Identification of Average Effects under Magnitude and Sign Restrictions on Confounding

Karim Chalak*†
University of Virginia

July 19, 2015

Abstract

This paper studies measuring the average effects of X on Y in structural systems without requiring (conditional) exogeneity of causes, treatment, or instruments. We study the identification of covariate-conditioned average random coefficients, average nonparametric discrete and marginal effects, local and marginal treatment effects as well as average treatment effects for the population, treated, and untreated. First, we characterize the omitted variable bias, due to confounders U, of regression and instrumental variables methods for the identification of these average effects, thereby generalizing the classic linear regression omitted variable bias representation. Then, we impose magnitude and sign restrictions on confounding to fully or partially identify these average effects. In particular, using proxies W for U, we ask how do the average direct effects of U on Y compare in magnitude and sign to those of U on W. Exogeneity and proportional confounding are examples of limit cases yielding full identification. Alternatively, the effects of X on Y are partially identified in sharp bounded intervals if W is sufficiently sensitive to U, and sharp upper or lower bounds may obtain otherwise. After studying estimation and inference, we apply this method to study the financial return to education and the black-white wage gap.

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

*Department of Economics, University of Virginia, P.O. Box 400182, Charlottesville, VA 22904-4182. Email: chalak@virginia.edu.

†Acknowledgments: I thank the participants in the Northwestern Junior Festival of Recent Developments in Microeconometrics, Harvard Causal Inference Seminar, 2012 California Econometrics Conference, 2013 BU-BC Green Line Econometrics Conference, 2013 North American Winter Meeting of the Econometric Society, New York Camp Econometrics VIII, 23rd annual meeting of the Midwest Econometrics Group, 9th Greater New York Metropolitan Area Econometrics Colloquium, Cowles Foundation Conference on Econometrics, and the seminars at BC, the Federal Reserve Bank of Cleveland, UCSD, UCLA, USC, University of Pittsburgh, IUPUI, University of Wisconsin-Milwaukee, Oxford, Royal Holloway University of London, University of Leicester, UVA, University of Montreal, and Georgetown as well as Kate Antonovics, Andrew Beauchamp, Stéphane Bonhomme, Federico Ciliberto, Donald Cox, Julie Cullen, Stefan Hoderlein, Arthur Lewbel, Matthew Masten, and Elie Tamer for helpful comments. I thank Rossella Calvi, Daniel Kim, and Tao Yang for excellent research assistance. Any errors are the author's responsibility.

1 Introduction

This paper studies identifying and estimating average causal effects in structural systems with omitted variables (unobserved confounders) without requiring conditional exogeneity of causes, treatment, or instruments given covariates. We study the identification of various average effects including covariate-conditioned average random coefficients, average nonparametric discrete and marginal effects, local and marginal treatment effects as well as average treatment effects for the population, treated, and untreated. In the case of a linear homogenous effect, the linear regression omitted variable bias 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 omitted variable bias formula "is one of the most important things to know about regression." First, this paper generalizes the classic linear regression omitted variable bias representation by characterizing the omitted variable bias of regression and instrumental variables (IV) methods for the identification of the parametric and nonparametric average effects described above. This enables reasoning about the direction of the omitted variable bias of IV and regression methods, including in nonparametric nonseparable structures. Then, using proxies for the omitted variables, we demonstrate how restrictions on the magnitude and sign of confounding yield full (point) or partial identification of these various average effects. As we demonstrate, restrictions on confounding are a weakening of common identifying assumptions employed in the literature such as exogeneity or perfect proxies. On the one hand, restrictions on confounding can thus be useful for identification when stronger assumptions are suspected to fail. On the other hand, restrictions on confounding can be used to study the sensitivity of estimates to deviations from stronger identifying assumptions.

The paper is organized as follows. Section 2 discusses the results of the paper and connects these to the literature. Section 3 introduces the data generation assumptions. We formally derive the omitted variable bias representations and sharp identification regions for the average effects of X on Y under magnitude and sign restrictions on confounding in Section 4 for random coefficients, in Section 5 for nonparametric discrete and marginal effects, and in Section 6 for local and marginal treatment effects as well as average treatment effects for the population, treated, and untreated. Section 7 studies estimation and inference. Section 8 applies these results to study the return to education and the black-white wage gap. Section 9 concludes. Mathematical proofs are gathered in Appendix A. Online Appendix B[1] contains additional extensions.

[1] Appendix B is available at http://people.virginia.edu/~kmc2yf/SignMagAppB.pdf

2 Motivation, Outline, and Discussion

This section discusses restrictions on confounding and provides an overview of this paper's results that motivates the formal results derived in subsequent sections.

2.1 Linear Equations with Homogenous Effects

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

$$Y = \alpha_Y + X'\bar{\beta} + U\bar{\delta}_Y, \tag{1}$$

where Y denotes the logarithm of hourly wage, X denotes observed determinants of wage including years of education, and the scalar U, commonly referred to as "ability" in the literature, denotes unobserved skill. While both X and U are potential structural determinants (causes) of Y, realizations of Y and X are observed by the econometrician whereas those of U are not. As demonstrated below, we emphasize that this paper's approach does not require a linear or parametric specification. Nevertheless, in order to introduce the main ideas in their simplest form, we let U be scalar and consider constant slope coefficients $\bar{\beta}$ and $\bar{\delta}_Y$ for now but allow for a random intercept α_Y which may be correlated with X. We also leave implicit conditioning on covariates. Our object of interest here is $\bar{\beta}$, the vector of (average) direct effects of the elements of X on Y, e.g. average financial return to education. Because U is freely associated with X and may cause Y (e.g. education choices and wage may depend on ability), we say that U is an unobserved "confounder" and X is "endogenous." The researcher observes realizations of a vector Z of potential instruments that are uncorrelated with α_Y but possibly freely correlated with ability U and, therefore, invalid. This allows for the possibility that a potential instrument for education, e.g. proximity to a college, may be correlated with ability U, e.g. due to unobserved parental characteristics or choices. We let Z and X have the same dimension; in particular, Z may equal X. Define $\tilde{Z} \equiv Z - E(Z)$ and let $E(\tilde{Z}X')$ be nonsingular. The classic (IV) regression omitted variable bias (or inconsistency) B in recovering $\bar{\beta}$ is given by

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

and this expression may be useful to reason about the direction of the omitted variable bias. Suppose that the researcher observes realizations of a proxy W for U that is possibly error-laden and given by

$$W = \alpha_W + U\bar{\delta}_W, \tag{2}$$

where, for now, we consider a constant slope coefficient $\bar{\delta}_W$ and random intercept α_W which may be correlated with α_Y, U, and X. 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 slope 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 wage and test score with respect to unobserved ability[2] (i.e. $100\bar{\delta}_Y\%$ and $100\bar{\delta}_W\%$ are the average approximate percentage changes in wage and test score respectively due directly to a unit or percentile increase in U). Alternatively, as we discuss shortly in the nonparametric case, one could consider quantile changes in wage and test score due to a quantile change in ability. It is common in applied work to condition on proxies, such as test scores, to control for endogeneity. Provided $\bar{\delta}_W \neq 0$, substituting for U in equation (1) gives

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

If α_W is degenerate, e.g. $\alpha_W = 0$, then W is a perfect (one to one) proxy for U and, provided $Cov(\alpha_Y, (Z, U)') = 0$, an IV regression of Y on $(1, X', W')'$ using instruments $(1, Z', W')'$ (recall Z may equal X) may identify $\bar{\beta}$ and $\frac{\bar{\delta}_Y}{\bar{\delta}_W}$. However, this result does not generally hold when α_W is nondegenerate since the possible conditional correlation[3] between Z (or X) and α_W given W leads to (IV) regression bias. Instead of α_W being degenerate, consider the weaker restriction $Cov(\alpha_W, Z) = 0$. Then the (IV) omitted variable bias is given by

$$B = E(\tilde{Z}X')^{-1}E(\tilde{Z}U)\bar{\delta}_Y = E(\tilde{Z}X')^{-1}E(\tilde{Z}W)\frac{\bar{\delta}_Y}{\bar{\delta}_W}$$

and $\bar{\beta}$ is therefore characterized by:

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

This expression for $\bar{\beta}$ involves two linear IV regression estimands $E(\tilde{Z}X')^{-1}E(\tilde{Z}Y)$ and $E(\tilde{Z}X')^{-1}$ $E(\tilde{Z}W)$. It also involves the unknown $\frac{\bar{\delta}_Y}{\bar{\delta}_W}$ denoting the ratio of the (average) direct effect of U on Y to that of U on W. Importantly, the IV regression omitted variable bias $E(\tilde{Z}X')^{-1}E(\tilde{Z}W)\frac{\bar{\delta}_Y}{\bar{\delta}_W}$ in measuring $\bar{\beta}$ is known up to this ratio. As we show, similar expressions for the average effects of X on Y obtain, under suitable assumptions, in the cases of random slope coefficients and nonparametric effects.

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

[3]From $W = \alpha_W + U\bar{\delta}_W$, we have that α_W is generally correlated with U given W. Since Z (or X) and U are freely correlated, it follows that α_W is generally correlated with Z (or X) given W.

2.2 Magnitude and Sign Restrictions on Confounding

We ask the following questions:

1. How does the average direct effect of U on Y compare in magnitude to that of U on W?
2. How does the average direct effect of U on Y compare in sign to that of U on W?

The answers to these questions impose restrictions on the magnitude and sign of confounding which fully or partially identify the (IV) regression bias and thus the average effects of X on Y. Section 2.4 discusses criteria that a researcher may consider to guide his or her answers to these questions in empirical applications. However, this paper does not require particular answers to these questions. Instead, it characterizes the mapping[4] from every possible answer to the corresponding identification region for the average effect of X on Y. To keep manageable the scope of this paper's study of the identification of average parametric and nonparametric effects, we focus here on deterministic answers to these questions, such as interval restrictions on $\frac{\bar{\delta}_Y}{\delta_W}$, rather than probabilistic answers (prior distributions). In particular, exogeneity is a limiting special case, which can obtain if the average direct effect $\bar{\delta}_Y$ of U on Y is zero, yielding full (point) identification. Proportional confounding is another limiting case, in which the average direct effect of U on Y equals a known proportion of that of U on W, also yielding full identification. Alternatively, weaker restrictions on how the average direct effect of U on Y compares in magnitude and/or sign to that of U on W partially identify elements of $\bar{\beta}$, yielding sharp bounded intervals when the proxy W is sufficiently sensitive to the confounder U, and sharp lower or upper bounds otherwise.

2.3 Nonparametric Nonseparable Equations

An advantage of this paper's approach is that it does not require a linear or parametric specification; Sections 5 and 6 give key general results.

2.3.1 Regression and Identification of Nonparametric Effects

Section 5 studies the structural equation:

$$Y = r(X, S, U, U_Y), \tag{4}$$

where the vectors of unobservables U_Y and the scalar confounder U interact nonseparably with X and a vector of observed covariates S to drive Y according to the unknown nonparametric structural function r. We assume[5] that $U_Y \perp (U, X)|S = s$ so that, unlike U_Y, the omitted

[4] Leamer (1983) suggests the slogan "the mapping is the message."

[5] Throughout, we use $A \perp B|S$ to denote conditional independence as in Dawid (1979). Further, we write $A \perp B|S = s$ to denote conditional independence at $S = s$.

variable U may statistically depend on X given the covariates (or subpopulation) S. The requirement that $U_Y \perp (U, X)|S$ is a weakening of the common assumption of conditional exogeneity $U_Y \perp X|S$ which would obtain if either (a) U is observed and thus could be subsumed into S or (b) $(U_Y, U) \perp X|S$ and thus U could be subsumed into U_Y. Thus, both U and U_Y generate heterogeneity in the response of Y to X but U is the only source of endogeneity of X (or of "essential heterogeneity" in the nomenclature of Heckman, Urzua, and Vytlacil (2006)). We study the identification of conditional average discrete and marginal effects of X on Y given covariates S, such as $E[\frac{\partial}{\partial x} r(x, s, U, U_Y)|X = x, S = s]$, without imposing exogeneity. As we show, these effects correspond to various average effects studied in the literature in special cases such as exogeneity. First, we characterize the omitted variable bias of nonparametric regression methods for recovering these average effects of X on Y thereby generalizing the classic linear regression omitted variable representation (discussed above when $Z = X$) to the nonparametric nonseparable case. For example, for scalar X and U and leaving covariates S implicit, Theorem 5.2 shows that the omitted variable bias $B(x)$ of the nonparametric regression $\frac{\partial}{\partial x} E(Y|X = x)$ in recovering the average marginal effect $E[\frac{\partial}{\partial x} r(x, U, U_Y)|X = x]$ is given by[6]:

$$B(x) \equiv \frac{\partial}{\partial x} E(Y|X = x) - E[\frac{\partial}{\partial x} r(x, U, U_Y)|X = x] = -\int_{\mathcal{U}_x} E[\frac{\partial}{\partial u} r(x, u, U_Y)] \ \frac{\partial}{\partial x} F_{U|X}(u|x) du.$$

Thus, the bias $B(x)$ is a weighted average of the average marginal effects $E[\frac{\partial}{\partial u} r(x, u, U_Y)]$ of U on Y at $X = x$, with "weights" $\frac{\partial}{\partial x} F_{U|X}(u|x)$. As discussed in Section 5, this characterization enables reasoning about the direction of the omitted variable bias in the general nonparametric nonseparable case, for example when $E[\frac{\partial}{\partial u} r(x, \cdot, U_Y)]$ and $\frac{\partial}{\partial x} F_{U|X}(\cdot|x)$ do not change sign for all u. $B(x)$ generalizes the standard expression $B = E(\tilde{X}X)^{-1} E(\tilde{X}U)\bar{\delta}_Y$ in the linear case.

Suppose that there is a proxy W for U generated by the equation

$$W = q(S, U, U_W), \tag{5}$$

where the unobserved vector U_W and U interact nonseparably with the covariates S to drive the proxy W according to the unknown nonparametric function q. Analogously to U_Y, we assume that $U_W \perp (U, X)|S = s$. Then, using the equation for W and leaving covariates implicit for simplicity, Theorem 5.2 shows that

$$\frac{\partial}{\partial x} E(W|X = x) = -\int_{\mathcal{U}_x} E[\frac{\partial}{\partial u} q(u, U_W)] \frac{\partial}{\partial x} F_{U|X}(u|x) du.$$

Similar to the linear case, Corollary 5.3 gives conditions under which contrasting the average effect $E[\frac{\partial}{\partial u} r(x, u, U_Y)]$ of U on Y to the average effect $E[\frac{\partial}{\partial u} q(u, U_W)]$ of U on W enables bounding

[6]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 $\mathcal{A}_b$.

the regression bias and therefore yields full or partial identification of $E[\frac{\partial}{\partial x}r(x,U,U_Y)|X=x]$. For instance, for continuous response, proxy, and confounder, one can apply a probability transformation to rewrite the response and proxy equations such that U, Y, and W have a standard uniform distribution[7]. Then, $E[\frac{\partial}{\partial u}r(x,u,U_Y)]$ and $E[\frac{\partial}{\partial u}q(u,U_W)]$ denote quantile changes in the response and proxy due to a quantile change in the confounder (e.g. a percentile increase in ability leads on average to a larger percentile increase in test score than in wage). Contrasting the sign and magnitude of these average effects may enable bounding $E[\frac{\partial}{\partial x}r(x,U,U_Y)|X=x]$. Such distributional (rank) comparisons may be convenient in economic contexts where Y and W are measured on different scales.

Intermediate cases between the linear and nonparametric nonseparable cases encompass leading specifications in the literature. For instance, consider the additively separable case:

$$Y = \ddot{r}(X,S,U_Y) + U'\delta_Y \quad \text{and} \quad W' = \alpha_W' + U'\delta_W, \tag{6}$$

where δ_Y is a vector of random coefficients that can depend on $(S', U_Y')'$ and where α_W is a vector, and δ_W a matrix, of random coefficients that can depend on $(S', U_W')'$. When $\delta_Y = 0$ and $U_Y \perp X|S$, we obtain the specification for the Y equation studied in e.g. Altonji and Matzkin (2005), Hoderlein and Mammen (2007), and Imbens and Newey (2009), yielding full identification of various average effects of X on Y. Section 5 studies the full and partial identification of conditional average effects of X on Y in systems with $U_Y \perp X|S=s$ but where U may depend on X given S, first in the additively separable case in equations (6) with the random vector δ_Y possibly nonzero, and second when the effect of U on Y in the nonseparable equations (4, 5) is possibly nonzero. In section 4, we focus on the linear special case in which $\ddot{r}(X,S,U_Y) = \ddot{r}_0(S,U_Y) + \sum_{j=1}^k X_j \ddot{r}_j(S,U_Y) \equiv \alpha_Y + X'\beta$, with α_Y and β denoting random intercept and slope coefficients. In this case, we study the identification of conditional averages of β under magnitude and sign restrictions on confounding, while allowing for X and potential instruments Z to be freely (conditionally) correlated with U and for δ_Y to be nonzero. Online Appendix B contains additional extensions in the linear random coefficients case to allow for a panel structure and to study proxies included in the Y equation.

2.3.2 IV Methods and Identification of Local and Marginal Effects

Section 6 studies treatment effects and augments equations (4, 5) and (6) from Section 5 with a threshold crossing equation generating the binary treatment X:

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

[7] If $Y^* = r^*(X,S,U^*,U_Y)$ where Y^* and U^* are any continuous random variables then $Y \equiv F_{Y^*}(Y^*)$ and $U \equiv F_{U^*}(U^*)$ have a standard uniform distribution and we have $F_{Y^*}(Y^*) = F_{Y^*}(r^*(X,S,F_{U^*}^{-1}(F_{U^*}(U^*)),U_Y))$. In particular, one can rewrite the response equation as $Y = r(X,S,U,U_Y)$. Similarly, if W^* is also continuous and $W^* = q^*(S,U^*,U_W)$ then a similar argument gives $W = q(S,U,U_W)$ where $W \equiv F_{W^*}(W^*)$.

where U_X is an unobserved variable and the function ν is unknown. For example, when $\delta_Y = 0$ in the separable equations (6) and $(U_X, U_Y) \perp Z|S$, we obtain the specification for the X and Y equations studied e.g. in Imbens and Angrist (1994) and Heckman and Vytlacil (2005). First, Theorem 6.2 characterizes the omitted variable bias of the Wald and local IV estimands for recovering conditional local and marginal treatment effects when $(U_X, U_Y) \perp Z|S = s$ but U may depend on Z given S in both the separable case in equations (6) with δ_Y possibly nonzero and the nonseparable case in equations (4, 5) when the effect of U on Y may be nonzero. Second, we study the full and partial identification of these local and marginal effects as well as the identification of conditional average treatment effects for the population, treated, and untreated. We refer the reader to Section 6 for details.

2.4 Magnitude and Sign Restrictions: Discussion and Connections to the Literature

What information may a researcher employ in practice when postulating answers to the above questions regarding magnitude and sign restrictions on confounding? In what follows, we discuss how economic theory and evidence may provide guidance to answering these questions. Then, we discuss stronger or alternative assumptions that lead to full or partial identification and give examples from the literature. On the one hand, we demonstrate how magnitude and sign restrictions weaken common identifying assumptions, thereby providing a means for identification when these assumptions fail. Moreover, even when stronger or alternative assumptions hold, valid restrictions on confounding may yield tighter identification regions and/or confidence intervals. On the other hand, we discuss how this paper's framework can be used to conduct a sensitivity analysis whereby a researcher examines the sensitivity of a study's empirical conclusions to deviations from identifying assumptions. For instance, a researcher may gain confidence in estimates of $\bar{\beta}$ that are not highly sensitive to deviations of $\frac{\bar{\delta}_Y}{\bar{\delta}_W}$ either from maintained point identifying assumptions such as exogeneity ($\frac{\bar{\delta}_Y}{\bar{\delta}_W} = 0$) or from estimates of $\frac{\bar{\delta}_Y}{\bar{\delta}_W}$ obtained under point identifying assumptions such as perfect proxies. Conversely, a researcher may ask: what restrictions on $\frac{\bar{\delta}_Y}{\bar{\delta}_W}$ are in accord with economic or qualitative features of elements of $\bar{\beta}$ or with estimates of $\bar{\beta}$ obtained under point identifying assumptions? We illustrate our discussion in the context of this paper's empirical application studying the financial return to education and the black-white wage gap as well as in the context of production function estimation.

2.4.1 Economic Theory and Evidence

Sometimes, economic theory and evidence can help shed light on sign and magnitude restrictions. For example, in the case of the earning equation, it may be reasonable to assume that, given the observables, wage is on average less elastic or sensitive to unobserved ability than

the test score is, i.e. $\left|\bar{\delta}_Y\right| \leq \left|\bar{\delta}_W\right|$. Specifically, given the observed characteristics, a change in U may, on average, directly cause a higher percentage change in the test score than in wage. Moreover, we sometimes further assume that ability, on average, directly affects wage and the test score 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 effects 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. Further, the empirical findings in this paper corroborate the assumption $\left|\bar{\delta}_Y\right| \leq \left|\bar{\delta}_W\right|$ since allowing $\left|\bar{\delta}_Y\right|$ to be larger than $\left|\bar{\delta}_W\right|$ often extends the estimated identification regions to include a negative average return to education and a black-white wage gap in favor of blacks, which is inconsistent with the general findings in the literature.

2.4.2 Weakening of Exogeneity

Magnitude and sign restrictions such as $\left|\bar{\delta}_Y\right| \leq \left|\bar{\delta}_W\right|$, $0 \leq \frac{\bar{\delta}_Y}{\bar{\delta}_W}$, or $0 \leq \frac{\bar{\delta}_Y}{\bar{\delta}_W} \leq 1$, are a weakening of the commonly employed assumption of exogeneity. For instance, in the linear equations (1, 2), exogeneity holds if $Cov(U'\bar{\delta}_Y + \alpha_Y, Z) = 0$. By definition, α_Y denotes the vector of unobservables that drive Y and are thought to be uncorrelated with Z (conditional on covariates), and we absorb into U the unobserved drivers of Y that are possibly freely (conditionally) correlated with Z. Exogeneity then holds if $Cov(\alpha_Y, Z) = 0$ and $\bar{\delta}_Y = 0$. We allow but do not require $\bar{\delta}_Y = 0$; instead, it suffices that $Cov((\alpha_Y, \alpha_W), Z) = 0$. Thus, this paper's method provides a practical alternative to IV methods when potential instruments may be weak or (conditionally) endogenous. Moreover, $\bar{\beta}$ is "under-identified" in equation (3) since Z and X have the same dimension and there are fewer exogenous instruments for $(X', W')'$ than is needed for full identification. Similar difficulties arise in more general nonlinear cases.

Manski and Pepper (2000) employ alternative assumptions to bound nonparametric average effects when exogeneity may fail. In particular, they assume known bounds on the range of Y and that $E[r(x, s, U, U_Y)|Z = z, S = s]$ is monotonic in z. They also consider the assumption that r is monotonic in x. Okumura and Usui (2014) combine the last two assumptions for $Z = X$ along with the assumption that r is concave in x. Also, several recent papers employ alternative assumptions to partially identify linear or parametric effects of endogenous variables. For example, Altonji, Conley, Elder, and Taber (2011) assume that the selection on unobservables

occurs similarly to that on observables. Also, Reinhold and Woutersen (2009) and Nevo and Rosen (2012) assume that the correlation between the potential instrument and U on the one hand and that between the endogenous variable and U on the other hand have the same sign and then further assume that the potential instrument is less correlated with U than the endogenous variable is. Conley, Hansen, and Rossi (2012) weaken the instrument "exclusion restriction" by allowing Z to enter the linear equation for Y and employing prior assumptions on the coefficient on Z that are weaker than requiring it to be zero. Klein and Vella (2009, 2010) and Lewbel (2012) impose restrictions on heteroskedasticity. Bontemps, Magnac, and Maurin (2012) provide additional examples and a general treatment of set identified linear models. We do not require the assumptions imposed in these papers. Instead, we employ proxies to identify average effects under magnitude and sign restrictions on confounding. Also, this paper's method does not require linearity and applies in the nonparametric nonseparable case. Of course, which (combination of) identifying assumption(s) is appropriate depends on the context.

2.4.3 Perfect Proxies

This paper's method provides a simple alternative to the common practice which assumes that conditioning on proxies for confounders ensures conditional exogeneity. As discussed above for equations (1, 2) and (3), the coefficient on X from a linear regression of Y on $(1, X', W')'$ need not identify $\bar{\beta}$. Indeed, more generally, conditioning on W may, but need not, attenuate the regression bias (see e.g. Wickens, 1972; Battistin and Chesher, 2014; and Ogburna and Vander-Weele, 2012). However, this regression consistently estimates $(\bar{\beta}', \frac{\bar{\delta}_Y}{\bar{\delta}_W})'$ if α_W is degenerate, so that the proxy W is a perfect rescaling of U, and $Cov(\alpha_Y, (X', U)') = 0$. Instead of assuming degenerate α_W, this paper shows that $Cov((\alpha_Y, \alpha_W)', X) = 0$ along with restrictions on $\frac{\bar{\delta}_Y}{\bar{\delta}_W}$ suffice to bound $\bar{\beta}$. Further, these restrictions may include the "prefect proxy" estimate of $\frac{\bar{\delta}_Y}{\bar{\delta}_W}$ and its e.g. 95% confidence interval.

More generally, conditioning on a "perfect proxy" W that is a one to one mapping of U may enable recovering the effect of X on Y. For example, this occurs when $W = q(S, U)$, so that U_W is degenerate, and q is strictly monotonic in U given S as in Olley and Pakes (1996) (see also Griliches and Mairesse, 1998; Levinsohn and Petrin, 2003). We do not require these assumptions and allow for an imperfect error-laden proxy $W = q(S, U, U_W)$ with U_W random and q possibly non-monotonic in u. Rather than conditioning on perfect proxies, this paper employs possibly error-laden proxies to bound nonparametric average effects (see Section 5).

2.4.4 Classical Measurement Error

In the linear case, additional assumptions may generate useful bounds on $\frac{\bar{\delta}_Y}{\bar{\delta}_W}$ and thus $\bar{\beta}$. See for example Klepper and Leamer (1984), Leamer (1987), and Bollinger (2003) who study linear systems where the measurement errors α_W are classical, i.e. uncorrelated with the explanatory variables $(X', U')'$ and with[8] α_Y (as well as sometimes among each other $Cov(\alpha_{W_h}, \alpha_{W_{h'}}) = 0$). We do not impose these assumptions in the linear case (e.g. we do not require $Cov(\alpha_W, U) = 0$ or $Cov(\alpha_W, \alpha_Y) = 0$) and, importantly, our analysis does not require linearity.

2.4.5 Multiple Proxies

Another means for full identification is to impose restrictions involving the components of α_W (or U_W) and U ensuring the availability of multiple proxies for U that are otherwise unrelated. For example, let scalars U, W_1, and W_2 be such that

$$Y = \alpha_Y + X'\bar{\beta} + U\bar{\delta}_Y, \quad W_1 = \alpha_{W_1} + U\bar{\delta}_{W_1}, \quad \text{and } W_2 = \alpha_{W_2} + U\bar{\delta}_{W_2},$$

with $\bar{\delta}_{W_1}, \bar{\delta}_{W_2} \neq 0$, so that there are two proxies for U, and assume $Cov[(Z', U, \alpha_{W_2})', (\alpha_Y, \alpha_{W_1})'] = 0$. Then $Y = \alpha_Y - \alpha_{W_1}\frac{\bar{\delta}_Y}{\bar{\delta}_{W_1}} + X'\bar{\beta} + W_1\frac{\bar{\delta}_Y}{\bar{\delta}_{W_1}}$ and $(\bar{\beta}', \frac{\bar{\delta}_Y}{\bar{\delta}_{W_1}})'$ may be fully identified from an IV regression of Y on $(1, X', W_1)'$ using instruments $(1, Z', W_2)'$ (see e.g. Blackburn and Neumark, 1992). We do not require such restrictions here, allowing for example, for components of α_W (e.g. test taking skills) to be correlated. (Online Appendix B studies the case of multiple proxies for U that are components of X.) More generally, in nonparametric systems, the availability of multiple proxies for U (e.g. repeated measurements) that are mutually independent given U (see e.g. Cunha, Heckman, and Schennach, 2010) could be useful in identifying the effects of X on Y. This paper does not require multiple proxies for U or a particular specification for the proxy equations. Further, when multiple proxies are available, we do not require the proxies to be mutually independent (or even uncorrelated) given U.

2.5 Empirical Application and Additional Examples

We illustrate our discussion of magnitude and sign restrictions in the contexts of this paper's empirical application as well as production function estimation.

[8] When U is scalar, the resulting 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 is not a scalar, in order for nontrivial bounds to exist, these methods require that all coefficients in regressions of any element of $(Y, W')'$ on the remaining elements and X have the same sign.

2.5.1 Empirical Application: Return to Education and Black-White Wage Gap

Section 8 applies this paper's method to study the return to education and the black-white wage gap using the data in Card (1995). We employ restrictions on confounding ($0 \leq \frac{\bar{\delta}_Y}{\bar{\delta}_W} \leq 1$ and $\left|\frac{\bar{\delta}_Y}{\bar{\delta}_W}\right| \leq 1$) to partially identify in sharp bounded intervals the covariate-conditioned average financial incremental return to each year of education as well as the average black-white wage gap. As discussed above, these restrictions are in accord with several theoretical and empirical findings. Importantly, we do not require that instruments or regressors are conditionally exogenous. Generally, we find that regression estimates, which would be consistent under exogeneity ($\frac{\bar{\delta}_Y}{\bar{\delta}_W} = 0$), provide an upper bound on the average return to education (e.g. 19.5% for the return to the 16^{th} year) and black-white wage gap (-17.8%) and that the regression-based bounds estimates are generally narrower than the IV-based ones, with especially narrower confidence intervals (CI). If one assumes that the proxy W ($\log(KWW)$) is a perfect rescaling of U then $\frac{\bar{\delta}_Y}{\bar{\delta}_W}$ is estimated to be 0.203 with 95% CI $[0.141, 0.264]$. Requiring that W has classical measurement error (e.g. Bollinger, 2003) yields very large bounds in this case that are inconsistent with the literature (e.g. with large negative return to education and large black-white wage gap in favor of blacks). Allowing for (possibly nonclassical) error-laden proxies and imposing magnitude and sign restrictions on confounding ($0 \leq \frac{\bar{\delta}_Y}{\bar{\delta}_W} \leq 1$) that weaken exogeneity and include the "perfect proxy" 95% CI for $\frac{\bar{\delta}_Y}{\bar{\delta}_W}$, the regression-based estimated sharp identification region for the black-white wage gap is relatively wide, $[-17.8\%,\ 1.9\%]$ with a 95% CI $[-21\%,\ 5.4\%]$. Thus, under these weaker than exogeneity assumptions on confounding, 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. Further, we find evidence suggesting a nonlinearity in the return to education, with the 12^{th}, 16^{th}, and 18^{th} years, corresponding to obtaining a high school, college, and possibly a graduate degree, yielding a high average return. For example, under sign and magnitude restrictions on confounding, the estimated identification region for the average return to the 16^{th} year is $[13.33\%,\ 19.5\%]$ with 95% CI $[7.5\%,\ 25.1\%]$ whereas that for the 13^{th} year is $[0.7\%, 7.8\%]$ with 95% CI $[-3.4\%, 11.6\%]$. This nonlinearity may partly explain why, contrary to the expected direction of ability bias, linear IV estimates of the average return to education often exceed linear regression estimates. In particular, both types of estimates are weighted averages of yearly incremental returns for different subpopulations and the large IV estimates may reflect the relatively high return to graduation years for the subpopulation whose graduation outcomes depends on instruments such as proximity to college (see e.g. Card 1995, 1999).

2.5.2 Additional Examples: Production Function

This paper's framework can also be applied in several other contexts in which proxies for omitted variables are available. Examples include studies of student achievement, health outcomes, and production functions. To illustrate, consider the following production function and proxy equations for a plant as in Olley and Pakes (1996) and Griliches and Mairesse (1998):

$$Y = \alpha_Y + X\bar{\beta} + S'\bar{\gamma} + U\bar{\delta}_Y \quad \text{and} \quad W = q(S, U), \tag{7}$$

where[9] Y is log of output and X is log of labor input. We focus here on identification of $\bar{\beta}$, the elasticity of output with respect to labor. S denotes logarithm of capital input and age in Olley and Pakes (1996) and logarithms of physical capital input and R&D capital in Griliches and Mairesse (1998). U denotes unobserved "productivity"[10] that, in contrast to α_Y, may be correlated with $(X, S')'$. Investment W is assumed to be a perfect one-to-one proxy for U, with q invertible given S so that $U = q^{-1}(S, W)$. Substituting for U in the Y equation gives the partially linear equation

$$Y = \alpha_Y + X\bar{\beta} + S'\bar{\gamma} + q^{-1}(S, W)\bar{\delta}_Y,$$

that may identify $\bar{\beta}$ if $Cov(\alpha_Y, X|U, S) = 0$. For example, if $W = S'\bar{\theta} + U\bar{\delta}_W$ with constants $\bar{\theta}$ and $\bar{\delta}_W \neq 0$ then substituting for U gives

$$Y = \alpha_Y + X\bar{\beta} + S'(\bar{\gamma} - \bar{\theta}\frac{\bar{\delta}_Y}{\bar{\delta}_W}) + W\frac{\bar{\delta}_Y}{\bar{\delta}_W}$$

and a linear regression of Y on $(1, X, S', W)'$ may recover $(\bar{\beta}, \frac{\bar{\delta}_Y}{\bar{\delta}_W})$ if $Cov(\alpha_Y, (X, U, S')') = 0$.

Our framework enables relaxing the perfect proxy assumption. In particular, we allow for $W = q(S, U, U_W)$ with U_W nondegenerate and q non-monotonic in u. Instead, weaker assumptions, such as monotonicity of $E[q(s, u, U_W)|S = s]$ in u, suffice to partially identify $\bar{\beta}$ (see Corollary 5.3). For example, suppose that investment is an imperfect proxy $W = \alpha_W + S'\bar{\theta} + U\delta_W$ where α_W is a random intercept and $\delta_W = \bar{\delta}_W + \eta_W$ with constant $\bar{\delta}_W \neq 0$ and random η_W. This removes the assumption that W is a deterministic function of (U, S) and allows for failure of monotonicity of W in U. Then, as discussed in Section 4, restrictions on $\frac{\bar{\delta}_Y}{\bar{\delta}_W}$, which may include the "perfect proxy" estimate and 95% CI for $\frac{\bar{\delta}_Y}{\bar{\delta}_W}$, yield bounds on $\bar{\beta}$ if e.g. $Cov[(\alpha_Y, \alpha_W)', (X, S')'] = 0$ and $E(\eta_W|X, U, S) = 0$. To illustrate this, we estimate the production function in equations (7) for U.S. R&D performing firms using the setup and

[9] For simplicity, we leave the time subscript implicit here and we ignore self-selection e.g. due to plants closing (below, we briefly consider a sample on U.S. R&D performing firms used in Griliches and Mairesse (1998) who suggest that the selection problem may not be very severe in this data).

[10] Olley and Pakes (1996), refer to $\omega = U\bar{\delta}_Y$ as productivity.

data[11] used in Griliches and Mairesse (1998). Assuming exogeneity of X given S, the estimate of the labor elasticity $\bar{\beta}$ from a linear regression of Y on $(1, X, S')'$ is 0.578 with 95% CI (using robust standard errors) $[0.548, 0.609]$ (see e.g. Table 6.4 in Griliches and Mairesse, 1998). Alternatively, assuming that the logarithm of investment is a perfect proxy for U ($W = S'\bar{\gamma} + U\bar{\delta}_W$), the estimate for $\bar{\beta}$ is 0.551 with 95% CI $[0.521, 0.582]$ and that for $\frac{\bar{\delta}_Y}{\bar{\delta}_W}$ is 0.110 with 95% CI $[0.085, 0.134]$, suggesting that the elasticity of investment to productivity is larger than that of sales. Requiring the measurement error in W to be classical yields a large upper bound on $\frac{\bar{\delta}_Y}{\bar{\delta}_W}$ with implausible bounds on $\bar{\beta}$ (containing negative labor elasticity). Applying this paper's methods to allow for the logarithm of investment to be an imperfect proxy ($W = \alpha_W + S'\bar{\theta} + U\delta_W$) as described above and imposing the restrictions $0 \leq \frac{\bar{\delta}_Y}{\bar{\delta}_W} \leq 1$, which weakens exogeneity and includes the perfect proxy estimate for $\frac{\bar{\delta}_Y}{\bar{\delta}_W}$ and its 95% CI, yields the bounds $[0.331, 0.578]$ on $\bar{\beta}$ with 95% CI $[0.291, 0.604]$. This suggests that the linear regression estimates of the labor input coefficient may be biased upward (see e.g. Olley and Pakes, 1996). We leave a detailed study of estimation of production functions to other work; our goal here is simply to illustrate the scope of this paper's approach.

3 Data Generation

The next assumption defines the data generating process.

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

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

where r *and* q *are unknown real- and vector-valued measurable functions respectively and* $E(Y, W')' < \infty$. *Realizations of* M *are observed whereas those of* L *are not.*

S.1(i) defines the notation for observables. S.1(ii) imposes structure on the data generating process. It distinguishes between the observed (or measured) variables M and unobserved (or latent) variables L. The vectors of unobservables U_Y and of confounders U may interact nonseparably with the causes of interest X and covariates S to impact the response Y according to the nonparametric structural function r. We observe realizations of a vector W of proxies

[11] I thank Jacques Mairesse for permitting me to use this data set. This sample has 2971 observations and is an unbalanced panel on firms in years 1973, 1978, 1983, and 1988. Following Griliches and Mairesse (1998), we include in S year indicators and their interactions with an indicator for the computer industry.

for U. We sometimes allow for W and X to have common elements. Analogously to U_Y, the vector U_W interacts with U and S nonseparably to generate the proxies W according to the nonparametric function q. We allow but do not require the availability of covariates S; if these are absent, set $S = 1$. Further, we allow but do not require elements of S to directly affect Y or W; if r or q does not directly depend on S, these may nevertheless serve as conditioning variables. Last, we also observe realizations of a vector of potential instruments Z possibly equal to, or containing elements of, X. Importantly, unlike U_W and U_Y, U may statistically depend on Z or X given S, thereby creating difficulties for the identification of the effects of X on Y. Thus, elements of Z may but need not be valid instruments since these may be included in the Y equation and are freely (conditionally) correlated with U.

We are interested in measuring effects of X on Y averaged over (conditional) distributions of the unobservables U and U_Y. At the outset of Sections 4, 5, and 6 we discuss several conditional average effects of interest and how these correspond to average effects that have been studied in the literature in special cases, such as exogeneity, separability in U, or linearity. After studying the omitted variable bias of traditional estimands, we obtain full or partial identification of the average effects of X on Y by imposing magnitude and sign restrictions on the average effects of U on Y and W. Here, the causal effects of X on Y and those of U on Y and W are features of the structural system that can derive from economic theory and are encoded in the structural functions r and q whereas the observability of X or U is an empirical matter.

4 Identification of Average Random Coefficients

Although the paper's method does not require a linear or parametric effect of X on Y, we find it instructive to begin our identification analysis by studying linear structures with random coefficients. We relax the linearity assumption when studying the identification of conditional average nonparametric effects in Section 5 and of local, marginal, and average treatment effects in Section 6. Specifically, we impose the following assumption in Section 4.

Assumption 2 (S.2) *Linearity: Assume S.1 with* $Cov[Z, (Y, W')'] < \infty$ *and*

$$Y = r(X, S, U, U_Y) = \ddot{r}_0(S, U_Y) + \sum_{j=1}^{k} X_j \ddot{r}_j(S, U_Y) + \sum_{g=1}^{l} U_g \breve{r}_g(S, U_Y) \equiv \alpha_Y + X'\beta + U'\delta_Y,$$

and let the h^{th} *component* q_h *of* q *be given by*

$$W_h = q_h(U, U_W) = q_{h,0}(S, U_W) + \sum_{g=1}^{l} U_g q_{h,g}(S, U_W) \equiv \alpha_{W_h} + U'\delta_{W_h},$$

so that stacking W_h, $h = 1, ..., m$, *into* W *gives*

$$W' = \alpha'_W + U'\delta_W.$$

Thus, under S.2, for each observation i in a sample, we have

$$Y_i = \alpha_{Y,i} + X_i'\beta_i + U_i'\delta_{Y,i} \quad \text{and} \quad W_i' = \alpha'_{W,i} + U_i'\delta_{W,i}.$$

This allows for random effects for each individual or unit and encompasses constant slope coefficients as a special case. Section 4 studies the identification of $\bar{\beta}(S) \equiv E(\beta|S)$, the average direct effect of X on Y given covariates (or subpopulation) S, and employs magnitude and sign restrictions on the average direct effects of U on Y and W given S, $\bar{\delta}_Y(S) \equiv E(\delta_Y|S)$ and $\bar{\delta}_W(S) \equiv E(\delta_W|S)$.

4.1 IV Regression Notation

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

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

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

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

so that by construction $Cov(C, \epsilon_{A.B|C}|S = s) = 0$. In the special case where $B = C$, we obtain the conditional linear regression coefficients $R_{A.B}(s) \equiv R_{A.B|B}(s)$ and residuals $\epsilon'_{A.B}(s) \equiv \epsilon'_{A.B|B}(s)$. When S is degenerate, $S = 1$, we leave S implicit. For example, we write $\bar{A} \equiv E(A)$ and $\tilde{A} \equiv A - \bar{A}$ and denote by $R_{Y.X|Z} \equiv Cov(Z, X)^{-1}Cov(Z, Y)$ the vector of slope coefficients associated with X in a linear IV regression of Y on $(1, X')'$ using instruments $(1, Z')'$.

4.2 Characterization and Full Identification

We begin by characterizing $\bar{\beta}(s)$ and studying conditions for full identification. Section 4.3 studies partial identification of elements of $\bar{\beta}(s)$ under sign and magnitude restrictions on confounding. For illustration, consider the linear example from Section 2.1 with a scalar confounder U and a scalar proxy W, degenerate covariates $S = 1$, and constant slope coefficients:

$$Y = \alpha_Y + X'\bar{\beta} + U\bar{\delta}_Y \quad \text{and} \quad W = \alpha_W + U\bar{\delta}_W.$$

Using the IV regression succinct notation, recall that, provided $Cov(Z, X)$ is nonsingular, $\bar{\delta}_W \neq 0$, and $Cov(Z, (\alpha_Y, \alpha_W)') = 0$, we have that:

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

Thus, the IV regression (omitted variable) bias B depends on the ratio $\frac{\bar{\delta}_Y}{\bar{\delta}_W}$ of the (average) direct effect of U on the response Y to that of U on the proxy W. Theorem 4.1 extends this result to allow for vectors U and W, covariates S, and random slope coefficients. First, it derives the conditional IV regression bias $B(s)$. Then, it employs W in characterizing[12] $B(s)$.

Theorem 4.1 *Assume S.2 with $\ell = k$ and $m = l$. Let $s \in \mathcal{S}$ with $Cov[Z, (Y, W')'|S = s] < \infty$. (i) If (i.a) $Cov(Z, X|S = s)$ is nonsingular and (i.b) $Cov(\alpha_Y, Z|S = s) = 0$, $\beta \perp_m (Z, X)|S = s$, and $\delta_Y \perp_m (U, Z)|S = s$ then*

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

(ii) If, in addition to (i.a), (ii.a) $\bar{\delta}_W(s)$ is nonsingular with $\bar{\delta}(s) \equiv \bar{\delta}_W^{-1}(s)\bar{\delta}_Y(s)$ and (ii.b) $Cov(\alpha_W, Z|S = s) = 0$ and $\delta_W \perp_m (U, Z)|S = s$ then

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

Using $B(s) = R_{U.X|Z}(s)\bar{\delta}_Y(s)$, a researcher can reason about the direction of the omitted variable bias under assumptions on the signs of $R_{U.X|Z}(s)$ and $\bar{\delta}_Y(s)$. When proxies are available, Theorem 4.1 gives that $B(s) = R_{W.X|Z}(s)\bar{\delta}(s)$ and thus

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

and information on how the average effects of U on Y and W compare in magnitude and sign can fully or partially identify $\bar{\beta}(s)$ as discussed below. When $Z = X$, Theorem 4.1 gives $\bar{\beta}(s) = R_{Y.X}(s) - R_{W.X}(s)\bar{\delta}(s)$. If $S = 1$, Theorem 4.1 characterizes $\bar{\beta} \equiv E(\beta)$. The covariate-conditioned conditions $(i.b)$ and $(ii.b)$ in Theorem 4.1 weaken their unconditional analogs. Also, conditioning on covariates S may render Z "closer" to exogeneity and the conditional IV regression bias smaller. If the conditions in Theorem 4.1 hold for almost every $s \in \mathcal{S}$ then an expression for $\bar{\beta}$ derives by averaging the expression for $\bar{\beta}(s)$ over the distribution of S.

Next, we discuss the conditions in Theorem 4.1. Condition $(i.a)$ requires $\ell = k$ with $Cov(Z, X|S = s)$ nonsingular and condition $(ii.a)$ that $m = l$ with $\bar{\delta}_W(s)$ nonsingular. More generally, $\ell \geq k$ and $m \geq l$ suffice for full or partial identification and one may use a weighted

[12] Throughout, we write $A \perp_m B|S$ to denote mean independence $E(A|B, S) = E(A|B)$ provided these moments exist. We write $A \perp_m B|S = s$ to denote conditional mean independence at $S = s$.

combination of the resulting moments. In particular, having $\ell \geq k + m$ may fully identify $\bar{\beta}(s)$. For example, if $S = 1$, $\ell = k + m$ and $m = l$ then $(\bar{\beta}', \bar{\delta}')' = R_{Y.(X',W')'|Z}$ provided $Cov[Z, (X', W')']$ is nonsingular. We do not require this many instruments here; as such $\bar{\beta}$ is "under-identified." In particular, we let $\ell = k$, with Z possibly equal to X.

The uncorrelation and mean independence conditions $(i.b)$ and $(ii.b)$ are implied by the assumption that the random coefficients are mean independent[13] of (U, Z, X) given $S = s$ or the stronger assumption that $(U_W, U_Y) \perp (U, Z, X)|S = s$. Rewriting the equations for Y and W as

$$\begin{aligned} Y &= \bar{\alpha}_Y(s) + X'\bar{\beta}(s) + U'\bar{\delta}_Y(s) + \eta_Y \quad \text{with} \quad \eta_Y \equiv \tilde{\alpha}_Y(s) + X'\tilde{\beta}(s) + U'\tilde{\delta}_Y(s), \text{ and} \\ W' &= \bar{\alpha}'_W(s) + U'\bar{\delta}_W(s) + \eta'_W \quad \text{with} \quad \eta'_W \equiv \tilde{\alpha}'_W(s) + U'\tilde{\delta}_W(s), \end{aligned}$$

gives that the conclusion of Theorem 4.1 holds provided $Cov((\eta_Y, \eta'_W)', Z|S = s) = 0$. Conditions $(i.b)$ and $(ii.b)$ imply this zero covariance and have a simple interpretation. Moreover, allowing for random slopes β, δ_Y, and δ_W generates heterogeneity, accommodating an error-laden proxy W that is nonmonotonic in U given S for example. Nevertheless, $(i.b)$ and $(ii.b)$ ensure that $\bar{\beta}(s)$ is characterized by the same expression that would obtain when the slope coefficients are constant or deterministic functions of S, in which case $(i.b)$ and $(ii.b)$ reduce to $Cov((\alpha_Y, \alpha'_W)', Z|S = s) = 0$. In particular, $(i.b)$ and $(ii.b)$ ensure that restrictions on the average effects $\bar{\delta}_Y(s)$ and $\bar{\delta}_W(s)$ (as opposed to other features of δ_Y and δ_W) suffice to fully or partially identify $\bar{\beta}(s)$ as we discuss below.

Roughly speaking, $(i.b)$ isolates U as the source of the difficulty in identifying $\bar{\beta}(s)$. While the slope coefficients are heterogenous, endogeneity or "essential heterogeneity" is due to U. Had U been observed with $\ell = k+l$ and $Cov(Z, (X', U')'|S = s)$ nonsingular, $(i.b)$ would permit identifying $\bar{\beta}(s)$ and $\bar{\delta}_Y(s)$ via conditional IV regression. Note that linearity and condition $(i.b)$ can (indirectly) restrict how the return to education β depends on ability U (see e.g. Card 1999). In the linear correlated random coefficient model, and if valid instruments are available, one can consider IV methods, e.g. Wooldridge (1997, 2003) and Heckman and Vytlacil (1998). Similarly, linearity and condition $(i.b)$ can restrict how δ_Y relates to Z (and X), e.g. in learning models where the return to ability can vary with experience and depend on educational attainment (e.g. Altonji and Pierret, 2001; Arcidiacono, Bayer, and Hizmo, 2010). Sections 5 and 6 weaken these restrictions in Theorem 4.1 and give identification results for the case in which X and U can interact possibly nonlinearly to determine Y via the specification in S.1. Importantly, however, the conditions in Theorem 4.1 do not restrict the joint distribution of $(U, Z', X')'$ given

[13] In the linear case, having (α_Y, α_W) be mean independent of Z given S in Theorem 4.1 may fully identify $\bar{\beta}$ by generating a sufficient number of instruments as functions of Z. Indeed, under such stronger (mean) independence assumptions involving Z, one can dispense with linearity as we show in Sections 5 and 6.

S other than requiring that $Cov(Z, X|S = s)$ is nonsingular. In particular, Z and X can be freely (conditionally) correlated with U and thus endogenous.

Condition $(ii.b)$ ensures that W is an informative proxy with $Cov(Z, W|S = s)$ arising solely due to U. Note that we do not directly restrict the dependence between α_W and U. Nevertheless, even when δ_W is constant and $\alpha_W \perp U|S = s$, recall that the coefficient on X from a regression of Y on $(1, X', W')'$ given $S = s$ need not identify $\bar{\beta}(s)$. Further, as discussed in Section 2.4.5, restrictions involving α_W, δ_W, and U, ensuring that there are multiple proxies per confounder that are otherwise unrelated, may enable identifying $(\bar{\beta}(s)', \bar{\delta}(s)')'$. We do not require such restrictions here, allowing for components of α_W (e.g. test taking skills) to be correlated for example. (Online Appendix B studies the case of multiple proxies for U that are components of X.)

If the conditions in Theorem 4.1 hold for all $s \in \mathcal{S}$ and $\bar{\delta}_W(S)$, $\bar{\beta}(S)$, and $\bar{\delta}_Y(S)$ are constant, so that the mean independence assumptions involving β, δ_Y, and δ_W hold unconditionally, then

$$Cov(Z, Y|S) = Cov(Z, X|S)\bar{\beta} + Cov(Z, W|S)\bar{\delta},$$

and $(\bar{\beta}', \bar{\delta}')'$ is fully identified if variation in S generates a system of at least $k + m$ linearly independent equations[14]. Even if this fails, applying the law of iterated expectations gives

$$\bar{\beta} = E(\tilde{Z}(S)\tilde{X}'(S))^{-1}E(\tilde{Z}(S)\tilde{Y}(S)) - E(\tilde{Z}(S)\tilde{X}'(S))^{-1}E(\tilde{Z}(S)\tilde{W}(S))\bar{\delta}.$$

In addition to $\bar{\delta}$, this expression involves two estimands from IV regressions of $\tilde{Y}(S)$ and $\tilde{W}(S)$ respectively on $\tilde{X}(S)$ using instrument $\tilde{Z}(S)$. Further, if $\bar{Z}(S)$ or/and the conditional expectations $\bar{X}(S)$ (if $Z \neq X$), $\bar{W}(S)$, and $\bar{Y}(S)$ are affine functions of S, we obtain

$$\bar{\beta} = E(\epsilon_{Z.S}\epsilon'_{X.S})^{-1}E(\epsilon_{Z.S}\epsilon'_{Y.S}) - E(\epsilon_{Z.S}\epsilon'_{X.S})^{-1}E(\epsilon_{Z.S}\epsilon'_{W.S})\bar{\delta}.$$

Using partitioned regressions (Frisch and Waugh, 1933), the two residual-based IV estimands in the above expression for $\bar{\beta}$ can be recovered from $R_{Y.(X',S')'|(Z',S')'}$ and $R_{W.(X',S')'|(Z',S')'}$ as the coefficients associated with X in IV regressions of Y and W respectively on $(1, X', S')'$ using instruments $(1, Z', S')'$.

To illustrate the consequences of Theorem 4.1 on full identification, consider the example from Section 2.1 with constant slope coefficients, scalars U and W, and $S = 1$. Observe that $R_{Y.X|Z}$ fully identifies $\bar{\beta}$ under exogeneity. In this case, the IV regression bias disappears either because U does not determine Y, and in particular $\bar{\delta}_Y = 0$, or because Z and U are uncorrelated, and thus $R_{W.X|Z} = 0$. Alternatively, restrictions on the effects of U on Y and W can fully identify $\bar{\beta}$. In particular, this occurs under signed proportional confounding, in

[14]Specifically, we have $E(\tilde{Z}(S)[Y - X'\bar{\beta} - W'\bar{\delta}]|S) = 0$ and $(\bar{\beta}', \bar{\delta}')'$ may be identified if interacting $\tilde{Z}(S)$ with functions of S generates sufficiently many instruments.

which case the sign of the ratio $\frac{\bar{\delta}_Y}{\bar{\delta}_W}$ of the average direct effect of U on Y to that of U on W is known and its magnitude equals a known constant $|d|$. For example, under equiconfounding, $|d| = 1$, and U directly affects Y and W equally on average. Then $\bar{\beta}$ is fully 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 $\delta_W = diag(\delta_{W_1}, ..., \delta_{W_m})$. In this case,

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

As before, under exogeneity $\bar{\beta} = R_{Y.X|Z}$ whereas under e.g. positive equiconfounding $\frac{\bar{\delta}_{Y,h}}{\bar{\delta}_{W_h}} = 1$ for $h = 1, ..., m$ and $\bar{\beta} = R_{Y.X|Z} - \sum_{h=1}^{m} R_{W_h.X|Z}$.

Corollary 4.2 extends these full identification results to allow for covariates S and a general matrix δ_W, with $\bar{\delta}(s)$ possibly equal to $d(s)$ a vector of known or estimable functions of s. However, it is useful throughout to keep in mind the leading one-to-one case where $\delta_W(s)$ is a diagonal matrix with straightforward interpretation. We use subscripts to denote vector elements. For example, $\bar{\beta}_j(s)$ and $R_{Y.X|Z,j}(s)$ are the j^{th} elements of $\bar{\beta}(s)$ and $R_{Y.X|Z}(s)$, and $\bar{\delta}_h(s)$ and $d_h(s)$ are the h^{th} elements of $\bar{\delta}(s)$ and $d(s)$, respectively.

Corollary 4.2 *Assume the conditions of Theorem 4.1 and let $j = 1, ..., k$. (i) If $B_j(s) = 0$ (exogeneity) then $\bar{\beta}_j(s) = R_{Y.X|Z,j}(s)$. (ii) If $\bar{\delta}(s) = d(s)$ (signed proportional confounding) then $\bar{\beta}_j(s) = R_{Y.X|Z,j}(s) - R_{W.X|Z,j}(s)d(s)$.*

In sum, it suffices for exogeneity that $\bar{\delta}_Y(s) = 0$ or $R_{W.X|Z}(s) = 0$. In particular, if one fails to reject the null hypothesis $R_{W.X|Z,j}(s) = 0$ against the alternative $R_{W.X|Z,j}(s) \neq 0$, say via a t-test in the scalar proxy case, then one cannot reject, under Theorem 4.1's assumptions, that $R_{Y.X|Z,j}(s)$ identifies $\bar{\beta}_j(s)$. Further, signed proportional confounding with known $|d_h(s)|$ and $sign(d_h(s)), h = 1, ..., m$, point identifies $\bar{\beta}(s)$.

4.3 Partial Identification

In the absence of conditions leading to full identification, magnitude and sign restrictions on confounding may partially identify the elements of $\bar{\beta}(s)$. To illustrate, consider the return to education example from Section 2.1 with $S = 1$, scalar U, and proxy W ($\log(KWW)$). As discussed above, exogeneity ($\bar{\delta}_Y = 0$) and signed proportional confounding ($\bar{\delta}_Y = d\bar{\delta}_W$) are limit cases securing full identification. Next, we derive sharp identification regions for the elements of $\bar{\beta}$ under weaker magnitude and sign restrictions. In particular, we ask how does 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 the magnitude of the average direct effect of U on Y is not larger than that of U on W. Here, W is, on average, at least as directly responsive to U than Y is. Assume further that the sign of $\bar{\delta}$ is known, e.g. $0 \leq \bar{\delta}$ so that U affects Y and W on average in the same direction. Then $\bar{\delta} \in \mathcal{D} = [0, 1]$ and the expression for $\bar{\beta}$ gives the following identification region for $\bar{\beta}_j$, $j = 1, ..., k$, which depends on the sign of $R_{W.X|Z,j}$:

$$\bar{\beta}_j \in \mathcal{B}_j([0,1] \mid R_{W.X|Z,j} \leq 0) = [R_{Y.X|Z,j},\ R_{Y.X|Z,j} - R_{W.X|Z,j}],$$
$$\bar{\beta}_j \in \mathcal{B}_j([0,1] \mid 0 \leq R_{W.X|Z,j}) = [R_{Y.X|Z,j} - R_{W.X|Z,j},\ R_{Y.X|Z,j}].$$

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

Instead, if $1 \leq \left|\bar{\delta}\right| \equiv \left|\frac{\bar{\delta}_Y}{\bar{\delta}_W}\right|$, so that W is on average at most as directly responsive to U than Y is, and $0 \leq \bar{\delta}$, so that U affects Y and W on average in the same direction, then $\bar{\delta} \in \mathcal{D} = [1, +\infty)$ and we obtain the following identification region for $\bar{\beta}_j$, $j = 1, ..., k$:

$$\bar{\beta}_j \in \mathcal{B}_j([1,+\infty) \mid R_{W.X|Z,j} \leq 0) = [R_{Y.X|Z,j} - R_{W.X|Z,j},\ +\infty),$$
$$\bar{\beta}_j \in \mathcal{B}_j([1,+\infty) \mid 0 \leq R_{W.X|Z,j}) = (-\infty,\ R_{Y.X|Z,j} - R_{W.X|Z,j}].$$

Note that this identification region excludes the IV estimand $R_{Y.X|Z,j}$. The mirror-image result obtains for $\bar{\delta} \in \mathcal{D} = (-\infty, -1]$.

Wider intervals obtain under either magnitude or sign (but not both) restrictions on the average direct effects $\bar{\delta}_Y$ and $\bar{\delta}_W$. In particular, if $\left|\bar{\delta}_Y\right| \leq \left|\bar{\delta}_W\right|$, W is on average at least as directly responsive as Y is to U, $\bar{\delta} \in \mathcal{D} = [-1, 1]$, and $\bar{\beta}_j$ is partially identified as follows:

$$\bar{\beta}_j \in \mathcal{B}_j([-1,\ 1]) = [R_{Y.X|Z,j} - \left|R_{W.X|Z,j}\right|,\ R_{Y.X|Z,j} + \left|R_{W.X|Z,j}\right|].$$

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 $\left|\bar{\delta}_W\right| \leq \left|\bar{\delta}_Y\right|$, W is, on average, less directly responsive to U than Y is, $\bar{\delta} \in \mathcal{D} = (-\infty,\ -1] \bigcup [1,\ +\infty)$, and

$$\bar{\beta}_j \in \mathcal{B}_j((-\infty,\ -1] \bigcup [1,\ +\infty)) = (-\infty,\ R_{Y.X|Z,j} - \left|R_{W.X|Z,j}\right|] \bigcup [R_{Y.X|Z,j} + \left|R_{W.X|Z,j}\right|,\ +\infty).$$

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 $\mathcal{B}_j((-\infty,\ -1] \bigcup [1,\ +\infty))$ are.

Alone, sign restrictions determine the direction of the IV regression omitted variable bias:

$$\mathcal{B}_j((-\infty,\ 0] \mid 0 \leq R_{W.X|Z,j}) = \mathcal{B}_j([0,\ +\infty) \mid R_{W.X|Z,j} \leq 0) = [R_{Y.X|Z,j},\ +\infty),$$
$$\mathcal{B}_j((-\infty,\ 0] \mid R_{W.X|Z,j} \leq 0) = \mathcal{B}_j([0,+\infty) \mid 0 \leq R_{W.X|Z,j}) = (-\infty,\ R_{Y.X|Z,j}].$$

The above identification regions for $\bar{\beta}_j$, under magnitude or sign restrictions on confounding (or both) with scalars U and W, are sharp. Thus, any point in these regions is feasible under the maintained assumptions. In particular, given S.2 and the distribution of the observables M, for each element b of $\mathcal{B}_j(\mathcal{D})$, one can construct constants d_Y and d_W (which, being constant, satisfy the conditions on δ_Y and δ_W in Theorem 4.1) such that $\frac{d_Y}{d_W} \in \mathcal{D}$. For example, for $R_{W.X|Z,j} \neq 0$, it suffices to set $\frac{d_Y}{d_W} = \frac{1}{R_{W.X|Z,j}}(R_{Y.X|Z,j} - b)$.

These identification regions obtain in part by asking how the average direct effects $\bar{\delta}_Y$ and $\bar{\delta}_W$ compare in magnitude. If this comparison is ambiguous, a researcher may be more confident imposing a lower bound d_L and upper bound d_H on $\frac{\bar{\delta}_Y}{\bar{\delta}_W}$ so that $\bar{\delta} \in \mathcal{D} = [d_L, d_H]$. In this case, similar sharp identification regions, involving $d_L R_{W.X|Z,j}$ and $d_H R_{W.X|Z,j}$, derive, with exogeneity or signed proportional confounding as limit cases. Further, suppose more generally 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 $\bar{\beta} = R_{Y.X|Z} - \sum_{h=1}^{m} \frac{\bar{\delta}_{Y,h}}{\bar{\delta}_{W_h}} R_{W_h.X|Z,j}$ and magnitude and/or sign restrictions on $\bar{\delta}_h = \frac{\bar{\delta}_{Y,h}}{\bar{\delta}_{W_h}} \in \mathcal{D}_h$, $h = 1, ..., m$, yield the sharp identification regions $\mathcal{B}_j(\times_{h=1}^{m} \mathcal{D}_h)$ for $\bar{\beta}_j$, $j = 1, ..., k$, defined next, which may depend on $sign(R_{W_h.X|Z,j})$, $h = 1, ..., m$.

Corollary 4.3 generalizes the above discussion and derives sharp identification regions allowing for covariates S, a general matrix $\delta_W(s)$, and restrictions[15] $\bar{\delta}_h(s) \in \mathcal{D}_h(s) = [d_{L,h,}(s), d_{H,h}(s)]$, $h = 1, ..., m$. We allow for $d_{L,h}(s) = -\infty$ and/or $d_{H,h}(s) = +\infty$ yielding identification regions that are either half open intervals or the real line.

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

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

and this identification region is sharp.

These sharp identification regions may but need not contain $R_{Y.X|Z,j}(s)$. Sharp identification regions under either magnitude or sign restrictions (or both) on confounding derive by setting the vectors $d_L(s)$ and $d_H(s)$ suitably. Last, note that different potential instruments or proxies may lead to different identification regions for $\bar{\beta}_j(s)$, in which case $\bar{\beta}_j(s)$ is identified in the intersection of these regions, provided it is nonempty.

Online Appendix B contains extensions of the results in Section 4 on identification of average coefficients under magnitude and sign restrictions on confounding. Section B.1 studies a panel structure with individual and time varying random coefficients without requiring the unobserved

[15] Here, we focus on interval restrictions. One can consider other types of restrictions on $\bar{\delta}_h$, including imposing 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 $\bar{\delta}_h$.

characteristics to have "fixed effects" over time. Section B.2 studies cases where the proxies W are a component X_1 of X, included in the Y equation.

5 Identification of Average Nonparametric Effects

We extend the analysis in Section 4 by removing the linearity assumption S.2 and letting $Y = r(X, S, U, U_Y)$ as specified in S.1. Here, we study the identification of the conditional average direct effect of X on Y at (x, x^*) given $X = x^*$ and $S = s$:

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

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

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

Our analysis also enables studying the identification of averages of $\bar{\beta}(x, x^*|x^*, s)$ and $\bar{\beta}(x|x, s)$ such as $\bar{\beta}(x, x^*|x^*) = E[\bar{\beta}(x, x^*|x^*, S)|X = x^*]$ and $\bar{\beta}(s) = E[\bar{\beta}(X|X, s)|S = s]$.

In studying the identification of $\bar{\beta}(x, x^*|x^*, s)$ and $\bar{\beta}(x|x, s)$, 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 of B given covariates $S = s$, we let

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

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

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

5.1 Additive Separability

We begin our analysis by studying the case in which U enters r separably as in the following assumption. We remove separability in Section 5.2.

Assumption 3 (S.3) *Additive Separability: Assume S.1 with*

$$Y = r(X, S, U, U_Y) = \ddot{r}(X, S, U_Y) + \sum_{g=1}^{l} U_g \breve{r}_g(S, U_Y) \equiv \ddot{r}(X, S, U_Y) + U'\delta_Y,$$

and W generated as in S.2:

$$W' = \alpha_W + U'\delta_W.$$

Under S.3, when $\delta_Y = 0$ and $U_Y \perp X|S$, we obtain the nonparametric specification for the Y equation, imposed e.g. in Altonji and Matzkin (2005), Hoderlein and Mammen (2007), and Imbens and Newey[16] (2009), in which case certain average effects of X on Y are point identified. We maintain that $U_Y \perp X|S = s$ but allow U to freely (conditionally) depend on X, with δ_Y possibly nonzero. We then characterize the resulting regression omitted variable bias and study the identification of average effects of X on Y under restrictions on confounding.

Under S.3 (separability) and $U_Y \perp X|S = s$, conditioning on X is irrelevant and we have:

$$\bar{\beta}(x, x^*|x^*, s) = \bar{\beta}(x, x^*|s) \equiv E[\ddot{r}(x^*, s, U_Y) - \ddot{r}(x, s, U_Y)|S = s],$$

and

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

If, in addition, the effect of X on Y is linear then these effects do not depend on X, and we obtain the random coefficient specification from Section 4.

Theorem 4.1 characterizes the nonparametric regression bias under S.3. Further, when proxies are available, it shows that $\bar{\beta}(x, x^*|s)$ and $\bar{\beta}(x|s)$ depend on the unknown $\bar{\delta}(s) \equiv \bar{\delta}_W^{-1}(s)\bar{\delta}_Y(s)$, involving the conditional average direct effects of U on Y and W.

Theorem 5.1 *Assume S.3 with $m = l$. Let $s \in \mathcal{S}$ and $x, x^* \in \mathcal{X}$ with $E[(Y, W')'|X = \ddot{x}, S = s] < \infty$ for $\ddot{x} = x, x^*$.*
(i.a) If $U_Y \perp X|S = s$ and $\delta_Y \perp_m (U, X)|S = s$ then

$$B(x, x^*|s) \equiv R^N_{Y.X}(x, x^*; s) - \bar{\beta}(x, x^*|s) = R^N_{U.X}(x, x^*; s)\bar{\delta}_Y(s).$$

(i.b) If (i.b.1) $\bar{\delta}_W(s)$ is nonsingular with $\bar{\delta}(s) \equiv \bar{\delta}_W^{-1}(s)\bar{\delta}_Y(s)$ and (i.b.2) $\alpha_W \perp_m X|S = s$ and $\delta_W \perp_m (U, X)|S = s$ then

$$B(x, x^*|s) = R^N_{W.X}(x, x^*; s)\bar{\delta}(s).$$

(ii.a) Set $k = 1$ and assume the conditions in (i.a). If (ii.a.1) $\frac{\partial}{\partial x}E(U'|X = x, S = s)$ exists and is finite and (ii.a.2) for all $x^\dagger \in \mathcal{N}(x) \subseteq \mathcal{X}$, a nonempty open neighborhood of x, $E[\ddot{r}(x^\dagger, s, U_Y)|S = s] < \infty$, $\frac{\partial}{\partial x}\ddot{r}(x^\dagger, s, u_y)$ exists for a.e.[17] (almost every) u_y, and there is a function $\Delta_s(U_Y)$ with $E(\Delta_s(U_Y)|S = s) < \infty$ such that $\left|\frac{\partial}{\partial x}\ddot{r}(x^\dagger, s, u_y)\right| < \Delta_s(u_y)$ for a.e. u_y then

$$B(x|s) \equiv R^N_{Y.X}(x; s) - \bar{\beta}(x|s) = R^N_{U.X}(x; s)\bar{\delta}_Y(s).$$

[16] Similar to Imbens and Newey (2009), one can consider covariates S_2 and a scalar unobserved S_1 recoverable from the choice equation $X = \ddot{q}(Z, S_2, S_1)$ with $\ddot{q}$ 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.

[17] Here, the qualifier "almost every" (a.e.) means that the condition can fail for u_y belonging to a measurable set $\mathcal{V}$ having $P[U_Y \in \mathcal{V} \mid S = s] = 0$ (see e.g. White and Chalak, 2013).

(ii.b) If the conditions in (i.b) and (ii.a.1) hold then

$$B(x|s) = R^N_{W.X}(x;s)\bar{\delta}(s).$$

Conditions $(i.a)$ and $(i.b.2)$ of Theorem 5.1 strengthen their counterparts in Theorem 4.1 with $Z = X$. They are implied by the stronger condition[18] $(U_W, U_Y) \perp (U, X)|S$. Condition $(i.a)$ allows for nonzero δ_Y and thus weakens the assumption $U_Y \perp X|S$ with $\delta_Y = 0$ often employed in the literature. Condition $(i.b.2)$ ensures that W is an informative proxy so that the mean dependence of W on X given $S = s$ arises solely due to U. Condition $(ii.a)$ ensures that derivatives exist and that $\frac{\partial}{\partial x}E[\ddot{r}(x, s, U_Y)|S = s] = E[\frac{\partial}{\partial x}\ddot{r}(x, s, U_Y)|S = s]$ (see e.g. White and Chalak, 2013).

The expressions in $(i.a)$ and $(ii.a)$ for the biases $B(x, x^*|s)$ and $B(x|s)$ enable reasoning about the direction of the omitted variable bias, which depends on the signs of $R^N_{U.X}(x, x^*; s)$ and $\bar{\delta}_Y(s)$. When proxies are available, Theorem 5.1 characterizes $\bar{\beta}(x, x^*|s)$ and $\bar{\beta}(x|s)$ by

$$\bar{\beta}(x, x^*|s) = R^N_{Y.X}(x, x^*; s) - R^N_{W.X}(x, x^*; s)\bar{\delta}(s) \quad \text{and} \quad \bar{\beta}(x|s) = R^N_{Y.X}(x; s) - R^N_{W.X}(x; s)\bar{\delta}(s).$$

As in the linear case, if $B(x, x^*|s) = 0$ (conditional exogeneity) then $\bar{\beta}(x, x^*|s) = R^N_{Y.X}(x, x^*; s)$ is fully identified. This obtains if $\bar{\delta}_Y(s) = 0$ or $U \perp_m X|S = s$. Alternatively, if $\bar{\delta}(s) = d(s)$ with $d(s)$ known or estimable (conditional signed proportional confounding) then $\bar{\beta}(x, x^*|s) = R^N_{Y.X}(x, x^*; s) - R^N_{W.X}(x, x^*; s)d(s)$ is fully identified. Analogous results hold for $\bar{\beta}(x|s)$.

From Theorem 5.1, observe that another avenue for full identification of $\bar{\beta}(x, x^*|s)$ or $\bar{\beta}(x|s)$ is to impose m restrictions on $\bar{\beta}(\cdot, \cdot|s)$ or $\bar{\beta}(\cdot|s)$. For example, if $m = l = 1$ and one assumes that $\bar{\beta}(x^\dagger, x^\ddagger|s) = 0$ for $x^\dagger, x^\ddagger \in \mathcal{X}$, as occurs e.g. if a nondegenerate component of X is excluded from r and thus the Y equation, then, provided $R^N_{W.X}(x^\dagger, x^\ddagger; s) \neq 0$, $\bar{\delta}(s) = \frac{R^N_{Y.X}(x^\dagger, x^\ddagger; s)}{R^N_{W.X}(x^\dagger, x^\ddagger; s)}$ and $\bar{\beta}(x, x^*|s)$ is thus fully identified. Analogous restrictions fully identify $\bar{\beta}(x|s)$. We do not require such restrictions.

In the absence of assumptions yielding full identification, restrictions on the magnitude and/or sign of confounding, $\bar{\delta}_h(s) \in \mathcal{D}_h(s) \equiv [d_{L,h}(s), d_{H,h}(s)]$, partially identify $\bar{\beta}(x, x^*|s)$ or $\bar{\beta}(x|s)$. Specifically, we obtain that $\bar{\beta}(x, x^*|s) \in \mathcal{B}(\times_{h=1}^m \mathcal{D}_h(s))$, with this sharp identification region defined analogously to Corollary 4.3 with $R^N_{Y.X}(x, x^*; s)$ replacing $R_{Y.X|Z}(s)$ and $R^N_{W.X}(x, x^*; s)$ replacing $R_{W.X|Z}(s)$. Analogous bounds obtain for $\bar{\beta}(x|s)$.

5.2 Nonseparability

Next, we remove the separability assumption in S.3 and let Y and W be generated as in S.1:

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

[18] The average effects of X on Y would be identified if either $U_Y \perp (U, X)|S$ with U observed or $(U, U_Y) \perp X|S$ with U unobserved. Here U is unobserved and can depend on X given S.

Theorem 5.2 characterizes the nonparametric bias $B(x, x^*|x^*, s)$ or $B(x|x, s)$ of the nonparametric regression estimand $R^N_{Y.X}(x, x^*; s)$ or $R^N_{Y.X}(x; s)$ in recovering the average effect $\bar{\beta}(x, x^*|x^*, s)$ or $\bar{\beta}(x|x, s)$. For brevity, we state the results in the case of a continuous scalar U with r and q differentiable in u. Theorem B.1 in Online Appendix B gives analogous results for discrete U, with sums replacing integrals. We obtain full or partial identification by imposing restrictions on the average marginal effects of U on Y and the scalar proxy W, denoted by:

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

Unlike for $\bar{\beta}(x, x^*|x^*, s)$ and $\bar{\beta}(x|x, s)$, we forgo conditioning on X in $\bar{\delta}_Y(u; x|s)$ and $\bar{\delta}_W(u|s)$ since this is irrelevant given our assumptions $U_Y \perp (U, X)|S = s$ and $U_W \perp (U, X)|S = s$.

As in Theorem 5.1$(ii.a)$, we impose regularity conditions to ensure that moments and derivatives exist and to justify interchanging the order of derivative and integral in expressions such as

$$\begin{aligned} R^N_{Y.X}(x; s) &= \frac{\partial}{\partial x} E\{E[r(x, s, U, U_Y)|X = x, U, S = s]|X = x, S = s\} \\ &= \frac{\partial}{\partial x} \int_{\mathcal{U}_{x,s}} E[r(x, s, u, U_Y)|S = s] f_{U|X,S}(u|x, s) du, \end{aligned}$$

where we make use of $U_Y \perp (U, X)|S = s$ in the last equality. For this, we also assume that the support $\mathcal{U}_{x,s}$ is constant in a neighborhood of x (or that $\mathcal{U}_{x,s} = \mathcal{U}_{x^*,s}$ in the discrete case), to remove the complication introduced by boundary terms. We collect these regularity conditions of Theorem 5.2 in Assumption A.1 of Appendix A.

Theorem 5.2 *Assume S.1 with $m = l = 1$, $s \in \mathcal{S}$, $x, x^* \in \mathcal{X}$, and that $F_{U|X,S}(\cdot|x^*, s)$ and $F_{U|X,S}(\cdot|x, s)$ are absolutely continuous.*
(i.a) If $U_Y \perp (U, X)|S = s$ and A.1.i(a,b,c,d) hold then

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

(i.b) If $U_W \perp (U, X)|S = s$ and A.1.i(b,e,f,g) hold then

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

(ii) Set $k = 1$. (ii.a) If, in addition to $U_Y \perp (U, X)|S = s$ and A.1.i(c,d), A.1.ii(a,b,c,d) hold then

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

(ii.b) If, in addition to $U_W \perp (U, X)|S = s$ and A.1.i(f,g), A.1.ii(a,d,e) hold then

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

Theorem 5.2 derives the nonparametric regression omitted variable bias $B(x, x^*|x^*, s)$ or $B(x|x, s)$ for the identification of $\bar{\beta}(x, x^*|x^*, s)$ or $\bar{\beta}(x|x, s)$ and shows that this depends on the average marginal effect $\bar{\delta}_Y(u; x|s)$ of U on Y as well as on the conditional distribution of U given X and S. This generalizes the classic linear regression omitted variable bias representation and provides insight into the sign of this bias in the nonparametric nonseparable case. For instance, if we assume that $\bar{\delta}_Y(u; x|s)$ is nonnegative for a.e. $u \in \mathcal{U}_{x,s}$ (e.g. the average marginal effect of ability on wage is nonnegative) and that the stochastic dominance relation $F_{U|X,S}(u|x^*, s) \leq F_{U|X,S}(u|x, s)$ for a.e. $u \in \mathcal{U}_{x,s}$ holds (e.g. the probability of low ability U is small when education is large $(x < x^*)$) then $B(x, x^*|x^*, s)$ is nonnegative.

Under conditional exogeneity, $B(x, x^*|x^*, s) = 0$ and $R^N_{Y.X}(x, x^*; s)$ fully identifies $\bar{\beta}(x, x^*|x^*, s)$. This occurs if $U \perp X|S = s$ or if $\bar{\delta}_Y(u; x|s) = 0$ for a.e. $u \in \mathcal{U}_{x,s}$. Alternatively, under conditional proportional confounding, $\bar{\delta}_Y(u; x|s) = d(x, s)\bar{\delta}_W(u|s)$ for a.e. $u \in \mathcal{U}_{x,s}$, with $d(x, s)$ known or estimable. In this case, $\bar{\beta}(x, x^*|x^*, s) = R^N_{Y.X}(x, x^*; s) - d(x, s)R^N_{W.X}(x, x^*; s)$. Analogous results obtain for $\bar{\beta}(x|x, s)$.

In the absence of point identifying assumptions, magnitude and sign restrictions on confounding yield sharp bounds.

Corollary 5.3 *Suppose that, for a.e.* $u \in \mathcal{U}_{x,s}$, $\bar{\delta}_Y(u; x|s) = d(u, x, s)\bar{\delta}_W(u|s)$ *with* $d(u, x, s) \in \mathcal{D}(x, s) \equiv [d_L(x, s), d_H(x, s)]$. *(i) Under the conditions of Theorem 5.2(i), if* $\bar{\delta}_W(u|s)[F_{U|X,S}(u|x^*, s) - F_{U|X,S}(u|x, s)]$ *is either nonpositive for a.e.* $u \in \mathcal{U}_{x,s}$ *or nonnegative for a.e.* $u \in \mathcal{U}_{x,s}$ *then*

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

and this identification region is sharp.
(ii) Under the conditions of Theorem 5.2(ii), if $\bar{\delta}_W(u|s)\frac{\partial}{\partial x}F_{U|X,S}(u|x, s)$ *is either nonpositive for a.e.* $u \in \mathcal{U}_{x,s}$ *or nonnegative for a.e.* $u \in \mathcal{U}_{x,s}$ *then*

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

and this identification region is sharp.

The conditions of Corollary 5.3 obtain if $E[q(s, u, U_W)|S = s]$ is monotonic in u and the stochastic dominance relation $F_{U|X,S}(u|x^*, s) \leq F_{U|X,S}(u|x, s)$ for a.e. $u \in \mathcal{U}_{x,s}$ holds[19]. This allows W to be an imperfect proxy that depends on U_W and is possibly nonmonotonic in u. The restriction $d(u, x, s) \in \mathcal{D}(x, s)$ allows but does not require the comparison between $\bar{\delta}_Y(u; x|s)$ and $\bar{\delta}_W(u|s)$ to be local at each u and to depend on x in addition to s (e.g. conditional

[19]Manski and Pepper (2009) use similar conditions in lemma 3.1 where they 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. We impose neither of these assumptions here.

average return to ability depends on education). This reduces to $\bar{\delta}(s) \in \mathcal{D}(s)$ in the linear separable case. Last, given these identification results for $\bar{\beta}(x, x^*|x^*, s)$ and $\bar{\beta}(x|x, s)$, full or partial identification results for various average effects (e.g. $\bar{\beta}(s) = E[\bar{\beta}(X|X, s)|S = s]$ or $\bar{\beta} = E[\bar{\beta}(X|X, S)]$) obtain by averaging the bounds over the relevant distributions.

6 Identification of Local, Marginal, and Average Treatment Effects

This section characterizes the omitted variable bias of IV methods for recovering local and marginal treatment effects, as well as average treatment effects for the population, treated, and untreated, in a discrete choice structural system without exogeneity, and studies their identification under magnitude and sign restrictions on confounding. Recall that, under assumption S.1, Y and W are generated by:

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

Next, we follow e.g. Imbens and Angrist (1994) and Heckman and Vytlacil (2005) in considering a binary treatment generated according to the following selection structural equation.

Assumption 4 (S.4) *Assume S.1 and suppose further that X is generated by*[20]

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

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

Thus, an individual i selects into treatment ($x_i = 1$) if and only if $u_{x,i} \leq \nu(z_i, s_i)$ realizes. As for r, ν may but need not depend on covariates S. When interest attaches to a scalar potential instrument, we set $\ell = 1$ to denote it by Z and we subsume, without loss of generality, into S the remaining potential instruments. As shown in Vytlacil (2002), the threshold crossing specification in S.4 ensures (and, under conditional exogeneity of Z, is equivalent to) the monotonicity assumption imposed e.g. in Imbens and Angrist (1994). In particular, consider the additively separable case $Y = \ddot{r}(X, S, U_Y) + U'\delta_Y$. When $\delta_Y = 0$ and $(U_X, U_Y) \perp Z|S$, we obtain the specification for the X and Y equations studied in e.g. Imbens and Angrist (1994) and Heckman and Vytlacil (2005). In Section 6.2, we maintain that Z satisfies $(U_X, U_Y) \perp Z|S = s$ (in contrast to $U_Y \perp X|S = s$ in Section 5) but allow U to freely depend on Z, with the random vector δ_Y possibly nonzero. We remove additive separability in Section 6.3 and study identification of

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

local and marginal treatment effects under S.1 and S.4. Last, we let $F_{U_X|S}(\cdot|s)$ be absolutely continuous to simplify the exposition. In particular, we employ a probability transform to write

$$X = \mathbf{1}\{U_X \leq \nu(Z,S)\} = \mathbf{1}\{F_{U_X|S}(U_X|s) \leq F_{U_X|S}(\nu(Z,S)|s)\} = \mathbf{1}\{V \leq P(Z,s)\}$$

where $V \backsim Unif[0,1]$ and $P(Z,s)$ denotes the propensity score $\Pr(X=1|Z,S=s)$ given $U_X \perp Z|S=s$. Sometimes, we employ the convenient representation $X = \mathbf{1}\{V \leq P\}$ with scalar potential instrument $P \equiv P(Z,s)$.

6.1 Local and Marginal Treatment Effects

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

$$\begin{aligned}\bar{\beta}(0,1|\ \nu(z,s) < U_X \leq \nu(z^*,s), z^*, s) \\ \equiv E[r(1,s,U,U_Y) - r(0,s,U,U_Y)|\nu(z,s) < U_X \leq \nu(z^*,s), Z=z^*, S=s].\end{aligned}$$

This is the average direct effect of the treatment for the subpopulation with instrument $Z=z^*$ and covariates $S=s$ and for whom $X=0$ if $Z=z$ whereas $X=1$ if $Z=z^*$. Given $U_X \perp Z|S=s$, averaging this local effect over the distribution of Z given $S=s$ yields LATE:

$$\bar{\beta}(0,1|\nu(z,s) < U_X \leq \nu(z^*,s), s) \equiv E[r(1,s,U,U_Y) - r(0,s,U,U_Y)|\nu(z,s) < U_X \leq \nu(z^*,s), S=s].$$

For example, when Z is binary, $\bar{\beta}(0,1|\nu(0,s) < U_X \leq \nu(1,s), s)$ is the average direct treatment effect in subpopulation $S=s$ for the "compliers" who receive the treatment ($X=1$) if and only if $Z=1$ (see e.g. Angrist, Imbens, and Rubin, 1996).

Similarly, define the conditional marginal treatment effect (MTE) where we condition on Z and S in addition to U_X:

$$\bar{\beta}(0,1|\nu(z,s),z,s) \equiv E[r(1,s,U,U_Y) - r(0,s,U,U_Y)|U_X = \nu(z,s), Z=z, S=s].$$

Given $U_X \perp Z|S=s$, averaging $\bar{\beta}(0,1|\nu(z,s),z,s)$ over the distribution of Z given $S=s$ yields MTE:

$$\bar{\beta}(0,1|\nu(z,s),s) \equiv E[r(1,s,U,U_Y) - r(0,s,U,U_Y)|U_X = \nu(z,s), S=s],$$

denoting the average direct effect of the treatment for individuals with $S=s$ who are indifferent toward receiving the treatment if $Z=z$. Using the representation $X = \mathbf{1}\{V \leq P\}$ with $p=(z,s)$, we can rewrite this marginal effect as

$$\bar{\beta}(0,1|p,s) \equiv E[r(1,s,U,U_Y) - r(0,s,U,U_Y)|V=p, S=s].$$

In studying the identification of local and marginal effects as well as weighted averages of MTEs, such as the average treatment effect $\bar{\beta}(0,1|s)$, we use the following succinct notation for the conditional Wald (1940) and local instrumental variable (LIV) estimands. In particular, for random variable B and vectors A and C with $E[(A',B)'] < \infty$, let c and c^* be in the support of C given covariates $S = s$, and, provided the denominator is nonzero, define

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

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

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

6.2 Additive Separability

We begin by studying the additively separable specification for the Y and W equations given in S.3. Section 6.3 removes the separability assumption. Thus, in this subsection, we let

$$Y = \ddot{r}(X,S,U_Y) + U'\delta_Y, \quad X = \mathbf{1}\{U_X \leq \nu(Z,S)\}, \quad \text{and} \quad W' = \alpha_W + U'\delta_W.$$

Given separability in S.3 and $(U_X,U_Y) \perp Z|S=s$, we have that:

$$\begin{aligned}\bar{\beta}(0,1|\nu(z,s) < U_X \leq \nu(z^*,s),z^*,s) &= \bar{\beta}(0,1|\nu(z,s) < U_X \leq \nu(z^*,s),s) \\ &= E[\ddot{r}(1,s,U_Y) - \ddot{r}(0,s,U_Y)|\nu(z,s) < U_X \leq \nu(z^*,s),S=s]\end{aligned}$$

and

$$\bar{\beta}(0,1|\nu(z,s),z,s) = \bar{\beta}(0,1|\nu(z,s),s) = E[\ddot{r}(1,s,U_Y) - \ddot{r}(0,s,U_Y)|U_X = \nu(z,s),S=s].$$

Theorem 6.1 extends the results in Imbens and Angrist (1994) and Heckman and Vytlacil (2005) first by characterizing the omitted variable bias of the Wald or LIV estimand for $\bar{\beta}(0,1|\nu(z,s) < U_X \leq \nu(z^*,s),s)$ or $\bar{\beta}(0,1|\nu(z,s),s)$ and then by showing that, when proxies are available, this bias depends on the unknown $\bar{\delta}(s) \equiv \bar{\delta}^{-1}_W(s)\bar{\delta}_Y(s)$, which involves the conditional average direct effects of U on Y and W.

Theorem 6.1 *Assume S.3 and S.4 with $m = l$. Let $s \in \mathcal{S}$ and $z,z^* \in \mathcal{Z}$ with $\Pr[\nu(z,s) < U_X \leq \nu(z^*,s) \ |S=s] > 0$ and $E[(Y,W')'|Z=\ddot{z},S=s] < \infty$ for $\ddot{z} = z,z^*$.*
(i.a) If $(U_X,U_Y) \perp Z|S=s$ and $\delta_Y \perp_m (U,Z)|S=s$ then

$$\begin{aligned}B(0,1|\nu(z,s) \leq U_X < \nu(z^*,s),s) &\equiv R^{Wald}_{Y.X|Z}(z,z^*;s) - \bar{\beta}(0,1|\nu(z,s) \leq U_X < \nu(z^*,s),s) \\ &= R^{Wald}_{U.X|Z}(z,z^*;s)\bar{\delta}_Y(s).\end{aligned}$$

(i.b) If (i.b.1) $\bar{\delta}_W(s)$ *is nonsingular with* $\bar{\delta}(s) \equiv \bar{\delta}_W^{-1}(s)\bar{\delta}_Y(s)$ *and (i.b.2)* $\alpha_W \perp_m Z|S = s$ *and* $\delta_W \perp_m (U, Z)|S = s$ *then*

$$B(0,1|\nu(z,s) \leq U_X < \nu(z^*,s),s) = R^{Wald}_{W.X|Z}(z,z^*;s)\bar{\delta}(s).$$

(ii.a) Set $\ell = 1$ *and assume the conditions in (i.a). If (ii.a.1)* $\frac{\partial}{\partial z}E(U'|Z = z, S = s)$ *exists and (ii.a.2)* $\nu(\cdot, s)$ *is differentiable at* z *with* $\frac{\partial}{\partial z}\nu(z,s) \neq 0$ *and* $\bar{\beta}(0,1|\cdot,s)$ *and* $f_{U_X|S}(\cdot|s)$ *are continuous at* $\nu(z,s)$ *with* $f_{U_X|S}(\nu(z,s)|s) > 0$ *then*

$$B(0,1|\nu(z,s),s) \equiv R^{LIV}_{Y.X|Z}(z;s) - \bar{\beta}(0,1|\nu(z,s),s) = R^{LIV}_{U.X|Z}(z;s)\bar{\delta}_Y(s).$$

(ii.b) If the conditions in (i.b) and (ii.a.1) hold then

$$B(0,1|\nu(z,s),s) = R^{LIV}_{W.X|Z}(z;s)\bar{\delta}(s).$$

Conditions $(i.a)$ and $(i.b.2)$ of Theorem 6.1 are implied by the stronger condition $(U_X, U_W, U_Y) \perp (U, Z)|S$. In particular, $(i.a)$ weakens the common assumption $(U_X, U_Y) \perp Z|S$ with $\delta_Y = 0$. Also, $(i.b.2)$ ensures that W is an informative proxy so that, given $S = s$, the conditional mean dependence of W on Z arises solely due to U. The regularity conditions in $(ii.a)$ enable applying theorems for the derivative of an integral.

The sign of the conditional Wald or LIV bias depends on the signs of $R^N_{U.Z}(z,z^*;s)$ or $R^N_{U.Z}(z;s)$ and $\bar{\delta}_Y(s)$. These biases vanish under conditional exogeneity, which occurs if $\bar{\delta}_Y = 0$ or $U \perp_m Z|S = s$. In this case, we obtain the results in e.g. Imbens and Angrist (1994) and Heckman and Vytlacil (2005) for point identification of LATE and MTE:

$$\bar{\beta}(0,1|\nu(z,s) \leq U_X < \nu(z^*,s),s) = R^{Wald}_{Y.X|Z}(z,z^*;s) \quad \text{and} \quad \bar{\beta}(0,1|\nu(z,s),s) = R^{LIV}_{Y.X|Z}(z;s).$$

Moreover, when proxies are available, Theorem 6.1 gives that

$$\bar{\beta}(0,1|\nu(z,s) \leq U_X < \nu(z^*,s),s) = R^{Wald}_{Y.X|Z}(z,z^*;s) - R^{Wald}_{W.X|Z}(z,z^*;s)\bar{\delta}(s)$$

and

$$\bar{\beta}(0,1|\nu(z,s),s) = R^{LIV}_{Y.X|Z}(z;s) - R^{LIV}_{W.X|Z}(z;s)\bar{\delta}(s).$$

Thus, conditional signed proportional confounding, $\bar{\delta}(s) = d(s)$ with $d_h(s)$, $h = 1, ..., m$, known or estimable, also yields point identification.

Last, Theorem 6.1 shows that another avenue for full identification of $\bar{\beta}(0,1|\nu(z,s) \leq U_X < \nu(z^*,s),s)$ or $\bar{\beta}(0,1|\nu(z,s),s)$ is to assume m restrictions involving this local or marginal effect. For example, if $m = l = 1$ and one assumes that $\bar{\beta}(0,1|\nu(z^\dagger,s) \leq U_X < \nu(z^\ddagger,s),s) = \bar{\beta}(0,1|\nu(\dot{z},s) \leq U_X < \nu(\ddot{z},s),s)$ for $z^\dagger, z^\ddagger, \dot{z}, \ddot{z} \in \mathcal{Z}$, as occurs e.g. if a nondegenerate component of Z is excluded from ν and thus the X equation, then, provided $R^{Wald}_{W.X|Z}(z^\dagger, z^\ddagger; s) \neq$

$R^{Wald}_{W.X|Z}(\dot{z},\ddot{z};s)$, $\bar{\delta}(s) = \frac{R^{Wald}_{Y.X|Z}(\dot{z},\ddot{z};s) - R^{Wald}_{Y.X|Z}(z^\dagger,z^\ddagger;s)}{R^{Wald}_{W.X|Z}(\dot{z},\ddot{z};s) - R^{Wald}_{W.X|Z}(z^\dagger,z^\ddagger;s)}$ and $\bar{\beta}(0,1|\nu(z,s) \leq U_X < \nu(z^*,s),s)$ is thus fully identified. Analogous restrictions fully identify $\bar{\beta}(0,1|\nu(z,s),s)$. We do not require such assumptions.

When the conditions for point identification do not hold, magnitude and sign restrictions on confounding partially identify LATE and MTE in sharp identification regions $\mathcal{B}(\times_{h=1}^m \mathcal{D}_h(s))$ defined analogously to Corollary 4.3 with $R^{Wald}_{Y.X|Z}(z,z^*;s)$ or $R^{LIV}_{Y.X|Z}(z;s)$ replacing $R_{Y.X|Z}(s)$ and with $R^{Wald}_{W.X|Z}(z,z^*;s)$ or $R^{LIV}_{W.X|Z}(z;s)$ replacing $R_{W.X|Z}(s)$.

Building on the results in Heckman and Vytlacil (2005), we also fully or partially identify the average treatment effects for the population, treated, and untreated under restrictions on confounding. In particular, applying Theorem 6.1 for almost every p using the representation $X = \mathbf{1}\{V \leq P\}$ with potential instrument $P \equiv P(Z,s)$, gives that $\bar{\beta}(0,1|p,s) = R^{LIV}_{Y.X|P}(p,s) - R^{LIV}_{W.X|P}(p,s)\bar{\delta}(s)$. If P given $S = s$ has the unit interval for support, we have that the conditional average treatment effect is characterized by:

$$\bar{\beta}(0,1|s) = \int_0^1 \bar{\beta}(0,1|p,s)dp = \int_0^1 [R^{LIV}_{Y.X|P}(p,s) - R^{LIV}_{W.X|P}(p,s)\bar{\delta}(s)]dp.$$

Similarly, the conditional average treatment effects for the treated and untreated are characterized respectively by

$$\bar{\beta}(0,1|X=1,s) = \int_0^1 [R^{LIV}_{Y.X|P}(p;s) - R^{LIV}_{W.X|P}(p;s)\bar{\delta}(s)]\frac{(1-F_{P|S}(p|s))}{E(P(Z,S)|S=s)}dp,$$

and

$$\bar{\beta}(0,1|X=0,s) = \int_0^1 [R^{LIV}_{Y.X|P}(p;s) - R^{LIV}_{W.X|P}(p;s)\bar{\delta}(s)]\frac{F_{P|S}(p|s)}{E(1-P(Z,S)|S=s)}\ dp.$$

As these expressions show, these effects are fully identified under conditional exogeneity, e.g. $B(0,1|p,s) = 0$, or conditional signed proportional confounding with $\bar{\delta}(s) = d(s)$. Otherwise, sharp identification regions obtain under sign and magnitude restrictions on confounding analogously to Corollary 4.3.

6.3 Nonseparability

Next, we remove the separability assumption S.3 and let Y and W be as in S.1 and S.4:

$$Y = r(X,S,U,U_Y), \quad X = \mathbf{1}\{U_X \leq \nu(Z,S)\}, \quad \text{and } W = q(S,U,U_W),$$

and study the identification of $\bar{\beta}(0,1|\nu(z,s) < U_X \leq \nu(z^*,s),z^*,s)$ and $\bar{\beta}(0,1|\nu(z,s),z,s)$. Theorem 6.2 characterizes the nonparametric bias of the Wald and LIV estimands for recovering these local and marginal effects. For brevity, we give the results in the case of a continuous

scalar U with r and q differentiable in u. Similar results obtain for discrete U, with sums replacing integrals. Assumption A.2 in Appendix A collects regularity conditions justifying interchanging the order of well-defined integral and derivative. A.2 also assumes that $\mathcal{U}_{z,s}$ is constant in a neighborhood of z (or that $\mathcal{U}_{z,s} = \mathcal{U}_{z^*,s}$ in the discrete case), to remove the complication introduced by boundary terms. We obtain full or partial identification by imposing restrictions on the following average marginal effect or U on Y and the scalar proxy W:

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

where we forgo conditioning on Z given $(U_X, U_Y) \perp (U, Z)|S = s$ and $U_W \perp (U, Z)|S = s$.

Theorem 6.2 *Assume S.1 and S.4 with $m = l = 1$, $s \in \mathcal{S}$, $z, z^* \in \mathcal{Z}$, $\Pr[\nu(z,s) < U_X \leq \nu(z^*,s) \ |S = s] > 0$, and that $F_{U|Z,S}(\cdot|z^*, s)$ and $F_{U|Z,S}(\cdot|z, s)$ are absolutely continuous.*
(i.a) If $(U_X, U_Y) \perp (U, Z)|S = s$ and A.2.i(a,b,c,d) hold then

$$B(0,1|\nu(z,s) < U_X \leq \nu(z^*,s), z^*, s) \equiv R^{Wald}_{Y.X|Z}(z, z^*; s) - \bar{\beta}(0,1|\nu(z,s) < U_X \leq \nu(z^*,s), z^*, s)$$
$$= -\frac{1}{R^N_{X.Z}(z, z^*; s)} \int_{\mathcal{U}_{z,s}} \bar{\delta}_Y(u; z|s)[F_{U|Z,S}(u|z^*, s) - F_{U|Z,S}(u|z, s)]du.$$

(i.b) If $U_W \perp (U, Z)|S = s$ and A.2.i(b,e,f,g) hold then

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

(ii) Set $\ell = 1$. (ii.a) If, in addition to $(U_X, U_Y) \perp (U, Z)|S = s$ and A.2.i(c,d), A.2.ii(a,b,c,d,e) hold then

$$B(0,1|\nu(z,s), z, s) \equiv R^{LIV}_{Y.X|Z}(z; s) - \bar{\beta}(0,1|\nu(z,s), z, s) = -\frac{1}{R^N_{X.Z}(z; s)} \int_{\mathcal{U}_{z,s}} \bar{\delta}_Y(u; z|s) \frac{\partial}{\partial z} F_{U|Z,S}(u|z, s)du.$$

(ii.b) If, in addition to $U_W \perp (U, Z)|S = s$ and A.2.i(f,g), A.2.ii(a,c,e,f) hold then

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

Theorem 6.2 demonstrates that the nonparametric omitted variable bias of the Wald or LIV estimand for the identification of a local or marginal treatment effect depends on the average marginal effect of U on Y as well as on the distribution of U conditional on Z and S. For example, if we assume that $\bar{\delta}_Y(u; z|s)$ is nonnegative for a.e. $u \in \mathcal{U}_{z,s}$ (e.g. the average marginal effect of ability on wage is nonnegative) and that the stochastic dominance relation

$F_{U|Z,S}(u|z^*,s) \leq F_{U|Z,S}(u|z,s)$ for a.e. $u \in \mathcal{U}_{z,s}$ holds (e.g. the probability of low ability U is small when in proximity to a college) then $B(0,1|\nu(z,s) < U_X \leq \nu(z^*,s),z^*,s)$ is nonnegative.

Under conditional exogeneity, $B(0,1|\nu(z,s),z,s) = 0$ and therefore $R^{LIV}_{Y.X|Z}(z;s)$ identifies $\bar{\beta}(0,1|\nu(z,s),z,s)$. This occurs if $U \perp Z|S = s$ or $\bar{\delta}_Y(u;z|s) = 0$ for a.e. $u \in \mathcal{U}_{z,s}$. Alternatively, under conditional proportional confounding, $\bar{\delta}_Y(u;z|s) = d(z,s)\bar{\delta}_W(u|s)$ for a.e. $u \in \mathcal{U}_{z,s}$, with $d(z,s)$ known or estimable. In this case, $\bar{\beta}(0,1|\nu(z,s),z,s) = R^{LIV}_{Y.X|Z}(z;s) - R^{LIV}_{W.X|Z}(z;s)d(z,s)$. Analogous results hold for $\bar{\beta}(0,1|\nu(z,s) < U_X \leq \nu(z^*,s),z^*,s)$.

In the absence of conditions sufficient for point identification, magnitude and/or sign restrictions on confounding yield sharp bounds.

Corollary 6.3 *Suppose that, for a.e.* $u \in \mathcal{U}_{z,s}$, $\bar{\delta}_Y(u;z|s) = d(u,z,s)\bar{\delta}_W(u|s)$ *with* $d(u,z,s) \in \mathcal{D}(z,s) \equiv [d_L(z,s),d_H(z,s)]$. *(i) Under the conditions of Theorem 6.2(i), if* $\bar{\delta}_W(u|s)[F_{U|Z,S}(u|z^*,s) - F_{U|Z,S}(u|z,s)]$ *is either nonpositive for a.e.* $u \in \mathcal{U}_{z,s}$ *or nonnegative for a.e.* $u \in \mathcal{U}_{z,s}$ *then*

$$\bar{\beta}(0,1|\nu(z,s) < U_X \leq \nu(z^*,s),z^*,s) \in \mathcal{B}(\mathcal{D}(z,s)) \equiv \{R^{Wald}_{Y.X|Z}(z,z^*;s) - R^{Wald}_{W.X|Z}(z,z^*;s)d : d \in \mathcal{D}(z,s)\},$$

and this identification region is sharp.
(ii) Under the conditions of Theorem 6.2(ii), if $\bar{\delta}_W(u|s)\frac{\partial}{\partial z}F_{U|Z,S}(u|z,s)$ *is either nonpositive for a.e.* $u \in \mathcal{U}_{z,s}$ *or nonnegative for a.e.* $u \in \mathcal{U}_{z,s}$ *then*

$$\bar{\beta}(0,1|\nu(z,s),z^*,s) \in \mathcal{B}(\mathcal{D}(z,s)) \equiv \{R^{LIV}_{Y.X|Z}(z;s) - R^{LIV}_{W.X|Z}(z;s)d : d \in \mathcal{D}(z,s)\},$$

and this identification region is sharp.

Using Corollary 6.3, full or partial identification of various average effects obtain by averaging over the relevant distributions. For example, averaging the bounds for $\bar{\beta}(0,1|\nu(z,s),z^*,s)$ over the distribution of Z given $S = s$ (recall $U_X \perp Z|S = s$) yields bounds for $\bar{\beta}(0,1|\nu(z,s),s)$. In turn, the latter bounds can be used to fully or partial identify e.g the average treatment effects for the population, treated, and untreated, as discussed in Section 6.2.

To conclude this section, we note that one can build on these nonparametric IV results to study the identification of various average effects under restrictions on confounding in structural systems with discrete or continuous X and possibly mismeasured potential instruments (see e.g. Schennach, White, and Chalak, 2012; Chalak, 2013). We omit the details of these extensions for brevity.

7 Estimation and Inference

7.1 Asymptotic Normality

We obtain a uniformly consistent set estimator $\hat{\mathcal{B}}$ for an identification region $\mathcal{B}$ obtained under restrictions on confounding by using uniformly consistent estimators for its bounds. In par-

ticular, consider a sharp identification region of the form $[b_L, b_H]$, a bounded interval of finite width. Note that b_L and b_H are linear transformations of covariate-conditioned (IV) regression estimands of Y and W on X (using instruments Z). For example, the bounds in the identification regions $\mathcal{B}([0,1])$ and $\mathcal{B}([-1,1])$ for $\bar{\beta}(x, x^*|x^*, s)$ are $R^N_{Y.X}(x, x^*; s)$, $R^N_{Y-W.X}(x, x^*; s)$, or $R^N_{Y+W.X}(x, x^*; s)$, and are therefore a linear transformation of $E(Y|X, S)$ and $E(W|X, S)$ evaluated at (x^*, s) and (x, s). Thus, one can construct estimators $(\hat{b}_L, \hat{b}_H)'$ for $(b_L, b_H)'$, and derive their joint asymptotic distribution, as a linear transformation of uniformly consistent and jointly asymptotically normal parametric, semiparametric, or nonparametric (e.g. kernel[21]) estimators for the underlying (IV) regression estimands.

To proceed, we focus on the specification in our empirical application in Section 8, which encompasses common specifications from the literature (e.g. Card, 1995). In this case, the elements of the vector $(X', Z', S')'$ are binary or discrete variables and the response Y and proxy W for scalar U are generated by

$$Y = \alpha_Y + g_X(X)'\gamma + U\delta_Y \quad \text{and} \quad W = \alpha_W + U\delta_W, \tag{8}$$

where we subsume the covariates into α_Y and α_W. In particular, 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. Thus, an average effect $\bar{\beta}(x, x^*)$ is encoded here by the linear transformation $[g_X(x^*) - g_X(x)]'\bar{\gamma}$ of $\bar{\gamma}$. As discussed in Section 4, when $\bar{\gamma}(S)$, $\bar{\delta}_Y(S)$, and $\bar{\delta}_W(S)$ are constant and $\bar{H}_Z(S)$ or/and $\bar{G}_X(S)$, $\bar{W}(S)$, and $\bar{Y}(S)$ are affine functions of G_S, applying 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 $H \equiv (H'_Z, G'_S)'$. In this case, the bounds on $\bar{\beta}(x, x^*)$ under restrictions on confounding are linear transformations of $(R'_{Y.G|H}, R'_{W.G|H})'$. For instance, if X_j is the j^{th} component of G_X and the effect β_j of X_j on Y is linear then $\bar{\beta}_j = \bar{\gamma}_j$ and the bounds in the identification regions $\mathcal{B}_j([0,1])$ and $\mathcal{B}_j([-1,1])$ for $\bar{\beta}_j$ are $R_{Y.G|H,j}$, $R_{Y-W.G|H,j}$, or $R_{Y+W.G|H,j}$. Thus, we begin by deriving the joint asymptotic distribution of the plug-in estimators $(\hat{R}'_{Y.G|H}, \hat{R}'_{W.G|H})'$ for $(R'_{Y.G|H}, R'_{W.G|H})'$; this encompasses the case in which $H = G$.

Given observations $\{A_i\}_{i=1}^n$ of a $d \times 1$ vector A, we let $\tilde{A}_i \equiv A_i - \frac{1}{n}\sum_{i=1}^n A_i$. Further, for observations $\{A_i, B_i, C_i\}_{i=1}^n$ corresponding to A and random vectors B and C of equal dimension, we denote the linear IV regression estimator and sample residuals by:

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

The next theorem employs standard arguments to derive the asymptotic distribution of

[21]Sometimes, the vector $(X', S')'$ (or $(Z', S')'$) may be a high-dimensional vector of binary and discrete variables, as in our empirical application, in which case one can "smooth" discrete regressors (e.g. Li and Racine, 2007) since there may be cells with few or no data points in a given sample.

$\sqrt{n}((\hat{R}'_{Y.G|H}, \hat{R}'_{W.G|H})' - (R'_{Y.G|H}, R'_{W.G|H})')$. For this, we let $Q \equiv diag(E(\tilde{H}\tilde{G}'), E(\tilde{H}\tilde{G}'))$. We maintain that W is scalar, as in the empirical application, to simplify the notation.

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

$$\Xi = \begin{bmatrix} E(\tilde{H}\epsilon^2_{Y.G|H}\tilde{H}') & E(\tilde{H}\epsilon_{Y.G|H}\epsilon_{W.G|H}\tilde{H}') \\ E(\tilde{H}\epsilon_{W.G|H}\epsilon_{Y.G|H}\tilde{H}') & E(\tilde{H}\epsilon^2_{W.G|H}\tilde{H}') \end{bmatrix}$$

is finite and positive definite uniformly in $P \in \mathcal{P}$.

Then $\Lambda \equiv Q^{-1}\Xi Q'^{-1}$ is finite and positive definite uniformly in $P \in \mathcal{P}$, and uniformly in $P \in \mathcal{P}$

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

We refer the reader to e.g. Shorack (2000) and Imbens and Manski (2004, lemma 5) for primitive conditions ensuring the uniform law of large numbers and central limit theorem in assumptions (i, ii) of Theorem 7.1.

As discussed above, the joint asymptotic distribution of the estimators for the bounds on $\bar{\beta}(x, x^*) = [g_X(x^*) - g_X(x)]'\bar{\gamma}$ under restrictions on confounding obtains as a linear transformation of that of $\sqrt{n}(\hat{R}'_{Y.G|H}, \hat{R}'_{W.G|H})'$. For example, consider a linear effect $\bar{\beta}_j = \bar{\gamma}_j = R_{Y.G|H,j} - R_{W.G|H,j}\bar{\delta}$ and include in $R_1 \equiv (R'_{Y.G|H}, R'_{Y-W.G|H})'$ the bounds of the identification region $\mathcal{B}_j([0, 1])$ and in $R_2 \equiv (R'_{Y-W.G|H}, R'_{Y+W.G|H})'$ those of $\mathcal{B}_j([-1, 1])$. Since the corresponding IV plug-in estimators $\hat{R}_1$ and $\hat{R}_2$ are linear transformations of $(\hat{R}'_{Y.G|H}, \hat{R}'_{W.G|H})'$, it follows from Theorem 7.1 that, uniformly in $P \in \mathcal{P}$,

$$\sqrt{n}(\hat{R}_1 - R_1) \xrightarrow{d} N(0, \Sigma_1) \quad \text{and} \quad \sqrt{n}(\hat{R}_2 - R_2) \xrightarrow{d} N(0, \Sigma_2),$$

with Σ_1 and Σ_2 finite and positive definite uniformly in $P \in \mathcal{P}$, and given by

$$\Sigma_1 = \begin{bmatrix} \sigma^2_{11} & \sigma^2_{12} \\ \sigma^2_{21} & \sigma^2_{22} \end{bmatrix} \quad \text{and} \quad \Sigma_2 = \begin{bmatrix} \sigma^2_{22} & \sigma^2_{23} \\ \sigma^2_{32} & \sigma^2_{33} \end{bmatrix}$$

with $\sigma^2_{ab} \equiv E(\tilde{H}\tilde{G}')^{-1}E(\tilde{H}\epsilon_a\epsilon_b\tilde{H}')E(\tilde{G}\tilde{H}')^{-1}$ and $(\epsilon_1, \epsilon_2, \epsilon_3) \equiv (\epsilon_{Y.G|H}, \epsilon_{Y-W.G|H}, \epsilon_{Y+W.G|H})$.

Under regularity conditions (e.g. White, 1980, 2001), a uniformly in $P \in \mathcal{P}$ consistent heteroskedasticity robust estimator for a block σ^2_{ab} of an asymptotic covariance matrix is given by $\hat{\sigma}^2_{ab} \equiv (\frac{1}{n}\sum_{i=1}^n \tilde{H}_i\tilde{G}'_i)^{-1}(\frac{1}{n}\sum_{i=1}^n \tilde{H}_i\hat{\epsilon}_{a,i}\hat{\epsilon}_{b,i}\tilde{H}')(\frac{1}{n}\sum_{i=1}^n \tilde{G}_i\tilde{H}'_i)^{-1}$.

Uniformly consistent and jointly asymptotically normal parametric, semiparametric, or nonparametric estimators $(\hat{b}_L, \hat{b}_H)'$ for $(b_L, b_H)'$ can be derived analogously.

7.2 Confidence Intervals

This subsection discusses constructing a $1-\alpha$ confidence interval (CI) for an average effect $\bar{\beta}$ that is partially identified in a sharp bounded interval $[b_L, b_H]$ of finite width. Consider the estimators $(\hat{b}_L, \hat{b}_H)'$ such that $\sqrt{n}((\hat{b}_L, \hat{b}_H)' - (b_L, b_H)') \xrightarrow{d} N(0, \Sigma)$ uniformly in $P \in \mathcal{P}$, as follows from e.g. Theorem 7.1. Let $\hat{\sigma}_L^2$ denote the uniformly in $P \in \mathcal{P}$ consistent estimator for $\sigma_L^2 \equiv Avar(\sqrt{n}(\hat{b}_L - b_L))$ and define $\hat{\sigma}_H^2$ similarly. We construct[22] a $1-\alpha$ CI for $\bar{\beta}$ using:

$$[\hat{b}_L - c_\alpha \frac{\hat{\sigma}_L}{\sqrt{n}},\ \hat{b}_H + c_\alpha \frac{\hat{\sigma}_H}{\sqrt{n}}],$$

with the critical value c_α, discussed next, such that

$$\Phi(c_\alpha + \frac{\sqrt{n}(\hat{b}_H - \hat{b}_L)}{\max\{\hat{\sigma}_L, \hat{\sigma}_H\}}) - \Phi(-c_\alpha) = 1 - \alpha. \tag{9}$$

To illustrate, consider a linear average effect $\bar{\beta}_j = \bar{\gamma}_j$ in equations (8) from the empirical application, and consider a $1-\alpha$ CI for the average effect $\bar{\beta}_j \in \mathcal{B}_j([-1,1])$. Then $[b_L, b_H] = [R_{Y.G|H,j} - \left|R_{W.G|H,j}\right|,\ R_{Y.G|H,j} + \left|R_{W.G|H,j}\right|]$. Picking c_α such that $\Phi(c_\alpha) - \Phi(-c_\alpha) = 1-\alpha$, where Φ denotes the standard normal cumulative density function, (e.g. $c_{0.05} = 1.96$) yields a $1-\alpha$ confidence interval $CI_{\mathcal{B}_j, 1-\alpha}$ for the identification region $\mathcal{B}_j([-1,1])$. However, $CI_{\mathcal{B}_j, 1-\alpha}$ is a conservative CI for $\bar{\beta}_j \in \mathcal{B}_j$ since when $\mathcal{B}_j$ has positive width, $\bar{\beta}_j$ can be close to at most b_L or b_H. Further, as discussed in Imbens and Manski (2004), picking c_α such that $\Phi(c_\alpha) - \Phi(-c_\alpha) = 1 - 2\alpha$ (e.g. $c_{0.05} = 1.645$) yields a CI whose coverage probabilities do not converge to $1-\alpha$ uniformly across different widths of $\mathcal{B}_j([-1,1])$, e.g. for $R_{W.G|H,j} = 0$ with point identification. Instead, to account for the estimated width of the identification interval, we construct the uniformly valid confidence interval $CI_{\bar{\beta}_j, 1-\alpha}$ for $\bar{\beta}_j \in \mathcal{B}_j$ with c_α as in equation (9). For $\bar{\beta}_j = \bar{\gamma}_j$ in $\mathcal{B}_j([-1,1])$, by construction, $\hat{b}_H - \hat{b}_L = 2\left|R_{W.G|H,j}\right| \geq 0$ and it follows from lemma 4 in Imbens and Manski (2004) and lemma 3 and proposition 1 in Stoye (2009) that the confidence interval $CI_{\bar{\beta}_j, 1-\alpha}$ is uniformly valid for $\bar{\beta}_j$ in $\mathcal{B}_j([-1,1])$.

In the empirical application, in addition to $\hat{\mathcal{B}}_j([-1,1])$, we report an estimate of the half as large sharp identification region $\mathcal{B}_j([0,1])$ obtained under magnitude and sign restrictions on confounding, and a CI for $\bar{\beta}_j = \bar{\gamma}_j$ that is partially identified in this set. Note that, unlike for $\mathcal{B}_j([-1,1])$, this identification region depends on $sign(R_{W.G|H,j})$ which can be estimated. We leave studying the consequences of estimating $R_{W.G|H,j}$ to other work to keep the scope of this paper manageable. Here, we follow the literature (e.g. Reinhold and Woutersen, 2009; Nevo and Rosen, 2012) and report estimated identification intervals such as $\hat{\mathcal{B}}_j([0,1]\,|sign(R_{W.G|H,j}) = sign(\hat{R}_{W.G|H,j}))$ for $\bar{\beta}_j = \bar{\gamma}_j$ and the confidence interval $CI_{\bar{\beta}_j, 1-\alpha}(sign(R_{W.G|H,j}) = sign(\hat{R}_{W.G|H,j}))$

[22] An alternative method considers the union over confidence intervals for $\bar{\beta}(\bar{\delta})$ generated for each $\bar{\delta} \in [d_L, d_H]$ as in Chernozhukov, Rigobon, and Stoker (2010).

under the assumption that $sign(R_{W.G|H,j}) = sign(\hat{R}_{W.G|H,j})$. In addition, we indicate the p-value for a t-test for the null hypothesis $R_{W.G|H,j} = 0$ against the alternative hypothesis $sign(R_{W.G|H,j}) = sign(\hat{R}_{W.G|H,j})$. When the p-value for this one-sided test is larger than $\frac{1}{2}\alpha$, one cannot reject the null hypothesis $R_{W.G|H,j} = 0$ against the alternative $R_{W.G|H,j} \neq 0$ at the α significance level, or that, under the maintained assumptions, $R_{Y.G|H,j}$ identifies $\bar{\beta}_j = \bar{\gamma}_j$.

8 Return to Education and the Black-White Wage Gap

We apply this paper's method to study the financial return to education and the black-white wage gap. Card (1999) surveys several studies measuring the causal effect of education on earning. 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%. Although these IV estimates are higher than the surveyed regression estimates, which range from 5.2% to 8.5%, they are less precise with standard errors sometimes as large as nearly half the IV point estimates. 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 measurement error in reported education) ranging from 2.4% to 11%. See Card (1999, section 4 and tables 4 and 6) for a detailed account. Many studies document a black-white wage gap and try to understand 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 primarily reflects 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 education as well as test scores 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) for studies of the black-white wage gap and its causes.

We consider the wage and proxy equations (8) discussed in Section 7:

$$Y = \alpha_Y + g_X(X)'\gamma + U\delta_Y \quad \text{and} \quad W = \alpha_W + U\delta_W,$$

where $G_X \equiv g_X(\cdot)$ is a vector of flexible functions (e.g. power, threshold crossing) of X discussed below, so that the average effect $\bar{\beta}(x, x^*)$ is encoded by the linear transformation $[g_X(x^*) - g_X(x)]'\bar{\gamma}$ of $\bar{\gamma}$. Here, Y denotes the logarithm of hourly wage and X consists of completed years of education, years of experience, and a binary variable taking the value 1 if a person is black and 0 otherwise. We subsume into the random intercepts α_Y and α_W the covariates S that we discuss shortly. The confounder U denotes unobserved skill or "ability" and is potentially correlated with elements of G_X given S. The proxy W for U denotes the logarithm of the Knowledge of the

World of Work (KWW) test score, a test of occupational information. We use data drawn from the 1976 subset of the National Longitudinal Survey of Young Men (NLSYM), described in[23] Card (1995). The sample used in Card (1995) contains 3010 observations on individuals who reported valid wage and education. We drop 47 observations (1.56% of the total observations) with missing KWW score[24], as in some results in Card (1995), leading to a sample size of 2963. As in Card (1995), the covariates S consist of an indicator for living in the South and another for living in a metropolitan area (SMSA), 8 indicators for region of residence in 1966 and 1 for residence in SMSA in 1966, imputed[25] father and mother education plus 2 indicators for missing father and mother education respectively, 1 indicator for the presence of the father and mother at age 14 and another for having a single mother at age 14. We employ a vector $G_S \equiv g_S(S)$ of functions of covariates that contains, in addition to S, 8 binary indicators for interacted mother and father high school, college, or post graduate education. The sample also contains data on potential instruments Z that we consider below.

Although our identification results do not require this, we find it instructive to employ the specification in equations (8) since it generalizes specifications for the wage equation that are common in the literature (e.g. Card, 1995) by allowing for unobserved confounders and nonlinear random effects, thereby facilitating comparing our findings to the literature. Further, this parsimonious specification facilitates comparing the slope coefficients on the unobserved confounder U in the Y and W equations while maintaining the commonly used (e.g. Card, 1995) log-level specification for the wage equation. Specifically, $100\bar{\delta}_Y\%$ and $100\bar{\delta}_W\%$ are semi-elasticities, i.e. the ceteris paribus average approximate percentage changes in wage and KWW respectively due to a unit or percentile increase in U.

We apply Theorem 4.1 to equations (8) with G_X and a vector of instruments $H_Z \equiv h_Z(Z)$ (recall H_Z may equal G_X) replacing X and Z respectively. We maintain that $\bar{\gamma}(S)$, $\bar{\delta}_Y(S)$, and $\bar{\delta}_W(S)$ are constants and that $\bar{H}_Z(S)$ or/and $\bar{G}_X(S)$ (if $G_X \neq H_Z$), $\bar{W}(S)$, and $\bar{Y}(S)$ are affine functions of G_S. Putting $G \equiv (G_X', G_S')'$ and $H \equiv (H_Z', G_S')'$, this characterizes the components $\bar{\gamma}_j = R_{Y.G|H,j} - R_{W.G|H,j}\bar{\delta}$ of $\bar{\gamma}$, and therefore the effects $\bar{\beta}$. When studying magnitude and sign restrictions on confounding, we maintain the following assumptions. First, we assume that KWW is, on average, at least as directly elastic or responsive to ability than wage is,

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

[24] The sample also contains IQ score. However, we do not employ IQ as a proxy here since 949 observations (31.5% of the total observations) report missing IQ score. Using the available observations, the sample correlation between IQ and KWW is 0.43 and is strongly significant. Further, using the available observations, employing $\log(IQ)$ instead of $\log(KWW)$ as proxy often leads to tighter bounds and confidence intervals. This, however, could be partly due to sample selection.

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

$\left|\bar{\delta}_Y\right| \leq \left|\bar{\delta}_W\right|$. Thus, a ceteris paribus change in U leads, on average, to a direct percentage change in KWW that is at least as large as that in wage. This is a weakening of exogeneity, which would require $\bar{\delta}_Y = 0$ when U depends freely on G_X (or H_Z) given S. As discussed in Section 2.4.1, the assumption $\left|\bar{\delta}_Y\right| \leq \left|\bar{\delta}_W\right|$ is in accord with several theoretical and empirical findings that suggest that the average direct effects of U on Y may be modest. One may further weaken this assumption by assuming $\left|\bar{\delta}_Y\right| \leq d\left|\bar{\delta}_W\right|$ for known $d > 1$, leading to qualitatively similar but larger and implausible identification regions as we discuss below. Second, we sometimes assume that, on average, ability directly affects KWW and wage in the same direction, $0 \leq \frac{\bar{\delta}_Y}{\bar{\delta}_W}$. 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 conditional correlation between $\log(KWW)$ and education is positive, which often holds.

We begin by letting $G_X \equiv g_X(X)$ consists of education, experience, experience squared, and the black binary indicator, as in the specification in Card (1995, e.g. table 2, column 5). This assumes a linear return to education β_1 encoded in the components γ_1 of γ whereas γ_4 encodes the black-white wage gap β_3. In this case, the average approximate financial return to education is $100\bar{\gamma}_1\%$ and the average approximate black-white wage gap[26] 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 regression estimates using $\hat{R}_{Y.G,j}$, which consistently estimates $\bar{\gamma}_j$ under conditional exogeneity ($\frac{\bar{\delta}_Y}{\bar{\delta}_W} = 0$), along with heteroskedasticity-robust standard errors (s.e.) and 95% confidence intervals (denoted by $CI_{0.95}$). The regression estimates for the return to education and the black-white wage gap[27], with robust s.e. in parentheses, are 7.2%, (0.4%), and -18.7%, (2.0%), respectively. Column 2 reports estimates $\widehat{\mathcal{G}}_j([0,1] \mid sign(R_{W.G,j}) = sign(\hat{R}_{W.G,j}))$ of the sharp identification region for $\bar{\gamma}_j$ obtained under magnitude and sign restrictions on confounding, $0 \leq \frac{\bar{\delta}_Y}{\bar{\delta}_W} \leq 1$ (that are weaker than exogeneity $\frac{\bar{\delta}_Y}{\bar{\delta}_W} = 0$) along with the uniformly valid 95% confidence interval $CI_{\bar{\gamma}_j,0.95}(sign\ (R_{W.G,j}) = sign(\hat{R}_{W.G,j}))$ for $\bar{\gamma}_j$. The estimated identification region for the return to education is $[0.1\%, 7.2\%]$ with $CI_{\bar{\gamma}_1,0.95}$ $[-0.6\%, 7.8\%]$ and that for the black-white wage gap is $[-18.7\%, 1.5\%]$ with $CI_{\bar{\gamma}_4,0.95}$ $[-21.9\%, 5.0\%]$. We also report $\hat{R}_{W.G,j}$, whose magnitude is the estimated width of the identification region, along with its robust standard error and indicate whether a t-test rejects the null hypothesis $R_{W.G,j} = 0$ against the alternative hypothesis $sign(R_{W.G,j}) = sign(\hat{R}_{W.G,j})$ at the 10%, 5%, or 1% level. Last, column 4 reports estimates $\widehat{\mathcal{G}}_j([-1,1])$ of the twice as large identification region $\mathcal{G}_j([-1,1])$ for

[26]The coefficient on a binary (non-continuous) variable in a log-linear equation does not 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).

[27]In all tables, we also report point and interval estimates for the coefficients associated with experience and experience squared. For brevity, we don't discuss these in detail.

$\bar{\gamma}_j$ obtained under magnitude restrictions only, along with the uniformly valid 95% confidence intervals $CI_{\bar{\gamma}_j,0.95}$. Note that weakening the assumptions $\bar{\delta} \in [0,1]$ to $\bar{\delta} \in [0,d]$ (or $\bar{\delta} \in [-1,1]$ to $\bar{\delta} \in [-d,d]$) for $d > 1$, thereby allowing wage to be on average more sensitive to ability than the test score is, would extend the estimated identification regions to include a negative average return to education $\bar{\beta}_1$ and a black-white wage gap $\bar{\beta}_3$ in favor of blacks, which is inconsistent with the general findings in the literature. Thus, the empirical findings in this paper corroborate the assumption $\left|\bar{\delta}_Y\right| \leq \left|\bar{\delta}_W\right|$. Specifically, we have:

$$\hat{\mathcal{B}}_1([0,d] \mid sign(R_{W.G,1}) = sign(\hat{R}_{W.G,1})) \approx [7.2\% - (d \times 7\%),\ 7.2\%], \text{ and}$$
$$\hat{\mathcal{B}}_3([0,d] \mid sign(R_{W.G,4}) = sign(\hat{R}_{W.G,4})) \approx [-18.7\%,\ -18.7\% + (d \times 20.1\%)].$$

Last, we note that similar results (with slightly wider bounds for the average black-white wage gap) obtain when, as in Card (1995, table 2, column 1), we let $G_S = S_1$ with S_1 a subset of S consisting of two indicators for living in the South and SMSA respectively. In sum, we find that regression estimates provide an upper bound for the average (assumed linear for now) return to education as well as for the average black-white wage gap.

We briefly turn attention away from the results obtained under restrictions on confounding to report some complementary results. First, recall that conditioning on the proxy W for the confounder U, as is commonly done, may attenuate the regression bias but does not generally ensure recovering $\bar{\beta}$ from a regression of Y on $(1, G', W)'$ except in special cases such as when W is a perfect rescaling[28] of U given S. Table 2 reports the "perfect proxy" estimates from a linear regression of Y on $(1, G', W')'$. These estimates and the corresponding confidence intervals lie respectively within the identification regions and $CI_{\bar{\gamma}_j,0.95}$ reported in Table 1. In particular, 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. Further, the coefficient on W estimates $\frac{\bar{\delta}_Y}{\bar{\delta}_W}$ to be 0.2 with robust s.e. 0.03. Note that this estimate is also the lower bound on $\frac{\bar{\delta}_Y}{\bar{\delta}_W}$ if one assumes that W measures U with classical measurement error[29] (see e.g. Klepper and Leamer (1984) and Bollinger (2003)) but the corresponding estimate for the upper bound on $\frac{\bar{\delta}_Y}{\bar{\delta}_W}$ is very large in this case, allowing for $\bar{\beta}_j$ values that are inconsistent with the literature (e.g. large negative return to education and a large wage gap in favor of blacks). Last, as discussed in Section 4 and footnote 14, we also study fully identifying $(\bar{\gamma}', \frac{\bar{\delta}_Y}{\bar{\delta}_W})'$ (and thus $\bar{\beta}$) via conditional IV regressions of Y on $(1, G_X', W)'$ using functions of $(G_X', S')'$ as

[28] For example, it suffices in equations (8) that γ, δ_Y, and δ_W are constants, and $\alpha_Y = G_S'\bar{\psi}_Y + \eta_Y$ and $\alpha_W = G_S'\bar{\psi}_W$ where $\bar{\psi}_Y$ and $\bar{\psi}_W$ are constant vectors and $Cov(\eta_Y, (G', U)') = 0$.

[29] Specifically, this assumes that γ, δ_Y, and δ_W are constants and that $\alpha_Y = G_S'\bar{\psi}_Y + \eta_Y$ and $\alpha_W = G_S'\bar{\psi}_W + \eta_W$ where $\bar{\psi}_Y$ and $\bar{\psi}_W$ are constant vectors and $Cov(\eta_Y, (G_X', G_S', U)') = 0$ (correctly specified Y equation) and $Cov(\eta_W, (G_X', G_S', U, \eta_Y)') = 0$ (classical measurement error). The bounds on $\frac{\bar{\delta}_Y}{\bar{\delta}_W}$ are the coefficient on W in a linear regression or Y on $(1, G', W)'$ and the inverse of the coefficient on Y in a linear regression of W on $(1, G', Y)'$.

instruments. In particular, we consider excluding parental education variables (e.g. mother or father education) from G_S and using these as instruments for W. However, this yields unstable (e.g. varying depending on whether the mother's or father's education is used as instrument) and imprecise estimates. Also, we estimate $(\bar{\gamma}', \frac{\bar{\delta}_Y}{\bar{\delta}_W})'$ via a two stage least squares regression of Y on $(1, G_X', W, G_S)'$ with G_S restricted to S_1, consisting of the South and SMSA indicators, and the instruments for W consisting of the interactions of G_X and S_1, i.e. the product of G_X with each of the indicators $1\{S_1 = s_1\}$ for $s_1 = (0,1), (1,0)$, and $(1,1)$. This estimates $\frac{\bar{\delta}_Y}{\bar{\delta}_W}$ to be 0.54 with $CI_{0.95}$ $[0.18, 0.90]$, the return to education to be 3.4% with $CI_{0.95}$ $[0.6\%, 6.1\%]$, and the black-white wage gap to be -6.6% with $CI_{0.95}$ $[-15.6\%, 2.3\%]$. Similar results obtain when we instrument for (G_X,W) using the interactions of G_X and S_1 or when conditioning on the fuller set of covariates G_S and augmenting the vector of instruments with the product of G_X and an indicator for low parental education[30].

Returning to the restrictions on confounding results, we augment G_X to include, as its second component, an interaction term $(Education - 12) \times Black$ which multiplies the black binary indicator with years of education minus 12. Table 3 reports[31] the conditional on G_S results. Under magnitude and sign restrictions on confounding, the estimates for the sharp identification region for the average return to education for non-blacks is $[0.4\%, 6.8\%]$ with $CI_{\bar{\gamma}_1, 0.95}$ $[-0.4\%, 7.5\%]$, that for the average black-white return to education differential is $[-1.2\%, 1.7\%]$ with $CI_{\bar{\gamma}_2, 0.95}$ $[-2.4\%, 2.8\%]$, and that for the average black-white wage gap, corresponding to individuals with 12 years of education, is $[-19.3\%, 1.8\%]$ with $CI_{\bar{\gamma}_5, 0.95}$ $[-22.5\%, 5.5\%]$. Thus, the average return to education for the black subpopulation may differ slightly from the non-black subpopulation, if at all. We follow Card (1995) and maintain that these average returns are equal.

When $H_Z = G_X$, Theorem 4.1 assumes that education is conditionally uncorrelated with unobserved determinants of wage and KWW other than ability. This assumption can fail if X is mismeasured or if e.g. test taking skills drive KWW and are correlated with education given the covariates. As in Card (1995), we also employ an indicator for the presence of a four year college in the local labor market, age, and age squared as potential instruments[32] for education, experience, and experience squared in the specification from Table 1. For this paper's method, this assumes for example that proximity to a college is conditionally uncorrelated with unobserved determinants of wage and KWW other than ability. This can fail if e.g. access to counseling drives KWW and is correlated with proximity to a college given the covariates. However, this paper's method does not require H_Z to be conditionally uncorrelated with abil-

[30] This indicator is 1 if neither parent has 12 or more years of education, and 0 otherwise.

[31] Similar results obtain when conditioning on a subset of the covariates, $G_S = S_1$.

[32] 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.

ity (for example, Carneiro and Heckman (2002) provide evidence suggesting that distance to college may be endogenous). As reported in Table 4, under magnitude and sign restrictions on confounding, the covariate-conditioned 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 $CI_{\bar{\gamma}_1,0.95}$ [−6.1%, 22%]. Similarly, the estimated identification region for the average black-white wage gap is [−16.2%, 2.6%], which is slightly tighter than the regression-based estimate albeit with comparable $CI_{\bar{\gamma}_4,0.95}$ [−20.9%, 7.6%]. Further, 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, age, and age squared as potential instruments for education, experience, and experience squared. Last, in both IV specifications, conditioning on a subset $G_S = S_1$ of the covariates yields generally similar results. This sometimes leads to tighter identification regions albeit with possibly wider confidence intervals (e.g. [−10.1%, 2.4%] with $CI_{\bar{\gamma}_4,0.95}$ [−22.4%, 15%] for the average black-white wage gap in the first IV specification and [−0.6%, 8.1%] with $CI_{\bar{\gamma}_1,0.95}$ [−2.6%, 9.9%] for the average return to education in the second IV specification). In sum, the IV-based estimated identification regions are generally wider than, or comparable to, the above regression-based ones and have especially wider confidence intervals.

Last, this paper's method does not require linear or parametric effects of X on Y. Next, we relax the linearity of the return to education in the previous specification and let G_X contain binary indicators for having at least t years of education, where $t = 2, ..., 18$ as in the sample, instead of total years of education, thus allowing for year-specific incremental return to education. In particular, γ_t encodes the incremental return $\beta(t, t+1)$ to year $t+1$ of education. As reported in Table 5, here too regression estimates generally give an upper bound on the average return to education and the average black-white wage gap[33]. We find evidence[34] for 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, under magnitude and sign restrictions on confounding, the estimate of the identification region for the average return to the 12^{th} year is [1.6%, 14.6%] with $CI_{\bar{\gamma}_{11},0.95}$ [−4.2%, 20%] and that for the 16^{th} year is [13.3%, 19.5%] with $CI_{\bar{\gamma}_{15},0.95}$ [7.5%, 25.1%]. Similarly, the estimated identification region for the return to the 18^{th} year is [13.9%, 14.9%] with $CI_{\bar{\gamma}_{17},0.95}$ [5.5%, 23.3%] and we cannot reject at comfortable significance levels the hypothesis that the width of this region is zero or, under the maintained assumptions, that regression consistently estimates this return

[33] Similar results obtain when we let $G_S = S_1$.

[34] We don't provide a formal test for linearity in partial identification setups to maintain a manageable scope of this paper. However, we note that, under magnitude and sign restrictions on confounding, the 95% CI for the partially identified return to the 16^{th} or 18^{th} year of education does not overlap with the 95% CI for the partially identified return to e.g. the 15^{th} year and barely overlaps with that of the 17^{th} year.

(the regression estimate is 14.9% with robust s.e. 4.5%). In contrast, the estimated identification region for the return to the 13^{th} year is smaller, $[0.7\%, 7.8\%]$ with $CI_{\bar{\gamma}_{12},0.95}$ $[-3.4\%, 11.6\%]$. Graphs 1 and 2 illustrate the nonlinearity in the return to education; these plot the estimates of the sharp identification regions 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 magnitude and sign restrictions on confounding (Graph 1) as well as magnitude restrictions only (Graph 2). Last, the estimate of the sharp identification region for the black-white wage gap under magnitude and sign restrictions using this specification is similar to that in Table 1 and given by $[-17.8\%,\ 1.9\%]$ with $CI_{\bar{\gamma}_{20},0.95}$ $[-21\%,\ 5.4\%]$.

In sum, the estimated bounds for the black-white wage gap are relatively wide, suggesting that, under the imposed weaker than exogeneity assumptions, this data 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. This nonlinearity may partly explain why, contrary to the expected direction of ability bias, linear IV estimates of the average return to education often exceed linear regression estimates. In particular, both types of estimates are weighted averages of yearly incremental returns for different subpopulations and the large IV estimates may reflect the relatively high return to graduation years for the subpopulation whose graduation outcomes depend on potential instruments such as proximity to college (see e.g. Card 1995, 1999).

This empirical analysis imposes assumptions including, at times, linearity or separability among observables and the confounder, restrictions on the random coefficients, the presence of one confounder U denoting "ability" which we proxy using $\log(KWW)$, and the assumptions $0 \leq \frac{\bar{\delta}_Y}{\bar{\delta}_W} \leq 1$ 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. In general, if other confounders are present, and strong valid instruments or proxies for these are not available, then additional assumptions are needed to (partially) identify average effects. Nevertheless, this empirical analysis does not require several commonly employed assumptions. In particular, (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 test scores to be error-laden proxies for unobserved ability, with possibly nonclassical measurement error.

9 Conclusion

This paper studies measuring average causal effects in structural systems without conditional exogeneity of causes, treatment, or instruments given covariates. In particular, we study full or partial identification of covariate-conditioned average random coefficients, average nonpara-

metric discrete and marginal effects, local and marginal treatment effects as well as average treatment effects for the population, treated, and untreated. First, the paper characterizes the omitted variable bias, due to unobserved confounders U, of regression and IV methods for the identification of these various average effects, thereby generalizing the classic linear regression omitted variable bias representation. This enables reasoning about the direction of the omitted variable bias in general structures. Second, using proxies W for the confounders U, the paper demonstrates how restrictions on the magnitude and sign of confounding can be used to fully or partially identity average effects when stronger identifying assumptions are suspected to fail or to conduct a sensitivity analysis to deviations from stronger assumptions otherwise. In particular, we ask how do the average direct effects of U on Y compare in magnitude and sign to those of U on W. Exogeneity (zero average direct effect) and proportional confounding (equal to a known proportion direct effects) are examples of limiting cases yielding full identification of the average effects of X on Y. Alternatively, the effects of X on Y are partially identified in sharp bounded intervals when W is sufficiently sensitive to U, and sharp upper or lower bounds may obtain otherwise. After studying estimation and inference, the paper applies its methods, using data from the 1976 subset of NLSYM used in Card (1995), to partially identify in sharp bounded intervals the average financial incremental return to education as well as the average black-white wage gap. Under magnitude and sign restrictions on confounding, we find that regression estimates provide an upper bound on the average return to education and the black-white wage gap. Further, the regression-based bounds estimates are generally narrower than the IV-based ones, with especially narrower confidence intervals. Our findings suggest a nonlinearity in the return to education with the 12^{th}, 16^{th}, and 18^{th} years, corresponding to obtaining a high school, college, and possibly a graduate degree, yielding a high average return. Also, we find that, under the imposed weaker than exogeneity assumptions, this data 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. Extensions for future work include imposing distributional restrictions on confounding (e.g. imposing a prior distribution on $\bar{\delta}$) as well as employing restrictions on confounding to identify the distribution of the effect of X on Y or features of it other than the mean.

A Appendix A: Mathematical Proofs

Proof of Theorem 4.1 (i) By $(i.b)$ we have

$$Cov(Z, Y|S = s) = Cov(Z, X|S = s)\bar{\beta}(s) + Cov(Z, U|S = s)\bar{\delta}_Y(s),$$

and thus by $(i.a)$

$$R_{Y.X|Z}(s) = \bar{\beta}(s) + R_{U.X|Z}(s)\bar{\delta}_Y(s).$$

(ii) By $(ii.b)$ we have

$$Cov(Z, W|S = s) = Cov(Z, U|S = s)\bar{\delta}_W(s),$$

and thus by $(i.a)$ and $(ii.a)$

$$R_{W.X|Z}(s)\bar{\delta}(s) = R_{U.X|Z}(s)\bar{\delta}_Y(s).$$

Proof of Corollary 4.2 The proof is immediate.

Proof of Corollary 4.3 The result follows since $\bar{\beta}_j$ is generated via the linear mapping $L : \mathcal{D}_1 \times ... \times \mathcal{D}_m \to \mathcal{B}_j$ given by $b = R_{Y.X|Z,j} - R_{W.X|Z,j}d$. The region $\mathcal{B}_j$ is sharp, i.e. for every $b \in \mathcal{B}_j$ there exists a degenerate vector $d = d_W^{-1}d_Y \in \times_{h=1}^m \mathcal{D}_h$ where d_Y and d_W, being degenerate, satisfy the conditions on δ_Y and δ_W in Theorem 4.1 respectively; e.g. set $d_W = I$ so that $d = d_Y$. In particular, since $\mathcal{D}_1 \times ... \times \mathcal{D}_m$ is connected, $\mathcal{B}_j$ is totally ordered, and L is continuous, the generalized intermediate value theorem gives that for every $b \in \mathcal{B}_j$ there exists $d \in \times_{h=1}^m \mathcal{D}_h$ such that $L(d) = b$ (see e.g. Rudin, 1976, p. 93).

Proof of Theorem 5.1 Let $x, x^* \in \mathcal{X}$. (i) By $(i.a)$

$$E(Y|X = x, S = s) = E[\ddot{r}(x, s, U_Y)|S = s] + E(U'|X = x, S = s)\ \bar{\delta}_Y(s),$$

and thus

$$R_{Y.X}^N(x, x^*; s) = \bar{\beta}(x, x^*|s) + R_{U.X}^N(x, x^*; s)\bar{\delta}_Y(s).$$

By $(i.b.2)$ we have

$$E(W'|X = x, S = s) = E(\alpha'_W|S = s) + E(U'|X = x, S = s)\bar{\delta}_W(s),$$

and thus, by $(i.b.1)$, we have

$$R_{W.X}^N(x, x^*; s)\bar{\delta}(s) = R_{U.X}^N(x, x^*; s)\bar{\delta}_Y(s).$$

(ii) By $(ii.a.2)$, $\frac{\partial}{\partial x}E[\ddot{r}(x, s, U_Y)|S = s] = E[\frac{\partial}{\partial x}\ddot{r}(x, s, U_Y)|S = s]$ (see e.g. White and Chalak, 2013, Theorem 4.2), and $(i.a)$ and $(ii.a.1)$ yield

$$R_{Y.X}^N(x; s) = \bar{\beta}(x|s) + R_{U.X}^N(x; s)\bar{\delta}_Y(s).$$

By $(i.b)$ and $(ii.a.1)$, we have

$$R^N_{W.X}(x;s)\bar{\delta}(s) = R^N_{U.X}(x;s)\bar{\delta}_Y(s).$$

We make use of the following regularity conditions in the proof of Theorem 5.2. For this, we let $\bar{r}(x,u|s) \equiv E[r(x,s,u,U_Y)|S=s]$ and $\bar{q}(u|s) \equiv E[q(s,u,U_W)|S=s]$. It is implicitly assumed that referenced derivatives exist.

Assumption 5 (A.1) *Let $s \in \mathcal{S}$, $x, x^* \in \mathcal{X}$, and denote by $\mathcal{N}(u) \subseteq \mathcal{U}$ and $\mathcal{N}(x) \subseteq \mathcal{X}$ nonempty open neighborhoods of u and x respectively.*

(i.a) $E[r(x,s,U,U_Y)|X=x^*,S=s] < \infty$ and $E(Y|X=\ddot{x},S=s) < \infty$ for $\ddot{x} = x, x^*$,

(i.b) $\mathcal{U}_{x^*,s} = \mathcal{U}_{x,s}$,

(i.c) $\bar{r}(x,\cdot|s)$ is absolutely continuous on $\mathcal{U}_{x,s}$,

(i.d) for a.e. u and all $u^\dagger \in \mathcal{N}(u)$, $\bar{r}(x,u^\dagger|s) < \infty$ and there is a function $\Delta_{1,u}(u_y)$ with $E[\Delta_{1,u}(U_Y)|S=s] < \infty$ such that $\left|\frac{\partial}{\partial u} r(x,s,u^\dagger,u_y)\right| \leq \Delta_{1,u}(u_y)$ for a.e. u_y,

(i.e) $E(W|X=\ddot{x},S=s) < \infty$ for $\ddot{x} = x, x^*$,

(i.f) $\bar{q}(\cdot|s)$ is absolutely continuous on $\mathcal{U}_{x,s}$,

(i.g) for a.e. u and all $u^\dagger \in \mathcal{N}(u)$, $\bar{q}(u^\dagger|s) < \infty$ and there is a function $\Gamma_{1,u}(u_w)$ with $E[\Gamma_{1,u}(U_W)|S=s] < \infty$ such that $\left|\frac{\partial}{\partial u} q(s,u^\dagger,u_w)\right| \leq \Gamma_{1,u}(u_w)$ for a.e. u_w,

(ii.a) for all $x^\dagger \in \mathcal{N}(x)$, $\mathcal{U}_{x^\dagger,s} = \mathcal{U}_{x,s}$,

(ii.b) for all $x^\dagger \in \mathcal{N}(x)$, $\int_{\mathcal{U}_{x,s}} \bar{r}(x^\dagger,u|s) f_{U|X,S}(u|x^\dagger,s)du < \infty$ and there is a function $\Delta_2(u)$ with $\int_{\mathcal{U}_{x,s}} \Delta_2(u)du < \infty$ such that $\left|\frac{\partial}{\partial x}\{\bar{r}(x^\dagger,u|s) f_{U|X,S}(u|x^\dagger,s)\}\right| \leq \Delta_2(u)$ for a.e. u,

(ii.c) for a.e. u and all $x^\dagger \in \mathcal{N}(x)$, $\bar{r}(x^\dagger,u|s) < \infty$ and there is a function $\Delta_{3,u}(u_y)$ with $E[\Delta_{3,u}(U_Y)|S=s] < \infty$ such that $\left|\frac{\partial}{\partial x} r(x^\dagger,s,u,u_y)\right| \leq \Delta_{3,u}(u_y)$ for a.e. u_y,

(ii.d) for all $x^\dagger \in \mathcal{N}(x)$, $F_{U|X,S}(\cdot|x^\dagger,s)$ is absolutely continuous on $\mathcal{U}_{x,s}$ and there is a function $\Delta_4(u)$ with $\int_{\mathcal{U}_{x,s}} \Delta_4(u)du < \infty$ such that $\left|\frac{\partial}{\partial x} f_{U|X,S}(u|x^\dagger,s)\right| \leq \Delta_4(u)$ for a.e. u,

(ii.e) for all $x^\dagger \in \mathcal{N}(x)$, $\int_{\mathcal{U}_{x,s}} \bar{q}(u|s) f_{U|X,S}(u|x^\dagger,s)du < \infty$ and there is a function $\Gamma_2(u)$ with $\int_{\mathcal{U}_{x,s}} \Gamma_2(u)du < \infty$ such that $\left|\bar{q}(u|s)\frac{\partial}{\partial x} f_{U|X,S}(u|x^\dagger,s)\right| \leq \Gamma_2(u)$ for a.e. u.

The absolute continuity of $\bar{r}(x,\cdot|s)$ and $\bar{q}(\cdot|s)$ on $\mathcal{U}_{x,s}$ in A.1 ensures that $\frac{\partial}{\partial u}\bar{r}(x,\cdot|s)$ and $\frac{\partial}{\partial u}\bar{q}(\cdot|s)$ exist for a.e. u and are integrable. Assuming that derivatives are bounded almost everywhere by an integrable function justifies the interchange of derivative and integral.

Proof of Theorem 5.2: $(i.a)$ By A.1$(i.a)$, we have

$$\begin{aligned}
\bar{\beta}(x,x^*|x^*,s) &\equiv E[r(x^*,s,U,U_Y) - r(x,s,U,U_Y)|X=x^*,S=s] \\
&= R^N_{Y.X}(x,x^*;s) - \{E[r(x,s,U,U_Y)|X=x^*,S=s] - E[r(x,s,U,U_Y)|X=x,S=s]\} \\
&\equiv R^N_{Y.X}(x,x^*;s) - B(x,x^*|x^*,s).
\end{aligned}$$

Since $U_Y \perp (U,X)|S=s$, we have for $\ddot{x}=x,x^*$:

$$\begin{aligned}E[r(x,s,U,U_Y)|X=\ddot{x},S=s] &= E\{E[r(x,s,U,U_Y)|X=\ddot{x},U,S=s]|X=\ddot{x},S=s\}\\ &= E[\bar{r}(x,U|s)\ |X=\ddot{x},S=s].\end{aligned}$$

By A.1$(i.b)$ and since $F_{U|X,S}(\cdot|x^*,s)$ and $F_{U|X,S}(\cdot|x,s)$ are absolutely continuous:

$$B(x,x^*|x^*,s)=\int_{\mathcal{U}_{x,s}}\bar{r}(x,u|s)\ [f_{U|X,S}(u|x^*,s)-f_{U|X,S}(u|x,s)]du.$$

Given A.1$(i.c)$, integration by parts applies and gives

$$\begin{aligned}B(x,x^*|x^*,s)=\bar{r}(x,u|s)[F_{U|X,S}(u|x^*,s)-F_{U|X,S}(u|x,s)]\Big|_{\underline{u}}^{\bar{u}}\\ -\int_{\mathcal{U}_{x,s}}\frac{\partial}{\partial u}\bar{r}(x,u|s)[F_{U|X,S}(u|x^*,s)-F_{U|X,S}(u|x,s)]du,\end{aligned}$$

with $\underline{u}$ and $\bar{u}$ the (possibly infinite) infimum and supremum over $\mathcal{U}_{x,s}$. The first term vanishes and the result obtains since, by A.1$(i.d)$, $\frac{\partial}{\partial u}\bar{r}(x,u|s)=\bar{\delta}_Y(u;x|s)$ for a.e. u (see e.g. Corbae, Stinchcombe, and Zeman (2009, Theorem 7.5.17) or Bartle (1966, corollary 5.9)).
$(i.b)$ Similarly, $U_W \perp (U,X)|S=s$ and A.1$.i(b,e,f,g)$ give

$$\begin{aligned}R^N_{W.X}(x,x^*;s) &= \int_{\mathcal{U}_{x,s}}\bar{q}(u|s)\ [f_{U|X,S}(u|x^*,s)-f_{U|X,S}(u|x,s)]du\\ &= -\int_{\mathcal{U}_{x,s}}\bar{\delta}_W(u|s)[F_{U|X,S}(u|x^*,s)-F_{U|X,S}(u|x,s)]du.\end{aligned}$$

$(ii.a)$ Using $U_Y \perp (U,X)|S=s$, we have

$$\begin{aligned}R^N_{Y.X}(x;s) &= \frac{\partial}{\partial x}E\{E[r(x,s,U,U_Y)|X=x,U,S=s]|X=x,S=s\}\\ &= \frac{\partial}{\partial x}E[\bar{r}(x,U|s)|X=x,S=s]=\frac{\partial}{\partial x}\int_{\mathcal{U}_{x,s}}\bar{r}(x,u|s)\ f_{U|X,S}(u|x,s)du.\end{aligned}$$

By A.1$.ii(a,b)$, we interchange the order of derivative and integral and we apply the product rule since the derivatives exist by A.1$.ii(c,d)$:

$$R^N_{Y.X}(x;s)=\int_{\mathcal{U}_{x,s}}[\frac{\partial}{\partial x}\bar{r}(x,u|s)]f_{U|X,S}(u|x,s)du+\int_{\mathcal{U}_{x,s}}\bar{r}(x,u|s)[\frac{\partial}{\partial x}f_{U|X,S}(u|x,s)]du\equiv T_1+T_2.$$

By A.1$(ii.c)$ and $U_Y \perp (U,X)|S=s$, we have

$$T_1=\int_{\mathcal{U}_{x,s}}E[\frac{\partial}{\partial x}r(x,s,u,U_Y)|S=s]\ f_{U|X,S}(u|x,s)\ du=\bar{\beta}(x|x,s).$$

By A.1($ii.d$) and A.1($i.c$), integration by parts gives

$$T_2 = \bar{r}(x,u|s)\ \frac{\partial}{\partial x}F_{U|X,S}(u|x,s)\Big|_{\underline{u}}^{\bar{u}} - \int_{\mathcal{U}_{x,s}} \frac{\partial}{\partial u}\bar{r}(x,u|s)\frac{\partial}{\partial x}F_{U|X,S}(u|x,s)du$$
$$= -\int_{\mathcal{U}_{x,s}} \bar{\delta}_Y(u;x|s)\frac{\partial}{\partial x}F_{U|X,S}(u|x,s)du = B(x|x,s),$$

with $\underline{u}$ and $\bar{u}$ the (possibly infinite) infimum and supremum over $\mathcal{U}_{x,s}$ and where we use A.1($i.d$) and A.1($ii.a$) in the second equality.
($ii.b$) Similarly, $U_W \perp (U,X)|S=s$, A.1.$i(f,g)$, and A.1.$ii(a,d,e)$ give

$$R^N_{W.X}(x;s) = \frac{\partial}{\partial x}\int_{\mathcal{U}_{x,s}} \bar{q}(u|s)\ f_{U|X,S}(u|x,s)\ du = -\int_{\mathcal{U}_{x,s}} \bar{\delta}_W(u|s)\frac{\partial}{\partial x}F_{U|X,S}(u|x,s)du$$

Proof of Corollary 5.3: (i) Since $\bar{\delta}_Y(u;x|s) = d(u,x,s)\bar{\delta}_W(u|s)$ we have

$$B(x,x^*|x^*,s) = -\int_{\mathcal{U}_{x,s}} d(u,x,s)\bar{\delta}_W(u|s)[F_{U|X,S}(u|x^*,s) - F_{U|X,S}(u|x,s)]du.$$

Since $\bar{\delta}_W(u|s)[F_{U|X,S}(u|x^*,s) - F_{U|X,S}(u|x,s)]$ does not change sign for a.e. $u \in \mathcal{U}_{x,s}$, and

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

$d_L(x,s) \le d(u,x,s) \le d_H(x,s)$ gives

$$B(x,x^*|x^*,s) \in \{R^N_{W.X}(x,x^*;s)d : d \in \mathcal{D}(x,s)\},$$

The bounds then follow from

$$\bar{\beta}(x,x^*|x^*,s) = R^N_{Y.X}(x,x^*;s) - B(x,x^*|x^*,s).$$

For sharpness, for $R^N_{W.X}(x,x^*;s) \neq 0$ and each $b \in \mathcal{B}(\mathcal{D}(x,s))$, one can set $d(u,x,s)$ equal to $d(x,s) = \frac{1}{R^N_{W.X}(x,x^*;s)}(R^N_{Y.X}(x,x^*;s) - b) \in \mathcal{D}(x,s)$.

(ii) The proof is analogous to (i) and omitted.

Proof of Theorem 6.1: ($i.a$) By ($i.a$), we have

$$R^N_{Y.Z}(z,z^*;s) = E[\ddot{r}(\mathbf{1}\{U_X \le \nu(z^*,s)\},s,U_Y) - \ddot{r}(\mathbf{1}\{U_X \le \nu(z,s)\},s,U_Y)|S=s] + R^N_{U.Z}(z,z^*;s)\bar{\delta}_Y(s).$$

Arguments similar to those in the proof of Theorem 6.2($i.a$) give

$$E[\ddot{r}(\mathbf{1}\{U_X \le \nu(z^*,s)\},s,U_Y) - \ddot{r}(\mathbf{1}\{U_X \le \nu(z,s)\},s,U_Y)|S=s]$$
$$= \bar{\beta}(0,1|\nu(z,s) < U_X \le \nu(z^*,s),s) \times R^N_{X.Z}(z,z^*;s)$$

with $R^N_{X.Z}(z,z^*;s) = \Pr[\nu(z,s) < U_X \leq \nu(z^*,s) \,|S=s] > 0$, permitting dividing by $R^N_{X.Z}(z,z^*;s)$.
$(i.b)$ By $(i.b.2)$, for $\ddot{z} = z, z^*$:

$$E(W'|Z=\ddot{z}, S=s) = E(\alpha'_W|S=s) + E(U'|Z=\ddot{z}, S=s)\ \bar{\delta}_W(s).$$

Differencing this expression and applying $(i.b.1)$ gives

$$R^N_{W.Z}(z,z^*;s)\bar{\delta}(s) = R^N_{U.Z}(z,z^*;s)\bar{\delta}_Y(s).$$

The results obtains from division by $R^N_{X.Z}(z,z^*;s) \neq 0$.
$(ii.a)$ To characterize $\bar{\beta}(0,1|\nu(z,s),s)$, note that by $(ii.a.1)$

$$R^N_{Y.Z}(z;s) = \frac{\partial}{\partial z} E[\ddot{r}(\mathbf{1}\{U_X \leq \nu(z,s)\}, s, U_Y)|S=s] + R^N_{U.Z}(z;s)\bar{\delta}_Y(s)$$

The result obtains from division by $R^N_{X.Z}(z;s) \neq 0$ and since $(ii.a.2)$ and arguments similar to those in the proof of Theorem 6.2$(ii.a)$ give

$$\frac{\partial}{\partial z} E[\ddot{r}(\mathbf{1}\{U_X \leq \nu(z,s)\}, s, U_Y)|S=s] = \bar{\beta}(0,1|\nu(z,s),s) \times R^N_{X.Z}(z;s).$$

$(ii.b)$ The result obtains, after division by $R^N_{X.Z}(z;s) \neq 0$, since $(i.b.1)$ and $(ii.a.1)$ give

$$R^N_{W.Z}(z;s)\bar{\delta}(s) = R^N_{U.Z}(z;s)\bar{\delta}_Y(s).$$

For Theorem 6.2, we make use of regularity conditions collected in Assumption A.2. In what follows, we slightly abuse notation and write $\bar{r}(z,u|s) \equiv E[r(\mathbf{1}\{U_X \leq \nu(z,s)\}, s, u, U_Y)|S=s]$. It is implicitly assumed that referenced derivatives exist.

Assumption 6 (A.2) *Let $s \in \mathcal{S}$, $z, z^* \in \mathcal{Z}$, and denote by $\mathcal{N}(u) \subseteq \mathcal{U}$ and $\mathcal{N}(z) \subseteq \mathcal{Z}$ non-empty open neighborhoods of u and z respectively.*

(i.a) $E[r(\mathbf{1}\{U_X \leq \nu(z,s)\}, s, U, U_Y)|Z=z^*, S=s] < \infty$ and $E(Y|Z=\ddot{z}, S=s) < \infty$ for $\ddot{z} = z, z^*$,

(i.b) $\mathcal{U}_{z^*,s} = \mathcal{U}_{z,s}$,

(i.c) $\bar{r}(z,\cdot|s)$ is absolutely continuous on $\mathcal{U}_{z,s}$,

(i.d) for a.e. u and all $u^\dagger \in \mathcal{N}(u)$, $\bar{r}(z,u^\dagger|s) < \infty$ and there is a function $\Phi_{1,u}(u_x,u_y)$ with $E[\Phi_{1,u}(U_X,U_Y)|S=s] < \infty$ such that $\left|\frac{\partial}{\partial u} r(\mathbf{1}\{u_x \leq \nu(z,s)\}, s, u^\dagger, u_y)\right| \leq \Phi_{1,u}(u_x,u_y)$ for a.e. (u_x,u_y),

(i.e) $E(W|Z=\ddot{z}, S=s) < \infty$ for $\ddot{z} = z, z^*$,

(i.f) $\bar{q}(\cdot|s)$ is absolutely continuous on $\mathcal{U}_{z,s}$,

(i.g) for a.e. u and all $u^\dagger \in \mathcal{N}(u)$, $\bar{q}(u^\dagger|s) < \infty$ and there is a function $\Upsilon_{1,u}(u_w)$ with $E[\Upsilon_{1,u}(U_W)|S=s] < \infty$ such that $\left|\frac{\partial}{\partial u} q(s,u^\dagger,u_w)\right| \leq \Upsilon_{1,u}(u_w)$ for a.e. u_w,

(ii.a) for all $z^\dagger \in \mathcal{N}(z)$, $\mathcal{U}_{z^\dagger,s} = \mathcal{U}_{z,s}$,

(ii.b) for all $z^\dagger \in \mathcal{N}(z)$, $\int \bar{r}(z^\dagger, u|s) f_{U|Z,S}(u|z^\dagger, s)du < \infty$ *and there is a function* $\Phi_2(u)$ *with* $\int_{\mathcal{U}_{z,s}} \Phi_2(u)du < \infty$ *such that* $\left|\frac{\partial}{\partial z}\{\bar{r}(z^\dagger, u|s) f_{U|Z,S}(u|z^\dagger, s)\}\right| \leq \Phi_2(u)$ *for a.e.* u,

(ii.c) $\frac{\partial}{\partial z}\nu(z,s) \neq 0$ *and* $f_{U_X|S}(\cdot|s)$ *is continuous at* $\nu(z,s)$ *with* $f_{U_X|S}(\nu(z,s)|s) > 0$,

(ii.d) for a.e. u, $E[\beta(0,1;s,u,U_Y)|U_X = \cdot, S = s]$ *is continuous at* $\nu(z,s)$,

(ii.e) for all $z^\dagger \in \mathcal{N}(z)$, $F_{U|Z,S}(\cdot|z^\dagger, s)$ *is absolutely continuous on* $\mathcal{U}_{z,s}$ *and there is a function* $\Phi_3(u)$ *with* $\int_{\mathcal{U}_{z,s}} \Phi_3(u)du < \infty$ *such that* $\left|\frac{\partial}{\partial z} f_{U|Z,S}(u|z^\dagger, s)\right| \leq \Phi_3(u)$ *for a.e.* u,

(ii.f) for all $z^\dagger \in \mathcal{N}(z)$, $\int \bar{q}(u|s)\frac{\partial}{\partial z} f_{U|Z,S}(u|z^\dagger, s)du < \infty$ *and there is a function* $\Upsilon_2(u)$ *with* $\int_{\mathcal{U}_{z,s}} \Upsilon_2(u)du < \infty$ *such that* $\left|\bar{q}(u|s)\frac{\partial}{\partial z} f_{U|Z,S}(u|z^\dagger, s)\right| \leq \Upsilon_2(u)$ *for a.e.* u.

The absolute continuity of $\bar{r}(z,\cdot|s)$ and $\bar{q}(\cdot|s)$ on $\mathcal{U}_{z,s}$ in A.2 ensures that $\frac{\partial}{\partial u}\bar{r}(z,\cdot|s)$ and $\frac{\partial}{\partial u}\bar{q}(\cdot|s)$ exist for a.e. u and are integrable. Assuming that derivatives are bounded almost everywhere by an integrable function justifies the interchange of derivative and integral.

Proof of Theorem 6.2: $(i.a)$ By A.1$(i.a)$, adding and subtracting $E(Y|Z = z, S = s)$ gives

$$\begin{aligned}\gamma(z,z^*|z^*,s) &\equiv E[r(\mathbf{1}\{U_X \leq \nu(z^*,s)\},s,U,U_Y) - r(\mathbf{1}\{U_X \leq \nu(z,s)\},s,U,U_Y)|Z = z^*, S = s] \\ &= R^N_{Y.Z}(z,z^*;s) \\ -\{E[r(\mathbf{1}\{U_X \leq \nu(z,s)\},s,U,U_Y)|Z = z^*, S = s] &- E[r(\mathbf{1}\{U_X \leq \nu(z,s)\},s,U,U_Y)|Z = z, S = s]\} \\ &\equiv R^N_{Y.Z}(z,z^*;s) - B_\gamma(z,z^*|z^*,s).\end{aligned}$$

Further, for $\ddot{z} = z, z^*$,

$$\begin{aligned}E[r(\mathbf{1}\{U_X \leq \nu(\ddot{z},s)\},s,U,U_Y)|Z = z^*, S = s] \\ = E[r(0,s,U,U_Y)|Z = z^*, S = s] \\ + E[\mathbf{1}\{U_X \leq \nu(\ddot{z},s)\}[r(1,s,U,U_Y) - r(0,s,U,U_Y)]\ |Z = z^*, S = s].\end{aligned}$$

$\Pr[\nu(z,s) < U_X \leq \nu(z^*,s)\ |S = s] > 0$ gives that $\nu(z,s) < \nu(z^*,s)$ and thus

$$\begin{aligned}\gamma(z,z^*|z^*,s) &= E[\ \mathbf{1}\{\nu(z,s) < U_X \leq \nu(z^*,s)\}\ [r(1,s,U,U_Y) - r(0,s,U,U_Y)]\ |Z = z^*, S = s] \\ &= E[r(1,s,U,U_Y) - r(0,s,U,U_Y)\ |\ \nu(z,s) < U_X \leq \nu(z^*,s), Z = z^*, S = s] \\ &\qquad \times \Pr[\nu(z,s) < U_X \leq \nu(z^*,s)\ |Z = z^*, S = s].\end{aligned}$$

By $U_X \perp Z|S = s$, we have

$$R^N_{X.Z}(z,z^*;s) = E[\mathbf{1}\{\nu(z,s) < U_X \leq \nu(z^*,s)\}|S = s] = \Pr[\nu(z,s) < U_X \leq \nu(z^*,s)|Z = z^*, S = s].$$

Dividing $\gamma(z,z^*|z^*,s)$ by $R^N_{X.Z}(z,z^*;s) = \Pr[\nu(z,s) < U_X \leq \nu(z^*,s)\ |S = s] > 0$ gives:

$$\begin{aligned}\bar{\beta}(0,1|\nu(z,s) < U_X \leq \nu(z^*,s), z^*, s) &= R^{Wald}_{Y.X|Z}(z,z^*;s) - \frac{1}{R^N_{X.Z}(z,z^*;s)} B_\gamma(z,z^*|z^*,s) \\ &\equiv R^{Wald}_{Y.X|Z}(z,z^*;s) - B(0,1|\nu(z,s) < U_X \leq \nu(z^*,s), z^*, s).\end{aligned}$$

To derive the expression for $B_\gamma(z, z^*|z^*, s)$, $(U_X, U_Y) \perp (U, Z)|S = s$ gives for $\ddot{z} = z, z^*$:

$$\begin{aligned} E[r(\mathbf{1}\{U_X \leq \nu(z,s)\}, s, U, U_Y)|Z = \ddot{z}, S = s] \\ &= E\{\ E[r(\mathbf{1}\{U_X \leq \nu(z,s)\}, s, U, U_Y)|U, Z = \ddot{z}, S = s]\ |Z = \ddot{z}, S = s\} \\ &= E[\bar{r}(z, U|s)|Z = \ddot{z}, S = s]. \end{aligned}$$

By A.2.$i(b, c)$ and since $F_{U|Z,S}(\cdot|z^*, s)$ and $F_{U|Z,S}(\cdot|z, s)$ are absolutely continuous, integration by parts gives

$$\begin{aligned} B_\gamma(z, z^*|z^*, s) &= \int_{\mathcal{U}_{z,s}} \bar{r}(z, u|s)[f_{U|Z,S}(u|z^*, s) - f_{U|Z,S}(u|z, s)]du \\ &= \bar{r}(z, u|s)[F_{U|Z,S}(u|z^*, s) - F_{U|Z,S}(u|z, s)]\Big|_{\underline{u}}^{\bar{u}} \\ &\quad - \int_{\mathcal{U}_{z,s}} \frac{\partial}{\partial u}\bar{r}(z, u|s)[F_{U|Z,S}(u|z^*, s) - F_{U|Z,S}(u|z, s)]du, \end{aligned}$$

with $\underline{u}$ and $\bar{u}$ the (possibly infinite) infimum and supremum over $\mathcal{U}_{z,s}$. The first term vanishes and the result then follows since A.2$(i.d)$ gives $\frac{\partial}{\partial u}\bar{r}(z, u|s) = \bar{\delta}_Y(u; z|s)$ for a.e. u.
$(i.b)$ Similarly, $U_W \perp Z|S = s$, A.2.$i(b, e, f, g)$, and integration by parts give

$$\begin{aligned} R^N_{W.Z}(z, z^*; s) &= \int_{\mathcal{U}_{z,s}} \bar{q}(u|s)[f_{U|Z,S}(u|z^*, s) - f_{U|Z,S}(u|z, s)du \\ &= -\int_{\mathcal{U}_{z,s}} \bar{\delta}_W(u|s)[F_{U|Z,S}(u|z^*, s) - F_{U|Z,S}(u|z, s)]du. \end{aligned}$$

Division by $R^N_{X.Z}(z, z^*; s) > 0$ gives the result.
$(ii.a)$ By $(U_X, U_Y) \perp (U, Z)|S = s$ and A.2.$ii(a)$, we have

$$\begin{aligned} R^N_{Y.Z}(z; s) &= \frac{\partial}{\partial z}E[\ E[r(\mathbf{1}\{U_X \leq \nu(z,s)\}, s, U, U_Y)|U, Z = z, S = s]\ |Z = z, S = s] \\ &= \frac{\partial}{\partial z}\int_{\mathcal{U}_{z,s}} \bar{r}(z, u|s)\ f_{U|Z,S}(u|z, s)du \\ &= \int_{\mathcal{U}_{z,s}} \frac{\partial}{\partial z}\bar{r}(z, u|s)\ f_{U|Z,S}(u|z, s)du + \int_{\mathcal{U}_{z,s}} \bar{r}(z, u|s)\ \frac{\partial}{\partial z}f_{U|Z,S}(u|z, s)\ du \equiv T_1 + T_2, \end{aligned}$$

where we interchange the derivative and integral by A.2.$ii(b)$, and apply the product rule since the derivatives exist by A.2.$ii(c, d, e)$. In particular, to examine T_1 note that

$$\begin{aligned} &\bar{r}(z, u|s) \equiv E[r(\mathbf{1}\{U_X \leq \nu(z,s)\}, s, u, U_Y)|S = s] \\ &= E[r(0, s, u, U_Y)|S = s] + E\{E[\mathbf{1}\{U_X \leq \nu(z,s)\}\ [r(1, s, u, U_Y) - r(0, s, u, U_Y)]\ |U_X, S = s]|S = s\} \\ &= E[r(0, s, u, U_Y)|S = s] + \int_{-\infty}^{\nu(z,s)} E[r(1, s, u, U_Y) - r(0, s, u, U_Y)|U_X = t, S = s]f_{U_X|S}(t|s)dt. \end{aligned}$$

A.2.$ii(c,d)$, the Lebesgue differentiation theorem, and the chain rule give

$$T_1 = f_{U_X|S}(\nu(z,s)|s)\frac{\partial}{\partial z}\nu(z,s)\int_{\mathcal{U}_{z,s}} E[r(1,s,u,U_Y) - r(0,s,u,U_Y)|U_X = \nu(z,s), S = s]f_{U|Z,S}(u|z,s)du$$
$$= f_{U_X|S}(\nu(z,s)|s)\frac{\partial}{\partial z}\nu(z,s)\ \bar{\beta}(0,1|\nu(z,s),z,s),$$

where we make use of $(U_X, U_Y) \perp (U,Z)|S = s$ in the last equality.
Further, note that, by A.2$(ii.c)$, we have

$$R^N_{X.Z}(z;s) \equiv \frac{\partial}{\partial z}E(X|Z=z,S=s) = \frac{\partial}{\partial z}\Pr(U_X \leq \nu(z,s)|S=s)$$
$$= \frac{\partial}{\partial z}\int_{-\infty}^{\nu(z,s)} f_{U_X|S}(t|s)dt = f_{U_X|S}(\nu(z,s)|s)\frac{\partial}{\partial z}\nu(z,s) \neq 0.$$

To examine T_2, A.2$(ii.e)$ and A.2$(i.c)$ enable integration by parts which gives:

$$T_2 = \bar{r}(z,u|s)\frac{\partial}{\partial z}F_{U|Z,S}(u|z,s)\Big|_{\underline{u}}^{\bar{u}} - \int_{\mathcal{U}_{z,s}} \frac{\partial}{\partial u}\bar{r}(z,u|s)\frac{\partial}{\partial z}F_{U|Z,S}(u|z,s)du$$
$$= -\int_{\mathcal{U}_{z,s}} \bar{\delta}_Y(u;z|s)\frac{\partial}{\partial z}F_{U|Z,S}(u|z,s)du,$$

with $\underline{u}$ and $\bar{u}$ the (possibly infinite) infimum and supremum over $\mathcal{U}_{z,s}$ and where we use A.2$(ii.a)$ and A.2$(i.d)$ in the last equality. Dividing $R^N_{Y.Z}(z;s)$ by $R^N_{X.Z}(z;s)$ gives

$$B(0,1|\nu(z,s),z,s) = R^{LIV}_{Y.X|Z}(z;s) - \bar{\beta}(0,1|\nu(z,s),z,s) = -\frac{1}{R^N_{X.Z}(z;s)}\int_{\mathcal{U}_{z,s}} \bar{\delta}_Y(u;z|s)\frac{\partial}{\partial z}F_{U|Z,S}(u|z,s)du.$$

$(ii.b)$ Similarly, $U_W \perp (U,Z)|S = s$, A.2$i(f,g)$ and A.2$ii(a,c,e,f)$, and integration by parts give

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

Proof of Corollary 6.3: The proof is analogous to that of Corollary 5.3 and omitted.

Proof of Theorem 7.1 Let $\hat{Q} \equiv diag(\frac{1}{n}\sum_{i=1}^n \tilde{H}_i\tilde{G}_i', \frac{1}{n}\sum_{i=1}^n \tilde{H}_i\tilde{G}_i')$ and $\hat{M} \equiv \frac{1}{n}\sum_{i=1}^n(\tilde{H}_i'\epsilon_{Y.G|H,i}, \tilde{H}_i'\epsilon_{W.G|H,i})'$. By (i) and since $E(\tilde{H}\tilde{G}')$, and thus Q, is finite and nonsingular uniformly in $P \in \mathcal{P}$,

$$\sqrt{n}((\hat{R}_{Y.G|H}', \hat{R}_{W.G|H}')' - (R_{Y.G|H}', R_{W.G|H}')') = \hat{Q}^{-1}\sqrt{n}\hat{M} = (\hat{Q}^{-1} - Q^{-1})\sqrt{n}\hat{M} + Q^{-1}\sqrt{n}\hat{M},$$

exists in probability for all n sufficiently large uniformly in $P \in \mathcal{P}$. The result then obtains since (i) gives $\hat{Q}^{-1} - Q^{-1} = o_p(1)$ uniformly in $P \in \mathcal{P}$ and (ii) gives $\sqrt{n}\hat{M} \xrightarrow{d} N(0,\Xi)$, with Ξ finite and positive definite, uniformly in $P \in \mathcal{P}$.

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

j		$\hat{R}_{Y.G,j}$	$\hat{\mathcal{G}}_j([0,1])$	$\hat{R}_{W.G,j}$	$\hat{\mathcal{G}}_j([-1,1])$
1	Education	0.072	[0.001,0.072]	0.070***	[0.001,0.142]
	Robust s.e.	(0.004)	-	(0.002)	-
	$CI_{.95}$ and $CI_{\bar{\gamma}_j,.95}$	[0.064,0.079]	[-0.006,0.078]	-	[-0.006,0.150]
2	Experience	0.083	[0.035,0.083]	0.048***	[0.035,0.131]
	Robust s.e.	(0.007)	-	(0.004)	-
	$CI_{.95}$ and $CI_{\bar{\gamma}_j,.95}$	[0.070,0.096]	[0.022,0.094]	-	[0.022,0.145]
3	$\frac{1}{100}$Experience^2	-0.220	[-0.220,-0.133]	-0.087***	[-0.307,-0.133]
	Robust s.e.	(0.032)	-	(0.023)	-
	$CI_{.95}$ and $CI_{\bar{\gamma}_j,.95}$	[-0.283,-0.156]	[-0.273,-0.070]	-	[-0.374,-0.070]
4	Black indicator	-0.187	[-0.187,0.015]	-0.201***	[-0.388,0.015]
	Robust s.e.	(0.020)	-	(0.013)	-
	$CI_{.95}$ and $CI_{\bar{\gamma}_j,.95}$	[-0.226,-0.148]	[-0.219,0.050]	-	[-0.430,0.050]

Notes: Y denotes the logarithm of hourly wage and log(KWW) is used as predictive proxy W. $G = (G_X', G_S')'$ where G_X consists of education, experience, experience squared, and a binary indicator taking the value 1 if a person is black. G_S consist of two binary indicators taking the value 1 if a person lives in the South and in a metropolitan area (SMSA) respectively, 8 indicators for region of residence in 1966, an indicator for residence in SMSA in 1966, imputed father and mother education plus 2 indicators for missing father and mother education respectively, 8 binary indicators for interacted parental high school, college, or post graduate education, an indicator for father and mother being present at age 14, and an indicator for a single mother at age 14. 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. The estimates $\hat{\mathcal{G}}_j([0,1])$ and the corresponding $CI_{\bar{\gamma}_j,.95}$ obtain under the assumption $sign(R_{W.G,j}) = sign(\hat{R}_{W.G,j})$. The *, **, or *** next to $\hat{R}_{W.G,j}$ indicate that the p-value associated with a t-test for the null hypothesis $R_{W.G,j} = 0$ against the alternative hypothesis $sign(R_{W.G,j}) = sign(\hat{R}_{W.G,j})$ is less than 0.1, 0.05, or 0.01 respectively.

Table 2: Regression Estimates of Log Wage Equation Conditioning on Covariates and log(KWW)

	Education	Experience	$\frac{1}{100}$Experience^2	Black indicator	log(KWW)
$\hat{R}_{Y.(G',W')',j}$	0.057	0.073	-0.202	-0.146	0.203
Robust s.e.	(0.004)	(0.007)	(0.032)	(0.021)	(0.031)
$CI_{.95}$	[0.049, 0.066]	[0.060,0.087]	[-0.266,-0.139]	[-0.187,-0.104]	[0.141,0.264]

Notes: This table reports estimates $\hat{R}_{Y.(G',W')',j}$ from a linear regression of Y on G_X and covariates $(G_S', W')'$ with Y, W, and $G = (G_X', G_S')'$ defined as in Table 1.

Table 3: Regression-Based Estimates of Log Wage Equation with an Education and Race Interaction Term Conditioning on Covariates under Restrictions on Confounding

j		$\hat{R}_{Y.G,j}$	$\hat{\mathcal{G}}_j([0,1])$	$\hat{R}_{W.G,j}$	$\hat{\mathcal{G}}_j([-1,1])$
1	Education	0.068	[0.004,0.068]	0.064***	[0.004,0.131]
	Robust s.e.	(0.004)	-	(0.003)	-
	$CI_{.95}$ and $CI_{\bar{\gamma}_j,.95}$	[0.060,0.076]	[-0.004,0.075]	-	[-0.004,0.140]
2	(Education-12)×Black	0.017	[-0.012,0.017]	0.029***	[-0.012,0.046]
	Robust s.e.	(0.006)	-	(0.005)	-
	$CI_{.95}$ and $CI_{\bar{\gamma}_j,.95}$	[0.005,0.030]	[-0.024,0.028]	-	[-0.024,0.060]
3	Experience	0.081	[0.036,0.081]	0.045***	[0.036,0.127]
	Robust s.e.	(0.007)	-	(0.005)	-
	$CI_{.95}$ and $CI_{\bar{\gamma}_j,.95}$	[0.068,0.095]	[0.023,0.093]	-	[0.023,0.140]
4	$\frac{1}{100}$Experience^2	-0.210	[-0.210,-0.139]	-0.070***	[-0.280,-0.139]
	Robust s.e.	(0.033)	-	(0.023)	-
	$CI_{.95}$ and $CI_{\bar{\gamma}_j,.95}$	[-0.273,-0.146]	[-0.263,-0.075]	-	[-0.347,-0.075]
5	Black indicator	-0.193	[-0.193,0.018]	-0.211***	[-0.403,0.018]
	Robust s.e.	(0.020)	-	(0.013)	-
	$CI_{.95}$ and $CI_{\bar{\gamma}_j,.95}$	[-0.231,-0.154]	[-0.225,0.055]	-	[-0.445,0.055]

Notes: The results obtain using the specification in Table 1 after augmenting G_X with an interaction term that multiplies years of education minus 12 with a black binay indicator. The remaining notes in Table 1 apply analogously here.

Table 4: IV Regression-Based Estimates of Log Wage Equation Conditioning on Covariates under Restrictions on Confounding

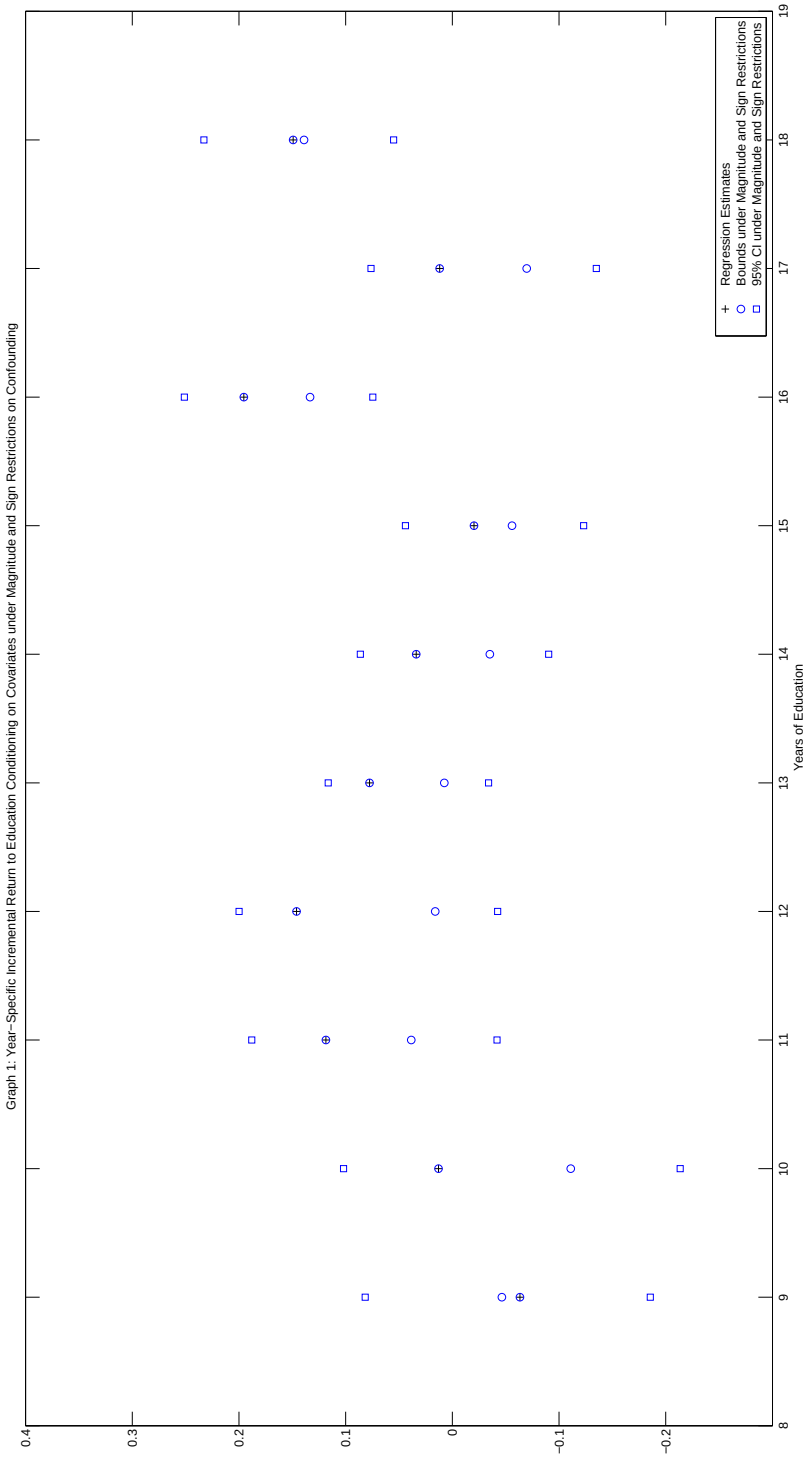
j		$\hat{R}_{Y.G\|H,j}$	$\hat{\mathcal{G}}_j([0,1])$	$\hat{R}_{W.G\|H,j}$	$\hat{\mathcal{G}}_j([-1,1])$
1	Education	0.134	[0.029,0.134]	0.106***	[0.029,0.240]
	Robust s.e.	(0.052)	-	(0.032)	-
	$CI_{.95}$ and $CI_{\bar{\gamma}_j,.95}$	[0.032,0.237]	[-0.061,0.220]	-	[-0.061,0.351]
2	Experience	0.061	[0.006,0.061]	0.054***	[0.006,0.115]
	Robust s.e.	(0.025)	-	(0.015)	-
	$CI_{.95}$ and $CI_{\bar{\gamma}_j,.95}$	[0.011,0.110]	[-0.036,0.102]	-	[-0.036,0.168]
3	$\frac{1}{100}$Experience^2	-0.113	[-0.113,0.009]	-0.122**	[-0.235,0.009]
	Robust s.e.	(0.123)	-	(0.072)	-
	$CI_{.95}$ and $CI_{\bar{\gamma}_j,.95}$	[-0.354,0.129]	[-0.316,0.216]	-	[-0.495,0.216]
4	Black indicator	-0.162	[-0.162,0.026]	-0.189***	[-0.351,0.026]
	Robust s.e.	(0.029)	-	(0.018)	-
	$CI_{.95}$ and $CI_{\bar{\gamma}_j,.95}$	[-0.218,-0.107]	[-0.209,0.076]	-	[-0.412,0.076]

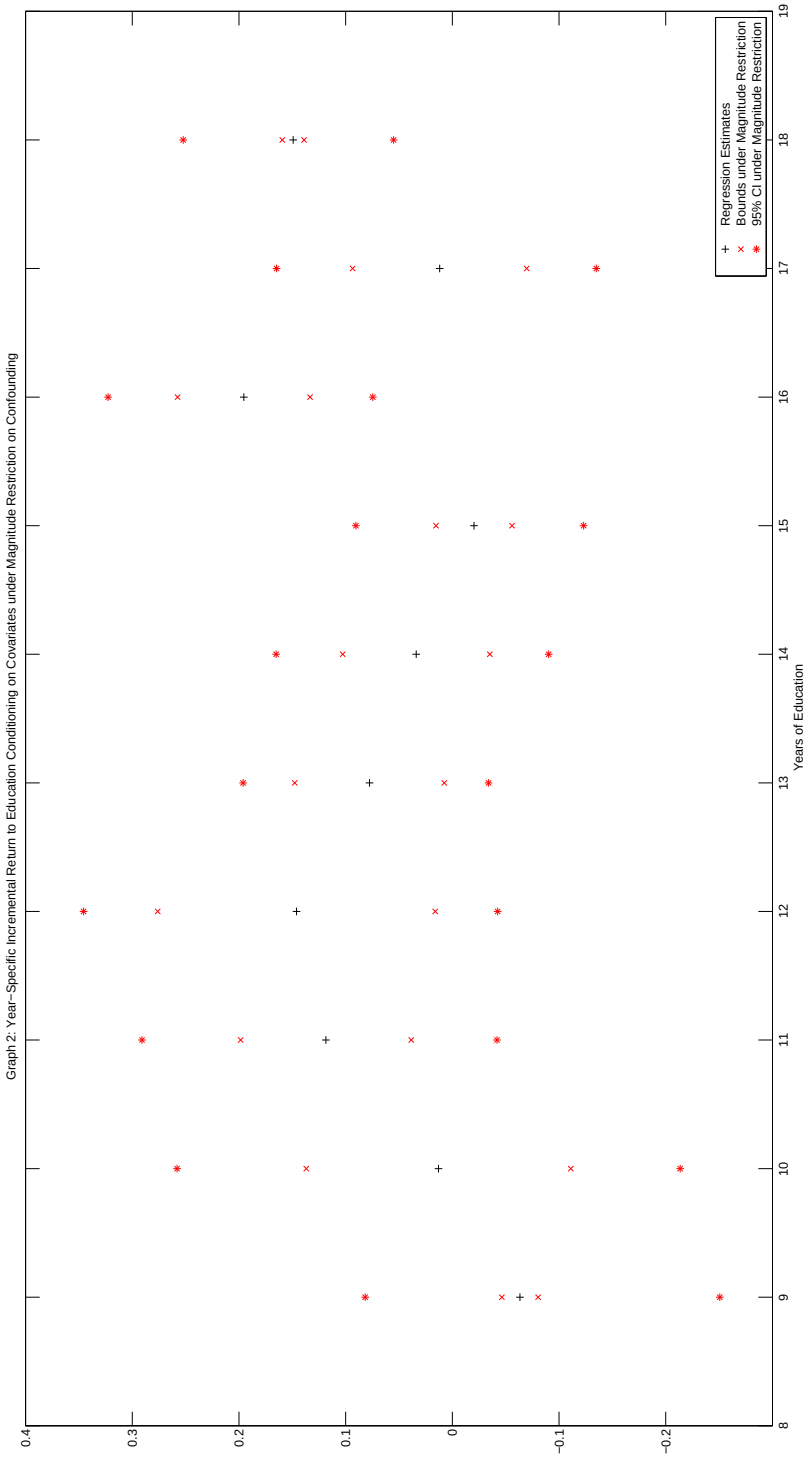
Notes: The results obtain by employing the specification in Table 1 and using an indicator for whether there is a four year college in the local labor market, age, age squared as instruments for education, experience, and experience squared. In particular, the instruments H_Z for G_X consist of the college proximity indicator, 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.

Table 5: Regression-Based Estimates of Log Wage Equation with Year-Specific Education Indicators Conditioning on Covariates under Restrictions on Confounding

j		$\hat{R}_{Y.G,j}$	$\hat{\mathcal{G}}_j([0,1])$	$\hat{R}_{W.G,j}$	$\hat{\mathcal{G}}_j([-1,1])$
10	Educ $\geq$ 11 years	0.118	[0.039,0.118]	0.080***	[0.039,0.198]
	Robust s.e.	(0.042)	-	(0.031)	-
	$CI_{.95}$ and $CI_{\bar{\gamma}_j,.95}$	[0.036,0.201]	[-0.042,0.188]	-	[-0.042,0.291]
11	Educ $\geq$ 12 years	0.146	[0.016,0.146]	0.130***	[0.016,0.276]
	Robust s.e.	(0.033)	-	(0.021)	-
	$CI_{.95}$ and $CI_{\bar{\gamma}_j,.95}$	[0.082,0.210]	[-0.042,0.200]	-	[-0.042,0.346]
12	Educ $\geq$ 13 years	0.078	[0.007,0.078]	0.070***	[0.007,0.148]
	Robust s.e.	(0.024)	-	(0.014)	-
	$CI_{.95}$ and $CI_{\bar{\gamma}_j,.95}$	[0.032,0.124]	[-0.034,0.116]	-	[-0.034,0.196]
13	Educ $\geq$ 14 years	0.034	[-0.035,0.034]	0.069***	[-0.035,0.103]
	Robust s.e.	(0.032)	-	(0.016)	-
	$CI_{.95}$ and $CI_{\bar{\gamma}_j,.95}$	[-0.029,0.096]	[-0.090,0.086]	-	[-0.090,0.165]
14	Educ $\geq$ 15 years	-0.020	[-0.056,-0.020]	0.036**	[-0.056,0.015]
	Robust s.e.	(0.038)	-	(0.018)	-
	$CI_{.95}$ and $CI_{\bar{\gamma}_j,.95}$	[-0.095,0.054]	[-0.123,0.044]	-	[-0.123,0.090]
15	Educ $\geq$ 16 years	0.195	[0.133,0.195]	0.062***	[0.133,0.258]
	Robust s.e.	(0.034)	-	(0.016)	-
	$CI_{.95}$ and $CI_{\bar{\gamma}_j,.95}$	[0.129,0.262]	[0.075,0.251]	-	[0.075,0.323]
16	Educ $\geq$ 17 years	0.012	[-0.070,0.012]	0.082***	[-0.070,0.093]
	Robust s.e.	(0.039)	-	(0.014)	-
	$CI_{.95}$ and $CI_{\bar{\gamma}_j,.95}$	[-0.065,0.088]	[-0.135,0.076]	-	[-0.135,0.165]
17	Educ $\geq$ 18 years	0.149	[0.139,0.149]	0.010	[0.139,0.159]
	Robust s.e.	(0.045)	-	(0.016)	-
	$CI_{.95}$ and $CI_{\bar{\gamma}_j,.95}$	[0.061,0.237]	[0.055,0.233]	-	[0.055,0.252]
18	Experience	0.087	[0.052,0.087]	0.035***	[0.052,0.122]
	Robust s.e.	(0.008)	-	(0.005)	-
	$CI_{.95}$ and $CI_{\bar{\gamma}_j,.95}$	[0.072,0.102]	[0.038,0.099]	-	[0.038,0.137]
19	$\frac{1}{100}$Experience^2	-0.241	[-0.241,-0.227]	-0.014	[-0.256,-0.227]
	Robust s.e.	(0.037)	-	(0.025)	-
	$CI_{.95}$ and $CI_{\bar{\gamma}_j,.95}$	[-0.314,-0.169]	[-0.309,-0.150]	-	[-0.341,-0.150]
20	Black indicator	-0.178	[-0.178,0.019]	-0.196***	[-0.374,0.019]
	Robust s.e.	(0.020)	-	(0.013)	-
	$CI_{.95}$ and $CI_{\bar{\gamma}_j,.95}$	[-0.216,-0.139]	[-0.210,0.054]	-	[-0.415,0.054]

Notes: The results obtain by extending the specification in Table 1 to include in G_X indicators for having at least t years of education, where $t = 2, ..., 18$ as in the sample, instead of total years of education. For brevity, we don't report in Table 5 the estimated identification regions for the average return to education for $t < 11$; these are often relatively imprecise with wide $CI_{\bar{\gamma}_j,.95}$. The remaining notes in Table 1 apply analogously here.


Graph 1: Year−Specific Incremental Return to Education Conditioning on Covariates under Magnitude and Sign Restrictions on Confounding
Regression Estimates
Bounds under Magnitude and Sign Restrictions
95% CI under Magnitude and Sign Restrictions
Years of Education
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


Graph 2: Year−Specific incremental Return to Education Conditioning on Covariates under Magnitude Restriction on Confounding
Years of Education
Regression Estimates
Bounds under Magnitude Restriction
95% CI under Magnitude Restriction
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

References

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.

Baltagi, B. (1999). *Econometrics, 2nd Edition*. Springer-Verlag, Berlin.

Bartle, R. (1966), *Elements of Integration*. New York: Wiley.

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.

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.

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. (2013), "Instrumental Variables Methods with Heterogeneity and Mismeasured Instruments," Boston College Department of Economics Working Paper.

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.

Corbae D., M. Stinchcombe, and J. Zeman (2009). *An Introduction to Mathematical Analysis for Economic Theory and Econometrics*. Princeton University Press.

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.

Frisch, R. and F. Waugh (1933), "Partial Regressions as Compared with Individual Trends," *Econometrica*, 1, 939-953.

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 (1998), "Instrumental Variables Methods for the Correlated Random Coefficient Model: Estimating the Average Rate of Return to Schooling When the Return is Correlated with Schooling," *Journal of Human Resources*, 33, 974-987.

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

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.

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 (2010) "Estimating a Class of Triangular Simultaneous Equations Models without Exclusion Restrictions," *Journal of Econometrics*, 154, 154-164.

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.

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.

Leamer, E. (1987), "Errors in Variables in Linear Systems," *Econometrica*, 55, 893-909.

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.

Li, Q. and J. Racine (2007). *Nonparametric Econometrics: Theory and Practice*. Princeton: Princeton University Press.

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.

Rudin, W. (1976). *Principles of Mathematical Analysis*. McGraw-Hill International.

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.

Shorack, G. (2000). *Probability for Statisticians*. New York: Springer-Verlag.

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. (1997), "On Two Stage Least Squares Estimation of the Average Treatment Effect in a Random Coefficient Model," *Economics Letters*, 56, 129–133.

Wooldridge, J. (2003), "Further Results on Instrumental Variables Estimation of Average Treatment Effects in the Correlated Random Coefficient Model," *Economics Letters*, 79, 185–191.

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