Hoderlein, Stefan; Sasaki, Yuya

Working Paper

Outcome conditioned treatment effects

cemmap working paper, No. CWP39/13

Provided in Cooperation with:
Institute for Fiscal Studies (IFS), London

Suggested Citation: Hoderlein, Stefan; Sasaki, Yuya (2013) : Outcome conditioned treatment effects, cemmap working paper, No. CWP39/13, Centre for Microdata Methods and Practice (cemmap), London,
https://doi.org/10.1920/wp.cem.2013.3913

This Version is available at:
https://hdl.handle.net/10419/97388

cemmap
centre for microdata methods and practice

Outcome conditioned treatment effects

Stefan Hoderlein
Yuya Sasaki

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP39/13

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
An ESRC Research Centre

Outcome Conditioned Treatment Effects

Stefan Hoderlein* Yuya Sasaki

Boston College Johns Hopkins University

First Draft: July 15, 2009

This Draft: August 17, 2013

Abstract

This paper introduces average treatment effects conditional on the outcome variable in an endogenous setup where outcome Y, treatment X and instrument Z are continuous. These objects allow to refine well studied treatment effects like ATE and ATT in the case of continuous treatment (see Florens et al (2008)), by breaking them up according to the rank of the outcome distribution. For instance, in the returns to schooling case, the outcome conditioned average treatment effect on the treated (ATTO), gives the average effect of a small increase in schooling on the subpopulation characterized by a certain treatment intensity, say 16 years of schooling, and a certain rank in the wage distribution. We show that IV type approaches are better suited to identify overall averages across the population like the average partial effect, or outcome conditioned versions thereof, while selection type methods are better suited to identify ATT or ATTO. Importantly, none of the identification relies on rectangular support of the errors in the identification equation. Finally, we apply all concepts to analyze the nonlinear heterogeneous effects of smoking during pregnancy on infant birth weight.

Keywords: Continuous Treatment, Average Treatment Effect on the Treated, Marginal Treatment Effect, Average Partial Effect, Local Instrumental Variables, 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: Johns Hopkins University, Department of Economics, 440 Mergenthaler Hall, 3400 N. Charles St., Baltimore, MD 21218, USA, email: sasaki@jhu.edu. We have benefited from discussions and comments by Alberto Abadie, Guido Imbens, Arthur Lewbel, Rosa Matzkin, Whitney Newey, Jim Stock, Ed Vytlacil and seminar participants of the 2010 Cowles Foundation conference on Endogenous Nonseparable Models. This paper incorporates small part of a meanwhile retired paper entitled "Continuous Treatments".

1 Introduction

Motivation: Unobserved heterogeneity in preferences and other complex unobservable objects arises naturally if microeconomic models are taken to the data. Moreover, when determining the causal effect of one variable of interest $X \in \mathcal{X} \subseteq \mathbb{R}$ on an outcome $Y \in \mathcal{Y} \subseteq \mathbb{R}$, these unobservable determinants $A \in \mathcal{A} \subseteq \mathbb{R}^\infty$, are commonly correlated with X, even after controlling for all observable determinants $S \in \mathcal{S} \subseteq \mathbb{R}^K$. Given that unobserved heterogeneity is so pervasive and leads to fundamentally different outcomes, it is natural to consider analyzing causal effects separately for different level of the outcome variable, which has lead to the great popularity of quantile regression and other distributional methods. This paper presents a framework to analyze causal effects largely using quantiles in a setup with an endogenous continuous explanatory variable 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

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

where g 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 addition to the excluded factor Z, this first stage could also depend on S, however, we suppress the dependence on S in what follows.

Parameters of Interest: To define the causal effects, it is useful to consider the well studied case when X is binary, i.e., classical treatment effects. Without loss of generality, we can 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 then the treatment effect $\beta(A) = g(1, A) - g(0, A)$, 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

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

where $\mathcal{F}$ is frequently either the trivial sigma algebra (i.e., $\mathcal{F} = \{\emptyset, \Omega\}$), in which case we obtain the average treatment effect (ATE), $\mathbb{E}\left[g(1,A) - g(0,A)\right]$, or 2. $\mathcal{F} = \sigma(X)$, and $X = 1$, in which case we obtain the average treatment on the treated (ATT), $\mathbb{E}\left[g(1,A) - g(0,A)|X = 1\right]$.

In this paper, we focus mainly on the case where X and Z are continuously distributed as has been common in the literature on nonseparable models (e.g., Altonji and Matzkin (2005), Chesher (2003), Florens, Heckman, Meghir and Vytlacil (2008, FHMV henceforth), Hoderlein and Mammen (2007), Imbens and Newey (2009, IN henceforth). 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. A natural parameter of interest is a natural generalization of the binary treatment setting: It is $\partial_x g(x,a)$, the partial effect of a marginal change in x, which we denote $\beta(x,a) = \frac{\partial}{\partial x} g(x,a)$, 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 $\beta(x,a)$. A very commonly analyzed marginal effect is, e.g., a demand or labor supply elasticity with respect to price, resp. the wage rate.

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. Starting with work by Chamberlain (1984), an analog to the ATE is the average partial effect (APE), i.e., $\mathbb{E}\left[\beta(X,A)\right]$,which extends the notion of average derivative studied extensively in the semiparametric literature by averaging across unobserved heterogeneity as well. In the special case of a (correlated) linear random coefficients structure, i.e., $\beta(x,a) = \beta(a)$, this parameter becomes the average random coefficient $\mathbb{E}\left[\beta(A)\right]$, an object extensively studied in the recent panel data literature, see e.g., Graham and Powell (2012).

However, there are many instances where the overall average marginal effect may not be the only parameter of interest, and where one would like to fix the position x, at which the effect of interest is analyzed, and study this effect at different values, say, $x_1, .., x_k$, to obtain

an idea about heterogeneity in responses depending on the level of X chose. Formally, we are interested in:

$$\mathbb{E}\left[\beta(x, A)|X = x\right] = \int \beta(x, a) f_{A|X}(a|x) da,$$

where, for simplicity, we assume that $f_{A|X}$ is a density. In this setup, 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. It is instructive to think of this object in terms of an economic example: Suppose X were schooling (stylized, since we are talking about continuous quantities), and A were ability. Then, $\beta(10, a)$ would be the effect of an exogenous marginal increase in schooling for individuals having attended 10 units of schooling, and having ability $A = a$. $f_{A|X}(a|10)$ in turn gives the density of ability given 10 years of schooling. Since ability and schooling are believed to be correlated, $f_{A|X}(a|x) \neq f_A(a)$. $\mathbb{E}\left[\beta(x, A)|X = x\right]$ is then the average marginal effect of an exogenous marginal increase in schooling for individuals having attended 10 years (units) of schooling, weighted with the density of ability of those individuals who attend ten years of schooling. This the subpopulation directly affected by the exogenous policy change.

This effect is considered by FHMV (2008), Altonji and Matzkin (2005), and Hoderlein and Mammen (2007), amongst many others. FHMV (2008) call this effect the ATT, because of the obvious parallel with the treatment effects literature, and we adapt this terminology[1].

[1]Other weighting schemes are imaginable. For instance, following Blundell and Powell (2004), one could form

$$\mathbb{E}_A\left[\beta(x, A)\right] = \int \beta(x, a) f_A(a) da,$$

which gives the average marginal effect, the derivative of the ASF, of an exogenous marginal increase in schooling for individuals having attended 10 years of schooling, weighted by the density of ability of **all** individuals (such an approach was recently advocated by Kasy (2013) in a related setup that unlike us does not assume monotonicity in the unobservable in the selection equation, but assumes monotonicity in the instrument). We, however, focus on the LAR for the following reason: suppose a policy maker were able to perform such an exogenous marginal increase for everybody who has attended 10 years of schooling. Then it seems that she should primarily be concerned with the average effect **on this subpopulation**. Put reversely, there are many individuals, presumably with lower ability, whose unobserved ability is such that they will never reach 10 years of schooling. The natural question is then why these people should be included in the averaging, if they are

Causal effects conditional on a continuous covariate $S = s$ are also analyzed in the regression discontinuity approach, specifically, $\mathbb{E}[Y_1 - Y_0|S = s]$, where s is the threshold at which the discontinuity appears, see any standard textbook, e.g., Wooldridge (2002). Finally, observe that the overall APE is a weighted average of the subpopulation $X = x$ specific ATT, weighted with their density $f_X(x)$.

While we present some new results for APE and ATT in this setup, the main innovation in this paper is to present outcome conditioned treatment effects. To stay in the returns-to-schooling example, a policy maker may want to focus in his decision not just on different levels of X, say, high school drop-outs, but also on levels of Y, say, low wage individuals. In this paper, we therefore introduce outcome conditioned treatment effects. In particular, we propose natural generalizations of the APE and ATT. The first generalization is the *average partial effect conditional on the outcome*, abbreviated APEO, and defined formally through

$$\mathbb{E}[\beta(X, A)|Y = y] = \int \beta(x, a) f_{AX|Y}(a, x|y) da dx,$$

and the second generalization is the *average treatment effect on the treated conditional on the outcome,* abbreviated ATTO,

$$\mathbb{E}[\beta(x, A)|X = x, Y = y] = \int \beta(x, a) f_{A|XY}(a|x, y) da,$$

Both effects allow the policy maker to split the overall population, respectively the subpopulation defined by the treatment intensity $X = x$, according to the values of the outcome variable Y. In the special case where $X = 10$ denotes 10 years of schooling and $Y = y$ is subsequent wage, say, y is half median wage, a typical measure of poverty, $f_{A|XY}(a|x, y)$ gives the density of ability A given 10 years of schooling and half median wage. Note that $\beta(x, a)$ is still the (heterogeneous) exogenous causal effect of a marginal increase in schooling for individuals having attended 10 units of schooling, and having ability $A = a$. The weighting with $f_{A|XY}(a|x, y)$, however, reflects now the different objective of the policy maker, who wants to focus on the subpopulation with ten years of schooling *and* low subsequent wages y.

Contributions: Since the focus of this paper is on the introduction of two causal effects, we separate the paper accordingly. We are first concerned with obtaining the APE and the APEO.

never affected by this measure. Of course, there may be other applications for which the unconditional average may be more sensible. Note that in either of these approaches, one could condition on observables S as well.

We will consider a local instrumental variable (LIV) approach to identification analogously to Heckman and Vytlacil (1999, 2005, 2007, henceforth HV)) in the binary choice case that is based on mean regression. Interestingly, the LIV ceases to identify the MTE in the continuous case, i.e., a causal effect conditional on a first stage unobservable. Instead, we establish that the LIV identifies an average structural causal effect, $\mathbb{E}\left[\beta(X,A)|Z=z\right]$, which allows to obtain the overall APE by integrating with respect to $F_Z(z)$. This results in sample counterpart estimators that are averages, and hence sensitive to large values, however, it does not require that we have an instrument that has large support in that it pushes the selection probabilities all the way to one or zero. Surprisingly, if we replace mean by quantile regressions, the same principle allows to identify the APEO under similar unrestrictive support conditions.

After having established identification for APEs, we turn to the question under which conditions ATT and ATTO are identified We start out by establishing that straightforward mean regression based generalizations of the MTE are point identified. This is very similar to FHMV (2008), and extends Heckman and Vytlacil (2007). We then show that we can generalize these results further, using distributional information as embodied in regression quantiles, to obtain averages which also involve the dependent variable Y, i.e., $\mathbb{E}\left[\beta(x,A)|X=x,Y=y\right]$. These effects can be derived from a general identification theorem that can be seen as s generalization of the Heckman (1979) selection principle.

Relationship to the Literature: This work aims at integrating two different strands of literature: 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), due in particular to the continuity of the instrument Z. Our approach is also related to the LATE framework of Imbens and Angist (1994), in particular to Angrist, Grady and Imbens (2000), though to a lesser degree because of the discreteness of Z. Unlike Imbens and Angrist (1994), we do not assume monotonicity of the FS equation in Z. Also, Abadie et al (2002) consider identification 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.

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. From Hoderlein and Mammen (2007, 2009), we adopt the notion that quantile derivatives identify average marginal effects. Less closely related to our work is Chesher (2003) who assumes triangularity as well as monotonicity in the outcome equation.

Other recent work that exploits distributional assumptions includes Chernozhukov and Hansen (2005), who propose estimation of a quantile IV model based on moment restrictions that assume that the individuals have either the same unobservable in both the 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 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 in the outcome equation are multivariate, as is also the case in the related work by Torgovitsky (2011), and D'Haultfoeuille and Février (2011).

As already mentioned, like us FHMV (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. 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 outlines the model and some basic assumptions, and then establishes that identification of APE and APEO through integrated local IV methods. In the third section, we focus on the treatment on the treated parameters, i.e., ATT and ATTO, and we show how regressions on X and Z serve to identify this object. 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 The Local Instrumental Variable Estimator

Throughout this paper, we assume to observe variables $(Y, X, Z) \in \mathbb{R}^3$, where Y is the outcome of interest, X is the endogenous first-stage choice that causally affects Y, and Z is an instrument that causally affects X. Without further mentioning, we will assume that they are defined on a complete probability space, $(\Omega, \mathcal{F}, P)$, and that there exist an absolutely continuous derivative of the measure P with respect to Lebesgue measure, with the density f_{YXZ} its Radon Nikodym derivative, i.e., all variables are continuous. Finally, by $Q_{Y|K}(\tau \mid k) := \inf\{y \mid F_{Y|K}(y \mid k) \geqslant \tau\}$ we denote the τ-th quantile regression of Y on K.

To provide examples, 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 Z drive X, but they are excluded from affecting Y through any other channel than through X, and they are also unrelated to any unobservable factor that affects Y.

This structure leads to the general class of causal models defined by

$$\begin{aligned} Y &= g(X, A), \\ X &= h(Z, V) \end{aligned} \tag{2.1}$$

where h is a smooth function that relates the choice of treatment intensity X, to the observed instruments Z, and unobservables V. In the same vein, g is smooth function that relates Y to X and an infinite dimensional vector of unobservables A. The unobservables (A, V) characterize the individual; all functions could depend on a set of observable exogenous covariates S as well, but we suppress them for transparency of exposition. The instruments Z represent external factors that an individual with (A, V) takes as given when making their first stage choice of treatment intensity, and subsequently, given this first stage choice, chooses or obtains Y.

As discussed above, our interest centers on averages of the structural marginal effect $\beta(x, a) := \frac{\partial}{\partial x} g(x, a)$ across a heterogeneous population. A natural candidate to identify average effects is the LIV, i.e.,

$$\frac{\frac{\partial}{\partial z}\mathbb{E}\left[Y|Z=z\right]}{\frac{\partial}{\partial z}\mathbb{E}\left[X|Z=z\right]}, \tag{2.2}$$

which identifies causal effects in both the classical linear model with endogenous regressors, and in the endogenous binary treatment model with heterogeneous effects. Surprisingly, in the endogenous continuous treatment case, the LIV itself does not turn out to identify interesting causal effects, and it is only integrals of these quantities that identify interesting causal effects. We require an instrument independence assumption, i.e.,

Assumption 1 (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., that joint independence of the instruments from all unobservables in the system holds. This is somewhat stronger than, but similar in spirit 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 $X = 1\{\pi(Z) > V\}$, see e.g., Heckman and Vytlacil (2007). Since $Y_j = g(j, A)$ in this setup, $(Y_1, Y_0, V) = (g(1, A), g(0, A), V)$, and obviously $(A, V) \perp\!\!\!\perp Z$ implies the weaker conditions $(Y_1, Y_0, V) \perp\!\!\!\perp Z$. This stronger full independence condition has been assumed in the recent work of Torgovitsky (2011), and Fevrier and d'Haultfoeuille (2011) on the continuous treatment case. However, the continuous case allows to weaken the full joint independence. While (i) $A \perp\!\!\!\perp Z \mid V$ is retained throughout this paper, in some of our results we may in particular weaken (ii) $V \perp\!\!\!\perp Z$ to a location normalization like $\mathbb{E}\left[V|Z\right] = 0$.

We next provide an equivalent to a rank condition, which is formulated for an arbitrary fixed position $z \in \mathcal{Z}$; if interest centers later on averages like the APE, they would have to hold for every point $z \in \mathcal{Z}$, except possibly on a set of measure zero :

Assumption 2 (Local Rank Condition for LIV)**.** $\frac{d}{dz}\mathbb{E}[X \mid Z = z] \neq 0$.

Assumption 2 is a generalization of the conventional rank condition. In the linear model where $X = \gamma + \delta Z + V$, for instance, this holds iff $\delta \neq 0$. These two assumptions, however, are not sufficient to point identify the effect of interest, and we have to add a condition on the way the errors enter:

Assumption 3 (Separable First Stage)**.** $h(z, v) = \pi(z) + v$.

In addition to this set of structural assumptions, we use the following set of regularity conditions for our analysis:

Assumption 4 (Regularity)**.** *The distributions of the random variables are absolutely continuous. The structural and density functions are continuously differentiable, and are dominated in absolute value by functions with finite first moments (L^1 domination).*

Under these assumptions, we obtain the first main theorem in this paper. It discusses the identification of APE

Theorem 1 (APE)**.** *(i) Suppose that Assumptions 1, 3 and 4 are satisfied for the model (2.1). If in addition Assumption 2 hold for the model (2.1) $[F_Z]$-a.s., then the average partial effect (APE) is given by*

$$APE = \mathbb{E}\left[\beta(X, A)\right] = \int_{\mathcal{Z}} \frac{\partial}{\partial p}\mathbb{E}\left[Y|P = p\right]\Big|_{p=\pi(z)} f_Z(z)dz.$$

(ii) If π is continuously differentiable and Assumption 2 is satisfied for all z in the support of $Z \mid Y = y_0$, then the average partial effect conditional on outcomes (APEO) is given by

$$APEO(y_0) = \mathbb{E}\left[\beta(X, A)|Y = y_0\right] = \int_{\mathcal{Z}} \frac{\partial}{\partial p} Q_{Y|P}(F_{Y|Z}(y_0 \mid z) \mid p)\Big|_{p=\pi(z)} f_{Z|Y}(z|y_0)dz.$$

Discussion of theorem 1 : 1. This theorem provides equalities that equate APE and APEO to quantities on the right hand side that are directly estimable from data. Both objects

take the form of average derivatives, weighted by a density. Observe the differences in these results, in particular the fact that the former uses mean regression where the latter uses quantile regressions, and that integration is with respect to the unconditional, resp. conditional, distribution of Z. These differences are owed to the fact that APEO considers the effect around y_0 only, and hence it uses the rank associated with y_0, i.e., $F_{Y|Z}(y_0 \mid z)$, for every z, as well as the conditional distribution $f_{Z|Y}(z|y_0)$ at y_0. Note that we can relax the rank condition A 2 in the latter case, too, because we are only concerned with the subpopulation for which $Y = y_0$.

2. This right hand side structures admit natural sample counterparts estimators, i.e.,

$$\begin{aligned}\widehat{\mathbb{E}}\left[\beta(X,A)\right] &= n^{-1}\sum_i \frac{\partial}{\partial p}\widehat{m}_{Y|P}(\widehat{\pi}(Z_i)),\\ \widehat{\mathbb{E}}\left[\beta(X,A)|Y=y_0\right] &= n^{-1}\sum_i \frac{\partial}{\partial p}\widehat{Q}_{Y|P}(\widehat{F_{Y|Z}}(Y_i,Z_i),\widehat{\pi}(Z_i))\widehat{f}_Y(Y_i)^{-1}K_h(Y_i-y_0)\end{aligned}$$

where the hats denote nonparametric estimators, e.g., kernel based local polynomials, and K_h is a standard kernel.

3. Support conditions: since we are integrating with respect to the conditional distribution of Z, we are not subject to identification at infinity. Also, we do not place any restrictions on the support of the first stage unobservable V; in particular, we are not assuming rectangular support, or any comparably restrictive assumption.

4. Assuming additivity and full independence of (A, V) from Z together, i.e., Assumptions 1, and 3, is quite restrictive. However, both assumptions may be weakened. In the next section, we show an alternative way to obtain APE and APEO that only requires monotonicity in a scalar error term, but retains full independence. If, on the other side, additivity in the first stage equation is acceptable, but heteroskedasticity is suspected, the full independence assumption 1 (ii) can be weakened to a location restriction, e.g., mean independence $E[V|Z=z]=0$. In this case, we obtain that

$$APE = \mathbb{E}\left[\beta(X,A)\right] = \int_{\mathcal{Z}} \frac{\partial}{\partial p}\mathbb{E}\left[Y|P=p\right]\Big|_{p=\pi(z)} f_Z(z)dz - \int_{\mathcal{Z}} \mathbb{E}\left[YS|P=p\right]|_{p=\pi(z)}\, f_Z(z)dz,$$

where S is the score $\frac{\partial}{\partial p}\log f_{V|P}(V;p)$. In this expression, the second term corrects for the bias that arises from the fact that V and Z are not fully independent. A similar expression can be obtained for APEO in this case. Straightforward sample counterpart estimation is again possible, and the support conditions are similarly unrestrictive.

3 Average Treatment Effect on the Treated

As mentioned above, in this section we show that we can identify the average treatment effect on the treated (ATT, see FHMV (2008)) $\mathbb{E}\left[\beta(x,A)|X=x\right]$, and the outcome conditioned average treatment effect on the treated (ATTO), i.e., $\mathbb{E}\left[\beta(x,A)|X=x,Y=y\right]$, which is one of the major innovations in this paper. We obtain these quantities by objects that much closer resemble integrating a "selection" structure, where the second term is akin to a selection correction. We establish in particular that a similar approach can also be pursued using the entire distribution of Y, and in our mind is even preferable to existing methods as it allows combining high dimensional unobservables in the outcome equation with distributional information. We replace Assumptions 2 and 3 by the following more general assumptions. They hold again for almost all $z \in \mathcal{Z}$. The first assumption replaces the additive separability assumed in section 2.

Assumption 5 (Invertibility)**.** *$h(z,\cdot\,)$ is invertible in its second argument Z-a.s, and $\upsilon(z,\cdot\,)$ denotes the inverse.*

This obviously contains assumption 3 as special case, and will play an equally prominent role in our analysis. Similar to before, we require a rank condition, which comes in the following form:

Assumption 6 (Local Rank Condition for MTE)**.** *$\frac{\partial}{\partial z}\upsilon(z,x) \neq 0$ for all $x \in \mathcal{X}$, Z-a.s.*

Observe that this is a stronger condition, as it now requires z to be informative at all values of x, however, this condition is satisfied in the standard specifications like linear IV. To illustrate the content of this assumption, we provide three examples: 1. structural models, 2. mean regression, 3. quantile regression

Example 1: Suppose that $Y = g(X,A)$ models the demand for a single good Y as a function of wealth X and preferences A . Wealth X can be taken to be within-period log total expenditure, which is justified under separability restrictions on preferences (see Lewbel, (1999)). The first-stage endogenous choice $X = h(Z,V)$ is modeled as a result of optimizing behavior: given a life cycle income stream, individuals first decide on how much to allocate to period t. Then they decide on how much to spend for the single good in question.

More formally, suppose that the dynamic consumption decision of economic agent at time t is given by

$$\max_{\{X_{t+\tau}\}_{\tau=0}^{\bar{T}-t}} \mathbb{E}_t \left[\sum_{\tau=0}^{\bar{T}-t} \beta^\tau u(X_{t+\tau}; \theta) \right] \quad \text{s.t.} \quad M_{t+1} = (M_t - X_t)R + Z_{t+1}$$

where $u(\cdot;\theta)$ is the CARA utility function with parameter θ, Z_t is the consumer's idiosyncratic labor income, $\bar{T}$ is the terminal period, M_t denotes assets, and R is the interest factor which, for simplicity, is fixed and deterministic. If the growth of Z_t is stochastic, with iid Gaussian innovations having variance σ^2, then the first-stage function for individuals with no initial assets is given by

$$X_t = \log\left(\delta(\bar{T}-t)Z_t - \frac{\sigma^2\theta}{2\beta}\right),$$

where individuals have heterogeneous structural parameters (β, θ), and $\delta(\cdot)$ is a known function of the remaining life span. In this case, $V = (2\beta)^{-1}\sigma^2\theta$, and Assumption 5 is trivially satisfied. Moreover, if $\delta(\bar{T}-t) \neq 0$, which is satisfied under mild conditions, Assumption 6 is satisfied as well.

Example 2: Suppose that the first-stage model is a nonparametric mean regression of the form

$$h(z,v) = \pi(z) + v \qquad \text{where } \mathbb{E}[V \mid Z] = 0.$$

In this case, Assumption 5 is trivially satisfied. Furthermore, the traditional local rank condition $\pi'(z) \neq 0$ satisfies Assumption 6.

Example 3: Another important special case is when the first-stage model can be represented by a nonparametric quantile regression. This is the case when we assume to have a model of the form

$$h(z,v) = Q_{X|Z}(v \mid z) \qquad \text{where } Z \perp\!\!\!\perp V,$$

which happens for instance if h is strictly monotonic in a scalar unobservable, w.l.o.g. normalized to be $U[0,1]$. In this case, Assumption 5 is satisfied if $X \mid Z = z$ is continuously distributed. Furthermore, the commonly assumed local rank condition $\frac{\partial}{\partial z} F_{X|Z}(x \mid z) \neq 0$ satisfies Assumption 6.

We now provide the results for the average treatment effect on the treated, for a given level of treatment intensity $X = x_0$. For these results, we can weakend the above conditions to hold only locally; however, their economic interpretation remains materially unchanged.

Assumption 7. *The following conditions hold at $X = x_0$:*

(i) $h(z, \cdot)$ is invertible in its second argument for a.s. z in the support of $Z \mid X = x_0$.

(ii) $\frac{\partial}{\partial z}\nu(z, x_0) \neq 0$ for a.s. z in the support of $Z \mid X = x_0$.

Moreover, by (i′) and (ii′) we denote the same conditions, but change the support to a.s. z in the support of $Z \mid Y = y_0, X = x_0$.

Parts (i) and (ii) of this assumption are restatements of Assumptions 5 and 6, respectively, at various locations of z in the support of $Z \mid X = x_0$, and parts (i) and (ii) of the same assumptions at various locations of z in the support of $Z \mid Y = y_0, X = x_0$. Equipped with these assumptions and the notation $\rho(z, x) := \left[\frac{\partial \upsilon(z,x)}{\partial x}\right] / \left[\frac{\partial \upsilon(z,x)}{\partial z}\right]$, we obtain the following theorem, whose proof we delegate to the appendix:

Theorem 2 (ATT and ATTO)**.** *(i) Suppose that Assumptions 1 (i), 4, and 7 (i), (ii) are satisfied for the model (2.1). Then, the average treatment effect on the treated at $X = x_0$ is given by*

$$\begin{aligned}\mathbb{E}\left[\beta(x_0, A)|X = x_0\right] &= \int \frac{\partial}{\partial \xi}\mathbb{E}\left[Y|X = \xi, Z = z\right]\Big|_{\xi=x_0} f_{Z|X}(z|x_0)dz \\ &\quad - \int \rho(z, x_0)\, \frac{\partial}{\partial \zeta}\mathbb{E}\left[Y|X = x_0, Z = \zeta\right]\Big|_{\zeta=z} f_{Z|X}(z|x_0)dz\end{aligned}$$

In addition, if Assumption 7 (i), (ii) holds for a.s. x_0 in the support of X, then the average partial effect (APE) is given by integrating the right hand side with respect to f_X.

(ii) Suppose that Assumptions 1 (i), 4, and 7 (i′),(ii′) are satisfied for the model (2.1). Then, the outcome conditioned average treatment effect on the treated at $(Y, X) = (y_0, x_0)$ (ATTO) is given by

$$\begin{aligned}\mathbb{E}\left[\beta(x_0, A)|Y = y_0, X = x_0\right] &= \int \frac{\partial}{\partial \xi} Q_{Y|XZ}(F_{Y|XZ}(y_0 \mid x_0, z) \mid \xi, z)\Big|_{\xi=x_0} f_{Z|YX}(z|y_0, x_0)dz \\ &\quad - \int \rho(z, x_0) \cdot \frac{\partial}{\partial \zeta} Q_{Y|XZ}(F_{Y|XZ}(y_0 \mid x_0, z) \mid x_0, \zeta)\Big|_{\zeta=z} f_{Z|YX}(z|y_0, x_0)dz.\end{aligned}$$

If in addition Assumption 7 (i′),(ii′) is satisfied for a.s. x_0 in the support of $X \mid Y = y_0$, then the outcome conditioned average partial effect (APEO) is given by integrating the right hand side with respect to $f_{X|Y}(\cdot|y_0)$.

Discussion of Theorem 2 : 1. This theorem has a similar structure than theorem 1; however, unlike theorem 1, the object of interest is a conditional effect, either the effect on the treated subpopulation (with treatment intensity x_0, say 10 years of schooling), or the effect on the treated subpopulation with a certain level of the outcome variable, i.e., $Y = y_0$, say, individuals with 10 years of schooling, and income at the poverty line. Both objects take again the form of average derivatives, weighted by a density. The differences in the results (i) and (ii) are again that the former uses mean regression while the latter uses quantile regressions, but now that integration is with respect to the distribution of Z conditional on X, resp., on X and Y. Note that compared to theorem 1 we can relax the rank condition further in the latter case, too, because we are only concerned with the subpopulation for which $X = x_0$, resp., $X = x_0, Y = y_0$.

2. Similar to theorem 1, we can devise sample counterpart estimators. Let $\frac{\partial}{\partial x}\widehat{m}_{Y|XZ}(x,z)$ denote a nonparametric estimator for the derivative with respect to x of the mean regression of Y on X and Z at x, z, and analogously for $\frac{\partial}{\partial z}$. Let $\frac{\partial}{\partial x}\widehat{Q}_{Y|XZ}$, $\frac{\partial}{\partial z}\widehat{Q}_{Y|XZ}$ denote the analogous objects for quantile regression, and let $\hat{\rho}$ be a nonparametric estimator for ρ. Then, natural estimators for right hand side structures are:

$$\begin{aligned}
\widehat{\mathbb{E}}\left[\beta(X,A)|X=x_0\right] &= n^{-1}\sum_i \left\{ \frac{\partial}{\partial x}\widehat{m}_{Y|XZ}(X_i,Z_i) - \hat{\rho}(Z_i,X_i)^{-1}\frac{\partial}{\partial x}\widehat{m}_{Y|XZ}(X_i,Z_i)\right\} \\
&\quad \times \widehat{f}_X(X_i)^{-1}K_h(X_i-x_0), \\
\widehat{\mathbb{E}}\left[\beta(X,A)|X=x_0,Y=y_0\right] &= n^{-1}\sum_i \left\{ \frac{\partial}{\partial x}\widehat{Q}_{Y|XZ}(\widehat{F_{Y|XZ}}(Y_i|X_i,Z_i)|X_i,Z_i) \right. \\
&\quad \left. - \hat{\rho}(Z_i,X_i)^{-1}\frac{\partial}{\partial x}\widehat{Q}_{Y|XZ}(\widehat{F_{Y|XZ}}(Y_i|X_i,Z_i)|X_i,Z_i)\right\} \\
&\quad \times \widehat{f}_{XY}(X_i,Y_i)^{-1}K_h(X_i-x_0)K_h(Y_i-y_0)
\end{aligned}$$

where K_h is a standard kernel. Similar remarks apply again for the estimation of APE and APEO under these assumptions. Section A.8 in the appendix provides more details of estimation.

3. Support conditions - as before: since we are integrating with respect to the conditional distribution of Z, we are not subject to identification at infinity. Also, we do not place any restrictions on the support of the first stage unobservable V; in particular, we are not assuming rectangular support.

4. Examples for different choices of $\rho(z,x)$ and theorem 2. **Example 2 (cont.):** First, note that theorem 2 does not require full independence between the instrument Z and the first-stage unobservable V. Therefore, heteroskedasticity in the first stage, $\mathbb{E}[V^2 \mid Z] \neq$ constant, is admissible. Applying the result to this special case where $\rho(z,x) = -1/[\pi'(z)]$ yields the following result:

$$\begin{aligned}\mathbb{E}\left[\beta(x_0,A)|X=x_0\right] &= \int \frac{\partial}{\partial \xi}\mathbb{E}\left[Y|X=\xi,Z=z\right]\Big|_{\xi=x_0} f_{Z|X}(z|x_0)dz \\ &\quad - \int \pi'(z)^{-1}\,\frac{\partial}{\partial \zeta}\mathbb{E}\left[Y|X=x_0,Z=\zeta\right]\Big|_{\zeta=z} f_{Z|X}(z|x_0)dz,\end{aligned}$$

and analogously for ATTO. Moreover, in the case of **Example 3**, $\rho(z,x) = -\left[\frac{\partial}{\partial z}Q_{X|Z}(v \mid z)\right]^{-1}$ where $v = F_{X|Z}(x \mid z)$, and the results continue to hold with the obvious modifications.

5. From closer inspection of the proof, it is obvious that a key building block is a parameter that is closely related to the marginal treatment effect (MTE) of Heckman and Vytlacil (2005, 2007). Consequently, one could use this quantity to construct policy relevant treatment effects as in Heckman and Vytlacil (2005), see also the illuminating discussion in Heckman and Vytlacil (2007). To keep the focus of this paper, we leave such an approach for future research.

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 Ringel (1999); Lien and Evans (2005)). Most papers, including those in the medical literature, have suggested that the effect of smoking (as a 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 effect of smoking, in order to better assess potential effects of a policy that encourages pregnant woman to reduce the number of cigarettes smoked per day by one.

To start out with, it is instructive to formulate hypotheses that specifies what one would expect. Recall that the relationship of interest is $Y = g(X, A)$, where Y denotes the outcome of interest, in our case the birth weight, X denotes the number of cigarettes, and A denotes other unobserved factors that affect the birth weight. Examples for A include such things as the diet of the mother, overall health conditions, whether she consumes alcohol, how conducive her social environment is, etc. Note that these variables are heavily correlated with X, individuals who live a healthy lifestyle are less likely to smoke a lot during pregnancy. For fixed $A = a$, we would expect cigarettes to have a negative effect, though probably with decreasing returns to scale. If $A = a$ is such that the birth weight is at the lower end of the distribution of birth weights, we would expect the effect still to be negative, but perhaps less large in absolute value, simple because there is not as much weight the newborn might be able to loose; i.e., in economic terms we have decreasing returns to scale. Put differently, if the other factors have contributed to a lot of harm already, the additional harm of a single cigarette might be less big. Conversely, if $A = a'$ such that the other conditions are perfect for the optimal growth of the fetus, the harm an additional (marginal) cigarette causes may be highest.

To isolate the causal effects of cigarettes on infant birth weight, we extend an idea of Evans and Ringel (1999). We allow for arbitrary nonlinear, endogenous and heterogeneous effects of smoking, and want to obtain APEO and ATTO. 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

$$\begin{cases} Y = g(X, A) \\ X = h(Z, V) \end{cases}$$

where A captures other unobserved factors related to the physiological characteristics and the

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
Births	1.97	1.00	Number of live births experienced

Table 1: Descriptive statistics of the data of the repeated cross sections from 1989 to 1999 from the Natality Vital Statistics System of the National Center for Health Statistics.

lifestyle of the mother that impact the child's birth weight, and V is a scalar summarizing first stage factors that impact the choice of number of cigarettes. The relevant independence restriction we use is $Z \perp\!\!\!\perp (A, V)$, but note that A and V are generally correlated, and cause endogeneity of X. The structural features of interest are the APEO and ATTO, which, as we established are identified under the conditions of Theorem 2.

In terms of data, we use a repeated 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 1. From this data set we extract a random sample in the time period between 1989 to 1999.

Due to the point mass of the distribution of X at $X = 0$ which conflicts the assumption of absolute continuity, we focus on the subsample with $X > 0$, and our analysis is hence about the intensity of smoking, conditional on a pregnant mother being a smoker.

In terms of details of implementation: We use the straightforward sample-counterpart estimators proposed in Section A.8 in the appendix, with standard normal kernels. We select the bandwidth to be smaller than the cross validated bandwidth,[2] to account for the fact that we are integrating out Z. We report our results in graphs, as they summarize the gist of our result in ways a numerical representation or table could not do. In addition to showing the results in the original scale, due to the fact that some effects are close to zero and others are compara-

[2]For all the results to be presented, we use the bandwidths of $h_y = 250$, $h_x = 5$ and $h_z = 25$ for the random variables Y, X and Z, respectively. Perturbations of these values do not qualitatively change our results.

bly large, we also plot the results on a logarithmic scale; more precisely, since the results are negative, we show the negative logarithm of the absolute value.

When presenting the results, since the APEO can be obtained from ATTO by integrating out X, we start with the latter. Estimates of the ATTO at $X = 5$, $X = 10$ and $X = 20$ are plotted in Figure 1. Each graph should be read as a collection of average causal effects of subpopulations defined by the value on abscissa, e.g., the subpopulation with birth weight $Y = 3000$. The three dashed curves are the 2.5-th, 50-th, and 97.5-th percentiles of 500 bootstrap resampled estimates of the ATTOs. Note first that the subpopulations may (and actually will) differ according to their distribution of A, the subpopulation of newborns with $Y = 2000$ is very different from the one with $Y = 3500$, as the mothers in the former subpopulation have, at least on average, either worse health conditions or a exhibit a behavior that is much less conducive to healthy child growth. In either case, the effect of an additional cigarette is as we would expect much larger on the subpopulation that is more healthy, for the subpopulations with smaller than average birth weight the effect gradually tapers off, until we observe that subpopulations with $Y = 2000$ exhibit only a small effect of an additional cigarette. We would like to emphasize, however, two things: First, in this case the damage has already been done by other factors - recall in that respect also that the distribution of A may change. Second, a counterfactual marginal reduction in smoking intensity would still improve birth weights for all levels of Y, as is evident from switching to the log scale This general reduction of the causal effect is true regardless of the value of X, however, it is more pronounced for lower vales of X, i.e., when mothers smoke less. Observe how heterogeneous the causal effects are, and that it is important to have a tool that is able to disentangle these details.

This brings us to our second comparison: When comparing across the different values of X, we see that the average effect of an additional cigarette is much stronger for $X = 5$. It implies that counterfactually exposing pregnant mothers to an additional cigarette causes much more harm if the actual smoking intensity is small; it is the first cigarettes that have the strongest effect, and the effect tapers of gradually. If the fetus is already exposed to a lot of smoking, a satiation effect kicks in. This is in accordance with the hypothesis that, in economic terms, that cigarettes exhibit a negative, but diminishing marginal product in the birth weight production

function. This effect gets smaller the more other factors have already contributed to low birth weight. This means also that for these individuals, the policy maker must provide different incentives: mothers living an unhealthy lifestyle should not just have to change their smoking behavior, but also give up other unhealthy habits, while for the healthy mothers the effect of (only) giving up smoking is much more pronounced, and should hence be at the focus of the incentives.

Next, we repeat the same exercise, but condition in addition on different observed characteristics to understand whether there are additional factors (denoted S above) that aggravate the effect of smoking. First, we consider subgroups defined mother's age, and distinguish between first and later births, to see whether there are some aspects of maturity or life cycle planning visible. The top and bottom rows of Figure 2 show the ATTO of the first birth for mothers in their teens and twenties, respectively, at $X = 5$. We do not see a major difference in these results from the ATTO at $X = 5$ that used the entire sample. This suggests that the outcome-conditioned partial effects are not heterogeneous across birth or and mother's age.

We then repeat the exercise using the subpopulation of mothers who have regular alcohol intake. Figure 3 shows the ATTO, again at $X = 5$, for the subpopulation of drinking mothers giving their first births. We now see that the magnitudes of the outcome-conditioned partial effects are somewhat larger than before, implying that alcohol and smokes are complementary negative factors of the birth weight production function. The qualitative pattern of the ATTO remains robust even with this subpopulation. A policy implication derived from this observation is that it is even more effective to encourage drinking mothers to reduce the number of cigarettes than non-drinking mothers, or, that the policy makers should encourage a healthy lifestyle during pregnancy in general, not just a reduction in tobacco consumption.

The final step is to aggregate the ATTO, to obtain the APEO and APE. Recall that the APEO can be obtained by integrating ATTO with the density $f_{X|Y}(\cdot|y_0)$. The density itself is not a causal object, but it tends to weigh higher areas with high X and low Y, and vice versa, as newborns who are exposed to a lot of smoking are much more likely to be underweight. Since all individual ATTOs share, however, similar qualitative features, in particular much stronger causal effects for large values of Y than for small values, the results for the APEO are not

qualitatively different.

Estimates of the APEO are plotted in Figure 4 (a). The three dashed curves are again the 2.5-th, 50-th, and 97.5-th percentiles of 500 bootstrap resampled estimates of the APEO. Figure 4 (b) shows the same result in the log scale. We make the following two observations. First, note that the APEO is significantly negative for any subpopulation $Y = y$. Second, the magnitude of APEO is negligible when the actual birth weights are lower (< 2000 grams), but becomes substantial when the actual birth weights are normal to high (> 3000 grams), corroborating previous findings. Finally, the overall APE (integrated over y) is -11.82 grams, in line with other results in the literature.[3]

5 Summary

In this paper we propose outcome conditioned treatment effects as an alternative object of interest in the case of a continuous endogenous treatment. These effects use the entire distribution of the data, to obtain them we do not require assumptions like identification at infinity or strong support conditions. Our focus is on the average partial effect, respectively the average treatment effect on the treated, but both conditional on (levels of) the outcome variable. An application to data on the effect of smoking on the birth weight of newborns shows that they allow a detailed analysis that makes

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

[3]This overall average is slightly smaller in magnitude than the back-of-envelope calculation presented at the beginning of this section since our analysis focuses only of smokers.

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, Vol. 139, No. 1, pp. 4–14.

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

de Chaisemartin, Clément (2012) "Late Again, without Monotonicity," Working Paper, Paris School of Economics.

D'Haultfoeuille, X. and P Février (2011) "Identification of Nonseparable Models with Endogeneity and Discrete Instruments," Working Paper.

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.

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. and Richard Robb Jr. (1985) "Alternative Methods for Evaluating the Impact of Interventions: An Overview," *Journal of Econometrics,* Vol. 30, No. 1, pp. 239–267.

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.

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

Hoderlein, Stefan and Enno Mammen (2009) "Identification and Estimation of Local Average Derivatives in Non-Separable Models without Monotonicity," *Econometrics Journal,* Vol. 12, No. 1, pp. 1–25.

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. (2007) "Nonadditive Models with Endogenous Regressors," in *Advances in Economics and Econometrics: Theory and Applications,* R. Blundell, W. Newey, T. Persson, eds. Cambridge University Press.

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 (2011) "Tighter Bounds in Triangular Systems," *Journal of Econometrics,* Vol. 161, pp. 122–128.

Kasy, Maximilian (2013) "Identification in Continuous Triangular Systems with Unrestricted Heterogeneity," Working Paper, Harvard University.

Lehman, E.L. (1974) "Nonparametrics: Statistical Methods Based on Ranks." Holden-Day, San Francisco.

Lewbel, A. (1999); "Consumer Demand Systems and Household Expenditure", in Pesaran, H. and M. Wickens (Eds.), Handbook of Applied Econometrics, Blackwell Handbooks in economics.

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.

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 .

Torgovitsky, Alexander (2011) "Identification and Estimation of Nonparametric Quantile Regressions with Endogeneity," Working Paper.

A Appendix

This appendix starts out by presenting auxilliary Lemmata that are used in the proofs of the main results, alongside with all the proofs.

A.1 LIV Lemma

The following lemma states that the assumptions intrdocued above are essential for LIV as defined in formula (2.2) to identify a local average partial effect which we later integrate to obtain the causal effects.

Lemma 1 (LIV: Necessary and Sufficient Condition)**.** *Suppose that Assumptions 1, 2, 3 and 4 are satisfied for the model (2.1). Then, the LIV identity*

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

holds. If $h(z,v)=p+v$*, where* $p=\pi(z)$*, and* $\mathbb{E}[V]=0$*, this specializes to*

$$\mathbb{E}\left[\beta(X,A)|P=\pi(z)\right] \;=\; \frac{\partial}{\partial p}\mathbb{E}\left[Y|P=p\right]\Big|_{p=\pi(z)}.$$

Moreover, under the same set of assumptions

$$\mathbb{E}\left[\beta(X,A)|Y=Q_{Y|P}(\tau|\pi(z)),Z=z\right]= \;\frac{\partial}{\partial p}Q_{Y|P}(\tau\mid p)\Big|_{p=\pi(z)}$$

holds.

It is instructive to observe two things about this result: First, the average effect which is identified by LIV is defined for the subpopulation for which $Z=z$. This is different from the binary result, in which case the causal effect is the MTE, and the subpopulation is defined by first stage preference parameters V. This makes LIV in this setup subject to the critique that the subpopulations for which statements can be made are specific to the instrument (e.g., the data at hand and the experimental variation in question, cf. Heckman and Vytlacil (2007)). Second, if we allow the second stage structural function g to take more general forms than the simple linear coefficient model, the necessary and sufficient condition for identification of a causal effect generally amounts to requiring separability for the first stage, i.e., $h(z,v)=\pi(z)+v$, in general.

A.2 Proof of Lemma 1

Using the definition (2.1) of the structural model, we have

$$\mathbb{E}\left[Y|Z=z\right] \;=\; \int\int g(h(z,v),a)f_{AV|Z}(a,v\mid z)dadv \;=\; \int\int g(h(z,v),a)f_{AV}(a,v)dadv$$

where the last equality is due to Assumption 1. Taking derivatives on the both sides produces

$$\begin{aligned}
\frac{d}{dz}\mathbb{E}\left[Y|Z=z\right] &= \int\int \beta(h(z,v),a)\left[\frac{\partial}{\partial z}h(z,v)\right] f_{AV}(a,v)dadv \\
&= \int\int \beta(h(z,v),a)\left[\frac{\partial}{\partial z}h(z,v)\right] f_{AV|Z}(a,v \mid z)dadv \\
&= \mathbb{E}\left[\beta(X,A)\cdot\frac{\partial}{\partial z}h(Z,V)\middle| Z=z\right] \\
&= \mathbb{E}\left[\beta(X,A)|Z=z\right]\cdot\mathbb{E}\left[\frac{\partial}{\partial z}h(Z,V)\middle| Z=z\right] + \mathrm{Cov}\left(\beta(X,A),\ \frac{\partial}{\partial z}h(Z,V)\middle| Z=z\right)
\end{aligned}$$

where the first equality is due to the differentiability of g and h with respect their first arguments as well as the L^1 dominance of the integrand, and the second equality is again due to Assumption 1. The instrument independence also yields

$$\begin{aligned}
\mathbb{E}\left[\frac{\partial}{\partial z}h(Z,V)\middle| Z=z\right] &= \frac{d}{dz}\mathbb{E}\left[X|\,Z=z\right] \qquad \text{and} \\
\mathrm{Cov}\left(\beta(X,A),\ \frac{\partial}{\partial z}h(Z,V)\middle| Z=z\right) &= \mathrm{Cov}\left(\beta(h(z,V),A),\ \frac{\partial}{\partial z}h(z,V)\right).
\end{aligned}$$

Substituting these equalities and rearranging terms under Assumption 2, we obtain

$$\mathbb{E}\left[\beta(X,A)|Z=z\right] = \frac{\frac{d}{dz}\mathbb{E}\left[Y|Z=z\right] - \mathrm{Cov}\left(\beta(h(z,V),A),\ \frac{\partial}{\partial z}h(z,V)\right)}{\frac{d}{dz}\mathbb{E}\left[X|\,Z=z\right]}$$

Therefore, the desired equality holds if and only if Assumption 3 is true.

The last identifying equality in the lemma is proved as follows. We derive the following three auxiliary equations. First,

$$\begin{aligned}
&\Pr[g(X,A)\leqslant Q_{Y|P}(\tau \mid p+\delta) \mid P=p+\delta] - \Pr[g(X,A)\leqslant Q_{Y|P}(\tau \mid p) \mid P=p+\delta] \\
=\ & F_{Y|P}(Q_{Y|P}(\tau \mid p+\delta) \mid p+\delta) - F_{Y|P}(Q_{Y|P}(\tau \mid p) \mid p+\delta) \\
=\ & \delta\left[\frac{\partial}{\partial p}Q_{Y|P}(\tau \mid p)\right] f_{Y|P}(Q_{Y|P}(\tau \mid p) \mid p+\delta) + o(\delta) \qquad \text{(A.1)}
\end{aligned}$$

holds under the differentiability of $F_{Y|P}$ and $Q_{Y|P}$ with respect to y and p, respectively. Second,

$$\begin{aligned}
&\Pr[g(X,A)\leqslant Q_{Y|P}(\tau \mid p) \mid P=p+\delta] - \Pr[g(X+\delta,A)\leqslant Q_{Y|P}(\tau \mid p) \mid P=p] \\
=\ & \Pr[g(p+\delta+V,A)\leqslant Q_{Y|P}(\tau \mid p) \mid P=p+\delta] - \Pr[g(p+\delta+V,A)\leqslant Q_{Y|P}(\tau \mid p) \mid P=p] \\
=\ & \Pr[g(p+\delta+V,A)\leqslant Q_{Y|P}(\tau \mid p)] - \Pr[g(p+\delta+V,A)\leqslant Q_{Y|P}(\tau \mid p)] \ =\ 0, \qquad \text{(A.2)}
\end{aligned}$$

where the second equality is due to Assumption 1. Third, using the short-hand notation $B = \beta(X, A)$, we have

$$\begin{aligned}
& \Pr[g(X+\delta, A) \leqslant Q_{Y|P}(\tau \mid p) \mid P = p] - \Pr[g(X, A) \leqslant Q_{Y|P}(\tau \mid p) \mid P = p] \\
= \; & \Pr[Q_{Y|P}(\tau \mid p) < Y \leqslant Q_{Y|P}(\tau \mid p) - (g(X+\delta, A) - Y) \mid P = p] \\
& - \Pr[Q_{Y|P}(\tau \mid p) - (g(X+\delta, p) - Y) < Y \leqslant Q_{Y|P}(\tau \mid p) \mid P = p] \\
= \; & \Pr[Q_{Y|P}(\tau \mid p) \leqslant Y \leqslant Q_{Y|P}(\tau \mid p) - \delta B \mid P = p] \\
& - \Pr[Q_{Y|P}(\tau \mid p) - \delta B \leqslant Y \leqslant Q_{Y|P}(\tau \mid p) \mid P = p] + o(\delta) \\
= \; & \int_{Q_{Y|P}(\tau|p)}^{\infty} \int_{-\infty}^{-\delta^{-1}[y - Q_{Y|P}(\tau|p)]} f_{YB|P}(y, b \mid p) db dy \\
& - \int_{-\infty}^{Q_{Y|P}(\tau|p)} \int_{-\delta^{-1}[y - Q_{Y|P}(\tau|p)]}^{\infty} f_{YB|P}(y, b \mid p) db dy + o(\delta) \\
= \; & -\delta \int_{-\infty}^{0} b f_{YB|P}(Q_{Y|P}(\tau \mid p), b \mid p) db - \delta \int_{0}^{\infty} b f_{YB|P}(Q_{Y|P}(\tau \mid p), b \mid p) db + o(\delta) \\
= \; & -\delta \mathbb{E}[B \mid Y = Q_{Y|P}(\tau \mid p), P = p] \cdot f_{Y|P}(Q_{Y|P}(\tau \mid p) \mid p) + o(\delta), \qquad \text{(A.3)}
\end{aligned}$$

where the second equality is due to the differentiability of g and $F_{Y|P}$ with respect to x and y, respectively, and the fourth equality is due to change of variables and integration by parts (see Hoderlein and Mammen (2007) for details of this step of computation). Add these three equations (A.1), (A.2) and (A.3) together to get

$$\begin{aligned}
0 \;=\; & \delta \left[\frac{\partial}{\partial p} Q_{Y|P}(\tau \mid p) \right] \cdot f_{Y|P}(Q_{Y|P}(\tau \mid p) \mid p + \delta) \\
& - \delta \mathbb{E}[B \mid Y = Q_{Y|P}(\tau \mid p), P = p] \cdot f_{Y|P}(Q_{Y|P}(\tau \mid p) \mid p) + o(\delta).
\end{aligned}$$

Under the condition that $f_{Y|P}$ is continuous in p, letting $\delta \to 0$ yields

$$\mathbb{E}[\beta(X, A) \mid Y = Q_{Y|P}(\tau \mid p),\; P = p] \;=\; \frac{\partial}{\partial p} Q_{Y|P}(\tau \mid p).$$

Since $Z = z$ and $P = \pi(z)$ are the same events under Assumption 2, setting $p = \pi(z)$ yields the result.

A.3 Proof of Theorem 1

Proof. Part (i) of the theorem follows immediately from the second equation in Lemma 1. Part (ii) of the theorem follows from the last equation in Lemma 1 through the following lines of

argument. Given Assumption 2, applying Lemma 1 with $\tau = F_{Y|Z}(y_0 \mid z)$ yields

$$\mathbb{E}\left[\beta(X,A)|Y=y_0,Z=z\right] \;=\; \frac{\partial}{\partial p}Q_{Y|P}(F_{Y|Z}(y_0 \mid z) \mid p)\Big|_{p=\pi(z)}$$

for all z in the support of $Z \mid Y = y_0$. Integrating both sides of this equality with respect to $F_{Z|Y}(\cdot \mid y_0)$ yields

$$\mathbb{E}\left[\beta(X,A)|Y=y_0\right] \;=\; \int \frac{\partial}{\partial p}Q_{Y|P}(F_{Y|Z}(y_0 \mid p) \mid z)\Big|_{p=\pi(z)} \cdot F_{Z|Y}(dz \mid y_0)$$

as desired. □

A.4 MTE Lemma

In this section, we show that we can identify causal effects of the form $\mathbb{E}\left[\beta(x,A)|X=x,Z=z\right]$ and $\mathbb{E}\left[\beta(x,A)|X=x,Y=y,Z=z\right]$, which involve all the conditioning variables (Z,X), respectively the entire distribution of the data (X,Y,Z).

Lemma 2 (Mean MTE)**.** *Suppose that Assumptions 1 (i), 4, 5 and 6 are satisfied for the model (2.1). Then, the MTE is given by:*

$$\mathbb{E}\left[\beta(x,A)|X=x,Z=z\right] \;=\; \frac{\partial}{\partial x}\mathbb{E}\left[Y|X=x,Z=z\right] - \rho(z,x)\frac{\partial}{\partial z}\mathbb{E}\left[Y|X=x,Z=z\right],$$

where $v = \upsilon(z,x)$.

The LAR relates this conditioning on unobservables to an observable conditioning set (X,Z). Finally, the last equality relates the LAR to an object that is entirely a function of the joint distribution of the data, thus permitting sample counterparts estimation. Note the structure on the right hand side as the derivative of the mean regression $\frac{\partial}{\partial x}\mathbb{E}\left[Y|X=x,Z=z\right]$, which due to the correlation between A, X and Z is not equal to the structural effect of interest, and a second term, which accounts for the selection distortion, as in Heckman (1979). Note that this selection distortion crucially depends on $\rho(z,x) = \left[\frac{\partial \upsilon(z,x)}{\partial x}\right] / \left[\frac{\partial \upsilon(z,x)}{\partial z}\right]$, which relates X and Z via the first stage structure.

The following lemma generalizes LAR and MTE by linking it similar structures as in the previous lemma, but now involving the entire distribution of the data:

Lemma 3 (Quantile MTE)**.** *Suppose that Assumptions 1 (i), 4, 5 and 6 are satisfied for the model (2.1). Then, the quantile LAR is given by*

$$\mathbb{E}\left[\beta(x,A)|Y=Q_{Y|XZ}(\tau\mid x,z),X=x,Z=z\right] \;=\; \frac{\partial}{\partial x}Q_{Y|XZ}(\tau\mid x,z)-\rho(z,x)\cdot\frac{\partial}{\partial z}Q_{Y|XZ}(\tau\mid x,z),$$

and the quantile MTE is

$$\begin{aligned}\mathbb{E}\left[\beta(x,A)|Y=Q_{Y|XZ}(\tau\mid x,z),X=x,V=v\right] \;=\; \mathbb{E}\Bigg[&\frac{\partial}{\partial x}Q_{Y|XZ}(F_{Y|XZ}(Y\mid X,Z)\mid x,Z)\Big|_{x=X}\\ &-\rho(Z,X)\cdot\frac{\partial}{\partial z}Q_{Y|XZ}(F_{Y|XZ}(Y\mid X,Z)\mid X,z)\Big|_{z=Z}\Bigg|\,Y=Q_{Y|XZ}(\tau\mid x,z),X=x,V=\nu\Bigg],\end{aligned}$$

where $v=\upsilon(z,x)$.

The first equality characterizes the LAR through an expression involving the entire distribution of the data. The LAR has not direct interpretation, because it depends on the exogenous incentive Z. However, recall that by integrating out Z from the LAR we have previously obtained an analogous effect to the average treatment effect on the treated (ATT). Likewise, integrating out Z from the second object yields a quantile version of this effect. If we compensate τ such that $Q_{Y|XZ}(\tau\mid x,z)=y$ for any value of z, by integrating out Z we obtain the effect $\mathbb{E}\left[\beta(x,A)|Y=y,X=x\right]$. This is the ATT for the population choosing treatment intensity X, and having, say, a large value of Y. Moreover, observe the parallels between this equality and the right hand side of Lemma 2. This parallel is not trivial, and in our mind an expression of the deep structure of the problem. It is not easily explained, because differentiating quantile regressions does not directly transform into moments of derivatives, and the proofs follow very different steps.

The second part of this lemma provides the MTE. It follows from the first part and the fact that $\sigma(X,Y,Z)\supseteq\sigma(X,Y,V)$. In an ideal world, we would like to have $\beta(x,a)$, i.e., know $A=a$. However, due to the complex nature of A this effect is not available. However, parts of A are given by V, and other parts of A may be obtained, if for various values of X, the response Y is evaluated. As a simple example, suppose that $Y=\phi(X,A_1,A_2)$ and assume that $A_1=V$. as well as $A_2\perp X|V$, $A_2\backsim U[0,1]$, as well as ϕ strictly monotonic in V. Then,

$$\mathbb{E}\left[\beta(x,A_1,A_2)|Y=Q_{Y|XZ}(\tau\mid x,z),X=x,V=v\right]=\beta(x,\tau,v).$$

However, if A has more components, this will generally be an average.

As with all lemmas in this paper, the main results equate an average structural derivative on the left hand side with a statistical object on the right hand side. The first (and consequently the second) equality is, to the best of our knowledge, novel. The right-hand side of the first equation in Lemma 3 turns out to be similar to the equality

$$\frac{\partial}{\partial x}Q_{Y|XV}(\tau \mid x, v) \;=\; \frac{\partial}{\partial x}Q_{Y|XZ}(\tau \mid x, z) + \frac{\frac{\partial}{\partial z}Q_{Y|XZ}(\tau \mid x, z)}{\frac{\partial}{\partial z}Q_{X|Z}(v \mid z)} \tag{A.4}$$

derived by Imbens and Newey (2009; Theorem 2). In their model, it is directly the object $\frac{\partial}{\partial x}Q_{Y|XV}(y \mid x, v)$ which is of interest, whereas in our approach identifies a local average of the structural objects $\beta(x, A)$. Note that quantile partial effects do not generally represent structural partial effects unless rank invariance is assumed. Our result therefore adds a structural interpretation to (A.4) in the case of a high dimensional structural unobservable.

A.5 Proof of Lemma 2

Proof. We derive the following two auxiliary equations. First,

$$\begin{aligned}\frac{\partial}{\partial x}\mathbb{E}[Y \mid X = x, Z = z] \;&=\; \partial_x \int g(x,a) f_{A|XZ}(a \mid x, z)da \;=\; \frac{\partial}{\partial x}\int g(x,a) f_{A|VZ}(a \mid v(z,x), z)da \\ &=\; \mathbb{E}[\beta(X, A) \mid V = v(z,x), Z = z] \\ &+\; \frac{\partial}{\partial x}v(z,x) \cdot \mathbb{E}\left[g(X,A)\frac{\partial}{\partial v}\log f_{A|VZ}(A \mid V, Z)\middle| V = v(z,x), Z = z\right],\end{aligned}$$

where the second equality is due to Assumption 5 and the third equality is due to the differentiability of g with respect to x as well as the L^1 dominance of the integrand. Second, a similar calculation yields

$$\frac{\partial}{\partial z}\mathbb{E}[Y \mid X = x, Z = z] = \frac{\partial}{\partial z}v(z,x) \cdot \mathbb{E}\left[g(X,A)\frac{\partial}{\partial v}\log f_{A|VZ}(A \mid V, Z)\middle| V = v(z,x), Z = z\right],$$

where the instrumental independence in model (2.1) was used to vanish $\frac{\partial}{\partial z}f_{A|VZ}$. Combining the above two equations and rearranging by Assumption 6 yield the desired result. □

A.6 Proof of Lemma 3

Proof. Assumptions 5 and 6 provide the parameterized curve $h \mapsto (h, \delta_z(h)) =: (\delta_x, \delta_z)$ that solves the implicit function equation $v(z + \delta_z, x + \delta_x) - v(z, x) = 0$ of a smooth submanifold

in a neighborhood of $h = 0$. Furthermore, $\delta_z(0) = 0$ and $(\delta_x, \delta_z) \to 0$ as $h \to 0$. By these properties, we have

$$\frac{\delta_z}{\delta_x} = -\frac{\frac{\partial}{\partial x}\upsilon(z,x)}{\frac{\partial}{\partial z}\upsilon(z,x)} + o(1) \quad \text{as } h \to 0. \tag{A.5}$$

Next, we derive the following four auxiliary equations. First,

$$\begin{aligned}
& \Pr[g(x+\delta_x, A) \leqslant Q_{Y|XZ}(\tau \mid x+\delta_x, z+\delta_z) \mid X = x+\delta_x, Z = z+\delta_z] \\
& - \Pr[g(x+\delta_x, A) \leqslant Q_{Y|XZ}(\tau \mid x, z+\delta_z) \mid X = x+\delta_x, Z = z+\delta_z] \\
= \; & F_{Y|XZ}(Q_{Y|XZ}(\tau \mid x+\delta_x, z+\delta_z) \mid x+\delta_x, z+\delta_z) \\
& - F_{Y|XZ}(Q_{Y|XZ}(\tau \mid x, z+\delta_z) \mid x+\delta_x, z+\delta_z) \\
= \; & \delta_x \frac{\partial}{\partial x} Q_{Y|XZ}(\tau \mid x, z) f_{Y|XZ}(Q_{Y|XZ}(\tau \mid x, z) \mid x+\delta_x, z+\delta_z) + o(\delta_x),
\end{aligned} \tag{A.6}$$

where the last equality is due to the differentiability of $Q_{Y|XZ}$ and $F_{Y|XZ}$ with respect to x and y, respectively. Second, similar lines of calculations yield

$$\begin{aligned}
& \Pr[g(x+\delta_x, A) \leqslant Q_{Y|XZ}(\tau \mid x, z+\delta_z) \mid X = x+\delta_x, Z = z+\delta_z] \\
& - \Pr[g(x+\delta_x, A) \leqslant Q_{Y|XZ}(\tau \mid x, z) \mid X = x+\delta_x, Z = z+\delta_z] \\
= \; & F_{Y|XZ}(Q_{Y|XZ}(\tau \mid x, z+\delta_z) \mid x+\delta_x, z+\delta_z) \\
& - F_{Y|XZ}(Q_{Y|XZ}(\tau \mid x, z) \mid x+\delta_x, z+\delta_z) \\
= \; & \delta_z \frac{\partial}{\partial z} Q_{Y|XZ}(\tau \mid x, z) f_{Y|XZ}(Q_{Y|XZ}(\tau \mid x, z) \mid x+\delta_x, z+\delta_z) + o(\delta_z),
\end{aligned} \tag{A.7}$$

under the differentiability of $Q_{Y|XZ}$ with respect to z. Third,

$$\begin{aligned}
& \Pr[g(x+\delta_x, A) \leqslant Q_{Y|XZ}(\tau \mid x, z) \mid X = x+\delta_x, Z = z+\delta_z] \\
& - \Pr[g(x+\delta_x, A) \leqslant Q_{Y|XZ}(\tau \mid x, z) \mid X = x, Z = z] \\
= \; & \Pr[g(x+\delta_x, A) \leqslant Q_{Y|XZ}(\tau \mid x, z) \mid Z = z+\delta_z, V = \upsilon(z+\delta_z, x+\delta_x)] \\
& - \Pr[g(x+\delta_x, A) \leqslant Q_{Y|XZ}(\tau \mid x, z) \mid Z = z, V = \upsilon(z, x)] \\
= \; & \Pr[g(x+\delta_x, A) \leqslant Q_{Y|XZ}(\tau \mid x, z) \mid Z = z+\delta_z, V = \upsilon(z, x)] \\
& - \Pr[g(x+\delta_x, A) \leqslant Q_{Y|XZ}(\tau \mid x, z) \mid Z = z, V = \upsilon(z, x)] \;=\; 0,
\end{aligned} \tag{A.8}$$

where the first equality is due to Assumption 5, the second equality is due to the definition of (δ_x, δ_z), and the last equality is due to Assumption 1 (i). Fourth, with the short-hand notation

$B := \beta(X, A)$, we have

$$
\begin{aligned}
& \Pr[g(x+\delta_x, A) \leqslant Q_{Y|XZ}(\tau \mid x, z) \mid X = x, Z = z] \\
& - \Pr[g(x, A) \leqslant Q_{Y|XZ}(\tau \mid x, z) \mid X = x, Z = z] \\
= \; & \Pr[Q_{Y|XZ}(\tau \mid x, z) < Y \leqslant Q_{Y|XZ}(\tau \mid x, z) - (g(x+\delta_x, A) - Y) \mid X = x, Z = z] \\
& - \Pr[Q_{Y|XZ}(\tau \mid x, z) - (g(x+\delta_x, z) - Y) < Y \leqslant Q_{Y|XZ}(\tau \mid x, z) \mid X = x, Z = z] \\
= \; & \Pr[Q_{Y|XZ}(\tau \mid x, z) \leqslant Y \leqslant Q_{Y|XZ}(\tau \mid x, z) - \delta_x B \mid X = x, Z = z] \\
& - \Pr[Q_{Y|XZ}(\tau \mid x, z) - \delta_x B \leqslant Y \leqslant Q_{Y|XZ}(\tau \mid x, z) \mid X = x, Z = z] + o(\delta_x) \\
= \; & \int_{Q_{Y|XZ}(\tau|x,z)}^{\infty} \int_{-\infty}^{-\delta_x^{-1}[y - Q_{Y|XZ}(\tau|x,z)]} f_{YB|XZ}(y, b \mid x, z) db dy \\
& - \int_{-\infty}^{Q_{Y|XZ}(\tau|x,z)} \int_{-\delta_x^{-1}[y - Q_{Y|XZ}(\tau|x,z)]}^{\infty} f_{YB|XZ}(y, b \mid x, z) db dy + o(\delta_x) \\
= \; & -\delta_x \int_{-\infty}^{0} b f_{YB|XZ}(Q_{Y|XZ}(\tau \mid x, z), b \mid x, z) db \\
& -\delta_x \int_{0}^{\infty} b f_{YB|XZ}(Q_{Y|XZ}(\tau \mid x, z), b \mid x, z) db + o(\delta_x) \\
= \; & -\delta_x \mathbb{E}[B \mid Y = Q_{Y|XZ}(\tau \mid x, z), X = x, Z = z] f_{Y|XZ}(Q_{Y|XZ}(\tau \mid x, z) \mid x, z) + o(\delta_x), \quad \text{(A.9)}
\end{aligned}
$$

where the second equality is due to the differentiability of g and $F_{Y|XZ}$ with respect to x and y, respectively, and the fourth equality is due to change of variables and integration by parts (see Hoderlein and Mammen (2007) for details of this step of computation). Add the above four equations (A.6), (A.7), (A.8) and (A.9) together to get

$$
\begin{aligned}
0 \; = \; & \delta_x \frac{\partial}{\partial x} Q_{Y|XZ}(\tau \mid x, z) f_{Y|XZ}(Q_{Y|XZ}(\tau \mid x, z) \mid x + \delta_x, z + \delta_z) \\
& + \delta_z \frac{\partial}{\partial z} Q_{Y|XZ}(\tau \mid x, z) f_{Y|XZ}(Q_{Y|XZ}(\tau \mid x, z) \mid x + \delta_x, z + \delta_z) \\
& - \delta_x \mathbb{E}[B \mid Y = Q_{Y|XZ}(\tau \mid x, z), X = x, Z = z] f_{Y|XZ}(Q_{Y|XZ}(\tau \mid x, z) \mid x, z) + o(\delta_x) + o(\delta_z).
\end{aligned}
$$

The desired result follows from this equation together with Equation (A.5), Assumption 5, and the differentiability of $f_{Y|XZ}$ with respect to the conditioning variables (x, z). □

A.7 Proof of Theorem 2

Proof. Part (i) of the theorem follows immediately from Lemma 2. Part (ii) of the theorem follows from Lemma 3 through the following lines of argument. Under Assumption 7 (i′) and

(ii′), applying Lemma 3 with $x = x_0$ and $\tau = F_{Y|XZ}(y_0 \mid x_0, z)$ yields

$$\begin{aligned}
&\mathbb{E}\left[\beta(x_0, A)|Y = y_0, X = x_0, Z = z\right] \\
= \;& \frac{\partial}{\partial \xi} Q_{Y|XZ}(F_{Y|XZ}(y_0 \mid x_0, z) \mid \xi, z)\Big|_{\xi = x_0} - \rho(z, x_0) \cdot \frac{\partial}{\partial \zeta} Q_{Y|XZ}(F_{Y|XZ}(y_0 \mid x_0, z) \mid x_0, \zeta)\Big|_{\zeta = z},
\end{aligned}$$

for a.s. z in the support of $Z \mid Y = y_0, X = x_0$. Integrating both sides of this equality with respect to $F_{Z|YX}(\cdot \mid y_0, x_0)$ yields

$$\begin{aligned}
\mathbb{E}\left[\beta(x_0, A)|Y = y_0, X = x_0\right] \;=\; & \int \frac{\partial}{\partial \xi} Q_{Y|XZ}(F_{Y|XZ}(y_0 \mid x_0, z) \mid \xi, z)\Big|_{\xi = x_0} \cdot F_{Z|YX}(dz \mid y_0, x_0) \\
& - \int \rho(z, x_0) \cdot \frac{\partial}{\partial \zeta} Q_{Y|XZ}(F_{Y|XZ}(y_0 \mid x_0, z) \mid x_0, \zeta)\Big|_{\zeta = z} \cdot F_{Z|YX}(dz \mid y_0, x_0)
\end{aligned}$$

which proves the first equality in the theorem. The second equality immediately follows from this result. □

A.8 Estimation of the Outcome Conditioned Treatment Effects

We take the identification results of Theorem 2 (ii) to propose sample-counterpart estimators of the outcome conditioned ATE and ATT. For simplicity, we specialize in the case of $\rho(z, x) \equiv 1/\frac{\partial}{\partial \zeta} Q_{X|Z}(F_{X|Z}(x \mid z) \mid \zeta)|_{\zeta = z}$, but similar arguments continue to apply to other cases.

A.8.1 Estimation of the Outcome Conditioned ATT

First, consider the leave-one-out Nadaraya-Watson estimators

$$\begin{aligned}
\widehat{F}_{Y|XZ}(y_0 \mid x_0, Z_i) &= \frac{\sum_{j \neq i} \mathbb{1}\{Y_j \leqslant y_0\} K\left(\frac{X_j - x_0}{h_x}\right) K\left(\frac{Z_j - Z_i}{h_z}\right)}{\sum_{j \neq i} K\left(\frac{X_j - x_0}{h_x}\right) K\left(\frac{Z_j - Z_i}{h_z}\right)} \qquad \text{and} \\
\widehat{F}_{X|Z}(x_0 \mid Z_i) &= \frac{\sum_{j \neq i} \mathbb{1}\{X_j \leqslant x_0\} K\left(\frac{Z_j - Z_i}{h_z}\right)}{\sum_{j \neq i} K\left(\frac{Z_j - Z_i}{h_z}\right)}
\end{aligned}$$

of the conditional CDF values $F_{Y|XZ}(y_0 \mid x_0, Z_i)$ and $F_{X|Z}(x_0 \mid Z_i)$ for each i. Given these conditional CDFs, we in tern let $\widehat{B}_i$, $\widehat{C}_i$, and $\widehat{b}_i$ denote estimators of

$$\begin{aligned}
B_i &= \frac{\partial}{\partial \xi} Q_{Y|XZ}(F_{Y|XZ}(y_0 \mid x_0, Z_i) \mid \xi, Z_i)\Big|_{\xi = x_0}, \\
C_i &= \frac{\partial}{\partial \zeta} Q_{Y|XZ}(F_{Y|XZ}(y_0 \mid x_0, Z_i) \mid x_0, \zeta)\Big|_{\zeta = Z_i}, \qquad \text{and} \\
b_i &= \frac{\partial}{\partial \zeta} Q_{X|Z}(F_{X|Z}(x_0 \mid Z_i) \mid \zeta)\Big|_{\zeta = Z_i} \qquad \left(= \frac{1}{\rho(Z_i, x_0)}\right),
\end{aligned}$$

respectively. They can be consistently estimated by the leave-one-out local-linear quantile regression estimators defined by

$$\begin{aligned}\left(\widehat{A}_i, \widehat{B}_i, \widehat{C}_i\right) &= \arg\max_{(A,B,C)} \sum_{j\neq i} K\left(\frac{X_j - x_0}{h_x}\right) K\left(\frac{Z_j - Z_i}{h_z}\right) (Y_j - A - BX_j - CZ_j) \times \\ &\qquad \left[\widehat{F}_{Y|XZ}(y_0 \mid x_0, Z_i) - \mathbb{1}\left\{Y_j - A - BX_j - CZ_j < 0\right\}\right]\end{aligned}$$

and

$$\begin{aligned}\left(\widehat{a}_i, \widehat{b}_i\right) &= \arg\max_{(a,b)} \sum_{j\neq i} K\left(\frac{Z_j - Z_i}{h_z}\right) (X_j - a - bZ_j) \times \\ &\qquad \left[\widehat{F}_{X|Z}(x_0 \mid Z_i) - \mathbb{1}\left\{X_j - a - bZ_j < 0\right\}\right]\end{aligned}$$

Using these local linear estimates, we can consistently estimate the outcome conditioned ATT by the Nadaraya-Watson method:

$$\widehat{\mathbb{E}}\left[\beta(x_0, A) \mid Y = y_0, X = x_0\right] = \frac{\sum_i \left[\widehat{B}_i - \frac{\widehat{C}_i}{\widehat{b}_i}\right] K\left(\frac{Y_i - y_0}{h_y}\right) K\left(\frac{X_i - x_0}{h_x}\right)}{\sum_i K\left(\frac{Y_i - y_0}{h_y}\right) K\left(\frac{X_i - x_0}{h_x}\right)}.$$

A.8.2 Estimation of the Outcome Conditioned ATE

Consider the leave-one-out Nadaraya-Watson estimators

$$\begin{aligned}\widehat{F}_{Y|XZ}(y_0 \mid X_i, Z_i) &= \frac{\sum_{j\neq i} \mathbb{1}\{Y_j \leqslant y_0\} K\left(\frac{X_j - X_i}{h_x}\right) K\left(\frac{Z_j - Z_i}{h_z}\right)}{\sum_{j\neq i} K\left(\frac{X_j - X_i}{h_x}\right) K\left(\frac{Z_j - Z_i}{h_z}\right)} \qquad \text{and} \\ \widehat{F}_{X|Z}(X_i \mid Z_i) &= \frac{\sum_{j\neq i} \mathbb{1}\{X_j \leqslant X_i\} K\left(\frac{Z_j - Z_i}{h_z}\right)}{\sum_{j\neq i} K\left(\frac{Z_j - Z_i}{h_z}\right)}\end{aligned}$$

of the conditional CDF values $F_{Y|XZ}(y_0 \mid X_i, Z_i)$ and $F_{X|Z}(X_i \mid Z_i)$ for each i. Given these conditional CDFs, we in tern let $\widehat{B}'_i$, $\widehat{C}'_i$, and $\widehat{b}'_i$ denote estimators of

$$\begin{aligned}B'_i &= \frac{\partial}{\partial \xi} Q_{Y|XZ}(F_{Y|XZ}(y_0 \mid X_i, Z_i) \mid \xi, Z_i)\Big|_{\xi = X_i}, \\ C'_i &= \frac{\partial}{\partial \zeta} Q_{Y|XZ}(F_{Y|XZ}(y_0 \mid X_i, Z_i) \mid X_i, \zeta)\Big|_{\zeta = Z_i}, \qquad \text{and} \\ b'_i &= \frac{\partial}{\partial \zeta} Q_{X|Z}(F_{X|Z}(X_i \mid Z_i) \mid \zeta)\Big|_{\zeta = Z_i} \qquad \left(= \frac{1}{\rho(Z_i, X_i)}\right),\end{aligned}$$

respectively. They can be consistently estimated by the leave-one-out local-linear quantile regression estimators defined by

$$\begin{aligned}\left(\widehat{A}_i', \widehat{B}_i', \widehat{C}_i'\right) &= \arg\max_{(A,B,C)} \sum_{j\neq i} K\left(\frac{X_j - X_i}{h_x}\right) K\left(\frac{Z_j - Z_i}{h_z}\right) (Y_j - A - BX_j - CZ_j) \times \\ &\qquad \left[\widehat{F}_{Y|XZ}(y_0 \mid X_i, Z_i) - \mathbb{1}\left\{Y_j - A - BX_j - CZ_j < 0\right\}\right]\end{aligned}$$

and

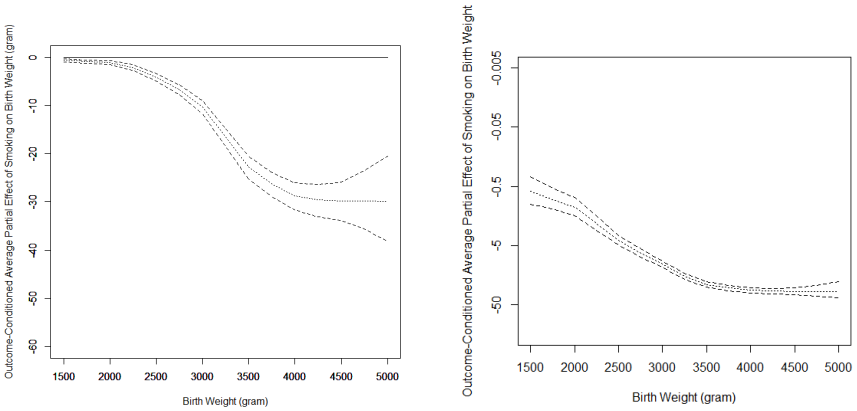
$$\begin{aligned}\left(\widehat{a}_i', \widehat{b}_i'\right) &= \arg\max_{(a,b)} \sum_{j\neq i} K\left(\frac{Z_j - Z_i}{h_z}\right) (X_j - a - bZ_j) \times \\ &\qquad \left[\widehat{F}_{X|Z}(X_i \mid Z_i) - \mathbb{1}\left\{X_j - a - bZ_j < 0\right\}\right]\end{aligned}$$

Using these local linear estimates, we can consistently estimate the outcome conditioned ATE by the Nadaraya-Watson method:

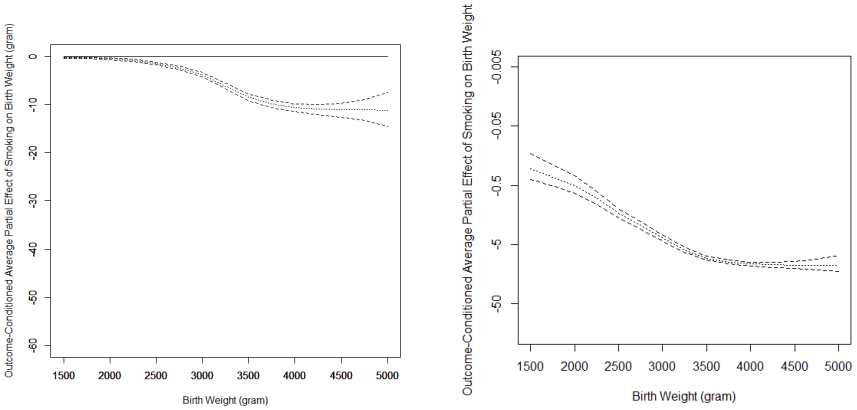
$$\widehat{\mathbb{E}}\left[\beta(X, A) \mid Y = y_0\right] = \frac{\sum_i \left[\widehat{B}_i' - \frac{\widehat{C}_i'}{\widehat{b}_i'}\right] K\left(\frac{Y_i - y_0}{h_y}\right)}{\sum_i K\left(\frac{Y_i - y_0}{h_y}\right)}.$$

A.9 Figures

ATTO at $X = 5$ Log-Scaled ATTO at $X = 5$



ATTO at $X = 10$ Log-Scaled ATTO at $X = 10$



ATTO at $X = 20$ Log-Scaled ATTO at $X = 20$

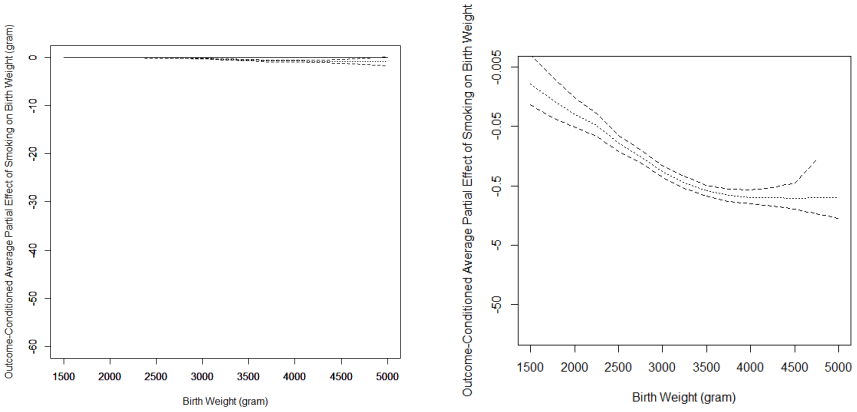


Figure 1: ATTO $\mathbb{E}[\beta(X, A) \mid Y = y, X = x]$ across $y \in [1500, 5000]$ and $x \in \{5, 10, 20\}$. The dashed curves represent the 2.5-th, 50-th, and 97.5-th percentiles of bootstrap estimates. The left figures are plotted in raw values, whereas the right figures are plotted in the logarithmic scale of the raw values.

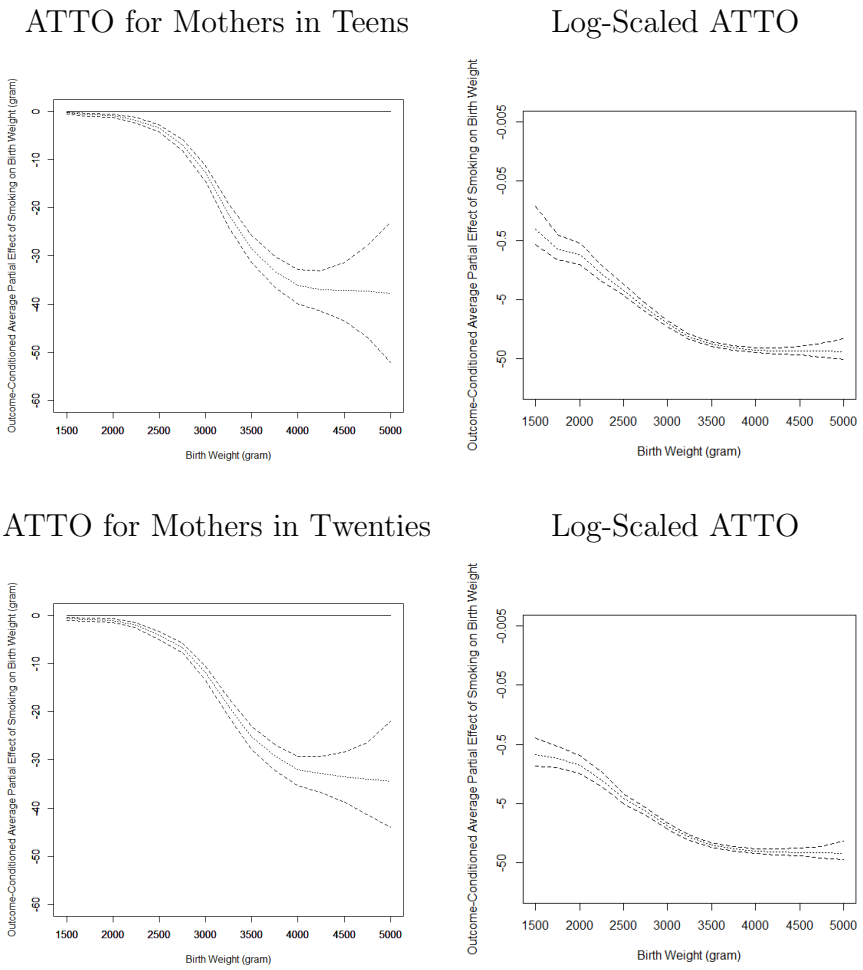
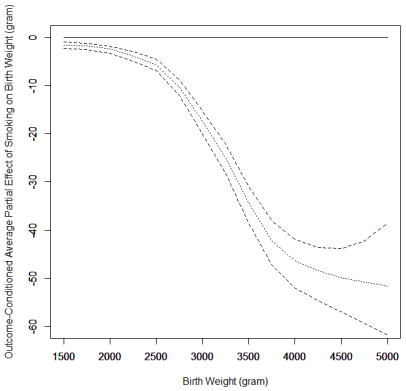


Figure 2: ATTO $\mathbb{E}[\beta(X, A) \mid Y = y, X = x]$ across $y \in [1500, 5000]$ and $x = 5$ for the subpopulation of first births. The top and bottom rows restrict to mothers aged teens and twenties, respectively. The dashed curves represent the 2.5-th, 50-th, and 97.5-th percentiles of bootstrap estimates. The left figures are plotted in raw values, whereas the right figures are plotted in the logarithmic scale of the raw values.

ATTO for Drinking Mothers



Log-Scaled ATTO

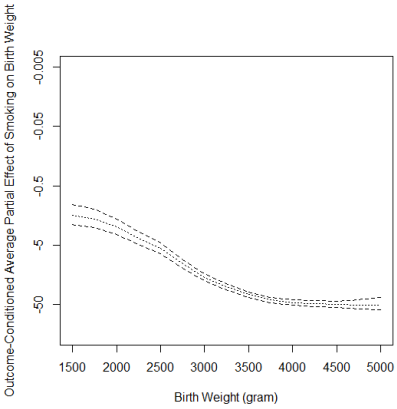
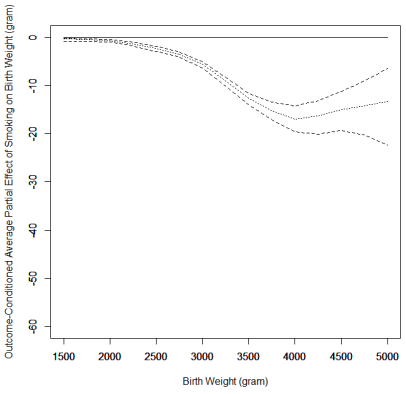


Figure 3: ATTO $\mathbb{E}[\beta(X, A) \mid Y = y, X = x]$ across $y \in [1500, 5000]$ and $x = 5$ for the subpopulation of first births and drinking mothers. The dashed curves represent the 2.5-th, 50-th, and 97.5-th percentiles of bootstrap estimates. The left figures are plotted in raw values, whereas the right figures are plotted in the logarithmic scale of the raw values.

(a) APEO in Raw Values



(b) APEO in Log Scale

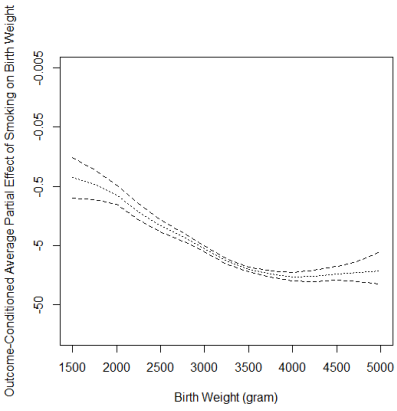


Figure 4: APEO $\mathbb{E}[\beta(X, A) \mid Y = y]$ across $y \in [1500, 5000]$. The dashed curves represent the 2.5-th, 50-th, and 97.5-th percentiles of bootstrap estimates. (a) is plotted in raw values, whereas (b) is plotted in the logarithmic scale of the raw values.