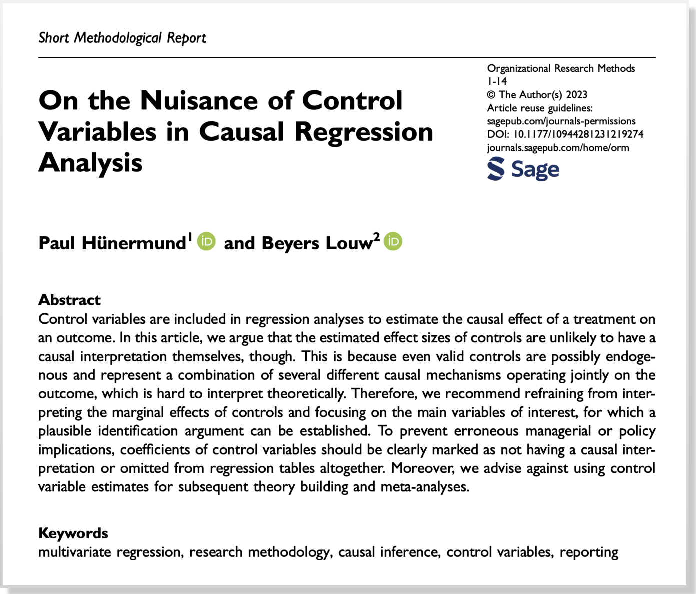
On the Nuisance of Control Variables in Causal Regression Analysis

CARMA Webcast Lecture, April 12th, 2024

Paul Hünermund

CBS COPENHAGEN BUSINESS SCHOOL
HANDELSHØJSKOLEN



Short Methodological Report

Organizational Research Methods
1-14
© The Author(s) 2023
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/10944281231219274
journals.sagepub.com/home/orm
Sage

On the Nuisance of Control Variables in Causal Regression Analysis

Paul Hünermund[1] and **Beyers Louw**[2]

Abstract

Control variables are included in regression analyses to estimate the causal effect of a treatment on an outcome. In this article, we argue that the estimated effect sizes of controls are unlikely to have a causal interpretation themselves, though. This is because even valid controls are possibly endogenous and represent a combination of several different causal mechanisms operating jointly on the outcome, which is hard to interpret theoretically. Therefore, we recommend refraining from interpreting the marginal effects of controls and focusing on the main variables of interest, for which a plausible identification argument can be established. To prevent erroneous managerial or policy implications, coefficients of control variables should be clearly marked as not having a causal interpretation or omitted from regression tables altogether. Moreover, we advise against using control variable estimates for subsequent theory building and meta-analyses.

Keywords

multivariate regression, research methodology, causal inference, control variables, reporting

Introduction

- The main purposes of regression analyses is to control for confounding influence factors between a *treatment* and an *outcome* in order to obtain consistent *causal effect* estimates
- In practice scholars often overstate the role of control variables
- In this paper we discuss why control variables do not necessarily have a causal interpretation themselves and should therefore be interpreted carefully
- 47% of papers published in *Organization Science* and *Strategic Management Journal* 2015–2020 and using regression methods explicitly discussed the estimated effect sizes of controls
  - Most common formulations such as *"control variables have expected signs"* or *"it is worth noting the coefficients of our control variables"*
- We argue that this could lead to potentially misleading conclusions and a false sense of accumulating empirical evidence

Recommendations in Previous Literature

- Control variables should carry the same importance as the main independent variables (Becker, 2005; Spector and Brannick, 2011; Carlson and Wu, 2012; Atinc et al., 2012)
- Report all regression coefficients of control variables as well as their significance levels (Becker, 2005)
- Controls should be given equal status to the main treatment variable in the analysis (Spector and Brannick, 2011)
- Provide an ex-ante prediction of the sign of the relationship between the controls and outcome variable based on theory (Atinc et al., 2012)
- Overall consensus in the organizational literature seems to be:
  - Interpreting control variable estimates is safe
  - They add to the body of cumulative evidence regarding a particular effect size

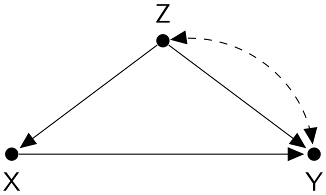
Structural Causal Models

$$
\begin{aligned}
z &\leftarrow f_Z(u_z) \\
x &\leftarrow f_X(z, u_x) \\
y &\leftarrow f_Y(x, z, u_Y)
\end{aligned}
$$

- The f_i's denote the causal mechanisms in the model
  - Are not restricted to be linear as in traditional SEM
- The u_i's refer to background factors that are determined outside of the model
- Assignment operator ($\leftarrow$) captures asymmetry of causal relationships
  - $x \leftarrow a \cdot z \;\neq\; z \leftarrow x/a$
- Similar to definition of “structure” according to Cowles foundation

Directed Acyclic Graphs

$$
\begin{aligned}
z &\leftarrow f_z(u_z) \\
x &\leftarrow f_x(z, u_x) \\
y &\leftarrow f_y(x, z, u_y)
\end{aligned}
$$



- In a fully specified SCM, every counterfactual quantity is computable
- In most social science contexts it's hard to know the causal mechanisms f_i and distribution of background factors $P(U)$
- Therefore, restrict attention to qualitative causal information of the model, which can be encoded by a graph G
  - Nodes V: variables in the model
  - Directed edges E: causal relationships in the model

Directed Acyclic Graphs

- No functional form or distributional assumptions means that framework remains fully nonparametric
  - Particularly helpful in fields where theory is purely qualitative and no shape restrictions on can be derived (Matzkin, 2007)
- $Z \dashleftarrow\dashrightarrow Y$ is a shortcut notation for unobserved common causes $Z \leftarrow U \rightarrow Y$
- Acyclicity
  - Directed cycles such as $A \rightarrow B \rightarrow C \rightarrow A$ are excluded
  - This means there are no feedback loops
  - Otherwise A could be a cause of itself
  - Gives rise to what economists call a *recursive* model (Wold, 1954)
  - Extensions of the SCM framework to cyclic graphs exist (Bongers et al., 2021)

Structural Econometrics vs. PO
Recursive vs. Interdependent Systems

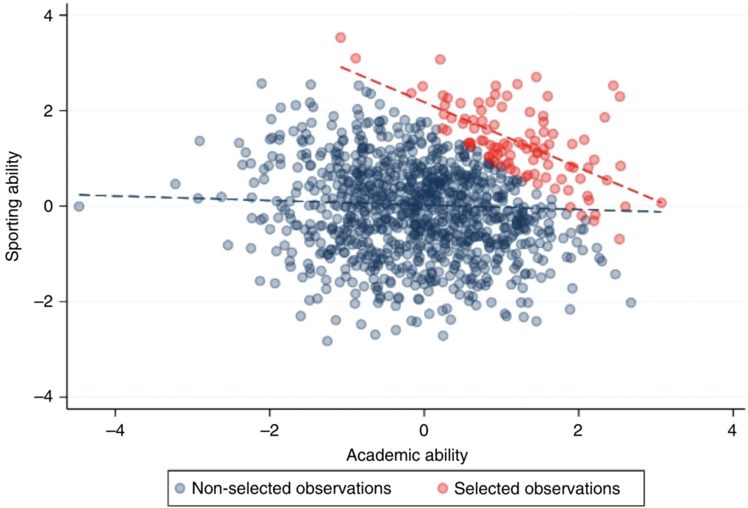
D-Separation

- DAGs are such a useful tool because they are able to efficiently encode conditional independence relationships:

Chain:	$A \rightarrow B \rightarrow C$	$\Rightarrow$	$A \not\perp\!\!\!\perp C$ and $A \perp\!\!\!\perp C \vert B$
Fork:	$A \leftarrow B \rightarrow C$	$\Rightarrow$	$A \not\perp\!\!\!\perp C$ and $A \perp\!\!\!\perp C \vert B$
Collider:	$A \rightarrow B \leftarrow C$	$\Rightarrow$	$A \perp\!\!\!\perp C$ and $A \not\perp\!\!\!\perp C \vert B$

- The same holds for longer paths in the graph
  - Conditioning on a variable along a chain or fork blocks (*"d-separates"*) the path
  - Conditioning on a collider opens the path

Collider Bias Example



Source: "Collider bias undermines our understanding of COVID-19 disease risk and severity" (2020, Nature Comm.)

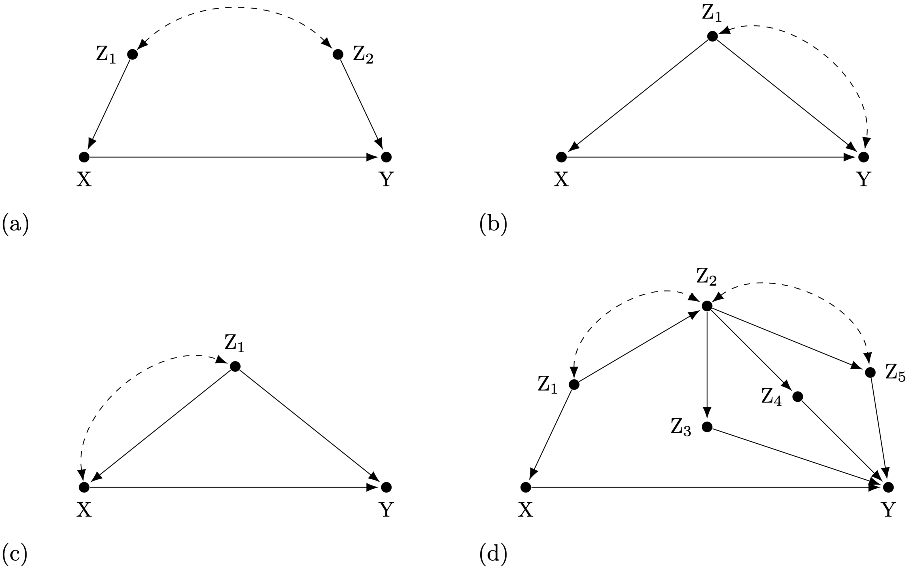
Backdoor Adjustment

Definition: The Backdoor Criterion (Pearl et al., 2016, p. 61)

Given an ordered pair of of variables (X, Y) in a directed acyclic graph G, a set of variables Z satisfies the backdoor criterion relative to (X, Y) if no node in Z is a descendant of X, and Z blocks every path between X and Y that contains an arrow into X.

- Intuition: block all spurious paths between X and Y while leaving direct paths unperturbed and creating no new spurious paths
  - Avoid introducing collider bias (*"bad controls"*)
- Finding suitable adjustment sets Z can be easily automated (Textor and Liśkiewicz, 2011)

Examples with Valid Control Variable Z_1



(a)

(b)

(c)

(d)

Simulation Setup

- For illustration we simulate data from the SCMs implied by (a)–(d)
- For simplicity we use linear relationships and effect sizes normalized to one; e.g., for (a)

  $$\begin{aligned} z_1 &\leftarrow u + \varepsilon_1, \\ z_2 &\leftarrow u + \varepsilon_2, \\ x &\leftarrow z_1 + \varepsilon_3, \\ y &\leftarrow x + z_2 + \varepsilon_4, \end{aligned}$$

  with $n = 10,000$, and U, ε_i being standard normal
- Standard errors are bootstrapped with $1,000$ replications

Simulation Results

	Figure 1a			Figure 1b	Figure 1c	Figure 1d		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment Variable:								
X	1.017	1.004	1.015	0.993	1.001	0.991	1.006	1.003
	(0.015)	(0.006)	(0.010)	(0.012)	(0.008)	(0.057)	(0.007)	(0.010)
Control Variables **(not to be interpreted causally)**:								
Z_1	0.499		-0.019	1.503	1.004	4.565		0.004
	(0.018)		(0.013)	(0.014)	(0.014)	(0.069)		(0.016)
Z_2		0.993	0.997					0.009
		(0.008)	(0.008)					(0.019)
Z_3							0.994	0.991
							(0.008)	(0.010)
Z_4							0.991	0.988
							(0.008)	(0.010)
Z_5							1.011	1.009
							(0.006)	(0.008)

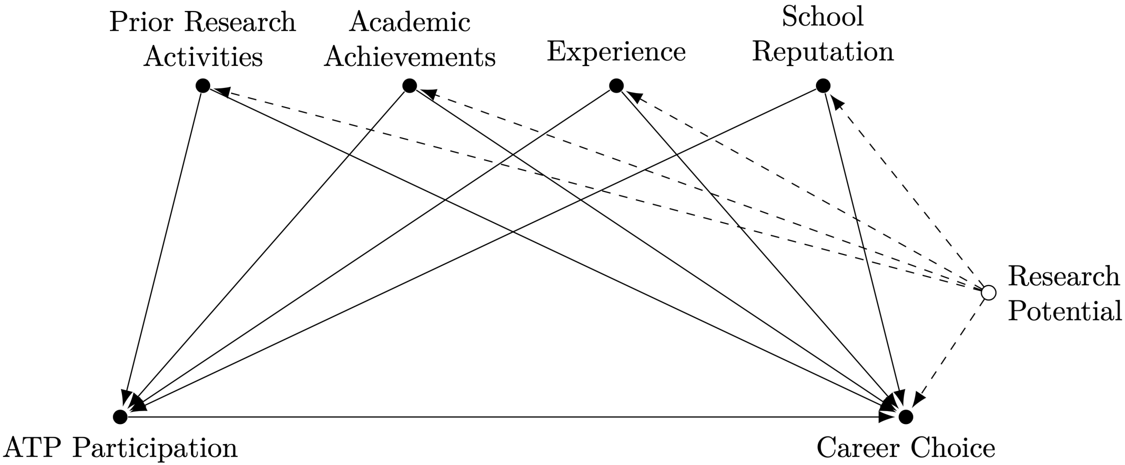
Example #1: Hoffman and Strezhnev (2023)

- Research question: effect of longer travel time on default judgments in eviction cases
- OLS analysis of 200,000+ eviction proceedings in Philadelphia (2005–2021)
- Main finding: One-hour increase in travel time raises default judgment likelihood by 3.8–8.6%
- They control for neighborhood characteristics and building types, among other things
- The find a positive and significant effect of multi-unit apartment buildings (compared to row houses or single family dwelling) on the probability of default judgements
  - Unlikely to have a causal interpretation
  - Building characteristics might be correlated with other factors such as unfavorable terms in residential leases or the geographical distribution of dwellings within the city

Example #2: Azoulay et al. (2021)

- Investigate effect of early career exposure to frontier research on the career trajectory of potential innovators
- Empirical setting: Associate Training Program (ATP) of NIH
- ATP started in 1953 for recent MD graduates
- Participants received 2–3 years research training at NIH intramural campus
- Popular among young physicians during Vietnam War period
- Screening criteria for ATP applicants: research activities, academic achievements, experience, institutional reputation
- Selection based on observable characteristics at interview stage
    - Difficulty in predicting future research potential beyond observable markers
- ATP participants twice as likely to pursue research-focused career, leading to more publications, citations, grant funding, prestigious awards, and membership in National Academy of Sciences

Causal Diagram Capturing Assumptions by Azoulay et al.



Example #3: Analyst Coverage & Innovation

- He and Tian (2013): Negative relationship between analyst coverage and patenting
  - Study of U.S. public firms from 1993 to 2005
  - Utilizes difference-in-differences and instrumental variable approach
- Theorized: Analyst pressure may worsen managerial myopia and impede innovation investment
- Analyst coverage commonly used as control variable in studies of R&D activities in publicly listed firms
- Less stringent identification strategies may yield unexpected results
- Chen et al. (2016) and Huang et al. (2022): Find positive effects of analyst coverage on patents, seemingly contradicting He and Tian (2013)
  - Interpreting positive regression coefficients as evidence against He and Tian (2013) would be a mistake
  - No Bayesian updating about the effect of analyst coverage should take place

Summary

- The purpose of regression analysis in organizational research is typically to build and test theories that explain the causal mechanisms (Sutton and Staw, 1995)
- Attaching substantive meaning to the marginal effects of biased control variables can lead to erroneous managerial and policy conclusions
- Control variables can be endogenous and will likely be so in practice (Frölich, 2008)
- Controls should be chosen to close all backdoor paths between a treatment and outcome, based on a theoretical model of the context under study (Bono and McNamara, 2011)
- It is thereby not necessary to include all causal influence factors of the outcome variable in a regression
- It might be easier to control the treatment assignment mechanism, if institutional knowledge is richer about what determines treatment take-up (see Azoulay et al.)

Our Recommendations

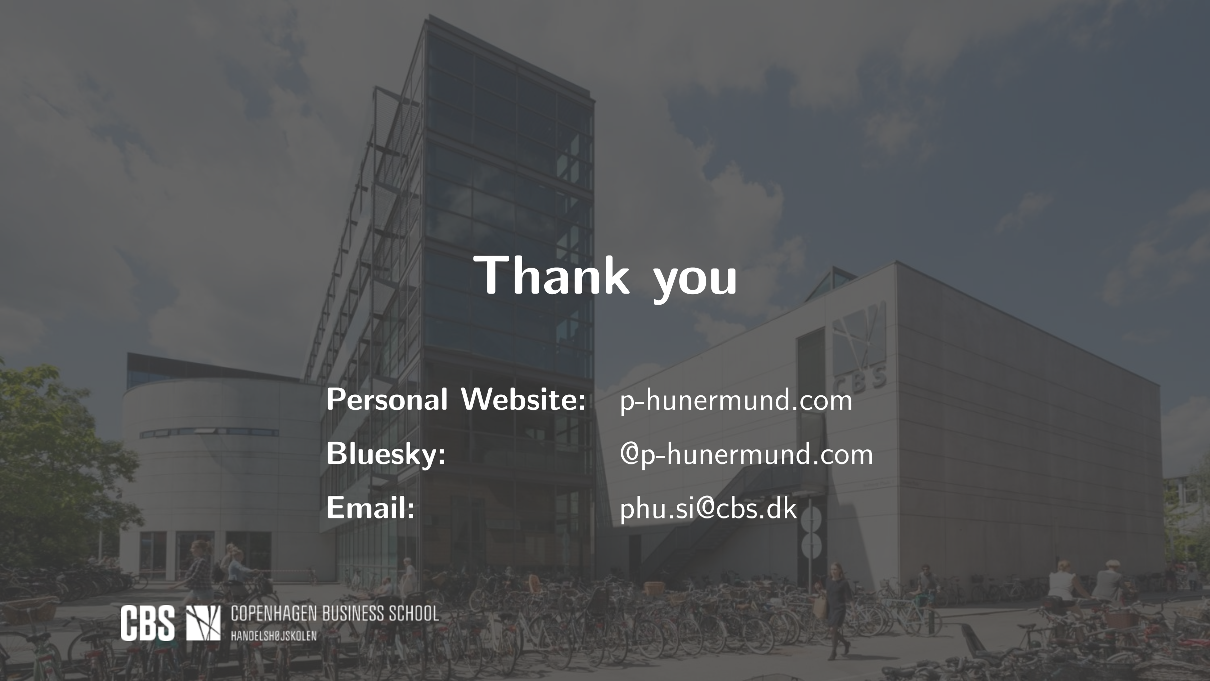
- Since accounting for all influence factors of the outcome might be unrealistic in many contexts, interpreting control variables in light of theory is potentially dangerous
- We recommend to treat controls as *nuisance parameters*, which are included in the analysis for identification purposes (and discussed as such) but their effects are not interpreted
    - Corresponds to the way control variables are treated by matching estimators (Heckman et al., 1998) and modern ML methods (Chernozhukov et al., 2018; Hünermund et al., 2023)
- Control variable should not be promoted to have equal status with the other variables and formulations such as *"estimates of control variables have expected signs"* should be avoided
- As a "nudge", we find it appropriate that authors omit their coefficients entirely from regression tables or relegate them to an appendix

Our Preferred Regression Table Format

	Figure 1a			Figure 1b	Figure 1c	Figure 1d		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
X	1.017	1.004	1.015	0.993	1.001	0.991	1.006	1.003
	(0.015)	(0.006)	(0.010)	(0.012)	(0.008)	(0.057)	(0.007)	(0.010)
Z_1	✓		✓	✓	✓	✓		✓
Z_2		✓	✓					✓
Z_3							✓	✓
Z_4							✓	✓
Z_5							✓	✓

A "Compromise"

	Figure 1a			Figure 1b	Figure 1c	Figure 1d		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment Variable:								
X	1.017	1.004	1.015	0.993	1.001	0.991	1.006	1.003
	(0.015)	(0.006)	(0.010)	(0.012)	(0.008)	(0.057)	(0.007)	(0.010)
Control Variables **(not to be interpreted causally)**:								
Z_1	0.499		-0.019	1.503	1.004	4.565		0.004
	(0.018)		(0.013)	(0.014)	(0.014)	(0.069)		(0.016)
Z_2		0.993	0.997					0.009
		(0.008)	(0.008)					(0.019)
Z_3							0.994	0.991
							(0.008)	(0.010)
Z_4							0.991	0.988
							(0.008)	(0.010)
Z_5							1.011	1.009
							(0.006)	(0.008)



Thank you

Personal Website:	p-hunermund.com
Bluesky:	@p-hunermund.com
Email:	phu.si@cbs.dk

CBS COPENHAGEN BUSINESS SCHOOL
HANDELSHØJSKOLEN

References I

Guclu Atinc, Marcia J. Simmering, and Mark J. Kroll. Control variable use and reporting in macro and micro management research. *Organizational Research Methods*, 15(1):57–74, 2012. doi: 10.1177/1094428110397773.

Pierre Azoulay, Wesley H. Greenblatt, and Misty L. Heggeness. Long-term effects from early exposure to research: Evidence from the NIH "yellow berets". *Research Policy*, 50(9):104332, 2021. doi: 10.1016/j.respol.2021.104332.

Thomas E. Becker. Potential problems in the statistical control of variables in organizational research: A qualitative analysis with recommendations. *Organizational Research Methods*, 8(3):274–289, July 2005. doi: 10.1177/1094428105278021.

R. Bentzel and B. Hansen. On recursiveness and interdependency in economic modeks. *The Review of Economic Studies*, 22(3):153–168, 1954 - 1955.

Stephan Bongers, Patrick Forré, Jonas Peters, and Joris M. Mooij. Foundations of sturctural causal models with cycles and latent variables. *The Annals of Statistics*, 49(5):2885–2915, 2021.

Joyce E Bono and Gerry McNamara. Publishing in amj–part 2: Research design. *Academy of management journal*, 54(4): 657–660, 2011. doi: 10.5465/amj.2011.64869103.

Kevin D. Carlson and Jinpei Wu. The illusion of statistical control: Control variable practice in management research. *Organizational Research Methods*, 15(3):413–435, 2012. doi: 10.1177/1094428111428817.

Nancy Cartwright. *Hunting Causes and Using Them*. Cambridge University Press, 2007.

Chen Chen, Yangyang Chen, Pso-Huan Hsu, and Edward J. Podolski. Be nice to your innovators: Employee treatment and corporate innovation performance. *Journal of Corporate Finance*, 39:78–98, 2016. doi: 10.1016/j.jcorpfin.2016.06.001.

References II

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1): C1–C68, 2018. doi: 10.1111/ectj.12097.

Markus Frölich. Parametric and nonparametricregression in the presence of endogenous control variables. *International Statistical Review*, 76(2):214–227, aug 2008. doi: 10.1111/j.1751-5823.2008.00045.x.

Trygve Haavelmo. The statistical implications of a system of simultaneous equations. *Econometrica*, 11(1):1–12, 1943.

Jie Jack He and Xuan Tian. The dark side of analyst coverage: The case of innovation. *Journal of Financial Economics*, 109(3):856–878, 2013. doi: 10.1016/j.jfineco.2013.04.001.

James J. Heckman and Rodrigo Pinto. Causal Analysis after Haavelmo. *Econometric Theory*, 31:115–151, 2013.

James J. Heckman and Sergio Urzua. Comparing IV with Structural Models: What Simple IV Can And Cannot Identify. NBER Working Paper 14706, 2009.

James J. Heckman and Edward J. Vytlacil. Econometric evaluation of social programs, part 1: Causal models, structural models and econometric policy evaluation. In *Hanbook of Econometrics*, volume 6B. Elsevier B.V., 2007.

James J. Heckman, Hidehiko Ichimura, and Petra Todd. Matching as an econometric evaluation estimator. *The Review of Economic Studies*, 65(2):261–294, 4 1998. doi: 10.1111/1467-937X.00044.

David A. Hoffman and Anton Strezhnev. Longer trips to court cause evictions. *Proceedings of the National Academy of Sciences of the United States of America*, 120(2):e2210467120, 2023. doi: doi.org/10.1073/pnas.221046712.

References III

Yi-Hou Huang, Woan lih Liang, Quang-Thai Truong, and Yanzhi Wang. No new tricks for old dogs? old directors and innovation performance. *Technological Forecasting & Social Change*, 179:121659, 2022. doi: 10.1016/j.techfore.2022.121659.

Paul Hünermund, Beyers Louw, and Itamar Caspi. Double machine learning and automated confounder selection — a cautionary tale. *Journal of Causal Inference*, 11(1):20220078, 2023. doi: 10.1515/jci-2022-0078.

Guido W. Imbens. Instrumental variables: An econometrician's perspective. *Statistical Science*, 29(3):323–358, 2014.

Guido W. Imbens and Donal B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge University Press, 2015.

Rosa L. Matzkin. Nonparametric identification. In *Handbook of Econometrics*, volume 6B, 2007.

Rosa L. Matzkin. Nonparametric identification in structural economic models. *Annual Review of Economics*, 5:457–486, 2013.

Judea Pearl, Madelyn Glymour, and Nicholas P. Jewell. *Causal Inference in Statistics: A Primer*. John Wiley & Sons Ltd, West Sussex, United Kingdom, 2016.

Mark R. Rosenzweig and Kenneth I. Wolpin. Natural ”Natural Experiments” in Economics. *Journal of Economic Literature*, 38:827–874, December 2000.

D. B. Rubin. Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology*, 66:688–701, 1974.

References IV

Paul E. Spector and Michael T. Brannick. Methodological urban legends: The misuse of statistical control variables. *Organizational Research Methods*, 14(2):287–305, 2011. doi: 10.1177/1094428110369842.

Robert H. Strotz and Herman O. A. Wold. Recursive vs. nonrecursive systems: An attempt at synthesis (part i of a triptych on causal chain systems). *Econometrica*, 28(2):417–427, 1960.

Robert I. Sutton and Barry M. Staw. What theory is not. *Administrative Science Quarterly*, 40(3):371, 1995. doi: 10.2307/2393788.

Johannes Textor and Maciej Liśkiewicz. Adjustment criteria in causal diagrams: An algorithmic perspective. In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence*, pages 681–688. AUAI press, 2011. doi: 10.5555/3020548.3020627.

Herman Wold. Causality and econometrics. *Econometrica*, 22(2):162–177, 1954.

Herman O. A. Wold. A generalization of causal chain models (part iii of a triptych on causal chain systems). *Econometrica*, 28(2):443–463, 1960.

James Woodward. *Making Things Happen*. Oxford Studies in Philosophy of Science. Oxford University Press, 2003.

Structural Econometrics vs. Potential Outcomes

- Econometrics is currently dominated by two competing streams
- Structural econometrics
    - In practice, relies on distributional assumptions and (parametric) shape restrictions
    - Work by, e.g., Matzkin (2007) that aims to relax parametric assumptions, but
        - still relies on (weaker) shape restrictions, and is not widely adopted in applied work
- Potential outcomes framework (Rubin, 1974; Imbens and Rubin, 2015)
    - Does impose crucial identifying assumptions (e.g., ignorability) without reference to an underlying model ("black box character")
        - A feature that has been frequently criticized by the structural camp (e.g., by Rosenzweig and Wolpin, 2000 and Heckman and Urzua, 2009)
    - In practice, causal inference in PO boils down to the four "tricks of the trade" (matching, IV, RDD, difference-in-differences)

⇒ DAGs are a perfect "middle ground" between structural econometrics and PO

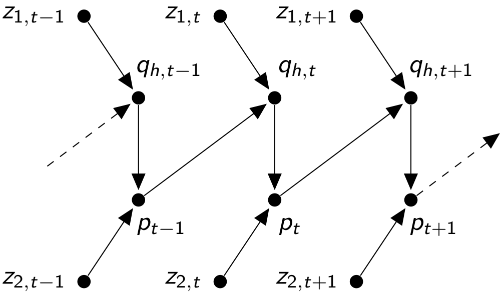
Back

Recursive versus Interdependent Systems

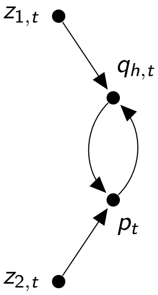
- DAGs represent recursive systems, but many standard models in economics are interdependent (Marshallian cross, game theory, etc.)
- This connects to an old debate within econometrics about the causal interpretation of recursive versus interdependent models that emerged in the aftermath of Haavelmo's celebrated 1943 paper
- One central argument (Bentzel and Hansen, 1954 - 1955; Strotz and Wold, 1960):
  - Individual equations in an interdependent model do not have a causal interpretation *in the sense of a stimulus-response relationship* (Strotz and Wold, 1960, p. 417)
  - Interdependent systems with equilibrium conditions are regarded as a *shortcut* (Wold, 1960; Imbens, 2014) description of the underlying dynamic behavioral processes

Recursive versus Interdependent Systems

- In this context, Strotz and Wold (1960) discuss the example of the cobweb model:



$$p_{t-1} = p_t$$
$$\Rightarrow$$



$$\begin{aligned} q_{h,t} &\leftarrow \gamma + \delta p_{t-1} + \nu z_{1,t} + u_{1,t}, \\ p_t &\leftarrow \alpha - \beta q_{h,t} + \varepsilon z_{2,t} + u_{2,t}. \end{aligned}$$

$$\begin{aligned} q_{h,t} &\leftarrow \gamma + \delta p_t + \nu z_{1,t} + u_{1,t} \\ p_t &\leftarrow \alpha - \beta q_{h,t} + \varepsilon z_{2,t} + u_{2,t} \end{aligned}$$

Recursive versus Interdependent Systems

- However, equilibrium assumption $p_{t-1} = p_t$ carries no behavioral interpretation
- Individual equations in interdependent system do not represent autonomous causal relationships in the stimulus-response sense (Heckman and Pinto, 2013)
  - Endogenous variables are determined jointly by all equations in the system (Matzkin, 2013)
  - Not possible, e.g., to directly manipulate p_t to bring about a desired change in $q_{h,t}$
- Equilibrium models can of course still be useful for learning about causal parameters
- But, if individual mechanisms are supposed to be interpreted as stimulus-response relationships, cyclic patterns need to be excluded (Woodward, 2003; Cartwright, 2007)
  - For this reason, potential outcomes framework (Rubin, 1974; Imbens and Rubin, 2015) also implicitly maintains the assumption of acyclicity (Heckman and Vytlacil, 2007)

Back