Assumption 7 is standard from the series (or sieve) nonparametric estimation literature.

Theorem 10 *(Nonparametric Estimation) Let Assumptions 1(b) and 7 hold. Then, conditions (5.1) and (5.2) in Theorem 5, and conditions (6.1), (6.2) and (6.3) in Theorem 6 are satisfied by the nonparametric estimators introduced in this section if* $K = n^{\nu}$ *with*

$$\frac{1}{4s/d_x - 4\eta - 2} < \nu < \frac{1}{4\eta + 2}$$

where $\eta = 1$ *or* $\eta = 1/2$ *depending on whether power series or splines are used as basis functions, respectively.*

3.3 Empirical Illustration

To show how the procedures work in practice, this section reports a brief empirical exercise that studies the effect of maternal smoking during pregnancy on birth weight. In a recent paper, Almond, Chay, and Lee (2005) (ACL hereafter) present detailed empirical evidence on the economic costs of low birth weight (LBW). In their paper,

www.manaraa.com

the authors estimate the direct economic costs imposed by LBW on society and also study the possible causes of LBW using different nonexperimental techniques. In particular, ACL present empirical evidence on the effect of maternal smoking on birth weight for a rich database of singletons in Pennsylvania and find a strong effect of about 200-250 gram reduction in birth weight using both subclassification on the propensity score and regression adjusted methods.

The application presented here extends the results of ACL by considering the effect of maternal smoking *intensity* during pregnancy on birth weight. The database used by ACL not only includes almost half a million singleton births and many pre-intervention covariates, but also records the mother's declared number of cigarettes-per-day smoked during pregnancy. This additional information allows to consider multi-valued treatment effects and address several interesting questions, particularly relevant from a policy-making perspective. For example, it is assessed whether the effect of smoking is constant across levels of smoking, whether there exist differential and/or heterogeneous treatment effects, and whether the variability in birth weight is affected by smoking intensity.

The empirical illustration uses the same database, response variable and pre-intervention variables as ACL. In this sample, approximately 80% of mothers did not smoke during pregnancy, while for the remaining 20% inspection of the empirical distribution of smoked cigarettes reveals important mass points approximately every 5 cigarettes ranging from 1 to 25. This feature suggests considering 5-cigarette bins

www.manaraa.com

as a starting point for the empirical analysis. The number of smoked cigarettes were collapsed into 6 categories ($J = 5$) {0, 1-5, 6-10, 11-15, 16-20, 21+} and joint estimation of five quantiles (.9,.75,.5,.25,.1), the mean and standard deviation for each potential outcome, leading to 42 treatment effects, was considered. For $t \in \mathcal{T}$, the identifying moment function in this case is given by the vector-valued function $m(y; \beta_t) = ((\mathbf{1}\{y \leq \beta_{1t}\} - 0.95), (\mathbf{1}\{y \leq \beta_{2t}\} - 0.75), (\mathbf{1}\{y \leq \beta_{3t}\} - 0.5), (y - \beta_{4t}), (\mathbf{1}\{y \leq \beta_{5t}\} - 0.25), (\mathbf{1}\{y \leq \beta_{6t}\} - 0.1), (y^2 - \beta_{7t}))'$ for $\beta_t = (\beta_{1t}, \beta_{2t}, \beta_{3t}, \beta_{4t}, \beta_{5t}, \beta_{6t})'$. For the implementation, first β^* was jointly estimated using both the IPWE and EIFE and then the marginal population parameters of interest were recovered by means of the delta method.

To ensure comparability the same pre-intervention covariates as in ACL were used. In particular, these variables include 43 dummy variables (mother's demographics, father's demographics, prenatal care, alcohol use, pregnancy history, month of birth and county of residency) and 6 "continuous" covariates (mother's age and education, father's age and education, number of prenatal visits, months since last birth and order of birth).[1] For the estimation of both nonparametric nuisance parameters, cubic B-splines with knots ranging from 1 to 3 depending on the continuous covariate were used, and to reduce the computational burden an additive separability assumption on the approximating functions was imposed. Other choices of smoothing parameters for the splines as well as different interactions between the dummies and the smoothed

[1] A full description of the variables used is given in footnote 36 of ACL. The analysis does not include maternal medical risk factors in the analysis; see also footnote 39 of ACL.

www.manaraa.com

covariates were also considered. In all the cases, the results appeared to be robust to the particular specification of the nonparametric estimators.[2]

Because in this case the model is exactly identified, it is possible to estimate each treatment effect separately and then form the full EIF to estimate the SPEB. Table 1 presents the point and uncertainty estimates for the 42 treatment effects using three estimators: a simple dummy regression estimator (DRE), the IPWE and the EIFE. In this sample, estimates from the (inefficient, possibly inconsistent) DRE appear to be very similar to those obtained from the (consistent and efficient) IPWE and EIFE. This result is consistent with the findings in ACL. The standard errors of the estimators IPWE and EIFE appear to be very similar to each other and considerably lower than those of the DRE in the case of the mean, while for the quantiles the standard errors are slightly higher.[3]

A simple way to present the information in Table 1 is by means of Figure 1, which gives important qualitative information about the treatment effects. This figure shows the point estimates and their 95% (marginal) confidence intervals for the case of the MMTE and MQTE when estimated using the IPWE. Interestingly, a parallel shift in

[2]This is consistent with the available literature on semiparametric estimation suggesting that the choice of basis or smoothing parameters are relatively unimportant (see for example Newey (1994), Ai and Chen (2003), Chen, Hong, and Tamer (2005), or Chen, Hong, and Tarozzi (2007)). Based on these results, and for computational simplicity, data-driven procedures (such as cross-validation) were not considered for the selection of the smoothing parameters.

[3]In the quantile dummy regression case the standard errors were calculated using a kernel density estimator with bandwidth set by Silverman's rule-of-thumb. In the case of IPWE and EIFE, the gradient matrix Γ_* was estimated using its exact form (implemented by a weighted kernel density estimator with bandwidth set by Silverman's rule-of-thumb as in Firpo (2007)). The general numerical derivative approach (implemented by a simple numerical difference) was also considered, which led to very similar estimates.

www.manaraa.com

the entire distribution of birth weight along smoking intensity is observed. In particular, there is a large reduction of about 150 grams when the mother starts to smoke (1-5 cigarettes), an additional reduction of approximately 70 grams when changing from 1-5 to 6-10 cigarettes-per-day, and no additional effects once the mother smokes at least 11 cigarettes. These findings provide qualitative evidence that differential treatment effects are non-linear and approximately homogeneous along the distribution of the potential outcomes. In particular, a close to symmetric distribution with approximately constant dispersion (as measured by both interquartile ranges and standard deviation) is observed.

The qualitative results summarized in Figure 1 may be formally tested. Since the 42 marginal treatment effects are jointly estimated, it is straightforward to test the hypotheses suggested by Figure 1 as well as other hypotheses of interest. Table 2 presents a collection of hypothesis tests regarding pairwise differences and difference-in-differences of marginal mean treatment effects. On the diagonal, pairwise differences across treatment levels are reported. For example, the reduction in birth weight induced by increasing maternal smoking from 0 to 1-5 cigarettes is 146 grams (statistically significant), while the corresponding reduction induced by increasing maternal smoking from 6-10 to 11-15 cigarettes is 37 grams (not statistically significant). This table also reports the difference-in-differences comparisons which may be used to test for non-linearities. For example, increasing maternal smoking from 0 to 1-5 cigarettes induces an additional 75 gram reduction in birth weight when compared to

www.manaraa.com

the corresponding reduction induced by increasing maternal smoking from 1-5 to 6-10 cigarettes. This differential effect is statistically significant and provides formal evidence of non-linear treatment effects. Importantly, non-linearities disappear beyond the tenth cigarette smoked during pregnancy. Similar results are obtained when analyzing the MQTE.

Table 3 illustrates additional multiple-hypotheses tests of interest. In the first row, it is reported the joint test for the hypothesis of no treatment effect (as measured by mean, quantile and spread) for the highest three treatment levels, while in the second and third rows analogous tests considering the highest four and highest five treatment levels, respectively, are considered. As shown in this table, increasing smoking intensity beyond 10 cigarettes per day has no further effect on birth weight. The remaining rows in Table 3 test for different hypotheses involving possible distributional effects across and within treatment levels. Small but statistically significant differences on the interquantile ranges are found.

Finally, based on the main finding that most of the effect of smoking on birth weight appears to be concentrated on the first 10 cigarettes-per-day smoked, it is of interest to replicate the analysis for the subpopulation of mothers who smoked between 0 and 10 cigarettes-per-day breaking up the treatment variable into 2-cigarette bins.[4] To conserve space, only qualitative results in Figure 2 are presented. According to this figure, the treatment effects continue to be non-linear and approximately

[4]Unfortunately, 1-cigarette bins could not be used due to sample size restrictions.

www.manaraa.com

homogenous at all quantile levels. Interestingly, the main reduction in birth weight appears to be caused by increasing the number of cigarettes smoked from 0 to 1-2. This effect appears constant until the fourth cigarette. Increasing smoking beyond the fourth cigarette has an additional negative effect on birth weight, although this effect is smaller than the effect from 0 to 1-2.

www.manaraa.com

Chapter 4

Block Regression Estimators

This chapter derives the optimal rates of convergence of a nonparametric estimator of the regression function. In particular, this estimator is a generalization of the nonparametric estimator known as Partitioning in the statistical literature (see, e.g. Kohler, Krzyzak, and Walk (2006) and references therein). The potential usefulness of this estimator in the context of program evaluation is discussed below. After describing the motivating example, L_2 and uniform optimal rates of convergence are derived followed by a brief discussion of their applicability and how the findings of this chapter contribute to both the literature on program evaluation and, more generally, the literature of nonparametric regression.

www.manaraa.com

4.1 Motivating Example: Subclassification on the Propensity Score

The estimators proposed in Chapter 2, as well as many other alternative estimators available in the literature of program evaluation, achieve identification of the estimand of interest by means of Assumption 1. However, as originally discussed in Rosenbaum and Rubin (1983) for the binary treatment case and in Imbens (2000) for the multi-valued treatment case, identification may also be obtained by conditioning on the (Generalized) Propensity Score.[1] Assuming $\mathcal{T} = \{0, 1\}$ and focusing on the ATE for simplicity, this idea leads to a well-known estimator in program evaluation that conditions on the (estimated) propensity score to remove bias due to endogenous selection. This estimator was originally suggested in Rosenbaum and Rubin (1983) and is generically referred to as Subclassification on the Propensity Score.

To describe the estimator, first note that Assumption 1 implies

$$Y(t) \perp\!\!\!\perp D_t \mid e^*(X), \tag{4.1}$$

for $t \in \{0, 1\}$ and where $e^*(X) = p_1^*(X)$ for notational simplicity. Consequently, an alternative estimation procedure would proceed by conditioning on the (estimated) propensity score rather than on the observable characteristics directly, because

$$\begin{aligned} \Delta^{ATE} &= \mathbb{E}[Y(1) - Y(0)] \\ &= \mathbb{E}\left[\mathbb{E}[Y \mid T = 1, e^*(X)] - \mathbb{E}[Y \mid T = 0, e^*(X)]\right]. \end{aligned}$$

[1] For a recent discussion on this topic see, e.g., Imai and Dyk (2004).

www.manaraa.com

In general, a fully nonparametric implementation of this idea leads to

$$\hat{\Delta} = \widehat{\mathbb{E}}\left[\widehat{\mathbb{E}}\left[Y \mid T = 1, \hat{e}\left(X\right)\right] - \widehat{\mathbb{E}}\left[Y \mid T = 0, \hat{e}\left(X\right)\right]\right],$$

where $\hat{e}\left(\cdot\right)$ is a (non-)parametric estimator of the propensity score and $\widehat{\mathbb{E}}\left[Y \mid T = t, \cdot\right]$ is a corresponding nonparametric estimator of the regression function $\mathbb{E}\left[Y \mid T = t, \cdot\right]$ for $t \in \{0, 1\}$. In particular, in a very influential paper, Rosenbaum and Rubin (1983) propose the following nonparametric estimator: given the estimator $\hat{e}\left(\cdot\right)$ (parametric or nonparametric), partition the support of $\hat{e}\left(\cdot\right)$ in J blocks, $j = 1, \cdots, J$, compute the within-block difference in means between treatment and control groups, and then estimate the ATE by a weighted average across blocks, where the weights are the proportion of observations in each block. More formally,

$$\hat{\Delta}_1 = \sum_{j=1}^{J} \frac{\hat{N}_j}{n} \left(\bar{Y}_{1j} - \bar{Y}_{0j}\right),$$

where $\hat{N}_j = \hat{N}_{1j} + \hat{N}_{0j}$,

$$\hat{N}_{tj} = \sum_{i=1}^{n} \mathbf{1}\left\{\hat{e}\left(X_i\right) \in \mathcal{W}_j\right\} D_{t,i},$$

$\{\mathcal{W}_j : j = 1, \cdots J\}$ forms a partition of $\text{supp}\left(e\left(X\right)\right)$, and $\bar{Y}_{tj}$ is the mean of the outcome variable for group t in block j. Intuitively, this estimator can be regarded as nonparametric in the sense that (under some conditions) when $J \to \infty$ the mean of the outcome variable within-block approximates the underlying regression function (nonparametrically), which then is averaged out across blocks to obtain the overall estimator. As it is standard in nonparametric regression problems, J can be regarded

www.manaraa.com

as the smoothing parameter in the sense that the higher J the lower the bias and the higher the variance. This idea motivated Rosenbaum and Rubin (1983) to suggest a modification of this estimator, given by

$$\hat{\Delta}_2 = \sum_{j=1}^{J} \frac{\hat{N}_j}{n} \left(\hat{\alpha}_{1j} + \hat{\beta}_{1j} \bar{e}_j - \hat{\alpha}_{0j} - \hat{\beta}_{0j} \bar{e}_j \right),$$

where

$$\left(\hat{\alpha}_{tj}, \hat{\beta}_{tj} \right) = \arg\min_{\alpha,\beta} \sum_{i=1}^{n} \mathbf{1} \left\{ \hat{e}\left(X_i\right) \in \mathcal{W}_j \right\} D_{t,i} \left(Y_i - \alpha - \beta \hat{e}\left(X_i\right) \right)^2,$$

and

$$\bar{e}_j = \frac{1}{\hat{N}_j} \sum_{i=1}^{n} \mathbf{1} \left\{ \hat{e}\left(X_i\right) \in \mathcal{W}_j \right\} \hat{e}\left(X_i\right).$$

Intuitively, $\hat{\Delta}_2$ removes further bias within-block by estimating a linear regression rather than just the mean for each group.

It is not difficult to verify that these two estimators are also given by $\hat{\Delta}_1 = \hat{\Delta}_{1,1} - \hat{\Delta}_{0,1}$ and $\hat{\Delta}_2 = \hat{\Delta}_{1,2} - \hat{\Delta}_{0,2}$ with

$$\hat{\Delta}_{t,K} = \frac{1}{n} \sum_{i=1}^{n} \hat{\mu}_{t,K} \left(\hat{e}\left(X_i\right) \right),$$

where $R_K\left(e\right) = \left[1, e, e^2, \cdots, e^{K-1}\right]$, $K \in \mathbb{N}$,

$$\hat{\mu}_{t,K}\left(e\right) = \sum\nolimits_{j=1}^{J} \mathbf{1} \left\{ e \in \mathcal{W}_j \right\} R_K\left(e\right)' \hat{\gamma}_{t,K}\left(j\right),$$

and

$$\hat{\gamma}_{t,K}\left(j\right) = \arg\min_{\gamma} \sum_{i=1}^{n} \mathbf{1} \left\{ \hat{e}\left(X_i\right) \in \mathcal{W}_j \right\} D_{t,i} \left(Y_i - R_K\left(\hat{e}\left(X_i\right) \right)' \gamma \right)^2.$$

www.manaraa.com

As a consequence, it is of interest to develop the asymptotic properties of the non-parametric estimator $\hat{\mu}_{t,K}(e)$ since it enters directly in the definitions of $\hat{\Delta}_1$ and $\hat{\Delta}_2$. In this chapter $\hat{\mu}_{t,K}(e)$ is generically referred to as the Block Regression Estimator.

Somehow surprising, even though the estimators $\hat{\Delta}_1$ and $\hat{\Delta}_2$ have been widely used in applications in different fields of study, there is no formal statistical theory that addresses their properties. One possible justification for the lack of formal theory underlying these estimators is their complexity. In particular, note that a preliminary (non-)parametric estimator for the propensity score is plugged-in in a non-smooth function at the time that an increasing number of regressions within blocks (which are in fact determined by the random function $\hat{e}(\cdot)$) are averaged-out. Moreover, note that additional complications may arise from the fact that the estimator involves random denominators that may take a value of zero. In a first attempt to develop formal large sample results for $\hat{\Delta}_1$, Cattaneo, Imbens, Pinto, and Ridder (2008) proceed by assuming that $e(\cdot)$ is known and derive a mean squared error expansion of the estimator and establish its large sample distribution under some regularity conditions.

This chapter derives optimal rates of convergence for $\hat{\mu}_{t,K}(\cdot)$, which constitutes a first step in the characterization of the large sample properties of objects such as $\hat{\Delta}_1$ and $\hat{\Delta}_2$. Even though this motivation for studying the large sample properties of the Block Regression Estimator comes from the program evaluation literature, the results derived in the next section may be of independent interest for the literature of nonparametric regression and in particular for splines regression, as discussed in

www.manaraa.com

more detail below.

4.2 Optimal Rates of Convergence for Block Regression Estimators

This section derives the optimal rates of convergence for the Block Regression Estimator, a particular nonparametric estimator of a regression function that is implicitly used when estimating the ATE by subclassification on the propensity score. The notation used in the preceding section is maintained whenever possible to facilitate the comparison across sections. However, it is assumed here that the propensity score is known, so that $E_i = e(X_i)$, and that there is only one group (e.g., $T = 1$).

To describe the estimator, recall that $Y \in \mathbb{R}$, and $E \in \mathcal{W} = [p_{\min}, 1 - p_{\min}]$ (note that $|\mathcal{W}| = 1 - 2p_{\min}$), and (Y_i, E_i), $i = 1, \cdots, n$, i.i.d. Let $\mu^*(e) = \mathbb{E}[Y \mid E]$, and consider the estimator

$$\hat{\mu}_K(e) = \sum_{j=1}^{J} \mathbf{1}_{N_j} \mathbf{1}_{\mathcal{W}_j}(e) R_K(e)' \hat{\gamma}_{j,K},$$

where $\mathbf{1}_{N_j} = \mathbf{1}\{N_j \geq K\}$, $\mathbf{1}_{\mathcal{W}_j}(e) = \mathbf{1}\{e \in \mathcal{W}_j\}$,

$$N_j = \sum_{i=1}^{n} \mathbf{1}\{E_i \in \mathcal{W}_j\},$$

and

$$\hat{\gamma}_{j,K} \in \arg\min_{\gamma} \sum_{i=1}^{n} \mathbf{1}\{E_i \in \mathcal{W}_j\} \left(Y_i - R_K(E_i)'\gamma\right)^2,$$

www.manaraa.com

with $R_K(e) = \left[1, e, e^2, \cdots, e^{K-1}\right]'$ (this is, $R_K(e)$ is a polynomial basis of approximation) and $(\mathcal{W}_j : 1, \cdots, J)$ forms a partition on $[p_{\min}, 1 - p_{\min}]$. To save notation, it is assumed that $|\mathcal{W}_j| = |\mathcal{W}|/J$ for all $j = 1, \cdots, J$, this is, all blocks have the same length over the support of the random variable E.[2] Note that random sampling gives $N_j \sim \text{Bin}(q_j, n)$, $q_j = \mathbb{P}[E \in \mathcal{W}_j]$.

The main difference between the informal discussion of the previous section and the formal description of the nonparametric estimator given here is the truncation indicator $\mathbf{1}_{N_j}$, which is introduced to account for the fact that there may be less observations than needed to obtain a unique solution for the (least squares) minimization problem within block. Nonetheless, as shown below, under a mild restriction imposed on the rate of growth for J this event occurs with probability approaching zero (exponentially fast).

The following assumption is sufficient to establish the L_2 and L_∞ rates of convergence for $\hat{\mu}_K(e)$:

Assumption 8 *On* $\mathcal{W} = [p_{\min}, 1 - p_{\min}]$,

(a) the density of E*, denoted* $f(\cdot)$*, is bounded and bounded away from zero;*

(b) $\mu^*(e)$ *is* K *times continuously differentiable; and*

(c) $\mathbb{V}[Y \mid E = e]$ *and* $\mathbb{E}\left[(Y - \mu^*(e))^4 \mid E = e\right]$ *are bounded and bounded away from zero.*

[2]The results presented in this section can in fact be extended to higher dimensions and/or other basis of approximation. Similarly, it is easy to see that (under some regularity conditions) other configurations of blocks will not affect the large sample results.

www.manaraa.com

Assumption 8(a) leads directly to $C_1/J \leq q_j \leq C_2/J$ for some positive constants C_1 and C_2, uniformly in j. Moreover, if $J^{-1}n \to \infty$ it follows by Chernoff's Inequality that

$$\mathbb{P}\left[\mathbf{1}_{N_j} = 0\right] = \mathbb{P}\left[N_j < K\right] \leq C \exp\left\{-J^{-1}n\right\},$$

and thus

$$\mathbb{P}\left[\min_{1\leq j\leq J} \mathbf{1}_{N_j} = 0\right] \leq CJ \exp\left\{-CJ^{-1}n\right\},$$

that is, the truncation indicators decline exponentially fast to zero uniformly in j.

The following theorem is the main result of this chapter.

Theorem 11 *If Assumption 8 holds and $J^{-1}n \to \infty$, then for $K \in \mathbb{N}$,*

$$\int \left(\hat{\mu}_K(e) - \mu^*(e)\right)^2 \, dF(e) = O_p\left(J/n + J^{-2K}\right).$$

If Assumption 8 holds and $J^2 n^{-1} = O(1)$, then for $K \in \mathbb{N}$,

$$\sup_{e\in\mathcal{W}} \left|\hat{\mu}_K(e) - \mu^*(e)\right| = O_p\left(J^{1/2}\left(\log(n)/n\right)^{1/2} + J^{-K}\right)$$

The results in Theorem 11 show that the nonparametric rates of convergence of the Block Regression estimator are optimal. It is easy to verify that these rates attain the optimal bound derived in Stone (1982). The rate of convergence in L_2-norm has been already established in the literature when $K = 1$ (see, e.g., Kohler, Krzyzak, and Walk (2006) and references therein), while the optimal rate of convergence in L_∞-norm appears to be new.

It is interesting to note that the nonparametric estimator considered here is very similar in spirit to the Regression Splines Series Estimator (see, e.g., Newey (1997)).

www.manaraa.com

In particular, both estimators approximate nonparametrically a regression function by fitting a polynomial (of some degree) over an increasing sequence of shrinking "blocks" or "partitions", where the end-points of these intervals are some times referred to as knots. The only substantive difference between Regression Splines and Block Regression is that the former imposes a certain degree of smoothness in the overall fit (in the sense that the piece-wise polynomials are required to be continuous at the end-points and admit certain degree of differentiability), while the latter leaves completely unrestricted how the estimated polynomials in each block are related. Somehow surprising, Regression Splines do not attain the optimal uniform rate of convergence while the estimator considered in this chapter does. Although it may be an artifact of the proofs available (for a recent proof of this result see, e.g., de Jong (2002)), it is reasonable to conjecture that the increase in speed of uniform convergence enjoyed by Block Regression is achieved because of the relaxation of the restrictions imposed by Splines at the end-points.

As mentioned before, the results of Theorem 11 coupled with the proofs presented in Appendix C may be used to establish the large sample properties of semiparametric estimators that use this particular nonparametric estimator for the nonparametric component of the model. In particular, these results can be used to analyze the asymptotic behavior of the estimators introduced in the previous section, generically known as Subclassification on the Propensity Score, in the context of program evaluation. Undoubtedly, establishing these results formally would require additional

www.manaraa.com

technical work and therefore this analysis is left for future research.

www.manaraa.com

Chapter 5

Conclusion

Chapter 2 studied the efficient estimation of a large class of multi-valued treatment effects implicitly defined by a possibly over-identified non-smooth collection of moment conditions. Two alternative estimators based on standard GMM arguments combined with the corresponding modifications needed to circumvent the fundamental problem of causal inference were proposed. Under regularity conditions, these estimators were shown to be root-n consistent, asymptotically normal and efficient for the general population parameter of interest. Using these estimators it was shown how other estimands of interest may also be efficiently estimated, allowing the researcher to recover a rich class of population parameters.

Chapter 3 discussed particular examples of multi-valued treatment effects covered by the general results presented in Chapter 2. It was shown that important results in the literature of program evaluation with binary treatment assignments may be seen

www.manaraa.com

as particular cases of the procedures discussed here when the treatment is dichotomous. Considering multi-valued treatment assignments provides the opportunity for a better characterization of the program under study. As illustrated in the empirical application also included in Chapter 3, collapsing a multiple treatment into a binary indicator may prevent the researcher from detecting the presence of important non-linear effects. More generally, in many applications it would not be surprising to have multiple differential impacts within and across treatments, which highlights the relevance of considering multi-valued treatments, when possible, for making informed policy decisions.

The theoretical results presented in Chapter 2 were obtained under the assumption of finite multi-valued treatments, which leads to a statistical model where many estimands of interest are regular, this is, they can be estimated at the parametric rate. A natural extension would be to relax this assumption to continuous treatment assignments. This may be appealing from an empirical perspective, but would make many population parameters of interest irregular. Nonetheless, when treatments are continuous, it may be possible to consider relevant regular estimands such as specific functionals of the treatment effect process or, more interestingly, alternative restrictions on the underlying statistical model that may deliver regular population parameters.

The results presented in this chapter could also be extended based on the developments available in the literature of binary treatment effects. For example, in

www.manaraa.com

applications it may be of interest to consider the multi-valued analogue of weighted treatment effects (Hirano, Imbens, and Ridder (2003)), including average and quantile treatment effects for a given treatment level as particular cases. Efficiency calculations and the corresponding efficient estimation procedures for these estimands may be derived by following and extending the work discussed here.

Note that the two estimators proposed in Chapter 2 are first-order efficient. However, as in the binary treatment case, other efficient estimators may also be considered, which implies that an important open question for future research is how to rank the large class of first-order efficient estimators available. Although it seems unclear how to rank these estimators, the results of this paper justify focusing on the marginal treatment effects as the target estimand when ranking the competing first-order efficient estimators.

Finally, Chapter 4 presented optimal rates of convergence for the Block Regression Estimator, a nonparametric estimator of the regression function. These results may be of particular importance for the literature of program evaluation since a commonly used estimator for ATE, known as Subclassification on the Propensity Score, can be written as a semiparametric estimator that uses the Block Regression Estimator. In addition, the results of this chapter contribute to the literature of nonparametric estimation of a regression function.

www.manaraa.com

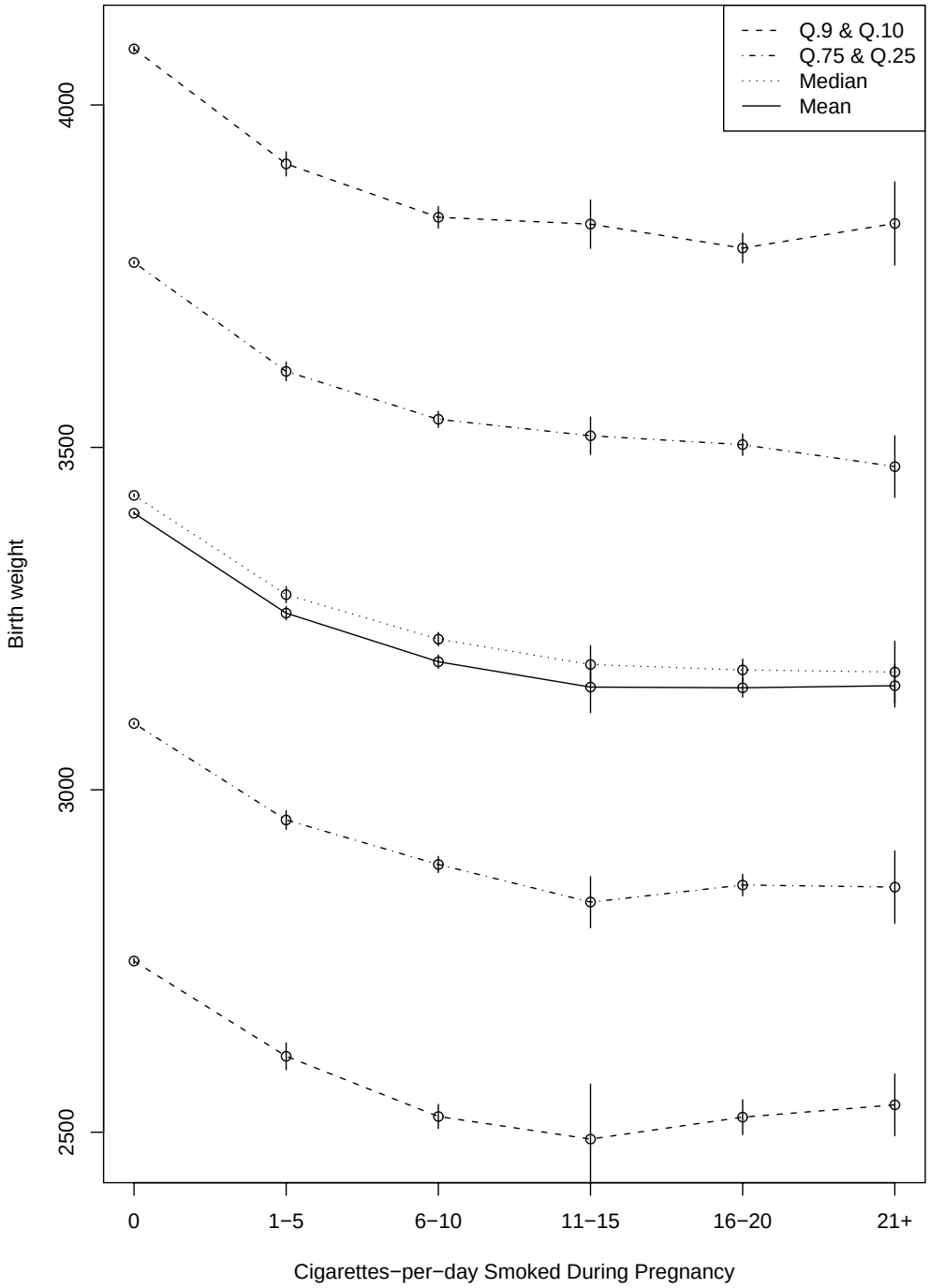

Figure 1: Effect of Maternal Smoking Intensity on Birth Weight (5-cigarette bins)

www.manaraa.com

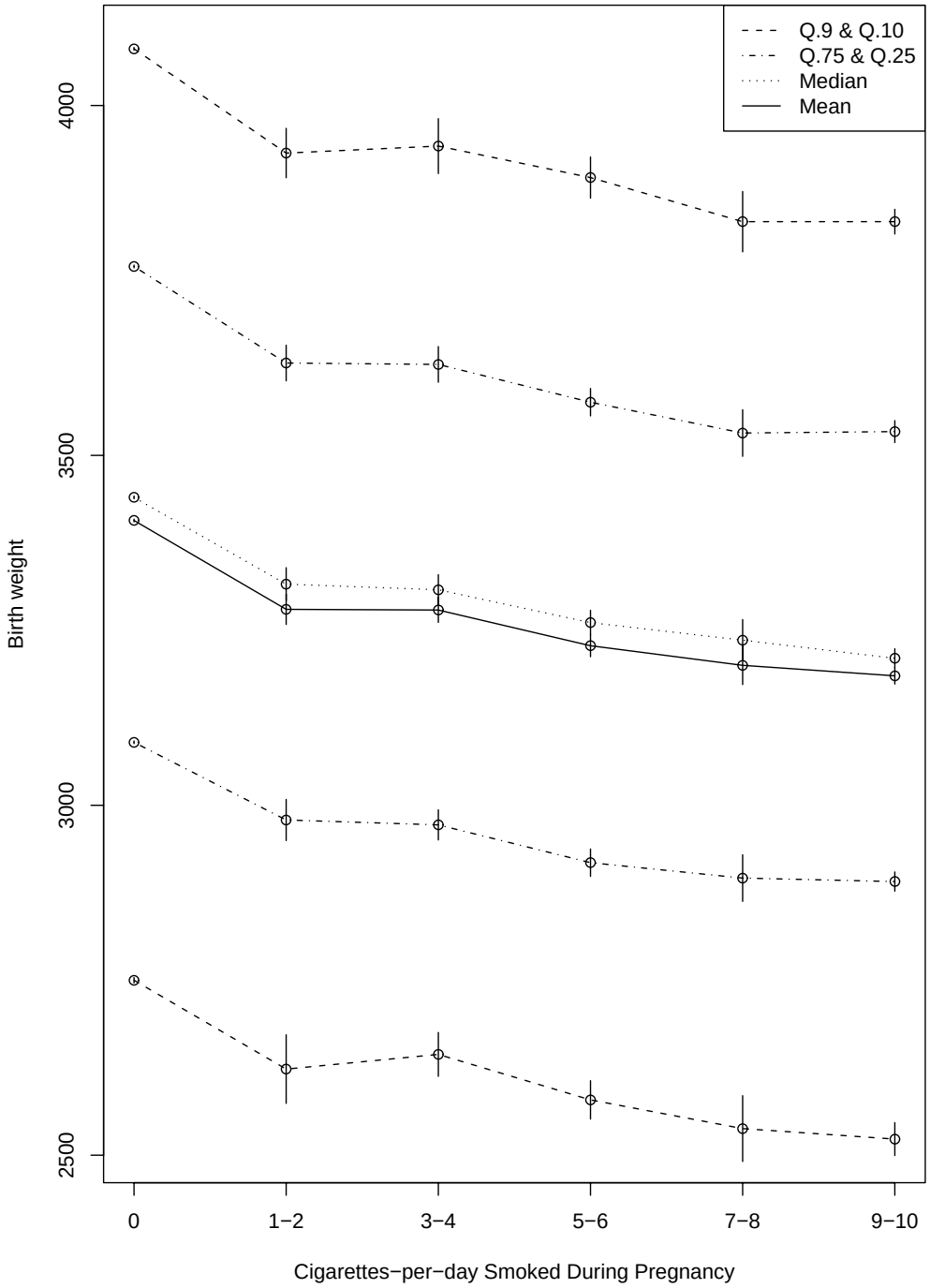

Figure 2: Effect of Maternal Smoking Intensity on Birth Weight (2-cigarette bins)

www.manaraa.com

المنارة للاستشارات

Table 1: Effect of Maternal Smoking Intensity on Birth Weight

	DRE							**IPWE**						
	Q.9	Q.75	Q.5	Mean	Q.25	Q.1	SD	Q.9	Q.75	Q.5	Mean	Q.25	Q.1	SD
0	4082	3771	3430	3417	3119	2778	576	4082	3770	3430	3404	3097	2750	586
	(1)	(1)	(1)	(2)	(1)	(2)	n.a.	(2)	(1)	(1)	(1)	(1)	(2)	(1)
1-5	3872	3572	3232	3189	2892	2520	601	3914	3611	3285	3258	2956	2611	577
	(7)	(5)	(5)	(25)	(5)	(9)	n.a.	(9)	(7)	(6)	(5)	(7)	(10)	(5)
6-10	3814	3503	3175	3133	2835	2466	593	3836	3541	3220	3187	2891	2523	568
	(4)	(3)	(3)	(17)	(4)	(6)	n.a.	(8)	(6)	(5)	(5)	(6)	(9)	(4)
11-15	3800	3515	3175	3161	2863	2523	565	3826	3517	3183	3150	2836	2490	598
	(12)	(9)	(8)	(44)	(10)	(14)	n.a.	(18)	(14)	(14)	(19)	(19)	(41)	(26)
16-20	3780	3487	3153	3119	2807	2460	581	3791	3504	3175	3149	2861	2522	565
	(5)	(4)	(4)	(21)	(5)	(8)	n.a.	(11)	(8)	(8)	(7)	(8)	(13)	(8)
21+	3799	3459	3147	3105	2780	2438	585	3827	3472	3172	3152	2858	2540	560
	(12)	(9)	(9)	(46)	(10)	(15)	n.a.	(31)	(23)	(23)	(16)	(27)	(23)	(16)

Notes: (i) DRE = Dummy Regression Estimator, IPWE = Inverse Probability Weighting Estimator.
(ii) Q.9, Q.75, Q.5, Q.25 and Q.1 are the 90%, 75%, 50%, 25% and 10% quantiles, respectively, and SD is the standard deviation.
(iii) standard errors in parentheses.

www.manaraa.com

Table 1 (cont.): Effect of Maternal Smoking Intensity on Birth Weight

	EIFE						
	Q.9	Q.75	Q.5	Mean	Q.25	Q.1	SD
0	4081	3770	3431	3405	3091	2750	582
	(1)	(1)	(1)	(1)	(1)	(2)	(1)
1-5	3914	3611	3285	3258	2954	2608	578
	(9)	(7)	(6)	(5)	(7)	(10)	(5)
6-10	3836	3549	3231	3189	2891	2529	567
	(8)	(6)	(5)	(5)	(6)	(9)	(5)
11-15	3826	3535	3201	3162	2865	2526	594
	(18)	(13)	(13)	(19)	(19)	(35)	(26)
16-20	3805	3503	3183	3156	2869	2540	560
	(11)	(8)	(8)	(7)	(8)	(12)	(8)
21+	3845	3487	3175	3165	2877	2549	546
	(31)	(24)	(23)	(16)	(25)	(22)	(17)

Notes: (i) EIFE = Eficient Influence Function Estimator.
(ii) Q.9, Q.75, Q.5, Q.25 and Q.1 are the 90%, 75%, 50%, 25% and 10% quantiles, respectively, and SD is the standard deviation.
(iii) standard errors in parentheses.

www.manaraa.com

Table 2: Hypothesis Tests for Pairwise Differences and Difference-in-Differences Effects

	T1-T0	T2-T0	T3-T0	T4-T0	T5-T0	T2-T1	T3-T1	T4-T1	T5-T1	T3-T2	T4-T2	T5-T2	T4-T3	T5-T3	T5-T4
T1-T0	-146*					75*	38	37*	40*	109*	108*	111*	145*	148*	149*
T2-T0		-217*				146*	109*	108*	111*	180*	179*	182*	216*	219*	220*
T3-T0			-254*			183*	146*	145*	148*	217*	216*	219*	253*	256*	257*
T4-T0				-255*		184*	147*	146*	149*	218*	217*	220*	254*	257*	258*
T5-T0					-252*	181*	144*	143*	146*	215*	214*	217*	251*	254*	255*
T2-T1						-71*				34	33*	36	70*	73*	74*
T3-T1							-108*			71*	70*	73*	107*	110*	111*
T4-T1								-109*		72*	71*	74*	108*	111*	112*
T5-T1									-106*	69*	68*	71*	105*	108*	109*
T3-T2										-37			36	39	40
T4-T2											-38*		37	40	41
T5-T2												-35*	34	37	38*
T4-T3													-1		4
T5-T3														2	1
T5-T4															3

Notes: (i) treatments T0, T1, T2, T3, T4 and T5 are 0, 1-5, 6-10, 11-15, 16-20 and 21+ cigarettes-per-day smoked, respectively.
(ii) pairwise differences are reported on the diagonal, and difference-in-differences are reported outside the diagonal.
(iii) in all cases the null hypothesis is zero differential effect; (iv) * significant at 5%.

www.manaraa.com

Table 3: Joint Hypotheses Tests (IPWE)

Joint Null Hypotheses	Restrictions	Wald Test	p-value
Equal treatment effects (mean, quantiles, spread) for (11-15,16-20,21+)	14	16.60	0.2781
Equal treatment effects (mean, quantiles, spread) for (6-10,11-15,16-20,21+)	21	55.86	0.0001
Equal treatment effects (mean, quantiles, spread) for (1-5,6-10,11-15,16-20,21+)	28	246.88	0.0000
Equal mean and median for each treatment	6	1402.62	0.0000
Equal mean-median difference (MMD) across treatments	5	3.78	0.5809
Equal standard deviation across treatments	5	25.38	0.0001
Equal interquartile range (IQR) across treatments	5	25.32	0.0001
Equal Q.9-Q.1 range (Q.9-Q.1) across treatments	5	21.98	0.0005
Equal MMD, IQR and Q.9-Q.1 across treatments	15	38.59	0.0007

Note: all tests have been computed using the IPWE and its corresponding limiting distribution.

www.manaraa.com

Bibliography

Abadie, A. (2005): "Semiparametric Difference-in-Differences Estimators," *Review of Economic Studies*, 72(1), 1–19.

Abadie, A., and G. W. Imbens (2006): "Large Sample Properties of Matching Estimators for Average Treatment Effects," *Econometrica*, 7(1), 235–267.

Ai, C., and X. Chen (2003): "Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions," *Econometrica*, 71(6), 1795–1843.

Almond, D., K. Y. Chay, and D. S. Lee (2005): "The Costs of Low Birth Weight," *Quarterly Journal of Economics*, 120(3), 1031–1083.

Andrews, D. W. K. (1994): "Empirical Process Methods in Econometrics," in *Handbook of Econometrics, Volumn IV*, ed. by R. F. Engle, and D. L. McFadden, pp. 2247–2294. Elsevier Science B.V.

www.manaraa.com

——— (2002): "Generalized Method of Moments Estimation When a Parameter Is on a Boundary," *Journal of Business and Economic Statistics*, 20(4), 530–544.

Bang, H., and J. M. Robins (2005): "Doubly Robust Estimation in Missing Data and Causal Inference Models," *Biometrics*, 61, 962–972.

Bickel, P. J., C. A. J. Klaasen, Y. Ritov, and J. A. Wellner (1993): *Efficient and Adaptive Estimation for Semiparametric Models*. Springer, New York.

Cattaneo, M. D., G. W. Imbens, C. Pinto, and G. Ridder (2008): "Subclassification on the Propensity Score: Large Sample Properties," Work in progress.

Chen, X. (2007): "Large Sample Sieve Estimation of Semi-Nonparametric Models," in *Handbook of Econometrics, Volumn VI*, ed. by J. Heckman, and E. Leamer. Elsevier Science B.V.

Chen, X., H. Hong, and E. Tamer (2005): "Measurement Error Models with Auxiliary Data," *Review of Economic Studies*, 72, 343–366.

Chen, X., H. Hong, and A. Tarozzi (2007): "Semiparametric Efficiency in GMM Models With Auxiliary Data," *The Annals of Statistics*, forthcoming.

Chen, X., O. Linton, and van Keilegom (2003): "Estimation of Semiparametric Models when The Criterion Function Is Not Smooth," *Econometrica*, 71(5), 1591–1608.

www.manaraa.com

Crump, R. K., V. Hotz, G. W. Imbens, and O. A. Mitnik (2007): "Dealing with Limited Overlap in Estimation of Average Treatment Effects," Working Paper.

de Jong, R. M. (2002): "A Note on "Convergence Rates and Asymptotic Normality for Series Estimators": Uniform Convergence Rates," *Journal of Econometrics*, 111, 1–9.

DeVore, R. A., and G. G. Lorentz (1993): *Constructive Approximation*. Springer, New York.

Firpo, S. (2007): "Efficient Semiparametric Estimation of Quantile Treatment Effects," *Econometrica*, 75(1), 259–276.

Frölich, M. (2004): "Programme Evaluation With Multiple Treatments," *Journal of Economic Surveys*, 18(2), 181–224.

Hahn, J. (1998): "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica*, 66(2), 315–331.

Heckman, J., H. Ichimura, and P. Todd (1998): "Matching as an Econometric Evaluation Estimator," *The Review of Economic Studies*, 65(2), 261–294.

Hirano, K., G. W. Imbens, and G. Ridder (2003): "Efficient Estimation of Average Treatment Effects using the Estimated Propensity Score," *Econometrica*, 71(4), 1161–1189.

www.manaraa.com

Horowitz, J. L., and C. F. Manski (2000): "Nonparametric Analysis of Randomized Experiments with Missing Covariate and Outcome Data," *Jounal of the American Statistical Association*, 95(449), 77–84.

Horvitz, D. G., and D. J. Thompson (1952): "A Generalization of Sampling Without Replacement from a Finite Population," *Journal of the American Statistical Association*, 47(260), 663–685.

Imai, K., and D. A. V. Dyk (2004): "Causal Inference With General Treatment Regimes: Generalizing the Propensity Score," *Journal of the American Statistical Association*, 99(467), 854–866.

Imbens, G. W. (2000): "The Role of the Propensity Score in Estimating Dose-Response Functions," *Biometrika*, 87(3), 706–710.

——— (2004): "Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review," *Review of Economics and Statistics*, 86(1), 4–29.

Imbens, G. W., W. K. Newey, and G. Ridder (2006): "Mean-Squared-Error Calculations for Average Treatment Effects," Working Paper.

Kohler, M., A. Krzyzak, and H. Walk (2006): "Rates of Convergence for Partitioning and Nearest Neighbor Regression Estimates with Unbounded Data," *Journal of Multivariate Analysis*, 97(1), 311–323.

Lechner, M. (2001): "Identification and Estimation of Causal Effects of Multiple

www.manaraa.com

Treatments Under The Conditional Independence Assumption," in *Econometric Evaluation of Labour Market Policies*, ed. by M. Lechner, and F. Pfeiffer, pp. 43–58. Physica/Springer, Heidelberg.

LEE, M. J. (2005): *Micro-Econometrics for Policy, Program and Treatment Effects*. Oxford University Press, Oxford.

NEWEY, W. K. (1990): "Semiparametric Efficiency Bounds," *Journal of Applied Econometrics*, 5(2), 99–135.

——— (1994): "The Asymptotic Variance of Semiparametric Estimators," *Econometrica*, 62(6), 1349–1382.

——— (1997): "Convergence Rates and Asymptotic Normality for Series Estimators," *Journal of Econometrics*, 79, 147–168.

NEWEY, W. K., AND D. MCFADDEN (1994): "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics, Volumn IV*, ed. by R. F. Engle, and D. L. McFadden, pp. 2112–2245. Elsevier Science B.V.

PAKES, A., AND D. POLLARD (1989): "Simulation and the Asymptotics of Optimization Estimators," *Econometrica*, 57(5), 1027–1057.

POLLARD, D. (1984): *Convergence of Stochastic Processes*. Springer, New York.

ROBINS, J. M., AND A. ROTNITZKY (1995): "Semiparametric Efficiency in Multi-

www.manaraa.com

variate Regression Models with Missing Data," *Journal of the American Statistical Association*, 90(429), 122–129.

ROBINS, J. M., A. ROTNITZKY, AND L. ZHAO (1994): "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed," *Journal of the American Statistical Association*, 89(427), 846–866.

——— (1995): "Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data," *Journal of the American Statistical Association*, 90(429), 846–866.

ROSENBAUM, P. R. (2002): *Observational Studies*. Springer, New York.

ROSENBAUM, P. R., AND D. B. RUBIN (1983): "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70(1), 41–55.

RUBIN, D. B. (1974): "Estimating Causal Effects of Treatments in Randomized and Non-Randomized Studies," *Journal of Educational Psychology*, 66, 688–701.

STONE, C. J. (1982): "Optimal Global Rates of Convergence for Nonparametric Regression," *Annals of Statistics*, 10(4), 1040–1053.

TANABE, K., AND M. SAGAE (1992): "An Exact Cholesky Decomposition and the Generalized Inverse of the Variance-Covariance Matrix of the Multinomial Distribution, with Applications," *Journal of the Royal Statistical Society. Series B Methodological*, 54(1), 211–219.

www.manaraa.com

TSIATIS, A. A. (2006): *Semiparametric Theory and Missing Data*. Springer, New York.

VAN DER VAART, A. W. (1998): *Asymptotic Statistics*. Cambridge University Press, New York.

VAN DER VAART, A. W., AND J. A. WELLNER (1996): *Weak Convergence and Empirical Processes*. Springer, New York.

——— (2000): "Preservation Theorems for Glivenko-Cantelli and Uniform Glivenko-Cantelli Classes," in *High Dimensional Probability II*, ed. by E. Giné, D. Mason, and J. A. Wellner, pp. 115–134. Bikhäuser, Boston.

WOOLDRIDGE, J. M. (2007): "Inverse probability weighted estimation for general missing data problems," *Journal of Econometrics*, forthcoming.

www.manaraa.com

Appendix A

Proof of Theorems in Chapter 1

Throughout all the appendixes C denotes a generic positive constant which may vary depending on the context. Also, for any vector v, let $v_{[t]}$ denote its t-th element, and for any matrix A denote its (i,j)-th element by $A_{[i,j]}$. $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ are the minimum and maximum eigenvalue of the matrix A, respectively.

Proof of Theorem 1 (EIF AND SPEB)**:** the proof given is based on the theoretical approach described in Bickel, Klaasen, Ritov, and Wellner (1993) and Newey (1990), and follows the results presented in Hahn (1998) and Chen, Hong, and Tarozzi (2007). The derivation is completed in three steps: characterization of the tangent space, verification of pathwise differentiability of the parameter of interest, and SPEB computation. Let $L_0^2(F_W)$ be the usual Hilbert space of zero-mean, square-integrable functions with respect to the distribution function F_W.

First, consider a (regular) parametric submodel of the joint distribution of (Y,T,X),

www.manaraa.com

the observed data model, with c.d.f. $F(y,t,x;\theta)$ and log-likelihood given by

$$\log f(y,t,x;\theta) = \sum\nolimits_{j\in\mathcal{T}} \mathbf{1}\{t=j\}\left[\log f_j(y \mid x;\theta) + \log p_j(x;\theta)\right] + \log f_X(x;\theta),$$

which equals $\log f(y,t,x)$ when $\theta=\theta_0$, and where $f_j(y \mid x;\theta)$ corresponds to the density of $Y(j)|X$, $p_j(x;\theta) = \mathbb{P}[D_j=1|x;\theta]$ and $p_j(x;\theta_0) = p_j^*(x)$ for all $j\in\mathcal{T}$. The corresponding score is given by

$$S(y,t,x;\theta_0) = \frac{d}{d\theta}\log f(y,t,x;\theta)\Big|_{\theta_0} = S_y(y,t,x) + S_p(t,x) + S_x(x),$$

where

$$\begin{aligned}
S_y(y,t,x) &= \sum\nolimits_{j\in\mathcal{T}} \mathbf{1}\{t=j\}\, s_j(y,x), & s_j(y,x) &= \frac{d}{d\theta}\log f_j(y \mid x;\theta)\Big|_{\theta_0},\\
S_p(t,x) &= \sum\nolimits_{j\in\mathcal{T}} \mathbf{1}\{t=j\}\frac{\dot{p}_j^*(x)}{p_j^*(x)}, & \dot{p}_j^*(x) &= \frac{d}{d\theta}p_j(x;\theta)\Big|_{\theta_0},\\
S_x(x) &= \frac{d}{d\theta}\log f_X(x;\theta)\Big|_{\theta_0}.
\end{aligned}$$

Therefore, the tangent space of this statistical model is characterized by the set of functions $\mathcal{T} \equiv \mathcal{T}_y + \mathcal{T}_p + \mathcal{T}_x$, where

$$\begin{aligned}
\mathcal{T}_y &= \left\{S_y(Y,T,X) : s_j(Y(t),X) \in L_0^2\left(F_{Y(t)|X}\right),\ \forall j\in\mathcal{T}\right\},\\
\mathcal{T}_p &= \left\{S_p(T,X) : S_p(T,X) \in L_0^2\left(F_{T|X}\right)\right\},\\
\mathcal{T}_x &= \left\{S_x(X) : S_x(X) \in L_0^2(F_X)\right\}.
\end{aligned}$$

In particular, observe that

$$\mathbb{E}[S_p(T,X) \mid X] = \mathbb{E}\left[\sum\nolimits_{t\in\mathcal{T}} D_j\frac{\dot{p}_t^*(X)}{p_t^*(X)} \;\middle|\; X\right] = \sum\nolimits_{t\in\mathcal{T}}\dot{p}_t(X;\theta_0),$$

www.manaraa.com

and

$$\mathbb{E}\left[S_p\left(T,X\right)^2 \mid X\right] = \mathbb{E}\left[\sum\nolimits_{i\in\mathcal{T}}\sum\nolimits_{j\in\mathcal{T}} D_i\frac{\dot{p}_i^*\left(X\right)}{p_i^*\left(X\right)}D_j\frac{\dot{p}_j^*\left(X\right)}{p_j^*\left(X\right)} \;\middle|\; X\right] = \sum\nolimits_{t\in\mathcal{T}}\frac{\dot{p}_t^*\left(X\right)^2}{p_t^*\left(X\right)},$$

and hence it is required that $p_t^*\left(x\right)$ and $\dot{p}_t\left(x;\theta_0\right)$ are measurable functions such that $\sum_{t\in\mathcal{T}}\dot{p}_t^*\left(X\right)=0$ and $\sum_{t\in\mathcal{T}}\dot{p}_t^*\left(X\right)^2/p_t^*\left(X\right)<\infty$, almost surely. Notice that the first condition implies that by varying the model the probabilities should change in such a way that they still add up to one. The second condition is verified by Assumption 1(b) and the fact that T is finite.

Next, define

$$\underline{\mathrm{m}}\left(\beta\right) = \left[m\left(Y\left(0\right);\beta_0\right)',\cdots,m\left(Y\left(J\right);\beta_J\right)'\right]$$

and let A be any $d_\beta\left(J+1\right)\times d_m\left(J+1\right)$ positive semi-definite matrix. Then the population parameter of interest satisfies $A\mathbb{E}[\underline{\mathrm{m}}(\beta)]=0$ if and only if $\beta=\beta^*$, and using the implicit function theorem,

$$\frac{\partial}{\partial\theta}\beta^*\left(\theta\right) = -\left(A\Gamma_*\right)^{-1}A\Upsilon\left(\theta_0\right),$$

where

$$\Gamma_* = \frac{\partial}{\partial\beta}\mathbb{E}\left[\underline{\mathrm{m}}\left(\beta\right)\right]\Big|_{\beta=\beta^*},$$

$$\Upsilon\left(\theta_0\right) = \frac{\partial}{\partial\theta}\mathbb{E}_\theta\left[\underline{\mathrm{m}}\left(\beta^*\right)\right]\Big|_{\theta=\theta_0} = \frac{\partial}{\partial\theta}\int\underline{\mathrm{m}}\left(\beta^*\right)dF\left(y,t,x;\theta\right)\Big|_{\theta=\theta_0},$$

and observe that

$$\Upsilon\left(\theta_0\right) = \left[\frac{\partial}{\partial\theta}\mathbb{E}_\theta\left[m\left(Y\left(0\right);\beta_0\right)'\right]\Big|_{\theta=\theta_0},\cdots,\frac{\partial}{\partial\theta}\mathbb{E}_\theta\left[m\left(Y\left(J\right);\beta_J\right)'\right]\Big|_{\theta=\theta_0}\right]$$

www.manaraa.com

with typical element $j \in \mathcal{T}$,

$$\frac{\partial}{\partial\theta}\mathbb{E}_\theta\left[m\left(Y\left(j\right);\beta_j^*\right)'\right]\Big|_{\theta=\theta_0} = \mathbb{E}\left[m\left(Y\left(j\right);\beta_j^*\right)s_j\left(Y\left(j\right)\mid X\right)\right]+\mathbb{E}\left[e_j^*\left(X;\beta_j^*\right)S_x\left(X\right)\right].$$

Now, to show that the parameter is pathwise differentiable it is needed to find a $d_\beta\left(J+1\right)$-valued function $\Psi_\beta\left(y,t,x;A\right) \in \mathcal{T}$ such that for all regular parametric submodels

$$\frac{\partial}{\partial\theta}\beta^*\left(\theta\right) = \mathbb{E}\left[\Psi_\beta\left(Y,T,X;A\right)S\left(Y,T,X;\theta_0\right)\right].$$

It is not difficult to verify that the function satisfying such condition is given by

$$\Psi_\beta\left(Y,T,X;A\right) = -\left(A\Gamma_*\right)^{-1}A\psi\left(Y,T,X;\beta^*,p^*,e^*\left(\beta^*\right)\right),$$

for a fixed choice of the matrix A.

Finally, it follows from semiparametric efficiency theory and standard GMM arguments that the EIF is obtained when $A = \Gamma_*'V_*^{-1}$, which leads to the SPEB given by $V^* = (\Gamma_*V_*^{-1}\Gamma_*')^{-1}$. ■

Proof of Theorem 3 (CONSISTENCY OF IPWE): the proof applies Corollary 3.2 in Pakes and Pollard (1989) after setting $\theta = \beta$, $\theta_0 = \beta^*$, $G_n\left(\beta\right) = A_nM_n^{IPW}\left(\beta,\hat{p}\right)$, $G\left(\beta\right) = AM^{IPW}\left(\beta,p^*\right)$, and verifying their three sufficient conditions (i), (ii), and (iii). First observe that conditions (i) and (ii) are satisfied by construction of the estimator and the model considered. Next, because $A_n - A = o_p\left(1\right)$, to verify condition

www.manaraa.com

(iii) it is enough to show

$$\begin{aligned}
&\sup_{\beta\in\mathcal{B}}\left|M^{IPW}_{[t],n}\left(\beta,\hat{p}_t\right)-M^{IPW}_{[t]}\left(\beta,p^*_t\right)\right| \\
&\leq \sup_{\beta\in\mathcal{B}}\left|M^{IPW}_{[t],n}\left(\beta,\hat{p}_t\right)-M^{IPW}_{[t],n}\left(\beta,p^*_t\right)\right|+\sup_{\beta\in\mathcal{B}}\left|M^{IPW}_{[t],n}\left(\beta,p^*_t\right)-M^{IPW}_{[t]}\left(\beta,p^*_t\right)\right| \\
&= o_p\left(1\right),
\end{aligned}$$

for all $t\in\mathcal{T}$. Now the result follows because for n large enough,

$$\begin{aligned}
&\sup_{\beta\in\mathcal{B}}\left|M^{IPW}_{[t],n}\left(\beta,\hat{p}_t\right)-M^{IPW}_{[t],n}\left(\beta,p^*_t\right)\right| \\
&\leq C\left\|\hat{p}_t-p^*_t\right\|_\infty\frac{1}{n}\sum_{i=1}^{n}\frac{D_{t,i}}{p^*_t\left(X_i\right)}\sup_{\beta_t\in\mathcal{B}}\left|m\left(Y_i;\beta_t\right)\right| \\
&= o_p\left(1\right),
\end{aligned}$$

by Assumption 3(b), and

$$\begin{aligned}
&\sup_{\beta\in\mathcal{B}}\left|M^{IPW}_{[t],n}\left(\beta,p^*_t\right)-M^{IPW}_{[t]}\left(\beta,p^*_t\right)\right| \\
&= \sup_{\beta_t\in\mathcal{B}}\left|\frac{1}{n}\sum_{i=1}^{n}\frac{D_{t,i}m\left(Y_i;\beta_t\right)}{p^*_t\left(X_i\right)}-\mathbb{E}\left[\frac{D_t m\left(Y;\beta_t\right)}{p^*_t\left(X\right)}\right]\right| \\
&= o_p\left(1\right)
\end{aligned}$$

because (assuming $d_m=1$ or applying the following argument element by element) the class of functions $\mathcal{F}_t=\left\{\mathbf{1}\left\{\cdot=t\right\}\cdot m\left(\cdot;\beta\right)/p^*_t\left(\cdot\right):\beta\in\mathcal{B}\right\}$ is Glivenko-Cantelli by Assumptions 1(b) and 3 (van der Vaart and Wellner (2000)). ■

Proof of Theorem 4 (CONSISTENCY OF EIFE): the proof of this theorem follows the same logic as the proof of Theorem 3. It is applied Corollary 3.2 in

www.manaraa.com

Pakes and Pollard (1989) after setting $\theta = \beta$, $\theta_0 = \beta^*$, $G_n(\theta) = A_n M_n^{EIF}(\beta, \hat{p}, \hat{e})$, $G(\theta) = AM^{EIF}(\beta, p^*, e^*)$, and verifying their three sufficient conditions (i), (ii), and (iii). Using the same arguments in the proof and the conclusion of Theorem 3, it is sufficient to show

$$\sup_{\beta \in \mathcal{B}} \left| \frac{1}{n} \sum_{i=1}^{n} \hat{e}_t(X_i; \beta) \frac{D_{t,i} - \hat{p}_t(X_i)}{\hat{p}_t(X_i)} \right| = o_p(1),$$

for all $t \in \mathcal{T}$. To establish this result, first notice that $\mathbb{E}\left[\sup_{\beta \in \mathcal{B}} |e_t^*(X; \beta)|\right] < \infty$ for all $t \in \mathcal{T}$ by Assumption 3(b). Now, for n large enough,

$$\begin{aligned}
&\sup_{\beta \in \mathcal{B}} \left| \frac{1}{n} \sum_{i=1}^{n} \hat{e}_t(X_i; \beta) \frac{D_{t,i} - \hat{p}_t(X_i)}{\hat{p}_t(X_i)} \right| \\
&\leq C \sup_{\beta \in \mathcal{B}} \left\| \hat{e}_t(\beta) - e_t^*(\beta) \right\|_\infty + \sup_{\beta_t \in \mathcal{B}} \left| \frac{1}{n} \sum_{i=1}^{n} e_t^*(X_i; \beta) \frac{D_{t,i} - p_t^*(X_i)}{p_t^*(X_i)} \right| + o_p(1) \\
&= o_p(1),
\end{aligned}$$

because (assuming $d_m = 1$ or applying the argument element by element) the class of functions $\mathcal{F}_t = \{e_t^*(\cdot; \beta)(\mathbf{1}\{\cdot = t\} - p_t^*(\cdot))/p_t^*(\cdot) : \beta \in \mathcal{B}\}$ is Glivenko-Cantelli by Assumptions 1(b) and 3 (van der Vaart and Wellner (2000)). ■

Proof of Theorem 5 (Asymptotic Linear Representation of IPWE): it is applied Theorem 3.3 and Lemma 3.5 in Pakes and Pollard (1989) after setting $\theta = \beta$, $\theta_0 = \beta^*$, $G_n(\beta) = A_n M_n^{IPW}(\beta, \hat{p})$, $G(\beta) = AM^{IPW}(\beta, p^*)$, and verifying their five sufficient conditions (i)-(v). First, observe that conditions (i), (ii), (iv) and (v) hold by the construction of the estimator, Assumptions 2 and 5, and condition (3.1). Thus it only remains to show the stochastic equicontinuity condition (iii). To

www.manaraa.com

establish this condition, it suffices to show (see, e.g., Lemma 3.5 in Pakes and Pollard (1989) and Lemma 1 in Andrews (2002)) for all sequences $\delta_n = o(1)$ that

$$\sup_{|\beta_t - \beta_t^*| \leq \delta_n} \frac{n^{1/2} \left| M_{[t],n}^{IPW}(\beta, \hat{p}) - M_{[t]}^{IPW}(\beta^*, p^*) - M_{[t],n}^{IPW}(\beta^*, \hat{p}) \right|}{1 + Cn^{1/2} |\beta_t - \beta_t^*|} = o_p(1),$$

for all $t \in \mathcal{T}$. Now, to verify this final condition define

$$\Delta_{[t],n}(\beta, p - p^*) = -\frac{1}{n} \sum_{i=1}^{n} \frac{D_{t,i} m(Y_i; \beta_t)}{p_t^*(X_i)^2} (p_t(X_i) - p_t^*(X_i)),$$

and consider the following decomposition

$$\begin{aligned}
&\left| M_{[t],n}^{IPW}(\beta, \hat{p}) - M_{[t]}^{IPW}(\beta, p^*) - M_{[t],n}^{IPW}(\beta^*, \hat{p}) \right| \\
&\leq \left| M_{[t],n}^{IPW}(\beta, p^*) - M_{[t]}^{IPW}(\beta, p^*) - M_{[t],n}^{IPW}(\beta^*, p^*) \right| && \text{(A.1)} \\
&\quad + \left| M_{[t],n}^{IPW}(\beta, \hat{p}) - M_{[t],n}^{IPW}(\beta, p^*) - \Delta_{[t],n}(\beta, \hat{p} - p^*) \right| && \text{(A.2)} \\
&\quad + \left| M_{[t],n}^{IPW}(\beta^*, \hat{p}) + M_{[t],n}^{IPW}(\beta^*, p^*) - \Delta_{[t],n}(\beta^*, \hat{p} - p^*) \right| && \text{(A.3)} \\
&\quad + \left| \Delta_{[t],n}(\beta, \hat{p} - p^*) - \Delta_{[t],n}(\beta^*, \hat{p} - p^*) \right|. && \text{(A.4)}
\end{aligned}$$

Now, for n large enough and using the first term (A.1),

$$\sup_{|\beta_t - \beta_t^*| \leq \delta_n} \frac{n^{1/2} \left| M_{[t],n}^{IPW}(\beta, p^*) - M_{[t]}^{IPW}(\beta^*, p^*) - M_{[t],n}^{IPW}(\beta^*, p^*) \right|}{1 + Cn^{1/2} |\beta_t - \beta_t^*|} = o_p(1)$$

because (assuming $d_m = 1$ or applying the following argument element by element) the class of functions $\mathcal{F}_t = \{\mathbf{1}\{\cdot = t\} m(\cdot; \beta) / p_t^*(\cdot) : |\beta - \beta_t^*| \leq \delta\}$ is Donsker with finite integrable envelope by Assumption 5 (Theorem 2.10.6 of van der Vaart and Wellner (1996)) and L_2 continuous by Assumptions 2 and 5 (compare to Lemma 2.17 in Pakes and Pollard (1989)).

www.manaraa.com

For the second term (A.2),

$$\begin{aligned}
&\sup_{|\beta_t-\beta_t^*|\leq\delta_n} \frac{n^{1/2}\left|M_{[t],n}^{IPW}(\beta,\hat{p})-M_{[t],n}^{IPW}(\beta,p^*)-\Delta_{[t],n}(\beta,\hat{p}-p^*)\right|}{1+Cn^{1/2}|\beta_t-\beta_t^*|} \\
&\leq Cn^{1/2}\|\hat{p}_t-p_t^*\|_\infty^2 \frac{1}{n}\sum_{i=1}^n \frac{D_i(t)\sup_{|\beta_t-\beta_t^*|\leq\delta_n}|m(Y_i;\beta_t)|}{p_t^*(X_i)} \\
&= o_p(1),
\end{aligned}$$

by condition (3.1) and Assumption 3.

For the third term (A.3),

$$\begin{aligned}
&\sup_{|\beta_t-\beta_t^*|\leq\delta_n} \frac{n^{1/2}\left|M_{[t],n}^{IPW}(\beta^*,\hat{p})+M_{[t],n}^{IPW}(\beta^*,p^*)-\Delta_{[t],n}(\beta^*,\hat{p}-p^*)\right|}{1+Cn^{1/2}|\beta_t-\beta_t^*|} \\
&\leq Cn^{1/2}\|\hat{p}_t-p_t^*\|_\infty^2 \frac{1}{n}\sum_{i=1}^n \frac{D_i(t)|m(Y_i;\beta_t^*)|}{p_t^*(X_i)} \\
&= o_p(1),
\end{aligned}$$

by condition (3.1) and Assumption 3.

Finally, for the last term (A.4) define

$$\upsilon_{t,i}(\beta_t)=\frac{D_{t,i}|m(Y_i;\beta_t)-m(Y_i;\beta_t^*)|}{p_t^*(X_i)}-\mathbb{E}\left[\frac{D_{t,i}|m(Y_i;\beta_t)-m(Y_i;\beta_t^*)|}{p_t^*(X_i)}\right],$$

and note that

$$\begin{aligned}
&\sup_{|\beta_t-\beta_t^*|\leq\delta_n} \frac{n^{1/2}\left|\Delta_{[t],n}(\beta,\hat{p}-p^*)-\Delta_{[t],n}(\beta^*,\hat{p}-p^*)\right|}{1+Cn^{1/2}|\beta_t-\beta_t^*|} \\
&\leq Cn^{1/2}\|\hat{p}_t-p_t^*\|_\infty \sup_{|\beta_t-\beta_t^*|\leq\delta_n}\left|\frac{1}{n}\sum_{i=1}^n \upsilon_{t,i}(\beta_t)\right| \\
&+C\|\hat{p}_t-p_t^*\|_\infty \sup_{|\beta_t-\beta_t^*|\leq\delta_n} \frac{n^{1/2}\mathbb{E}\left[|m(Y(t),X;\beta_t)-m(Y(t),X;\beta_t^*)|\right]}{1+Cn^{1/2}|\beta_t-\beta_t^*|} \\
&= o_p(1),
\end{aligned}$$

www.manaraa.com

because (assuming $d_m = 1$ or applying the following argument element by element) the class of functions $\mathcal{F}_t = \{\mathbf{1}\{\cdot = t\} |m(\cdot;\beta) - m(\cdot;\beta_t^*)| / p_t^*(\cdot) : |\beta - \beta^*| \leq \delta\}$ is Donsker with finite integrable envelop by Assumption 5 (Theorem 2.10.6 of van der Vaart and Wellner (1996)) and L_2 continuous by Assumptions 2 and 5.

This establishes condition (iii) of Theorem 3.3 in Pakes and Pollard (1989). ■

Proof of Theorem 6 (ASYMPTOTIC LINEAR REPRESENTATION OF EIFE): the proof of this theorem follows the same logic as the proof of Theorem 5. It is applied Theorem 3.3 and Lemma 3.5 in Pakes and Pollard (1989) after setting $\theta = \beta$, $\theta_0 = \beta^*$, $G_n(\theta) = A_n M_n^{EIF}(\beta, \hat{p}, \hat{e})$, $G(\theta) = AM^{EIF}(\beta, p^*, e^*)$, and verifying their five sufficient conditions (i)-(v). Like in the proof of Theorem 5, conditions (i), (ii), (iv) and (v) are already satisfied, thus it only remains to establish the stochastic equicontinuity condition (iii), which is implied by the following condition: for all sequences $\delta_n = o(1)$,

$$\sup_{|\beta_t - \beta_t^*| \leq \delta_n} \frac{n^{1/2} \left| M_{[t],n}^{EIF}(\beta, \hat{p}, \hat{e}) - M_{[t]}^{EIF}(\beta, p^*, e^*(\beta)) - M_{[t],n}^{EIF}(\beta^*, \hat{p}, \hat{e}) \right|}{1 + Cn^{1/2} |\beta_t - \beta_t^*|} = o_p(1),$$

for all $t \in \mathcal{T}$. Now, using the results in Theorem 5, it only remains to show that

$$\sup_{|\beta_t - \beta_t^*| \leq \delta_n} \frac{n^{1/2} \left| \frac{1}{n} \sum_{i=1}^n \left(\hat{e}_t(X_i;\beta_t) - \hat{e}_t(X_i;\beta_t^*) \right) \left(D_{t,i} - \hat{p}_t(X_i) \right) / \hat{p}_t(X_i) \right|}{1 + Cn^{1/2} |\beta_t - \beta_t^*|} = o_p(1).$$

www.manaraa.com

Now, for n large enough,

$$\sup_{|\beta_t-\beta_t^*|\leq\delta_n}\frac{n^{1/2}\left|\frac{1}{n}\sum_{i=1}^n\left(\hat{e}_t\left(X_i;\beta_t\right)-\hat{e}_t\left(X_i;\beta_t^*\right)\right)\left(D_{t,i}-\hat{p}_t\left(X_i\right)\right)/\hat{p}_t\left(X_i\right)\right|}{1+Cn^{1/2}\left|\beta_t-\beta_t^*\right|}$$

$$\leq\sup_{|\beta_t-\beta_t^*|\leq\delta_n,\left\|e_t-e_t^*\right\|_\infty\leq\delta_n}\frac{n^{1/2}\left|\frac{1}{n}\sum_{i=1}^n\left(e_t\left(X_i;\beta\right)-e_t\left(X_i;\beta_t^*\right)\right)\left(D_{t,i}-\hat{p}_t\left(X_i\right)\right)/\hat{p}_t\left(X_i\right)\right|}{1+Cn^{1/2}\left|\beta_t-\beta_t^*\right|}$$

$$\leq\sup_{|\beta_t-\beta_t^*|\leq\delta_n,\left\|e_t-e_t^*\right\|_\infty\leq\delta_n}\frac{n^{1/2}\left|\frac{1}{n}\sum_{i=1}^n\upsilon_{t,i}\right|}{1+Cn^{1/2}\left|\beta_t-\beta_t^*\right|} \tag{A.5}$$

$$+\sup_{|\beta_t-\beta_t^*|\leq\delta_n}\frac{n^{1/2}\left|\frac{1}{n}\sum_{i=1}^n\frac{\partial}{\partial\beta}e_t^*\left(X_i;\beta_t^*\right)\left(\beta_t-\beta_t^*\right)\left(D_{t,i}-\hat{p}_t\left(X_i\right)\right)/\hat{p}_t\left(X_i\right)\right|}{1+Cn^{1/2}\left|\beta_t-\beta_t^*\right|}, \tag{A.6}$$

for some convex linear combination $\tilde{\beta}$ (between β_t and β_t^*), and where

$$\upsilon_{t,i}=\left(\frac{\partial}{\partial\beta}e_t(X_i;\tilde{\beta})-\frac{\partial}{\partial\beta}e_t^*\left(X_i;\beta_t^*\right)\right)\left(\beta_t-\beta_t^*\right)\left(D_{t,i}-\hat{p}_t\left(X_i\right)\right)/\hat{p}_t\left(X_i\right).$$

Next, for the first term (A.5) and for n large enough,

$$\sup_{|\beta_t-\beta_t^*|\leq\delta_n,\left\|e_t-e_t^*\right\|_\infty\leq\delta_n}\frac{n^{1/2}\left|\frac{1}{n}\sum_{i=1}^n\upsilon_{t,i}\right|}{1+Cn^{1/2}\left|\beta_t-\beta_t^*\right|}$$

$$\leq C\sup_{|\beta_t-\beta_t^*|\leq\delta_n,\left\|e_t-e_t^*\right\|_\infty\leq\delta_n}\frac{1}{n}\sum_{i=1}^n\left|\frac{\partial}{\partial\beta}e_t\left(X_i;\beta_t\right)-\frac{\partial}{\partial\beta}e_t^*\left(X_i;\beta\right)\right|$$

$$+C\sup_{|\beta_t-\beta_t^*|\leq\delta_n}\left|\frac{1}{n}\sum_{i=1}^n\left(\frac{\partial}{\partial\beta}e_t^*\left(X_i;\beta_t\right)-\frac{\partial}{\partial\beta}e_t^*\left(X_i;\beta_t^*\right)\right)\frac{D_{t,i}-p_t^*\left(X_i\right)}{p_t^*\left(X_i\right)}\right|$$

$$+C\frac{1}{n}\sum_{i=1}^n\sup_{|\beta_t-\beta_t^*|\leq\delta_n}\left|\frac{\partial}{\partial\beta}e_t^*\left(X_i;\beta_t\right)\right|\left|\frac{D_{t,i}-\hat{p}_t\left(X_i\right)}{\hat{p}_t\left(X_i\right)}-\frac{D_{t,i}-p_t^*\left(X_i\right)}{p_t^*\left(X_i\right)}\right|$$

$$=o_p\left(1\right),$$

because the first term is $o_p\left(1\right)$ by Assumption 6(b), the second term is $o_p\left(1\right)$ because (assuming $d_m=1$ or applying the argument element by element) the class of functions $\mathcal{F}_t=\left\{\left(\partial_\beta e_t^*\left(\cdot;\beta\right)-\partial_\beta e_t^*\left(\cdot;\beta_t^*\right)\right)\left(\mathbf{1}\left\{\cdot=t\right\}-\ p_t^*\left(\cdot\right)\right)/p_t^*\left(\cdot\right):\left|\beta-\beta_t^*\right|\leq\delta\right\}$ is Glivenko-

www.manaraa.com

Cantelli for some $\delta > 0$ by Assumption 6(a) (van der Vaart and Wellner (2000)), and the third term is $o_p(1)$ by Assumption 6(a).

The second term (A.6) is

$$\begin{aligned}
&\sup_{|\beta_t - \beta_t^*| \leq \delta_n} \frac{n^{1/2} \left| \frac{1}{n} \sum_{i=1}^n \frac{\partial}{\partial \beta} e_t^* (X_i; \beta_t^*) (\beta_t - \beta_t^*) (D_{t,i} - \hat{p}_t (X_i)) / \hat{p}_t (X_i) \right|}{1 + C n^{1/2} |\beta_t - \beta_t^*|} \\
&\leq \left| \frac{1}{n} \sum_{i=1}^n \frac{\partial}{\partial \beta} e_t^* (X_i; \beta_t^*) \frac{D_{t,i} - p_t^* (X_i)}{p_t^* (X_i)} \right| \\
&\qquad + \frac{1}{n} \sum_{i=1}^n \left| \frac{\partial}{\partial \beta} e_t^* (X_i; \beta_t^*) \right| \left| \frac{D_{t,i} - \hat{p}_t (X_i)}{\hat{p}_t (X_i)} - \frac{D_{t,i} - p_t^* (X_i)}{p_t^* (X_i)} \right| \\
&= o_p (1),
\end{aligned}$$

by Assumption 6(a).

This establishes condition (iii) of Theorem 3.3 in Pakes and Pollard (1989). ■

Proof of Theorem 8 (CONSISTENT ESTIMATOR OF V^*): first it is useful to establish the following two results. For all sequences $\delta_n = o(1)$ and for all $t \in \mathcal{T}$,

$$\frac{1}{n} \sum_{i=1}^n \left| m(Y_i, T_i, X_i; \hat{\beta}, \hat{p}) - m (Y_i, T_i, X_i; \beta^*, p^*) \right|^2 = o_p (1) \tag{A.7}$$

and

$$\frac{1}{n} \sum_{i=1}^n \left| \alpha(T_i, X_i; \hat{p}, \hat{e}(\hat{\beta})) - \alpha (T_i, X_i; p^*, e^* (\beta^*)) \right|^2 = o_p (1). \tag{A.8}$$

www.manaraa.com

The first result (A.7) follows because for n large enough and for all $t \in \mathcal{T}$,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\left|\frac{D_{t,i}m(Y_i;\hat{\beta}_t)}{\hat{p}_t\left(X_i\right)}-\frac{D_{t,i}m\left(Y_i;\beta_t^*\right)}{p_t^*\left(X_i\right)}\right|^2 \\
&\leq \left\|\hat{p}_t-p_t^*\right\|_\infty^2\frac{C}{n}\sum_{i=1}^{n}\frac{D_{t,i}\sup_{|\beta-\beta_t^*|\leq\delta_n}\left|m\left(Y_i;\beta\right)\right|^2}{p_t^*\left(X_i\right)} \\
&\qquad +\frac{C}{n}\sum_{i=1}^{n}\frac{D_{t,i}}{p_t^*\left(X_i\right)}\left|m(Y_i;\hat{\beta}_t)-m\left(Y_i;\beta_t^*\right)\right|^2 \\
&= o_p\left(1\right),
\end{aligned}$$

by the same arguments and assumptions used in Theorem 5 and an application of Theorem 2.10.14 of van der Vaart and Wellner (1996). The second result (A.8) follows because for n large enough and for all $t \in \mathcal{T}$,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{e}_t(X_i;\hat{\beta}_t)}{\hat{p}_t\left(X_i\right)}\left(D_{t,i}-\hat{p}_t\left(X_i\right)\right)-\frac{e_t^*\left(X_i;\beta_t^*\right)}{p_t^*\left(X_i\right)}\left(D_{t,i}-p_t^*\left(X_i\right)\right)\right|^2 \\
&\leq \frac{C}{n}\sum_{i=1}^{n}\left|e_t^*(X_i;\hat{\beta}_t)-e_t^*\left(X_i;\beta_t^*\right)\right|^2+o_p\left(1\right) \\
&\leq \frac{C}{n}\sum_{i=1}^{n}\sup_{|\beta_t-\beta_t^*|\leq\delta_n}\left|\frac{\partial}{\partial\beta}e_t^*\left(X_i;\beta_t\right)\right|^2\left|\hat{\beta}_t-\beta_t^*\right|+o_p\left(1\right)=o_p\left(1\right).
\end{aligned}$$

Now, define

$$V_n=\frac{1}{n}\sum_{i=1}^{n}\psi\left(Y_i,T_i,\beta^*,p^*,e^*\left(\beta^*\right)\right)\psi\left(Y_i,T_i,\beta^*,p^*,e^*\left(\beta^*\right)\right)',$$

and notice that $V_n - V_* = o_p\left(1\right)$. Next, using Holder's Inequality,

$$\left|\hat{V}_n-V_*\right|\leq\left|\hat{V}_n-V_n\right|+\left|V_n-V_*\right|\leq R_{1,n}+R_{2,n}+R_{3,n}+R_{4,n}+R_{5,n}+o_p\left(1\right),$$

www.manaraa.com

where

$$
\begin{aligned}
R_{1,n} &= \frac{1}{n}\sum_{i=1}^{n}\left|m(Y_i,T_i,X_i;\hat{\beta},\hat{p})-m\left(Y_i,T_i;\beta^*,p^*\right)\right|^2,\\
R_{2,n} &= \frac{1}{n}\sum_{i=1}^{n}\left|\alpha(T_i,X_i;\hat{p},\hat{e}(\hat{\beta}))-\alpha\left(Y_i,T_i;p^*,e^*\left(\beta^*\right)\right)\right|^2,\\
R_{3,n} &= 2R_{1,n}^{1/2}R_{2,n}^{1/2},\\
R_{4,n} &= 2R_{1,n}^{1/2}\left(\frac{1}{n}\sum_{i=1}^{n}\left|\psi\left(Y_i,T_i,\beta^*,p^*,e^*\left(\beta^*\right)\right)\right|^2\right)^{1/2},\\
R_{5,n} &= 2R_{2,n}^{1/2}\left(\frac{1}{n}\sum_{i=1}^{n}\left|\psi\left(Y_i,T_i,\beta^*,p^*,e^*\left(\beta^*\right)\right)\right|^2\right)^{1/2},
\end{aligned}
$$

and using (A.7) and (A.8) the result follows. ■

Proof of Theorem 9 (CONSISTENT ESTIMATOR OF Γ^*): follows directly by the same arguments given in the proof of Theorem 6. ■

Proof of Theorem 10 (NONPARAMETRIC ESTIMATION): first, for power series and splines $\zeta(K)=K^{\eta}$, with $\eta=1$ and $\eta=1/2$, respectively, and using Assumption 7 (which for these cases implies Assumption B-1 in Appendix B) $\alpha=s/d_x$ (Newey (1997)). Now Theorem B-1 in Appendix B implies

$$
n^{1/4}\sup_{x\in\mathcal{X}}\left|\hat{p}(x)-p^*(x)\right|=n^{1/4}O_p\left(K^{\eta}K^{1/2}n^{-1/2}+K^{\eta}K^{1/2}K^{-s/d_x}\right)=o_p(1),
$$

under the assumptions of the theorem and therefore condition (5.1) in Theorem 5 holds.

Next, consider condition (5.2) in Theorem 5. It is enough to show the result for a

www.manaraa.com

typical t-th component of the vector. Thus,

$$\sqrt{n}\left|M_{[t],n}^{IPW}\left(\beta_t^*,\hat{p}_t\right)-M_{[t],n}^{EIF}\left(\beta_t^*,p_t^*,e_t^*\left(\beta_t^*\right)\right)\right|$$

$$\leq\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\frac{D_{t,i}m\left(Y_i;\beta_t^*\right)}{\hat{p}_t\left(X_i\right)}-\frac{D_{t,i}m\left(Y_i;\beta_t^*\right)}{p_t^*\left(X_i\right)}+\frac{D_{t,i}m\left(Y_i;\beta_t^*\right)}{p_t^*\left(X_i\right)^2}\left(\hat{p}_t\left(X_i\right)-p_t^*\left(X_i\right)\right)\right\}\right| \tag{A.9}$$

$$+\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{-\frac{D_{t,i}m\left(Y_i;\beta_t^*\right)}{p_t^*\left(X_i\right)^2}\left(\hat{p}_t\left(X_i\right)-p_t^*\left(X_i\right)\right)+\frac{e_t^*\left(X_i;\beta_t^*\right)}{p_t^*\left(X_i\right)}\left(\hat{p}_t\left(X_i\right)-p_t^*\left(X_i\right)\right)\right\}\right| \tag{A.10}$$

$$+\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{-\frac{e_t^*\left(X_i;\beta_t^*\right)}{p_t^*\left(X_i\right)}\left(\hat{p}_t\left(X_i\right)-p_t^*\left(X_i\right)\right)+\frac{e_t^*\left(X_i;\beta_t^*\right)}{p_t^*\left(X_i\right)}\left(D_{t,i}-p_t^*\left(X_i\right)\right)\right\}\right|. \tag{A.11}$$

The bound of term (A.9) is given by (for n large enough)

$$\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\frac{D_{t,i}m\left(Y_i;\beta_t^*\right)}{\hat{p}_t\left(X_i\right)}-\frac{D_{t,i}m\left(Y_i;\beta_t^*\right)}{p_t^*\left(X_i\right)}+\frac{D_{t,i}m\left(Y_i;\beta_t^*\right)}{p_t^*\left(X_i\right)^2}\left(\hat{p}_t\left(X_i\right)-p_t^*\left(X_i\right)\right)\right\}\right|$$

$$\leq C\sqrt{n}\left\|\hat{p}_t-p_t^*\right\|_\infty^2\frac{1}{n}\sum_{i=1}^{n}\frac{D_{t,i}\left|m\left(Y_i;\beta_t^*\right)\right|}{p_t^*\left(X_i\right)}$$

$$=\sqrt{n}O_p\left(\left(K^\eta K^{1/2}n^{-1/2}+K^\eta K^{1/2}K^{-s/d_x}\right)^2\right).$$

The bound of term (A.10) is given by

$$\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{-\frac{D_{t,i}m\left(Y_i;\beta_t^*\right)}{p_t^*\left(X_i\right)^2}\left(\hat{p}_t\left(X_i\right)-p_t^*\left(X_i\right)\right)+\frac{e_t^*\left(X_i;\beta_t^*\right)}{p_t^*\left(X_i\right)}\left(\hat{p}_t\left(X_i\right)-p_t^*\left(X_i\right)\right)\right\}\right|$$

$$\leq\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{e_t^*\left(X_i;\beta_t^*\right)}{p_t^*\left(X_i\right)}-\frac{D_{t,i}m\left(Y_i;\beta_t^*\right)}{p_t^*\left(X_i\right)^2}\right)\left(\hat{p}_t\left(X_i\right)-p_{K,t}^0\left(X_i\right)\right)\right| \tag{A.12}$$

$$+\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{e_t^*\left(X_i;\beta_t^*\right)}{p_t^*\left(X_i\right)}-\frac{D_{t,i}m\left(Y_i;\beta_t^*\right)}{p_t^*\left(X_i\right)^2}\right)\left(p_{K,t}^0\left(X_i\right)-p_t^*\left(X_i\right)\right)\right|, \tag{A.13}$$

using the notation introduced in Appendix B. Now, to obtain a bound on the term (A.12), first notice that by a second order Taylor expansion and using the results in

www.manaraa.com

Appendix B it is obtained for some $\tilde{\gamma}_K$ such that $|\tilde{\gamma}_K - \gamma_K^0| \leq |\hat{\gamma}_K - \gamma_K^0|$ and n large enough,

$$\begin{aligned}
&\hat{p}_t(x) - p_{K,t}^0(x) \\
&= \left[\dot{\mathbf{L}}_t\left(g_{-0}\left(x, \gamma_K^0\right)\right) \otimes R_K(x)'\right]\left(\hat{\gamma}_K - \gamma_K^0\right) \\
&\qquad + \frac{1}{2}\left(\hat{\gamma}_K - \gamma_K^0\right)'\left[\mathbf{H}\left(x, \tilde{\gamma}_K\right) \otimes R_K(x) R_K(x)'\right]\left(\hat{\gamma}_K - \gamma_K^0\right) \\
&\leq \left[\dot{\mathbf{L}}_t\left(g_{-0}\left(x, \gamma_K^0\right)\right) \otimes R_K(x)'\right]\left(\hat{\gamma}_K - \gamma_K^0\right) \\
&\qquad + C\left(\hat{\gamma}_K - \gamma_K^0\right)'\left[\mathbf{I}_J \otimes R_K(x) R_K(x)'\right]\left(\hat{\gamma}_K - \gamma_K^0\right),
\end{aligned}$$

which implies that

$$\begin{aligned}
&\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n\left(\frac{e_t^*\left(X_i;\beta_t^*\right)}{p_t^*(X_i)} - \frac{D_{t,i}m\left(Y_i;\beta_t^*\right)}{p_t^*(X_i)^2}\right)\left(\hat{p}_t(X_i) - p_{K,t}^0(X_i)\right)\right| \\
&\leq \left|\hat{\gamma}_K - \gamma_K^0\right|\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n\left(\frac{e_t^*\left(X_i;\beta_t^*\right)}{p_t^*(X_i)} - \frac{D_{t,i}m\left(Y_i;\beta_t^*\right)}{p_t^*(X_i)^2}\right)\left[\dot{\mathbf{L}}_t\left(g_{-0}\left(X_i, \gamma_K^0\right)\right) \otimes R_K(X_i)'\right]\right| \\
&\qquad + \left|\hat{\gamma}_K - \gamma_K^0\right|^2\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n\left(\frac{e_t^*\left(X_i;\beta_t^*\right)}{p_t^*(X_i)} - \frac{D_{t,i}m\left(Y_i;\beta_t^*\right)}{p_t^*(X_i)^2}\right)\left[\mathbf{I}_J \otimes R_K(X_i) R_K(X_i)'\right]\right| \\
&= O_p\left(K^{1/2}n^{-1/2} + K^{1/2}K^{-s/d_x}\right) O\left(K^{1/2}\right),
\end{aligned}$$

where the bound follows because the random variables inside the sums are mean zero and variance bounded by K.

Now, for the term (A.13)

$$\begin{aligned}
\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n\left(\frac{e_t^*\left(X_i;\beta_t^*\right)}{p_t^*(X_i)} - \frac{D_{t,i}m\left(Y_i;\beta_t^*\right)}{p_t^*(X_i)^2}\right)\left(p_{K,t}^0(X_i) - p_t^*(X_i)\right)\right| &= O_p\left(K^{-s/(2d_x)}\right) \\
&= o_p(1).
\end{aligned}$$

www.manaraa.com

Finally, the bound of term (A.11) is given by

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{-\frac{e_t^*\left(X_i;\beta_t^*\right)}{p_t^*\left(X_i\right)}\left(\hat{p}_t\left(X_i\right)-p_t^*\left(X_i\right)\right)+\frac{e_t\left(X_i;\beta_t^*\right)}{p_t^*\left(X_i\right)}\left(D_{t,i}-p_t^*\left(X_i\right)\right)\right\}$$
$$=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{e_t^*\left(X_i;\beta_t^*\right)}{p_t^*\left(X_i\right)}-R_K\left(X_i\right)'\theta\right)\left(D_{t,i}-\hat{p}_t\left(X_i\right)\right),$$

using the first order condition for MLSE, which implies that

$$\sum_{i=1}^{n}(D_{t,i}-\hat{p}_t\left(X_i\right))R_K\left(X_i\right)=\mathbf{0},$$

and where $\theta\in\mathbb{R}^K$ is any vector. Now, by choosing θ appropriately, verify for n large enough that

$$\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{e_t^*\left(X_i;\beta_t^*\right)}{p_t^*\left(X_i\right)}-R_K\left(X_i\right)'\theta\right)\left(D_{t,i}-\hat{p}_t\left(X_i\right)\right)\right|$$
$$\leq\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{e_t^*\left(X_i;\beta_t^*\right)}{p_t^*\left(X_i\right)}-R_K\left(X_i\right)'\theta\right)\left(D_{t,i}-p_t^*\left(X_i\right)\right)\right|$$
$$+\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{e_t^*\left(X_i;\beta_t^*\right)}{p_t^*\left(X_i\right)}-R_K\left(X_i\right)'\theta\right)\left(p_t^*\left(X_i\right)-\hat{p}_t\left(X_i\right)\right)\right|$$
$$\leq O_p\left(K^{-s/(2d_x)}\right)+n^{1/2}O\left(K^{-s/d_x}\right)O_p\left(K^{\eta}K^{1/2}n^{-1/2}+K^{\eta}K^{1/2}K^{-s/d_x}\right).$$

Using the bounds derived and under the assumptions of Theorem 10,

$$\left|M_n^{IPW}\left(\beta^*,\hat{p}\right)-M_n^{EIF}\left(\beta^*,p^*,e^*\left(\beta^*\right)\right)\right|=o_p\left(n^{-1/2}\right),$$

which verifies condition (5.2) in Theorem 5 as desired.

Next, consider Theorem 6. Conditions (6.1) and (6.2) follow directly from the previous calculations and the first part of Proposition A1 in Chen, Hong, and Tamer

www.manaraa.com

(2005), respectively. It remains only to show condition (6.3) in Theorem 6. From Newey (1997) it follows immediately that

$$n^{1/4}\sup_{x\in\mathcal{X}}\left|\hat{e}\left(x;\beta^{*}\right)-e^{*}\left(x;\beta^{*}\right)\right|=n^{1/4}O_{p}\left(K^{\eta}K^{1/2}n^{-1/2}+K^{\eta}K^{-s/d_{x}}\right)=o_{p}\left(1\right).$$

Now, to establish the final condition is enough to show the result for the typical t-th component. From the previous calculations

$$\begin{aligned}&\frac{1}{\sqrt{n}}\sum\nolimits_{i=1}^{n}\left\{\frac{D_{t,i}m\left(Y_{i};\beta_{t}^{*}\right)}{\hat{p}_{t}\left(X_{i}\right)}-\frac{D_{t,i}m\left(Y_{i};\beta_{t}^{*}\right)}{p_{t}^{*}\left(X_{i}\right)}\right\}\\&=-\frac{1}{\sqrt{n}}\sum\nolimits_{i=1}^{n}\frac{D_{t,i}m\left(Y_{i};\beta_{t}^{*}\right)}{p_{t}^{*}\left(X_{i}\right)^{2}}\left(\hat{p}_{t}\left(X_{i}\right)-p_{t}^{*}\left(X_{i}\right)\right)+o_{p}\left(1\right),\end{aligned}$$

and using the identity

$$\frac{\hat{a}}{\hat{b}}=\frac{a}{b}+\frac{1}{b}\left(\hat{a}-a\right)-\frac{a}{b^{2}}\left(\hat{b}-b\right)+\frac{a}{b^{2}\hat{b}}\left(\hat{b}-b\right)^{2}-\frac{1}{b\hat{b}}\left(\hat{a}-a\right)\left(\hat{b}-b\right),$$

$$\begin{aligned}&\frac{1}{\sqrt{n}}\sum\nolimits_{i=1}^{n}\left\{\frac{D_{t,i}\hat{e}_{t}\left(X_{i};\beta_{t}^{*}\right)}{\hat{p}_{t}\left(X_{i}\right)}-\frac{D_{t,i}e_{t}^{*}\left(X_{i};\beta_{t}^{*}\right)}{p_{t}^{*}\left(X_{i}\right)}\right\}\\&=\frac{1}{\sqrt{n}}\sum\nolimits_{i=1}^{n}\frac{D_{t,i}\left(\hat{e}_{t}\left(X_{i};\beta_{t}^{*}\right)-e_{t}^{*}\left(X_{i};\beta_{t}^{*}\right)\right)}{p_{t}^{*}\left(X_{i}\right)}\\&\quad-\frac{1}{\sqrt{n}}\sum\nolimits_{i=1}^{n}\frac{D_{t,i}e_{t}^{*}\left(X_{i};\beta_{t}^{*}\right)}{p_{t}^{*}\left(X_{i}\right)^{2}}\left(\hat{p}_{t}\left(X_{i}\right)-p_{t}^{*}\left(X_{i}\right)\right)+o_{p}\left(1\right).\end{aligned}$$

Putting these results together,

$$\begin{aligned}&\sqrt{n}\left|M_{[t],n}^{EIF}\left(\beta_{t}^{*},\hat{p}_{t},\hat{e}_{t}\left(\beta_{t}^{*}\right)\right)-M_{[t],n}^{EIF}\left(\beta_{t}^{*},p_{t}^{*},e_{t}^{*}\left(\beta_{t}^{*}\right)\right)\right|\\&\leq\left|\frac{1}{\sqrt{n}}\sum\nolimits_{i=1}^{n}\frac{D_{t,i}\left(m\left(Y_{i};\beta_{t}^{*}\right)-e_{t}^{*}\left(X_{i};\beta_{t}^{*}\right)\right)}{p_{t}^{*}\left(X_{i}\right)^{2}}\left(\hat{p}_{t}\left(X_{i}\right)-p_{t}^{*}\left(X_{i}\right)\right)\right|\\&\quad+\left|\frac{1}{\sqrt{n}}\sum\nolimits_{i=1}^{n}\frac{D_{t,i}-p_{t}^{*}\left(X_{i}\right)}{p_{t}^{*}\left(X_{i}\right)}\left(\hat{e}_{t}\left(X_{i};\beta_{t}^{*}\right)-e_{t}^{*}\left(X_{i};\beta_{t}^{*}\right)\right)\right|\\&\quad+\left|\frac{1}{\sqrt{n}}\sum\nolimits_{i=1}^{n}\left(\hat{e}_{t}\left(X_{i};\beta_{t}^{*}\right)-e_{t}^{*}\left(X_{i};\beta_{t}^{*}\right)\right)\right|.\end{aligned}$$

www.manaraa.com

Finally, observe that by the same arguments as those used for term (A.10) above,

$$\frac{1}{\sqrt{n}}\sum\nolimits_{i=1}^{n}\frac{D_{t,i}\left(m\left(Y_i;\beta_t^*\right)-e_t^*\left(X_i;\beta_t^*\right)\right)}{p_t^*\left(X_i\right)^2}\left(\hat{p}_t\left(X_i\right)-p_t^*\left(X_i\right)\right)=o_p\left(1\right),$$

and by analogous arguments, but for the case of series (linear sieves) it is possible to verify that

$$\frac{1}{\sqrt{n}}\sum\nolimits_{i=1}^{n}\frac{D_{t,i}-p_t^*\left(X_i\right)}{p_t^*\left(X_i\right)}\left(\hat{e}_t\left(X_i;\beta_t^*\right)-e_t^*\left(X_i;\beta_t^*\right)\right)=o_p\left(1\right),$$

and

$$\frac{1}{\sqrt{n}}\sum\nolimits_{i=1}^{n}\left(\hat{e}_t\left(X_i;\beta_t^*\right)-e_t^*\left(X_i;\beta_t^*\right)\right)=o_p\left(1\right),$$

under the assumptions of this theorem. Therefore

$$\sqrt{n}\left|M_n^{EIF}\left(\beta^*,\hat{p},\hat{e}\left(\beta^*\right)\right)-M_n^{EIF}\left(\beta^*,p^*,e^*\left(\beta^*\right)\right)\right|=o_p\left(1\right),$$

which gives condition (6.3) in Theorem 6 as needed. ■

www.manaraa.com

Appendix B

Multinomial Logistic Series Estimator

This appendix derives uniform rates of convergence for the non-linear sieve estimator proposed for the estimation of the GPS. The results presented here generalize those in Hirano, Imbens, and Ridder (2003) by allowing for arbitrary number of outcomes, arbitrary choice of approximating basis, and less stringent requirements in terms of smoothness of the underlying conditional expectation.

It is important to introduce some normalizations and notation. Under some conditions imposed below and by choosing an appropriate non-singular linear transformation assume without loss of generality that $\mathbb{E}\left[R_K(X) R_K(X)'\right] = \mathbf{I}_K$, where $\mathbf{I}_K$ is the $(K \times K)$ identity matrix (see Newey (1997) for details). Let $\zeta(K) = \sup_{x \in \mathcal{X}} |R_K(x)|$, and observe that in general this bound will depend on the approximating func-

www.manaraa.com

tions chosen. To reduce notational burden this appendix uses the same number of approximating functions for each conditional probability, a feature that may be relaxed at the expense of only additional notation. To deal with all the relevant probabilities simultaneously define $p_{-0}(X) = (p_1(X), \cdots, p_J(X))' \in \mathbb{R}^J$, $\gamma_{-0,K} = \left(\gamma'_{K,1}, \cdots, \gamma'_{K,J}\right)' \in \mathbb{R}^{JK}$, and $g_{-0}(X, \gamma_K) = [R_K(X)'\gamma_{K,1}, \cdots, R_K(X)'\gamma_{K,J}]' \in \mathbb{R}^J$. Recall that $p_0^*(X) = 1 - \sum_{j=1}^J p_j^*(X)$.

Next, define for a vector $z \in \mathbb{R}^J$, $z = [z_1, \cdots, z_J]'$, the functions $L_t : \mathbb{R}^J \to \mathbb{R}$ and $L_t^{-1} : \mathbb{R}^J \to \mathbb{R}$, for all $t = 1, 2, \cdots, J$,

$$L_t(z) = \frac{\exp\{z_t\}}{1 + \sum_{j=1}^J \exp\{z_j\}}, \qquad \text{and} \qquad L_t^{-1}(z) = \log\left\{\frac{z_t}{1 - \sum_{j=1}^J z_j}\right\}.$$

and set $L_0(z) = 1 - \sum_{j=1}^J L_j(z)$. The gradient of $L_t : \mathbb{R}^J \to \mathbb{R}$ is given by

$$\dot{L}_t(z) = [-L_t(z) L_1(z), \ \cdots, \ -L_t(z) L_{t-1}(z), \\ L_t(z)(1 - L_t(z)), \ -L_t(z) L_{t+1}(z), \ \cdots, -L_t(z) L_J(z)]',$$

and observe that $\sup_z |\dot{L}_t(z)| < C$ since $|L_t(z) L_j(z)| < 1$ and $L_t(z)(1 - L_t(z)) < 1/4$. Also define the vector-valued functions $\mathbf{L}(z) = [L_1(z), \cdots, L_J(z)]'$ and $\mathbf{L}^{-1}(z) = [L_1^{-1}(z), \ \cdots, \ L_J^{-1}(z)]'$ and observe that the function $\mathbf{L}(\cdot)$ is differentiable with gradient (matrix) $\dot{\mathbf{L}}(z) = [\dot{L}_1(z), \ \cdots, \ \dot{L}_J(z)] \in \mathbb{R}^{J \times J}$ and notice that $\sup_z |\dot{\mathbf{L}}(z)| < C$, for some constant C that only depends on J. With this notation, $p\left(X; \gamma_{t,K}\right) = L_t(g_{-0}(X, \gamma_K))$ for $t \in \mathcal{T}$ (recall $\gamma_{K,0} = \mathbf{0}_K$ for identification purposes).

The multinomial logistic log-likelihood is given by

$$\ell_n(\gamma_K) = \sum_{i=1}^n \sum_{t=0}^J D_{t,i} \log\left(L_t\left(g_{-0}(X_i, \gamma_K)\right)\right),$$

www.manaraa.com

with solution $\hat{\gamma}_K = \arg\max_{\gamma_K} \ell_n(\gamma_K)$ and estimated probabilities given by $\hat{p}_t(X) = L_t(g_{-0}(X_i, \hat{\gamma}_K))$ for all $t \in \mathcal{T}$. Verify that

$$\begin{aligned}
\frac{\partial}{\partial \gamma_{K,t}} \ell_n(\gamma_K) &= \sum_{i=1}^{n} \left[D_{t,i} - L_t(g_{-0}(X_i, \gamma_K))\right] R_K(X_i), \\
\frac{\partial^2}{\partial \gamma_{K,t} \partial \gamma'_{K,l}} \ell_n(\gamma_K) &= -\sum_{i=1}^{n} L_l(g_{-0}(X_i, \gamma_K)) \\
&\qquad \cdot \left[\mathbf{1}\{t = l\} - L_t(g_{-0}(X_i, \gamma_K))\right] R_K(X_i) R_K(X_i)',
\end{aligned}$$

for $t = 1, 2, \cdots, J$, $l = 1, 2, \cdots, J$, and in matrix notation

$$\begin{aligned}
\frac{\partial}{\partial \gamma_K} \ell_n(\gamma_K) &= \sum_{i=1}^{n} \left[\mathbf{D}_i - \mathbf{L}(g_{-0}(X_i, \gamma_K))\right] \otimes R_K(X_i), \\
\frac{\partial^2}{\partial \gamma_K \partial \gamma'_K} \ell_n(\gamma_K) &= -\sum_{i=1}^{n} \mathbf{H}(X_i, \gamma_K) \otimes R_K(X_i) R_K(X_i)',
\end{aligned}$$

where $\mathbf{D}_i = (D_{1,i}, D_{2,i} \cdots, D_{J,i})'$ and

$$\mathbf{H}(X_i, \gamma_K) = \operatorname{diag}(\mathbf{L}(g_{-0}(X_i, \gamma_K))) - \mathbf{L}(g_{-0}(X_i, \gamma_K)) \mathbf{L}(g_{-0}(X_i, \gamma_K))'.$$

To derive the uniform rates of convergence, the followings conditions is sufficient:

Assumption B-1. (a) The smallest eigenvalue of $\mathbb{E}\left[R_K(X) R_K(X)'\right]$ is bounded away from zero uniformly in K; (b) there is a sequence of constants $\zeta(K)$ satisfying $\sup_{x \in \mathcal{X}} |R_K(x)| \leq \zeta(K)$, for $K = K(n) \to \infty$ and $\zeta(K) K^{1/2} n^{-1/2} \to 0$, as $n \to \infty$; and (c) for all $t \in \mathcal{T}$ there exists $\gamma^0_{t,K} \in \mathbb{R}^K$ and $\alpha > 0$ such that

$$\sup_{x \in \mathcal{X}} \left| \log\left(\frac{p^*_t(x)}{p^*_0(x)}\right) - R_K(x)' \gamma^0_{t,K} \right| = O\left(K^{-\alpha}\right),$$

and $\zeta(K) K^{1/2} K^{-\alpha} \to 0$.

www.manaraa.com

Assumption B-1 is automatically satisfied in the case of power series or splines if the GPS is smooth enough. Parts (a) and (b) are standard in the literature (Newey (1997)), while Part (c) is slightly stronger than its counterpart for linear series because it imposes a lower bound in $\alpha > 0$. Part (c) guarantees the existence of an approximating sequence that can approximate the function uniformly well. For notational simplicity, denote such sequence by $p^0_{t,K}(X) = L_t(g_{-0}(X, \gamma^0_K))$, for all $t \in \mathcal{T}$, and define $p^0_K = [p^0_{0,K}, \cdots, p^0_{J,K}]'$.

The following theorem provides the uniform rate of convergence for the MLSE.

Theorem B-1. (Uniform Rate of Convergence of MLSE) If Assumptions 1(b) and B-1 hold, then

(i) $\|p^0_K - p^*\|_\infty = O(K^{-\alpha})$,

(ii) $|\hat{\gamma}_K - \gamma^0_K| = O_p\left(K^{1/2}n^{-1/2} + K^{1/2}K^{-\alpha}\right)$,

and consequently $\|\hat{p} - p^*\|_\infty = O_p(\zeta(K) K^{1/2}n^{-1/2} + \zeta(K) K^{1/2}K^{-\alpha})$.

Proof of Theorem B-1 (Uniform Rate of Convergence of MLSE):

First, Assumption B-1(c) implies that

$$\sup_{x\in\mathcal{X}} \left|\mathbf{L}^{-1}\left(p^*_{-0}(x)\right) - g_{-0}\left(x, \gamma^0_K\right)\right| = O\left(K^{-\alpha}\right).$$

Since the mapping $\mathbf{L}(\cdot)$ is differentiable with $\sup_z |\dot{\mathbf{L}}(z)| < C$, an application of the mean value theorem gives

$$\sup_{x\in\mathcal{X}} \left|p^*_{-0}(x) - \mathbf{L}\left(g_{-0}\left(x, \gamma^0_K\right)\right)\right| \leq C \sup_{x\in\mathcal{X}} \left|\mathbf{L}^{-1}\left(p^*_{-0}(x)\right) - g_{-0}\left(x, \gamma^0_K\right)\right|,$$

www.manaraa.com

and since $p_0^*(x) = 1 - \sum_{j=1}^{J} p_j^*(x)$ and $L_0(g_{-0}(x, \gamma_K^0)) = 1 - \sum_{j=1}^{J} L_j(g_{-0}(x, \gamma_K^0))$ part (i) follows directly.

For part (ii), first recall that $L_t(g_{-0}(x, \gamma)) > 0$, for all $t = 1, 2, \cdots, J$, and $\sum_{t=1}^{J} L_t(g_{-0}(x, \gamma)) < 1$. The special structure of the matrix $\mathbf{H}(x, \gamma)$ and Theorem 1 in Tanabe and Sagae (1992) shows that $\mathbf{H}(x, \gamma)$ is symmetric positive definite with $0 < \lambda_{\min}(\mathbf{H}(x, \gamma)) \leq \lambda_{\max}(\mathbf{H}(x, \gamma)) < 1$, which implies that

$$\mathbf{H}(x, \gamma) \geq \lambda_{\min}(\mathbf{H}(x, \gamma)) \mathbf{I}_J \geq \det(\mathbf{H}(x, \gamma)).$$

These results and the exact Cholesky decomposition of $\mathbf{H}(x, \gamma)$ gives

$$\inf_{x \in \mathcal{X}} \mathbf{H}(x, \gamma) \geq \inf_{x \in \mathcal{X}} \prod_{t=0}^{J} L_t(g_{-0}(x, \gamma)) \mathbf{I}_J,$$

in a positive semidefinite sense.

Now, let $\hat{\Omega}_K = n^{-1} \sum_{i=1}^{n} R_K(X_i) R_K(X_i)'$, and observe that (Newey (1997)) $|\hat{\Omega}_K - \mathbf{I}_K| = O_p(\zeta(K) K^{1/2} n^{-1/2})$. Define the event $\mathcal{A}_n = \{\lambda_{\min}(\hat{\Omega}_K) > 1/2\}$ and by Assumption B-1(b) $O_p\left(\zeta(K) K^{1/2} n^{-1/2}\right) = o_p(1)$, which implies $\mathbb{P}[\mathcal{A}_n] \to 1$.

www.manaraa.com

Next,

$$\begin{aligned}
\mathbb{E}\left[\left|\frac{1}{n}\frac{\partial}{\partial\gamma}\ell_n\left(\gamma_K^0\right)\right|\right] &= \mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n}\left[\mathbf{D}_i - \mathbf{L}\left(g_{-0}\left(X_i,\gamma_K^0\right)\right)\right]\otimes R_K\left(X_i\right)\right|\right] \\
&\leq \left(\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n}\left[\mathbf{D}_i - p_{-0}^*\left(X_i\right)\right]\otimes R_K\left(X_i\right)\right|^2\right]\right)^{1/2} \\
&\quad + \mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n}\left[p_{-0}^*\left(X_i\right) - \mathbf{L}\left(g_{-0}\left(X_i,\gamma_K^0\right)\right)\right]\otimes R_K\left(X_i\right)\right|\right] \\
&\leq C\left(\frac{1}{n}\mathbb{E}\left[\left|\left[\mathbf{D}_i - p_{-0}^*\left(X_i\right)\right]\otimes R_K\left(X_i\right)\right|^2\right]\right)^{1/2} \\
&\quad + C\sup_{x\in\mathcal{X}}\left|p_{-0}^*\left(x\right) - \mathbf{L}\left(g_{-0}\left(x,\gamma_K^0\right)\right)\right|\mathbb{E}\left[\left|R_K\left(X\right)\right|\right] \\
&= O\left(K^{1/2}n^{-1/2} + K^{1/2}K^{-\alpha}\right),
\end{aligned}$$

and by Markov's Inequality

$$\left|\frac{1}{n}\frac{\partial}{\partial\gamma}\ell_n\left(\gamma_K^0\right)\right| = O_p\left(K^{1/2}n^{-1/2} + K^{1/2}K^{-\alpha}\right),$$

which implies that for any fixed constant $\varsigma > 0$ the probability of the event

$$\mathcal{B}_n\left(\varsigma\right) = \left\{\left|\frac{1}{n}\frac{\partial}{\partial\gamma}\ell_n\left(\gamma_K^0\right)\right| < \varsigma\left(K^{1/2}n^{-1/2} + K^{-\alpha+1/2}\right)\right\}$$

approaches one, i.e., $\mathbb{P}\left[\mathcal{B}_n\left(\varsigma\right)\right] \to 1$.

Let $\delta = \inf_{x\in\mathcal{X}}\prod_{t=0}^{J}L_t\left(g_{-0}\left(x,\gamma_K^0\right)\right)$ and observe that for K large enough $\delta > 0$ by part (i) and the assumption that the true probabilities are strictly between zero and one. Define the sets

$$\Gamma_K^\delta = \left\{\gamma\in\mathbb{R}^{JK} : \inf_{x\in\mathcal{X}}\prod_{t=0}^{J}L_t\left(g_{-0}\left(x,\gamma\right)\right) > \frac{\delta}{2}\right\},$$

www.manaraa.com

and $\Gamma_K^0(\varrho) = \left\{\gamma \in \mathbb{R}^{JK} : |\gamma - \gamma_K^0| \leq \varrho\left(K^{1/2}n^{-1/2} + K^{1/2}K^{-\alpha}\right)\right\}$ for any $\varrho > 0$, and because (for some intermediate point $\tilde{\gamma}_K$),

$$\begin{aligned}
&\sup_{x\in\mathcal{X},\gamma\in\Gamma_K^0(\varrho)} \left|\mathbf{L}\left(g_{-0}\left(x,\gamma\right)\right) - \mathbf{L}\left(g_{-0}\left(x,\gamma_K^0\right)\right)\right| \\
&\leq \sup_{x\in\mathcal{X},\gamma\in\Gamma_K^0(\varrho),\tilde{\gamma}_K} \left|\dot{\mathbf{L}}\left(g_{-0}\left(x,\tilde{\gamma}_K\right)\right) \otimes R_K\left(X_i\right)'\right| \left|\gamma - \gamma_K^0\right| \\
&\leq C\zeta\left(K\right) \sup_{\gamma\in\Gamma_K^0(\varrho)} \left|\gamma - \gamma_K^0\right| \\
&= O\left(\zeta\left(K\right)K^{1/2}n^{-1/2} + \zeta\left(K\right)K^{1/2}K^{-\alpha}\right) = o\left(1\right)
\end{aligned}$$

by Assumptions B-1(b) and B-1(c), and for n for large enough it follows that $\Gamma_K^\delta \subset \Gamma_K^0(\varrho)$.

To finish the argument, choose n large enough so that $\Gamma_K^\delta \subset \Gamma_K^0(C)$, $\mathbb{P}\left[\mathcal{A}_n\right] \geq 1 - \varepsilon/2$ and $\mathbb{P}\left[\mathcal{B}_n\left(\delta C/8\right)\right] \geq 1 - \varepsilon/2$, for some $C > 0$. Then for any $\gamma_K \in \Gamma_K^0$,

$$\begin{aligned}
-\frac{\partial}{\partial\gamma\partial\gamma'}\ell_n\left(\gamma_K\right) &= \frac{1}{n}\sum_{i=1}^n \mathbf{H}\left(X_i,\gamma_K\right) \otimes R_K\left(X_i\right) R_K\left(X_i\right)' \\
&\geq \frac{1}{n}\sum_{i=1}^n \left[\inf_{x\in\mathcal{X}} \prod\nolimits_{t=0}^J L_t\left(g_{-0}\left(x,\gamma_K\right)\right) \mathbf{I}_J\right] \otimes R_K\left(X_i\right) R_K\left(X_i\right)' \\
&\geq \frac{\delta}{2}\left[\mathbf{I}_J \otimes \hat{\Omega}_K\right],
\end{aligned}$$

which implies that with probability at least $(1 - \varepsilon)$,

$$\lambda_{\min}\left(-\frac{\partial}{\partial\gamma\partial\gamma'}\ell_n\left(\gamma_K\right)\right) \geq \frac{\delta}{4}.$$

Moreover, under the same conditions (i.e., also with probability at least $(1 - \varepsilon)$)

www.manaraa.com

verify that for any $\gamma_K \in \Gamma_K^0 \backslash \{\gamma_K^0\}$,

$$\begin{aligned}
&\ell_n(\gamma_K) - \ell_n(\gamma_K^0) \\
&= \frac{\partial}{\partial \gamma}\ell_n(\gamma_K^0)(\gamma_K - \gamma_K^0) - \frac{1}{2}(\gamma_K - \gamma_K^0)'\left[-\frac{\partial}{\partial \gamma \partial \gamma'}\ell_n(\tilde{\gamma}_K)\right](\gamma_K - \gamma_K^0) \\
&\leq \left|\frac{\partial}{\partial \gamma}\ell_n(\gamma_K^0)\right||\gamma_K - \gamma_K^0| - \frac{\delta}{8}|\gamma_K - \gamma_K^0|^2 \\
&\leq \left(\left|\frac{\partial}{\partial \gamma}\ell_n(\gamma_K^0)\right| - \frac{\delta}{8}C\left(K^{1/2}n^{-1/2} + K^{1/2}K^{-\alpha}\right)\right)|\gamma_K - \gamma_K^0| < 0,
\end{aligned}$$

for some $\tilde{\gamma}_K$ such that $|\tilde{\gamma}_K - \gamma_K^0| \leq |\gamma_K - \gamma_K^0|$. Since $\ell_n(\gamma_K)$ is continuous and concave, it follows that $\hat{\gamma}_K$ maximizes $\ell_n(\gamma_K)$ and $\hat{\gamma}_K$ satisfies the first order condition with probability approaching one.

Now the result follows directly. ■

www.manaraa.com

Appendix C

Block Regression Estimator

Recall that $r(e) = \left[1, e, e^2, \cdots, e^{K-1}\right] \in \mathbb{R}^K$. Let

$$\Omega_j = \mathbb{E}\left[R_K(E) R_K(E)' \mid \mathbf{1}_{\mathcal{W}_j}(E) = 1\right]$$

be the second moment matrix. The best linear predictor within block is defined as $\tilde{\mu}_j(e) = 1_{\mathcal{W}_j}(e) r(e)' \tilde{\gamma}_j$ where

$$\begin{aligned}
\tilde{\gamma}_j &= \arg\min_{\gamma} \mathbb{E}\left[\left(Y - R_K(E)'\gamma\right)^2 \mid \mathbf{1}_{\mathcal{W}_j}(E) = 1\right] \\
&= \left(\mathbb{E}\left[R_K(E) R_K(E)' \mid \mathbf{1}_{\mathcal{W}_j}(E) = 1\right]\right)^{-1} \left(\mathbb{E}\left[R_K(E) Y \mid \mathbf{1}_{\mathcal{W}_j}(E) = 1\right]\right) \\
&= \Omega_j^{-1} \mathbb{E}\left[R_K(E) Y \mid \mathbf{1}_{\mathcal{W}_j}(E) = 1\right].
\end{aligned}$$

www.manaraa.com

while the regression estimator is given by $\hat{\mu}_j(e) = \mathbf{1}_{N_j} R_K(E_i)(e)' \hat{\gamma}_j$ where (on $\{\mathbf{1}_{N_j} = 1\}$)

$$\begin{aligned}\hat{\gamma}_j &= \arg\min_{\gamma} \sum_{i=1}^{n} \mathbf{1}_{\mathcal{W}_j}(E_i)\left(Y_i - R_K(E_i)'\gamma\right)^2 \\ &= \left(\frac{1}{N_j}\sum_{i=1}^{n} \mathbf{1}_{\mathcal{W}_j}(E_i) R_K(E_i) R_K(E_i)'\right)^{-1}\left(\frac{1}{N_j}\sum_{i=1}^{n} \mathbf{1}_{\mathcal{W}_j}(E_i) R_K(E_i) Y_i\right) \\ &= \hat{\Omega}_j^{-1}\left(\frac{1}{N_{1j}}\sum_{i=1}^{n} \mathbf{1}_{\mathcal{W}_j}(E_i) R_K(E_i) Y_i\right).\end{aligned}$$

Next, for fixed K, define the diagonal matrix

$$\mathbf{J}^{-1} = \operatorname{diag}\left[\left(\frac{|\mathcal{W}|}{J}\right)^k : k = 0, 1, \cdots, K-1\right]$$

and define the lower triangular matrix $\mathbf{L}_j$ as

$$\mathbf{L}_j(k,l) = \binom{k-1}{l-1}\left(\frac{b_{j-1}}{|\mathcal{W}_j|}\right)^{k-l} \mathbf{1}\{1 \le l \le k \le K+1\},$$

where b_j is the upper limit of the j-th block, and let $\mathbf{U}_j = \mathbf{L}_j'$, the corresponding upper triangular matrix.

The proof of Theorem 11 will follow directly from the following Lemmas:

Lemma 12 *(Decomposition for Change of Variables of Regression Basis) For $d, l \in \mathbb{R}$, $R_K(d(e+l)) = \mathbf{D}\mathbf{L}R_K(e)$, where $\mathbf{D} = \operatorname{diag}\left[d^k : k = 0, 1, \cdots, K\right]$ and the matrix $\mathbf{L}$ is a lower triangular matrix with typical element,*

$$\mathbf{L}(k,i) = \binom{k-1}{i-1} l^{k-i} \mathbf{1}\{1 \le i \le k \le K+1\}$$

and observe that $\mathbf{L}(k,k) = 1$. Define $\mathbf{U} = \mathbf{L}'$.

www.manaraa.com

Lemma 13 *(Uniform Bound of Normalized Regression Basis)*

$$\sup_{e\in\mathcal{W}_j}\left|\mathbf{L}_j^{-1}\mathbf{J}R_K(e)\right| = K+1, \qquad \textit{uniformly in } j \textit{ and } J.$$

Lemma 14 *(Eigenvalues for Normalized Second Moment Matrix)*

$$C_*(K) \leq \lambda_{\max}\left(\mathbf{L}_j^{-1}\mathbf{J}\Omega_j\mathbf{J}\mathbf{U}_j^{-1}\right) \leq C^*(K), \quad \textit{uniformly in } j \textit{ and } J,$$

and where $0 < C_*(K) \leq C^*(K) < \infty$.

Proof. *First, let* $\mathbf{H}$ *be the Hilbert matrix of order* K *and define the matrix* $\mathbf{H}_j \equiv \int_{\mathcal{W}_j} R_K(e) R_K(e)' de$, *and observe that by change of variables*

$$\mathbf{H}_j \equiv \int_{\mathcal{W}_j} R_K(e) R_K(e)' de = |\mathcal{W}_j| \int_0^1 R_K(u|\mathcal{W}_j| + b_{j-1}) R_K(u|\mathcal{W}_j| + b_{j-1})' du.$$

Recall that $|\mathcal{W}_j| = |\mathcal{W}|/J$ *and observe that using the previous Lemmas,*

$$R_K(u|\mathcal{W}_j| + b_{j-1}) = R_K(|\mathcal{W}_j|(u + b_{j-1}/|\mathcal{W}_j|)) = \mathbf{J}^{-1}\mathbf{L}_j R_K(u).$$

Putting these results together,

$$J\mathbf{H}_j = |\mathcal{W}| \int_0^1 \mathbf{J}^{-1}\mathbf{L}_j R_K(u) R_K(u)' \mathbf{U}_j\mathbf{J}^{-1} du = C\mathbf{J}^{-1}\mathbf{L}_j\mathbf{H}\mathbf{U}_j\mathbf{J}^{-1}.$$

Next, observe that

$$\begin{aligned}
\Omega_j &= \mathbb{E}\left[R_K(E) R_K(E)' \;\middle|\; \mathbf{1}_{\mathcal{W}_j}(E) = 1\right] \\
&= \frac{1}{q_j}\mathbb{E}\left[\mathbf{1}_{\mathcal{W}_j}(E) R_K(E) R_K(E)'\right] \\
&= \frac{1}{q_j}\int_{\mathcal{W}_j} R_K(e) R_K(e)' f(e) de,
\end{aligned}$$

www.manaraa.com

which implies that

$$C_* J\mathbf{H}_j \equiv C_* J \int_{\mathcal{W}_j} R_K(e) R_K(e)' de \leq \Omega_j \leq C^* J \int_{\mathcal{W}_j} R_K(e) R_K(e)' de \equiv C^* J\mathbf{H}_j,$$

which leads to

$$C_*\mathbf{H} = C_*\mathbf{L}_j^{-1}\mathbf{J}(J\mathbf{H}_j)\mathbf{J}\mathbf{U}_j^{-1} \leq \mathbf{L}_j^{-1}\mathbf{J}\Omega_j\mathbf{J}\mathbf{U}_j^{-1} \leq C^*\mathbf{L}_j^{-1}\mathbf{J}(J\mathbf{H}_j)\mathbf{J}\mathbf{U}_j^{-1} = C^*\mathbf{H}.$$

Finally, to get the first result, note that $\lambda_{\max}\left(\mathbf{L}_j^{-1}\mathbf{J}\Omega_j\mathbf{J}\mathbf{U}_j^{-1}\right) \leq C^*\lambda_{\max}(\mathbf{H}) \equiv C^*(K)$ *which holds uniformly in* j *and* J*. To verify the second result, observe that* $\lambda_{\min}\left(\mathbf{L}_j^{-1}\mathbf{J}\Omega_j\mathbf{J}\mathbf{U}_j^{-1}\right) \geq C_*\lambda_{\min}(\mathbf{H}) \equiv C_*(K)$ *which holds uniformly in* j *and* J*.* ■

Lemma 15 *(*Rates of Convergence for Normalized Sample Second Moment Matrix*)*

$$\begin{aligned}
\mathbb{E}\left[\left|\mathbf{1}_{N_j}\mathbf{L}_j^{-1}\mathbf{J}\hat{\Omega}_j\mathbf{J}\mathbf{U}_j^{-1} - \mathbf{L}_j^{-1}\mathbf{J}\Omega_j\mathbf{J}\mathbf{U}_j^{-1}\right|^2\right] &= O\left(Jn^{-1}\right), \text{ uniformly in } j, \text{ and} \\
\mathbb{E}\left[\max_{1\leq j\leq J}\left|\mathbf{1}_{N_j}\mathbf{L}_j^{-1}\mathbf{J}\hat{\Omega}_j\mathbf{J}\mathbf{U}_j^{-1} - \mathbf{L}_j^{-1}\mathbf{J}\Omega_j\mathbf{J}\mathbf{U}_j^{-1}\right|^2\right] &= O\left(J^2n^{-1}\right).
\end{aligned}$$

As a consequence,

$$\begin{aligned}
\left|\mathbf{1}_{N_j}\mathbf{L}_j^{-1}\mathbf{J}\hat{\Omega}_j\mathbf{J}\mathbf{U}_j^{-1} - \mathbf{L}_j^{-1}\mathbf{J}\Omega_j\mathbf{J}\mathbf{U}_j^{-1}\right| &= O_p\left(J^{1/2}n^{-1/2}\right), \text{ uniformly in } j, \\
\max_{1\leq j\leq J}\left|\mathbf{1}_{N_j}\mathbf{L}_j^{-1}\mathbf{J}\hat{\Omega}_j\mathbf{J}\mathbf{U}_j^{-1} - \mathbf{L}_j^{-1}\mathbf{J}\Omega_j\mathbf{J}\mathbf{U}_j^{-1}\right| &= O_p\left(Jn^{-1/2}\right).
\end{aligned}$$

www.manaraa.com

Proof. *First, observe that*

$$\begin{aligned}
&\mathbb{E}\left[\left|\mathbf{1}_{N_j}\mathbf{L}_j^{-1}\mathbf{J}\hat{\Omega}_j\mathbf{J}\mathbf{U}_j^{-1}-\frac{N_{1j}}{Nq_j}\mathbf{L}_j^{-1}\mathbf{J}\hat{\Omega}_j\mathbf{J}\mathbf{U}_j^{-1}\right|^2\right] \\
&=\mathbb{E}\left[\left(\mathbf{1}_{N_j}\frac{1}{N_j}-\frac{1}{q_jn}\right)^2\left|\sum_{i=1}^n\mathbf{1}_{\mathcal{W}_j}(E_i)\,\mathbf{L}_j^{-1}\mathbf{J}R_K(E_i)\,R_K(E_i)'\,\mathbf{J}\mathbf{U}_j^{-1}\right|^2\right] \\
&\leq\left(\sup_{e\in\mathcal{W}_j}\left|\mathbf{L}_j^{-1}\mathbf{J}r(e)\right|^2\right)\mathbb{E}\left[\left(\mathbf{1}_{N_j}\frac{1}{N_j}-\frac{1}{q_jn}\right)^2N_j^2\right] \\
&=\frac{C}{q_j^2n^2}\mathbb{E}\left[\left(N_j-\mathbf{1}_{N_j}q_jn\right)^2\right]=O\left(Jn^{-1}\right).
\end{aligned}$$

Second, define

$$v_{ij}=\frac{1}{q_j}\mathbf{1}_{\mathcal{W}_j}(E_i)\,\mathbf{L}_j^{-1}\mathbf{J}R_K(E_i)\,R_K(E_i)'\,\mathbf{J}\mathbf{U}_j^{-1},$$

and note that

$$\begin{aligned}
&\mathbb{E}\left[\left|\frac{N_{1j}}{q_jn}\mathbf{L}_j^{-1}\mathbf{J}\hat{\Omega}_j\mathbf{J}\mathbf{U}_j^{-1}-\mathbf{L}_j^{-1}\mathbf{J}\Omega_j\mathbf{J}\mathbf{U}_j^{-1}\right|^2\right] \\
&=\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^n v_{ij}-\mathbb{E}\left[v_{ij}\right]\right|^2\right] \\
&\leq\frac{1}{n}\mathbb{E}\left[\left|\frac{1}{q_j}\mathbf{1}_{\mathcal{W}_j}(E)\,\mathbf{L}_j^{-1}\mathbf{J}R_K(E)\,R_K(E)'\,\mathbf{J}\mathbf{U}_j^{-1}\right|^2\right] \\
&=\frac{C}{J^2n}\mathbb{E}\left[\mathbf{1}_{\mathcal{W}_j}(E)\left|\mathbf{L}_j^{-1}\mathbf{J}R_K(E)\,R_K(E)'\,\mathbf{J}\mathbf{U}_j^{-1}\right|^2\right] \\
&=O\left(Jn^{-1}\right).
\end{aligned}$$

Now, the first conclusion follows by putting these two results together and using the triangular inequality. The second conclusion follows by Boole's inequality, while the last two conclusions follow directly by Markov's inequality. ■

www.manaraa.com

Lemma 16 *(Normalized Sample Second Moment Matrix Eigenvalues)*

$$C_*(K) - O_p\left(J^2 n^{-1}\right) \leq \mathbf{1}_{N_j} \lambda_{\max}\left(\mathbf{L}_j^{-1}\mathbf{J}\hat{\Omega}_j\mathbf{J}\mathbf{U}_j^{-1}\right) \leq C^*(K) + O_p\left(J^2 n^{-1}\right),$$

uniformly in j and J.

Lemma 17 *(Uniform Bounded Tails) If $J^2 n^{-1} = O(1)$ then*

$$\mathbb{P}\left[\max_{1\leq j\leq J} \frac{1}{q_j n} \sum\nolimits_{i=1}^{n} \mathbf{1}_{\mathcal{W}_j}(E_i)\left(Y_i - \mu^*(E_i)\right)^2 > 2M\right] \longrightarrow 0, \qquad \text{as } M \longrightarrow \infty.$$

Proof. *First, define the stochastic process*

$$Z_n(j) = \frac{1}{q_j n} \sum\nolimits_{i=1}^{n} \mathbf{1}_{\mathcal{W}_j}(E_i)\left(Y_i - \mu^*(E_i)\right)^2,$$

and observe that for any $\alpha > 0$ and for all $j = 1, 2, \cdots, J$,

$$\begin{aligned}
\mathbb{P}\left[|Z_n(j)| \leq \alpha\right] &\geq 1 - \frac{1}{\alpha}\mathbb{E}\left[\frac{1}{q_j n}\sum\nolimits_{i=1}^{n} \mathbf{1}_{\mathcal{W}_j}(E_i)\left(Y_i - \mu^*(E_i)\right)^2\right] \\
&= 1 - \frac{1}{\alpha}\mathbb{E}\left[\left(Y - \mu^*(E)\right)^2 \;\middle|\; \mathbf{1}_{\mathcal{W}_j}(E) = 1\right] \\
&\geq 1 - \frac{v^*}{\alpha} \equiv \beta,
\end{aligned}$$

which implies that $\beta \longrightarrow 1$ as $\alpha \longrightarrow \infty$. In particular, let $\alpha = M$ and observe that for M large enough $\mathbb{P}\left[|Z_n(j)| \leq M\right] \geq 1/2$.

Next, apply Lemma 8 (First Symmetrization) of Pollard (1984) with $\alpha = M$ to obtain

$$\mathbb{P}\left[\max_{1\leq j\leq J} |Z_n(j)| \geq 2M\right] \leq 2\mathbb{P}\left[\max_{1\leq j\leq J} \left|Z_n(j) - \tilde{Z}_n(j)\right| \geq M\right],$$

www.manaraa.com

where $\tilde{Z}_n(j)$ *is an independent copy of* $Z_n(j)$ *as explained in Pollard (1984). To finish the argument, observe that by Markov's Inequality*

$$\begin{aligned}\mathbb{P}\left[\max_{1\leq j\leq J}\left|Z_n(j)-\tilde{Z}_n(j)\right|\geq M\right] &\leq \frac{1}{M^2}J\max_{1\leq j\leq J}\mathbb{E}\left[\left|Z_n(j)-\tilde{Z}_n(j)\right|^2\right]\\ &\leq \frac{C}{M}J\max_{1\leq j\leq J}\mathbb{E}\left[\left|Z_n(j)-\mathbb{E}\left[Z_n(j)\right]\right|^2\right]\\ &\leq \frac{C}{M^2}\frac{J^3}{n}\max_{1\leq j\leq J}\mathbb{E}\left[\mathbf{1}_{\mathcal{W}_j}(E)\left(Y-\mu^*(E)\right)^4\right]\\ &\leq \frac{C}{M^2}\frac{J^2}{n}=o(1),\end{aligned}$$

giving the result. ■

Lemma 18 *(*Uniform Convergence of Normalized Sum*) If* $J^2n^{-1}=O(1)$ *then*

$$\max_{1\leq j\leq J}\left|\frac{1}{q_jn}\sum\nolimits_{i=1}^{n}\mathbf{1}_{\mathcal{W}_j}(E_i)\left(\mathbf{L}_j^{-1}\mathbf{J}R_K(E_i)\right)\left(Y_i-\mu^*(E_i)\right)\right|=O_p\left(J^{1/2}n^{-1/2}\log(n)^{1/2}\right).$$

Proof. *The proof of this result also follows the seminal work of Pollard (1984) (see Theorem 37, pages 34-35). First, observe that*

$$\begin{aligned}&\max_{1\leq j\leq J}\left|\frac{1}{q_jn}\sum\nolimits_{i=1}^{n}\mathbf{1}_{\mathcal{W}_j}(E_i)\left(\mathbf{L}_j^{-1}\mathbf{J}R_K(E_i)\right)\left(Y_i-\mu^*(E_i)\right)\right|\\ &=\max_{1\leq j\leq J}\left[\sum\nolimits_{k=0}^{K}\left[\frac{1}{q_jn}\sum\nolimits_{i=1}^{n}\mathbf{1}_{\mathcal{W}_j}(E_i)\left(\left[\mathbf{L}_j^{-1}\mathbf{J}R_K(E_i)\right]_k\right)\left(Y_i-\mu^*(E_i)\right)\right]^2\right]^{1/2}\\ &\leq C\max_{0\leq k\leq K}\max_{1\leq j\leq J}\left|\frac{1}{q_jn}\sum\nolimits_{i=1}^{n}\mathbf{1}_{\mathcal{W}_j}(E_i)\left(\left[\mathbf{L}_j^{-1}\mathbf{J}R_K(E_i)\right]_k\right)\left(Y_i-\mu^*(E_i)\right)\right|,\end{aligned}$$

where the term inside the absolute value is a sum of iid mean zero random variables and $[v]_k$ *denotes the* k*-th component of the vector* v*. To save notation, define*

$$v_{ijk}=\frac{1}{q_j}\mathbf{1}_{\mathcal{W}_j}(E_i)\left(\left[\mathbf{L}_j^{-1}\mathbf{J}R_K(E_i)\right]_k\right)\left(Y_i-\mu^*(E_i)\right).$$

www.manaraa.com

Next, fix $k = 0, 1, ..., K$ *and let* $\zeta_n^2 = Jn^{-1}\log(n)$. *Then, for fixed* $\alpha > 0$ *and by the Markov's Inequality*

$$\begin{aligned}
&\mathbb{P}\left[\zeta_n^{-1}\left|\frac{1}{n}\sum\nolimits_{i=1}^{n} \upsilon_{ijk}\right| \geq \alpha\right] \\
&\leq \frac{1}{\alpha^2}\zeta_n^{-2}\mathbb{E}\left[\left|\frac{1}{n}\sum\nolimits_{i=1}^{n} \upsilon_{ijk}\right|^2\right] \\
&\leq \frac{1}{\alpha^2}\zeta_n^{-2}\frac{1}{n}\mathbb{E}\left[\left|\frac{1}{q_j}\mathbf{1}_{\mathcal{W}_j}(E)\left(\left[\mathbf{L}_j^{-1}\mathbf{J}R_K(E)\right]_k\right)(Y-\mu^*(E))\right|^2\right] \\
&\leq \frac{CJ^{-1}}{\alpha^2\log(n)\,q_j}\mathbb{E}\left[\left(\left[\mathcal{L}_j^{-1}\mathcal{J}r(E)\right]_k\right)^2(Y-\mu(E))^2 \;\middle|\; \mathbf{1}_{\mathcal{B}_j}(E)=1\right] \\
&\leq \frac{C}{\alpha^2\log(n)},
\end{aligned}$$

and hence for any constant $\alpha > 0$ *and* n *large enough it follows that*

$$\mathbb{P}\left[\zeta_n^{-1}\left|\frac{1}{n}\sum\nolimits_{i=1}^{n}\frac{1}{q_j}\mathbf{1}_{\mathcal{W}_j}(E_i)\left(\left[\mathbf{L}_j^{-1}\mathbf{J}R_K(E_i)\right]_k\right)(Y_i-\mu^*(E_i))\right| \leq \alpha\right] \geq \frac{1}{2},$$

which holds for each $j = 1, 2, ..., J$. *Using this result and applying the two symmetrizations discussed in Pollard (1984) (pp. 14-15) with* $\alpha = M$ *and* $\beta = 1/2$, *gives*

$$\begin{aligned}
&\mathbb{P}\left[\max_{1\leq j\leq J}\left|\frac{1}{n}\sum\nolimits_{i=1}^{n}\upsilon_{ijk}\right| > \zeta_n 3M\right] \\
&\leq 4\mathbb{P}\left[\max_{1\leq j\leq J}\left|\frac{1}{n}\sum\nolimits_{i=1}^{n}\upsilon_{ijk}\sigma_i\right| > \zeta_n M\right] \\
&= 4\mathbb{E}\left[\mathbb{P}\left[\max_{1\leq j\leq J}\left|\frac{1}{n}\sum\nolimits_{i=1}^{n}\upsilon_{ijk}\sigma_i\right| > \zeta_n M \;\middle|\; \{E_i, Y_i\}\right]\right],
\end{aligned}$$

where σ_i *are independent Rademacher random variables as explained in Pollard (1984).*

Now, using the Boole's inequality, Hoeffding's inequality and previous bounds it

www.manaraa.com

follows that

$$\begin{aligned}
&\mathbb{P}\left[\max_{1\leq j\leq J}\left|\frac{1}{n}\sum\nolimits_{i=1}^{n}\upsilon_{ijk}\sigma_i\right|>\zeta_n M \;\middle|\; \{E_i,Y_i\}\right]\\
&\leq\sum\nolimits_{j=1}^{J}\mathbb{P}\left[\left|\frac{1}{n}\sum\nolimits_{i=1}^{n}\upsilon_{ijk}\sigma_i\right|>\zeta_n M \;\middle|\; \{E_i,Y_i\}\right]\\
&\leq 2J\max_{1\leq j\leq J}\exp\left\{-\frac{2\left(q_j n\zeta_n M\right)^2}{\sum_{i=1}^{n}\left[2\mathbf{1}_{\mathcal{W}_j}\left(E_i\right)\left[\mathbf{L}_j^{-1}\mathbf{J}R_K\left(E_i\right)\right]_k\left(Y_i-\mu^*\left(E_i\right)\right)\right]^2}\right\}\\
&=2J\max_{1\leq j\leq J}\exp\left\{-\frac{1}{2}\frac{J^{-1}n^2\log\left(n\right)M^2}{\sum_{i=1}^{n}\mathbf{1}_{\mathcal{W}_j}\left(E_i\right)\left[\mathbf{L}_j^{-1}\mathbf{J}R_K\left(E_i\right)\right]_k^2\left(Y_i-\mu^*\left(E_i\right)\right)^2}\right\}\\
&\leq 2J\exp\left\{-\frac{C\log\left(N\right)M^2}{\max_{1\leq j\leq J}q_j^{-1}n^{-1}\sum_{n=1}^{N}\mathbf{1}_{\mathcal{B}_j}\left(E_n\right)\left(Y_n-\mu\left(E_n\right)\right)^2}\right\}.
\end{aligned}$$

To finish the argument, using the previous results, note that

$$\begin{aligned}
&\mathbb{P}\left[\max_{1\leq j\leq J}\left|\frac{1}{n}\sum\nolimits_{i=1}^{n}\frac{1}{q_j}\mathbf{1}_{\mathcal{W}_j}\left(E_i\right)\left(\left[\mathbf{L}_j^{-1}\mathbf{J}R_K\left(E_i\right)\right]_k\right)\left(Y_i-\mu^*\left(E_i\right)\right)\right|>\zeta_n 3M\right]\\
&\leq 2J\exp\left\{-\frac{C\log\left(N\right)M^2}{2M}\right\}+\mathbb{P}\left[\max_{1\leq j\leq J}\frac{1}{q_j n}\sum\nolimits_{i=1}^{n}\mathbf{1}_{\mathcal{W}_j}\left(E_i\right)\left(Y_i-\mu^*\left(E_i\right)\right)^2>2M\right],
\end{aligned}$$

where the first term is $o\left(1\right)$ *(for example, if* $J=N^{\delta}$ *then this term is* $o\left(1\right)$ *for any* $M>2\delta/\left(C\varepsilon\right)$*), and the second term is* $o\left(1\right)$ *according the previous Lemmas. This establishes that for fixed* k*,*

$$\mathbb{P}\left[\max_{1\leq j\leq J}\left|\frac{1}{n}\sum\nolimits_{i=1}^{n}\frac{1}{q_j}\mathbf{1}_{\mathcal{W}_j}\left(E_i\right)\left(\left[\mathbf{L}_j^{-1}\mathbf{J}R_K\left(E_i\right)\right]_k\right)\left(Y_i-\mu^*\left(E_i\right)\right)\right|>\zeta_n 3M\right]\longrightarrow 0,$$

as $M\longrightarrow\infty$ *for* n *large enough. Now the conclusion follows because the final bound is not a function of* $k=0,1,...,K$*.* ■

Proof of Theorem 11: For the uniform rate of convergence observe that it is

www.manaraa.com

sufficient to show,

$$\sup_{e\in\mathcal{W}_j}\left|\mu^0_{K,j}(e)-\mu^*(e)\right| = O\left(J^{-K}\right) \tag{C.1}$$

for some $\mu^0_{K,j}(e) = R_K(e)'\gamma^0_{K,j}$ and uniformly in j,

$$\sup_{e\in\mathcal{W}_j}\left|\tilde{\mu}_{K,j}(e)-\mu^0_{K,j}(e)\right| = O\left(J^{-K}\right), \tag{C.2}$$

for $\tilde{\mu}_{K,j}(e) = R_K(e)'\tilde{\gamma}_{K,j}$ and uniformly in j, and

$$\sup_{e\in\mathcal{W}_j}\left|\hat{\mu}_{K,j}(e)-\tilde{\mu}_{K,j}(e)\right| = O_p\left(J^{1/2}n^{-1/2}\log(n)^{1/2}\right)+O_p\left(J^{-K}\right), \tag{C.3}$$

uniformly in j.

Now, Equation (C.1) can be verified by using Theorem 4.2 (page 183) jointly with the results on page 46 of DeVore and Lorentz (1993), which implies that there exists a vector $\gamma^0_j\in\mathbb{R}^K$ such that

$$\sup_{e\in\mathcal{W}_j}\left|R_K(e)'\gamma^0_j-\mu^*(e)\right| \le C\left(\frac{1}{J}\right)^K \sup_{e\in\mathcal{B}_j}\left|\frac{\partial^K}{\partial e^K}\mu^*(e)\right| = O\left(J^{-K}\right).$$

Next, observe that

$$\begin{aligned}
&\sup_{e\in\mathcal{W}_j}\left|R_K(e)'\tilde{\gamma}_j - R_K(e)'\gamma^0_j\right| \\
&= \sup_{e\in\mathcal{W}_j}\left|R_K(e)'\left[\Omega_j^{-1}\mathbb{E}\left[R_K(E)\,Y \mid \mathbf{1}_{\mathcal{W}_j}(E)=1\right]-\gamma^0_j\right]\right| \\
&= C\sup_{e\in\mathcal{W}_j}\left|\mathbb{E}\left[\mathbf{L}_j^{-1}\mathbf{J}R_K(E)\left(\mu^*(E)-R_K(E)'\gamma^0_j\right) \mid \mathbf{1}_{\mathcal{W}_j}(E)=1\right]\right| \\
&\le C\sup_{e\in\mathcal{W}_j}\left|\mu^*(e)-R_K(e)'\gamma^0_j\right| \\
&= O\left(J^{-K}\right),
\end{aligned}$$

www.manaraa.com

which gives Equation (C.2).

Finally,

$$\begin{aligned}
&\sup_{e\in\mathcal{W}_j}\left|\mathbf{1}_{N_j}R_K(e)'\hat{\gamma}_j - R_K(e)'\gamma_j^*\right| \\
&= \sup_{e\in\mathcal{W}_j}\left|\mathbf{1}_{N_j}R_K(e)'\hat{\Omega}_j^{-1}\left(\frac{1}{N_j}\sum\nolimits_{i=1}^{n}\mathbf{1}_{\mathcal{W}_j}(E_i)R_K(E_i)\left(Y_i - R_K(E_i)'\gamma_j^*\right)\right)\right| + o_p\left(J^{-K}\right) \\
&\leq O_p(1)\left|\frac{1}{N_j}\sum\nolimits_{n=1}^{N}\mathbf{1}_{\mathcal{W}_j}(E_i)\left(\mathbf{L}_j^{-1}\mathbf{J}R_K(E_i)\right)\left(Y_i - R_K(E_i)'\gamma_j^*\right)\right| + o_p\left(J^{-K}\right) \\
&\leq O_p(1)\max_{1\leq j\leq J}\left|\frac{1}{q_j n}\sum\nolimits_{n=1}^{N}\mathbf{1}_{\mathcal{W}_j}(E_i)\left(\mathbf{L}_j^{-1}\mathbf{J}R_K(E_i)\right)\left(Y_i - \mu^*(E_i)\right)\right| + O_p\left(J^{-K}\right) \\
&= O_p\left(Jn^{-1/2}\log(n)^{1/2}\right) + O_p\left(J^{-K}\right),
\end{aligned}$$

where the last result follows by the previous calculations and lemmas.

Using Equations (C.1), (C.2), and (C.3), the uniform rate of converge follows directly by the triangular inequality after noting that

$$\begin{aligned}
\sup_{e\in\mathcal{W}}\left|\hat{\mu}(e) - \mu^*(e)\right| &= \sup_{e\in\mathcal{W}}\left|\sum\nolimits_{j=1}^{J}\mathbf{1}_{\mathcal{W}_j}(e)\hat{\mu}_j(e) - \mu^*(e)\right| \\
&= \max_{1\leq j\leq J}\sup_{e\in\mathcal{W}_j}\left|\hat{\mu}_j(e) - \mu^*(e)\right|.
\end{aligned}$$

Next, to verify the L_2 rate of convergence, using the results of the previous Lemmas it follows that

$$\begin{aligned}
&\mathbf{L}_j^{-1}\mathbf{J}\left(\mathbf{1}_{N_j}\hat{\gamma}_j - \gamma_j^*\right) \\
&= C\frac{1}{q_j n}\sum\nolimits_{i=1}^{n}\mathbf{1}_{\mathcal{W}_j}(E_i)\mathbf{L}_j^{-1}\mathbf{J}R_K(E_i)\left(Y_i - R_K(E_i)'\gamma_j^*\right) + o_p\left(Jn^{-1}\right) \\
&= O_p\left(J^{1/2}n^{-1/2}\right),
\end{aligned}$$

www.manaraa.com

uniformly in j since

$$\mathbb{E}\left[\left|\frac{1}{q_j n}\sum\nolimits_{i=1}^{n}\mathbf{1}_{\mathcal{W}_j}(E_i)\,\mathbf{L}_j^{-1}\mathbf{J}R_K(E_i)\left(Y_i - R_K(E_i)'\gamma_j^*\right)\right|^2\right] = O\left(Jn^{-1}\right).$$

This result leads to

$$\begin{aligned}
&\int \left(\hat{\mu}_K(e) - \mu_0(e)\right)^2 \, dF(e) \\
&\leq \left(\mathbf{1}_{N_j}\mathbf{L}_j^{-1}\mathbf{J}\hat{\gamma}_j - \mathbf{L}_j^{-1}\mathbf{J}\gamma_j^0\right)^2 \int \left(\sum\nolimits_{j=1}^{J}\mathbf{1}_{\mathcal{W}_j}(e)\left(\mathbf{L}_j^{-1}\mathbf{J}R_K(e)\right)'\right)^2 \, dF(e) \\
&\qquad + \int \left(\sum\nolimits_{j=1}^{J}\mathbf{1}_{\mathcal{W}_j}(e)\,\mu_j^0(e) - \mu^*(e)\right)^2 \, dF(e) \\
&\leq O_p\left(Jn^{-1}\right) + O\left(J^{-2K}\right),
\end{aligned}$$

and this completes the proof. ■

www.manaraa.com