Estimation of Dose-Response Functions and Optimal Treatment Doses with a Continuous Treatment

by

Carlos Arturo Flores

GRAD (Monterrey Institute of Technology) 1998
M.A. (University of California at Berkeley) 2003

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 at BERKELEY

Committee in charge:

Professor Guido W. Imbens, Chair
Professor Deborah Nolan
Professor James L. Powell

Fall 2005

UMI Number: 3210582

Estimation of Dose-Response Functions and Optimal Treatment Doses with a Continuous Treatment

The dissertation of Carlos Arturo Flores is approved:

Chair Date

Date

Date

University of California at Berkeley

Fall 2005

Abstract

Estimation of Dose-Response Functions and Optimal Treatment Doses with a Continuous Treatment

by

Carlos Arturo Flores

Doctor of Philosophy in Economics

University of California at Berkeley

Professor Guido W. Imbens, Chair

Most of the recent program evaluation literature that uses the selection-on-observables assumption focuses on the estimation of average treatment effects of a binary treatment on a scalar outcome. In practice, however, units in a study can often be exposed to different levels or doses of the treatment. Analyzing the impacts of such treatment as if it were binary can mask important features of it. Moreover, with a continuous treatment many more parameters than the usual average treatment effect ones can be of interest. In this dissertation, I focus on estimation of three objects that are of interest in this continuous-treatment setting: (i) the entire curve of average potential outcomes; (ii) the treatment level at which that curve is maximized; and (iii) the maximum value achieved by that curve. In Chapter 2, I discuss nonparametric estimation of these objects under the assumption that units in a study are randomly assigned to different doses of the treatment. Then, I

propose estimators of our objects of interest under the assumption that selection by units into different treatment levels is made based on an observed set of covariates. In both settings, I show that the estimators are asymptotically normally distributed. Regardless of the nature of the treatment, estimation of average potential outcomes with a large number of covariates makes nonparametric estimation problematic. When the treatment is binary, a common approach in the literature is the use of propensity score methods. In this chapter I discuss the use of the generalized propensity score for estimation of our three objects of interest. Different approaches are discussed, such as regression, matching and weighting. This chapter also discusses how to extend the results presented for average dose-response functions to a more general class of functions, such as quantile dose-response functions.

In Chapter 3, I illustrate the utility of our approach by presenting an empirical application. Since the paper by Grossman and Krueger (1991) a large number of studies have documented an inverted U-type relationship between some indicators of environmental degradation and income per capita, known in this literature as the "Environmental Kuznets Curve" (EKC). In this literature, a lot of emphasis is given to estimating the turning point of this relation, that is, the level of income at which the different pollutants reach their peak and start decreasing. Here, I use the methodology presented in this dissertation to estimate non-parametrically the turning points of the EKCs for two pollutants, nitrogen oxide and sulfur dioxide. This empirical application also illustrates what can go wrong when using parametric models to estimate turning points.

Finally, Chapter 4 focuses on a Monte Carlo study to analyze the finite properties of our estimators. To gain insight into the behavior of our estimators in situations actually found in empirical research, I partly base our simulation design on the same data set used in Chapter 3. This final chapter also illustrates the applicability of the estimators developed in this dissertation.

Professor Guido W. Imbens
Dissertation Committee Chair

To God, for all He has given me.

To Melina, the love of my life and my best friend, for always being there for me.

To Arena, for keep reminding me what really matters in life.

To my parents, for all their love, guidance and their longstanding support.

Contents

List of Figures

List of Tables

Acknowledgements

I would like to express my gratitude to my dissertation advisor, Professor Guido W. Imbens, for his patience, guidance, support and many suggestions. He always provided very insightful comments and was very helpful and encouraging at all stages of this project. I would also like to thank Professor James L. Powell for his comments and suggestions at different stages of my dissertation, and for his patience in listening to my ideas and sharing his knowledge in a very friendly way. I am specially indebt to Kenneth Y. Chay. I learned a lot from him at the classroom and during the long conversations at his office. He is a great faculty member to have around. Also, his support during my time at Berkeley was very important in the completion of my degree. I am also grateful to Max Auffhammer, David Card, Michael Jansson, Deborah Nolan, Thomas Rothenberg, and Paul Ruud for their valuable comments and suggestions.

During my time at Berkeley I also had the opportunity to meet extraordinary people and develop strong friendships with many of them. Some of these good friends are: Andres Aradillas and Laura, Miguel Fuentes and Veronica, Pablo Ibarraran and Cathy Kettlewell, Alejandro Moreno, Raul Razo and Lili Cruz, Jose Antonio Rodriguez and Paty Diaz, and Gerardo Zuniga. I would like to thank each of them for their help during our time at Berkeley, for the conversations and words of encouragement, and most of all, for offering my wife and me their unconditional friendship.

I would also like to thank my brother Alfonso Flores-Lagunes. As my older brother, he has always guided me through my life and has made each step much easier. I want to thank him for all the long talks and for always finding the time to help me. I am also grateful to my parents for their love, help, and many sacrifices during my education. My younger sister also helped to the completion of this project by letting me use her computer whenever I ran out of places where to perform my simulations. Definitely, I would not have been here without my family.

Melina always supported me in my academic goals since we were teenagers, even when that meant living away from each other during eight years and sometimes seeing each other just twice a year. For always being there, and for all her love and understanding, I will always be grateful to her. Arena came to my life in the final year of this project. In addition to making the final step more interesting, she also gave me the motivation and strength to give it. I am sure she will continue to illuminate my life.

Finally, I am thankful to CONACYT, UCMEXUS, Banco de Mexico and the Institute for Labor and Employment for their generous financial support.

Chapter 1

Introduction

Most of the recent literature on program evaluation has focused on the analysis of the effect of a binary treatment on a scalar outcome. In practice, however, units in a study can often be exposed to multiple levels or doses of the treatment. Analyzing the effects of such treatment as being binary can mask important features of it. Moreover, even when some studies address the effects of different treatment doses on the outcome, they often do so by creating a discrete number of categories (e.g., Royer, 2003). However, the definition of the groups is typically arbitrary and we lose information about the effects of the treatment within each group. In this dissertation, I propose a method to estimate and carry out inference for different parameters that may be of interest when we have a continuous dose of the treatment.

The main focus when evaluating a binary treatment is often on estimation of average treatment effects (ATE). The two most common are the population ATE and the ATE on the treated. Each one of these parameters is relevant depending on the question one wants to answer. When the treatment is continuous many more parameters and questions can be of interest. For example, we may be interested in learning about the form of the entire function of average potential outcomes over all possible values of the treatment. Also, a policy maker may be interested in finding the level of the treatment that maximizes (or minimizes) some average outcome, as well as the value of the average outcome at that level. In some other applications, the average outcome may increase and then decrease (or vice versa) with the level of the treatment. In this case, a policy maker may be interested in knowing the "turning point" of this relationship. Or we could be interested in the derivative of the average potential outcome curve, or in knowing if there is a dose level at which the curve of average outcomes has a jump or discontinuity, as well as the size of the jump (e.g., if we think of education as our treatment, is there a discrete effect of graduation on average wages?). As these examples suggest, a lot may be learned from analyzing continuous treatments. Moreover, even in cases where the treatment is not strictly continuous but can take many values, a continuous treatment approximation to the problem could be useful.

In this dissertation, I focus on estimating three objects of interest, namely, the entire curve of average potential outcomes or dose-response function, the treatment dose at which that curve is maximized, and the maximum value achieved by that curve[1],[2] . I estimate these objects non-parametrically and establish asymptotic normality for the estimators. The importance of the first parameter is obvious from a policy maker perspective. In contrast to the approach that defines groups based on different treatment doses, this parameter gives the average outcome for all possible values of the treatment. The second

[1]The rest of the parameters mentioned on the previous paragraph are left for future work.

[2]In the rest of the dissertation I will also refer to the last two parameters, respectively, as the location and size of the optimal treatment dose.

and third parameters are important when a policy maker wants to apply or recommend a particular treatment dose to a population. For example, it could be of interest for an agency to know the level of training that maximizes the average net benefits of a given program; or for a health provider to have an estimate of the maternal age at which health outcomes of the newborn are optimized. Moreover, in some cases estimating the maximum or minimum of a given relation would be equivalent to estimating its turning point. The latter parameter is also important in economics. For example, many studies in economics have documented an inverted U-shaped relationship between some measures of pollution and per capita income (e.g., Grossman and Krueger, 1991). These studies also focus on estimation of the turning point, that is, the level of per capita income at which a particular pollution indicator reaches its peak and starts decreasing. This example is further analyzed in the empirical part of this dissertation.

The non-parametric approach presented in this dissertation for estimation of the optimum treatment has some advantages over previous approaches found in the economics literature. One approach that has been previously used is to discretize the treatment, estimate average outcomes for each group (or range of the treatment) and conclude which group is best (e.g., Royer 2003). The problem with this approach is that often discretization is arbitrary, so the best range or group depends on the way the treatment is discretized. Moreover, confidence bounds for the best group are rarely provided, since this would require multiple comparison procedures. Another approach found in the economics literature is to assume a parametric form for the relationship between the treatment and the outcome of interest and estimate the optimal treatment or turning point from there (e.g., Grossman and Krueger, 1991). However, results may be quite sensitive to model specifications (e.g., Harbaugh et al. 2002). For example, as will be further discussed in the empirical part of this dissertation, in some cases using a quadratic model for estimation of optimal treatments or turning points may be misleading. Finally, even when some authors use non-parametric methods for estimating optimal treatments or turning points (e.g., Millimet et al. 2003), they do not provide standard errors for their estimators. The estimators developed in this dissertation are shown to have an asymptotically normal distribution and therefore, they can be used to construct confidence intervals and undertake statistical inference. Moreover, the estimators of the location and size of the optimal dose are shown to be jointly asymptotically normal and uncorrelated.

In Chapter 2, I lay out the problem and formally define the parameters to be estimated in this dissertation. I continue by discussing estimation of our parameters of interest in the experimental case, in which units are assumed to be randomly assigned to different doses of the treatment (i.e., treatment doses are exogenous). In this case, the assumption of randomization allows us to identify our parameters. Then, I move to the non-experimental case and assume that selection by individuals into different treatment levels is made based on an observed set of covariates and on unobserved components not correlated with the potential outcomes. This is a straightforward extension to the continuous treatment case of the "unconfoundedness" or "selection-on-observables" assumption used in the binary-treatment literature. Under this assumption, I present estimators of the parameters of interest based on a regression approach and derive their asymptotic distribution. Even though in this case we need to control for observed covariates, it is shown that the scaling factors for asymptotic normality of the estimators of location and size of the optimal dose in this nonexperimental case are the same as the ones for the corresponding estimators in the experimental case. However, for calculation of the estimators under the unconfoundedness assumption we need first to estimate the nonparametric regression of the observed outcome on the treatment dose and the covariates. This may be a problem if the dimension of the covariates is large, as is usually the case for the unconfoundedness assumption to be more plausible. Because of this problem, which also appears when the treatment is binary, I consider the role the generalized propensity score plays for estimation of our parameters of interest.

In the binary treatment case, a very useful result due to Rosenbaum and Rubin (1983) states that if assignment to the treatment is independent of potential outcomes conditional on a set of pre-treatment variables (i.e., if assignment to treatment is unconfounded), then it is also unconfounded given the propensity score, where they define the propensity score as the probability of receiving the treatment conditional on pre-treatment variables. Hence, in order to control the bias due to imbalances in pre-treatment variables we only need to adjust for a scalar variable (the propensity score), as opposed to adjusting for a possibly high dimensional vector of pre-treatment variables. Imbens (2000) extends the propensity score methodology to the case where the treatment of interest takes on integer values between 0 and L. He shows that the dimension of the conditioning set in this case can again be reduced to one, just as in the binary case. Hirano and Imbens (2004) apply the results from Imbens (2000) to the continuous treatment case to obtain a similar reduction in the conditioning set. With a continuous treatment, the generalized propensity score is defined as the conditional density of the treatment level given pretreatment variables. In chapter

2, I use the results from Hirano and Imbens (2004) to estimate our parameters of interest. Analogous to the implementation of the propensity score methodology in the binary and multiple integer-valued cases, the first step involves estimation of the generalized propensity score. In the second step, I use this estimated generalized propensity score to estimate our objects of interest. In this context, I also discuss how the generalized propensity score can be used following other approaches such as matching or weighting.

Although the main focus of the dissertation is on average dose-response functions, sometimes one may want to consider more general types of functions. For instance, the *p-th* quantile dose-response function gives us, for each dose of the treatment, the *p-th* quantile of the potential outcomes. This could be particularly useful when one is more concerned about the effects of the treatment on the upper or lower regions of the distribution of potential outcomes. Moreover, even when the focus is on a measure of the center of the distribution, it is well known that the median is more robust than the mean. In chapter 2, I also discuss how we can generalize the results we obtain for the estimation of mean dose-response function, its maximum and its value at the maximum, to a more general class of dose-response functions using non-parametric methods. Average and quantile dose-response functions are special cases of that class of functions.

In Chapter 3, I illustrate the use of the techniques developed in this dissertation by presenting an empirical application. Since the paper by Grossman and Krueger (1991) a large number of studies have documented an inverted U-type relationship between some indicators of environmental degradation and income per capita, known in this literature as the "Environmental Kuznets Curve" (EKC). In this literature, a lot of emphasis is given to estimating the turning point of this relation, that is, the level of income at which the different pollutants reach their peak and start decreasing. Here, I use the methodology presented in this dissertation to estimate non-parametrically the turning points of the EKCs for two pollutants, nitrogen oxide and sulfur dioxide.

Finally, in Chapter 4 I analyze the finite properties of the estimators presented in Chapter 2 through a Monte Carlo Study. In order to gain insight into the behavior of our estimators in situations empirical researchers may find in their work, I partly base my simulation study on a real data set. In particular, the simulations are partly based on the data set used in Chapter 3. This chapter also illustrates the applicability of the estimators developed in this dissertation.

Chapter 2

Econometric Theory on the Estimation of Dose-Response Functions and Optimal Doses with a Continuous Treatment

2.1 Introduction

This chapter proposes nonparametric estimators for three objects of interest: the entire curve of average potential outcomes or dose-response function, the treatment dose at which the dose-response is maximized and the maximum value achieved by this curve. These objects are first estimated assuming random assignment of the units in a study to different doses of the treatment. This experimental case is helpful to gain intuition about the problem at hand. In this chapter I show that in this case the proposed estimators of the location and size of the optimal dose are jointly asymptotically normal and uncorrelated. I also discuss similar results, but in different settings, that are available statistics literature. This chapter then considers the case when units are assigned to different doses of the treatment based on an observed set of covariates and on unobserved components not correlated with potential outcomes. This is a straightforward extension of the unconfoundedness assumption commonly used in the binary-treatment literature. In this case, I propose estimators of the parameters of interest based on a regression approach. I show that the estimators for the location and size of the optimal dose are also jointly normal and asymptotically uncorrelated in this non-experimental case. In addition I show that, even though in this case one needs to control for observed covariates, the scaling factors for asymptotic normality of the estimators of location and size of the optimal dose in this nonexperimental case are the same as the ones for the corresponding estimators in the experimental case.

The proposed estimators under the unconfoundedness assumption require estimation of a possibly high dimensional object in a first step. Hence, these estimators are prone to the "curse of dimensionality" problem. A similar problem is also present when the treatment is binary and one needs to control for a large number of covariates. With a binary treatment, a popular approach has been the use of propensity score methods. In this chapter, I discuss how the generalized propensity score (GPS) can be used to estimate our parameters of interest using regression, matching or weighting techniques. I also point out that some other techniques such as the use of additive and partially linear model can be used to reduce the dimensionality problem.

The main focus of this chapter is on average dose-response functions. However, I also discuss how one can extend the results in this chapter to a more general class of functions, for which average and quantile dose-response functions are special cases. In addition, this chapter shows how one can identify that class of dose-response functions under an unconfoundedness assumption.

This chapter is organized as follows. In the following section I lay out the problem and present the parameters to be estimated in the rest of the chapter. In Section 2.3 I discuss estimation of our parameters in the experimental case. In Section 2.4 I present results for the non-experimental case under the unconfoundedness or selection-on-observables assumption. In the next section I discuss some techniques to reduce the dimensionality of the problem considered in the preceding section. First, I show how we could use the generalized propensity score to estimate dose-response functions using different techniques such as regression, matching and weighting. These estimators involve two-stage nonparametric estimation. Second, I briefly discuss how to use our methods in a semiparametric setting. In Section 2.6 I discuss how we can generalize the results obtained in this chapter for estimation of the average dose-response function and its maximum to a more general type of functions, for which average and quantile dose-response functions are special cases.

2.2 Model

I base my model in the potential outcome approach developed by Rubin (1974) and now widely used in the program evaluation literature when analyzing a binary treatment[1].

[1] See, for instance, the surveys by Heckman, Lalonde and Smith (1999) and Imbens (2004).

Assume we have a random sample of size N from a large population. We are interested in how the units in our sample respond to different doses of some treatment with the response measured by some outcome variable Y. The treatment levels, t, take on values in a set $\mathcal{T}$. In the continuous treatment case $\mathcal{T}$ is an interval, e.g. $[0,1]$. Let $Y_i(t)$ denote the potential outcome of unit i under dose t; that is, the outcome unit i would received if exposed to treatment level t. Also, let T_i be the actual treatment dose received by unit i. For each unit, out of all possible values $Y_i(t), t \in \mathcal{T}$, only $Y_i = Y_i(T_i)$ is observed, which leads to a missing-data problem[2].

In the binary case we have that $\mathcal{T} = \{0,1\}$, so that $T_i = 1$ denotes that unit i received the treatment and $T_i = 0$ denotes it did not. This is the case that has recently received most of the attention in the program evaluation literature. The two most common parameters of interest are:

$$E(\Delta) = E[Y(1) - Y(0)] \tag{2.1}$$

$$E(\Delta|T = 1) = E[Y(1) - Y(0)|T = 1] \tag{2.2}$$

The first one is the population average treatment effect[3], which gives the expected effect of the treatment for a unit randomly drawn from the population. The second is the

[2]As noted in Hirano, Imbens and Ridder (2003), the stable-unit-treatment-value assumption (SUTVA) is implicitly assumed in this notation. SUTVA is the assumption that the potential outcome for unit i at treatment level t is not affected either by the mechanism used to assign treatment level t or by the treatment received by other units (Rubin 1978, 1986, 1991). Note that in the binary case this assumption implies that there is only one version of the treatment, and to the extend that treated units receive different doses of the treatment and those doses can be seen as "different treatments", the SUTVA assumption will be violated if such treatment is analyzed as being binary.

[3]For a discussion of concepts of causality in this context see for example Holland (1986) and Heckman (2000).

average treatment effect on the treated, which gives the mean effect of the treatment for the subpopulation of all treated units. In randomized experiments, both parameters are equal.

As discussed in chapter 1, in the case of a continuous treatment we can move beyond pairwise differences as the ones in (2.1)-(2.2) and consider more parameters of interest[4]. In this dissertation, we focus on estimation of three objects:

$$\mu(t) = E\{Y(t)\} \qquad \text{for all } t \in \mathcal{T} \tag{2.3}$$

$$\alpha_0 = \arg\max_{t \in \mathcal{T}} E\{Y(t)\} \tag{2.4}$$

and

$$\mu(\alpha_0) = E\{Y(\alpha_0)\} \tag{2.5}$$

The first one is the entire curve of average potential outcomes or average dose-response function, which gives the average potential outcome at every possible level or dose of the treatment. Note that from this curve we could calculate pairwise treatment effects of the form $E(\Delta^{st}) = E[Y(s) - Y(t)]$ for $s, t \in \mathcal{T}$. However, as opposed to only showing pairwise differences, the dose-response function shows us how average responses vary along the domain of treatment doses. The second parameter is the treatment dose at which the average dose-response function is maximized. For instance, if a policy maker were to choose or recommend a treatment dose to be applied to a population to maximize their expected

[4]Behrman, Cheng and Todd (2004) analyze estimation of treatment effects similar to (2.2) allowing for continuous doses of the treatment. They use matching methods as those studied in Heckman, Ichimura and Todd (1998). This paper will be discussed in Section 2.5.

potential outcome, then she would be interested in knowing this parameter. This parameter is also useful in some applications where the objective is to estimate the turning point of a given relationship. For example, when estimating the turning point of the Environmental Kuznets Curve (EKC) for a particular pollutant, that is, the level of per capita income at which emissions of the pollutant reach their peak and start decreasing. In this case, the turning point of this relationship would be equivalent to a parameter such as (2.4). Finally, our third parameter gives the expected potential outcome at the optimal treatment dose. Note that this latter parameter could be combined with (2.3) to calculate, for example, the maximum expected gain from the treatment, or $E\{Y(\alpha_0) - Y(0)\}$.

In the binary treatment case the answers to the questions parameters (2.3)-(2.5) address are very direct. The dose-response function in this case is given by $E\{Y(1)\}$ and $E\{Y(0)\}$, which are always considered either explicitly or implicitly when analyzing binary treatments. Also, with a binary treatment the average treatment effect given by (2.1) tell us directly which of those two treatments (i.e., treated or non-treated) is optimal on average depending on the sign of $E(\Delta)$. Hence, whenever we are testing hypotheses about the sign of $E(\Delta)$ we are testing which of those treatments is optimal. Likewise, depending on the sign of $E(\Delta)$ our parameter in (2.5) would be given by either $E\{Y(1)\}$ or $E\{Y(0)\}$. Analysis of (2.3)-(2.5) becomes more complicated in the continuous treatment case. Here, the dose-response function is defined at each treatment level, and the answer to which treatment is optimal is not as direct as when the treatment is binary. Moreover, just as in the binary treatment case we are interested in testing hypotheses about which of the two treatments is optimal (i.e., about the sign of $E(\Delta)$), when the treatment is continuous we also want to be able to test hypotheses about α_0, or create confidence intervals for it. The estimators presented later in this chapter will allow us to do so.

After having defined our parameters, the next point to consider is their estimation. A common approach found in the applied literature when faced with a continuous treatment is to create a discrete number of categories. For example, when analyzing the effect of maternal age on birth outcomes, Royer (2003) creates the following maternal age groups: <18, 18-21, 22-25, 26-29, 30-33, 34-37 and >38 years. Some drawbacks of this approach are that very often discretization is arbitrary and that we lose information about the effects of the treatment within each of those groups. These problems are more severe if we follow that approach to estimate the location of the optimal treatment. For instance, Royer (2003) estimates that the best age for first births in terms of minimizing the likelihood of a premature birth is between 22 and 25 years old. First, we would not be able to say if giving birth between the ages of 22 and 25 is truly better than doing so between the ages of 23 and 26, or 21 and 24, since the best range of 22-25 depends on the way the maternal age variable was discretized. Second, given that such conclusion is based on estimates of the relevant parameters, we would expect a confidence level to be assigned to it; however, this is rarely done in practice. Another possibility for estimation of (2.3)-(2.5) is to assume a parametric form for the dose-response function. Unfortunately, we rarely have a clear idea of the form of the relationship between our treatment and outcome, and assuming an incorrect functional form would lead the estimated dose-response function to be inconsistent. Moreover, if the estimators of (2.4) and (2.5) are based on this estimated dose-response function, then they would also be inconsistent.

In this chapter, I use nonparametric methods for estimation of (2.3)-(2.5), which avoids imposing functional form restrictions on the relation between our treatment and outcome. Also, asymptotic normality of the estimators is shown so that we can create confidence intervals for our parameters. Nonparametric methods are based on the idea that if the function to be estimated is sufficiently smooth around a given point, then observations in its neighborhood can provide us with information about the value of the function at that particular point. In the next section, estimators for (2.3)-(2.5) are presented under the assumption that treatment doses are randomly assigned to the units in our sample. In section 2.4 I relax this assumption and consider the case when assignment to different treatment doses is random conditional on a set of observed covariates.

Before concluding this section, it is important to point out that the parameters in (2.3)-(2.5) are also well defined in the case when the treatment can take on a finite number of values, for example, when $\mathcal{T} = \{0, 1, ..., L\}$. It is possible to construct estimators for those parameters in this case. However, construction of confidence bounds for an optimal treatment is more difficult than when the treatment is binary because it would require making multiple comparisons. In this case, and provided that we have a reasonable number of different treatment values, we can think of our methods as a continuous approximation to the problem. Therefore, the methods presented in the rest of the chapter can also be useful even if the treatment is not strictly continuous.

2.3 Experimental Design

In this section I estimate (2.3)-(2.5) under the assumption that units are randomly assigned to different doses of the treatment, so that the set of potential outcomes for unit i is independent of the treatment assignment.

Assumption 2.3.1. $\{Y_i(t)\}_{t\in\mathcal{T}} \perp T_i$.

This assumption implies that $E[Y(t)] = E[Y|T=t]$ and therefore, in this experimental case estimating the dose response function is equivalent to estimating the unknown regression function of the outcome Y on the treatment level T. Similarly, estimating α_0 and $E\{Y(\alpha_0)\}$ is the same as estimating the location and size of the maximum of the unknown regression function $E[Y|T=t]$. Let $g_0(t) = E[Y|T=t]$ and assume we observe n pairs (y_i, t_i), $i = 1, \ldots, n$. Then, I estimate (2.3)-(2.5) respectively as

$$\widehat{g}_h(t) = \frac{\sum_{i=1}^{n} Y_i K(\frac{t-t_i}{h})}{\sum_{i=1}^{n} K(\frac{t-t_i}{h})} \quad \text{for all} \quad t \in \mathcal{T} \tag{2.6}$$

$$\widehat{\alpha} = \arg\max_{t\in\mathcal{T}} \widehat{g}_{h_1}(t) \tag{2.7}$$

$$\widehat{E}\{Y(\alpha)\} = \widehat{g}_{h_2}(\widehat{\alpha}) \tag{2.8}$$

where $K(\cdot)$ is a kernel function and $\widehat{g}_h$ is based on the bandwidth h. (2.6) is the Nadaraya-Watson (NW) estimator, which has been widely studied in the literature. Let $f_0(t)$ be the density of t and $\sigma_0^2(t) = Var[Y|T=t]$. Then, in this literature is shown that

$$\sqrt{nh}\left(\widehat{g}_h\left(t\right)-g_0\left(t\right)\right)\xrightarrow{d}\mathcal{N}\left(dB,\frac{\sigma^2(t)}{f(t)}\int[K(z)]^2dz\right) \tag{2.9}$$

provided that $h = h(n)$ satisfies $h \to 0$, $nh \to \infty$, and $nh^5 \to d^2$ (for some $0 \leq d < \infty$) as $n \to \infty$, among other conditions; and where B is a function of $g_0(t)$, $f_0(t)$ and the kernel function[5]. The optimal bandwidth for the NW estimator is proportional to $n^{-1/5}$, and the speed of convergence of the mean squared error using this bandwidth is $n^{-4/5}$. However, it is common in econometrics[6] to remove the bias in (2.9) by allowing $nh^5 \to 0$, which implies undersmoothing.

Estimators of the location and size of an optimal of a regression function based on kernel estimators of $g_0(t)$, such as the ones considered here, have been previously studied in the statistics literature. Müeller (1985) was the first one to analyze this type of estimators in the context of the non-random regressors model and using the Gasser-Müeller nonparametric estimator of $g_0(t)$. There, Müeller shows that his estimators of location and size of the peak of $g_0(t)$ are asymptotically jointly normal and uncorrelated. Later, Müeller (1989) allowed for data-dependent (i.e., random) bandwidths, and showed that the asymptotic distribution of those estimators using consistent estimates of the optimal bandwidths is the same as the one using optimal bandwidths. In contrast to the papers by Müeller, Ziegler (2000) focuses on the random regressor model and employs the NW estimator. Like Müeller (1989), Ziegler allows for random bandwidths and establishes a functional central limit theorem for the joint distribution of the estimators of location and size of the maximum. In contrast to Müeller, his conditions are imposed locally on a neighborhood of the location of the maximum

[5]See, for example, Bierens (1987) or Pagan and Ullah (1999).

[6]See, for example, Newey (1994), Ahn (1995) and Pagan and Ullah (1999).

rather than globally on a compact interval. In general, Ziegler's results are similar to those in Müeller (1989). Both authors consider the case with a single covariate, but it is straightforward to extend their results to maximization over more than one dimension[7].

An important conclusion from analyzing the asymptotic behavior of estimators such as (2.7) and (2.8) is that if we want both of them to be jointly asymptotically normal, then they should not be based on the same estimator of $g_0(\cdot)$. Intuitively, for a given estimator $\widehat{g}(\cdot)$ of $g_0(\cdot)$, $\widehat{\alpha}$ will solve $\partial\widehat{g}(t)/\partial t = 0$ (assuming an interior maximum), so the asymptotic behavior of $\widehat{\alpha}$ will be closely related to that of the estimator of the first derivative of $g_0(\cdot)$. On the other hand, the asymptotic behavior of (2.8) will be closer to that of the usual NW estimator of a regression function. As will be discussed later, one of our conditions on the bandwidth used for estimation of the location of the maximum requires $nh_1^6 \to \infty$. If we were to use this bandwidth in (2.8) for estimation of the size of the maximum, then the asymptotic bias of the estimator would explode, as illustrated by (2.9). The way Müeller (1989) deals with this is by allowing the bandwidths of his estimators of location and size of the maximum to go to zero at different speeds. On the other hand, Ziegler (2000) uses kernels of different orders for both parameters while using bandwidths of the same order for both. As suggested by the notation in (2.7) and (2.8), this section follows an approach similar to the one by Müeller, but now using the NW estimator and allowing the regressor

[7]Although the literature on estimation of the maximum of a regression function is not large, the opposite is true for the related problem of estimating the mode of a density using kernel methods (e.g., Parzen 1962, Eddy 1980, 1982; Romano 1988). On the other hand, there are other approaches in the statistics literature for estimating the location of a maximum of a regression function in this experimental case. Some involve algorithms detecting peaks (e.g., Heckman 1992) and the use of extreme order statistics (e.g., Chen et al. 1996). We prefer our approach because its extension to the non-experimental case is more natural. However, the other approaches will be taken into consideration for future research.

to be random[8,9].

This section presents a result which is similar in spirit to those by Müeller (1989) and Ziegler (2000). This result is helpful to get the intuition of the problem at hand and extend the results to the case where one needs to control for non-random selection into different doses based on observables, as will be discussed in the next section. For simplicity, and in order to highlight the main ideas, a second order kernel is used in the theorem below[10]. Our set up here is similar to the one considered by Ziegler; however, as previously mentioned, instead of using different kernels for estimation of the location and size of the maximum as he does, I employ bandwidths of different order as in Müeller. Another slight difference of the results presented below and those by Müeller, in addition to the set up and the kernel estimator used, is that Müeller requires the use of higher order kernels while our results are also valid for second order kernels. The next theorem shows that the estimators of the location and size of the optimum dose presented in (2.7) and (2.8) are asymptotically jointly normal and uncorrelated, so that they can be used to construct confidence bounds and perform statistical inference in this experimental case. The asymptotic distribution of the estimators has been centered around zero by choice of bandwidth, so the result implies undersmoothing.

Theorem 1 *Assume*

(i) Assumption 2.3.1. holds.

[8]As previously mentioned, Müeller considers the fixed design model and uses the Gasser-Müeller estimator.

[9]Whether is preferable in practice to use bandwidths or kernels of different orders for estimation of the location and size of the maximum of a regression function is unknown. This question will be analyzed in the future using simulation methods.

[10]The extension of Theorem 1 to higher order kernels is straightforward. Moreover, the theorem presented in the next section under the selection-on-observables assumption and using higher order kernels can be easily reduced to this experimental case.

(ii) $\alpha \in \mathcal{T}$, *where* $\mathcal{T}$ *is compact and* α_0 *is in the interior of* $\mathcal{T}$.

(iii) $g_0(t)$ *is uniquely maximized at* α_0 *and is three times continuously differentiable and its derivatives up to the third order are bounded. Also,* $g_0^{(2)}(\alpha_0) < 0$.

(iv) $f_0(t)$ *(density of* t*) is continuous and bounded away from zero uniformly in* $\mathcal{T}$. *Also, partial derivatives of* $f_0(t)$ *exists up to the third order and are bounded, and* $f_0^{(1)}(t)$ *is continuous at* α_0.

(v) $\sigma_0^2(\alpha_0) = Var\,[Y|T = \alpha_0]$ *is bounded and its derivative is continuous at* α_0. *Also, partial derivatives of* $\sigma_0^2(t)$ *exists up to the third order and are bounded.*

(v) $E\,|Y|^{2+\delta} < \infty$ *for some* $\delta > 0$.

(vi) Let $h_1 = h_1(n)$, $h_2 = h_2(n)$ *be such that:* $h_1 \to 0$, $nh_1^6 \to \infty$, $nh_1^7 \to 0$; *and* $h_2 \to 0$, $nh_2^4 \to \infty$, $nh_2^5 \to 0$, *as* $n \to \infty$.

(vii) The kernel $K(\cdot)$ *satisfies: (a)* $\int K(u)du = 1$; *(b)* K *is symmetric and three times continuously differentiable; (c)* $\int u^2 K(u)du < \infty$ *and* $\int |uK(u)|\,du < \infty$; *(d)* $|u^3|\,|K(u)| \to 0$ *as* $|u| \to \infty$; *(e) Let* $\psi(w)$ *be the characteristic function of* $K(\cdot)$ *so that* $\psi(w) = \int e^{iwu}K(u)du$, *with* $i^2 = -1$. *Assume that* $\int |\psi(w)|\,dw < \infty$ *and* $\int |w^2\psi(w)|\,dw < \infty$; *(f) For some* $\delta > 0$, *and for* $H(u) = K(u)$ *and* $H(u) = K^{(1)}(u)$ *we have that* $\int |H(u)|^{2+\delta}\,du < \infty$, $\sup_u |H(u)|^{2+\delta} < \infty$, *and* $|u|\,|H(u)|^{2+\delta} \to 0$ *as* $|u| \to \infty$.

Then,

$$\begin{pmatrix} \sqrt{nh_1^3}(\widehat{\alpha} - \alpha_0) \\ \sqrt{nh_2}(\widehat{g}_{h_2}(\widehat{\alpha}) - g_0(\alpha_0)) \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} \right) \tag{2.10}$$

with $\sigma_1^2 = \frac{\sigma_0^2(\alpha_0)}{[g_0^{(2)}(\alpha_0)]^2\; f_0(\alpha_0)} \int [K^{(1)}(z)]^2 dz$ *and* $\sigma_2^2 = \frac{\sigma_0^2(\alpha_0)}{f_0(\alpha_0)} \int [K(z)]^2 dz$; *and where the superscript* (j) *denotes the* j*th derivative with respect to its argument.*

PROOF. See appendix.

Asymptotic normality of the estimator of the location of the maximum follows from noting that under assumption (i) $\widehat{\alpha}$ satisfies $\widehat{g}_{h_1}^{(1)}(\widehat{\alpha}) = 0$, so that expanding this expression around α_0 and solving for $\widehat{\alpha} - \alpha_0$ gives

$$\widehat{\alpha} - \alpha_0 = -\frac{\widehat{g}_{h_1}^{(1)}(\alpha_0)}{\widehat{g}_{h_1}^{(2)}(\alpha^*)} \tag{2.11}$$

where α^* is a mean value. Thus, the three key ingredients in showing asymptotic normality of $\sqrt{nh_1^3}(\widehat{\alpha} - \alpha_0)$ are consistency of $\widehat{\alpha}$, uniform convergence in probability of $\widehat{g}_{h_1}^{(2)}(t)$ to $g_0^{(2)}(t)$ for all $t \in \mathcal{T}$, and asymptotic normality of $\sqrt{nh_1^3}\ \widehat{g}_{h_1}^{(1)}(\alpha_0)$. Consistency follows easily by noting that $\widehat{\alpha}$ is an extremum estimator with objective function $\widehat{g}_{h_1}(t)$, $t \in \mathcal{T}$, so standard results can be used here (e.g., Theorem 2.1 in Newey and McFadden,1994). Results regarding asymptotic normality of the numerator and uniform convergence of the denominator in (2.11) can also be found in the literature (e.g., Schuster and Yakowits, 1979; Ahmad and Ullah, 1987). The requirements on the kernel along with the smoothness conditions imposed on $g_0(t)$, $f_0(t)$ and $\sigma_0^2(t)$ in Theorem 1 are useful when applying those results. The assumption that $nh_1^6 \to \infty$ is used for showing uniform convergence of the denominator, and $nh_1^7 \to 0$ is used for obtaining asymptotic normality of the numerator. Like in the usual NW estimator (see (2.9)), this latter assumption implies undersmoothing and the use of a suboptimal bandwidth for estimating the first derivative of $g_0(t)$[11]

The proof of the joint asymptotic normality result proceeds by using the Cramér-Wold device to show that for every real numbers λ_1 and λ_2 we have:

[11]The optimal bandwidth for estimating the first derivative of $g_0(\cdot)$ is of order $n^{-1/7}$, in which case the speed of convergence of the MSE is $n^{-4/7}$.

$$\lambda_1\sqrt{nh_1^3}(\widehat{\alpha}-\alpha_0)+\lambda_2\sqrt{nh_2}(\widehat{g}_{h_2}(\widehat{\alpha})-g_0(\alpha_0)) \xrightarrow{d}$$

$$\mathcal{N}\left(0,\frac{\lambda_1^2\sigma_0^2(\alpha_0)}{[g_0^{(2)}(\alpha_0)]^2\ f_0(\alpha_0)}\int[K^{(1)}(z)]^2dz+\frac{\lambda_2^2\sigma_0^2(\alpha_0)}{f_0(\alpha_0)}\int[K(z)]^2dz\right) \tag{2.12}$$

In obtaining (2.12), it is important to show that $\sqrt{nh_2}(\widehat{g}_{h_2}(\widehat{\alpha})-g_0(\alpha_0))$ is asymptotically equivalent to $\sqrt{nh_2}(\widehat{g}_{h_2}(\alpha_0)-g_0(\alpha_0))$. It is in this step where Müeller requires the use of higher order kernels, while Theorem 1 shows that the same also holds true even when using second order kernels. For showing this result, Müeller requires uniform convergence of $\widehat{g}_{h_2}^{(2)}(\cdot)$ to $g_0^{(2)}(\cdot)$, which in turn requires $nh_2^6\to\infty$ and hence the use of higher order kernels to reduce the asymptotic bias. On the other hand, the proof of Theorem 1 only requires $\sqrt{h_2}\widehat{g}_{h_2}^{(2)}(\cdot)$ to be asymptotically bounded in a neighborhood of α_0, which imposes weaker conditions on h_2 and on the order of the kernel. Using the mean value theorem we can write for some α^* between $\widehat{\alpha}$ and α_0,

$$\sqrt{nh_2}\,(\widehat{g}_{h_2}(\widehat{\alpha})-g_0(\alpha_0))=\sqrt{nh_2}\,(\widehat{g}_{h_2}(\alpha_0)-g_0(\alpha_0))+\sqrt{nh_2}\,\widehat{g}_{h_2}^{(1)}(\alpha^*)(\widehat{\alpha}-\alpha_0)$$

Hence, we need to show that the second term to the right of the above equation is $o_p(1)$. Again, for a suitable mean value α^{**} between α_0 and α^* we can write

$$\begin{aligned}\sqrt{nh_2}\widehat{g}_{h_2}^{(1)}(\alpha^*)(\widehat{\alpha}-\alpha_0) &= \frac{1}{\sqrt{nh_1^3h_2^2}}\sqrt{nh_2^3}\,\widehat{g}_{h_2}^{(1)}(\alpha_0)\sqrt{nh_1^3}(\widehat{\alpha}-\alpha_0)\\ &\quad+\frac{1}{\sqrt{nh_1^6}}\sqrt{h_2}\,\widehat{g}_{h_2}^{(2)}(\alpha^{**})\sqrt{nh_1^3}(\alpha^*-\alpha_0)\sqrt{nh_1^3}(\widehat{\alpha}-\alpha_0)\end{aligned} \tag{2.13}$$

As previously discussed, $\sqrt{nh_1^3}(\widehat{\alpha}-\alpha_0)=O_p(1)$. Similarly, using standard results it can be shown that $\sqrt{nh_2^3}\,\widehat{g}_{h_2}^{(1)}(\alpha_0)=O_p(1)$ (e.g., Ahmad and Ullah, 1988). The conditions on h_2 for the latter result are that $nh_2^3 \to \infty$ and $nh_2^7 \to 0$, which clearly do not contradict our assumptions on h_2. Given our assumptions on h_1 and h_2, the first term to the right of (2.13) is $o_p(1)$. Note that given that $nh_1^6 \to \infty$, in order to show that the second term to the right of (2.13) is $o_p(1)$ we only require $\sqrt{h_2}\ \widehat{g}_{h_2}^{(2)}(\alpha^{**}) = O_p(1)$. This imposes weaker restrictions on h_2 than requiring $\widehat{g}_{h_2}^{(2)}(\alpha^{**})$ to converge in probability to $g_0^{(2)}(\alpha_0)$, as Müeller does. Note that the assumption $nh_2^4 \to \infty$ is used in showing that the terms to the right of (2.13) are $o_p(1)$, and the one requiring $nh_2^5 \to 0$ is used to show asymptotic normality of the estimator of the size of the maximum, as in (2.9).

The requirement of the kernel being symmetric is key for the asymptotic uncorrelatedness of $\widehat{\alpha}$ and $\widehat{g}_{h_2}(\widehat{\alpha})$. Specifically, symmetry of $K(\cdot)$ implies that $\int K^{(1)}(u) K(u)\, du = 0$, which along with the bandwidth conditions and the smoothness assumptions on $g_0(t)$, $f_0(t)$ and $\sigma_0^2(\alpha_0)$ implies that the covariance term goes to zero asymptotically (see Lemma 8 in appendix A).

The asymptotic variances of the estimators of the location and size obtained in Theorem 1 are very intuitive. For the estimator of the size, the asymptotic variance is exactly the same as for the usual NW estimator evaluated at the maximum, α_0 (see (2.9)). As for the estimator of the location of the maximum, the asymptotic variance is the same as the variance of the estimator of the first derivative of $g_0(\cdot)$ evaluated at α_0 plus the additional term $[g_0^{(2)}(\alpha_0)]^{-2}$. This last term is a measure of the curvature of $g_0(\cdot)$ at α_0; so, as one would expect, the greater the curvature of $g_0(\cdot)$ at α_0 the easier would be to estimate the location of the maximum. In other words, if the maximum is located in a region of $g_0(\cdot)$ where it increases and decreases very rapidly so that $g_0^{(2)}(\alpha_0)$ is large (in absolute value), then the asymptotic variance of our estimator of α_0 will be small.

Also, note the bandwidth for estimation of the location of the maximum converges to zero slower than the one for estimation of the size of the maximum. In other words, in the first case we oversmooth as compared to the second case. The intuition is that if our estimate of $g_0(\cdot)$ is not sufficiently smooth, then it could be very difficult to estimate the first derivative of $g_0(\cdot)$ and consequently α_0. The conditions on h_1 and h_2 imply that they should be proportional to n^{δ} and n^{γ}, where $\delta \in (-1/6, -1/7)$ and $\gamma \in (-1/4, -1/5)$, respectively.

Finally, it is important to point out that the estimators of the location and size of the maximum based on the usual NW estimator (i.e., using a bandwidth of order $n^{-1/5}$) are both consistent. Say, as is usually done in practice (e.g., Millimet et al. 2003), that we were to use the usual NW estimator based on bandwidth h to estimate the location and size of the maximum. As previously mentioned $\widehat{\alpha}$ is an extremum estimator, so the only conditions imposed on h for consistency of $\widehat{\alpha}$ are those needed for uniform convergence of $\widehat{g}_h(t)$ to $g_0(t)$, which are that $h \to 0$ and $nh^2 \to \infty$ as $n \to \infty$ (e.g., Bierens, 1987). The same restrictions on h make the estimator of the size of the maximum consistent in this case. Therefore, the advantage of the estimators presented in this section over those based on the usual NW estimator is that the former ones are not only consistent but also asymptotically normal, so they can be used to create confidence intervals and test hypothesis regarding our parameters of interest.

2.4 Non-experimental Design: Selection on Observables

In economics usually we do not have an experiment at hand to evaluate the effects of a given treatment. A common approach in the binary-treatment literature and a natural "next step" when analyzing the effects of a given treatment is to assume that selection into treatment is based on a given set of observed covariates[12]. In this section I follow a similar approach and assume that assignment into different levels of the treatment is unconfounded given a set of covariates X with dimension equal to k, that is, I assume that selection is based on observables.

Assumption 2.4.1. $\{Y_i(t)\}_{t\in\mathcal{T}} \perp T_i|X$.

As discussed in Hirano and Imbens (2004) and Imbens (2000), assumption 2.4.1 is stronger than needed and can be replaced by a weaker version of unconfoundedness in which all that is required is pairwise conditional independence of each of the potential outcomes at a given treatment dose t with the treatment assignment, or $Y_i(t) \perp T_i|X$ for all $t \in \mathcal{T}$. However, as also pointed out in Imbens (1999), in practice can be difficult to find applications in which the latter may be plausible but the stronger form in assumption 2.4.1 may not. Because of this, the stronger form of the unconfoundedness assumption is maintained in this section. Assumption 2.4.1 implies that we can write the dose-response function as

$$E[Y(t)] = E_X[E[Y(t)|X=x]] = E_X[E[Y(t)|T=t, X=x]] = E_X[E[Y|T=t, X=x]] \tag{2.14}$$

[12]See for example Heckman, Lalonde and Smith (1999) and Imbens (2004).

for all $t \in \mathcal{T}$, where the unconfoundedness assumption is used in the second equality. Hence, equation (2.14) states that, under assumption 2.4.1, we can express the dose response function as a function of the observed data. In the experimental case discussed in the previous section, randomization controlled for observed and unobserved confounders by not allowing their values to differ systematically across different treatment doses, and therefore, we were able to write the dose response function as the regression function of the observed outcome, Y, on the treatment level received, T. On the other hand, in the non-experimental case considered in this section, and under assumption 2.4.1., we need to control for systematic differences in the observed covariates across treatment doses, and $E_X [E[Y|T = t, X = x]]$ does so by averaging over them[13]. Note that this latter expression suggests calculating the dose-response function by first computing the regression function of the observed outcome (Y) on the observed treatment (T) and covariate values (X) and then taking its expectation over the covariates. Hence, this suggests a regression approach to estimating the dose-response function. However, just as in the binary case we could apply other methodologies such as matching on the covariates or weighting by the propensity score to estimate the average treatment effect, here we can think of other ways to estimate the dose-response function. Extensions to the continuous treatment case of the methods previously mentioned for the binary case will be discussed later in section 2.5.1 in the context of the role the generalized propensity score plays when the treatment is continuous. For the moment, the focus is on the properties of the regression approach suggested by (2.14).

The last term in (2.14) is what Newey (1994) calls a partial mean, which is an average over some conditioning variables while holding others fixed. Thus, the estimators of

[13]Note that in general $E_X [E[Y|T = t, X = x]] \neq E[Y|T = t]$.

the location and size of the optimal treatment dose analyzed in this section are also useful in the more general context of estimating the location and size of the maximum of a partial mean.

Assume we observe (y_i, t_i, x_i), $i = 1, \ldots, n$. Also, let $\tau(\cdot)$ be a trimming function used to avoid the "denominator problem" by keeping a denominator bounded away from zero. Based on the sample analogue of the last term in (2.14), the non-parametric estimators of the parameters of interest given by (2.3)-(2.5) are defined, respectively, as

$$\widehat{E}\{Y(t)\} = \frac{1}{n}\sum_{i=1}^{n} \tau(x_i)\ \widehat{g}_h(t, x_i) \quad \text{for all} \quad t \in \mathcal{T} \tag{2.15}$$

$$\widehat{\alpha} = \arg\max_{t \in T} \frac{1}{n}\sum_{i=1}^{n} \tau(x_i)\ \widehat{g}_{h_1}(t, x_i) \tag{2.16}$$

$$\widehat{E}\{Y(\alpha_0)\} = \widehat{\mu}_{h_2}(\widehat{\alpha}) = \frac{1}{n}\sum_{i=1}^{n} \tau(x_i)\ \widehat{g}_{h_2}(\widehat{\alpha}, x_i) \tag{2.17}$$

where $\widehat{g}_h(t, x)$ is the NW multiple regression estimator

$$\widehat{g}_h(t, x) = \frac{\sum_{j=1}^{n} Y_j K(\frac{t-t_i}{h}, \frac{x_1 - x_{1i}}{h}, \ldots, \frac{x_k - x_{ki}}{h})}{\sum_{j=1}^{n} K(\frac{t-t_i}{h}, \frac{x_1 - x_{1i}}{h}, \ldots, \frac{x_k - x_{ki}}{h})} \tag{2.18}$$

and as before, $K(\cdot)$ is a kernel function and $\widehat{g}_h$ is based on the bandwidth h.

The results presented in this section are based on the asymptotic theory developed by Newey (1994) on functionals of kernels estimators. Specifically, Newey considers two-step estimators where the first step is a vector of kernel estimators, say $\widehat{s}(t, x)$, and the second step is a m-estimator that depends on $\widehat{s}(t, x)$. He shows that under some regularity conditions these estimators are asymptotically normal. Newey also applies his general results to derive asymptotic normality of a partial mean estimator as the one in (2.15), which is an average of a kernel regression estimator over some components while holding others fixed. The general results in Newey (1994) are also useful for showing joint asymptotic normality of the estimators of location and size of the optimal treatment in (2.16) and (2.17), as discussed below.

Theorem 2 below, which is a special case of Theorem 4.1 in Newey (1994), shows asymptotically normality of the estimator in (2.15). Redefine $E\left[Y(t)\right]$ as $E\left[Y(t)\right] = E_X\left[\tau(x)\right.$ $\left.E\left[Y|T=t, X=x\right]\right]$. Also, let $w = (t,x)$, r be the order of the kernel, $\widetilde{f}_0(x)$ be the marginal density of x and $f_0(w) = f_0(t,x)$ be the joint density of t and x. Finally, remember the dimension of X is given by k.

Theorem 2 *(Newey, 1994). Assume*

(i) $E\left[|y|^4\right] < \infty$; $E\left[|y|^4|w\right] f_0(w)$ *and* $f_0(w)$ *are bounded.*

(ii) Let $K(\cdot)$ *be such that* $\int K(u)du = 1$; $K(u)$ *is zero outside a bounded set;* $K(u)$ *is continuously differentiable, with Lipschitz derivative; and, there is a positive integer* r *such that for all* $j < r$, $\int K(u)\left[\ {}_{\ell=1}^{j} u^{\ell}\right] = 0$.

(iii) There is a non-negative integer $d \geq r$ *and an extension of* $s_0(w)$ *to all of* $\Re^{k+1}$ *that is continuously differentiable to order* d *on* $\Re^{k+1}$.

(iv) $\tau(x)$ *is bounded and zero except on a compact set where* $f_0(t, x_i)$ *is bounded away from zero.*

(v) $\tau(x)$ *and* $\widetilde{f}_0(x)$ *are continuous a.e.,* $\widetilde{f}_0(x)$ *is bounded,* $E[y|w]$ *and* $E[y^2|w]$ *are continuous; and, for some* $\varepsilon > 0$,

$$\int \sup_{||\eta||<\varepsilon}\left[\left\{1 + E\left[y^4|w = (t+\eta, x)\right]\right\} f_0(t+\eta, x)\right] dx < \infty.$$

(vi) For $h = h(n)$*,* $nh^{2k+1}/[\ln(n)]^2 \to \infty$ *and* $nh^{2r+1} \to 0$*, as* $n \to \infty$*.*

Then,

$$\sqrt{nh}(\widehat{E}\{Y(t)\} - E\{Y(t)\}) \xrightarrow{d} \mathcal{N}(0, V_0)$$

with $V_0 = \left[\int \{\int K(u,v)dv\}^2 du\right] \times \int f_0(t,x)^{-1}\tau^2(x)\,\widetilde{f}_0^2(x)\sigma_0^2(t,x)\,dx$*; where* $K(w)$ *is partitioned according to* $w = [t,x]$ *and* $\sigma^2(t,x) = var\,[y|t,x]$*.*

As discussed in Newey (1994), since $E_X[E[Y|T=t, X=x]]$ is only a function of T, its nonparametric estimators will converge faster than estimators of $E[Y|T=t, X=x]$. This is reflected in the conclusion of Theorem 2, where the non-parametric estimator of the dose-response function defined in (2.15) has a convergence rate of $\sqrt{nh}$, which is the same rate at which the corresponding estimator in the experimental case considered in the previous section converged. The bandwidth conditions in theorem 2 imply undersmoothing, which is reflected in the fact that the limiting distribution is centered around zero. Finally, note that the theorem requires the use of higher order kernels, which also helps ensuring the limiting distribution is centered at the true value by reducing the bias of the estimator. Specifically, theorem 2 requires $r > k$.

The variance of the limiting distribution in Theorem 2 can be simplified in some cases. For example, if we use a product kernel for calculation of the estimator, then the first term in V_0 reduces to $\int [K(u)]^2 du$, which is the same kernel term that appears in the limiting distribution of the usual NW estimator (see 2.9). Moreover, ignore for the moment the trimming function in V_0. Then, we can write the second term in V_0 as $E_X\left[\sigma_0^2(t,x) \;/\; f_0(t|x)\right]$, where $f(t|x)$ is the conditional density of t given x. This term is similar to the other term appearing in the variance of the limiting distribution of the NW estimator in (2.9), but now allowing the conditional variance of Y to include also the covariates and dividing by the conditional density of t given x (instead of dividing by only $f_0(t)$ as in (2.9)), and then taking expectation over X. Hence, V_0 seems like a natural extension of the asymptotic variance for the NW estimator to the partial mean case.

Next, I consider the limiting joint distribution of the estimators of location and size of the maximum of the dose-response function in (2.16) and (2.17). As mentioned before, here the general results on functionals of kernel estimators presented in Newey (1994) are used. In order to highlight the main steps in the proofs of the theorems in this section we focus first on the asymptotic distribution of the estimator of the optimal dose in (2.16). Moreover, in some applications only the location of the optimal treatment dose may be of interest, so in this case theorem 3 can be used directly.

To use the general results in Newey (1994) the estimator in (2.16) needs to be written as a two-step estimator for which the first step involves a vector of kernel estimators. To simplify notation, ignore for the moment the trimming function $\tau(\cdot)$. Letting $q = [1\ y]'$, and using the same notation as before, define $s_0(w)$ as $s_0(w) = E[q|w]f_0(w) = [f_0(w)\ \ E[y|w]f_0(w)]' = [s_{10}(w)\ \ s_{20}(w)]'$. Let $z_i = (q_i, w_i)$, $i = 1, \ldots, n$, denote data observations on q and w. Then, a kernel estimator of $s_0(w)$ is

$$\widehat{s}(w) = \frac{1}{n}\sum_{j=1}^{n} q_j K_h(w - w_j) = \begin{bmatrix} \frac{1}{n}\sum_{j=1}^{n} K_h(w - w_j) \\ \frac{1}{n}\sum_{j=1}^{n} y_j K_h(w - w_j) \end{bmatrix} = \begin{bmatrix} \widehat{s}_1(w) \\ \widehat{s}_2(w) \end{bmatrix} \tag{2.19}$$

where $K_h(u) = h^{-(k+1)}K(u/h)$. This is the first step kernel estimator. Now, let $g_0(t, x) = g_0(w) = E[y|w] = s_{20}(w)/s_{10}(w)$. Then, by definition of α_0 in (2.2), and assuming an interior maximum for the dose-response function, α_0 solves

$$\left.\frac{\partial E\left[Y(t)\right]}{\partial t}\right|_{t=\alpha_0} = \left.\frac{\partial E_X\left[g_0(t,x)\right]}{\partial t}\right|_{t=\alpha_0} = E_X\left[\left.\frac{\partial g_0(t,x)}{\partial t}\right|_{t=\alpha_0}\right] = 0 \tag{2.20}$$

where (2.14) was used in the first equality. Let $m(z,\alpha_0,s_0) = \partial g_0(\alpha_0,x)/\partial t$. Then, the moment condition implied by (2.20) is $E\left[m(z,\alpha_0,s_0)\right] = 0$. In this case the sample moment function becomes $m(z_i,\alpha,\widehat{s}) = \partial\widehat{g}(\alpha,x_i)/\partial t$, so the second-step m-estimator of α_0 solves:

$$\frac{1}{n}\sum_{i=1}^{n} m(z_i,\alpha,\widehat{s}) = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial\widehat{g}(\alpha,x_i)}{\partial t} = 0 \tag{2.21}$$

where $\widehat{g}(w) = \widehat{s}_2(w)/\widehat{s}_1(w)$ is the NW estimator in (2.18). Therefore, under the assumption of an interior maximum of the dose-response function the estimator of the optimal dose $\widehat{\alpha}$ in (2.16) can be written as a two-step m-estimator that solves (2.21).

As usual for m-estimators, to derive the limiting distribution of $\widehat{\alpha}$ the left side of (2.21) is expanded around the true value α_0 to get

$$\widehat{\alpha} - \alpha_0 = -\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial m(z_i,\alpha^*,\widehat{s})}{\partial\alpha}\right]^{-1}\left[\frac{1}{n}\sum_{i=1}^{n} m(z_i,\alpha_0,\widehat{s})\right] \tag{2.22}$$

where α^* is a mean value. Consistency of $\widehat{\alpha}$ and uniform convergence in probability of the Jacobian term in (2.22) are useful in showing convergence in probability of the denominator in (2.22) to some matrix M, which is assumed to be invertible. Then, the limiting distribution of $\widehat{\alpha} - \alpha_0$ is determined by the behavior of the numerator in (2.22). Note that in this latter term the moment function depends on the kernel estimator $\widehat{s}$ in (2.19). Therefore, its limiting distribution is derived in two steps. The first one involves a linearization around s_0, and the second entails asymptotic normality of such linearization.

The following theorem specifies conditions for asymptotic normality of $\widehat{\alpha}$. As in the case for estimation of the dose-response function described before, in this case higher order kernels are also used to reduce the bias of the estimator and center its asymptotic distribution around zero. As before, let $E\left[Y(t)\right]=E_X\left[\tau\left(x\right)E\left[Y|T=t,X=x\right]\right]$, where $\tau\left(\cdot\right)$ is the trimming function previously defined.

Theorem 3 *Assume*

(i) $\alpha\in\mathcal{T}$, *where* $\mathcal{T}$ *is compact and* α_0 *is in the interior of* $\mathcal{T}$.

(ii) Let $K(\cdot)$ *be such that* $\int K(u)du=1$; $K(u)$ *is zero outside a bounded set;* $K(u)$ *is twice continuously differentiable, with Lipschitz derivatives; and, there is a positive integer* r *such that for all* $j<r$, $\int K(u)\left[\ _{\ell=1}^{j}u^{\ell}\right]=0$.

(iii) There is a non-negative integer $d\geq r+1$ *and an extension of* $s_0(w)$ *to all of* $\Re^{k+1}$ *that is continuously differentiable to order* d *on* $\Re^{k+1}$.

(iv) $\tau\left(x\right)$ *is bounded, continuous almost everywhere and zero except on a compact set where* $f_0\left(t,x_i\right)$ *is bounded away from zero.*

(v) $E\left[||m(z,\alpha_0,s_0)||^2\right]<\infty$; $E\left[|y|^4\right]<\infty$; $E\left[|y|^4|w\right]f_0(w)$ *and* $f_0(w)$ *are bounded; .*

(vi) $\widetilde{f}_0(x)$ *is zero outside a compact set* $\mathcal{X}$ *and is continuous almost everywhere and bounded;* $E[y|w]$ *and* $E[y^2|w]$ *are continuous; and, for some* $\varepsilon>0$,

$$\int\sup\nolimits_{||\eta||<\varepsilon}\left[\left\{1+E\left[y^4|w=(\alpha_0+\eta,x)\right]\right\}f_0\left(\alpha_0+\eta,x\right)\right]dx<\infty.$$

(vii) $\partial^2E\left[y|w=(\alpha,x)\right]/\partial t^2$ *is continuous at each* $t\in\mathcal{T}$ *with probability one;* $E\left[\left\{\partial^2\quad E(y|\alpha_0,x)/\partial t^2\right\}^2\right]<\infty$; *and,* $E\left\{\sup\limits_{\alpha\in\mathcal{T}}\left|\partial^2E\left[y|w=(\alpha,x)\right]/\partial t^2\right|\right\}<\infty$.

(viii) $E\left[\tau\left(x\right)\partial^2E\left[y|w=(\alpha_0,x)\right]/\partial^2t\right]$ *is invertible.*

(ix) $E_X[\tau(x) E[Y|T=t, X=x]]$ *is continuous and uniquely maximized at* α_0*;* $E\left[\sup_{\alpha\in\mathcal{T}} |E[y|w=(\alpha,x)]|\right] < \infty$*;* $E\left[\{E(y|\alpha_0,x)\}^2\right] < \infty$.

(x) Let $h_1 = h_1(n)$ *be such that:* $h_1 \to 0$, $nh_1^{k+5}/\ln(n) \to \infty$, $nh_1^{2k+3}/[\ln(n)]^2 \to \infty$, $nh_1^{2r+3} \to 0$.

Then,

$$\sqrt{nh_1^3}(\widehat{\alpha}-\alpha) \xrightarrow{d} \mathcal{N}(0, V_1)$$

with $V_1 = d^{-2}\left[\int\{\int K^{(1)}(u,v)dv\}^2 du\right] \times \int f_0(\alpha_0,x)^{-1}\tau^2(x)\widetilde{f}_0^2(x)\sigma^2(\alpha_0,x)\,dx$ *and* $d = \partial^2 E\{\tau(x) E[Y|T=\alpha_0, X=x]\}/\partial t^2$*; and where* $K(w)$ *is partitioned according to* $w=[t,x]$*, and* $K^{(1)}(\cdot)$ *means the partial derivative with respect to* t.

PROOF. See appendix.

Assumptions (vii) and (ix) are useful in showing uniform convergence in probability of the Jacobian term in (2.22) and of the objective function that $\widehat{\alpha}$ maximizes when seen as an extremum estimator. These two results are in turn useful to prove convergence in probability of the denominator in (2.22) to $E\left[\tau(x)\partial^2 E[y|w=(\alpha_0,x)]/\partial^2 t\right]$, which by assumption (viii) is invertible. The dominance condition in assumption (vi) integrates over the covariates and is used when showing asymptotic normality of the linearization of $n^{-1}\sum m(z_i,\alpha_0,\widehat{s})$ in (2.22) around s_0. The condition requiring $nh_1^{k+5}/\ln(n) \to \infty$ is important in showing uniform convergence in probability of the Jacobian term in (2.22). Note that if $k=0$ then this condition is analogous to the one used in the previous section to show uniform convergence of the second derivative of the NW estimator. The assumption $nh_1^{2k+3}/[\ln(n)]^2 \to \infty$ is used for linearization of the numerator in (2.22) around s_0. The last bandwidth condition, $nh_1^{2r+3} \to 0$, implies undersmoothing and is used to center the asymptotic distribution around the true value α_0. In general, the conditions on the bandwidth imply that the order of the kernel used has to be strictly greater than the number of covariates used, i.e., $r > k$. Hence, the result in theorem 3 holds true even when using a second order kernel in the presence of a single covariate; however, as is always the case, the use of higher order kernels would reduce bias and increase the speed of convergence of the mean square error. Finally, the conditions on the bandwidth imply that if h_1 is proportional to n^{γ}, then $\gamma \in (\min(-1/(2k+3), -1/(k+5)), -1/(2r+3))$. If $k = 0$, this interval reduces to the one we had before in the experimental case[14].

Note that as in the case for partial means discussed before, in estimation of the optimal treatment the rate achieved by the estimator is the same as in the experimental case (i.e., as in the case without covariates). As before, this happens because of the averaging over the covariates of the non-parametric regression of Y on T and X. The variance of the limiting distribution in theorem 3 is not very different from the one obtained in theorem 1, but now it allows for the presence of covariates in the estimation. For example, in theorem 1 we had the square of the second derivative of the objective function for α in the denominator; and in theorem 3 that term is given by d. Also, note that if product kernels are employed then the kernel term in theorem 3 reduces to the one in theorem 1. Finally, as for partial means in theorem 2, the last term in V_1 can be written as $E_X\left[\sigma_0^2(t,x) \; / \; f_0(t|x)\right]$, which is an extension to the case with covariates of the corresponding term in theorem 1.

Now consider the joint limiting distribution of the estimators of location and size of the optimal treatment in (2.16) and (2.17). As in section 2.3, to obtain the asymptotic joint

[14]The interval is not exactly the same as before since in this non-experimental case we are using bounded support kernels, and in the experimental case they had unbounded support. Therefore, the conditions in the former have an extra "$\log n$" term.

normality result shown below both estimators should not be based on the same estimator of the dose-response function. The intuition is the same as in the experimental case, that is, if the same estimator of the dose-response function used for estimation of the optimal treatment is used to estimate its size, then the asymptotic bias of the estimator of the size would explode. On the other hand, if the estimator of the dose-response function used for estimation of the size is used to estimate the location of the optimal treatment, then the Jacobian term in (2.22) would not converge in probability to $\partial^2 E\{\tau(x) E[Y|T=\alpha_0, X=x]\}/\partial t^2$. As in section 2.3, we use bandwidths of different order and the same kernel function for both estimators[15]. The following theorem shows that the estimators in (2.16) and (2.17) are jointly asymptotically normal and uncorrelated.

Theorem 4 *Assume*

(i) Conditions in Theorem 3 hold.

(ix) Let $h_2 = h_2(n)$ *be such that:* $h_2 \to 0$, $nh_2^{k+4}/\ln(n) \to \infty$, $nh_2^{2k+3}/[\ln(n)]^2 \to \infty$, *and* $nh_2^{2r+1} \to 0$, *as* $n \to \infty$. *Also, for* h_1 *in Theorem 3, let* $nh_1^6 \to \infty$.

Then,

$$\begin{pmatrix} \sqrt{nh_1^3}(\widehat{\alpha} - \alpha_0) \\ \sqrt{nh_2}\left(\widehat{\mu}_{h_2}(\widehat{\alpha}) - E\{Y(\alpha_0)\}\right) \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} V_1 & 0 \\ 0 & V_2 \end{bmatrix} \right)$$

with V_1 *as in theorem 3 and* $V_2 = \left[\int \left\{\int K(u,v)dv\right\}^2 du\right] \times \int f_0(\alpha_0, x)^{-1} \tau^2(x) \widetilde{f}_0^2(x) \sigma^2(\alpha_0, x)\, dx$; *and where* $K(w)$ *is partitioned according to* $w = [t, x]$ *and* $K^{(1)}(\cdot)$ *means the partial derivative with respect to* t.

PROOF. See appendix.

[15] As discussed in section 2.3, another alternative is to use the same order of bandwidth for both and use different order kernels for each estimator. Whether one approach is preferable to the other is unknown.

Clearly, if $k > 0$ the bandwidth assumptions $nh_1^{k+5}/\ln(n) \rightarrow \infty$ and $nh_2^{2k+3}/[\ln(n)]^2 \rightarrow \infty$ imply, respectively, the conditions that $nh_1^6 \rightarrow \infty$ and $nh_2^{k+4}/\ln(n) \rightarrow \infty$. The latter conditions are kept to make Theorem 4 also hold when $k = 0$.

The proof of this theorem follows similar steps as the ones in the proof of theorem 1. For example, the proof of theorem 4 uses the Cramér-Wold device to show that for every real numbers λ_1 and λ_2 we have:

$$\lambda_1\sqrt{nh_1^3}(\widehat{\alpha}-\alpha_0)+\lambda_2\sqrt{nh_2}\left(\widehat{E}\left[Y\left(\widehat{\alpha}\right)\right]-E\left[Y\left(\alpha_0\right)\right]\right)\right) \xrightarrow{d} \mathcal{N}(0,\lambda_1 V_1+\lambda_2 V_2)$$

Also, as in theorem 1, the condition that $nh_2^{k+4}/\ln(n) \rightarrow \infty$ is important in showing that the term $\widehat{E}\left[Y\left(\widehat{\alpha}\right)\right]$ in the above equation can be replaced by $\widehat{E}\left[Y\left(\alpha_0\right)\right]$ asymptotically. As in previous results, theorem 4 requires the asymptotic bias of the estimators to go to zero faster than their variance in order to center their limiting distribution around zero. Thus, the conditions on the bandwidths imply undersmoothing. Specifically, they imply that if h_2 is proportional to n^{δ} then $\delta \in \left(\min\left(-1/\left(k+4\right), -1/\left(2k+3\right)\right), -1/\left(2r+1\right)\right)$, and for h_1 is the same as discussed in theorem 3. Regarding the order of the kernel used, the assumptions on h_1 and h_2 imply that $r > k+1$. Therefore, in the presence of covariates the joint asymptotic normality result in theorem 4 requires the use of higher order kernels. Note that the restrictions imposed on h_2 and on the order of the kernel for asymptotic normality of the estimator of the size of the maximum are stronger than those for the estimator of the dose-response function in theorem 2. For example, while in the presence of a single covariate the estimator of the size in theorem 4 requires higher order kernels, the result in theorem 2 for the estimator of the dose-response function holds even when using a second order kernel.

2.5 Dimension Reduction Techniques

The results in the previous section show that the scaling factors for asymptotic normality of the estimators (2.15)-(2.17) considered in the non-experimental case under the selection on observables assumption are the same as the ones for the corresponding estimators in the experimental case considered in section 2.3. This is so because of the second averaging used for calculation of the former ones. However, for calculation of the estimators in the non-experimental case we need first to estimate $g_0(t,x)$ with some precision, and this may be a problem if the dimension of X is large, as is usually the case in practice for the unconfoundedness assumption to be more plausible. This section discusses some approaches to dealing with this "curse of dimensionality" problem, which is common when employing nonparametric methods. The first subsection considers the role the propensity score plays in this continuous treatment case; and the second one discusses other dimension reduction devices often used econometrics.

2.5.1 The Role of the Propensity Score

A very useful result due to Rosenbaum and Rubin (1983) in the binary treatment case states that if the two potential outcomes are independent of the treatment assignment conditional on X, then they are also independent conditional on the propensity score, $p(x)$, defined as the probability of being in the treatment group conditional on X. Hence, this result reduces the dimensionality of the problem by requiring adjustment of only one scalar variable, as opposed to adjusting for all pretreatment variables. Several techniques have been used to adjust for the propensity score in the econometrics and statistics literature, such as matching, weighting, stratification and regression[16]. Since the propensity score is rarely known in practice, it is usually estimated using a logit model with higher order terms in X, which can provide a relatively good approximation to the true model (see for example Rosenbaum and Rubin, 1983; Dehejia and Wahba, 1995).

Imbens (2000) and Lechner (2001a) extend the results from Rosenbaum and Rubin (1983) to the case where the treatment of interest can take on integer values from 0 to L, so that $\mathcal{T} = \{0, 1, ..., L\}$[17]. However, while for estimation of the dose-response function Lechner (2001a) shows that in this case we can reduce the dimension of the conditioning set from $\dim(X)$ to L, Imbens (2000) shows that a reduction of the dimension to one is possible, just as in the binary case. Based on Imbens (2000), Hirano and Imbens (2004) extend the propensity score methodology to the continuous treatment case. Letting $r(t, x)$ denote the conditional density of the treatment given the covariates (i.e., $r(t, x) = f_{T|X}(t|x)$), they define the generalized propensity score (GPS) as: $R = r(T, X)$. Then they show that, given Assumption 2.4.1 (i.e., the unconfoundedness or selection on observables assumption), we can write

$$\begin{aligned} E[Y(t)] &= E_X[E[Y(t)|r(t, X) = r]] = E_X[E[Y(t)|T = t, r(t, X) = r]] \qquad (2.23) \\ &= E_X[E[Y|T = t, R = r]] \end{aligned}$$

for all $t \in \mathcal{T}$. Note that the last expression is directly estimable from observed data. Thus,

[16]For discussion of such techniques see for example Rosenbaum and Rubin (1983), Dehejia and Wahba (1995, 1998, 1999), Heckman, Ichimura and Todd (1998) and Hirano, Imbens and Ridder (2000). For discussion on efficiency issues from conditioning on the propensity score rather than on X see Hahn (1998), Heckman, Ichimura and Todd (1998) and Hirano, Imbens and Ridder (2003). For a discussion of both, techniques and efficiency issues, see for example Imbens (2004).

[17]Note that this specification also includes the case with multiple treatments.

according to (2.23) the results from Section 2.4 can be used with X replaced by the GPS, which is one-dimensional. In this case, estimation of $g_0(t,r)$ would involve only two continuous regressors, as opposed to $k+1$. It is important to note that, as emphasized by Hirano and Imbens (2004), the outer average is taken over the covariate distribution (or equivalently, over the score evaluated at t, $r(t,X)$), as opposed to averaging over $r(T,X)$. Also, contrary to the binary treatment case, in the continuous case the regression $E\left[Y|T=t,R=r\right]$ has no causal interpretation.

Unfortunately, as in the binary case, the GPS is rarely known in practice and has to be estimated. In the binary treatment case, Hirano, Imbens and Ridder (2003) consider nonparametric estimation of the propensity score using series estimators. For consistency with the rest of the chapter, in this subsection the GPS is estimated using nonparametric kernel estimators. Assume we observe (y_i, t_i, x_i), $i = 1, \ldots, n$ and let $\widehat{R} = \widehat{r}(t,x)$ be the non-parametric estimator of the GPS defined by

$$\widehat{r}(t,x) = \frac{1}{nh_r^{k+1}} \sum_{j=1}^{n} K\left(\frac{t-t_j}{h_r}, \frac{x-x_j}{h_r}\right) \tag{2.24}$$

where h_r is a given bandwidth. Also, let $\widehat{r}_i = \widehat{r}(t,x_i)$ and, as before, let $\tau(\cdot)$ be a trimming function. Then, using (2.23), the estimators of the parameters of interest in (2.3)-(2.5) can be defined, respectively, as

$$\widehat{E}\{Y(t)\} = \frac{1}{n}\sum_{i=1}^{n} \tau(\widehat{r}_i)\ \widehat{g}_h(t,\widehat{r}_i) \quad \text{for all} \quad t \in \mathcal{T} \tag{2.25}$$

$$\widehat{\alpha} = \arg\max_{t \in T} \frac{1}{n}\sum_{i=1}^{n} \tau(\widehat{r}_i)\ \widehat{g}_{h_1}(t,\widehat{r}_i) \tag{2.26}$$

$$\widehat{E}\{Y(\alpha_0)\} = \frac{1}{n}\sum_{i=1}^{n} \tau(\widehat{r}_i)\ \widehat{g}_{h_2}(\widehat{\alpha}, \widehat{r}_i) \tag{2.27}$$

where $\widehat{g}(t, x)$ is the NW multiple regression estimator,

$$\widehat{g}(t, \widehat{r}) = \frac{\sum_{j=1}^{n} Y_j K(\frac{t-t_j}{h}, \frac{\widehat{r}-\widehat{r}_j}{h})}{\sum_{j=1}^{n} K(\frac{t-t_j}{h}, \frac{\widehat{r}-\widehat{r}_j}{h})}, \tag{2.28}$$

and $\widehat{g}_h$ is based on bandwidth h.

As previously discussed, note that the outer averaging is taken over the distribution of the covariates while holding the value at t constant. The estimators (2.25)-(2.27) are two-step nonparametric estimators, where the GPS is estimated nonparametrically in the first step. A theorem that presents an asymptotic normality result for these estimators is left for future work.

So far the use of the GPS has been discussed only in a regression context. However, as is the case with a binary treatment, the GPS can be used in different ways when the treatment is continuous. Consider for example employing matching methods; and suppose for the moment that the GPS is known. For estimation of the dose-response function at t, it would be difficult in practice to find observations with a dose value of exactly t because of the continuous nature of the treatment. Thus, now the matching has to be done not only on the GPS, but also in the treatment level. Two sources of bias can be distinguished in this case. The first one comes from not having a treatment level of exactly t and having to use observations in a neighborhood to get information about the mean potential outcomes at t; and the second one comes from not matching exactly on the GPS. A way in which the matching method could be done is by matching observations on $||t - T_j, r(t, X_i) - r(t, X_j)||$, with $||\cdot||$ being a given metric, such as the Mahalanobis distance. A disadvantage of this type of matching is that one can end up predicting the dose-response function at t using observations that received doses which are very far from t and that consequently do not provide much information about the potential outcomes at t. Another way to implement a matching estimator with a continuous treatment is the following. For estimation of $E[Y(t)]$ at t, consider a window of size $2\delta_n$ around t, where as usual δ_n is a sequence of positive real numbers tending to zero as $n \to \infty$. Then, observed outcomes of observations with $T_i \in [t - \delta_n, t + \delta_n]$ can be thought as an approximation to the potential outcome of those observations at t. For observations with $T_i \notin [t - \delta_n, t + \delta_n]$, we look for matches based on the GPS to impute their missing potential outcomes at the treatment level t. However, given the continuous nature of the treatment, the search for matches is restricted to an interval around t, so that the treatment values of the matches are not too far from the treatment level of interest, t. Note that when the treatment is binary, the imputation of missing potential outcomes takes place by matching observations that received the opposite treatment based on X or the propensity score; however, when the treatment is continuous we need matches that received a dose sufficiently close to t in order for them to be informative about the potential outcomes at t.

Let $S_M(i)$ be the set of indices for the M closest matches for unit i in terms of $|r(t, X_i) - r(t, X_j)|$, with $i \neq j$ and $T_j \in \left[t - \widetilde{\delta}_n, t + \widetilde{\delta}_n\right]$. As before, $\widetilde{\delta}_n$ is a sequence tending to zero as $n \to \infty$. Here, the sequences δ_n and $\widetilde{\delta}_n$ are allowed to be different, since it could be desirable for $\widetilde{\delta}_n$ to go to zero slower than δ_n in order to improve the quality of the matches in terms of the GPS[18]. Then, the matching estimator of $E[Y(t)]$ can be

[18]The relative rates of convergence of δ_n and $\widetilde{\delta}_n$ are left for future work.

written as

$$\widehat{E}\left[Y\left(t\right)\right]=\frac{1}{n}\sum_{i=1}^{n}\widehat{Y}_{i}\left(t\right) \tag{2.29}$$

with

$$\widehat{Y}_i\left(t\right)=\begin{cases} Y_i,\text{ if } T_i\in\left[t-\delta_n,t+\delta_n\right] \\ \frac{1}{M}\sum\limits_{l\in S_M(i)}Y_l,\text{ if } T_i\notin\left[t-\delta_n,t+\delta_n\right]\end{cases}$$

Note that as in the binary treatment case, even if the covariates from a given observation $j \neq i$ differ from the covariates for observation i, it could serve as a match for the latter if $r(t, X_j)$ is close enough to $r(t, X_i)$ (and T_j to t with a continuous treatment). This is an important point of propensity score methods, since this tries to replicate randomization in the sense that those observations have the same probability of receiving treatment level t (if $r(t, X_i) = r(t, X_i)$) and therefore, there are no systematic differences in the values of their covariates. If the GPS is not known, an estimator of it such as (2.24) could be used in the above discussion. Also, note that a similar matching approach can be followed for matching on the entire set of covariates X, instead of matching on the GPS. Finally, as in the binary treatment case and regardless of whether the matching is done on X or on the GPS, it is always important in practice to check the quality of the matches used [19].

A weighting approach could also be considered to estimate the dose-response function nonparametrically when the treatment is continuous. In the binary treatment case this approach is carefully analyzed in Hirano, Imbens and Ridder (2003). As before, let δ_n be

[19]When calculating the average in $\widehat{E}\left[Y\left(t\right)\right]=\frac{1}{n}\sum_{i=1}^{n}\widehat{Y}_i\left(t\right)$, we could also consider the use of different weights depending on how close the treatment levels of the observations used are to t. For example, we could use $\widehat{E}\left[Y\left(t\right)\right]=\frac{1}{n}\sum_{i=1}^{n}w\left(\left|T_i-t\right|\right)\widehat{Y}_i\left(t\right)$ with $w\left(\cdot\right)$ being a weight function depending on the distance of T_i to t.

a sequence of positive real numbers tending to zero and let $\Delta = [t - \delta_n, t + \delta_n]$. Then, as will be shown in detail in Section 2.6, the dose-response function can be estimated as

$$\widehat{E}\left[Y\left(t\right)\right] = \frac{1}{n} \sum_{i=1}^{n} \frac{I\left(T_i \in \Delta\right) \cdot Y_i}{2\delta_n \widehat{r}\left(t, X_i\right)} \tag{2.30}$$

where $I\left(\cdot\right)$ is the indicator function. This seems to be a natural extension to the continuous treatment case of the weighting-by-the-propensity-score approach used in the literature when the treatment is binary. In this latter case, for example, the estimator of $\widehat{E}\left[Y\left(1\right)\right]$ can be written as $n^{-1} \sum_{i=1}^{n} I\left(T_i = 1\right) \cdot Y_i \left[p\left(x\right)\right]^{-1}$, with $p\left(x\right) = \Pr\left(T = 1|X\right)$ being the propensity score. (2.30) has this same structure, where for small enough δ_n it uses the approximation $\Pr\left(T \in \Delta|X\right) \approx 2\delta_n f_{T|X}(t|x)$. As in the matching approach previously described, here δ_n is the smoothing parameter. A nice feature of the weighting approach is that it can be extended to estimation of more general types of dose-response functions, such as quantile dose-response functions. In Section 2.6 we briefly discuss such extensions, and show that the estimator in (2.30) can be thought as a particular case of a more general class of estimators of the mean dose-response function that use the GPS as weights.

Once we have an estimator of the dose-response function either by matching or weighting, we could based on it the estimation of the location and size of the optimum treatment, just as previously done when using the GPS in a regression approach. The asymptotic properties of the estimators (2.29), (2.30) and their corresponding estimators of location and size of the optimum treatment will not be discussed here and are left for future work.

Similar to the binary treatment case, and regardless of the way in which the GPS is employed, it is still true that we need an estimate of the GPS. If the dimension of the covariates is large this can be as difficult as estimating the dose-response function in first place. In some sense, the dimensionality problem from estimating the nonparametric regression of Y on T and X in (2.14) is just shifted to estimating the GPS in (2.24). As mentioned in Imbens (2004), unless we have additional information on the GPS, it is difficult to argue why estimation of the latter object may be preferred. However, the approach based on the GPS can still be useful in practice. First, when analyzing the effects of a treatment on an outcome under the presence of selection on observables it is always informative to know how the selection is taking place. In this sense, it seems natural that the GPS per se should be studied or at least estimated in this case. Second, if the dimension of the covariates is such that we find difficult to implement fully nonparametric methods and we have to impose some restrictions such as additivity, single index sufficiency or simply make part of the model parametric, it could be preferable to impose those restrictions on the relation of the treatment and the covariates -that is, on the GPS-, as opposed to impose them directly on the relation of the outcome of interest and the treatment and covariates (e.g., on the regression of Y on T and X). It is possible that in the former case results are less sensitive to imposing that kind of restrictions. This conjecture will be analyzed in detail in future work.

Before ending this subsection it may be appropriate to relate the above discussion to a recent paper by Behrman, Cheng and Todd (2004) (BCT hereafter) in which they study estimation of treatment effects allowing for continuous doses of the treatment. They do so in the context of studying the effect of a preschool development program targeted toward disadvantage children between the ages of 6 and 72 months in Bolivia on some child outcome measures related to health, psycho-social skill and cognitive developments. They focus on estimating parameters analogous to the average treatment effect on the treated in the binary case. For example, letting $l \in \mathcal{T}$ denote time spent in the program (with $l = 0$ for non-participants), one of the parameters they estimate is[20] $E(\Delta^{L0}|l > 0) = E[Y(l) - Y(0)|l > 0]$. The key identifying assumption they use for estimation of this parameter is that

$$E[Y(0)|l_i = l, X = x] = E\left[Y(0)|l_i = 0, X = x\right], \text{ for all } \; l \in \mathcal{T}. \tag{2.31}$$

A stronger version of this assumption is that $Y(0) \perp l|X$, for all $l \in \mathcal{T}$[21]. Under this assumption, for estimation of $E(\Delta^{L0}|l > 0)$ they propose the estimator

$$\widehat{E}(\Delta^{L0}|l > 0) = \frac{1}{n} \sum_{i \in \{l_i > 0\} \cap \{l_i \in S_p\}} \{\widehat{E}[Y(l_i)|x_i, l_i > 0] - \widehat{E}[Y(0)|x_i, l_i = 0]\} \tag{2.32}$$

where S_p is the region of common or overlapping support and n is the cardinality of the set $\{l_i > 0\} \cap \{l_i \in S_p\}$. They estimate the two conditional expectations that follow the summation sign in (2.32) using local nonparametric regression methods as those studied in Heckman, Ichimura and Todd (1998)[22].

[20]BCT also allow the program impact to depend in a flexible way on the age of the child by writing the potential outcomes as $Y(a, l)$ and the treatment effects as $\Delta(A, L, 0)$ and $\Delta(A, l_1, l_0)$, where a denotes children age, and $a \in A$. A can be a singleton set or a range of ages. Since the focus here is on continuous treatments, that additional complication is not considered.

[21]As discussed in Imbens (2001), the latter assumption may be preferred since it makes (2.31) valid for all transformations of the outcome, and it may be hard to argue in practice why the conditional mean independence may hold while conditional independence may not.

[22]Those estimators can be written as

$$\widehat{E}[Y(0)|x_i, l_i = 0] = \sum_{k \in \{l_i = 0\}} Y_k(0) W_k(\|X_k - X_i\|)$$

$$\widehat{E}[Y(l_i)|x_i, l_i > 0] \sum_{k \in \{l_i > 0\}} Y_k(l_k) W_k(\|l_k - l_i\|, \|X_k - X_i\|)$$

To reduce the dimensionality problem when estimating the expectations in (2.32) they assume that their conditional mean independence assumption in (2.31) holds with X replaced by the propensity score, $p(x) = p(L > 0|X = x)$. However, note that this last assumption does not follow from (2.31)[23]. On the other hand, as shown by Hirano and Imbens (2004), in the case with continuous treatments unconfoundedness given X implies weak unconfoundedness given the GPS, which is enough for estimation of dose-response functions and average treatment effects. Thus, in this sense, the assumption about $p(x)$ made by BCT has no relation to the unconfoundedness-given-X assumptions discussed so far.

2.5.2 Other Dimension Reduction Techniques

One way of reducing the dimensionality problem caused by nonparametric estimation of functions in high dimensions is to impose additional assumptions on them. Examples are additive and projection pursuit models[24]. Any of those type of models can be used in the context of this chapter to reduce the dimensionality problem. Also, it is important to point out that, as discussed in the previous subsection, such restrictions can be imposed either on the dose-response function directly or on the GPS. In practice, it may be preferable to impose those restrictions on the GPS and estimate the dose-response function and the location and size of its optimum based on the estimated GPS approach. This avoids imposing restrictions directly on the function of interest, and it may be the case that results

where $W_k(\|X_k - X_i\|)$ and $W_k(\|l_k - l_i\|, \|X_k - X_i\|)$ are weights that add to one and come from the local nonparametric regression of $Y_k(0)$ on X, and of $Y_k(l_k)$ on l and X, respectively; and $\|\cdot\|$ is the euclidean distance.

[23]Or, written in its stronger form, it is not the case that $Y(0) \perp l|X$ implies $Y(0) \perp l|p(x)$, for all $l \in \mathcal{T}$

[24]For additive models see, for example, Hastie and Tibshirani (1990) and for projection pursuit Friedman and Stuetzle (1981).

are more robust in this case. As previously mentioned, the careful analysis of this conjecture is left for future work.

Another common way of reducing the dimensionality problem is to allow some parametric components into the models, such as in partially linear models. This is the approach followed in the empirical application in chapter 3.

2.6 General Approach to Estimating Dose-Response Functions and their Maximum

This section presents a general approach for estimation of a more general class of dose-response functions, such as quantile dose-response functions, and their location and size of optimum treatments. Many times is useful to think beyond mean dose-response functions. For example, the *p-th* quantile dose-response function, which gives us for each dose of the treatment the *p-th* quantile of $Y(t)$, can be useful when one is more concerned about the effects of the treatment on the upper or lower part of the distribution of the potential outcomes than about its mean. Moreover, even when the focus is on a measure of the center of the distribution of potential outcomes, it is well known that the median is more robust than the mean.

Let the function of interest be $\phi\left(Y\left(t\right)\right)$; so for example, for the mean dose-response function $\phi\left(\cdot\right)$ is given by $E\left(\cdot\right)$, while in the *p-th* quantile case is given by $F_{Y(t)}^{-1}\left(p\right)$, with $F_{Y(t)}\left(\cdot\right)$ being the cumulative density function of $Y\left(t\right)$. As before, it is convenient to start by considering the case in which assignment to different treatment doses is made randomly. In this case, nonparametric estimators of $\phi\left(\cdot\right)$ are usually available. In section 2.3, the

NW regression estimator was used for estimation of the mean dose-response function, and similar nonparametric estimators are available, for example, for conditional quantiles (e.g., Chaudhuri, 1991a, 1991b; Bhattacharya and Gangopadhyay, 1990; Samanta, 1989). Let the location and size of the maximum of $\phi(\cdot)$ be given, respectively, by θ_0 and $\phi(\theta_0)$. Also, let the nonparametric estimator of $\phi(\cdot)$ be $\widehat{\phi}(\cdot)$; then, an estimator of the location and size of its optimum treatment can be defined, respectively, as

$$\widehat{\theta} = \arg\max_{t \in \mathcal{T}} \widehat{\phi}(t) \tag{2.33}$$

$$\widehat{\phi}(\theta_0) = \widehat{\phi}(\widehat{\theta}) \tag{2.34}$$

where to simplify notation $\phi(Y(t))$ is written only as a function of t. Now, a brief sketch on how to derive the asymptotic properties of estimators (2.33) and (2.34) is presented. Assuming an interior maximum, (2.33) solves[25] $\widehat{\phi}^{(1)}(t) = 0$. Expanding this expression around the true maximum of $\phi(t)$, θ_0, and solving for $\widehat{\theta} - \theta_0$ gives

$$\widehat{\theta} - \theta_0 = -\frac{\widehat{\phi}^{(1)}(\theta_0)}{\widehat{\phi}^{(2)}(\theta^*)} \tag{2.35}$$

for some mean value θ^*. Thus, asymptotic normality for a scaled version of $\widehat{\theta} - \theta_0$ can be obtained by multiplying $\widehat{\phi}^{(1)}(\theta_0)$ by the appropriate function of n to get asymptotic normality of the numerator, and by requiring uniform convergence in probability of the denominator. As for estimation of the size of the optimum treatment, expanding $\widehat{\phi}(\widehat{\theta})$ around θ_0 and subtracting $\phi(\theta_0)$ from both sides gives

[25] All derivatives in this section are taken with respect to t.

$$\widehat{\phi}(\widehat{\theta}) - \phi(\theta_0) = \left[\widehat{\phi}(\theta_0) - \phi(\theta_0)\right] + \widehat{\phi}^{(1)}(\theta^*) \cdot (\widehat{\theta} - \theta_0) \tag{2.36}$$

Asymptotic normality of $\widehat{\phi}(\widehat{\theta}) - \phi(\theta_0)$ then follows by multiplying both sides by the appropriate function of n and making sure the second term on the right side of the equality is $o_p(1)$. Sometimes the same estimator $\widehat{\phi}(\cdot)$ would not satisfy all the conditions needed for obtaining joint asymptotic normality of the location and size of the optimum treatment of the dose-response function of interest. This was the case in previous sections when focusing on the mean dose-response function using the NW estimator. In that specific case, two estimators of $\phi(\cdot)$ were used based on bandwidths of different order for estimation of the location and size of the optimum treatment.

Now, consider estimation of optimal treatments in the nonexperimental case under an unconfoundedness or selection-on-observables assumption similar to that used in section 2.4 (Assumption 2.4.1). For the case of the mean dose-response function one could, as in section 2.4, compute first the expectation of the outcome conditional on the treatment received and the covariates, and then integrate it over the distribution of the covariates to recover $E\{Y(t)\}$. However, this approach does not work for general $\phi(\cdot)$. For example, it is well known that one cannot identify the *p-th* quantile, $F_{Y(t)}^{-1}(p)$, in this way since the quantile of the mean is generally different to the mean of the quantile. Instead, we could use a weighting approach to identify the more general class of dose-response functions $\phi(Y(t))$. The approach presented below can be seen as an extension of the one used in the binary treatment case by Hirano, Imbens and Ridder (2003) and Firpo (2002) to estimate average and quantile treatment effects, respectively.

To further simplify notation, let $\phi(Y(t)) = \phi_t$, and let ϕ_t^o be the true value of ϕ_t at $Y(t)$. Also, let $\psi(Y(t), \phi_t)$ be a function such that $E[\psi(Y(t), \phi_t^o)] = 0$. For example, in the mean (p-th quantile) dose-response case $\psi(Y(t), \phi_t) = Y(t) - \mu_t$ ($\psi(Y(t), \phi_t) = I(Y(t) \leq q_{p,t}) - p)$[26], where $\phi_t = \mu_t$ ($= q_{p,t}$) is the mean (p-th quantile) of the potential outcome at t. The next Theorem shows that ϕ_t can be identified from observed data (Y, T, X).

Theorem 5 *Let $\omega(T, X)$ be a function of the treatment and the covariates such that $E[\omega(T, X)|X] < \infty$ and $E[\omega(T, X)|X] \neq 0$, and assume $\{Y(t)\}_{t \in \mathcal{T}} \perp T|X$. Then,*

$$E\left[\frac{\omega(T, X) \cdot \psi(Y, \phi_t^o)}{E[\omega(T, X)|X]}\right] = 0 \tag{2.37}$$

PROOF. See appendix.

Thus, Theorem 5 enable us to write ϕ_t as an implicit function of the data, and of the function $\omega(T, X)$. The generalized propensity score discussed in section 2.5, $r(t, x) = f_{T|X}(t|x)$, is implicit in the denominator of (2.37). Note that the identification result in theorem 5 holds for a general class of functions $\omega(T, X)$. In section 2.5., in the context of estimating the mean dose-response function using the GPS, this function was chosen as $\omega(T, X) = I(T_i \in \Delta)$, with $\Delta = [t - \delta_n, t + \delta_n]$ and δ_n a sequence of positive real numbers tending to zero. Hence, that estimator is a special case of the ones suggested by (2.37). The question regarding the best choice of $\omega(T, X)$ is left for future work.

Let $\widehat{\varpi}(X) = \widehat{E}[\omega(T, X)|X]$. Then, a sample analog estimator of ϕ_t^o, $\widehat{\phi}_t$, can be written implicitly as a solution to

[26]Where, as before, $I(\cdot)$ is the indicator function.

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\omega\left(T_i,X_i\right)\cdot\psi\left(Y_i,\widehat{\phi}_t\right)}{\widehat{\varpi}\left(X_i\right)}=0 \tag{2.38}$$

It is often the case that the estimator $\widehat{\phi}_t$ in (2.38) can be written as the solution to the problem of minimizing a weighted objective function with weights given by $\omega\left(T,X\right)/\widehat{\varpi}\left(X\right)$. For example, consider estimation of the p-th quantile dose-response, $q_{p,t}$, and assume we use a local polynomial quantile regression approach similar to the one studied in Chaudhuri (1991a, 1991b). Let $\rho_p(r)=|r|+(2p-1)\,r$ be the check function, $\widehat{w}_i=\widehat{w}_i\left(T_i,X_i\right)=\omega\left(T_i,X_i\right)/\widehat{\varpi}\left(X_i\right)$ and $\tau\left(\cdot\right)$ be a trimming function. Also, for a nonnegative integer a let $\boldsymbol{\beta}$ be a vector of dimension $a\times 1$, and let $\widehat{\boldsymbol{\beta}}(t)$ be

$$\widehat{\boldsymbol{\beta}}(t)=\arg\min_{\boldsymbol{\beta}}\sum_{i=1}^{n}\tau\left(X_i\right)\ \widehat{w}_i\ \rho_\alpha\left\{Y_i-P_n\left(\boldsymbol{\beta},t,T_i\right)\right\} \tag{2.39}$$

where $P_n\left(\boldsymbol{\beta},t,T_i\right)$ is the polynomial given by $\sum_{j=0}^{a}\beta_j\left(\frac{t-T_i}{\delta_n}\right)^j$, and as before, δ_n is a sequence of positive real numbers tending to zero. Then, the estimator of $q_{t,p}$ is given by $\widehat{q}_{t,p}=\widehat{\beta}_0$ (i.e., the first element of $\widehat{\boldsymbol{\beta}}(t)$)[27].

Note that given a nonparametric estimator of ϕ_t^o for $t\in\mathcal{T}$ is possible to estimate the location and size of the optimal treatment based on that estimator. Thus, it is possible to extend the estimation of the location and size of optimal treatments to more general types of dose-response functions, such as those based on M-estimators. The asymptotic properties of these estimators are left for future work.

[27] If as in section 2.5 we set $w(T,X)=I\left(T_i\in\Delta_n\left(t\right)\right)$ with $\Delta_n\in\left[t-\delta_n,t+\delta_n\right]$ and δ_n a sequence tending to zero as $n\to\infty$, then (2.39) could be written as
$\widehat{\boldsymbol{\beta}}(t)=\arg\min_{\boldsymbol{\beta}}\sum_{i\in\Delta_n}\tau\left(X_i\right)\ \widehat{w}_i\ \rho_\alpha\left\{Y_i-P_n\left(\boldsymbol{\beta},t,T_i\right)\right\}$, with weights given by $\widehat{w}_i=1/\ 2\delta_n\widehat{f}\left(t|X_i\right)$. Here, $\widehat{f}\left(t|X_i\right)$ is an estimator of the conditional density of t given X.

2.7 Conclusions

In this chapter we developed nonparametric estimators based on the Nadaraya-Watson estimator for three objects of interest: the entire curve of average potential outcomes or dose-response function, the treatment dose at which the dose-response is maximized and the maximum value achieved by this curve. I presented a joint-asymptotic-normality result for our estimators of location and size of the maximum under the assumption that units in a study are randomly assigned to different doses of the treatment. These results are very helpful to gain intuition about the problem at hand. Then, I proposed estimators for the case in which we assume that units are assigned to different doses of the treatment based on an observed set of covariates, which is a straightforward extension of the unconfoundedness assumption commonly used in the binary-treatment literature. I showed that in this case our estimators of the location and size of the optimal dose are jointly normal and asymptotically uncorrelated, and that their scaling factors for asymptotic normality are the same as the ones for the corresponding estimators when random assignment is assumed. On the other hand, estimation of average potential outcomes (and hence of the optimal dose) with a large number of covariates makes nonparametric estimation problematic. When the treatment is binary, a common approach in the literature is the use of propensity score methods. This chapter discusseed the use of the generalized propensity score for estimation of our three objects of interest. Different approaches are considered, such as regression, matching and weighting. This chapter also discusseed how to extend the results presented for average dose-response functions to a more general class of functions, such as quantile dose-response functions.

Finally, this chapter also set ground for future research on the evaluation of continuous treatments. As previously mentioned, this chapter discussed how one can use methods such as matching and weighting for estimation of dose-response functions and optimal doses with continuous treatments. However, the asymptotic properties of these estimators are left for future work. Also, comparison of these estimation approaches could be very valuable. Likewise, as mentioned above, this chapter showed how to identify more general classes of dose-response functions (e.g., quantile dose-response functions) using an unconfoundedness assumption. Here, specific asymptotic results are also left for future work. Note that throughout this chapter we have assumed that the unique maximum of the dose-response function is in the interior of the treatment domain, $\mathcal{T}$. A natural next step is to consider the asymptotic theory for the case when the maximum is in the boundary of $\mathcal{T}$. Also, development of a nonparametric test for existence of an interior maximum could be very valuable. Another important extension is to allow for the case when selection into different levels of the treatment is based on unobservables and we have a continuous instrument. In this case, we can use a model similar to the one in Newey, Powell and Vella (1999) to estimate the objects of interest analyzed in this chapter.

Chapter 3

Empirical Application: the Environmental Kuznets Curve

3.1 Introduction

This chapter uses some of the tools developed in the previous chapter to analyze the relation between two indicators of environmental degradation and per capita income. The two indicators considered are per capita emissions of sulfur dioxide (SO_2) and nitrogen oxide (NO_x). Since the path breaking paper by Grossman and Krueger (1991), a large number of studies have focus on the relation between diverse environmental indicators and income per capita. Many of them have documented an inverted U-type relationship known in this literature as the environmental Kuznets curve (EKC)[1]. In this literature, a lot of emphasis is given to estimating the turning point of the EKC, that is, the level of per capita

[1]Some examples are Grossman and Krueger, 1991; Selden and Song, 1994; Kaufmann et al., 1998; Shafik, 1994; Cropper and Griffiths, 1994; List and Gallet, 1999; among others. For a critical review of the literature see Stern (1998), and most recently Stern (2004).

income at which the level of emissions or concentration of a particular pollutant reaches its peak and starts decreasing. Correct estimation of the turning point is critical for creating optimal regulatory policies at both local and worldwide level, and for predicting future levels of pollution. For example, if most of the countries are below the turning point for a given pollutant, then we can expect large increases on the global level of that pollutant in the future. Moreover, if the turning point is located at an extremely high level of income, then the benefits of economic growth on the environment may be unachievable for many countries, and global emissions may increase consistently in the future. Estimation of EKCs and their turning points for different pollutants has been at the center of discussions on worldwide organizations such as the World Bank, World Trade Organization, and environmental organizations in general, since they raise doubt on the argument that progress invariably means more pollution[2]. Finally, other reason for the importance of the correct estimation of turning points is that in this literature they are used to summarize results from different studies (e.g., Stern, 1998).

Several reasons have been considered in the literature for the eventual decline in environmental degradation as income raises. Some of them are a negative income elasticity for pollution, increased levels of education, environmental awareness and openness of the political system; changes in the composition of consumption and production; better technologies, among others[3]. Arrow et al. (1995) argue that pollutants that have local effects are more likely to have an EKC and a lower turning point than those that have only global effects, such as carbon dioxide. Finally, Khanna and Plassmann (2003) consider the

[2]See, for instance, the World Bank's World Development Report 1992: Development and the Environment (IBRD, 1992).

[3]See, for instance, Selden and Song (1994), Stern (1998) and Dasgupta et al. (2002).

importance of the ability to spatially separate the production and consumption of goods and services. They argue that when spatial separation is not possible, the turning point of the EKC is likely to be higher.

This chapter focuses on two indicators of environmental degradation, per capita emissions of sulfur dioxide (SO_2) and nitrogen oxide (NO_x)[4]. Both of them have negative health impacts. For instance, SO_2 irritates the respiratory system and lowers the respiratory's system defenses, making it more vulnerable to bacteria. If combined with high levels of particulate matter and long-term exposures it can aggravate existing cardiovascular and respiratory diseases. Likewise, short term exposure to high concentrations of SO_2 can cause temporary breathing impairment, specially for children, the elderly, those with chronic lung disease, asthmatics and people who are exercising. For NO_2, short-term exposures can decrease lung function and increase respiratory illness, specially in children; and long-term exposure can increase vulnerability to respiratory infections, destroy lung tissue and cause emphysema. Also, SO_2 and NO_x are very important components of the acid rain problem, which is associated with the acidification of soil (which has negative effects on vegetation), lakes and rivers, and with the accelerated corrosion of buildings and monuments. NO_x is also an important component of the ground-ozone level (smog) problem[5]. Given the negative consequences of high levels of these pollutants, it is not surprising they receive considerably public policy attention and are two of the most studied pollutants in this literature[6].

[4]Nitrogen oxides (NO_x) is a generic term for a group of gases that contain nitrogen and oxygen in various amounts. Nitrogen oxide (NO) and nitrogen dioxide (NO_2) are the most important ones.

[5]Source: Environmental and Protection Agency, www.epa.gov.

[6]See, for example, Grossman and Krueger (1995); Selden and Song (1994); Stern (1998); and Stern and Common (2001). The latter provides a good review for SO_2.

Finally, it is worth mentioning that the main anthropogenic source of SO_2 is fuel (coal and oil) combustion from electricity generation. For example, about 73% of SO_2 emissions in Wisconsin comes from coal-burning electrical utilities[7]. Other sources of SO_2 are burning of fossil fuels during metal smelting, other industrial processes and domestic heating. The major sources of NO_x are fuel combustion in power plants and automobiles; and some processes used in chemical plants[8].

3.2 Estimation of the location and size of the turning point of the EKC

The typical paper in this literature uses panel data with measures of some pollutants (either on emissions or concentrations) in various locations (usually countries or cities) over time. The relation is usually specified using a location and time fixed effects model, which can be written as

$$y_{it} = \gamma_i + \lambda_t + g(x_{it}) + \varepsilon_{it} \tag{3.1}$$

where i stands for a given location and t for time, y is an indicator of environmental degradation, x is per capita income, γ_i and λ_t are the corresponding individual and time fixed effects, and ε_{it} is a random error term. The function $g(\cdot)$ is almost always specified as a quadratic or cubic function of per capita income. The variables in (3.1) are usually used in levels, although some authors argue in favor or working with logarithms (e.g., Stern, 1998). An obvious problem of working with parametric specifications such as those considered

[7]Source: Wisonsin Department of Natural Resources, www.dnr.state.wi.us.
[8]Source: Environmental and Protection Agency homepage, www.epa.gov.

in the literature is that the estimated function, as well as the estimate of the turning point, will heavily depend on the assumed functional form. There have been some recent attempts to allow $g(\cdot)$ to depend on x in a more flexible way. For example, Schmalensee et al. (1998) consider a piecewise linear specification with 10 segments, while Millimet et al. (2003) estimate (3.1) as a partially linear model (PLM) with the fixed effects as the linear part of the model. Azomahou and Phu (2001) go even further and model an EKC for carbon dioxide as a complete nonparametric regression by first using the nonpoolability test developed by Baltagi et al. (1996) and then, given that they are not able to reject the null hypothesis that $g(\cdot)$ changes over time, pooling their panel data and using the Nadaraya-Watson estimator. However, those studies that document the existence of a EKC for particular pollutants using nonparametric methods do not assign standard errors to their estimators of the turning point, so they cannot be used to create confidence intervals. In this section, the nonparametric methods described in the previous chapter are used to estimate the location and size of the turning point of the EKC for the two pollutants considered, SO_2 and NO_x, and provide standard errors for the estimators.

Some papers in this literature consider controlling for additional covariates when estimating EKCs for SO_2 and NO_x. For example, Selden and Song (1994) estimate a model like (3.1) but also controlling for population density. They argue that more densely populated areas are more likely to be concerned about reducing per capita emissions than areas where the population is more sparse. Some other authors consider variables such as GDP/area, steel exports/GDP (Kaufmann et al., 1998), technological level (Cole et al., 1997), policy variables (Panayotou, 1997), among others[9]. In order to illustrate how the

[9] For a table summarizing many of the studies on estimation of EKCs for SO_2, see Stern and Common

procedures presented in Section 2.4 work when controlling for an additional covariate on an average way, this chapter also considers estimation of turning points controlling for population density. As mentioned before, some of the papers in this literature include population density as an additional explanatory variable, such as Panayotou (1993), Selden and Song (1994), Grossman and Krueger (1995), among others.

The data for emissions of SO_2, NO_x and income analyzed in this section is the same as the one used in List and Gallet (1999) and Millimet et al. (2003). It comes originally from the US EPA in their National Air Pollutant Emission Trends, 1900-1994[10]. It consists of data on emissions and income for 48 US states from 1929 to 1994. As discussed in both papers, there are at least two major advantages of this data. First, since the data comes from only the US, it is likely to be of higher quality than cross-country data, such as the Global Environmental Monitoring System (GEMS) data used in many studies (e.g., Grossman and Krueger, 1995; Panayotou, 1997). Another advantage is that it covers a long period of time, so it is more likely to cover both, the increasing and decreasing parts of the EKC[11].

In order to illustrate the methodology described in the previous chapter we first estimate a reduced form relation (i.e., without covariates), and then, population density is also included in the model. Two models are considered in this chapter. In the first one, the data for all states and years is pooled together, and a nonparametric regression is performed. The second one is a partially linear model with state and time effects as in (3.1), and where $g(\cdot)$ is left unspecified and the nonparametric methods described in the

(2001).

[10]I thank John List and Daniel Millimet for kindly providing me with a copy of their data.

[11]For more information on the data, including emission estimation methodologies, see List and Gallet (1999), and Millimet et al. (2003).

previous chapter are used for estimation of the turning point. For comparison purposes, we also consider quadratic and cubic specifications for $g(\cdot)$, as well as the Nadaraya-Watson estimator with bandwidths of the usual order[12].

Table 3.1 presents basic statistics of the variables used in this section. Emissions of SO_2 and NO_x are measured in thousands of short tons, income in thousands of 1987 US dollars, and population density in habitants per square mile. As previously mentioned, the per capita income levels cover a wide range of values, being the lowest 1,160 and the highest 22,460, both in 1987 US dollars[13].

Consider first the simplest case of estimating the reduced form relation by pooling the data. Figures 3.1 and 3.2 present scatterplots of the corresponding pollutants against per capita income for the pooled data. These scatterplots suggest that the pooled data follows the EKC hypothesis and that a within sample turning point may be estimable. Also, note that the data for SO_2 looks more disperse than the one for NO_x. This is also reflected in the higher standard deviation of SO_2 in table 3.1. Figures 3.3 and 3.4 present results for NO_x and SO_2 using the four specifications considered. Both NW estimators are based on a second-order Gaussian kernel[14]. The first one is the usual NW estimator with a bandwidth of order $n^{-(1/5)-\eta}$, where $\eta > 0$ is chosen to undersmooth and center the limiting distribution around zero. From here on, I will refer to this estimator as "the usual NW". As discussed in section 2.3, the estimator of the location of the turning point is based on

[12]The quadratic and cubic models using fixed effects are the same as the ones previously estimated in List and Gallet (1999) and Millimet et al. (2003).

[13]In the rest of the dissertation all money figures are in 1987 US dollars, unless otherwise noted.

[14]In order to avoid boundary problems the nonparametric estimation is done in $[\min(x+h), \max(x-h)]$, where x is per capita income and h is the bandwidth used. Other ways to procede could be either the use of boundary kernels (e.g., Gasser and Müeller, 1979), or consider local polynomial kernel estimators instead of the NW estimators considered here, for which no boundary adjustment is necessary.

a bandwidth of order $n^{-(1/7)-\eta}$, with $\eta > 0$. This is done in order to use the asymptotic mean-zero normal approximation to the distribution of the estimator presented in Theorem 1. The choice of bandwidth is always an issue when using nonparametric methods; and the calculation of an optimal bandwidth for the cases discussed in this dissertation is left for future work. In this application, the bandwidth is chosen as $h = ax_{sd}n^{-(1/\delta)-\eta}$, where $a = 1$, x_{sd} is the sample standard deviation of x, δ is the corresponding order of the bandwidth and η is a small number used for undersmoothing. This type of bandwidth has been previously used in the literature (e.g., Baltagi et al., 1996; Pagan and Ullah, 1999), and for our purposes it has the advantage that the order of the bandwidth can be specified directly. Later, we analyze the sensitivity of the results to the choice of bandwidth by varying a.

The four specifications used in figures 3.3 and 3.4 are close to each other, even on the location of the turning point. Tables 3.2 and 3.3 present estimates of the location and size of the turning point using the pooled data. The estimated turning points for NO_x and SO_2 using the proposed bandwidth are, respectively, 12,970 and 7,200 dollars. As point of reference, per capita income in Illinois was 12, 970 dollars in 1971, and 7240 dollars in 1942. As suggested from the graphs, the estimates of the location of the turning point in all four models for both pollutants are close to each other. However, the quadratic specifications underestimate the size of the turning points. Note that, based on Theorem 1, the last column in these tables presents the estimates based on the NW with a bandwidth of order $n^{-(1/7)-\eta}$ for estimation of the location of the turning point, and the NW with a bandwidth of order $n^{-(1/5)-\eta}$ for estimation of the level of per capita emissions at the turning point. It is not surprising that both estimates of the location and size of the turning point based on the NW estimators in the third and fourth columns are close to each other, since in both cases they are consistent. However, only the ones in the last column are shown to be asymptotically normal. Hence, for the estimators in the last column, one is able to compute asymptotic standard errors and confidence intervals based on Theorem 1. Specifically, we use plug-in estimators to estimate the asymptotic variances in Theorem 1. As expected, the standard errors of the nonparametric estimators are greater than the ones from the parametric models; however, the former estimators are more robust to functional form misspecifications. The nonparametric models used on the pooled data suggest that the turning point for NO_x is between 11,937 and 14,002; and for SO_2 between 4,965 and 9,434 dollars; each with a 95% confidence level. Given the highest dispersion of the SO_2 data in figure 3.2, it is not surprising that the confidence level of the turning point for this pollutant is wider than the one for NO_x. Also, an approximate confidence interval for the size of the turning point based on the NW estimator can be calculated from the tables, as well as joint confidence intervals for the location and size of the turning points.

Now consider the PLM with fixed state and time effects as described in (3.1). In this context, the γ_i's from equation (3.1) are state specific intercepts that control for persistent differences across states that affect emissions (e.g., fossil fuel availability, tastes, etc.); and the λ_t's are time specific intercepts that account for time varying factors that are common to all US states (e.g., federal environmental policies and standards, macroeconomic effects, changes in technology used, etc.). Figure 3.5 shows estimated emissions of NO_x as a function of per capita income[15]. The curves for the PLMs are evaluated at the average state

[15]There are different ways to estimate the nonparametric part of a PLM (e.g., Stock, 1984; Robinson, 1988; Yatchew, 1997). Here the approach followed is the one described in Härdle (1990). First, the outcome variable and the fixed effects dummies are smoothed against per capita income. Then, the residuals from the

and year effects to avoid scaling issues. The first point to note is that for all four models considered the estimated curves have an inverted U shape. The curves from the PLMs in Figure 3.5 are to the left of the ones from the pooled data in Figure 3.3, which means the estimated truning points are lower in the PLM case. Also, in Figure 3.5 there seems to be a little more variation across models in the location of the turning point as compared to figure 3.3. Note that in this case the quadratic specification is very different from the nonparametric ones. Table 3.4 presents the point estimates for the location and size of the turning point for NO_x using the fixed effect model in (3.1). All of the point estimates are below the ones from the pooled data in Table 3.2. In Table 3.4, both estimates of the location of the turning point based on parametric models are above the one from the nonparametric model that uses a bandwidth of order $n^{-(1/7)-\eta}$ in the last column. The latter estimate of the turning point is 8,150 dollars, and the estimated level of emissions of NO_x at this point is 113.3 short tons. An approximate 95% confidence interval for the turning point in this case is given by [7548, 8751][16]. As a point of reference, in 1966, per capita income in Texas was 8,155 dollars.

Selden and Song (1994) estimate a turning point of 12,041, in 1985 dollars, for NO_x using twenty two OECD and eight developing countries from 1979 to 1987. They use the same specification in (3.1) with $g(\cdot)$ being a quadratic function of per capita income. Likewise, Cole et al. (1997) use data on eleven OECD countries from 1970 to 1992 and

outcome are regressed on the residuals from the fixed effects dummies to estimate the fixed effects. Finally, the estimated fixed effects are substracted from the outcome, and theses outcome residuals are smoothed against per capita income.

[16]It may seem strange that the estimated standard error for the nonparametric estimator of the location of the turning point in Table 2.5 is less than the ones using the parametric models. A bootstrap exercise would be helpful in this case to evaluate the performance of the asymptotic approximation to the variance of the nonparametric estimator in this partially linear model. This exercise is left for future work.

a fixed country effect model which is quadratic in per capita income and includes a linear trend. Cole et al. (1997) also include a trade intensity variable in their specification. They estimate a turning point of 15,100 (in 1985 US dollars) for NO_2[17]. It is always difficult to compare estimated turning points across studies because of the different data sets used and units of observation (e.g., cities, countries, states). However, given the results in Table 3.4 in which the quadratic model gives higher estimated turning points, it is possible that in those papers the estimates are upward biased because of the use of an incorrect functional form. Of course, a careful analysis of this conjecture would entail working with the data used in those studies.

Now consider the fixed effects model for SO_2. The fixed effects model in Figure 3.6 presents a very different picture from the one using the pooled data. The only two models for which there seems to be a turning point are the quadratic and the usual NW estimator, and for the other two models the curve is increasing over the income levels in the sample. From Table 3.6 can be seen that the turning points for the quadratic and the usual NW estimator are 20,140 and 19,780, respectively. For the former model is possible to create a confidence interval for the turning point; however, given the different results obtained for the four models is very likely the quadratic model is misspecified. Using the NW estimators is possible to estimate a turning point for the one with the usual bandwidth; however, in this case it is not possible to apply Theorem 1 to obtain confidence bounds. On the other hand, when the NW estimator suggested by Theorem 1 is used, the curve no longer has a maximum. Two points are worth mentioning here. First, in this case the

[17]In this case they report a standard error of 758. Selden and Song (1994) do not report standard errors for their estimators.

estimate of the turning point using the usual NW estimator is very sensitive to the choice of bandwidth, and this may raise doubts on the existence of a turning point in this relation. For example, the bandwidth used for the usual NW estimator in figure 3.5 using standardized data is 0.1840 (see Table 3.4). If instead we use 0.18 as bandwidth, then the usual NW estimator no longer has a maximum within the sample. Second, it is generally the case that nonparametric estimation is more difficult near the boundaries, so it is not surprising that estimating the location of a maximum when the maximum is near the boundary is also difficult. In this context, developing a statistical test for the existence of a maximum could be valuable.

Although many authors have documented an EKC for SO_2 (e.g., Grossman and Krueger, 1991; Selden and Song, 1994; among others) using different data sets, some recent papers have also failed to find such a relation. For example, Millimet et al. 2003 failed to find a turning point for SO_2 with a cubic model using the same data as here, and Stern and Common (2001) also found a monotonic relation between sulfur emissions and GDP per capita using global data including developed and developing countries. In this context, the results obtained here for SO_2 seem to confirm these recent findings. Moreover, note that if we had only considered the quadratic model, then we would have concluded that per capita emissions of SO_2 also followed an inverted U-shaped relation with a turning point of 20,140 dollars. Given that many of the studies that have found an EKC for SO_2 use a quadratic specification, one possibility for the finding of an EKC for SO_2 in other studies is the use of an incorrect functional form. This highlights the importance of considering more flexible model specifications when estimating turning points.

To illustrate the results presented in section 2.4, we now estimate the models controlling for an additional available covariate, population density. Therefore, in this case $g(\cdot)$ in (3.1) is a function of x and z, where z is the population density. As before, we first estimate the models using the pooled data, and then we include fixed state and year effects. Theorem 4 requires the order of the kernel used to be greater than $k+1$, where k is the number of covariates used. In this case $k=1$, and the use of higher order kernels is required. Here, a sixth order Gaussian kernel is employed. Specifically, we use the product kernel $K(u,v) = \widetilde{K}(u)\widetilde{K}(v)$, with $\widetilde{K}(\zeta) = \frac{1}{8}\left(15 - 10\zeta^2 + \zeta^4\right)\phi(\zeta)$, and $\phi(\zeta)$ the standard normal density function. For estimation of the location of the turning point the order of the of bandwidth used is $n^{-(1/15)-\eta}$, where, as before, η is used to undersmooth [18]. For estimation of the size of the turning point we use the same sixth order kernel previously described, but with a bandwidth of order $n^{-(1/13)-\eta}$, as required by Theorem 4. Finally, as in the case without covariates, we also estimate the location and size of the turning point using a second order Gaussian kernel with a bandwidth of order $n^{-(1/5)-\eta}$. In all cases, we selected the bandwidth as previously discussed.

Figure 3.7 presents the partial mean estimator used for estimation of the location of the turning point for NO_x based on the pooled data. Also, figure 3.7 shows the partial mean estimator based on the second order kernel. In this figure, the quadratic and cubic models presented are evaluated at the average population density. Note that for estimation of the size of the turning point one needs to calculate the partial mean estimator only at one particular point (i.e., at the estimated location of the turning point). Thus, I do not

[18]The optimal bandwidth for a NW regression estimator used to estimate the d derivative of a regression function with q regressors and using a kernel of order r is of order $n^{-1/[2(d+r)+q]}$. However, note that since in this case a partial mean is used with one variable not being averaged out (per capita income), then $q=1$.

show in figure 3.7 the curve of the partial mean estimator used to estimate the size of the turning point[19]. The curves in figure 3.7 are very similar to the ones for the reduced form models presented in figure 3.3. However, the curve we use for nonparametric estimation of the location of the turning point (the ones based on the sixth order kernel) is not as smooth as in the reduced form model (see figure 3.3). This is in part because of the sixth order kernel used, but also because of the second averaging in (2.15).

Table 3.6 presents the point estimates for the location and size of the turning point of the EKC for NO_x after controlling for population density. The estimated turning point based on our nonparametric estimator is 13,200 dollars, with a standard error of 735.8; and the estimated size of the turning point is 129.4, with a standard error of 4.4. These results are not very different from the ones obtained before in the reduced-form models. Selden and Song (1994), using parametric methods, also obtain that conditioning on population density does not affect results very much. As in the reduced form models, note that the quadratic specification underestimates the size of the turning point.

Figure 3.8 shows the corresponding curves for SO_2. As for NO_x, the curves are very close to the ones calculated based on the reduced form model (see figure 3.4). From table 3.7, the estimate of the turning point for SO_2 controlling for population density is 7,420; and the estimated size is 202.9 short tons. The corresponding standard errors are 1,365 and 9, respectively.

As in the reduced-form models, we now consider a fixed state and year effects model. The results from this model are very similar to the ones in the reduced form case.

[19]Note that in the reduced form models presented before the estimator of the size of the turning point is based on the usual NW estimator.

The estimated curves for NO_x in figure 3.9 still keep the inverted U-shape obtained when pooling the data. However, the results in figure 3.10 suggest an increasing relation between per capita emissions of SO_2 and per capita income. Like in the reduced form model, if we were to use a quadratic model in per capita income we would have concluded that the turning point is 20,130 dollars. However, according to the nonparametric models, the results seem to be driven by the quadratic-in-income assumption. Also, note that for NO_x the quadratic specification overestimates the location of the turning point. From Table 3.8, the estimate of the turning point using the partial mean estimator presented in this paper is 8,690 dollars, with a standard error of 336.5. This point estimate is close to the result obtained in the reduced form model in Table 3.4, which is 8,150 dollars. The estimated size of the turning point in this case is 113.6 short tons, with a standard error of 1.4.

In the previous analysis the bandwidth was chosen as $h = ax_{sd}n^{-(1/\delta)-\eta}$, with $a = 1$. In order to check the sensitivity of the results to different choices of h, we now vary a in the interval $[0.5, 2]$. We perform this analysis for the estimators presented in sections 2.3 and 2.4. Table 3.10 shows the estimated location and size of the turning point for different values of a when using the reduced form model on the pooled data. In this case, results are fairly stable for the estimated location of the turning point across different values of h. Table 3.11 presents the results for the reduced form model estimated using fixed time and state effects. Although the estimates of the location of the turning point vary more in this case, they are still reasonably close to each other. They go from 7,580 when $a = 1.4$ to 9,240 when $a = 0.6$. The estimated size of the turning point does not vary much with a. In general, the results in this reduced form model are not drastically changed by the choice of bandwidth.

Table 3.12 presents the results for the pooled data when controlling for population density. Clearly, the choice of bandwidth becomes more important in this case. The estimated turning point for NO_x goes from 10,790 with $a = 0.5$ to 16,400 with $a = 1.1$. It is also important to point out how the estimator of the turning point changes from 13,200 to 16,400 when changing a from 1 to 1.1. The estimates of the turning point for SO_2 are not as sensitive as the ones for NO_x. In fact, if we drop the extreme cases $a = 2$ and $a = 0.5$, all but one value are between 7,420 and 8,520. Table 3.13 shows the results using the fixed state and year model. The estimates of the location and size of the turning point in this case do not seem to vary much for values of a between 0.6 and 1.3. However, outside of this range the estimates vary considerably, specially at the extreme values $a = 0.5$ and $a = 2$.

For comparison purposes, the same sensitivity analysis is performed for the estimators based on a second order kernel used in tables 3.6 to 3.9. In this case, the results do not change much when the a is varied between 0.5 and 2. From table 3.14, the turning point estimates for NO_x using the pooled data go from 12,780 to 13,330 dollars; and the ones for SO_2 go from 6,950 to 7,490 dollars. The turning point estimates using the partially linear model vary slightly more, going from a 7,810 to 9,340 dollars (see table 3.15). However, they are still fairly stable. Therefore, it seems that most of the sensitivity of the estimates of location and size in tables 3.12 and 3.13 comes from using a sixth order kernel. A simulation study could be valuable to determine if this is a special feature of our data. Hence, bandwidth choice for the proposed estimators in section 2.4 is an important topic to be considered in the future. Finally, the fact that the point estimates from the estimators using higher order kernels in tables 3.6 to 3.9 are close to the ones using a second order kernel, along with the results from tables 3.14 and 3.15, gives us some confidence on our previous results.

3.3 Conclusions

This chapter showed how one can use the results presented in chapter 2 to estimate turning points (and optimal doses in general) nonparametrically and create confidence bounds. In this chapter I estimated the location and size of the turning point of the environmental Kuznets curve (EKC) for two pollutants, per capita emissions of NO_x and SO_2, using data for 48 US states from 1929 to 1994. Using a reduced form model with fixed state and year effects, I estimated the turning point for emissions of NO_x to be equal to 8,150 dollars with a 95% confidence interval of [7,548, 8751]. The estimated size of the turning point at the average state and year effect is 113.6 short tons, with a 95% confidence level of [110.9, 116.3]. Using the same model, I found emissions of SO_2 to increase monotonically over the range of income values in the sample. The results are very similar after conditioning for population density. This empirical application also illustrated what can go wrong when using parametric models to estimate turning points. For example, the quadratic-in-income specification for SO_2 using state and year fixed effects suggested the existence of an EKC with a turning point equal to 20,000 dollars. This conclusion was not supported by our nonparametric methods. Moreover, in all the models considered in the application the quadratic specification underestimated the size of the turning point. This chapter concluded with a sensitivity analysis which suggests that the choice of bandwidth is important for the estimators considered in section 2.4.

Some caveats should be mentioned regarding our empirical application. While I approached the problem of the functional form assumed for $g(\cdot)$ in (3.1), some other things can go wrong with a specification such as (3.1). For example, the assumption that the errors are independent is strong, and it may be violated by the data. A more careful analysis would also look at the behavior of the errors over time. This is beyond the scope of this dissertation and is left for future work. Also, one should be careful not to make causality conclusions from the models estimated in this empirical application. At most, they can be seen as suggesting that the conjecture that economic progress inevitably leads to more pollution should be reconsidered, and that the quadratic models commonly used in this literature to estimate EKCs may be misspecified in some cases. Finally, it is important to point out that the literature on EKC has focused on estimating regression functions. However, as discussed in section 1.6, it could be very valuable to also look at quantiles. For example, the focus could be on estimating the median regression function of emissions of NO_x on per capita income; and to estimate the level of income at which the median level of emissions of NO_x starts decreasing. The focus could also be on an upper quantile, such as the 90th percentile[20].

[20]For further discussion on EKC see for example, Stern and Common (2001), Dasgupta et al. (2002), Stern (2004) and Copeland and Taylor (2004).

3.4 Figures

Figure 3.1. Scatterplot of pooled data for Nitrogen Oxide.

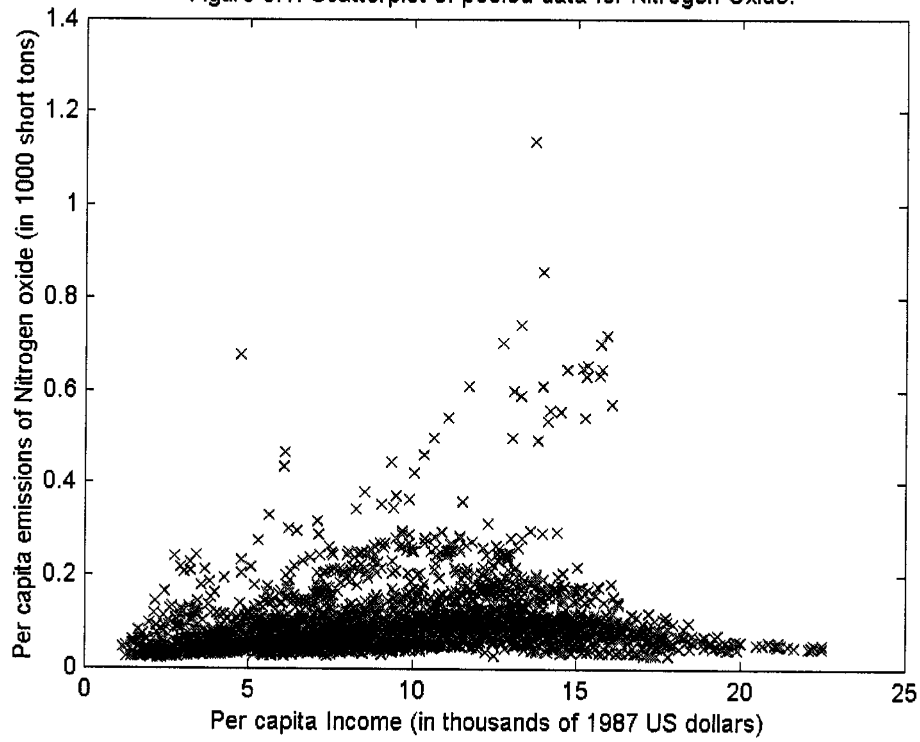


Figure 3.2. Scatterplot of pooled data for Sulfur Dioxide.

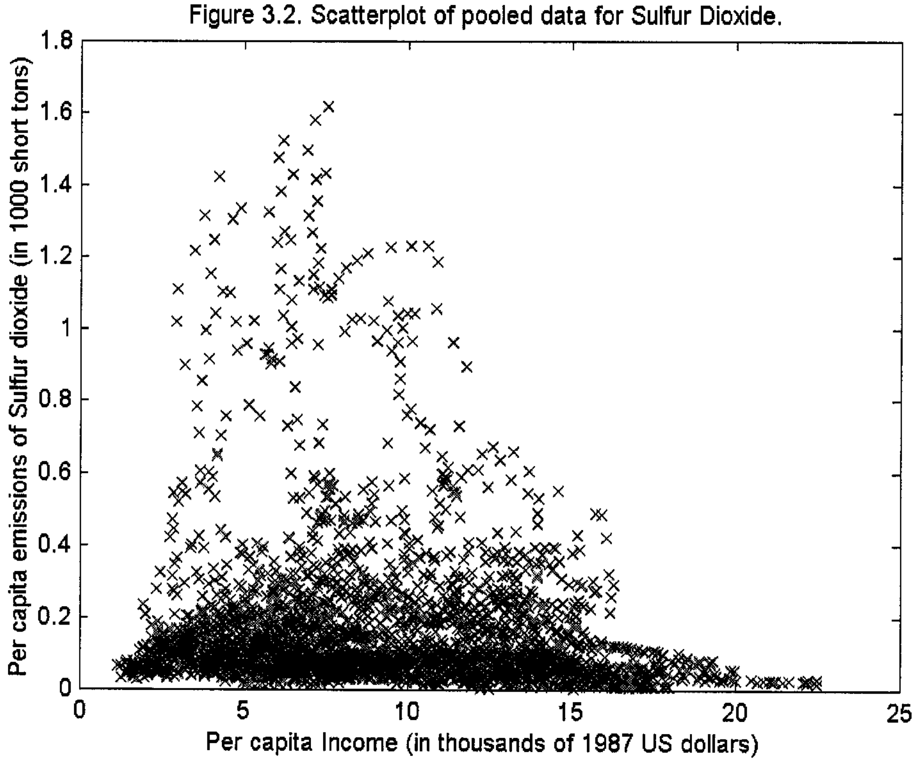


Figure 3.3. Pooled data. Nitrogen Oxide.

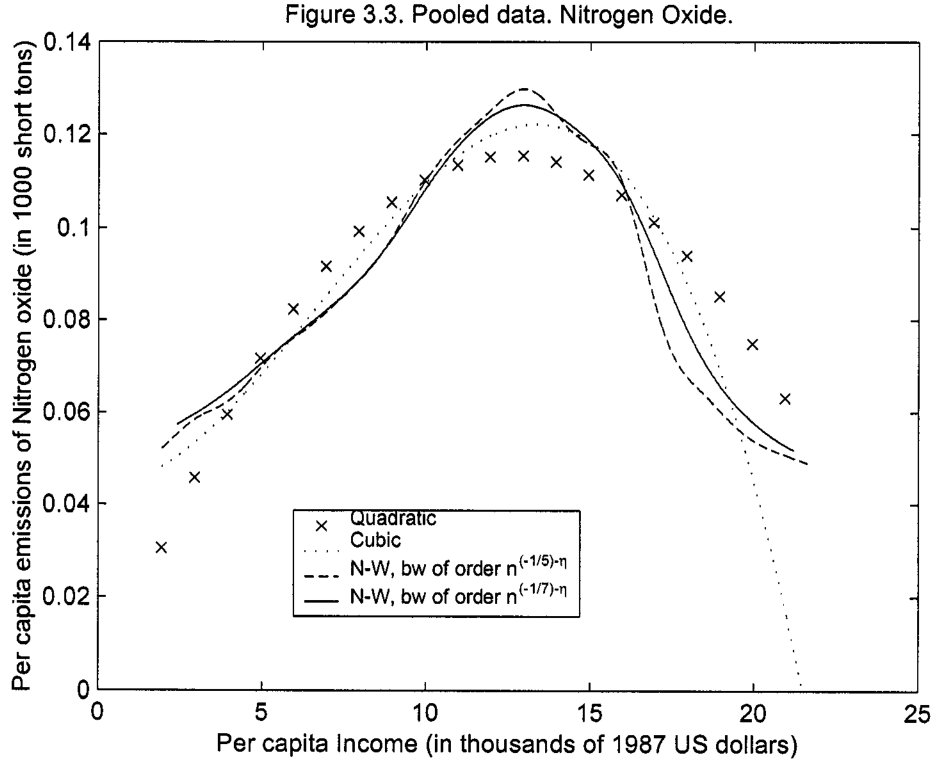


Figure 3.4. Pooled data. Sulfur Dioxide.

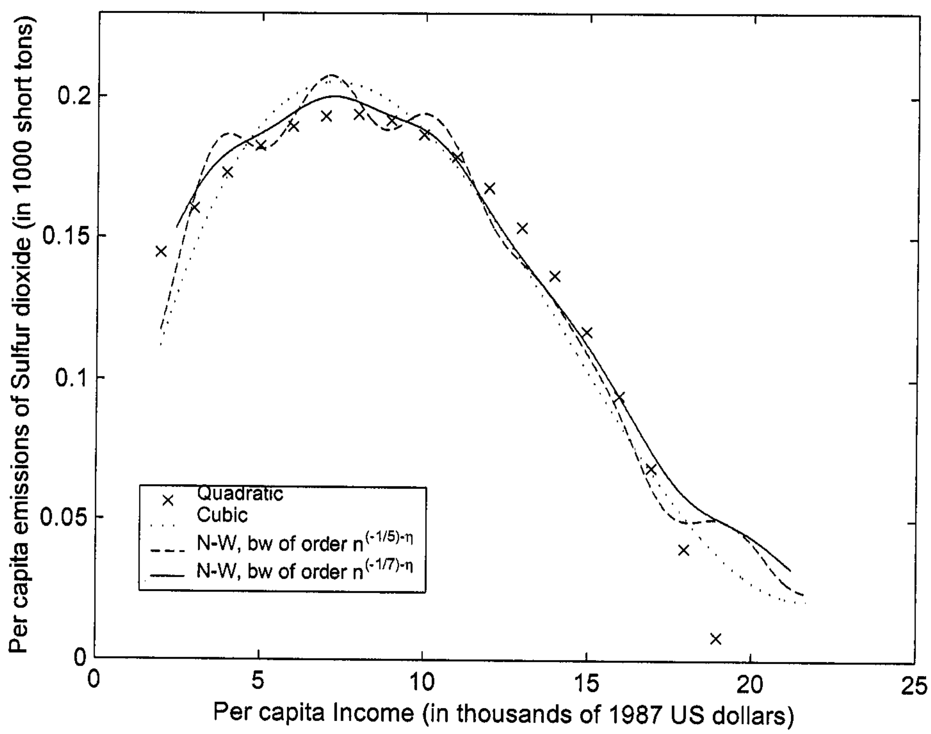


Figure 3.5. Partially Linear Model with state and time fixed effects, evaluated at average fixed effects. Nitrogen Oxide.

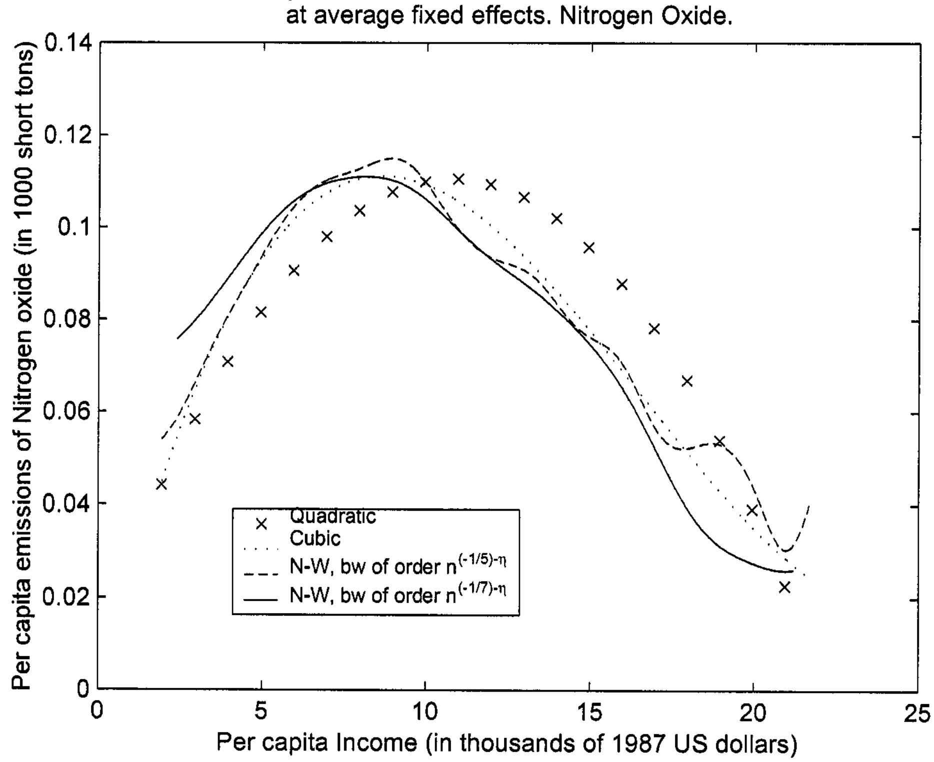


Figure 3.6. Partially Linear Model with state and time fixed effects, evaluated at average fixed effects. Sulfur Dioxide.

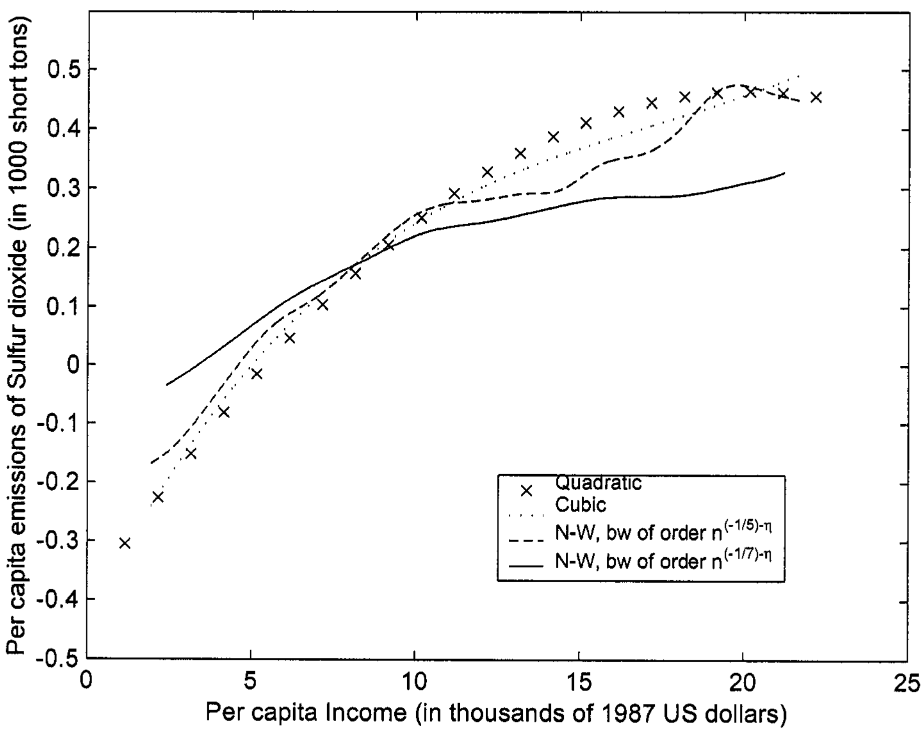


Figure 3.7. Pooled data. Controlling for Population Density. Nitrogen Oxide.

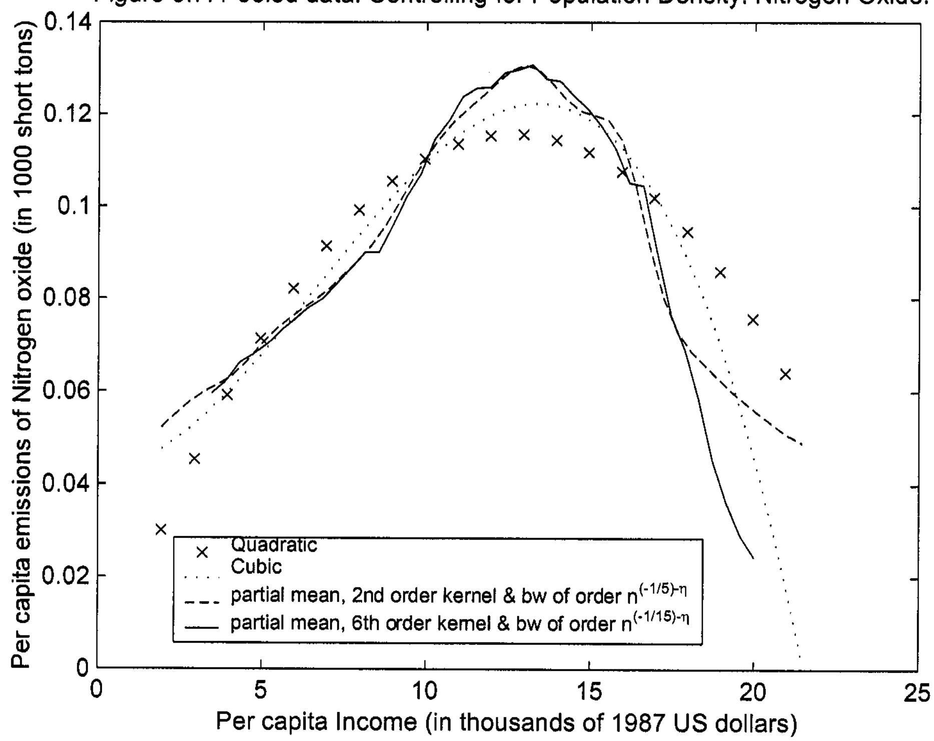


Figure 3.8. Pooled data. Controlling for Population Density. Sulfur Dioxide.

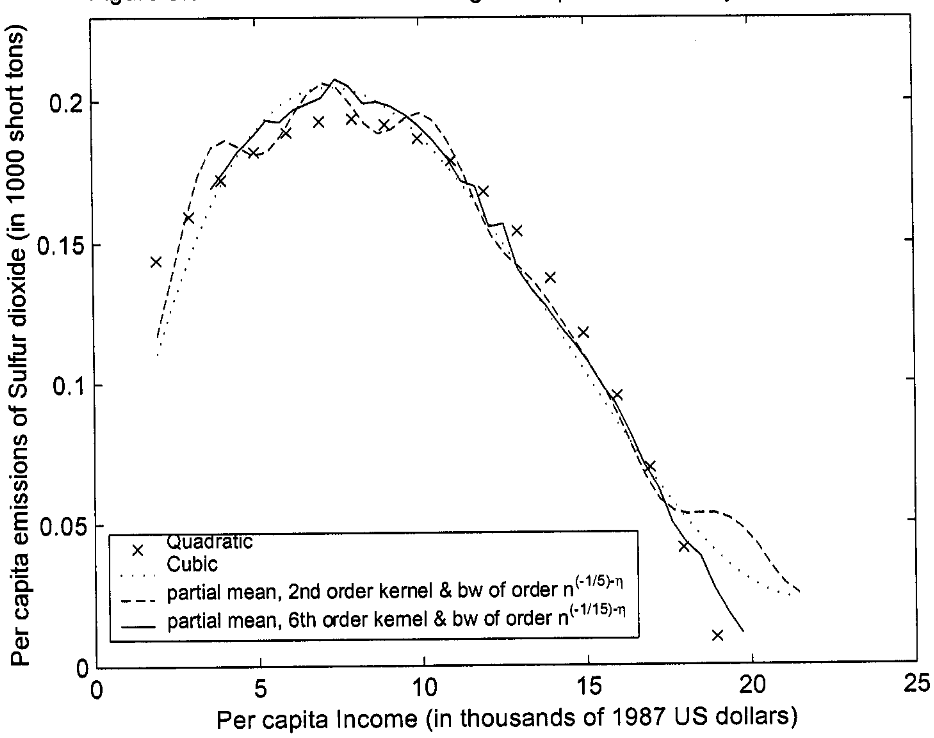


Figure 3.9. PLM controlling for Population Density. Evaluated at average fixed effects. Nitrogen Oxide.

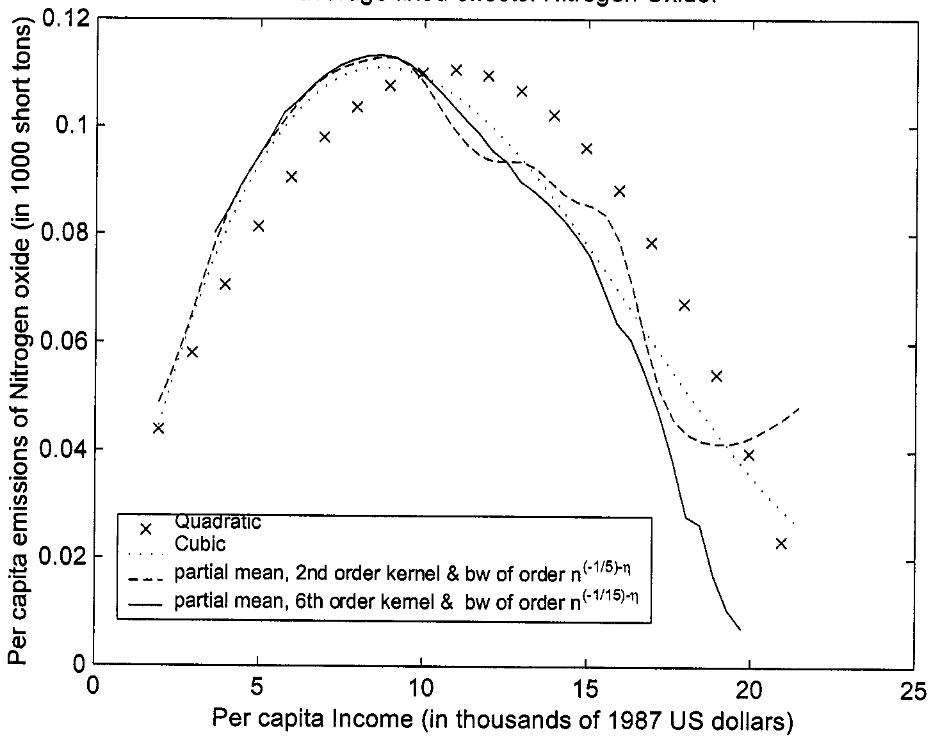
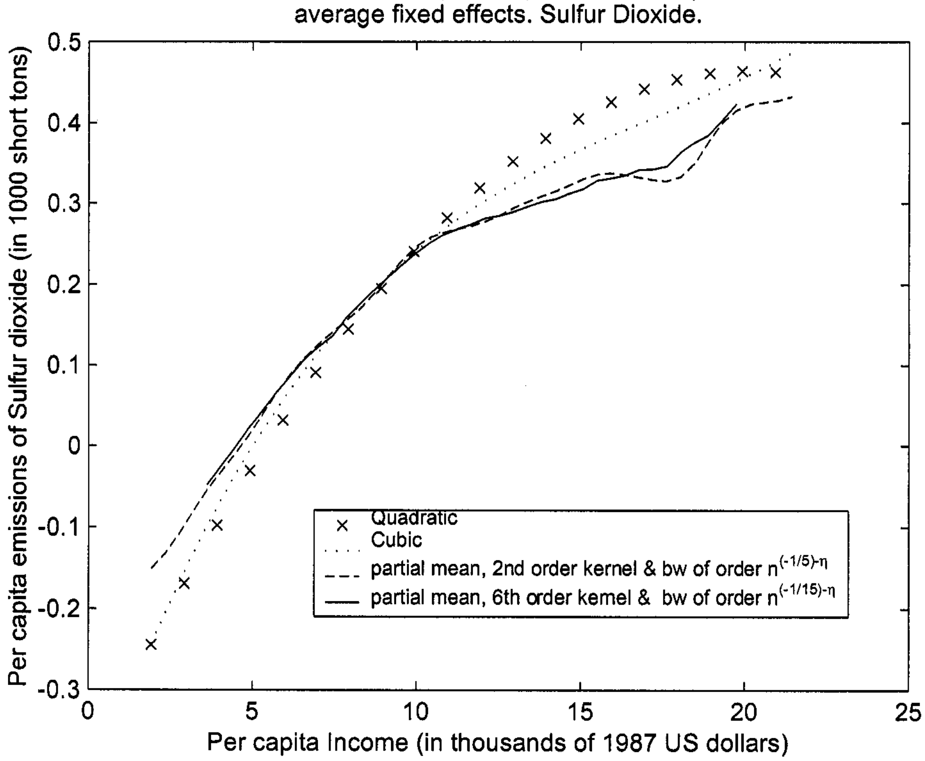


Figure 3.10. PLM controlling for Population Density. Evaluated at average fixed effects. Sulfur Dioxide.



3.5 Tables

Table 3.1. Basic Statistics. Number of Observations:3168.

	Mean	Std. Deviation	Minimum	Maximum
Per capita SO_2 emissionss	0.16	0.21	0.002	1.62
Per capita NO_x emissions	0.09	0.07	0.02	1.14
Per capita Income	9.09	4.24	1.16	22.46
Population density	230.6	843.4	0.64	17,574

Emissions in thousand of short tons; income in thousands of 1987 US dollars and population density in habitants per square mile.

Table 3.2. Pooled data, NO_x

	Models			
	Quadratic in income	Cubic in income	NW, usual bw (order $n^{-(1/5)-\eta}$)	NW,suggested order of bw for location (order $n^{-(1/7)-\eta}$) *and* size (order $n^{-(1/5)-\eta}$)
Turning point	12.59 (0.2970)	13.36 (0.2081)	12.96	12.97 (0.5267)
Level at turning point	0.1154 (0.0017)	0.1221 (0.0021)	0.1298	0.1298 (0.0042)
Bandwidth[a]			0.1840	0.2916, 0.1840

*Standard errors in parenthesis
a. Bandwidth for standardized data.

Table 3.3. Pooled data, SO_2

	Models			
	Quadratic in income	Cubic in income	NW, usual bw (order $n^{-(1/5)-\eta}$)	NW,suggested order of bw for location (order $n^{-(1/7)-\eta}$) *and* size (order $n^{-(1/5)-\eta}$)
Turning point	7.72 (0.3846)	7.24 (0.2423)	7.10	7.20 (1.14)
Level at turning point	0.1939 (0.0045)	0.2053 (0.0053)	0.2075	0.2073 (0.0092)
Bandwidth[a]			0.1840	0.2916, 0.1840

*Standard errors in parenthesis
a. Bandwidth for standardized data.

Table 3.4. Partially Linear Model with state and year fixed effects, evaluated at average fixed effects. NO_x.

	Models			
	Quadratic in income	Cubic in income	NW, usual bw (order $n^{-(1/5)-\eta}$)	NW,suggested order of bw for location (order $n^{-(1/7)-\eta}$) *and* size (order $n^{-(1/5)-\eta}$)
Turning point	10.79 (0.8118)	8.66 (0.7223)	8.94	8.15 (0.3070)
Level at turning point	0.1105 (0.0032)	0.1111 (0.0014)	0.1149	0.1133 (0.0015)
Bandwidth[a]			0.1840	0.2916, 0.1840

*Standard errors in parenthesis
a. Bandwidth for standardized data.

Table 3.5. Partially Linear Model with state and year fixed effects, evaluated at average fixed effects. SO_2.

	Models			
	Quadratic in income	Cubic in income	NW, usual bw (order $n^{-(1/5)-\eta}$)	NW,suggested order of bw for location (order $n^{-(1/7)-\eta}$) *and* size (order $n^{-(1/5)-\eta}$)
Turning point	20.14 (0.8243)	-----	19.78	-----
Level at turning point	0.4638 (0.0364)	-----	0.4744	-----
Bandwidth[a]			0.1840	0.2916, 0.1840

*Standard errors in parenthesis
a. Bandwidth for standardized data.

Table 3.6. Pooled data. Controlling for Population Density. NO_X.

	Models			
	Quadratic in income	Cubic in income	Partial mean, 2^{nd} order kernel & usual bw (order $n^{-(1/5)-\eta}$)	Partial mean, 6^{th} order kernel & suggested order of bw for location (order $n^{-(1/13)-\eta}$) *and* size (order $n^{-(1/15)-\eta}$)
Turning point	12.64 (0.3005)	13.38 (0.2079)	12.97	13.20 (0.7358)
Level at turning point	0.1157 (0.0017)	0.1224 (0.0021)	0.1307	0.1294 (0.0044)
Bandwidth[a]			0.1840	0.5389, 0.4963

*Standard errors in parenthesis
a. Bandwidth for standardized data.

Table 3.7. Pooled data. Controlling for Population Density. SO_2.

	Models			
	Quadratic in income	Cubic in income	Partial mean, 2^{nd} order kernel & usual bw (order $n^{-(1/5)-\eta}$)	Partial mean, 6^{th} order kernel & suggested order of bw for location (order $n^{-(1/13)-\eta}$) *and* size (order $n^{-(1/15)-\eta}$)
Turning point	7.76 (0.3823)	7.27 (0.2430)	7.03	7.42 (1.3650)
Level at turning point	0.1937 (0.0045)	0.2051 (0.0053)	0.2065	0.2029 (0.0090)
Bandwidth[a]			0.1840	0.5761, 0.4963

*Standard errors in parenthesis
a. Bandwidth for standardized data.

Table 3.8. Partially Linear Model with state and year fixed effects, and controlling for Population Density. Evaluated at average fixed effects. NO_x

	Models			
	Quadratic in income	Cubic in income	Partial mean, 2^{nd} order kernel & usual bw (order $n^{-(1/5)-\eta}$)	Partial mean, 6^{th} order kernel & suggested order of bw for location (order $n^{-(1/13)-\eta}$) *and* size (order $n^{-(1/15)-\eta}$)
Turning point	10.80 (0.8118)	8.68 (0.7254)	8.73	8.69 (0.3365)
Level at turning point	0.1106 (0.0033)	0.1110 (0.0014)	0.1129	0.1136 (0.0014)
Bandwidth[a]			0.1840	0.5739, 0.4963

*Standard errors in parenthesis
a. Bandwidth for standardized data.

Table 3.9. Partially Linear Model with state and year fixed effects, and controlling for Population Density. Evaluated at average fixed effects. SO_2.

	Models			
	Quadratic in income	Cubic in income	Partial mean, 2^{nd} order kernel & usual bw (order $n^{-(1/5)-\eta}$)	Partial mean, 6^{th} order kernel & suggested order of bw for location (order $n^{-(1/13)-\eta}$) *and* size (order $n^{-(1/15)-\eta}$)
Turning point	20.13 (0.8248)	-----	-----	-----
Level at turning point	0.4638 (0.0364)	-----	-----	-----
Bandwidth[a]			0.1840	0.5793, 0.4963

*Standard errors in parenthesis
a. Bandwidth for standardized data.

Table 3.10. Estimates of location and size of the turning point using bandwidth $h = an^{-(1/\delta)-\eta}$, with $\delta = 7$ for location, $\delta = 5$ for size, $n = 3168$, and for various values of a. Pooled data.

	NO_x		SO_2	
a	Location	Size	Location	Size
2	13.66	0.1238	7.20	0.1975
1.75	13.43	0.1251	7.33	0.1989
1.5	13.23	0.1265	7.39	0.2006
1.4	13.17	0.1270	7.39	0.2014
1.3	13.12	0.1277	7.37	0.2025
1.2	13.05	0.1283	7.32	0.2038
1.1	13.01	0.1290	7.26	0.2054
1	**12.97**	**0.1298**	**7.20**	**0.2073**
0.9	12.94	0.1306	7.15	0.2094
0.8	12.94	0.1316	7.11	0.2116
0.7	12.95	0.1327	7.09	0.2138
0.6	12.97	0.1340	7.10	0.2164
0.5	12.98	0.1354	7.14	0.2204

Table 3.11. Estimates of location and size of the turning point using bandwidth $h = an^{-(1/\delta)-\eta}$, with $\delta = 7$ for location, $\delta = 5$ for size, $n = 3168$, and for various values of a. Partially Linear Model with state and year fixed effects, evaluated at average fixed effects.

	NO_x		SO_2	
a	Location	Size	Location	Size
2	7.81	0.1107	---	---
1.75	7.63	0.1103	---	---
1.5	7.71	0.1108	---	---
1.4	7.58	0.1106	---	---
1.3	7.77	0.1106	---	---
1.2	8.07	0.1117	---	---
1.1	8.02	0.1118	---	---
1	**8.15**	**0.1133**	**---**	**---**
0.9	8.19	0.1123	---	---
0.8	8.82	0.1126	---	---
0.7	8.89	0.1138	---	---
0.6	9.24	0.1139	---	---
0.5	9.17	0.1139	---	---

Table 3.12. Estimates of location and size of the turning point using a sixth order kernel and bandwidth $h = an^{-(1/\delta)-\eta}$, with $\delta = 15$ for location, $\delta = 13$ for size, $n = 3168$, and for various values of a. Pooled data, and controlling for Population Density.

	NO_x		SO_2	
a	Location	Size	Location	Size
2	15.50	0.1181	6.90	0.1960
1.75	13.65	0.1275	7.56	0.1959
1.5	15.62	0.1170	8.22	0.1976
1.4	12.42	0.1340	7.98	0.1992
1.3	13.46	0.1297	6.88	0.2000
1.2	12.39	0.1320	8.34	0.2008
1.1	16.40	0.1039	8.52	0.1940
1	**13.20**	**0.1294**	**7.42**	**0.2029**
0.9	15.10	0.1206	8.45	0.1988
0.8	13.17	0.1298	8.21	0.2015
0.7	14.64	0.1215	7.54	0.2054
0.6	12.29	0.1239	8.14	0.1391
0.5	10.79	0.1173	4.80	0.1853

Table 3.13. Estimates of location and size of the turning point using a sixth order kernel and bandwidth $h = an^{-(1/\delta)-\eta}$, with $\delta = 15$ for location, $\delta = 13$ for size, $n = 3168$, and for various values of a. Partially Linear Model evaluated at average state and year fixed effects, and controlling for Population Density.

	NO_x		SO_2	
a	Location	Size	Location	Size
2	17.48	0.0521	---	---
1.75	12.21	0.0927	---	---
1.5	12.87	0.0886	---	---
1.4	11.36	0.0989	---	---
1.3	9.42	0.1115	---	---
1.2	8.75	0.1125	---	---
1.1	9.35	0.1073	---	---
1	**8.69**	**0.1136**	**---**	**---**
0.9	8.87	0.1118	---	---
0.8	8.62	0.1095	---	---
0.7	8.38	0.1117	---	---
0.6	8.99	0.1149	---	---
0.5	6.62	0.1081	---	---

Table 3.14. Estimates of location and size of the turning point using a second order kernel and bandwidth $h = an^{-(1/\delta)-\eta}$, with $\delta = 5$ for both location and size, $n = 3168$, and for various values of a. Pooled data, and controlling for Population Density.

	NO_x		SO_2	
a	Location	Size	Location	Size
2	13.33	0.1250	7.39	0.1975
1.75	13.13	0.1262	7.19	0.1986
1.5	12.94	0.1275	7.42	0.2003
1.4	12.86	0.1280	7.34	0.2012
1.3	12.78	0.1284	7.27	0.2024
1.2	13.13	0.1291	7.19	0.2036
1.1	13.05	0.1299	7.11	0.2050
1	**12.97**	**0.1307**	**7.03**	**0.2065**
0.9	12.89	0.1315	6.95	0.2077
0.8	13.24	0.1322	7.30	0.2099
0.7	13.16	0.1336	7.22	0.2127
0.6	13.08	0.1350	7.14	0.2150
0.5	13.00	0.1366	7.49	0.2162

Table 3.15. Estimates of location and size of the turning point using a second order kernel and bandwidth $h = an^{-(1/\delta)-\eta}$, with $\delta = 5$ for both location and size, $n = 3168$, and for various values of a. Partially Linear Model evaluated at average state and year effects, and controlling for Population Density.

	NO_x		SO_2	
a	Location	Size	Location	Size
2	7.81	0.1091	---	---
1.75	8.04	0.1095	---	---
1.5	8.27	0.1113	---	---
1.4	8.19	0.1103	---	---
1.3	8.96	0.1106	---	---
1.2	8.89	0.1129	20.32	0.4375
1.1	9.23	0.1076	---	---
1	**8.73**	**0.1129**	**---**	**---**
0.9	9.07	0.1119	20.09	0.4447
0.8	8.99	0.1128	---	---
0.7	9.34	0.1136	---	---
0.6	9.27	0.1135	19.43	0.5666
0.5	9.19	0.1149	19.78	0.4703

Chapter 4

Simulation Study

4.1 Introduction

In this chapter I analyze the finite properties of the estimators presented in Chapter 2 through a Monte Carlo study. A few simulation results for estimators similar to the ones presented in this dissertation can be found in the statistics literature. For example, Müeller (1985) studies the performance of his estimators of location and size of the maximum. As mentioned in Chapter 2, he considers the fixed design case with equidistant points in the [0,1] interval, and his estimators are based on the Gasser-Müeller nonparametric estimator. In his simulations, he works with two functional forms, one having a sharp Gaussian peak and the other a smooth peak, and adds a mean-zero Gaussian error with variance $\sigma^2 = 0.2$ to them. The true location of the maximum in both cases is at $1/2$, and the sizes are 4 and 2, respectively. Müeller reports results for 100 repetitions with sample sizes 25 and 100. Based on this set up, he concludes that in the location of asymmetric peaks the bias is directed towards the flat part of the peak and that there is usually a negative bias in the estimators of the size of the maximum, which is greater for small and high peaks than for flat peaks. He also concludes that higher order kernels improve the performance of his estimators.

Müeller (1989) also performs a simulation study similar to the one in Müeller (1985) but now focusing on the use by his estimators of a global versus a local bandwidth. For this purpose, he considers a function with a sharp Gaussian peak and different variances. He performs his simulations using 50 observations equidistantly in $[0, 1]$ and 200 repetitions. He concludes that there is a clear advantage of using local bandwidths for estimation of the location and size of the maximum in terms of smaller average squared errors.

Finally, there are some papers which simulations focus on comparing the nonparametric-regression based estimators in Müeller (1985, 1989) with the so-called best-r-point-average (BRPA) estimators (e.g., Chen, Lo Huang and Huang, 1996; Bai and Lo Huang, 1999; Bai, Chen and Wu, 2003). These later estimators, which focus mainly on estimation of the location of the maximum, are based on picking the r observations pairs with the highest values of y (outcome variable) and taking the average over their corresponding values of x (the treatment variable) as the estimate. Consistency results and rates of convergence for these estimators can be found in the literature (e.g., Chen, Lo Huang and Huang, 1996), but asymptotic approximations to their distribution are yet to be developed[1]. Those simulation studies usually consider different error distributions and regression functions, including the one analyzed by Müeller(1985, 1989). They also consider different values of r (e.g., 1, 5, 8), and the treatment variable is either taken to be equally

[1]Two advantages of these BRPA estimators are that they are easy to compute and they do not require continuity at the maximum (e.g., see Bai and Lo Huang, 1999). However, as mentioned in Chapter 2, in this dissertation I focus on nonparametric-regression estimators of the location and size of the maximum because it is easier to extend them to the case where we need to control for additional covariates.

spaced or uniformly distributed in a given interval, usually [0,1]. The BRPA estimators have the disadvantage of being inconsistent when the right tail of the error distribution is heavy, and this is confirmed in the simulations. However, it is also shown that even when the error distribution has a heavy right tail the BRPA estimators sometimes perform better (in terms of smaller average absolute deviations) than the ones in Müeller (1985) for some choices of r and some sample sizes (e.g., 100, 200, 400). In general, these simulation results suggest the use of the BRPA estimator when the error distribution has a lighter tail than the normal distribution (e.g., Bai et al., 2003).

In this Chapter, as in the rest of the dissertation, I consider two settings. First, I consider the case when the treatment level is assumed to be randomly assigned; and second, the case when we control for an additional covariate in an average way. This latter case has not been considered before in the literature. In the first case the simulations presented here differ from the ones in the current literature in that we intend our simulation design to be closer to the situations actually found in empirical research by basing our design on a real data set. Specifically, I partly base the simulation design on the same data set used in Chapter 3 to analyze the relationship between per-capita income and pollution. Also, I consider larger sample sizes, a larger number of repetitions and I present a larger set of summary statistics of the simulation results including those regarding estimation of the asymptotic variance of our estimator, which have not been done before.

4.2 Experimental Design

As in Chapter 2, let $Y(t)$ denote the potential outcome under treatment level t, where $t \in \mathcal{T}$. Then, we can write our objects of interest as:

$$\alpha_0 = \arg\max_{t \in \mathcal{T}} E\{Y(t)\} \tag{4.1}$$

and

$$\mu(\alpha_0) = E\{Y(\alpha_0)\} \tag{4.2}$$

where the first one is the location of the maximum and the second one its size. As discussed in Chapter 2, when assignment to different treatment levels is independent of potential outcomes (Assumption 2.3.1.) we can write the dose-response function, $E\{Y(t)\}$, as the regression function of the observed outcome Y on the observed treatment level T, or $E[Y|T=t]$. Thus, estimating α_0 and $\mu(\alpha_0)$ in this case is equivalent to estimating the location and size of the regression function $E[Y|T=t]$.

The data generating process is partly chosen to mimic the data used in Chapter 3. As discussed in that chapter, the data consists of measures of per-capita income and per-capita emissions of NO_x and SO_2 for 48 US states from 1929 to 1994. In this case our treatment variable is given by per-capita income (in thousands of 1989 dollars)[2]. Table 4.1 shows some basic statistics of the per-capita income variable.

The functional forms $g(t) = E[Y|T=t]$ considered in this experimental case are given by:

[2]In this Chapter we ignore the panel-data nature of the original data and pool all observations. A Monte Carlo study that takes this into account is left for future work.

$$g_1(t) = 0.07 + 0.025\sin(0.5t) + 0.15e^{\{-0.15(t-8)^2\}} \tag{4.3}$$

$$g_2(t) = 0.2 + 0.005\sin(0.75t - 5) - 0.001(t - 11)^2 \tag{4.4}$$

$$g_3(t) = 0.09 + 0.05\sin(0.5t - 13) + 0.15e^{\{-0.02(4t-35)^2\}} \tag{4.5}$$

Each function has parameter values $(\alpha_0, \mu(\alpha_0))$: $(7.7968, 0.2019)$, $(9.7418, 0.2021)$ and $(8.5480, 0.2059)$, respectively. These parameter values were chosen to be somehow close to the estimated turning point (i.e. peak) for the relation between per-capita emissions of SO_2 and income obtained when pooling the data in Chapter 3 (see Table 3.3). Figures 4.1-4.3 show graphs for these functions. The first function has a sharp and symmetric peak at 7.7968. This function is similar to the one analyzed in Müeller (1985, 1989). The second one has a smooth and asymmetric peak at 9.7418. This second function represents a difficult case for our estimator of the location of the maximum since, given the very low curvature of the function, it would be difficult to detect the place where the function attains its maximum. Finally, the third function has also a sharp peak at 8.5480, but the function is relatively highly nonlinear.

I add a Gaussian error with standard deviation $\sigma = 0.1$ to the functions in (4.3)-(4.5). For reference, when we calculate the empirical errors from a nonparametric regression of per-capita emissions of NO_x and SO_2 on per-capita income using the original data set, the sample standard deviation of these errors is 0.07 and 0.2, respectively[3]. In this experimental

[3]As can be seen from Table 4.1 and Figure 4.1, per-capita emissions of SO_2 have some values which are large as compared to the rest and they lead to large and positive empirical errors. The sample standard deviation for the SO_2 empirical errors when we drop their top 5 percent values is 0.1. Considering different variances and adding other type of errors (such as the estimated errors from a nonparametric regression of Y on T) to the functions in (4.3)-(4.5) could be very valuable. For brevity, we leave these extensions for future work.

case we consider five sample sizes: 100, 300, 500, 1000 and 3000. In order to have a better idea of the noise-to-signal ratios in our simulations, figures 4.1-4.3 also show representative simulated samples of size 500 for each of the models considered. Note that by partly basing our simulated samples on an actual data set and by considering a variance level similar to the one found in our original data, we expect our simulations to be useful to empirical researches.

As discussed in Chapter 2 our estimators of α_0 and $\mu(\alpha_0)$ are

$$\widehat{\alpha} = \arg\max_{t \in \mathcal{T}} \widehat{g}_{h_L}(t) \tag{4.6}$$

$$\widehat{\mu}(\alpha_0) = \widehat{g}_{h_S}(\widehat{\alpha}) \tag{4.7}$$

where $\widehat{g}_{h_L}$ and $\widehat{g}_{h_S}$ are the Nadaraya-Watson (NW) kernel estimators based on bandwidths h_L and h_S, respectively. The NW estimator is given by

$$\widehat{g}_h(t) = \frac{\sum_{i=1}^{n} Y_i K(\frac{t-t_i}{h})}{\sum_{i=1}^{n} K(\frac{t-t_i}{h})} \text{ , for } t \in \mathcal{T} \tag{4.8}$$

In our simulations we use a second order Gaussian kernel and choose the bandwidths h_L and h_S to satisfy our conditions in Theorem 1. In particular, as in the empirical application in Chapter 3, we set the bandwidths equal to $h_L = t_{SD} n^{-(1/7)-\eta}$ and $h_S = t_{SD} n^{-(1/5)-\eta}$ for estimation of the location and size of the maximum, respectively. Here, t_{SD} is the sample standard deviation of per-capita income at each simulated sample and η is a small quantity chosen to undersmooth so that we can use our asymptotic approximation to the distribution of our estimators presented in Theorem 1. As in Chapter

3, I restrict the nonparametric estimation (and hence the search for the maximum) to the interval $[\min(t_i) + h_L, \max(t_i) - h_L]$ in order to avoid boundary problems. Here, t_i are the per-capita-income observations from a given the simulated sample.

At each replication, we estimate the asymptotic variance of our estimators in Theorem 1 using a plug-in estimator in the same way we did in our empirical application in Chapter 3. Specifically, we substitute nonparametric estimates of each of the quantities appearing in the asymptotic variances in Theorem 1. Finally, for comparison purposes, we also present results for estimation of α_0 and $\mu(\alpha_0)$ based on a cubic model of per-capita income. In order to make a better comparison between the cubic and our nonparametric model, I also restrict the search for the maximum in the cubic model to the interval $[\min(t_i) + h_L, \max(t_i) - h_L]$.

Tables 4.2-4.4 present the results for each of the models in (4.3)-(4.5) based on 10,000 repetitions. For each estimator and sample size I report the number of times the estimated function is monotonic, the mean and median bias, the square root of the mean squared error (mse), the median absolute error, the standard deviation of the estimates, the range of the estimates, the mean and median estimated standard error, and the coverage rates for nominal 95% and 90% confidence intervals. For the purposes of this chapter, I defined an estimated function to be "monotonic" when using our nonparametric approach if either $\widehat{g}_{h_L}(\min(t_i) + h_L) > \widehat{g}_{h_L}(t)$ or $\widehat{g}_{h_L}(\max(t_i) - h_L) > \widehat{g}_{h_L}(t)$ for all $t \in (\min(t_i) + h_L, \max(t_i) - h_L)$; where, as before, t_i are the per-capita-income observations from a given the simulated sample. That is, I defined a nonparametric estimated function to be monotonic if the value of the function when using the estimator $\widehat{g}_{h_L}(\cdot)$ is greater at either one of the boundaries than in the interior of the interval $[\min(t_i) + h_L, \max(t_i) - h_L]$. Note that this definition would classify as "monotonic" a case in which the estimated function has a local maximum but the function evaluated at one of the boundaries is larger. Figure 3.9 presents an example of this situation. For the cubic case, I defined an estimated function to be monotonic if: i) the maximum of the estimated cubic function is outside $[\min(t_i) + h_L, \max(t_i) - h_L]$; or ii) the value of the estimated cubic function at either $\min(t_i) + h_L$ or $\max(t_i) - h_L$ is greater than at any other point in the interior of $[\min(t_i) + h_L, \max(t_i) - h_L]$[4].

In all cases the results are as expected in terms of the mse, bias and variance statistics of the estimators decreasing as we increase the sample size. In the sharp case in Table 4.2 the location estimator has a relatively low mean and median bias, and its coverage rates are higher than the nominal ones. However, the size estimator does not seem to perform as well as the location one in this case. There we can see a negative bias in the estimator of the size, as reported in Müeller (1985), and the coverage rates for nominal 95% and 90% confidence intervals are relatively low and decreasing as the sample size increases. Our size estimator performs much better than the cubic one, though. It is also worth pointing out the relatively good performance of the location and size estimators in this case as compared to the cubic ones even for a sample size of 100. Also, note that even for relatively small sample sizes (e.g., $n = 100$) the fraction of times the estimated function is monotonic using our nonparametric approach is very small. For sample sizes of 1000 all estimated maximands are in the interior. On the other hand, the number of monotonic

[4]In principle we could find a case in which the estimated location of the maximum using the cubic model is outside the interval $[\min(t_i) + h_L, \max(t_i) - h_L]$ but inside $[\min(t_i), \max(t_i)]$. This would be a within-sample estimated maximum. I allowed for this possibility in the simulations below; however, this particular case never happened.

cases when using a cubic model is greater than when using our nonparametric approach. This is because our nonparametric approach is better in capturing the sharp-peak nature of the true function.

Table 4.3 presents the results for our second model, which represents a more difficult case for estimation of the location of the maximum since the peak is smooth and is therefore more difficult to estimate the place where the function is maximized. As expected, we see a larger bias and mean squared error than when the peak is sharp. Moreover, as discussed in Chapter 2 in relation to the asymptotic variance obtained in Theorem 1, when the peak of the function is smooth then the second derivative of the function at the maximum is smaller and we can expect a higher asymptotic variance. This is confirmed by the results in Table 4.3. However, note that even in this more difficult case the coverage rates for the location estimator are greater than the nominal ones. Also note that, as pointed out by Müeller (1985), the bias in this case is directed towards the flat part of the peak. As compared to the cubic-based location estimator, note that in terms of root mse the superiority of the nonparametric location estimator shows up only for large sample sizes ($n = 3000$); however, in terms of median absolute error it shows up even for smaller sample sizes ($n = 300$). Regarding the size estimator, this seems to perform much better with a smooth rather than with a sharp peak. In this smooth case even its coverage rates are larger than the nominal ones. On the other hand, note that in terms of root mse and median absolute error the cubic estimator outperforms the nonparametric one for the sample sizes considered. This is not that surprising given the smooth nature of the function to be estimated and the much lower variance of the cubic estimator. In this smooth-peak case the number of times the estimated function is monotonic when using our nonparametric approach is greater than in the sharp-peak case. In addition, we need larger sample sizes in order to have all estimates for the location of the peak in the interior. These results also show that is more difficult to estimate the location of the maximum when the function has a smooth peak. Finally, note that the number of monotonic cases is very similar when using either our nonparametric approach or the cubic one.

Table 4.4 presents the case of a highly nonlinear function with a sharp peak. As in the first case, our location estimator performs reasonably well in this setting, specially as compared to the cubic-based location estimator for which the number of cases in which the estimated function is monotonic increases with the sample size. Nevertheless, the size estimator does not perform very well in this case, showing a relatively large negative bias. On the other hand, the size estimator still performs much better than its cubic counterpart.

Regarding the three models, note that the mean and median estimated standard errors of our location estimators are usually higher than the standard deviation of the estimators, specially for the models with sharp peaks. This may suggest that our estimates of the asymptotic variance of our location estimators tend to overestimate their actual variance. On the other hand, note that the mean and median estimated errors of the size estimator are close to their standard deviation, which may suggest our estimated variances are doing a reasonable job. By looking at the asymptotic variances obtained in Theorem 1, this may suggest that our plug-in estimator of the second derivative of the regression function evaluated at the estimated location is not very accurate, and is actually underestimating its true value.

Given the relatively poor performance of the size estimators in the presence of sharp peaks, and following Müeller (1989), I investigate the performance of our estimators when a local bandwidth is employed. The intuition for the use of a local bandwidth is that if we choose a global bandwidth which is too large as compared to the optimal local one, then we would be oversmoothing and we would decrease the estimated size of the peak. This is likely to be the case in the presence of a sharp peak, as will be confirmed below. As always, we are faced with the question of how to select the local bandwidth. Müeller (1989) relies on a modified version of the procedure developed by Müeller and Stadtmüeller (1987) to choose his local bandwidths[5]. Here, for illustrative purposes, I use an estimate of the optimal bandwidth of the NW estimator at a particular point t[6]. It is well know in the literature that this optimal bandwidth is given by[7]

$$h_{opt} = \left(\frac{\sigma^2(t) \int K^2(v)\, dv}{\mu_2^2\, f(t) \left[g^{(2)}(t)\right]^2} \right)^{1/5} n^{-1/5} \tag{4.9}$$

where $\sigma^2(t)$ is the conditional variance of Y at t, $K(\cdot)$ is the kernel used, $g^{(2)}(t)$ is the second derivative of the regression function at t, $f(t)$ is the density of the regressor evaluated at t and $\mu_2 = \int v^2 K(v)\, dv$. Here I use two estimates of the optimal bandwidth in (4.9) to evaluate the importance of correctly selecting the local bandwidth to be used. Note that the unknown quantities in (4.9) are $\sigma^2(t)$, $f(t)$ and $g^{(2)}(t)$, from which the last one is the most difficult to estimate. Because of this our first estimate of the optimal bandwidth uses the true

[5]In short, this procedure is based on an asymptotic relation between the optimal global and local bandwidths. Then, Müeller (1989) determines the global bandwidth based on the criteria discussed in Rice (1984).

[6]We could have used some other ways of selecting the local bandwidth such as the one used by Müeller (1989) or procedures based on bootstrap methods. The problem of bandwidth selection is left for future work.

[7]See, for instance, Pagan and Ullah (1999).

second-derivative function, which can be derived from the corresponding functions in (4.3)-(4.5), evaluated at the estimated location of the maximum, $\widehat{\alpha}$. The second estimate of (4.9) uses a plug-in estimator of $g^{(2)}(\widehat{\alpha})$. In both cases, the optimal bandwidth is undersmoothed in order to satisfy our assumptions in Theorem 1.

Tables 4.5-4.7 present the results from estimation of the size for the three functional forms in (4.3)-(4.5) using both estimators of the bandwidth in (4.9) and the global one used before. For the first model, the use of the estimated bandwidth based on the true value of the second derivative of the regression function improves results significantly. In this case, the bias as well as the root mse and median absolute error are decreased, and the coverage rates improve a lot, although they are still below the nominal 95% and 90% levels. Note that this improvement is not as much when we use an estimate of the second derivative of the regression function to estimate (4.9). Here, although the results improve as compared to the use of a global bandwidth, the results do not improve as much as when we use the true value of the second derivative of the regression function. Note that in this model, for a sample size of 3000, the mean global bandwidth used is 0.7892, while when using the local bandwidth based on the true second derivative the mean bandwidth used is 0.3398. Finally, the mean local bandwidth used when estimating the second derivative is 0.5455[8].

Table 4.6 show the results for the case with a smooth peak. Here the use of a local bandwidth does not improve, and actually seems to negatively affect, the performance of the size estimator. Thus, using a local bandwidth for estimation of the size may not always be the best approach to follow. Finally, Table 4.7 shows the results for the highly nonlinear

[8]Note that this discussion regarding estimation of the second derivative of the regression function at the estimated value reinforces our previuos observation that our plug-in estimator of this quantity may not be very accurate.

case with a sharp peak. In this case the use of a local bandwidth improves the results significantly. For example, for the local bandwidth that uses the true value of the second derivative of the regression function and for a sample size of 3000, the bias is reduced by more than 75%, and the root mse by more than a half. However, the coverage rates of the estimator remained very low[9].

In general, we can draw the following conclusions from the simulations presented in this section: i). The nonparametric estimator of the location of the peak performs better for sharp than for smooth peaks; ii). The more non-linear the true regression function is the better is to use our nonparametric estimators as compared to the ones based on a cubic specification, even for relatively small sample sizes (e.g., 100, 300); iii). For smooth peaks as the one considered here, our location estimator needs a larger sample size for its superiority over a cubic-based model to show up; iv). Our size estimator performs better for smooth rather than for sharp peaks; v). The use of local bandwidths improves the performance of the size estimator when the peak is sharp, but it may affect negatively for smooth peaks; vi). In this regard, how much the performance of the size estimator improves with the use of a local bandwidth depends on how well the optimal local bandwidth is estimated.

4.3 Non-experimental Design

Now we turn our attention to the case when we need to control for additional covariates. As discussed in Chapter 2, in this framework we assume that assignment to different treatment levels and potential outcomes are independent conditional on a set of

[9]As pointed out by Müeller (1985), the use of higher order kernels improved the performance of his nonparametric estimators of location and size of the maximum. The Monte Carlo analysis of how the use of higher order kernels affects our estimators is left for future work.

covariates (Assumption 2.4.1). In this case, we showed in Chapter 2 that we can write the dose-response function as the expectation over the covariates, X, of the regression function of Y on T and X. Using our notation, we can write our estimators as a function of the observed data as

$$\alpha_0 = \arg\max_{t \in T} E_X \left[g\left(t, x\right) \right] \tag{4.10}$$

and

$$\mu\left(\alpha_0\right) = E_X \left[g\left(\alpha_0, x\right) \right] \tag{4.11}$$

where $g\left(t, x\right) = E\left[Y|T = t, X = x\right]$. Based on these equations we estimate our quantities of interest as

$$\widehat{\alpha} = \arg\max_{t \in T} \frac{1}{n} \sum_{i=1}^{n} \tau\left(x_i\right)\ \widehat{g}_{h_L}(t, x_i) \tag{4.12}$$

$$\widehat{E}\{Y(\alpha_0)\} = \widehat{\mu}_{h_2}\left(\widehat{\alpha}\right) = \frac{1}{n} \sum_{i=1}^{n} \tau\left(x_i\right)\ \widehat{g}_{h_S}(\widehat{\alpha}, x_i) \tag{4.13}$$

where $\widehat{g}_{h_L}(t, x)$ and $\widehat{g}_{h_S}(t, x)$ are the NW multivariate regression estimators of the regression function $g\left(t, x\right)$ based on bandwidths h_L and h_S, respectively; and $\tau\left(x_i\right)$ is a trimming function used to keep the denominator bounded away from zero. As discussed in Chapter 2 estimators (4.12) and (4.13) are based on partial mean estimators, which control for additional covariates by averaging over them.

In this chapter I consider the case with one additional covariate. As discussed in Sections 2.4 and 2.5, the case with multiple covariates is the same in principle, although it is more difficult in practice because of the dimensionality problem when using nonparametric methods. As in Chapter 2, the covariate used in this section is population density. Table 4.1 presents summary statistics for population density, and Table 4.8 reports the correlation matrix for per-capita income, population density and emissions of NO_x and SO_2. In our simulations we resample from the joint empirical distributions of T and X. Figure 4.4 shows a scatterplot of per-capita income and population density from our original data set. In general, we can see that the population density variable is more disperse for values over 200, and it reaches levels over 1000. Also, note that for values of per-capita income below 5000 dollars we do not have observations with high population density. Similarly, for values over 15,000 dollars the amount of data is less, and for very large values of per-capita income we do not have observations with low population densities. As we can expect, our nonparametric methods will be negatively affected in these regions.

In this nonexperimental case we choose the true regression functions to somehow reflect the observed relationship between population density and per-capita emissions of NO_x and SO_2. Here we consider two models, one having a sharp and symmetric peak and another one with a smooth and asymmetric peak. The true regression functions in this case are given by

$$
\begin{aligned}
g_1(t,x) &= -0.25 + 0.15e^{\left\{-0.15(t-9.5)^2\right\}} + 0.175e^{\left\{-0.025(t-0.1x)^2\right\}} \\
&\quad +10000e^{\{-0.01x-10\}} \qquad (4.14) \\
g_2(t,x) &= 0.01\sin(0.75t-5) - 0.002(t-0.01x-9)^2 + 10000e^{\{-0.01x-10\}} \qquad (4.15)
\end{aligned}
$$

Following (4.10) and (4.11), the dose-response functions are given by $E\{Y(t)\} =$

$E_X[g(t,x)]$, where E_X is the empirical expectation of the population density variable based on the original data set. Using this approach, the true values of the parameters $(\alpha_0, \mu(\alpha_0))$ for the models based on (4.14) and (4.15) are given by $(9.2982, 0.2107)$ and $(9.4262, 0.2354)$, respectively. Figures 4.5 and 4.6 show graphs for these functions. As in the experimental case, we expect the second function to be a more difficult case for our location estimator.

In order to gain more insight into the functions we are considering and our available data, figure 4.7 presents a scatterplot of per-capita income and the output generated from $g_1(t,x)$ in (4.14) using our original data (without adding an error term), as well as the true dose-response function. Figure 4.8 shows the corresponding scatterplot for $g_2(t,x)$ in (4.15). Note that, as we mentioned before, for values of per-capita income below 5,000 dollars we do not have observations with high population densities. Hence, most of the observed values of $g_1(t,x)$ and $g_2(t,x)$ appear at the top of the graph. Thus, we may find difficult to apply a partial mean estimator like the ones appearing in (4.12) and (4.13) in this region. In fact, one would expect a partial mean estimator to become noisier and to overestimate the true values of the dose-response function as we approach to this region.

As in the experimental case, we add a Gaussian error term with standard deviation $\sigma = 0.1$ to the models in (4.14) and (4.15). The sample sizes analyzed are 100, 300, 500, 1000 and 2000. Figures 4.5 and 4.6 show a scatterplot of a representative simulated sample of size 500 for each of the functions analyzed. Just by looking at these figures, and based on our previous discussion regarding figures 4.7 and 4.8, we can see that in this case is going to be more difficult to estimate the true location and size of the peak than in the experimental case.

In order to satisfy our assumptions in Theorem 4 we base our estimates on the same six-order Gaussian kernel used in our empirical application[10] and choose the bandwidths using standardized data as $h_L = n^{-(1/15)-\eta}$ and $h_S = n^{-(1/13)-\eta}$ for estimation of the location and size of the maximum, respectively. As before, η is used to undersmooth. Finally, we trim those observations with estimated joint density lower than 0.01. Finally, for reference, we also estimate a cubic model in per-capita income and evaluate it at the sample mean population density[11].

Tables 4.9 and 4.10 present our simulation results based on 1000 repetitions. The summary statistics shown are the same as the ones shown in Tables 4.5-4.7. Table 4.9 reports the results for the sharp-peak case. For the location estimator, the mean bias and the median absolute error decrease as the sample size increases. The root mse also decreases as sample size increases except for a sample size of 2000, in which case there is an increase in the standard deviation of the estimators which leads to an increase in root mse. I will come back to this point later. Here is important to note the difference between the mean and median estimated standard errors, which suggests the presence of some large estimated standard errors in some simulated samples. Also note that the coverages for large sample sizes are below the nominal ones. As compared to the cubic-based location estimator, the nonparametric estimator performs better in terms of root mse and median absolute error even for small sample sizes; however, for very small sample sizes (e.g. $n = 100$) the large estimated standard errors can make difficult to create meaningful confidence intervals.

[10]Specifically, the kernel is given by $K(u,v) = \widetilde{K}(u)\widetilde{K}(v)$, with $\widetilde{K}(\zeta) = \frac{1}{8}\left(15 - 10\zeta^2 + \zeta^4\right)\phi(\zeta)$, and $\phi(\zeta)$ the standard normal density function.

[11]As before, to avoid boundary problems in the nonparametric estimation we look for the location of the maximum in the interval $[\min(ts_i) + h_L, \max(ts_i) - h_L]$, where ts_i is the standardized per capita income from a given simulated sample.

Regarding estimation of the size in this sharp case, the bias as well as the root mse and median absolute error decrease as the sample size increases. As compared to the sharp cases presented in the experimental case, the coverage rates seem reasonably good, although they are still below the nominal values. In general, the size estimator performs much better than its cubic counterpart, even for small sample sizes (e.g., $n = 100$).

Table 4.10 presents the case with a smooth peak, which as we have mentioned poses a difficult problem for our location estimator. For the location estimator, the root mse and median absolute error decrease when the sample size increases. Here, it is important to note the discrepancy between the mean and median estimated standard errors, which suggest the presence of some large estimated values. Moreover, note that in this case even the standard deviation of the estimates is relatively large. The location estimator based on the cubic model shows a larger bias. However, its root mse and median absolute error are smaller than for our location estimator even for large sample sizes ($n = 2000$) because of its much lower variance. Hence, in this setting we may need very large sample sizes for our location estimator to outperform the cubic one. This is not surprising given that estimation of the location of the maximum is like estimating the first derivative of the function. On the other hand, the nonparametric estimator of the size performs reasonably well, and outperforms the cubic-based one for all sample sizes considered.

So far we have ignore the fact that for our nonparametric estimator as the sample size increases the number of cases in which the estimated function is monotonic also increases. This result is mainly an artifact of the way we deal with the boundary problem in nonparametric estimation. Specifically, remember that to avoid boundary problems we restrict our search for the maximum to the points between min $(ts_i) + h_L$ and max $(ts_i) - h_L$, where ts_i stands for standardized-per-capita-income observations from a given simulated sample. Hence, as n increases and h decreases, we allow ourselves to look for the maximum closer to the boundaries. As previously discussed, we do not have enough data close to the boundary, so our estimation there becomes very noisy and we can end up with some high estimates of the dose-response function which our estimator identifies as the maximum (see for example figure 4.9). This illustrates the importance of having enough data and overlap between our treatment and covariates in order for our nonparametric estimators to work properly. In Table 4.9, this also increased the standard deviation of the estimator for a sample size of 2000, and we ended up with a larger root mse than with a sample size of 1000.

To evaluate the performance of our estimators when i) the interval where we look for the maximum does not increase with the sample size and; ii) we have "enough" data points and a "reasonably" overlap between our treatment and covariates, I simulate our models again but now restricting the search for the maximum between the $25th$ and $75th$ sample percentiles of the treatment.

Table 4.11 presents the results for the case with a sharp peak. Now, as expected, the number of monotonic cases using our nonparametric estimator decreases as the sample size increases. The root mse and the median absolute error for the location estimator also decrease with the sample size. Note that in this case, for a sample size of 2000, the standard deviation of the location estimates is much smaller than the one shown in Table 4.9. Also, it is important to point out that in this case the discrepancy between the mean and median estimated standard errors is not as large as in Table 4.9 and for large sample sizes ($n = 2000$) they are very close to each other. As for the coverage rates, they are still below the nominal ones. Regarding the size estimator, the results are similar to the ones presented in Table 4.9. Finally, it is worth pointing out that in this sharp-peak case the root mse and median absolute error are much more smaller for our nonparametric estimators than for the cubic-based ones for large sample sizes.

The results for the case with a smooth peak presented in Table 4.12 show that the root mse as well as the median absolute error for the location estimator decrease with the sample size. In this case the mean of the estimated standard deviation of the location estimator is smaller than the ones shown in Table 4.10, and for a sample size of 2000 it is reasonably close to the median. Moreover, in this case the standard deviations of the location estimator are smaller than the ones shown in Table 4.10. The cubic-based location estimator still performs better than the nonparametric one in terms of root mse and median absolute error for the sample sizes considered. As for the nonparametric size estimator, the results do not change much from those reported in Table 4.10.

From our simulations in this non-experimental case we can draw the following conclusions: i). As in the experimental case, our nonparametric estimator of location performs better for sharp rather than smooth peaks; ii). For both models and all sample sizes considered, our size estimator performs better than the one based on the cubic model in terms of lower mse, median absolute error, bias and coverage rates; iii). For the sharp-peak case considered here, the performance of our location estimator was better than the one based on the cubic model in terms of lower root mse and median absolute error; and in some cases the differences were large; iv). For the smooth case considered here, the location estimator based on the cubic model has a lower root mse and median absolute error than our location estimator for all sample sizes considered. This may suggest that in this case we need a large amount of data for our location estimator to perform better than the cubic-based one; v). In general, in this non-experimental case we need larger sample sizes than in the experimental case for our asymptotic results to work properly. This is not surprising given that in the non-experimental case we need to estimate nonparametrically a high dimensional function (i.e., $E\left[Y|T,X\right]$) with some precision. Moreover, as we have mentioned, estimation of the location of the maximum is like estimating the first derivative of the function. vi). It is important to have enough data and overlap of our treatment with the additional covariate in order for our estimators to perform adequately. Otherwise, we may have to rely on parametric assumptions.

4.4 Conclusions

In this chapter we analyzed the performance of our estimators in finite samples through a simulation study. In order to gain insight into the behavior of our estimators in situations empirical researchers may find in their work, we partly based our simulation exercise on a real data set. Specifically, our simulations partly rely on the data set we used in Chapter 3, and we added to our models an error term with variance similar to the ones we could find in actual data sets. Also, we faced our estimators with difficult situations, such as very smooth peaks.

In general, our location estimator performs better the sharper is the peak of the function of interest. On the other hand, the size estimators find more difficult to estimate the true size of the peak the sharper is the peak. Our results in the experimental setting show that the use of local bandwidths improves the performance of the size estimator; however, the level of improvement depends on how well the optimal local bandwidth is estimated. Although we did not investigate the use of local bandwidths in our non-experimental setting, this conclusion is very likely to be extended to this case. As compared to the location and size estimators based on a cubic model, our estimators usually performed better in terms of a lower root mse and median absolute error, sometimes even for relatively small sample sizes (e.g., 100). This conclusion is stronger the sharper is the peak and the more non-linear the function of interest is. Focusing in our non-experimental setting, we talked about the importance of having enough data and overlap between our treatment and covariate for estimation of the location and size of the maximum. In the absence of such overlap, one may need to rely on parametric assumptions to extrapolate to those regions with not enough data and/or poor overlap.

In this chapter I only considered the case with one additional covariate. As discussed in section 2.4, the case with more covariates can be approached in the same way as with a single covariate. However, as discussed in section 2.5 and as illustrated in this chapter, the more covariates we add the larger sample sizes we need for the adequate performance of our estimators. For the case with many covariates our estimators may become intractable. In section 2.5 I argued that in this case the use of dimension reduction techniques such as additive models, partially linear models and other semiparametric techniques becomes relevant. Given the good performance of our estimators in the experimental design (i.e., with only one regressor), the dimension-reduction techniques just mentioned are likely to work properly. The analysis of different dimension-reduction techniques using simulation methods is left for future work.

Finally, studying the performance of our estimators under different variance levels and error distributions, as well as analyzing of the use of higher order kernels and semiparametric models, can be very useful. In addition, it would be very enlightening to compare our estimators to estimators based on higher-than-cubic polynomials, and to consider the panel data structure of our original data. I left these considerations for future work.

4.5 Figures

Figure 4.1. Regression curve g_1 and a representative simulated sample of size 500.

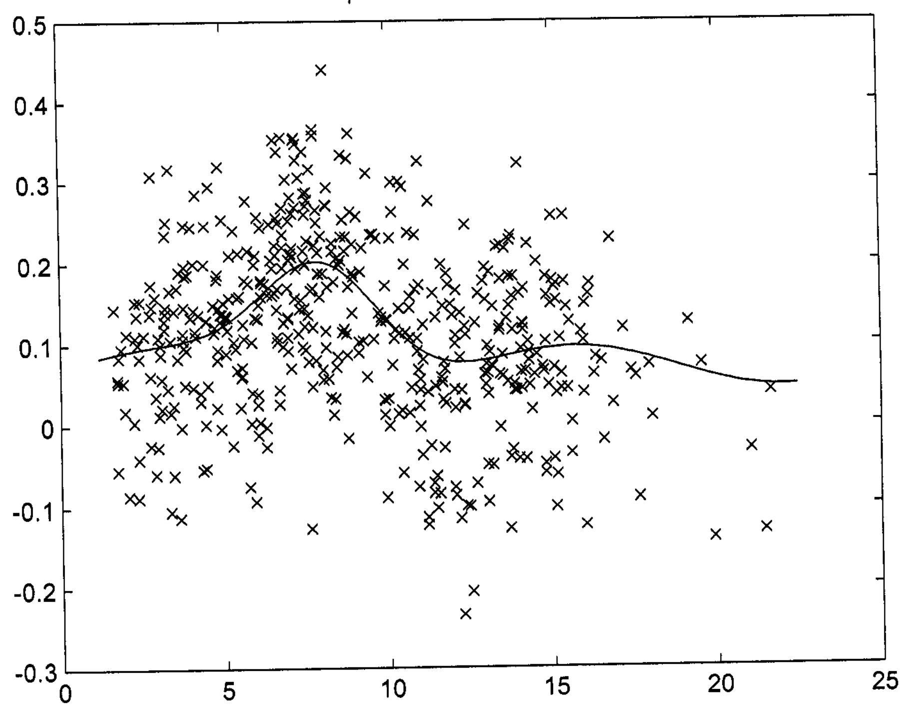


Figure 4.2. Regression curve g_2 and a representative simulated sample of size 500.

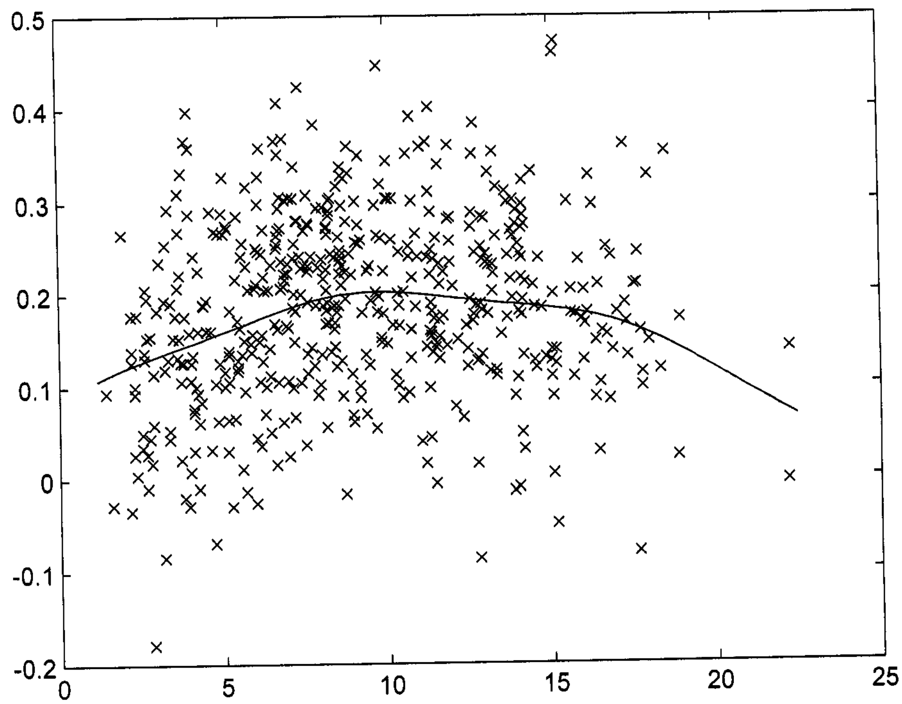


Figure 4.3. Regression curve g_3 and a representative simulated sample of size 500.

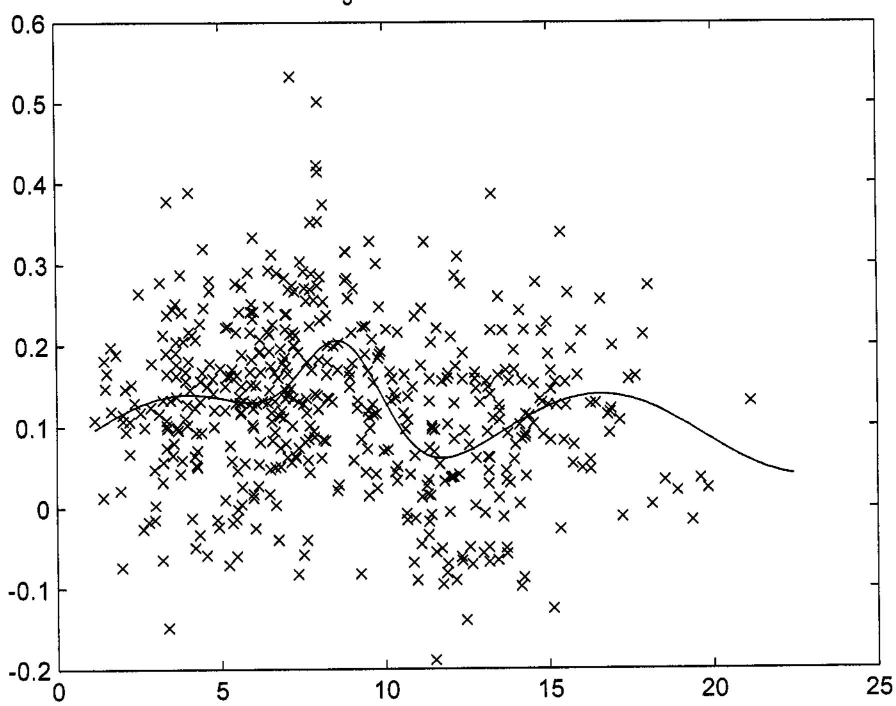


Figure 4.4. Scatterplot of per-capita income and population density from original data.

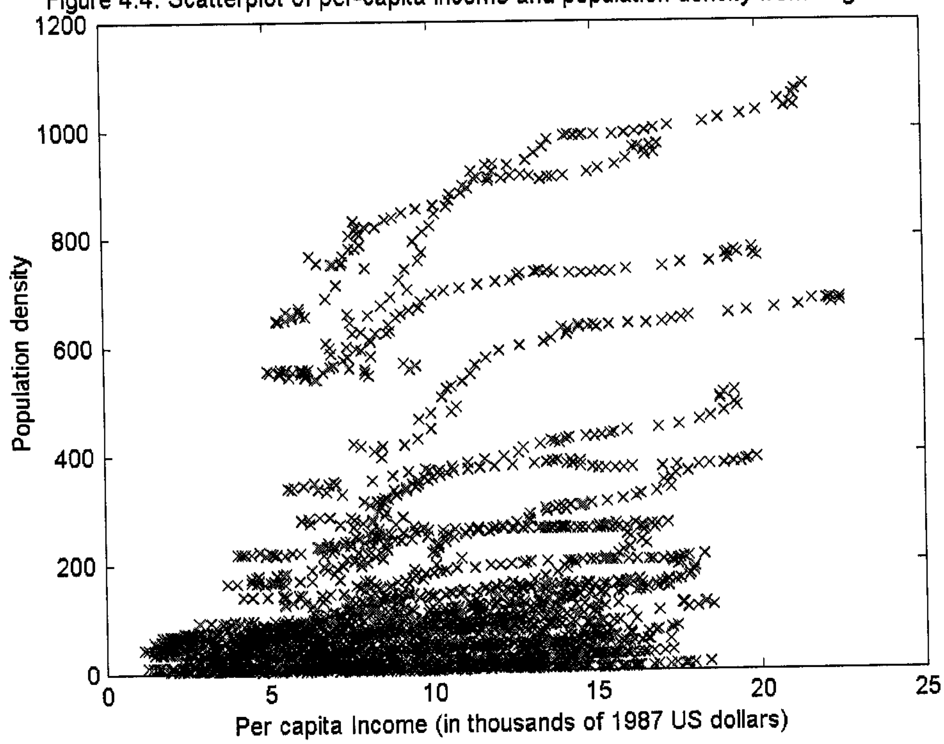


Figure 4.5. True dose-response function based on $g_1(t,x)$, along with a representative simulated sample of size 500.

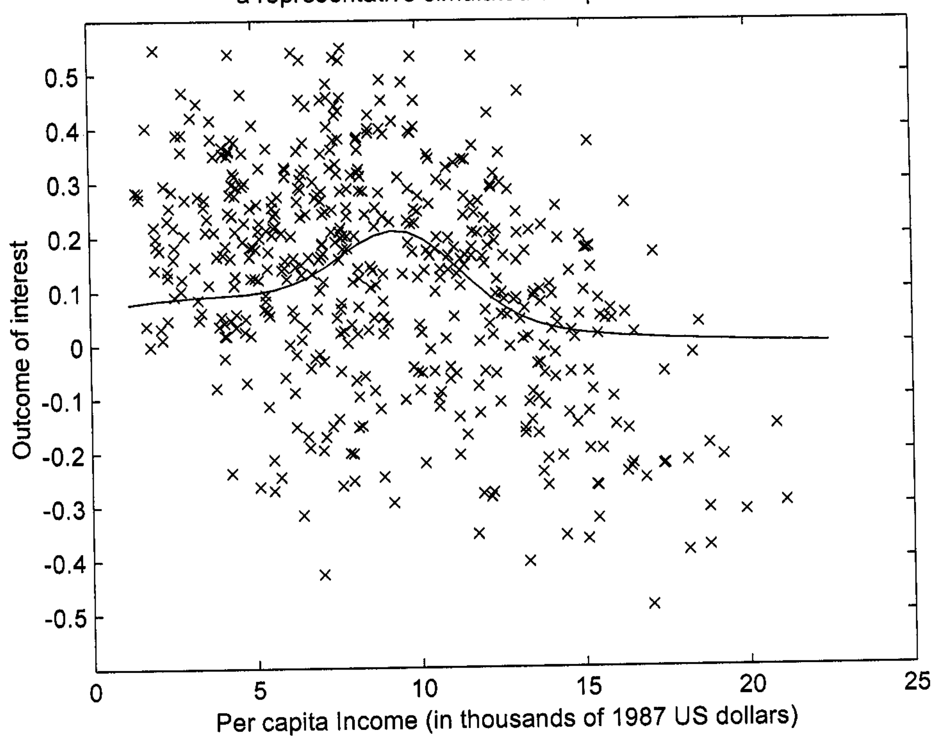


Figure 4.6. True dose-response function based on $g_2(t,x)$, along with a representative simulated sample of size 500.

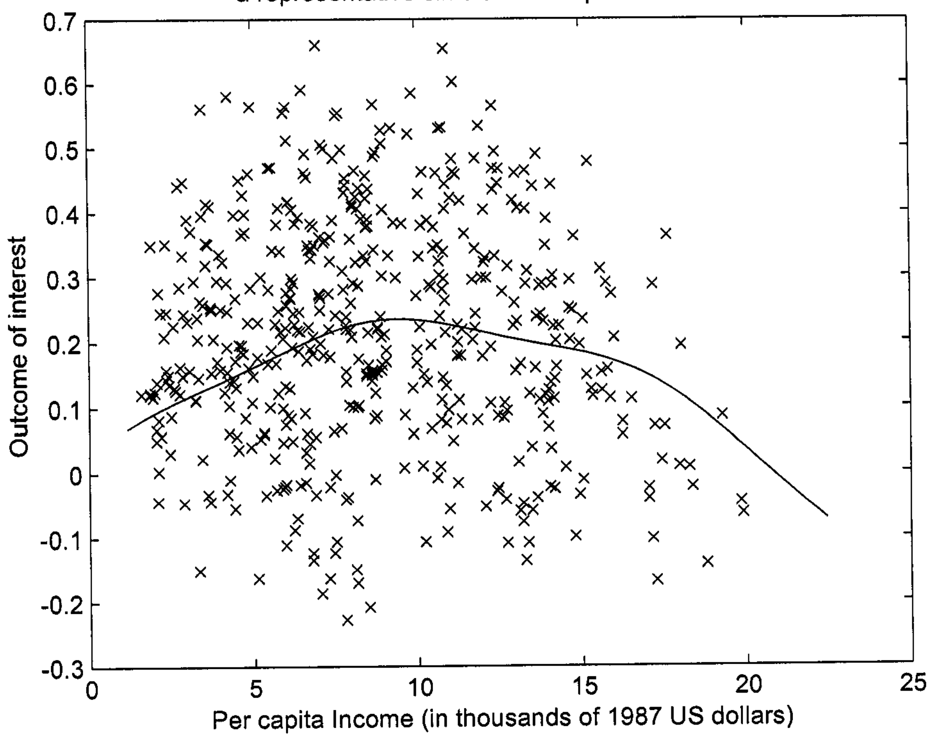


Figure 4.7. Scatterplot of per-capita income and output generated from (4.14) based on the original data (no error added), along with true dose-response function.

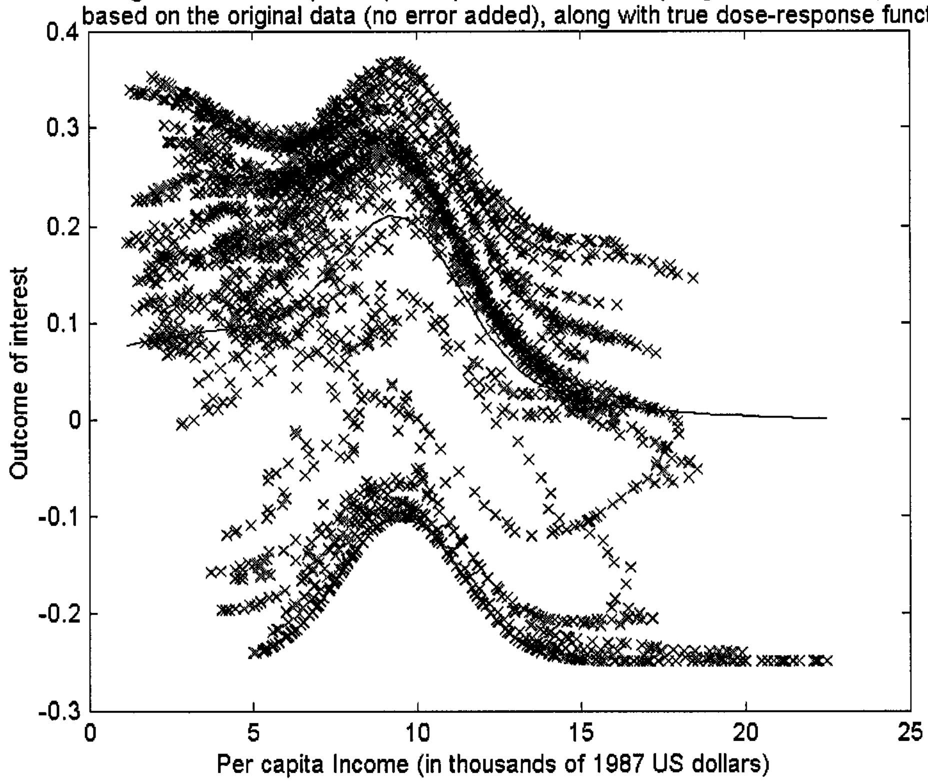


Figure 4.8. Scatterplot of per-capita income and output generated from (4.15) based on the original data (no error added), along with true dose-response function.

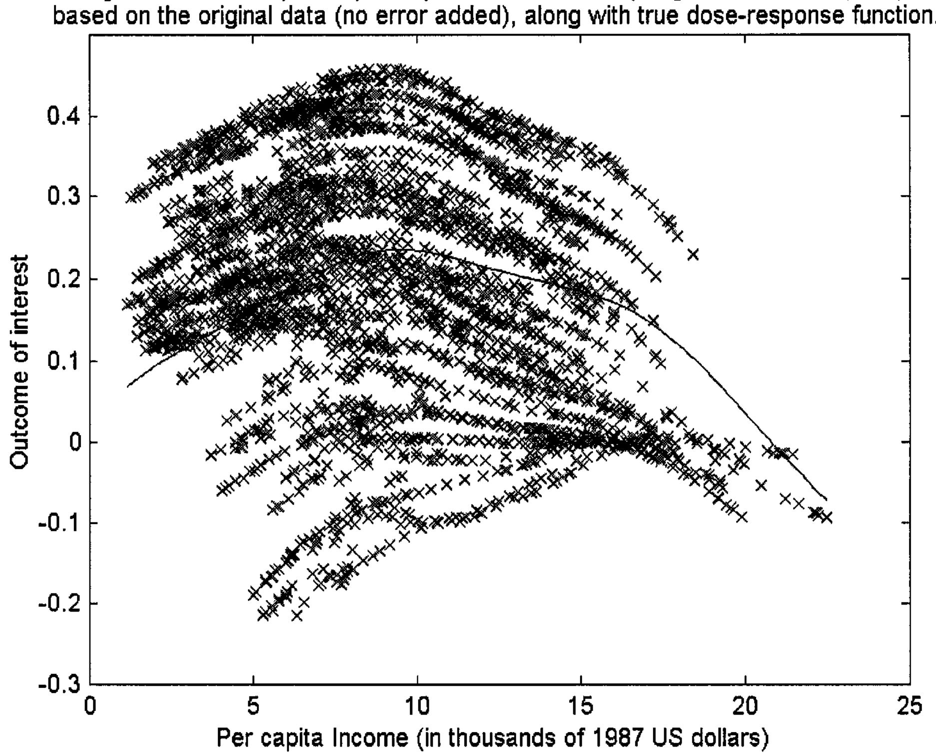
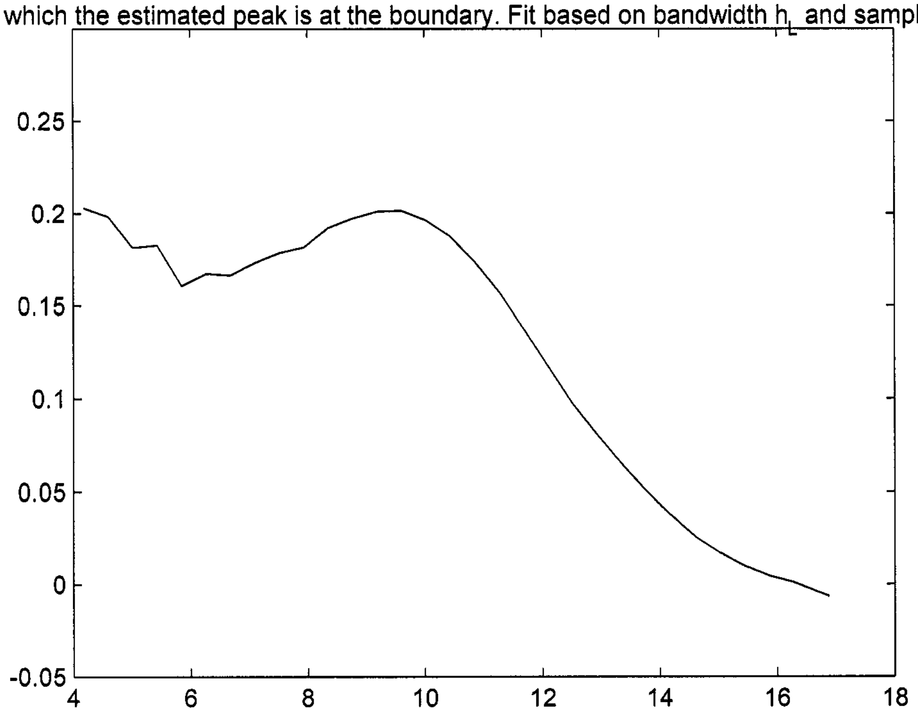


Figure 4.9. Example of a nonparametric fit of the dose-response function based on $g_1(t,x)$ and for which the estimated peak is at the boundary. Fit based on bandwidth h_L and sample size 500.



4.6 Tables

Table 4.1. Basic Statistics. Number of observations: 3168.

	Per-capita			
	Income	Emissions of NO_X	Emissions of SO_2	Population density
Mean	9.0893	0.0928	0.1647	132.3179
Std. Deviation	4.2415	0.0735	0.2056	198.7439
Minimum	1.1621	0.023	0.0021	0.8196
Maximum	22.4625	1.136	1.6179	1080.3
Percentie:				
1%	1.8439	0.028	0.0096	2.3678
5%	2.8135	0.0347	0.0254	4.7204
10%	3.7694	0.0398	0.0357	8.2616
15%	4.4771	0.044	0.0459	9.8849
20%	5.2159	0.0479	0.0529	17.6644
25%	5.8475	0.0514	0.0591	25.4093
30%	6.346	0.0555	0.0667	32.3025
40%	7.3529	0.0659	0.0807	45.7986
50%	8.4346	0.0759	0.0968	58.7491
60%	9.9725	0.0874	0.1193	75.9248
70%	11.5933	0.0999	0.1588	101.1822
75%	12.3803	0.1066	0.1839	131.8647
80%	13.1763	0.1147	0.2158	166.4229
85%	14.0093	0.1323	0.2614	247.1094
90%	14.9382	0.1624	0.3357	361.3737
95%	16.2717	0.2064	0.5451	647.8338
99%	19.1976	0.3807	1.1668	956.3541

Table 4.2. Simulation results for regression function g_1 with a sharp peak at 7.7968 and size 0.2019. Number of repetitions: 10,000.

Sample size	Number of monotonic fits[a]	Mean bias	Median bias	Root MSE	Median abs. error	St. dev. of estimators	Range of estimators	Mean Std. Error	Median Std. Error	Coverage Rate nom. 95%	Coverage Rate nom. 90%
					Nonparametric Estimator of Location						
100	76	-0.2882	-0.2644	0.6716	0.4087	0.6067	[4.0033, 16.8752]	1.0972	1.0022	0.9866	0.9733
300	6	-0.158	-0.1538	0.3287	0.2196	0.2882	[6.0702, 8.7516]	0.6298	0.615	0.9987	0.9945
500	2	-0.1178	-0.1136	0.2482	0.1657	0.2185	[6.744, 8.4803]	0.4946	0.4866	0.9995	0.9976
1000	0	-0.0725	-0.0717	0.1685	0.1151	0.1521	[7.0848, 8.2812]	0.3567	0.3539	0.9997	0.9988
3000	0	-0.0193	-0.0203	0.0937	0.063	0.0917	[7.4094, 8.1134]	0.2154	0.2149	1	0.9998
					Cubic Estimator of Location						
100	146	-0.8758	-0.8869	1.1483	0.9057	0.7427	[3.8122, 11.4044]	0.7543	0.6286	0.6995	0.5985
300	141	-0.8156	-0.8177	0.9075	0.8179	0.398	[4.5898, 8.791]	0.3892	0.3716	0.4221	0.3145
500	194	-0.8043	-0.8071	0.8589	0.8071	0.3014	[5.5749, 8.153]	0.2963	0.2885	0.2227	0.1504
1000	200	-0.7899	-0.7922	0.8173	0.7922	0.2102	[6.1521, 7.8928]	0.2068	0.2042	0.0409	0.0217
3000	84	-0.7853	-0.7852	0.7944	0.7852	0.1204	[6.4995, 7.4827]	0.1183	0.1178	0	0
					Nonparametric Estimator of Size						
100	76	-0.0269	-0.0268	0.0311	0.0268	0.0156	[0.1145, 0.235]	0.0146	0.0146	0.5441	0.4305
300	6	-0.0203	-0.0205	0.0225	0.0205	0.0096	[0.1467, 0.2204]	0.0093	0.0093	0.4104	0.298
500	2	-0.0177	-0.0176	0.0193	0.0176	0.0079	[0.1486, 0.2166]	0.0076	0.0076	0.3644	0.2519
1000	0	-0.0141	-0.014	0.0153	0.014	0.0059	[0.1641, 0.2094]	0.0057	0.0057	0.3119	0.2096
3000	0	-0.0097	-0.0098	0.0104	0.0098	0.0038	[0.1784, 0.2075]	0.0037	0.0037	0.2477	0.1616
					Cubic Estimator of Size						
100	146	-0.0339	-0.0339	0.0374	0.0339	0.0159	[0.1092, 0.2282]	0.0156	0.0156	0.4121	0.2999
300	141	-0.0367	-0.0368	0.0378	0.0368	0.009	[0.132, 0.2054]	0.0089	0.0089	0.0172	0.0083
500	194	-0.0372	-0.0372	0.0379	0.0372	0.007	[0.1391, 0.1923]	0.0069	0.0069	0.0005	0.0003
1000	200	-0.0376	-0.0376	0.0379	0.0376	0.0049	[0.1459, 0.1823]	0.0049	0.0049	0	0
3000	84	-0.0378	-0.0378	0.0379	0.0378	0.0028	[0.1537, 0.1755]	0.0028	0.0028	0	0

a. For an explanation of what we mean by "monotonic fit" see Section 4.2.

Table 4.3. Simulation results for regression function g_2 with a smooth peak at 9.7418 and size 0.2021. Number of repetitions: 10,000.

Sample size	Number of monotonic fits[a]	Mean bias	Median bias	Root MSE	Median abs. error	St. dev. of estimators	Range of estimators	Mean Std. Error	Median Std. Error	Coverage Rate nom. 95%	Coverage Rate nom. 90%
Nonparametric Estimator of Location											
100	983	1.4983	1.1561	2.811	1.5476	2.3785	[3.9957, 19.6356]	2.69	1.8667	0.8036	0.7643
300	251	1.3202	0.9101	2.3459	1.1028	1.9392	[6.0027, 19.994]	1.9085	1.6136	0.8507	0.8163
500	77	1.134	0.7776	2.011	0.928	1.6609	[6.7419, 19.1878]	1.6164	1.4557	0.8767	0.8462
1000	7	0.849	0.6197	1.5004	0.7129	1.2372	[7.2079, 19.2959]	1.3079	1.2347	0.9218	0.8923
3000	0	0.5184	0.4321	0.8718	0.4841	0.701	[8.4736, 14.8575]	0.9246	0.8901	0.9742	0.9451
Cubic Estimator of Location											
100	921	1.1186	1.1191	2.2035	1.5387	1.8986	[4.2016, 19.7164]	1.9042	1.5334	0.7945	0.7305
300	222	1.2033	1.1506	1.7083	1.1975	1.2127	[7.1026, 18.8452]	1.2282	1.1112	0.8012	0.7276
500	67	1.1861	1.1688	1.5205	1.1771	0.9514	[7.3492, 16.416]	0.9492	0.8975	0.7509	0.6582
1000	11	1.1781	1.1718	1.361	1.1718	0.6814	[8.6094, 13.665]	0.6767	0.6596	0.6021	0.4766
3000	0	1.1753	1.1668	1.2386	1.1668	0.391	[9.5347, 12.4526]	0.3954	0.3911	0.148	0.0859
Nonparametric Estimator of Size											
100	983	0.0054	0.0046	0.0156	0.01	0.0147	[0.1563, 0.2985]	0.0167	0.016	0.9589	0.9183
300	251	0.0022	0.0019	0.0096	0.0062	0.0093	[0.1711, 0.2949]	0.0109	0.0106	0.9723	0.9355
500	77	0.0014	0.0012	0.0077	0.0051	0.0076	[0.1799, 0.2373]	0.0088	0.0087	0.9694	0.9366
1000	7	0.0004	0.0003	0.0058	0.0039	0.0058	[0.182, 0.2253]	0.0067	0.0066	0.973	0.9405
3000	0	-0.00004	-0.0001	0.0039	0.0027	0.0039	[0.184, 0.2187]	0.0043	0.0043	0.9718	0.9329
Cubic Estimator of Size											
100	921	0.0034	0.003	0.0146	0.0096	0.0142	[0.1584, 0.2583]	0.0155	0.0151	0.961	0.9168
300	222	-0.0005	-0.000501	0.0082	0.0055	0.0081	[0.1701, 0.2386]	0.0085	0.0083	0.9592	0.9135
500	67	-0.0011	-0.0012	0.0065	0.0043	0.0064	[0.1793, 0.2317]	0.0065	0.0064	0.9501	0.899
1000	11	-0.0017	-0.0017	0.0048	0.0032	0.0045	[0.184, 0.219]	0.0045	0.0045	0.9354	0.8764
3000	0	-0.002	-0.002	0.0033	0.0023	0.0026	[0.1895, 0.2095]	0.0026	0.0026	0.8772	0.8003

a. For an explanation of what we mean by "monotonic fit" see Section 4.2.

Table 4.4. Simulation results for regression function g_3 with peak at 8.5480 and size 0.2059. Number of repetitions: 10,000.

Sample size	Number of monotonic fits[a]	Mean bias	Median bias	Root MSE	Median abs. error	St. dev. of estimators	Range of estimators	Mean Std. Error	Median Std. Error	Coverage Rate nom. 95%	Coverage Rate nom. 90%
					Nonparametric Estimator of Location						
100	1719	-0.9605	-0.8996	1.5344	0.9182	1.1967	[3.1597, 19.5984]	2.6788	1.4978	0.937	0.9048
300	590	-0.6578	-0.6149	0.811	0.6152	0.4744	[2.9875, 17.4201]	1.0967	0.9924	0.9868	0.9713
500	198	-0.5115	-0.4959	0.5844	0.4959	0.2828	[3.4566, 17.8204]	0.8216	0.7905	0.9963	0.9877
1000	13	-0.373	-0.3676	0.4041	0.3676	0.1557	[7.4337, 8.708]	0.58	0.5697	0.9995	0.9968
3000	0	-0.2188	-0.2175	0.2324	0.2175	0.0785	[8.0223, 8.6256]	0.3357	0.3341	1	0.9996
					Cubic Estimator of Location						
100	2027	-2.1707	-2.2853	2.5535	2.3137	1.345	[3.2992, 19.057]	1.4614	1.0434	0.4524	0.3586
300	2183	-2.4295	-2.3697	2.5786	2.3706	0.8643	[2.9406, 11.6721]	1.0033	0.7607	0.2239	0.1465
500	2650	-2.4422	-2.364	2.5434	2.364	0.7103	[2.7918, 8.679]	0.7708	0.6284	0.0759	0.0399
1000	3566	-2.3994	-2.3457	2.4505	2.3457	0.4977	[3.0267, 7.5747]	0.5092	0.4603	0.0016	0.0005
3000	4706	-2.3738	-2.3599	2.3884	2.3599	0.2635	[4.7437, 6.9549]	0.2736	0.266	0	0
					Nonparametric Estimator of Size						
100	1719	-0.0405	-0.0406	0.0431	0.0406	0.0146	[0.1094, 0.2218]	0.015	0.0148	0.2255	0.1505
300	590	-0.0366	-0.0367	0.0379	0.0367	0.01	[0.1304, 0.2091]	0.0096	0.0095	0.0414	0.0204
500	198	-0.0332	-0.0332	0.0342	0.0332	0.0084	[0.1452, 0.204]	0.0078	0.0078	0.0195	0.0091
1000	13	-0.028	-0.028	0.0287	0.028	0.0063	[0.1547, 0.2022]	0.006	0.006	0.0053	0.0025
3000	0	-0.0203	-0.0203	0.0207	0.0203	0.0041	[0.1699, 0.2006]	0.0039	0.0039	0.0011	0.0003
					Cubic Estimator of Size						
100	2027	-0.0463	-0.0468	0.0487	0.0468	0.0152	[0.1118, 0.2222]	0.0161	0.016	0.1751	0.1064
300	2183	-0.0517	-0.0518	0.0525	0.0518	0.0091	[0.1209, 0.1919]	0.0092	0.0091	0.0001	0.0001
500	2650	-0.0531	-0.0531	0.0536	0.0531	0.0071	[0.1242, 0.1835]	0.0071	0.007	0	0
1000	3566	-0.0542	-0.0542	0.0544	0.0542	0.005	[0.1313, 0.1712]	0.0049	0.0049	0	0
3000	4706	-0.0547	-0.0546	0.0547	0.0546	0.0028	[0.142, 0.1623]	0.0028	0.0028	0	0

a. For an explanation of what we mean by "monotonic fit" see Section 4.2.

Table 4.5. Simulation results for regression function g1 when using global and local bandwidths in estimation. Number of repetitions: 10,000.

Sample size	No. of monotonic fits[a,b]	mean bandwidth	Mean bias	Median bias	Root MSE	Median abs. error	St. dev. of estimators	Range of estimators	Mean Std. Error	Median Std. Error	Coverage Rate nom. 95%	Coverage Rate nom. 90%
Nonparametric Estimator of Size using a global bandwidth												
100	93	1.61	-0.0268	-0.0267	0.0309	0.0267	0.0153	[0.1044, 0.2305]	0.0146	0.0145	0.5498	0.4313
300	2	1.2792	-0.0203	-0.0203	0.0226	0.0203	0.0098	[0.1415, 0.2166]	0.0093	0.0093	0.4182	0.3059
500	0	1.1499	-0.0175	-0.0176	0.0192	0.0176	0.0078	[0.1567, 0.2156]	0.0076	0.0076	0.3637	0.2624
1000	0	0.994	-0.0141	-0.0141	0.0153	0.0141	0.0058	[0.1657, 0.2076]	0.0057	0.0057	0.3113	0.212
3000	0	0.7892	-0.0097	-0.0097	0.0104	0.0097	0.0037	[0.1789, 0.2056]	0.0037	0.0037	0.2505	0.162
Nonparametric estimator of size based on local bandwidth and using true value of second derivative at estimated maximum [a]												
100	93	0.8195	-0.009	-0.0089	0.023	0.0156	0.0212	[0.1028, 0.2729]	0.021	0.0211	0.8061	0.6859
300	2	0.5736	-0.0053	-0.0054	0.0147	0.01	0.0137	[0.1436, 0.2444]	0.0139	0.0139	0.7588	0.605
500	0	0.5068	-0.0041	-0.0043	0.0119	0.0081	0.0111	[0.1547, 0.2368]	0.0114	0.0114	0.7324	0.559
1000	0	0.4327	-0.0031	-0.0032	0.0091	0.0062	0.0085	[0.1635, 0.2313]	0.0087	0.0087	0.6855	0.5106
3000	0	0.3398	-0.002	-0.002	0.0058	0.004	0.0054	[0.1786, 0.2194]	0.0056	0.0056	0.632	0.4436
Nonparametric estimator of size based on local bandwidth and using an estimate of second derivative at estimated maximum [a]												
100	93	1.5787	-0.0249	-0.0249	0.0306	0.025	0.0178	[0.1007, 0.245]	0.0149	0.0148	0.5598	0.4455
300	2	1.0896	-0.0158	-0.0158	0.0196	0.0159	0.0116	[0.1403, 0.2284]	0.0101	0.0101	0.482	0.3585
500	0	0.9283	-0.0124	-0.0126	0.0155	0.0127	0.0093	[0.159, 0.2278]	0.0084	0.0084	0.4471	0.3239
1000	0	0.752	-0.0088	-0.0089	0.0113	0.009	0.007	[0.1666, 0.2189]	0.0066	0.0066	0.4187	0.2938
3000	0	0.5455	-0.005	-0.005	0.0067	0.0052	0.0045	[0.181, 0.2132]	0.0044	0.0044	0.3884	0.2602

a. As local bandwidth we use an estimate of the optimal bandwidth for the Nadayara-Watson estimator at the estimated location of the maximum. This optimal bandwidth depends on the second derivative of the regression function at the estimated location of the maximum. For comparison purposes here we present results for two estimated optmial bandwidths: one that uses the true value of the second derivative at the maximum and one that uses an estimate of this quantity.

b. For an explanation of what we mean by "monotonic fit" see Section 4.2.

Table 4.6. Simulation results for regression function g2 when using global and local bandwidths in estimation. Number of repetitions: 10,000.

Sample size	No. of monotonic fits[a,b]	mean bandwidth	Mean bias	Median bias	Root MSE	Median abs. error	St. dev. of estimators	Range of estimators	Mean Std. Error	Median Std. Error	Coverage Rate nom. 95%	Coverage Rate nom. 90%
Nonparametric Estimator of Size using a global bandwidth												
100	964	1.6074	0.0061	0.0054	0.0161	0.0102	0.0149	[0.1595, 0.2907]	0.0168	0.0161	0.9585	0.916
300	258	1.2796	0.0022	0.0019	0.0096	0.0063	0.0093	[0.175, 0.2613]	0.0109	0.0106	0.9706	0.9371
500	87	1.1496	0.0013	0.0011	0.0077	0.0051	0.0076	[0.1767, 0.2618]	0.0088	0.0087	0.9729	0.9372
1000	8	0.9938	0.0005	0.0004	0.0058	0.0039	0.0058	[0.182, 0.227]	0.0067	0.0066	0.9721	0.9397
3000	0	0.7892	-5×10^{-5}	-0.0001	0.0039	0.0026	0.0039	[0.1903, 0.2179]	0.0043	0.0043	0.9696	0.9323
Nonparametric estimator of size based on local bandwidth and using true value of second derivative at estimated maximum [a]												
100	964	3.0505	-0.003	-0.0034	0.0146	0.0098	0.0143	[0.1494, 0.267]	0.0132	0.0132	0.8796	0.8164
300	258	2.3487	-0.0036	-0.0035	0.0101	0.0068	0.0094	[0.1613, 0.2388]	0.0087	0.0089	0.9094	0.8521
500	87	2.0618	-0.0032	-0.0032	0.0084	0.0055	0.0078	[0.1697, 0.2256]	0.0072	0.0075	0.9174	0.8619
1000	8	1.6583	-0.0025	-0.0022	0.0067	0.0042	0.0062	[0.172, 0.2236]	0.0055	0.0058	0.9242	0.8732
3000	0	1.1265	-0.0015	-0.0013	0.0043	0.0027	0.0041	[0.1749, 0.2168]	0.0038	0.0039	0.9321	0.8763
Nonparametric estimator of size based on local bandwidth and using an estimate of second derivative at estimated maximum [a]												
100	964	2.3227	0.0017	0.0009	0.0154	0.01	0.0153	[0.1531, 0.2777]	0.0144	0.014	0.9336	0.8743
300	258	1.814	-0.0011	-0.0015	0.0097	0.0065	0.0096	[0.1715, 0.2426]	0.0092	0.0091	0.9463	0.8915
500	87	1.6193	-0.0014	-0.0017	0.008	0.0054	0.0078	[0.1738, 0.2548]	0.0075	0.0075	0.9455	0.8931
1000	8	1.3786	-0.0015	-0.0016	0.0062	0.0042	0.006	[0.1789, 0.228]	0.0057	0.0057	0.9459	0.8921
3000	0	1.058	-0.0012	-0.0013	0.0041	0.0028	0.0039	[0.1874, 0.2174]	0.0038	0.0038	0.9372	0.8816

a. As local bandwidth we use an estimate of the optimal bandwidth for the Nadayara-Watson estimator at the estimated location of the maximum. This optimal bandwidth depends on the second derivative of the regression function at the estimated location of the maximum. For comparison purposes here we present results for two estimated optmial bandwidths: one that uses the true value of the second derivative at the maximum and one that uses an estimate of this quantity.

b. For an explanation of what we mean by "monotonic fit" see Section 4.2.

Table 4.7. Simulation results for regression function g3 when using global and local bandwidths in estimation. Number of repetitions: 10,000.

Sample size	No. of monotonic fits[a,b]	mean bandwidth	Mean bias	Median bias	Root MSE	Median abs. error	St. dev. of estimators	Range of estimators	Mean Std. Error	Median Std. Error	Coverage Rate nom. 95%	Coverage Rate nom. 90%
						Nonparametric Estimator of Size using a global bandwidth						
100	1703	1.6092	-0.0408	-0.0411	0.0433	0.0411	0.0147	[0.1178, 0.2194]	0.015	0.0148	0.2227	0.1416
300	548	1.2805	-0.0363	-0.0364	0.0377	0.0364	0.01	[0.1272, 0.2154]	0.0096	0.0095	0.0456	0.0203
500	182	1.1498	-0.0333	-0.0333	0.0343	0.0333	0.0082	[0.1448, 0.2046]	0.0078	0.0078	0.0165	0.007
1000	6	0.9941	-0.028	-0.028	0.0287	0.028	0.0063	[0.1537, 0.2009]	0.006	0.006	0.0058	0.0018
3000	0	0.7893	-0.0203	-0.0203	0.0207	0.0203	0.004	[0.1707, 0.2014]	0.0039	0.0039	0.0009	0.0001
						Nonparametric estimator of size based on local bandwidth and using true value of second derivative at estimated maximum [a]						
100	1703	0.971	-0.0266	-0.0292	0.0372	0.0306	0.0259	[0.1079, 0.2878]	0.0211	0.0212	0.5089	0.3558
300	548	0.6975	-0.0207	-0.021	0.0284	0.0217	0.0194	[0.1214, 0.2603]	0.0139	0.0143	0.2352	0.1285
500	182	0.5277	-0.0157	-0.0155	0.022	0.0162	0.0154	[0.1216, 0.2513]	0.012	0.0123	0.1504	0.0728
1000	6	0.374	-0.0096	-0.0098	0.0145	0.0107	0.011	[0.1531, 0.2408]	0.0098	0.0098	0.0926	0.0355
3000	0	0.2723	-0.0046	-0.0047	0.0082	0.0058	0.0068	[0.1749, 0.2284]	0.0066	0.0066	0.0374	0.0119
						Nonparametric estimator of size based on local bandwidth and using an estimate of second derivative at estimated maximum [a]						
100	1703	2.0292	-0.0418	-0.0425	0.045	0.0425	0.0168	[0.1037, 0.2345]	0.0143	0.0142	0.2156	0.1438
300	548	1.2293	-0.0341	-0.0345	0.0362	0.0345	0.0121	[0.1257, 0.2275]	0.0099	0.0099	0.059	0.0309
500	182	1.0129	-0.0293	-0.0295	0.0311	0.0295	0.0103	[0.1422, 0.2207]	0.0083	0.0083	0.0283	0.012
1000	6	0.7961	-0.0218	-0.0218	0.0232	0.0218	0.0081	[0.1536, 0.2175]	0.0067	0.0067	0.013	0.0049
3000	0	0.5531	-0.0122	-0.0123	0.0133	0.0123	0.0052	[0.1747, 0.212]	0.0046	0.0046	0.004	0.0011

a. As local bandwidth we use an estimate of the optimal bandwidth for the Nadayara-Watson estimator at the estimated location of the maximum. This optimal bandwidth depends on the second derivative of the regression function at the estimated location of the maximum. For comparison purposes here we present results for two estimated optmial bandwidths: one that uses the true value of the second derivative at the maximum and one that uses an estimate of this quantity.

b. For an explanation of what we mean by "monotonic fit" see Section 4.2.

Table 4.8. Matrix of correlation coefficients.

	Per-capita			
	Income	Emissions of NO_X	Emissions of SO_2	Population density
Income	1			
Emissions of NO_x	0.2363	1		
Emissions of SO_2	-0.1296	0.3014	1	
Population density	0.3292	-0.2275	-0.1858	

Table 4.9. Simulation results for model based on regression function $g_1(t,x)$. In this case the dose-response function has a sharp peak at 9.2982 with size 0.2107. Number of repetitions: 1,000.

Sample size	Number of monotonic fits[a]	Mean bias	Median bias	Root MSE	Median abs. error	St. dev. of estimators	Range of estimators	Mean Std. Error	Median Std. Error	Coverage Rate nom. 95%	Coverage Rate nom. 90%
					Nonparametric Estimator of Location						
100	243	-0.4885	-0.3501	1.1934	0.6305	1.0896	[5.08, 11.70]	2.413	2.065	1	1
300	280	-0.2834	-0.0501	1.0607	0.4154	1.0228	[5.04, 11.04]	1.4537	0.9693	0.9944	0.9736
500	323	-0.1192	0.0547	0.9947	0.3564	0.9883	[4.98, 10.92]	1.4402	0.6813	0.9764	0.9439
1000	471	0.1287	0.2265	0.8415	0.3039	0.8324	[4.98, 10.80]	0.7761	0.4346	0.896	0.8563
2000	513	0.0909	0.2362	0.9229	0.2846	0.9194	[4.89, 10.80]	0.3211	0.2752	0.807	0.7413
					Cubic Estimator of Location						
100	44	-2.1043	-2.2371	2.3375	2.2406	1.0183	[4.48, 11.63]	0.98	0.8073	0.3065	0.2469
300	1	-2.3075	-2.3423	2.3679	2.3423	0.5316	[4.37, 8.89]	0.5042	0.4429	0.042	0.026
500	0	-2.3091	-2.3221	2.3446	2.3221	0.4064	[4.19, 8.47]	0.3628	0.343	0.005	0
1000	0	-2.3481	-2.3432	2.3617	2.3432	0.2529	[6.14, 7.83]	0.2442	0.2386	0	0
2000	0	-2.3429	-2.3401	2.3487	2.3401	0.1661	[6.41, 7.61]	0.1696	0.1685	0	0
					Nonparametric Estimator of Size						
100	243	-0.0021	-0.0027	0.0273	0.018	0.0273	[0.1074, 0.2918]	0.0205	0.0203	0.8666	0.79
300	280	-0.0048	-0.0051	0.0192	0.0128	0.0186	[0.1560, 0.2717]	0.0122	0.0122	0.7875	0.7111
500	323	-0.0071	-0.0069	0.0165	0.0114	0.0149	[0.1563, 0.2588]	0.0095	0.0095	0.7371	0.6632
1000	471	-0.0069	-0.0065	0.0125	0.0081	0.0105	[0.1707, 0.2330]	0.0069	0.0069	0.7391	0.6616
2000	513	-0.004	-0.0036	0.0086	0.0055	0.0076	[0.1779, 0.2462]	0.005	0.005	0.7515	0.6571
					Cubic Estimator of Size						
100	44	-0.1016	-0.0996	0.1066	0.0996	0.0325	[-0.017, 0.2008]	0.02	0.0198	0.022	0.0115
300	1	-0.1006	-0.1003	0.1023	0.1003	0.0187	[0.0582, 0.1725]	0.0113	0.0113	0	0
500	0	-0.1022	-0.1019	0.1033	0.1019	0.0146	[0.0669, 0.1549]	0.0087	0.0088	0	0
1000	0	-0.102	-0.1017	0.1026	0.1017	0.0103	[0.0726, 0.1415]	0.0062	0.0062	0	0
2000	0	-0.102	-0.1022	0.1023	0.1022	0.0069	[0.0865, 0.1297]	0.0044	0.0044	0	0

a. For an explanation of what we mean by "monotonic fit" see Section 4.2.

Table 4.10. Simulation results for model based on regression function $g_2(t,x)$. In this case the dose-response function has a smooth peak at 9.4262 with size 0.2354. Number of repetitions: 1,000.

Sample size	Number of monotonic fits[a]	Mean bias	Median bias	Root MSE	Median abs. error	St. dev. of estimators	Range of estimators	Mean Std. Error	Median Std. Error	Coverage Rate nom. 95%	Coverage Rate nom. 90%
Nonparametric Estimator of Location											
100	195	0.2735	0.1245	2.1628	1.1908	2.1468	[5.51, 18.47]	9.3641	3.4181	0.9963	0.9913
300	174	-0.1552	0.042	1.8021	0.9117	1.7965	[5.06, 18.36]	3.0951	1.9762	0.9867	0.9649
500	212	-0.1778	0.082	1.6302	0.8884	1.6215	[5.15, 18.36]	1.9679	1.5423	0.9772	0.9454
1000	213	-0.128	0.2711	1.5987	0.7282	1.5946	[4.99, 12.26]	2.8759	1.1504	0.9454	0.9098
2000	237	0.0579	0.3899	1.4788	0.6464	1.4787	[4.84, 12.01]	9.4162	0.805	0.9292	0.8938
Cubic Estimator of Location											
100	102	-0.1392	-0.3879	1.5116	1.1085	1.506	[5.71, 15.61]	1.4713	1.2082	0.8185	0.7639
300	86	-0.3748	-0.5154	0.954	0.7236	0.8777	[6.80, 12.76]	0.8332	0.7448	0.8162	0.744
500	74	-0.3948	-0.4879	0.7763	0.6073	0.6688	[7.45, 11.63]	0.6334	0.6007	0.797	0.7289
1000	60	-0.4681	-0.5006	0.6545	0.5269	0.4576	[7.67, 10.99]	0.4344	0.4206	0.7223	0.6447
2000	13	-0.5064	-0.5249	0.5978	0.5295	0.3178	[8.17, 10.36]	0.3018	0.2967	0.5542	0.4519
Nonparametric Estimator of Size											
100	195	0.0249	0.0249	0.0365	0.0263	0.0267	[0.10, 0.3416]	0.0227	0.0219	0.7478	0.6621
300	174	0.0184	0.0183	0.0246	0.0186	0.0163	[0.2092, 0.3054]	0.0132	0.013	0.6828	0.5787
500	212	0.0142	0.0142	0.0198	0.0147	0.0138	[0.1968, 0.2940]	0.0102	0.0102	0.6751	0.5736
1000	213	0.0116	0.0116	0.0157	0.0117	0.0105	[0.198, 0.2816]	0.0074	0.0074	0.6188	0.526
2000	237	0.0107	0.0105	0.0133	0.0106	0.0079	[0.2176, 0.2731]	0.0053	0.0054	0.4889	0.3932
Cubic Estimator of Size											
100	102	-0.0652	-0.0658	0.0715	0.0658	0.0292	[0.0739, 0.2704]	0.0185	0.0182	0.1626	0.1036
300	86	-0.0656	-0.0657	0.0679	0.0657	0.0175	[0.1173, 0.2391]	0.0103	0.0102	0.0044	0.0044
500	74	-0.0673	-0.0671	0.0686	0.0671	0.0132	[0.1302, 0.2062]	0.0079	0.0079	0	0
1000	60	-0.0675	-0.0672	0.0682	0.0672	0.0095	[0.1394, 0.1986]	0.0056	0.0056	0	0
2000	13	-0.0683	-0.0684	0.0686	0.0684	0.0064	[0.1459, 0.1862]	0.0039	0.0039	0	0

a. For an explanation of what we mean by "monotonic fit" see Section 4.2.

Table 4.11. Simulation results for model based on regression function $g_1(t,x)$ and for which the dose-response function has a sharp peak at 9.2982 with size 0.2107. In this case we restrict the search of the peak to the points between the 25th and 75th sample percentile of the simulated per-capita income. Number of repetitions: 1,000.

Sample size	Number of monotonic fits[a]	Mean bias	Median bias	Root MSE	Median abs. error	St. dev. of estimators	Range of estimators	Mean Std. Error	Median Std. Error	Coverage Rate nom. 95%	Coverage Rate nom. 90%
					Nonparametric Estimator of Location						
100	116	-0.5039	-0.2978	1.2923	0.635	1.1907	[5.20, 11.65]	2.7647	2.1197	1	0.9989
300	72	-0.2228	-0.0771	0.943	0.4366	0.9168	[5.78, 11.21]	1.2022	1.0034	0.9989	0.9806
500	55	-0.0975	0.0558	0.8542	0.3525	0.8491	[5.77, 11.23]	0.8072	0.705	0.9778	0.9503
1000	42	0.0655	0.0989	0.697	0.2763	0.6943	[6.02, 11.16]	0.5175	0.4457	0.9113	0.8622
2000	3	0.2148	0.2059	0.423	0.2628	0.3645	[6.39, 10.81]	0.2862	0.2831	0.8485	0.7904
					Cubic Estimator of Location						
100	104	-1.9593	-2.098	2.1765	2.098	0.9483	[4.46, 11.22]	0.9862	0.8015	0.3828	0.2991
300	23	-2.2953	-2.3391	2.3477	2.3391	0.4937	[5.66, 8.89]	0.4894	0.4535	0.0358	0.0225
500	7	-2.3244	-2.3385	2.3527	2.3385	0.3639	[5.81, 8.44]	0.3544	0.3356	0.003	0
1000	0	-2.3508	-2.355	2.3648	2.355	0.2571	[6.17, 7.98]	0.2434	0.2383	0	0
2000	0	-2.3487	-2.3522	2.3548	2.3522	0.17	[6.23, 7.57]	0.1692	0.1668	0	0
					Nonparametric Estimator of Size						
100	116	-0.0062	-0.0055	0.0292	0.0197	0.0285	[0.1236, 0.2966]	0.0205	0.0205	0.8167	0.75
300	72	-0.0115	-0.0118	0.0223	0.0153	0.0191	[0.1399, 0.2765]	0.0121	0.012	0.708	0.6164
500	55	-0.0127	-0.0125	0.0201	0.0143	0.0156	[0.1428, 0.2445]	0.0095	0.0095	0.6106	0.5407
1000	42	-0.0113	-0.0115	0.0165	0.0123	0.012	[0.161, 0.2448]	0.0069	0.0069	0.5449	0.4562
2000	3	-0.0091	-0.0093	0.0125	0.0096	0.0085	[0.174, 0.2272]	0.005	0.005	0.5045	0.4173
					Cubic Estimator of Size						
100	104	-0.1014	-0.0997	0.1065	0.0997	0.0325	[0.0134, 0.2113]	0.0198	0.0197	0.0201	0.0156
300	23	-0.1024	-0.103	0.104	0.103	0.0178	[0.0434, 0.1642]	0.0113	0.0113	0	0
500	7	-0.1023	-0.102	0.1033	0.102	0.0145	[0.0568, 0.1527]	0.0087	0.0087	0	0
1000	0	-0.1014	-0.1011	0.1019	0.1011	0.0101	[0.0767, 0.1383]	0.0062	0.0062	0	0
2000	0	-0.1022	-0.1023	0.1025	0.1023	0.0074	[0.0844, 0.1318]	0.0044	0.0044	0	0

a. For an explanation of what we mean by "monotonic fit" see Section 4.2.

Table 4.12. Simulation results for model based on regression function $g_2(t,x)$ and for which the dose-response function has a smooth peak at 9.4262 with size 0.2354. In this case we restrict the search of the peak to the points between the 25th and 75th sample percentile of the simulated per-capita income. Number of repetitions: 1,000.

Sample size	Number of monotonic fits[a]	Mean bias	Median bias	Root MSE	Median abs. error	St. dev. of estimators	Range of estimators	Mean Std. Error	Median Std. Error	Coverage Rate nom. 95%	Coverage Rate nom. 90%
					Nonparametric Estimator of Location						
100	109	-0.253	-0.1978	1.7197	1.2668	1.7019	[4.67, 13.67]	5.4938	3.5125	0.9989	0.9989
300	99	-0.085	0.1026	1.4777	0.9019	1.476	[5.24, 12.83]	2.3707	2.0502	0.9945	0.98
500	77	0.1089	0.283	1.344	0.8122	1.3404	[5.89, 12.73]	1.7566	1.5791	0.9772	0.9523
1000	79	0.2315	0.3705	1.2182	0.6814	1.1966	[5.88, 12.36]	1.2328	1.1311	0.9446	0.911
2000	57	0.1954	0.4204	1.1654	0.6215	1.1495	[5.94, 11.94]	0.8518	0.7976	0.9109	0.8717
					Cubic Estimator of Location						
100	95	-0.5062	-0.697	1.3661	1.0799	1.2695	[5.60, 12.50]	1.431	1.1519	0.7901	0.737
300	11	-0.4032	-0.5036	0.9522	0.7385	0.863	[7.24, 12.17]	0.841	0.7625	0.8079	0.7371
500	0	-0.4081	-0.4721	0.788	0.6129	0.6745	[7.32, 11.60]	0.6445	0.6048	0.784	0.717
1000	0	-0.4791	-0.5017	0.668	0.5353	0.4657	[7.74, 11.25]	0.4364	0.4233	0.701	0.62
2000	0	-0.4976	-0.5132	0.5921	0.5173	0.321	[8.01, 10.30]	0.3019	0.2959	0.568	0.471
					Nonparametric Estimator of Size						
100	109	0.021	0.0213	0.0346	0.0247	0.0276	[0.1682, 0.3423]	0.0216	0.0214	0.7699	0.6723
300	99	0.014	0.0143	0.0223	0.0159	0.0174	[0.1892, 0.3064]	0.013	0.013	0.7314	0.6349
500	77	0.0113	0.0112	0.0184	0.0128	0.0146	[0.1995, 0.2951]	0.0102	0.0102	0.7075	0.6381
1000	79	0.0102	0.0101	0.0149	0.0111	0.0108	[0.2079, 0.2791]	0.0074	0.0074	0.6363	0.5581
2000	57	0.0091	0.0089	0.0124	0.0094	0.0085	[0.2165, 0.2726]	0.0053	0.0053	0.5493	0.474
					Cubic Estimator of Size						
100	95	-0.0643	-0.0635	0.071	0.0635	0.0303	[0.077, 0.2714]	0.0185	0.0184	0.168	0.1249
300	11	-0.068	-0.0683	0.0702	0.0683	0.0176	[0.1053, 0.2218]	0.0103	0.0103	0.002	0.001
500	0	-0.0681	-0.0685	0.0694	0.0685	0.0132	[0.1243, 0.2106]	0.0079	0.0079	0	0
1000	0	-0.0681	-0.0681	0.0687	0.0681	0.0096	[0.139, 0.1956]	0.0056	0.0056	0	0
2000	0	-0.0682	-0.0683	0.0685	0.0683	0.0065	[0.148, 0.1875]	0.0039	0.0039	0	0

a. For an explanation of what we mean by "monotonic fit" see Section 4.2.

Bibliography

[1] Ahmad, I. A. and Ullah, A. (1988). "Nonparametric Estimation of the p-th Order Derivative of a Regression Function", Research Report 8903, University of Western Ontario.

[2] Ahn, Hyungtaik (1995). "Nonparametric Two-Stage Estimation of Conditional Choice Probabilities in a Binary Choice Model Under Uncertainty", *Journal of Econometrics,* 67, 337-378.

[3] Arrow, K.; Bolin, B.; Costanza, R.; Dasgupta, P.; et al. (1995). "Economic Growth, Carrying Cpacity, and the Environment", *Science,* 268, 520-521.

[4] Azomahou, T. and Nguyen, V.P. (2001). "Economic Growth and CO_2 Emissions" A Nonparametric Approach", *Working Paper*, Université Louis Pasteur.

[5] Bai, Z.; Chen, Z. and Wu, Y. (2003)."Convergence Rate of the Best-r-Point-Average Estimator for the Maximizer of a Nonparametric Regression Function", *Journal of Multivariate Analysis*, 84, 319-334.

[6] Bai, Z. and Huang, M. (1999)."On Consistency of the Best-r-Point-Average Estimator

for the Maximizer of a Nonparametric Regression Function", *Sankhya: The Indian Journal of Statistics*, A, 61, 208-217.

[7] Baltagi, B.; Hidalgo, J. and Li, Q. (1996). "A Nonparametric Test for Poolability using Panel Data", *Journal of Econometrics*, 75, 345-367.

[8] Behrman, J.; Cheng, Y. and Todd, P. (2004). "Evaluating Preschool Programs when Length of Exposure to the Program Varies: A Nonparametric Approach", *Review of Economics and Statistics,* 86(1), 108-132.

[9] Bhattacharya, P. K. and Gangopadhyay, A. K. (1990). "Kernel and Nearest-Neighbor Estimation of Conditional Quantile", *The Annals of Statistics,* 18(3), 1400-1415

[10] Bierens, H. J. (1983). "Uniform Consistency of Kernel Estimators of a Regression Function Under Generalized Conditions", *Journal of the American Statistical Association*, 78(383), 699-707.

[11] Bierens, H. J. (1987). "Kernel Estimators of Regression Functions", Ch. 3 in T. F. Bewley ed., *Advances in Econometrics*, Cambridge University Press, Vol. I, 99-144.

[12] Brodaty, T.; Crépon, B. and Fougère, D. (2001). "Using Matching Estimators to Evaluate Alternative Youth Employment Programs: Evidence from France, 1986-1988", in M. Lechner and F. Pfeiffer, eds., *Econometric Evaluation of Labour Market Policies*, Physica-Verlag, 85-123.

[13] Chaudhuri, P. (1991a). "Nonparametric Estimates of Regression Quantiles and Their Local Bahadur Representation", *The Annals of Statistics,* 19(2), 760-777.

[14] Chaudhuri, P. (1991b). "Global Nonparametric Estimation of Conditional Quantile Functions and Their Derivatives", *Journal of Multivariate Analysis,* 39, 246-269.

[15] Chen, H.; Huang, M. and Huang, W. (1996). "Estimation of the Location of the Maximum of a Regression Function using Extreme Order Statistics", *Journal of Multivariate Analysis*, 57, 191-214.

[16] Cole, M. A.; Rayner, A. J. and Bates, J.M. (1997). "The Environmental Kuznets Curve: An Empirical Analysis", *Environment and Development Economics*, 2, 401-416.

[17] Copeland, B. R. and Taylor, M. S. (2004). "Trade, Growth, and the Environment", *Journal of Economic Literature,* 42(1), 7-71.

[18] Cropper, M. and Griffiths, C. (1994). "The Interaction of Population Growth and Environmental Quality", *American Economic Review,* 84, 250-254.

[19] Dasgupta, S.; Laplante, B.; Wang, H. and Wheeler, D. (2002). "Confronting the Environmental Kuznets Curve", *Journal of Economic Perspectives,* 16, 147-168.

[20] Dehejia, R. and Wahba, S. (1995). "Causal Effects in Non-Experimental Studies: Reevaluating the Evaluation of Training Programs", manuscript, Harvard University.

[21] Dehejia, R. and Wahba, S. (1998). "Propensity Score Matching Methods for Nonexperimental Causal Studies", *NBER Technical Working Paper*, No. 6829.

[22] Dehejia, R. and Wahba, S. (1999). "Causal Effects in Non-Experimental Studies: Reevaluating the Evaluation of Training Programs", *Journal of the American Statistical Association*, 94(448), 1053-62.

[23] Eddy, W. (1980). "Optimum Kernel Estimators of the Mode", *Annals of Statistics*, 8(4), 870-882.

[24] Eddy, W. (1982). "The Asymptotic Distributions of Kernel Estimators of the Mode", *Z. Wahrsch. Verw. Gabiete*, 59, 279-290.

[25] Firpo, S. (2002). "Efficient Semiparametric Estimation of Quantile Treatment Effects", Working Paper, Department of Economics, University of California, Berkeley.

[26] Friedman, J. H. and Stuetzle, W. (1981). "Projection Pursuit Regression", *Journal of the American Statistical Association,* 76, 817-823.

[27] Gasser, T. and Müller, H.-G (1979). "Kernel Estimation of Regression Functions", in T. Gasser and M. Rosenblatt eds., *Smoothing Techniques for Curve Estimation*, Springer-Verlag, Heidelberg, 23-68.

[28] Gerfin, M. and Lechner, M. (2002). "A Microeconometric Evaluation of the Active Labour Market Policy in Switzerland", *The Economic Journal*, 112 (October), 854-893.

[29] Grossman, G. M. and Krueger, A. (1991). "Environmental Impacts of a North American Free Trade Agreement", *NBER Working Paper*, No. 3914.

[30] Grossman, G. M. and Krueger, A. (1995). "Economic Growth and the Environment", *Quarterly Journal of Economics*, 110(2), 353-377.

[31] Hahn, J. (1998). "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effets", *Econometrica*, 66(2), 315-331.

[32] Harbaugh, W. T.; Levinson, A. and Wilson, D.M. (2002). "Reexamining the Empirical Evidence for an Environmental Kuznetz Curve", *Review of Economics and Statistics,* 84(3), 541-551.

[33] Hastie, T. and Tibshirani, R. (1990). *General Additive Models,* New York, Chapman and Hall.

[34] Heckman, J. (1992). "Randomization and Social Program Evaluation" in C. Manski and I. Garfinkel, eds., *Evaluating Welfare and Training Programs*, Boston: Harvard University Press, 201-230.

[35] Heckman, J. (2000). "Causal Parameters and Policy Analysis in Economics: A Twentieth Century Retrospective", *Quarterly Journal of Economics*, 115, 45-97.

[36] Heckman, J.; Hohmann, N. and Smith, J., with M. Khoo (2000). "Substitution and Dropout Bias in Social Experiments: A Study of an Influential Social Experiment", *Quarterly Journal of Economics*, May 2000, 651-694.

[37] Heckman, J.; Ichimura, H. and Todd, P. (1998). "Matching as an Econometric Estimator", *Review of Economic Studies*, 65(2), 261-94.

[38] Heckman, J.; LaLonde, R. and Smith, J. (1999). "The Economics and Econometrics of Active Labor Market Programs", in O. Ashenfelter and D. Card, eds., *Handbook of Labor Economics*, Vol. 3A, Amsterdam: North-Holland, 1865-2097.

[39] Heckman, J. and Smith, J. (1995). "Assessing the Case for Social Experiments", *Journal of Economic Perspectives*, 9(2), 85-110.

[40] Heckman, N.E. (1992). "Bump Hunting in Regression Analysis", *Statistics and Probability Letters*, 14, 141-152.

[41] Hirano, K. and Imbens, G. W. (2004). "The Propensity Score with Continuous Treatments", *Working Paper*, Department of Economics, University of California at Berkeley.

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

[43] Holland, P. W. (1986). "Statistics and Causal Inference"(with Discussion and Reply), *Journal of the American Statistical Association*, 81(396), 945-970.

[44] Ichimura, H. (2004). "Computation of Asymptotic Distribution for Semiparametric GMM Estimators", *Working Paper,* University College London.

[45] Imbs, J. and Wacziarg, R. (2003). "Stages of Diversification", *American Economic Review*, 93(1), 63-86.

[46] Imbens, G. W. (1999). "The Role of the Propensity Score in Estimating Dose-Response Functions", *NBER Technical Working Paper*, No. 237.

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

[48] Imbens, G. W. (2001). "Some Remarks on Instrumental Variables", in M. Lechner and

F. Pfeiffer, eds., *Econometric Evaluation of Labour Market Policies*, Physica-Verlag, 17-42.

[49] Imbens, G. W. (2004). "Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review", *Review of Economics and Statistics,* 86(1), 4-29.

[50] Imbens, G. and Angrist, J. (1994). "Identification and Estimation of Local Average Treatment Effects", *Econometrica*, 62(2), 467-475.

[51] Joffe, M. and Rosenbaum, P. R. (1999). "Invited Commentary: Propensity Scores", *American Journal of Epidemiology*, 150, 327-333.

[52] Kaufmann, R. K.; Davidsdottir, B.; Garnham, S. and Pauly, P. (1998). "The Determinants of Atmospheric SO_2 Concentrations: Reconsidering the Environmental Kuznets Curve", *Ecological Economics, 25, 209-220.*

[53] Khanna, N. and Plassmann, F. (2003). "On the Future of World Pollution: The Demand for Environmental Quality and the Environmental Kuznets Curve Hypothesis", *Working Paper,* Binghamton University.

[54] List, J.A. and Gallet, C.A. (1999). "The Environmental Kuznets Curve: Does One Size Fit All?", *Ecological Economics,* (31), 409-423.

[55] Lechner, M. (2001a). "Identification and Estimation of Causal Effects of Multiple Treatments under the Conditional Independence Assumption", in M. Lechner and F. Pfeiffer, eds., *Econometric Evaluation of Labour Market Policies*, Physica-Verlag, 43-58.

[56] Lechner, M. (2001b). "Program Heterogeneity and Propensity Score Matching: An

Application to the Evaluation of Active Labor Market Policies", Discussion paper, Department of Economics, University of St. Gallen.

[57] Manski, C. F. (1997). "The Mixing Problem in Programme Evaluation", *Review of Economic Studies*, 64(4), 537–553.

[58] Manski, C. F. (2000a). "Identification Problems and Decisions Under Ambiguity: Empirical Analysis of Treatment Response and Normative Analysis of Treatment Chioce", *Journal of Econometrics*, 95(2), 415-442.

[59] Manski, C. F. (2000b). "Using Studies of Treatment Response To Inform Treatment Choice in Heterogeneous Populations", *NBER Technical Working Paper*, No. 263.

[60] Manski, C. F. (2001). "Designing Programs for Heterogeneous Populations: The Value of Covariate Information", *American Economic Review*, Papers and Proceedings, 91(2), 103-106.

[61] Millimet, D. L.; List, J. A. and Stengos, T. (2003). "The Environmental Kuztens Curve: Real Progress or Misspecified Models", *Review of Economics and Statistics*, 85(4), 1038-1047.

[62] Misra, D. P. and Ananth, C.V. (2002). "Infant Mortality Among Singletons and Twins in the United States During 2 Decades: Effects of Maternal Age", *Pediatrics, 110(6), 1163-1168*

[63] Müeller, H.-G. (1985). "Kernel Estimators of Zeros and Location and Size of Extrema of Regression Functions", *Scandinavian Journal of Statistics*, 12, 221-232.

[64] Müeller, H.-G. (1989). "Adaptive Nonparametric Peak Estimation", *Annals of Statistics*, 17(3), 1053-1069.

[65] Nadaraya, É. A. (1964). "On Estimating Regression", *Theory of Probability and Its Applications*, 9, 141-142.

[66] Newey, W. K. (1994). "Kernel Estimation of Partial Means and a General Variance Estimator", *Econometric Theory*, 10, 233-253.

[67] Newey, W. K. and Hausman, J. A. (1995). "Nonparametric Estimation of Exact Consumers Surplus and Deadweight Loss", *Econometrica*, 63(6), 1145-1476.

[68] Newey, W. K. and McFadden (1994). "Large Sample Estimation and Hypothesis Testing", in R. F. Engle and D. L. McFadden, eds., *Handbook of Econometrics*, Vol. 4, North-Holland, 2111-2245.

[69] Newey, W. K.; Powell, J. L. and Vella, F. (1999). "Nonparametric Estimation of Triangular Simultaneous Equations Models", *Econometrica*, 67(3), 565-603.

[70] Newey, W. K. and Ruud, P. A. (2001). "Density Weighted Linear Least Squares", *Working Paper*, Department of Economics, University of California at Berkeley.

[71] Pagan, A. and Ullah A. (1999). *Nonparametric Econometrics*, Cambridge University Press.

[72] Panayotou, T. (1993). "Empirical Test and Policy Analysis of Environmental Degradation at Different Stages of Economic Development", *Working Paper,* WP238, Technology and Employment Programme, International Labour Office, Geneva.

[73] Panayotou, T. (1997). "Demystifying the Environmental Kuznets Curve: Turning a Black Box into a Policy Tool", *Environment and Development Economics*, 2, 465-484.

[74] Parzen, E. (1962). "On Estimation of a Probability Density Function and Mode", *Annals of Mathematical Statistics*, 33(3), 1065-1076.

[75] Rilstone, P. and Ullah, A. (1989). "Nonparametric Estimation of Response Coefficients", *Communications in Statistics, Theory and Methods*, 18(7), 2615-2627.

[76] Romano, J. P. (1988). "On Weak Convergence and Optimality of Kernel Density Estimates of the Mode", *Annals of Statistics*, 16(2), 629-647.

[77] Rosenbaum, P. R. and Rubin, D. B. (1983). "The Central Role of the Propensity Score in Observational Studies for Causal Effects", *Biometrika*, 70(1), 41-55.

[78] Royer, H. (2003). "The Question Every Woman (and Some Men) Ponder: Does Maternal Age Affect Infant Health?", Unpublished manuscript, Department of Economics, University of California at Berkeley.

[79] Rubin, D. B. (1974). "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies", *Journal of Educational Psychology*, 66(5), 688-701.

[80] Rubin, D. B. (1978). "Bayesian Inference for for Causal Effects: The Role of Randomization", *Annals of Statistics*, 6(1), 34-58.

[81] Rubin, D. B. (1986). "Which Ifs Have Causal Answers?" Discussion of Paul Holland's "Statistics and Causal Inference", *Journal of the American Statistical Association*, 81(396), 961-962.

[82] Rubin, D. B. (1991). "Practical Implications of Modes of Statistical Inference for Causal Effects and the Critical Role of the Assignment Mechanism", *Biometrics*, 47(4), 1213-1234.

[83] Samanta, M. (1989). "Non-parametric Estimation of Conditional Quantiles", *Statistics and Probability Letters,* 7, 407-412.

[84] Schmalensee, R.; Stoker, T. M. and Judson, R. A. (1998). "World Carbon Dioxide Emissions: 1950-2050", *Review of Economics and Statistics,* 80(1), 15-27.

[85] Schuster, E. (1969). "Estimation of a Probability Density Function and Its Derivatives", *The Annals of Mathematical Statistics, 40(4), 1187–1195.*

[86] Schuster, E. and Yakowitz, S. (1979). "Contributions to the Theory of Nonparametric Regression, with Application to System Identification", *The Annals of Statistics, 7(1), 139-149.*

[87] Selden T. M. and Song, D. (1994). "Environmental Quality and Development: Is there a Kuznets Curve for Air Pollution Emissions?", *Journal of Environmental Economics and Management,* 27, 147-162.

[88] Shafik, N. (1994). "Economic Development and Environmental Quality: An Econometric Analysis", *Oxford Economic Papers*, (46), 757-773.

[89] Stern, D. I. (1998). "Progress on the Environmental Kuznets Curve?", *Environment and Development Economics,* 3, 173-196.

[90] Stern, D. I. (2004). "The Rise and Fall of the Environmental Kuznets Curve", *World Development,* 32(8), 1419-1439.

[91] Stern, D. I. and Common, M. S. (2001). "Is There an Environmental Kuznets Curve for Sulfur?", *Journal of Environmental Economics and Management,* 41, 162-178.

[92] Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing*, Chapman and Hall, London, New York.

[93] Watson, G. S. (1964). "Smooth Regression Analysis", *Sankhya*, A, 26, 359-372.

[94] Ziegler, K. (2000). "Nonparametric Estimation of Location and Size of Maxima of Regression Functions in the Random Design Case based on the Nadaraya-Watson Estimator with Data-Dependent Bandwidths", PhD Thesis, University of Munich, Germany.

Appendix A

Proofs: Experimental Design

Let $g_0(t) = E[Y|T = t]$; then, the parameters of interest in Theorem 1 are: $\alpha_0 = \arg\max_{t \in \mathcal{T}} g_0(t)$ and $g_0(\alpha_0)$. These parameters are estimated respectively as

$$\widehat{\alpha} = \arg\max_{t \in \mathcal{T}} \ \widehat{g}_{h_1}(t)$$

$$\widehat{g}_0(\alpha_0) = \widehat{g}_{h_2}(\widehat{\alpha})$$

where $g_l(t)$ is the Nadayara-Watson estimator

$$\widehat{g}_{h_l}(t) = \frac{\sum_{i=1}^{n} Y_i K(\frac{t-t_i}{h_l})}{\sum_{i=1}^{n} K(\frac{t-t_i}{h_l})}$$

based on the bandwidth h_l and where $K(\cdot)$ is a Kernel satisfying the conditions on Theorem 1.

The result in Theorem 1 states that

$$\begin{pmatrix} \sqrt{nh_1^3}(\widehat{\alpha} - \alpha_0) \\ \sqrt{nh_2}(\widehat{g}_2(\widehat{\alpha}) - g_0(\alpha_0)) \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} \right) \tag{A.1}$$

with $\sigma_1^2 = \frac{\sigma_0^2(\alpha_0)}{[g_0^{(2)}(\alpha_0)]^2\ f_0(\alpha_0)} \int [K^{(1)}(z)]^2 dz$ and $\sigma_2^2 = \frac{\sigma_0^2(\alpha_0)}{f_0(\alpha_0)} \int [K(z)]^2 dz$; and where the superscript (j) denotes the jth derivative with respect to t.

The general approach of the proof is similar in spirit to those in Müeller (1985) and Ziegler (2000), but is closer to the latter since Ziegler (2000) also considers the random design case and the NW estimator. However, as mentioned in the text, rather than using kernels of different order for estimation of location and size of the maximum, we use bandwidths of different orders for their estimation.

First, we briefly sketch the steps involved in proving Theorem 1. To simplify notation let $\widehat{g}_{h_1}(\cdot) = \widehat{g}_1(\cdot)$ and $\widehat{g}_{h_2}(\cdot) = \widehat{g}_2(\cdot)$. According to the Cramér-Wold device we need to show that for every real numbers λ_1 and λ_2 we have:

$$\lambda_1\sqrt{nh_1^3}(\widehat{\alpha} - \alpha_0) + \lambda_2\sqrt{nh_2}(\widehat{g}_2(\widehat{\alpha}) - g_0(\alpha_0)) \xrightarrow{d}$$

$$\mathcal{N}\left(0, \frac{\lambda_1^2\sigma_0^2(\alpha_0)}{[g_0^{(2)}(\alpha_0)]^2\ f_0(\alpha_0)} \int [K^{(1)}(z)]^2 dz + \frac{\lambda_2^2\sigma_0^2(\alpha_0)}{f_0(\alpha_0)} \int [K(z)]^2 dz\right) \tag{A.2}$$

As usual in the nonparametric literature (e.g., Bierens, 1987; Pagan and Ullah, 1999), and similar to Ziegler (2000), the first step involves showing that we can write

$$\begin{aligned} & \lambda_1\sqrt{nh_1^3}(\widehat{\alpha} - \alpha_0) + \lambda_2\sqrt{nh_2}(\widehat{g}_2(\widehat{\alpha}) - g_0(\alpha_0)) \\ = & -\frac{1}{f_0(\alpha)}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[Y_i - g_0(\alpha_0)]K^*(t_i; h_1, h_2)\right) + o_p(1) \end{aligned} \tag{A.3}$$

with $K^*(t_i; h_1, h_2) = \frac{\lambda_1}{g_0^{(2)}(\alpha_0)\sqrt{h_1}} K^{(1)}\left(\frac{\alpha_0 - t_i}{h_1}\right) - \frac{\lambda_2}{\sqrt{h_2}} K\left(\frac{\alpha_0 - t_i}{h_2}\right)$. Lemmas 6 and 7 will be helpful in showing (A.3). The first one shows that $\sqrt{nh_1^3}(\widehat{\alpha} - \alpha_0) = O_p(1)$, while the second one shows that $\sqrt{nh_2}(\widehat{g}_2(\widehat{\alpha}) - g_0(\alpha_0)) = \sqrt{nh_2}(\widehat{g}_2(\alpha_0) - g_0(\alpha_0)) + o_p(1)$.

The next step is to show that the term inside the parenthesis in (A.3) is asymptotically normal. As is usually done in the literature, we derive the limiting distribution in two stages. Letting $W_{n,i} = [Y_i - g_0(\alpha_0)]K^*(t_i; h_1, h_2)$, we first show in Lemma 9 that $\frac{1}{\sqrt{n}} \sum\limits_{i=1}^{n} \{W_{n,i} - E(W_{n,i})\}$ is asymptotically normal. Then, Lemma 10 shows that $\frac{1}{\sqrt{n}} \sum\limits_{i=1}^{n} E(W_{n,i}) \longrightarrow 0$. In order to highlight the importance of the kernel being symmetric for the asymptotically independece result of the estimators of location and size, we calculate the asymptotic variance of W_i in Lemma 8.

We now present the lemmas to be used in the proof of Theorem 1.

Lemma 6 *(Asymptotic Normality of the Location of the Maximum). Assume:*

(i) $\{(y_1, t_1), \ldots, (y_n, t_n)\}$ *is an i.i.d. sample.*

(ii) $E\left[|Y|^{2+\delta}\right] < \infty$ *for some* $\delta > 0$.

(iii) $\alpha \in \mathcal{T}$, *where* $\mathcal{T}$ *is compact and* α_0 *is in the interior of* $\mathcal{T}$.

(iv) $g_0(t)$ *is uniquely maximized at* α_0 *and is three times continuously differentiable and its derivatives up to the third order are bounded. Also,* $g_0^{(2)}(\alpha_0) \neq 0$.

(v) $f_0(t)$ *is continuous and bounded away from zero uniformly in* $\mathcal{T}$. *Also, partial derivatives of* $f_0(t)$ *at* α_0 *exists up to the third order and are bounded.*

(vi) Partial derivatives of $\sigma_0^2(t)$ *exists up to the third order and are bounded.*

(vii) $h_1 \to 0$, $nh_1^2 \to \infty$, $nh_1^3 \to \infty$, $nh_1^6 \to \infty$, $nh_1^7 \to 0$, *as* $n \to \infty$.

(vii) The kernel $K(\cdot)$ *satisfies:*

(a) $\int K(u)du = 1$, $\int u^2 K(u)du < \infty$, *and* $\int |K(u)|\, du < \infty$.

(b) K *is symmetric and three times continuously differentiable.*

(c) $|u|\,|K(u)| \to 0$ *as* $|u| \to \infty$.

(d) $\sup_u |K(u)| < \infty$.

(e) $\int \left[K^{(1)}(u)\right]^2 du < \infty$.

(f) For some $\delta > 0$ *we have that* $\int |K(u)|^{2+\delta} du < \infty$.

(g) Finally, let $\psi(w)$ *be the characteristic function of* $K(\cdot)$ *so that* $\psi(w) = \int e^{iwu} K(u) du$, *with* $i^2 = -1$. *Then, assume that* $\int |\psi(w)| dw < \infty$ *and* $\int \left|w^2 \psi(w)\right| dw < \infty$.

Then,

$$\sqrt{nh_1^3}(\widehat{\alpha} - \alpha_0) \xrightarrow{d} \mathcal{N}\left(0, \frac{\sigma_0^2(\alpha_0)}{[g_0^{(2)}(\alpha_0)]^2 \ f_0(\alpha_0)} \int [K^{(1)}(z)]^2 dz\right)$$

Proof. Under our assumptions we have that $\widehat{\alpha}$ satisfies $\widehat{g}_1^{(1)}(\widehat{\alpha}) = 0$. Expanding this expression around α_0 and solving for $\widehat{\alpha} - \alpha_0$ gives

$$\widehat{\alpha} - \alpha_0 = -\frac{\widehat{g}_1^{(1)}(\alpha_0)}{\widehat{g}_1^{(2)}(\alpha^*)} \tag{A.4}$$

where α^* is a mean value. Thus, the three key ingredients in showing asymptotic normality of $\sqrt{nh_1^3}(\widehat{\alpha} - \alpha_0)$ are (i) consistency of $\widehat{\alpha}$, (ii) uniform convergence in probability of $\widehat{g}_1^{(2)}(t)$ to $g_0^{(2)}(t)$; and (iii) asymptotic normality of $\sqrt{nh_1^3}\widehat{g}_1^{(1)}(\alpha_0)$.

First, we consider the consistency of $\widehat{\alpha}$. This follows easily by noting that $\widehat{\alpha}$ is an extremum estimator with objective function $\widehat{g}_1(t)$, $t \in \mathcal{T}$. Hence, by Theorem 2.1 in Newey and McFadden (1994), the result follows from continuity of $g_0(t)$, $g_0(t)$ being uniquely maximized at α_0, $\mathcal{T}$ being compact and from uniform convergence in probability of $\widehat{g}_1(t)$ to $g_0(t)$. This latter result is standard in the literature (e.g., Theorem 2.3.1. in Bierens 1987, or Theorem 3.7 in Pagan and Ullah 1999) and follows from our assumptions. Thus, $\widehat{\alpha} \xrightarrow{p} \alpha_0$.

Next, consider uniform convergence in probability of $\widehat{g}_1^{(2)}(t)$ to $g_0^{(2)}(t)$. Results of this type can also be found in the literature (e.g., Schuster and Yakowitz, 1979; Ahmad and Ullah, 1988). Instead of using any of those results, in Lemma 11 we obtain a uniform convergence result for higher order derivatives of the NW estimator based on the approach followed by Bierens (1987). The rate of uniform convergence derived in Lemma 11 will be useful in proving Lemma 7 below. For the moment, note that our assumptions in Lemma 6 satisfy all conditions in Lemma 11 for the estimator of the second derivative of $g_0(\cdot)$. Hence, $\widehat{g}_1^{(2)}(t)$ converges uniformly in probability to $g_0^{(2)}(t)$. This latter result, along with consistency of $\widehat{\alpha}$, continuity of $g_0^{(2)}(t)$ and the fact that α^* is a mean value between $\widehat{\alpha}$ and α_0, imply that $\widehat{g}_1^{(2)}(\alpha^*) \xrightarrow{p} g_0^{(2)}(\alpha_0)$.

Finally, consider asymptotic normality of $\sqrt{nh_1^3}\widehat{g}_1^{(1)}(\alpha_0)$. As in previous cases this result is also standard in the literature. This result is presented, for example, in Theorems 4.1 and 4.2 from Pagan and Ullah (1999). It is easy to verify that our assumptions impliy the result of those theorems, so that $\sqrt{nh_1^3}\widehat{g}_1^{(1)}(\alpha_0) \xrightarrow{p} \mathcal{N}\left(0, [f_0(\alpha_0)]^{-1}\sigma_0^2(\alpha_0)\int \left[K^{(1)}(z)\right]^2 dz\right)$. Therefore, the result in Lemma 6 follows by using Slutzky theorem. ∎

Lemma 7 *Assume*

(i) Same set of assumptions as in Lemma 6.

(ii) $h_2 \to 0$, $nh_2^3 \to \infty$, $nh_2^4 \to \infty$, $nh_2^5 \to 0$, *as* $n \to \infty$.

(iii) K *is symmetric and three times continuously differentiable.*

Then,

$$\sqrt{nh_2}(\widehat{g}_2(\widehat{\alpha}) - g_0(\alpha_0)) = \sqrt{nh_2}(\widehat{g}_2(\alpha_0) - g_0(\alpha_0)) + o_p(1) \qquad \text{(A.5)}$$

Proof. Using the mean value theorem we can write $\widehat{g}_2(\widehat{\alpha}) = \widehat{g}_2(\alpha_0) + \widehat{g}_2^{(1)}(\alpha^*)(\widehat{\alpha} -$

$\alpha_0)$, for some α^* between $\widehat{\alpha}$ and α_0. Subtracting $g_0(\alpha_0)$ from both sides and multiplying by $\sqrt{nh_2}$ we have:

$$\sqrt{nh_2}\left(\widehat{g}_2(\widehat{\alpha}) - g_0(\alpha_0)\right) = \sqrt{nh_2}\left(\widehat{g}_2(\alpha_0) - g_0(\alpha_0)\right) + \sqrt{nh_2}\widehat{g}_2^{(1)}(\alpha^*)(\widehat{\alpha} - \alpha_0) \tag{A.6}$$

Hence, we need to show that the second term to the right of (A.6) is $o_p(1)$. Again, for a suitable mean value α^{**} between α_0 and α^* we can write $\widehat{g}_2^{(1)}(\alpha^*) = \widehat{g}_2^{(1)}(\alpha_0) + \widehat{g}_2^{(2)}(\alpha^{**})(\alpha^* - \alpha_0)$. Thus we have that:

$$\begin{aligned}\sqrt{nh_2}\widehat{g}_2^{(1)}(\alpha^*)(\widehat{\alpha} - \alpha_0) &= \frac{1}{\sqrt{nh_1^3h_2^2}}\sqrt{nh_2^3}\widehat{g}_2^{(1)}(\alpha_0)\sqrt{nh_1^3}(\widehat{\alpha} - \alpha_0) \\ &+\frac{1}{\sqrt{nh_1^6}}\sqrt{h_2}\widehat{g}_2^{(2)}(\alpha^{**})\sqrt{nh_1^3}(\alpha^* - \alpha_0)\sqrt{nh_1^3}(\widehat{\alpha} - \alpha_0)\end{aligned} \tag{A.7}$$

Consider the first term to the right side of (A.7). Note that from Lemma 6 we know that $\sqrt{nh_1^3}(\widehat{\alpha} - \alpha_0) = O_p(1)$. Also, as in Lemma 6, we can use Theorems 4.1 and 4.2 in Pagan and Ullah (1999) to show that $\sqrt{nh_2^3}\widehat{g}_2^{(1)}(\alpha) = O_p(1)$. The conditions of those theorems require h_2 to be such that $nh_2^3 \to \infty$ and $nh_2^7 \to 0$, which clearly do not contradict our assumptions on h_2. Finally, observe that

$$\frac{1}{\sqrt{nh_1^3h_2^2}} = \frac{1}{\sqrt{\sqrt{nh_1^6}\sqrt{nh_2^4}}} \longrightarrow 0$$

since by assumption $nh_1^6 \longrightarrow \infty$ and $nh_2^4 \longrightarrow \infty$. Thus, the first term to the right of (A.7) is $o_p(1)$.

Now consider the second term to the right of (A.7). Again, from Lemma 6 we know that $\sqrt{nh_1^3}(\widehat{\alpha} - \alpha_0) = O_p(1)$. Also, given that $|\alpha^* - \alpha_0| \leq |\widehat{\alpha} - \alpha_0|$ we have $\sqrt{nh_1^3}(\alpha^* - \widehat{\alpha}) = O_p(1)$. Therefore, in order for the expression in (A.7) to be $o_p(1)$, we require $\left(nh_1^6\right)^{-1/2}\sqrt{h_2}\widehat{g}_2^{(2)}(\alpha^{**}) \xrightarrow{p} 0$. Write

$$\frac{\sqrt{h_2}}{\sqrt{nh_1^6}}\left|\widehat{g}_2^{(2)}(\alpha^{**})\right| \leq \frac{\sqrt{h_2}}{\sqrt{nh_1^6}}\sup_t\left|\widehat{g}_2^{(2)}(t)\right| \leq \frac{\sqrt{h_2}}{\sqrt{nh_1^6}}\sup_t\left|\widehat{g}_2^{(2)}(t) - g_0^{(2)}(t)\right| + \frac{\sqrt{h_2}}{\sqrt{nh_1^6}}\sup_t\left|g_0^{(2)}(t)\right|$$

Given our assumptions, we know that $\left(nh_1^6\right)^{-1/2}\sqrt{h_2}\sup_t\left|g_0^{(2)}(t)\right| = o(1)$. From Lemma 11 we know that $\sup_t\left|\widehat{g}_2^{(2)}(t) - g_0^{(2)}(t)\right| = O_p\left(\left(nh_2^6\right)^{-1/2}\right)$. Thus, we obtain that $\left(nh_1^6\right)^{-1/2}$ $\left|\sqrt{h_2}\widehat{g}_2^{(2)}(\alpha^{**})\right| \leq O_p\left(\left(nh_1^6\right)^{-1/2}\ \left(nh_2^5\right)^{-1/2}\right)$. Note that if we use a badwidth with the optimal order $n^{-1/5}$, then $nh_2^5 \rightarrow d^2 > 0$ and $\left(nh_1^6\right)^{-1/2}\left(nh_2^5\right)^{-1/2} \rightarrow 0$. Hence, in this case $\left(nh_1^6\right)^{-1/2}\sqrt{h_2}\widehat{g}_2^{(2)}(t)$ converges uniformly to zero, which combined $\alpha^{**} \xrightarrow{p} \alpha_0$ and continuity of $g_0^{(2)}(t)$ imply that $\left(nh_1^6\right)^{-1/2}\sqrt{h_2}\widehat{g}_2^{(2)}(\alpha^{**}) \xrightarrow{p} 0$. However, in Theorem 1 we decided to center the distribution around zero by choice of bandwidth. Let h_1 and h_2 be proportional, respectively, to $n^{-(1+\varepsilon)/7}$ and $n^{-(1+\eta)/5}$, where ε and η are small positive numbers used for undersmoothing. Let "$\backsim$" denote proportionality. Then, $\left(nh_1^6\right)^{-1/2}\left(nh_2^5\right)^{-1/2} \backsim \left(n^{-(1-6\varepsilon-7\eta)}\right)$. Thus, if ε and η are chosen such that $1 > 6\varepsilon + 7\eta$, then $\left(nh_1^6\right)^{-1/2}\left(nh_2^5\right)^{-1/2} \rightarrow 0$ and consequently $\left(nh_1^6\right)^{-1/2}\sqrt{h_2}\widehat{g}_2^{(2)}(\alpha^{**}) \xrightarrow{p} 0$. Hence, the second term to the right of (A.7) is $o_p(1)$[1].

Thus, both terms to the right hand of (A.7) are $o_p(1)$, and the conclusion of Lemma 7 follows from (A.6). ■

Now, in order to highlight the importance of the kernel being symmetric for the asymptotically independece result of the estimators of location and size, we calculate the asymptotic variance of $W_{n,i} = [Y_i - g_0(\alpha_0)]K^*(t_i;h_1,h_2)$, with $K^*(t_i;h_1,h_2)$ as defined in (A.3), in the following Lemma.

Lemma 8 *Let* $W_{n,i} = [Y_i - g_0(\alpha_0)]K^*(t_i;h_1,h_2)$, *with* $K^*(t_i;h_1,h_2) = \frac{\lambda_1}{g_0^{(2)}(\alpha_0)\sqrt{h_1}}K^{(1)}\left(\frac{\alpha_0-t_i}{h_1}\right) - \frac{\lambda_2}{\sqrt{h_2}}\ K\left(\frac{\alpha_0-t_i}{h_2}\right)$, *and assume that:*

[1]This part of the proof works nicely if we use kernels with bounded support. In this case we would have (e.g., Newey (1994)) $\sup_t\left|\widehat{g}_2^{(2)}(t) - g_0^{(2)}(t)\right| = O_p\left(\left[\ln(n)/nh_2^5\right]^{-1/2}\right)$. Then, it would follow that $\left|\sqrt{h_2}\widehat{g}_2^{(2)}(\alpha^{**})\right| \leq O_p\left(\left[\ln(n)/nh_2^4\right]^{-1/2}\right)$; so that $\sqrt{h_2}\widehat{g}_2^{(2)}(\alpha^{**}) \rightarrow 0$ given the assumption that $nh_2^4/\ln(n) \rightarrow \infty$.

(i) $E\left[Y^2\right] < \infty$.

(ii) $g_0(t)$ *is continuous at* α_0*, its derivative exits and is continuous at* α_0*. Also,* $g_0(\alpha_0)$ *is bounded and* $g_0^{(2)}(\alpha_0) \neq 0$.

(iii) $f_0(t)$ *is continuous and bounded away from zero uniformly in* $\mathcal{T}$*. Also, the derivative of* $f_0(t)$ *at* α_0 *exists and is continuous; and* $f_0(\alpha_0)$ *and* $f_0^{(1)}(\alpha_0)$ *are both bounded.*

(iv) The derivative of $E\left[Y^2|T=t\right]$ *exists and is continuous at* $T=\alpha_0$*. Also,* $\sigma_0^2(t)$ *is continuous at* α_0 *and* $\sigma_0^2(\alpha_0)$ *and its first derivative evaluated at* α_0 *are both bounded.*

(v) $h_1 \to 0$*,* $h_2 \to 0$*,* $nh_1^2 \to \infty$*,* $nh_2^4 \to \infty$*,* $nh_2^5 \to a^2$ *for some constant* $a \geq 0$*, as* $n \to \infty$.

(vi) Let $K(\cdot)$ *be symmetric with* $\int |uK(u)|\,du < \infty$*; and for* $H(u) = K(u)$ *and* $H(u) = K^{(1)}(u)$ *we have*

(a) $\int |H(u)|^2\,du < \infty$

(b) $|u|\,|H(u)|^2 \to 0$ *as* $|u| \to \infty$

(c) $\sup_u |H(u)|^2 < \infty$

Then,

$$Var\,(W_i) \longrightarrow \frac{\lambda_1^2\sigma_0^2(\alpha_0)f_0(\alpha_0)}{[g_0^{(2)}(\alpha_0)]^2}\int [K^{(1)}(z)]^2dz + \lambda_2^2\sigma_0^2(\alpha_0)f_0(\alpha_0)\int [K(z)]^2dz \qquad \text{(A.8)}$$

Proof. Using the definition of W_i first write:

$$\begin{aligned} Var\,(W_i) &= E\left(\{(Y_i - g_0(\alpha_0))\,K^*(t_i;h_1,h_2)\}^2\right) \\ &\quad - \left(E\left[\{Y_i - g_0(\alpha_0)\}\,K^*(t_i;h_1,h_2)\right]\right)^2 \end{aligned} \qquad \text{(A.9)}$$

First we focus on the first term to the right hand of (A.9). Let $\phi(t_i) = E\left[(Y_i - g_0(\alpha_0))^2\,|T = t_i\right]$, then using iterated expectations and the definition of $K^*(t_i;h_1,h_2)$ we obtain

$$
\begin{aligned}
& E\left(\left\{\left(Y_i - g_0(\alpha_0)\right) K^*(t_i; h_1, h_2)\right\}^2\right) \\
= \; & E\left(K^{2*}(t_i; h_1, h_2)\phi(t_i)\right) \\
= \; & \frac{\lambda_1^2}{\left[g_0^{(2)}(\alpha_0)\right]^2 h_1} E\left(\phi(t_i) K^{2(1)}\left(\frac{\alpha_0 - t_i}{h_1}\right)\right) + \frac{\lambda_2^2}{h_2} E\left(\phi(t_i) K^2\left(\frac{\alpha_0 - t_i}{h_2}\right)\right) \\
& -2 \frac{\lambda_1 \lambda_2}{g_0^{(2)}(\alpha_0)\sqrt{h_1 h_2}} E\left(\phi(t_i) K^{(1)}\left(\frac{\alpha_0 - t_i}{h_1}\right) K\left(\frac{\alpha_0 - t_i}{h_2}\right)\right) \qquad \text{(A.10)}
\end{aligned}
$$

By bounded convergence[2], and noting that by definition $\phi(\alpha_0) = \sigma_0^2(\alpha_0) = Var[Y|T = \alpha_0]$, we obtain:

$$
\frac{1}{h_1} E\left(\phi(t_i) K^{2(1)}\left(\frac{\alpha_0 - t_i}{h_1}\right)\right) \longrightarrow \sigma_0^2(\alpha_0) f_0(\alpha_0) \int [K^{(1)}(z)]^2 dz \qquad \text{(A.11)}
$$

and

$$
\frac{1}{h_2} E\left(\phi(t_i) K^2\left(\frac{\alpha_0 - t_i}{h_2}\right)\right) \longrightarrow \sigma_0^2(\alpha_0) f_0(\alpha_0) \int [K(z)]^2 dz \qquad \text{(A.12)}
$$

Now we need an approximation for $\frac{1}{\sqrt{h_1 h_2}} E\left(\phi(t_i) K^{(1)}\left(\frac{\alpha_0 - t_i}{h_1}\right) K\left(\frac{\alpha_0 - t_i}{h_2}\right)\right)$ as $n \to \infty$. Since now the term we want to approximate depends on two different bandwidths, h_1 and h_2, we cannot directly use the bounded convergence result as usually stated in the literature (e.g., Parzen, 1962 Theorem 1A). Let $s(t) = \phi(t) f_0(t)$, $a = (h_2 h_1^{-1})^{1/2}$, and $b = (h_2^{-1} h_1)^{1/2}$, then we can write:

$$
\begin{aligned}
& \frac{1}{\sqrt{h_1 h_2}} E\left(\phi(t_i) K^{(1)}\left(\frac{\alpha_0 - t_i}{h_1}\right) K\left(\frac{\alpha_0 - t_i}{h_2}\right)\right) \\
= \; & \frac{1}{\sqrt{h_1 h_2}} \int \phi(t_i) K^{(1)}\left(\frac{\alpha_0 - t_i}{h_1}\right) K\left(\frac{\alpha_0 - t_i}{h_2}\right) f_0(t_i) dt_i \\
= \; & \int K^{(1)}\left(\frac{\sqrt{h_2}}{\sqrt{h_1}} z\right) K\left(\frac{\sqrt{h_1}}{\sqrt{h_2}} z\right) s(\alpha_0 - \sqrt{h_1 h_2} z) dz \\
= \; & s(\alpha_0) \int K^{(1)}(az) K(bz) \, dz - s^{(1)}(\alpha^*) \sqrt{h_1 h_2} \int z K^{(1)}(az) K(bz) \, dz \qquad \text{(A.13)}
\end{aligned}
$$

[2]See, for instance, Theorem 1A in Parzen (1962).

where we made the change of variable $z = \frac{\alpha_0 - t_i}{\sqrt{h_1 h_2}}$ in the third line and used the mean value theorem in the last one, for some α^* that lies between $\alpha_0 - \sqrt{h_1 h_2} z$ and α_0. By symmetry of the kernel we know that $K(z) = K(-z)$ and consequently $K^{(1)}(z) = -K^{(1)}(-z)$. Hence we have that:

$$\begin{aligned}
\int K^{(1)}(az)\, K(bz)\, dz &= \int_{-\infty}^{0} K^{(1)}(az)\, K(bz)\, dz + \int_{0}^{\infty} K^{(1)}(az)\, K(bz)\, dz \\
&= \int_{-\infty}^{0} K^{(1)}(az)\, K(bz)\, dz - \int_{0}^{\infty} K^{(1)}(-az)\, K(-bz)\, dz \\
&= \int_{-\infty}^{0} K^{(1)}(az)\, K(bz)\, dz - \int_{-\infty}^{0} K^{(1)}(az)\, K(bz)\, dz = 0
\end{aligned}$$

Thus, the first term in (A.13) is zero because of the symmetry of the kernel. As for the second term, observe that $\alpha^* \to \alpha_0$ as $n \to \infty$, so that by continuity of $s^{(1)}(\cdot)$ we have that $s^{(1)}(\alpha^*) \longrightarrow s^{(1)}(\alpha_0)$, which is less than infinity by assumption. Also, letting C be the constant bounding $K^{(1)}(\cdot)$, note that

$$\begin{aligned}
\left| \sqrt{h_1 h_2} \int z K^{(1)}(az)\, K(bz)\, dz \right| &\leq \sqrt{h_1 h_2} \int \left| z K^{(1)}(az)\, K(bz) \right| dz \\
&\leq C\sqrt{h_1 h_2} \int \left| z K(bz) \right| dz \\
&= C\sqrt{h_1 h_2} \frac{1}{b^2} \int \left| w K(w) \right| dw
\end{aligned}$$

Remember that by definition $b = (h_2^{-1} h_1)^{1/2}$, so that

$$\sqrt{h_1 h_2} \frac{1}{b^2} = \frac{\sqrt{h_2^3}}{\sqrt{h_1}} = \frac{\sqrt{n h_2^5}}{\sqrt{\sqrt{n h_1^2} \sqrt{n h_2^4}}} \longrightarrow 0$$

given that $nh_2^5 \to a^2$, $nh_1^2 \to \infty$ and $nh_2^4 \to \infty$. This, along with the assumption that $\int |wK(w)|\, dw < \infty$, implies that $\left|\sqrt{h_1 h_2} \int z K^{(1)}(az)\, K(bz)\, dz\right| \longrightarrow 0$. Hence, the second term in (A.13) goes to zero as $n \to \infty$, so we conclude that:

$$\frac{1}{\sqrt{h_1 h_2}} E\left(\phi(t_i) K^{(1)}\left(\frac{\alpha_0 - t_i}{h_1}\right) K\left(\frac{\alpha_0 - t_i}{h_2}\right)\right) \longrightarrow 0 \tag{A.14}$$

Plugging (A.11), (A.12) and (A.14) into (A.10) we obtain:

$$\begin{aligned} & E\left(\{(Y_i - g_0(\alpha_0)) K^*(t_i; h_1, h_2)\}^2\right) \\ \longrightarrow \quad & \frac{\lambda_1^2 \sigma_0^2(\alpha_0) f_0(\alpha_0)}{[g_0^{(2)}(\alpha_0)]^2} \int [K^{(1)}(z)]^2 dz + \lambda_2^2 \sigma_0^2(\alpha_0) f_0(\alpha_0) \int [K(z)]^2 dz \end{aligned} \tag{A.15}$$

To complete our proof we show that the second term in (A.9) goes to zero as $n \to \infty$. Let $r(t) = g_0(t) - g_0(\alpha_0)$, then we have:

$$\begin{aligned} & E\left[\{Y_i - g_0(\alpha)\} K^*(t_i; h_1, h_2)\right] \\ = \quad & E\left[K^*(t_i; h_1, h_2)\{E(Y_i|T = t_i) - g_0(\alpha_0)\}\right] = E\left[K^*(t_i; h_1, h_2) r(t_i)\right] \\ = \quad & \int r(t_i) \left[\frac{\lambda_1}{g_0^{(2)}(\alpha_0)\sqrt{h_1}} K^{(1)}\left(\frac{\alpha_0 - t_i}{h_1}\right) - \frac{\lambda_2}{\sqrt{h_2}} K\left(\frac{\alpha_0 - t_i}{h_2}\right)\right] f_0(t_i) dt_i \\ = \quad & \frac{\lambda_1 \sqrt{h_1}}{g_0^{(2)}(\alpha_0)} \frac{1}{h_1} \int r(t_i) K^{(1)}\left(\frac{\alpha_0 - t_i}{h_1}\right) f_0(t_i) dt_i \\ & -\lambda_2 \sqrt{h_2} \frac{1}{h_2} \int r(t_i) K\left(\frac{\alpha_0 - t_i}{h_2}\right) f_0(t_i) dt_i \end{aligned} \tag{A.16}$$

By bounded convergence and noting that $r(\alpha_0) = 0$ we obtain:

$$\frac{1}{h_1} \int r(t_i) K^{(1)}\left(\frac{\alpha_0 - t_i}{h_1}\right) f_0(t_i) dt_i \longrightarrow r(\alpha_0) f_0(\alpha_0) \int K^{(1)}(z) dz = 0$$

$$\frac{1}{h_2} \int r(t_i) K\left(\frac{\alpha_0 - t_i}{h_2}\right) f_0(t_i) dt_i \longrightarrow r(\alpha_0) f_0(\alpha_0) = 0$$

Hence we obtain from (A.16) that[3]

$$E\left[\{Y_i - g_0(\alpha_0)\} K^*(t_i; h_1, h_2)\right] \longrightarrow 0 \tag{A.17}$$

[3]Note also that $\sqrt{h_1} \to 0$ and $\sqrt{h_2} \to 0$ in the last equality in (A.16).

By plugging (A.15) and (A.17) into (A.9) the result follows. ■

We now present the lemmas that will be useful in showing asymptotic normality of the term inside the parenthesis in (A.3). As previously mentioned, we first show asymptotic normality of this term centered at its expectation, and then show that under our conditions the normalized expectation goes to zero asymptotically.

Lemma 9 *Let* $W_{n,i} = [Y_i - g_0(\alpha_0)]K^*(t_i; h_1, h_2)$, *with* $K^*(t_i; h_1, h_2) = \frac{\lambda_1}{g_0^{(2)}(\alpha_0)\sqrt{h_1}} K^{(1)} \left(\frac{\alpha_0 - t_i}{h_1}\right) - \frac{\lambda_2}{\sqrt{h_2}}\ K\left(\frac{\alpha_0 - t_i}{h_2}\right)$. *Assume*

(i) Same set of assumptions as in Lemma 8.

(ii) $\{(y_1, t_1), \ldots, (y_n, t_n)\}$ *is an i.i.d. sample.*

(iii) $E\left[|Y|^{2+\delta}\right] < \infty$ *for some* $\delta > 0$.

(iv) $g_0^{(2)}(\alpha_0)$ *is bounded.*

(v) For some $\delta > 0$, *and for* $H(u) = K(u)$ *and* $H(u) = K^{(1)}(u)$ *we have that*

(a) $\int |H(u)|^{2+\delta} du < \infty$

(b) $\sup_u |H(u)|^{2+\delta} < \infty$

(c) $|u|\,|H(u)|^{2+\delta} \to 0$ *as* $|u| \to \infty$.

Then,

$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{W_{n,i} - E(W_{n,i})\} \xrightarrow{d}$
$\mathcal{N}\left(0, \frac{\lambda_1^2 \sigma_0^2(\alpha_0) f_0(\alpha_0)}{[g_0^{(2)}(\alpha_0)]^2} \int [K^{(1)}(z)]^2 dz + \lambda_2^2 \sigma_0^2(\alpha_0) f_0(\alpha_0) \int [K(z)]^2 dz\right)$.

Proof. Let

$$\sum_{i=1}^{n} L_{n,i} = \sum_{i=1}^{n} \frac{W_{n,i} - E(W_{n,i})}{[nVar\,(W_{n,i})]^{1/2}}$$

Note that $L_{n,i}$ is a traingular array of i.i.d. random variables such that $E\,(L_{n,i}) = 0$, $Var\,(L_{n,i}) = 1/n < \infty$ and $Var\left(\sum_{i=1}^{n} L_{n,i}\right) = 1$. Hence, we have by Liapunov's Central

Limit Theorem that $\sum_{i=1}^{n} L_{n,i} \xrightarrow{d} \mathcal{N}(0,1)$ if for some $\delta > 0$:

$$\lim_{n\to\infty} \sum_{i=1}^{n} E\left(|L_{n,i}|^{2+\delta}\right) = 0 \tag{A.18}$$

In our case we have:

$$\begin{aligned}
& \sum_{i=1}^{n} E\left(|L_{n,i}|^{2+\delta}\right) \\
= \;& nE\left(\left|\frac{W_{n,i}-E(W_{n,i})}{[nVar\,(W_{n,i})]^{1/2}}\right|^{2+\delta}\right) \\
= \;& n^{-\frac{\delta}{2}}\left[Var\,(W_i)\right]^{-\left(1+\frac{\delta}{2}\right)} E\left|W_i - E\,(W_i)\right|^{2+\delta} \\
\leq \;& 2^{2+\delta} n^{-\frac{\delta}{2}}\left[Var\,(W_i)\right]^{-\left(1+\frac{\delta}{2}\right)} E\left|W_i\right|^{2+\delta} \\
= \;& 2^{2+\delta} n^{-\frac{\delta}{2}}\left[Var\,(W_i)\right]^{-\left(1+\frac{\delta}{2}\right)} E\Bigg|[Y_i - g_0(\alpha_0)]\frac{\lambda_1}{g_0^{(2)}(\alpha_0)\sqrt{h_1}}K^{(1)}\left(\frac{\alpha_0-t_i}{h_1}\right) \\
& -[Y_i - g_0(\alpha_0)]\frac{\lambda_2}{\sqrt{h_2}}K\left(\frac{\alpha_0-t_i}{h_2}\right)\Bigg|^{2+\delta} \\
\leq \;& 2^{3+2\delta} n^{-\frac{\delta}{2}}\left[Var\,(W_i)\right]^{-\left(1+\frac{\delta}{2}\right)} \Bigg\{E\left|[Y_i - g_0(\alpha_0)]K^{(1)}\left(\frac{\alpha_0-t_i}{h_1}\right)\right|^{2+\delta} \\
& \frac{\lambda_1^{2+\delta} h_1^{-\left(1+\frac{\delta}{2}\right)}}{\left[g_0^{(2)}(\alpha_0)\right]^{2+\delta}} + E\left|[Y_i - g_0(\alpha_0)]K\left(\frac{\alpha_0-t_i}{h_2}\right)\right|^{2+\delta} \lambda_2^{2+\delta} h_2^{-\left(1+\frac{\delta}{2}\right)}\Bigg\}
\end{aligned} \tag{A.19}$$

where in the fourth and sixth lines we used the C_r-inequality, while in the fifth we used the definitions of $W_{n,i}$ and $K^*(t_i; h_1, h_2)$. Given that $E\,|Y_i|^{2+\delta} < \infty$, we have that $E[|Y_i|^{2+\delta} \mid T = t_i] < \infty$. Let C_1 be the constant bounding the latter, and let C_2 be the constant bounding $g_0(\cdot)$. Then,

$$\begin{aligned}
& E\left|[Y_i - g_0(\alpha_0)]K^{(1)}\left(\frac{\alpha_0-t_i}{h_1}\right)\right|^{2+\delta} \\
= \;& E\left|Y_i K^{(1)}\left(\frac{\alpha_0-t_i}{h_1}\right) - g_0(\alpha_0)K^{(1)}\left(\frac{\alpha_0-t_i}{h_1}\right)\right|^{2+\delta} \\
\leq \;& 2^{1+\delta}\left[E\left|Y_i K^{(1)}\left(\frac{\alpha_0-t_i}{h_1}\right)\right|^{2+\delta} + E\left|g_0(\alpha_0)K^{(1)}\left(\frac{\alpha_0-t_i}{h_1}\right)\right|^{2+\delta}\right]
\end{aligned}$$

$$
\begin{aligned}
&< E\left(\left|K^{(1)}\left(\frac{\alpha_0 - t_i}{h_1}\right)\right|^{2+\delta} E\left[|Y_i|^{2+\delta}|T = t_i\right]\right) \\
&\quad +|g_0(\alpha_0)|^{2+\delta} E\left|K^{(1)}\left(\frac{\alpha_0 - t_i}{h_1}\right)\right|^{2+\delta} \\
&\leq C_1 E\left|K^{(1)}\left(\frac{\alpha_0 - t_i}{h_1}\right)\right|^{2+\delta} + C_2^{2+\delta} E\left|K^{(1)}\left(\frac{\alpha_0 - t_i}{h_1}\right)\right|^{2+\delta} \\
&= C_3 E\left|K^{(1)}\left(\frac{\alpha_0 - t_i}{h_1}\right)\right|^{2+\delta} \qquad (A.20)
\end{aligned}
$$

where we again used the C_r-inequality in the third line, and we let C_3 be another constant. Following the same steps as before we get:

$$
E\left|[Y_i - g_0(\alpha_0)] K\left(\frac{\alpha_0 - t_i}{h_2}\right)\right|^{2+\delta} < C_4 E\left|K\left(\frac{\alpha_0 - t_i}{h_2}\right)\right|^{2+\delta} \qquad (A.21)
$$

for a constant C_4. Plugging (A.20) and (A.21) into (A.19) we obtain

$$
\begin{aligned}
&\sum_{i=1}^{n} E\left(|L_{n,i}|^{2+\delta}\right) \\
&< (nh_1)^{-\frac{\delta}{2}} \left[Var\left(W_i\right)\right]^{-\left(1+\frac{\delta}{2}\right)} \frac{\lambda_1^{2+\delta} C_3}{\left[g_0^{(2)}\left(\alpha_0\right)\right]^{2+\delta}} h_1^{-1} E\left|K^{(1)}\left(\frac{\alpha_0 - t_i}{h_1}\right)\right|^{2+\delta} \\
&\quad + (nh_2)^{-\frac{\delta}{2}} \left[Var\left(W_i\right)\right]^{-\left(1+\frac{\delta}{2}\right)} \lambda_2^{2+\delta} C_4 h_2^{-1} E\left|K\left(\frac{\alpha_0 - t_i}{h_2}\right)\right|^{2+\delta} \qquad (A.22)
\end{aligned}
$$

Now, by bounded convergence we have that

$$
\begin{aligned}
h_1^{-1} E\left|K^{(1)}\left(\frac{\alpha_0 - t_i}{h_1}\right)\right|^{2+\delta} &= h_1^{-1} \int \left|K^{(1)}\left(\frac{\alpha_0 - t_i}{h_1}\right)\right|^{2+\delta} f_0(t_1) dt_1 \\
&\longrightarrow f_0(\alpha_0) \int \left|K^{(1)}\left(z\right)\right|^{2+\delta} dz
\end{aligned}
$$

and

$$
\begin{aligned}
h_1^{-1} E\left|K\left(\frac{\alpha_0 - t_i}{h_2}\right)\right|^{2+\delta} &= h_1^{-1} \int \left|K\left(\frac{\alpha_0 - t_i}{h_2}\right)\right|^{2+\delta} f_0(t_1) dt_1 \\
&\longrightarrow f_0(\alpha_0) \int \left|K\left(z\right)\right|^{2+\delta} dz
\end{aligned}
$$

where by assumption $\int |K(z)|^{2+\delta} dz < \infty$ and $\int \left|K^{(1)}(z)\right|^{2+\delta} dz < \infty$ for some $\delta > 0$.

Also, by Lemma 8 we have that

$$Var(W_i) \longrightarrow \frac{\lambda_1^2 \sigma_0^2(\alpha_0) f_0(\alpha_0)}{[g_0^{(2)}(\alpha_0)]^2} \int [K^{(1)}(z)]^2 dz + \lambda_2^2 \sigma_0^2(\alpha_0) f_0(\alpha_0) \int [K(z)]^2 dz \qquad \text{(A.23)}$$

Hence, given that $g_0^{(2)}(\alpha_0) < \infty$ we have that all quantities to the right hand side of (A.22) are bounded as $n \longrightarrow \infty$, and given that $nh_1 \longrightarrow \infty$, $nh_2 \longrightarrow \infty$ and $\delta > 0$, we have from (A.22) that Liapunov's condition (A.18) is satisfied, so that $\sum_{i=1}^{n} L_{n,i} \xrightarrow{d} \mathcal{N}(0,1)$.

Now, let $D = \frac{\lambda_1^2\sigma_0^2(\alpha_0)f_0(\alpha_0)}{[g_0^{(2)}(\alpha_0)]^2} \int [K^{(1)}(z)]^2 dz + \lambda_2^2\sigma_0^2(\alpha_0)f_0(\alpha_0) \int [K(z)]^2 dz$. Note that since $\frac{Var(W_i)}{D} \to 1$ by (23), $\sum_{i=1}^{n} L_{n,i} = \sum_{i=1}^{n} \frac{W_{n,i}-E(W_{n,i})}{[nVar(W_{n,i})]^{1/2}}$ and $\sum_{i=1}^{n} \frac{W_{n,i}-E(W_{n,i})}{[nD]^{1/2}}$ have the same limiting distribution, and therefore:

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{W_{n,i} - E(W_{n,i})\} \xrightarrow{d}$$
$$\mathcal{N}\left(0, \frac{\lambda_1^2\sigma_0^2(\alpha_0)f_0(\alpha_0)}{[g_0^{(2)}(\alpha_0)]^2} \int [K^{(1)}(z)]^2 dz + \lambda_2^2\sigma_0^2(\alpha_0)f_0(\alpha_0) \int [K(z)]^2 dz\right)$$

as required ∎

Lemma 10 *Let $W_{n,i}$ and $K^*(t_i; h_1, h_2)$ be as in Lemma 8 and 9. Assume,*

(i) $\{(y_1, t_1), \ldots, (y_n, t_n)\}$ is an i.i.d. sample.

(ii) α_0 is in the interior of $\mathcal{T}$, so that $g_0^{(1)}(\alpha_0) = 0$.

(iii) $g_0(t)$ three times continuously differentiable and its derivatives up to the third order are bounded. Also, $g_0^{(2)}(\alpha_0) \neq 0$.

(iv) $f_0(t)$ is continuous and bounded away from zero uniformly in $\mathcal{T}$. Also, partial derivatives of $f_0(t)$ exists up to the third order and are bounded, and $f_0^{(1)}(t)$ is continuous at α_0.

(v) $h_1 \to 0$, $h_2 \to 0$, $nh_1^7 \to 0$, $nh_2^5 \to 0$, *as* $n \to \infty$.

(vi) Finally, let $K(\cdot)$ *be symmetric and satisfy* $\int u^2 K(u)du < \infty$ *and* $|u^3| \cdot |K(u)| \to 0$ *as* $|u| \to \infty$.

Then,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} E(W_{n,i}) \longrightarrow 0$$

Proof. As before, let $r(t) = g_0(t) - g_0(\alpha_0)$. Also, let $\phi(t) = r(t) f_0(t)$. Then, using iterated expectations along with the fact that the $W_{n,i}$ are i.i.d., and given the definitions of $W_{n,i}$ and $K^*(t_i; h_1, h_2)$ we observe that $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} E(W_{n,i})$ can be written as:

$$\begin{aligned}
&\frac{1}{\sqrt{n}} \sum_{i=1}^{n} E(W_{n,i}) \\
= \; & \sqrt{n} E\left([Y_i - g_0(\alpha_0)] K^*(t_i; h_1, h_2)\right) = \sqrt{n} E\left[K^*(t_i; h_1, h_2) r(t_i)\right] \\
= \; & \frac{\lambda_1}{g_0^{(2)}(\alpha_0)} \frac{\sqrt{n}}{\sqrt{h_1}} \int K^{(1)}\left(\frac{\alpha_0 - t_i}{h_1}\right) \phi(t_i) dt_i \\
& -\lambda_2 \frac{\sqrt{n}}{\sqrt{h_2}} \int K\left(\frac{\alpha_0 - t_i}{h_2}\right) \phi(t_i) dt_i \\
= \; & \frac{\lambda_1}{g_0^{(2)}(\alpha_0)} \sqrt{nh_1^7} \frac{1}{h_1^4} \int K^{(1)}\left(\frac{\alpha_0 - t_i}{h_1}\right) \phi(t_i) dt_i \\
& -\lambda_2 \sqrt{nh_2^5} \frac{1}{h_2^3} \int K\left(\frac{\alpha_0 - t_i}{h_2}\right) \phi(t_i) dt_i
\end{aligned} \tag{A.24}$$

where in the last equality we multiplied the numerator and denominator of the first term by $\sqrt{h_1^7}$ and of the second term by $\sqrt{h_2^5}$. Given our assumptions that $nh_1^7 \longrightarrow 0$ and $nh_2^5 \longrightarrow 0$, our proof will be completed by showing that the terms $\frac{1}{h_1^4} \int K^{(1)}\left(\frac{\alpha_0 - t_i}{h_1}\right) \phi(t_i) dt_i$ and $\frac{1}{h_2^3} \int K\left(\frac{\alpha_0 - t_i}{h_2}\right) \phi(t_i) dt_i$ are both $O(1)$. First, consider the term $\frac{1}{h_1^4} \int K^{(1)}\left(\frac{\alpha_0 - t_i}{h_1}\right) \phi(t_i) dt_i$. Making the usual change of variable $z = (\alpha_0 - t_i)/h_1$ we find that

$$\frac{1}{h_1^4} \int K^{(1)}\left(\frac{\alpha_0 - t_i}{h_1}\right) \phi(t_i) dt_i = \frac{1}{h_1^3} \int K^{(1)}(z) \phi(\alpha_0 - h_1 z) dz \tag{A.25}$$

Next, we make a Taylor expansion of $\phi(\alpha_0 - h_1 z)$ around α_0. Before doing so, note that by definition of $\phi(t)$ (i.e. $\phi(t) = [g_0(t) - g_0(\alpha_0)]\, f_0(t)$) and using the fact that $g_0^{(1)}(\alpha_0) = 0$ we have :

$$\begin{aligned}
\phi(\alpha_0) &= 0 \qquad\qquad (A.26)\\
\phi^{(1)}(\alpha_0) &= 0\\
\phi^{(2)}(\alpha_0) &= g_0^{(2)}(\alpha_0) f_0(\alpha_0)\\
\phi^{(3)}(\alpha_0) &= g_0^{(3)}(\alpha_0) f_0(\alpha_0) + 3 g_0^{(2)}(\alpha_0) f_0^{(1)}(\alpha_0)
\end{aligned}$$

where all quantities are less than infinity by assumption. Also observe that: $\int K^{(1)}(z)dz = 0$, $\int zK^{(1)}(z)dz = -1$, and $\int z^2K^{(1)}(z)dz = 0$; where the first and last equality follows from the fact that $K(\cdot)$ is symmetric, and in the second one we used integration by parts and the assumption that $|zK(z)| \to 0$ as $z \to \infty$. Hence, for some α^* between $\alpha_0 - h_1 z$ and α_0 we can write (A.25) as

$$\frac{1}{h_1^4}\int K^{(1)}\left(\frac{\alpha_0 - t_i}{h_1}\right)\phi(t_i)dt_i = -\frac{1}{6}\phi^{(3)}(\alpha^*)\int z^3K^{(1)}(z)dz$$

Given our assumptions on $g_0(\cdot)$ and $f_0(\cdot)$ we have that $\phi^{(3)}(\cdot)$ is continuous at α_0, and given our assumptions on h_1 we have that $\phi^{(3)}(\alpha^*) \to \phi^{(3)}(\alpha_0)$. Moreover, note that by using integration by parts again and the assumption that $|z^3K(z)| \to 0$ as $z \to \infty$ we obtain that $\int z^3K^{(1)}(z)dz = -3\int z^2K(z)dz$, with $\int z^2K(z)dz < \infty$ by assumption. Therefore, we obtain that:

$$\begin{aligned}
&\frac{1}{h_1^4}\int K^{(1)}\left(\frac{\alpha_0 - t_i}{h_1}\right)\phi(t_i)dt_i\\
\longrightarrow\ &\frac{1}{2}\left[g_0^{(3)}(\alpha_0) f_0(\alpha_0) + 3 g_0^{(2)}(\alpha_0)\, f_0^{(1)}(\alpha_0)\right]\int z^2K(z)dz \qquad (A.27)
\end{aligned}$$

Now consider the term $\frac{1}{h_2^3}\int K\left(\frac{\alpha_0-t_i}{h_2}\right)\phi(t_i)dt_i$. Following the same steps as above, and using (A.26) and our assumptions that $\int zK(z)dz = 0$ and $\int z^2K(z)dz < \infty$, we find that

$$\frac{1}{h_2^3}\int K\left(\frac{\alpha_0-t_i}{h_2}\right)\phi(t_i)dt_i \longrightarrow \frac{g_0^{(2)}(\alpha_0)\,f_0(\alpha_0)}{2}\int z^2K(z)dz \tag{A.28}$$

Therefore, the proof is completed by plugging (A.27) and (A.28) into (A.24) and using the fact that $nh_1^7 \longrightarrow 0$ and $nh_2^5 \longrightarrow 0$. ■

Given lemmas 6 to 10, we now proceed to prove Theorem 1.

Proof of Theorem 1. As mentioned before, according to the Cramér-Wold device we need to show that for every real numbers λ_1 and λ_2 we have:

$$\lambda_1\sqrt{nh_1^3}(\widehat{\alpha}-\alpha_0)+\lambda_2\sqrt{nh_2}(\widehat{g}_2(\widehat{\alpha})-g_0(\alpha_0)) \xrightarrow{d}$$

$$\mathcal{N}\left(0, \frac{\lambda_1^2\sigma_0^2(\alpha_0)}{[m_0^{(2)}(\alpha_0)]^2\ f_0(\alpha_0)}\int[K^{(1)}(z)]^2dz+\frac{\lambda_2^2\sigma_0^2(\alpha_0)}{f_0(\alpha_0)}\int[K(z)]^2dz\right) \tag{A.29}$$

From Lemma 7 we have that $\sqrt{nh_2}(\widehat{g}_2(\widehat{\alpha})-g_0(\alpha_0))$ is asymptotically equivalent to $\sqrt{nh_2}(\widehat{g}_2(\alpha_0)-g_0(\alpha_0))$, so we will focus on the asymptotic distribution of $\lambda_1\sqrt{nh_1^3}(\widehat{\alpha}-\alpha_0)+\lambda_2\sqrt{nh_2}(\widehat{g}_2(\alpha_0)-g_0(\alpha_0))$.

Let $\widehat{s}_l(t) = \frac{1}{nh_l}\sum\limits_{i=1}^{n} Y_iK\left(\frac{t-t_i}{h_l}\right)$ and $\widehat{f}_l(t) = \frac{1}{nh_l}\sum\limits_{i=1}^{n} K\left(\frac{t-t_i}{h_l}\right)$, so that $\widehat{g}_l(t) = \left[\widehat{f}_l(t)\right]^{-1}\widehat{s}_l(t)$, for $l=1,2$. From Lemma 6 we know that we can write $(\widehat{\alpha}-\alpha_0)$ as (see (4)) $-\widehat{g}_1^{(1)}(\alpha_0)\left[\widehat{g}_1^{(2)}(\alpha^*)\right]^{-1}$ for some α^* between α_0 and $\widehat{\alpha}$. Hence, we have that:

$$\lambda_1\sqrt{nh_1^3}(\widehat{\alpha}-\alpha_0)+\lambda_2\sqrt{nh_2}(\widehat{g}_2(\alpha_0)-g_0(\alpha_0)) = \tag{A.30}$$

$$\lambda_1\sqrt{nh_1^3}\left(-\frac{\widehat{g}_1^{(1)}(\alpha_0)}{\widehat{g}_1^{(2)}(\alpha^*)}\right)+\frac{\lambda_2\sqrt{nh_2}}{\widehat{f}_2(\alpha_0)}(\widehat{s}_2(\alpha_0)-g_0(\alpha_0)\widehat{f}_2(\alpha_0)) \tag{A.31}$$

As in Ziegler (2000), it is convenient to write $\widehat{g}^{(1)}(\alpha_0)$ as:

$$\begin{aligned}\widehat{g}_1^{(1)}(\alpha_0) &= \frac{\widehat{s}_1^{(1)}(\alpha_0) - \widehat{g}_1(\alpha_0)\widehat{f}_1^{(1)}(\alpha_0)}{\widehat{f}_1(\alpha_0)} \\ &= \frac{\widehat{s}_1^{(1)}(\alpha_0) - g_0(\alpha_0)\widehat{f}_1^{(1)}(\alpha_0)}{\widehat{f}_1(\alpha_0)} - \frac{\widehat{f}_1^{(1)}(\alpha_0)}{\widehat{f}_1(\alpha_0)}[\widehat{g}_1(\alpha_0) - g_0(\alpha_0)]\end{aligned}$$

Plugging this into (A.31) we get that (A.30) is equal to:

$$\begin{aligned}&\frac{\lambda_1\sqrt{nh_1^3}}{\widehat{g}_1^{(2)}(\alpha^*)}\left(-\frac{\widehat{s}_1^{(1)}(\alpha_0) - g_0(\alpha_0)\widehat{f}_1^{(1)}(\alpha_0)}{\widehat{f}_1(\alpha_0)} + \frac{\widehat{f}_1^{(1)}(\alpha_0)}{\widehat{f}_1(\alpha_0)}[\widehat{g}_1(\alpha_0) - g_0(\alpha_0)]\right) \\ &+\frac{\lambda_2\sqrt{nh_2}}{\widehat{f}_2(\alpha_0)}(\widehat{s}_2(\alpha_0) - g_0(\alpha_0)\widehat{f}_2(\alpha_0)) \qquad (A.32)\end{aligned}$$

Note that given our assumptions we have that: $\widehat{f}_l(\alpha_0) \xrightarrow{p} f_0(\alpha_0)$ for $l = 1, 2$ (e.g., Parzen 1962) and $\widehat{f}_1^{(1)}(\alpha_0) \xrightarrow{p} f_0^{(1)}(\alpha_0)$ (e.g., Schuster 1969). Also, as discussed in the proof of Lemma 6 we have that $\widehat{g}_1^{(2)}(\alpha^*) \xrightarrow{p} g_0^{(2)}(\alpha_0)$. Thus, (A.30) is asymptotically equivalent to

$$\begin{aligned}&-\frac{\lambda_1}{g_0^{(2)}(\alpha_0)f_0(\alpha_0)}\sqrt{nh_1^3}[\widehat{s}_1^{(1)}(\alpha_0) - g_0(\alpha_0)\widehat{f}_1^{(1)}(\alpha_0)] \qquad (A.33) \\ &+\frac{\lambda_1 f_0^{(1)}(\alpha_0)}{g_0^{(2)}(\alpha_0)f_0(\alpha_0)}\sqrt{nh_1^3}[\widehat{g}_1(\alpha_0) - g_0(\alpha_0)] + \frac{\lambda_2\sqrt{nh_2}}{f_0(\alpha_0)}(\widehat{s}_2(\alpha_0) - g_0(\alpha_0)\widehat{f}_2(\alpha_0))\end{aligned}$$

Note that we can write $\sqrt{nh_1^3}[\widehat{g}_1(\alpha_0) - g_0(\alpha_0)]$ from the second term as

$$\begin{aligned}&\sqrt{nh_1^3}[\widehat{g}_1(\alpha_0) - g_0(\alpha_0)] \\ =\ &\sqrt{nh_1^3}\,[\widehat{g}_1(\alpha_0) - E\,(\widehat{g}_1(\alpha_0)|t_1, t_2, \ldots, t_n)] + \sqrt{nh_1^3}\,[E\,(\widehat{g}_1(\alpha_0)|t_1, t_2, \ldots, t_n) - g_0(\alpha_0)]\end{aligned}$$

It is a standard result that, given our assumptions, $\sqrt{nh_1}[\widehat{g}_1(\alpha_0) - E\,(\widehat{g}_1(\alpha_0)|$ $t_1, t_2, \ldots, t_n)]$ is asymptotically bounded (e.g., Theorem 3.5 in Pagan and Ullah, 1999), so that $\sqrt{nh_1^3}\,[\widehat{g}_1(\alpha_0) - E\,(\widehat{g}_1(\alpha_0)|t_1, t_2, \ldots, t_n)] = h_1\sqrt{nh_1}\,[\widehat{g}_1(\alpha_0) - E\,(\widehat{g}_1(\alpha_0)|t_1, t_2, \ldots, t_n)]$ $\xrightarrow{P} 0$. Also, given our assumption that $nh_1^7 \to 0$, we obtain that $\sqrt{nh_1^3}[E(\widehat{g}_1(\alpha_0)|t_1, t_2, \ldots,$

$t_n) - g_0(\alpha_0)] \xrightarrow{P} 0$. The proof of this latter result follows basically the same proof of Theorem 3.6 in Pagan and Ullah (1999), which shows that $\sqrt{nh}\left[E\left(\widehat{g}(\alpha_0)|t_1, t_2, \ldots, t_n\right) - g_0(\alpha_0)\right] \xrightarrow{P} 0$ for the usual Nadayara-Watson estimator when $nh^5 \to 0$. However, the fact that the term in brackets in our case is multiplied by $\sqrt{nh_1^3}$ instead of $\sqrt{nh}$ allows us to obtain the result even though in our case $nh_1^5 \to \infty$. Intuitively, the expected value of the term in brackets is $O\left(h_1^2\right)$, so that when multiplied by $\sqrt{nh_1^3}$ and using our assumption that $nh_1^7 \to 0$ gives that the expected value of that term goes to zero asymptotically. Also, it is straightforward to show that its variance vanishes asymptotically, so the convergence to zero in probability follows. Therefore, $\sqrt{nh_1^3}[\widehat{g}_1(\alpha_0) - g_0(\alpha_0)] \xrightarrow{P} 0$, and the second term in (A.33) converges to zero in probability.

Now, given the definitions of $\widehat{s}_l(t)$ and $\widehat{f}_l(t)$, we have that: $\widehat{s}_l^{(1)}(t) = \frac{1}{nh_l^2}\sum_{i=1}^{n} Y_i K^{(1)}\left(\frac{t-t_i}{h_l}\right)$ and $\widehat{f}_l^{(1)}(t) = \frac{1}{nh_l^2}\sum_{i=1}^{n} K^{(1)}\left(\frac{t-t_i}{h_l}\right)$. Substituing these into (A.33) and using the fact that the second term in equation (A.33) is asymptotically negligible we obtain that (A.30) is asymptotically equivalent to:

$$\begin{aligned}
& \frac{1}{f_0(\alpha_0)}\left[\frac{-\lambda_1\sqrt{nh_1^3}}{g_0^{(2)}(\alpha_0)\cdot nh_1^2}\left\{\sum_{i=1}^{n} Y_i K^{(1)}\left(\frac{\alpha_0 - t_i}{h_1}\right) - g_0(\alpha_0)\sum_{i=1}^{n} K^{(1)}\left(\frac{\alpha_0 - t_i}{h_1}\right)\right\}\right. \\
& \left. +\lambda_2\frac{\sqrt{nh_2}}{nh_2}\left\{\sum_{i=1}^{n} Y_i K\left(\frac{\alpha_0 - t_i}{h_2}\right) - g_0(\alpha_0)\sum_{i=1}^{n} K\left(\frac{\alpha_0 - t_i}{h_2}\right)\right\}\right] \\
= \;& -\frac{1}{f_0(\alpha_0)}\left[\frac{\lambda_1}{g_0^{(2)}(\alpha_0)}\frac{1}{\sqrt{nh_1}}\left\{\sum_{i=1}^{n}[Y_i - g_0(\alpha_0)]K^{(1)}\left(\frac{\alpha_0 - t_i}{h_1}\right)\right\}\right. \\
& \left. -\frac{\lambda_2}{\sqrt{nh_2}}\left\{\sum_{i=1}^{n}[Y_i - g_0(\alpha_0)]K\left(\frac{\alpha_0 - t_i}{h_2}\right)\right\}\right] \\
= \;& -\frac{1}{f_0(\alpha_0)}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[Y_i - g_0(\alpha_0)]\left(\frac{\lambda_1}{g_0^{(2)}(\alpha_0)\sqrt{h_1}}K^{(1)}\left(\frac{\alpha_0 - t_i}{h_1}\right)\right.\right. \\
& \left.\left. -\frac{\lambda_2}{\sqrt{h_2}}K\left(\frac{\alpha_0 - t_i}{h_2}\right)\right)\right]
\end{aligned}$$

$$= -\frac{1}{f_0(\alpha_0)}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[Y_i - g_0(\alpha_0)]K^*(t_i; h_1, h_2)\right) \tag{A.34}$$

where we define $K^*(t_i; h_1, h_2) = \frac{\lambda_1}{g_0^{(2)}(\alpha_0)\sqrt{h_1}}K^{(1)}\left(\frac{\alpha_0 - t_i}{h_1}\right) - \frac{\lambda_2}{\sqrt{h_2}}K\left(\frac{\alpha_0 - t_i}{h_2}\right)$. Let $W_{n,i} = [Y_i - g_0(\alpha_0)]K^*(t_i; h_1,\, h_2)$. Then, from (A.34) we have that (A.30) is asymptotically equivalent to:

$$-\frac{1}{f_0(\alpha_0)}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{W_{n,i} - E(W_{n,i})\} + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}E(W_{n,i})\right] \tag{A.35}$$

In Lemma 9 we showed that:

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{W_{n,i} - E(W_{n,i})\} \xrightarrow{d}$$
$$\mathcal{N}\left(0, \frac{\lambda_1^2\sigma_0^2(\alpha_0)f_0(\alpha_0)}{[g_0^{(2)}(\alpha_0)]^2}\int[K^{(1)}(z)]^2dz + \lambda_2^2\sigma_0^2(\alpha_0)f_0(\alpha_0)\int[K(z)]^2dz\right) \tag{A.36}$$

while in Lemma 10 we showed that:

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}E(W_{n,i}) \longrightarrow 0 \tag{A.37}$$

Using Slutsky's Theorem we find that (A.35), and consequently (A.30) is asymptotically distributed as: $\mathcal{N}\left(0, \frac{\lambda_1^2\sigma_0^2(\alpha_0)}{[g_0^{(2)}(\alpha_0)]^2\, f_0(\alpha_0)}\int[K^{(1)}(z)]^2dz + \frac{\lambda_2^2\sigma_0^2(\alpha_0)}{f_0(\alpha_0)}\int[K(z)]^2dz\right)$, as required. ∎

The next Lemma was used in the proofs of Lemmas 6 and 7.

Lemma 11 *(Uniform rate of convergence for higher order derivatives of the NW estimator). Let $\widehat{g}(x)$ be the Nadayara-Watson estimator and let the p-th derivative of $g_0(x) = E[Y|X]$ be estimated by $\partial^p\widehat{g}(x)/\partial x^p$. Assume:*

(i) $\{(y_1, x_1), \ldots, (y_n, x_n)\}$ is an i.i.d. sample.

(ii) $f_0(x)$ (density of x) is continuous at x. Also, partial derivatives of $f_0(x)$ exists up to the $p+1$ order, with the $p+1$ derivative continuous at x.

(iii) $g_0(x)$ is continuous at x. Also, partial derivatives of $g_0(x)$ exists up to the $p+1$ order, with the p-th derivative uniformly continuous and the $p+1$ derivative continuous at x.

(vii) Let $h = h(n)$ be such that: $h \to 0$ as $n \to \infty$.

(viii) The kernel $K(\cdot)$ satisfies: (a) $\int K(u)du = 1$; (b) K is p times continuously differentiable; (c) $\int |K(u)| du < \infty$; (d) $\sup_u |K(u)| < \infty$; (e) $|u^p| |K(u)| \to 0$ as $|u| \to \infty$. Finally, let $\psi(w)$ be the characteristic function of $K(\cdot)$ so that $\psi(w) = \int e^{iwu} K(u)du$, with $i^2 = -1$. Assume that $\int |w^p \psi(w)| dw < \infty$.

Then,

$$\sup_x \left|\widehat{g}^{(p)}(x) - g_0^{(p)}(x)\right| = O_p\left(\frac{1}{\sqrt{nh^{2(p+1)}}}\right)$$

Proof. The proof of this Lemma closely follows the one in Bierens (1987) for uniform consistency of the Nadayara-Watson estimator. Let $\widehat{s}(x) = \frac{1}{nh}\sum_{j=1}^{n} Y_j K\left(\frac{x-x_j}{h}\right)$, so that $\widehat{s}^{(p)}(x) = \frac{1}{nh^{p+1}}\sum_{j=1}^{n} Y_j K^{(p)}\left(\frac{x-x_j}{h}\right)$ is the p-*th* derivative estimator of $s_0^{(p)}(x) = \partial^p E[Y|X] f_o(x) / \partial x^p$. Also, let $\psi(t) = \int \exp(itx) \cdot K(x)\, dx$, with $i^2 = -1$, and assume $\int |t^p \psi(t)|\, dt < \infty$. By the well known inversion formula for Fourier transforms the kernel can be written as $K(x) = \frac{1}{2\pi}\int \exp(-itx) \cdot \psi(t)\, dt$, so that $K^{(p)}(x) = \frac{1}{2\pi}\int (-it)^p \exp(-itx) \cdot \psi(t)\, dt$. Hence, $\widehat{s}^{(p)}(x)$ can be written as:

$$\begin{aligned}
\widehat{s}^{(p)}(x) &= \frac{1}{nh^{p+1}}\sum_{j=1}^{n} Y_j K^{(p)}\left(\frac{x-x_j}{h}\right) \\
&= \frac{1}{nh^{p+1}}\sum_{j=1}^{n} Y_j \left[\frac{1}{2\pi}\int (-it)^p \exp\left(-it\left(\frac{x-x_j}{h}\right)\right) \cdot \psi(t)\, dt\right] \\
&= \frac{1}{2\pi n}\sum_{j=1}^{n} Y_j \int (-is)^p \exp(-isx) \cdot \exp(isx_j) \cdot \psi(hs)\, ds
\end{aligned}$$

$$\widehat{s}^{(p)}(x) = \frac{1}{2\pi}\int\left[\frac{1}{n}\sum_{j=1}^{n}Y_j\exp(isx_j)\right](-is)^p\exp(-isx)\cdot\psi(hs)\,ds \tag{A.38}$$

where in the third equality we have used the change of variable $t = hs$. In this case, and given our i.i.d. assumption, we have

$$E\left[\widehat{s}^{(p)}(x)\right] = \frac{1}{2\pi}\int(-is)^p\exp(-isx)\cdot\psi(hs)\,E\left[Y_1\exp(isx_1)\right]ds$$

Then, we can write

$$\begin{aligned}
&\left|\widehat{s}^{(p)}(x) - E\left[\widehat{s}^{(p)}(x)\right]\right| \\
= &\left|\frac{1}{2\pi}\int\left[\frac{1}{n}\sum_{j=1}^{n}Y_j\exp(isx_j)\right](-is)^p\exp(-isx)\cdot\psi(hs)\,ds-\right. \\
&\left.\frac{1}{2\pi}\int(-is)^p\exp(-isx)\cdot\psi(hs)\,E\left[Y_1\exp(isx_1)\right]ds\right| \\
= &\left|\frac{1}{2\pi}\int(-is)^p\exp(-isx)\cdot\psi(hs)\left[\frac{1}{n}\sum_{j=1}^{n}Y_j\exp(isx_j) - E\left[Y_1\exp(isx_1)\right]\right]ds\right| \\
\leq &\frac{1}{2\pi}\int\left|s^p\psi(hs)\right|\left|\frac{1}{n}\sum_{j=1}^{n}Y_j\exp(isx_j) - E\left[Y_1\exp(isx_1)\right]\right|ds
\end{aligned} \tag{A.39}$$

Since the right side of the inequality in (A.39) does not depend on x we have that:

$$\sup_x\left|\widehat{s}^{(p)}(x) - E\left[\widehat{s}^{(p)}(x)\right]\right| \leq \frac{1}{2\pi}\int\left|s^p\psi(hs)\right|\left|\frac{1}{n}\sum_{j=1}^{n}Y_j\exp(isx_j) - E\left[Y_1\exp(isx_1)\right]\right|ds \tag{A.40}$$

Taking expectations to both sides of (A.40) we find that:

$$\begin{aligned}
&E\left[\sup_x\left|\widehat{s}^{(p)}(x) - E\left[\widehat{s}^{(p)}(x)\right]\right|\right] \\
\leq &\frac{1}{2\pi}\int\left|s^p\psi(hs)\right|E\left\{\left|\frac{1}{n}\sum_{j=1}^{n}Y_j\exp(isx_j) - E\left[Y_1\exp(isx_1)\right]\right|\right\}ds
\end{aligned} \tag{A.41}$$

Consider for the moment the expectation inside the integral. It can be writen as:

$$E\left\{\left|\frac{1}{n}\sum_{j=1}^{n}Y_j\exp(isx_j) - E\left[Y_1\exp(isx_1)\right]\right|\right\} \tag{A.42}$$

$$
\begin{aligned}
&= E\left\{\left|\frac{1}{n}\sum_{j=1}^{n}Y_j\left[\cos\left(-sx_j\right)-i\sin\left(-sx_j\right)\right]-E\left[Y_1\left[\cos\left(-sx_j\right)-i\sin\left(-sx_j\right)\right]\right]\right|\right\}\\
&= E\left\{\left|\frac{1}{n}\sum_{j=1}^{n}Y_j\left[\cos\left(sx_j\right)-E\left[Y_1\cos\left(sx_1\right)\right]\right]+i\frac{1}{n}\sum_{j=1}^{n}Y_j\sin\left(sx_j\right)-iE\left[Y_1\sin\left(sx_1\right)\right]\right|\right\}
\end{aligned}
$$

From this last expression, let $A=\frac{1}{n}\sum_{j=1}^{n}Y_j\left[\cos\left(sx_j\right)-E\left[Y_1\cos\left(sx_1\right)\right]\right]$ and $B=\frac{1}{n}\sum_{j=1}^{n}Y_j\sin\left(sx_j\right)-iE\left[Y_1\sin\left(sx_1\right)\right]$. Then, note that $E\left[A\right]=E\left[B\right]=0$, so that $E\left|A+iB\right|=E\left[\left(A^2+B^2\right)^{-1/2}\right]\leq E\left[\left(A^2\ \ +B^2\right)\right]^{-1/2}=\left[var\left(A\right)+var\left(B\right)\right]^{-1/2}$. Plugging this into the last equation in (A.42) and using the fact that we have i.i.d. observations we find:

$$
\begin{aligned}
& E\left\{\left|\frac{1}{n}\sum_{j=1}^{n}Y_j\exp\left(isx_j\right)-E\left[Y_1\exp\left(isx_1\right)\right]\right|\right\}\\
\leq\ & \frac{1}{\sqrt{n}}\left[var\left\{Y_j\cos\left(sx_j\right)\right\}+var\left\{Y_j\sin\left(sx_j\right)\right\}\right]^{-1/2}\\
\leq\ & \frac{1}{\sqrt{n}}\left[E\left(Y_j^2\right)\right]^{-1/2} \qquad \text{(A.43)}
\end{aligned}
$$

Plugging the result in (A.43) into (A.41) we obtain:

$$
\begin{aligned}
E\left[\sup_x\left|\widehat{s}^{(p)}\left(x\right)-E\left[\widehat{s}^{(p)}\left(x\right)\right]\right|\right] &\leq \frac{1}{2\pi}\frac{1}{\sqrt{n}}\left[E\left(Y_j^2\right)\right]^{-1/2}\int\left|s^p\psi\left(hs\right)\right|ds\\
&= \frac{1}{h^{p+1}\sqrt{n}}\frac{1}{2\pi}\left[E\left(Y_j^2\right)\right]^{-1/2}\int\left|u^p\psi\left(u\right)\right|du\\
&= O\left(\frac{1}{\sqrt{nh^{2(p+1)}}}\right) \qquad \text{(A.44)}
\end{aligned}
$$

since by assumption $E\left(Y^2\right)<\infty$ and $\int\left|u^p\psi\left(u\right)\right|du<\infty$. Thus, by Markov inequality we have that

$$
\sup_x\left|\widehat{s}^{(p)}\left(x\right)-E\left[\widehat{s}^{(p)}\left(x\right)\right]\right|\xrightarrow{p}0 \qquad \text{(A.45)}
$$

By the triangle inequality we have

$$
\sup_x\left|\widehat{s}^{(p)}\left(x\right)-s_0^{(p)}\left(x\right)\right|\leq\sup_x\left|\widehat{s}^{(p)}\left(x\right)-E\left[\widehat{s}^{(p)}\left(x\right)\right]\right|+\sup_x\left|E\left[\widehat{s}^{(p)}\left(x\right)\right]-s_0^{(p)}\left(x\right)\right| \qquad \text{(A.46)}
$$

It has been previously shown in the literature that $\sup_x \left| E\left[\widehat{s}^{(p)}(x)\right] - s_0^{(p)}(x)\right| = O\left(h^2\right)$ (e.g., Corollary 1 in Schuster and Yakowitz, 1979). Then, it follows from (A.45) that

$$\sup_x \left| \widehat{s}^{(p)}(x) - s_0^{(p)}(x)\right| \xrightarrow{p} 0 \tag{A.47}$$

Finally, a similar result holds for the derivative density estimator $\widehat{f}^{(p)}(x) = \frac{1}{nh^{p+1}} \sum_{j=1}^{n} K^{(p)}\left(\frac{x-x_i}{h}\right)$. Then, it follows that

$$\sup_x \left| \widehat{g}^{(p)}(x) - g_0^{(p)}(x)\right| = O_p\left(\frac{1}{\sqrt{nh^{2(p+1)}}}\right) \tag{A.48}$$

∎

Appendix B

Proofs: Non-experimental Design

In this case we assume that assignment into different levels of the treatment is unconfounded given a set of covariates X with dimension equal to k. That is, using the notation from chapter 2 we have:

$$\{Y(t)\}_{t\in\mathcal{T}} \perp T|X$$

As discussed in chapter 2, let $E[Y(t)] = E_X[\tau(x)\,E[Y(t)|X=x]]$, where $\tau(\cdot)$ is a trimming function used to avoid the "denominator problem" by keeping a denominator bounded away from zero. Under the unconfoundedness assumption we can write the dose-response function as:

$$\begin{aligned} E[Y(t)] &= E_X[\tau(x)\,E[Y(t)|X=x]] = E_X[\tau(x)\,E[Y(t)|T=t,X=x]] \\ &= E_X[\tau(x)\,E[Y|T=t,X=x]]\,, \end{aligned} \tag{B.1}$$

for all $t\in\mathcal{T}$; where in the first equality we have used iterated expectations, and in the second one our unconfoundedness assumption. Writing the maximum of the dose-response as

$$\alpha_0 = t^* = \arg\max_{t \in T} E\{Y(t)\} \tag{B.2}$$

and using (B.1), we have that we can estimate α_0 as

$$\begin{aligned} \widehat{\alpha} &= \arg\max_{t \in T} \widehat{E}\{Y(t)\} = \arg\max_{t \in T} \widehat{E}_X \left[\tau(x)\, \widehat{E}\,[Y|T=t, X=x] \right] \\ &= \arg\max_{t \in T} \frac{1}{n} \sum_{i=1}^{n} \tau(x_i)\, \widehat{g}_{h_1}(t, X_i), \end{aligned} \tag{B.3}$$

for all $t \in \mathcal{T}$; where $\widehat{g}(t, X_i)$ is the Nadayara-Watson multiple-regression estimator

$$\widehat{g}(t, X_i) = \frac{\sum_{i=1}^{n} Y_i K(\frac{t-t_i}{h_1}, \frac{x_1 - x_{1i}}{h_1}, \ldots, \frac{x_q - x_{qi}}{h_1})}{\sum_{i=1}^{n} K(\frac{t-t_i}{h_1}, \frac{x_1 - x_{1i}}{h_1}, \ldots, \frac{x_q - x_{qi}}{h_1})}$$

and $K(u)$ is a kernel function satisfying some conditions specified below, and $h > 0$ is the bandwidth.

Proof of Theorem 3. In analyzing the asymptotic behavior of $\widehat{\alpha}$, we use results from Newey (1994) on the asymptotic theory of functionals of kernel estimators. Specifically, Newey considers two-step estimators where the first step is a vector of kernel estimators, say $\widehat{s}(x)$, and the second is an m-estimator that depends on $\widehat{s}(x)$. To analyze the asymptotic behavior of (B.3), first we write the estimator as a two-step m-estimator that depends on kernel estimators. Let $q = [1, y]'$ and $w = [t\ x]$, so that w is formed by putting together the treatment level variable and the covariates. Given that the dimension of x is k, we have that the dimension of w is $k+1$. Also, let $f_0(w)$ be the density of w.

Then, we can write

$$s_0(w) = E[q|w]f_0(w) = \begin{bmatrix} f_0(w) \\ \\ E[y|w]f_0(w) \end{bmatrix} = \begin{bmatrix} s_{10}(w) \\ \\ s_{20}(w) \end{bmatrix}$$

where we defined $s_{10}(w) = f_0(w)$ and $s_{20}(w) = E[y|w]f_0(w)$. Let $z_i = (q_i, w_i)$, $i = 1, \ldots, n$, denote data observations on q and w, and let $K(u)$ be a kernel function and $h_1 > 0$ be the bandwidth. To simplify notation, in what follows we let $h_1 = h$. Define $K_h(u)$ as $K_h(u) = h^{-(k+1)}K(u/h)$. Then, we estimate $s_0(w)$ as

$$\widehat{s}(w) = \frac{1}{n}\sum_{j=1}^{n} q_j K_h(w - w_j) = \begin{bmatrix} \frac{1}{n}\sum\limits_{j=1}^{n} K_h(w - w_j) \\ \\ \frac{1}{n}\sum\limits_{j=1}^{n} y_j K_h(w - w_j) \end{bmatrix} = \begin{bmatrix} \widehat{s}_1(w) \\ \\ \widehat{s}_2(w) \end{bmatrix} \tag{B.4}$$

where we have let $\widehat{s}_1(w) = \frac{1}{n}\sum\limits_{j=1}^{n} K_h(w - w_j)$ and $\widehat{s}_2(w) = \frac{1}{n}\sum\limits_{j=1}^{n} y_j K_h(w - w_j)$. Having described the first-step estimator, we now describe the second-step m-estimator. Let $g_0(w) = E[y|w] = s_{20}(w)/s_{10}(w)$. Then, by definition of α_0 in (B.2), and given the assumption that α_0 is in the interior of $\mathcal{T}$, we have that $\alpha_0 = t^*$ solves

$$\begin{aligned} \frac{\partial E\left[Y(t)\right]}{\partial t} &= \frac{\partial E_X\left[\tau\left(x\right)E\left[Y|T=t, X=x\right]\right]}{\partial t} = \frac{\partial E_X\left[\tau\left(x\right)g_0(t,x)\right]}{\partial t} \\ &= E_X\left[\tau\left(x\right)\frac{\partial g_0(\alpha_0, x)}{\partial t}\right] = 0 \end{aligned} \tag{B.5}$$

where we have used (B.1) in the first equality. Hence, if we let $m(z, \alpha_0, s_0) = \tau\left(x\right)\partial g_0(\alpha_0, x)$ $/\partial t$, we have that $E\left[\tau\left(x\right)m(z, \alpha_0, s_0)\right] = 0$. In this case, the sample moment function is $m(z_i, \alpha, \widehat{s}) = \tau\left(x_i\right)\partial\widehat{g}(\alpha, x_i)/\partial t$. Then the second-step estimator $\widehat{\alpha}$ solves:

$$\frac{1}{n}\sum_{i=1}^{n} m(z_i, \alpha, \widehat{s}) = \frac{1}{n}\sum_{i=1}^{n}\tau\left(x_i\right)\frac{\partial\widehat{g}(\alpha, x_i)}{\partial t} = 0 \tag{B.6}$$

where we let $\widehat{g}(w) = \widehat{E}[y|w] = \widehat{s}_2(w)/\widehat{s}_1(w)$. This estimator is equivalent to the one in (B.3).

If we expand the left-hand side of (B.6) around α_0 and solve for $\widehat{\alpha}-\alpha_0$ we get

$$\begin{aligned}\widehat{\alpha}-\alpha_0 &= -\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial m(z_i,\alpha^*,\widehat{s})}{\partial\alpha}\right]^{-1}\left[\frac{1}{n}\sum_{i=1}^{n}m(z_i,\alpha_0,\widehat{s})\right]\\ &= -\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial m(z_i,\alpha^*,\widehat{s})}{\partial\alpha}\right]^{-1}\widehat{m}_n(\alpha_0) \qquad \text{(B.7)}\end{aligned}$$

where α^* is a mean value and, as in Newey (1994), we let $\widehat{m}_n(\alpha)=\frac{1}{n}\sum_{i=1}^{n}m(z_i,\alpha,\widehat{s})$. Newey (1994) presents an uniform (in α) convergence result (Lemma 5.1) that is useful for showing consistency of $\widehat{\alpha}$ and also of the Jacobian term in (B.7). Given this latter result, and assuming nonsingularity of the probability limit of the Jacobian term in (B.7), asymptotic normality of $\widehat{m}_n(\alpha_0)$ will imply asymptotic normality of $\alpha_0-\widehat{\alpha}$. Newey (1994) gives conditions under which $\sqrt{n}h^{\delta}\widehat{m}_n(\alpha_0)\xrightarrow{d}\mathcal{N}(0,V)$, for some $\delta\geqslant 0$. There are two steps for obtaining this asymptotic normality result for $\widehat{m}_n(\alpha_0)$. The fisrt one involves a linearization around s_0, and the second entails the asymptotic normality of such linearization. Newey (1994) provides Lemmas for each of those steps.

In what follows, let $||s||_j=\max_{\ell\leq j}\sup_{t\in\mathcal{T}}\left|\left|\frac{\partial^{\ell}s(t,x)}{\partial t^{\ell}}\right|\right|$. This is a Sobolev supremum norm of order j, and is the norm used by Newey (1994) to impose smoothness conditions on $m(z,\alpha,s)$ as a function of s.

We start by showing convergence in probability of the Jacobian term in (B.7). First, consider consistency of $\widehat{\alpha}$. As in the proof of Theorem 1, note that we can see $\widehat{\alpha}$ as an extremum estimator that maximizes the objective function $\frac{1}{n}\sum_{i=1}^{n}\tau(x_i)\widehat{g}(t,X_i)$. The assumptions in Theorem 3 directly satisfy all conditions for the uniform converge result (Lemma 5.1) in Newey (1994). Specifically, conditions (ix), (ii), (iii) and the bandwidth conditions in Theorem 3 directly imply conditions (i)-(iii) in Newey's Lemma 5.1. Hence,

$\frac{1}{n}\sum_{i=1}^{n} \tau(x_i)\,\widehat{g}(t, X_i)$ converges uniformly in probability to $E_X[\tau(x)\,E[Y|T = t, X = x]]$ for all $\alpha \in \mathcal{T}$. This latter result, along with assumptions (i) and (ix) in Theorem 3, guarantee that the assumptions of Theorem 2.1. in Newey and McFadden (1994) are satisfied, so that $\widehat{\alpha} \xrightarrow{p} \alpha_0$.

Given the definition of $m(z_i, \alpha, \widehat{s})$, we can write the Jacobian term in (B.7) as

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\partial m(z_i, \alpha^*, \widehat{s})}{\partial t} = \frac{1}{n}\sum_{i=1}^{n} \tau(x_i)\,\frac{\partial^2 \widehat{g}(\alpha^*, x_i)}{\partial t^2} \tag{B.8}$$

As before, we use Newey's Lemma 5.1 to show uniform convergence in probability of the Jacobian term. Assumptions (ii), (iii), (iv), (vii) and the bandwidth condition $nh^{k+5}/\ln(n)$ in Theorem 3 directly imply conditions (i) and (ii) in Newey's Lemma 5.1. Finally, by applying the quotient rule for derivatives is not difficult to show that condition (iii) in Lemma 5.1 is also satisfied. Then, the conclusion in Lemma 5.1 states that $\frac{1}{n}\sum_{i=1}^{n} \tau(x_i)\,\partial^2\widehat{g}(\alpha, x_i)/\partial t^2$ converges uniformly in probability to $E\left[\tau(x)\,\partial^2 g_0(\alpha, x)/\partial t^2\right]$ for all $\alpha \in \mathcal{T}$, and that $E\left[\tau(x)\,\partial^2 g_0(\alpha, x)/\partial t^2\right]$ is continuous on $\mathcal{T}$. This, along with consistency of $\widehat{\alpha}$, implies

$$\frac{1}{n}\sum_{i=1}^{n} \tau(x_i)\,\frac{\partial^2 \widehat{g}(\alpha^*, x_i)}{\partial t^2} \xrightarrow{P} E\left[\tau(x)\,\frac{\partial^2 g_0(\alpha_0, x)}{\partial t^2}\right] = \frac{\partial^2 E\left[\tau(x)\,E(Y|T = \alpha_0, X = x)\right]}{\partial t^2}$$

Let $M = E\left[\tau(x)\,\partial^2 g_0(\alpha_0, x)/\partial^2 t\right]$. Given our assumption of M being nonsingular (assumption (viii)), we obtain

$$\left[\frac{1}{n}\sum_{i=1}^{n} \tau(x_i)\,\frac{\partial m(z_i, \alpha^*, \widehat{s})}{\partial t}\right]^{-1} \xrightarrow{P} M^{-1} \tag{B.9}$$

The next step in the proof is to derive asymptotic normality of $\sqrt{nh^3}\,\widehat{m}_n(\alpha_0)$, where $\widehat{m}_n(\alpha_0) = \frac{1}{n}\sum_{i=1}^{n} m(z_i, \alpha_0, \widehat{s})$ (See (B.7)). Let $w_0 = [\alpha_0\ x]$. Then we can write $m(z_i, \alpha_0, \widehat{s})$ as

$$m(z, \alpha_0, \widehat{s}) \;=\; \frac{\partial \widehat{g}(\alpha_0, x_i)}{\partial t} = \tau(w_0)\,\frac{\partial\left[\widehat{s}_2(w_0)/\widehat{s}_1(w_0)\right]}{\partial t}$$

$$= \frac{\tau(w_0)\left\{[\partial\widehat{s}_2(w_0)/\partial t]\,\widehat{s}_1(w_0) - [\partial\widehat{s}_1(w_0)/\partial t]\,\widehat{s}_2(w_0)\right\}}{[\widehat{s}_1(w_0)]^2} \tag{B.10}$$

Given the form of $m(z, \alpha_0, \widehat{s})$, we first use Lemma 5.4 in Newey (1994) to linearize it. Let

$$D(z, s; \widetilde{s}) = \frac{\tau(w_0)}{\widetilde{s}_1(w_0)}\left[\frac{\partial s_2(w_0)}{\partial t} - \left\{\frac{\widetilde{s}_2(w_0)}{\widetilde{s}_1(w_0)}\right\}\frac{\partial s_1(w_0)}{\partial t}\right] \tag{B.11}$$

and, to simplify notation, let $D(z, s) = D(z, s; s_0)$. Using Newey's Lemma 5.4 we obtain that for some $\delta \geq 0$

$$\sqrt{n}h^{\delta}\frac{1}{n}\sum_{i=1}^{n}[m(z_i, \alpha_0, \widehat{s}) - m(z_i, \alpha_0, \widehat{s})] = \sqrt{n}h^{\delta}\,[m(\widehat{s}) - m(s_0)] + o_p(1) \tag{B.12}$$

where $m(s) = \int D(z, s)dF(z)$ and $m(s)$ satisfies the conditions in Newey's Lemma 5.3. This latter lemma specifies conditions under which $\sqrt{n}h^{\delta}\,[m(\widehat{s}) - m(s_0)] \xrightarrow{d} \mathcal{N}(0, V)$, for some V.

We now verify the hypotheis of Lemma 5.4 in Newey (1994). Note that our assumptions (ii), (iii) and (iv) are equivalent to assumptions K, H and Y in Newey (1994). Let $\delta = 3/2$, $\Delta = \Delta_1 = \Delta_2 = 1$, where Δ, Δ_1, and Δ_2 are defined in Newey (1994), and note that by assumption (iii) we have $d \geq r + 1$. Note that $D(z, s)$ is linear in s on the set $\{s : ||s||_1 < \infty\}$, and $||D(z, s)|| \leq ||s||_1$. As before, it is straightforward algebra to show that for all s with $||s - s_0||_1 < \varepsilon$ we have $||m(z, s) - m(z, s_0) - D(z, s - s_0)|| \leq C$ $||s - s_0||_1^2$, for a given constant C. Now we consider the rate hypothesis in Newey's Lemma 5.4. Here one needs to show that for $\eta_n = \left[\ln(n)/(nh^{k+3})\right]^{1/2} + h^r$ we have: a) $\eta_n \to 0$; b) $\sqrt{n}h^{3/2}E\,[b(z)]\,\eta_n^2 \to 0$; and c) $\sqrt{n}h^{k+1/2} \to \infty$. From assumption (x) we have that

$$\frac{nh^{2k+3}}{[\ln(n)]^2} = \frac{nh^{k+3}}{\ln(n)} \cdot \frac{h^k}{\ln(n)} \to \infty$$

Since $k \geq 0$ and $h \to 0$, we have that the second term goes to 0, so our assumption implies that $nh^{k+3}/\ln(n) \to \infty$. Hence, $\eta_n \to 0$ follows from this last result and the fact that $r \geq 0$ and $h \to 0$. Now, to verify the next condition we note that

$$\sqrt{n}h^{3/2}\eta_n^2 = \frac{\ln(n)}{\sqrt{n}h^{k+3/2}} + 2\left[\ln(n)\right]^{1/2} h^{r-k/2} + \sqrt{n}h^{2r+3/2} \tag{B.13}$$

As for the first term note that $\frac{\ln(n)}{\sqrt{n}h^{k+3/2}} = \left[\frac{[\ln(n)]^2}{nh^{2k+3}}\right]^{1/2} \to 0$, since by assumption (x) the term inside the parenthesis in the second expression goes to zero. For the second term note that our asumptions require $3 + 2r > 2k + 3$, which combined with the fact that $k \geq 0$, implies that $r > k/2$. Let $b = r - k/2$ (note that $b > 0$), and let h be of the form $h \sim n^{-\delta}$ with $\delta > 0$[1]. Then we can write the second term as $2\left[\ln(n)\right]^{1/2} h^b \sim \left[\ln(n)\right]^{1/2} n^{-b\delta}$. Using L'Hospital's rule in this last expression, and using the fact that $b\delta > 0$, we obtain that $2\left[\ln(n)\right]^{1/2} h^{r-k/2} \to 0$. As for the last term, remember that for centering the asymptotic distribution of $\alpha_0 - \alpha$ around zero we assumed that $nh^{2r+3} \to 0$. Hence, we have that $\sqrt{n}h^{2r+3/2} = \left[nh^{4r+3}\right]^{1/2} = \left[h^{2r} \cdot nh^{2r+3}\right]^{1/2} \to 0$, since $h^{2r} \to 0$ and $nh^{2r+3} \to 0$ by assumption. Therefore, since each of the terms in (B.13) goes to zero asymptotically, then we have that $\sqrt{n}h^{3/2}\eta_n^2 \to 0$.

Finally, we show that $\sqrt{n}h^{k+1/2} \to \infty$. This follows from writing $\sqrt{n}h^{k+1/2} = \left[nh^{2k+1}\right]^{1/2} = \left[nh^{2k+3} \cdot h^{-2}\right]^{1/2} \to \infty$, since $h^{-2} \to \infty$ and, by assumption (x), $nh^{2k+3} \to \infty$. Thus, the rate hypothesis of Newey's Lemma 5.4 are satisfied.

Given that all conditions of Lemma 5.4 are satisfied, we have that (B.12) holds with $\delta = 3/2$ and

$$m(s) \;=\; \int D(z, s; s_0) dF(z)$$

[1]Here "$\sim$" stands for "proportional to".

$$= \int \frac{\tau\left(w(\upsilon)\right)}{s_{10}(w(\upsilon))} \left[\frac{\partial s_2(w(\upsilon))}{\partial t} - \left\{\frac{s_{20}(w(\upsilon))}{s_{10}(w(\upsilon))}\right\} \frac{\partial s_1(w(\upsilon))}{\partial t}\right] \widetilde{f}_0(\upsilon)\, d\upsilon \quad \text{(B.14)}$$

for $\upsilon = x$ and $w(\upsilon) = (\alpha_0, \upsilon)$.

We now turn our attention to showing asymptotic normality of $\sqrt{n}h^{3/2}[m(\widehat{s}) - m(s_0)]$. As previously disccused, here we use Newey's Lemma 5.3. Here, we need a matrix of functions $\phi(\upsilon)$ with domain $\Re^k$ and a vector of functions $t(\upsilon)$ in $\Re$ such that $m(s) = \int \phi(\upsilon)\,[\partial s(w(\upsilon))/\partial t]\, d\upsilon$ for $w(\upsilon) = (t(\upsilon), \upsilon)$. Let

$$\phi(\upsilon) = \frac{\tau\left(w(\upsilon)\right)}{s_{10}(w(\upsilon))} \widetilde{f}_0(\upsilon)\left[-g_0(w(\upsilon)), 1\right] \quad \text{(B.15)}$$

where $t(\upsilon) = \alpha_0$. Note that given our definition of $m(s)$ used in (B.14), we can write

$$m(s) = \int \phi(\upsilon)\,[\partial s(w(\upsilon))/\partial t]\, d\upsilon = \int \frac{\tau\left(w(\upsilon)\right)}{s_{10}(w(\upsilon))} \widetilde{f}_0(\upsilon)\left[-g_0(w(\upsilon)), 1\right] \begin{bmatrix} \partial s_1(w(\upsilon))/\partial t \\ \\ \partial s_2(w(\upsilon))/\partial t \end{bmatrix} d\upsilon$$

for $\upsilon = x$ anf $w(\upsilon) = (\alpha_0, \upsilon)$. Also, given our assumptions on $\tau\left(\cdot\right)$, f_0, $\widetilde{f}_0$ and g_0, $\phi(\upsilon)$ is bounded and cotinuous almost everywhere and zero outside a compact set Υ (where $\upsilon \in \Upsilon$). Thus, conditions (i) and (ii) in Assumption 5.1. in Newey (1994), which is needed for Lemma 5.3 in Newey (1994), are satisfied. Part (iii) in Newey's Assumption 5.1 and assumptions K,H, and Y in Newey's Lemma 5.3 are satisfied directly by our assumptions in Theorem 3. Finally, we verify the rate hypothesis in Newey's Lemma 5.3. Here, we need to show that $\sqrt{n}h^{1/2} \rightarrow \infty$ and $\sqrt{n}h^{r+3/2} \rightarrow 0$. For the first part note that $\sqrt{n}h^{1/2} = [nh]^{1/2} = \left[nh^{2k+3} \cdot h^{-2(k+1)}\right]^{1/2} \rightarrow \infty$, since $h^{-2(k+1)} \rightarrow \infty$ and by our assumption (vi) we have that $nh^{2k+3} \rightarrow \infty$. As for the second part, write $\sqrt{n}h^{r+3/2} = \left[nh^{2r+3}\right]^{1/2} \rightarrow 0$, since $nh^{2r+3} \rightarrow 0$ by assumption (vi).

Therefore, the conclusion of Lemma 5.3 implies that

$$\sqrt{n}h^{3/2}\left[m(\widehat{s}) - m(s_0)\right] \xrightarrow{d} \mathcal{N}(0, V) \quad \text{(B.16)}$$

where for $\widetilde{K}(u_1, \upsilon) = \int \partial K(u + [\partial t(\upsilon)/\partial \tau]\, \upsilon, \upsilon)/\partial u \, d\upsilon$ and $\Sigma(x) = E\,[qq'|x]$ we have

$$V = \int \phi(\upsilon) \left[\Sigma\left(w(\upsilon)\right) \quad \left\{ \int \widetilde{K}(u_1, \upsilon)\widetilde{K}(u_1, \upsilon)' \, du_1 \right\} \right] \phi(\upsilon)' f_0(w\left(\upsilon\right)) \, d\upsilon \tag{B.17}$$

We now derive the form of V for our case. First note that,

$$\Sigma(w(\upsilon)) = E\left[qq'|w(\upsilon)\right] = E\left[\left[\begin{array}{c} 1 \\ y \end{array} \right] \left[\begin{array}{cc} 1 & y \end{array} \right] \middle| w(\upsilon) \right] = \left[\begin{array}{cc} 1 & E\left[y|w(\upsilon)\right] \\ E\left[y|w(\upsilon)\right] & E\left[y^2|w(\upsilon)\right] \end{array} \right]$$

so that

$$\Sigma(w(\upsilon)) = \left[\begin{array}{cc} 1 & g_0\left(w(\upsilon)\right) \\ g_0\left(w(\upsilon)\right) & E\left[y^2|w(\upsilon)\right] \end{array} \right] \tag{B.18}$$

Also, note that given that in our case $\partial t(\upsilon)/\partial \tau = 0$, we have that

$$\widetilde{K}(u_1, \upsilon) = \int \frac{\partial K(u, \upsilon)}{\partial u} d\upsilon = \int K'(u, \upsilon) d\upsilon \tag{B.19}$$

where we have defined $K'(u, \upsilon) = \partial K(u, \upsilon)/\partial u$. Hence, plugging (B.15), (B.18) and (B.19) into (B.17) we find that

$$\begin{aligned}
V &= \int \frac{\tau\left(\upsilon\right) \widetilde{f}_0(\upsilon)}{s_{10}(w(\upsilon))} \left[-g_0(w(\upsilon)), 1\right] \left[\left[\begin{array}{cc} 1 & g_0\left(w(\upsilon)\right) \\ g_0\left(w(\upsilon)\right) & E\left[y^2|w(\upsilon)\right] \end{array} \right] \right. \\
&\quad \left. \left\{ \int \left[\int K'(u, \upsilon) d\upsilon \right]^2 du \right\} \right] \times \frac{\tau\left(\upsilon\right) \widetilde{f}_0(\upsilon)}{s_{10}(w(\upsilon))} \left[\begin{array}{c} -g_0(w(\upsilon)) \\ 1 \end{array} \right] f_0(w\left(\upsilon\right)) \, d\upsilon \\
&= \left\{ \int \left[\int K'(u, \upsilon) d\upsilon \right]^2 du \right\} \int \frac{\tau^2\left(\upsilon\right) \widetilde{f}_0^2(\upsilon)}{f_0(w(\upsilon))} \left[-g_0(w(\upsilon)), 1\right] \\
&\quad \left[\begin{array}{cc} 1 & g_0\left(w(\upsilon)\right) \\ g_0\left(w(\upsilon)\right) & E\left[y^2|w(\upsilon)\right] \end{array} \right] \left[\begin{array}{c} -g_0(w(\upsilon)) \\ 1 \end{array} \right] d\upsilon \\
&= \left\{ \int \left[\int K'(u, \upsilon) d\upsilon \right]^2 du \right\} \int \frac{\tau^2\left(\upsilon\right) \widetilde{f}_0^2(\upsilon)}{f_0(w(\upsilon))} \left[E\left[y^2|w(\upsilon)\right] - \left\{g_0\left(w(\upsilon)\right)\right\}^2 \right] d\upsilon \\
&= \left\{ \int \left[\int K'(u, \upsilon) d\upsilon \right]^2 du \right\} \int \frac{\tau^2\left(\upsilon\right) \widetilde{f}_0^2(\upsilon)}{f_0(w(\upsilon))} Var\left[y|w = (\alpha_0, \upsilon)\right] \, d\upsilon
\end{aligned} \tag{B.20}$$

where in the second equality we use the fact that $f_0(w) = s_{10}(w)$ by definition, and in the last equality we use the definition of $w(v)$ and the fact that $Var[y|w] = E[y^2|w] - \{E[y|w]\}^2$.

Given the previuos results, we now show asymptotic normality of $\sqrt{nh^3}\widehat{m}_n(\alpha_0)$, where $\widehat{m}_n(\alpha_0) = \frac{1}{n}\sum_{i=1}^{n} m(z_i,\, \alpha_0, \widehat{s})$ (See (B.7)). Write

$$\begin{aligned}
\sqrt{n}h^{3/2}\widehat{m}_n(\alpha_0) &= \sqrt{n}h^{3/2}\frac{1}{n}\sum_{i=1}^{n} m(z_i,\alpha_0,\widehat{s}) - \sqrt{n}h^{3/2}\frac{1}{n}\sum_{i=1}^{n} m(z_i,\alpha_0,s_0) + \\
&\quad \sqrt{n}h^{3/2}\frac{1}{n}\sum_{i=1}^{n} m(z_i,\alpha_0,s_0) \\
&= \sqrt{n}h^{3/2}\frac{1}{n}\sum_{i=1}^{n}[m(z_i,\alpha_0,\widehat{s}) - m(z_i,\alpha_0,s_0)] + o_p(1) \\
&= \sqrt{n}h^{3/2}\,[m(\widehat{s}) - m(s_0)] + o_p(1) \xrightarrow{d} \mathcal{N}(0,V)
\end{aligned} \tag{B.21}$$

for V in (B.17). The second equality follows by noting that $h^{3/2}\sum_{i=1}^{n} m(z_i, \alpha_0, s_0)/\sqrt{n} \xrightarrow{p} 0$[2]. In the third equality we have used the result we derived using Newey's (1994) Lemma 5.4 (see (B.12)), and finally, to obtain asymptotic normality in the last line we have used the conclusion from Newey's Lemma 5.3 (see (B.16)).

Therefore, multiplying both sides of (B.7) by $\sqrt{n}h^{3/2}$and using our results in (B.9) and the last expression in (B.21) along with Slutzky's Theorem, we find that

$$\sqrt{n}h^{3/2}(\widehat{\alpha} - \alpha_0) = -\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial m(z_i,\alpha^*,\widehat{s})}{\partial\alpha}\right]^{-1}\sqrt{n}h^{3/2}\widehat{m}_n(\alpha_0) \xrightarrow{d} \mathcal{N}(0,\ \ ) \tag{B.22}$$

where

$$= M^{-2}V = \frac{1}{M^2}\left\{\int\left[\int K'(u,v)dv\right]^2 du\right\}\int\frac{\tau^2(v)\,\widetilde{f}_0^2(v)}{f_0(w(v))}Var\,[y|w=(\alpha_0,v)]\ dv \tag{B.23}$$

[2]In this case note that $m(z_i, \alpha_0, s_0)$ is only a function of z_i, since α_0 and s_0 are held fixed. Thus, we can write $\psi(z_i) = m(z_i,\alpha_0,s_0)$, and using an appropriate Central Limit Theorem we would find that $\sum_{i=1}^{n}\psi(z_i)/\sqrt{n} \xrightarrow{d} \mathcal{N}(E(\psi(z)),\Gamma)$, for some variance Γ. This, along with the fact that $h^{3/2} \to 0$ implies that $h^{3/2}\sum_{i=1}^{n} m(z_i,\alpha_0,s_0)/\sqrt{n} \xrightarrow{p} 0$

and $M = \partial^2 E\left[\tau^2(x) E(Y|T = \alpha_0, X = x)\right]/\partial t^2$. ∎

Now, we prove the joint asymptotic normality result in Theorem 4. Let $\widehat{\alpha}$ be as in (B.3) and let the estimator of the size be given by

$$\widehat{E}\{Y(\alpha_0)\} = \widehat{\mu}_{h_2}(\widehat{\alpha}) = \frac{1}{n}\sum_{i=1}^{n}\tau(x_i)\ \widehat{g}_{h_2}(\widehat{\alpha}, x_i) \tag{B.24}$$

where $\widehat{g}_{h_2}(\widehat{\alpha}, x_i)$ is the Nadayara-Watson estimator based on bandwidth h_2. To simplify notation, let $E\{Y(t)\} = \mu_0(t)$. Then, the result in Theorem 4 states that:

$$\begin{pmatrix} \sqrt{nh_1^3}(\widehat{\alpha} - \alpha_0) \\ \sqrt{nh_2}\left(\widehat{\mu}_{h_2}(\widehat{\alpha}) - \mu_0(\alpha_0)\right) \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} V_1 & 0 \\ 0 & V_2 \end{bmatrix}\right) \tag{B.25}$$

with V_1 as in theorem 3 and $V_2 = \left[\int \left\{\int K(u,v)dv\right\}^2 du\right] \times \int f_0(\alpha_0, x)^{-1}\tau^2(x)\, \widehat{f}_0^2(x)\sigma^2(\alpha_0, x)$ dx; and where $K(w)$ is partitioned according to $w = [t, x]$ and $K^{(1)}(\cdot)$ means the partial derivative with respect to t.

Proof of Theorem 4.. According to the Cramér-Wold device we need to show that for every real numbers λ_1 and λ_2 we have:

$$\lambda_1\sqrt{nh_1^3}(\widehat{\alpha} - \alpha_0) + \lambda_2\sqrt{nh_2}\left(\widehat{\mu}_{h_2}(\widehat{\alpha}) - \mu_0(\alpha_0)\right) \xrightarrow{d} \mathcal{N}\left(0, \lambda_1^2 V_1 + \lambda_2^2 V_2\right) \tag{B.26}$$

where V_1 and V_2 are as above.

As in the proof of Theorem 1, first we show that $\sqrt{nh_2}\left(\widehat{\mu}_{h_2}(\widehat{\alpha}) - \mu_0(\alpha_0)\right)$ is asymptotically equivalent to $\sqrt{nh_2}\left(\widehat{\mu}_{h_2}(\alpha_0) - \mu_0(\alpha_0)\right)$, so that we could focus on the asymptotic distribution of $\lambda_1\sqrt{nh_1^3}(\widehat{\alpha} - \alpha_0) + \lambda_2\sqrt{nh_2}\left(\widehat{\mu}_{h_2}(\alpha_0) - \mu_0(\alpha_0)\right)$.

Using the mean value theorem we can write $\widehat{\mu}_{h_2}(\widehat{\alpha}) = \widehat{\mu}_2(\alpha_0) + \widehat{\mu}_2^{(1)}(\alpha^*)(\widehat{\alpha} - \alpha_0)$, for some α^* between $\widehat{\alpha}$ and α_0. Subtracting $\mu_0(\alpha_0)$ from both sides and multiplying by

$\sqrt{nh_2}$ we have:

$$\sqrt{nh_2}\left(\widehat{\mu}_2(\widehat{\alpha}) - \mu_0(\alpha_0)\right) = \sqrt{nh_2}\left(\widehat{\mu}_2(\alpha_0) - \mu_0(\alpha_0)\right) + \sqrt{nh_2}\widehat{\mu}_2^{(1)}(\alpha^*)(\widehat{\alpha} - \alpha_0) \qquad \text{(B.27)}$$

Hence, we need to show that the second term to the right of (B.27) is $o_p(1)$. Again, for a suitable mean value α^{**} between α_0 and α^* we can write $\widehat{\mu}_2^{(1)}(\alpha^*) = \widehat{\mu}_2^{(1)}(\alpha_0) + \widehat{\mu}_2^{(2)}(\alpha^{**})(\alpha^* - \alpha_0)$. Thus we have that:

$$\begin{aligned} \sqrt{nh_2}\widehat{\mu}_2^{(1)}(\alpha^*)(\widehat{\alpha} - \alpha_0) &= \frac{1}{\sqrt{nh_1^3 h_2^2}}\sqrt{nh_2^3}\widehat{\mu}_2^{(1)}(\alpha_0)\sqrt{nh_1^3}(\widehat{\alpha} - \alpha_0) \\ &\quad + \frac{1}{\sqrt{nh_1^6}}\sqrt{h_2}\widehat{\mu}_2^{(2)}(\alpha^{**})\sqrt{nh_1^3}(\alpha^* - \alpha_0)\sqrt{nh_1^3}(\widehat{\alpha} - \alpha_0) \end{aligned} \qquad \text{(B.28)}$$

Consider the first term to the right side of (B.28). Note that from Theorem 3 we know that $\sqrt{nh_1^3}(\widehat{\alpha} - \alpha_0) = O_p(1)$. Also, by following the same steps as in Theorem 3 is straightforward to show that given our conditions $\sqrt{nh_2^3}\widehat{\mu}_2^{(1)}(\alpha_0) = O_p(1)$ (see, for example, (B.21)). Specifically, the conditions on h_2 for the latter result require that $nh_2^{2k+3}/\left[\ln(n)\right]^2 \longrightarrow \infty$ and $nh_2^{2r+3} \longrightarrow 0$. The first condition is directly assumed in Theorem 4, and the second one follows from $nh_2^{2r+1} \longrightarrow 0$. Finally, given our assumptions on h_1 and h_2 we have that

$$\frac{1}{\sqrt{nh_1^3 h_2^2}} = \frac{1}{\sqrt{\sqrt{nh_1^6}\sqrt{nh_2^4}}} \longrightarrow 0$$

Thus, the first term to the right of (B.28) is $o_p(1)$. As for the second term to the right of (B.28), note that since $|\alpha^* - \alpha_0| \leq |\widehat{\alpha} - \alpha_0|$, then $\sqrt{nh_1^3}(\alpha^* - \widehat{\alpha}) = O_p(1)$. Now, consider the term $\sqrt{h_2}\widehat{\mu}_2^{(2)}(\alpha^{**})$. Write

$$\sqrt{h_2}\left|\widehat{\mu}_2^{(2)}(\alpha^{**})\right| \leq \sqrt{h_2}\sup_t\left|\widehat{\mu}_2^{(2)}(t)\right| \leq \sqrt{h_2}\sup_t\left|\widehat{\mu}_2^{(2)}(t) - \mu_0^{(2)}(t)\right| + \sqrt{h_2}\sup_t\left|\mu_0^{(2)}(t)\right|$$

Given our assumptions, we know that $\sqrt{h_2}\ \sup_t\ \left|g_0^{(2)}(t)\right| = o(1)$. As discussed in the proof of Theorem 3, the assumptions from Lemma 5.1 in Newey (1994) are satisfied.

Then, we have that $\sup_t \left|\widehat{\mu}_2^{(2)}(t) - \mu_0^{(2)}(t)\right| = O_p\left(\left(\ln(n)/nh_2^{k+5}\right)^{-1/2}\right)$. Thus, we have $\sqrt{h_2}\left|\sqrt{h_2}\widehat{\mu}_2^{(2)}(\alpha^{**})\right| \leq O_p\left(\left(\ln(n)/nh_2^{k+4}\right)^{-1/2}\right)$. Given our assumptions on h_2 and h_1, we have that $\sqrt{h_2}\left|\sqrt{h_2}\widehat{\mu}_2^{(2)}(\alpha^{**})\right| \xrightarrow{p} 0$ and $1/nh_1^6 \to 0$. Therefore, the second term to the right of (B.27) is $o_p(1)$, which implies that $\sqrt{nh_2}\,(\widehat{\mu}_2(\widehat{\alpha}) - \mu_0(\alpha_0))$ and $\sqrt{nh_2}\,(\widehat{\mu}_2(\alpha_0) - \mu_0(\alpha_0))$ are asymptotically equivalent. Hence, we focus on the asymptotic distribution of $\lambda_1\sqrt{nh_1^3}(\widehat{\alpha}-\alpha_0) + \lambda_2\sqrt{nh_2}\left(\widehat{\mu}_{h_2}(\widehat{\alpha}) - \mu_0(\alpha_0)\right)$. As in the proof of Theorem 3, the first step to show asymptotic normality of $\lambda_1\sqrt{nh_1^3}(\widehat{\alpha} - \alpha_0) + \lambda_2\sqrt{nh_2}\left(\widehat{\mu}_{h_2}(\widehat{\alpha}) - \mu_0(\alpha_0)\right)$ involves writing this expression as a linear function of a kernel estimator. Let $\widehat{s}^* = [\widehat{s}_1^* \ \ \widehat{s}_2^*]'$, with $\widehat{s}_1^*(w) = \frac{1}{n}\sum_{j=1}^{n} K_h^*(w-w_j)$ and $\widehat{s}_2^*(w) = \frac{1}{n}\sum_{j=1}^{n} y_j K_h^*(w-w_j)$; where we define $K_h^*(u) = -\frac{\lambda_1}{M\, h_1^k\sqrt{h_1}}K^{(1)}(u/h) + \frac{\lambda_2}{h_2^k\sqrt{h_2}}K(u/h)$ and, as in Theorem 3, we let $w = [t\ x]$ and $M = E\left[\tau(x)\,\partial^2 g_0(\alpha_0, x)/\partial^2 t\right]$. Following similar steps as in the proof of Theorem 3, and using the same notation, we find that we can write

$$\lambda_1\sqrt{nh_1^3}(\widehat{\alpha} - \alpha_0) + \lambda_2\sqrt{nh_2}\left(\widehat{\mu}_{h_2}(\widehat{\alpha}) - \mu_0(\alpha_0)\right) = \sqrt{n}\left[m(\widehat{s}^*) - m(s_0)\right] + o_p(1) \quad \text{(B.29)}$$

where $m(s)$ is as defined in (B.14). The next step involves showing asymptotic normality of the first term to the right of (B.29). Following similar steps as in the proofs of Theorems 1 and 3, and using Lemma 5.3 in Newey (1994), we obtain that

$$\sqrt{n}\left[m(\widehat{s}^*) - m(s_0)\right] \xrightarrow{d} \mathcal{N}\left(0, \lambda_1^2 V_1 + \lambda_2^2 V_2\right) \quad \text{(B.30)}$$

Therefore, the joint asymptotic normality result in (B.25) follows. ∎

Appendix C

Proofs: General Approach to Estimating Dose-Response Functions and their Maximum

Finally, we present the proof of Theorem 5, which is an identification result for ϕ_t^o, where $\phi(Y(t)) = \phi_t$ and there is a function $\psi(Y(t), \phi_t)$ such that $E[\psi(Y(t), \phi_t^o)] = 0$. This proof follows similar steps as those found in Hirano, Imbens and Ridder (2003) and Firpo (2002) regarding estimation of average and quantile treatment effects, respectively, in the binary-treatment case.

Proof of Theorem 5.

$$\begin{aligned} & E\left[\frac{\omega(T,X) \cdot \psi(Y, \phi_t^o)}{E[\omega(T,X)|X]}\right] \\ = \; & E\left[\frac{1}{E[\omega(T,X)|X]} E[\omega(T,X) \cdot \psi(Y, \phi_t^o)|X = x]\right] \end{aligned}$$

$$
\begin{aligned}
&= E\left[\frac{1}{E\left[\omega\left(T,X\right)|X\right]}E\left[\omega\left(T,X\right)E\left\{\psi\left(Y,\phi_t^o\right)|T=t,X=x\right\}|X=x\right]\right] \\
&= E\left[\frac{1}{E\left[\omega\left(T,X\right)|X\right]}E\left[\omega\left(T,X\right)E\left\{\psi\left(Y\left(t\right),\phi_t^o\right)|T=t,X=x\right\}|X=x\right]\right] \\
&= E\left[\frac{1}{E\left[\omega\left(T,X\right)|X\right]}E\left[\omega\left(T,X\right)E\left\{\psi\left(Y\left(t\right),\phi_t^o\right)|X=x\right\}|X=x\right]\right] \\
&= E\left[E\left\{\psi\left(Y\left(t\right),\phi_t^o\right)|X=x\right\}\right] \\
&= E\left[\psi\left(Y\left(t\right),\phi_t^o\right)\right] = 0
\end{aligned}
$$

The first and second equalities use iterated expectations. The fourth line uses the unconfoundedness assumption, and the last one uses the definition of $\psi\left(Y\left(t\right),\phi_t\right)$. ■