Chapter 3

Policy Evaluation II: Identifiability in parametric models

In Chapter 2, we considered identification where we make only *non-parametric* assumptions on our causal model: assumptions about the factorization of the distribution $\mathbb{P}_{\mathcal{X}}$, but not about any of the individual terms $\mathbb{P}_{\mathcal{X}}(X_i \mid \mathrm{pa}_{\mathcal{G}}(X_i))$. In econometrics and other fields, it is common to achieve better identifiability guarantees by making *parametric* assumptions on our models. Such assumptions restrict the functional form of the causal mechanisms f_{X_i} to a class of functions that can be indexed by a finite-dimensional set of parameters. Common parametric models include linear models, where

$$f_{X_i}(\mathrm{pa}_{\tilde{\mathcal{G}}}(X_i)) = \sum_{V_i \in \mathrm{pa}_{\tilde{\mathcal{G}}}(X_i)} \beta_{V_i, X_i} V_i,$$

binary choice models, i.e.,

$$f_{X_i}(\mathrm{pa}_{\tilde{\mathcal{G}}}(X_i)) = \mathbb{1}_{U_i \geq 0} \qquad \text{where } U_i = \sum_{V_i \in \mathrm{pa}_{\tilde{\mathcal{G}}}(X_i)} \beta_{V_i, X_i} V_i,$$

exponential family models, and more. In the lecture, we will focus on linear models.

3.1 Instrumental variables

One of the most commonly used models for achieving identifiability via parametric assumptions is the **instrumental variable model**.

Definition 3.1. *We say that (W, A, Y) satisfy the* **instrumental variable model** *if there exists Z such that (Z, W, A, Y) are generated according to a linear structural causal model with causal graph $\mathcal{G}$ in Figure 3.1, i.e.,*

$$\begin{aligned} Z &= \varepsilon_z \\ W &= \varepsilon_w \\ A &= \beta_{za} Z + \beta_{wa} W + \varepsilon_a \\ Y &= \beta_{zy} Z + \beta_{ay} A + \varepsilon_y \end{aligned}$$

all zero means Z, W

for $\varepsilon_z, \varepsilon_w, \varepsilon_a, \varepsilon_y$ independent. The variable W is called an **instrumental variable**.

Remark 3.1. *An important feature of the instrumental variable model is that W does not have a direct effect on Y, i.e., all effects of W on Y are mediated via A. This feature is called the* **exclusion restriction**.

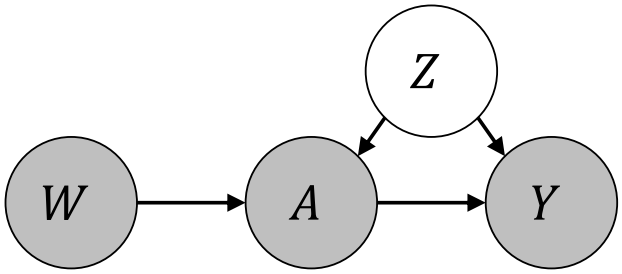

Figure 3.1: The instrumental variable model. W is called the **instrumental variable** for A.

Example 3.1. *A canonical example of the instrumental variable model is for the estimation of the effect of smoking (A) on cancer (Y). The association between smoking and higher rates of cancer has been well-known for over a century. However, in the 1950's, there was vigorous debate over whether smoking* caused *higher rates of cancer. Statisticians aligned with the tobacco industry noted that there may be some confounding factors (U), such as socioeconomic status, that explained the correlation between smoking and cancer. Possible instrumental variables for smoking include a tax on tobacco: higher taxes are likely to decrease the prevalence of smoking, but are not expected to have a direct effect on the rate of cancer.*

Theorem 3.1. *Let (W, A, Y) satisfy the instrumental variable model with $\beta_{wa} \neq 0$. Then β_{ay} is identifiable from $\mathbb{P}_{\mathcal{X}}(W, A, Y)$.*

Proof.

$$Y = (\beta_{zy} + \beta_{za}\beta_{ay})Z + \beta_{wa}\beta_{ay}W + \beta_{ay}\varepsilon_a + \varepsilon_y$$

We have $\mathbb{E}[A \mid W = \mathbf{w}] = \beta_{wa}\mathbf{w}$ by independence of W from Z and ε_a. Thus, β_{wa} is identifiable by linear regression of A on W.

Further,

$$Y = (\beta_{zy} + \beta_{za}\beta_{ay})Z + \beta_{wa}\beta_{ay}W + \beta_{ay}\varepsilon_a + \varepsilon_y$$

Thus, $\mathbb{E}[Y \mid W = \mathbf{w}] = \beta_{wa}\beta_{ay}\mathbf{w}$ by independence of W from Z, ε_a, and ε_y.

Thus, $\beta_{wa}\beta_{ay}$ is identifiable by linear regression of Y on W. Therefore, we can identify β_{ay} as the ratio of these two regression coefficients. □

Remark 3.2. *Note that our method of identifying β_{ay} uses two linear regressions. When the linear regression coefficients are estimated from data, the resulting estimator of β_{ay} is called the* **two-stage least squares** *estimator. Note that this estimator requires dividing by β_{wa}, which can lead to an estimator with high variance if the value of β_{wa} is near zero. In such cases, W is called a* **weak instrument**.

3.2 Proxy variables

Definition 3.2. *We say that (V, W, A, Y) satisfy the* **proxy variable model** *if there exists Z such that (Z, V, W, A, Y) is generated according to a linear structural causal model with causal graph $\mathcal{G}$ in Figure 3.2, i.e.,*

$$\begin{aligned} Z &= \varepsilon_z & Var(\varepsilon_z) &= \sigma_z^2 \\ V &= \beta_{zv}Z + \varepsilon_v & Var(\varepsilon_z) &= \sigma_v^2 \\ W &= \beta_{zw}Z + \varepsilon_w & Var(\varepsilon_z) &= \sigma_w^2 \\ A &= \beta_{za}Z + \varepsilon_a & Var(\varepsilon_z) &= \sigma_a^2 \\ Y &= \beta_{zy}Z + \beta_{ay}A + \varepsilon_y & Var(\varepsilon_z) &= \sigma_y^2 \end{aligned}$$

for $\varepsilon_z, \varepsilon_v, \varepsilon_w, \varepsilon_a, \varepsilon_y$ independent. The variables V and W are called an **proxies** *for Z.*

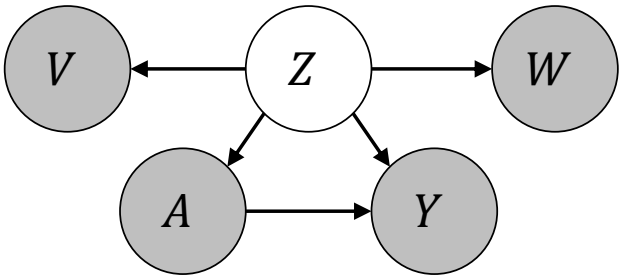

Figure 3.2: The proxy variable setup. V and W are proxies for Z.

Example 3.2. *Suppose once again that we are trying to measure the impact of smoking on the rate of cancer. These might be confounded by Z, an individual's attitude toward health-related behaviors. While we might not have data about Z, researchers may have proxies for Z including V, the amount an individual exercises, and W, the frequency with which the individual consumes multivitamins.*

Theorem 3.2. *Let (V, W, A, Y) satisfy a proxy variable model. Assume that $\beta_{zv} \neq 0$, $\beta_{zw} \neq 0$, $\sigma_a^2 \neq 0$, $\sigma_z^2 \neq 0$. Let $\Sigma_{x_1,x_2} := \mathrm{Cov}(X_1, X_2)$*

$$\beta_{ay} = \frac{t_{av,yw}}{t_{av,aw}} = \frac{\Sigma_{ay}\Sigma_{vw} - \Sigma_{aw}\Sigma_{vy}}{\Sigma_{aa}\Sigma_{vw} - \Sigma_{aw}\Sigma_{va}}$$

where $t_{ij,uv} := \det(\Sigma_{[ij],[uv]})$.

Proof. We have

$$\begin{aligned} V &= \beta_{zv}\varepsilon_z + \varepsilon_z \\ W &= \beta_{zw}\varepsilon_z + \varepsilon_w \\ A &= \beta_{za}\varepsilon_z + \varepsilon_a \\ Y &= (\beta_{zy} + \beta_{za}\beta_{ay})\varepsilon_z + \beta_{ay}\varepsilon_a + \varepsilon_y \end{aligned}$$

By the assumed independences,

$$\begin{aligned} \Sigma_{vw} &= \sigma_z^2\beta_{zv}\beta_{zw} \quad \text{<- how to derive this?} \\ \Sigma_{va} &= \sigma_z^2\beta_{za}\beta_{zv} \\ \Sigma_{aw} &= \sigma_z^2\beta_{za}\beta_{zw} \\ \Sigma_{vy} &= \sigma_z^2(\beta_{zy}\beta_{zv} + \beta_{za}\beta_{ay}\beta_{zv}) \\ \Sigma_{ay} &= \sigma_z^2(\beta_{zy}\beta_{za} + \beta_{za}^2\beta_{ay}) + \sigma_a^2\beta_{ay} \\ \Sigma_{aa} &= \sigma_z^2\beta_{za}^2 + \sigma_a^2 \end{aligned}$$

Thus,

$$\begin{aligned} \Sigma_{ay}\Sigma_{vw} - \Sigma_{aw}\Sigma_{vy} &= \sigma_z^4(\beta_{zy}\beta_{za} + \beta_{za}^2\beta_{ay})\beta_{zv}\beta_{zw} + \sigma_z^2\sigma_a^2\beta_{ay}\beta_{zv}\beta_{zw} - \sigma_z^4(\beta_{zy}\beta_{zv} + \beta_{za}\beta_{ay}\beta_{zv})\beta_{za}\beta_{zw} \\ &= \sigma_z^2\sigma_a^2\beta_{ay}\beta_{zv}\beta_{zw} \end{aligned}$$

and

$$\begin{aligned} \Sigma_{aa}\Sigma_{vw} - \Sigma_{aw}\Sigma_{va} &= \left(\sigma_z^4\beta_{za}^2\beta_{zv}\beta_{zw} + \sigma_z^2\sigma_a^2\beta_{zv}\beta_{zw}\right) - \sigma_z^4\beta_{za}^2\beta_{zv}\beta_{zw} \\ &= \sigma_z^2\sigma_a^2\beta_{zv}\beta_{zw} \end{aligned}$$

□

Remark 3.3. *Theorem 3.2 can be generalized to the case where $V \to A$ and $W \to Y$, see e.g. Kuroki and Pearl (2014). In such a case, where A does not have a causal effect on W, W is sometimes referred to as a* **negative control outcome**, *since it is an outcome that should not be effected by the intervention on A.*

3.3 Additional Reading

- **Nonlinear instrumental variables models and proxy variable models.** In general non-parametric models (those with no restrictions on the probability distributions $\mathbb{P}_{\mathcal{X}}(X_i \mid \text{pa}_{\mathcal{G}}(X_i))$), $\mathbb{P}_{\mathcal{X}}(Y \mid \texttt{do}(A = a))$ is not identified for the causal graph in Figure 3.1. However, we do not need to go all the way to parametric models in order to guarantee identifiability: we may retain significant expressivity in these conditional distributions and still identify the interventional quantity, see e.g. Newey and Powell (2003) and Singh et al. (2019). Similar comments apply to the proxy variable model, see e.g. Kallus et al. (2021).

- **Identifiability in linear models.** There is a large literature on identifying causal effects in linear structural causal models, see for example Kumor et al. (2020), Barber et al. (2022), Drton et al. (2016). In linear models, identifiability can always be checked using techniques from computational algebra, but these are computationally intractable for large graphs. Thus, these papers seek to develop graphical conditions to check identifiability. Thus far, most papers focus on sufficient conditions for identifiability, though a necessary condition is given in Foygel et al. (2012). To the best of my knowledge, there is not yet a necessary and sufficient graphical characterization for identifiability.

- **Estimation of treatment effects.** In the past two lectures, we have been concerned with determining whether the interventional quantities of interest are identifiable from observational data. This is intrinsically a question about something that happens in the *infinite data* limit. This leaves open a major question: what should we do in practice, where we always have a finite amount of data?

  One intuitive answer is suggested by the observation that our proofs of identifiability often result in *identification formulas* which express the target interventional quantity in terms of observational quantities. This suggests that we could estimate the observational quantities from data (e.g., $\mathbb{P}_{\mathcal{X}}(Y \mid A = a, \mathbf{S} = \mathbf{s})$ and $\mathbb{P}_{\mathcal{X}}(\mathbf{S} = \mathbf{s})$ in the context of backdoor adjustment), and then plug these estimates into our identification formula. This approach is called the **plug-in** approach, and works well in parametric models (including linear models or models where all variables are discrete). Indeed, if one estimates all parameters of a structural causal model using maximum likelihood estimation, then the plug-in approach can be shown to be (asymptotically) better than any other approach for estimating the target interventional quantity, using classical parametric efficiency theory.

  However, if one does not make such parametric assumptions, then the plug-in approach can be shown to be quite suboptimal. For example, if $\mathbf{S}$ is continuous-valued in the setting of backdoor adjustment, then $\mathbb{P}_{\mathcal{X}}(Y \mid A = a, \mathbf{S} = \mathbf{s})$ can be a complicated function which takes a large amount of data to estimate. The plug-in approach propagates errors in the estimation of these so-called **nuisance functions** to errors in the estimation of the target intervention quantity. More sophisticated methods can avoid this error propagation, see e.g. Kennedy (2022).

  Unfortunately, in this lecture series, we will not have time to properly address these statistical questions.

Bibliography

Barber, R. F., Drton, M., Sturma, N., and Weihs, L. (2022). Half-trek criterion for identifiability of latent variable models. *arXiv preprint arXiv:2201.04457*.

Drton, M. et al. (2016). Algebraic problems in structural equation modeling. *The 50th anniversary of Gröbner bases*, pages 35–86.

Foygel, R., Draisma, J., and Drton, M. (2012). Half-trek criterion for generic identifiability of linear structural equation models. *The Annals of Statistics*, pages 1682–1713.

Kallus, N., Mao, X., and Uehara, M. (2021). Causal inference under unmeasured confounding with negative controls: A minimax learning approach. *arXiv preprint arXiv:2103.14029*.

Kennedy, E. H. (2022). Semiparametric doubly robust targeted double machine learning: a review. *arXiv preprint arXiv:2203.06469*.

Kumor, D., Cinelli, C., and Bareinboim, E. (2020). Efficient identification in linear structural causal models with auxiliary cutsets. In *International Conference on Machine Learning*, pages 5501–5510. PMLR.

Kuroki, M. and Pearl, J. (2014). Measurement bias and effect restoration in causal inference. *Biometrika*, 101(2):423–437.

Newey, W. K. and Powell, J. L. (2003). Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578.

Singh, R., Sahani, M., and Gretton, A. (2019). Kernel instrumental variable regression. *Advances in Neural Information Processing Systems*, 32.