Applied Causal Inference Powered by ML and AI

Victor Chernozhukov* Christian Hansen† Nathan Kallus‡
Martin Spindler§ Vasilis Syrgkanis¶

July 28, 2024

Publisher: Online
Version 0.1.1

* MIT
† Chicago Booth
‡ Cornell University
§ Hamburg University
¶ Stanford University

Unobserved Confounders, Instrumental Variables, and Proxy Controls

12

"Without Philip Wright
would there have been causal DAGs?
Who can really say?"

– Kei Hirano*

https://keihirano.github.io/haiku.html

In this chapter, we discuss various models with unobserved confounders where the adjustment strategies based on conditioning that we have discussed no longer work. We start with sensitivity analysis of causal inference to the presence of unobserved confounders. Then we discuss identification of causal effects when instrumental variables or proxy controls are available.

* Sewall Wright, son, and Philip Wright, father, were responsible for some of the greatest ideas in causal inference. Sewall Wright invented causal path diagrams (linear DAGs), and Philip Wright wrote down DAGs for supply-demand equations, proposed IV methods for their identification, and even proposed weather conditions as instruments. Just one of these contributions would probably have been enough to get a QJE publication in the 1970s and later, but it was not good enough in 1926 or so. Philip Wright is a (causal) parent of Sewall Wright, so he is one of the causes of DAGs (hence the haiku).

12.1 The Difficulty of Causal Inference with an Unobserved Confounder

> "All happy statisticians are happy in their own way; but all the unhappy ones are all alike — they all do causal inference with observational data". L. Tolstoy in Anna Karenina (Source: Twitter)

Here we consider models with unobserved confounding variables. The presence of unobserved confounding variables complicates identification of causal effects. Without further assumptions, it is impossible to identify causal effects in a setting with unobserved confounding variables.

For example, consider the following two basic models shown in Figures 12.1 and 12.2, where we can think of Y as wages, D as education, and A as latent ability.

In the first model, D has a causal effect on Y; and in the second, it does not. However, the two models in Figures 12.1 and 12.2 are statistically indistinguishable from each other if A is not observed. Even with strong restrictions, as in Gaussian linear SEMs, the observed correlation between D and Y can always be rationalized either as a causal effect of D on Y or the result of a common cause A.

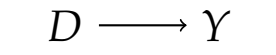

Figure 12.1: D causes Y

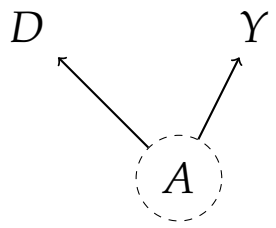

Figure 12.2: D and Y are caused by a latent factor A

The observation that Figures 12.1 and 12.2 are statistically indistinguishable applies more generally. While we cannot precisely pin down causal effects in such cases, we can still learn about causal effects by performing sensitivity analysis if we are willing to assume a bound on the strength of unobserved confounders. We discuss a practical and intuitive approach to sensitivity analysis in Section 12.2.

We may also make progress in learning causal effects in the presence of unobserved confounders by considering the use of instrumental variables (IVs) – additional random vectors Z that create exogenous variation in D – as illustrated in Figure 12.3. This approach was introduced by Philip Wright in 1928 [1]. The use of instruments renders many linear ASEMs identifiable, allowing us to perform inference on structural effects $D \to Y$. Some nonlinear ASEMs also become identifiable, though identification still fails for completely unrestricted nonlinear models. We discuss the use of instruments in Sections 12.3-12.4.

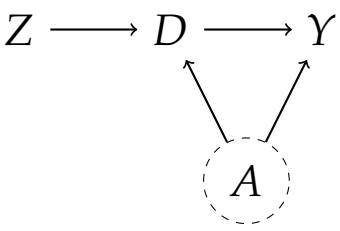

Figure 12.3: A DAG with Latent Confounder A and Instrument Z.

A related set of problems is when we observe multiple proxy measurements of the latent confounder A. For example, we may observe S, the SAT score, and Q, the ACT score, which may

both be proxies for latent confounder, A, ability as illustrated in Figure 12.4. Note that conditioning on Q and S does not block the backdoor path $Y \leftarrow A \rightarrow D$. Hence we cannot use the regression adjustment method for identification of $D \rightarrow Y$. However, this problem is related to IVs, because we can effectively use one measurement in place of A and instrument it with another measurement to deal with the measurement error. This process can provide identification of the main effect $D \rightarrow Y$. In other words, we can use instrumental variable regression of Y on D and S, using D and Q as technical instrumental variables. This approach was introduced by Zvi Griliches in 1977 [2]. This model has also been extensively studied for nonlinear models as well, e.g., Miao et al. [3] and Deaner [4], especially in the recent literature. We discuss proxy approaches in Section 12.6.

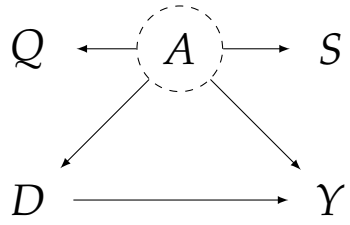

Figure 12.4: A DAG with two proxies for latent confounders.

12.2 Impact of Confounders on Causal Effect Identification and Sensitivity Analysis

Example 12.2.1 (Partially Linear SEM) Consider the SEM (illustrated in Figure 12.5)

$$\begin{aligned} Y &:= \alpha D + \delta A + f_Y(X) + \epsilon_Y, \\ D &:= \gamma A + f_D(X) + \epsilon_D, \\ A &:= f_A(X) + \epsilon_A, \\ X &:= \epsilon_X, \end{aligned}$$

where, conditional on X, $\epsilon_Y, \epsilon_D, \epsilon_A$ are mean zero and mutually uncorrelated. We further normalize

$$\mathrm{E}[\epsilon_A^2] = 1.$$

The key structural parameter is α:

$$\alpha = \partial_d Y(d)$$

where

$$Y(d) := (Y : do(D = d)).$$

To give context to our example, we can interpret Y as earnings, D as education, A as ability, and X as a set of observed background variables. In this example, we can interpret α as the returns to schooling.

We start by applying the partialling out operator to get rid of the

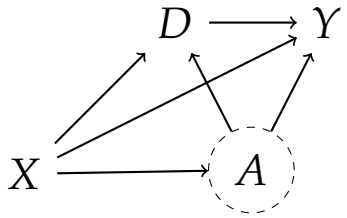

Figure 12.5: X are observed confounders, and A are unobserved confounders.

X's in all of the equations. Define the partialling out operation of any random vector V with respect to another random vector X as the residual that is left after subtracting the best predictor of V given X:

$$\tilde{V} = V - \mathrm{E}[V \mid X].$$

If f's are linear, we can replace $\mathrm{E}[V \mid X]$ by linear projection. After partialling out, we have a simplified system:

$$\begin{aligned} \tilde{Y} &:= \alpha\tilde{D} + \delta\tilde{A} + \epsilon_Y, \\ \tilde{D} &:= \gamma\tilde{A} + \epsilon_D, \\ \tilde{A} &:= \epsilon_A, \end{aligned}$$

where ϵ_Y, ϵ_D, and ϵ_A are uncorrelated.

Then the projection of $\tilde{Y}$ on $\tilde{D}$ recovers

$$\beta = \mathrm{E}[\tilde{Y}\tilde{D}]/\mathrm{E}[\tilde{D}^2] = \alpha + \phi,$$

where

$$\phi = \delta\gamma/\mathrm{E}\left[(\gamma^2 + \epsilon_D^2)\right]$$

is the omitted confounder bias.

Omitted confounder bias is also often referred to as omitted variables bias.

The formula follows from inserting the expression for $\tilde{D}$ into the definition of β and then simplifying the resulting expression using the assumptions on the ϵ's.

We can use this formula to bound ϕ directly by making assumptions on the size of δ and γ. An alternative approach can be based on the following characterization, based on partial R^2's. This characterization essentially follows from Cinelli and Hazlett [5], with the slight difference that we have adapted the result to the partially linear model.

Theorem 12.2.1 (Omitted Confounder Bias in Terms of Partial R^2's) *In the setting given in Example 12.2.1,*

$$\phi^2 = \frac{R^2_{\tilde{Y}\sim\tilde{A}|\tilde{D}} R^2_{\tilde{D}\sim\tilde{A}}}{(1 - R^2_{\tilde{D}\sim\tilde{A}})} \frac{\mathrm{E}\left[(\tilde{Y} - \beta\tilde{D})^2\right]}{\mathrm{E}\left[(\tilde{D})^2\right]},$$

where $R^2_{V\sim W|X}$ denotes the population R^2 in the linear regression of V on W, after partialling out X from V and W linearly.

Therefore, if we place bounds on how much of the variation in $\tilde{Y}$ and in $\tilde{D}$ the unobserved confounder $\tilde{A}$ is able to explain, we

can bound the omitted confounder bias by

$$\sqrt{\phi^2}.$$

Example 12.2.2 Here, we consider the empirical example from Cinelli and Hazlett [5], which is based on the original analysis from Hazlett [6]. In this example, we are interested in estimating the effect of having experienced direct war violence (`violence`) on attitudes towards peace (`peace`). To obtain our estimates, we use data from a survey on attitudes of Darfurian refugees in eastern Chad. For further details regarding the data and historical context, see the original paper [6].

As we suspect that experiencing direct war violence is not as good as randomly assigned, we control for observed demographics to hopefully captured sources of confounding. The R^2 of the regression of `peace` on the controls after partialling out `violence` is 0.13, and the R^2 of the regression of `violence` on the controls is 0.01. Based on these observed values, suppose we are willing to accept that

$$R^2_{\tilde{Y}\sim\tilde{A}|\tilde{D}} \leq 0.13, \quad R^2_{\tilde{D}\sim\tilde{A}} \leq 0.01.$$

That is, we are willing to assume that any latent confounder is no stronger than the observed controls for predicting Y and for predicting D.

We can now apply Theorem 12.2.1 to obtain a point estimate of the bias using these benchmark values for $R^2_{\tilde{Y}\sim\tilde{A}|\tilde{D}}$ and $R^2_{\tilde{D}\sim\tilde{A}}$. Filling these values in, we obtain a point estimate of the bias as

$$\hat{\phi} = \sqrt{\frac{0.13 * 0.01}{0.99}}\frac{0.0597}{0.1402} \approx 0.0236$$

where we estimate $\mathrm{E}\left[(\tilde{Y} - \beta\tilde{D})^2\right] \approx 0.0597$ using the sample average of the squared residuals from the regression of the residuals from partialling the controls out from Y onto the residuals from partialling out the controls from D and $\mathrm{E}\left[(\tilde{D})^2\right] \approx 0.1402$ as the sample average of the squared residuals from partialling out the controls from D. Finally, the point estimate of β from the data is 0.1003, so we can obtain point estimates of the upper and lower bounds on α as $\hat{\beta} \pm \hat{p}hi = 0.1003 \pm 0.0236$. That is, our point estimate for

DML Sensitivity R Notebook analyses the sensitivity of the DML estimate in the Darfur wars in R. DML Sensitivity Python Notebook does the same analysis in Python.

Benchmarking the association of unobserved confounders to that of observed confounders can be argued for on the basis that the original controls were chosen in an effort to find good variables to capture confounding, so any remaining source of confounding is likely less predictive of the outcome and variable of interest than the observed controls. While a nice story, one should be cautious of such arguments as controls are often based on convenience and what is readily available and measurable rather than careful consideration. Here, we adopt this story as a benchmark for illustrative purposes.

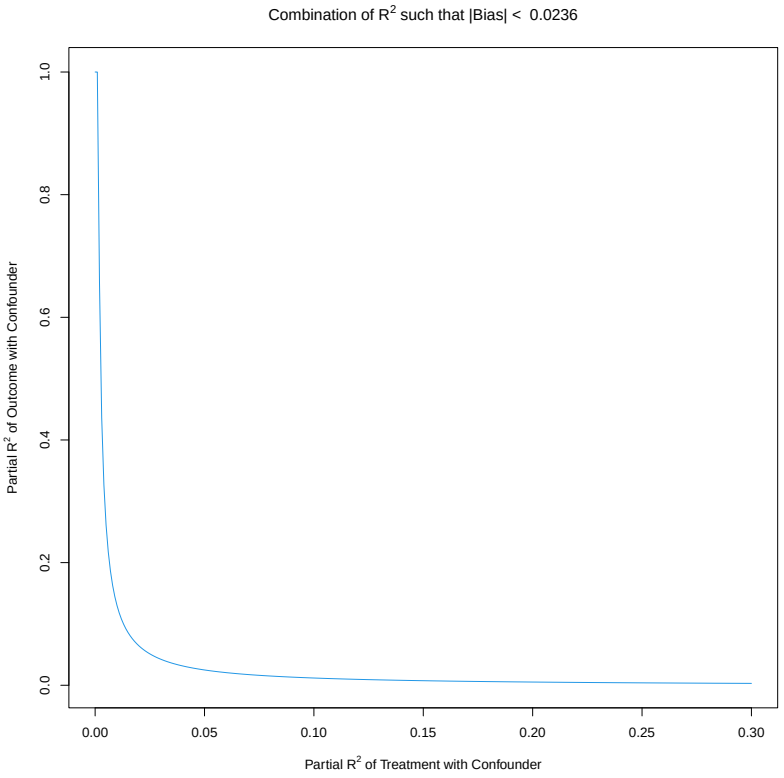

Figure 12.6: Sensitivity contour plot: The graph shows values of $R^2_{\tilde{Y}\sim\tilde{A}|\tilde{D}}$ and $R^2_{\tilde{D}\sim\tilde{A}}$ that give a given value of the bias $|\hat{\phi}| = 0.0236$. in the Darfur example. The value 0.0236 was chosen as this corresponds to the estimated bias in our benchmark scenario for $R^2_{\tilde{Y}\sim\tilde{A}|\tilde{D}}$ and $R^2_{\tilde{D}\sim\tilde{A}}$. All points below the plotted contour would correspond to estimated bias smaller than 0.0236.

the lower bound for the causal effect of `violence` on `peace` under our stated beliefs about unobserved confounding is 0.0767.

We show all combinations of $R^2_{\tilde{Y}\sim\tilde{A}|\tilde{D}}$ and $R^2_{\tilde{D}\sim\tilde{A}}$ that would lead to estimated bias of 0.0236 in Figure 12.6 and note that all combinations below the curve would lead to smaller bias estimates. We see, for example, that we need to believe there is almost no relationship between unobserved confounds and the outcome (after partialling out other variables) to keep bias smaller than 0.0236 if we believe there that $R^2_{\tilde{D}\sim\tilde{A}}$ could be 0.05 or larger.

Finally, we note that we have focused on point estimation here under benchmark beliefs but that one might be interested in other quantities. For example, one might wish to understand how large $R^2_{\tilde{Y}\sim\tilde{A}|\tilde{D}}$ and $R^2_{\tilde{D}\sim\tilde{A}}$ can be before one would draw qualitatively different conclusions from point estimates. For example, one might wish to understand when the lower bound in the Darfur example becomes 0. Alternatively, one might wish to understand sensitivity of inferential statements rather than point estimates. Such extensions are readily accommodated; see, e.g. [5].

The result of Theorem 12.2.1 extends beyond the linear setting. For example, [7] provides similar results for fully nonlinear

models. In the general PLM setting with

$$Y = \alpha D + g(X, A) + \varepsilon_Y,\ \mathrm{E}[\varepsilon_Y | D, X, A] = 0$$

and ϕ continuing to denote bias, [7] show that one can use

$$\phi^2 = S^2 C_Y^2 C_D^2$$

where

$$\begin{aligned} S^2 &= \mathrm{E}\left[(Y - E[Y|D, X])^2\right] \mathrm{E}\left[\left(\frac{D - \mathrm{E}[D|X]}{\mathrm{E}[(D - \mathrm{E}[D|X])^2]}\right)^2\right] \\ C_Y^2 &= \eta^2_{Y \sim A|D,X} \\ C_D^2 &= \frac{\eta^2_{D \sim A|X}}{1 - \eta^2_{D \sim A|X}}. \end{aligned}$$

Here,

$$\eta^2_{V \sim A|W} = \frac{\mathrm{Var}(\mathrm{E}[V|A, W]) - \mathrm{Var}(\mathrm{E}[V|W])}{\mathrm{Var}(V) - \mathrm{Var}(\mathrm{E}[V|W]}$$

denotes the nonparametric partial R^2 of variable V with variable A conditional on variable W. For example, $\eta^2_{Y \sim A|D,X}$ is the fraction of the variation in Y that is explained by A holding fixed D and W which measures the additional explanatory for Y that A adds beyond what is already explained by D and X. In this expression, S^2 can be estimated from data. Values for C_Y^2 and C_D^2 must be reasoned about but cannot be estimated from the observed data.

12.3 Partially Linear IV Models

When instrumental variables are available, it becomes possible to point identify causal effects in partially linear models and certain types of causal effects in nonlinear models. Here we begin with partially linear models.

A Wage Equation with Unobserved Ability

Example 12.3.1 (Returns to Education with Omitted Ability;

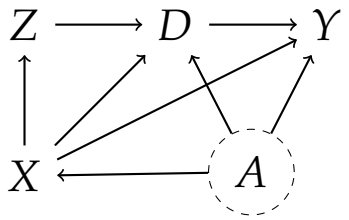

Figure 12.7: An IV model with observed and unobserved confounders.

Generalization of Griliches, 1977 [2]) Consider the ASEM

$$\begin{aligned} Y &:= \alpha D + \delta A + f_Y(X) + \epsilon_Y, \\ D &:= \beta Z + \gamma A + f_D(X) + \epsilon_D, \\ Z &:= f_Z(X) + \epsilon_Z, \\ A &:= f_A(X) + \epsilon_A, \\ X &:= \epsilon_X, \end{aligned}$$

where, conditional on X, $\epsilon_Y, \epsilon_D, \epsilon_Z, \epsilon_A$ have mean zero and are mutually uncorrelated.

We can interpret Y as earnings, D as education, A as ability, Z as an observed shifter of education, and X as a set of observed background variables. The key structural parameter is α, the returns to schooling, i.e.

$$\alpha = \partial_d Y(d),$$

where

$$Y(d) = Y : do(D = d).$$

Examples of instruments for schooling, Z, that have appeared in the literature include

- distance to college (Card [8]),
- compulsory schooling laws (Angrist [9]),
- offer to participate/offer to treat in a training program (Bloom et al. [10]), and
- local earnings and unemployment at age 17 (Cameron and Heckman [11]).

We apply the partialling-out operator to get rid of the X's in all of the equations. As before, we define the partialling out operation of any random vector V with respect to another random vector X as the residual that is left after subtracting the best predictor of V given X:

$$\tilde{V} = V - \mathrm{E}[V \mid X].$$

If f's are linear, we replace $\mathrm{E}[V \mid X]$ with linear projection.

After partialling-out, we have a simplified system.

$$\begin{aligned} \tilde{Y} &:= \alpha \tilde{D} + \delta \tilde{A} + \epsilon_Y, \\ \tilde{D} &:= \beta \tilde{Z} + \gamma \tilde{A} + \epsilon_D, \\ \tilde{Z} &:= \epsilon_Z, \\ \tilde{A} &:= \epsilon_A, \end{aligned}$$

where ϵ_Y, ϵ_D, ϵ_Z, and ϵ_A are uncorrelated.

We immediately obtain the following result:

Theorem 12.3.1 *In Example 12.3.1, we can rewrite an econometric measurement model for identification of α:*

$$\tilde{Y} \;:=\; \alpha\tilde{D} + U, \quad U \perp \tilde{Z},$$

where $U = \delta\tilde{A} + \epsilon_Y$. Alternatively, we can equivalently identify α using the moment restriction

$$\mathrm{E}\left[(\tilde{Y} - \alpha\tilde{D})\tilde{Z}\right] = 0.$$

The identification of α follows from solving this equation,

$$\alpha = \mathrm{E}[\tilde{Y}\tilde{Z}]/\mathrm{E}[\tilde{D}\tilde{Z}],$$

provided the instruments are relevant: $\mathrm{E}[\tilde{D}\tilde{Z}] \neq 0$ *or* $\beta \neq 0$.

Remark 12.3.1 (Neyman Orthgonality and DML) The target parameter α is Neyman orthogonal with respect to nuisance parameters – the regression functions $\mathrm{E}[Y \mid X]$, $\mathrm{E}[D \mid X]$, and $\mathrm{E}[Z \mid X]$. Therefore we can use DML for learning and performing statistical inference on the parameter α.

Wright's Causal Path Derivation

Starting from the DAG given in Figure 12.7, we obtain Figure 12.8 after partialling out.

Philip Wright (1928) [1] observed that the structural parameter $\beta\alpha$, the effect $\tilde{Z} \to \tilde{Y}$, is identified from the projection of $\tilde{Y} \sim \tilde{Z}$:

$$\beta\alpha = \mathrm{E}[\tilde{Y}\tilde{Z}]/\mathrm{E}[\tilde{Z}^2].$$

The structural parameter β, the effect of $Z \to D$, is identified from the projection of $\tilde{D} \sim \tilde{Z}$:

$$\beta = \mathrm{E}[\tilde{D}\tilde{Z}]/\mathrm{E}[\tilde{Z}^2].$$

α, the effect of $D \to Y$, is then identified by the ratio of the two provided $\beta \neq 0$:

$$\alpha = \frac{\beta\alpha}{\beta} = \mathrm{E}[\tilde{Y}\tilde{Z}]/\mathrm{E}[\tilde{D}\tilde{Z}].$$

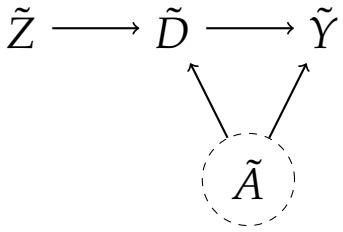

Figure 12.8: DAG corresponding to Figure 12.7 after partialling out observed confounder X.

We provide a thorough discussion of using DML to estimate parameters within the instrumental variables framework along with example applications in Chapter 13.

Aggregate Market Demand

Let's apply our approach to a canonical example in economics: the identification of the price elasticity of demand using a supply shifter as an instrument.

Example 12.3.2 (Market Demand; Generalization of P. Wright, 1928 [1]) Consider the ASEM

$$\begin{aligned} Y &:= \alpha D + f_Y(X) + \epsilon^d, \\ D &:= \beta Z + f_D(X) + \rho\epsilon^d + \gamma\epsilon^s, \\ Z &:= f_Z(X) + \epsilon_Z \end{aligned}$$

where ϵ^d, ϵ^s and ϵ_Z are mean zero and uncorrelated conditional on X. In this example, Y is (log) demand, D is (log) price, Z is an observed supply shifter, X is a vector of observed demand shifters, ϵ^d is a demand shock, and ϵ^s is a supply shock. The key parameter is α, the price elasticity of demand:

$$\alpha = \partial_d Y(d),$$

where $Y(d) := (Y : do(D = d))$. Here we focus on only the demand side of the market and do not attempt to explicitly model the supply side.

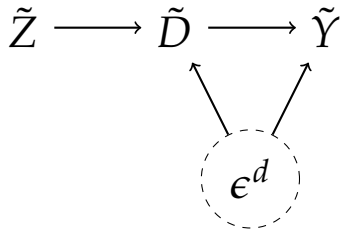

Figure 12.9: A DAG for aggregate demand, with the latent node ϵ^d representing the demand shock

In econometrics, the set-up here is sometimes referred to as a *limited information* model or formulation because we are focusing on identifying only a single equation in a more complicated underlying system.

Example 12.3.2 is equivalent to the previous Example 12.3.1 – set $A = \epsilon^d$, $\epsilon_Y = 0$, $\epsilon^s = \epsilon_D$, and so on. Hence, the identification method is the same as before.

Limits of Average Causal Effect Identification under Partial Linearity

The result in Theorem 12.3.1 extends beyond the partially linear setting presented in Example 12.3.1 to the following non-linear structural equation model:

Example 12.3.3 (Partially Linear Outcome IV Model) Consider the ASEM

$$\begin{aligned} Y &:= g_Y(\epsilon_Y)D + f_Y(A, X, \epsilon_Y), \\ D &:= f_D(Z, X, A, \epsilon_D), \\ Z &:= f_Z(X, \epsilon_Z), \\ A &:= f_A(X, \epsilon_A), \\ X &:= \epsilon_X, \end{aligned}$$

where $\epsilon_Y, \epsilon_D, \epsilon_Z, \epsilon_A$ are exogenous and mutually independent. The key structural parameter is:

$$\alpha := \mathrm{E}[\partial_d Y(d)] = \mathrm{E}[g_Y(\epsilon_Y)],$$

where

$$Y(d) = Y : do(D = d).$$

This parameter is typically referred to as the average marginal effect of the treatment.

Theorem 12.3.1 extends almost as is to this more general non-linear structural equation model.

Theorem 12.3.2 *In Example 12.3.3, we can identify α using the moment restriction*

$$\mathrm{E}\left[(\tilde{Y} - \alpha\tilde{D})\tilde{Z}\right] = 0.$$

The identification of α follows from solving this equation,

$$\alpha = \mathrm{E}\left[\tilde{Y}\tilde{Z}\right] / \mathrm{E}\left[\tilde{D}\tilde{Z}\right],$$

provided the instruments are relevant: $\mathrm{E}[\tilde{D}\tilde{Z}] \neq 0$.

Note that the non-linear structural equation model in Example 12.3.3 imposes extra assumptions on the structural response function of the outcome Y. Thus our identification argument imposes more conditions on the structural equations than the ones that can be encoded via a DAG. Such auxiliary assumptions are required for identification of average treatment effects with instruments.

In particular, the identification argument relies on the fact that the unobserved confounder A enters in an additively separable manner in the outcome equation. If for instance, A was an input to the function g, i.e. $Y := g_Y(A, \epsilon_Y)D + f_Y(A, X, \epsilon_Y)$, then the quantity identified by the moment restriction in Theorem 12.3.2 would not correspond to an average treatment effect. In this case,

the unobserved confounder creates heterogeneity in the treatment effect and also heterogeneity in the effect of the instrument on the treatment, typically referred to as the "compliance" (i.e., the correlation between Z and D varies with A). This property is what renders the ratio quantity $\alpha = \mathrm{E}\left[\tilde{Y}\tilde{Z}\right]/\mathrm{E}\left[\tilde{D}\tilde{Z}\right]$ invalid for the causal estimand of interest.

In fact, it is the joint heterogeneity in both the outcome relationship and the compliance relationship that causes the problem. We show next that we could allow for a much more complex outcome model as long as the effect of the instrument on the treatment (compliance) is not heterogeneous in A or X.

Example 12.3.4 (Partially Linear Compliance IV Model) Consider the ASEM

$$\begin{aligned} Y &:= g_Y(A, X, \epsilon_Y)D + f_Y(A, X, \epsilon_Y), \\ D &:= g_D(\epsilon_D)Z + f_D(X, A, \epsilon_D), \\ Z &:= f_Z(X) + \epsilon_Z, \\ A &:= f_A(X, \epsilon_A), \\ X &:= \epsilon_X, \end{aligned}$$

where, $\epsilon_Y, \epsilon_D, \epsilon_Z, \epsilon_A$ are exogenous and mutually independent. The key structural parameter is:

$$\alpha := \mathrm{E}[\partial_d Y(d)] = \mathrm{E}[g(A, X, \epsilon_Y)],$$

where

$$Y(d) = Y : do(D = d).$$

Theorem 12.3.3 *In Example 12.3.4, we can identify α using the moment restriction*

$$\mathrm{E}\left[(\tilde{Y} - \alpha\tilde{D})\tilde{Z}\right] = 0.$$

The identification of α follows from solving this equation,

$$\alpha = \mathrm{E}\left[\tilde{Y}\tilde{Z}\right]/\mathrm{E}\left[\tilde{D}\tilde{Z}\right],$$

provided the instruments are relevant: $\mathrm{E}[\tilde{D}\tilde{Z}] \neq 0$.

Thus, we see that we need that either the effect of education on wages is not heterogeneous in the unobserved ability variable A or that the effect of the observed education shifter Z (e.g. distance to college) on education D is not heterogeneous in the unobserved ability variable to use the identification strategies presented in this section in the context of our education exam-

ple. In Section 12.4, we will investigate what causal quantities are identifiable even in non-linear structural equation models, where the unobserved confounder creates heterogeneity in both the treatment effect and in the compliance behavior.

Remark 12.3.2 (Effect heterogeneity based on observables) We note that allowing for X to enter the g_Y or g_D function in Example 12.3.3 and Example 12.3.4 (i.e. allowing for the treatment effect or compliance – the effect of the instrument on treatment – to vary with X) is a more benign extension because X is an observed variable. In this case, we can repeat the identification strategies in this section, conditional on X, and we can show with similar arguments that

$$\beta(X) := \mathrm{E}[\partial_d Y(d) \mid X] = \frac{\mathrm{E}[\tilde{Y}\tilde{Z} \mid X]}{\mathrm{E}[\tilde{D}\tilde{Z} \mid X]}. \tag{12.3.1}$$

We can simply average these conditional effects to get the average marginal effect:

$$\alpha = \mathrm{E}[\beta(X)]. \tag{12.3.2}$$

Such an identification strategy was initiated in [12, 13] and was also recently used in the context of DML estimators [14–16]. In particular, the following moment condition that identifies α,

$$\mathrm{E}\left[\beta(X) + \frac{(\tilde{Y} - \beta(X)\tilde{D})\tilde{Z}}{\mathrm{E}[\tilde{D}\tilde{Z} \mid X]} - \alpha\right] = 0, \tag{12.3.3}$$

is Neyman orthogonal with respect to the nuisance functions $\beta(X)$ and $\gamma(X) := \mathrm{E}[\tilde{D}\tilde{Z} \mid X]$. We note that this identification strategy remains valid even if in Example 12.3.4 the instrument equation is fully non-linear, i.e. $Z := f_Z(X, \epsilon_Z)$.

12.4 Nonlinear IV Models

Once we consider nonlinear models, identification becomes a much more delicate matter. We first consider the local average treatment effect (LATE) model, and then we turn to quantile models.

The LATE Model

An important nonlinear IV model in the case of a binary treatment variable and a binary instrumental variable is the local average treatment effect model (LATE) proposed by Imbens and Angrist [17].

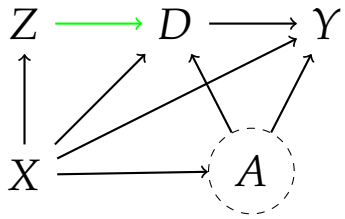

Figure 12.10: LATE models. Green arrow denotes a monotone functional relation.

Example 12.4.1 (LATE) Consider the SEM where

$$\begin{aligned} Y &:= f_Y(D, X, A, \epsilon_Y) \\ D &:= f_D(Z, X, A, \epsilon_D) \in \{0, 1\}, \\ Z &:= f_Z(X, \epsilon_Z) \in \{0, 1\}, \\ X &:= \epsilon_X, \quad A = \epsilon_A, \end{aligned}$$

where the ϵ's are mutually independent, and

$$z \mapsto f_D(z, A, X, \epsilon_D)$$

is weakly increasing (weakly monotone).

Suppose the instrument Z is an offer to participate in a training program and that D is the actual endogenous participation in the training program. Participation in the program may depend on unobservables A, such as ability or perseverence, that also affect the eventual outcome Y. We can also have background exogenous covariates X in the model.

Define

$$Y(d) := f_Y(d, X, A, \epsilon_Y) \text{ and } D(z) := f_D(z, X, A, \epsilon_D)$$

as the potential outcomes that result from applying fix-interventions in the corresponding equations from Example 12.4.1.

The model allows us to identify the local average treatment effect (LATE), defined as

$$\theta = \mathrm{E}[Y(1) - Y(0) \mid D(1) > D(0)].$$

Here, $\{D(1) > D(0)\}$ is the *compliance event* which corresponds to the case where exogenously switching the instrument value from $Z = 0$ to $Z = 1$ induces a switch from the control state to the treatment state. Therefore, the LATE measures the average treatment effect conditional on compliance.

Theorem 12.4.1 *In the LATE model, we have that θ is identified by the ratio of two statistical parameters,*

$$\theta = \theta_1/\theta_2,$$

where

$$\theta_1 := \mathrm{E}\left[\mathrm{E}[Y \mid X, Z = 1] - \mathrm{E}[Y \mid X, Z = 0]\right],$$

and

$$\theta_2 := \mathrm{E}\left[\mathrm{E}[D \mid X, Z = 1] - \mathrm{E}[D \mid X, Z = 0]\right],$$

provided that the instrument Z is relevant, $\theta_2 > 0$, and Z has full conditional support – namely $0 < P(Z = 1 \mid X) < 1$. Moreover, θ_2 identifies the probability of compliance:

$$\theta_2 = \mathrm{P}[D(1) > D(0)].$$

The result has an intuitive interpretation.[1] In the event of compliance, the instrument moves the treatment as if experimentally, which induces quasi-experimental variation in the outcome. We measure the probability of compliance with θ_2 and the average induced changes in outcome by θ_1. Taking the ratio is then like conditioning on the compliance event. See the proof in Section 12.A for details.

1: In the model with no X the ratio θ_1/θ_2 is equivalent to Wright's [1] IV estimand.

The ratio can be recognized as the ratio of average treatment effects of Z on Y and D,

$$\theta_1 = ATE(Z \to Y),$$

$$\theta_2 = ATE(Z \to D).$$

This assertion follows from the application of the backdoor criterion. Therefore, we can simply re-use the tools for performing inference on the two ATEs to perform inference on the LATE.

Remark 12.4.1 (DML for θ_1/θ_2) We can apply DML to obtain $\hat{\theta}_1$ and $\hat{\theta}_2$ and then construct the estimator $\hat{\theta} = \hat{\theta}_1/\hat{\theta}_2$ via the plug-in principle. This approach has the Neyman orthogonality property.

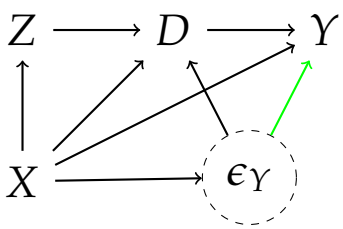

Figure 12.11: IV Quantile Model. The green arrow represents a strictly monotonic effect.

The IV Quantile Model*

Another nonlinear IV model is the following model that exploits monotonicity in the unobservable shock in the outcome equation to obtain identification.

Example 12.4.2 (IV Quantile Model) Consider the SEM

$$\begin{aligned} Y &= f_Y(D, X, \epsilon_Y), \\ D &= f_D(Z, X, \epsilon_Y, \epsilon_D), \\ Z &= f_Z(X, \epsilon_Z), \\ X &= \epsilon_X, \end{aligned}$$

where the ϵ's are mutually independent,

$$f_Y(D, X, \cdot) : [0, 1] \mapsto \mathbb{R} \text{ is strictly increasing,}$$

and ϵ_Y is normalized to have uniform distribution on $(0, 1)$.

As a concrete example, suppose we are interested in estimating demand. In this setting, Y would denote quantity sold of some product, D would denote the product's price, ϵ_Y would denote a demand shock, and ϵ_D a supply shock. X would then denote a set of observed variables, such as product characteristics, that potentially relate to both price and quantity sold, and Z would be a set of instrumental variables. The function $f_Y(d, x, u)$ is then the u-th quantile of the structural function of $f_Y(d, x, \epsilon_Y)$, which gives us the demand with price fixed to d and characteristics fixed to x. For example, $f_Y(d, x, 1/2)$ would denote the median quantity demanded at fixed price d and characteristics x.

The testable implication of the IV Quantile Model is the following.

Theorem 12.4.2 *In the IV Quantile Model, the testable moment restriction is*

$$\mathrm{P}[Y \leq f_Y(D, X, u) \mid Z, X] = u,$$

for each $u \in (0, 1)$. There exist regularity conditions, analogous to instrument relevance, under which the structural function f_Y is identified from this restriction.

In practice, linear forms $f_Y(D, X, u) = \alpha(u)'D + \beta(u)'X$ are often used. Adopting a linear functional form leads to method

of moments approaches such as the IV quantile regression for performing inference on structural quantile functions.

Code for IV Quantile Models can be found here.

Remark 12.4.2 (DML for IVQR Models) Neyman orthogonal approaches for partially linear IVQR models are sketched out in the review [18]. Exploring DML in more general settings may be an interesting area for further work.

12.5 Partially Linear SEMs with Griliches-Chamberlain Proxy Controls

Suppose we are interested in the causal effect of college education on earnings in the presence of an unobserved confounder – individual ability. Here we show that we can recover the effect of college education on earnings in the presence of latent ability using proxies for ability, but not the effect of ability itself.

Example 12.5.1 (Earnings with Omitted Ability; Griliches, 1977 [2]; Griliches and Chamberlain, 1977 [19]) Consider the ASEM

$$\begin{aligned} Y &:= \alpha D + \delta A + \iota S + f_Y(X) + \epsilon_Y, \\ D &:= \gamma A + \beta Q + f_D(X) + \epsilon_D, \\ Q &:= \eta A + f_Q(X) + \epsilon_Q, \\ S &:= \phi A + f_S(X) + \epsilon_S, \\ A &:= f_A(X) + \epsilon_A, \\ X &:= \epsilon_X, \end{aligned}$$

where $\epsilon_Y, \epsilon_D, \epsilon_Q, \epsilon_S, \epsilon_A, \epsilon_X$ have mean zero and are uncorrelated conditional on X.

As an example, one might interpret Y as earnings, D as college degree, A as ability, Q and S as proxies of ability, and X as a set of observed background variables. Example proxies Q and S are

- might be test scores or grades in some period t_0 (Q) and test scores or grades at a later period t_1 (S).

The key structural parameter is α, the returns to schooling; i.e.

$$\alpha = \partial_d Y(d),$$

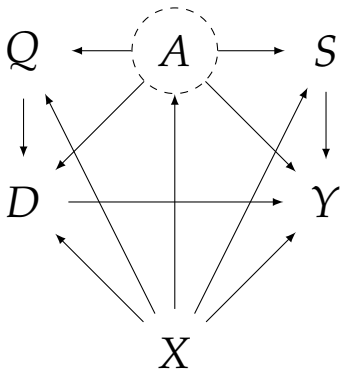

Figure 12.12: A DAG with Controls and Proxy Controls

where $Y(d) = Y : do(D = d)$.

After partialling out we are left with the DAG in Figure 12.13:

$$\begin{aligned}
\tilde{Y} &:= \alpha\tilde{D} + \delta\tilde{A} + \iota\tilde{S} + \epsilon_Y, \\
\tilde{D} &:= \gamma\tilde{A} + \beta\tilde{Q} + \epsilon_D, \\
\tilde{Q} &:= \eta\tilde{A} + \epsilon_Q, \\
\tilde{S} &:= \phi\tilde{A} + \epsilon_S, \\
\tilde{A} &:= \epsilon_A,
\end{aligned}$$

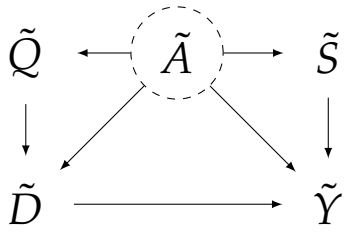

Figure 12.13: A DAG with Proxy Controls After Partialling Out

where $\epsilon_Y, \epsilon_D, \epsilon_Q, \epsilon_S, \epsilon_A$ are uncorrelated. The idea now is to replace $\tilde{A}$ in the equation for $\tilde{Y}$ with $\tilde{S}$. Note that because S enters the Y equation directly, we cannot consider using $\tilde{Q}$ to proxy for $\tilde{A}$. We still cannot learn α from the regression of $\tilde{Y}$ on $\tilde{D}$ and $\tilde{S}$ though as S is an imperfect proxy for A. The following result, which provides an IV approach to identify α, is immediate via substitution.[2]

Theorem 12.5.1 *Assume that all variables in Example 12.5.1 are square-integrable. Then we have the following measurement equation:*

$$\tilde{Y} = \alpha\tilde{D} + \bar{\delta}\tilde{S} + U, \quad \mathrm{E}[U(\tilde{D}, \tilde{Q})] = 0,$$

$$U = -\delta\epsilon_S/\phi + \epsilon_Y; \quad \bar{\delta} = \iota + \delta/\phi.$$

Here α is identified from the moment condition $\mathrm{E}[U(\tilde{D}, \tilde{Q})] = 0$, which is equivalent to using $\tilde{Q}$ as an instrument for $\tilde{S}$, provided that $\tilde{D}$ and the best linear predictor of $\tilde{S}$ using $\tilde{Q}$ and $\tilde{D}$ have non-degenerate covariance matrix.

2: Prove the result as a reading exercise. Substitute $\tilde{A} = (\tilde{S} - \epsilon_S)/\phi$ in the first equation and use the assumptions on the disturbances.

Note that $\tilde{Q}$ here plays the role of a *technical* instrument for $\tilde{S}$. This approach recovers α, but not δ. For inference, we can employ the DML method for IV models; see Chapter 13.

Remark 12.5.1 (Neyman Orthogonality and DML) The formulation of the target parameter given above is Neyman orthogonal, and high-quality estimation and statistical inference can be carried out using DML. In essence, we just residualize the system, using cross-fitted residuals, and then apply standard instrumental variable methods from econometrics to perform inference on the structural parameter of interest.

Example 12.5.2 Here, we consider a stylized empirical exam-

DML for Linear Proxy Controls R Notebook and DML for Linear Proxy Controls Python Notebook provide code for the birth weight proxy controls example using data downloaded from Stat Labs.

ple where we wish to estimate the causal effect of a mother's smoking on infant birth weight. We posit a partially linear structural system exactly as in Example 12.5.1 where the outcome of interest Y is infant birth weight in ounces, and the treatment D is a categorical representation of number of cigarettes smoked per day. We abstract from issues related to the categorical nature of D and simply include it in its raw form. In our stylized example, we pretend we do not observe income and treat it as the unobserved confounder, A.

We then divide the other observed variables into

- proxy treatment control Q: mother's education
- proxy outcome control S: parity (total number of previous pregnancies)
- other observed covariates X: mother's race and age.

We might believe that education serves as a proxy treatment control Q because it reflects unobserved confounding due to household income A but has no direct medical effect on birth weight. Parity may serve as a proxy outcome control because family size may reflect household income A but is plausibly not directly caused by smoking D or education Q.

Assuming this structure, it is then easy to estimate the coefficient on D within the partially linear structural model. We obtain an estimate of -1.68 with standard error of 0.026. That is, *within the context of the posited model*, there is substantial evidence that smoking causally leads to lower infant birth weights.

The posited model is very stylized, but illustrates the main ideas and thought process.

12.6 Nonlinear Models with Proxy Controls⋆

A relatively recent literature considers *proximal causal inference*, which generalizes early work by Griliches and Chamberlain [19]. See, among others, [20], [21], [22], [23], [3]. Here we describe some results specialized to the discrete case.

Example 12.6.1 (Miao, Geng, and Tchetgen Tchetgen [3]) We consider the following model encoded in the DAG in Figure

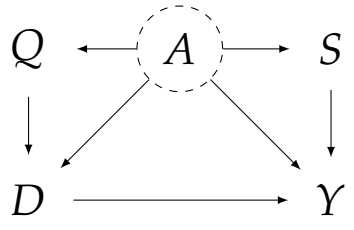

Figure 12.14: A SEM with Proxy Controls Q and S. Note that conditioning on Q and S does not block the backdoor path $Y \leftarrow A \rightarrow D$, hence we cannot use the regression adjustment method for identification of $D \rightarrow Y$.

12.14:

$$\begin{aligned} Y &:= f_Y(D, S, A, \epsilon_Y), \\ D &:= f_D(A, Q, \epsilon_D), \\ Q &:= f_Q(A, \epsilon_Q), \\ S &:= f_S(A, \epsilon_S), \\ A &:= \epsilon_A, \end{aligned}$$

where ϵ's are mutually independent. We can endow the same context to this model as in Example 12.5.1.

Here we can introduce background exogenous controls X in each of the equations, but we don't do so to save notation. Notice that the model in Example 12.6.1 generalizes the Example 12.5.1 to the nonparametric case.

Assumption 12.6.1 *In Example 12.6.1, assume*

(a) Variables Q, S and A are finitely discrete and take on the same number of values.
(b) The matrix $\Pi(S \mid Q, d)$, whose s^{th} row and q^{th} column is $p(s \mid q, d)$, is invertible for each value d.

Condition (b) is analogous to the usual relevance condition in IV and basically says that the two proxies S and Q have sufficient joint variation at any value of d to allow Q to serve as an "instrument" for S. The discreteness assumption can be generalized to a more general completeness condition; see, e.g., Miao et al. [3] and Deaner [4]. As with the usual IV relevance condition, Condition (b) is testable from the data. In contrast, the DAG itself and the other conditions involve an unobserved variable A and are therefore generally untestable. The validity of these untestable conditions must be assessed using contextual knowledge about the empirical problem.

Theorem 12.6.1 *Under Assumption 12.6.1, $p(y : do(d))$ is identifiable by the proximal formula:*

$$p(y : do(d)) = \Pi(y \mid d, Q)\, \Pi(S \mid Q, d)^{-1}\, \Pi(S)\,, \qquad (12.6.1)$$

where $\Pi(y \mid d, Q)$ and $\Pi(S)$ are row and column vectors whose entries are of the form $p(y \mid d, q)$ and $p(s)$.

The mnemonic way to think about the formula above is that we are doing a kind of instrumental variable regression of Y on S,

while instrumenting S with Q, which is exactly how we dealt with the linear version of this problem in Section 12.6.

Remark 12.6.1 [23] and [22] provide moment functions defined in terms of efficient influence functions, which possess the Neyman orthogonality property, for estimating the average treatment effect within this proxy control setting in the presence of a high-dimensional set of control variables. These moment functions can thus serve as the foundation for the use of DML inference methods for the average treatment effect in such settings.

Notebooks

- DML Sensitivity R Notebook analyses the sensitivity of the DML estimate in the Darfur wars example to unobserved confounders using the Sensemakr package in R. DML Sensitivity Python Notebook does the same analysis in Python.

- DML for Linear Proxy Controls R Notebook and DML for Linear Proxy Controls Python Notebook provide an application of using linear instrumental variables estimation withing the proxy controls framework to estimate the effect of smoking on birth weight.

Study Problems

1. Explain omitted confounder bias to a fellow student (one paragraph). Explore using sensitivity analysis to aid in understanding robustness of economic conclusions to the presence of unobserved confounders in an empirical example of your choice. The notebooks DML Sensitivity R Notebook and DML Sensitivity Python Notebook can be a helpful starting point, but be sure to apply the ideas to a different empirical example. (You could use any of the previous examples we have analyzed).

2. Write a brief explanation of the idea of the instrumental variables regression model that would be appropriate for educating a fellow student. Discuss the idea of identifying the causal effect in this setting via path analysis in the spirit of what Philip Wright did. Illustrate your discussion within a concrete empirical setting.

3. (Simulation.) Create a notebook to simulate one of the linear IV or proxy controls models that we've described. Assume there are no X's for simplicity. Demonstrate numerically why using least squares may not be appropriate due to unobserved confounding. Demonstrate numerically how using instrumental variable regression overcomes the issue.

4. (LATE etc.) Write a verbal explanation of one of the nonlinear models (e.g. LATE, IV quantile model, or the nonlinear model with proxy controls) that would be understandable by a fellow student. Be sure that the explanation includes an intuitive discussion of how causal parameters in these models are identified.

12.A Proofs

Latent Confounder Bias Result: Theorem 12.2.1

The proof heavily relies on the Frisch-Waugh-Lovel partialling out theorem (FWL) and the normalization on the variance of the latent confounder:

$$\mathrm{E}[\tilde{A}^2] = 1. \tag{12.A.1}$$

The proof also relies on the properties of $R^2_{U\sim V}$ which measures the proportion of variance of centered random variable U that is linearly explained by another centered random variable V:

$$R^2_{U\sim V} = \frac{E[\beta^2 V^2]}{\mathrm{E}[U^2]} = 1 - \frac{\mathrm{E}[\epsilon^2]}{\mathrm{E}[U^2]} = \frac{(\mathrm{E}[UV])^2}{\mathrm{E}[U^2]\mathrm{E}[V^2]} = [\mathrm{Cor}(U,V)]^2,$$

where $\beta = \mathrm{E}[VU]/\mathrm{E}[V^2]$ is the coefficient of the best linear projection of U onto V, $\epsilon = U - \beta V$ is the projection residual, and $\mathrm{Cor}(U,V)$ denotes the correlation between U and V. Note that R^2 is symmetric in U and V: $R^2_{U\sim V} = R^2_{V\sim U}$.

By FWL and the normalization (12.A.1), we have

$$\gamma = \mathrm{E}[\tilde{A}\tilde{D}], \quad \delta = \mathrm{E}[\bar{A}\bar{Y}]/\mathrm{E}[\bar{A}^2],$$

where

$$\bar{Y} = \tilde{Y} - \beta\tilde{D}; \quad \beta = \mathrm{E}[\tilde{Y}\tilde{D}]/\mathrm{E}[\tilde{D}^2];$$
$$\bar{A} = \tilde{A} - \tilde{\beta}\tilde{D}; \quad \tilde{\beta} = \mathrm{E}[\tilde{A}\tilde{D}]/\mathrm{E}[\tilde{D}^2].$$

It follows that

$$\phi^2 = \frac{\gamma^2\delta^2}{(\mathrm{E}[\tilde{D}^2])^2} = \frac{(\mathrm{E}[\tilde{A}\tilde{D}])^2}{(\mathrm{E}[\tilde{D}^2])^2}\frac{(\mathrm{E}[\bar{Y}\bar{A}])^2}{(\mathrm{E}[\bar{A}^2])^2}.$$

Then the result follows from the normalization (12.A.1);the relations

$$(\mathrm{E}[\tilde{D}\tilde{A}])^2 = [\mathrm{Cor}(\tilde{D},\tilde{A})]^2\mathrm{E}[\tilde{D}^2] = R^2_{\tilde{D}\sim\tilde{A}}\mathrm{E}[\tilde{D}^2],$$

$$(\mathrm{E}[\bar{Y}\bar{A}])^2 = [\mathrm{Cor}(\bar{Y},\bar{A})]^2\mathrm{E}[\bar{Y}^2]\mathrm{E}[\bar{A}^2] = R^2_{\bar{Y}\sim\tilde{A}}\mathrm{E}[\bar{Y}^2]\mathrm{E}[\bar{A}^2],$$

$$\mathrm{E}[\bar{A}^2] = 1 - R^2_{\tilde{A}\sim\tilde{D}} = 1 - R^2_{\tilde{D}\sim\tilde{A}};$$

and by noting that by definition $R^2_{\tilde{Y}\sim\bar{A}} = R^2_{\tilde{Y}\sim\tilde{A}|\tilde{D}}$.

Partially Linear Outcome IV Model: Theorem 12.3.2

First note that since $\mathrm{E}[\tilde{Z} \mid X] = 0$, we can re-write the moment condition as

$$\mathrm{E}[(Y - \alpha D)\tilde{Z}] = 0.$$

We can use the structural equation for Y to replace Y in the moment equation:

$$\mathrm{E}[(g_Y(\epsilon_Y)D + f_Y(A, X, \epsilon_Y) - \alpha D)\tilde{Z}] = 0.$$

Furthermore, since $\tilde{Z} \perp\!\!\!\perp A, \epsilon_Y \mid X$, we have that

$$\begin{aligned}\mathrm{E}[f_Y(A, X, \epsilon_Y)\tilde{Z}] &= \mathrm{E}[f_Y(A, X, \epsilon_Y)\mathrm{E}[\tilde{Z} \mid X, A, \epsilon_Y]] \\ &= \mathrm{E}[f_Y(A, X, \epsilon_Y)\mathrm{E}[\tilde{Z} \mid X]] = 0.\end{aligned}$$

Thus, we can re-write the moment equation as

$$\mathrm{E}[(g_Y(\epsilon_Y)D - \alpha D)\tilde{Z}] = 0.$$

Solving for α and using the fact that $\epsilon_Y \perp\!\!\!\perp \tilde{Z}$, we get

$$\alpha = \frac{\mathrm{E}[g_Y(\epsilon_Y)D\tilde{Z}]}{\mathrm{E}[D\tilde{Z}]} = \frac{\mathrm{E}[g_Y(\epsilon_Y)]\mathrm{E}[D\tilde{Z}]}{\mathrm{E}[D\tilde{Z}]} = \mathrm{E}[g_Y(\epsilon_Y)].$$

Partially Linear Compliance IV Model: Theorem 12.3.3

Using the exact same arguments as in the proof of Theorem 12.3.2, we can deduce that the solution to the moment restriction takes the form

$$\alpha = \frac{\mathrm{E}[g_Y(X, A, \epsilon_Y)D\tilde{Z}]}{\mathrm{E}[D\tilde{Z}]} = \frac{\mathrm{E}[g_Y(X, A, \epsilon_Y)\mathrm{E}[D\tilde{Z} \mid X, A, \epsilon_Y]]}{\mathrm{E}[D\tilde{Z}]}.$$

We now use the assumptions on the structural response functions of D and Z to argue that $\mathrm{E}[D\tilde{Z} \mid X, A, \epsilon_Y] = \mathrm{E}[D\tilde{Z}]$ – that is, to argue the covariance of D and Z (aka compliance) is independent of X, A, ϵ_Y. This independence then implies the theorem, since

$$\alpha = \mathrm{E}[g_Y(X, A, \epsilon_Y)].$$

First, we use the assumption on the structural response function

of D:

$$\mathrm{E}[D\tilde{Z} \mid X, A, \epsilon_Y] = \mathrm{E}[(g_D(\epsilon_D)Z + f_D(X, A, \epsilon_D))\tilde{Z} \mid X, A, \epsilon_Y].$$

Using the fact that $Z \perp\!\!\!\perp A, \epsilon_D, \epsilon_Y \mid X$ and that $\mathrm{E}[\tilde{Z} \mid X] = 0$, we can remove the term $f_D(X, A, \epsilon_D)$ from the above equation:

$$\mathrm{E}[D\tilde{Z} \mid X, A, \epsilon_Y] = \mathrm{E}[g_D(\epsilon_D)Z\tilde{Z} \mid X, A, \epsilon_Y].$$

Using the additively separable assumption on the structural response of Z and the fact that ϵ_Z is an exogenous independent variable, we have

$$\begin{aligned}\mathrm{E}[D\tilde{Z} \mid X, A, \epsilon_Y] &= \mathrm{E}[g_D(\epsilon_D)Z\epsilon_Z \mid X, A, \epsilon_Y] \\ &= \mathrm{E}[g_D(\epsilon_D)\epsilon_Z^2 \mid X, A, \epsilon_Y] = \mathrm{E}[g_D(\epsilon_D)\epsilon_Z^2]\end{aligned}$$

where we used the fact that all noise variables $\epsilon_D, \epsilon_Y, \epsilon_Z$ are exogenous and mutually independent.

Linear Proxy Model: Theorem 12.5.1.

We substitute $\tilde{A} = (\tilde{S} - \epsilon_S)/\phi$ in the equation $\tilde{Y} = \alpha\tilde{D} + \delta\tilde{A} + \iota\tilde{S} + \epsilon_Y$ to obtain

$$\tilde{Y} = \alpha\tilde{D} + \bar{\delta}\tilde{S} + U$$

where

$$U = -\delta\epsilon_S/\phi + \epsilon_Y \text{ and } \bar{\delta} = \iota + \delta/\phi.$$

To verify

$$\mathrm{E}[U] = 0,$$

we observe using repeated substitutions that

- $\tilde{D}$ is a linear combination of $(\epsilon_A, \epsilon_Q, \epsilon_D)$,
- $\tilde{Q}$ is a linear combination of ϵ_A and ϵ_Q.
- U is a linear combination of (ϵ_S, ϵ_Y).

The conclusion follows from the assumption that

$$(\epsilon_A, \epsilon_Q, \epsilon_D, \epsilon_S, \epsilon_Y)$$

are all uncorrelated. The conclusion that α is identified provided that $\tilde{D}$ and the best linear predictor of $\tilde{S}$ using $\tilde{Q}$ and $\tilde{D}$ have non-degenerate covariance matrices is left as an exercise.

The LATE Result: Theorem 12.4.1

We can use, for example, the backdoor criterion to conclude that

$$\mathrm{E}[\mathrm{E}[D \mid Z = z, X]] = \mathrm{E}[\mathrm{E}[D(z) \mid X]] = \mathrm{E}[D(z)].$$

Similarly,

$$\mathrm{E}[\mathrm{E}[Y \mid Z = z, X]] = \mathrm{E}[\mathrm{E}[Y(D(z)) \mid X]] = \mathrm{E}[Y(D(z))].$$

Furthermore, by monotonicity, we have both

$$\theta_2 = \mathrm{E}[D(1) - D(0)] = \mathrm{P}(D(1) > D(0))$$

and

$$\begin{aligned}\theta_1 &= \mathrm{E}[Y(D(1)) - Y(D(0))] \\ &= \mathrm{E}[(Y(1) - Y(0))1\{D(1) > D(0)\}].\end{aligned}$$

Therefore

$$\theta_1/\theta_2 = \mathrm{E}[Y(1) - Y(0) \mid D(1) > D(0)].$$

Testable Restriction for the IV Quantile Model: Theorem 12.4.2

The result is immediate from (i) the equivalence of the event $Y \leq f_Y(D, X, u)$ and the event $\epsilon_Y \leq u$, which holds under the strict monotoniticity assumption, and (ii) the independence of ϵ_Y from Z and X which follows from the stated independence conditions. Using (i) and (ii), we have

$$\begin{aligned}\mathrm{P}[Y \leq f_Y(D, X, u) \mid Z, X] &= \mathrm{P}[\epsilon_Y \leq u \mid Z, X] \\ &= \mathrm{P}[\epsilon_Y \leq u] = \mathrm{P}[U(0,1) \leq u] = u.\end{aligned}$$

Identification in the Nonlinear Proxy Variables Model: Theorem 12.6.1

To sketch a proof, the DAG implies that the observed variables D, Y, Q, S and the unobserved variable A obey the two conditional independence relations:

$$(i) \quad S \perp\!\!\!\perp (Q, D) \mid A \quad (ii) \quad Q \perp\!\!\!\perp Y \mid (A, D). \qquad (12.A.2)$$

These in turn imply

$$\begin{aligned}\Pi(S \mid Q, d) &= \Pi(S \mid Q, A, d)\Pi(A \mid Q, d) \\ &= \Pi(S \mid A)\Pi(A \mid Q, d)\end{aligned}$$

and

$$\begin{aligned}\Pi(y \mid Q, d) &= \Pi(y \mid Q, A, d)\Pi(A \mid Q, d) \\ &= \Pi(y \mid A, d)\Pi(A \mid Q, d).\end{aligned}$$

We now want to solve these equations for $\Pi(y \mid A, d)$ in terms of quantities that could be learned in the data.

We will need invertibility of $\Pi(S \mid Q, d)$ which requires invertibility of both $\Pi(S \mid A)$ and $\Pi(A \mid Q, d)$. Under these invertibility conditions, we have

$$\Pi(A \mid Q, d) = \Pi(S \mid A)^{-1}\Pi(S \mid Q, d)$$

and

$$\Pi(y \mid Q, d) = \Pi(y \mid A, d)\Pi(S \mid A)^{-1}\Pi(S \mid Q, d),$$

which yield

$$\Pi(y \mid A, d) = \Pi(y \mid Q, d)\Pi(S \mid Q, d)^{-1}\Pi(S \mid A).$$

Next, because A blocks backdoor paths between D and Y, we have that

$$p(y \mid a : do(d)) = p(y \mid a, d) \tag{12.A.3}$$

or, after integrating out a,

$$p(y : do(d)) = \Pi(y \mid A, d)\Pi(A),$$

which can be further expressed as

$$\Pi(y \mid d, Q)\, \Pi(S \mid Q, d)^{-1}\, \Pi(S)\,, \tag{12.A.4}$$

using the derivations above.

Bibliography

[1] Philip G. Wright. *The Tariff on Animal and Vegetable Oils*. New York: The Macmillan company, 1928 (cited on pages 324, 331, 332, 337).

[2] Zvi Griliches. 'Estimating the returns to schooling: Some econometric problems'. In: *Econometrica* 45.1 (1977), pp. 1–22 (cited on pages 325, 330, 339).

[3] Wang Miao, Zhi Geng, and Eric J. Tchetgen Tchetgen. 'Identifying causal effects with proxy variables of an unmeasured confounder'. In: *Biometrika* 105.4 (2018), pp. 987–993 (cited on pages 325, 341, 342).

[4] Ben Deaner. 'Proxy controls and panel data'. In: *arXiv preprint arXiv:1810.00283* (2018) (cited on pages 325, 342).

[5] Carlos Cinelli and Chad Hazlett. 'Making sense of sensitivity: Extending omitted variable bias'. In: *Journal of the Royal Statistical Society: Series B* 82.1 (2020), pp. 39–67 (cited on pages 326–328).

[6] Chad Hazlett. 'Angry or Weary? How Violence Impacts Attitudes toward Peace among Darfurian Refugees'. In: *Journal of Conflict Resolution* 64.5 (2020), pp. 844–870. DOI: 10.1177/0022002719879217 (cited on page 327).

[7] Victor Chernozhukov, Carlos Cinelli, Whitney Newey, Amit Sharma, and Vasilis Syrgkanis. 'Long Story Short: Omitted Variable Bias in Causal Machine Learning'. In: *arXiv preprint arXiv:2112.13398* (2023) (cited on pages 328, 329).

[8] David Card. 'Using Geographic Variation in College Proximity to Estimate the Return to Schooling'. In: *Aspects of Labor Market Behavior: Essays in Honor of John Vanderkamp*. Ed. by L. N. Christofides and R. Swidinsky. Toronto: University of Toronto Press, 1995, pp. 201–222 (cited on page 330).

[9] Joshua D. Angrist and Alan B. Krueger. 'Does Compulsory School Attendance Affect Schooling and Earnings?' In: *Quarterly Journal of Economics* 106.4 (1991), pp. 979–1014 (cited on page 330).

[10] Howard S. Bloom, Larry L. Orr, Stephen H. Bell, George Cave, Fred Doolittle, Winston Lin, and Johannes M. Bos. 'The Benefits and Costs of JTPA Title II-A Programs: Key Findings from the National Job Training Partnership Act Study'. In: *The Journal of Human Resources* 32.3 (1997), pp. 549–576. (Visited on 04/01/2024) (cited on page 330).

[11] Stephen V. Cameron and James J. Heckman. 'Life Cycle Schooling and Dynamic Selection Bias: Models and Evidence for Five Cohorts of American Males'. In: *Journal of Political Economy* 106.2 (1998), pp. 262–333. (Visited on 04/01/2024) (cited on page 330).

[12] Alberto Abadie. 'Semiparametric instrumental variable estimation of treatment response models'. In: *Journal of Econometrics* 113.2 (2003), pp. 231–263 (cited on page 335).

[13] Peter M. Aronow and Allison Carnegie. 'Beyond LATE: Estimation of the Average Treatment Effect with an Instrumental Variable'. In: *Political Analysis* 21.4 (2013), pp. 492–506. (Visited on 02/27/2023) (cited on page 335).

[14] Ryo Okui, Dylan S. Small, Zhiqiang Tan, and James M. Robins. 'Doubly Robust Instrumental Variable Regression'. In: *Statistica Sinica* 22.1 (2012), pp. 173–205. (Visited on 02/27/2023) (cited on page 335).

[15] Susan Athey and Stefan Wager. 'Policy learning with observational data'. In: *Econometrica* 89.1 (2021), pp. 133–161 (cited on page 335).

[16] Vasilis Syrgkanis, Victor Lei, Miruna Oprescu, Maggie Hei, Keith Battocchi, and Greg Lewis. 'Machine learning estimation of heterogeneous treatment effects with instruments'. In: *Advances in Neural Information Processing Systems* 32 (2019) (cited on page 335).

[17] Guido W. Imbens and Joshua D. Angrist. 'Identification and Estimation of Local Average Treatment Effects'. In: *Econometrica* 62.2 (1994), pp. 467–475 (cited on page 336).

[18] Victor Chernozhukov, Christian Hansen, and Kaspar Wuthrich. 'Instrumental variable quantile regression'. In: *arXiv preprint arXiv:2009.00436* (2020) (cited on page 339).

[19] Gary Chamberlain and Zvi Griliches. *More on brothers. In "Kinometrics: Determinants of Socioeconomic Success Within and Between Families "(P. Taubman, Ed.)* 1977 (cited on pages 339, 341).

[20] Marc Lipsitch, Eric Tchetgen Tchetgen, and Ted Cohen. 'Negative controls: a tool for detecting confounding and bias in observational studies'. In: *Epidemiology* 21.3 (2010), pp. 383–388 (cited on page 341).

[21] Ben Deaner. 'Proxy controls and panel data'. In: *arXiv preprint arXiv:1810.00283* (2018) (cited on page 341).

[22] Nathan Kallus, Xiaojie Mao, and Masatoshi Uehara. 'Causal inference under unmeasured confounding with negative controls: A minimax learning approach'. In: *arXiv preprint arXiv:2103.14029* (2021) (cited on pages 341, 343).

[23] Xu Shi Wang Miao Yifan Cui Hongming Pu and Eric Tchetgen Tchetgen. 'Semiparametric Proximal Causal Inference'. In: *Journal of the American Statistical Association* 0.0 (2023), pp. 1–12. DOI: 10.1080/01621459.2023.2191817 (cited on pages 341, 343).