Quantitative Economics 6 (2015), 749–794 1759-7331/20150749

Estimating nonseparable models with mismeasured endogenous variables

Suyong Song
Department of Economics, University of Iowa

Susanne M. Schennach
Department of Economics, Brown University

Halbert White
Department of Economics, University of California, San Diego

We study the identification and estimation of covariate-conditioned average marginal effects of endogenous regressors in nonseparable structural systems when the regressors are mismeasured. We control for the endogeneity by making use of covariates as control variables; this ensures conditional independence between the endogenous causes of interest and other unobservable drivers of the dependent variable. Moreover, we recover distributions of the underlying true causes from their error-laden measurements to deliver consistent estimators. We obtain uniform convergence rates and asymptotic normality for estimators of covariate-conditioned average marginal effects, faster convergence rates for estimators of their weighted averages over instruments, and root-n consistency and asymptotic normality for estimators of their weighted averages over control variables and regressors. We investigate their finite-sample behavior using Monte Carlo simulation and apply new methods to study the impact of family income on child achievement measured by math and reading scores, using a matched mother–child subsample of the National Longitudinal Survey of Youth. Our findings suggest that these effects are considerably larger than previously recognized, and depend on parental abilities and family income. This underscores the importance of measurement errors, endogeneity of family income, nonlinearity of income effects, and interactions between causes of child achievement.

Suyong Song: suyong-song@uiowa.edu
Susanne M. Schennach: smschenn@brown.edu
Hal White passed away before the first version was submitted to *Quantitative Economics*, and we were immensely saddened by his passing. He was a great advisor, colleague and friend, and we deeply miss him. Hal was a Chancellor's Associates Distinguished Professor of Economics at University of California, San Diego, a Fellow of the American Academy of Arts and Sciences and the Econometric Society and a Guggenheim Fellow. The authors would like to express their appreciation to Jin Seo Cho, Gordon Dahl, Graham Elliott, Anthony Gamst, Stephen Hoderlein, Joel Horowitz, Arthur Lewbel, Chuck Manski, Dimitris Politis, James Powell, Andres Santos, Yixiao Sun, and Quang Vuong, as well as seminar audiences at various universities, at the Midwest Econometrics Group 2010, and at the Econometric Society 2011 Winter Meetings for helpful comments and discussions.

DOI: 10.3982/QE275

Keywords. Causal effects, child development, endogeneity, measurement error, nonparametric estimation, nonseparable.

JEL classification. C13, C14, C31.

1. Introduction

Increasing attention is being devoted to more realistic economic models in which the simplifying assumptions of linearity, separability, or exogeneity are not imposed (e.g., Chesher (2005), Altonji and Matzkin (2005), Chernozhukov and Hansen (2005), Imbens (2007), Hoderlein and Mammen (2007), Florens, Heckman, Meghir, and Vytlacil (2008), Chernozhukov, Imbens, and Newey (2007), Imbens and Newey (2009), Chalak and White (2011a, 2011b), Hahn and Ridder (2011), Hoderlein (2011), and the references therein). At the same time, increasingly powerful techniques to properly handle data problems such as measurement error are being developed (e.g., Hausman, Newey, Ichimura, and Powell (1991), Hausman, Newey, and Powell (1995), Hsiao and Wang (2000), Li (2002), Schennach (2004a, 2004b, 2007), Hu and Ridder (2012), Chen, Hong, and Tamer (2005), Mahajan (2006), Lewbel (2007), Chen, Hong, and Tarozzi (2008), Hu (2008), Hu and Schennach (2008)). However, there is still a striking absence of methods able to simultaneously accommodate both endogeneity in nonseparable models and measurement error. This is the gap this paper aims to fill. It is well known since the work of Amemiya (1985) and Hsiao (1989) that correcting for both endogeneity and measurement error in the same variable is not straightforward, since valid instruments to handle endogeneity are typically insufficient to control for measurement error in nonlinear models.

To obtain our results, we rely on a conditional independence assumption between the endogenous causes and other unobservable drivers of the dependent variable to ensure structural identification, as considered, for example, by Altonji and Matzkin (2005), Hoderlein and Mammen (2007), Chalak and White (2011a, 2011b), and Hoderlein (2011). This framework shows that the identification of effects with a genuine structural interpretation can be achieved via a suitable average of a nonparametric regression of the outcome variable on the endogenous regressor and the "conditioning instruments." We generalize this approach to allow for mismeasured regressors, which we handle via repeated measurements (e.g., Hausman et al. (1991), Hausman, Newey, and Powell (1995), Hsiao and Wang (2000), Li (2002), Schennach (2004a, 2004b, 2007)). The key goals of the present work are (i) to determine the specific assumptions that enable identification of structural effects (a task made nontrivial due to the interaction between the measurement errors and the various disturbances in the model) and (ii) to develop suitable estimators and corresponding asymptotic theory for inference (which is challenging due to the large number of mutually interacting nonparametric quantities involved).

Examples of economic applications where endogeneity, nonseparability, and measurement error are simultaneously present abound. Engel curves provide one typical example, since purchased quantities may be endogenously determined and misreported, and consumers may have preferences that do not obey a simple additive structure. More generally, demand systems with heterogenous consumers exhibit similar features. The study of treatment effects under potentially nonseparable heterogeneity is another area where our framework applies, in the case where the possible levels of treatment are continuous, are potentially mismeasured, and cannot be considered exogenous (because the outcomes are correlated with determinants of the treatment level). We illustrate the use of our methodology in this context via an application to the study of the impact of family income on child achievement. Endogeneity arises because both child achievement and parental income are in part determined by innate abilities, which exhibit correlation across generations. This problem is addressed by using a measure of parental abilities as a conditioning instrument. In this application, income data are also notoriously mismeasured and outcomes may depend on individual heterogeneity in a nonseparable way.

Our framework is perhaps most related to that proposed in Altonji and Matzkin (2005) for estimating nonseparable models with observable endogenous regressors and unobservable errors in cross-section and panel data. One of their objects of interest is a local average response, a type of average marginal effect. Here, we consider a similar structure for cross-section data, but allow for the true endogenous cause of interest to be unobservable. Instead, we suppose we have available two error-laden measurements of the underlying true causes. We then recover distributions of the true causes from them. Even though our structural relations are nonparametric and nonseparable, we show that we can identify and estimate objects of interest, specifically, covariate-conditioned average marginal effects and weighted averages of covariate-conditioned average marginal effects.

In a related work, Schennach, White, and Chalak (2012) (SWC hereafter) study identification and estimation of average marginal effects in nonseparable structural systems. They consider estimating causal effects from a nonseparable data generating process using either an observed standard exogenous instrument or an unobserved exogenous instrument for which two error-laden measurements are available. However, they assume that the endogenous cause is observable. As SWC show, however, in the absence of certain separability assumptions, local indirect least squares methods generally cannot recover the average marginal effect. We complement the analysis of SWC by treating the case in which the instrument is no longer exogenous, but is instead a control variable, and by allowing for measurement errors in an endogenous cause. Here, this control variable is observable, but the endogenous cause of interest is unobservable. Our approach is, therefore, complementary to SWC along more than one conceptual dimension. It is worthwhile noting that, in contrast to the negative result obtained by SWC for exogenous instruments, we show that use of the control variable enables the recovery of various structurally meaningful objects. Furthermore, because of the use of the conditional independence, our asymptotic treatment for the proposed estimator is more challenging than SWC or Schennach (2004a, 2004b) for additively separable models; conditioning has nontrivial implications on the asymptotic treatment so that one cannot merely invoke their approaches.

This paper is organized as follows. We first study nonparametric estimation of generic aspects of the structure of interest; we then construct specific objects of interest from them. This includes such objects as the average counterfactual response function, the covariate-conditioned average marginal effect, Altonji and Matzkin's (2005) "local average response," corresponding to the effect of treatment on the treated for continuous treatments (Florens et al. (2008)), and the average treatment effect. We establish uniform convergence rates and asymptotic normality for estimators of covariate-conditioned average marginal effects, faster convergence rates for estimators of their weighted averages over instruments, and $\sqrt{n}$ consistency and asymptotic normality for estimators of their weighted averages over instruments and regressors. We also propose a practical procedure to select the optimal bandwidths.

The practical usefulness of the proposed estimator is illustrated by both Monte Carlo experiments and an economic application to the study of the impact of family income on child achievement. We estimate the effect of family income on child achievement using our proposed estimator. Our findings suggest that correcting for endogeneity and measurement errors in family income as well as considering a nonseparable model in the relationship is critical to obtaining accurate estimates of the effect.

2. The model

We first specify the data generating process (DGP) for the structural system. There is an inherent ordering of the variables in such systems: in the language of White and Chalak (2009), "predecessor" variables may determine "successor" variables, but not vice versa. For instance, when X causes Y, then Y cannot cause X. In such cases, we say that Y *succeeds* X, and we write $Y \Leftarrow X$ as a shorthand notation. (See also Chalak and White (2011a, 2011b) and SWC.) Throughout, random variables are defined on a complete probability space $(\Omega, \mathcal{F}, P)$. We denote the support of a random variable by $\text{supp}(\cdot)$. By convention, we take the value of any referenced function to be zero except when the indicated random variable lies in $\text{supp}(\cdot)$. Also, we assume that any referenced conditional density is regular.

Assumption 2.1. (i) *Let* (U, W, X, Y) *be random variables such that* $E(|Y|) < \infty$. (ii) *The set* (U, W, X, Y) *is generated by a recursive structural system such that* $Y \Leftarrow (U, X)$ *and* $X \Leftarrow (U, W)$ *with* Y *generated by the structural equation*

$$Y = r(X, U_y),$$

where r *is an unknown measurable scalar-valued function and* $U_y \equiv v_y(U)$ *is a random vector of countable dimension* l_y, *with* v_y *a measurable function.* (iii) *The realizations of* Y *and* W *are observed, whereas those of* U, X, *and* U_y *are not.*

In our general approach, U, X, and W can be viewed as random vectors, while we let Y be scalar. Although X and W have finite dimension, the dimensions of U and U_y may be countably infinite. The restrictions that $Y \Leftarrow (U, X)$ and $X \Leftarrow (U, W)$ explicitly impose the allowed recursive structure without having to specify its details. The specified structural relations are directional causal links; thus, variations in X and U_y structurally determine Y, as in Strotz and Wold (1960). We do not assume that r is linear or monotone in its arguments or is separable between X and U_y. In line with a long tradition in the sciences, we refer to r as a *structural function* or a *response function* (e.g., Manski

(1997)).[1] As there is no restriction to the contrary, X and U_y are generally dependent, so that X is endogenous.[2] Moreover, standard instrumental variables are absent here, as the covariates W are also generally endogenous. However, identification of certain average marginal effects is possible when X satisfies a particular conditional form of exogeneity. To state this, we follow Dawid (1979) and write $X \perp U_y \mid W$ to denote that X is independent of U_y given $W = w$.[3]

Assumption 2.2. *We have* $X \perp U_y \mid W$.

Assumption 2.2 is analogous to structure imposed by Altonji and Matzkin (2005), Hoderlein and Mammen (2007), White and Chalak (2009, 2013), Chalak and White (2011a, 2011b), and Hoderlein (2011). Following Chalak and White (2011a), we call W "conditioning instruments." We also synonymously employ the term "control variables" (following the treatment effect or control function literature).

To fix the ideas, in the context of our empirical application to the impact of family income on child achievement, Y represents child scholastic achievement on standardized tests, while X is family income and U_y contains unobserved causes (individual heterogeneity) of child achievement. As a conditioning instrument W, we use a measure of parental abilities via a standardized cognitive test. Thanks to the nonseparable model, we can allow for nonlinearity in income effects and interactions between family income and unobserved drivers of child achievement such as parental skills, ability, or personality. In general, family income and unobserved drivers of child achievement are correlated because earning potential and child ability passed from parental abilities share common causes. This endogeneity makes it harder to recover features of the causal relationship. When two causes X and U_y are independent given parental cognitive test, however, the causal impact of family income on child achievement can be identified from the nonseparable model. In fact, we control for the indirect effect via parental and child abilities by conditioning on the underlying cause of endogeneity (parental abilities). It is thus plausible to expect that any other individual heterogeneity in child achievement inside U_y are independent from family income.

Primary parameters of interest are the conditional expectation of the response given X and W,

$$\mu(X, W) \equiv E(Y \mid X, W) \tag{1}$$

and its derivative $\beta(x, w) \equiv D_x\mu(x, w)$, where $D_x \equiv (\partial/\partial x)$. The functions exist whenever $E(|Y|) < \infty$, as ensured by Assumption 2.1(i), regardless of underlying structural rela-

[1] This also comports with the notion of "best response function" in game theory. Thus, we refer to Y as the response, since this is what the response function determines. This should not be confused with the term "local response," which refers to certain derivatives of r (e.g., Altonji and Matzkin (2005)).

[2] In classical treatments, the effects of endogenous variables are identified with the aid of instrumental variables. These are "standard" or "proper" when they are (i) correlated with X and (ii) exogenous (i.e., uncorrelated with or independent of unobservables, corresponding to U_y here).

[3] Conditional independence implies a similar "common support assumption" in Imbens and Newey (2009). We can see this from the argument $\text{supp}(U_y \mid X = x, W = w) \equiv \bigcap\{S \in \mathcal{F} : P[U_y \in S \mid X = x, W = w] = 1\} = \bigcap\{S \in \mathcal{F} : P[U_y \in S \mid W = w] = 1\} \equiv \text{supp}(U_y \mid W = w)$, where the second equality follows by $X \perp U_y \mid W$.

tions. With no further restrictions, these are purely stochastic objects. They provide no information about the causal effect of X on Y. When $X \Leftarrow (U, W)$ and Assumption 2.2 hold, $\mu(x, w)$ and $\beta(x, w)$ acquire causal interpretations from a particular conditional expectation and its derivative that have clear counterfactual meanings. Such objects are defined as the average counterfactual response at x given $W = w$,

$$\rho(x \mid w) \equiv E\bigl(r(x, U_y) \mid W = w\bigr),$$

and the covariate-conditioned average marginal effect of X on Y at x given $W = w$,

$$\beta^*(x \mid w) \equiv D_x \rho(x \mid w).$$

That is, under $X \perp U_y \mid W = w$, we have

$$\int r(x, u_y)\, dF(u_y \mid x, w) = \int r(x, u_y)\, dF(u_y \mid w),$$

since $X \perp U_y \mid W = w$ implies $dF(u_y \mid x, w) = dF(u_y \mid w)$. As a result, we obtain $\mu(x, w) = \rho(x \mid w)$ so that $\beta(x, w) = \beta^*(x \mid w)$. Consistent with Hurwicz (1950), we call this a *structural identification* result because it identifies aspects of the causal structure, ρ and β^*, with μ and β, standard stochastic objects.[4] We formally state the result for structural identification.

Proposition 2.1. *Under Assumptions 2.1 and 2.2, $\mu(x, w)$ and $\beta(x, w)$ are structurally identified as $\mu(x, w) = \rho(x \mid w)$ and $\beta(x, w) = \beta^*(x \mid w)$, respectively.*

In contrast to some of the literature on nonseparable models (e.g., Altonji and Matzkin (2005)), we do not assume that X is observable. Instead, we suppose that we observe two error-contaminated measurements of X. The following assumption expresses this formally.

Assumption 2.3. *Observables X_1 and X_2 are determined by the structural equations $X_1 = X + U_1$ and $X_2 = X + U_2$, where $U_1 \equiv v_1(U)$ and $U_2 \equiv v_2(U)$ for measurable functions v_1 and v_2.*

[4]Specifically, we have the representation

$$\mu(x, w) = \int r(x, u_y)\, dF(u_y \mid x, w),$$

where $dF(u_y \mid x, w)$ denotes the conditional density of U_y given $X = x$ and $W = w$. This represents $\mu(X, W)$ as the average response Y given $(X, W) = (x, w)$. The average counterfactual response at x given $W = w$ is

$$\rho(x \mid w) \equiv E\bigl(r(x, U_y) \mid W = w\bigr) = \int r(x, u_y)\, dF(u_y \mid w),$$

where $dF(u_y \mid w)$ denotes the conditional density of U_y given $W = w$. The function $\rho(x \mid w)$ is a conditional analog of the average structural function of Blundell and Powell (2004). The covariate-conditioned average marginal effect of X on Y at x given $W = w$ is

$$\beta^*(x \mid w) = D_x \int r(x, u_y)\, dF(u_y \mid w) = \int D_x r(x, u_y)\, dF(u_y \mid w),$$

provided the derivative and integral can be interchanged. This function is a weighted average of the unobservable marginal effect $D_x r(x, u_y)$ over unobserved causes, given observed covariates. As described in Appendix C, it can be used as a stepping stone to the analysis of various causally informative quantities.

If X were observable, we could estimate the covariate-conditioned average marginal effect $\beta^*(x \mid w)$ through $\mu(x, w)$ or $\beta(x, w)$. Since X is, however, not observable, such a direct approach is not available. Instead, we estimate $\mu(x, w)$ and its derivatives using the Fourier transform approach based on X_1 and X_2. Given mild restrictions on the relations between the measurement errors and the other variables of the system, two measurements of X are sufficient to identify parameters of interest. In our empirical application, family income (X) is usually observed with errors due to the limitation of surveys such as recall and rounding errors. We utilize the panel structure of the National Longitudinal Survey of Youth (NLSY) data by making use of 2-year reported family income $(X_1$ and $X_2)$. Given mild conditions introduced in the next section, it can be shown that two error-laden measurements of true family income are sufficient to identify the causal relationship.

3. Identification and estimation

In the section, we provide identification conditions under which the parameters of interest can be identified by observables, although true X is unobservable. We then propose a consistent nonparametric estimator.

3.1 *Identification*

In what follows, we take X and W to be scalars for simplicity. Analogous to the approach taken in SWC, we first focus on estimating quantities of the general form

$$g_{V,\lambda}(x, w) \equiv D_x^\lambda\big(E[V \mid X = x, W = w] f_{X|W}(x \mid w)\big),$$

where $D_x^\lambda \equiv (\partial^\lambda/\partial x^\lambda)$ denotes the derivative operator of degree λ, V is a generic random variable that will stand either for Y or for the constant $(V \equiv 1)$, and $f_{X|W}$ is the conditional density of X given W. For example, special cases of the general form above are $f_{X|W}(x \mid w) = g_{1,0}(x, w)$, $E[Y \mid X = x, W = w] f_{X|W}(x \mid w) = g_{Y,0}(x, w)$, and $\mu(x, w) = g_{Y,0}(x, w)/g_{1,0}(x, w)$. Thus, with structural identification, the covariate-conditioned average marginal effect of X on Y at x given $W = w$ is

$$\beta(x, w) = \frac{g_{Y,1}(x, w)}{g_{1,0}(x, w)} - \frac{g_{Y,0}(x, w)}{g_{1,0}(x, w)} \frac{g_{1,1}(x, w)}{g_{1,0}(x, w)}.$$

We analyze the asymptotic properties of estimators of $g_{V,\lambda}$ with generic V when we observe two error-contaminated measurements of X, as in Assumption 2.3. We can then straightforwardly obtain the asymptotic properties of estimators of $\beta(x, w)$ and weighted averages of $\beta(x, w)$. We impose the following conditions on Y, X, W, U_1, and U_2.

Assumption 3.1. *We have* $E[|X|] < \infty$ *and* $E[|U_1|] < \infty$.

Assumption 3.2. *We have* (i) $E[U_1 \mid X, U_2] = 0$, (ii) $U_2 \perp (X, W)$, *and* (iii) $E[Y \mid X, U_2, W] = E[Y \mid X, W]$.

Assumption 3.3. *We have* (i) $\inf_{w\in \text{supp}(W)} f_W(w) > 0$ *and* (ii) $\sup_{(x,w)\in\text{supp}(X,W)} f_{X|W}(x \mid w) < \infty$.

Assumption 3.4. *For any finite* $\zeta \in \mathbb{R}$, $|E[\exp(i\zeta X_2)]| > 0$.

Assumption 3.1 imposes mild conditions regarding the existence of the first moments of the cause of interest and the measurement error of the first error-laden observation. Assumption 3.4 is commonly imposed in the deconvolution literature (e.g., Fan (1991a), Fan and Truong (1993), Li and Vuong (1998), Li (2002), Schennach (2004a, 2004b)), which requires a nonvanishing characteristic function for X_2. Assumptions 3.1, 3.3, and 3.4 jointly ensure that $g_{V,0}(x, w)$ is well defined.

Assumption 3.2 has been imposed in a similar fashion in the repeated measurements literature (e.g., Hausman et al. (1991), Schennach (2004a, 2004b)); however, the presence of W is new here. Assumption 3.2(i) imposes a mild conditional moment restriction, while Assumption 3.2(ii) is crucial but plausible. The independence in Assumption 3.2(ii) seems difficult to avoid, given the nonlinearity of the model. On the one hand, the first measurement only needs a conditional mean restriction (Assumption 3.2(i)). Note that $E[U_1 \mid U_2] = E[E[U_1 \mid X, U_2] \mid U_2] = 0$, so that U_1 is mean independent of U_2. On the other hand, the mean of U_2 does not have to be zero. These relatively mild requirements on the measurement errors are plausible for many practical applications, but are asymmetric between U_1 and U_2. If both U_1 and U_2 satisfy Assumption 3.2(i) and (ii), one can obtain analogous estimators, interchanging the roles of X_1 and X_2 and suitably averaging the two estimators for higher efficiency. In the example of family income, the condition rules out the correlation between the error on the first measurement and true family income, and the error on the second measurement. It allows, nevertheless, other form of dependence between them. This structure of measurement errors is particularly useful when each survey participant experiences common shocks over surveys that affect her surveyed family income (e.g., habit persistence). It is also interesting to note that the imposed condition tolerates a systematic drift in family income (e.g., positive trend) since the error on the second measurement is allowed to have nonzero mean.

Let $\mathbb{N} \equiv \{0, 1, \dots\}$ and $\overline{\mathbb{N}} \equiv \mathbb{N} \cup \{\infty\}$.

Assumption 3.5. *For* $V \in \{1, Y\}$, $g_{V,0}(\cdot, w)$ *is continuously differentiable of order* $\Lambda \in \overline{\mathbb{N}}$ *on* $\mathbb{R}$ *for each* $w \in \text{supp}(W)$.

This assumption imposes smoothness on $g_{V,0}$. If $g_{V,\lambda}$ can be defined solely in terms of the joint distribution of observable variables V, X_1, and X_2, we say it is *stochastically identified*. This is shown in the next lemma.[5]

[5] Derivation of a part of the expression for ϕ_V is similar to that of an identity due to Kotlarski (see Prakasa Rao (1992, p. 21)), which enables one to recover the densities of X, U_1, and U_2 from the joint density of X_1 and X_2 under the assumption that X, U_1, and U_2 are independent. Our identification strategy for the density of X relies on weaker assumptions than independence. In fact, we only require $E[U_1 \mid X, U_2] = 0$ and $U_2 \perp X$ for the result, instead of mutual independence of X, U_1, and U_2. As a result, our setup allows dependence between X and U_1, and between U_1 and U_2.

Lemma 3.1. *Under Assumptions 2.1(i), 2.3, and 3.1–3.5, the set of observables* $\{V, X_1, X_2, W\}$ *stochastically identify*

$$g_{V,\lambda}(x, w) = \frac{1}{2\pi} \int (-i\zeta)^{\lambda} \phi_V(\zeta, w) \exp(-i\zeta x)\, d\zeta$$

for $V \in \{1, Y\}$ *and for each* $\lambda \in \{0, \ldots, \Lambda\}$ *and* $(x, w) \in \operatorname{supp}(X, W)$, *where for each real* ζ *and* $w \in \operatorname{supp}(W)$,

$$\phi_V(\zeta, w) \equiv E\big[Ve^{i\zeta X} \mid W = w\big] = \frac{E[Ve^{i\zeta X_2} \mid W = w]}{E[e^{i\zeta X_2}]} \exp\biggl(\int_0^{\zeta} \frac{iE[X_1 e^{i\xi X_2}]}{E[e^{i\xi X_2}]}\, d\xi\biggr).$$

Thus, knowledge of $E[Ve^{i\zeta X_2} \mid W = w]$, $E[e^{i\zeta X_2}]$, and $E[X_1 e^{i\xi X_2}]$ is sufficient to obtain stochastic identification of $g_{V,\lambda}$. The identification result shows that the causal effect of family income on child achievement can be identified by observables, that is, child test scores, 2-year reported family incomes, and a measure of parental skills, although true family income is endogenous and mismeasured.

3.2 *Estimation*

Our nonparametric estimators of $g_{V,\lambda}$ make use of the following class of flat-top kernels of infinite order proposed by Politis and Romano (1999).

Assumption 3.6. *A real-valued kernel* $x \to k(x)$ *is measurable and symmetric, satisfying* $\int k(x)\, dx = 1$, *and its Fourier transform* $\xi \to \kappa(\xi)$ *is compactly supported and bounded.*

Flat-top kernels of infinite order have the property that their Fourier transforms are "flat" over an open neighborhood of the origin. When a flat-top kernel of infinite order is used, the smoothness of the function to be estimated is the only factor controlling the rate of decrease of the bias, whereas when a finite-order kernel is used, both the smoothness of the function and the order of the kernel affect the rate of decrease of the bias. Compact support of the Fourier transform of the kernel is a weak requirement because one can transform any given kernel $\tilde{k}$ into a modified kernel k with compact Fourier support, having most of the properties of the original kernel. To construct the modified Fourier transform κ from the original Fourier transform $\tilde{\kappa}$ of $\tilde{k}$, put

$$\kappa(\xi) = \mathcal{W}(\xi)\tilde{\kappa}(\xi),$$
$$\mathcal{W}(\xi) = \begin{cases} 1 & \text{if } |\xi| \le \bar{\xi}, \\ \bigl(1 + \exp\bigl((1 - \bar{\xi})\bigl((1 - |\xi|)^{-1} - (|\xi| - \bar{\xi})^{-1}\bigr)\bigr)\bigr)^{-1} & \text{if } \bar{\xi} < |\xi| \le 1, \\ 0 & \text{if } 1 < |\xi|. \end{cases} \tag{2}$$

Here $\mathcal{W}(\cdot)$ is a window function that is constant in the neighborhood of the origin and vanishes beyond a given frequency, determined by $\bar{\xi} \in (0, 1)$.

The next lemma incorporates the Fourier transform of the kernel into the expression for $g_{V,\lambda}(x, w)$.

Lemma 3.2. *For $V \in \{1, Y\}$ and for each $\lambda \in \{0, \dots, \Lambda\}$, $(x, w) \in \operatorname{supp}(X, W)$, and $h_1 > 0$, let*

$$g_{V,\lambda}(x, w; h_1) \equiv \int \frac{1}{h_1} k\left(\frac{\tilde{x} - x}{h_1}\right) g_{V,\lambda}(\tilde{x}, w)\, d\tilde{x},$$

where k satisfies Assumption 3.6. Then under Assumptions 2.1(i)*, 2.3, 3.1, and 3.3–3.5,*

$$g_{V,\lambda}(x, w; h_1) = \frac{1}{2\pi} \int (-i\zeta)^\lambda \kappa(h_1 \zeta) \phi_V(\zeta, w) \exp(-i\zeta x)\, d\zeta.$$

Note that the kernel bandwidth or smoothing parameter is h_1.[6] Because the denominator of our estimator contains an asymptotically vanishing characteristic function as frequency goes to infinity, we face the well known ill-posed inverse problem that occurs when one tries to invert a convolution operation. This problem can be regularized by estimating the associated numerator using the kernel whose Fourier transform is compactly supported, which guarantees that the numerator will decay to zero before the denominator causes the ratio to diverge.

We now define our estimator for $g_{V,\lambda}(x, w)$ based on Lemma 3.2 by replacing ϕ_V with a sample analog $\hat{\phi}_V$ as follows.

Definition 3.1. Let $h_n \equiv (h_{1n}, h_{2n}) \to 0$ as $n \to \infty$. The estimator for $g_{V,\lambda}(x, w)$ is defined as

$$\hat{g}_{V,\lambda}(x, w; h_n) \equiv \frac{1}{2\pi} \int (-i\zeta)^\lambda \kappa(h_{1n} \zeta) \hat{\phi}_V(\zeta, w; h_{2n}) \exp(-i\zeta x)\, d\zeta,$$

where

$$\hat{\phi}_V(\zeta, w; h_{2n}) \equiv \frac{\hat{E}[V e^{i\zeta X_2} \mid W = w]}{\hat{E}[e^{i\zeta X_2}]} \exp\left(\int_0^\zeta \frac{i\hat{E}[X_1 e^{i\xi X_2}]}{\hat{E}[e^{i\xi X_2}]}\, d\xi\right),$$

$$\hat{E}\left[V e^{i\zeta X_2} \mid W = w\right] \equiv \frac{\hat{E}[V e^{i\zeta X_2} k_{h_{2n}}(W - w)]}{\hat{E}[k_{h_{2n}}(W - w)]},$$

and where $k_{h_{2n}}(\cdot) = h_{2n}^{-1} k(\cdot / h_{2n})$ and $\hat{E}[\cdot]$ denotes a sample average.[7]

Here, $\phi_V(\zeta, w)$ is replaced with its sample analog, $\hat{\phi}_V(\zeta, w; h_{2n})$, and $\hat{E}[V e^{i\zeta X_2} \mid W = w]$ is a kernel estimator of $E[V e^{i\zeta X_2} \mid W = w]$. Note that $\hat{\phi}_V$ uses a bandwidth h_2 distinct from h_1. The proposed estimator can be extended to multivariate settings. When

[6]We define $g_{V,\lambda}(x, w; 0) \equiv g_{V,\lambda}(x, w)$ because $\lim_{h_1 \to 0} g_{V,\lambda}(x, w; h_1) = g_{V,\lambda}(x, w)$, as can be seen by a direct application of Lebesgue's dominated convergence theorem, under our assumptions ($\kappa(\zeta)$ is bounded and the remaining terms are bounded by an integrable function).

[7]There are two kernels in the expression for $\hat{g}_{V,\lambda}(x, w, h_n)$: one is associated with the cause X and the other with the conditioning instrument W. Although for notational convenience we do not explicitly use different notations, these could be different (note that we do use different bandwidths for different kernels). Thus, $\kappa(\cdot)$ is the Fourier transform of a flat-top kernel associated with X and $k(\cdot)$ is another flat-top kernel for W. Indeed, we employ different flat-top kernels in the empirical section.

X is multivariate, X_1, X_2, and ζ become multivariate as well. Then the identification and estimation of $g_{V,\lambda}(x, w)$ are the same as the univariate case, except that multivariate Fourier transforms should be used so that the integral of ζ is now multidimensional (except the integral over ξ, which becomes a path integral in analogy with Theorem 1 in Cunha, Heckman, and Schennach (2010)). When W is multivariate, the kernel estimation of $E[Ve^{i\zeta X_2} \mid W = w]$ involves a multidimensional covariate, in which case a curse of dimensionality is to be expected as in usual nonparametric estimation of the conditional mean. Practitioners would instead use semiparametric single index models if the dimension of W is large (see Ichimura (1993) for more details).

4. Asymptotics

We now establish asymptotic properties of the general form $\hat{g}_{V,\lambda}(x, w; h)$ in Section 4.1 and of covariate-conditioned average marginal effects $\hat{\beta}(x, w; h)$ in Section 4.2. In the empirical application, we estimate the impact of family income on child achievement by $\hat{\beta}(x, w; h)$ in the presence of endogenous and mismeasured family income. The section thus provides regularity conditions under which the proposed estimator exhibits nice asymptotic behaviors. The covariate-conditioned average marginal effects $\hat{\beta}(x, w; h)$ depend on levels of family income and a measure of parental skills. This estimator thus delivers a nonlinear shape of the causal effect over different levels of family income given the measure of parental skills. In many empirical examples, weighted averages have also been of interest. In Appendix C, we provide asymptotics for estimators of weighted averages of $\beta(x, w)$ such as the derivative of the average structural function in Blundell and Powell (2004), (the weighted average of) the local average response in Altonji and Matzkin (2005), and the average continuous treatment effect (on the treated) in Florens et al. (2008).[8] The results would be useful when an empirical researcher wants to estimate average effects over either a measure of parental skills or both family income and a measure of parental skills.

4.1 *Asymptotics for the general form*

SWC generalize Schennach (2004a, 2004b) to encompass (i) the $\lambda \neq 0$ case, (ii) uniform convergence results, and (iii) general semiparametric functionals of $g_{V,\lambda}$. Here, we use an approach similar to Schennach (2004a, 2004b) to obtain counterparts of these three results in the context of models where endogeneity is handled with conditional independence, as in the treatment effect literature, and where the cause of interest is contaminated by measurement error. The analysis of the properties of our estimator is, nevertheless, more complex due to the presence of the kernel estimator of the conditional expectation. Note that conditioning has nontrivial implications on the asymptotic treatment so that one cannot merely invoke Schennach (2004a, 2004b) or SWC. It adds a nonparametric aspect along another dimension, which interacts with the measurement error through numerous remainder terms that have to be carefully analyzed. We begin

[8] See also Heckman and Vytlacil (2005) for possible various treatment effect parameters in the literature.

by deriving the asymptotic behavior of the estimator for quantities of the general form $\hat{g}_{V,\lambda}(x, w; h_n)$.

Letting $\theta_A(\zeta) \equiv E[Ae^{i\zeta X_2}]$ for $A \in \{1, X_1\}$ and $\chi_V(\zeta, w) \equiv E[Ve^{i\zeta W X_2} \mid W = w] f_W(w)$, we define

$$\begin{aligned}
\Psi_{V,\lambda,1}(\zeta, x, w; h_1) &\equiv -\frac{1}{2\pi} \frac{i\theta_{X_1}(\zeta)}{(\theta_1(\zeta))^2} \int_\zeta^{\pm\infty} (-i\xi)^\lambda \kappa(h_1\xi) \exp(-i\xi x) \phi_V(\xi, w)\, d\xi \\
&\quad - \frac{1}{2\pi}(-i\zeta)^\lambda \kappa(h_1\zeta) \exp(-i\zeta x) \frac{\phi_V(\zeta, w)}{\theta_1(\zeta)}, \\
\Psi_{V,\lambda,X_1}(\zeta, x, w; h_1) &\equiv \frac{1}{2\pi} \frac{i}{\theta_1(\zeta)} \int_\zeta^{\pm\infty} (-i\xi)^\lambda \kappa(h_1\xi) \exp(-i\xi x) \phi_V(\xi, w)\, d\xi, \\
\Psi_{V,\lambda,\chi_V}(\zeta, x, w; h_1) &\equiv \frac{1}{2\pi}(-i\zeta)^\lambda \kappa(h_1\zeta) \exp(-i\zeta x) \frac{\phi_V(\zeta, w)}{\chi_V(\zeta, w)}, \\
\Psi_{V,\lambda,f_W}(\zeta, x, w; h_1) &\equiv -\frac{1}{2\pi}(-i\zeta)^\lambda \kappa(h_1\zeta) \exp(-i\zeta x) \frac{\phi_V(\zeta, w)}{f_W(w)},
\end{aligned}$$

where for a given function $\xi \to f(\xi)$, $\int_\zeta^{\pm\infty} f(\xi)\, d\xi \equiv \lim_{c\to+\infty} \int_\zeta^{c\zeta} f(\xi)\, d\xi$. Then $\bar{g}_{V,\lambda}(x, w; h)$ can be defined as the linearization of $\hat{g}_{V,\lambda}(x, w; h)$ in terms of $(\hat{E}[e^{i\zeta X_2}] - E[e^{i\zeta X_2}])$, $(\hat{E}[X_1 e^{i\zeta X_2}] - E[X_1 e^{i\zeta X_2}])$, $(\hat{E}[Ve^{i\zeta X_2} k_{h_2}(W - w)] - E[Ve^{i\zeta X_2} k_{h_2}(W - w)])$, and $(\hat{E}[k_{h_2}(W - w)] - E[k_{h_2}(W - w)])$. We then have

$$\bar{g}_{V,\lambda}(x, w; h) - g_{V,\lambda}(x, w; h_1) \equiv \hat{E}\big[\ell_{V,\lambda}(x, w, h; V, X_1, X_2, W)\big],$$

where

$$\begin{aligned}
&\ell_{V,\lambda}(x, w, h; v, x_1, x_2, \tilde{w}) \\
&\quad \equiv \int \Psi_{V,\lambda,1}(\zeta, x, w; h_1)\big(e^{i\zeta x_2} - E\big[e^{i\zeta X_2}\big]\big)\, d\zeta \\
&\qquad + \int \Psi_{V,\lambda,X_1}(\zeta, x, w; h_1)\big(x_1 e^{i\zeta x_2} - E\big[X_1 e^{i\zeta X_2}\big]\big)\, d\zeta \\
&\qquad + \int \Psi_{V,\lambda,\chi_V}(\zeta, x, w; h_1)\big(ve^{i\zeta x_2} k_{h_2}(\tilde{w} - w) - E\big[Ve^{i\zeta X_2} k_{h_2}(W - w)\big]\big)\, d\zeta \\
&\qquad + \int \Psi_{V,\lambda,f_W}(\zeta, x, w; h_1)\big(k_{h_2}(\tilde{w} - w) - E\big[k_{h_2}(W - w)\big]\big)\, d\zeta.
\end{aligned}$$

The first result decomposes the estimation error into a "bias term," a "variance term," and a "remainder term."

Lemma 4.1. *Suppose that* $\{U_i, W_i, X_i, Y_i\}$ *is an independent and identically distributed* (*IID*) *sequence satisfying Assumptions* 2.1(i), 2.3, *and* 3.1–3.6 *hold. Then for* $V \in \{1, Y\}$ *and for each* $\lambda \in \{0, \dots, \Lambda\}$, $(x, w) \in \operatorname{supp}(X, W)$, *and* $h \equiv (h_1, h_2) > 0$,

$$\hat{g}_{V,\lambda}(x, w; h) - g_{V,\lambda}(x, w) = B_{V,\lambda}(x, w; h_1) + L_{V,\lambda}(x, w; h) + R_{V,\lambda}(x, w; h),$$

where $B_{V,\lambda}(x, w; h_1)$ *is a nonrandom bias term defined as*

$$B_{V,\lambda}(x, w; h_1) \equiv g_{V,\lambda}(x, w; h_1) - g_{V,\lambda}(x, w),$$

$L_{V,\lambda}(x, w; h)$ *is a variance term admitting the linear representation*

$$L_{V,\lambda}(x, w; h) \equiv \bar{g}_{V,\lambda}(x, w; h) - g_{V,\lambda}(x, w; h_1),$$

and $R_{V,\lambda}(x, w; h)$ *is a remainder term*

$$R_{V,\lambda}(x, w; h) \equiv \hat{g}_{V,\lambda}(x, w; h) - \bar{g}_{V,\lambda}(x, w; h).$$

Because $\hat{g}_{V,\lambda}(x, w; h)$ is a nonlinear functional of the data generating process, the above linearization facilitates the analysis of the asymptotic behavior of the estimator. In fact, the limiting distribution of $\hat{g}_{V,\lambda}(x, w; h) - g_{V,\lambda}(x, w)$ is equivalent to that of $L_{V,\lambda}(x, w; h)$, as long as $B_{V,\lambda}(x, w; h_1)$ and $R_{V,\lambda}(x, w; h)$ are asymptotically negligible. Thus, we first establish bounds on the bias, the variance, and the remainder terms; we then establish the asymptotic normality of the variance term.

Bounds on the tail behavior of the Fourier transforms are needed to obtain the specific rate of convergence results for our kernel estimators. These conditions correspond to smoothness constraints on the corresponding densities. The deconvolution literature (e.g., Fan (1991a), Fan and Truong (1993), Li and Vuong (1998), Li (2002), Schennach (2004a), and Carroll, Ruppert, Stefanski, and Crainiceanu (2006)) commonly distinguishes between "ordinarily smooth" and "supersmooth" functions. Specifically, ordinarily smooth functions admit a finite number of continuous derivatives and have a Fourier transform whose tail decays to zero at a geometric rate, $|\zeta|^{\gamma}$, $\gamma < 0$, as the frequency, $|\zeta|$, goes to infinity (e.g., uniform, gamma, and double exponential), whereas supersmooth functions admit an infinite number of continuous derivatives and have a Fourier transform whose tail decays to zero at an exponential rate as $\exp(\alpha|\zeta|^{\nu})$, $\alpha < 0$, $\nu > 0$ as the frequency goes to infinity (e.g., Cauchy and normal). For conciseness, our smoothness restrictions encompass both the ordinarily smooth and supersmooth cases; thus, our regularity conditions are expressed in terms of $(1 + |\zeta|)^{\gamma} \exp(\alpha|\zeta|^{\nu})$. Let $\phi_1(\zeta) \equiv E[e^{i\zeta X}]$.

Assumption 4.1. (i) *There exist constants* $C_1 > 0$ *and* $\gamma_1 \geq 0$ *such that*

$$\left|D_{\zeta} \ln \phi_1(\zeta)\right| = \left|\frac{D_{\zeta}\phi_1(\zeta)}{\phi_1(\zeta)}\right| \leq C_1\left(1 + |\zeta|\right)^{\gamma_1}.$$

(ii) *There exist constants* $C_{\phi} > 0$, $\alpha_{\phi} \leq 0$, $\nu_{\phi} \geq 0$, *and* $\gamma_{\phi} \in \mathbb{R}$ *such that* $\nu_{\phi}\gamma_{\phi} \geq 0$ *and for* $V \in \{1, Y\}$,

$$\sup_{w \in \text{supp}(W)} \left|\phi_V(\zeta, w)\right| \leq C_{\phi}\left(1 + |\zeta|\right)^{\gamma_{\phi}} \exp\left(\alpha_{\phi}|\zeta|^{\nu_{\phi}}\right),$$

and if $\alpha_{\phi} = 0$, *then* $\gamma_{\phi} < -\lambda - 1$ *for given* $\lambda \in \{0, \dots, \Lambda\}$.

(iii) *There exist constants* $C_\theta > 0$, $\alpha_\theta \le 0$, $\nu_\theta \ge \nu_\phi \ge 0$, *and* $\gamma_\theta \in \mathbb{R}$ *such that* $\nu_\theta\gamma_\theta \ge 0$ *and for* $V \in \{1, Y\}$,

$$\min\Big\{\inf_{w\in\operatorname{supp}(W)}\big|\chi_V(\zeta, w)\big|, \big|\theta_1(\zeta)\big|\Big\} \ge C_\theta\big(1+|\zeta|\big)^{\gamma_\theta}\exp\big(\alpha_\theta|\zeta|^{\nu_\theta}\big).$$

We omit a term $\exp(\alpha_1|\zeta|^{\nu_1})$ in Assumption 4.1(i) with only a small loss of generality because $\ln\phi_1(\zeta)$ is typically a power of ζ for large ζ, even when the density of $\phi_1(\zeta)$ is supersmooth, as pointed out in Schennach (2004a) and SWC. Note that the rate of decay of $\phi_V(\zeta, w)$ is governed by the smoothness of $g_{V,0}(x, w) = E[V \mid X = x, W = w]f_{X|W}(x \mid w)$, as $\phi_V(\zeta, w) = \int g_{V,0}(x, w)e^{i\zeta x}\,dx$. Note that a lower bound, instead of an upper bound, is imposed on $\chi_V(\zeta, w)$ and $\theta_1(\zeta)$, because these appear in the denominator of the expression for $\hat{g}_{V,\lambda}(x, w; h)$. Individual lower bounds on the modulus of the characteristic functions of X and U_2 imply the lower bound on $\theta_1(\zeta)$, as $\theta_1(\zeta) = E[e^{i\zeta X_2}] = E[e^{i\zeta X}]E[e^{i\zeta U_2}]$ by Assumption 3.2(ii). We group together $\chi_V(\zeta, w)$ and $\theta_1(\zeta)$ (in fact, $E[e^{i\zeta X}]$ and $E[e^{i\zeta U_2}]$) in a single assumption for the lower bound for notational convenience. We explicitly impose $\nu_\theta \ge \nu_\phi$ because

$$\begin{aligned}
&C_\phi\big(1+|\zeta|\big)^{\gamma_\phi}\exp\big(\alpha_\phi|\zeta|^{\nu_\phi}\big)\\
&\quad\ge \sup_{w\in\operatorname{supp}(W)}\big|\phi_1(\zeta, w)\big| = \sup_{w\in\operatorname{supp}(W)}\big|E\big[e^{i\zeta X} \mid W = w\big]\big|\\
&\quad\ge \left|\int E\big[e^{i\zeta X} \mid W = w\big]f_W(w)\,dw\right| = \big|E\big[e^{i\zeta X}\big]\big|\\
&\quad\ge \big|E\big[e^{i\zeta X}\big]\big|\big|E\big[e^{i\zeta U_2}\big]\big| \ge \big|E\big[e^{i\zeta X_2}\big]\big|\\
&\quad= \big|\theta_1(\zeta)\big| \ge C_\theta\big(1+|\zeta|\big)^{\gamma_\theta}\exp\big(\alpha_\theta|\zeta|^{\nu_\theta}\big).
\end{aligned}$$

Requiring both upper and lower bounds on various Fourier transforms is a common occurrence in the deconvolution literature (e.g., Fan (1991a), Fan and Truong (1993), Li and Vuong (1998), Li (2002), Schennach (2004a, 2004b)). One condition that is perhaps slightly different here is the requirement that quantities related to conditional characteristic functions weighted by a density (such as $\chi_V(\zeta, w) \equiv E[Ve^{i\zeta WX_2} \mid W = w]f_W(w)$) be bounded below. This will typically require the weighting density (e.g., $f_W(w)$) to be bounded away from zero on its support.

Lemma B.1 in Appendix B describes the asymptotic properties of the bias term defined in Lemma 4.1. We note that the bias term behaves identically to that of a conventional kernel estimator employed when X is measurement error-free, because $B_{V,\lambda}(x, w; h_1)$ only involves the kernel and error-free variables.[9]

[9] When X is perfectly observed, one can estimate $g_{V,\lambda}$ using a similar Fourier transform as

$$\hat{g}_{V,\lambda}(x, w; h_n) \equiv \frac{1}{2\pi}\int(-i\zeta)^\lambda\kappa(h_{1n}\zeta)\hat{\phi}_V(\zeta, w; h_{2n})\exp(-i\zeta x)\,d\zeta$$

for $h_n \to 0$ as $n \to \infty$, where

$$\hat{\phi}_V(\zeta, w; h_{2n}) \equiv \hat{E}\big[Ve^{i\zeta X} \mid W = w\big] = \frac{\hat{E}[Ve^{i\zeta X}k_{h_{2n}}(W-w)]}{\hat{E}[k_{h_{2n}}(W-w)]}.$$

To establish a convergence rate and asymptotic normality for the variance term, $L_{V,\lambda}(x, w; h)$, we impose some regularity conditions. We first impose conditions implying finite variance of $L_{V,\lambda}(x, w; h)$.

Assumption 4.2. *We have* $E[|X_1|^2] < \infty$ *and* $E[|Y|^2] < \infty$.

We next impose conditional moment bounds that are important for establishing asymptotic normality of $L_{V,\lambda}(x, w; h)$.

Assumption 4.3. *For some* $\delta > 0$, $\sup_{x_2 \in \text{supp}(X_2)} E[|X_1|^{2+\delta} \mid X_2 = x_2] < \infty$ *and* $\sup_{w \in \text{supp}(W)} E[|Y|^{2+\delta} \mid W = w] < \infty$.

We also suitably control the bandwidth to establish asymptotic normality.

Assumption 4.4. *The bandwidth sequence* $h_n \to 0$ *as* $n \to \infty$, *such that if* $\nu_\theta \neq 0$ *in Assumption 4.1*(iii), *then* $h_{1n}^{-1} = O((\ln n)^{1/\nu_\theta - \eta})$ *and* $h_{2n}^{-1} = O(n^{1/4-\eta})$ *for some* $\eta > 0$; *otherwise, for each* $\lambda \in \{0, \dots, \Lambda\}$, $h_{1n}^{-1} = O(n^{-\eta} n^{(3/2)/(\gamma_\phi+\lambda+\gamma_1-\gamma_\theta+3)})$ *and* $h_{2n}^{-1} = O(n^{((\gamma_1+2)/2)(3/2)/(\gamma_\phi+\lambda+\gamma_1-\gamma_\theta+3)})$ *for some* $\eta > 0$.

The bandwidth sequences given above were designed to ensure that a regularity condition in Lemma B.3 holds (see Lemma B.3 and the proof of Lemma B.4 in the Appendix B). The bandwidth sequences for h_1 imply that if densities appearing in quantities in the denominator ($\chi_V(\zeta, w)$ and θ_1) are supersmooth, one must choose a larger bandwidth than in the case of ordinary smoothness. The achievable convergence rates will thus be slower than for ordinary smoothness. Similar but simpler results have also been observed in the kernel deconvolution literature (see Fan (1991a), Fan and Truong (1993), Li and Vuong (1998), Li (2002), Schennach (2004a)).

Lemma B.4 in Appendix B states a uniform rate and asymptotic normality for the variance term. The rate of divergence of the variance term is controlled by the smoothness of the density of the measurement error U_2 and $E[\varphi(x_2, w) \mid X_2 = x_2]$ (through γ_θ, α_θ, and ν_θ) as well as by the smoothness of the density of X and $E[V \mid X = x, W = w]$ (through γ_ϕ, α_ϕ, ν_ϕ, and γ_1), where $\varphi(x_2, w) = \int v f_{V,X_2,W}(v, x_2, w)\, dv$. As expected, the order of the variance term is larger than that of a traditional kernel estimator with error-free variables.[10] As a result, the rate of convergence of the estimator $\hat{g}_{V,\lambda}$ will be slower than that of a standard kernel estimator, because the bias term is identical to that of a standard kernel estimator with measurement error-free X.

We now establish a uniform convergence rate and asymptotic normality of the estimator $\hat{g}_{V,\lambda}(x, w; h_n)$. We first provide bounds on the remainder term that are used to

Then one can easily derive the order of the bias, which is the same as that in Lemma B.1. Note that this estimator for $g_{V,\lambda}$ has the same asymptotic properties as a traditional kernel estimator of $g_{V,\lambda}$ with the flat-top kernel of infinite order when X is perfectly observed; but this estimator using the Fourier transform approach makes easy comparisons possible with our estimator in Definition 3.1.

[10]With perfectly observed X, the order of the variance term of the estimator in footnote 6 can be derived as $n^{-1/2} h_{2n}^{-1} (h_{1n}^{-1})^{1+\gamma_\phi+\lambda} \exp(\alpha_\phi (h_{1n}^{-1})^{\nu_\phi})$. Thus, if $\nu_\phi > 0$, $\nu_L \equiv \nu_\theta \geq \nu_\phi$ by construction and if $\nu_L \equiv \nu_\theta = \nu_\phi = 0$, then $\gamma_{\lambda,L} \equiv 1 + \gamma_\phi - \gamma_\theta + \lambda > 1 + \gamma_\phi + \lambda$ since $(-\gamma_\theta) > 0$ and $\max\{(h_{1n}^{-1})^{\delta_L}, h_{2n}^{-1}\} \geq h_{2n}^{-1}$. Then the order of the variance term in Lemma B.4 is greater than that of the kernel estimator with perfectly observed variables.

obtain a convergence rate. The next assumption puts restrictions on the moments of X_2 (associated with X_1 and Y), which are useful for establishing a bound on the remainder term, $R_{V,\lambda}(x, w; h_n)$.

Assumption 4.5. *We have* $E[|X_2|] < \infty$, $E[|X_1X_2|] < \infty$, *and* $E[|YX_2|] < \infty$.

The following assumption requires a uniform convergence rate for the kernel density estimator, $\hat{f}_W(w)$, in the denominator of $\hat{g}_{V,\lambda}(x, w; h)$. This assumption is also used to get the bound on the remainder term and is satisfied by density estimation with conventional choice of kernel. Even though flat-top kernels of infinite order attain a faster convergence rate than that below (e.g., Politis and Romano (1999)), the faster rate is not necessary for our result. Primitive conditions for this assumption are given by Theorem 3.2 (eqs. (11) and (12)) in SWC.

Assumption 4.6. *We have* $\sup_{w\in \operatorname{supp}(W)} |\hat{f}_W(w) - f_W(w)| = O_p(\sqrt{\frac{\ln n}{nh_2}} + h_2^2)$.

The following assumption gives a lower bandwidth bound that slightly differs from that of Assumption 4.4. Note that neither Assumption 4.4 nor 4.7 is necessarily stronger than the other.

Assumption 4.7. *If* $\nu_\theta \neq 0$ *in Assumption* 4.1(iii), *then* $h_{1n}^{-1} = O((\ln n)^{1/\nu_\theta - \eta})$ *and* $h_{2n}^{-1} = O(n^{1/6-\eta})$ *for some* $\eta > 0$; *otherwise* $h_{1n}^{-1} = O(n^{-\eta} n^{1/(14\gamma_1 - 14\gamma_\theta)})$ *and* $h_{2n}^{-1} = O(n^{1/7-\eta})$ *for some* $\eta > 0$.

The bandwidth sequences above were designed to ensure that the nonlinear remainder term, $R_{V,\lambda}(x, w; h_n)$, is indeed asymptotically negligible so that the decomposition of the estimation error into bias, variance, and remainder terms is justified, thus implying that the linear approximation of $\hat{g}_{V,\lambda}(x, w; h_n) - g_{V,\lambda}(x, w)$ using the variance term, $L_{V,\lambda}(x, w; h_n)$, is appropriate. The basic intuition behind the selection of the bandwidth is similar to that for Assumption 4.4.

We provide uniform bounds on the nonlinear remainder in Lemma B.6 in Appendix B. Lemma B.6(i) is used to establish the asymptotic normality of $\hat{g}_{V,\lambda}$ and (ii) is relevant to obtaining a convergence rate. Conditioning on control variables in $\hat{g}_{V,\lambda}(x, w; h_n)$ interacts with the measurement error through numerous remainder terms so that it complicates the analysis of the asymptotic properties; contrast to Schennach (2004a, 2004b) or SWC.

The next theorem establishes a consistency and uniform convergence rate of the estimator that is a stepping stone to study covariate-conditioned average marginal effects.

Theorem 4.2. *Suppose the conditions of Lemma* 4.1, *together with Assumptions* 4.1–4.3 *and* 4.5–4.7 *hold. Then for* $V \in \{1, Y\}$ *and each* $\lambda \in \{0, \dots, \Lambda\}$,

$$\begin{aligned}
&\sup_{(x,w)\in \operatorname{supp}(X,W)} \left|\hat{g}_{V,\lambda}(x, w; h_n) - g_{V,\lambda}(x, w, 0)\right| \\
&\quad = O\big(\big(h_{1n}^{-1}\big)^{\gamma_{\lambda,B}} \exp\big(\alpha_B \big(h_{1n}^{-1}\big)^{\nu_B}\big)\big) \\
&\qquad + O_p\big(n^{-1/2}\big(\max\big\{\big(h_{1n}^{-1}\big)^{\delta_L}, h_{2n}^{-1}\big\}\big)\big(h_{1n}^{-1}\big)^{\gamma_{\lambda,L}} \exp\big(\alpha_L \big(h_{1n}^{-1}\big)^{\nu_L}\big)\big).
\end{aligned}$$

In the next assumption, we ensure that the bias term and the remainder term do not dominate the variance term for the linear representation.

ASSUMPTION 4.8. *The bandwidth sequence* $h_n \to 0$ *at a rate such that for* $V \in \{1, Y\}$ *and for each* $\lambda \in \{0, \dots, \Lambda\}$ *and* $(x, w) \in \text{supp}(X, W)$, *we have* (i) $\Omega_{V,\lambda}(x, w; h_n) > 0$ *for all* n *sufficiently large,* (ii) $n^{1/2}(\Omega_{V,\lambda}(x, w; h_n))^{-1/2}\ |B_{V,\lambda}(x, w; h_{1n})| \to 0$, *and* (iii) $n^{1/2}(\Omega_{V,\lambda}(x, w; h_n))^{-1/2}|R_{V,\lambda}(x, w; h_n)| \xrightarrow{p} 0$.

This assumption imposes a lower bound on $\Omega_{V,\lambda}(x, w; h_n)$ such that the bias $B_{V,\lambda}(x, w; h_{1n})$ and remainder $R_{V,\lambda}(x, w; h_n)$ are small relative to this lower bound. The bounds on the bias and the nonlinear remainder (provided by Lemmas B.1 and B.6, respectively) can be directly used to establish this assumption from more primitive conditions. However, note that the bound on $\Omega_{V,\lambda}(x, w; h_n)$ given in Lemma B.4(i) is only an upper bound on the convergence rate, so it is not sufficient to obtain our next result, Theorem 4.3. As a result, the bias term and the nonlinear remainder term must be asymptotically negligible relative to $n^{-1/2}(\Omega_{V,\lambda}(x, w; h_n))^{1/2}$, the standard deviation of $L_{V,\lambda}(x, w; h_n)$, so as to ensure that they have no effect on the limiting distribution of the estimator. Although we provide this condition in high-level form for the benefit of conciseness, more primitive conditions can be found via Theorem 3 in Schennach (2004a). Essentially, the exact asymptotic rate of convergence of $n^{-1/2}(\Omega_{V,\lambda}(x, w; h_n))^{1/2}$ can be derived from the assumption that the limiting behavior (for large frequencies) of the relevant Fourier transforms has a power law or an exponential form (instead of merely being bounded above by such functional forms).[11] A similar notion was also used in Fan (1991b) to establish asymptotic normality of the kernel deconvolution estimator and is directly related to the notion of functions that are "well behaved at infinity," as introduced by Lighthill (1962). One can construct simple examples where this assumption holds by considering cases where quantities such as $E[V \mid X_2 = x_2, W = w]$, $f_{X_2|W}(x_2|w)$, $E[V \mid X_2 = x_2]$, and $f_{X_2}(x_2)$ are extremely smooth (in x_2), so that their Fourier transform has compact support (uniformly in w). Functions such as powers of $\sin(u)/u$, polynomials, and products thereof are simple examples of such functions. In such cases, below a certain bandwidth, the bias term eventually vanishes while the variance and the nonlinear remainders lose their bandwidth dependence. It is then straightforward to see that the bias is negligible and that the nonlinear remainder simply has an extra power of $n^{-1/2}$ relative to the leading term of the asymptotic expansion. Typical functions, of course, often do not have compactly supported Fourier transforms, but one can reasonably expect that Assumption 4.8 would still hold for data generating processes in the neighborhood of this idealized case.

The following theorem establishes asymptotic normality.

THEOREM 4.3. *Suppose the conditions of Lemma* 4.1, *together with Assumptions* 4.1–4.6 *and* 4.8 *hold. Then for* $V \in \{1, Y\}$ *and each* $\lambda \in \{0, \dots, \Lambda\}$ *and* $(x, w) \in \text{supp}(X, W)$, *we have*

$$n^{1/2}\big(\Omega_{V,\lambda}(x, w; h_n)\big)^{-1/2}\big(\hat{g}_{V,\lambda}(x, w; h_n) - g_{V,\lambda}(x, w; 0)\big) \xrightarrow{d} N(0, 1).$$

[11]Other technical but nonrestrictive conditions are needed to ensure no fortuitous cancellations between the different terms in the estimator's asymptotic expansion.

4.2 *Asymptotics for covariate-conditioned average marginal effects*

We now apply our previous general results to obtain the asymptotic properties of estimators of the objects of interest here. First, consider the plug-in estimator for the covariate-conditioned average marginal effect,

$$\hat{\beta}(x, w; h) \equiv \frac{\hat{g}_{Y,1}(x, w; h)}{\hat{g}_{1,0}(x, w; h)} - \frac{\hat{g}_{Y,0}(x, w; h)}{\hat{g}_{1,0}(x, w; h)} \frac{\hat{g}_{1,1}(x, w; h)}{\hat{g}_{1,0}(x, w; h)}$$

for each $(x, w) \in \operatorname{supp}(X, W)$, where the nonparametric estimators $\hat{g}$ are as given above. In the empirical application, the covariate-conditioned average marginal effect of family income on child's reading and math scores can be estimated by $\hat{\beta}(x, w; h)$, which is a functional of $\hat{g}_{V,\lambda}(x, w; h)$. Based on the convergence rate and asymptotic normality for $\hat{g}_{V,\lambda}(x, w; h)$ from Theorems 4.2 and 4.3, we can obtain the asymptotic properties of $\hat{\beta}(x, w; h)$.

Define

$$s_{Y,1}(x, w) \equiv \frac{1}{g_{1,0}(x, w)},$$

$$s_{Y,0}(x, w) \equiv -\frac{g_{1,1}(x, w)}{g_{1,0}(x, w)} \frac{1}{g_{1,0}(x, w)},$$

$$s_{1,1}(x, w) \equiv -\frac{g_{Y,0}(x, w)}{g_{1,0}(x, w)} \frac{1}{g_{1,0}(x, w)},$$

$$s_{1,0}(x, w) \equiv \left(2\frac{g_{Y,0}(x, w)}{g_{1,0}(x, w)} \frac{g_{1,1}(x, w)}{g_{1,0}(x, w)} - \frac{g_{Y,1}(x, w)}{g_{1,0}(x, w)}\right) \frac{1}{g_{1,0}(x, w)}.$$

The results above and a straightforward Taylor expansion yield the following result.

Theorem 4.4. *Suppose the conditions of Theorem 4.2 hold for $\Lambda = 1$ and that $\max_{V \in \{1,Y\}} \max_{\lambda=0,1} \sup_{(x,w)\in\operatorname{supp}(X,W)} |g_{V,\lambda}(x, w)| < \infty$. Further, for $\tau = \tau_n > 0$, define*

$$\Gamma_\tau \equiv \{(x, w) \in \mathbb{R}^2 : f_{X|W}(x \mid w) \geq \tau_n\}.$$

Then

$$\begin{aligned} &\sup_{(x,w)\in\Gamma_\tau} \left|\hat{\beta}(x, w; h_n) - \beta(x, w)\right| \\ &\quad = O\left(\tau^{-3}\left(h_{1n}^{-1}\right)^{\gamma_{1,B}} \exp\left(\alpha_B\left(h_{1n}^{-1}\right)^{\nu_B}\right)\right) \\ &\qquad + O_p\left(\tau^{-3} n^{-1/2}\left(\max\left\{\left(h_{1n}^{-1}\right)^{\delta_L}, h_{2n}^{-1}\right\}\right)\left(h_{1n}^{-1}\right)^{\gamma_{1,L}} \exp\left(\alpha_L\left(h_{1n}^{-1}\right)^{\nu_L}\right)\right), \end{aligned}$$

and there exists a sequence $\{\tau_n\}$ such that $\tau_n > 0$, $\tau_n \to 0$ as $n \to \infty$, and

$$\sup_{(x,w)\in\Gamma_\tau} \left|\hat{\beta}(x, w; h_n) - \beta(x, w)\right| = o_p(1).$$

The delta method gives us the next result.

THEOREM 4.5. *Suppose the conditions of Theorem 4.3 hold for* $\Lambda = 1$ *and that*

$$\max_{V \in \{1,Y\}} \max_{\lambda=0,1} \sup_{(x,w)\in \mathrm{supp}(X,W)} \left|g_{V,\lambda}(x,w)\right| < \infty.$$

Then for all $(x,w) \in \mathrm{supp}(X,W)$,

$$n^{1/2}\big(\Omega_\beta(x,w;h_n)\big)^{-1/2}\big(\hat{\beta}(x,w;h_n) - \beta(x,w)\big) \xrightarrow{d} N(0,1),$$

provided that

$$\Omega_\beta(x,w;h_n) \equiv E\big[\big(\ell_\beta(x,w,h_n;V,X_1,X_2,W)\big)^2\big]$$

is finite and positive for all n sufficiently large, where

$$\begin{aligned}&\ell_\beta(x,w,h;v,x_1,x_2,\tilde{w})\\ &\quad= s_{Y,1}(x,w)\ell_{Y,1}(x,w,h;y,x_1,x_2,\tilde{w}) + s_{Y,0}(x,w)\ell_{Y,0}(x,w,h;y,x_1,x_2,\tilde{w})\\ &\qquad+ s_{1,1}(x,w)\ell_{1,1}(x,w,h;1,x_1,x_2,\tilde{w}) + s_{1,0}(x,w)\ell_{1,0}(x,w,h;1,x_1,x_2,\tilde{w})\end{aligned}$$

and where $\ell_{V,\lambda}$ *is as defined just before Lemma 4.1.*

5. MONTE CARLO SIMULATIONS

This section investigates the finite-sample properties of the proposed estimator through various Monte Carlo experiments. We consider the nonseparable data generating process

$$\begin{aligned}&Y = f_1(X)U_y, \qquad X = 0.5W + U_x, \qquad U_y = f_2(W) + U_u,\\ &X_1 = X + U_1, \qquad X_2 = X + U_2,\end{aligned}$$

where the distributions of each random variable and the explicit forms of f_1, f_2 are specified below, and where Y, W, X_1, and X_2 are standardized to have mean 0 and standard deviation 1. We assume $U_x \perp U_u \mid W$, which implies $X \perp U_y \mid W$.[12] The variables (Y, X_1, X_2, W) are used as an input for our estimator, and the variables (Y, X_1, W) are used for the local linear estimator that neglects the measurement error. We also use the variables (Y, X, W) to construct an infeasible local linear estimator, and (Y, X, X_2, W) and (Y, X_1, X, W) to construct other infeasible versions of our estimator for purposes of comparison. For those estimators, we consider flat-top kernels of infinite order. In our estimators,[13] the Fourier transform, $\kappa(\cdot)$, associated with X is given in eq. (2) with $\bar{\xi} = 0.5$. We use a different flat-top kernel for W, which was introduced in Politis and

[12]In the simulations, we assume $U_x \perp (U_u, W)$, which implies $U_x \perp U_u \mid W$ by Lemma 4.3 of Dawid (1979). Lemma 4.1 of Dawid then ensures that $U_x \perp U_u \mid W$ implies $X \perp U_y \mid W$.

[13]For the local linear estimator, the same flat-top kernel is used for X and W since estimation results are not sensitive to the choice of the kernel.

Romano (1999):

$$k_h(x) \equiv \frac{h}{2\pi} \frac{\sin^2(2\pi x/h) - \sin^2(\pi x/h)}{\pi^2 x^2}.$$

All estimates are constructed for the values $x = 0$ and $w = 1$. For our estimators, we scan a set of bandwidths[14] ranging from 7 to 12.5 for X and from 3.5 to 6 for W in increments of 0.05 so as to find the optimal bandwidth minimizing the root mean squared error (RMSE). For both local linear estimators, we scan a set of bandwidths ranging from 2.5 to 6 for X and from 1.5 to 3.5 for W, with the same increments. All simulations draw 500 samples of 1000, 2000, or 8000 observations.

We examine a total of 16 combinations of ordinary and supersmooth distributions for random variables and functions f_1 and f_2, as given in Table 1. As in Schennach (2004a), we consider the Laplace distribution as an example of an ordinarily smooth distribution. The Laplace distribution density, denoted by $L(t; \mu, \sigma^2)$, is defined by

$$\frac{1}{\sigma\sqrt{2}} \exp\bigl(-\sigma|t - \mu|\sqrt{2}\bigr)$$

for any $t \in \mathbb{R}$ with mean μ and variance σ^2. Its characteristic function has a tail of the form $|\zeta|^{-2}$. The normal distribution with variance σ^2 is used as an example of a supersmooth distribution. The tail of the characteristic function of the normal distribution is

Table 1. Monte Carlo simulation designs.

Example	U_x	W	U_1, U_2	U_u	$f_1(X)$	$f_2(W)$
1	$N(0, 0.5)$	$N(0, 1)$	$N(0, 0.25)$	$N(0, 0.09)$	$\text{erf}(x)$	$S(w)$
2	$N(0, 0.5)$	$N(0, 1)$	$N(0, 0.25)$	$N(0, 0.09)$	$\text{erf}(x)$	$\text{erf}(w)$
3	$N(0, 0.5)$	$N(0, 1)$	$L(0, 0.25)$	$N(0, 0.09)$	$\text{erf}(x)$	$\text{erf}(w)$
4	$N(0, 0.5)$	$N(0, 1)$	$L(0, 0.25)$	$N(0, 0.09)$	$\text{erf}(x)$	$S(w)$
5	$N(0, 0.5)$	$L(0, 1)$	$N(0, 0.25)$	$N(0, 0.09)$	$\text{erf}(x)$	$\text{erf}(w)$
6	$N(0, 0.5)$	$L(0, 1)$	$N(0, 0.25)$	$N(0, 0.09)$	$\text{erf}(x)$	$S(w)$
7	$N(0, 0.5)$	$L(0, 1)$	$L(0, 0.25)$	$N(0, 0.09)$	$\text{erf}(x)$	$\text{erf}(w)$
8	$N(0, 0.5)$	$L(0, 1)$	$L(0, 0.25)$	$N(0, 0.09)$	$\text{erf}(x)$	$S(w)$
9	$L(0, 0.5)$	$N(0, 1)$	$N(0, 0.25)$	$N(0, 0.09)$	$\text{erf}(x)$	$\text{erf}(w)$
10	$L(0, 0.5)$	$N(0, 1)$	$N(0, 0.25)$	$N(0, 0.09)$	$\text{erf}(x)$	$S(w)$
11	$L(0, 0.5)$	$N(0, 1)$	$L(0, 0.25)$	$N(0, 0.09)$	$\text{erf}(x)$	$\text{erf}(w)$
12	$L(0, 0.5)$	$N(0, 1)$	$L(0, 0.25)$	$N(0, 0.09)$	$\text{erf}(x)$	$S(w)$
13	$L(0, 0.5)$	$L(0, 1)$	$N(0, 0.25)$	$N(0, 0.09)$	$\text{erf}(x)$	$\text{erf}(w)$
14	$L(0, 0.5)$	$L(0, 1)$	$N(0, 0.25)$	$N(0, 0.09)$	$\text{erf}(x)$	$S(w)$
15	$L(0, 0.5)$	$L(0, 1)$	$L(0, 0.25)$	$N(0, 0.09)$	$\text{erf}(x)$	$\text{erf}(w)$
16	$L(0, 0.5)$	$L(0, 1)$	$L(0, 0.25)$	$N(0, 0.09)$	$\text{erf}(x)$	$S(w)$

[14]Note that the flat-top kernel has a very narrow central peak, so that even moderately large bandwidths result in highly local smoothing.

of the form $\exp(-(\sigma^2/2)|\zeta|^2)$. Our example of an ordinarily smooth function for $f_2(W)$ is a piecewise linear continuous function with a discontinuous first derivative

$$S(W) \equiv \begin{cases} -1 & \text{if } W < -1, \\ W & \text{if } W \in [-1, 1], \\ 1 & \text{if } W > 1, \end{cases}$$

whose Fourier transform decays at the rate $|\zeta|^{-2}$ as $|\zeta| \to \infty$. As an example of a supersmooth function for $f_1(X)$ or $f_2(W)$, we consider the error function

$$\operatorname{erf}(V) \equiv \frac{2}{\sqrt{\pi}} \int_0^V e^{-t^2}\,dt,$$

which has a Fourier transform decaying at the rate $|\zeta|^{-1}\exp(-\frac{1}{4}|\zeta|^2)$ as $|\zeta| \to \infty$ for $V = X$ or W.

Table 2 reports the bias squared, variance, and RMSE of the five estimators, which are functions of bandwidth for a sample size of 1000, for Example 1.[15] The headings "Fourier 1, 2, and 3" refer to our estimators, which are based on variables (Y, X_1, X_2, W), (Y, X, X_2, W), and (Y, X_1, X, W), respectively. The headings "Local Linear" and "No Measurement Error" refer to local linear estimators that use variables (Y, X_1, W) and (Y, X, W), respectively. We show results from only a subset of the bandwidths for conciseness. For each choice of bandwidth, the bias squared, variance, and RMSE are reported in the first, second, and third rows, respectively. The results for the optimal bandwidth are reported at the bottom of each estimator.

A few remarks are in order. We find that our estimator is as effective in reducing bias as the infeasible local linear estimator using the true covariate X. In contrast and as expected, the bias from the feasible local linear estimator ignoring the measurement error does not shrink toward zero as bandwidth decreases. Our estimator also gives smaller variance than the error-contaminated local linear estimator. As a result, our estimator outperforms the feasible local linear estimator in terms of RMSE. Interestingly, all Fourier estimators perform better than the infeasible local linear estimator using X. A useful direction for further research is to investigate under what conditions and why Fourier-based estimators outperform local linear estimators. Comparing the Fourier estimators, we can see the role of clean data as well as the asymmetry between two measurement errors in Assumption 3.2. Interestingly, Fourier 1 and Fourier 2 obtain quite similar estimation results, but Fourier 3 outperforms them. Thus, it appears that one would want to choose X_2 to be the less mismeasured of the two measures of X.

Table 3 reports Monte Carlo simulation results for the convergence rate as a function of sample size for each example. The RMSE's in all examples decrease as expected as sample size increases, corroborating our theoretical results.

[15] We only report this example due to space limitations, but the performance of the estimators is similar for all the examples.

TABLE 2. Monte Carlo simulation results for Example 1.

Fourier 1

$h_1 \setminus h_2$		4	4.25	4.5	4.75	5	5.25	5.5
9.5	Bias^2	0.01753	0.00433	0.00041	0.00289	0.02462	0.05118	0.11600
	Variance	0.48058	0.49628	0.42267	0.38406	0.38546	0.30834	0.31586
	RMSE	0.70577	0.70753	0.65044	0.62205	0.64038	0.59959	0.65716
9.75	Bias^2	0.00666	0.00226	0.00018	0.00834	0.03165	0.07544	0.13614
	Variance	0.46158	0.40620	0.37223	0.34772	0.32196	0.30496	0.27895
	RMSE	0.68428	0.63911	0.61025	0.59670	0.59465	0.61676	0.64427
10	Bias^2	0.00347	0.00036	0.00211	0.01624	0.04329	0.08934	0.15408
	Variance	0.42542	0.38354	0.36214	0.34860	0.31326	0.28803	0.26343
	RMSE	0.65490	0.61959	0.60353	0.60402	0.59712	0.61430	0.64615
10.25	Bias^2	0.00115	0.00010	0.00544	0.02394	0.05677	0.10781	0.17322
	Variance	0.40351	0.37447	0.35360	0.34020	0.31521	0.28968	0.25723
	RMSE	0.63613	0.61202	0.59920	0.60344	0.60990	0.63047	0.65609
10.5	Bias^2	0.00000	0.00281	0.01131	0.03279	0.06980	0.12437	0.20091
	Variance	0.39215	0.38599	0.34373	0.32113	0.29737	0.27343	0.25757
	RMSE	0.62622	0.62353	0.59586	0.59492	0.60595	0.63072	0.67712
10.75	Bias^2	0.00125	0.00762	0.02094	0.04758	0.08990	0.14961	0.23111
	Variance	0.36811	0.36257	0.33230	0.31047	0.28750	0.26437	0.24904
	RMSE	0.60775	0.60843	0.59434	0.59837	0.61433	0.64341	0.69293
11	Bias^2	0.00658	0.01626	0.03543	0.06770	0.11861	0.18107	0.26360
	Variance	0.35396	0.33839	0.31956	0.29858	0.28493	0.25426	0.23256
	RMSE	0.60045	0.59552	0.59581	0.60521	0.63525	0.65980	0.70439

Optimal

h_1	h_2	Bias^2	Variance	RMSE
9.75	4.9	0.02029	0.33234	0.59382

(*Continues*)

TABLE 2. *Continued.*

Local Linear

$h_1 \setminus h_2$		2	2.25	2.5	2.75	3	3.25	3.5
3	Bias^2	0.32857	0.46799	0.47779	0.55940	0.57653	0.80000	1.30221
	Variance	18.22160	23.44127	12.40588	12.52235	6.33969	9.20820	8.23076
	RMSE	4.30699	4.88971	3.58938	3.61687	2.62987	3.16357	3.08755
3.25	Bias^2	0.17816	0.13128	0.38641	0.59926	0.88147	0.87869	1.14468
	Variance	13.89843	8.14524	2.36372	2.23755	4.67790	0.96041	1.13249
	RMSE	3.75188	2.87689	1.65835	1.68428	2.35783	1.35613	1.50903
3.5	Bias^2	0.60058	0.76277	0.89476	0.87785	0.99725	0.99212	1.55899
	Variance	5.34878	0.65439	1.83879	0.66070	1.13176	3.89185	2.75340
	RMSE	2.43913	1.19044	1.65334	1.24038	1.45912	2.20997	2.07663
3.75	Bias^2	1.03708	0.82405	1.01636	0.99417	1.22282	1.41411	1.55216
	Variance	2.28356	2.11708	0.43360	0.56335	0.27968	0.56551	1.04877
	RMSE	1.82226	1.71497	1.20414	1.24801	1.22577	1.40699	1.61274
4	Bias^2	1.04103	1.06472	1.20429	1.18884	1.25602	1.56423	1.71529
	Variance	0.62326	1.20819	0.22532	0.48526	1.37113	0.60255	0.36498
	RMSE	1.29007	1.50762	1.19566	1.29387	1.62085	1.47200	1.44231
4.25	Bias^2	1.16685	1.27326	1.14203	1.46695	1.41502	1.39306	1.87139
	Variance	0.29115	0.16609	2.47443	1.10869	0.63847	2.81544	1.67206
	RMSE	1.20748	1.19973	1.90170	1.60488	1.43300	2.05146	1.88241
4.5	Bias^2	1.33409	1.35225	1.39334	1.38871	1.60060	1.74283	1.94063
	Variance	0.25767	0.26342	0.16547	0.85056	0.21279	0.64793	0.97151
	RMSE	1.26165	1.27109	1.24853	1.49642	1.34662	1.54621	1.70650

Optimal

h_1	h_2	Bias^2	Variance	RMSE
3.5	2.55	0.78695	0.44977	1.11208

(*Continues*)

Table 2. *Continued.*

No Measurement Error

$h_1 \setminus h_2$		2	2.25	2.5	2.75	3	3.25	3.5
3	Bias^2	0.86469	0.53839	0.42244	0.22939	0.12645	0.02125	0.00891
	Variance	10.73736	4.36472	3.03947	4.04071	3.18704	1.00666	0.90149
	RMSE	3.40618	2.21430	1.86062	2.06642	1.82030	1.01386	0.95415
3.25	Bias^2	0.03877	0.08256	0.04476	0.01467	0.00120	0.02553	0.12702
	Variance	5.37811	1.13461	0.70957	0.58603	1.61839	0.43638	0.86019
	RMSE	2.32742	1.10326	0.86852	0.77505	1.27263	0.67964	0.99358
3.5	Bias^2	0.00421	0.00020	0.00430	0.01584	0.03653	0.10939	0.20268
	Variance	4.98230	0.57077	0.41955	0.41517	0.68762	0.32481	0.79210
	RMSE	2.23305	0.75563	0.65104	0.65651	0.85097	0.65894	0.99738
3.75	Bias^2	0.03600	0.04366	0.07150	0.08675	0.13417	0.21329	0.32440
	Variance	0.54931	0.41812	0.68290	0.36188	0.30962	0.27867	0.45446
	RMSE	0.76505	0.67955	0.86856	0.66980	0.66618	0.70139	0.88253
4	Bias^2	0.08723	0.12965	0.13026	0.14728	0.19202	0.30910	0.44513
	Variance	1.40275	0.36412	0.29826	0.44403	0.84644	0.23775	0.54903
	RMSE	1.22064	0.70269	0.65461	0.76897	1.01905	0.73949	0.99708
4.25	Bias^2	0.18980	0.20088	0.23143	0.24291	0.24317	0.40123	0.42689
	Variance	0.46710	0.42852	0.34077	0.25060	2.37821	0.21447	1.71866
	RMSE	0.81049	0.79334	0.75644	0.70250	1.61907	0.78467	1.46477
4.5	Bias^2	0.17285	0.31214	0.29557	0.32417	0.39853	0.48788	0.71452
	Variance	4.72205	3.71663	0.22839	0.22542	0.21802	0.20143	3.08174
	RMSE	2.21244	2.00718	0.72386	0.74135	0.78521	0.83025	1.94840

Optimal

h_1	h_2	Bias^2	Variance	RMSE
3.7	2.55	0.04578	0.33558	0.61754

(*Continues*)

TABLE 2. *Continued.*

Fourier 2

$h_1 \setminus h_2$		4	4.25	4.5	4.75	5	5.25	5.5
9.5	Bias^2	0.01755	0.00434	0.00041	0.00288	0.02459	0.05114	0.11595
	Variance	0.48055	0.49628	0.42266	0.38406	0.38548	0.30835	0.31588
	RMSE	0.70576	0.70755	0.65044	0.62205	0.64037	0.59958	0.65714
9.75	Bias^2	0.00667	0.00227	0.00018	0.00832	0.03162	0.07540	0.13608
	Variance	0.46156	0.40618	0.37223	0.34772	0.32197	0.30497	0.27896
	RMSE	0.68428	0.63911	0.61025	0.59669	0.59463	0.61674	0.64424
10	Bias^2	0.00348	0.00036	0.00210	0.01622	0.04325	0.08929	0.15402
	Variance	0.42540	0.38353	0.36214	0.34860	0.31327	0.28805	0.26345
	RMSE	0.65489	0.61959	0.60352	0.60400	0.59709	0.61428	0.64611
10.25	Bias^2	0.00116	0.00009	0.00543	0.02391	0.05674	0.10776	0.17315
	Variance	0.40351	0.37447	0.35360	0.34021	0.31522	0.28970	0.25725
	RMSE	0.63613	0.61201	0.59919	0.60342	0.60988	0.63044	0.65605
10.5	Bias^2	0.00000	0.00279	0.01129	0.03276	0.06976	0.12432	0.20085
	Variance	0.39214	0.38599	0.34374	0.32114	0.29738	0.27345	0.25759
	RMSE	0.62621	0.62352	0.59584	0.59489	0.60592	0.63069	0.67708
10.75	Bias^2	0.00124	0.00760	0.02091	0.04754	0.08984	0.14954	0.23103
	Variance	0.36810	0.36257	0.33231	0.31048	0.28751	0.26438	0.24906
	RMSE	0.60774	0.60842	0.59432	0.59835	0.61430	0.64337	0.69288
11	Bias^2	0.00657	0.01624	0.03539	0.06766	0.11855	0.18100	0.26352
	Variance	0.35395	0.33839	0.31957	0.29859	0.28495	0.25428	0.23258
	RMSE	0.60043	0.59550	0.59579	0.60518	0.63522	0.65975	0.70434

Optimal

h_1	h_2	Bias^2	Variance	RMSE
9.7	4.95	0.02318	0.32942	0.59380

(*Continues*)

Table 2. *Continued.*

		Fourier 3						
$h_1 \setminus h_2$		4	4.25	4.5	4.75	5	5.25	5.5
9.5	Bias^2	0.01718	0.00402	0.00034	0.00309	0.02540	0.05223	0.11746
	Variance	0.47176	0.48737	0.41483	0.37669	0.37827	0.30180	0.30948
	RMSE	0.69925	0.70100	0.64434	0.61626	0.63535	0.59500	0.65341
9.75	Bias^2	0.00628	0.00209	0.00024	0.00872	0.03243	0.07676	0.13761
	Variance	0.45267	0.39809	0.36465	0.34060	0.31528	0.29859	0.27294
	RMSE	0.67746	0.63260	0.60406	0.59104	0.58966	0.61266	0.64074
10	Bias^2	0.00325	0.00029	0.00229	0.01677	0.04418	0.09067	0.15586
	Variance	0.41696	0.37567	0.35477	0.34165	0.30676	0.28192	0.25767
	RMSE	0.64824	0.61316	0.59754	0.59868	0.59240	0.61040	0.64306
10.25	Bias^2	0.00107	0.00013	0.00572	0.02445	0.05770	0.10913	0.17508
	Variance	0.39536	0.36680	0.34640	0.33360	0.30883	0.28377	0.25161
	RMSE	0.62962	0.60575	0.59340	0.59837	0.60542	0.62682	0.65322
10.5	Bias^2	0.00000	0.00300	0.01171	0.03351	0.07090	0.12590	0.20265
	Variance	0.38428	0.37847	0.33674	0.31457	0.29121	0.26763	0.25218
	RMSE	0.61990	0.61763	0.59030	0.58999	0.60176	0.62732	0.67441
10.75	Bias^2	0.00136	0.00792	0.02146	0.04843	0.09112	0.15124	0.23294
	Variance	0.36041	0.35536	0.32555	0.30413	0.28155	0.25876	0.24383
	RMSE	0.60147	0.60272	0.58908	0.59377	0.61047	0.64032	0.69048
11	Bias^2	0.00682	0.01668	0.03610	0.06869	0.11988	0.18283	0.26578
	Variance	0.34657	0.33147	0.31308	0.29249	0.27918	0.24887	0.22748
	RMSE	0.59446	0.59004	0.59091	0.60098	0.63171	0.65704	0.70233

Optimal

h_1	h_2	Bias^2	Variance	RMSE
9.75	4.9	0.02090	0.32549	0.58855

Table 3. Monte Carlo simulation results as a function of sample size.

Sample Size		1000	2000	8000
Example 1				
Bandwidth	h_1	9.75	9.55	9.85
	h_2	4.9	4.9	4.85
Bias^2		0.02029	0.00149	0.00004
Variance		0.33234	0.09040	0.01433
RMSE		0.59382	0.30314	0.11987
Example 2				
Bandwidth	h_1	11.7	11.35	11.05
	h_2	4.75	4.85	5
Bias^2		0.01025	0.00062	0.00049
Variance		0.20523	0.05342	0.04740
RMSE		0.46420	0.23247	0.21885
Example 3				
Bandwidth	h_1	11.2	11.55	11.25
	h_2	5.05	4.9	4.9
Bias^2		0.00968	0.00044	0.00016
Variance		0.19729	0.04240	0.02930
RMSE		0.45493	0.20699	0.17164
Example 4				
Bandwidth	h_1	9.5	9.55	9.7
	h_2	5	5	4.4
Bias^2		0.01226	0.00167	0.00004
Variance		0.25057	0.09960	0.01405
RMSE		0.51267	0.31823	0.11870
Example 5				
Bandwidth	h_1	9.75	9.45	9.3
	h_2	4.35	4.4	4.4
Bias^2		0.01764	0.00145	0.00070
Variance		0.28277	0.09061	0.05482
RMSE		0.54809	0.30342	0.23562
Example 6				
Bandwidth	h_1	8.4	8.6	8.75
	h_2	4.6	4.4	4.35
Bias^2		0.02251	0.00490	0.00031
Variance		0.34412	0.18187	0.04241
RMSE		0.60550	0.43216	0.20669
Example 7				
Bandwidth	h_1	9.1	9.25	9.4
	h_2	4.6	4.5	4.45
Bias^2		0.01697	0.00080	0.00010
Variance		0.28118	0.06201	0.02290
RMSE		0.54604	0.25061	0.15167
Example 8				
Bandwidth	h_1	8.25	8.25	8.25
	h_2	4.75	4.5	4.7
Bias^2		0.01510	0.00117	0.00009
Variance		0.29672	0.09851	0.01682
RMSE		0.55841	0.31571	0.13004

(*Continues*)

Table 3. *Continued.*

Sample Size		1000	2000	8000
Example 9				
Bandwidth	h_1	9.4	9.65	9.9
	h_2	5.05	4.95	5
Bias^2		0.00927	0.00310	0.00002
Variance		0.22173	0.12096	0.01247
RMSE		0.48062	0.35221	0.11172
Example 10				
Bandwidth	h_1	8.8	8.65	8.85
	h_2	5.05	5.05	5
Bias^2		0.01473	0.003151	0.00083
Variance		0.27797	0.155258	0.06535
RMSE		0.54101	0.398006	0.25726
Example 11				
Bandwidth	h_1	9.55	9.55	9.8
	h_2	5.1	5.05	5
Bias^2		0.00629	0.00120	0.00005
Variance		0.17604	0.07214	0.01288
RMSE		0.42700	0.27080	0.11372
Example 12				
Bandwidth	h_1	8.85	8.75	8.55
	h_2	4.95	5.2	5.15
Bias^2		0.02034	0.00231	0.00134
Variance		0.32799	0.12246	0.09321
RMSE		0.59020	0.35324	0.30749
Example 13				
Bandwidth	h_1	8.45	8.45	8.35
	h_2	4.7	4.75	4.8
Bias^2		0.01659	0.00259	0.00010
Variance		0.28442	0.11723	0.02484
RMSE		0.54864	0.34615	0.15793
Example 14				
Bandwidth	h_1	7.25	7.65	7.2
	h_2	4.7	4.6	4.8
Bias^2		0.02838	0.00128	0.00102
Variance		0.40722	0.09167	0.07542
RMSE		0.66000	0.30487	0.27648
Example 15				
Bandwidth	h_1	8.4	8.6	8.3
	h_2	4.7	4.6	4.7
Bias^2		0.01968	0.00131	0.00037
Variance		0.30630	0.07827	0.04837
RMSE		0.57094	0.28210	0.22078
Example 16				
Bandwidth	h_1	7.4	7.65	7.65
	h_2	4.8	4.65	4.6
Bias^2		0.01692	0.00159	0.00085
Variance		0.32160	0.10115	0.07403
RMSE		0.58182	0.32054	0.27364

6. Application: The impact of family income on child achievement

This section applies our estimator to study the causal effect of family income on child achievement. We also discuss practical methods for the optimal bandwidth choices in Appendix A, since estimation results depend crucially on this choice.

6.1 *The model*

The association between family income and child development is a contentious issue in economics, sociology, and developmental psychology. Even though it has been examined in a number of studies, there is no consensus on the relative effectiveness of income transfers and direct intervention in augmenting the human capital of children. Income transfers could have a significant impact on the economic well-being of children growing up in poor families if family income plays a substantial role in child development. If not, then direct interventions, such as the Head Start program, to improve child health, education, and parenting may be more effective.

Using data from the Panel Study of Income Dynamics (PSID), Duncan, Yeung, Brooks-Gunn, and Smith (1998) find that family income in early childhood has the greatest impact on completed schooling, especially among children in families with low incomes, regardless of whether they control for fixed family effects. Blau (1999) uses the matched mother–child subsample of the National Longitudinal Survey of Youth (NLSY) to estimate the impact of parental income on children's cognitive, social, and emotional development. He finds that ordinary least squares (OLS) estimates of income effects are generally statistically significant and positive, but that they are smaller and insignificant when he uses either random- or fixed-effect methods. In addition, his findings indicate that the effect of permanent income is much larger, but not large enough to make income transfer a feasible approach to achieving substantial improvements in child outcomes. He also finds that there is no evidence for any systematic indication of diminishing returns to income, that is, income effects that are larger at lower levels of income.

Aughinbaugh and Gittleman (2003) examine the relationship between child development and income in Great Britain and compare it with that in the United States. Using the NLSY and Great Britain's National Child Development Study, they find that the relationship between income and child development is quite similar in the two countries. Income tends to improve cognitive test scores, but the magnitude of the impact is small. Using participants from the National Institute of Child Health and Human Development (NICHD) study of early child care, Taylor, Dearing, and McCartney (2004) estimate the impact of family economic resources on developmental outcomes in early childhood. They find that economic resources are important when properly compared with other important variables, such as maternal verbal intelligence, and when compared with established interventions, such as Early Head Start. Their findings also indicate that there are significant nonlinear effects of permanent (but not current) income, implying that income effects are larger for children living in poor families.

Dahl and Lochner (2012) address both omitted variables bias and attenuation bias due to measurement error on family income using fixed-effect (parametric) instrumental variables estimation. They use panel data on over 6000 children matched to their mothers in the NLSY data. They find that estimates from the fixed-effect instrumental variables approach are larger than cross-section OLS or standard fixed-effects estimates, so that current income has a significant effect on a child's math and reading test scores.

Here we examine the effect of family income on child achievement, as measured by scores on math and reading assessments. We accommodate measurement errors, endogeneity of family income, nonlinearity of income effects, and interactions between drivers of child achievement by considering a data generating process of the form

$$Y = r(X, U_y),$$

where Y is the child's scholastic achievement, X is family income, and U_y represents other unobserved drivers of child achievement; r is an unknown measurable scalar-valued function. Because unobserved parental abilities could be common causes of both family incomes and child achievement, X is generally correlated with the unobserved U_y. Moreover, income is noisily measured in most surveys, and the data used here are no exception.

Figure 1 depicts structural relations consistent with Assumptions 2.1–2.3. Arrows denote direct causal relationships. Dashed circles denote unobservables and complete circles denote observables. A line without an arrow denotes dependence arising from a causal relation in either direction or the existence of an underlying common cause. Mother's cognitive ability is a common cause for family earning potential and child ability. The fact that earning potential and child ability share a common cause induces a correlation between family income and child ability. Nevertheless, the conditional inde-

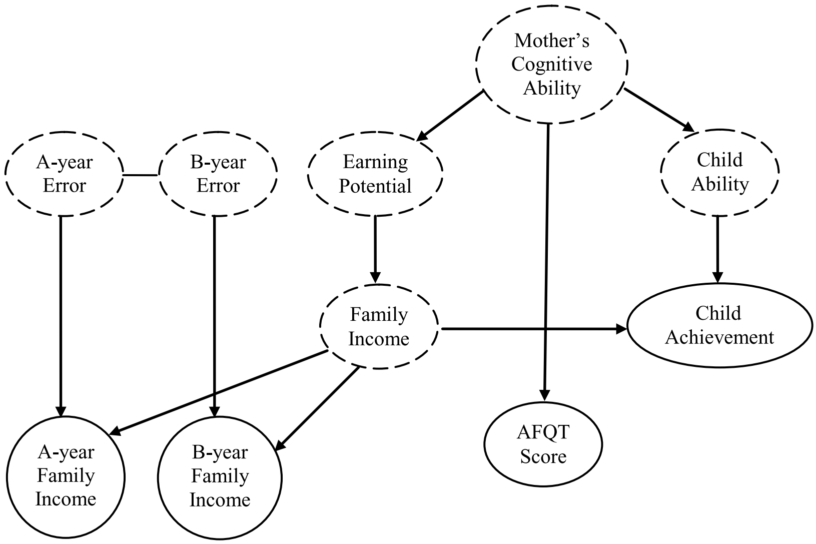

FIGURE 1. Impact of family income on child achievement. Due to a common cause, mother's cognitive ability, family income, and child achievement are correlated. The AFQT score, a proxy for the common cause, plays a key role as a control variable ensuring conditional independence between family income and child's achievement. Two error-laden measurements of family income are used to remove attenuation bias due to measurement error in family income.

pendence assumption makes it possible to recover features of the causal relationship. Because mother's Armed Forces Qualifying Test (AFQT) scores, a proxy for mother's cognitive ability, are observable, they can serve as control variables or conditioning instruments to ensure the conditional independence between family income and unobserved child ability. In essence, by fixing (or conditioning on) the underlying cause of endogeneity (parental ability), we can "turn off" the indirect effect via parental and children abilities, and study the direct effect of parental income on children's scholastic achievements. As a result, any residual heterogeneity in achievements (as modeled by U_y) would plausibly be expected to be unrelated (i.e., independent) to parental income. A separate issue is that true family income is unobservable because income is noisily measured in survey data. Without correcting for the measurement error, estimates would be biased. Fortunately, we observe two error-laden measurements of true family income. This permits us to recover the desired effect measures using our new estimator.

It is worth noting that the use of parental skills as control variables has been common in the literature on human development (e.g., Cunha and Heckman (2007), Cunha, Heckman, and Schennach (2010)). It would be ideal to include father's skills and even other control variables such as location and parental education level. Due to the limitation of the data availability, however, we only include mother's AFQT scores. To the extent that father's and mother's skills are correlated, the omission of father's skills is not too detrimental. Our limited choice of control variables is clearly not ideal, but it should nevertheless capture a large part of the endogeneity. Apparently, testing the conditional independence using control variables would be useful for the empirical works. Since it is beyond the scope of the current paper, we leave it to future research.

We use the matched mother–child subsample of the NLSY from Dahl and Lochner (2012) in the cross-sectional nonparametric model.[16] The dependent variables (i.e., child scholastic achievement (Y)) are measures of achievement in math and reading based on standardized scores of the Peabody Individual Achievement Tests (PIAT). Math achievement is measured by mathematics scores, and reading achievement is measured by a simple average of the reading recognition and reading comprehension scores. Our error-laden measurement of family income (X_1) is the natural logarithm (log) of reported family income in 1998. The log of reported family income in year 2000 is used as an additional error-laden measurement of family income (X_2). Income in each year is after tax and after transfer. Because the assumptions regarding the measurement error do allow for a systematic drift in income (Assumption 3.2(ii) does not impose zero conditional mean on the second measurement, X_2), family income in year 2000 (X_2) is allowed to have a nonzero deterministic trend. The control variable (W) is the mother's AFQT score; see Tables 4(a) and 4(b) for descriptive statistics, and see Dahl and Lochner (2012) for further details. We assume true family incomes and unobserved drivers of child achievement are independent, conditional on AFQT scores (i.e., $X \perp U_y \mid W$). Observe that this type of conditioning instrument is conceptually rather different from conventional instruments satisfying standard exclusion restrictions. We create standardized

[16]We thank Gordon Dahl for providing the NLSY data.

TABLE 4. Descriptive statistics.

	Mean	SD
(a) Impact of family income on math score		
Math score	0.07510	1.03763
Family income (year 1998)	2.42646	1.65861
Family income (year 2000)	2.51579	1.78155
Mother's AFQT	−0.07766	0.99484
Observations	1544	
(b) Impact of family income on reading score		
Reading score	0.06053	0.91971
Family income (year 1998)	2.37502	1.57405
Family income (year 2000)	2.47168	1.67452
Mother's AFQT	−0.08737	0.98670
Observations	1274	

TABLE 5. Optimal choice of smoothing parameters.

		α	h_1	h_2
Math	Fourier	10^{-13}	6.8	5.7
	Local linear	5×10^{-4}	6.8	7
Reading	Fourier	10^{-12}	7.5	7
	Local linear	10^{-1}	7.1	6.9

test scores, AFQT scores, and family incomes having mean 0 and standard deviation 1.[17] For later reference, let the scaling constant for X be denoted σ_x.

We apply leave-one-out cross-validation to estimate optimal bandwidths. Since true X is unobserved, modifications of the original method in Wahba (1990) are necessary. Details on the new method can be found in the Appendix A. Table 5 reports the estimated optimal bandwidths h and a smoothing parameter α associated with penalization on the smoothness of the nonparametric estimate of the expectation of response function $\hat{\mu}_h$, measured by the second derivative of $\hat{\mu}_h$. The selection rules of the optimal smoothing parameters are applied to our proposed Fourier estimator and conventional local linear estimator neglecting measurement errors. For the local linear methods, we use a second-order local polynomial estimator to obtain the smoothing parameters because this automatically estimates the second derivatives of $\hat{\mu}_h$, and both local linear and local polynomial estimators are first-order identical. Since the Fourier estimator allows greater roughness in $\hat{\mu}_h$, this results in a larger penalty for given α. The estimated choice of α here is consequently fairly small.

[17] Since the log of family incomes for 1998 and 2000 share similar sample means and standard deviations, we standardize them using the average of their means and the square root of the average of the their variances. This common standardization conforms to Assumption 2.3, whereas standardizing separately would not.

6.2 *Estimation results*

Tables 6 and 7 show estimation results obtained by our new estimator and a local linear estimator ignoring the family income measurement error. These results use the smoothing parameters given in Table 5. Each estimate is evaluated at given values of standardized family income (X) ranging from -2.8 to 1.6 and standardized mother's AFQT score (W) ranging from -1.2 to 1.2 in increments of 0.2. Estimates for only a subset of the covariate values are reported for conciseness. We report bootstrap standard error estimates. As Gonçalves and White (2005) note, one must formally justify using the bootstrap to compute standard errors, because the consistency of the bootstrap distribution does not guarantee the consistency of the variance of the bootstrap distribution as an estimator of the asymptotic variance. Nevertheless, the bootstrap should give us standard errors with first-order accuracy, sufficient for our purposes.

Table 6 reports the estimated impact of family income on children's math achievement. The covariate-conditioned average marginal effects[18] of family income on children's math achievement from our estimator are positive over all ranges of x and w, and are large and significant for most values. The average marginal effect is about 3.8099 at $x = -2.8$ and $w = -0.6$. This implies that the expected effect of a 1% increase in standardized family income is to increase a child's math score by about 2.2% of a standard deviation.[19] For a given mother's AFQT score, w, effects decrease as standardized family income, x, increases toward about 0, but increase again when standardized family income is above 0. Significantly, the covariate-conditioned average marginal effects estimated from the local linear estimator are much smaller than those from our estimator for all (x, w) values. The effects appear insignificant for most poor families.[20] The gap between the two estimators is especially large for low-income families with high standardized mother's AFQT score (around 0.6). Note that the average marginal effect from the local linear estimator is about -0.4724 at $x = -2.8$ and $w = -0.6$, whereas that from our estimator is 3.8099, a difference of about 4.2823. It follows that *measurement errors in family income have an important impact on estimated effects on child achievement; properly handling these errors is critical to obtaining accurate estimates of these effects.*

Table 7 shows the impact of family income on children's reading achievement. The covariate-conditioned average marginal effects of family income on reading score from our estimator are also much larger than those from the local linear estimator in all ranges of (x, w). The average marginal effect from our estimator, for instance, is about 3.0323 at $x = -2.8$ and $w = -0.6$, whereas that from the local linear estimator is -0.3037. Our Fourier estimate implies that the expected effect of a 1% increase in standardized family

[18]Note that because of the conditioning, covariate-conditioned average marginal effects provide more accurate estimates of unknown marginal effects than their unconditional analogs.

[19]With $\sigma_x = 1.72118$, $3.8099 \times \ln(1.01)/\sigma_x = 0.022025$.

[20]Near the boundary values for family incomes, the local linear estimated marginal effects are extremely irregular, showing huge fluctuations in marginal effects for small changes in family income.

TABLE 6. Impact of family income on children's math achievement.

$w \setminus x$		−2.8	−2.6	−2	−1.4	−0.8	−0.2	0.4	1	1.6
−1.2	Fourier	3.1105	2.7091	1.9599	1.5823	1.3978	1.3389	1.3876	1.5627	1.9358
		(0.3801)	(0.3524)	(0.3110)	(0.2837)	(0.2629)	(0.2501)	(0.2405)	(0.2348)	(0.2376)
	Local linear	0.0563	0.0784	−0.0164	0.1046	0.1396	0.1772	0.2324	0.4018	0.0527
		(0.0537)	(0.0516)	(0.1978)	(0.0571)	(0.0527)	(0.0614)	(0.0771)	(0.1905)	(0.2583)
−1	Fourier	3.3676	2.9194	2.0892	1.6731	1.4682	1.3979	1.4401	1.6113	1.9806
		(0.2593)	(0.2695)	(0.2773)	(0.2719)	(0.2646)	(0.2576)	(0.2513)	(0.2467)	(0.2458)
	Local linear	−0.0306	0.0224	−0.2071	0.0933	0.1355	0.1786	0.2379	0.3859	−0.1320
		(0.0769)	(0.0609)	(0.4812)	(0.0573)	(0.0535)	(0.0550)	(0.0719)	(0.1421)	(0.2494)
−0.6	Fourier	3.8099	3.2708	2.2882	1.8020	1.5598	1.4670	1.4933	1.6496	1.9973
		(0.2813)	(0.2918)	(0.3011)	(0.2999)	(0.2961)	(0.2921)	(0.2882)	(0.2851)	(0.2844)
	Local linear	−0.4724	−0.1210	0.2064	0.0670	0.1255	0.1802	0.2480	0.3689	25.6584
		(0.2267)	(0.0942)	(0.1884)	(0.0559)	(0.0472)	(0.0531)	(0.0616)	(0.1040)	(16.0568)
−0.2	Fourier	4.1645	3.5378	2.4162	1.8699	1.5959	1.4824	1.4911	1.6268	1.9419
		(0.3252)	(0.3347)	(0.3445)	(0.3438)	(0.3409)	(0.3375)	(0.3342)	(0.3312)	(0.3294)
	Local linear	10.4756	−0.3270	0.1124	0.0338	0.1144	0.1826	0.2578	0.3593	0.7088
		(3.4989)	(0.1338)	(0.0795)	(0.0563)	(0.0490)	(0.0490)	(0.0593)	(0.0893)	(0.2494)

(Continues)

Table 6. *Continued.*

$w \setminus x$		−2.8	−2.6	−2	−1.4	−0.8	−0.2	0.4	1	1.6
0.2	Fourier	4.4192	3.7109	2.4693	1.8761	1.5781	1.4475	1.4389	1.5511	1.8266
		(0.3640)	(0.3717)	(0.3781)	(0.3768)	(0.3740)	(0.3711)	(0.3683)	(0.3656)	(0.3632)
	Local linear	0.7923	−0.7679	0.0808	−0.0114	0.1023	0.1867	0.2680	0.3524	0.4907
		(0.2153)	(0.2685)	(0.0669)	(0.0593)	(0.0479)	(0.0471)	(0.0573)	(0.0794)	(0.1461)
0.6	Fourier	4.5305	3.7547	2.4277	1.8086	1.4985	1.3572	1.3335	1.4207	1.6517
		(0.3802)	(0.3836)	(0.3847)	(0.3827)	(0.3808)	(0.3791)	(0.3776)	(0.3760)	(0.3742)
	Local linear	0.3040	2.0879	0.0617	−0.0807	0.0887	0.1935	0.2790	0.3465	0.4104
		(0.0823)	(0.7247)	(0.0642)	(0.0695)	(0.0514)	(0.0540)	(0.0604)	(0.0737)	(0.1024)
1	Fourier	4.3967	3.5884	2.2466	1.6391	1.3374	1.1963	1.1625	1.2252	1.4078
		(0.3770)	(0.3781)	(0.3797)	(0.3809)	(0.3821)	(0.3833)	(0.3843)	(0.3853)	(0.3858)
	Local linear	0.0142	0.0647	0.0391	−0.2165	0.0722	0.2048	0.2917	0.3405	0.3634
		(0.0507)	(0.0516)	(0.0729)	(0.0937)	(0.0615)	(0.0633)	(0.0582)	(0.0685)	(0.0951)
1.2	Fourier	4.1714	3.3752	2.0767	1.4992	1.2144	1.0802	1.0448	1.0961	1.2533
		(0.9049)	(0.8068)	(0.6417)	(0.5463)	(0.4896)	(0.4484)	(0.4172)	(0.3931)	(0.3743)
	Local linear	−0.0890	−0.0635	0.5561	−0.3910	0.0605	0.2131	0.2991	0.3372	0.3449
		(0.0708)	(0.0606)	(0.6878)	(0.1540)	(0.0724)	(0.0660)	(0.0667)	(0.0747)	(0.0868)
	Observations					1544				

Note: Standard errors obtained by bootstrap methods are given in parentheses. Data are from the children of the NLSY linked to their mothers in the main NLSY79. All variables are standardized, having means of 0 and standard deviations of 1. Mother's AFQT is used as a control variable. Income is after tax and after transfer. The error-laden measurement of family income is family income in 1998. Family income in 2000 is used as an additional error-laden measurement of family income. Year refers to the NLSY survey year; income refers to the previous year's income.

Table 7. Impact of family income on children's reading achievement.

$w \setminus x$		−2.8	−2.6	−2	−1.4	−0.8	−0.2	0.4	1	1.6
−1.2	Fourier	2.5034	2.2364	1.7091	1.4258	1.2816	1.2335	1.2698	1.4029	1.6788
		(0.3243)	(0.3066)	(0.2756)	(0.2591)	(0.2477)	(0.2380)	(0.2289)	(0.2212)	(0.2170)
	Local linear	0.1189	0.1655	0.0599	0.1752	0.2068	0.2246	0.2403	0.2886	−0.6218
		(0.0714)	(0.0765)	(0.1607)	(0.0867)	(0.0785)	(0.0881)	(0.1147)	(0.2453)	(0.8483)
−1	Fourier	2.7152	2.4166	1.8302	1.5162	1.3548	1.2970	1.3279	1.4585	1.7331
		(0.2391)	(0.2487)	(0.2520)	(0.2457)	(0.2392)	(0.2333)	(0.2281)	(0.2236)	(0.2209)
	Local linear	0.0331	0.0972	0.0371	0.1691	0.2060	0.2301	0.2533	0.3009	0.0306
		(0.0988)	(0.0827)	(0.1949)	(0.0810)	(0.0802)	(0.0815)	(0.1026)	(0.2023)	(0.5488)
−0.6	Fourier	3.0323	2.6770	1.9881	1.6221	1.4309	1.3538	1.3702	1.4866	1.7416
		(0.2406)	(0.2475)	(0.2526)	(0.2505)	(0.2472)	(0.2440)	(0.2409)	(0.2381)	(0.2361)
	Local linear	−0.3037	−0.0429	−0.0883	0.1542	0.1975	0.2263	0.2502	0.2794	0.7104
		(0.2351)	(0.1197)	(0.4426)	(0.0773)	(0.0698)	(0.0793)	(0.0930)	(0.1549)	(2.0220)
−0.2	Fourier	3.2047	2.8040	2.0382	1.6359	1.4234	1.3303	1.3305	1.4256	1.6471
		(0.2523)	(0.2575)	(0.2620)	(0.2607)	(0.2584)	(0.2561)	(0.2538)	(0.2517)	(0.2499)
	Local linear	−3.8050	−0.2060	0.2788	0.1358	0.1851	0.2172	0.2405	0.2584	0.2808
		(1.9608)	(0.1758)	(0.5248)	(0.0764)	(0.0686)	(0.0771)	(0.0838)	(0.1280)	(0.3516)

(*Continues*)

TABLE 7. *Continued.*

$w \setminus x$		−2.8	−2.6	−2	−1.4	−0.8	−0.2	0.4	1	1.6
0.2	Fourier	3.1964	2.7691	1.9663	1.5509	1.3303	1.2277	1.2134	1.2845	1.4642
		(0.2638)	(0.2680)	(0.2711)	(0.2698)	(0.2680)	(0.2664)	(0.2648)	(0.2632)	(0.2616)
	Local linear	0.9619	−0.4772	0.1234	0.1125	0.1692	0.2049	0.2279	0.2388	0.2313
		(0.4082)	(0.3072)	(0.1663)	(0.0793)	(0.0736)	(0.0759)	(0.0878)	(0.1227)	(0.2047)
0.6	Fourier	2.9272	2.5080	1.7365	1.3450	1.1373	1.0369	1.0132	1.0606	1.1943
		(0.2696)	(0.2723)	(0.2734)	(0.2723)	(0.2711)	(0.2702)	(0.2694)	(0.2686)	(0.2676)
	Local linear	0.3479	−54.7943	0.0848	0.0800	0.1483	0.1888	0.2123	0.2191	0.2055
		(0.1425)	(31.5205)	(0.1143)	(0.0875)	(0.0780)	(0.0767)	(0.0853)	(0.1145)	(0.1721)
1	Fourier	2.2369	1.8945	1.2794	0.9750	0.8150	0.7362	0.7136	0.7412	0.8277
		(0.2710)	(0.2716)	(0.2714)	(0.2709)	(0.2708)	(0.2712)	(0.2718)	(0.2723)	(0.2725)
	Local linear	0.0592	0.1187	0.0573	0.0252	0.1174	0.1663	0.1925	0.1981	0.1848
		(0.0841)	(0.0874)	(0.1167)	(0.1171)	(0.0897)	(0.0863)	(0.0975)	(0.1204)	(0.1524)
	Fourier	1.6394	1.3825	0.9270	0.7048	0.5890	0.5325	0.5170	0.5377	0.6011
1.2		(0.2732)	(0.2734)	(0.2735)	(0.2741)	(0.2753)	(0.2768)	(0.2782)	(0.2796)	(0.2806)
	Local linear	−0.0373	−0.0112	−0.1813	−0.0351	0.0923	0.1500	0.1798	0.1864	0.1750
		(0.1184)	(0.0987)	(1.1729)	(0.1484)	(0.0937)	(0.0967)	(0.1031)	(0.1198)	(0.1461)
	Observations					1274				

Note: Standard errors obtained by bootstrap methods are given in parentheses. Data are from the children of the NLSY linked to their mothers in the main NLSY79. All variables are standardized, having means of 0 and standard deviations of 1. Mother's AFQT is used as a control variable. Income is after tax and after transfer. The error-laden measurement of family income is family income in 1998. Family income in 2000 is used as an additional error-laden measurement of family income. The reading score is obtained by taking a simple average of the reading recognition and reading comprehension scores. Year refers to the NLSY survey year; income refers to the previous year's income.

income is to increase a child's reading score by about 1.9% of a standard deviation.[21] Interestingly, estimated effects from the local linear estimator appear to be statistically insignificant over all ranges of (x, w), with few exceptions.[22]

Figure 2 shows a graph of the covariate-conditioned average marginal effect (top) and average counterfactual response (bottom) of family income on children's math scores at various values of standardized family income ranging from -2.8 to 1.6 and standardized mother's AFQT ranging from -1.2 to 1.2. These are obtained using our estimator with the bandwidths in Table 5. All estimates of the average marginal effect are positive over the ranges of both family income and AFQT score. In general, the impact of family income at a given AFQT value increases as standardized family income moves from 0 to -2.8 or 1.6, making a broad U-shape. As a result, we find slightly increasing returns to family income for children in higher-income families. However, diminishing returns to family income are also observed at standardized income levels below $x = 0$. We also note that the shape of the income effect varies over different levels of mother's AFQT score. For instance, at $w = 0.6$, the average marginal effect is quite variable, while that at -1.2 is flat. Thus, the average marginal effect depends on the level of mother's AFQT. Imposing separability would mask this feature of the relationship, illustrating the usefulness of the nonseparability allowed here.

Figure 3 shows a graph based on the local linear estimator of the apparent causal effect (top) and average counterfactual response (bottom) of family income on children's math scores.[23] This shows much smaller marginal effects than those from our estimator, and we see apparent negative marginal effects for poor families with high mother's AFQT. Indeed, the results from the local linear estimator indicate increasing returns to income, that is, income effects that are larger at higher levels of family income, at odds with economic intuition.

Figure 4 shows the covariate-conditioned average marginal effect (top) and average counterfactual response (bottom) of family income on children's reading scores at various values of standardized family income ranging from -2.8 to 1.6 and standardized AFQT ranging from -1.2 to 1.2. These are obtained using our estimator with the bandwidths of Table 5. The effects are positive over the full range of both family income and AFQT score. Children in poorer families are likely to have a higher effect of family income at a given AFQT value. Thus, the results show diminishing returns to income over all ranges of family income at a given value of AFQT. We also observe the dependence of the average marginal effect on mother's AFQT.

Figure 5 depicts the apparent causal effect (top) and average counterfactual response (bottom) of family income on children's reading scores obtained using the local linear estimator. The results indicate much smaller income effects than those from our estimator. We also see counterintuitive increasing returns to family income.

---

[21] With $\sigma_x = 1.62506$, $3.0323 \times \ln(1.01)/\sigma_x = 0.018567$.

[22] Again, marginal effects on children's reading achievement from the local linear estimator near boundary values of family income can be extremely irregular.

[23] Because of the irregular behavior of the local linear estimator, we truncate the results to remove the extreme values of family income and plot only the results for standardized family income ranging from -1.4 to 1.

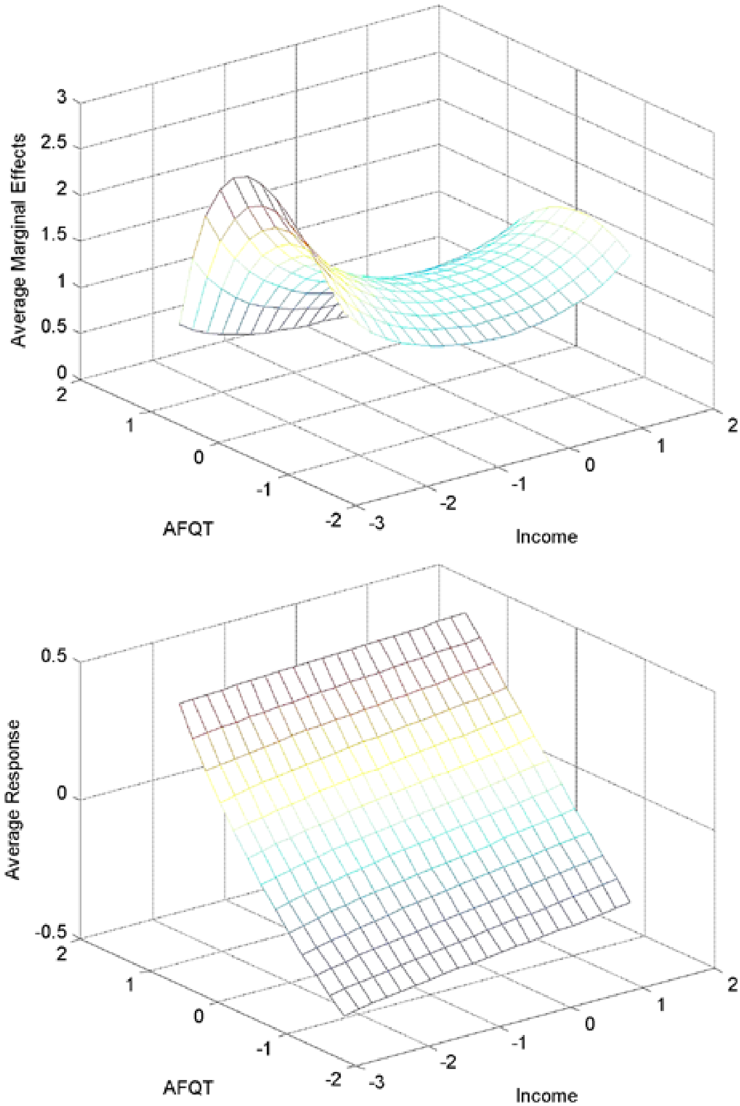

FIGURE 2. Impact of family income on children's math scores (Fourier). Our estimator is used for the covariate-conditioned average marginal effect (upper panel) and the average counterfactual response (lower panel). The error-laden measurement of family income is family income in 1998. Family income in 2000 is used as an additional error-laden measurement of family income.

Taken as a whole, these results suggest that measurement error in family income matters considerably and that using an estimator that properly accommodates the presence of measurement error reveals important effects that are obscured by estimators that ignore measurement error, such as the local linear estimator. For math scores, we find that the effects of family income are positive and that the magnitudes of the income effects are substantially larger than those obtained from the local linear estimator. Although the local linear estimates appear statistically significant for nonpoor families,

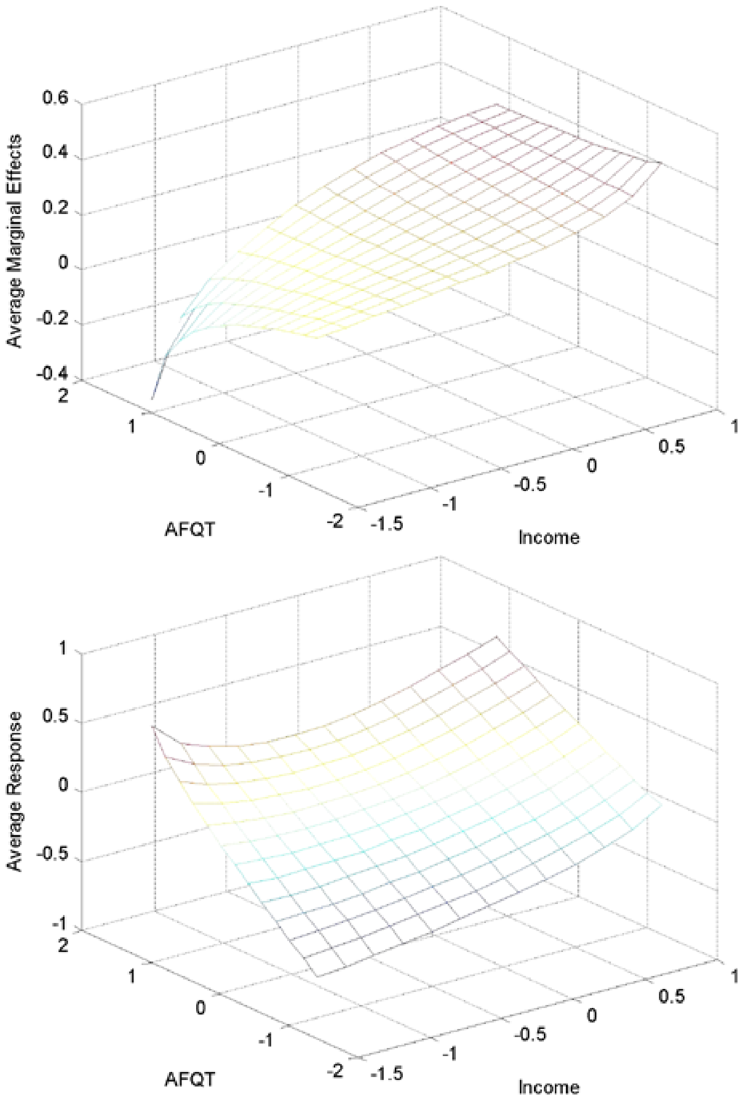

Figure 3. Impact of family income on children's math scores (local linear). The local linear estimator is used for the covariate-conditioned average marginal effect (upper panel) and the average counterfactual response (lower panel). The error-laden measurement of family income is family income in 1998.

they are rather modest, as seen in previous studies. For reading scores, the Fourier estimated effects are positive, large, and significant for poorer families, whereas those from the local linear estimator are tiny and statistically insignificant over most ranges of family income and mother's AFQT score. It follows that, contrary to previous thinking, income transfers could have a significant impact on the development of children growing up in poor families.

Our results also demonstrate that the relation between family income and child's achievement is nuanced. We find nonlinearity in income effects over family income.

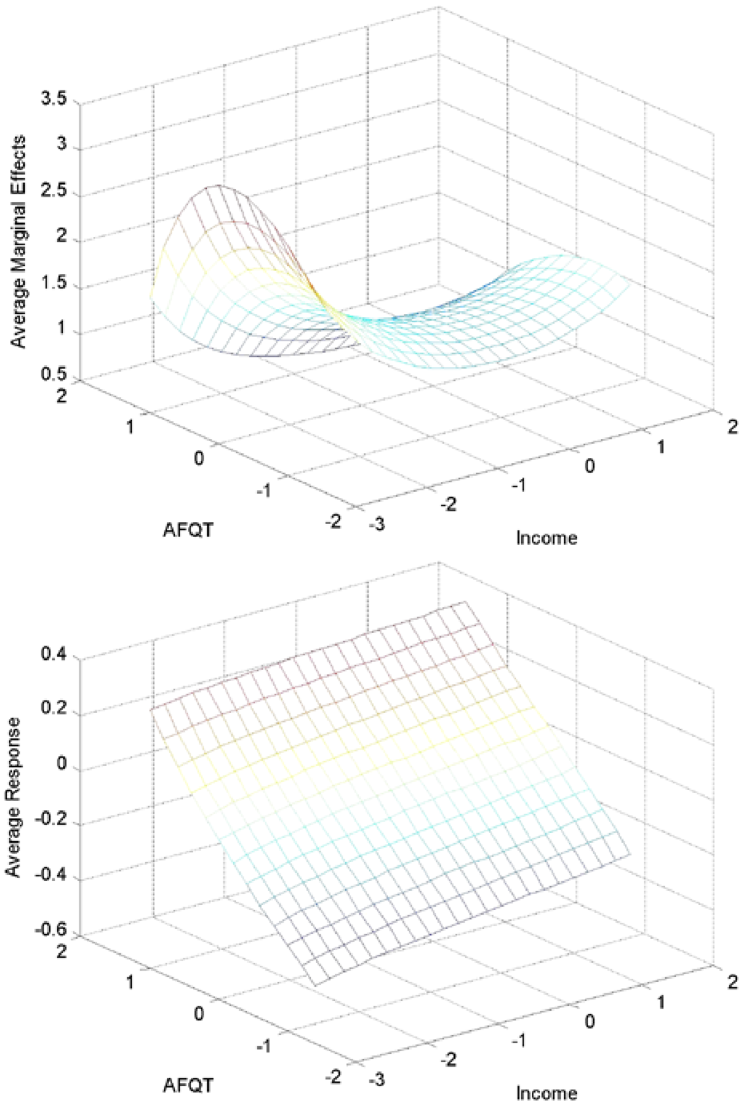

FIGURE 4. Impact of family income on children's reading scores (Fourier). Our estimator is used for the covariate-conditioned average marginal effect (upper panel) and the average counterfactual response (lower panel). The error-laden measurement of current family income is family income in 1998. Family income in 2000 is used as an additional error-laden measurement of family income.

Specifically, both effects on math scores and effects on reading scores show diminishing returns to income for families with standardized income levels below $x = 0$, but show a wide U-shape overall. Moreover, we observe that the expected income effect depends on the level of mother's AFQT score, a feature that would not be apparent using a method that enforced additive separability between the drivers of child achievement, thereby ruling out interactions.

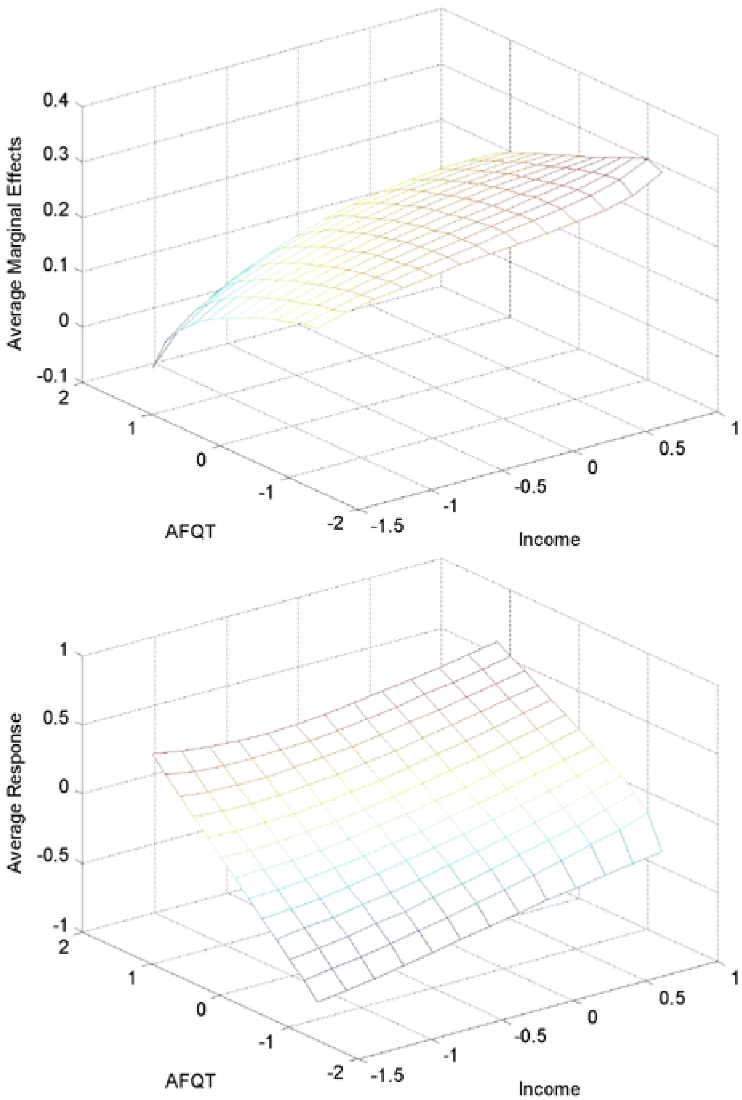

FIGURE 5. Impact of family income on children's reading scores (local linear). The local linear estimator is used for the covariate-conditioned average marginal effect (upper panel) and the average counterfactual response (lower panel). The error-laden measurement of current family income is family income in 1998.

7. SUMMARY AND CONCLUDING REMARKS

We examine the identification and estimation of covariate-conditioned average marginal effects in a nonseparable data generating process with an endogenous and mismeasured cause of interest. This is the first study to simultaneously address these issues. We use control variables to ensure the conditional independence between the cause of interest and other unobservable drivers, permitting identification of the causal effects of interest. Although the endogenous cause of interest is unobserved, two error-laden measurements are available. We extend methods of the deconvolution literature for nonlinear measurement errors to obtain estimates of the distribution functions of the underlying cause of interest from its error-laden measurements and to recover parameters of interest. These parameters include covariate-conditioned average marginal effects and weighted averages of them. We obtain uniform convergence rates and asymptotic normality for estimators of covariate-conditioned average marginal effects, faster convergence rates for estimators of their weighted averages over control variables, and $\sqrt{n}$ consistency and asymptotic normality for estimators of their weighted averages over control variables and causes. We investigate the finite-sample behavior of our estimators using Monte Carlo simulations, and we apply our new methods to study the impact of family income on child achievement. There we find interesting new results, suggesting that these effects are considerably larger than previously recognized.

Mathematical Appendix

Additional material and mathematical proofs of the results are presented in the Appendixes available in supplementary files on the journal website, http://qeconomics.org/supp/275/supplement.pdf and http://qeconomics.org/supp/275/code_and_data.zip.

References

Altonji, J. G. and R. L. Matzkin (2005), "Cross section and panel data estimators for nonseparable models with endogenous regressors." *Econometrica*, 73, 1053–1102. [750, 751, 753, 754, 759]

Amemiya, T. (1985), "Instrumental variable estimator for the nonlinear error-in-variables model." *Journal of Econometrics*, 28, 273–289. [750]

Aughinbaugh, A. and M. Gittleman (2003), "Does money matter? A comparison of the effect of income on child development in the United States and Great Britain." *Journal of Human Resources*, 38, 416–440. [777]

Blau, D. M. (1999), "The effect of income on child development." *Review of Economics and Statistics*, 81, 261–276. [777]

Blundell, R. W. and J. L. Powell (2004), "Endogeneity in semiparametric binary response models." *Review of Economic Studies*, 71, 655–679. [754, 759]

Carroll, R. J., D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu (2006), *Measurement Error in Nonlinear Models*. Chapman & Hall/CRC, Boca Raton, FL. [761]

Chalak, K. and H. White (2011a), "An extended class of instrumental variables for the estimation of causal effects." *Canadian Journal of Economics*, 44, 1–51. [750, 752, 753]

Chalak, K. and H. White (2011b), "Causality, conditional independence, and graphical separation in settable systems." Department of Economics Discussion paper, UC San Diego. [750, 752, 753]

Chen, X., H. Hong, and E. Tamer (2005), "Measurement error modes with auxiliary data." *Review of Economic Studies*, 72, 343–366. [750]

Chen, X., H. Hong, and A. Tarozzi (2008), "Semiparametric efficiency in GMM models with auxiliary data." *The Annals of Statistics*, 36, 808–843. [750]

Chernozhukov, V. and C. Hansen (2005), "An IV model of quantile treatment effects." *Econometrica*, 73, 245–261. [750]

Chernozhukov, V., G. W. Imbens, and W. K. Newey (2007), "Instrumental variable estimation of nonseparable models." *Journal of Econometrics*, 139, 4–14. [750]

Chesher, A. (2005), "Nonparametric identification under discrete variation." *Econometrica*, 73, 1525–1550. [750]

Cunha, F. and J. Heckman (2007), "The technology of skill formation." *American Economic Review*, 97, 31–47. [779]

Cunha, F., J. Heckman, and S. M. Schennach (2010), "Estimating the technology of cognitive and noncognitive skill formation." *Econometrica*, 78, 883–931. [759, 779]

Dahl, G. and L. Lochner (2012), "The impact of family income on child achievement: Evidence from changes in the earned income tax credit." *American Economic Review*, 102, 1927–1956. [777, 779]

Dawid, A. P. (1979), "Conditional independence in statistical theory." *Journal of the Royal Statistical Society, Series B*, 41, 1–31. [753, 767]

Duncan, G. J., W. J. Yeung, J. Brooks-Gunn, and J. R. Smith (1998), "How much does childhood poverty affect the life chances of children?" *American Sociological Review*, 63, 406–423. [777]

Fan, J. (1991a), "On the optimal rates of convergence for nonparametric deconvolution problems." *The Annals of Statistics*, 19, 1257–1272. [756, 761, 762, 763]

Fan, J. (1991b), "Asymptotic normality for deconvolution kernel density estimators." *Sankhyā: The Indian Journal of Statistics, Series A*, 53, 97–110. [765]

Fan, J. and Y. K. Truong (1993), "Nonparametric regression with errors in variables." *The Annals of Statistics*, 21, 1900–1925. [756, 761, 762, 763]

Florens, J. P., 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*, 76, 1191–1206. [750, 752, 759]

Gonçalves, S. and H. White (2005), "Bootstrap standard error estimates for linear regression." *Journal of the American Statistical Association*, 100, 970–979. [781]

Hahn, J. and G. Ridder (2011), "Conditional moment restrictions and triangular simultaneous equations." *Review of Economics and Statistics*, 93, 683–689. [750]

Hausman, J. A., W. K. Newey, H. Ichimura, and J. L. Powell (1991), "Identification and estimation of polynomial errors-in-variables models." *Journal of Econometrics*, 50, 273–295. [750, 756]

Hausman, J. A., W. K. Newey, and J. L. Powell (1995), "Nonlinear errors in variables estimation of some Engel curves." *Journal of Econometrics*, 65, 205–233. [750]

Heckman, J. J. and E. Vytlacil (2005), "Structural equations, treatment effects, and econometric policy evaluation." *Econometrica*, 73, 669–738. [759]

Hoderlein, S. (2011), "How many consumers are rational?" *Journal of Econometrics*, 164, 294–309. [750, 753]

Hoderlein, S. and E. Mammen (2007), "Identification of marginal effects in nonseparable models without monotonicity." *Econometrica*, 75, 1513–1518. [750, 753]

Hsiao, C. (1989), "Consistent estimation for some nonlinear errors-in-variables models." *Journal of Econometrics*, 41, 159–185. [750]

Hsiao, C. and K. Wang (2000), "Estimation of structural nonlinear errors-in-variables models by simulated least-squares method." *International Economic Review*, 41, 523–542. [750]

Hu, Y. (2008), "Identification and estimation of nonlinear models with misclassification error using instrumental variables: A general solution." *Journal of Econometrics*, 144, 27–61. [750]

Hu, Y. and G. Ridder (2012), "Estimation of nonlinear models with mismeasured regressors using marginal information." *Journal of Applied Econometrics*, 27, 347–385. [750]

Hu, Y. and S. M. Schennach (2008), "Instrumental variable treatment of nonclassical measurement error models." *Econometrica*, 76, 195–216. [750]

Hurwicz, L. (1950), "Generalization of the concept of identification." In *Statistical Inference in Dynamic Economic Models*, Cowles Commission Monograph, Vol. 10, Wiley, New York. [754]

Ichimura, H. (1993), "Semiparametric least squares (SLS) and weighted SLS estimation of single-index models." *Journal of Econometrics*, 58, 71–120. [759]

Imbens, G. W. (2007), "Nonadditive models with endogenous regressors." In *Advances in Economics and Econometrics, Ninth World Congress of the Econometric Society*, Vol. 3 (R. Blundell, W. Newey, and T. Persson, eds.), 17–46, Cambridge University Press, Cambridge. [750]

Imbens, G. W. and W. K. Newey (2009), "Identification and estimation of triangular simultaneous equations models without additivity." *Econometrica*, 77, 1481–1512. [750, 753]

Lewbel, A. (2007), "Estimation of average treatment effects with misclassification." *Econometrica*, 75, 537–551. [750]

Li, T. (2002), "Robust and consistent estimation of nonlinear errors-in-variables models." *Journal of Econometrics*, 110, 1–26. [750, 756, 761, 762, 763]

Li, T. and Q. Vuong (1998), "Nonparametric estimation of the measurement error model using multiple indicators." *Journal of Multivariate Analysis*, 65, 139–165. [756, 761, 762, 763]

Lighthill, M. J. (1962), *Introduction to Fourier Analysis and Generalized Function*. Cambridge University Press, Cambridge. [765]

Mahajan, A. (2006), "Identification and estimation of single index models with misclassified regressor." *Econometrica*, 74, 631–665. [750]

Manski, C. (1997), "Monotone treatment response." *Econometrica*, 65, 1311–1334. [752, 753]

Politis, D. N. and J. P. Romano (1999), "Multivariate density estimation with general flat-top kernels of infinite order." *Journal of Multivariate Analysis*, 68, 1–25. [757, 764, 767, 768]

Prakasa Rao, B. L. S. (1992), *Identifiability in Stochastic Models*. Academic Press, Boston, MA. [756]

Schennach, S. M. (2004a), "Nonparametric estimation in the presence of measurement error." *Econometric Theory*, 20, 1046–1093. [750, 751, 756, 759, 761, 762, 763, 764, 765, 768]

Schennach, S. M. (2004b), "Estimation of nonlinear models with measurement error." *Econometrica*, 72, 33–75. [750, 751, 756, 759, 762, 764]

Schennach, S. M. (2007), "Instrumental variable estimation of nonlinear errors-in-variables models." *Econometrica*, 75, 201–239. [750]

Schennach, S. M., H. White, and K. Chalak (2012), "Local indirect least squares and average marginal effects in nonseparable structural systems." *Journal of Econometrics*, 166, 282–302. [751]

Strotz, R. H. and H. O. A. Wold (1960), "Recursive vs. nonrecursive systems: An attempt at synthesis." *Econometrica*, 28, 417–427. [752]

Taylor, B. A., E. Dearing, and K. McCartney (2004), "Incomes and outcomes in early childhood." *Journal of Human Resources*, 39, 980–1007. [777]

Wahba, G. (1990), *Spline Models for Observational Data*. SIAM, Philadelphia, PA. [780]

White, H. and K. Chalak (2009), "Settable systems: An extension of Pearl's causal model with optimization, equilibrium, and learning." *Journal of Machine Learning Research*, 10, 1759–1799. [752, 753]

White, H. and K. Chalak (2013), "Identification and identification failure for treatment effects using structural systems." *Econometric Reviews*, 32, 273–317. [753]

Submitted May, 2012. Final version accepted March, 2015.