Continuous Treatments

Stefan Hoderlein[*] Yuya Sasaki

Boston College Brown University

First Draft: July 15, 2009
This Draft: January 19, 2011

Abstract

This paper explores the relationship between nonseparable models and treatment effect models when the causal variable of interest is endogenous. Like the treatment effect literature, our aim is to place no structure on the outcome equation, and establish necessary and sufficient conditions on the first stage equation for point identification of the causal effect of interest. We focus on the case of continuous causal variables and continuous instruments. It is the latter fact that makes our approach particularly close to the marginal treatment effect (MTE). We establish that the equality between MTE and LIV ceases to hold in the continuous case, and argue that the LIV does not identify a policy relevant object. Selection type structures in contrast identify the MTE and generalizations that do not suffer from limited support conditions. Moreover, we establish that these concepts extend naturally to quantile regression in the continuous treatment case, and that they allow to identify economically meaningful quantities without requiring that the structural unobservable be scalar or enter monotonically. Finally, we apply all concepts to analyze the nonlinear heterogeneous effects of smoking during pregnancy on infant birth weight.

Keywords: Marginal Treatment Effect, Nonseparable Model, Endogeneity, Quantiles.

[*] Stefan Hoderlein: Boston College, Department of Economics, 140 Commonwealth Avenue, Chestnut Hill, MA 02467, USA, email: stefan_hoderlein@yahoo.com. Yuya Sasaki: Brown University, Department of Economics, Box B, Providence, RI 02912, USA, email: yuya.sasaki@gmail.com. We have benefited from discussions and comments by Alberto Abadie, Andrew Chesher, Guido Imbens, Arthur Lewbel, Rosa Matzkin, Whitney Newey, Ed Vytlacil and seminar participants of the 2010 Cowles Foundation conference on Endogenous Nonseparable Models for comments and helpful suggestions.

1 Introduction

Unobserved heterogeneity in preferences and other complex but not directly observable objects arises naturally if microeconomic models are taken to the data. The influence of these unobservable objects on observed relationships are potentially large; yet they are still not completely understood. Recently, econometricians have proposed a plethora of methods to deal with this issue. In structural models for instance, heterogeneity is often modeled through random parameters. These approaches carry the risk of misspecification, but are useful if welfare analysis and other counterfactual exercises be performed that require structure for extrapolation.

However, often times interest centers only on determining the effect of one variable of interest, denoted $X \in \mathcal{X} \subseteq \mathbb{R}$ on an outcome $Y \in \mathcal{Y} \subseteq \mathbb{R}$, keeping other observable and unobservable determinants, denoted $S \in \mathcal{S} \subseteq \mathbb{R}^K$ and $A \in \mathcal{A} \subseteq \mathbb{R}^\infty$, constant. This is usually done in a linear regression framework but, as the binary treatment effect literature has pointed out, this generally leads to a bias in the presence of complex unobserved heterogeneity. This paper proposes a general framework to perform such standard analysis in the case of continuous X.

To allow for sufficient generality, we follow the recent econometric literature by modeling the relationship of interest through a nonseparable model, i.e., we let

$$Y = \phi(X, S, A), \tag{1.1}$$

where ϕ is smooth in x. Throughout this paper, we think of X as a continuous variable the individual chooses as part of a second economic decision which involves observable exogenous factors (instruments), denoted Z, and unobservable factors, denoted V. Logically, this second decision is chosen in a first stage (henceforth abbreviated FS), because there is no effect of Y on this decision, i.e., there is no simultaneity. In this paper, we are concerned with the effect that changes in this first step choice X have on the outcome of interest Y.

If X were binary, we can think of course of X as a treatment and of the first stage decision as whether the individuals select into treatment. Suppressing the dependence on S, in the binary X case we can without loss of generality rewrite the model to be a linear random coefficient model, i.e.

$$Y = \alpha(A) + \beta(A)X. \tag{1.2}$$

The object of interest is $\beta(A) = \phi(1, A) - \phi(0, A)$, usually denoted $Y_1 - Y_0$. In the absence of any structure on the complex unobservable A, this object is not identified. Instead, the aim of the binary treatment literature is to identify average effects. Specifically, interest centers on

$$\mathbb{E}\left[\beta(A)|\mathcal{F}\right] = \mathbb{E}\left[Y_1 - Y_0|\mathcal{F}\right] = \mathbb{E}\left[\phi(1, A) - \phi(0, A)|\mathcal{F}\right],$$

where typical choices of the conditioning set $\mathcal{F}$ include the following: 1. the trivial sigma algebra (i.e., $\mathcal{F} = \{\emptyset, \Omega\}$), in which case we obtain the average treatment effect (ATE), $\mathbb{E}\left[\phi(1, A) - \phi(0, A)\right]$. 2. $\mathcal{F} = \sigma(X)$, and $X = 1$, in which case we obtain the average treatment on the treated (ATT), $\mathbb{E}\left[\phi(1, A) - \phi(0, A) | X = 1\right]$. In addition, in the case of endogeneity, the selection into treatment has to be taken into account, and hence the quantities typically considered involve information from the FS equation as well. The discussion has largely focused on the following two information sets: First, $\mathcal{F} = \sigma(V)$, where V is the (scalar) first stage unobservable, and Z are continuous instruments. This yields the marginal treatment effect (MTE), which is the effect of treatment on the subpopulation at the margin of selecting into treatment, see Heckman and Vytlacil (2007) for an overview. Second, the LATE framework of Imbens and Angrist (1994), which focuses on the subpopulation of compliers, i.e., the subpopulation that selects into treatment as (usually binary) instruments Z change from z to z'[1].

In this paper, we focus mainly on the case where X is continuously distributed as has been common in the literature on nonseparable models (e.g., Angrist, Grady and Imbens (2000), Chesher (2003), Florens, Heckman, Meghir and Vytlacil (2008), Hoderlein and Mammen (2007), Imbens and Newey (2009), Matzkin (1994, 2003, 2007)), but we also aim at clarifying parallels and differences between the binary and the continuous treatment cases. Throughout this paper we assume that Z is continuous, as is plausible if the endogenous regressor X is continuous. Also, we do not want to make the assumption of monotonicity of the FS equation in Z. Both factors make our approach more directly comparable to MTE than to LATE in the binary treatment case. Therefore, we largely compare our results to those in Heckman and Vytlacil (2007).

If X is continuous, we may view the value of the regressor X as a chosen level of intensity of treatment, e.g., the choice of duration of participation in a training program, total length of schooling, the amount of nicotine or drug intake, the price of a good, etc. The first observation we make is that a linear random coefficient model as in equation (1.2) is restrictive once we move beyond the binary case. Our parameter of interest is a natural generalization of the regression setting: It is $\partial_x \phi(x, a)$, the partial effect of a marginal change in x, which we denote

$$\beta(x, a) = \partial_x \phi(x, a),$$

to emphasize the parallels to the random coefficients case. One can think of this quantity as the policy experiment of changing x to $x + h$, for small h. In this experiment, $\beta(x, a)h$ represents the (approximate) magnitude of the implied change in Y; hence our focus on $\beta(x, a)$.

Due to the high dimensionality of A, like in the binary treatment effect literature $\beta(x, a)$ is not identified. Therefore, we focus again on mean causal effects of a treatment, but now on a

[1] There is an ongoing debate about the policy relevance of these concepts; this paper does not contribute or comment on this issue.

treatment whose intensity the individual chooses, i.e.,

$$\mathbb{E}\left[\beta(X,A)|\mathcal{F}\right]$$

where averaging takes place over either the entire population or certain subpopulations of interests. Like in the binary treatment effect literature one can consider several choices of $\mathcal{F}$. The overall average partial effect (APE), $\mathbb{E}\left[\beta(X,A)\right]$, as proposed by Chamberlain (1984) is an obvious counterpart to the ATE. Moreover, instead of the ATT, it is natural to condition on the level of treatment intensity $X = x$, and obtain the average structural marginal effect of a treatment for individuals with treatment intensity $X = x$. We emphasize at this point that $X = x$ is kept fixed, as it characterizes the subpopulation, while $\beta(X,A)$ is the effect of interest. This is in perfect parallel to the binary treatment effect literature, where the (two) subpopulations kept fixed are treated and untreated, and the effect $\beta(X,A)$ corresponds to $Y_1 - Y_0$.

In this paper, we will allow for treatment to be endogenously determined, and we focus in particular on the identification of averages that we deem relevant as building blocks for policy analysis. Since the MTE is a key quantity in the binary treatment case with continuous instruments, we will be concerned with conditional averages that involve information about first stage preference parameters like V. However, we argue that a generalization of the MTE which we call local average structural derivative (LASD) is preferable, as it requires less stringent specification assumption on the first stage unobservables, and does in particular not suffer from effective restrictions on the support. We elaborate on this point in the application.

Contributions: In this wider picture our specific contributions are tied to specific questions that arise naturally. The first question is whether important insights from the binary treatment effects literature carry over to the continuous case. To this end, we consider the central identification equation in the binary setup with continuous instruments,

$$\mathbb{E}\left[Y_1 - Y_0|V = p\right] = \partial_p \mathbb{E}\left[Y|P = p\right],$$

which identifies the structural parameter of interest (the MTE) with the local instrumental variable (LIV), i.e., the derivative of the regression of Y on the probability to be treated, $P = p(Z)$ (Heckman and Vytlacil (1999, 2005, 2007, henceforth HV)). We show that this identity of MTE and LIV ceases to hold in the continuous case. As such, an interesting gap opens up between discrete and continuous treatments. The next question that arises naturally is the following: If the LIV does not identify the MTE any longer, what does it identify? We establish that it identifies a weighted average of MTEs, and argue that the subpopulation it defines are economically rather not interesting[2].

[2] Safe for the fact that they may be conveniently aggregated to compute the APE - in particular if p is known to have a parametric structure - and do not require "identification at infinity" for aggregation.

A natural next step is then to consider identification of economically meaningful quantities in the continuous case. We will argue that this includes both averages with respect to the regressor of interest, as well as the MTE. As it turns out, all of these quantities can be derived from a general identification theorem, which clarifies identification of average marginal effects conditional on X and Z by mean regression tools. This result can be seen as s generalization of the Heckman (1979) selection principle, and is generally related to contributions by Altonji and Matzkin (2005), Florens et al (2008), and Imbens and Newey (2009) in the literature on nonseparable models. We establish the minimal (i.e., necessary and sufficient) conditions we require for point identification in this setup, and show that identification of the MTE actually requires some additional assumptions on the first stage structure. We would like to emphasize the character of our approach as exploring the frontier of point identification.

Given these mean regressions results, we proceed to a largely open question: How can distributional information, as for instance summarized in all the conditional quantiles, be employed to identify parameters of interest in a population with high dimensional unobserved heterogeneity? To answer this question, we make use of results from the literature on nonseparable models, in particular Hoderlein and Mammen (2007, HM, for short), which links derivatives of quantile regressions to averages in the exogenous case, as well as Imbens and Newey (2009) who lay out the control function approach to identification. Based on these results, we pursue the question whether we can identify conditional averages of the effect of interest which involve the dependent variable and preference parameters, i.e., can we identify

$$\mathbb{E}\left[\beta(X,A)|V,Y\right] \quad \text{and} \quad \mathbb{E}\left[\beta(X,A)|X,V,Y\right],$$

and if yes, by which device? The answer we give for the continuous endogenous treatment case is affirmative. This is important because it establishes that quantile based structures can identify interesting structural effects even in cases where the unobservable is more complex than a scalar random variable that is identical between treatment and control group, as in Chernozhukov and Hansen (2005). The main identification results resemble the case without the dependent variable in the conditioning set, safe for the fact that we have replaced conditional expectations by conditional quantiles as identifying device. Moreover, we show that the same necessary and sufficient condition on the structure of the FS equation required for point identification of average effects in the mean regression case remains necessary and sufficient in the quantile case. An important conclusion we draw from this result is that the quantity of interest together with the model structure results in an identifying equation with a fairly similar structure regardless of whether one uses mean regression or quantile regression as means of identification.

Relationship to the Literature: This work aims at integrating two different strands of literatures: The literature on binary treatment effects, and the literature on nonseparable models. As already discussed, close in terms of the mathematical structure in the former literature is in particular the work of Heckman and Vytlacil (1999, 2005, 2007). Our approach is also related to the LATE framework of Imbens and Angist (1994), in particular to Angrist, Grady and Imbens (2000) and Hirano and Imbens (2004), though to a lesser degree: Unlike Imbens and Angrist (1994), we do not assume monotonicity of the FS equation in Z. Also, Abadie et al (2002) consider identifaction of quantiles under similar assumptions as Imbens and Angrist (1994); the same remark applies. Moreover, instead of a simultaneous equation model as in Angrist, Grady and Imbens (2000) we consider a triangular structure, and unlike Hirano and Imbens (2004), we consider IV type independence assumption and not exogeneity conditional on covariates S. Related is also the work of Nekipelov (2010), who considers a LATE setup with a discrete, but not necessarily binary, treatment variable.

The literature on nonseparable models is generally related. We are rather closely related to models that do not assume monotonicity in a scalar unobservable in the outcome equation, and allow for continuous endogenous regressors. This is in particular Altonji and Matzkin (2005), Florens, Heckman, Meghir, and Vytlacil (2008), Imbens and Newey (2009) and Hoderlein and Mammen (2007). The first reference in this list focuses largely on panel data, but introduces control functions to obtain APEs. Chesher (2003) and Imbens and Newey (2009) establish the similarity of the control function correction if one conditions on (X, V) and a combination of derivatives with respect to instruments and endogenous regressors in models where one conditions on (X, Z). This constitutes the main overlap between our work, in particular theorem 2, and this literature, and we show this link formally in theorems 4 and 8. But even confined to theorem 2, our work goes beyond these papers as we allow for multidimensional first stage error V. From Hoderlein and Mammen (2007, 2009), we adopt the notion that quantile derivatives identify marginal effects.

Less closely related to our work is Firpo, Fortin, and Lemieux (2009), who show how counterfactual distributional shifts reveal local marginal effects. Chesher (2003) and Jun (2009) consider similar right hand side structures as in theorem 5, however, they assume triangularity as well as monotonicity in the outcome equation. The remainder of our paper is generally novel for the nonseparable models literature. In particular, the various versions of the policy relevant MTE and LASD are novel, as is the interpretation of LIV in the continuous case.

Other recent work that exploits distributional assumptions includes Jun, Pinkse and Xu (2010) who propose a quantile treatment structure. Chernozhukov and Hansen (2005) propose estimation of a quantile IV model based on moment restrictions that assume that the individuals have either the same unobservable in both treatment and the control group, or assume a stationarity assumption on the distribution of unobservables that is difficult to motivate economically. Abadie et al. (2002) consider the effect of the treatment on the quantile, but do not relate it to complex unobservables. Finally, Chernozhukov, Imbens and Newey (2007) base their estimator for a nonseparable models with a scalar unobservable under endogeneity on a moment restriction. None of these papers relate quantile structures to causal effects if the unobservables are multivariate.

The approach of Florens, Heckman, Meghir, and Vytlacil (2008) is also related: Florens at al (2008) consider continuous endogenous regressors, but unlike us impose structure on the outcome equation and do not impose structure on the selection equations. As such, their model is different than ours, and more closely related to random coefficient models. In a similar vein is Heckman and Vytlacil (1998), who also consider the random coefficient model and provided restrictions on (1.2) to identify $\mathbb{E}[\beta(A)]$. An older reference is the work of Garen (1984), but this work is parametric in nature.

In our application, we have some overlap with Evans and Ringel (1999), whose essential economic argument for the validity of the instruments we follow. However, we extend their work in the several dimension that are in the focus of our paper: In our nonparametric setup, we consider policy relevant effects with high dimensional unobservables, using both mean regression and quantile regression tools. More widely related are Rosenzweig and Shultz (1983), and Lien and Evans (2005), as detailed below in the application.

Organization of the Paper: The second section discusses in detail the first two contributions sketched in the introduction. It outlines the model and some basic assumptions, and then discusses the identification of average structural marginal effects. We first establish what LIV, i.e., conditioning on Z in a mean regression, identifies in this setup. Then we consider the case where we condition on X and Z, which can be seen as generalization of the Heckman (1979) selection approach, introducing the concepts of MTE and LASD. In the third section we are concerned with extending all concepts to quantiles. Finally, to illustrate our results, we apply all concepts to data from health economics. In particular, we consider the effect of smoking on the birth weight of a child. A summary and an outlook conclude.

2 Mean Regression Based Identification of Causal Effects

In this section we discuss identification of average structural effects using mean regression tools. In the first subsection, we consider possible extensions of the HV concepts of LIV and MTE to the continuous case and show the sense in which both differ in this setup. Then we argue that generalizations of selection models indeed identify economically important objects. We start, however, our discussion with detailing the model we consider.

2.1 Basic Elements of the Model

Throughout this paper, we assume to observe variables (Y, X, Z), where Y is the outcome of interest, X is the endogenous FS choice that causally affects Y, and Z is a set of instruments (exogenous variables) that causally affect X. In our application, Y is birth weight of a child, X is nicotine intake, and Z denotes exogenous factors that determine this intake, e.g., the tax rate on tobacco. In labor economics, Y could be log wages, X total duration of schooling, and Z exogenous cost factors that affect school duration. In consumer demand, Y could be the quantity of fish consumed, X could be the own price, and Z supply side instruments, e.g., the maritime weather. In all of these examples, the exogenous factors drive X, but they are excluded from affecting Y through any other channel than through X, and they are also unrelated to any observable factor. The following assumptions formalizes this relationship between all variables:

Assumption 1 (DGP)**.** *Let $(\Omega, \mathcal{F}, P)$ be a complete probability space on which are defined the random vectors $(Y, X, Z, A, V) : \Omega \to \mathcal{Y} \times \mathcal{X} \times \mathcal{Z} \times \mathcal{A} \times \mathcal{V}$, $\mathcal{Y} \subseteq \mathbb{R}, \mathcal{X} \subseteq \mathbb{R}, \mathcal{Z} \subseteq \mathbb{R}^L$, $\mathcal{A} \subseteq \mathbb{R}^N, \mathcal{V} \subseteq \mathbb{R}^M$, where $L < \infty$ and $N, M \leqslant \infty$. The causal model is defined by*

$$\begin{aligned} Y &= \phi(X, A), \\ X &= \mu(Z, V), \end{aligned}$$

where $\phi : \mathcal{X} \times \mathcal{A} \to \mathcal{Y}$ and $\mu : \mathcal{Z} \times \mathcal{V} \to \mathcal{X}$ are Borel measurable function, and realizations of (Y, X, Z) are observable whereas those of (A, V) are not.

While unobservables A and V are potentially infinite-dimensional, some of the subsequent assumptions indirectly limit the dimensionality to $N, M < \infty$. Note that the outcome equation ϕ specifies Y to be a function of treatment choice X and an unobserved individual characteristic term A. The first-stage (FS) equation involves a causal function μ that summarizes the choice of treatment intensity given the exogenous factor Z. Note that an agent (or many agents on a market) may take the unobserved characteristics A into account when making this treatment intensity decision. This implies that V is generally correlated with A causing X to be endogenous. Second, like A, V is in general not likely to be a scalar, and monotonicity of μ in V is difficult to assume generally. Finally, observe that in this setup one can consider exogenous discrete or continuous covariates S entering both ϕ and μ, but we suppress them for ease of notation, and simply note that virtually all of the following analysis can be done by simply conditioning on S. While S could be either continuous or discrete, we will largely focus on the case where all other variables are continuous. More precisely:

Assumption 2 (Continuous Distribution)**.** *(i) All of the defined probability distributions (joint, marginal, and conditional) involving (Y, X, Z, A, V) are absolutely continuous with respect to*

Lebesgue measure, with bounded probability densities. (ii) Furthermore, these probability densities are continuously differentiable.

As already outlined, out of these variables it is in particular the exogenous instrumental variables Z which is essential in order to achieve identification. The following assumption clarifies the sense in which they are exogenous:

Assumption 3 (Instrument Independence). *(i) $A \perp\!\!\!\perp Z \mid V$, (ii) $V \perp\!\!\!\perp Z$.*

If we combine both parts of this assumption, we obtain $(A, V) \perp\!\!\!\perp Z$, i.e., joint independence of the instruments from all unobservables in the system holds. This is a little bit stronger than, but similar to, traditional treatment effects assumptions. E.g., consider the typical assumption that $(Y_1, Y_0, V) \perp\!\!\!\perp Z$, in the setup where X is binary, and $X = \mathbf{1}\{p(Z) > V\}$, see e.g., Heckman and Vytlacil (2007). Since $Y_j = \phi(j, A)$ in this setup, $(Y_1, Y_0, V) = (\phi(1, A), \phi(0, A), V)$, and obviously $(A, V) \perp\!\!\!\perp Z$ implies the weaker conditions $(Y_1, Y_0, V) \perp\!\!\!\perp Z$. Note that if ϕ is assumed to be strictly monotonic in A the conditions are equivalent. However, the continuous case allows to also weaken the full joint independence. While $A \perp\!\!\!\perp Z \mid V$ is retained throughout this paper, in some of our models we may in particular weaken $V \perp\!\!\!\perp Z$ to a location restriction like $\mathbb{E}[V|Z] = 0$.

As already mentioned above, in some cases we may also use regression quantiles, in both the first stage and the outcome equation.To this end, we denote the quantile regression of X on Z by $q^{\theta}_{X|Z}(z)$, i.e., $q^{\theta}_{X|Z}(z) := \inf\{x \in \mathcal{X} \mid F_{X|Z}(x \mid z) \geqslant \theta\}$. In addition, we invoke the following regularity assumptions.

Assumption 4 (Basic Regularity). *(i) ϕ is differentiable in x for every $(x, a) \in \mathcal{X} \times \mathcal{A}$, with continuous and bounded derivatives. (ii) $q^{\theta}_{X|Z}$ is continuously differentiable in z for every $(z, \theta) \in \mathcal{Z} \times (0, 1)$, with nonzero derivative (iii) $\mathbb{E}[X \mid Z = z]$ is continuously differentiable in z for every $z \in \mathcal{Z}$, with nonzero derivative. (iv) The derivative of every function to be integrated throughout the paper is dominated in absolute value by an L^1 function.*

Assumption 2 specifies all random variables to be continuously distributed. Thus, the quantiles are strictly increasing in their ranks, a property which we will use in order for $q^{\theta}_{X|Z}$ to be useful as identification device (see the auxiliary lemmas in the appendix section on this point). Assumption 4 contains otherwise standard differentiability and boundedness conditions.

Finally, to make the parallel to the binary case clear, we will give results for the special case of additively separability of the FS equation.

Assumption 5 (Separable FS). *$\mu(Z, V) = P + V$ where $P := \pi(Z)$ for a bounded function π.*

Observe that at this stage we do not specify what π is. The leading case is of course $\pi(z) = \mathbb{E}[X \mid Z = z]$, i.e., π is the mean regression of X on Z, in which case V is the mean regression residual, but π could also denote, say, the conditional median.

2.2 Projection on Z

This subsection starts by considering the extension of local IV to continuous treatment variables. We define the LIV to be $[\partial_z \mathbb{E}[X|Z=z]]^{-1} \partial_z \mathbb{E}[Y|Z=z]$. Does this quantity identify an interesting structural marginal effect at least among the subpopulation of individuals characterized by $Z=z$? The answer is given in the following theorem:

Theorem 1 (LIV). *(i) Suppose that Assumptions 1, 2, 3, and 4 hold with $L=1$. Also, assume that $\mathbb{E}[X|Z=z]$ and $\mathbb{E}[Y|Z=z]$ are differentiable with $\partial_z \mathbb{E}[X|Z=z] \neq 0$. Then,*

$$\mathbb{E}[\beta(X,A)|Z=z] = \frac{1}{\partial_z \mathbb{E}[X|Z=z]} \left\{ \partial_z \mathbb{E}[Y|Z=z] - B_0(z) \right\},$$

where the bias $B_0(z) = Cov(\beta(X,A);\ \partial_z \mu(Z,V) \mid Z=z)$.
(ii) Suppose that Assumptions 1, 2, 3, 4 and 5 hold with $L \geqslant 1$. Then

$$\mathbb{E}[\beta(X,A) \mid P=p] = \partial_p \mathbb{E}[Y \mid P=p].$$

(iii) Suppose that Assumptions 1, 2, 3 (i), 4 and 5 hold with $L \geqslant 1$. Then

$$\mathbb{E}[\beta(X,A) \mid P=p] = \partial_p \mathbb{E}[Y \mid P=p] - \mathbb{E}[Y \partial_p \log f_{V|P}(V \mid P) \mid P=p].$$

Part (i) states that LIV does in general not identify $\mathbb{E}[\beta(X,A)|Z=z]$: however, it does so iff $Cov(\beta(X,A); \partial_z \mu(Z,V)|Z=z) = 0$. Note that this bias-correcting covariance term is unobservable and cannot be used to correct the LIV. However, this covariance term obviously vanishes if either ϕ or μ are additively separable in the influence of the respective unobservable, or there is no endogeneity (i.e., $A \perp\!\!\!\perp V$). Since choosing an additive specification in general entails the risk of misspecification, if not suggested differently by economic theory one may perhaps be willing to rather impose structure on the FS equation. From this result it is obvious, however, that even if this additive structure is correct, it does not yield the equality of MTE and LIV that make the HV approach so appealing. To see this, suppose the additively separable FS has the form of a mean regression, which has the consequence that the LIV can be written as $\partial_p \mathbb{E}[Y \mid \pi(Z)=p]$, i.e., it is the derivative of the mean regression of Y on the "fitted value" P. While this closely resembles the quantity proposed by HV in the binary treatment case, it identifies, however, something different: $\mathbb{E}[\beta(X,A)|P=p]$, which is, as easily established, a conditional average over conditional expectations with respect to V (MTEs), i.e. $\mathbb{E}[\beta(X,A)|P=p] = \mathbb{E}[\mathbb{E}[\beta(X,A)|V,P]|P=p] = \mathbb{E}[\mathbb{E}[\beta(X,A)|V]|P=p]$, for the population for which $P=p$.

There is another difference between the binary and continuous case: In the latter case, we may under certain assumptions be able to solve for V and can thus identify $f_{V|P}$. This allows to weaken the independence assumptions required in the FS equation - instead of full independence we may actually weaken the assumption on the relationship between V and Z to, e.g., mean or median independence of V from Z. The flip side is that one specification has to be chosen, and there is only weak guidance from economics on this. In summary, our results suggests that conditioning on Z or P alone may not be as sensible in the continuous case as it is in the binary case.

What are arguments for nevertheless considering this quantity? In the case of several instruments, analogously to the treatment effect literature one may assume a certain parametric form for $p(z)$, say $p\,(z, \theta)$, where θ is a finite parameter. In this case, the nonparametric regression $\mathbb{E}[Y \mid P = p]$ is of lower dimension than $\mathbb{E}[Y \mid Z = z]$, and hence may suffer less from dimensionality issues. This may be an issue, if interest centers on ways to ultimately estimate $\mathbb{E}\,[\beta(X, A)]$. Else, there are few arguments that support considering this economically rather meaningless quantity, and we continue to proceed to economically more sensible objects.

2.3 Projection on (X, Z)

Given the conclusion in the previous subsection, a natural direction is to consider conditioning on X and Z (or P), which can be viewed as generalizations of Heckman (1979) selection. The question that remains to be answered is which quantities are sensibly considered, and by which objects they can be identified. This subsection will show that invertibility of the FS is a sufficient condition for identification of the MTE through a selection type quantity. We would like to stress, however, that invertibility of the FS equation is a rather stringent assumption. Thus, before we present a result on the identification of the MTE we consider how far we can relax this invertible FS condition and still point identify the local average structural derivative (LASD), the structural derivative projected on $\mathcal{F} = \sigma(X, Z)$, which, as we argue below, also constitutes an interesting object.

To understand the role of the first stage structure, we introduce a key assumption. As mentioned above, this assumption defines the frontier of point identification of a structural object.

Assumption 6 (Z Invariance). *$F_{X|Z}(\mu(z, v) \mid z) = F_{X|Z}(\mu(z', v) \mid z')$ holds for all $z, z' \in \mathcal{Z}$ and $v \in \mathcal{V}$*

This assumption, though substantial, is nonetheless weaker than index invertibility, i.e., the mapping $\mathit{index}(v) \mid z \mapsto x$ need *not* be for a given z under this assumption. It does not require restrictions on the dimensionality of V. To understand what this assumption encompasses, note that it nests index invertibility, which is of course weaker than index monotonicity, which in turn is weaker than monotonicity. As such, we can think of this assumption as the "necessary part" in the widely employed monotonicity assumption. In particular, this assumption turns out to be the weakest possible (i.e., necessary as well as sufficient) assumption to obtain point identification of the LASD via structure that involves a first stage quantile regression. Due to the fact that we consider classes of functions, there are two different notions of necessity of a condition C one could consider: The first notion of necessity is "weak", since it is logically equivalent to the statement that there exist "some generic" ϕ for which the bias does not vanish without satisfying the condition C. The second one is "strong" in the sense that condition C is required to vanish the bias for every single ϕ; in the absence of condition C the bias will not be zero. We mean by the necessity of Assumption 6 the weak notion. Assumption 6 generally cannot be necessary in the strong notion; given the averaging character and the complex structure involving high dimensional unobservables, it seems impossible to derive an universal condition that has to hold to exactly vanish the bias for every function ϕ, and without which the bias would never vanish. Observe that neither monotonic FS nor invertible FS are necessary in even the weak notion, but Assumption 6 is.

Equipped with this notion, we obtain the following result:

Theorem 2 (LASD). *(i) Suppose that Assumptions 1, 2, 3, and 4 hold with $L = 1$ and $M < \infty$. Then,*

$$\mathbb{E}\left[\beta(X,A)|X=x,Z=z\right] = \partial_x \mathbb{E}\left[Y|X=x,Z=z\right] + c\partial_z \mathbb{E}\left[Y|X=x,Z=z\right] - B(c,x,z)$$

holds for any $c \in \mathbb{R}$, where

$$B(c,x,z) \quad = \quad \mathbb{E}\left[Y\left\{c\partial_z \log f_{V|XZ}(V|x,z) + \partial_x \log f_{V|XZ}(V|x,z)\right\}|X=x,Z=z\right].$$

(ii) If in addition $\theta = F_{X|Z}(x \mid z)$ and $\partial_z q^{\theta}_{X|Z}(z) \neq 0$, then Assumption 6 is sufficient to make $B([\partial_z q^{\theta}_{X|Z}(z)]^{-1}, q^{\theta}_{X|Z}(z), z) = 0$ for all admissible[3] structural models $(\phi, F_{A|V})$.
(iii) Assumption 6 is also necessary to make $B([\partial_z q^{\theta}_{X|Z}(z)]^{-1}, q^{\theta}_{X|Z}(z), z) = 0$ for all admissible structural models $(\phi, F_{A|V})$.

In this theorem, we provide sufficient and necessary conditions for point identification of the parameter of interest, $\mathbb{E}\left[\beta(X,A)|X=x,Z=z\right]$. This effect is of interest, because it allows to identify average infinitisimal treatment effects, conditional on a subpopulation characterized by both X and Z. In our application for instance, we may obtain the average effect of a marginal unit, a cigarette, on the birth weight of the child for any value of the number of cigarettes and any level of the exogenous tax rate. Since, for a fixed value of cigarettes x, a higher value of z makes smoking more expensive, we can conclude indirectly that the individual has a revealed preference for smoking. But this preference may well be correlated with an unhealthy life style,

[3] By 'admissible' we mean the class of functions ϕ delineated by the imposed assumptions, which in this case are Assumptions 2, and 4 (i).

which is one of the unobserved factors in A. Hence different values of the exogenous factor Z may at least partially reflect differences in the high dimensional unobservables V or A.

To understand the role of our assumptions, observe that with our baseline assumptions only - in particular without the Z invariance - there is a bias term B. This emphasizes that as in the LIV case, point identification is not available in this setup in general. As is established in parts (ii) and (iii), it is exactly the above assumption that vanishes the bias. It states that the Z invariance assumption is necessary and sufficient to vanish B with the choice of $c = [\partial_z q^{\theta}_{X|Z}(z)]^{-1}$ under presence of reasonable endogeneity. This establishes the exact degree to which monotonicity in the FS can be weakened, if point identification using the conditional quantile of X given $Z = z$ is to be used.

We summarize the empirically relevant point identification result as Corollary:

Corollary 1 (LASD)**.** *Suppose that Assumptions 1, 2, 3, 4, and 6 hold with $L = 1$ and $M < \infty$. Then*

$$\mathbb{E}\left[\beta(X, A)|X = q^{\theta}_{X|Z}(z), Z = z\right] = \partial_x \mathbb{E}\left[Y|X = x, Z = z\right] + [\partial_z q^{\theta}_{X|Z}(z)]^{-1} \partial_z \mathbb{E}\left[Y|X = x, Z = z\right].$$

The object identified here parallels the treatment effect on the treated as in Florens, Heckman, Meghir, and Vytlacil (2008), however, observe the additional conditioning variable Z. There are also similarities with Chesher (2003) as well as Imbens and Newey (2009). We elaborate on this in detail when we consider identification via regression quantiles in section 3.

2.4 Projection on V

Corollary 1 suggests an object that we can identify under non-monotonic FS and outcome equation. However, the identified object $\mathbb{E}\left[\beta(X, A)|X = x, Z = z\right]$ in general fails to be the MTE, whereas it is certainly again a weighted average of the MTE. Under what condition does it coincide with something resembling a MTE? The key assumption is invertibility, the ability to recover V from observable (X, Z). Specifically, we formalize this notion in two ways: First, in unrestrictive form, second, in an "index sufficiency" form:

Assumption 7 (Invertibility)**.** *There exists a continuously differentiable function $\nu : \mathcal{X} \times \mathcal{Z} \to \mathcal{V}$ such that $v = \nu(\mu(z, v), z)$ and $x = \mu(z, \nu(x, z))$ for all $(x, z, v) \in \mathcal{X} \times \mathcal{Z} \times \mathcal{V}$ (i.e., this ν is the inverse of μ in V which is assumed to be smooth). Furthermore, $\partial_z \nu(x, z) \neq 0$.*

Assumption 8 (Invertibility)**.** *Let $\mathcal{P} \subseteq \mathbb{R}$. There exist functions $\pi : \mathcal{Z} \to \mathcal{P}$ and $\zeta : \mathcal{X} \times \mathcal{P} \to \mathcal{V}$, such that ζ is continuously differentiable and $\nu(X, Z) = \zeta(X, P)$ where $P = \pi(Z)$. Furthermore, $\partial_p \zeta(x, p) \neq 0$ for $p = \pi(z)$.*

While the former assumption provides a condition to obtain V as a function of X and Z, the latter specifies conditions under which we can obtain V as a function of X and P. Obviously, the model specified by Assumption 5 is a special instance of the latter invertibility assumption.

Theorem 3 (MTE)**.** *(i) Suppose that Assumptions 1, 2, 3, 4, and 7 hold with $L = 1$. Then*

$$\mathbb{E}\left[\beta(x, A)|V = \nu(x, z)\right] = \partial_x\mathbb{E}\left[Y|X = x, Z = z\right] - \rho_z(x, z)\partial_z\mathbb{E}\left[Y|X = x, Z = z\right],$$

where $\rho_z(x, z) = \partial_x\nu(x, z)/\partial_z\nu(x, z)$.
(ii) Suppose that Assumptions 1, 2, 3, 4, and 8 hold. Then

$$\mathbb{E}\left[\beta(x, A)|V = \zeta(x, p)\right] = \partial_x\mathbb{E}\left[Y|X = x, P = p\right] - \rho_p(x, p)\partial_p\mathbb{E}\left[Y|X = x, P = p\right],$$

where $\rho_p(x, p) = \partial_x\zeta(x, p)/\partial_p\zeta(x, p)$.

Because of the ρ_z and ρ_p terms, identification with the above result requires knowledge of the FS function in the sense that we have to know the precise independence condition that defines V. Under monotonicity and normalization, ρ_z and ρ_p can be constructed from the control function $F_{X|Z}$ as in Imbens and Newey (2009). In the case of the special additive FS structure displayed in assumption 5, the ρ_p term is exactly -1. In this case, Theorem 3 (ii) immediately yields the following.

Corollary 2 (MTE)**.** *Suppose that Assumptions 1, 2, 3, 4, and 5. Then*

$$\mathbb{E}\left[\beta(x, A)|V = x - p\right] = \partial_x\mathbb{E}\left[Y|X = x, P = p\right] + \partial_p\mathbb{E}\left[Y|X = x, P = p\right].$$

In analogy to the treatment case, we call this quantity the MTE, because we think of the subpopulation defined by a preference parameter V. Note that due to the continuous nature of X, there is no population who discretely jumps from treatment to non-treatment, and is at the margin of indifference. Instead, there is a constant, "smooth" adaptation, and to keep the individuals with same $V = v$ at the margin of indifference a marginal change in p would have to be compensated by a marginal change in x. Nevertheless, we consider this to be the closest continuous counterpart to the concept proposed by HV. Note that this identification does not rely on identification through LIV (Theorem 1).

The invertible FS as in Assumption 7 facilitates obtaining conditional expectations involving V via the control function (CF) approach as well. Under invertibility, assume that $V = F_{X|Z}(X \mid Z)$ without loss of generality. Then, the following equality can be derived from results in Imbens and Newey (2009).

Theorem 4 (CF)**.** *Suppose that Assumptions 1, 2, 3, 4, and 7 hold. Then*

$$\mathbb{E}[\beta(x, A) \mid V = v] = \mathbb{E}[\beta(X, A) \mid X = x, V = v] = \partial_x\mathbb{E}[Y \mid X = x, V = v].$$

3 Quantile Regression Based Identification of Causal Effects

The literature on nonseparable models has emphasized the importance of distributional features for the identification of causal effects, e.g., Matzkin (2003), Chesher (2003), Imbens and Newey (2009). These results rely crucially on scalar monotonicity of the unobservables, or/and a triangular structure. However, the treatment effect literature is concerned with allowing for causal models with several unobservables, each of which may not enter in any simple way. The question to be answered in any transfer from the binary treatment effect literature to the continuous case is therefore how to link high dimensional unobservables to distributional effects.

Hoderlein and Mammen (2007, HM, for short) provide such a link in the exogenous case by establishing that quantile derivatives are related to average marginal effects of the structural function, conditional on the dependent variable. Given the results in HM (2007), one may suspect that many results established above in the mean regression case extend to identification via quantiles. In this section we do precisely this: we construct averages with respect to a finer sigma algebra by expanding the conditioning set to include Y as well as X and Z. We use $q^\tau_{Y|XZ}(x,z) := \inf\{y \in \mathcal{Y} \mid F_{Y|XZ}(y \mid x, z) \geqslant \tau\}$, the quantile regression of Y on (X,Z), as a device for identification. Similarly defined are the quantile regression of Y on P denoted by $q^\tau_{Y|P}$, and the quantile regression of Y on (X,P), denoted by $q^\tau_{Y|XP}$. The regularity conditions for these quantiles are precisely stated as Assumptions 9, 10, and 11, in Section A.1 in the appendix. We follow the same structure as in the second section: we first consider projections involving only Z, and proceed to projections involving X and Z, but always add Y to these projections.

3.1 Projection on (Y,Z)

This section provides a quantile-extension of the local IV results derived above using mean regression tools. Evidently, the results of Section 2.2 extend in a very analogous fashion

Theorem 5 (Quantile LIV)**.** *Suppose that Assumptions 1, 2, 3, 4, 5, and 9 hold. Then,*

$$\mathbb{E}[\beta(X,A) \mid Y = q^\tau_{Y|P}(p), P = p] = \partial_p q^\tau_{Y|P}(p).$$

Observe the clear parallel to Theorem 1 (ii), making the advantages of this unified identification framework obvious[4]. The quantile LIV identifies an average marginal effect conditional on (Y,P). An important point of this result is the HM flavor of the conditioning set on the

[4] Indeed, we can produce analog result to the rest of theorem 1 but we desist because the additional insights do not warrant this.

left hand side. In our opinion, however, this added conditioning does not have any feature that removes the disadvantages associated with LIV in the mean regression case. In particular, the conditioning set is still incomplete in the sense that it does not allow to consider subpopulations defined by treatment intensity. As such we argue that unlike the binary choice case, LIVs do not identify structurally interesting quantities in the continuous case, whether one uses mean regressions or quantiles. Instead we argue now that quantile selection type structures identify interesting objects.

3.2 Projection on (Y, X, Z)

Given the obvious parallel between theorem 1 and theorem 5, we next attempt to generalize the results of section 2.3. Indeed, this is possible, again under roughly similar assumption. Recall, in particular assumption 6,

Assumption 6 (Z-Invariance)**.** $F_{X|Z}(\mu(z,v) \mid z) = F_{X|Z}(\mu(z',v) \mid z')$ holds for all $z, z' \in \mathcal{Z}$ and $v \in \mathcal{V}$

We provide now a quantile-extension to the results of Theorem 2:

Theorem 6 (Quantile LASD)**.** *(i) Suppose that Assumptions 1, 2, 3, 4, and 10 hold with* $L = 1$ *and* $M < \infty$*. If* $q^{\tau}_{Y|XZ}(x,z)$ *is on the support of* $f_{Y|XZ}(\cdot \mid x,z)$*, then*

$$\begin{aligned}\mathbb{E}[\beta(X,A) \mid Y = q^{\tau}_{Y|XZ}(x,z), X = x, Z = z] &= \partial_x q^{\tau}_{Y|XZ}(x,z) + c\partial_z q^{\tau}_{Y|XZ}(x,z) \\ &\quad + B(c,x,z).\end{aligned}$$

holds for any constant c, where

$$\begin{aligned}B(c,x,z) &= \mathbb{E}\left[\frac{1\{Y \leqslant q^{\tau}_{Y|XZ}(x,z)\}\partial_x \log f_{V|XZ}(V \mid x,z)\}}{f_{Y|X,Z}(q^{\tau}_{Y|XZ}(x,z) \mid x,z)}\middle| X = x, Z = z\right] \\ &\quad + c\mathbb{E}\left[\frac{1\{Y \leqslant q^{\tau}_{Y|XZ}(x,z)\}\partial_z \log f_{V|XZ}(V \mid x,z)}{f_{Y|X,Z}(q^{\tau}_{Y|XZ}(x,z) \mid x,z)}\middle| X = x, Z = z\right].\end{aligned}$$

(ii) If in addition $\theta = F_{X|Z}(x \mid z)$ *and* $\partial_z q^{\theta}_{X|Z}(z) \neq 0$*, then Assumption 6 is sufficient to make* $B([\partial_z q^{\theta}_{X|Z}(z)]^{-1}, q^{\theta}_{X|Z}(z), z) = 0$ *for all admissible structural models* $(\phi, F_{A|V})$*.*
(iii) Assumption 6 is also necessary to make $B([\partial_z q^{\theta}_{X|Z}(z)]^{-1}, q^{\theta}_{X|Z}(z), z) = 0$ *for all admissible structural models* $(\phi, F_{A|V})$*.*

Observe again the parallels to the mean regression case, in particular the structure involving the bias term, and exactly the same condition under which these bias term vanishes (as before, the critical condition for vanishing the bias is the Z-invariance assumption A 6). In the following, we focus again on the point identified implication of this result:

Corollary 3 (Quantile LASD)**.** *Suppose that Assumptions 1, 2, 3, 4, 6, and 10 hold with $L = 1$ and $M < \infty$. Then,*

$$\mathbb{E}[\beta(X, A) \mid Y = q^{\tau}_{Y|XZ}(q^{\theta}_{X|Z}(z), z), X = q^{\theta}_{X|Z}(z), Z = z] = \partial_x q^{\tau}_{Y|XZ}(x, z) + [\partial_z q^{\theta}_{X|Z}(z)]^{-1}\partial_z q^{\tau}_{Y|XZ}(x, z).$$

The left hand side is now the structural object of interest, an average structural marginal effect, for subpopulations defined by all three observable random variables, Y, X and Z. This is the finest sigma algebra that can be generated from the data, and allows to consider conditional averages of the heterogeneous structural effect that are most detailed, and as close to the true heterogeneous marginal effect as possible, given the available information. The same right hand side objects that identifies this object was analyzed in Chesher (2003) and Imbens and Newey (2009), but our result allows now for non-scalar unobservables in both the selection and the outcome equation and adds an economic interpretation for this case. Part (ii) of the theorem shows that Assumption 6 is a necessary and sufficient condition for point identification to hold for all admissible structural functions, provided the quantile of X given Z is to be used. Notice the apparent parallel between Corollaries 1 and 3. That is, this result is obtained by replacing the mean regression of Y on (X, Z) in Corollary 1 by the τ-th quantile regression of Y on (X, Z). Analogously, we can now consider the MTE as special case of this structure.

3.3 Projection on $(Y, V, \,\cdot\,)$

The MTE follows by quantile extensions of theorem 3 and corollary 2. Invertibility of the first stage is again the crucial restriction, and assumption which may be questionable in some applications:

Theorem 7 (Quantile MTE)**.** *(i) Suppose that Assumptions 1, 2, 3 (i), 4, 7, and 10 hold with $L = 1$. Then*

$$\begin{aligned}\mathbb{E}\left[\beta(X, A)|Y = q^{\tau}_{Y|XZ}(x, z), V = \nu(x, z), Z = z\right] &= \mathbb{E}\left[\beta(X, A)|Y = q^{\tau}_{Y|XZ}(x, z), X = x, Z = z\right] \\ &= \partial_x q^{\tau}_{Y|XZ}(x, z) - \rho_z(x, z)\partial_z q^{\tau}_{Y|XZ}(x, z),\end{aligned}$$

where $\rho_z(x, z) = \partial_x \nu(x, z)/\partial_z \nu(x, z)$.
(ii) Suppose that Assumptions 1, 2, 3 (i), 8 and 11 hold, then

$$\begin{aligned}\mathbb{E}\left[\beta(X, A)|Y = q^{\tau}_{Y|XP}(x, p), V = \zeta(x, p), P = p\right] &= \mathbb{E}\left[\beta(X, A)|Y = q^{\tau}_{Y|XP}(x, p), X = x, P = p\right] \\ &= \partial_x q^{\tau}_{Y|XP}(x, p) - \rho_p(x, p)\partial_p q^{\tau}_{Y|XP}(x, p),\end{aligned}$$

where $\rho_p(x, p) = \partial_x \zeta(x, p)/\partial_p \zeta(x, p)$.

Theorem 7 (ii) immediately yields the following in the case of additive first stage regressions:

Corollary 4 (Quantile MTE)**.** *Suppose that Assumptions 1, 2, 3 (i), 4, 5 and 11 hold. Then,*

$$\begin{aligned}\mathbb{E}\left[\beta(X,A) \mid Y = q^{\tau}_{Y|XP}(x,p), V = x-p, P = p\right] &= \mathbb{E}[\beta(X,A) \mid Y = q^{\tau}_{Y|XP}(x,p), X = x, P = p] \\ &= \partial_x q^{\tau}_{Y|XP}(x,p) + \partial_p q^{\tau}_{Y|XP}(x,p)\end{aligned}$$

holds.

Note the parallel between Corollaries 2 and 4, the only difference being that the mean regression has been replaced by the corresponding quantile regression. Again, a selection type structure identifies an average structural marginal effect. Note that the same remark about smooth adaptation applies here, but in addition the information in Y is used, which maps out information in A. Note, moreover, that information in P is also used now, because there are now complicated trivariate relationships, i.e., conditional on Y, the unobservables (A, V) are not independent from the instruments Z any more.

The projection on (Y, V) can also be identified via the control function (CF) approach, if one is willing to assume a invertible FS as in Assumption 7. Under this assumption, V plays the role of a control function on which X and A are independent, see again Imbens and Newey (2009). A combination of arguments in Imbens and Newey (2009) with the method of Hoderlein and Mammen (2007) yields a quantile extension of Theorem 4.

3.4 Summary of Identified Treatment Parameters

At this point, we summarize the main identified treatment parameters in Table 1. One issue that distinguishes LASD from LIV and MTE is that it circumvents assuming separability (Assumption 5) or invertibility (Assumption 7 or 8) in the first-stage. Identification of LASD therefore imposes milder structural restrictions than required for the identification of the other parameters. We also present two versions of the MTE, namely the MTE based on additive separability and the MTE* based on invertibility, where obviously the latter nests the former as a special case. This table clearly emphasizes something that has already been pointed out: the passage from mean regression based structures to the corresponding quantile regression structures q^{τ} very much implies adding Y as an additional conditioning variable to the respective conditioning set defining the subpopulations; this points to the deep common causal relationships specified by the same structural model and the respective independence assumptions.

4 Nonlinear Heterogeneous Effects of Smoking

Adverse effects of smoking during pregnancy on infant birth weights have been extensively studied in the health economic literature (e.g., Rosenzweig and Schultz (1983); Evans and

	Projection	Device	Statement	Assumptions	FS Structure
LIV	Z	Mean Regression	Theorem 1 (ii)	1, 2, 3, 4, 5	Additively Separable
	(Y, Z)	Quantile Regression	Theorem 5	1, 2, 3, 4, 5, 9	
LASD	(X, Z)	Mean Regression	Corollary 1	1, 2, 3, 4, 6	Nonseparable and Non-invertible
	(Y, X, Z)	Quantile Regression	Corollary 3	1, 2, 3, 4, 6, 10	
MTE*	V	Mean Regression	Theorem 3	1, 2, 3, 4, 8	Nonseparable but Invertible
	(Y, V, Z)	Quantile Regression	Theorem 7	1, 2, 3, 4, 8, 11	
MTE	V	Mean Regression	Corollary 2	1, 2, 3, 4, 5	Additively Separable
	(Y, V, Z)	Quantile Regression	Corollary 4	1, 2, 3, 4, 5, 11	

	Mean Regression	Quantile Regression
LIV	$\mathbb{E}[\beta(X, A) \mid P = p]$ $= \partial_p m_{Y\|P}(p)$	$\mathbb{E}[\beta(X, A) \mid Y = q^{\tau}_{Y\|P}(p), P = p]$ $= \partial_p q^{\tau}_{Y\|P}(p)$
LASD	$\mathbb{E}[\beta(X, A) \| X = q^{\theta}_{X\|Z}(z), Z = z]$ $= \partial_x m_{Y\|XZ}(x, z)$ $+[\partial_z q^{\theta}_{X\|Z}(z)]^{-1} \partial_z m_{Y\|XZ}(x, z)$	$\mathbb{E}[\beta(X, A) \mid Y = q^{\tau}_{Y\|XZ}(q^{\theta}_{X\|Z}(z), z), X = q^{\theta}_{X\|Z}(z), Z = z]$ $= \partial_x q^{\tau}_{Y\|XZ}(x, z)$ $+[\partial_z q^{\theta}_{X\|Z}(z)]^{-1} \partial_z q^{\tau}_{Y\|XZ}(x, z)$
MTE	$\mathbb{E}[\beta(x, A) \| V = x - p]$ $= \partial_x m_{Y\|XP}(x, p)$ $+\partial_p m_{Y\|XP}(x, p)$	$\mathbb{E}[\beta(X, A) \mid Y = q^{\tau}_{Y\|XP}(x, p), V = x - p, P = p]$ $= \partial_x q^{\tau}_{Y\|XP}(x, p)$ $+\partial_p q^{\tau}_{Y\|XP}(x, p)$

Table 1: Summary of required assumptions and identified treatment parameters. The symbol m denotes the mean regression, e.g., $m_{Y|X}(x) = \mathbb{E}[Y \mid X = x]$.

Variable	Mean	Std. Dev.	Description
Birth Weight	3330	606	Infant birth weight measured in grams
Cigarette	1.75	5.51	Number of cigarettes smoked per day
Tax	30.4	15.5	Excise tax rate on cigarettes in percentage
Age	26.7	6.0	Maternal age
Drinks	0.04	0.75	Number of times of drinking per week
Visits	11.3	4.1	Number of prenatal care visits
Births	1.97	1.00	Number of live births

Table 2: Descriptive statistics of the data.

Ringel (1999); Lien and Evans (2005)). Most papers, including those in the medical literature, have suggested that the effect of smoking (as binary variable) on infant birth weights ranges from -200 grams to -400 grams. Given that the average number of cigarettes smoked by smoking pregnant women is 12 between years 1989 and 1999, average effects of one cigarette on infant birth weight thus ranges from -17 grams to -33 grams. The goal of our analysis is to provide a much more detailed assessment of the effect of smoking, in particular we consider the heterogeneous marginal effects of a single cigarette as opposed to these coarse average effects.

Specifically, we analyze the effects of the number of cigarettes on infant birth weight, extending an older idea of Evans and Ringel (1999). We allow for arbitrary nonlinear, endogenous and heterogeneous effects of smoking, and want to obtain averages of causal marginal effects for various subpopulations defined by treatment intensity, as well as other variables that proxy for unobserved heterogeneity as detailed below. Evans and Ringel use cigarette excise tax rate as source of exogenous variation to mitigate confounding factors in identifying the effects of smoking. We follow this idea; in our framework tax rates hence play the role of Z, while number of cigarettes per day and infant birth weight are X and Y, respectively. The causal model is then given by

$$\begin{cases} Y = \phi(X, A, S) \\ X = \mu(Z, V, S) \end{cases}$$

where A captures other unobserved factors related to the lifestyle of the mother that impact the child's birth weight. Other observed characteristics of the mother, denoted S, are also controlled for, including maternal age, alcohol intake, number of prenatal visits, and number of live births experienced. We use a cross section of the natality data from the Natality Vital Statistics System of the National Center for Health Statistics. The main variables in the data are summarized in Table 2. From this data set we extract a random sample of size 100,000 from the time period between 1989 to 1999.

The structural features of interest are the average marginal effect of a cigarette, and we start out by considering the quantile LASD, i.e., using subpopulation defined by various combinations of values y, x, z of Y, X, Z. Formally, we estimate the left hand side in

$$\mathbb{E}[\beta(X,A) \mid Y = q^{\tau}_{Y|XZ}(q^{\theta}_{X|Z}(z), z), X = q^{\theta}_{X|Z}(z), Z = z] = \partial_x q^{\tau}_{Y|XZ}(x,z) + [\partial_z q^{\theta}_{X|Z}(z)]^{-1} \partial_z q^{\tau}_{Y|XZ}(x,z).$$

by replacing the components $\partial_x q^{\tau}_{Y|XZ}$ and $\partial_z q^{\tau}_{Y|XZ}$ by same counterparts estimators, specifically, the slope coefficients of a local quadratic estimator of $q^{\tau}_{Y|XZ}$, see Fan and Gijbels (1996). Similarly, the component $\partial_z q^{\theta}_{X|Z}$ is estimated by the slope of a local quadratic estimator of $q^{\theta}_{Y|Z}$. We have experimented with several bandwidth and we picked the ones we deemed most plausible, however, automated methods like cross validation produced roughly comparable results. We construct bootstrap confidence bands of the quantile LASD (Figures 1–6) by using bootstrap samples and forming the same estimator with an undersmoothed choice of bandwidth. Due to the point mass of the distribution of X at $X = 0$ which conflicts with Assumption 2 (i), our analysis focuses on the domain outside of this locality. With this framework in place, we make the following observations:

1. Comparing the graphs with lower Z (e.g., Figure 1) and higher Z (e.g., Figure 5), we observe ceteris paribus a great deal of heterogeneity in overall effects. In particular, the marginal effects under higher tax rates are relatively larger in magnitude. In other words, pregnant women who still choose to smoke despite facing higher tax rates exhibit larger marginal effects of smoking on infant birth weights. We will discuss this phenomenon in more detail below.

2. Comparing the marginal effects across X, we observe a common tendency for marginal effects to diminish towards $x = 20$ (e.g., Figures 2–6). That is, the negatively sloped structural function ϕ will eventually flatten on average as x increases. This phenomenon reflects the reduction in harm of an additional cigarette as the number of cigarettes increases. It is imperative to keep in mind, however, that a woman who smoked 20 cigarettes a day has already inflicted a large cumulative effect on her child; unsurprisingly, the remaining effect diminishes. In this light, one should rather emphasize that even at high levels of smoking there is still a significantly negative effect of an additional cigarette.

3. Comparing the graphs with different values of Y (e.g., Figures 1 and 2), we observe some differences in marginal effects across quantiles of Y, especially at lower tax rates $z = 30$. Marginal effects of smoking on birth weights tend to be smaller for lower quantiles of Y. This makes sense as it is more difficult to reduce a birth weight that is already low by the same absolute value (though a similar percentage reduction seems conceivable). These quantile differences are milder at higher tax rates $z \geqslant 40$. However, the differences in Y are not pronounced in this application, and as a consequence it may be justified to focus on the difference across (x, z) by integrating out Y, i.e., to consider the simpler mean regression based

LASDs of Section 2.3. These effects are illustrated in Figures 7–9, and they reinforce nicely the observations made in the first two points above.

Whether one employs mean or quantile regressions, it is instructive to examine the first point in more detail and provide likely causal explanations. As the graphs indicate, the magnitude of partial effects tends to be negatively related to Z for each fixed value of X. Suppose now that $z' > z$. The subpopulation who smokes x cigarettes when the taxes are z' is then characterized by a higher preference for smoking than the subpopulation that smokes x cigarettes at the lower price (tax) z. What causes endogeneity is now precisely the correlation between this preference for smoking and other factors in A, in particular adverse ones, say, a preference for an unhealthy lifestyle, and/or a partner who also smokes. The graphical results imply that the magnitude of partial effects tends to be positively related to higher taxes in excess of the effect already incurred through X, suggesting this revealed preference for a negative lifestyle as explanation. Moreover, it implies that the magnitude of partial effects tends to be positively related with unhealthy factors in A, other things fixed. Lastly, this implies

$$0 < \frac{\partial}{\partial a}\left|\frac{\partial}{\partial x}\phi(x,a)\right| = -\frac{\partial^2}{\partial a \partial x}\phi(x,a).$$

We therefore conclude that smoking X and other unhealthy behavioral inputs A are complementary negative inputs in the birth weight "production" function ϕ. So, based on our results, policy should not just discourage smoking, but also the negative and unhealthy life style associated with it that exacerbates its effect.

Finally, other than the LASD, we can also consider the MTE, i.e., $\mathbb{E}[\beta(x,A) \mid V = v]$ discussed in section 2.4, or its quantile version discussed in section 3.3. When estimating the MTE, we need to carefully delineate its effective domain in the first stage, i.e. $X = P + V$ (cf. Assumption 8 and Corollary 2). The left graph in Figure 10 shows a kernel density estimate of the empirical distribution of P for the entire sample. Most of the mass is concentrated around $1 \leqslant P \leqslant 2$, implying that the MTE can (only) be effectively computed across (x,v) for which $1 \leqslant x - v \leqslant 2$. Extrapolating the MTE outside of the domain where we do not have any observations generally requires additional and substantial assumptions. For this reason, we present the MTE only for a narrow domain of $v \in [x-2, x-1]$ given a fixed value of x. In this interval, heterogeneity in effects may be severely limited. This practical limitation is analogous to the narrow domain over which their common support assumption is satisfied in the application of Imbens and Newey (2009), and is significant drawback of the MTE.

The first stage with the added additive separability restriction is displayed in the left graph of Figure 11, which shows the cigarette excise tax as a discouraging factor for smoking. Figures 12 (a), (b), and (c) in the top row displays the MTEs across $v \in [3,4]$ for $x = 5$, across $v \in [8,9]$ for $x = 10$, and across $v \in [13,14]$ for $x = 15$, respectively. As may be anticipated, the MTEs within each of the three narrow domains fails to exhibit large variations. Note that v is the control variable for a residual measure of willingness to smoke, which may be positively correlated with other unhealthy lifestyles. According to our previous findings, larger v should be associated with larger marginal effects. On the other hand, the aforementioned domain restriction entails larger x for larger v within the effective domain. Because of this fact, mutually offsetting effects of x and v are likely to cause the apparent homogeneity across the three figures: larger marginal effects at lower x accompanied by smaller marginal effects at lower v (Figure 12 (a)) yield similar results compared to smaller marginal effects at higher x offset by larger marginal effects at higher v (Figure 12 (c)). This apparent homogeneity is an artifact of the support limitation that masks heterogeneous marginal effects across either dimension of x or v.

Since the entire sample contains a large share of non-smoking population who may possess distinct unobserved characteristics from smoking population, we repeat the estimation of the MTE focusing now on the subsample of smoking individuals only. The right graph of Figure 10 shows a kernel density estimate of the empirical distribution of P among smokers, which again exhibits a narrow support except now concentrating around the higher values $11 \leqslant P \leqslant 12.5$. Nonparametric first-stage estimates are presented in the right graph of Figure 11, and estimates of MTE are shown in the bottom row of Figure 12. The MTE estimates based on smokers are qualitatively similar to those based on the entire population.

The issue of narrow domain of the MTE implies that the control variable V is capable of producing results in a small portion of the population (in terms of unobserved heterogeneity) only. The size of this population is determined by the size of variations in mean X induced by Z, which is only between 1 and 2 in our application. Binary treatment models share a similar issue in that the instrumental strength determines the domain on which the MTE is identifiable, e.g., global identification of the HV MTE requires identification at infinity. The LASDs provides more information about heterogeneity as it does directly exploit the link between the support of X, Z and the unobservables. In addition to these limited information and narrow effective domain issues, the stronger structural assumption ($X = P+V$, Assumption 8) of the continuous-treatment MTE entails the risk of misspecification, whereas the LASD is less prone to misspecification. The upshot of this discussion is that the more general LASD is better suited than the MTE to identify economically interesting effects in a continuous treatment context.

5 Conclusions

This paper is concerned with providing a framework in which causal effects of continuous variables, or treatments, can be discussed in the presence of endogenous selection of the continuous treatment intensity. We aim at providing a framework that relates to the recent treatment effect literature, as in Imbens and Angrist (1994), and Heckman and Vytlacil (2007), in order to understand parallels and differences between this literature and our approach, which is closer related to nonseparable models. Since we are dealing with continuous endogenous variables we assume continuous instruments; it is in particular this feature that makes our results comparable to the work of HV, whose notion of preference conditioned average treatment effects (marginal treatment effect, MTE) we adopt.

We establish that the identity between MTE and local IV that serves as a key building block in the binary treatment effect literature with continuous instruments does not hold any longer. We clarify what Local IV identifies in the continuous treatment case, and argue that this object is not economically interesting. Instead, we argue that "selection type" structures identify the MTE and a more general object, the LASD (local average structural derivative) which we both consider to be more sensible on economic grounds. In both cases, we provide sufficient and necessary conditions for point identification, which allow for infinite dimensional unobservables in both the first stage, as well as the outcome equation.

The analysis thus far is largely based on mean regressions. However, we also consider quantile regression based identification. In particular, we show that the results mentioned are straightforward to extend to quantiles. We characterize what quantile local IV identifies, and argue that this is again in general not an interesting quantity. An alternative in the continuous case is easily devised, and given by the natural quantile generalizations of the selection objects considered in the second section. We establish that these quantities identify quantile MTE and quantile LASDs, and we explore the limits of point identification in this setup. We illustrate some of our contributions with an application to the effect of smoking on the birth weight of children, and find that a policy maker should not just target smoking, but also the unhealthy lifestyle that is associated with it.

References

Abadie, Alberto, Joshua Angrist, and Guido Imbens (2002) "Instrumental Variables Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings," *Econometrica*, Vol. 70, No. 1, pp. 91-117

Altonji, Joseph G. and Rosa L. Matzkin (2005) "Cross Section and Panel Data Estimators for Nonseparable Models with Endogenous Regressors," *Econometrica*, Vol. 73, No. 4, pp. 1053-1102

Chernozhukov, Victor and Christian Hansen (2005) "An IV Model of Quantile Treatment Effects," *Econometrica,* Vol. 73, No. 1, pp. 245-261.

Chernozhukov, Victor, Imbens, Guido, and Newey, Whitney, 2007. "Instrumental variable estimation of nonseparable models," Journal of Econometrics, Elsevier, vol. 139(1), pages 4-14, July

Chesher, Andrew (2003) "Identification in Nonseparable Models," *Econometrica*, Vol. 71, No. 5, pp. 1405-1441

Chesher, Andrew (2005) "Nonparametric Identification under Discrete Variation," *Econometrica*, Vol. 73, No. 5, pp. 1525-1550

Evans, William N. and Jeanne Ringel (1999) "Can Higher Cigarette Taxes Improve Birth Outcomes?" *Journal of Public Economics*, Vol. 72, No. 1, pp. 135–54.

Fan, Jianqing and Irène Gijbels (1996) "Local Polynomial Modelling and Its Applications" Chapman & Hall.

Firpo, Sergio, Nicole M. Fortin, and Thomas Lemieux (2009) "Unconditional Quantile Regressions," *Econometrica,* Vol. 77, No. 3, pp. 953-

Florens, J.P., J.J. Heckman, C. Meghir, and E. Vytlacil (2008) "Identification of Treatment Effects Using Control Functions in Models With Continuous, Endogenous Treatment and Heterogeneous Effects," *Econometrica*, Vol. 76, No. 5, pp. 1191-1206

Garen, John (1984) "The Returns to Schooling: A Selectivity Bias Approach with a Continuous Choice Variable," *Econometrica*, Vol. 52, No. 5, pp 1199-1218

Heckman, James J. (1979) "Sample Selection Bias as a Specification Error," *Econometrica,* Vol. 47, No. 1, pp. 153–161.

Heckman, James J. (1990) "Varieties of Selection Bias," *American Economic Review*, Vol. 80, No. 2, pp. 313–318.

Heckman, James J. and Edward Vytlacil (1998) "Instrumental Variables Methods for the Correlated Random Coefficient Model: Estimating the Average Rate of Return to Schooling When the Return is Correlated with Schooling," *The Journal of Human Resources,* Vol. 33, No. 4, pp. 974–987.

Heckman, James J. and Edward Vytlacil (1999) “Local Instrumental Variables and Latent Variable Models for Identifying and Bounding Treatment Effects,” *Proceedings of the National Academy of Science*, USA, Vol. 96, pp. 4730-4734

Heckman, James J. and Edward Vytlacil (2005) “Structural Equations, Treatment Effects, and Econometric Policy Evaluation,” *Econometrica*, Vol. 73, No. 3, pp. 669-738

Heckman, James J. and Edward Vytlacil (2007) “Econometric Evaluation of Social Programs, Part II: Using the Marginal Treatment Effect to Organize Alternative Econometric Estimators to Evaluate Social Programs, and to Forecast their Effects in New Environments” in J.J. Heckman and E.E. Leamer (ed.) *Handbook of Econometrics*, Vol. 6, Ch. 71.

Hirano, K., and G. W. Imbens (2004), ”The Propensity Score with Continuous Treatments”, Working Paper, University of California, Berkeley.

Hoderlein, Stefan and Enno Mammen (2007) “Identification of Marginal Effects in Nonseparable Models Without Monotonicity,” *Econometrica*, Vol. 75, No. 5, pp. 1513-1518

Imbens, Guido, W. and Joshua D. Angrist (1994) “Identification and Estimation of Local Average Treatment Effects,” *Econometrica*, Vol. 62, No. 2, pp. 467-475

Imbens, Guido W. and Whitney K. Newey (2009) “Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity” *Econometrica*, Vol. 77, No. 5, pp 1481-1512

Jun, Sung Jae (2009) “Local Structural Quantile Effects in a Model with a Nonseparable Control Variable,” *Journal of Econometrics,* Vol. 151, No. 1, pp. 82–97.

Jun, Sung Jae, Joris Pinkse, and Haiqing Xu (2010) “Tighter Bounds in Triangular Systems,” Working Paper

Lien, Diana S. and William N. Evans (2005) “Estimating the Impact of Large Cigarette Tax Hikes: The Case of Maternal Smoking and Infant Birth Weight.” *Journal of Human Resources*, Vol. 40, No. 2, pp. 373–392.

Matzkin, Rosa, L. (1994) “Restrictions of Economic Theory in Nonparametric Methods.” in Robert F. Engle and Daniel L. McFadden (ed.) *Handbook of Econometrics*, Vol. 4, Ch. 41

Matzkin, Rosa, L. (2003) “Nonparametric Estimation of Nonadditive Random Functions,” *Econometrica*, Vol. 71, No. 5, pp. 1339-1375

Matzkin, Rosa, L. (2007) "Nonparametric Identification" in J.J. Heckman and E.E. Leamer (ed.) *Handbook of Econometrics*, Vol. 6, Ch. 73.

Nekipelov, D. (2010), Identification and Efficient Estimation in a Class of Models with Discrete Endogenous Regressors, Working paper, Berkeley.

Rosenzweig, Mark R. and T. Paul Schultz (1983) "Estimating a Household Production Function: Heterogeneity, the Demand for Health Inputs, and Their Effects on Birth Weight," *Journal of Political Economy*, Vol. 91, No. 5, pp. 723–746 .

Vytlacil, Edward (2002) "Independence, Monotonicity, and Latent Index Models: An Equivalence Result," *Econometrica*, Vol. 70, No. 1, pp. 331-341

A Appendix

A.1 Technical Assumptions

This section lists basic regularity assumptions concerning three quantile regressions.

Assumption 9 (Regularity). *(i) $f_{Y|P}$ is continuous with respect to the conditioning variable P in a neighborhood of p for $Y = q^{\tau}_{Y|P}(p)$. (ii) $f_{Y,Y'|P}$ is uniformly bounded, has a bounded support, and is partially differentiable wrt Y in a neighborhood of $Y = q^{\tau}_{Y|P}(p)$. (iii) $q^{\tau}_{Y|P}$ is partially differentiable in a neighborhood of p. (iv) $\beta(X, A) \mid Y, P$ has a finite first moment at $(Y, P) = (q^{\tau}_{Y|P}(p), p)$.*

Assumption 10 (Regularity). *(i) $f_{Y|XZ}$ is continuous with respect to the conditioning variables (X, Z) in a neighborhood of (x, z) at $Y = q^{\tau}_{Y|XZ}(x, z)$. (ii) $f_{Y,Y'|X,Z}$ is uniformly bounded, has a bounded support, and is partially differentiable with respect to Y in a neighborhood of $Y = q^{\tau}_{Y|XZ}(x, z)$. (iii) $f_{Y'|YXZ}$ is uniformly bounded in a neighborhood of $X = x$ and is continuous in (Y', X). (iv) $q^{\tau}_{Y|XZ}$ is partially differentiable and $q^{\theta}_{X|Z}$ is differentiable. (v) $\beta(X, A) \mid Y, X, Z$ has a finite first moment at $(Y, X, Z) = (q^{\tau}_{Y|XZ}(x, z), x, z)$.*

Assumption 11 (Regularity). *(i) $f_{Y|XP}$ is continuous with respect to the conditioning variables (X, P) in a neighborhood of (x, p) at $Y = q^{\tau}_{Y|XP}(x, p)$. (ii) $f_{Y,Y'|X,P}$ is uniformly bounded, has a bounded support, and is partially differentiable wrt Y in a neighborhood of $Y = q^{\tau}_{Y|XP}(x, p)$. (iii) $q^{\tau}_{Y|XP}$ is partially differentiable in a neighborhood of (x, p). (iv) $\beta(X, A) \mid Y, X, P$ has a finite first moment at $(Y, X, P) = (q^{\tau}_{Y|XP}(x, p), x, p)$.*

A.2 Auxiliary Lemmas

Lemma 1. *Suppose that Assumptions 1, 2, 3 (ii), and 6 hold with $L = 1$. If $\partial_z q^\theta_{X|Z}(z)$ is defined and is nonzero, then for the choice of $c := [\partial_z q^\theta_{X|Z}(z)]^{-1}$,*

$$\partial_x \log f_{V|XZ}(v \mid q^\theta_{X|Z}(z), z) + c\partial_z \log f_{V|XZ}(v \mid q^\theta_{X|Z}(z), z) = 0$$

holds for all v on the support of $f_{V|XZ}(\cdot \mid q^\theta_{X|Z}(z), z)$.

Let Θ denote the random variable that provides the conditional rank of X given Z, i.e.,

$$\Theta = F_{X|Z}(\mu(Z, V) \mid Z).$$

Then, Assumption 6 implies that Θ does not depend on Z, thus $\Theta = g(V)$ for some function g. Since $(V, \Theta) = (V, g(V))$, we have $(V, \Theta) \perp\!\!\!\perp Z$ by Assumption 3 (ii). Using this independence restriction yields

$$f_{VZ|\Theta} = \frac{f_{V\Theta|Z}}{f_\Theta} f_Z = \frac{f_{V\Theta}}{f_\Theta} f_Z = f_{V|\Theta} f_Z = f_{V|\Theta} f_{Z|\Theta},$$

thus showing that $V \perp\!\!\!\perp Z \mid \Theta$.

Now, note that the map $(\theta, z) \mapsto (q^\theta_{X|Z}(z), z)$ is well-defined and injective, owing to the well-definition and injectivity of the map $\theta \mapsto q^\theta_{X|Z}(z)$ by Assumption 2 (note that the absolute continuity of the measure $P_{X|Z}$ in particular implies that there is no singular part in its Radon-Nikodym decomposition, thus the quantile is strictly increasing in θ). Therefore, we have $f_{V|XZ}(v \mid q^\theta_{X|Z}(z), z) = f_{V|\Theta Z}(v \mid \theta, z)$ for all z and θ in their respective domains. Finally, use the independence condition $V \perp\!\!\!\perp Z \mid \Theta$ obtained in the last paragraph to conclude that $f_{V|XZ}(v \mid q^\theta_{X|Z}(z), z) = f_{V|\Theta}(v \mid \theta)$, which is constant in z.

Since $f_{V|XZ}(v \mid q^\theta_{X|Z}(z), z)$ is constant in z, we have

$$\begin{aligned} 0 &= \frac{d}{dz} f_{V|XZ}(v \mid q^\theta_{X|Z}(z), z) \\ &= \partial_z q^\theta_{X|Z}(z) \cdot \partial_x f_{V|XZ}(v \mid q^\theta_{X|Z}(z), z) + \partial_z f_{V|XZ}(v \mid q^\theta_{X|Z}(z), z) \end{aligned}$$

by Assumption 2. Divide by $[\partial_z q^\theta_{X|Z}(z)] \cdot f_{V|XZ}(v \mid q^\theta_{X|Z}(z), z)$ to prove the lemma.

Lemma 2. *Suppose that Assumptions 1, 2, and 4 hold with $L = 1$ and $M < \infty$. If Assumption 6 does not hold, then there exists a set $\bar{\mathcal{V}} \subset \mathcal{V}$ of positive measure such that*

$$\frac{d}{dz} f_{V|XZ}(v \mid q^\theta_{X|Z}(z), z) \neq 0. \tag{A.1}$$

holds for some $z \in \mathcal{Z}$ and for all $v \in \bar{\mathcal{V}}$.

Let Θ denote the random variable for the conditional rank of X given Z, i.e.,

$$\Theta = F_{X|Z}(\mu(Z, V) \mid Z).$$

Write

$$H(z,v) = F_{X|Z}(\mu(z,v) \mid z).$$

Assumptions 2 (ii) and 4 (i) imply $H \in C^1(\mathbb{R}^{M+1}, \mathbb{R})$. As Assumption 6 does not hold, we have $\partial_z H(\bar{z}, \bar{v}) \neq 0$ for some $(\bar{z}, \bar{v}) \in \mathcal{Z} \times \mathcal{V}$. Let j be a nontrivial coordinate of V in μ. Then, by Assumption 2 (i), we have $\partial_{v_j} H(\bar{z}, \bar{v}) \neq 0$. Thus we have sufficient conditions to invoke the Implicit Function Theorem to obtain a continuous function $\lambda : \mathcal{Z} \supset \mathcal{B}_\varepsilon(\bar{z}) \to \mathcal{V}$ defined in a neighborhood of $\bar{z}$ such that

$$H(z, \lambda(z)) = H(\bar{z}, \bar{v}) =: \bar{\theta}$$

It follows that a continuum of the level set of $\Theta = \bar{\theta}$ exist in a neighborhood of $z = \bar{z}$. But this level set does not contain arbitrarily close horizontal or vertical displacements $(z \pm \varepsilon, v)$ and $(z, v \pm \varepsilon e_j)$, due to $\partial_z H(\bar{z}, \bar{v}) \neq 0$ and $\partial_{v_j} H(\bar{z}, \bar{v}) \neq 0$ as well as continuous differentiability of H. These two facts (i.e., existence of a continuum of the level set and no containment of arbitrarily close horizontal or vertical displacements) imply that $\text{supp}[(Z, V_j) \mid \Theta = \bar{\theta}] \neq \text{supp}[Z \mid \Theta = \bar{\theta}] \times \text{supp}[V_j \mid \Theta = \bar{\theta}]$, i.e., the support of $(Z, V_j) \mid \Theta = \bar{\theta}$ is not rectangular. Since a rectangular support of the joint distribution is a necessary condition for independence, this implies $Z \not\perp\!\!\!\perp V_j \mid \Theta = \bar{\theta}$, which in turn implies $Z \not\perp\!\!\!\perp V \mid \Theta = \bar{\theta}$.

Keeping the last result in mind, we now want to prove that (A.1) holds for some $z \in \mathcal{Z}$ and for all $v \in \bar{\mathcal{V}}$ with $\bar{\mathcal{V}}$ a set of positive measure. But by Assumptions 2 (ii) and 4 (ii), it suffices to show that (A.1) holds for some $(z, v) \in \mathcal{Z} \times \mathcal{V}$, since the continuity of the derivatives then yields the corresponding result throughout a neighborhood of such v. Suppose, by way of contradiction, that

$$\frac{d}{dz} f_{V|XZ}(v \mid q^\theta_{X|Z}(z), z) = 0$$

holds for all $(z, v) \in \mathcal{Z} \times \mathcal{V}$. As in the proof of Lemma 1, Assumptions 2 and 4 yield $f_{V|XZ}(v \mid q^\theta_{X|Z}(z), z) = f_{V|\Theta Z}(v \mid \theta, z)$. Hence,

$$\frac{d}{dz} f_{V|\Theta Z}(v \mid \theta, z) = 0$$

holds for all $(z, v) \in \mathcal{Z} \times \mathcal{V}$, showing that $V \perp\!\!\!\perp Z \mid \Theta$. This is a contradiction with the conclusion of the previous paragraph.

A.3 Proof of Theorem 1

(i): Compute

$$\begin{aligned} \mathbb{E}[Y|Z = z] &= \mathbb{E}[\phi(\mu(z, V), A)|Z = z] \\ &= \int\int \phi(\mu(z, v), a) f_{AV}(a, v) da dv, \end{aligned}$$

where the last equality is due to Assumption 3. Taking derivatives by Assumption 4 (iv) produces

$$\partial_z \mathbb{E}\left[Y|Z=z\right] = \int\int \beta(\mu(z,v),a)\partial_z\mu(z,v)f_{AV}(a,v)dadv,$$

But this is

$$\begin{aligned}\mathbb{E}\left[\beta(X,A)\partial_z\mu(Z,V)|Z=z\right] &= \mathbb{E}\left[\beta(X,A)|Z=z\right]\mathbb{E}\left[\partial_z\mu(Z,V)|Z=z\right]\\ &\quad +Cov\left(\beta(X,A),\partial_z\mu(Z,V)|Z=z\right).\end{aligned}$$

Using $\partial_z\mathbb{E}[\mu(Z,V)\mid Z=z]=\mathbb{E}[\partial_z\mu(Z,V)\mid Z=z]$ that follows from Assumption 4 (iv), and rearranging terms, we obtain the result.

(ii): If $\mu(Z,V)=\pi(Z)+V$, then $Cov\left(\beta(X,A),\partial_z\mu(Z,V)|Z\right)=Cov\left(\beta(X,A),\pi'(Z)|Z\right)=0$, hence the result follows.

(iii): First note that by Assumption 3 (i) we have $F_{AV|P}=F_{A|VP}F_{V|P}=F_{A|V}F_{V|P}$. Using this equality, conclude

$$\begin{aligned}\partial_p\mathbb{E}[Y\mid P=p] &= \partial_p\int\int\phi(p+v,a)f_{A|V}(a\mid v)f_{V|P}(v\mid p)dadv\\ &= \int\int\beta(p+v,a)f_{A|V}(a\mid v)f_{V|P}(v\mid p)dadv\\ &\quad +\int\int\phi(p+v,a)f_{A|V}(a\mid v)\partial_p f_{V|P}(v\mid p)dadv\\ &= \mathbb{E}[\beta(X,A)\mid P=p]+\mathbb{E}[Y\partial_p\log f_{V|P}(V\mid P)\mid P=p]\end{aligned}$$

as desired.

A.4 Proof of Theorem 2

We have

$$\begin{aligned}\mathbb{E}\left[Y|X=x,Z=z\right] &= \mathbb{E}\left[\phi(x,A)|X=x,Z=z\right]\\ &= \int\int\phi(x,a)f_{A|VXZ}(a\mid v,x,z)daf_{V|XZ}(v\mid x,z)dv\\ &= \int\int\phi(x,a)f_{A|V}(a\mid v)daf_{V|XZ}(v\mid x,z)dv,\end{aligned}$$

where the last equality is due to assumption 3 (i). Take derivatives to obtain

$$\begin{aligned}\partial_x\mathbb{E}\left[Y|X=x,Z=z\right] &= \int\int\beta(x,a)f_{A|V}(a\mid v)daf_{V|XZ}(v\mid x,z)dv\\ &\quad +\int\int\phi(x,a)f_{A|V}(a\mid v)da\partial_x f_{V|XZ}(v\mid x,z)dv\\ &= \mathbb{E}\left[\beta(x,A)|X=x,Z=z\right]\\ &\quad +\mathbb{E}\left[Y\partial_x\log f_{V|XZ}(V\mid x,z)|X=x,Z=z\right],\end{aligned}$$

where the first equality uses Assumptions 2 (ii), 3 (i), and 4 (iv). By similar arguments,

$$c\partial_z \mathbb{E}\left[Y|X=x,Z=z\right] \;=\; \mathbb{E}\left[Yc\partial_x \log f_{V|XZ}(V\mid x,z)|X=x,Z=z\right].$$

Combining these two inequalities yields

$$\mathbb{E}\left[\beta(X,A)|X=x,Z=z\right] = \partial_x \mathbb{E}\left[Y|X=x,Z=z\right] + c\partial_z \mathbb{E}\left[Y|X=x,Z=z\right] - B(c,x,z)$$

where

$$B(c,x,z) \;=\; \mathbb{E}\left[Y\left\{c\partial_z \log f_{V|XZ}(V|x,z) + \partial_x \log f_{V|XZ}(V|x,z)\right\}|X=x,Z=z\right].$$

This proves part (i) of the theorem. Apply Lemma 1 to prove part (ii).

Lastly, we prove part (iii) of the theorem by applying Lemma 2. We prove the contrapositive statement, that if Assumption 6 does not hold then there exists an admissible structure $(\phi, F_{A|V})$ such that $B([\partial_z q^{\theta}_{X|Z}(z)]^{-1}, q^{\theta}_{X|Z}(z), z) \neq 0$. By Lemma 2, there exists a set $\bar{\mathcal{V}} \subset \mathcal{V}$ of positive measure such that

$$\frac{d}{dz} f_{V|XZ}(v \mid q^{\theta}_{X|Z}(z), z) \neq 0.$$

holds for some $z \in \mathcal{Z}$ and for all $v \in \bar{\mathcal{V}}$. There exists a subset of $\bar{V}$ with positive measure on which the sign is positive or negative throughout. Without loss of generality, assume that the above expression is positive on $\bar{V}$. Pick a structure $(\phi, F_{A|V})$ such that $\int \phi(x,a) f_{A|V}(a\mid v)da$ is positive on $\bar{V}$ and zero outside $\bar{V}$ for some x, this is certainly an element of the space of admissible functions and distributions (e.g., if $Y(\omega) > 0$ for all $\omega \in \Omega_{\bar{V}}$, where $\Omega_{\bar{V}}$ is the set of states corresponding to $\bar{V}$, and $Y(\omega) = 0$ for all other ω). Then, we have

$$\begin{aligned}
& \partial_z q^{\theta}_{X|Z}(z) B([\partial_z q^{\theta}_{X|Z}(z)]^{-1}, q^{\theta}_{X|Z}(z), z) \\
= \;& \mathbb{E}\left[Y\left\{\partial_z \log f_{V|XZ}(V\mid x,z) + \partial_z q^{\theta}_{X|Z}(z) \cdot \partial x \log f_{V|XZ}(V\mid x,z)\right\} \mid X=x, Z=z\right] \\
= \;& \int \left[\int \phi(x,a) f_{A|V}(a\mid v)da\right]\left[\frac{d}{dz} f_{V|XZ}(v\mid q^{\theta}_{X|Z}(z), z)\right] dv \;>\; 0
\end{aligned}$$

This shows that $B([\partial_z q^{\theta}_{X|Z}(z)]^{-1}, q^{\theta}_{X|Z}(z), z) \neq 0$.

A.5 Proof of Theorem 3

(i): First, we have the equality

$$\begin{aligned}
\partial_x \mathbb{E}[Y \mid X=x, Z=z] \;&=\; \partial_x \int \phi(x,a) f_{A|XZ}(a\mid x,z)da \\
&\overset{(1)}{=}\; \partial_x \int \phi(x,a) f_{A|VZ}(a\mid \nu(x,z), z)da \\
&\overset{(2)}{=}\; \mathbb{E}[\beta(X,A) \mid V=\nu(x,z), Z=z] \\
&+\; \partial_x \nu(x,z) \mathbb{E}[\phi(X,A)\partial_v \log f_{A|VZ}(A\mid V,Z) \mid V=\nu(x,z), Z=z].
\end{aligned}$$

where step (1) is due to Assumption 7, and step (2) is due to Assumptions 2 (ii) and 4 (i) and (iv). Similarly, we have

$$\partial_z \mathbb{E}[Y \mid X = x, Z = z] = \partial_z \nu(x,z) \mathbb{E}[\phi(X,A) \partial_v \log f_{A|VZ}(A \mid V, Z) \mid V = \nu(x,z), Z = z],$$

where Assumptions 3 (i) was used to vanish $\partial_z f_{A|VZ}$. But then, Assumption 7 allows this to be rewritten as

$$\mathbb{E}[\phi(X,A) \partial_v \log f_{A|VZ}(A \mid V, Z) \mid V = \nu(x,z), Z = z] = \frac{1}{\partial_z \nu(x,z)} \partial_z \mathbb{E}[Y \mid X = x, Z = z],$$

which we can substitute in the first equation to get part (i) of the theorem.

(ii): The procedure is similar to the proof of part (i). First, we have the equality

$$\begin{aligned} \partial_x \mathbb{E}[Y \mid X = x, P = p] &= \partial_x \int \phi(x,a) f_{A|XP}(a \mid x, p) da \\ &\overset{(1)}{=} \partial_x \int \phi(x,a) f_{A|VP}(a \mid \zeta(x,p), p) da \\ &\overset{(2)}{=} \mathbb{E}[\beta(X,A) \mid V = \zeta(x,p), P = p] \\ &+ \partial_x \zeta(x,p) \mathbb{E}[\phi(X,A) \partial_v \log f_{A|VP}(A \mid V, P) \mid V = \zeta(x,p), P = p]. \end{aligned}$$

where step (1) is due to Assumption 8, and step (2) is due to Assumptions 2 (ii) and 4 (i) and (iv). Similarly, we have

$$\partial_p \mathbb{E}[Y \mid X = x, P = p] = \partial_p \zeta(x,p) \mathbb{E}[\phi(X,A) \partial_v \log f_{A|VP}(A \mid V, P) \mid V = \zeta(x,p), P = p],$$

where Assumptions 3 (i) was used to vanish $\partial_p f_{A|VP}$. But then, Assumption 8 allows this to be rewritten as

$$\mathbb{E}[\phi(X,A) \partial_v \log f_{A|VP}(A \mid V, P) \mid V = \zeta(x,p), P = p] = \frac{1}{\partial_p \zeta(x,p)} \partial_p \mathbb{E}[Y \mid X = x, P = p],$$

which we can substitute in the first equation to get part (ii) of the theorem.

A.6 Proof of Theorem 5

First, we have the equality

$$\begin{aligned} & P[\phi(X,A) \leqslant q^{\tau}_{Y|P}(p+\delta) \mid P = p+\delta] - P[\phi(X,A) \leqslant q^{\tau}_{Y|P}(p) \mid P = p+\delta] \\ = \;& F_{Y|P}(q^{\tau}_{Y|P}(p+\delta) \mid p+\delta) - F_{Y|P}(q^{\tau}_{Y|P}(p) \mid p+\delta) \\ = \;& \delta \partial_p q^{\tau}_{Y|P}(p) f_{Y|P}(q^{\tau}_{Y|P}(p) \mid p+\delta) + o(\delta), \end{aligned} \tag{A.2}$$

where the last equality is due to Assumptions 2 (i) and 9 (iii). Next,

$$\begin{aligned} & P[\phi(X,A) \leqslant q^{\tau}_{Y|P}(p) \mid P = p+\delta] - P[\phi(X+\delta,A) \leqslant q^{\tau}_{Y|P}(p) \mid P = p] \\ = \;& P[\phi(p+\delta+V,A) \leqslant q^{\tau}_{Y|P}(p) \mid P = p+\delta] - P[\phi(p+\delta+V,A) \leqslant q^{\tau}_{Y|P}(p) \mid P = p] \\ = \;& 0, \end{aligned} \tag{A.3}$$

where the last equality is due to Assumption 3 (i) and (ii). Lastly, using the notation $Y' := \beta(X, A)$ which is guaranteed to exist by Assumption 4 (i), we compute

$$
\begin{aligned}
& P[\phi(X+\delta, A) \leqslant q^{\tau}_{Y|P}(p) \mid P = p] - P[\phi(X, A) \leqslant q^{\tau}_{Y|P}(p) \mid P = p] \\
= \;& P[q^{\tau}_{Y|P}(p) < Y \leqslant q^{\tau}_{Y|P}(p) + (\phi(X+\delta, A) - Y) \mid P = p] \\
& - P[q^{\tau}_{Y|P}(p) - (\phi(X+\delta, p) - Y) < Y \leqslant q^{\tau}_{Y|P}(p) \mid P = p] \\
\overset{(1)}{=} \;& P[q^{\tau}_{Y|P}(p) \leqslant Y \leqslant q^{\tau}_{Y|P}(p) - \delta Y' \mid P = p] \\
& - P[q^{\tau}_{Y|P}(p) - \delta Y' \leqslant Y \leqslant q^{\tau}_{Y|P}(p) \mid P = p] + o(\delta) \\
\overset{(2)}{=} \;& \int_{q^{\tau}_{Y|P}(p)}^{\infty} \int_{-\infty}^{-\delta^{-1}[y - q^{\tau}_{Y|P}(p)]} f_{YY'|P}(y, y' \mid p) dy' dy \\
& - \int_{-\infty}^{q^{\tau}_{Y|P}(p)} \int_{-\delta^{-1}[y - q^{\tau}_{Y|P}(p)]}^{\infty} f_{YY'|P}(y, y' \mid p) dy' dy + o(\delta) \\
\overset{(3)}{=} \;& -\delta \int_{-\infty}^{0} y' f_{YY'|P}(q^{\tau}_{Y|P}(p), y' \mid p) dy' \\
& -\delta \int_{0}^{\infty} y' f_{YY'|P}(q^{\tau}_{Y|P}(x), y' \mid p) dy' + o(\delta) \\
\overset{(4)}{=} \;& -\delta \mathbb{E}[Y' \mid Y = q^{\tau}_{Y|P}(p), P = p] f_{Y|P}(q^{\tau}_{Y|P}(p) \mid p) + o(\delta),
\end{aligned}
\tag{A.4}
$$

where step (1) is due to Assumptions 4 (i), step (2) is due to Assumption 2 (i), step (3) is due to several steps of calculations using change of variables and integration by parts together with Assumption 9 (ii), and step (4) is due to Assumption 9 (iv). Now, add (A.2), (A.3), and (A.4) together to get

$$
0 = \delta \partial_p q^{\tau}_{Y|P}(p) f_{Y|P}(q^{\tau}_{Y|P}(p) \mid p + \delta) - \delta \mathbb{E}[Y' \mid Y = q^{\tau}_{Y|P}(p), P = p] f_{Y|P}(q^{\tau}_{Y|P}(p) \mid p) + o(\delta),
$$

and the result follows by Assumption 9 (i).

A.7 Proof of Theorem 6

(i): First, we have the equality

$$
\begin{aligned}
& P[\phi(x+\delta, A) \leqslant q^{\tau}_{Y|XZ}(x+\delta, z) \mid X = x+\delta, Z = z] \\
& - P[\phi(x+\delta, A) \leqslant q^{\tau}_{Y|XZ}(x, z) \mid X = x+\delta, Z = z] \\
= \;& F_{Y|XZ}(q^{\tau}_{Y|XZ}(x+\delta, z) \mid x+\delta, z) \\
& - F_{Y|XZ}(q^{\tau}_{Y|XZ}(x, z) \mid x+\delta, z) \\
= \;& \delta \partial_x q^{\tau}_{Y|XZ}(x, z) f_{Y|XZ}(q^{\tau}_{Y|XZ}(x, z) \mid x+\delta, z) + o(\delta),
\end{aligned}
\tag{A.5}
$$

where the last equality is due to Assumptions 2 (i) and 10 (iv). Using the notation $Y' := \beta(X, A)$ which is guaranteed to exist by Assumption 4 (i), compute

$$\begin{aligned}
& P[\phi(x+\delta, A) \leqslant q^{\tau}_{Y|XZ}(x,z) \mid X = x+\delta, Z = z] \\
& -P[\phi(x, A) \leqslant q^{\tau}_{Y|XZ}(x,z) \mid X = x+\delta, Z = z] \\
= \; & P[q^{\tau}_{Y|XZ}(x,z) < Y \leqslant q^{\tau}_{Y|XZ}(x,z) - (\phi(x+\delta, A) - Y) \mid X = x+\delta, Z = z] \\
& -P[q^{\tau}_{Y|XZ}(x,z) - (\phi(x+\delta, A) - Y) < Y \leqslant q^{\tau}_{Y|XZ}(x,z) \mid X = x+\delta, Z = z] \\
\overset{(1)}{=} \; & P[q^{\tau}_{Y|XZ}(x,z) \leqslant Y \leqslant q^{\tau}_{Y|XZ}(x,z) - \delta Y' \mid X = x+\delta, Z = z] \\
& -P[q^{\tau}_{Y|XZ}(x,z) - \delta Y' \leqslant Y \leqslant q^{\tau}_{Y|XZ}(x,z) \mid X = x+\delta, Z = z] + o(\delta) \\
\overset{(2)}{=} \; & \int_{q^{\tau}_{Y|XZ}(x,z)}^{\infty} \int_{-\infty}^{-\delta^{-1}[y - q^{\tau}_{Y|XZ}(x,z)]} f_{YY'|XZ}(y, y' \mid x+\delta, z) dy' dy \\
& - \int_{-\infty}^{q^{\tau}_{Y|XZ}(x,z)} \int_{-\delta^{-1}[y - q^{\tau}_{Y|XZ}(x,z)]}^{\infty} f_{YY'|XZ}(y, y' \mid x+\delta, z) dy' dy + o(\delta) \\
\overset{(3)}{=} \; & -\delta \int_{-\infty}^{0} y' f_{YY'|XZ}(q^{\tau}_{Y|XZ}(x,z), y' \mid x+\delta, z) dy' \\
& -\delta \int_{0}^{\infty} y' f_{YY'|XZ}(q^{\tau}_{Y|XZ}(x,z), y' \mid x+\delta, z) dy' + o(\delta) \\
\overset{(4)}{=} \; & -\delta \mathbb{E}[Y' \mid Y = q^{\tau}_{Y|XZ}(x,z), X = x+\delta, Z = z] f_{Y|XZ}(q^{\tau}_{Y|XZ}(x,z) \mid x+\delta, z) + o(\delta)
\end{aligned} \tag{A.6}$$

where step (1) is due to Assumption 4 (i), step (2) is due to Assumption 2 (i), step (3) is due to several steps of calculations via change of variables and integration by parts together with Assumption 10 (ii), and step (4) is due to Assumption 10 (v). Now, Assumption 10 (iii) implies that $y' f_{Y'|YXZ}(y' \mid q^{\tau}_{Y|XZ}(x,z), x+\delta, z) \to y' f_{Y'|YXZ}(y' \mid q^{\tau}_{Y|XZ}(x,z), x, z)$ pointwise for each $Y' = y'$. Moreover, Assumptions 4 (i) and 10 (ii) and (iii) yield a dominating $\mathcal{L}^1$ function g such that $\mid y' f_{Y'|YXZ}(y' \mid q^{\tau}_{Y|XZ}(x,z), x+\delta, z) \mid \leqslant g(y')$ for all y' and for all δ in a neighborhood of zero. But then, we have

$$\lim_{\delta \to 0} \mathbb{E}[Y' \mid Y = q^{\tau}_{Y|XZ}(x,z), X = x+\delta, Z = z] = \mathbb{E}[Y' \mid Y = q^{\tau}_{Y|XZ}(x,z), X = x, Z = z] \tag{A.7}$$

by the Lebesgue Dominated Convergence Theorem. Next, compute

$$\begin{aligned}
& P[\phi(x,A) \leqslant q^{\tau}_{Y|XZ}(x,z) \mid X = x+\delta, Z = z] \\
& -P[\phi(x,A) \leqslant q^{\tau}_{Y|XZ}(x,z) \mid X = x, Z = z] \\
\overset{(1)}{=} & \int\int 1\{\phi(x,a) \leqslant q^{\tau}_{Y|XZ}(x,z)\} f_{A|V}(a \mid v) f_{V|XZ}(v \mid x+\delta, z) dvda \\
& -\int\int 1\{\phi(x,a) \leqslant q^{\tau}_{Y|XZ}(x,z)\} f_{A|V}(a \mid v) f_{V|XZ}(v \mid x, z) dvda \\
\overset{(2)}{=} & \delta \int\int 1\{\phi(x,a) \leqslant q^{\tau}_{Y|XZ}(x,z)\} f_{A|V}(a \mid v) \partial_x f_{V|XZ}(v \mid x, z) dvda + o(\delta) \\
= & \delta \mathbb{E}[1\{\phi(x,A) \leqslant q^{\tau}_{Y|XZ}(x,z)\} \partial_x \log f_{V|XZ}(V \mid x,z) \mid X = x, Z = z] + o(\delta) \\
= & \delta \mathbb{E}[1\{Y \leqslant q^{\tau}_{Y|XZ}(x,z)\} \partial_x \log f_{V|XZ}(V \mid x,z) \mid X = x, Z = z] + o(\delta). \qquad \text{(A.8)}
\end{aligned}$$

where step (1) is due to Assumptions 2 (i) and 3 (i), and step (2) is due to Assumption 2 (i) and (ii). Now add Equations (A.5), (A.6), and (A.8) together and use Assumption 10 (i) and Equation (A.7) to get

$$\begin{aligned}
0 = & \partial_x q^{\tau}_{Y|XZ}(x,z) \\
& -\mathbb{E}[Y' \mid Y = q^{\tau}_{Y|XZ}(x,z), X = x, Z = z] \\
& +\mathbb{E}\left[\frac{1\{Y \leqslant q^{\tau}_{Y|XZ}(x,z)\} \partial_x \log f_{V|XZ}(V \mid x,z)}{f_{Y|XZ}(q^{\tau}_{Y|XZ}(x,z) \mid x,z)} \mid X = x, Z = z\right] \qquad \text{(A.9)}
\end{aligned}$$

Second, we have the equality

$$\begin{aligned}
& P[\phi(x,A) \leqslant q^{\tau}_{Y|XZ}(x,z+\delta) \mid X = x, Z = z+\delta] \\
& -P[\phi(x,A) \leqslant q^{\tau}_{Y|XZ}(x,z) \mid X = x, Z = z+\delta] \\
= & F_{Y|XZ}(q^{\tau}_{Y|XZ}(x,z+\delta) \mid x, z+\delta) \\
& -F_{Y|XZ}(q^{\tau}_{Y|XZ}(x,z) \mid x, z+\delta) \\
= & \delta \partial_z q^{\tau}_{Y|XZ}(x,z) f_{Y|XZ}(q^{\tau}_{Y|XZ}(x,z) \mid x, z+\delta) + o(\delta), \qquad \text{(A.10)}
\end{aligned}$$

where the last equality is due to Assumptions 2 (i) and 10 (iv). Also,

$$\begin{aligned}
& P[\phi(x,A) \leqslant q^{\tau}_{Y|XZ}(x,z) \mid X = x, Z = z+\delta] \\
& -P[\phi(x,A) \leqslant q^{\tau}_{Y|XZ}(x,z) \mid X = x, Z = z] \\
\overset{(1)}{=} & \int\int 1\{\phi(x,a) \leqslant q^{\tau}_{Y|XZ}(x,z)\} f_{A|V}(a \mid v) f_{V|XZ}(v \mid x, z+\delta) dvda \\
& -\int\int 1\{\phi(x,a) \leqslant q^{\tau}_{Y|XZ}(x,z)\} f_{A|V}(a \mid v) f_{V|XZ}(v \mid x, z) dvda \\
\overset{(2)}{=} & \delta \int\int 1\{\phi(x,a) \leqslant q^{\tau}_{Y|XZ}(x,z)\} f_{A|V}(a \mid v) \partial_z f_{V|XZ}(v \mid x, z) dvda + o(\delta) \\
= & \delta \mathbb{E}[1\{\phi(x,A) \leqslant q^{\tau}_{Y|XZ}(x,z)\} \partial_z \log f_{V|XZ}(V \mid x,z) \mid X = x, Z = z] + o(\delta) \\
= & \delta \mathbb{E}[1\{Y \leqslant q^{\tau}_{Y|XZ}(x,z)\} \partial_z \log f_{V|XZ}(V \mid x,z) \mid X = x, Z = z] + o(\delta), \qquad \text{(A.11)}
\end{aligned}$$

where step (1) is due to Assumptions 2 (i) and 3 (i), and step (2) is due to Assumption 2 (i) and (ii). Now, add (A.10) and (A.11) together and use Assumption 10 (i) to get

$$\begin{aligned} 0 &= \partial_z q^{\tau}_{Y|XZ}(x,z) \\ &+\mathbb{E}\left[\frac{1\{Y \leqslant q^{\tau}_{Y|XZ}(x,z)\}\partial_z \log f_{V|XZ}(V \mid x,z)}{f_{Y|XZ}(q^{\tau}_{Y|XZ}(x,z) \mid x,z)} \mid X=x, Z=z\right]. \end{aligned} \tag{A.12}$$

Prove part (i) from (A.9) and (A.12). Proofs of parts (ii) and (iii) follow from the same arguments as in the proofs of Theorem 2 (ii) and (iii) by applying Lemmas 1 and 2, respectively.

A.8 Proof of Theorem 7

(i): Assumption 7 provides the parameterized curve $h \mapsto (h, \delta_z(h)) =: (\delta_x, \delta_z)$ that solves the implicit function equation $\nu(x+\delta_x, z+\delta_z) - \nu(x,z) = 0$ of a smooth submanifold in a neighborhood of $h = 0$. Furthermore, under Assumption 7, $\delta_z(0) = 0$ and $(\delta_x, \delta_z) \to 0$ as $h \searrow 0$. By these properties, we have

$$\frac{\delta_z}{\delta_x} = -\frac{\nu_x(x,z)}{\nu_z(x,z)} + o(1) \quad \text{as } h \searrow 0. \tag{A.13}$$

Now, compute

$$\begin{aligned} & P[\phi(x+\delta_x, A) \leqslant q^{\tau}_{Y|XZ}(x+\delta_x, z+\delta_z) \mid X = x+\delta_x, Z = z+\delta_z] \\ & -P[\phi(x+\delta_x, A) \leqslant q^{\tau}_{Y|XZ}(x, z+\delta_z) \mid X = x+\delta_x, Z = z+\delta_z] \\ =\ & F_{Y|XZ}(q^{\tau}_{Y|XZ}(x+\delta_x, z+\delta_z) \mid x+\delta_x, z+\delta_z) \\ & -F_{Y|XZ}(q^{\tau}_{Y|XZ}(x, z+\delta_z) \mid x+\delta_x, z+\delta_z) \\ =\ & \delta_x \partial_x q^{\tau}_{Y|XZ}(x,z) f_{Y|XZ}(q^{\tau}_{Y|XZ}(x,z) \mid x+\delta_x, z+\delta_z) + o(\delta_x), \end{aligned} \tag{A.14}$$

where the last equality is due to Assumptions 2 (i) and 10 (iv). Similarly,

$$\begin{aligned} & P[\phi(x+\delta_x, A) \leqslant q^{\tau}_{Y|XZ}(x, z+\delta_z) \mid X = x+\delta_x, Z = z+\delta_z] \\ & -P[\phi(x+\delta_x, A) \leqslant q^{\tau}_{Y|XZ}(x, z) \mid X = x+\delta_x, Z = z+\delta_z] \\ =\ & F_{Y|XZ}(q^{\tau}_{Y|XZ}(x, z+\delta_z) \mid x+\delta_x, z+\delta_z) \\ & -F_{Y|XZ}(q^{\tau}_{Y|XZ}(x, z) \mid x+\delta_x, z+\delta_z) \\ =\ & \delta_z \partial_z q^{\tau}_{Y|XZ}(x,z) f_{Y|XZ}(q^{\tau}_{Y|XZ}(x,z) \mid x+\delta_x, z+\delta_z) + o(\delta_z), \end{aligned} \tag{A.15}$$

where the last equality is again due to Assumptions 2 (i) and 10 (iv). Next,

$$\begin{aligned}
& P[\phi(x+\delta_x, A) \leqslant q^{\tau}_{Y|XZ}(x,z) \mid X = x+\delta_x, Z = z+\delta_z] \\
& -P[\phi(x+\delta_x, A) \leqslant q^{\tau}_{Y|XZ}(x,z) \mid X = x, Z = z] \\
= \; & P[\phi(x+\delta_x, A) \leqslant q^{\tau}_{Y|XZ}(x,z) \mid Z = z+\delta_z, V = \nu(x+\delta_x, z+\delta_z)] \\
& -P[\phi(x+\delta_x, A) \leqslant q^{\tau}_{Y|XZ}(x,z) \mid Z = z, V = \nu(x,z)] \\
\overset{(1)}{=} \; & P[\phi(x+\delta_x, A) \leqslant q^{\tau}_{Y|XZ}(x,z) \mid Z = z+\delta_z, V = \nu(x,z)] \\
& -P[\phi(x+\delta_x, A) \leqslant q^{\tau}_{Y|XZ}(x,z) \mid Z = z, V = \nu(x,z)] \\
\overset{(2)}{=} \; & 0, \qquad \text{(A.16)}
\end{aligned}$$

where step (1) is due to the definition of (δ_x, δ_z) and step (2) is due to Assumption 3 (i). Lastly,

$$\begin{aligned}
& P[\phi(x+\delta_x, A) \leqslant q^{\tau}_{Y|XZ}(x,z) \mid X = x, Z = z] \\
& -P[\phi(x, A) \leqslant q^{\tau}_{Y|XZ}(x,z) \mid X = x, Z = z] \\
= \; & P[q^{\tau}_{Y|XZ}(x,z) < Y \leqslant q^{\tau}_{Y|XZ}(x,z) - (\phi(x+\delta_x, A) - Y) \mid X = x, Z = z] \\
& -P[q^{\tau}_{Y|XZ}(x,z) - (\phi(x+\delta_x, z) - Y) < Y \leqslant q^{\tau}_{Y|XZ}(x,z) \mid X = x, Z = z] \\
\overset{(1)}{=} \; & P[q^{\tau}_{Y|XZ}(x,z) \leqslant Y \leqslant q^{\tau}_{Y|XZ}(x,z) - \delta_x Y' \mid X = x, Z = z] \\
& -P[q^{\tau}_{Y|XZ}(x,z) - \delta_x Y' \leqslant Y \leqslant q^{\tau}_{Y|XZ}(x,z) \mid X = x, Z = z] + o(\delta_x) \\
\overset{(2)}{=} \; & \int_{q^{\tau}_{Y|XZ}(x,z)}^{\infty} \int_{-\infty}^{-\delta_x^{-1}[y - q^{\tau}_{Y|XZ}(x,z)]} f_{YY'|XZ}(y, y' \mid x, z) dy' dy \\
& - \int_{-\infty}^{q^{\tau}_{Y|XZ}(x,z)} \int_{-\delta_x^{-1}[y - q^{\tau}_{Y|XZ}(x,z)]}^{\infty} f_{YY'|XZ}(y, y' \mid x, z) dy' dy + o(\delta_x) \\
\overset{(3)}{=} \; & -\delta_x \int_{-\infty}^{0} y' f_{YY'|XZ}(q^{\tau}_{Y|XZ}(x,z), y' \mid x, z) dy' \\
& -\delta_x \int_{0}^{\infty} y' f_{YY'|XZ}(q^{\tau}_{Y|XZ}(x,z), y' \mid x, z) dy' + o(\delta_x) \\
\overset{(4)}{=} \; & -\delta_x \mathbb{E}[Y' \mid Y = q^{\tau}_{Y|XZ}(x,z), X = x, Z = z] f_{Y|XZ}(q^{\tau}_{Y|XZ}(x,z) \mid x, z) + o(\delta_x),
\end{aligned}$$

(A.17)

where step (1) is due to Assumption 4 (i), step (2) is due to Assumption 2 (i), step (3) is due to several steps of calculations using chage of variables and integration by parts together with Assumption 10 (ii), and step (4) is due to Assumption 10 (v). Now, add (A.14), (A.15), (A.16), and (A.17) together to get

$$\begin{aligned}
0 \; = \; & \delta_x \partial_x q^{\tau}_{Y|XZ}(x,z) f_{Y|XZ}(q^{\tau}_{Y|XZ}(x,z) \mid x+\delta_x, z+\delta_z) \\
& +\delta_z \partial_z q^{\tau}_{Y|XZ}(x,z) f_{Y|XZ}(q^{\tau}_{Y|XZ}(x,z) \mid x+\delta_x, z+\delta_z) \\
& -\delta_x \mathbb{E}[Y' \mid Y = q^{\tau}_{Y|XZ}(x,z), X = x, Z = z] f_{Y|XZ}(q^{\tau}_{Y|XZ}(x,z) \mid x, z) + o(\delta_x) + o(\delta_z),
\end{aligned}$$

and part (i) of the theorem now follows by Equation (A.13) and Assumptions 7 and 10 (i).

(ii): Assumption 8 provides the parameterized curve $h \mapsto (h, \delta_p(h)) =: (\delta_x, \delta_p)$ that solves the implicit function equation $\zeta(x + \delta_x, p + \delta_p) - \zeta(x, p) = 0$ of a smooth sub-manifold in a neighborhood of $h = 0$. Furthermore, under Assumption 8, $\delta_p(0) = 0$ and $(\delta_x, \delta_p) \to 0$ as $h \searrow 0$. By these properties, we have

$$\frac{\delta_p}{\delta_x} = -\frac{\zeta_x(x,p)}{\zeta_z(x,p)} + o(1) \quad \text{as } h \searrow 0. \tag{A.18}$$

Now, compute

$$\begin{aligned}
& P[\phi(x+\delta_x, A) \leqslant q^\tau_{Y|XP}(x+\delta_x, p+\delta_p) \mid X = x+\delta_x, P = p+\delta_p] \\
& -P[\phi(x+\delta_x, A) \leqslant q^\tau_{Y|XP}(x, p+\delta_p) \mid X = x+\delta_x, P = p+\delta_p] \\
= \; & F_{Y|XP}(q^\tau_{Y|XP}(x+\delta_x, p+\delta_p) \mid x+\delta_x, p+\delta_p) \\
& -F_{Y|XP}(q^\tau_{Y|XP}(x, p+\delta_p) \mid x+\delta_x, p+\delta_p) \\
= \; & \delta_x \partial_x q^\tau_{Y|XP}(x,p) f_{Y|XP}(q^\tau_{Y|XP}(x,p) \mid x+\delta_x, p+\delta_p) + o(\delta_x),
\end{aligned} \tag{A.19}$$

where the last equality is due to Assumptions 2 (i) and 11 (iii). Similarly,

$$\begin{aligned}
& P[\phi(x+\delta_x, A) \leqslant q^\tau_{Y|XP}(x, p+\delta_p) \mid X = x+\delta_x, P = p+\delta_p] \\
& -P[\phi(x+\delta_x, A) \leqslant q^\tau_{Y|XP}(x, p) \mid X = x+\delta_x, P = p+\delta_p] \\
= \; & F_{Y|XP}(q^\tau_{Y|XP}(x, p+\delta_p) \mid x+\delta_x, p+\delta_p) \\
& -F_{Y|XP}(q^\tau_{Y|XP}(x, p) \mid x+\delta_x, p+\delta_p) \\
= \; & \delta_p \partial_p q^\tau_{Y|XP}(x,p) f_{Y|XP}(q^\tau_{Y|XP}(x,p) \mid x+\delta_x, p+\delta_p) + o(\delta_p),
\end{aligned} \tag{A.20}$$

where the last equality is again due to Assumptions 2 (i) and 11 (iv). Next,

$$\begin{aligned}
& P[\phi(x+\delta_x, A) \leqslant q^\tau_{Y|XP}(x, p) \mid X = x+\delta_x, P = p+\delta_p] \\
& -P[\phi(x+\delta_x, A) \leqslant q^\tau_{Y|XP}(x, p) \mid X = x, P = p] \\
= \; & P[\phi(x+\delta_x, A) \leqslant q^\tau_{Y|XP}(x, p) \mid P = p+\delta_p, V = \zeta(x+\delta_x, p+\delta_p)] \\
& -P[\phi(x+\delta_x, A) \leqslant q^\tau_{Y|XP}(x, p) \mid P = p, V = \zeta(x, p)] \\
\overset{(1)}{=} \; & P[\phi(x+\delta_x, A) \leqslant q^\tau_{Y|XP}(x, p) \mid P = p+\delta_p, V = \zeta(x, p)] \\
& -P[\phi(x+\delta_x, A) \leqslant q^\tau_{Y|XP}(x, p) \mid P = p, V = \zeta(x, p)] \\
\overset{(2)}{=} \; & 0,
\end{aligned} \tag{A.21}$$

where step (1) is due to the definition of (δ_x, δ_p) and step (2) is due to Assumptions 3 (i) and

8. Lastly,

$$\begin{aligned}
& P[\phi(x+\delta_x, A) \leqslant q^{\tau}_{Y|XP}(x,p) \mid X=x, P=p] \\
& -P[\phi(x, A) \leqslant q^{\tau}_{Y|XP}(x,p) \mid X=x, P=p] \\
= \;& P[q^{\tau}_{Y|XP}(x,p) < Y \leqslant q^{\tau}_{Y|XP}(x,p) - (\phi(x+\delta_x, A) - Y) \mid X=x, P=p] \\
& -P[q^{\tau}_{Y|XP}(x,p) - (\phi(x+\delta_x, p) - Y) < Y \leqslant q^{\tau}_{Y|XP}(x,p) \mid X=x, P=p] \\
\overset{(1)}{=} \;& P[q^{\tau}_{Y|XP}(x,p) \leqslant Y \leqslant q^{\tau}_{Y|XP}(x,p) - \delta_x Y' \mid X=x, P=p] \\
& -P[q^{\tau}_{Y|XP}(x,p) - \delta_x Y' \leqslant Y \leqslant q^{\tau}_{Y|XP}(x,p) \mid X=x, P=p] + o(\delta_x) \\
\overset{(2)}{=} \;& \int_{q^{\tau}_{Y|XP}(x,p)}^{\infty} \int_{-\infty}^{-\delta_x^{-1}[y-q^{\tau}_{Y|XP}(x,p)]} f_{YY'|XP}(y,y' \mid x,p) dy' dy \\
& - \int_{-\infty}^{q^{\tau}_{Y|XP}(x,p)} \int_{-\delta_x^{-1}[y-q^{\tau}_{Y|XP}(x,p)]}^{\infty} f_{YY'|XP}(y,y' \mid x,p) dy' dy + o(\delta_x) \\
\overset{(3)}{=} \;& -\delta_x \int_{-\infty}^{0} y' f_{YY'|XP}(q^{\tau}_{Y|XP}(x,p), y' \mid x,p) dy' \\
& -\delta_x \int_{0}^{\infty} y' f_{YY'|XP}(q^{\tau}_{Y|XP}(x,p), y' \mid x,p) dy' + o(\delta_x) \\
\overset{(4)}{=} \;& -\delta_x \mathbb{E}[Y' \mid Y = q^{\tau}_{Y|XP}(x,p), X=x, P=p] f_{Y|XP}(q^{\tau}_{Y|XP}(x,p) \mid x,p) + o(\delta_x),
\end{aligned} \tag{A.22}$$

where step (1) is due to Assumptions 4 (i), step (2) is due to Assumption 2 (i), step (3) is due to several steps of calculations using change of variables and integration by parts together with Assumption 11 (ii), and step (4) is due to Assumption 11 (iv). Now, add (A.19), (A.20), (A.21), and (A.22) together to get

$$\begin{aligned}
0 \;=\; & \delta_x \partial_x q^{\tau}_{Y|XP}(x,p) f_{Y|XP}(q^{\tau}_{Y|XP}(x,p) \mid x+\delta_x, p+\delta_p) \\
& +\delta_p \partial_z q^{\tau}_{Y|XP}(x,p) f_{Y|XP}(q^{\tau}_{Y|XP}(x,p) \mid x+\delta_x, p+\delta_p) \\
& -\delta_x \mathbb{E}[Y' \mid Y = q^{\tau}_{Y|XP}(x,p), X=x, P=p] f_{Y|XP}(q^{\tau}_{Y|XP}(x,p) \mid x,p) + o(\delta_x) + o(\delta_p),
\end{aligned}$$

and part (ii) of the theorem now follows by Equation (A.18) and Assumptions 8 and 11 (i).

A.9 Figures

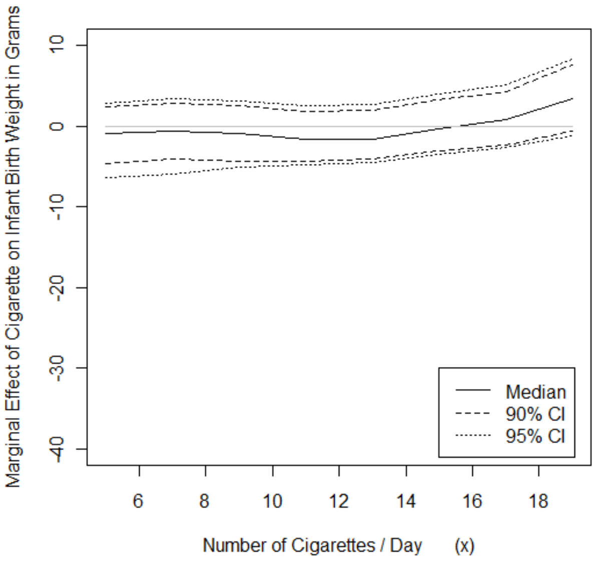


Figure 1: Confidence intervals of $\mathbb{E}[\beta(X,A) \mid Y = 2500, X = x, Z = 0.30, S = \bar{s}]$.

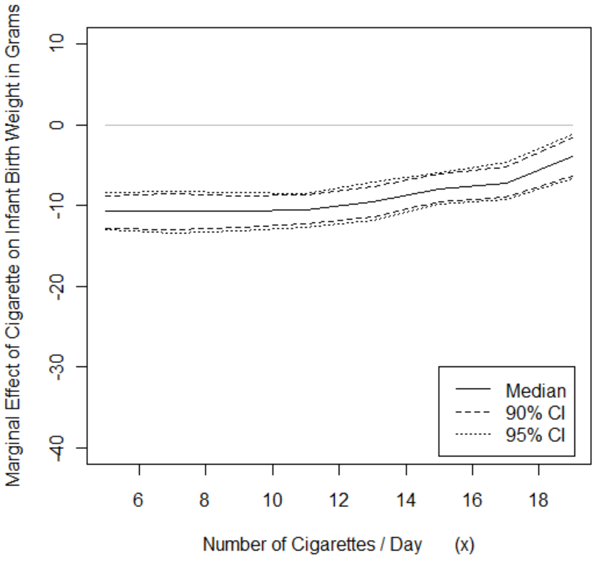


Figure 2: Confidence intervals of $\mathbb{E}[\beta(X,A) \mid Y = 3000, X = x, Z = 0.30, S = \bar{s}]$.

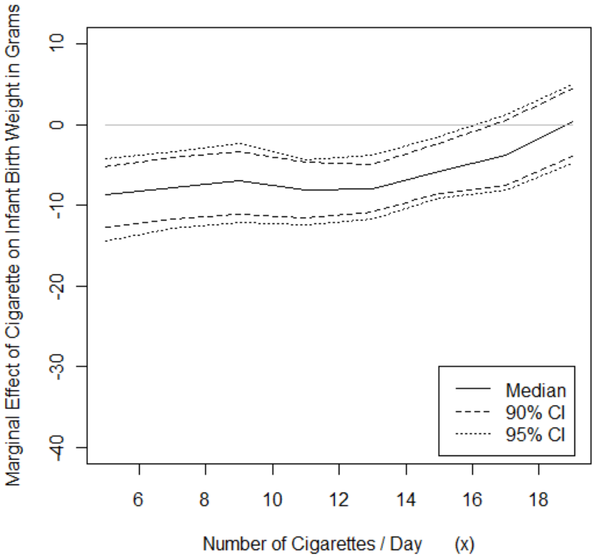


Figure 3: Confidence intervals of $\mathbb{E}[\beta(X, A) \mid Y = 2500, X = x, Z = 0.40, S = \bar{s}]$.

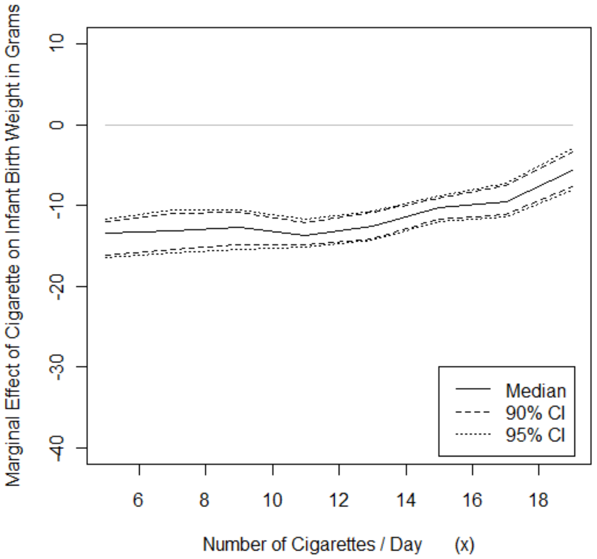


Figure 4: Confidence intervals of $\mathbb{E}[\beta(X, A) \mid Y = 3000, X = x, Z = 0.40, S = \bar{s}]$.

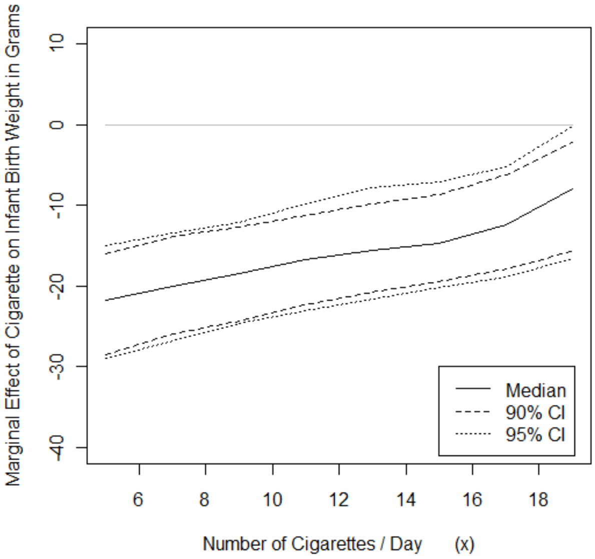


Figure 5: Confidence intervals of $\mathbb{E}[\beta(X, A) \mid Y = 2500, X = x, Z = 0.50, S = \bar{s}]$.

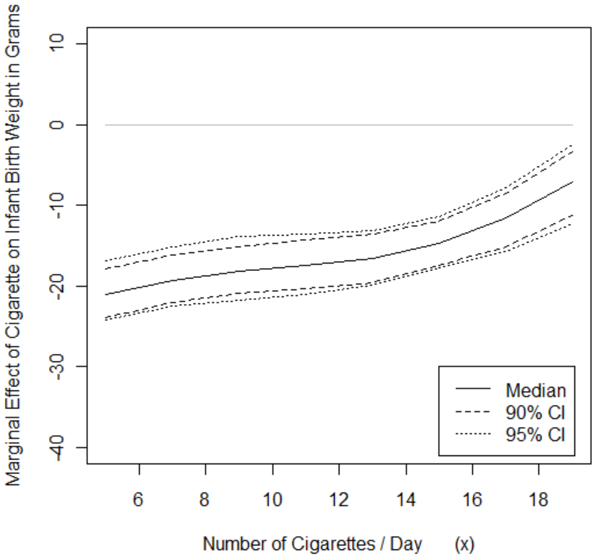


Figure 6: Confidence intervals of $\mathbb{E}[\beta(X, A) \mid Y = 3000, X = x, Z = 0.50, S = \bar{s}]$.

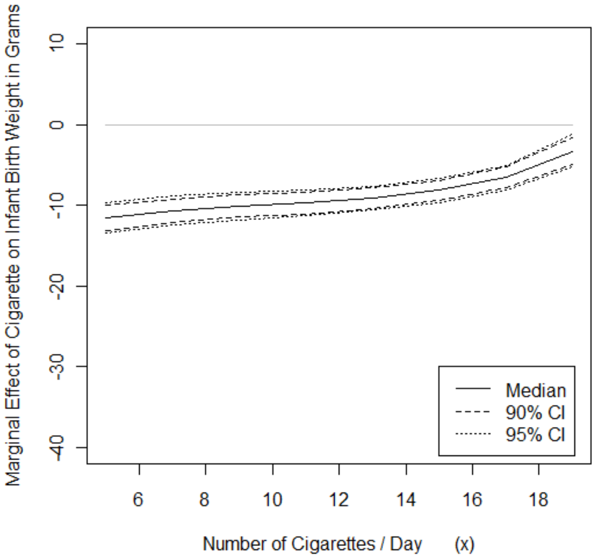


Figure 7: Confidence intervals of $\mathbb{E}[\beta(X, A) \mid X = x, Z = 0.30, S = \bar{s}]$.

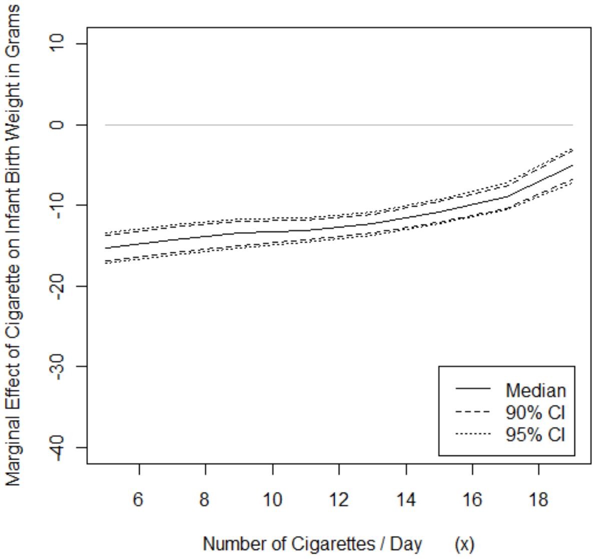


Figure 8: Confidence intervals of $\mathbb{E}[\beta(X, A) \mid X = x, Z = 0.40, S = \bar{s}]$.

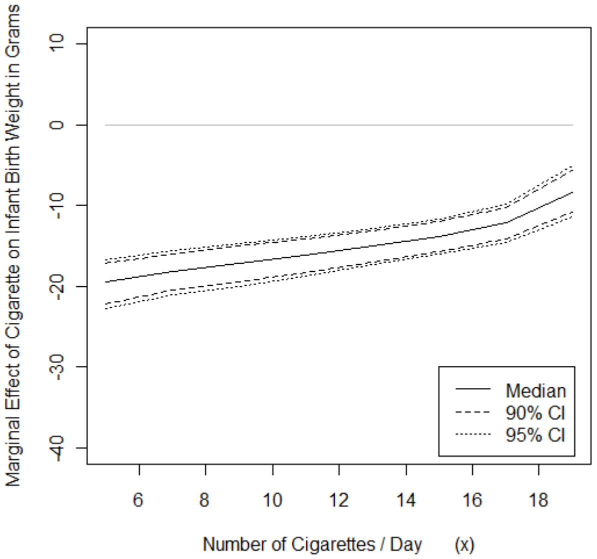


Figure 9: Confidence intervals of $\mathbb{E}[\beta(X, A) \mid X = x, Z = 0.50, S = \bar{s}]$.

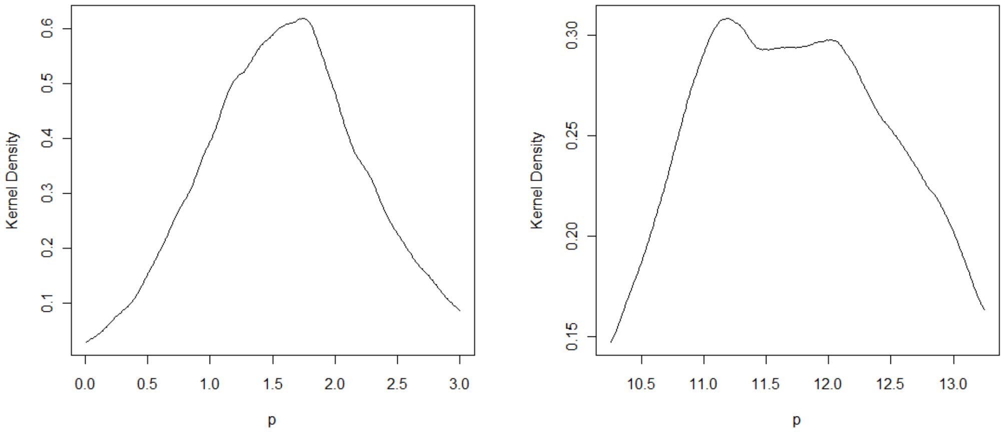


Figure 10: Kernel density estimates of the empirical distribution of P. Left: entire sample. Right: smokers only.

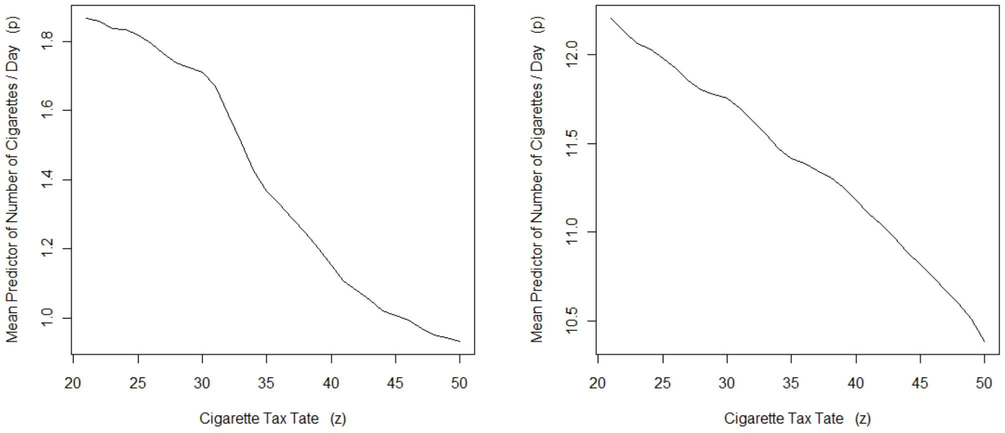
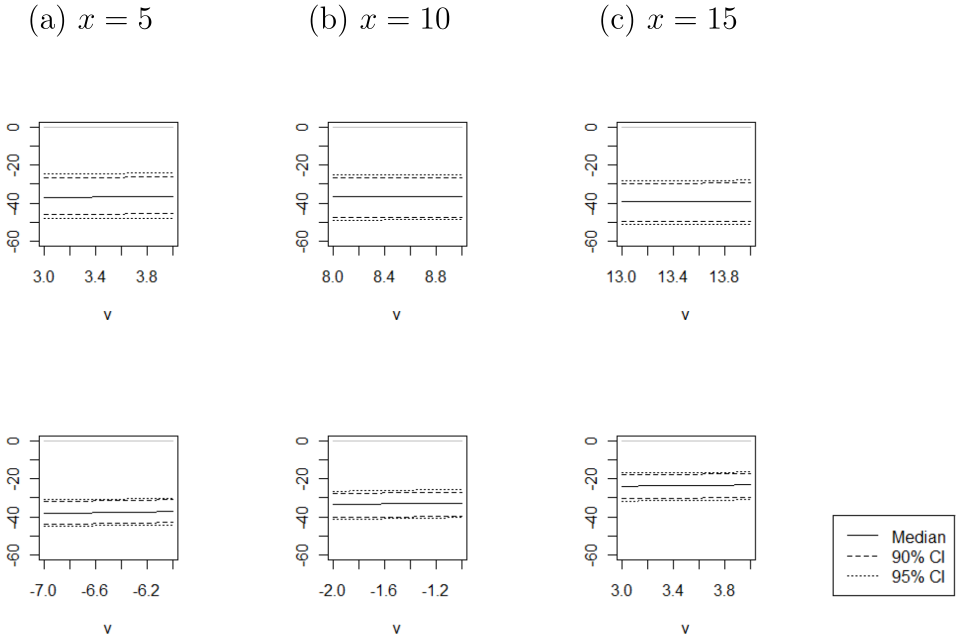


Figure 11: Nadaraya-Watson estimates of the mean regression $\mathbb{E}[X \mid Z = z, S = \bar{s}]$. Left: entire sample. Right: smokers only.


(a) $x = 5$

(b) $x = 10$

(c) $x = 15$

Figure 12: Confidence intervals of $\mathbb{E}[\beta(x, A) \mid V = v, S = \bar{s}]$ for x=5, 10, and 15. Top: entire sample. Bottom: smokers only.