Identification of Average Effects under Magnitude and Sign Restrictions on Confounding

Karim Chalak[*][†]
University of Virginia

April 3, 2019

Abstract

This paper studies measuring various average effects of X on Y in general structural systems with unobserved confounders U, a potential instrument Z, and a proxy W for U. We do not require X or Z to be exogenous given the covariates or W to be a perfect one-to-one mapping of U. We study the identification of coefficients in linear structures as well as covariate-conditioned average nonparametric discrete and marginal effects (e.g. average treatment effect on the treated), and local and marginal treatment effects. First, we characterize the bias, due to the omitted variables U, of (nonparametric) regression and instrumental variables estimands, thereby generalizing the classic linear regression omitted variable bias formula. We then study the identification of the average effects of X on Y when U may statistically depend on X and Z. These average effects are point identified if the average direct effect of U on Y is zero, in which case exogeneity holds, or if W is a perfect proxy, in which case the ratio (contrast) of the average direct effect of U on Y to the average effect of U on W is also identified. More generally, restricting how the average direct effect of U on Y compares in magnitude and/or sign to the average effect of U on W can partially identify the average effects of X on Y. These restrictions on confounding are weaker than requiring benchmark assumptions, such as exogeneity or a perfect proxy, and enable a sensitivity analysis. After discussing estimation and inference, we apply this framework to study earnings equations.

Keywords: *causality, confounding, endogeneity, omitted variable bias, partial identification, proxy, sensitivity analysis.* **JEL Codes:** C31, C35, C36.

[*] Department of Economics, UVA, P.O. Box 400182, Charlottesville, VA 22904-4182 (chalak@virginia.edu).

[†] Acknowledgments: I thank the participants in the Northwestern Junior Festival on New Developments in Microeconometrics, Harvard Causal Inference Seminar, 2012 CA Econometrics Conference, 2013 BU-BC Green Line Econometrics Conference, 2013 North American Winter Meeting of the Econometric Society, NY Camp Econometrics VIII, 23rd meeting of the Midwest Econometrics Group, 9th Greater NY Metropolitan Area Econometrics Colloquium, Cowles Foundation Conference on Econometrics, and the seminars at BC, Cleveland Fed, UCSD, UCLA, USC, Pitt, IUPUI, UW-Milwaukee, Oxford, Royal Holloway, University of Leicester, UVA, University of Montreal, Georgetown, Virginia Tech, Vanderbilt, and UNC as well as Kate Antonovics, Andrew Beauchamp, Stéphane Bonhomme, Federico Ciliberto, Donald Cox, Julie Cullen, Stefan Hoderlein, Arthur Lewbel, Matthew Masten, Elie Tamer, and especially John Pepper for helpful comments. I acknowledge support from the Boston College Research Incentive and Expense Grants. I thank Rossella Calvi, Daniel Kim, and Tao Yang for excellent research assistance. Any errors are the author's responsibility.

1 Introduction

When measuring causal effects in observational studies, researchers often consider the unobserved variables that may jointly drive the cause and response of interest. For example, when estimating the financial return to education, researchers consider the unobserved individual "ability" that may jointly affect educational attainment and wage. Similarly, when estimating the elasticity of output with respect to the labor input, researchers consider the unobserved firm productivity that may jointly affect the input demands (e.g. capital and labor) and the output. A standard assumption that is useful to point identify average effects is the exogeneity (unconfoundedness) of the treatment or the instrument given the covariates. For example, to estimate the return to education researchers sometime assume that educational attainment, or an instrumental variable that is related to educational attainment such as the distance to a college, does not depend on ability given the covariates (see e.g. Card, 1995). Similarly, to estimate production functions, researchers may consider using the prices of inputs as instrumental variables that are related to the inputs and unrelated to productivity (see e.g. the discussion in Griliches and Mairesse, 1998). A second common assumption that is useful to point identify average effects requires that there is a perfect one-to-one proxy for the unobserved confounders. For example, a researcher may rely on a test score as a measure of ability (see e.g. the discussion in Neal and Johnson, 1996). Similarly, a researcher may assume that, conditional on the capital input, investment or an intermediate input is a perfect proxy for the firm's productivity (see e.g. Olley and Pakes (1996) and Levinsohn and Petrin (2003)).

These standard assumptions are not directly testable and researchers often ponder their validity. In particular, researchers sometimes question whether selection on unobservables leads conditional exogeneity to fail. For example, Carneiro and Heckman (2002) provide evidence suggesting that several commonly employed instruments in the ability literature may be endogenous and Griliches and Mairesse (1998) discuss how input prices may fail to be valid instruments when estimating production functions. Also, researchers often question whether a proxy is a perfect coding of the unobserved confounders. For example, a test score may be an error-laden measure of ability (see e.g. Bollinger, 2003). Similarly, the firm's investment or intermediate input may fail to be a strictly monotonic function of its productivity if there is an optimization or measurement error in how this proxy variable is determined or coded[1].

Given the important role that the assumptions of conditional exogeneity and perfect proxy play in estimating average effects in a variety of empirical settings, it is useful to study the consequences of a possible departure from these benchmark assumptions. In order to do so, this paper characterizes the bias that standard estimands of various average effects would incur in the presence of omitted variables (unobserved confounders). It then demonstrates how restrictions on confounding that are weaker than requiring conditional exogeneity or a perfect proxy can partially identify these average effects. This enables a sensitivity analysis through which a researcher may gain confidence in a causal effect estimate that is not highly sensitive to deviations from a maintained assumption. In particular, the paper studies identifying and estimating various conditional average effects of the treatment X on the response Y in structural systems with unobserved confounders U, a potential (possibly invalid) instrument Z, and a proxy W for U. We do not require X or Z to be conditionally exogenous, and thus U may statistically depend on X and Z. Further, we do not require W to be a perfect proxy, i.e. a one-to-one mapping of U. The framework encompasses general specifications; we study the identification of coefficients in a linear structure as well as of covariate-conditioned average nonparametric discrete and marginal effects (e.g. average treatment effect on the treated), local average treatment effect, and marginal treatment effect.

The analysis proceeds in two steps. The first step studies the consequences of omitted variables on the identification of average effects via standard estimands in the general specifications that this paper considers. In the case of linear homogenous effects, the linear regression omitted variable bias (OVB) representation is a classic result in econometrics (see e.g. Stock and Watson, 2010, ch. 6; Wooldridge, 2012, ch. 3). For instance, Angrist and Pischke (2009, p. 62) state that the linear regression OVB formula "is one of the most important things to know about regression." What is the analogue of the OVB formula in the cases of standard estimands, such as nonparametric regression and instrumental variables (IV) estimands (e.g. Wald (1940) or local IV estimands), for the various average effects described above? The first contribution of this paper is to characterize the OVB formula in these cases thereby generalizing the classic linear regression OVB representation. This enables studying the direction of the

[1] Also, Ackerberg, Caves, and Frazer (2015) discuss how the perfect proxy assumption can sometimes render the input variables functionally dependent, complicating the identification of their effects on the output.

OVB, including in nonparametric nonseparable structures with heterogenous effects.

The second step of the analysis demonstrates how an imperfect proxy W for U can be used to either point or partially identify various average effects of X on Y. In particular, these average effects are point identified in two special cases. The first occurs if X or Z is exogenous. It suffices for exogeneity that U is unassociated (in a precise statistical sense) with X or Z. This condition is testable under our assumptions since it implies that W is unassociated with X or Z. Alternatively, when U may statistically depend on X and Z as we allow, exogeneity holds if one assumes that the average direct (i.e. holding X fixed) effect of U on Y is zero (e.g. the average effect of ability on wage is zero). The second special case in which these average effects of X on Y are point identified occurs if the proxy W is a perfect one-to-one mapping of U (e.g. a test score is a perfect proxy for ability). In this case, the ratio (contrast) of the average direct effect of U on Y to the average effect of U on W is also point identified. More generally, a researcher may impose restrictions on how the average direct effect of U on Y compares in magnitude and/or sign to the average effect of U on W that are weaker than the restrictions obtained when assuming exogeneity or a perfect proxy. Moreover, this comparison may be informed by economic theory and evidence, as illustrated in the paper's empirical application. The second contribution of this paper is to demonstrate how these magnitude and/or sign restrictions on confounding can point or partially identify the various average effects of X on Y. This enables a researcher to analyze the sensitivity of the causal effects estimates to deviations from the benchmark assumptions of exogeneity and perfect proxy and can help clarify the extent to which the empirical estimates hinge on these identifying assumptions.

The paper is organized as follows. Section 2 describes the paper's basic framework. Section 3 states the data generation assumption. We derive the OVB formulas and characterize the sharp identification regions under restrictions on confounding for constant coefficients in a linear structure in Section 4, nonparametric average discrete and marginal effects in Section 5, and local and marginal treatment effects in Section 6. Section 7 discusses estimation and inference. Section 8 applies this paper's framework to study the return to education and the black-white wage gap. Section 9 concludes. Online Appendix[2] A contains extensions. Mathematical proofs

[2] The Online Appendix is available at http://people.virginia.edu/˜kmc2yf/SignMagApp.pdf. Section A.1 extends the constant coefficients analysis to discuss exogenous random coefficients and conditioning on covariates, additional empirical estimates, a panel structure, and proxies included in the Y equation. Section A.2 studies the special case where U enters the Y and W equations additively separably. Section A.3 studies the nonparametric nonseparable case with discrete U. Appendix B collects the mathematical proofs.

are gathered in Online Appendix B.

2 Basic Framework and Overview

2.1 Linear Equations with Homogenous Effects

To illustrate the paper's main ideas, consider an earnings structural equation (see e.g. Mincer, 1974; Card, 1999), frequently employed in empirical work, given by

$$Y = X'\bar{\beta} + U\bar{\delta}_Y + U_Y'\bar{\alpha}_Y. \tag{1}$$

The researcher observes realizations of the logarithm of hourly wage, Y, and of determinants X of wage, such as years of education (and the level and square of years of experience). U, commonly referred to as "ability" in the literature, denotes unobserved skill, and U_Y collects additional unobservables (disturbances). To introduce the main ideas in their simplest form, we let U be a scalar and consider homogenous (constant) linear effects $\bar{\beta}$, $\bar{\delta}_Y$, and $\bar{\alpha}_Y$. Also, we leave any additional covariates implicit. However, as discussed below, we emphasize that this paper's approach does not require homogenous effects or a parametric or separable specification.

Our object of interest is the (average) effect $\bar{\beta}$ of X on Y, e.g. the (average) financial return to education. Although the return to education is homogenous in this example, and thus does not depend on U, ability U is freely associated with X and may cause Y (e.g. the educational attainment and wage may depend on ability). Thus, U is an unobserved "confounder" or "omitted variable" and X is potentially "endogenous." The researcher observes realizations of a vector Z of potential instrumental variables. By definition, U_Y collects the unobservables that may drive Y and are assumed to be uncorrelated with Z (given the covariates) whereas U (which may be a vector more generally in Section 4) is freely correlated with Z. This allows a potential instrument for education, e.g. the proximity to a college, to be invalid if it is correlated with ability U, e.g. due to unobserved parental characteristics or choices. We let Z and X have the same dimension and $Cov(Z,X)$ be nonsingular. In particular, the most basic case arises when Z equals X. The linear IV regression OVB (or inconsistency) B in recovering $\bar{\beta}$ is given by

$$B \equiv Cov(Z,X)^{-1}Cov(Z,Y) - \bar{\beta} = Cov(Z,X)^{-1}Cov(Z,U)\bar{\delta}_Y,$$

and this expression may help clarify the direction of the OVB. Exogeneity requires $Cov(U\bar{\delta}_Y + U_Y'\bar{\alpha}_Y, Z) = 0$ in which case $B = 0$. By definition of U and U_Y, $Cov(U_Y, Z) = 0$ and we allow

$Cov(U, Z) \neq 0$. Thus, exogeneity is guaranteed to hold only if $\bar{\delta}_Y = 0$, i.e. the average direct (holding X fixed) effect of U on Y is zero. The expression for B is the IV analogue of the classic regression OVB formula and reduces to it in the special case when $Z = X$.

The researcher may observe realizations of an error-laden proxy W for U given by

$$W = U\bar{\delta}_W + U_W'\bar{\alpha}_W, \tag{2}$$

where the unobservables U_W may be correlated with U_Y, U, and X. For now, we consider a linear equation for W with constant $\bar{\delta}_W$ and $\bar{\alpha}_W$. For example, W may denote the logarithm of a test score commonly used as a proxy for ability, such as IQ (Intelligence Quotient) or KWW (Knowledge of the World of Work). This parsimonious specification facilitates comparing the coefficients on U in the Y and W equations while maintaining the commonly used log-level specification for the wage equation. In particular, $\bar{\delta}_Y$ and $\bar{\delta}_W$ are the semi-elasticities of the wage and test score with respect to ability[3], i.e. $100\bar{\delta}_Y\%$ and $100\bar{\delta}_W\%$ are the (average) approximate percentage changes in the wage (with X fixed) and test score due to a unit increase in U.

Sometimes, researchers consider conditioning on the proxy to control for endogeneity. Provided $\bar{\delta}_W \neq 0$, substituting for U in equation (1) gives

$$Y = X'\bar{\beta} + W\frac{\bar{\delta}_Y}{\bar{\delta}_W} + U_Y'\bar{\alpha}_Y - U_W'\bar{\alpha}_W\frac{\bar{\delta}_Y}{\bar{\delta}_W}. \tag{3}$$

If U_W is degenerate then W is a perfect one-to-one proxy for U and, provided $Cov[U_Y, (Z', U)'] = 0$, a linear IV regression of Y on $(1, X', W')'$ using instruments $(1, Z', W')'$ may point identify the (average) effect $\bar{\beta}$ of X on Y as well as $\frac{\bar{\delta}_Y}{\bar{\delta}_W}$, the ratio of the (average) direct effect of U on Y to the (average) effect of U on W. This result fails to hold when U_W is nondegenerate because a nonzero[4] $Cov(U_W, Z|W)$ can lead to an IV regression bias[5] in recovering $\bar{\beta}$. Moreover, $\bar{\beta}$ is "under-identified" in equation (3) since Z and X have the same dimension (recall Z may equal X) and there are fewer exogenous instruments for $(X', W')'$ than is needed for an IV regression to point identify $(\bar{\beta}', \frac{\bar{\delta}_Y}{\bar{\delta}_W})'$.

[3] One could also consider standardizing the variables in equations (1) and (2), in which case the slope coefficients on the standardized ability denote standard deviation shifts in wage (holding X fixed) and the test score respectively due to a standard deviation shift in ability.

[4] From $W = U\bar{\delta}_W + U_W'\bar{\alpha}_W$, $Cov(U_W, U|W)$ is generally nonzero. Since Z (or X) and U are freely correlated, it follows that U_W is generally correlated with Z (or X) given W.

[5] In general models, conditioning on W may, but need not, attenuate the regression bias (see e.g. Wickens, 1972; Battistin and Chesher, 2014; and Ogburna and VanderWeele, 2012).

Instead of assuming that W is a perfect proxy with U_W degenerate and $Cov[U_Y, (Z', U)'] = 0$, we consider the weaker restriction $Cov[(U'_Y, U'_W)', Z] = 0$. Then the IV OVB is given by

$$B = Cov(Z, X)^{-1} Cov(Z, U)\bar{\delta}_Y = Cov(Z, X)^{-1} Cov(Z, W)\frac{\bar{\delta}_Y}{\bar{\delta}_W}.$$

Note that the condition $Cov(U, Z) = 0$, which ensures exogeneity ($B = 0$), is testable under these assumptions since it implies $Cov(W, Z) = 0$. Importantly, if $Cov(U, Z) \neq 0$ then the IV OVB is known up to the ratio $\frac{\bar{\delta}_Y}{\bar{\delta}_W}$ of the (average) direct effect of U on Y to the (average) effect of U on W. In particular, $\bar{\beta}$ is characterized by:

$$\bar{\beta} = Cov(Z, X)^{-1} Cov(Z, Y) - Cov(Z, X)^{-1} Cov(Z, W)\frac{\bar{\delta}_Y}{\bar{\delta}_W}.$$

This expression for $\bar{\beta}$ involves two linear IV regression estimands. It also involves the unknown ratio $\frac{\bar{\delta}_Y}{\bar{\delta}_W}$. As we show, analogous expressions obtain for various nonparametric average effects.

2.2 Magnitude and Sign Restrictions on Confounding

In the linear homogenous case above as well as the nonparametric heterogenous cases discussed below, we ask the following question: How does the average direct effect of U on Y compare in magnitude or sign (or both) to the average effect of U on W? The paper demonstrates how the answer to this question imposes restrictions on the magnitude and sign of confounding (e.g. on $\frac{\bar{\delta}_Y}{\bar{\delta}_W}$ in equations (1,2)) that can point or partially identify the average effect of X on Y (e.g. $\bar{\beta}$ in (1,2)). We do not require a particular answer to this question. Instead, we characterize the mapping[6] from every possible answer to the corresponding identification region for the average effect of X on Y. To keep the scope of the paper manageable, we focus on restricting the support of e.g. $\frac{\bar{\delta}_Y}{\bar{\delta}_W}$ rather than imposing more general prior distributions. In particular, when U may statistically depend on X and Z, the average effect of X on Y (e.g. $\bar{\beta}$) is point identified in the following three special cases. The first case is exogeneity which obtains when one assumes that the average direct effect of U on Y is zero (e.g. $\frac{\bar{\delta}_Y}{\bar{\delta}_W} = 0$). The second special case assumes that W is a perfect proxy, in which case the ratio (e.g. $\frac{\bar{\delta}_Y}{\bar{\delta}_W}$) of the average direct effect of U on Y to the average effect of U on W is also point identified. The third special case is proportional confounding which assumes that the average direct effect of U on Y is equal to a known proportion of the average effect of U on W (e.g. $\frac{\bar{\delta}_Y}{\bar{\delta}_W} = d$). The paper demonstrates

[6] Leamer (1983) suggests the slogan “the mapping is the message.”

that weaker restrictions on how the average direct effect of U on Y compares in magnitude and/or sign to the average effect of U on W (e.g. $\left|\bar{\delta}_Y\right| \leq \left|\bar{\delta}_W\right|$, $0 \leq \frac{\bar{\delta}_Y}{\bar{\delta}_W}$, $0 \leq \frac{\bar{\delta}_Y}{\bar{\delta}_W} \leq 1$, or $\frac{\bar{\delta}_Y}{\bar{\delta}_W} \in [d_L, d_H]$ where $[d_L, d_H]$ contains the perfect proxy estimate of $\frac{\bar{\delta}_Y}{\bar{\delta}_W}$ or its 95% confidence interval) can partially identify the average effect of X on Y and it characterizes the resulting sharp identification region. In this sense, restrictions on confounding may be used to weaken standard assumptions, such as exogeneity or a perfect proxy.

Sometimes, economic theory and/or evidence can provide guidance on sign and magnitude restrictions on confounding. For example, in the earnings and proxy equations (1,2), it may be reasonable to assume that, given the observables, a change in ability may cause an average direct percentage change (elasticity) in wage that is smaller in magnitude than the resulting average percentage change in the test score, i.e. $\left|\bar{\delta}_Y\right| \leq \left|\bar{\delta}_W\right|$. Moreover, these average effects may be in the same direction, i.e. $0 \leq \frac{\bar{\delta}_Y}{\bar{\delta}_W}$. These assumptions are in accord with several theoretical and empirical findings. For instance, Cawley, Heckman, and Vytlacil (2001) find that the fraction of wage variance explained by measures of cognitive ability is modest and that personality traits are correlated with earnings primarily through schooling attainment. Provided that ability measures, such as IQ or KWW, are sufficiently associated with unobserved ability U, this suggests that the average direct effect of U on Y may be modest. Second, when ability is not revealed to employers, they may statistically discriminate based on observables such as education (see e.g. Altonji and Pierret, 2001; Arcidiacono, Bayer, and Hizmo, 2010). This also suggests a modest average direct effect of U on Y. Last, recall that if one assumes $\frac{\bar{\delta}_Y}{\bar{\delta}_W} \in [d_L, d_H]$ then an estimate for $\bar{\beta}$ corresponds to each $d \in [d_L, d_H]$. Thus, in determining restrictions on confounding, a researcher may ask: what restrictions on $\frac{\bar{\delta}_Y}{\bar{\delta}_W}$ are in accord with plausible features of $\bar{\beta}$? For example, the empirical findings in this paper corroborate the assumption $\frac{\bar{\delta}_Y}{\bar{\delta}_W} \leq 1$ in the earnings equation since values of $\frac{\bar{\delta}_Y}{\bar{\delta}_W}$ that exceed 1 lead to negative estimates of the average return to education and to an estimated black-white wage gap in favor of blacks, which is unlikely and inconsistent with the general findings in the literature.

In a nutshell, magnitude and sign restrictions on confounding serve as a means for identification when stronger assumptions, such as exogeneity or a perfect proxy, may fail to hold[7] and they enable examining the sensitivity of a study's estimates to deviations from these standard

[7] Even if stronger or alternative assumptions hold, restrictions on confounding may yield tighter confidence intervals.

assumptions. Of course, a particular restriction on confounding may in turn fail to hold. For example, an imposed sign and/or magnitude restriction on $\frac{\bar{\delta}_Y}{\bar{\delta}_W}$ may be invalid or there may be additional omitted variables unaccounted for in the analysis. Thus, researchers may want to carefully consider a range of restrictions on confounding in a sensitivity analysis. Nevertheless, our goal here is to provide a framework in which restrictions on confounding can be used to weaken the benchmark assumptions of exogeneity and a perfect proxy which are employed in a vast literature, leading to more robust and credible causal estimates.

2.3 Nonparametric Nonseparable Equations

An advantage of this paper's approach is that it does not require a parametric or separable specification. Sections 5 and 6 give key nonparametric results. Section 5 focuses on the case in which there are no excluded instruments ($Z = X$) and then generalizes the linear specification of Section 4 to let the outcome Y and proxy W be generated by:

$$Y = r(X, S, U, U_Y) \qquad \text{and} \qquad W = q(S, U, U_W). \tag{4}$$

Here, the vector of observed covariates S, the scalar[8] confounder U, and the vector of unobservables U_Y interact nonseparably with X to drive Y according to the unknown nonparametric structural function r. For example, this generalizes the linear specification in the "correlated random coefficient" model[9]. Similarly, U and the unobserved vector U_W interact nonseparably with S to drive W according to the unknown nonparametric function q. Suppose that[10] $U_Y \perp (U, X)|S$ so that the component U of the "exogenous treatment" (U, X) is unobserved and, unlike U_Y, may statistically depend on X given the covariates S. In particular, if $U|S$ is degenerate then there are no omitted variables and we obtain the standard assumption of conditional exogeneity (or unconfoundedness) $U_Y \perp X|S$. First, we characterize the OVB of nonparametric regression methods in recovering the covariate-conditioned average discrete or marginal effects of X on Y, thereby generalizing the classic linear regression OVB formula to the nonparametric nonseparable case. Second, we show that if W is a perfect proxy for U (with U_W degenerate and q strictly monotonic in U given S) then the conditional average effect of X

[8] Online Appendix A.2 considers a vector U that enters r additively separably.

[9] The correlated random coefficient model restricts r such that $Y = r(X, U, U_Y) = \beta(U, U_Y)X + \alpha_Y(U, U_Y)$, with a random intercept $\alpha_Y(U, U_Y)$ and slope $\beta(U, U_Y)$. For example, in this special case ability U affects the wage Y through both $\alpha_Y(\cdot)$ and the linear return $\beta(\cdot)$ to education X (see e.g. Card, 2001).

[10] $A \perp B|S = s$ denotes conditional independence given $S = s$ as in Dawid (1979). $\not\perp$ denotes dependence.

on Y is point identified, as is the ratio of the conditional average effect of U on Y to that of U on W. More generally, if the proxy W is imperfect and $U_W \perp (U, X)|S$ then restrictions on how the magnitude or sign (or both) of the conditional average effect of U on Y contrasts with that of U on W can partially identify the conditional average effect of X on Y.

Section 6 focuses on the case where X is binary and then generalizes the specification in Section 5 by allowing Z to differ from X. Specifically, it augments equations (4) with a treatment selection equation where U_X is an unobserved variable and the function ν is unknown:

$$X = \mathbf{1}\{U_X \leq \nu(Z, S)\}. \tag{5}$$

Suppose $(U_X, U_Y) \perp (Z, U)|S$. If $U|S$ is degenerate then there are no omitted variables and we obtain the standard assumptions of "monotonicity" and exogeneity, $(U_X, U_Y) \perp Z|S$, of Z. Otherwise, U may statistically depend on the potential instrument Z given S. First, we characterize the OVB of the Wald and local IV estimands for the conditional local and marginal treatment effects (LATE and MTE). Then we use restrictions on confounding, that are weaker than exogeneity or a perfect proxy, to partially identify the conditional LATE and MTE.

3 Data Generation

The next assumption defines the data generating process. To ease the exposition, we leave the covariates S implicit hereafter except when necessary - the identification analysis can be readily generalized to be made conditional on covariates.

Assumption 1 (S.1) *(i) Let* $M \equiv (\underset{\ell\times 1}{Z}', \underset{k\times 1}{X}', \underset{m\times 1}{W}', \underset{1\times 1}{Y})'$ *be a random vector whose realizations are observed (by the researcher). (ii) Let a structural system generate the unobserved vectors* U_W *and* U_Y *of countable dimension and the unobserved confounders* $\underset{l\times 1}{U}$ *collected in* $L \equiv (U_W', U_Y', U')'$*, potential instruments* Z*, causes* X*, proxies* W*, and response* Y *such that*

$$Y = r(X, U, U_Y) \quad \text{and} \quad W = q(U, U_W),$$

where r *and* q *are unknown real- and vector-valued functions respectively and* $E(Y, W')' < \infty$*.*

S.1(i) introduces the observed (or measured) variables M. S.1(ii) introduces the unobserved (or latent) variables $L \equiv (U_W', U_Y', U')'$ and assumes that Y and W are generated according to the unknown nonparametric nonseparable structural functions r and q. The causal effects of X on Y and those of U on Y and W are features of the structural functions r and q whereas the observability of X or U is an empirical matter. Last, the vector of potential instruments Z may equal to, or contain elements of, X. Whereas we restrict how U_Y and U_W depend on Z below, we allow U to statistically depend on Z which complicates the identification of the effects of X on Y. Thus, the elements of Z may, but need not, be valid instruments.

4 Identification of Coefficients in a Linear Structure

Although the paper's framework does not require a linear or parametric effect of X on Y, it is instructive to begin the identification analysis by studying the linear specification S.2, with constant coefficients. Appendix A.1 builds on this section's analysis to consider random, albeit exogenous, coefficients (that depend on U_Y or U_W respectively) in the Y and W equations and to explicitly accommodate the covariates.

Assumption 2 (S.2) *Linearity: Assume S.1 and let*

$$Y = r(X, U, U_Y) = X'\bar{\beta} + U'\bar{\delta}_Y + U_Y'\bar{\alpha}_Y \quad \textit{and} \quad W' = q(U, U_W)' = U'\bar{\delta}_W + U_W'\bar{\alpha}_W.$$

Section 4 studies the identification of the (average) effect $\bar{\beta}$ of X on Y in the benchmark specification S.2, where the omitted variables U enters the Y and W equations linearly. Sections 5 and 6 study two nonseparable models that allow the effect of X on Y to depend on U.

4.1 IV Regression Notation

To shorten the notation throughout, for a random vector A with a finite mean, we write:

$$\bar{A} \equiv E(A) \quad \text{and} \quad \tilde{A} \equiv A - \bar{A}.$$

For example, $\bar{\beta} \equiv E(\beta)$. Further, for random vectors B and C of equal dimension with $Cov(C, A)$ finite and $Cov(C, B)$ finite and nonsingular, we use the following succinct notation for the linear IV regression estimand and residual

$$R_{A.B|C} \equiv Cov(C, B)^{-1} Cov(C, A) \quad \text{and} \quad \epsilon'_{A.B|C} \equiv \tilde{A}' - \tilde{B}' R_{A.B|C}.$$

By construction, $E(\epsilon_{A.B|C}) = 0$ and $Cov(C, \epsilon_{A.B|C}) = 0$. Thus, $R_{A.B|C}$ is the vector of slope coefficients associated with B in a linear IV regression of A on $(1, B')'$ using instruments $(1, C')'$. If $B = C$, we obtain the linear regression estimand $R_{A.B} \equiv R_{A.B|B}$ and residual $\epsilon_{A.B} \equiv \epsilon_{A.B|B}$.

4.2 Characterization and Point Identification

Next, we formalize the discussion illustrated in equations (1,2) and extend it to allow U and W to be vectors. First, Theorem 4.1 derives the IV regression bias B for $\bar{\beta}$. Then, it uses W in characterizing the expression for B.

Theorem 4.1 *Assume S.2 with $\ell = k$ and $m = l$. Let $Cov[Z, (Y, W')'] < \infty$.*
(i) If (i.a) $Cov(Z, X)$ is nonsingular and (i.b) $Cov(U_Y, Z) = 0$ then

$$B \equiv R_{Y.X|Z} - \bar{\beta} = R_{U.X|Z}\bar{\delta}_Y.$$

(ii) If, in addition, (ii.a) $\bar{\delta}_W$ is nonsingular with $\bar{\delta} \equiv \bar{\delta}_W^{-1}\bar{\delta}_Y$ and (ii.b) $Cov(U_W, Z) = 0$ then

$$B = R_{W.X|Z}\bar{\delta}.$$

The most basic version of Theorem 4.1 obtains when $Z = X$. In this case, (i) yields the standard linear regression OVB formula $B = R_{U.X}\bar{\delta}_Y$. More generally, Theorem 4.1 derives the IV OVB when Z may differ from X and shows how the direction of the bias depends on $R_{U.X|Z}$ and $\bar{\delta}_Y$. Last, if a proxy is available then $\bar{\beta}$ is characterized in (ii) by

$$\bar{\beta} = R_{Y.X|Z} - R_{W.X|Z}\bar{\delta}.$$

The uncorrelation condition $(i.b)$ would suffice for $R_{Y.X|Z}$ to point identify $\bar{\beta}$ had $Cov(U, Z) = 0$, as would occur if U is degenerate and there are not omitted variables. Theorem 4.1 characterizes the bias of the estimand $R_{Y.X|Z}$ for $\bar{\beta}$ under condition $(i.b)$ when U is unobserved and freely correlated with Z. Thus, condition $(i.b)$ isolates the omitted variable U as the source of the bias of $R_{Y.X|Z}$. Similarly, condition $(ii.b)$ ensures that W is an informative proxy, with $Cov(Z, W)$ arising solely due to U. Last, the nonsingularity conditions $(i.a)$ and $(ii.a)$ require that $\ell = k$ and $m = l$. More generally, one may consider $\ell \geq k$ and $m \geq l$, and having $\ell \geq k+m$ may point identify $\bar{\beta}$. For example, if $\ell = k + m$, $m = l$, and $Cov[Z, (X', W')']$ is nonsingular then $(\bar{\beta}', \bar{\delta}')' = R_{Y.(X',W')'|Z}$. This paper does not require this many instruments and thus $\bar{\beta}$ is "under-identified." In particular, we let $\ell = k$, as would obtain when $Z = X$.

To illustrate how Theorem 4.1 applies to point identify $\bar{\beta}$, consider equations (1,2) where $m = l = 1$. Under exogeneity, the IV OVB disappears and $R_{Y.X|Z} = \bar{\beta}$. This occurs in the event that Z and U are uncorrelated, in which case $R_{W.X|Z} = 0$, or if one assumes that U does not determine Y, and in particular $\bar{\delta}_Y = 0$. Alternatively, if one assumes that W is a perfect proxy with degenerate U_W then both $\bar{\beta}$ and the ratio $\frac{\bar{\delta}_Y}{\bar{\delta}_W}$ of the average direct effect of U on Y to that of U on W may be point identified. Specifically, provided $Cov[U_Y, (Z', U)'] = 0$ and $Cov[(Z', W)', (X', W)']$ is nonsingular, we have $(\bar{\beta}', \frac{\bar{\delta}_Y}{\bar{\delta}_W})' = R_{Y.(X',W')'|(Z',W')'}$. When W is error-laden as in Theorem 4.1, $\bar{\beta} = R_{Y.X|Z} - R_{W.X|Z}\frac{\bar{\delta}_Y}{\bar{\delta}_W}$ is point identified under proportional confounding, when the average effects $\bar{\delta}_Y$ and $\bar{\delta}_W$ are assumed to be of a known proportion, $\frac{\bar{\delta}_Y}{\bar{\delta}_W} = d$. For example, under equiconfounding, these average effects are of equal magnitude $\left|\frac{\bar{\delta}_Y}{\bar{\delta}_W}\right| = 1$. Then $\bar{\beta}$ is point identified under positive ($\bar{\delta}_Y = \bar{\delta}_W$) or negative ($\bar{\delta}_Y = -\bar{\delta}_W$) equiconfounding by $\bar{\beta} = R_{Y-W.X|Z}$ or $\bar{\beta} = R_{Y+W.X|Z}$ respectively (see Chalak, 2012). More generally, U may be a vector of potential confounders. Often, to each confounder U_h corresponds a proxy $W_h = \alpha_{W_h} + U_h \delta_{W_h}$ so that $W' = \alpha'_W + U'\delta_W$ with $\delta_W = diag(\delta_{W_1}, ..., \delta_{W_m})$. Then

$$\bar{\beta} = R_{Y.X|Z} - R_{W.X|Z}\bar{\delta} = R_{Y.X|Z} - \textstyle\sum_{h=1}^{m} \displaystyle\frac{\bar{\delta}_{Y,h}}{\bar{\delta}_{W_h}} R_{W_h.X|Z}.$$

As before, $\bar{\beta} = R_{Y.X|Z}$ under exogeneity, $(\bar{\beta}', \bar{\delta}')' = R_{Y.(X',W')'|(Z',W')'}$ if W is a perfect proxy, and $\bar{\beta} = R_{Y.X|Z} - \sum_{h=1}^{m} R_{W_h.X|Z} d_h$ under proportional confounding with $\frac{\bar{\delta}_{Y,h}}{\bar{\delta}_{W_h}} = d_h$ for $h = 1, ..., m$.

Corollary 4.2 collects these point identification results and allows for a general matrix δ_W. However, it is useful to keep in mind the leading case where δ_W is a diagonal matrix with straightforward interpretation. Subscripts denote vector elements; e.g. $\bar{\beta}_j$ and $R_{Y.X|Z,j}$ are the j^{th} elements of $\bar{\beta}$ and $R_{Y.X|Z}$ respectively.

Corollary 4.2 *Assume the conditions of Theorem 4.1 and let $j = 1, ..., k$. (i) If $B_j = 0$ (exogeneity) then $\bar{\beta}_j = R_{Y.X|Z,j}$. (ii) If U_W is degenerate (perfect proxy), $Cov(U_Y, U) = 0$, and $Cov[(Z', W')', (X', W')']$ is nonsingular then $(\bar{\beta}', \bar{\delta}')' = R_{Y.(X',W')'|(Z',W')}$. (iii) If $\bar{\delta} = d$ where d is a known vector (proportional confounding) then $\bar{\beta}_j = R_{Y.X|Z,j} - R_{W.X|Z,j} d$.*

In sum, it suffices for exogeneity that $R_{U.X|Z,j} = 0$, in which case $R_{W.X|Z,j} = 0$, or $\bar{\delta}_Y = 0$. Moreover, if one fails to reject the null hypothesis $R_{W.X|Z,j} = 0$ against the alternative $R_{W.X|Z,j} \neq 0$, e.g. via a t-test in the scalar proxy case, then one cannot reject, under Theorem 4.1's assumptions, that $R_{Y.X|Z,j}$ point identifies $\bar{\beta}_j$. Last, assuming either a perfect proxy or proportional confounding point identifies $\bar{\beta}$.

4.3 Partial Identification

Magnitude and sign restrictions on confounding that are weaker than requiring exogeneity ($\bar{\delta}_Y = 0$ when U may depend on Z), a perfect proxy (with $\bar{\delta}$ identified), or proportional confounding ($\bar{\delta} = d$) can partially identify the elements of $\bar{\beta}$. To illustrate, consider the earnings ($Y = \log(Wage)$) and proxy ($W = \log(KWW)$) equations (1,2). We consider how the average direct effect $\bar{\delta}_Y$ of U on Y compares in magnitude and sign to the average effect $\bar{\delta}_W$ of U on W.

Suppose that $\left|\bar{\delta}\right| \equiv \left|\frac{\bar{\delta}_Y}{\bar{\delta}_W}\right| \leq 1$ so that (on average) the response of W to U is at least as large as the direct response of Y to U. For example, this assumes that the elasticity of the test score with respect to ability is at least as large as the elasticity of wage with respect to ability. Assume further that $0 \leq \bar{\delta}$ so that these average responses have the same sign. Then $\bar{\delta} \in \mathcal{D} = [0, 1]$ and the expression for $\bar{\beta}$ from Theorem 4.1 implies that, for $j = 1, ..., k$, $\bar{\beta}_j$ is partially identified in the region $\mathcal{B}_j([0, 1])$ given by:

$$\mathcal{B}_j([0,1]) = \begin{cases} [R_{Y.X|Z,j},\ R_{Y.X|Z,j} - R_{W.X|Z,j}] & \text{if } R_{W.X|Z,j} \leq 0 \\ [R_{Y.X|Z,j} - R_{W.X|Z,j},\ R_{Y.X|Z,j}] & \text{if } 0 \leq R_{W.X|Z,j} \end{cases}.$$

The mirror-image identification region for $\bar{\beta}_j$ obtains if one assumes that $\bar{\delta} \in \mathcal{D} = [-1, 0]$.

Instead, if $1 \leq \left|\bar{\delta}\right| \equiv \left|\frac{\bar{\delta}_Y}{\bar{\delta}_W}\right|$ so that (on average) the response (elasticity) of W to U is at most as large as the direct response of Y to U, and $0 \leq \bar{\delta}$ then $\bar{\delta} \in \mathcal{D} = [1, +\infty)$ and we obtain the identification region $\mathcal{B}_j([1, +\infty))$ for $\bar{\beta}_j$:

$$\mathcal{B}_j([1,+\infty)) = \begin{cases} [R_{Y.X|Z,j} - R_{W.X|Z,j},\ +\infty) & \text{if } R_{W.X|Z,j} \leq 0 \\ (-\infty,\ R_{Y.X|Z,j} - R_{W.X|Z,j}] & \text{if } 0 \leq R_{W.X|Z,j} \end{cases}.$$

Note that $\mathcal{B}_j([1, +\infty))$ excludes the IV estimand $R_{Y.X|Z,j}$. The mirror-image result obtains when $\bar{\delta} \in \mathcal{D} = (-\infty, -1]$.

Wider identification regions obtain under either magnitude or sign (but not both) restrictions on $\bar{\delta}_Y$ and $\bar{\delta}_W$. In particular, if $\left|\bar{\delta}_Y\right| \leq \left|\bar{\delta}_W\right|$ then $\bar{\delta} \in \mathcal{D} = [-1, 1]$ and

$$\mathcal{B}_j([-1,1]) = \begin{cases} [R_{Y.X|Z,j} + R_{W.X|Z,j},\ R_{Y.X|Z,j} - R_{W.X|Z,j}] & \text{if } R_{W.X|Z,j} \leq 0 \\ [R_{Y.X|Z,j} - R_{W.X|Z,j},\ R_{Y.X|Z,j} + R_{W.X|Z,j}] & \text{if } 0 \leq R_{W.X|Z,j} \end{cases}.$$

Note that $\mathcal{B}_j([-1, 1])$ is twice as large as $\mathcal{B}_j([0, 1])$ or $\mathcal{B}_j([-1, 0])$. Also, the "closer" Z is to exogeneity, the smaller $\left|R_{W.X|Z,j}\right|$ is, and the tighter these three identification regions are. Alternatively, if $\left|\bar{\delta}_W\right| \leq \left|\bar{\delta}_Y\right|$ then $\bar{\delta} \in \mathcal{D} = (-\infty, -1] \cup [1, +\infty)$ and $\bar{\beta}_j$ is partially identified in

$$\mathcal{B}_j((-\infty,-1]\cup[1,+\infty)) = \begin{cases} (-\infty, R_{Y.X|Z,j} + R_{W.X|Z,j}] \cup [R_{Y.X|Z,j} - R_{W.X|Z,j}, +\infty) & \text{if } R_{W.X|Z,j} \leq 0 \\ (-\infty, R_{Y.X|Z,j} - R_{W.X|Z,j}] \cup [R_{Y.X|Z,j} + R_{W.X|Z,j}, +\infty) & \text{if } 0 \leq R_{W.X|Z,j} \end{cases}.$$

In this case, the "farther" Z is from exogeneity, the larger $\left|R_{W.X|Z,j}\right|$ is, and the more informative $\mathcal{B}_j((-\infty,-1])$, $\mathcal{B}_j([1,+\infty))$, and $\mathcal{B}_j((-\infty,-1]\cup[1,+\infty))$ are.

Alone, sign restrictions on $\frac{\bar{\delta}_Y}{\delta_W}$ yield the following identification region which determines the direction of the IV regression OVB. In particular, we have

$$\mathcal{B}_j([0,+\infty)) = \begin{cases} [R_{Y.X|Z,j}, +\infty) & \text{if } R_{W.X|Z,j} \leq 0 \\ (-\infty, R_{Y.X|Z,j}] & \text{if } 0 \leq R_{W.X|Z,j} \end{cases}$$

and the mirror-image result obtains for $\mathcal{B}_j((-\infty,0])$.

These identification regions obtain by restricting the sign and/or magnitude of $\frac{\bar{\delta}_Y}{\delta_W}$. More generally, a researcher may impose a lower bound d_L and an upper bound d_H on $\frac{\bar{\delta}_Y}{\delta_W}$ so that $\bar{\delta} \in \mathcal{D} = [d_L, d_H]$. For example, $\mathcal{D}$ may contain the perfect proxy estimate or 95% confidence interval for $\frac{\bar{\delta}_Y}{\delta_W}$. In this case, similar identification regions that involve $R_{Y.X|Z,j}$, $d_L R_{W.X|Z,j}$, and $d_H R_{W.X|Z,j}$ obtain. Moreover, suppose that U is a vector and there is a proxy $W_h = \alpha_{W_h} + U_h \delta_{W_h}$ for each confounder U_h, $h = 1, ..., m$, so that $\delta_W = diag(\delta_{W_1}, ..., \delta_{W_m})$. Then $\bar{\beta}_j = R_{Y.X|Z,j} - \sum_{h=1}^{m} R_{W_h.X|Z,j} \frac{\bar{\delta}_{Y,h}}{\delta_{W_h}}$ and magnitude and/or sign restrictions on $\bar{\delta}_h = \frac{\bar{\delta}_{Y,h}}{\delta_{W_h}} \in \mathcal{D}_h$, $h = 1, ..., m$, yield the identification region $\mathcal{B}_j(\times_{h=1}^{m}\mathcal{D}_h)$ for $\bar{\beta}_j$, $j = 1, ..., k$, discussed next.

Corollary 4.3 formalizes this discussion for a general matrix δ_W and interval restrictions[11] on $\bar{\delta}_h$. Identification regions under magnitude or sign restrictions (or both) on confounding obtain by setting the vectors d_L and d_H suitably, including possibly $d_{L,h} = -\infty$ or $d_{H,h} = +\infty$.

Corollary 4.3 *Assume the conditions of Theorem 4.1 and that* $\bar{\delta}_h \in \mathcal{D}_h = [d_{L,h,}, d_{H,h}]$, $h = 1, ..., m$. *Then, for* $j = 1, ..., k$,

$$\bar{\beta}_j \in \mathcal{B}_j(\times_{h=1}^{m}\mathcal{D}_h) \equiv \{R_{Y.X|Z,j} - R_{W.X|Z,j} d : d_h \in \mathcal{D}_h, h = 1, ..., m\},$$

and this identification region is sharp.

The bounds $\mathcal{B}_j(\times_{h=1}^{m}\mathcal{D}_h)$ in Corollary 4.3 are sharp. Specifically, Corollary 4.3's proof shows that for each average effect in $\mathcal{B}_j(\times_{h=1}^{m}\mathcal{D}_h)$ there corresponds unobservables (V, V_Y, V_W) and functions r^* and q^* for the Y and W equations that satisfy all the conditions imposed on (U, U_Y, U_W) and r and q in Corollary 4.3 and that could have generated the observables (X, Y, W). Last, different potential instruments or proxies may lead to different identification regions, in which case $\bar{\beta}_j$ is identified in the intersection of these regions, provided it is nonempty.

[11] One can consider other types of restrictions, including a prior distribution on $\bar{\delta}_h$ as in e.g. Conley, Hansen, and Rossi (2012). The interval restriction considered here can be viewed as a restriction on the support of $\bar{\delta}_h$.

4.4 Discussion and Connections to the Literature

To conclude Section 4, we comment on how its results complement several related papers. First, the results relate to the literature which imposes assumptions on the "measurement error" U_W in the proxy W. In particular, requiring that U_W is classical[12], i.e. $Cov[U_W, (X', U', U_Y)'] = 0$, may yield bounds[13] on $\bar{\delta}$ and thus $\bar{\beta}$ (see e.g. Klepper and Leamer, 1984; Bollinger, 2003). Here, we do not require $Cov[U_W, (U, U_Y')'] = 0$. Also, some papers use multiple proxies for identification. For example, consider two scalar proxies $W_h = U\bar{\delta}_{W_h} + U_{W_h}'\bar{\alpha}_{W_h}$, $h = 1, 2$, for U with $\bar{\delta}_{W_1}, \bar{\delta}_{W_2} \neq 0$ and $Cov[(Z', U, U_{W_2})', (U_Y, U_{W_1})'] = 0$. Then an IV regression of Y on $(1, X', W_1)'$ using instruments $(1, Z', W_2)'$ may point identify $(\bar{\beta}', \frac{\bar{\delta}_Y}{\bar{\delta}_{W_1}})'$ (see e.g. Blackburn and Neumark, 1992). This paper's method doesn't require multiple proxies for U. Further, if multiple proxies are available[14], the measurement error is not required to be uncorrelated across proxies. For example, U_{W_1} and U_{W_2} (e.g. test taking skills) may be correlated. Second, this section's results also add to a growing literature that employs several alternative assumptions to partially identify the coefficients in a linear or parametric model when exogeneity may fail. For example, Altonji, Conley, Elder, and Taber (2011) assume that the selection on unobservables occurs similarly to that on observables, restricting how X depends on U and the covariates. Reinhold and Woutersen (2009) and Nevo and Rosen (2012) assume that the correlation between Z and U and that between X and U have the same sign and that Z is less correlated with U than X is. Conley, Hansen, and Rossi (2012) allow Z to enter the linear equation for Y and impose a prior distribution on the coefficient on Z that is weaker than requiring it to be zero. Klein and Vella (2009, 2010) and Lewbel (2012) impose restrictions on the second moments (heteroskedasticity). Bontemps, Magnac, and Maurin (2012) provide additional examples and a general treatment of set identified linear models. Our framework complements these papers since it does not require their identifying assumptions. Further, as demonstrated in Sections 5 and 6, this paper's framework does not require a linear, parametric, or separable specification.

[12] Recall that classical measurement error in a scalar X induces an attenuation bias in the linear regression estimand $R_{Y.X}$ for $\bar{\beta}$ whereas the direction of the bias that results from an omitted variable U depends on $Cov(X, U)$ and the coefficient $\bar{\delta}_Y$ on U.

[13] When U is a scalar, the bounds on $\frac{\bar{\delta}_Y}{\bar{\delta}_W}$ are the coefficient on W in a linear regression or Y on $(1, X', W)'$ and the inverse of the coefficient on Y in a linear regression of W on $(1, X', Y)'$. When U and W are $m \times 1$ vectors, in order for the identification region for $\bar{\delta}$ to be bounded, all linear regressions of any element of $(Y, W')'$ on the remaining elements (after projecting on X) must yield (properly rearranged) coefficient vectors that lie in the same orthant (see Klepper and Leamer, 1984).

[14] Online Appendix B.1 studies the linear case with multiple proxies for U that are also components of X.

Instead, we use imperfect proxies to identify average effects under magnitude and sign restrictions on confounding, including in nonparametric nonseparable structures. Of course, which (combination of) identifying assumption(s) is appropriate depends on the empirical application.

5 Identification of Average Nonparametric Effects

Section 5 focuses on the case in which there are no excluded instruments, i.e. $Z = X$, and then extends Section 4's analysis by removing the linearity assumption[15] S.2. Here, we study the identification of the conditional average effect of X on Y when changing x to x^* given $X = x^*$:

$$\bar{\beta}(x, x^*|x^*) \equiv E[r(x^*, U, U_Y) - r(x, U, U_Y)|X = x^*].$$

For instance, for binary X, $\bar{\beta}(0, 1|1)$ is the average treatment effect on the treated. If r is differentiable in a scalar cause of interest, we set $k = 1$ to denote this variable by X and we subsume, without loss of generality, the remaining causes into the implicit covariates. We then study the identification of the conditional average marginal effect of X on Y at x given $X = x$:

$$\bar{\beta}(x|x) \equiv E[\frac{\partial}{\partial x} r(x, U, U_Y)|X = x].$$

In studying the identification of $\bar{\beta}(x, x^*|x^*)$ and $\bar{\beta}(x|x)$, we use a shorthand notation for the difference and derivative of a nonparametric regression. Specifically, for random vectors A and B with $E(A)$ finite and b and b^* in the support[16] of B, we let

$$R^N_{A.B}(b, b^*) \equiv E(A'|B = b^*) - E(A'|B = b).$$

Further, when B is a scalar and the derivative exists, we write

$$R^N_{A.B}(b) \equiv \frac{\partial}{\partial b} E(A'|B = b).$$

Theorem 5.1 characterizes the nonparametric bias $B(x, x^*|x^*)$ or $B(x|x)$ of the nonparametric regression estimand $R^N_{Y.X}(x, x^*)$ or $R^N_{Y.X}(x)$ in recovering the average effect $\bar{\beta}(x, x^*|x^*)$ or $\bar{\beta}(x|x)$ in the presence of an omitted variable U. While both U and U_Y can generate heterogeneity in the response of Y to X, U is the only source of endogeneity of X (or of "essential

[15] Online Appendix B.2 studies an intermediate case in which U enters r and q additively separably.

[16] Throughout, for random vectors A and B, we denote the cumulative distribution function (cdf) of A by $F_A(\cdot)$ and that of A conditional on $B = b$ by $F_{A|B}(\cdot|b)$. We let the corresponding probability density or mass functions be $f_A(\cdot)$ and $f_{A|B}(\cdot|b)$ respectively. We denote the support of A by $\mathcal{A}$ and that of $A|B = b$ by $\mathcal{A}_b$.

heterogeneity" in the nomenclature of Heckman, Urzua, and Vytlacil (2006)). Specifically, Theorem 5.1 imposes the local (at x and x^*) mean independence condition[17]:

$$E[r(x^\dagger, u, U_Y)|U = u, X = \ddot{x}] = E[r(x^\dagger, u, U_Y)] \quad \text{for } x^\dagger, \ddot{x} \in \{x, x^*\} \text{ and all } u \in \mathcal{U}_{\ddot{x}}. \tag{6a}$$

In the case of the marginal effect $\bar{\beta}(x|x)$, Theorem 5.1 further imposes the local (at x) condition

$$E[\frac{\partial}{\partial x} r(x, u, U_Y)|U = u, X = x] = E[\frac{\partial}{\partial x} r(x, u, U_Y)] \quad \text{for all } u \in \mathcal{U}_x. \tag{6b}$$

Note that $U_Y \perp (U, X)$ implies[18] (6a,6b). If U is degenerate then there are no omitted variables and $U_Y \perp (U, X)$ reduces to the standard exogeneity condition $U_Y \perp X$ (or more generally $U_Y \perp X|S$ as in e.g. Altonji and Matzkin, 2005; Hoderlein and Mammen, 2007; and Imbens and Newey, 2009[19]). In this special case, the weaker local mean independence conditions[20] (6a,6b) suffice for $R^N_{Y.X}(x, x^*)$ or $R^N_{Y.X}(x)$ to point identify $\bar{\beta}(x, x^*|x^*)$ or $\bar{\beta}(x|x)$.

Similarly, Theorem 5.1 imposes the analogous condition to (6a) for the proxy equation:

$$E[q(u, U_W)|U = u, X = \ddot{x}] = E[q(u, U_W)] \quad \text{for } \ddot{x} \in \{x, x^*\} \text{ and } u \in \mathcal{U}_{\ddot{x}}, \tag{7}$$

so that W is an informative proxy with the mean dependence of W on X at x and x^* arising solely due to U. Here, $U_W \perp (U, X)$ implies the local mean independence condition (7).

Using Theorem 5.1's characterization, Corollary 5.2 partially identifies $\bar{\beta}(x, x^*|x^*)$ or $\bar{\beta}(x|x)$ by imposing magnitude and sign restrictions on the average marginal effects of the omitted variable U on the response Y at (x, u) and on the proxy variable W at u, denoted by:

$$\bar{\delta}_Y(u; x) \equiv E[\frac{\partial}{\partial u} r(x, u, U_Y)] \quad \text{and} \quad \bar{\delta}_W(u) \equiv E[\frac{\partial}{\partial u} q(u, U_W)].$$

For brevity, Theorem 5.1 states the results in the case where r and q are differentiable in u and the distribution of U given X is continuous or, as a limiting case, degenerate[21]. Theorem A.7

[17] Mean independence conditions are often employed to identify average causal effects. See. e.g. Manski (1990) and Heckman, Ichimura, and Todd (1998).

[18] $U_Y \perp (U, X)$ is not necessary - for instance, Theorem 5.1 permits $Var(U_Y|X)$ to depend on X.

[19] Similar to Imbens and Newey (2009), one can consider covariates S_2 and a scalar unobserved S_1 recoverable from a choice equation $X = \tilde{p}(Z, S_2, S_1)$ with $\tilde{p}$ monotonic in S_1, such that $(U_Y, S_1) \perp Z\ |S_2$, yielding $U_Y \perp X|S$ with $S = (S_1, S_2')'$. We allow but do not require this possibility.

[20] For example, if U is degenerate at u and X is binary then condition (6a) states that the potential outcomes $r(0, u, U_Y)$ and $r(1, u, U_Y)$ are mean independent of the treatment X.

[21] It may be convenient to view the case in which $U|X = x$ is degenerate at $u(x)$ as a limiting case for a sequence of absolutely continuous $F^\tau_{U|X}(u|x)$ as $\tau \to 0$. In particular, one can set $F_{U|X}(u|x) = H(u - u(x))$ where $H(\cdot)$ is the Heaviside step function. Then, when $u(x)$ is a differentiable function, $\frac{\partial}{\partial x} F_{U|X}(u|x) = -\frac{\partial}{\partial x} u(x) \delta(u - u(x))$ where $\delta(\cdot)$ is the Dirac delta function, with an impulse concentrated at $u(x)$ (see e.g., Bracewell, 1986).

in Appendix A.3 gives the results for discrete U, with sums replacing integrals. To proceed, we collect in Assumption B.1 regularity conditions that ensure that the moments and derivatives exist and justify interchanging the order of the derivative and integral in expressions such as $R^N_{Y.X}(x) = \frac{\partial}{\partial x}\int_{\mathcal{U}_x} E[r(x,u,U_Y)]f_{U|X}(u|x)du$, where we use (6a). For this, B.1 also lets $\mathcal{U}_x$ be constant in a neighborhood of x (or $\mathcal{U}_x = \mathcal{U}_{x^*}$ in the case of $\bar{\beta}(x,x^*|x^*)$) to remove the complication introduced by the boundary terms. To give a stronger simpler condition, given (6a,6b,7), it suffices for B.1 that *(i)* $\mathcal{U}$ is compact and $\mathcal{U}_x{=}\,\mathcal{U}$ for all x in $\mathcal{X}$ and that, for all values of the fixed argument(s) in *(ii-iv)*, *(ii)* $r(x,\cdot,u_y)$ and $q(\cdot,u_w)$ (resp. $f_{U|X}(\cdot|x)$ and, for $k=1$, $\frac{\partial}{\partial x}f_{U|X}(\cdot|x)$ and $\frac{\partial}{\partial x}r(x,\cdot,u_y)$) are continuously differentiable (resp. continuous) on $\mathcal{U}$, *(iii)* $E[r(x,u,U_Y),q(u,U_W)'] < \infty$, and *(iv)* $\frac{\partial}{\partial u}r(x,u,\cdot)$, $\frac{\partial}{\partial x}r(x,u,\cdot)$ for $k=1$, and $\frac{\partial}{\partial u}q(u,\cdot)$ are each bounded in absolute value by integrable functions of u_y and u_w respectively. Assumption B.1 in Appendix B gives weaker local regularity conditions

Theorem 5.1 *Assume S.1 with* $m = l = 1$, $x, x^* \in \mathcal{X}$, *and that* $F_{U|X}(\cdot|x)$ *and* $F_{U|X}(\cdot|x^*)$ *are absolutely continuous or, in the limit, degenerate.*

(i.a) If conditions B.1.i(a,b,c,d) and (6a) hold then

$$B(x,x^*|x^*) \equiv R^N_{Y.X}(x,x^*) - \bar{\beta}(x,x^*|x^*) = -\int_{\mathcal{U}_x} \bar{\delta}_Y(u;x)[F_{U|X}(u|x^*) - F_{U|X}(u|x)]du.$$

(i.b) If conditions B.1.i(b,e,f,g) and (7) hold then

$$R^N_{W.X}(x,x^*) = -\int_{\mathcal{U}_x} \bar{\delta}_W(u)[F_{U|X}(u|x^*) - F_{U|X}(u|x)]du.$$

(ii) Set $k = 1$. *(ii.a) If conditions B.1.i(c,d), B.1.ii(a,b,c,d), and (6a,6b) hold then*

$$B(x|x) \equiv R^N_{Y.X}(x) - \bar{\beta}(x|x) = -\int_{\mathcal{U}_x} \bar{\delta}_Y(u;x)\frac{\partial}{\partial x}F_{U|X}(u|x)du.$$

(ii.b) If conditions B.1.i(f,g), B.1.ii(a,d,e), and (7) hold then

$$R^N_{W.X}(x) = -\int_{\mathcal{U}_x} \bar{\delta}_W(u)\frac{\partial}{\partial x}F_{U|X}(u|x)du.$$

$B(x,x^*|x^*)$ and $B(x|x)$ generalize the classic linear regression OVB formula to the nonparametric nonseparable case. These biases depend on the average marginal effect $\bar{\delta}_Y(u;x)$ of U on Y and on the conditional distribution of $U|X$. This provides insight into the sign of the OVB. For instance, if $\bar{\delta}_Y(u;x)$ is nonnegative for a.e. $u \in \mathcal{U}_x$ (e.g. the average marginal effect of ability on wage is nonnegative) and the stochastic dominance relation $F_{U|X}(u|x^*) \leq F_{U|X}(u|x)$ for a.e. $u \in \mathcal{U}_x$ holds (e.g. the probability of low ability U is small when education is high $(x < x^*)$) then $B(x, x^*|x^*)$ is nonnegative.

Under exogeneity, $B(x, x^*|x^*) = 0$ and $R^N_{Y.X}(x, x^*)$ point identifies $\bar{\beta}(x, x^*|x^*)$. This occurs if $U \perp X$, in which case $R^N_{W.X}(x, x^*) = 0$, or if $\bar{\delta}_Y(u; x) = 0$ for a.e. $u \in \mathcal{U}_x$. Alternatively, suppose that $W = q(U, U_W) \equiv \tilde{q}(U)$ is a perfect proxy, with U_W degenerate and $\tilde{q}$ strictly monotonic in u (see e.g. Olley and Pakes, 1996; Griliches and Mairesse, 1998). Then, substituting for $U = \tilde{q}^{-1}(W)$ in r and using condition (6a), we have that $\bar{\beta}(x, x^*|x^*)$ is point identified (see e.g. White and Chalak, 2013, theorem 4.2):

$$E[E(Y|X = x^*, W) - E(Y|X = x, W)|X = x^*] = \bar{\beta}(x, x^*|x^*).$$

In this case, under $U_Y \perp (U, X)$, the ratio $\frac{\bar{\delta}_Y(u;x)}{\bar{\delta}_W(u)}$ is also point identified by

$$\frac{\partial}{\partial w}E(Y|X = x, W = w) = E[\frac{\partial}{\partial u}r(x, u, U_Y)\frac{\partial}{\partial w}\tilde{q}^{-1}(w)] = \frac{\bar{\delta}_Y(u; x)}{\bar{\delta}_W(u)}.$$

Last, if W is an imperfect proxy as in Theorem 5.1 then $\bar{\beta}(x, x^*|x^*)$ is point identified by $R^N_{Y.X}(x, x^*) - d(x)R^N_{W.X}(x, x^*)$ under proportional confounding, when $\bar{\delta}_Y(u; x) = d(x)\bar{\delta}_W(u)$ for a.e. $u \in \mathcal{U}_x$ and $d(x)$ is known. In this case, the average effects of U on Y (at x) and W are assumed to be of a known proportion $d(x)$ for a.e. $u \in \mathcal{U}_x$. Analogous results hold for $\bar{\beta}(x)$.

Corollary 5.2 characterizes the sharp identification regions for $\bar{\beta}(x, x^*|x^*)$ and $\bar{\beta}(x|x)$ that obtain under weaker magnitude and/or sign restrictions on confounding.

Corollary 5.2 *Suppose that, for a.e.* $u \in \mathcal{U}_x$, $\bar{\delta}_Y(u; x) = d(u, x)\bar{\delta}_W(u)$ *with* $d(u, x) \in \mathcal{D}(x) \equiv [d_L(x), d_H(x)]$.
(i) Under the conditions of Theorem 5.1(i), if $\bar{\delta}_W(u)[F_{U|X}(u|x^*) - F_{U|X}(u|x)]$ *is either nonpositive for a.e.* $u \in \mathcal{U}_x$ *or nonnegative for a.e.* $u \in \mathcal{U}_x$ *then*

$$\bar{\beta}(x, x^*|x^*) \in \mathcal{B}(\mathcal{D}(x)) \equiv \{R^N_{Y.X}(x, x^*) - R^N_{W.X}(x, x^*)d : d \in \mathcal{D}(x)\},$$

and this identification region is sharp.
(ii) Under the conditions of Theorem 5.1(ii), if $\bar{\delta}_W(u)\frac{\partial}{\partial x}F_{U|X}(u|x)$ *is either nonpositive for a.e.* $u \in \mathcal{U}_x$ *or nonnegative for a.e.* $u \in \mathcal{U}_x$ *then*

$$\bar{\beta}(x|x) \in \mathcal{B}(\mathcal{D}(x)) \equiv \{R^N_{Y.X}(x) - R^N_{W.X}(x)d : d \in \mathcal{D}(x)\},$$

and this identification region is sharp.

If U is degenerate then exogeneity holds and Corollary 5.2's bounds collapse to the nonparametric regression estimand. More generally, the conditions of Corollary 5.2 obtain if $E[q(u, U_W)]$ is monotonic in u (e.g., on average, the test score is monotonic in ability) and $F_{U|X}(u|x^*) \leq F_{U|X}(u|x)$ for a.e. $u \in \mathcal{U}_x$ (e.g. the probability of low ability U is small when education is high). In particular, Corollary 5.2 analyzes the consequences of deviating from the exogeneity and perfect proxy assumptions by letting the set $\mathcal{D}(x)$ contain zero (recall that $\bar{\delta}_Y(\cdot; x) = 0$ ensures exogeneity) and/or estimates of $\frac{\bar{\delta}_Y(u;x)}{\bar{\delta}_W(u)}$ for $u \in \mathcal{U}_x$ that obtain under the perfect proxy assumption. The identification region $\mathcal{B}(\mathcal{D}(x))$ is sharp under the conditions[22] in Corollary 5.2. We leave studying the consequences of imposing stronger assumptions, such as $U_Y \perp (U, X)$ and $U_W \perp (U, X)$, on the identification of $\bar{\beta}(x, x^*|x^*)$ and $\bar{\beta}(x|x)$ to other work.

To conclude, Section 5's analysis removes the requirement that $U|S$ is degenerate in the conditional exogeneity condition $U_Y \perp (X, U)|S$ where U is an omitted component of the "exogenous treatment" (X, U) given the covariates S. This analysis complements the results in Imbens (2003) who proposes a sensitivity analysis under an alternative weakening of exogeneity that views U as an omitted covariate, with $U_Y \perp X|(U, S)$ and $U \perp S$. Also, the results in Section 5 relate to the literature that uses an error-laden measure[23] W of U to point identify the effect of U on Y under auxiliary assumptions. Recent examples include Hu, Shiu, and Woutersen (2015, 2016) who point identify the coefficient on a mismeasured endogenous variable in a single index model with exogenous instruments and either a separable equation for the latent variable or structural functions that are monotonic in certain unobservables. Last, multiple proxies for U that are mutually independent given U (see e.g. Cunha, Heckman, and Schennach, 2010) may help identify the average nonparametric effect of X on Y. Corollary 5.2's bounds may be useful when multiple proxies are unavailable or mutually dependent given U.

6 Identification of Local and Marginal Treatment Effects

Section 6 focuses on the case where X is binary and then extends Section 5's analysis to allow Z to differ from X. Here, both Z and X may be endogenous. Specifically, we let Y and W be

[22] As can be seen from the proof, $\mathcal{B}(\mathcal{D}(x))$ remains sharp if one strengthens the local conditions (6a,6b) and (7) to require the stronger global mean independence conditions $E[r(x, u, U_Y)|U, X] = E[r(x, u, U_Y)]$ for all $(x, u) \in \mathcal{X} \times \mathcal{U}$ and $E[q(u, U_W)|U, X] = E[q(u, U_W)]$ for all $u \in \mathcal{U}$.

[23] This paper's analysis does not require that the "measurement error" U_W obeys $U_W \perp U_Y$.

as in[24] S.1 and consider a treatment X generated via a threshold crossing selection equation, as in e.g. Heckman and Vytlacil (2005). As shown in Vytlacil (2002), under exogeneity of Z, S.3 is equivalent to the monotonicity assumption in e.g. Imbens and Angrist (1994).

Assumption 3 (S.3) *Assume S.1 and suppose further that X is generated by*[25]

$$X = \mathbf{1}\{U_X \leq \nu(Z)\},$$

where ν is an unknown real-valued function and U_X is an unobserved random variable with $F_{U_X}(\cdot)$ absolutely continuous. We augment $L \equiv (U_X', U_W', U_Y', U')'$ with U_X.

Under S.3, selection into treatment ($X = 1$) holds if and only if $\nu(Z)$ exceeds U_X. When interest attaches to a scalar potential instrument, we set $\ell = 1$ to denote it by Z and we subsume, without loss of generality, the remaining potential instruments into the implicit covariates. We let $F_{U_X}(\cdot)$ be absolutely continuous to simplify the exposition.

It is convenient to rewrite the equation for Y in its random coefficients form

$$Y = [r(1, U, U_Y) - r(0, U, U_Y)]X + r(0, U, U_Y) \equiv \beta(U, U_Y)X + \alpha_Y(U, U_Y). \tag{8}$$

Following the literature (e.g. lmbens and Angrist, 1994; Heckman and Vytlacil, 2005), we study the identification of the conditional local average treatment effect (LATE)

$$\bar{\beta}(\nu(z) < U_X \leq \nu(z^*), z^*) \equiv E[\beta(U, U_Y)|\nu(z) < U_X \leq \nu(z^*), Z = z^*].$$

This is the average treatment effect for the subpopulation with instrument $Z = z^*$ and for whom $X = 0$ if $Z = z$ whereas $X = 1$ if $Z = z^*$. Under $U_X \perp Z$, averaging this local effect over the distribution of Z yields the LATE $\bar{\beta}(\nu(z) < U_X \leq \nu(z^*))$. When Z is binary, this latter effect is the average treatment effect for the "compliers" who receive the treatment ($X = 1$) if and only if $Z = 1$ (see e.g. Angrist, Imbens, and Rubin, 1996). Similarly, we study the identification of the conditional marginal treatment effect (MTE)

$$\bar{\beta}(\nu(z), z) \equiv E[\beta(U, U_Y)|U_X = \nu(z), Z = z].$$

Under $U_X \perp Z$, averaging $\bar{\beta}(\nu(z), \cdot)$ over the distribution of Z yields the MTE $\bar{\beta}(\nu(z))$, the average treatment effect for those who are indifferent toward receiving the treatment if $Z = z$.

[24] Online Appendix A.2 studies the case in which U enters r and q additively separably.

[25] $\mathbf{1}\{A\} = 1$ if A is true and equals 0 otherwise.

In studying the identification of LATE and MTE, we use the following succinct notation for the Wald and local instrumental variable (LIV) estimands. In particular, for random variable B and vectors A and C, provided the means exist and the denominator is nonzero, define

$$R^{Wald}_{A.B|C}(c, c^*) \equiv \frac{R^N_{A.C}(c, c^*)}{R^N_{B.C}(c, c^*)} \equiv \frac{E(A'|C = c^*) - E(A'|C = c)}{E(B|C = c^*) - E(B|C = c)}.$$

Further, when C is a scalar and the derivatives exist with nonzero denominator, let

$$R^{LIV}_{A.B|C}(c) \equiv \frac{R^N_{A.C}(c)}{R^N_{B.C}(c)} \equiv \frac{\frac{\partial}{\partial c}E(A'|C = c)}{\frac{\partial}{\partial c}E(B|C = c)}.$$

Theorem 6.1 characterizes the OVB of $R^{Wald}_{Y.X|Z}(z, z^*)$ or $R^{LIV}_{Y.X|Z}(z)$ in recovering $\bar{\beta}(\nu(z) < U_X \leq \nu(z^*), z^*)$ or $\bar{\beta}(\nu(z), z)$ in the presence of an omitted variable U. It maintains that

$$U_X \perp (U, Z). \tag{9}$$

Further, analogously to Theorem 5.1, Theorem 6.1 restricts the local (at z and z^*) mean dependence of the random coefficients on (U, Z) so that:

$$\begin{aligned} E[\alpha(u, U_Y)|U = u, Z = \ddot{z}] &= E[\alpha(u, U_Y)] \quad \text{and} \\ E[\beta(u, U_Y)|U_X, U = u, Z = \ddot{z}] &= E[\beta(u, U_Y)|U_X] \quad \text{for } \ddot{z} = z, z^* \text{ and all } u \in \mathcal{U}_{\ddot{z}}. \end{aligned} \tag{10}$$

Note that $(U_X, U_Y) \perp (U, Z)$ implies conditions (9,10). Thus, Theorem 6.1 characterizes the bias $B(\nu(z) < U_X \leq \nu(z^*), z^*)$ or $B(\nu(z), z)$ that arises when the exogenous component U of the treatment (U, X) is unobserved and possibly stochastically dependent on the instrument Z for the endogenous treatment component X. If U is degenerate then there are no omitted variables and $(U_X, U_Y) \perp (U, Z)$ reduces to the standard instrument exogeneity condition $(U_X, U_Y) \perp Z$ which ensures the assumptions in e.g. Heckman and Vytlacil (2005) and Imbens and Angrist (1994). In this special case, condition (9) and the local mean independence condition (10) suffice for $R^{Wald}_{Y.X|Z}(z, z^*)$ or $R^{LIV}_{Y.X|Z}(z)$ to point identify $\bar{\beta}(\nu(z) < U_X \leq \nu(z^*), z^*)$ or $\bar{\beta}(\nu(z), z)$. Analogously to $\alpha(u, U_Y)$, we let the local mean dependence of W on Z arise solely due to U:

$$E[q(u, U_W)|U = u, Z = \ddot{z}] = E[q(u, U_W)] \quad \text{for } \ddot{z} = z, z^* \text{ and all } u \in \mathcal{U}_{\ddot{z}}. \tag{11}$$

Here, $U_W \perp (U, Z)$ implies the local mean independence condition (11).

Theorem 6.1 considers the case where r and q are differentiable in u and the distribution of U given Z is continuous or, as a limiting case, degenerate. Here too, Assumption B.2 collects regularity conditions that justify the operations involving derivatives and integrals and lets $\mathcal{U}_z$ be constant in a neighborhood of z (or $\mathcal{U}_z = \mathcal{U}_{z^*}$ in the case of LATE). For example, given (9,10,11) (and provided the denominator $R^N_{X.Z}(z) = f_{U_X}(\nu(z))\frac{\partial}{\partial z}\nu(z)$ is nonzero) it suffices for B.2 that (i) $\mathcal{U}$ is compact and $\mathcal{U}_z = \mathcal{U}$ for all z in $\mathcal{Z}$ and that, for all values of the fixed argument(s) in $(ii\text{-}v)$, (ii) $r(\mathbf{1}\{u_x \leq \nu(z)\}, \cdot, u_y)$ and $q(\cdot, u_w)$ (resp. $f_{U|Z}(\cdot|z)$, $E[\beta(\cdot, U_Y)|U_X = u_x]$, and $\frac{\partial}{\partial z}f_{U|Z}(\cdot|z)$ for $\ell = 1$) are continuously differentiable (resp. continuous) on $\mathcal{U}$, (iii) $E[r(\mathbf{1}\{U_X \leq \nu(z)\}, u, U_Y), q(u, U_W)'] < \infty$, (iv) $E[\beta(u, U_Y)|U_X = \cdot]$ and $f_{U_X}(\cdot)$ are continuous on $\mathcal{U}_\mathcal{X}$ and $\nu(\cdot)$ is continuously differentiable on $\mathcal{Z}$, and (v) $\frac{\partial}{\partial u}r(\mathbf{1}\{\cdot \leq \nu(z)\}, u, \cdot)$ and $\frac{\partial}{\partial u}q(u, \cdot)$ are bounded in absolute value by an integrable function of (u_x, u_y) and u_w respectively. Assumption B.2 in Appendix B gives weaker local regularity conditions. We slightly abuse the previous $\bar{\delta}_Y(u; x)$ notation and denote the average marginal effects of U on Y at (z, u) and W at u by:

$$\bar{\delta}_Y(u; z) \equiv E[\frac{\partial}{\partial u}r(\mathbf{1}\{U_X \leq \nu(z)\}, u, U_Y)] \quad \text{and} \quad \bar{\delta}_W(u) \equiv E[\frac{\partial}{\partial u}q(u, U_W)].$$

Theorem 6.1 *Assume S.1 and S.3 with* $m = l = 1$, $z, z^* \in \mathcal{Z}$, $\Pr[\nu(z) < U_X \leq \nu(z^*)] > 0$, *and that* $F_{U|Z}(\cdot|z)$ *and* $F_{U|Z}(\cdot|z^*)$ *are absolutely continuous or, in the limit, degenerate.*
(i.a) If conditions B.2.i(a,b,c,d) and (9,10) hold then

$$\begin{aligned} B(\nu(z) < U_X \leq \nu(z^*), z^*) &\equiv R^{Wald}_{Y.X|Z}(z, z^*) - \bar{\beta}(\nu(z) < U_X \leq \nu(z^*), z^*) \\ &= -\frac{1}{R^N_{X.Z}(z, z^*)}\int_{\mathcal{U}_z} \bar{\delta}_Y(u; z)[F_{U|Z}(u|z^*) - F_{U|Z}(u|z)]du. \end{aligned}$$

(i.b) If conditions B.2.i(b,e,f,g) and (11) hold then

$$R^{Wald}_{W.X|Z}(z, z^*) = -\frac{1}{R^N_{X.Z}(z, z^*)}\int_{\mathcal{U}_z} \bar{\delta}_W(u)[F_{U|Z}(u|z^*) - F_{U|Z}(u|z)]du.$$

(ii) Set $\ell = 1$. *(ii.a) If conditions B.2.i(c,d), B.2.ii(a,b,c,d,e), and (9,10) hold then*

$$B(\nu(z), z) \equiv R^{LIV}_{Y.X|Z}(z) - \bar{\beta}(\nu(z), z) = -\frac{1}{R^N_{X.Z}(z)}\int_{\mathcal{U}_z} \bar{\delta}_Y(u; z)\frac{\partial}{\partial z}F_{U|Z}(u|z)du.$$

(ii.b) If conditions B.2.i(f,g), B.2.ii(a,c,e,f), and (11) hold then

$$R^{LIV}_{W.X|Z}(z) = -\frac{1}{R^N_{X.Z}(z)}\int_{\mathcal{U}_z} \bar{\delta}_W(u)\frac{\partial}{\partial z}F_{U|Z}(u|z)du.$$

Theorem 6.1 shows how the OVB of the Wald or LIV estimand for the conditional LATE or MTE depends on the average marginal effect of U on Y and on the distribution of $U|Z$. For example, if $\bar{\delta}_Y(u;z)$ is nonnegative for a.e. $u \in \mathcal{U}_z$ (e.g. the average marginal effect of ability on wage is nonnegative) and $F_{U|Z}(u|z^*) \leq F_{U|Z}(u|z)$ for a.e. $u \in \mathcal{U}_z$ (e.g. the probability of low ability is small when in proximity to a college) then $B(\nu(z) < U_X \leq \nu(z^*), z^*)$ is nonnegative.

The OVB $B(\nu(z), z)$ vanishes under exogeneity (e.g. when $U \perp Z$ (and thus $R^{LIV}_{W.X|Z}(z) = 0$) or $\bar{\delta}_Y(u;z) = 0$ for a.e. $u \in \mathcal{U}_z$). Alternatively, if $W = q(U, U_W) = \tilde{q}(U)$ is a perfect proxy, with U_W degenerate and $\tilde{q}$ strictly monotonic, then using $U = \tilde{q}^{-1}(W)$ and (9,10) gives:

$$\frac{E[\frac{\partial}{\partial z}E(Y|Z=z,W)|Z=z]}{\frac{\partial}{\partial z}E(X|Z=z)} = \bar{\beta}(\nu(z), z).$$

In this case, under $(U_X, U_Y) \perp (U, Z)$, the ratio $\frac{\bar{\delta}_Y(u;z)}{\bar{\delta}_W(u)}$ is also point identified by

$$\frac{\partial}{\partial w}E(Y|Z=z, W=w) = E[\frac{\partial}{\partial w}r(\mathbf{1}\{U_X \leq \nu(z)\}, \tilde{q}^{-1}(w), U_Y)] = \frac{\bar{\delta}_Y(u;z)}{\bar{\delta}_W(u)}.$$

Last, when W is an imperfect proxy and proportional confounding holds (i.e. $\bar{\delta}_Y(u;z) = d(z)\bar{\delta}_W(u)$ for a.e. $u \in \mathcal{U}_z$ with $d(z)$ known) then $R^{LIV}_{Y.X|Z}(z) - R^{LIV}_{W.X|Z}(z)d(z)$ point identifies $\bar{\beta}(\nu(z), z)$. Analogous results hold for the LATE $\bar{\beta}(\nu(z) < U_X \leq \nu(z^*), z^*)$.

Restrictions on confounding that are weaker than setting $\frac{\bar{\delta}_Y(u;z)}{\bar{\delta}_W(u)}$ to 0 (exogeneity) or to the perfect proxy estimate can partially identify $\bar{\beta}(\nu(z) < U_X \leq \nu(z^*), z^*)$ or $\bar{\beta}(\nu(z), z)$.

Corollary 6.2 *Suppose that, for a.e.* $u \in \mathcal{U}_z$, $\bar{\delta}_Y(u;z) = d(u,z)\bar{\delta}_W(u)$ *with* $d(u,z) \in \mathcal{D}(z) \equiv [d_L(z), d_H(z)]$. *(i) Under the conditions of Theorem 6.1(i), if* $\bar{\delta}_W(u)[F_{U|Z}(u|z^*) - F_{U|Z}(u|z)]$ *is either nonpositive for a.e.* $u \in \mathcal{U}_z$ *or nonnegative for a.e.* $u \in \mathcal{U}_z$ *then*

$$\bar{\beta}(\nu(z) < U_X \leq \nu(z^*), z^*) \in \mathcal{B}(\mathcal{D}(z)) \equiv \{R^{Wald}_{Y.X|Z}(z, z^*) - R^{Wald}_{W.X|Z}(z, z^*)d : d \in \mathcal{D}(z)\},$$

and this identification region is sharp.
(ii) Under the conditions of Theorem 6.1(ii), if $\bar{\delta}_W(u)\frac{\partial}{\partial z}F_{U|Z}(u|z)$ *is either nonpositive for a.e.* $u \in \mathcal{U}_z$ *or nonnegative for a.e.* $u \in \mathcal{U}_z$ *then*

$$\bar{\beta}(\nu(z), z) \in \mathcal{B}(\mathcal{D}(z)) \equiv \{R^{LIV}_{Y.X|Z}(z) - R^{LIV}_{W.X|Z}(z)d : d \in \mathcal{D}(z)\},$$

and this identification region is sharp.

$\mathcal{B}(\mathcal{D}(z))$ is sharp under the conditions in Corollary 6.2. We leave studying the consequences of imposing stronger assumptions, such as $(U_X, U_Y) \perp (U, Z)$ and $U_W \perp (U, Z)$, on the identification of $\bar{\beta}(\nu(z) < U_X \leq \nu(z^*), z^*)$ and $\bar{\beta}(\nu(z), z)$ to other work[26]. Last, Appendix A.2.2 discusses how one may use the bounds on MTE to partially identify various average effects.

In closing, the analysis in Sections 5 and 6 contributes to the literature on partial identification of nonparametric average effects when X or Z are endogenous. In particular, Manski and Pepper (2000) assume known bounds on the range of Y and that $E[r(x, U, U_Y)|Z = z]$ is monotonic in z. They also consider having r be monotonic in x. Okumura and Usui (2014) combine these assumptions for $Z = X$ along with having r be concave in x. Further, the conditions in Corollaries 5.2 and 6.2 resemble those in Manski and Pepper (2009, lemma 3.1) who show that if r is monotonic in u and $F_{U|W}(u|w^*) \leq F_{U|W}(u|w)$ for all $w \leq w^*$ and u then W is a monotone IV. Sections 5 and 6 do not impose any of the above assumptions. Instead they use restrictions on confounding to partially identify various average effects. Last, one can build on the results in Section 6 to study the identification of various average effects under restrictions on confounding in systems with discrete (nonbinary) or continuous X and possibly mismeasured potential instruments (see e.g. Schennach, White, and Chalak, 2012; Chalak, 2017).

7 Estimation and Inference

The identification regions in Corollaries 4.3, 5.2, and 6.2 are of the form $\mathcal{B}(\mathcal{D})$ =$\{L(R; d) : d \in \mathcal{D}\}$ where the function $L(R; d)$ is known up to a nuisance parameter d, which is partially identified in a known set $\mathcal{D}$, and R collects (IV) regression estimands of Y and W on X (using instruments Z). For example, if $\mathcal{D} = [d_L, d_H]$ then the identification region for $\bar{\beta}(x, x^*|x^*)$ is

$$\mathcal{B}([d_L, d_H]) \equiv \{R^N_{Y.X}(x, x^*) - R^N_{W.X}(x, x^*)d : d \in [d_L, d_H]\},$$

and each element of $\mathcal{B}([d_L, d_H])$ is a linear transformation of $E(Y|X)$ and $E(W|X)$ evaluated at x^* and x. We can estimate the (IV) regression estimands R, underlying each element $\bar{b}(d) = L(R; d)$ of $\mathcal{B}(\mathcal{D})$, using consistent and asymptotically normal parametric, semiparametric, or nonparametric (e.g. kernel) standard estimators $\hat{R}$. We can then estimate the identification region $\mathcal{B}(\mathcal{D})$ consistently using $\hat{\mathcal{B}}(\mathcal{D})$ =$\{L(\hat{R}; d) : d \in \mathcal{D}\}$. Further, for each $d \in \mathcal{D}$, we can derive the asymptotic distribution of $L(\hat{R}; d)$ as a linear transformation of $\hat{R}$ and construct a $1 - \alpha$ (e.g. 95%) confidence interval $C_{1-\alpha}(d)$ for $L(R; d)$. Using proposition 2 of Chernozhukov,

[26] See the comments following the proof of Corollary 6.2 on the sharpness of $\mathcal{B}(\mathcal{D}(z))$ if one strengthens the local conditions (10,11) to the global mean independence conditions $E[\alpha(u, U_Y)|U, Z] = E[\alpha(u, U_Y)]$, $E[\beta(u, U_Y)|U_X, U, Z] = E[\beta(u, U_Y)|U_X]$, and $E[q(u, U_Y)|U, Z] = E[q(u, U_Y)]$ for all $u \in \mathcal{U}$.

Rigobon, and Stoker (2010), a $1-\alpha$ confidence region $CI_{\bar{\beta},1-\alpha}$ for a partially identified parameter $\bar{\beta} \in \mathcal{B}(\mathcal{D})$ then obtains by forming the union[27]:

$$CI_{\bar{\beta},1-\alpha} = \underset{d \in \mathcal{D}}{\cup} C_{1-\alpha}(d).$$

We illustrate the above discussion in the context of the earnings equation specification used in Section 8.1. In this case, X, Z, and the covariates S are binary or discrete variables, and Y and W (here U and W are scalar) are generated by

$$Y = g_X(X)'\bar{\gamma} + U\bar{\delta}_Y + U_Y'\bar{\alpha}_Y \qquad \text{and} \qquad W = U\bar{\delta}_W + U_W'\bar{\alpha}_W. \tag{12}$$

We collect into the vectors $G_X \equiv g_X(X)$, $H_Z \equiv h_Z(Z)$, and $G_S \equiv g_S(S)$ known flexible (e.g. power and threshold crossing) functions of X, Z, and S respectively. Here, the average effect $\bar{\beta}(x, x^*)$ is encoded by the linear transformation $[g_X(x^*) - g_X(x)]'\bar{\gamma}$ of $\bar{\gamma}$. As discussed in Section A.1.1 in Appendix A, when $E(H_Z|S)$ and/or $E[(G_X', W, Y)'|S]$ is affine in G_S, applying Theorem A.1 (the conditional on S version of Theorem 4.1), with G_X and H_Z replacing X and Z, yields $\bar{\gamma}_j = R_{Y.G|H,j} - R_{W.G|H,j}\bar{\delta}$ where we put $G \equiv (G_X', G_S')'$ and $H \equiv (H_Z', G_S')'$. The same characterization for $\bar{\gamma}_j$ obtains under the following specification, with $Cov[H, (U_Y', U_W')'] = 0$,

$$Y = G_X'\bar{\gamma} + G_S'\bar{\psi}_Y + U\bar{\delta}_Y + U_Y'\bar{\alpha}_Y \qquad \text{and} \qquad W = G_S'\bar{\psi}_W + U\bar{\delta}_W + U_W'\bar{\alpha}_W. \tag{13}$$

In either representation, each element in the identification region for $\bar{\beta}(x, x^*)$, obtained under restrictions on confounding, is a linear transformation of $(R_{Y.G|H}', R_{W.G|H}')'$.

To proceed, we first derive the asymptotic distribution of the plug-in estimator $(\hat{R}_{Y.G|H}', \hat{R}_{W.G|H}')'$ for $(R_{Y.G|H}', R_{W.G|H}')'$. This allows for $H = G$. For observations $\{A_i, B_i, C_i\}_{i=1}^n$ corresponding to generic random vector A and random vectors B and C of equal dimension, let $\tilde{A}_i \equiv A_i - \frac{1}{n}\sum_{i=1}^n A_i$ and denote the linear IV regression estimator and sample residuals by:

$$\hat{R}_{A.B|C} \equiv (\frac{1}{n}\sum_{i=1}^n \tilde{C}_i\tilde{B}_i')^{-1}(\frac{1}{n}\sum_{i=1}^n \tilde{C}_i\tilde{A}_i') \qquad \text{and} \qquad \hat{\epsilon}_{A.B|C,i}' \equiv \tilde{A}_i' - \tilde{B}_i'\hat{R}_{A.B|C}.$$

The asymptotic distribution of $\sqrt{n}(\hat{R}_{Y.G|H}', \hat{R}_{W.G|H}')'$ obtains using standard arguments[28]. For this, we put $Q \equiv diag(E(\tilde{H}\tilde{G}'), E(\tilde{H}\tilde{G}'))$.

[27] Alternatively, one can consider adapting the procedures in e.g. Imbens and Manski (2004) and Stoye (2009).

[28] See e.g. White (2001) for primitive (sampling and moment) conditions that ensure the law of large numbers and central limit theorem in conditions (i, ii) of Theorem 7.1.

Theorem 7.1 *Assume S.1(i) with $m = 1$ and that $E[\tilde{H}(\tilde{G}', \tilde{Y}, \tilde{W})]$ is finite and $E(\tilde{H}\tilde{G}')$ is nonsingular. Suppose further that*
(i) $\frac{1}{n}\sum_{i=1}^{n} \tilde{H}_i \tilde{G}_i' \xrightarrow{p} E(\tilde{H}\tilde{G}')$, and
(ii) $n^{-1/2}\sum_{i=1}^{n}(\tilde{H}_i'\epsilon_{Y.G|H,i}, \tilde{H}_i'\epsilon_{W.G|H,i})' \xrightarrow{d} N(0, \Xi)$, where

$$\Xi \equiv \left[\begin{array}{cc} E(\tilde{H}\epsilon^2_{Y.G|H}\tilde{H}') & E(\tilde{H}\epsilon_{Y.G|H}\epsilon_{W.G|H}\tilde{H}') \\ E(\tilde{H}\epsilon_{W.G|H}\epsilon_{Y.G|H}\tilde{H}') & E(\tilde{H}\epsilon^2_{W.G|H}\tilde{H}') \end{array} \right]$$

is finite and positive definite.

Then $\Lambda \equiv Q^{-1}\Xi Q'^{-1}$ is finite and positive definite and

$$\sqrt{n}((\hat{R}'_{Y.G|H}, \hat{R}'_{W.G|H})' - (R'_{Y.G|H}, R'_{W.G|H})') \xrightarrow{d} N(0, \Lambda).$$

The asymptotic distribution of the estimator for each element of the identification region for $\bar{\beta}(x, x^*)$ then obtains as a linear transformation of that of $\sqrt{n}(\hat{R}'_{Y.G|H}, \hat{R}'_{W.G|H})'$. For instance, suppose that X_j is the j^{th} component of G_X and that the effect β_j of X_j on Y is linear so that $\bar{\beta}_j = \bar{\gamma}_j$. Then $\bar{\beta}_j$ is partially identified in $\mathcal{G}_j([d_L, d_H]) = \{R_{Y.G|H,j} - R_{W.G|H,j}d : d \in [d_L, d_H]\}$. One can estimate each element $\bar{\gamma}_j(d)$ of $\mathcal{G}_j([d_L, d_H])$ by $\hat{\gamma}_j(d) = \hat{R}_{Y.G|H,j} - \hat{R}_{W.G|H,j}d$. Using Theorem 7.1, the asymptotic distribution of $\sqrt{n}\hat{\gamma}_j(d)$ obtains as the j^{th} component of

$$\sqrt{n}(\hat{\gamma}(d) - \bar{\gamma}(d)) \xrightarrow{d} N(0, \Sigma(d)),$$

where $\Sigma(d)$ can be written as (I is the identity matrix)

$$\Sigma(d) = \left[\begin{array}{cc} I & -dI \end{array} \right] \Lambda \left[\begin{array}{cc} I & -dI \end{array} \right]' = E(\tilde{H}\tilde{G}')^{-1}E(\tilde{H}\epsilon^2_{(Y-dW).G|H}\tilde{H}')E(\tilde{G}\tilde{H}')^{-1}.$$

Under regularity conditions (e.g. White, 1980, 2001), we can consistently estimate Λ and therefore $\Sigma(d)$. In particular, we use the heteroskedasticity-robust estimator

$$\hat{\Sigma}(d) \equiv (\frac{1}{n}\sum_{i=1}^{n}\tilde{H}_i\tilde{G}_i')^{-1}(\frac{1}{n}\sum_{i=1}^{n}\tilde{H}_i\hat{\epsilon}^2_{(Y-dW).G|H,i}\tilde{H}')(\frac{1}{n}\sum_{i=1}^{n}\tilde{G}_i\tilde{H}_i')^{-1}.$$

For each $d \in \mathcal{D} = [d_L, d_H]$, we use $\hat{\Sigma}(d)$ to construct a $1 - \alpha$ (e.g. 95%) confidence interval $C_{1-\alpha}(d)$ for $\bar{\gamma}_j(d)$ and obtain a $1 - \alpha$ confidence region $CI_{\bar{\gamma}_j, 1-\alpha} = \underset{d \in \mathcal{D}}{\cup} C_{1-\alpha}(d)$ for the partially identified parameter $\bar{\beta}_j = \bar{\gamma}_j \in \mathcal{G}_j(\mathcal{D})$. More generally, the interpretation of $\bar{\gamma}_j$ depends on the functional form of $g_X(\cdot)$. In the empirical application, we report estimates for $\mathcal{G}_j([0, 1])$ and $\mathcal{G}_j([-1, 1])$ as well as the $CI_{\bar{\gamma}_j, 1-\alpha}$ for $\bar{\gamma}_j$ that is partially identified in these regions.

8 Return to Education and the Black-White Wage Gap

Card (1999, section 4 and tables 4 and 6) surveys several papers that estimate the effect of educational attainment on earnings. Among these, studies using institutional features as instruments for education report estimates for the return to a year of education ranging from 6% to 15.3%. While these IV estimates are higher than the surveyed regression estimates, which range from 5.2% to 8.5%, they are much less precise. On the other hand, the surveyed twins studies report smaller within-family differenced estimates for the return to education, with regression estimates ranging from 2.2% to 7.8% and IV estimates (to correct for any error in reported education) ranging from 2.4% to 11%. Similarly, many studies document a black-white wage gap and study its causes. For example, Neal and Johnson (1996) employ a test score to control for unobserved skill and argue that the black-white wage gap reflects primarily a skill gap rather than labor market discrimination (see also Bollinger (2003) who allows the test score to measure human capital with classical measurement error). Lang and Manove (2011) provide a model which suggests that one should control for both education and the test score when comparing the earnings of blacks and whites and document a substantial black-white wage gap in this case. See also Carneiro, Heckman, and Masterov (2005) and Fryer (2011).

We apply the paper's framework to study the financial return to education and the black-white wage gap. Although our framework does not require it, we use the specification in (12) (or alternatively (13)). This generalizes common specifications for the wage equation (e.g. Card, 1995) by allowing for unobserved confounders and nonlinear effects, thereby facilitating comparing our findings to the literature. Here, Y denotes the logarithm of hourly wage and X consists of completed years of education, years of experience, and a binary variable that takes the value 1 if a person is black and is 0 otherwise. $G_X \equiv g_X(X)$ is a vector of flexible (e.g. power, threshold crossing) functions of X discussed below. The confounder U denotes unobserved skill or "ability" and is potentially correlated with elements of G_X given the covariates S. The proxy W for U denotes the logarithm of KWW, a test of occupational information. We use data drawn from the 1976 subset of the National Longitudinal Survey of Young Men (NLSYM), described in[29] Card (1995). The sample used in Card (1995) contains 3010 observations on individuals who reported valid wage and education. We drop 47 observations with missing KWW score[30],

[29] This sample is reported at http://davidcard.berkeley.edu/data_sets.html and in Wooldridge (2012).

[30] The sample also contains IQ score. However, 949 observations report missing IQ score. Using the available

as in some results in Card (1995), leading to a sample size of 2963. The covariates S consist of 2 indicators for living in the South and in a metropolitan area (SMSA), 8 indicators for region of residence in 1966 and 1 for residence in SMSA in 1966, imputed[31] father and mother education and 2 indicators for missing father and mother education respectively, 1 indicator for the presence of the father and mother at age 14 and another indicator for having a single mother at age 14. We employ a vector $G_S \equiv g_S(S)$ of functions of S that contains, in addition to S, 8 binary indicators for interacted mother and father high school, college, or post-graduate education. Throughout, we also consider restricting G_S to a subset S_1 of S, consisting of the two indicators for living in the South and SMSA, as in Card (1995, table 2, column 1). Except when otherwise noted, this generally leads to similar results. The sample also contains data on potential instruments Z - we consider a vector $H_Z \equiv h_Z(Z)$ of functions of Z below.

As discussed in Section 7, using Theorem A.1 and equations (12) (or (13)), we can express the components of $\bar{\gamma}$ by $\bar{\gamma}_j = R_{Y.G|H,j} - R_{W.G|H,j}\bar{\bar{\delta}}$ where $G \equiv (G'_X, G'_S)'$ and $H \equiv (H'_Z, G'_S)'$ (recall that H_Z may equal G_X). The average effect of X on Y is then given by $\bar{\beta}(x, x^*) = [g_X(x^*) - g_X(x)]'\bar{\gamma}$. We consider the magnitude restriction on confounding $\bar{\delta} \equiv \left|\frac{\bar{\delta}_Y}{\bar{\delta}_W}\right| \leq 1$. Recall that in (12), $100\bar{\delta}_Y\%$ and $100\bar{\delta}_W\%$ denote semi-elasticities. Thus, $\left|\bar{\delta}_Y\right| \leq \left|\bar{\delta}_W\right|$ assumes that, given the covariates, an increase in U leads to an average approximate direct percentage change in wage that is at most as large as the percentage change in KWW. This is weaker than imposing $\bar{\delta}_Y = 0$, which would ensure exogeneity when $R_{W.G|H,j} \neq 0$ and U depends on H_Z given S. Further, this admits, but does not require, the perfect proxy estimates of $\frac{\bar{\delta}_Y}{\bar{\delta}_W}$ reported below. The assumption $\left|\bar{\delta}_Y\right| \leq \left|\bar{\delta}_W\right|$ is also in accord with the several findings discussed in Section 2.3 that suggest that the elasticity of wage with respect to ability may be modest. As we discuss shortly, weakening this restriction to $\left|\frac{\bar{\delta}_Y}{\bar{\delta}_W}\right| \leq d$ for $d > 1$ leads to wide identification regions, in which the values of d that exceed 1 correspond to implausible estimates of the average return to education and the black-white wage gap. Sometimes, we also restrict the average effects of ability on wage and KWW to have the same sign, $0 \leq \frac{\bar{\delta}_Y}{\bar{\delta}_W}$. Alone, this sign restriction determines the direction of the (IV) regression bias. For example, it implies that a regression estimand gives an upper bound on the average return to education when the

observations, the sample correlation between IQ and KWW is 0.43 and is strongly significant. Using $\log(IQ)$ instead of $\log(KWW)$ as a proxy often leads to tighter bounds and confidence intervals. However, this could be partly due to sample selection.

[31] Among the 2963 observations, 11.68% (22.78%) are missing the mother's (father's) education. We follow Card (1995) and impute the missing values using the averages of the reported observations.

Table 1: Regression-Based Estimates of the Log Wage Equation Conditional on Covariates under Restrictions on Confounding

j		$\hat{R}_{Y.G,j}$	$\hat{R}_{Y.(G',W')',j}$	$\hat{\mathcal{G}}_j([0,1])$	$\hat{\mathcal{G}}_j([-1,1])$
1	Education	0.072	0.057	[0.001,0.072]	[0.001,0.142]
	(s.e.) and [p-value]	(0.004)	(0.004)	[0.000]	-
	$CI_{.95}$ and $CI_{\bar{\gamma}_j,.95}$	[0.064,0.079]	[0.049, 0.066]	[-0.007,0.079]	[-0.007,0.151]
2	Experience	0.083	0.073	[0.035,0.083]	[0.035,0.131]
	(s.e.) and [p-value]	(0.007)	(0.007)	[0.000]	-
	$CI_{.95}$ and $CI_{\bar{\gamma}_j,.95}$	[0.070,0.096]	[0.060,0.087]	[0.019,0.096]	[0.019,0.148]
3	$\frac{1}{100}$Experience^2	-0.220	-0.202	[-0.220,-0.133]	[-0.307,-0.133]
	(s.e.) and [p-value]	(0.032)	(0.032)	[0.000]	-
	$CI_{.95}$ and $CI_{\bar{\gamma}_j,.95}$	[-0.283,-0.156]	[-0.266,-0.139]	[-0.283,-0.057]	[-0.389,-0.057]
4	Black indicator	-0.187	-0.146	[-0.187,0.015]	[-0.388,0.015]
	(s.e.) and [p-value]	(0.020)	(0.021)	[0.000]	-
	$CI_{.95}$ and $CI_{\bar{\gamma}_j,.95}$	[-0.226,-0.148]	[-0.187,-0.104]	[-0.226,0.057]	[-0.438,0.057]
30	log(KWW)		0.203		
	(s.e.)	-	(0.031)	-	-
	$CI_{.95}$		[0.141,0.264]		

Notes: Y is the logarithm of hourly wage. The proxy W is log(KWW). $G = (G'_X, G'_S)'$ where G_X consists of education, experience, experience squared, and a binary indicator taking the value 1 if a person is black. G_S is the function of the covariates S described in the text. The sample size is 2963. It's a subset of the 3010 observations used in Card (1995) and drawn from the 1976 subset of NLSYM. Robust standard errors (s.e.) appear in parentheses. The p-value associated with a t-test for the null hypothesis $R_{W.G,j} = 0$ against the alternative hypothesis $R_{W.G,j} \neq 0$ appears in brackets below $\hat{\mathcal{G}}_j([0,1])$.

conditional correlation between $\log(KWW)$ and education is positive, which often holds.

8.1 Linear Return to Education

We begin by setting $H_Z = G_X$ where $G_X \equiv g_X(X)$ consists of education, experience, experience squared, and the black binary indicator, as in the specification in Card (1995, table 2, column 5). In this case, the average approximate[32] linear return to education is $100\bar{\gamma}_1\%$ and the average approximate black-white wage gap is $100\bar{\gamma}_4\%$. Below, we consider more general G_X configurations that allow for nonlinear effects. Table 1 reports the results. Column 1 reports the results for the regression estimator $\hat{R}_{Y.G,j}$, which consistently estimates $\bar{\gamma}_j$ under

[32] The coefficient on a binary variable in a log-linear equation does not exactly correspond to a semi-elasticity but we employ this approximation here since it's relatively accurate when the magnitude of the coefficient is small, as is the case for our estimates (see e.g. Halvorsen and Palmquist, 1980).

conditional exogeneity, along with heteroskedasticity-robust standard errors (s.e.) and 95% confidence intervals ($CI_{0.95}$). The regression estimates for the return to education and the black-white wage gap[33], with robust s.e. in parentheses, are 7.2% (0.4%) and -18.7% (2.0%) respectively. Column 2 reports the results for the linear regression estimator $\hat{R}_{Y.(G',W')',j}$ which consistently estimates $\bar{\gamma}$ and $\frac{\bar{\delta}_Y}{\bar{\delta}_W}$ if W is a perfect proxy[34] for U given S. The coefficient on W in $\hat{R}_{Y.(G',W')'}$ estimates $\frac{\bar{\delta}_Y}{\bar{\delta}_W}$ to be 0.2, with robust s.e. 0.03, and the perfect proxy estimates of the average return to education and black-white wage gap, with robust s.e. in parentheses, are 5.7% (0.4%) and -14.6% (2.1%) respectively. Next, we impose the magnitude and sign restriction $0 \leq \frac{\bar{\delta}_Y}{\bar{\delta}_W} \leq 1$ or the magnitude restriction $\left|\frac{\bar{\delta}_Y}{\bar{\delta}_W}\right| \leq 1$ on confounding that are weaker than imposing $\frac{\bar{\delta}_Y}{\bar{\delta}_W} = 0$ (which ensures exogeneity) or requiring the perfect proxy estimate for $\frac{\bar{\delta}_Y}{\bar{\delta}_W}$ (with $CI_{0.95}$ $[0.14, 0.26]$). This enables studying the consequences of reasonable deviations from these standard assumptions. Column 3 reports estimates $\widehat{\mathcal{G}}_j([0,1])$ of the identification region for $\bar{\gamma}_j$ obtained under $0 \leq \frac{\bar{\delta}_Y}{\bar{\delta}_W} \leq 1$, along with the 95% confidence interval $CI_{\bar{\gamma}_j,0.95}$ for $\bar{\gamma}_j$. The estimated bounds on the return to education are $[0.1\%, 7.2\%]$ with $CI_{\bar{\gamma}_1,0.95}$ $[-0.7\%, 7.9\%]$ and those for the black-white wage gap are $[-18.7\%, 1.5\%]$ with $CI_{\bar{\gamma}_4,0.95}$ $[-22.6\%, 5.7\%]$. We also report the p-value associated with a t-test of the null hypothesis that the width of this identification region is zero, $R_{W.G,j} = 0$, against the alternative hypothesis $R_{W.G,j} \neq 0$. Under our assumptions, if we cannot reject that $R_{W.G,j} = 0$ in favor of $R_{W.G,j} \neq 0$ then we cannot reject that $R_{U.G,j} = 0$ and thus that $\bar{\gamma}_j = R_{Y.G|H,j}$. Last, column 4 reports estimates $\widehat{\mathcal{G}}_j([-1,1])$ of the identification region for $\bar{\gamma}_j$ obtained under $\left|\bar{\delta}_Y\right| \leq \left|\bar{\delta}_W\right|$, along with $CI_{\bar{\gamma}_j,0.95}$. Note that weakening $\bar{\delta} \in [0,1]$ to $\bar{\delta} \in [0,d]$ for $d > 1$ to allow the wage to be on average more sensitive to ability than the test score is, extends the estimated identification regions as follows:

$$\begin{aligned} \text{Education: } & \widehat{\mathcal{G}}_1([0,d]) \approx [7.2\% - (d \times 7\%),\ 7.2\%], \quad \text{and} \\ \text{Wage Gap: } & \widehat{\mathcal{G}}_4([0,d]) \approx [-18.7\%,\ -18.7\% + (d \times 20.1\%)]. \end{aligned}$$

In particular, the values of d that are larger than 1 correspond to mostly negative estimates of the average return to education and to a black-white wage gap in favor of blacks, which is unlikely and inconsistent with the general findings in the literature. In this sense, the empirical findings in this paper corroborate the assumption $\left|\bar{\delta}_Y\right| \leq \left|\bar{\delta}_W\right|$. In sum, the regression estimates

[33] The tables also report point estimates and bounds for the coefficients associated with experience and experience squared. For brevity, we don't discuss these in detail.

[34] For instance, this hold if $\bar{\alpha}_W = 0$ and $Cov(U_Y, (G', U)') = 0$ in equations (13).

Table 2: Regression-Based Estimates of the Log Wage Equation with an Education and Race Interaction Term Conditional on Covariates under Restrictions on Confounding

j		$\hat{R}_{Y.G,j}$	$\hat{R}_{Y.(G',W')',j}$	$\hat{\mathcal{G}}_j([0,1])$	$\hat{\mathcal{G}}_j([-1,1])$
1	Education	0.068	0.055	[0.004,0.068]	[0.004,0.131]
	(s.e.) and [p-value]	(0.004)	(0.005)	[0.000]	-
	$CI_{.95}$ and $CI_{\bar{\gamma}_j,.95}$	[0.060,0.076]	[0.046 0.064]	[-0.005,0.076]	[-0.005,0.141]
2	(Education-12)×Black	0.017	0.012	[-0.012,0.017]	[-0.012,0.046]
	(s.e.) and [p-value]	(0.006)	(0.006)	[0.000]	-
	$CI_{.95}$ and $CI_{\bar{\gamma}_j,.95}$	[0.005,0.030]	[-0.001 0.024]	[-0.027,0.030]	[-0.027,0.063]
3	Experience	0.081	0.072	[0.036,0.081]	[0.036,0.127]
	(s.e.) and [p-value]	(0.007)	(0.007)	[0.000]	-
	$CI_{.95}$ and $CI_{\bar{\gamma}_j,.95}$	[0.068,0.095]	[0.059 0.086]	[0.020,0.095]	[0.020,0.143]
4	$\frac{1}{100}$Experience^2	-0.210	-0.196	[-0.210,-0.139]	[-0.280,-0.139]
	(s.e.) and [p-value]	(0.033)	(0.033)	[0.003]	-
	$CI_{.95}$ and $CI_{\bar{\gamma}_j,.95}$	[-0.273,-0.146]	[-0.260 -0.132]	[-0.273,-0.062]	[-0.360,-0.062]
5	Black indicator	-0.193	-0.152	[-0.193,0.018]	[-0.403,0.018]
	(s.e.) and [p-value]	(0.020)	(0.021)	[0.000]	-
	$CI_{.95}$ and $CI_{\bar{\gamma}_j,.95}$	[-0.231,-0.154]	[-0.193 -0.110]	[-0.231,0.062]	[-0.454,0.062]
31	log(KWW)		0.194		
	(s.e.)	-	(0.032)	-	-
	$CI_{.95}$		[0.132 0.257]		

Notes: The results use the specification in Table 1 and augment G_X with an interaction term (Education-12)×Black. The remaining notes in Table 1 apply analogously here.

provide an upper bound for the average (assumed linear for now) return to education and the average black-white wage gap. Further, the estimate of the black-white wage gap is particularly sensitive to deviations from either the regressor exogeneity or the perfect proxy assumption.

8.2 Black-White Return to Education Differential

We augment G_X to include, as its second component, the interaction term $(Education - 12) \times Black$. Table 2 reports the conditional on G_S results. Columns 1 and 2 report the exogeneity and perfect proxy estimates. Under the weaker restriction $0 \leq \frac{\bar{\delta}_Y}{\delta_W} \leq 1$ on confounding, the bounds on the average return to education for non-blacks is $[0.4\%, 6.8\%]$ with $CI_{\bar{\gamma}_1,0.95}$ $[-0.5\%, 7.6\%]$, those on the average black-white return to education differential is $[-1.2\%, 1.7\%]$ with $CI_{\bar{\gamma}_2,0.95}$ $[-2.7\%, 3.0\%]$, and those on the average black-white wage gap for individuals with 12 years of education is $[-19.3\%, 1.8\%]$ with $CI_{\bar{\gamma}_5,0.95}$ $[-23.1\%, 6.2\%]$. Thus, the average return

Table 3: IV Regression-Based Estimates of the Log Wage Equation Conditional on Covariates under Restrictions on Confounding

j		$\hat{R}_{Y.G\|H,j}$	$\hat{R}_{Y.(G',W')'\|(H',W')',j}$	$\hat{\mathcal{G}}_j([0,1])$	$\hat{\mathcal{G}}_j([-1,1])$
1	Education	0.134	0.147	[0.029,0.134]	[0.029,0.240]
	(s.e.) and [p-value]	(0.052)	(0.091)	[0.001]	-
	$CI_{.95}$ and $CI_{\bar{\gamma}_j,.95}$	[0.032,0.237]	[-0.030,0.324]	[-0.078,0.237]	[-0.078,0.372]
2	Experience	0.061	0.064	[0.006,0.061]	[0.006,0.115]
	(s.e.) and [p-value]	(0.025)	(0.019)	[0.000]	-
	$CI_{.95}$ and $CI_{\bar{\gamma}_j,.95}$	[0.011,0.110]	[0.027,0.102]	[-0.044,0.110]	[-0.044,0.178]
3	$\frac{1}{100}$Experience^2	-0.113	-0.119	[-0.113,0.009]	[-0.235,0.009]
	(s.e.) and [p-value]	(0.123)	(0.110)	[0.090]	-
	$CI_{.95}$ and $CI_{\bar{\gamma}_j,.95}$	[-0.354,0.129]	[-0.334 0.097]	[-0.354,0.256]	[-0.545,0.256]
4	Black indicator	-0.162	-0.178	[-0.162,0.026]	[-0.351,0.026]
	(s.e.) and [p-value]	(0.029)	(0.041)	[0.000]	-
	$CI_{.95}$ and $CI_{\bar{\gamma}_j,.95}$	[-0.218,-0.107]	[-0.258,-0.099]	[-0.218,0.085]	[-0.424,0.085]
30	log(KWW)		-0.090		
	(s.e.)	-	(0.298)	-	-
	$CI_{.95}$		[-0.673,0.494]		

Notes: The results use the specification in Table 1 with instruments H_Z for G_X that consist of an indicator for whether there is a four year college in the local labor market, age, age squared, and the black indicator, with $H = (H'_Z, G'_S)'$. The remaining notes in Table 1 apply analogously for the IV-based results here.

to education for the black subpopulation may differ slightly from the nonblack subpopulation, if at all. Below, we follow Card (1995) and maintain that these average returns are equal.

8.3 Distance to School Instrument

A useful solution to the endogeneity problem assumes the availability of valid instruments. For example, Card (1995) uses an indicator for the presence of a four year college in the local labor market, age, and age squared as instruments for education, experience, and experience squared in the specification from Table 1. What if the college proximity instrument is in turn correlated with ability and thus invalid? For example, Carneiro and Heckman (2002) provide evidence suggesting that distance to college may be endogenous. To study this possibility, we apply our framework to relax the assumption that the instrument is conditionally exogenous by allowing H_Z to be correlated with ability[35] given the covariates. As reported in Table 3, the

[35] In particular, we let G_X and G_S be as in the specification from Table 1 and let $H = (H'_Z, G'_S)'$ where $H_Z = h_Z(Z)$ consists of the proximity to college indicator, age, age squared, and the black indicator.

estimates (and robust s.e.) for the average return to education and black-white wage gap are respectively 13.4% (5.2%) and -16.2% (2.9%) under exogeneity and 14.7% (9.1%) and -17.8% (4.1%) when assuming a perfect proxy. Note that, when employing this IV specification, the perfect proxy estimate -0.09 (0.30) for $\frac{\bar{\delta}_Y}{\bar{\delta}_W}$ has an unlikely sign and is imprecisely estimated. For instance, when using this IV specification and conditioning on the subset S_1 of G_S only, the perfect proxy estimate for $\frac{\bar{\delta}_Y}{\bar{\delta}_W}$ becomes 0.21 (0.08), leading to the estimates 5.7% (2%) and -14% (2%) for the average return to education and black-white wage gap respectively. Under $0 \leq \frac{\bar{\delta}_Y}{\bar{\delta}_W} \leq 1$, the conditional on G_S IV-based identification region estimate for the average return to education is [2.9%, 13.4%], which is wider than the regression-based one, with wider $CI_{\bar{\gamma}_1,0.95}$ $[-7.8\%,\ 23.7\%]$. Similarly, the estimated bounds $[-16.2\%,\ 2.6\%]$ on the average black-white wage gap are slightly tighter than the regression-based estimate albeit with comparable $CI_{\bar{\gamma}_4,0.95}$ $[-21.8\%,\ 8.5\%]$. Similar but less precise results obtain when, as in Card (1995), we augment G_S with an indicator for a four year college in the local labor market and employ the product of this indicator with an indicator for low parental education as an instrument instead. Last, in both IV specifications, conditioning on the subset $G_S = S_1$ of the covariates yields generally similar bounds[36]. In sum, the IV-based bounds are generally wider than, or comparable to, the above regression-based ones and yield especially wider confidence intervals.

8.4 Nonlinear Return to Education

Returning to the regression-based estimates with $H_Z = G_X$, we allow for nonlinear year-specific incremental return to education. Specifically, we let G_X contain binary indicators for having at least t years of education, where $t = 2, ..., 18$ as in the sample, instead of the total years of education. Thus, γ_t encodes the incremental return $\beta(t, t+1)$ to year $t+1$ of education. Table 4 reports the results. Column 1 reports the results of the regression estimator which is consistent under exogeneity. Column 2 reports the perfect proxy results , yielding the estimate 0.21 for $\frac{\bar{\delta}_Y}{\bar{\delta}_W}$ with s.e. 0.03. Under the weaker restriction $0 \leq \frac{\bar{\delta}_Y}{\bar{\delta}_W} \leq 1$, we find evidence[37] for

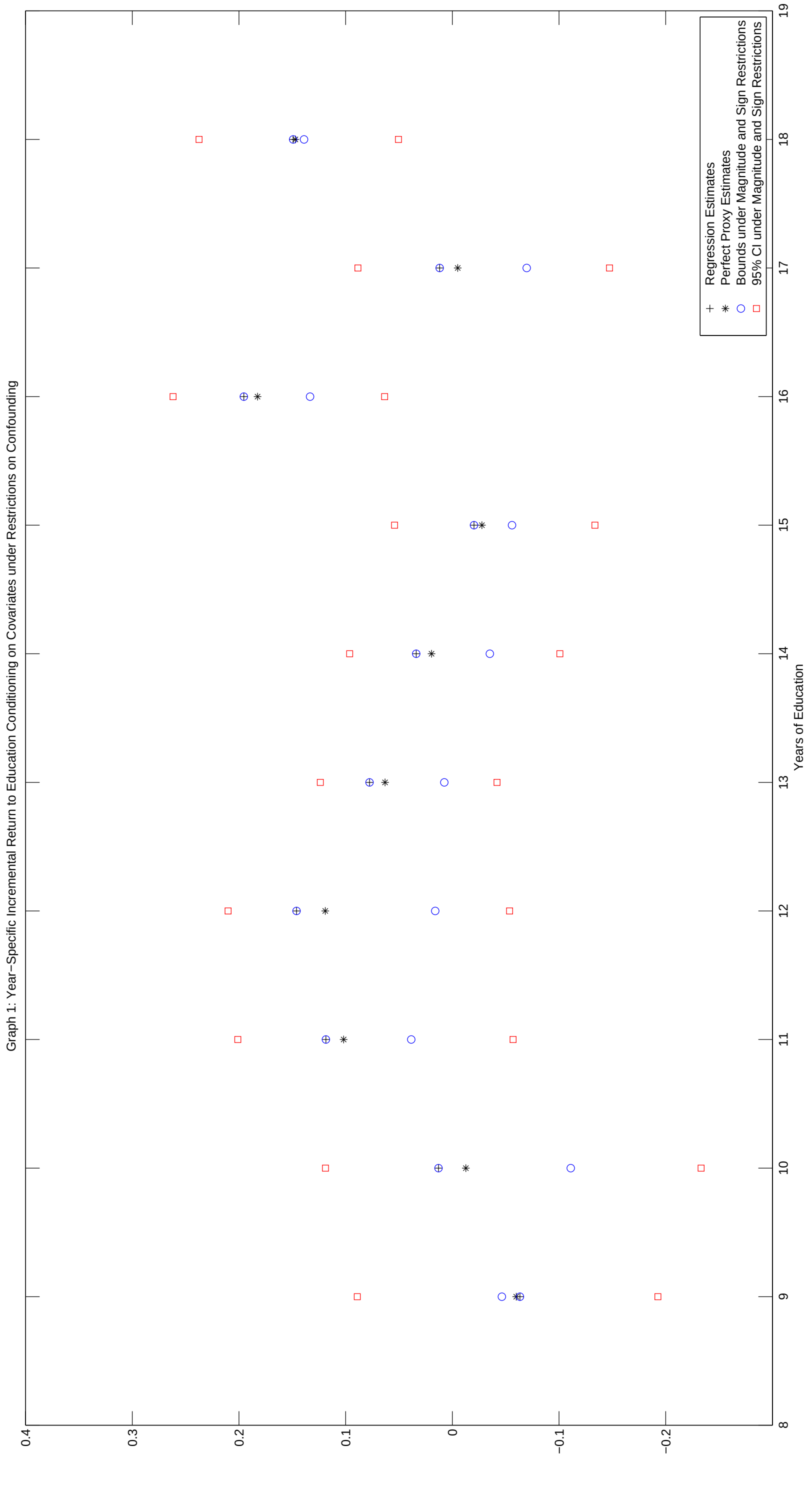
[36] Setting $G_S = S_1$ sometimes leads to tighter identification regions albeit with possibly wider confidence intervals (e.g. $[-10.1\%,\ 2.4\%]$ with $CI_{\bar{\gamma}_4,0.95}$ $[-24.8\%,\ 17.4\%]$ for the average black-white wage gap in the first IV specification and $[-0.6\%,\ 8.1\%]$ with $CI_{\bar{\gamma}_1,0.95}$ $[-3.0\%, 10.3\%]$ for the average return to education in the second IV specification).

[37] Although we don't conduct a formal test for linearity, we note that, under the restriction $0 \leq \frac{\bar{\delta}_Y}{\bar{\delta}_W} \leq 1$, the 95% CI for the partially identified return to the 16^{th} year of education does not overlap with the 95% CI for the partially identified return to e.g. the 15^{th} year and overlaps with that of the 17^{th} year slightly.

Table 4: Regression-Based Estimates of the Log Wage Equation with Year-Specific Education Indicators Conditional on Covariates under Restrictions on Confounding

j		$\hat{R}_{Y.G,j}$	$\hat{R}_{Y.(G',W')',j}$	$\hat{\mathcal{G}}_j([0,1])$	$\hat{\mathcal{G}}_j([-1,1])$
10	Educ $\geq$ 11 years	0.118	0.102	[0.039,0.118]	[0.039,0.198]
	(s.e.) and [p-value]	(0.042)	(0.042)	[0.011]	-
	$CI_{.95}$ and $CI_{\bar{\gamma}_j,.95}$	[0.036,0.201]	[0.020,0.184]	[-0.057,0.201]	[-0.057,0.308]
11	Educ $\geq$ 12 years	0.146	0.119	[0.016,0.146]	[0.016,0.276]
	(s.e.) and [p-value]	(0.033)	(0.032)	[0.000]	-
	$CI_{.95}$ and $CI_{\bar{\gamma}_j,.95}$	[0.082,0.210]	[0.056,0.182]	[-0.054,0.210]	[-0.054,0.359]
12	Educ $\geq$ 13 years	0.078	0.063	[0.007,0.078]	[0.007,0.148]
	(s.e.) and [p-value]	(0.024)	(0.023)	[0.000]	-
	$CI_{.95}$ and $CI_{\bar{\gamma}_j,.95}$	[0.032,0.124]	[0.017,0.109]	[-0.042,0.124]	[-0.042,0.205]
13	Educ $\geq$ 14 years	0.034	0.020	[-0.035,0.034]	[-0.035,0.103]
	(s.e.) and [p-value]	(0.032)	(0.032)	[0.000]	-
	$CI_{.95}$ and $CI_{\bar{\gamma}_j,.95}$	[-0.029,0.096]	[-0.042,0.081]	[-0.101,0.096]	[-0.101,0.177]
14	Educ $\geq$ 15 years	-0.020	-0.028	[-0.056,-0.020]	[-0.056,0.015]
	(s.e.) and [p-value]	(0.038)	(0.038)	[0.048]	-
	$CI_{.95}$ and $CI_{\bar{\gamma}_j,.95}$	[-0.095,0.054]	[-0.101,0.046]	[-0.134,0.054]	[-0.134,0.102]
15	Educ $\geq$ 16 years	0.195	0.183	[0.133,0.195]	[0.133,0.258]
	(s.e.) and [p-value]	(0.034)	(0.034)	[0.000]	-
	$CI_{.95}$ and $CI_{\bar{\gamma}_j,.95}$	[0.129,0.262]	[0.117,0.248]	[0.063,0.262]	[0.063,0.335]
16	Educ $\geq$ 17 years	0.012	-0.005	[-0.070,0.012]	[-0.070,0.093]
	(s.e.) and [p-value]	(0.039)	(0.039)	[0.000]	-
	$CI_{.95}$ and $CI_{\bar{\gamma}_j,.95}$	[-0.065,0.088]	[-0.081,0.071]	[-0.147,0.088]	[-0.147,0.178]
17	Educ $\geq$ 18 years	0.149	0.147	[0.139,0.149]	[0.139,0.159]
	(s.e.) and [p-value]	(0.045)	(0.045)	[0.512]	-
	$CI_{.95}$ and $CI_{\bar{\gamma}_j,.95}$	[0.061,0.237]	[0.060,0.234]	[0.050,0.237]	[0.050,0.257]
20	Black indicator	-0.178	-0.137	[-0.178,0.019]	[-0.374,0.019]
	(s.e.) and [p-value]	(0.020)	(0.021)	[0.000]	-
	$CI_{.95}$ and $CI_{\bar{\gamma}_j,.95}$	[-0.216,-0.139]	[-0.178,-0.096]	[-0.216,0.061]	[-0.423,0.061]
46	log(KWW)		0.207		
	(s.e.)	-	(0.032)	-	-
	$CI_{.95}$		[0.145,0.269]		

Notes: The results extend the specification in Table 1 to include in G_X indicators for having at least t years of education, where $t = 2, ..., 18$ corresponding to the sample, instead of total years of education. For brevity, Table 4 does not report the estimated bounds for the average return to education for $t < 11$; these are often relatively imprecise with wide $CI_{\bar{\gamma}_j,.95}$. Also, Table 4 omits the estimates associated with experience; these are similar to those reported in Table 1. The remaining notes in Table 1 apply analogously here.


Graph 1: Year−Specific Incremental Return to Education Conditioning on Covariates under Restrictions on Confounding
0.4
0.3
0.2
0.1
0
−0.1
−0.2
8
9
10
11
12
13
14
15
16
17
18
19
Years of Education
+ Regression Estimates
* Perfect Proxy Estimates
○ Bounds under Magnitude and Sign Restrictions
□ 95% CI under Magnitude and Sign Restrictions

nonlinearity in the return to education, with the 12^{th}, 16^{th}, and 18^{th} year, corresponding to obtaining a high school, college, and possibly a graduate degree, yielding a high average return. For example, the estimated bounds for the average return to the 12^{th} year are [1.6%, 14.6%] with $CI_{\bar{\gamma}_{11},0.95}$ [−5.4%, 21%] and those for the 16^{th} year are [13.3%, 19.5%] with $CI_{\bar{\gamma}_{15},0.95}$ [6.3%, 26.2%]. Similarly, the estimated bounds for the return to the 18^{th} year are [13.9%, 14.9%] with $CI_{\bar{\gamma}_{17},0.95}$ [5%, 23.7%] and we cannot reject at comfortable significance levels that the width of this region is zero or, under the maintained assumptions, that a regression consistently estimates this return by 14.9% with robust s.e. 4.5%. In contrast, the estimated bounds for the return to the 13^{th} year are [0.7%, 7.8%] with $CI_{\bar{\gamma}_{12},0.95}$ [−4.2%, 12.4%]. Graph 1 illustrates the nonlinearity in the return to education. In addition to the regression and perfect proxy estimates, it plots the estimated bounds and $CI_{\bar{\gamma}_j,0.95}$ for the incremental average returns to the 9^{th} up to the 18^{th} year of education under the restriction $0 \leq \frac{\bar{\delta}_Y}{\bar{\delta}_W} \leq 1$. Last, using this specification, the estimate of the identification region for the black-white wage gap under $0 \leq \frac{\bar{\delta}_Y}{\bar{\delta}_W} \leq 1$ is similar to that in Table 1 and given by [−17.8%, 1.9%] with $CI_{\bar{\gamma}_{20},0.95}$ [−21.6%, 6.1%].

8.5 Discussion and Summary

This empirical analysis employs a parametric specification in which U enters additively separably. Further, it assumes that there is one confounder U denoting "ability," which we proxy using $\log(KWW)$, and that $0 \leq \frac{\bar{\delta}_Y}{\bar{\delta}_W} \leq 1$ or $\left|\bar{\delta}_Y\right| \leq \left|\bar{\delta}_W\right|$. Of course, one should interpret the results carefully if these assumptions are suspected to fail. For example, the analysis relaxes the assumption of exogeneity by allowing ability to act as a confounder. But if other confounders are present and strong valid instruments or proxies for these are not available then additional assumptions are needed to (partially) identify the average effects of X. Similarly, the analysis allows W to be an imperfect proxy, with U_W nondegenerate and conditionally uncorrelated with G_X (or H_Z) in (13). However, this can in turn fail e.g. if U_W denotes test taking skill (respectively access to counseling) and is conditionally correlated with education (respectively distance to school). Section A.1.2 in Appendix A reports complementary results that obtain under some alternative assumptions, such as assuming that the measurement error in the proxy is classical or using the W equation to substitute for U in the Y equation and then assuming that certain excluded covariates (e.g. parental education) from G_S are valid instruments for W. Nevertheless, an advantage of the above empirical analysis is that it does not require several commonly employed assumptions thereby enabling a sensitivity analysis. Specifically, (1) it does not require regressor or instrument exogeneity or restrict the dependence of U on X or Z (given S), (2) it does not require a linear return to education, and (3) it permits a test score to be an error-laden proxy for unobserved ability, with possibly nonclassical measurement error.

In sum, the estimated bounds for the black-white wage gap are relatively wide, suggesting that, under the imposed assumptions that are weaker than requiring exogeneity or a perfect proxy, this data set is inconclusive about the extent of discrimination in the labor market. In contrast, the average return to education for the black subpopulation may differ slightly from the nonblack subpopulation, if at all. Last, we find evidence suggesting a nonlinearity in the return to education, with graduation years yielding a high average return.

9 Conclusion

This paper studies measuring average causal effects in general structural systems with unobserved confounders (omitted variables). We study the identification of coefficients in a linear structure, covariate-conditioned average nonparametric discrete and marginal effects (e.g. average treatment effect on the treated), and local and marginal treatment effects. The first contribution of this paper is to characterize the OVB of common (nonparametric) regression and IV (e.g. Wald and LIV) estimands for these various average effects, thereby generalizing the classic linear regression OVB formula. Using an imperfect proxy for the unobserved confounders, this paper then introduces magnitude and sign restrictions on confounding that are weaker than standard assumptions such as the conditional exogeneity of the treatment or the instrument or requiring a perfect proxy. The paper's second contribution is to demonstrate how these restrictions on confounding can be used to partially identity average effects and to conduct a sensitivity analysis to deviations from the stronger benchmark assumptions. The paper discusses estimation and inference and applies its framework to study the return to education and the black-white wage gap. Extensions for future work include imposing distributional restrictions on confounding (e.g. a prior distribution on $\bar{\delta}$) and using restrictions on confounding to identify the distribution of a causal effect or features of it other than the mean. It is also of interest to apply this paper's framework to estimate production functions.

References

Ackerberg, D., K. Caves, and G. Frazer (2015), “Identification Properties of Recent Production Function Estimators,” *Econometrica*, 83, 2411–2451.

Altonji, J. and R. Matzkin (2005), “Cross Section and Panel Data Estimators for Nonseparable Models with Endogenous Regressors,” *Econometrica*, 73, 1053-1102.

Altonji, J. and C. Pierret (2001), “Employer Learning and Statistical Discrimination,” *Quarterly Journal of Economics*, 116, 313-350.

Altonji, J., T. Conley, T. Elder, and C. Taber (2011),“Methods for Using Selection on Observed Variables to Address Selection on Unobserved Variables,” Yale University Department of Economics Working Paper.

Angrist, J., G. Imbens, and D. Rubin (1996), “Identification of Causal Effects Using Instrumental Variables,” (with Discussion), *Journal of the American Statistical Association*, 91, 444-455.

Angrist, J. and J. Pischke (2009). *Mostly Harmless Econometrics: An Empiricist’s Companion.* Princeton University Press.

Arcidiacono, P., P. Bayer, and A. Hizmo (2010), “Beyond Signaling and Human Capital: Education and the Revelation of Ability,” *American Economic Journal: Applied Economics*, 2, 76–104.

Battistin, E. and A. Chesher (2014), “Treatment Effect Estimation with Covariate Measurement Error,” *Journal of Econometrics*, 178, 707–715.

Blackburn, M. and D. Neumark (1992), “Unobserved Ability, Efficiency Wages, and Interindustry Wage Differentials,” *Quarterly Journal of Economics*, 107, 1421-1436.

Bollinger, C. (2003) “Measurement Error in Human Capital and the Black-White Wage Gap,” *Review of Economics and Statistics*, 85, 578-585.

Bontemps, C., T. Magnac, and E. Maurin (2012), “Set Identified Linear Models,” *Econometrica*, 80, 1129-1155.

Bracewell R. (1986). *The Fourier Transform and Its Applications.* McGraw-Hill, Inc.

Card, D. (1995), “Using Geographic Variation in College Proximity to Estimate the Return to Schooling,” In L.N. Christofides, E.K. Grant, and R. Swidinsky, editors, *Aspects of Labor Market Behaviour: Essays in Honour of John Vanderkamp.* Toronto: University of Toronto Press.

Card, D. (1999), “The Causal Effect of Education on Earnings,” in Ashenfelter, O. and Card, D. eds., *Handbook of Labor Economics*, vol. 3, Part A, Elsevier.

Card, D. (2001), “Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems,” *Econometrica*, 69, 1127–1160.

Carneiro, P. and J. Heckman (2002), “The Evidence on Credit Constraints in Post Secondary Schooling,” *The Economic Journal*, 112, 705-734.

Carneiro, P., J. Heckman, and D. Masterov (2005), “Understanding the Sources of Ethnic and Racial Wage Gaps and Their Implications for Policy.” In: Nelson, R and Nielsen, L, (eds.) *Handbook of Employment Discrimination Research: Rights and Realities.* Springer: Amsterdam, 99-136.

Cawley J., J. Heckman, and E. Vytlacil (2001), “Three Observations on Wages and Measured Cognitive Ability,” *Labour Economics*, 8, 419-442.

Chalak, K. (2012), “Identification without Exogeneity under Equiconfounding in Linear Recursive Structural Systems,” in X. Chen and N. Swanson (eds.), *Causality, Prediction, and Specification Analysis: Recent Advances and Future Directions - Essays in Honor of Halbert L. White, Jr.*, Springer, 27-55.

Chalak, K. (2017), “Instrumental Variables Methods with Heterogeneity and Mismeasured Instruments,” *Econometric Theory*, 33, 69-104.

Chernozhukov, V., R. Rigobon, and T. Stoker (2010), “Set Identification and Sensitivity Analysis with Tobin Regressors,” *Quantitative Economics*, 1, 255–277.

Conley, T., C. Hansen, and P. Rossi (2012), “Plausibly Exogenous,” *Review of Economics and Statistics*, 94, 260–272.

Cunha, F., J. Heckman, and S. Schennach (2010), “Estimating the Technology of Cognitive and Noncognitive Skill Formation,” *Econometrica*, 78, 883-931.

Dawid, A.P. (1979), “Conditional Independence in Statistical Theory” (with Discussion), *Journal of the Royal Statistical Society*, Series B, 41, 1-31.

Fryer, R. (2011), “Racial Inequality in the 21st Century: The Declining Significance of Discrimination.” In O. Ashenfelter and D. Card (eds.). *Handbook of Labor Economics*. Elsevier, 4B, 855–971.

Griliches, Z. and J. Mairesse (1998), “Production Functions: The Search for Identification.” In: Steinar Strøm (ed.) *Econometrics and Economic Theory in the 20th Century: The Ragnar Frisch Centennial Symposium*, Cambridge University Press, 169-203.

Halvorsen, R. and R. Palmquist (1980) “The Interpretation of Dummy Variables in Semilogarithmic Equations,” *American Economic Review*, 70, 474-475.

Heckman, J. and E. Vytlacil (2005), “Structural Equations, Treatment Effects, and Econometric Policy Evaluation,” *Econometrica*, 73, 669-738.

Heckman, J., H. Ichimura, and P. Todd (1998), “Matching as an Econometric Evaluation Estimator,” *Review of Economic Studies*, 65, 261-294.

Heckman, J., S. Urzua, and E. Vytlacil (2006), “Understanding Instrumental Variables in Models with Essential Heterogeneity,” *Review of Economics and Statistics*, 88, 389–432.

Hoderlein, S. and E. Mammen (2007), “Identification of Marginal Effects in Nonseparable Models without Monotonicity,” *Econometrica*, 75, 1513-1518.

Hu, Y., J. Shiu, and T. Woutersen (2015), “Identification and Estimation of Single Index Models with Measurement Error and Endogeneity,” *Econometrics Journal*, 18, 347-362.

Hu, Y., J. Shiu, and T. Woutersen (2016), “Identification in Nonseparable Models with Measurement Error and Endogeneity,” *Economic Letters*, 144, 33-36.

Imbens, G. (2003), “Sensitivity to Exogeneity Assumptions in Program Evaluation,” *The American Economic Review*, 93, 126-132.

Imbens, G. and J. Angrist (1994), “Identification and Estimation of Local Average Treatment Effects,” *Econometrica*, 62, 467-476.

Imbens, G. and C. Manski (2004), “Confidence Intervals for Partially Identified Parameters,” *Econometrica*, 72, 1845–1857.

Imbens, G. and W. Newey (2009), “Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity,” *Econometrica*, 77, 1481-1512.

Klepper, S. and E. Leamer (1984), “Consistent Sets of Estimates for Regressions with Errors in All Variables,” *Econometrica*, 52, 163-184.

Klein, R. and F. Vella (2009) “A Semiparametric Model for Binary Response and Continuous outcomes under Index Heteroscedasticity,” *Journal of Applied Econometrics*, 24, 735–762.

Klein, R. and F. Vella (2010) “Estimating a Class of Triangular Simultaneous Equations Models without Exclusion Restrictions,” *Journal of Econometrics*, 154, 154-164.

Lang, K. and M. Manove (2011), “Education and Labor Market Discrimination,” *American Economic Review*, 101, 1467–1496.

Leamer, E. (1983), “Let’s Take the Con out of Econometrics,” *American Economic Review*, 73, 31-43.

Levinsohn, J. and A. Petrin (2003), “Estimating Production Functions Using Inputs to Control for Unobservables,” *Review of Economic Studies*, 70, 317–341.

Lewbel, A. (2012), “Using Heteroscedasticity to Identify and Estimate Mismeasured and Endogenous Regressor Models,” *Journal of Business and Economic Statistics*, 30, 67-80.

Manski, C. (1990), “Nonparametric Bounds on Treatment Effects,” *American Economic Review Papers and Proceedings*, 80, 319-323.

Manski, C. and J. Pepper (2000), “Monotone Instrumental Variables: With an Application to the Returns to Schooling,” *Econometrica*, 68, 997-1010.

Manski, C. and J. Pepper (2009), “More on Monotone Instrumental Variables,” *Econometrics Journal*, 12, S200–S216.

Mincer, J., (1974). *Schooling, Experience, and Earning.* New York: National Bureau of Economic Research.

Neal, D. and W. Johnson (1996), “The Role of Premarket Factors in Black-White Wage Differences,” *Journal of Political Economy*, 104, 869-895.

Nevo, A. and A. Rosen (2012), “Identification With Imperfect Instruments,” *Review of Economics and Statistics*, 94, 659–671.

Ogburna, E. and T. VanderWeele (2012), “On the Nondifferential Misclassification of a Binary Confounder,” *Epidemiology*, 23, 433–439.

Okumura T. and E. Usui (2014), “Concave-Monotone Treatment Response and Monotone Treatment Selection: With an Application to the Returns to Schooling,” *Quantitative Economics*, 5, 175–194.

Olley, G. and A. Pakes (1996), “The Dynamics of Productivity in the Telecommunications Equipment Industry,” *Econometrica*, 64, 1263-1297.

Reinhold, S. and T. Woutersen, (2009), “Endogeneity and Imperfect Instruments: Estimating Bounds for the Effect of Early Childbearing on High School Completion,” University of Arizona Department of Economics Working Paper.

Schennach, S., H. White, and K. Chalak (2012), “Local Indirect Least Squares and Average Marginal Effects in Nonseparable Structural Systems,” *Journal of Econometrics*, 166, 282-302.

Stock, J. and M. Watson (2010). *Introduction to Econometrics.* Addison-Wesley, 3rd Edition

Stoye, J. (2009), “More on Confidence Intervals for Partially Identified Parameters,” *Econometrica*, 77, 1299–1315.

Vytlacil, E. (2002), “Independence, Monotonicity, and Latent Index Models: An Equivalence Result,” *Econometrica*, 70, 331-341.

Wald, A. (1940), “The Fitting of Straight Lines if Both Variables Are Subject to Error,” *Annals of Mathematical Statistics*, 11, 284-300.

White, H. (1980), “A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity,” *Econometrica*, 48, 817-838.

White, H. (2001). *Asymptotic Theory for Econometricians.* New York: Academic Press.

White, H. and K. Chalak (2013), “Identification and Identification Failure for Treatment Effects using Structural Systems,” *Econometric Reviews*, 32, 273-317.

Wickens, M. (1972), “A Note on the Use of Proxy Variables,” *Econometrica*, 40, 759-761.

Wooldridge, J. (2012). *Introductory Econometrics: A Modern Approach.* South-Western College Publishing, 5th Edition.