Identification of Average Effects under Magnitude and Sign Restrictions on Confounding

Karim Chalak*†
Boston College

May 30, 2014

Abstract

This paper studies measuring the average effects of X on Y in structural systems by imposing magnitude and sign restrictions on confounding 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. We characterize the omitted variables bias, due to confounders U, of regression and IV methods for the identification of these various average effects, thereby extending the classic linear omitted variable bias representation. Then, 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 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, partial identification, proxy.*

*Assistant Professor, Department of Economics, Boston College, 140 Commonwealth Ave., Chestnut Hill, MA 02467, USA. Email: chalak@bc.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, the Research in Econometrics Workshop at Boston College, and the seminars at 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, University of Virginia, and University of Montreal as well as Kate Antonovics, Andrew Beauchamp, Stéphane Bonhomme, 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 by imposing restrictions on the magnitude and sign of confounding without requiring conditional exogeneity of causes, treatment, or instruments given covariates. In particular, we study the full (point) and partial 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.

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

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

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. Thus, both X and U are potential structural determinants (causes) of Y, albeit realizations of Y and X are observed whereas those of U are not. As discussed 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. Suppose that the researcher observes realizations of a proxy W for U that is possibly error-laden and given by

$$W = \alpha_W + U\bar{\delta}_W, \tag{1b}$$

where, for now, we consider a constant slope coefficient $\bar{\delta}_W$ and random intercept α_W which may be correlated with U. For example, W may denote the logarithm of a test score commonly used as a proxy for ability, such 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 $\bar{\delta}_W$ denote respectively the semi-elasticities of wage and test score with respect to unobserved ability (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, one could consider standardizing the variables in these two equations, 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. Let Z be uncorrelated with α_W; thus, the proxy W is informative, in the sense that any correlation between Z and W arises solely due to[1] U. Define $\tilde{Z} \equiv Z - E(Z)$. We then have

$$E(\tilde{Z}Y) = E(\tilde{Z}X')\bar{\beta} + E(\tilde{Z}U)\bar{\delta}_Y \quad \text{and} \quad E(\tilde{Z}W) = E(\tilde{Z}U)\bar{\delta}_W.$$

Provided $E(\tilde{Z}X')$ is nonsingular and $\bar{\delta}_W \neq 0$, we obtain

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

This expression for $\bar{\beta}$ involves two linear instrumental variables (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 (or inconsistency) $E(\tilde{Z}X')^{-1}E(\tilde{Z}W)\frac{\bar{\delta}_Y}{\bar{\delta}_W}$ 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.

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 average effects $\bar{\beta}$ of X on Y. The paper does not require particular answers to these questions. Instead, it characterizes the mapping from every possible answer to the corresponding identification region for the average effects of X on Y. 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

[1] More generally, we let U denote the vector of unobservables that drive Y or the vector of proxies W and are thought to be freely correlated with Z (conditional on covariates S), and we absorb into α_Y and α_W respectively the unobserved drivers of Y and W that are (conditionally) uncorrelated with Z. It then suffices to have as many proxies W as confounders U. Recall that Z may equal X.

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.

Sometimes, economic theory and evidence can provide guidance to answering these questions in particular contexts. For example, in the earning equation illustrative example, 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. Third, 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 negative average returns to education and a black-white wage gap in favor of blacks, which is inconsistent with the general findings in the literature. Last, the assumptions $\left|\bar{\delta}_Y\right| \leq \left|\bar{\delta}_W\right|$, $0 \leq \frac{\bar{\delta}_Y}{\bar{\delta}_W}$, or $0 \leq \frac{\bar{\delta}_Y}{\bar{\delta}_W} \leq 1$, underlying partial identification, are a weakening of the commonly employed assumption of exogeneity. Specifically, with Z and U freely correlated, exogeneity requires the coefficient $\bar{\delta}_Y$ on U to be zero; we allow but don't require this.

More generally, this paper's approach does not require a linear or parametric specification. In particular, Section 4 studies the structural system:

$$Y = r(X, S, U, U_Y) \quad \text{and} \quad W = q(S, U, U_W), \tag{2}$$

where the vectors of unobservables U_Y and of confounders U interact nonseparably with X and a vector of observed covariates S to drive Y according to the unknown nonparametric structural function r. Further, U interacts with the unobserved vector U_W and the covariates S to drive the vector of proxies W according to the unknown function q. Unlike U_Y and U_W, U may statistically depend on X given the covariates S. Here, we study the identification of conditional average discrete and marginal causal effects of X on Y given covariates S under magnitude and sign restrictions on confounding[2]. To illustrate, consider the additively

[2]Note that in the general nonparametric specifications studied in this paper, one can apply a probability

separable case in which

$$Y = \ddot{r}(X, S, U_Y) + U'\delta_Y \quad \text{and} \quad W' = \alpha'_W + U'\delta_W, \tag{3}$$

where δ_Y is a vector of random coefficients that can depend on S and U_Y and where α_W is a vector, and δ_W a matrix, of random coefficients that can depend on S and U_W. When $\delta_Y = 0$ and[3] $U_Y \perp X|S$, 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 4 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 (3) with the random vector δ_Y possibly nonzero, and second when the effect of U on Y in the nonseparable equations (2) is possibly nonzero. In section 3, we focus on the 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 the potential instruments Z to be freely (conditionally) correlated with U and for δ_Y to be nonzero. Appendix A contains additional extensions in the random coefficients case to allow for a panel structure and proxies included in the Y equation. Last, Section 5 studies treatment effects and augments the nonseparable equations (2) from Section 4 with a threshold crossing equation generating the binary treatment X:

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

where U_X is an unobserved variable and the function ν is unknown. For example, when $\delta_Y = 0$ in the separable equations (3) 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). Section 5 studies the full and partial identification of conditional local and marginal treatment effects as well as conditional average treatment effects for the population, treated, and untreated, when $(U_X, U_Y) \perp Z|S = s$ but U may depend on Z given S in both the separable case in equations (3) with δ_Y possibly nonzero and the nonseparable case in equations (2) when the effect of U on Y may be nonzero.

This paper's method provides a simple alternative to the common practice which informally assumes that conditioning on proxies for confounders ensures conditional exogeneity. For example, consider the illustrative linear example in equations (1a, 1b) with $Corr(Z, (\alpha_Y, \alpha_W)') = 0$.

transformation, e.g. for continuous response, proxy, and confounder, to reparametrize the Y and W equations such that U, Y, and W have uniform distributions. The researcher can then contrast the average sign and magnitude of the percentile changes in the response and proxy due to a percentile change in the confounder.

[3]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 $S = s$.

Then conditioning on W doesn't ensure that the coefficient on X from a regression of Y on $(1, X', W')'$ identifies $\bar{\beta}$. In particular, substituting for scalar U gives

$$Y = \alpha_Y - \frac{\bar{\delta}_Y}{\bar{\delta}_W}\alpha_W + X'\bar{\beta} + \frac{\bar{\delta}_Y}{\bar{\delta}_W}W, \tag{4}$$

and the possible conditional correlation[4] between Z and α_W given W leads to (IV) regression bias, including when $Z = X$. Indeed, more generally, conditioning on W may, but need not, attenuate the regression bias (see e.g. Wickens, 1972; Battistin and Chesher, 2009; and Ogburna and VanderWeele, 2012). Note, however, that this regression consistently estimates $\frac{\bar{\delta}_Y}{\bar{\delta}_W}$ if one assumes that $W = U\bar{\delta}_W$, i.e. that $\alpha_W = 0$ and the proxy W is a perfect rescaling of U. Here, we don't require this or that α_W and U are uncorrelated, allowing e.g. for nonclassical measurement error. Rather than conditioning on mismeasured proxies, this paper employs proxies to bound the nonparametric regression bias (see Section 4). Furthermore, this paper's method provides a practical alternative to IV methods when potential instruments may be weak or (conditionally) endogenous. In particular, we don't require $Cov(Z, (\alpha_Y, U)') = 0$ in equations (1a, 1b). Instead we need only impose $Cov(Z, (\alpha_Y, \alpha_W)') = 0$. Moreover, $\bar{\beta}$ is "under-identified" in equation (4) from the linear illustrative example since Z and X have the same dimension and thus there are fewer exogenous instruments for $(X', W')'$ than needed for full identification. Similar difficulties arise in more general nonlinear cases.

As discussed above, imposing sign and magnitude restrictions on confounding is a weakening of exogeneity. Manski and Pepper (2000) employ alternative assumptions to bound nonparametric average effects. 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 and that between the endogenous variable and U have the same sign and then further assume that the potential instrument is less correlated with U than the endogenous variable is. Lewbel (2012) restricts the covariance of Z and the product of heteroskedastic error terms. Bontemps, Magnac, and Maurin (2012) provide additional examples and a general treatment of set identified linear models. We don't require the assumptions in these papers. Instead, we employ proxies to identify average effects under magnitude and sign restrictions on confounding. Of

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

course, which identifying assumption is more appropriate depends on the context.

After deriving sharp identification regions for the average direct effects of X on Y in Sections 3, 4, and 5 under magnitude and sign restrictions on confounding, Section 6 studies estimation and inference. Last, Section 7 applies these results 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. Importantly, we don't 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 bound estimates are generally narrower than the IV-based ones, with especially narrower confidence intervals. Note that 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% confidence interval (CI) $[0.141, 0.264]$. Allowing for error-laden proxies in a semiparametric specification, the regression-based estimated sharp identification region for the black-white wage gap under sign and magnitude restrictions on confounding ($0 \leq \frac{\bar{\delta}_Y}{\bar{\delta}_W} \leq 1$) 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 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). Section 8 concludes and mathematical proofs are gathered in Appendix B.

2 Data Generation

The next assumption defines the data generating process.

Assumption 1 (S.1) *(i) Let* $M \equiv (\underset{c\times 1}{S}', \underset{\ell\times 1}{Z}', \underset{k\times 1}{X}', \underset{m\times 1}{W}', \underset{1\times 1}{Y})'$ *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. We distinguish 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 nonparametrically impact the response Y according to the structural function r. We allow but do not require the availability of covariates S; if these are absent we set $S = 1$. Further, we allow but don't require elements of S to directly affect Y; if r doesn't directly depend on S, these may nevertheless serve as conditioning variables. We observe realizations of a vector W of proxies for U. We sometimes allow for W and X to have common elements. Analogously to U_Y, the vector U_W interacts with S and U to generate the proxies W. 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.

2.1 Causal Effects

For[5] (x, s, u, u_y) and (x^*, s, u, u_y) in $\mathcal{X} \times \mathcal{S} \times \mathcal{U} \times \mathcal{U}_\mathcal{Y}$, the direct effect of X on Y at (x, x^*) given (s, u, u_y) is $\beta(x, x^*; s, u, u_y) \equiv r(x^*, s, u, u_y) - r(x, s, u, u_y)$. Further, when r is differentiable in a particular 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. Then $\beta(x; s, u, u_y) \equiv \frac{\partial}{\partial x} r(x, s, u, u_y)$ is the direct marginal effect of X on Y at x given (s, u, u_y). If U enters r separably from X then $\beta(x; s, u, u_y) = \beta(x; s, u_y)$ does not depend on U. Further, if the effect of X on Y is linear, $\beta(x; s, u_y)$ does not depend on X, and we obtain the random coefficient specification $\beta \equiv \beta(S, U_Y)$.

Similarly, for (x, s, u, u_y) and (x, s, u^*, u_y) in $\mathcal{X} \times \mathcal{S} \times \mathcal{U} \times \mathcal{U}_\mathcal{Y}$, the direct effect of U on Y at (u, u^*) given (x, s, u_y) is $\delta_Y(u, u^*; x, s, u_y) \equiv r(x, s, u^*, u_y) - r(x, s, u, u_y)$. Further, for $l = 1$ and

[5] Throughout, for random vectors A and B, we denote the support of A by $\mathcal{A}$ and that of A given $B = b$ by $\mathcal{A}_b$.

r differentiable in u, the direct marginal effect of U on Y at u given (x, s, u_y) is $\delta_Y(u; x, s, u_y) \equiv \frac{\partial}{\partial u} r(x, s, u, u_y)$. If U enters r separably from X and its effect on Y is linear, we obtain the random coefficient specification $\delta_Y \equiv \delta_Y(S, U_Y)$. Analogously, for (s, u, u_w) and (s, u^*, u_w) in $\mathcal{S} \times \mathcal{U} \times \mathcal{U}_\mathcal{W}$, the direct effect of U on W_h, for $h = 1, ..., m$, are $\delta_{W_h}(u, u^*; s, u_w) \equiv q_h(s, u^*, u_w) - q_h(s, u, u_w)$ and the direct marginal effect of U on W_h is $\delta_{W_h}(u; s, u_w) \equiv \frac{\partial}{\partial u} q_h(s, u, u_w)$.

Thus, causal effects, such as β, δ_Y, and δ_W, are features of the structural system and can derive from economic theory whereas the observability of X and U is an empirical matter. In this paper, we're interested in measuring certain (conditional) averages of $\beta(x, x^*; s, u, u_y)$ and $\beta(x; s, u, u_y)$ such as the conditional average direct effect of X on Y at (x, x^*) given $X = x^*$ and $S = s$:

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

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 $\bar{\beta}(0, 1|1)$. As another example, we're interested in measuring the conditional average direct marginal effect of X on Y at x given $S = s$:

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

It's also useful to give a succinct notation for the average direct effects of U on Y and W. For example,

$$\bar{\delta}_Y(u; x|s) \equiv E[\delta_Y(u; x, s, U_Y)|S = s] \equiv E[\frac{\partial}{\partial u} r(x, s, u, U_Y)|S = s].$$

Similarly, for scalar W,

$$\bar{\delta}_W(u, u^*|s) \equiv E[\delta_W(u, u^*; s, U_W)|S = s] \equiv E[q(s, u^*, U_W) - q(s, u, U_W)|S = s].$$

3 Identification of Average Random Coefficients

While the paper's method doesn't require a linear or parametric effect of X on Y, we find it instructive to begin our analysis of the identification of average effects under magnitude and sign restrictions on confounding by studying linear structures with random coefficients. We relax the linearity assumption when studying the identification of conditional average nonparametric effects in Section 4 and of local, marginal, and average treatment effects in Section 5. Specifically, we impose the following assumption in Section 3.

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*

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

We collect the random coefficients in $\theta \equiv (\underset{m\times 1}{\alpha_W}', vec(\underset{l\times m}{\delta_W})', \underset{1\times 1}{\alpha_Y}, \underset{k\times 1}{\beta}', \underset{l\times 1}{\delta_Y}')'$.

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. We suppress the index i when referring to the population. In particular, β is the vector of random direct effects of the elements of X on Y. To simplify the exposition, we set $S = 1$ in Sections 3.2 and 3.3 and study identifying the average effect $\bar{\beta} \equiv E(\beta)$ of X on Y. Section 3.4 accommodates covariates S and studies the identification of the conditional average effect $\bar{\beta}(S) \equiv E(\beta|S)$ of X on Y given S.

3.1 IV Regression Notation

Throughout this paper, for a generic random vector A with finite mean, we write:

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

For example, we write $\bar{\beta} \equiv E(\beta)$ and $\bar{\delta}_Y \equiv E(\delta_Y)$. Further, we employ the following succinct notation for IV regression coefficients and residuals. For generic random vector A and vectors B and C of equal dimension with $E(\tilde{C}\tilde{A}')$ finite and $E(\tilde{C}\tilde{B}')$ finite and nonsingular, we let

$$R_{A.B|C} \equiv E(\tilde{C}\tilde{B}')^{-1}E(\tilde{C}\tilde{A}') \quad \text{and} \quad \epsilon_{A.B|C}' \equiv \tilde{A}' - \tilde{B}'R_{A.B|C}$$

denote the linear IV regression estimand and residual respectively, so that by construction $E(\tilde{C}\epsilon_{A.B|C}') = 0$. For example, for $k = \ell$, $R_{Y.X|Z} \equiv E(\tilde{Z}\tilde{X}')^{-1}E(\tilde{Z}\tilde{Y})$ is the vector of slope coefficients associated with X in a linear IV regression of Y on $(1,X')'$ using instruments $(1,Z')'$. In the special case where $B = C$, we obtain the linear regression coefficients and residuals:

$$R_{A.B} \equiv R_{A.B|B} \equiv E(\tilde{B}\tilde{B}')^{-1}E(\tilde{B}\tilde{A}') \quad \text{and} \quad \epsilon_{A.B}' \equiv \epsilon_{A.B|B}' \equiv \tilde{A}' - \tilde{B}'R_{A.B}.$$

3.2 Characterization and Full Identification

We begin by characterizing $\bar{\beta}$ and studying conditions for full identification. Section 3.3 studies partial identification of elements of $\bar{\beta}$ under sign and magnitude restrictions on confounding. For illustration, consider the example from the Introduction with scalar proxy W and confounder U 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 $E(\tilde{Z}\tilde{X}')$ is nonsingular, $\bar{\delta}_W \neq 0$, and $Cov(Z, (\alpha_Y, \alpha_W)') = 0$, we have:

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

In particular, the IV regression (omitted variable) bias $R_{W.X|Z}\frac{\bar{\delta}_Y}{\bar{\delta}_W}$ 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 3.1 extends this result to allow for vectors U and W and for random slope coefficients. We set $S = 1$ here; Section 3.4 explicitly conditions on covariates.

Theorem 3.1 *Assume S.2 with $S = 1$, $\ell = k$, $m = l$, and that*

(i) $E(\tilde{Z}\tilde{X}')$ and $\bar{\delta}_W$ are nonsingular,

(ii) $Cov(\alpha_Y, Z) = 0$, $E(\tilde{\beta}|Z, X) = 0$, and $E(\tilde{\delta}_Y|U, Z) = 0$,

(iii) $Cov(\alpha_W, Z) = 0$ and $E(\tilde{\delta}_W|U, Z) = 0$.

Let $\bar{\delta} \equiv \bar{\delta}_W^{-1}\bar{\delta}_Y$ then

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

We let $B \equiv R_{Y.X|Z} - \bar{\beta} = R_{W.X|Z}\bar{\delta}$ denote the IV regression bias (or inconsistency) in measuring $\bar{\beta}$. When $Z = X$, Theorem 3.1 gives that $\bar{\beta} = R_{Y.X} - R_{W.X}\bar{\delta}$.

Next, we discuss the conditions in Theorem 3.1. Condition (i) requires $\ell = k$ and $m = l$, with $E(\tilde{Z}\tilde{X}')$ and $\bar{\delta}_W$ nonsingular. More generally, $\ell \geq k$ and $m \geq l$ suffice for full or partial identification and one may use a weighted combination of the resulting moments. In particular, having $\ell \geq k + m$ may fully identify $\bar{\beta}$. For example, if $\ell = k + m$ and $m = l$ then $(\bar{\beta}', \bar{\delta}')' = R_{Y.(X',W')'|Z}$ provided $E[\tilde{Z}(\tilde{X}', \tilde{W}')']$ is nonsingular[6]. 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.

Conditions (ii) and (iii) are implied by the assumption that the coefficients θ are mean independent[7] of (U, Z, X) or the stronger assumption that $(U_W, U_Y) \perp (U, Z, X)$. Note that

[6] For example, if δ_Y and δ_W are constant, substituting for U gives $Y = \alpha_Y - \alpha_W\bar{\delta} + X'\bar{\beta} + W'\bar{\delta}$.

[7] In the linear case, having Z be mean independent of (α_Y, α_W) 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 4 and 5.

when β, δ_Y, and δ_W are constants, conditions (ii) and (iii) reduce to $Cov(Z, (\alpha_Y, \alpha_W)') = 0$. In particular, condition (ii) imposes assumptions on the random coefficients in the Y equation. It requires that Z is uncorrelated with α_Y, β is mean independent[8] of (Z, X), and δ_Y is mean independent of (U, Z). Roughly speaking, (ii) isolates U as the source of the difficulty in identifying $\bar{\beta}$. Had U been observed with $\ell = k + l$ and $E(\tilde{Z}(\tilde{X}', \tilde{U}'))$ nonsingular, (ii) would permit identifying the average slope coefficients via IV regression. Note that linearity and condition (ii) can (indirectly) restrict how the return to education β depends on ability U (see e.g. Card 1999). In the linear case, and if valid instruments are available, one can consider IV methods for the correlated random coefficient model, e.g. Wooldridge (1997, 2003) and Heckman and Vytlacil (1998). Similarly, linearity and condition (ii) 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 4 and 5 weaken these restrictions in Theorem 3.1 and give identification results for the case in which X and U can interact to determine Y via the nonseparable specification in S.1. Importantly, however, the conditions in Theorem 3.1 do not restrict the joint distribution of $(U, Z', X')'$ other than requiring that $E(\tilde{Z}\tilde{X}')$ is nonsingular. In particular, Z and X can be freely correlated with U and thus endogenous.

Condition (iii) imposes restrictions on the random coefficients in the W equation. It requires that Z is uncorrelated with α_W and that the elements of δ_W are mean independent of (U, Z). This linearly relates $Cov(Z, W)$ to $Cov(Z, U)$ via the matrix $\bar{\delta}_W$. Note that we do not directly restrict the dependence between α_W and U, allowing W to be an error-laden proxy for U with correlated measurement error. Nevertheless, even when δ_W is constant and α_W is independent of U, recall from the Introduction that the coefficient on X from a regression of Y on $(1, X', W')'$ need not identify $\bar{\beta}$. We note that one means for full identification generates sufficiently many valid instruments by imposing restrictions involving the components of α_W, δ_W, and U. For example, let U, W_1, and W_2 be scalars and suppose that

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

so that there are two proxies for U, with $\bar{\delta}_{W_1}, \bar{\delta}_{W_2} \neq 0$ and $Corr[(U, \alpha_{W_2})', (\alpha_Y, \alpha_{W_1})'] = 0$. 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 don't require such restrictions here, allowing for example, for components of α_W (e.g. test taking skills) to be correlated. (Appendix A studies the case of multiple proxies for U that are components of X.)

[8] We employ the unnecessary mean independence assumptions in conditions (ii) and (iii) of Theorem 3.1 because of their simple interpretation. However, zero covariances $E[\tilde{Z}X'\bar{\beta}] = 0$, $E[\tilde{Z}U'\bar{\delta}_Y] = 0$, and $E[\tilde{Z}U'\bar{\delta}_W] = 0$ suffice.

To illustrate the consequences of Theorem 3.1, consider the example from the Introduction with scalar U and W. 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, shape restrictions on the effects of U on Y and W can fully identify $\bar{\beta}$. In particular, this occurs under signed proportional confounding, in 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. In this case, $\bar{\beta}$ 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,

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

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

Corollary 3.2 *Assume the conditions of Theorem 3.1 and let $j = 1, ..., k$. (i) If $B_j = 0$ (exogeneity) then $\bar{\beta}_j = R_{Y.X|Z,j}$. (ii) If $\bar{\delta} = d$ (signed proportional confounding) then $\bar{\beta}_j = R_{Y.X|Z,j} - \sum_{h=1}^{m} d_h R_{W_h.X|Z,j}$.*

Thus, it suffices for exogeneity that $\bar{\delta}_Y = 0$ or $R_{W.X|Z} = 0$. In particular, if one fails to reject the null hypothesis $R_{W.X|Z,j} = 0$ against the alternative $R_{W.X|Z,j} \neq 0$, say via a t-test in the scalar proxy case, then one cannot reject, under Theorem 3.1's assumptions, that $R_{Y.X|Z,j}$ identifies $\bar{\beta}_j$. Further, signed proportional confounding with known $|d_h|$ and $sign(d_h), h = 1, ..., m$, point identifies $\bar{\beta}$.

3.3 Partial Identification

In the absence of conditions leading to full identification, magnitude and sign restrictions on the *average* direct effects of U on Y and W partially identify the elements of $\bar{\beta}$. To illustrate, consider the example from the Introduction with scalar U and W. 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 sign and magnitude 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. For example, suppose that $0 \leq \bar{\delta}$ so that U affects Y and W on average in the same direction. Thus, in this case $\bar{\delta} \in \mathcal{D} = [0, 1]$. Then the expression for $\bar{\beta}$ gives the following identification regions for $\bar{\beta}_j$, $j = 1, ..., k$, which depend on the sign of $R_{W.X|Z,j}$:

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

Symmetrically, if we maintain the magnitude restriction $\left|\bar{\delta}\right| \leq 1$ and assume that $\bar{\delta} \leq 0$ so that $\bar{\delta} \in \mathcal{D} = [-1, 0]$, we obtain

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

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 regions 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 these identification regions exclude the IV estimand $R_{Y.X|Z,j}$. Symmetrically, if we assume that $1 \leq \left|\bar{\delta}\right|$ and $\bar{\delta} \leq 0$ so that $\bar{\delta} \in \mathcal{D} = (-\infty, -1]$, we obtain

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

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 $\mathcal{B}_j([0,1]\ |sign(R_{W.X|Z,j}))$ or $\mathcal{B}_j([-1,0]\ |sign(R_{W.X|Z,j}))$. 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 $\mathcal{B}_j((-\infty,-1]\ |sign(R_{W.X|Z,j}))$, $\mathcal{B}_j([1,+\infty)\ |sign(R_{W.X|Z,j}))$, and $\mathcal{B}_j((-\infty,\ -1] \bigcup [1,\ +\infty))$ are.

Alone, sign restrictions determine the direction of the IV regression omitted variable bias. In particular, we have

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

The above identification regions for $\bar{\beta}_j$, under magnitude and/or sign restrictions on confounding with scalars U and W, are sharp. Thus, any point in these regions is feasible under the maintained assumptions. In particular, given the distribution of the observables M, for each element b of $\mathcal{B}_j(\mathcal{D}\ |\ sign(R_{W.X|Z,j}))$ or $\mathcal{B}_j(\mathcal{D})$, one can construct constants d_Y and d_W (which, being constant, satisfy the conditions on δ_Y and δ_W in Theorem 3.1) such that $\frac{d_Y}{d_W} \in \mathcal{D}$. For example, for $R_{W.X|Z,j} \neq 0$, it suffices to let $\frac{d_Y}{d_W} = \frac{1}{R_{W.X|Z,j}}(R_{Y.X|Z,j} - b)$ according to $\mathcal{S}_1$.

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 $\mathcal{B}_j(\times_{h=1}^{m} \mathcal{D}_h\ |\ \underset{h=1,...,m}{sign}\ (R_{W_h.X|Z,j}))$ for $\bar{\beta}_j$, $j = 1, ..., k$, defined next.

Corollary 3.3 derives sharp identification regions for a general matrix δ_W and restrictions[9] $\bar{\delta}_h \in \mathcal{D}_h = [d_{L,h,}, d_{H,h}]$, $h = 1, ..., m$ (we allow for $d_{L,h} = -\infty$ and/or $d_{H,h} = +\infty$ yielding identification regions that are either half open intervals or the real line). To facilitate the exposition in subsequent sections, we let $A \equiv R_{W.X|Z}$ so that $A_{jh} \equiv R_{W_h.X|Z,j}$. Also, when some elements of $R_{W.X|Z,j}$ have different signs, we partition these, without loss of generality,

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

such that $R_{W_h.X|Z,j} \leq 0$ for $h = 1, ..., g$ and $0 \leq R_{W_h.X|Z,j}$ for $h = g+1, ..., m$. Otherwise, if all the elements of $R_{W.X|Z,j}$ have common sign, we omit the irrelevant inequalities and sums.

Corollary 3.3 *Assume the conditions of Theorem 3.1 and that* $\bar{\delta}_h \in \mathcal{D}_h = [d_{L,h,}, d_{H,h}]$, $h = 1, ..., m$. *Then, for* $j = 1, ..., k$, $\bar{\beta}_j \in \mathcal{B}_j(\times_{h=1}^m \mathcal{D}_h \mid \underset{h=1,...,m}{sign}\ (A_{jh}))$ *defined as follows, and these bounds are sharp:*

$$\mathcal{B}_j([d_{L,1,}, d_{H,1}] \times ... \times [d_{L,m,}, d_{H,m}] \mid A_{j1} \leq 0, ..., A_{jg} \leq 0, 0 \leq A_{j\ g+1}, ..., 0 \leq A_{jm})$$
$$= [R_{Y.X|Z,j} - \textstyle\sum_{h=1}^g A_{jh} d_{L,h} - \sum_{h=g+1}^m A_{jh} d_{H,h},\ R_{Y.X|Z,j} - \sum_{h=1}^g A_{jh} d_{H,h} - \sum_{h=g+1}^m A_{jh} d_{L,h}].$$

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

3.4 Conditioning on Covariates

We extend the results in Sections 3.2 and 3.3 to accommodate conditioning on covariates S. For this, for a generic random vector A with $E(A)$ finite, let

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

For example, for $s \in \mathcal{S}$, $\bar{\beta}(s) \equiv E(\beta|S = s)$ denotes the average direct effects of X on Y given covariates $S = s$. Further, for generic random vectors B and C of equal dimension with $E[\tilde{C}(S)(\tilde{A}'(S), \tilde{B}'(S))|S = s]$ finite and $E(\tilde{C}(S)\tilde{B}'(S)|S = s)$ nonsingular, we define the conditional linear IV regression estimand and residual

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

so that by construction $E(\tilde{C}(S)\epsilon'_{A.B|C}(S)|S = s) = 0$. For example, for $k = \ell$ we write $R_{Y.X|Z}(s) \equiv E(\tilde{Z}(S)\tilde{X}'(S)|S = s)^{-1} E(\tilde{Z}(S)\tilde{Y}(S)|S = s)$. When $B = C$, we obtain $R_{A.B}(s) \equiv R_{A.B|B}(s)$ and $\epsilon_{A.B}(s) \equiv \epsilon_{A.B|B}(s)$. This notation reduces to that defined in Section 3.1 when S is degenerate, $S = 1$, in which case we leave S implicit.

Theorem 3.4 *Assume S.2 with* $\ell = k$ *and* $m = l$, *and that, for* $s \in \mathcal{S}$,

(i) $E(\tilde{Z}(S)\tilde{X}'(S)|S = s)$ *and* $\bar{\delta}_W(s)$ *are nonsingular,*

(ii) $Cov(\alpha_Y, Z|S = s) = 0$, $E(\tilde{\beta}(S)|Z, X, S = s) = 0$, *and* $E(\tilde{\delta}_Y(S)|U, Z, S = s) = 0$,

(iii) $Cov(\alpha_W, Z|S = s) = 0$ *and* $E(\tilde{\delta}_W(S)|U, Z, S = s) = 0$.

Let $\bar{\delta}(s) \equiv \bar{\delta}_W^{-1}(s)\bar{\delta}_Y(s)$ *then*

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

Here, $B(s) \equiv R_{Y.X|Z}(s) - \bar{\beta}(s) = R_{W.X|Z}(s)\ \bar{\delta}(s)$ denotes the conditional IV regression bias in measuring $\bar{\beta}(s)$. When $Z = X$, Theorem 3.4 gives $\bar{\beta}(s) = R_{Y.X}(s) - R_{W.X}(s)\bar{\delta}(s)$.

Conditions (ii) and (iii) in Theorem 3.4 are implied by the stronger (unnecessary under linearity) condition $(U_W, U_Y) \perp (U, Z, X)|S$. In particular, the covariate-conditioned conditions (ii) and (iii) in Theorem 3.4 weaken their unconditional uncorrelation and mean independence analogs in Theorem 3.1. Further, conditioning on covariates S may render Z "closer" to exogeneity and the conditional IV regression bias smaller. Theorem 3.4 reduces to Theorem 3.1 when S is degenerate, $S = 1$.

If the conditions in Theorem 3.4 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. If, in addition, the conditional average effects $\bar{\delta}_W(S)$, $\bar{\beta}(S)$, and $\bar{\delta}_Y(S)$ are constant, so that the mean independence assumptions involving δ_W, β, and δ_Y in Theorem 3.4 hold unconditionally, the law of iterated expectations gives

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

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 the conditional expectations $\bar{Z}(S)$, $\bar{X}(S)$, $\bar{W}(S)$, 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 $\tilde{X}$ (i.e. as the coefficients on X in IV regressions of Y and W respectively on $(1, X', S')'$ using instruments $(1, Z', S')'$).

From Theorem 3.4, we have that $\bar{\beta}_j(s)$ is fully identified under conditional exogeneity $(B_j(s) = 0)$ or conditional signed proportional confounding ($\bar{\delta}(s) = d(s)$, a vector of known or estimable functions of s). Otherwise, the expression for $\bar{\beta}(s)$ can be used, along with sign and magnitude restrictions on elements of $\bar{\delta}(s) \equiv \bar{\delta}_W^{-1}(s)\bar{\delta}_Y(s)$, involving the conditional average direct effects of U on Y and W given $S = s$, to partially identify $\bar{\beta}_j(s)$. This yields the sharp identification regions $\mathcal{B}_j(\times_{h=1}^m \mathcal{D}_h(s) \mid \underset{h=1,...,m}{sign}\ (A_{jh}(s)))$ for $\bar{\beta}_j(s)$ defined as in Corollary 3.3 with $\mathcal{D}_h(s) \equiv [d_{L,h}(s), d_{H,h}(s)]$ where $d_{L,h}(s)$ and $d_{H,h}(s)$ are known or estimable functions of s for $h = 1, ..., m$ and with $R_{Y.X|Z}(s)$ replacing $R_{Y.X|Z}$ and $A(s) \equiv R_{W.X|Z}(s)$ replacing A.

Appendix A contains extensions of the results in Section 3 on identification of average coefficients under magnitude and sign restrictions on confounding. Section A.1 studies a panel structure with individual and time varying random coefficients without requiring "fixed effects." Section A.2 studies cases where the proxies W are a component X_1 of X, included in the Y equation.

4 Identification of Average Nonparametric Effects

We extend the analysis in Section 3 by removing the linearity assumption S.2 and studying the identification of average nonparametric effects, with $Y = r(X, S, U, U_Y)$ as specified in S.1. As discussed in Section 2.2, here we study the identification of the conditional average direct effect of X on Y at (x, x^*) given $X = x^*$ and $S = s$:

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

such as the conditional average effect of the treatment on the treated $\bar{\beta}(0, 1|1, s)$ as well as the conditional average direct marginal effect X on Y at x given $X = x$ and $S = s$:

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

where we set $k = 1$ to label the cause of interest by X. We also study the identification of averages of $\bar{\beta}(x, x^*|x^*, s)$ and $\bar{\beta}(x|x, s)$ such as the average direct effect of X on Y at (x, x^*) given $X = x^*$, $\bar{\beta}(x, x^*|x^*) = E[\bar{\beta}(x, x^*|x^*, S)|X = x^*]$, and the conditional average marginal effect $\bar{\beta}(s) = E[\bar{\beta}(X|X, s)|S = s]$.

In studying the identification of $\bar{\beta}(x, x^*|x^*, s)$ and $\bar{\beta}(x|x, s)$, we find it useful to employ the following notation for the difference and derivative of a nonparametric regression. Specifically, for generic 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 define

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

4.1 Additive Separability

We begin our analysis by studying the case in which U enters r separably. In particular, we impose the following assumption. We remove separability in Section 4.2.

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:

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

and adjust the random coefficients $\theta \equiv (\underset{m\times 1}{\alpha_W}', vec(\underset{l\times m}{\delta_W})', \underset{l\times 1}{\delta_Y}')'$ correspondingly.

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[10] (2009), in which case certain average effects of X on Y are point identified. Next, 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 employ proxies to fully or partially identify average effects of X on Y under magnitude and sign restrictions on confounding.

For $s \in \mathcal{S}$ and $x, x^* \in \mathcal{X}$, note that S.3 and $U_Y \perp X|S = s$ give that

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

and

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

Theorem 4.1 gives conditions under which $\bar{\beta}(x, x^*|s)$ 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 4.1 *Assume S.3 with $m = l$ and that, for $s \in \mathcal{S}$,*

(i) $\bar{\delta}_W(s)$ is nonsingular,

(ii) $U_Y \perp X|S = s$ and $E(\tilde{\delta}_Y(S)|U, X, S = s) = 0$,

(iii) $E(\tilde{\alpha}_W(S)|X, S = s) = 0$ and $E(\tilde{\delta}_W(S)|U, X, S = s) = 0$.

Let $\bar{\delta}(s) \equiv \bar{\delta}_W^{-1}(s)\bar{\delta}_Y(s)$. Then for $x, x^* \in \mathcal{X}$,

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

Set $k = 1$ and suppose further that

(iv.a) $\frac{\partial}{\partial x}E(W'|X = x, S = s)$ exists and is finite and (iv.b) for all $x^\dagger \in \mathcal{N}(x) \subseteq \mathcal{X}$, a nonempty open neighborhood of x, $\frac{\partial}{\partial x}\ddot{r}(x^\dagger, s, u_y)$ exists for a.e.[11] 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

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

Conditions (ii) and (iii) of Theorem 4.1 strengthen their counterparts in Theorem 3.4 with $Z = X$. They are implied by the stronger condition $(U_W, U_Y) \perp (U, X)|S$. Note that (ii) 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 (iii) ensures that W is an informative proxy so that the

[10] 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.

[11] In this context, 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).

mean dependence of W on X given $S = s$ arises solely due to U. Condition (iv) 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 conditional nonparametric regression biases for $\bar{\beta}(x, x^*|s)$ and $\bar{\beta}(x|s)$ are

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

As in the linear case, $\bar{\beta}(x, x^*|s)$ and $\bar{\beta}(x|s)$ are fully identified under conditional exogeneity or signed proportional confounding. In particular, if $B(x, x^*|s) = 0$ or $B(x|s) = 0$ (conditional exogeneity) then $\bar{\beta}(x, x^*|s) = R^N_{Y.X}(x, x^*; s)$ or $\bar{\beta}(x|s) = R^N_{Y.X}(x; s)$. This obtains e.g. if $\bar{\delta}_Y(s) = 0$ or U is conditionally mean independent of X, $E(\tilde{U}(S)|X, S = s) = 0$. Alternatively, if $\bar{\delta}(s) = d(s)$ with $d_h(s)$, $h = 1, ..., m$, known or estimable (conditional signed proportional confounding) then

$$\begin{aligned}\bar{\beta}(x, x^*|s) &= R^N_{Y.X}(x, x^*; s) - \textstyle\sum_{h=1}^m R^N_{W_h.X}(x, x^*; s)d_h(s) \text{ and}\\ \bar{\beta}(x|s) &= R^N_{Y.X}(x; s) - \textstyle\sum_{h=1}^m R^N_{W_h.X}(x; s)d_h(s).\end{aligned}$$

From Theorem 4.1, 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, with $R^N_{W.X}(x^\dagger, x^\ddagger; s) \neq 0$ then $\bar{\delta}(s) = \frac{R^N_{Y.X}(x^\dagger, x^\ddagger; s)}{R^N_{W.X}(x^\dagger, x^\ddagger; s)}$ and $\bar{\beta}(x, x^*|s)$ is thus fully identified. Analogous restrictions fully identify $\bar{\beta}(x|s)$. We don't require such restrictions.

In the absence of assumptions yielding full identification, restrictions on the magnitude and 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 $\bar{\beta}(x, x^*|s) \in \mathcal{B}(\times_{h=1}^m \mathcal{D}_h(s) \mid \underset{h=1,...,m}{sign}\ (A_h(x, x^*; s)))$ and $\bar{\beta}(x|s) \in \mathcal{B}(\times_{h=1}^m \mathcal{D}_h(s) \mid \underset{h=1,...,m}{sign}\ (A_h(x; s)))$, where these sharp identification regions are defined analogously to Corollary 3.3 with $R^N_{Y.X}(x, x^*; s)$ or $R^N_{Y.X}(x; s)$ replacing $R_{Y.X|Z}$ and $A(x, x^*; s) \equiv R^N_{W.X}(x, x^*; s)$ or $A(x; s) \equiv R^N_{W.X}(x; s)$ replacing A.

4.2 Nonseparability

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

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

Theorem 4.2 characterizes the nonparametric bias of $R^N_{Y.X}(x, x^*; s)$ and $R^N_{Y.X}(x; s)$ in recovering $\bar{\beta}(x, x^*|x^*, s)$ and $\bar{\beta}(x|x, s)$ in this case. 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 Appendix B gives analogous results for discrete U, with sums replacing integrals. As in Theorem 4.1(iv), we impose regularity conditions to ensure that derivatives exist and justify interchanging the order of integral and derivative in expressions such as[12]

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

where we make use of $U_Y \perp (U,X)|S=s$ in the last equality. We collect these regularity conditions in Assumption B.1 of Appendix B. Recall the notation

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

Theorem 4.2 *Assume S.1 with $m=l=1$, $s \in \mathcal{S}$, and $x, x^* \in \mathcal{X}$. Suppose that $(U_W, U_Y) \perp (U,X)|S=s$ and that $F_{U|X,S}(\cdot|x^*,s)$ and $F_{U|X,S}(\cdot|x,s)$ are absolutely continuous.*
(i.a) If B.1.i(a,b) hold then

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

where

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

(i.b) If B.1.i(c,d) hold then

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

(ii) Set $k=1$. (ii.a) If, in addition to the conditions in (i.a), B.1.ii(a,b,c,d) hold then

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

where

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

(ii.b) If, in addition to the conditions in (i.b), B.1.ii(a,d,e) hold then

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

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

Theorem 4.2 characterizes the nonparametric regression omitted variable biases $B(x, x^*|x^*, s)$ and $B(x|x, s)$ for the identification of $\bar{\beta}(x, x^*|x^*, s)$ and $\bar{\beta}(x|x, s)$ and demonstrates how these biases depend 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 signs of these biases. For instance, if we assume that $\bar{\delta}_Y(u; x|s)$ is nonnegative for a.e. $u \in \mathcal{U}_s$ (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. $u \in \mathcal{U}_s$ holds (e.g. the probability of low ability U weakly decreases 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 e.g. if $U \perp X|S = s$ or if $\bar{\delta}_Y(\cdot; x|s) = 0$ for a.e. $u \in \mathcal{U}_s$. Alternatively, under conditional proportional confounding, $\bar{\delta}_Y(u; x|s) = d(x, s)\bar{\delta}_W(u|s)$ for a.e. $u \in \mathcal{U}_s$, 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 conditions sufficient for point identification, magnitude and sign restrictions on confounding yield sharp bounds.

Corollary 4.3 *Suppose that, for a.e.* $u \in \mathcal{U}_s$, $\bar{\delta}_Y(u; x|s) = d(u, x, s)\bar{\delta}_W(u|s)$ *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 4.2(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.* $u \in \mathcal{U}_s$ *or nonnegative for a.e.* $u \in \mathcal{U}_s$ *then*

$$\bar{\beta}(x, x^*|x^*, s) \in \mathcal{B}(\mathcal{D}(x, s) \mid sign(R^N_{W.X}(x, x^*; s)),$$

defined analogously to Corollary 3.3 with $R^N_{Y.X}(x, x^*; s)$ *replacing* $R_{Y.X|Z}$. *This bound is sharp. (ii) Under the conditions of Theorem 4.2(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*

$$\bar{\beta}(x|x, s) \in \mathcal{B}(\mathcal{D}(x, s) \mid sign(R^N_{W.X}(x; s)),$$

defined analogously to Corollary 3.3 with $R^N_{Y.X}(x; s)$ *replacing* $R_{Y.X|Z}$. *This bound is sharp.*

For example, the conditions of Theorem 4.3 hold if the average equation $E[q(s, u, U_W)|S = s]$ for the proxy W 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, discussed above, holds[13]. 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.

[13]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 $w \leq w'$ and all u then W is a monotone IV. We impose neither of these assumptions here.

5 Identification of Local, Marginal, and Average Treatment Effects

This section studies the identification of local and marginal treatment effects, as well as average treatment effects for the population, treated, and untreated, in a discrete choice structural system under magnitude and sign restrictions on confounding. Recall that, under assumption S.1, Y and W are generated by:

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

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

Assumption 4 (S.4) *Assume S.1 and suppose further that $\mathcal{S}$ generates X such that*[14]

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

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 accordingly.

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 particular 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 in which $Y = \ddot{r}(X, S, U_Y) + U'\delta_Y$. 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 5.2, we maintain that the vector of potential instruments Z satisfies $(U_X, U_Y) \perp Z|S = s$ (in contrast to $U_Y \perp X|S = s$ in Section 4) but allow U to freely depend on Z, with the random vector δ_Y possibly nonzero. We remove additive separability in Section 5.3 and study identification of average treatment effects under S.4. Last, to simplify the exposition, we let $F_{U_X|S}(\cdot|s)$ be absolutely continuous and thus strictly increasing, enabling the second equality in

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

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

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

5.1 Local and Marginal Treatment Effects

Following the literature (e.g. lmbens and Angrist, 1994; Heckman and Vytlacil, 2005), for $s \in \mathcal{S}$ and $z, z^* \in \mathcal{Z}$, 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 covariates $S = s$ and instruments $Z = z$ 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 $S = s$ recovers LATE given by

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

For example, when Z is binary, $\bar{\beta}(0,1|\nu(0,s) < U_X \leq \nu(1,s), s)$ denotes the average direct treatment effect for the conditional "compliers" who, given $S = s$, 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 in addition to S:

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

Given $U_X \perp Z|S = s$, averaging this marginal effect over the distribution of Z given $S = s$ recovers the MTE given by

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

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 what follows, we study the identification of these various local and marginal effects as well as weighted averages of MTEs, such as $\bar{\beta}(0,1|s)$, under magnitude and sign restrictions on confounding. To facilitate this study, we find it useful to define the following notation for the conditional Wald (1940) and local instrumental variable (LIV) estimands. In particular, for generic random variable B and vectors A and C with $E(A', B)$ finite, 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

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

5.2 Additive Separability

We begin our analysis by studying the additively separable specification for the Y and W equations given in $S.3$. Thus, in this subsection, we let

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

Section 5.3 removes the separability assumption. Note that, under S.3 and $(U_X, U_Y) \perp Z|S = s$, we have

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

Similarly, S.3 and $(U_X, U_Y) \perp Z|S = s$ give

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

Theorem 5.1 extends the results in Imbens and Angrist (1994) and Heckman and Vytlacil (2005) by characterizing $\bar{\beta}(0,1|\nu(z,s) < U_X \leq \nu(z^*,s),s)$ and $\bar{\beta}(0,1|\nu(z,s),s)$ under confounding and showing that the biases of the Wald and LIV methods for recovering these effects depend 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 5.1 *Assume S.3 and S.4 with $m = l$ and that, for $s \in \mathcal{S}$,*

(i) $\bar{\delta}_W(s)$ is nonsingular,

(ii) $(U_X, U_Y) \perp Z|S = s$ and $E(\tilde{\delta}_Y(S)|U,Z,S=s) = 0$,

(iii) $E(\tilde{\alpha}_W(S)|Z,S=s) = 0$ and $E(\tilde{\delta}_W(S)|U,Z,S=s) = 0$.

Let $\bar{\delta}(s) \equiv \bar{\delta}^{-1}_W(s)\bar{\delta}_Y(s)$. Then, for $z, z^* \in \mathcal{Z}$ with $\Pr[\nu(z,s) < U_X \leq \nu(z^*,s)\ |S=s] > 0$, we have

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

Set $\ell = 1$ and suppose further that

(iv.a) $\frac{\partial}{\partial z}E(W'|Z=z,S=s)$ exists and (iv.b) $\nu(\cdot,s)$ is differentiable at z with $\frac{\partial}{\partial z}\nu(z,s) \neq 0$ and that $\bar{\beta}(0,1|\cdot,s)$ and $f_{U_X|S}(\cdot|s)$ are continuous at $\nu(z,s)$ with $f_{U_X|S}(\nu(z,s)|s) > 0$. Then

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

Conditions (ii) and (iii) of Theorem 5.1 are implied by the stronger condition $(U_X, U_W, U_Y) \perp (U, Z)|S$. As discussed above, condition (ii) weakens the restriction $(U_X, U_Y) \perp Z|S$ with $\delta_Y = 0$. Condition (iii) 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 (iv) enable applying theorems for the derivative of an integral. The conditional Wald and LIV biases for $\bar{\beta}(0,1|\nu(z,s) \leq U_X < \nu(z^*,s),s)$ and $\bar{\beta}(0,1|\nu(z,s),s)$ are given by

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

These biases vanish under conditional exogeneity, which occurs for example if $\bar{\delta}_Y = 0$ or U is conditionally mean independent of Z, $E(\tilde{U}(S)|Z, S = s) = 0$. 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:

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

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

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

Last, from Theorem 5.1, observe 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 $\bar{\delta}(s) = \frac{R^{Wald}_{Y.X|Z}(\dot{z},\ddot{z};s) - R^{Wald}_{Y.X|Z}(z^\dagger,z^\ddagger;s)}{R^{Wald}_{W.X|Z}(\dot{z},\ddot{z};s) - R^{Wald}_{W.X|Z}(z^\dagger,z^\ddagger;s)}$, provided $R^{Wald}_{W.X|Z}(z^\dagger,z^\ddagger;s) \neq R^{Wald}_{W.X|Z}(\dot{z},\ddot{z};s)$, and $\bar{\beta}(0,1|\nu(z,s) \leq U_X < \nu(z^*,s),s)$ is fully identified. Analogous restrictions fully identify the local effect $\bar{\beta}(0,1|\nu(z,s),s)$. We don't require such assumptions.

When the conditions for point identification do not hold, magnitude and sign restrictions on confounding partially identify LATE and MTE:

$$\bar{\beta}(0,1|\nu(z,s) \leq U_X < \nu(z^*,s),s) \in \mathcal{B}(\times_{h=1}^m \mathcal{D}_h(s) \mid \underset{h=1,...,m}{sign}\ (A_h(z,z^*;s))\ ),$$

and

$$\bar{\beta}(0,1|\nu(z,s),s) \in \mathcal{B}(\times_{h=1}^m \mathcal{D}_h(s) \mid \underset{h=1,...,m}{sign}\ (A_h(z;s))\ ),$$

where these sharp identification regions are defined analogously to Corollary 3.3 with $R^{Wald}_{Y.X|Z}(z, z^*; s)$ or $R^{LIV}_{Y.X|Z}(z; s)$ replacing $R_{Y.X|Z}$ and with $A(z, z^*; s) \equiv R^{Wald}_{W.X|Z}(z, z^*; s)$ or $A(z; s) \equiv R^{LIV}_{W.X|Z}(z; s)$ replacing A.

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 5.1 using $X = \mathbf{1}\{V \leq P\}$ with potential instrument P, gives that $\bar{\beta}(0, 1|p, s) = R^{LIV}_{Y.X|P}(p, s) - R^{LIV}_{W.X|P}(p, s)\ \bar{\delta}(s)$ for almost every p. If P has the unit interval for support given $S = s$, we have that the conditional average treatment effect is characterized by:

$$\bar{\beta}(0, 1|s) = \int_0^1 \bar{\beta}(0, 1|p, s)dp = \int_0^1 \left[R^{LIV}_{Y.X|P}(p, s) - \textstyle\sum_{h=1}^m R^{LIV}_{W_h.X|P}(p, s)\bar{\delta}_h(s)\right] dp.$$

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

$$\bar{\beta}(0, 1|X = 1, s) = \int_0^1 \left[R^{LIV}_{Y.X|P}(p; s) - \textstyle\sum_{h=1}^m R^{LIV}_{W_h.X|P}(p; s)\bar{\delta}_h(s)\right]\frac{(1 - F_{P|S}(p|s))}{E(P(Z, S)|S = s)}\ dp,$$

and

$$\bar{\beta}(0, 1|X = 0, s) = \int_0^1 \left[R^{LIV}_{Y.X|P}(p; s) - \textstyle\sum_{h=1}^m R^{LIV}_{W_h.X|P}(p; s)\bar{\delta}_h(s)\right]\frac{F_{P|S}(p|s)}{E(1 - P(Z, S)|S = s)}\ dp,$$

where $F_{P|S}(\cdot|s)$ denotes the conditional distribution of P given $S = s$. 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 3.3.

5.3 Nonseparability

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

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

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 5.2 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 B.2 in Appendix B collects regularity conditions justifying interchanging the order of well-defined integral and derivative. Here, recall the notation

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

Theorem 5.2 *Assume S.1 and S.4 with* $m = l = 1$, $s \in \mathcal{S}$, *and* $z, z^* \in \mathcal{Z}$ *such that* $\Pr[\nu(z,s) < U_X \leq \nu(z^*,s) \mid S = s] > 0$. *Suppose that* $(U_X, U_W, U_Y) \perp (U, Z)|S = s$ *and that* $F_{U|Z,S}(\cdot|z^*, s)$ *and* $F_{U|Z,S}(\cdot|z, s)$ *are absolutely continuous.*
(i.a) If B.2.i(a,b) hold then

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

where

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

(i.b) If B.2.i(c,d) hold then

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

(ii) Set $\ell = 1$. *(ii.a) If, in addition to the conditions in (i.a), B.2.ii(a,b,c,d) hold then*

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

where

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

(ii.b) If, in addition to the conditions in (i.b), B.2.ii(a,d,e) 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 5.2 characterizes the nonparametric omitted variable bias of the of Wald and LIV methods for the identification of local and marginal treatment effects. It demonstrates how these biases depend on the average marginal effect of U on Y as well as on the conditional distribution of U given Z and S. For example, if we assume that $\bar{\delta}_Y(u; z|s)$ is nonnegative for a.e. $u \in \mathcal{U}_s$ (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. $u \in \mathcal{U}_s$ holds (e.g. the probability of low ability U weakly decreases with proximity to a college) then $B(0,1|\nu(z,s) < U_X \leq \nu(z^*,s), z^*, s)$ is nonnegative.

Under conditional exogeneity, we have $B(0,1|\nu(z,s), z, s) = 0$ and thus $R^{LIV}_{Y.X|Z}(z; s)$ identifies $\bar{\beta}(0,1|\nu(z,s), z, s)$. This occurs e.g. if $U \perp Z|S = s$ or if $\bar{\delta}_Y(u; z|s) = 0$ 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, $\bar{\beta}(0,1|\nu(z,s),z,s) = R^{LIV}_{Y.X|Z}(z;s) - d(z,s)\ R^{LIV}_{W.X|Z}(z;s)$. Analogous results obtain 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 sign restrictions on confounding yield sharp bounds.

Corollary 5.3 *Suppose that, for a.e.* $u \in \mathcal{U}_s$, $\bar{\delta}_Y(u;z|s) = d(u,z,s)\bar{\delta}_W(u|s)$ *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 5.2(i), if* $\bar{\delta}_W(u|s)\ [F_{U|X,S}(u|z^*,s) - F_{U|X,S}(u|z,s)]$ *is either nonpositive for a.e.* $u \in \mathcal{U}_s$ *or nonnegative for a.e.* $u \in \mathcal{U}_s$ *then*

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

defined analogously to Corollary 3.3 with $R^{Wald}_{Y.X|Z}(z,z^*;s)$ *replacing* $R_{Y.X|Z}$. *This bound is sharp. (ii) Under the conditions of Theorem 5.2(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*

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

defined analogously to Corollary 3.3 with $R^{LIV}_{Y.X|Z}$ *replacing* $R_{Y.X|Z}$. *This bound is sharp.*

Using Corollary 5.3, 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 5.2.

To conclude this section, we remark 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.

6 Estimation and Inference

6.1 Asymptotic Normality

We obtain a uniformly consistent set estimator $\hat{\mathcal{B}}_j$ for an identification region $\mathcal{B}_j$ by using uniformly consistent estimators for its bounds. In particular, consider the sharp identification regions under restrictions on confounding of the form $[b_{L,j}, b_{H,j}]$, a bounded interval of finite width. As discussed below, $b_{L,j}$ and $b_{H,j}$ are linear combinations of covariate-conditioned (IV) regression estimands of Y and W on X (using instruments Z). Thus, one can construct

estimators $(\hat{b}'_{L,j},\hat{b}'_{H,j})'$ for $(b'_{L,j},b'_{H,j})'$ as a linear combination of uniformly consistent and jointly asymptotically normal parametric, semiparametric, or nonparametric (e.g. kernel) estimators for the underlying (IV) regression estimands.

To illustrate, consider our empirical application studying the wage equation. There, the vector $(X',S')'$ of causes X (e.g. education) and covariates S (e.g. family background) is a high-dimensional vector of binary and discrete variables. For example, here we're interested in the identification regions $\mathcal{B}([0,1] \mid sign(R^N_{W.X}(x,x^*;s))$ and $\mathcal{B}([-1,1])$ for effects such as $\bar{\beta}(x,x^*|x^*,s)$. The bounds in these identification regions 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 linear transformation of $E(Y|X,S)$ and $E(W|X,S)$ evaluated at (x^*,s) and (x,s). One can estimate these conditional means using nonparametric (e.g. kernel) regression that can "smooth" discrete regressors since, for high-dimensional $(X',S')'$, there may be cells with few or no data points in a given sample. Alternatively, it's convenient in this case to employ a flexible specification, which can be made fairly general given that $(X',S')'$ is discrete. In particular, consider the specification that we employ in the empirical application with e.g. scalar U (ability) and W (logarithm of test score) and vectors $G_X \equiv g_X(X)$ and $G_S \equiv g_S(S)$ of known functions of X and S such that

$$Y = \alpha_Y + g_X(X)'\gamma + U\delta_Y \quad \text{and} \quad W = \alpha_W + U\delta_W, \tag{5}$$

where we subsume G_S into the random coefficients α_Y and α_W. Provided that $\bar{\delta}_W(G_S)$, $\bar{\gamma}(G_S)$, and $\bar{\delta}_Y(G_S)$ are constant and $\bar{G}_X(G_S)$, $\bar{W}(G_S)$, and $\bar{Y}(G_S)$ are affine functions of G_S, applying Theorem 3.4 with G_X and G_S replacing X and S respectively yields $\bar{\gamma} = R_{Y.G} - R_{W.G}\ \bar{\delta}$ where we put $G \equiv (G'_X,G'_S)'$. More generally, we need not set $Z = X$ in this case, and letting $H_Z = h_Z(Z)$ with $\bar{H}_Z(G_S)$ affine in G_S yields $\bar{\gamma}_j = R_{Y.G|H,j} - R_{W.G|H,j}\bar{\delta}$ where $H = (H'_Z,G'_S)'$. The average effects of X on Y at (x,x^*) are then encoded by linear transformations $[g_X(x^*) - g_X(x)]'\bar{\gamma}$ of $\bar{\gamma}$. For instance, if X_j is the j^{th} component of G_X and the effect of X_j on Y is linear then $\bar{\beta}_j = \bar{\gamma}_j$ and the bounds $R_{Y.G|H,j}$, $R_{Y-W.G|H,j}$, or $R_{Y+W.G|H,j}$ in the identification regions $\mathcal{B}_j([0,1]\ |sign(R_{W.G|H,j}))$ and $\mathcal{B}_j([-1,1])$ for $\bar{\beta}_j$ are linear transformations of $(R'_{Y.G|H},R'_{W.G|H})'$. We thus derive the joint 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 $H = G$.

We stack the observations $\{A_i\}_{i=1}^n$ of a generic $d\times 1$ vector A into the $n\times d$ matrix $\mathbf{A}$. Also, we let $\tilde{A}_i \equiv A_i - \frac{1}{n}\sum_{i=1}^n A_i$. Further, for generic observations $\{A_i,B_i,C_i\}_{i=1}^n$ corresponding to A and the random vectors B and C of equal dimension, we let

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

denote the linear IV regression estimator and sample residuals respectively.

The next theorem employs standard arguments to derive the asymptotic distribution of $\sqrt{n}((\hat{R}'_{Y.G|H}, \hat{R}'_{W.G|H})' - (R'_{Y.G|H}, R'_{W.G|H})')$. For this, we let $Q \equiv diag(E(\tilde{H}\tilde{G}'), E(\tilde{H}\tilde{G}'))$. We let W be scalar here to simplify the notation.

Theorem 6.1 *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 $E(\tilde{H}\tilde{G}')$ 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 $P \in \mathcal{P}$

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

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 6.1.

As discussed above, the joint asymptotic distribution of the estimators for the bounds on the average effects of X on Y at (x, x^*) under restrictions on confounding obtains in this case as a linear transformations of that of $\sqrt{n}(\hat{R}'_{Y.G|H}, \hat{R}'_{W.G|H})'$. For example, consider the linear effect $\bar{\beta}_j = \bar{\gamma}_j = R_{Y.G|H,j} - R_{W.G|H,j}\ \bar{\delta}$ in equations (5) and include in $R_1 \equiv (R'_{Y.G|H}, R'_{Y-W.G|H})'$ the bounds of the identification region $\mathcal{B}_j([0,1] \mid sign(R_{W.G|H,j})$ 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 6.1 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

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

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 $\hat{\sigma}^2_{ab} \equiv (\frac{1}{n}\tilde{\mathbf{H}}'\tilde{\mathbf{G}})^{-1}(\frac{1}{n}\sum_{i=1}^{n}\tilde{H}_i\hat{\epsilon}_{a,i}\hat{\epsilon}_{b,i}\tilde{H}')(\frac{1}{n}\tilde{\mathbf{G}}'\tilde{\mathbf{H}})^{-1}$.

Uniformly consistent and jointly asymptotically normal parametric, semiparametric, or nonparametric estimators $(\hat{b}'_{L,j}, \hat{b}'_{H,j})'$ for $(b'_{L,j}, b'_{H,j})'$ can be derived analogously.

6.2 Confidence Intervals

This subsection discusses constructing a $1-\alpha$ confidence interval (CI) for an average effect that is partially identified in sharp bounded interval $[b_{L,j}, b_{H,j}]$ of finite width, such as when $\bar{\beta}_j = \bar{\gamma}_j$ and $\bar{\beta}_j \in \mathcal{B}_j([0,1] \mid sign(R_{W.G|H,j}))$ and $\mathcal{B}_j([-1,1])$ with scalar U and W as considered in the empirical application. Let $(\hat{b}'_{L,j}, \hat{b}'_{H,j})'$ denote estimators for $(b'_{L,j}, b'_{H,j})'$ that are jointly asymptotically normally distributed uniformly in $P \in \mathcal{P}$, as in Theorem 6.1. Let $\hat{\sigma}^2_{L,j}$ denote the uniformly in $P \in \mathcal{P}$ consistent estimator for $\sigma^2_{L,j} \equiv Avar(\sqrt{n}(\hat{b}_{L,j} - b_{L,j}))$ and define $\hat{\sigma}^2_{H,j}$ similarly. We construct[15] a $1-\alpha$ CI for $\bar{\beta}_j$ using:

$$[\hat{b}_{L,j} - c_\alpha \frac{\hat{\sigma}_{L,j}}{\sqrt{n}},\ \hat{b}_{H,j} + c_\alpha \frac{\hat{\sigma}_{H,j}}{\sqrt{n}}],$$

with the critical value c_α chosen appropriately as we discuss next.

Consider a $1-\alpha$ CI for the average effect $\bar{\beta}_j \in \mathcal{B}_j([-1,1])$ from equations (5) and suppose that this effect is linear, $\bar{\beta}_j = \bar{\gamma}_j$, so that $[b_{L,j}, b_{H,j}] = [R_{Y.G|H,j} - \left|R_{W.G|H,j}\right|,\ R_{Y.G|H,j} + \left|R_{W.G|H,j}\right|]$. Picking $c_\alpha = c_{1,\alpha}$ with $\Phi(c_{1,\alpha}) - \Phi(-c_{1,\alpha}) = 1-\alpha$, where Φ denotes the standard normal cumulative density function, (e.g. $c_{1,0.05} = 1.96$) 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 $b_{L,j}$ or $b_{H,j}$. Further, as discussed in Imbens and Manski (2004), picking $c_\alpha = c_{2,\alpha}$ with $\Phi(c_{2,\alpha}) - \Phi(-c_{2,\alpha}) = 1-2\alpha$ (e.g. $c_{2,0.05} = 1.645$) 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 $\bar{\beta}_j \in \mathcal{B}_j$ by setting $c_\alpha = c_{3,\alpha}$ with

$$\Phi(c_{3,\alpha} + \frac{\sqrt{n}(\hat{b}_{H,j} - \hat{b}_{L,j})}{\max\{\hat{\sigma}_{L,j}, \hat{\sigma}_{H,j}\}}) - \Phi(-c_{3,\alpha}) = 1-\alpha.$$

For $\mathcal{B}_j([-1,1])$, by construction, $\hat{b}_{H,j} - \hat{b}_{L,j} = 2\left|R_{W.G|H,j}\right| \geq 0$ 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 also report an estimate of the half as large sharp identification region $\mathcal{B}_j([0,1] \mid sign(R_{W.G|H,j}))$ and a CI for $\bar{\beta}_j$ 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

[15] An alternative method considers the union over confidence intervals for $\bar{\beta}_j(\bar{\delta})$ generated for each $\bar{\delta} \in [-1,1]$ or $\bar{\delta} \in [0,1]$ as in Chernozhukov, Rigobon, and Stoker (2010).

follow the literature (e.g. Reinhold and Woutersen, 2009; Nevo and Rosen, 2012) and report the estimated identification interval $\hat{\mathcal{B}}_j([0,1] \mid sign(R_{W.G|H,j}) = sign(\hat{R}_{W.G|H,j}))$ for $\bar{\beta}_j$ and the confidence interval $CI_{\bar{\beta}_j,1-\alpha}(sign(R_{W.G|H,j}) = sign(\hat{R}_{W.G|H,j}))$ 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 $\bar{\beta}_j$.

7 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 measurement error 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. Lang and Manove (2011) provide a model which suggests that one should control for test scores as well as education 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

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

where 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. The confounder U denotes unobserved skill or "ability" and is potentially correlated with elements of X given the covariates S discussed below. 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 Card (1995). The sample[16] 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[17], 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[18] 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. In addition to S, X, W, and Y, the sample contains data on potential instruments Z discussed below.

The identification regions under magnitude and sign restrictions on confounding for average effects such as $\bar{\beta}(x, x^*|x^*, s)$, corresponding e.g. to the return to education and black-white wage gap, have bounds that are linear transformations of $E(Y|X, S)$ and $E(W|X, S)$ evaluated at (x^*, s) and (x, s) where $(X', S')'$ is a high-dimensional vector of binary or discrete variables. It's convenient in this case to employ the specification in equations (5) discussed in Section 6. In particular, let $G \equiv (G_X', G_S')'$ where $G_X = g_X(X)$ and $G_S = g_S(S)$ be vectors of functions of X and S and consider the wage and proxy equations

$$Y = \alpha_Y + G_X'\gamma + U\delta_Y \quad \text{and} \quad W = \alpha_W + U\delta_W,$$

where we subsume into α_Y and α_W the vector G_S. In addition to S, we let G_S contain 8 binary indicators for interacted mother and father high school, college, or post graduate education. For example, when G_X consists of education, experience, experience squared, and the black binary indicator, we obtain the specification in Card (1995, e.g. table 2, column 5). Note that 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. Here, the average financial return to education is $100\bar{\gamma}_1\%$ and the average black-white wage gap is $100\bar{\gamma}_4\%$. More generally, we allow below for nonlinear effects in equations (5). 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 of the wage equation. Thus, $100\bar{\delta}_Y\%$

[16] This sample is reported at http://davidcard.berkeley.edu/data_sets.html and in Wooldridge (2008).

[17] 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.

[18] 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.

and $100\bar{\delta}_W\%$ denote 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.

While our identification results do not require the specification in equations (5), it's useful to employ these equations in what follows as they encompass specifications common in the literature and generalize these to allow for unobserved confounders and nonlinear random effects, thereby facilitating comparing our findings to the literature.

We apply Theorem 3.4 to equations (5) with G_X, a vector of instruments $H_Z = h_Z(Z)$, and G_S replacing X, Z, and S respectively, and maintain that $\bar{\delta}_W(G_S)$, $\bar{\gamma}(G_S)$, and $\bar{\delta}_Y(G_S)$ are constants, and $\bar{H}_Z(G_S)$, $\bar{G}_X(G_S)$, $\bar{W}(G_S)$, and $\bar{Y}(G_S)$ are affine functions of G_S. Putting $H = (H_Z', G_S')'$, this characterizes $\bar{\gamma} = R_{Y.G|H} - R_{W.G|H}\ \bar{\delta}$, and therefore the effects $\bar{\beta}$. In addition, we maintain two assumptions. First, we assume that KWW is, on average, at least as directly elastic or sensitive to ability as wage is, $\left|\bar{\delta}_Y\right| \leq \left|\bar{\delta}_W\right|$. Thus, a 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 X (or Z) given S. As discussed in the Introduction, 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 identification regions. Nevertheless, 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 negative average returns to education and a black-white wage gap in favor of blacks, which is inconsistent with the general findings in the literature. 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.

Table 1 reports results applying the methods discussed in Section 3 and 4.1, using $G_X \equiv g_X(X)$ and conditioning on $G_S \equiv g_S(S)$. In particular, following Card (1995, table 2, column 5), we begin by letting G_X consist of education, experience, experience squared, and the black indicator, with G_S as described above. Thus, under this specification the components γ_1 and γ_4 of γ encode the return to education and black-white wage gap respectively. Column 1 reports regression estimates using $\hat{R}_{Y.G,j}$ (which consistently estimates $\bar{\gamma}_j$ under conditional exogeneity) 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[19], with robust s.e. in parentheses, are 7.2%, (0.4%), and −18.7%, (2.0%), respectively.

[19] In all tables, we also report point and interval estimates for the coefficients associated with experience and

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 sign and magnitude restrictions on confounding, $0 \leq \frac{\bar{\delta}_Y}{\bar{\delta}_W} \leq 1$, 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 $CI_{\bar{\beta}_1,0.95}$ $[-0.6\%, 7.8\%]$ and that for the black-white wage gap is $[-18.7\%, 1.5\%]$ with $CI_{\bar{\beta}_3,0.95}$ $[-21.9\%, 5.0\%]$. 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 $\bar{\gamma}_j$ obtained under magnitude restrictions only, along with the uniformly valid 95% confidence intervals $CI_{\bar{\gamma}_j,0.95}$. Note that weakening the assumptions $\bar{\delta} \in [0,1]$ or $\bar{\delta} \in [-1,1]$ to $\bar{\delta} \in [0,d]$ or $\bar{\delta} \in [-d,d]$ 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 negative average returns to education and a black-white wage gap in favor of blacks. Specifically, in this case and subject to rounding error, the estimated identification regions for the average return to education and black-wage gap under sign and magnitude restrictions are respectively:

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

Last, we note that similar results (with slightly wider bounds for the average black-white wage gap) obtain when we let $G_S = S_1$, a subset of S consisting of two indicators for living in the South and SMSA respectively, as in Card (1995, table 2, column 1). 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 given the observables such as education.

As discussed above, conditioning on the proxy W for the confounder U, as is commonly done, does not generally ensure recovering $\bar{\beta}$ from a regression of Y on $(1, X', W, S')'$ except in special cases such as when W is a perfect rescaling[20] of U given S. Nevertheless, this may attenuate the regression bias (see e.g. Wickens, 1972; Battistin and Chesher, 2009; and Ogburna and VanderWeele, 2012). Table 2 reports estimates from a linear regression of Y on $(1, G', W')'$. Conditioning on KWW, instead of $W = \log(KWW)$, yields very similar estimates of $\bar{\gamma}$. These estimates and the corresponding confidence intervals lie respectively within the

experience squared. For brevity, we don't discuss these in detail.

[20]For example, it suffices in equations (5) that δ_W, γ, and δ_Y are constants, $Cov(\alpha_Y, (G'_X, W)'|G_S) = 0$, and α_W is a deterministic function of G_S

identification regions and $CI_{\bar{\gamma}_j,0.95}$ reported in Table 1. In particular, this regression's 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. The estimate of the coefficient on W (which, as discussed above, estimates $\frac{\bar{\delta}_Y}{\bar{\delta}_W}$ consistently in the special case when W is a perfect rescaling of U given S) is 0.2 with robust s.e. 0.03. We also 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.

We also 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[21] the conditional on G_S results. Under sign and magnitude 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 nonblack subpopulation, if at all. We follow Card (1995) and maintain that these average returns are equal.

When $Z = X$, Theorem 3.4 assumes that education is conditionally uncorrelated with unobserved determinants of wage and KWW other than ability. This assumption can fail 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[22] for education, experience, and experience squared in the specification from Table 1 with covariates S. 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. Note, however, that this paper's method does not require Z to be conditionally uncorrelated with ability (for example, Carneiro and Heckman (2002) provide evidence suggesting that distance to college may be endogenous). As reported in Table 4, under sign and magnitude restrictions on confounding, the conditional on S IV-based identification region estimate [2.9%, 13.4%] for the average return to education 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

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

[22] Specifically, we let $H = (H_Z', G_S')'$ here where $H_Z = h_Z(Z)$ consists of the proximity to college indicator, age, age squared, and the black indicator.

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 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[23], age, and age squared as potential instruments for education, experience, and experience squared. Last, in both IV specifications, conditioning on S_1 only yields generally similar results. However, in some cases, this leads to tighter identification regions albeit with wider confidence intervals at times (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 doesn't require linear or parametric effects of X on Y and can accommodate less restrictive specifications. Next, we relax the linearity of the return to education in the previous specification by letting G_X contain binary indicators for having at least t years of education, where $t = 2, ..., 18$ as in the sample, instead of the total years of education, thus 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[24]. We find 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 sign and magnitude 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 can't 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 (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 $CI_{\bar{\gamma}_{12},0.95}$ $[-3.4\%, 11.6\%]$. Graph 1 illustrates the nonlinearity in the return to education; it plots 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 sign and magnitude restrictions on confounding as well as magnitude restrictions only. Last, the estimate of the sharp identification region for the black-white wage

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

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

gap under sign and magnitude 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 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 doesn't require several commonly employed assumptions. In particular, (1) it doesn't require regressor or instrument exogeneity or restrict the dependence between X or Z and U (given S), (2) it doesn't require a linear return to education, and (3) it permits test scores to be error-laden proxies for unobserved ability.

8 Conclusion

This paper studies measuring average causal effects in structural systems under restrictions on the magnitude and sign of confounding, without conditional exogeneity of causes, treatment, or instruments given covariates. In particular, we study the partial and full 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. We characterize the omitted variables bias, due to confounders U, of regression and IV methods for the identification of these various average effects, thereby extending the classic linear omitted variable bias representation. Using proxies W for the confounders U, we ask how do the average direct effects of U on Y compare in magni-

tude and sign to those of U on W. Exogeneity (zero average direct effect) and proportional confounding (equal to a known proportion direct effects) are limiting cases yielding full identification of the average effects. Alternatively, we partially identify the effects of X on Y in sharp bounded intervals when W is sufficiently sensitive to U, and may obtain sharp upper or lower bounds otherwise. After studying estimation and confidence intervals, the paper applies its methods to study the return to education and the black-white wage gap using data from the 1976 subset of NLSYM used in Card (1995). Under the imposed weaker than exogeneity restrictions on confounding, we partially identify in sharp bounded intervals the average financial return to education as well as the average black-white wage. 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 bound estimates are generally narrower than the IV-based ones, with especially narrower confidence intervals. In particular, we find 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 the average return to education for the black subpopulation may differ slightly from the nonblack subpopulation, if at all, and that, under the imposed assumptions, this data is inconclusive about the extent of discrimination in the labor market. Extensions for future work include 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: Extensions

Appendix A extends the results in Section 3 on the identification of average random coefficients under sign and magnitude restrictions to study a panel structure and cases with proxies included in the Y equation. Throughout, we set $S = 1$ to simplify the exposition.

A.1 Panel with Individual and Time Varying Random Coefficients

We consider a panel structure whereby we index the variables M_t and coefficients θ_t in S.2 by $t = 1, 2$. Here, U may denote time-invariant unobserved individual characteristics. We allow the proxy W_t for U to be an element $\underset{k_1 \times 1}{X_{t1}}$ of X_t. Thus, for $t = 1, 2$:

$$Y_t = \alpha_{Y_t} + X_t'\beta_t + U'\delta_{Y_t} \quad \text{and} \quad X_{t1}' = \alpha_{X_{t1}}' + U'\delta_{X_{t1}}.$$

This is a panel structure with individual (we omit the index i for succinctness) and time varying random coefficients where we do not require "fixed effects" and thus δ_{Y_t} need not equal $\delta_{Y_{t'}}$.

For $t, t' = 1, 2$, $t \neq t'$, we apply Theorem 3.1, using $X_{t'1}$ as proxy, to derive an expression for $\bar{\beta}_t$. In this case, the conditions in Theorem 3.1 require that (i) $E(\tilde{Z}_t\tilde{X}_t')$ and $\bar{\delta}_{X_{t'1}}$ are nonsingular, (ii) $Cov(\alpha_{Y_t}, Z_t) = 0$, $E(\tilde{\beta}_t|Z_t, X_t) = 0$, $E(\tilde{\delta}_{Y_t}|U, Z_t) = 0$, and (iii) $Cov(\tilde{\alpha}_{X_{t'1}}, Z_t) = 0$, $E(\tilde{\delta}_{X_{t'1}}|U, Z_t) = 0$. Condition (iii) restricts the dependence of the random coefficients in the $X_{t'1}$ equation with U and non-contemporaneous Z_t. Then, with $\bar{\delta}_t \equiv \bar{\delta}_{X_{t'1}}^{-1}\bar{\delta}_{Y_t}$, Theorem 3.1 gives that for $t, t' = 1, 2$, $t \neq t'$,

$$\bar{\beta}_t = R_{Y_t.X_t|Z_t} - \bar{\delta}_t R_{X_{t'1}.X_t|Z_t}.$$

The IV regression bias is $B_t \equiv R_{Y_t.X_t|Z_t} - \bar{\beta}_t = \bar{\delta}_t R_{X_{t'1}.X_t|Z_t}$. Thus, $\bar{\beta}_t$ is fully identified under exogeneity ($B_t = 0$) or signed proportional confounding ($\bar{\delta}_{th} = d_{th}$, $h = 1, ..., k_1$, with known d_{th}). Applying Corollary 3.3 here with $A_t \equiv R_{X_{t'1}.X_t|Z_t}$ sharply partially identifies $\bar{\beta}_{tj}$ for $t = 1, 2$ and $j = 1, ..., k$ so that $\bar{\beta}_{tj} \in \mathcal{B}_{tj}(\times_{h=1}^{k_1}\mathcal{D}_{th} \mid \underset{h=1,...,k_1}{sign}(A_{tjh}))$. The restrictions on the magnitude and sign of the average direct effects of U on Y_t and $X_{t'1}$ may be plausible, for example, if one suspects that Y, at time t, is less directly responsive to U than X_1 is at times t and t'.

A.2 Included Proxies

Sometimes, a researcher may want to allow proxies W to directly impact the response Y. In this case, W is a component X_1 of X. While Theorem 3.1 does not rule out that W and X have common elements, its conditions entail restrictions on Z in this case and Appendix A.2 enables relaxing these. First, when $W = X_1$, conditions (i) and (iii) of Theorem 3.1 imply

that all elements of Z must be endogenous (i.e. correlated with U) since $E(\tilde{Z}\tilde{X}')$ is singular otherwise. Section A.2.1 relaxes this restriction by studying the "under-identification" case where there are fewer valid instruments than the dimension of X. Second, when $W = X_1$, the requirement that $Cov(\alpha_W, Z) = 0$ in condition (iii) generally rules out that Z contains elements of X_1. Section A.2.2 relaxes this restriction by studying the case of multiple proxies for U that are included in the Y equation and allowed to be elements of Z.

A.2.1 "Under-Identification" Using Valid Instruments

When $W = X_1$, Theorem 3.1 requires all elements of Z to be correlated with U. Sometimes a vector Z_1 of one or a few valid instruments may be available, albeit the dimension of X may exceed that of Z_1. Nevertheless, a researcher may wish to employ the exogenous instrument Z_1. The next Theorem allows for this possibility and provides an expression for $\bar{\beta}$ which depends on the average direct effects of U on Y and X_1.

Theorem A.1 *Assume S.2 with* $\underset{\ell\times 1}{Z} \equiv (\underset{\ell_1\times 1}{Z_1}', \underset{\ell_2\times 1}{Z_2}')'$, $\underset{k\times 1}{X} \equiv (\underset{k_1\times 1}{X_1}', \underset{k_2\times 1}{X_2}')'$, $W = X_1$, *with* $\ell_1, \ell_2 \geq 0$, $\ell = k$, $k_1 = l$, *and*

(i) $E(\tilde{Z}\tilde{X}')$ *and* $\bar{\delta}_{X_1}$ *are nonsingular,*

(ii) $Cov(U, Z_1) = 0$,

(iii) $Cov(\alpha_Y, Z) = 0$, $E(\tilde{\beta}|Z, X) = 0$, *and* $E(\tilde{\delta}_Y|U, Z) = 0$,

(iv) $Cov(\alpha_{X_1}, Z_2) = 0$ *and* $E(\tilde{\delta}_{X_1}|U, Z_2) = 0$.

Let $A \equiv E(\tilde{Z}\tilde{X}')^{-1}\left[\ 0',\ \ E(\tilde{Z}_2\tilde{X}_1')'\ \right]'$ *and* $\bar{\delta} \equiv \bar{\delta}_{X_1}^{-1}\bar{\delta}_Y$ *then*

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

The IV regression bias is $B \equiv R_{Y.X|Z} - \bar{\beta} = A\bar{\delta}$. The conditions in Theorem A.1 are analogous to those in Theorem 3.1, except that they assume that Z_1 is uncorrelated with U and let Z_1 freely depend on the coefficients in the proxy X_1 equation. Thus, if $Z = Z_2$, Theorem A.1 reduces to Theorem 3.1 with $W = X_1$. Instead, if $Z = Z_1$ then exogeneity holds. Here, $\bar{\beta}_j$ is fully identified under exogeneity ($B_j = 0$) or signed proportional confounding ($\bar{\delta}_h = d_h$, $h = 1, ..., k_1$, with d_h known). Otherwise, $\bar{\beta}_j$ is sharply partially identified in $\mathcal{B}_j(\times_{h=1}^{k_1}\mathcal{D}_h \mid \underset{h=1,...,k_1}{sign}\ (A_{jh}))$ under assumptions on how the average direct effects of U on X_1 compare in magnitude and sign to those of U on Y.

A.2.2 Multiple Included Proxies

When $W = X_1$, the assumption $Cov(\alpha_W, Z) = 0$ in condition (iii) of Theorem 3.1 generally rules out that X_1 is a component of Z and therefore that $Z = X$. We relax this requirement

and let $W = (X_1', X_2')'$ with X_1 and X_2 two vectors of proxies included in the equation for Y and where X_1, and possibly X_2, is a component of Z.

The next Theorem derives an expression for $\bar{\beta}$ which depends on the unknowns $\bar{\delta}_{X_1}^{-1}\bar{\delta}_Y$ and $\bar{\delta}_{X_2}^{-1}\bar{\delta}_Y$ involving the average direct effects of U on Y and those of U on X_1 and X_2. Here, we let $Z_1 = X_1$, with Z potentially equal to X.

Theorem A.2 *Assume S.2 and let* $W = (\underset{k_1\times 1}{X_1}', \underset{k_2\times 1}{X_2}')'$ *with* $X_g' = \alpha_{X_g}' + U'\delta_{X_g}$, *for* $g = 1, 2$, $\underset{k\times 1}{X} = (W', \underset{k_3\times 1}{X_3}')'$, $Z_1 = X_1$, $\underset{\ell\times 1}{Z} \equiv (\underset{\ell_1\times 1}{Z_1}', \underset{\ell_2\times 1}{Z_2}')'$, $k_1 = k_2 = l$, $k_3 \geq 0$, $\ell = k$, *and that*

(i) $E(\tilde{Z}\tilde{X}')$, $\bar{\delta}_{X_1}$, $\bar{\delta}_{X_2}$ *are nonsingular,*

(ii) $Cov(\alpha_Y, Z) = 0, E(\tilde{\beta}|Z, X) = 0,$ *and* $E(\tilde{\delta}_Y|U, Z) = 0,$

(iii) $Cov(\alpha_{X_1}, (U', Z_2', X_2')') = 0$ *and* $E(\tilde{\delta}_{X_1}|U, Z_2, X_2) = 0,$

(iv) $Cov(\alpha_{X_2}, U) = 0$ *and* $E(\tilde{\delta}_{X_2}|U) = 0.$

Let $\bar{\delta}_1 \equiv \bar{\delta}_{X_1}^{-1}\bar{\delta}_Y$ *and* $\bar{\delta}_2 \equiv \bar{\delta}_{X_2}^{-1}\bar{\delta}_Y$ *then*

$$\bar{\beta} = R_{Y.X|Z} - E(\tilde{Z}\tilde{X}')^{-1}\begin{bmatrix} E(\tilde{Z}_1\tilde{X}_2')\bar{\delta}_2 \\ E(\tilde{Z}_2\tilde{X}_1')\bar{\delta}_1 \end{bmatrix}.$$

The IV regression bias is

$$B \equiv R_{Y.X|Z} - \bar{\beta} = E(\tilde{Z}\tilde{X}')^{-1}\begin{bmatrix} E(\tilde{Z}_1\tilde{X}_2')\bar{\delta}_2 \\ E(\tilde{Z}_2\tilde{X}_1')\bar{\delta}_1 \end{bmatrix}.$$

The conditions in Theorem A.2 extend those in Theorem 3.1 to allow the proxies X_1 and X_2 to be components of Z but they restrict the dependence between the proxy X_2 and the coefficients α_{X_1} and δ_{X_1} in the equation for the proxy X_1 as well as the dependence between U and $(\alpha_{X_1}', \delta_{X_1}', \alpha_{X_2}', \delta_{X_2}')'$. Here too, Z and X may be associated with U.

We use the expression for $\bar{\beta}$ in Theorem A.2 to fully or partially identify the elements of $\bar{\beta}$.

Corollary A.3 *Assume the conditions of Theorem A.2 and let* $X_{2,3} \equiv (X_2', X_3')'$ *and* $j = 1, ..., k$. *(i) If* $B_j = 0$ *(exogeneity) then* $\bar{\beta}_j = R_{Y.X|Z,j}$. *(ii) If* $\bar{\delta}_1 = c_1$ *and* $\bar{\delta}_2 = c_2$ *(signed proportional confounding) then*

$$\bar{\beta} = R_{Y.X|Z} - E(\tilde{Z}\tilde{X}')^{-1}\begin{bmatrix} \sum_{h=1}^{l} E(\tilde{Z}_1\tilde{X}_{2h}')c_{2h} \\ \sum_{h=1}^{l} E(\tilde{Z}_2\tilde{X}_{1h}')c_{1h} \end{bmatrix}.$$

In particular, let $d = (c_1', c_2')'$, $\bar{\delta} = (\bar{\delta}_1', \bar{\delta}_2')'$, $\underset{k_1\times k_1}{P_1} \equiv E(\epsilon_{Z_1.Z_2|X_{2,3}}\tilde{X}_1')$, $\underset{(k_2+k_3)\times(k_2+k_3)}{P_2} \equiv E(\epsilon_{Z_2.Z_1|X_1}\tilde{X}_{2,3}')$, *and*

$$\underset{k\times(k_1+k_2)}{A} \equiv \begin{bmatrix} -R_{X_{2,3}.X_1|Z_1}P_2^{-1}E(\tilde{Z}_2\tilde{X}_1'), & P_1^{-1}E(\tilde{Z}_1\tilde{X}_2') \\ P_2^{-1}E(\tilde{Z}_2\tilde{X}_1'), & -R_{X_1.X_{2,3}|Z_2}P_1^{-1}E(\tilde{Z}_1\tilde{X}_2') \end{bmatrix}.$$

Then

$$\bar{\beta} = R_{Y.X|Z} - B = R_{Y.X|Z} - A\bar{\delta},$$

and

$$\bar{\beta}_j = R_{Y.X|Z,j} - \sum_{h=1}^{2l} A_{jh} d_h \quad \textit{for } j = 1, ..., k.$$

Thus, $\bar{\beta}_j$ is fully identified under exogeneity ($B_j = 0$) or signed proportional confounding ($\bar{\delta}_h = d_h$, $h = 1, ..., 2l$, with d_h known). Otherwise, $\bar{\beta}_j$ is sharply partially identified in $\mathcal{B}_j(\times_{h=1}^{2l} \mathcal{D}_h \mid \underset{h=1,...,2l}{sign}\ (A_{jh}))$, defined analogously to Corollary 3.3, under assumptions on how the average direct effects of U on X_1 and X_2 compare in magnitude and sign to those of U on Y.

B Appendix B: Mathematical Proofs

Proof of Theorem 3.1 Apply Theorem 3.4 with $S = 1$.

Proof of Corollary 3.2 The proof is immediate.

Proof of Corollary 3.3 We have the following bounds for $h = 1, ..., m$:

$$\begin{aligned} -A_{jh} d_{L,h} \leq -A_{jh} \bar{\delta}_h \leq -A_{jh} d_{H,h} \quad & \text{if } A_{jh} \leq 0 \\ -A_{jh} d_{H,h} \leq -A_{jh} \bar{\delta}_h \leq -A_{jh} d_{L,h} \quad & \text{if } 0 \leq A_{jh} \end{aligned}$$

The identification regions then follow from $\bar{\beta}_j = R_{Y.X|Z,j} - \sum_{h=1}^{m} A_{jh} \bar{\delta}_h$ with $A_{jh} \leq 0$ for $h = 1, ..., g$ and $0 \leq A_{jh}$ for $h = g + 1, ..., m$. For sharpness, note that the identification region $\mathcal{B}_j$ for $\bar{\beta}_j$ is generated via the linear mapping $L : \mathcal{D}_1 \times ... \times \mathcal{D}_m \to \mathcal{B}_j$ given by $b = R_{Y.X|Z,j} - \sum_{h=1}^{m} A_{jh} d_h$. 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 3.1 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 3.4 By (ii) we have

$$\begin{aligned} E(\tilde{Z}(S)\tilde{Y}(S)|S = s) &= E(\tilde{Z}(S)Y|S = s) = E(\tilde{Z}(S)(\alpha_Y + X'\beta + U'\delta_Y)|S = s) \\ &= E(\tilde{Z}(S)\tilde{X}'(S)|S = s)\bar{\beta}(s) + E(\tilde{Z}(S)U'|S = s)\bar{\delta}_Y(s) \end{aligned}$$

and by (iii) we have

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

Since $E(\tilde{Z}(S)\tilde{X}'(S)|S = s)$ and $\bar{\delta}_W(s)$ are nonsingular, we have

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

Proof of Theorem 4.1 For $x \in \mathcal{X}$, (ii) gives

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

and by (iii) we have

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

Since $\bar{\delta}_W(s)$ is nonsingular, we have

$$E(Y|X=x,S=s) = E[\ddot{r}(x,s,U_Y)|S=s] + [E(W'|X=x,S=s) - E(\alpha'_W|S=s)]\ \bar{\delta}_W^{-1}(s)\bar{\delta}_Y(s).$$

It follows that for $x, x^* \in \mathcal{X}$,

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

Further, by (iv),

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

and $\frac{\partial}{\partial x}E[\ddot{r}(x,s,U_Y)|S=s] = E[\frac{\partial}{\partial x}\ddot{r}(x,s,U_Y)|S=s]$ (see e.g. White and Chalak, 2013, Theorem 4.2), yielding

$$\bar{\beta}(x|s) = \frac{\partial}{\partial x}E(Y|X=x,S=s) - \frac{\partial}{\partial x}E(W'|X=x,S=s)\bar{\delta}(s).$$

We make use of the following regularity conditions in the proof of Theorem 4.2. For this, we put $\bar{r}(x,u|s) \equiv E[r(x,s,u,U_Y)|S=s]$ and $\bar{q}(u|s) \equiv E[q(s,u,U_W)|S=s]$. It is implicitly assumed that referenced derivatives exist.

Assumption 5 (B.1) *Let $s \in \mathcal{S}$, $x \in \mathcal{X}$, 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) $\bar{r}(x,\cdot|s)$ is absolutely continuous on $\mathcal{U}_s$,

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

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

(i.d) for a.e. u there is a function $\Gamma_{1,u}(u_w)$ with $E_{U_W|S}[\Gamma_{1,u}(U_W)|S=s] < \infty$ such that for all $u^\dagger \in \mathcal{N}(u)$, $\left|\frac{\partial}{\partial u}q(s,u^\dagger,u_w)\right| \leq \Delta_{1,u}(u_w)$ 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) there is a function $\Delta_2(u)$ with $\int_{\mathcal{U}_{x,s}} \Delta_2(u)du < \infty$ such that for all $x^\dagger \in \mathcal{N}(x)$,

$\left|\frac{\partial}{\partial x}\{\bar{r}(x^\dagger, u|s)\ f_{U|X,S}(u|x^\dagger, s)\}\right| \leq \Delta_2(u)$ *for a.e.* u,

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

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

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

The absolute continuity of $\bar{r}(x, \cdot|s)$ and $\bar{q}(\cdot|s)$ on $\mathcal{U}_s$ in B.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 4.2: (i) Adding and subtracting $E[r(x, s, U, U_Y)|X = x, S = s]$, 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

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

Since $F_{U|X,S}(\cdot|x^*, s)$ and $F_{U|X,S}(\cdot|x, s)$ are absolutely continuous:

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

B.1$(i.a)$ justifies integration by parts, which gives

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

with $\underline{u}$ and $\bar{u}$ the (possibly infinite) infimum and supremum over $\mathcal{U}_s$. The first term vanishes and, by B.1$(i.b)$ and $U_Y \perp (U, X)|S = s$, we obtain

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

Similarly, $U_W \perp (U,X)|S=s$ and B.1$(i.c, i.d)$ give

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

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

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

By B.1$(ii.a)$ and $(ii.b)$, we interchange the order of derivative and integral and apply the product rule:

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

By $(ii.c)$ 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 $(ii.d)$ $\frac{\partial}{\partial x}F_{U|X,S}(\cdot|x,s)$ is absolutely continuous on $\mathcal{U}_{x,s}$. By $(i.a)$, integration by parts gives

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

where we use $(i.b)$ and $U_Y \perp (U,X)|S=s$ in the second to last equality.

Similarly, $U_W \perp (U,X)|S=s$ and B.1$i(c,d)$ and B.1$ii(a,d,e)$ give

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

The next Theorem characterizes the nonparametric regression bias for $\bar{\beta}(x,x^*|x^*,s)$ and $\bar{\beta}(x|x,s)$ in the case of discrete U.

Theorem B.1 *Assume S.1 with $m = l = 1$, $s \in \mathcal{S}$, and $x, x^* \in \mathcal{X}$, and suppose that $(U_W, U_Y) \perp (U,X)|S=s$. Suppose that $\mathcal{U}_s = \{u_0, u_1, ..., u_L\}$ with $u_{g-1} < u_g$ for $g = 1, ..., L$.*
(i) Then

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

where

$$B(x, x^*|x^*, s) = -\sum_{g=1}^{L} \bar{\delta}_Y(u_{g-1}, u_g; x|s)\ [F_{U|X,S}(u_{g-1}|x^*, s) - F_{U|X,S}(u_{g-1}|x, s)].$$

Further, we have

$$R^N_{W.X}(x, x^*; s) = -\sum_{g=1}^{L} \bar{\delta}_W(u_{g-1}, u_g|s)[F_{U|X,S}(u|x^*, s) - F_{U|X,S}(u|x, s)].$$

(ii) Set $k = 1$ and suppose further that $\frac{\partial}{\partial x}\bar{r}(x, u_g|s)$ and $\frac{\partial}{\partial x}f_{U|X,S}(u_g|x, s)$ exist for all $u_g \in \mathcal{U}_s$. Then

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

where

$$B(x|x, s) = -\sum_{g=1}^{L} \bar{\delta}_Y(u_{g-1}, u_g; x|s)\frac{\partial}{\partial x}F_{U|X,S}(u_{g-1}|x, s).$$

Further, we have

$$R^N_{W.X}(x; s) = -\sum_{g=1}^{L} \bar{\delta}_W(u_{g-1}, u_g|s)\frac{\partial}{\partial x}F_{U|X,S}(u_{g-1}|x, s).$$

Proof of Theorem B.1: From the proof of Theorem 4.2, $U_Y \perp (U, X)|S = s$ gives

$$B(x, x^*|x^*, s) = E_{U|X,S}[\ \bar{r}(x, U|s)\ |X = x^*, S = s] - E_{U|X,S}[\ \bar{r}(x, U|s)\ |X = x, S = s].$$

The expression for $B(x, x^*|x^*, s)$ follows from

$$\begin{aligned}
E_{U|X,S}[\bar{r}(x, U|s)\ |X = x^*, S = s] &= \sum_{h=0}^{L} \bar{r}(x, u_h|s)f_{U|X,S}(u_h|x^*, s) \\
&= \bar{r}(x, u_0|s)[1 - \sum_{h=1}^{L} f_{U|X,S}(u_h|x^*, s)] + \sum_{h=1}^{L} \bar{r}(x, u_h|s)f_{U|X,S}(u_h|x^*, s) \\
&= \bar{r}(x, u_0|s) + \sum_{h=1}^{L} f_{U|X,S}(u_h|x^*, s)[\bar{r}(x, u_h|s) - \bar{r}(x, u_0|s)] \\
&= \bar{r}(x, u_0|s) + \sum_{h=1}^{L} f_{U|X,S}(u_h|x^*, s)[\sum_{g=1}^{h} \bar{r}(x, u_g|s) - \bar{r}(x, u_{g-1}|s)] \\
&= \bar{r}(x, u_0|s) + \sum_{g=1}^{L} [\bar{r}(x, u_g|s) - \bar{r}(x, u_{g-1}|s)] \sum_{h=g}^{L} f_{U|X,S}(u_h|x^*, s) \\
&= \bar{r}(x, u_0|s) + \sum_{g=1}^{L} [\bar{r}(x, u_g|s) - \bar{r}(x, u_{g-1}|s)][1 - F_{U|X,S}(u_{g-1}|x^*, s)].
\end{aligned}$$

A similar derivation gives the expression for $R^N_{W|X}(x, x^*|x^*, s)$.

(ii) From the proof of Theorem 4.2, we have

$$R^N_{Y.X}(x; s) = \frac{\partial}{\partial x} E[r(x, s, U, U_Y)|X = x, S = s] = \frac{\partial}{\partial x} E_{U|X,S}[\bar{r}(x, U|s)|X = x, S = s].$$

Since $\frac{\partial}{\partial x}\bar{r}(x, u_l|s)$ and $\frac{\partial}{\partial x} f_{U|X,S}(u_l|x, s)$ exist for for all $u_l \in \mathcal{U}_s$,

$$R^N_{Y.X}(x; s) = \sum_{g=0}^{L} \frac{\partial}{\partial x}\bar{r}(x, u_g|s) f_{U|X,S}(u_g|x, s) + \sum_{g=0}^{L} \bar{r}(x, u_g|s) \frac{\partial}{\partial x} f_{U|X,S}(u_g|x, s) \equiv \bar{\beta}(x|x, s) + B(x|x, s),$$

where we use $U_Y \perp (U, X)|S = s$. Further,

$$\begin{aligned}
B(x|x, s) &= \sum_{h=0}^{L} \bar{r}(x, u_h|s) \frac{\partial}{\partial x} f_{U|X,S}(u_h|x, s) \\
&= \bar{r}(x, u_0|s) \frac{\partial}{\partial x}[1 - \sum_{h=1}^{L} f_{U|X,S}(u_h|x, s)] + \sum_{h=1}^{L} \bar{r}(x, u_h|s) \frac{\partial}{\partial x} f_{U|X,S}(u_h|x, s) \\
&= \sum_{h=1}^{L} \frac{\partial}{\partial x} f_{U|X,S}(u_h|x, s)[\bar{r}(x, u_h|s) - \bar{r}(x, u_0|s)] \\
&= \sum_{h=1}^{L} \frac{\partial}{\partial x} f_{U|X,S}(u_h|x, s)[\sum_{g=1}^{h} \bar{r}(x, u_g|s) - \bar{r}(x, u_{g-1}|s)] \\
&= \sum_{g=1}^{L} [\bar{r}(x, u_g|s) - \bar{r}(x, u_{g-1}|s)] \sum_{h=g}^{L} \frac{\partial}{\partial x} f_{U|X,S}(u_h|x, s) \\
&= -\sum_{g=1}^{L} \bar{\delta}_Y(u_{g-1}, u_g; x|s) \frac{\partial}{\partial x} F_{U|X,S}(u_{g-1}|x, s).
\end{aligned}$$

A similar derivation gives the expression for $R^N_{W|X}(x|x, s)$.

Proof of Corollary 4.3: Since $\bar{\delta}_Y(u; x|s) = d(u, x, s)\bar{\delta}_W(u|s)$ we have

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

Let $\omega(u) \equiv \bar{\delta}_W(u|s)\ [F_{U|X,S}(u|x^*, s) - F_{U|X,S}(u|x, s)]$ then

$$\begin{aligned}
d_H(x, s)\omega(u) \le d(u, x, s)\omega(u) \le d_L(x, s)\omega(u) \quad &\text{if } \omega(u) \le 0, \text{ and} \\
d_L(x, s)\omega(u) \le d(u, x, s)\omega(u) \le d_H(x, s)\omega(u) \quad &\text{if } 0 \le \omega(u).
\end{aligned}$$

Since $\omega(u)$ does not change sign for a.e. $u \in \mathcal{U}_s$, its sign is that of $\int_{\mathcal{U}_s} \omega(u)du = -R^N_{W.X}(x, x^*; s)$. The result then follows from

$$\begin{aligned}
&-d_L(x, s)R^N_{W.X}(x, x^*; s) \le -B(x, x^*|x^*, s) \le -d_H(x, s)R^N_{W.X}(x, x^*; s) \quad \text{if } R^N_{W.X}(x, x^*; s) \le 0 \text{ and} \\
&-d_H(x, s)R^N_{W.X}(x, x^*; s) \le -B(x, x^*|x^*, s) \le -d_L(x, s)R^N_{W.X}(x, x^*; s) \quad \text{if } 0 \le R^N_{W.X}(x, x^*; s).
\end{aligned}$$

For sharpness, for $R^N_{W.X}(x,x^*;s) \neq 0$ and each $b \in \mathcal{B}(\mathcal{D}(x,s) \mid sign(R^N_{W.X}(x,x^*;s))$, 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 then analogous to (i) and omitted.

Proof of Theorem 5.1: By (ii), we have

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

$\Pr[\nu(z,s) < U_X \leq \nu(z^*,s)\ |S=s] > 0$ and arguments similar to those in the proof of Theorem 5.2 give

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

Further, by (iii),

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

and (i) then gives

$$\{E[U'|Z=z^*,S=s] - E[U'|Z=z,S=s]\}\ \bar{\delta}_Y(s) = R^N_{W.Z}(z,z^*;s)\ \bar{\delta}^{-1}_W(s)\bar{\delta}_Y(s).$$

Dividing by $R^N_{X.Z}(z,z^*;s) \neq 0$ then yields

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

To characterize $\bar{\beta}(0,1|\nu(z,s),s)$, note that

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

By $(iv.b)$, arguments similar to those in the proof of Theorem 5.2 give

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

Further, by (iii) and $(iv.a)$,

$$R^N_{W.Z}(z;s) = \frac{\partial}{\partial z}E(U'|Z=z,S=s)\ \bar{\delta}_W(s).$$

Dividing by $R^N_{X.Z}(z;s) \neq 0$ and using (i) gives

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

For Theorem 5.2, we make use of regularity conditions collected in Assumption B.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 (B.2) *Let $s \in \mathcal{S}$, $z \in \mathcal{Z}$, and denote by $\mathcal{N}(u) \subseteq \mathcal{U}$ and $\mathcal{N}(z) \subseteq \mathcal{Z}$ nonempty open neighborhoods of u and z respectively.*

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

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

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

(i.d) for a.e. u there is a function $\Upsilon_{1,u}(u_w)$ with $E_{U_W|S}[\Upsilon_{1,u}(U_W)|S = s] < \infty$ such that for all $u^\dagger \in \mathcal{N}(u)$, $\left|\frac{\partial}{\partial u} q(s, u^\dagger, u_w)\right| \leq \Upsilon_{1,u}(u_w)$ 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) there is a function $\Phi_2(u)$ with $\int_{\mathcal{U}_{x,s}} \Phi_2(u)du < \infty$ such that for all $z^\dagger \in \mathcal{N}(z)$, $\left|\frac{\partial}{\partial z}\{\bar{r}(z^\dagger, u|s)\ f_{U|Z,S}(u|z^\dagger, s)\}\right| \leq \Phi_2(u)$ for a.e. u,

(ii.c) $\frac{\partial}{\partial z}\nu(z,s) \neq 0$ and $\bar{\beta}(0, 1; u|U_X = \cdot, S = s)$ for a.e. u and $f_{U_X|S}(\cdot|s)$ are continuous at $\nu(z,s)$ with $f_{U_X|S}(\nu(z,s)|s) > 0$,

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

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

The absolute continuity of $\bar{r}(z, \cdot|s)$ and $\bar{q}(\cdot|s)$ on $\mathcal{U}_s$ in B.2 ensure 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 5.2: Adding and subtracting $E(Y|Z = z, S = s)$ gives

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

where

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

Further,

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

$\Pr[\nu(z,s) < U_X \leq \nu(z^*,s) \mid S=s] > 0$ 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

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

Thus, dividing $\gamma(z,z^*|z^*,s)$ by $R^N_{X.Z}(z,z^*;s)$ gives

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

By $(U_X,U_Y) \perp (U,Z)|S=s$, we have

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

Since $F_{U|Z,S}(\cdot|z^*,s)$ and $F_{U|Z,S}(\cdot|z,s)$ are absolutely continuous then

$$B_\gamma(z,z^*|z^*,s) = \int_{\mathcal{U}_s} \bar{r}(z,u|s)[f_{U|Z,S}(u|z^*,s) - f_{U|X,S}(u|z,s)]du.$$

By B.2$(i.a)$, integration by parts gives

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

with $\underline{u}$ and $\bar{u}$ the (possibly infinite) infimum and supremum over $\mathcal{U}_s$. The first term vanishes and the result then follows by noting that B.2$(i.b)$ and $(U_X,U_Y) \perp (U,Z)|S=s$ give

$$B_\gamma(z,z^*|z^*,s) = -\int_{\mathcal{U}_s} \bar{\delta}_Y(u;z|s)[F_{U|X,S}(u|z^*,s) - F_{U|X,S}(u|z,s)]du.$$

Similarly, B.2.$i(c,d)$, integration by parts, and $U_W \perp Z|S=s$ give

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

(ii) By $(U_X, U_Y) \perp (U, Z)|S = s$, we have

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

Further, by B.2.$ii(a,b)$

$$R^N_{Y.Z}(z;s) = \frac{\partial}{\partial z} \int_{\mathcal{U}_{z,s}} \bar{r}(z,u|s)\ f_{U|Z,S}(u|z,s)du$$
$$= \int_{\mathcal{U}_{z,s}} \frac{\partial}{\partial z}\bar{r}(z,u|s)\ f_{U|Z,S}(u|z,s)du + \int_{\mathcal{U}_{z,s}} \bar{r}(z,u|s)\ \frac{\partial}{\partial z} f_{U|Z,S}(u|z,s)\ du.$$

Note that

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

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

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

where we make use of $(U_X, U_Y) \perp (U, Z)|S = s$ in the last equality.

Similarly, using B.2($ii.c$)

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

B.2($ii.d$) gives that $\frac{\partial}{\partial z}F_{U|Z,S}(\cdot|z,s)$ is absolutely continuous on $\mathcal{U}_{z,s}$. B.2($i.a$) enables integration by parts which gives

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

where we use B.2($i.b$) in the last equality. Dividing $R^N_{Y.Z}(z;s)$ by $R^N_{X.Z}(z;s)$ gives

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

Similarly, B.2.$i(c,d)$ and B.2.$ii(a,d,e)$, integration by parts, and $U_W \perp Z|S=s$ give

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

Proof of Corollary 5.3: The proof is analogous to that of Corollary 4.3 and omitted.

Proof of Theorem 6.1 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 $\hat{M} = \frac{1}{n}\sum_{i=1}^n (\tilde{H}_i' \epsilon_{Y.G|H,i}, \tilde{H}_i' \epsilon_{W.G|H,i})'$. 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, by (i), $\hat{Q}^{-1} - Q^{-1} = o_p(1)$ uniformly in $P \in \mathcal{P}$ and, by (ii), $\sqrt{n}\hat{M} \xrightarrow{d} N(0,\Xi)$, with Ξ finite and positive definite, uniformly in $P \in \mathcal{P}$.

Proof of Theorem A.1 S.2 ensures finiteness of moments. By (iii), we have

$$E(\tilde{Z}\tilde{Y}) = E[\tilde{Z}(\alpha_Y + X'\beta + U'\delta_Y)] = E(\tilde{Z}\tilde{X}')\bar{\beta} + E(\tilde{Z}U')\bar{\delta}_Y.$$

By (ii) and since $E(\tilde{Z}\tilde{X}')$ is nonsingular, we have

$$\bar{\beta} = R_{Y.X|Z} - E(\tilde{Z}\tilde{X}')^{-1}E(\tilde{Z}U')\bar{\delta}_Y = R_{Y.X|Z} - E(\tilde{Z}\tilde{X}')^{-1}\left[\ 0',\quad E(\tilde{Z}_2U')'\ \right]'\bar{\delta}_Y.$$

By (iv), we have

$$E(\tilde{Z}_2\tilde{X}_1') = E(\tilde{Z}_2X_1') = E[\tilde{Z}_2(\alpha_{X_1}' + U'\delta_{X_1})] = E(\tilde{Z}_2U')\bar{\delta}_{X_1},$$

so that by (i)

$$E(\tilde{Z}_2U') = E(\tilde{Z}_2\tilde{X}_1')\bar{\delta}_{X_1}^{-1}.$$

It follows that

$$\bar{\beta} = R_{Y.X|Z} - E(\tilde{Z}\tilde{X}')^{-1}\left[\ 0',\quad E(\tilde{Z}_2\tilde{X}_1')'\ \right]'\bar{\delta}_{X_1}^{-1}\bar{\delta}_Y.$$

Proof of Theorem A.2 S.2 ensures finiteness of moments. By (ii), we have

$$E(\tilde{Z}\tilde{Y}) = E(\tilde{Z}Y) = E(\tilde{Z}\tilde{X}')\bar{\beta} + E(\tilde{Z}U')\bar{\delta}_Y$$

Further, (iii) gives

$$E(\tilde{Z}_2\tilde{X}_1') = E[\tilde{Z}_2(\alpha_{X_1}' + U'\delta_{X_1})] = E(\tilde{Z}_2U')\bar{\delta}_{X_1},$$

and $Z_1 = X_1$, (iii), and (iv) give

$$\begin{aligned} E(\tilde{Z}_1\tilde{X}_2') &= \bar{\delta}_{X_1}'E(U\tilde{X}_2') = \bar{\delta}_{X_1}'E[\tilde{U}(\alpha_{X_2}' + U'\delta_{X_2})] = \bar{\delta}_{X_1}'E(\tilde{U}\tilde{U}')\bar{\delta}_{X_2}, \text{ and} \\ E(\tilde{Z}_1U') &= E[(\alpha_{X_1} + \delta_{X_1}'U)\tilde{U}'] = \bar{\delta}_{X_1}'E(\tilde{U}\tilde{U}'). \end{aligned}$$

Since $E(\tilde{Z}\tilde{X}')$, $\bar{\delta}_{X_1}$, and $\bar{\delta}_{X_2}$ are nonsingular, we have

$$\bar{\beta} = R_{Y.X|Z} - E(\tilde{Z}\tilde{X}')^{-1}\begin{bmatrix} E(\tilde{Z}_1\tilde{X}_2')\bar{\delta}_{X_2}^{-1}\bar{\delta}_Y \\ E(\tilde{Z}_2\tilde{X}_1')\bar{\delta}_{X_1}^{-1}\bar{\delta}_Y \end{bmatrix}.$$

Proof of Corollary A.3 The result follows from the expression for $\bar{\beta}$ in Theorem A.2. For the expression for $\bar{\beta}_j$, recall that $E(\tilde{Z}\tilde{X}')^{-1}$ is given by (e.g. Baltagi, 1999, p. 185):

$$E(\tilde{Z}\tilde{X}')^{-1} = \begin{bmatrix} E(\tilde{Z}_1\tilde{X}_1'), & E(\tilde{Z}_1\tilde{X}_{2,3}') \\ E(\tilde{Z}_2\tilde{X}_1'), & E(\tilde{Z}_2\tilde{X}_{2,3}') \end{bmatrix}^{-1} = \begin{bmatrix} P_1^{-1}, & -R_{X_{2,3}.X_1|Z_1}P_2^{-1} \\ -R_{X_1.X_{2,3}|Z_2}P_1^{-1}, & P_2^{-1} \end{bmatrix},$$

where

$$\begin{aligned} P_1 &\equiv E(\tilde{Z}_1\tilde{X}_1') - E(\tilde{Z}_1\tilde{X}_{2,3}')E(\tilde{Z}_2\tilde{X}_{2,3}')^{-1}E(\tilde{Z}_2\tilde{X}_1') = E(\epsilon_{Z_1.Z_2|X_{2,3}}\tilde{X}_1') \\ P_2 &\equiv E(\tilde{Z}_2\tilde{X}_{2,3}') - E(\tilde{Z}_2\tilde{X}_1')E(\tilde{Z}_1\tilde{X}_1')^{-1}E(\tilde{Z}_1\tilde{X}_{2,3}') = E(\epsilon_{Z_2.Z_1|X_1}\tilde{X}_{2,3}'). \end{aligned}$$

The result then follows from

$$\bar{\beta} = R_{Y.X|Z} - \begin{bmatrix} P_1^{-1}E(\tilde{Z}_1\tilde{X}_2')\bar{\delta}_2 - R_{X_{2,3}.X_1|Z_1}P_2^{-1}E(\tilde{Z}_2\tilde{X}_1')\bar{\delta}_1 \\ -R_{X_1.X_{2,3}|Z_2}P_1^{-1}E(\tilde{Z}_1\tilde{X}_2')\bar{\delta}_2 + P_2^{-1}E(\tilde{Z}_2\tilde{X}_1')\bar{\delta}_1 \end{bmatrix}.$$

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

		$\hat{R}_{Y.G,j}$	$\hat{\mathcal{G}}_j([0,1])$	$\hat{R}_{W.G,j}$	$\hat{\mathcal{G}}_j([-1,1])$
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

		$\hat{R}_{Y.G,j}$	$\hat{\mathcal{G}}_j([0,1])$	$\hat{R}_{W.G,j}$	$\hat{\mathcal{G}}_j([-1,1])$
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 including in G_X an interaction term multiplying 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

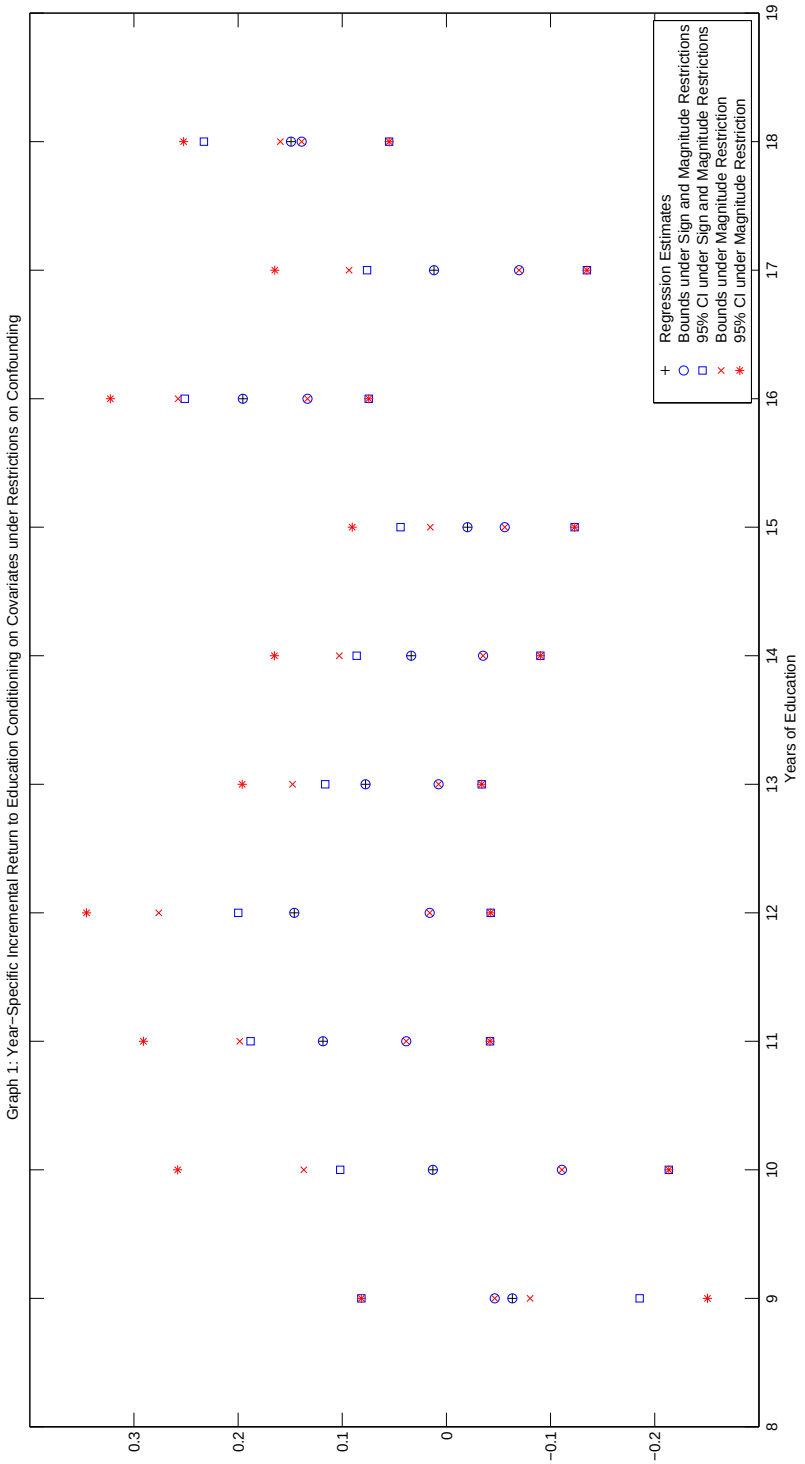
		$\hat{R}_{Y.G\|H,j}$	$\hat{\mathcal{G}}_j([0,1])$	$\hat{R}_{W.G\|H,j}$	$\hat{\mathcal{G}}_j([-1,1])$
1	Education	0.134	[0.029,0.134]	0.106***	[0.029,0.240]
	Robust s.e.	(0.052)	-	(0.032)	-
	$CI_{.95}$ and $CI_{\bar{\gamma}_j,.95}$	[0.032,0.237]	[-0.061,0.220]	-	[-0.061,0.351]
2	Experience	0.061	[0.006,0.061]	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, and age squared as instruments H_Z for education, experience, and experience squared 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

		$\hat{R}_{Y.G,j}$	$\hat{\mathcal{G}}_j([0,1])$	$\hat{R}_{W.G,j}$	$\hat{\mathcal{G}}_j([-1,1])$
1	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]
2	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]
3	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]
4	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]
5	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]
6	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]
7	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]
8	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]
9	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]
10	$\frac{1}{100}$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]
11	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 Restrictions on Confounding
Years of Education
Regression Estimates
Bounds under Sign and Magnitude Restrictions
95% CI under Sign and Magnitude Restrictions
Bounds under Magnitude Restriction
95% CI under Magnitude Restriction
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," *The 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.

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.

Battistin, E. and A. Chesher (2009), "Treatment Effect Estimation with Covariate Measurement Error," cemmap working paper CWP25/09.

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

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, pp. 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, pp. 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. M. Stoker (2010), "Set Identification and Sensitivity Analysis with Tobin Regressors," *Quantitative Economics*, 1, 255–277.

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, pp. 855–971.

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," *The 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.

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.

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

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

Manski, C. 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. R. Johnson (1996), "The Role of Premarket Factors in Black-White Wage

Differences," *Journal of Political Economy*, 104, 869-895.

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

Ogburna, E. L. and T. J. 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.

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

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

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. R. (1972), "A Note on the Use of Proxy Variables," *Econometrica*, 40, 759-761.

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

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

Wooldridge, J. M. (2008). *Introductory Econometrics: A Modern Approach*. South-Western College Publishing, 4th edition.