Applied Causal Inference Powered by ML and AI

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

February 5, 2026

Publisher: Online
Version 0.1.2

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

Causal Inference via Linear Structural Equations

6

"the scientific [. . .] problem of causality is essentially a problem regarding our way of thinking, not a problem regarding the nature of the exterior world."

– Ragnar Frisch [1].

Here we present the linear structural equation model framework and causal diagrams. The advantage of these models is they are closely related to underlying structural models commonly used in economics and other fields. They allow for transparent derivation of the conditional ignorability assumption from the structure of the model. While linearity is imposed in this chapter, it will be dispensed with in later chapters.

6.1 Structural Equation Modelling and Conditional Exogeneity

Basic ideas that appeared in econometrics between the 20s and 40s (P. Wright [2], S. Wright [3], J. Tinbergen [4], T. Haavelmo [5]) provide another take on and language for causality that is closely related to the potential outcomes framework.

A Simple Triangular Structural Equation Model (TSEM)

We illustrate the basic ideas using a simple model of a household's (say weekly) demand for gasoline, motivated by Hausman and Newey [6].

We start with a log-linear (Cobb-Douglas [7]) model for log-demand y given the log-price p

$$y(p) := \delta p,$$

where δ is the elasticity of demand. Demand is random across households, and we may model this randomness as

$$Y(p) := \delta p + U, \quad \mathrm{E}[U] = 0, \tag{6.1.1}$$

where U is a stochastic shock that describes variation of demand across households (or across time, but assume that we are just looking at a particular time point). We immediately recognize that $Y(p)$ plays the same role as a potential outcome in Rubin's potential outcome model.[1]

The stochastic function

$$p \mapsto Y(p)$$

describes a household's log-demand at a given log-price p. The expected log-demand at log-price p is given by $\mathrm{E}[Y(p)] = \delta p$. The function encodes various structural causal effects: If we change p from p_0 to p_1, the expected demand change would be

$$\mathrm{E}[Y(p_1)] - \mathrm{E}[Y(p_0)] = \delta(p_1 - p_0).$$

Model (6.1.1) is very simple, and we may want to introduce covariates to capture other observable factors that may be associated with demand. That is, we may think there are observable parts of the stochastic shock, characterized by X, which help us predict household demand. Leading examples are household

1: The subtle difference here is that U does not depend on the index p, though we could make U be indexed by p at the cost of more complicated exposition. The distinction drawn is not superficial. Later on, when we discuss models with instruments, the dependence of U on p can create non-trivial problems which are not present in this section.

characteristics. For example, we may think demand is associated with features such as family size, income, number of cars, or geographical location. We can incorporate these features by modelling $U = X'\beta + \epsilon_Y$, where ϵ_Y is independent of X and has mean zero. Employing this model structure, we can write our augmented model as

$$Y(p) := \delta p + X'\beta + \epsilon_Y, \quad \epsilon_Y \perp\!\!\!\perp X. \tag{6.1.2}$$

Equation (6.1.2) is a structural stochastic model of economic outcomes. This model has nothing to do with regression or a statistical predictive model. Rather, it is a model that provides counterfactual predictions: If log-price is set to p, then a household with characteristics X can be predicted to purchase

$$\delta p + X'\beta$$

log-units. Here p is not a random variable – it is an index describing potential values of the price.

Then we ask the question:

- What data (Y, P, X) on quantities, prices, and characteristics should we collect to allow us to estimate the structural parameter δ?

Assumption 6.1.1 (Conditional Exogeneity) *(i) (Consistency) Suppose the observed variables* (Y, P, X) *are such that*

$$Y = Y(P)$$

i.e. the outcome is generated from the structural model, (ii) (Conditional Exogeneity) The observed P *is determined outside of the model, independently of* ϵ_Y *conditional on* X*:*

$$P \perp\!\!\!\perp \epsilon_Y \mid X \implies P \perp\!\!\!\perp \{Y(p)\}_{p\in\mathbb{R}} \mid X$$

Assumption 6.1.1 is the econometric analog of ignorability.[2] In the context of household demand, this condition requires that P is determined independently of a household's demand shock ϵ_Y, conditional on characteristics X. This assumption seems plausible for household level decisions, especially if we include geography in the set of covariates X.

2: At a general level, gasoline prices are determined by aggregate supply and demand conditions, with small local geographic adjustments (e.g., gasoline prices in areas with higher prices of land may be higher than in other areas to reflect the higher land costs for gasoline stations). Conditional on being in a given small geographic region, we may think of price fluctuations as independent of household-specific demand shocks.

If the conditional exogeneity condition holds, then

$$Y = Y(P) = \delta P + X'\beta + \epsilon_Y, \quad \epsilon_Y \perp (P, X).$$

This means that the projection parameters of Y on P and X coincide with the structural parameters δ and β.

We stress that our parameters δ and β are not defined by regression; they are defined by the model. Under the conditional exogeneity condition, these parameters coincide with the projection parameters.[3]

We might further postulate a structural equation for log-prices:

$$P(x) := x'\nu + \epsilon_P,$$

where $P(x)$ is the stochastic price process indexed by a household characteristics and ϵ_P describes the centered stochastic price shock. We assume that observed X is independent of price shock ϵ_P,

$$X \perp\!\!\!\perp \epsilon_P.$$

Independence between ϵ_P and observed X implies that ν coincides with the projection coefficient of P on X.

The price process $P(x)$ captures the belief that prices faced by households may differ depending on household characteristics. Note that this notation allows for only a subset of household characteristics to be systematically related to price; that is, we can have $P(x) = P(x_1)$ for some subvector x_1 of x. For example, it seems reasonable that households located in different regions would experience different prices, in which case x_1 could represent a household's geographic characteristics. Independence of the price shock ϵ_P from observed X may be plausible if household characteristics are determined well before gasoline prices faced by individual households in any specific time period are set.

Putting the equations together, we have a triangular structural equation model (TSEM):

$$\begin{aligned} Y &:= \delta P + X'\beta + \epsilon_Y, \\ P &:= X'\nu + \epsilon_P, \\ X&, \end{aligned} \tag{6.1.3}$$

where ϵ_Y, ϵ_P, and X are mutually independent (or at least uncorrelated) and determined outside of the model. They

3: A weaker starting condition than the conditional exogeneity condition for the above result is simply

$$(P, X) \perp \epsilon_Y.$$

That is, the observed P and X are orthogonal to the structural error ϵ_Y.

are called exogenous variables. Y and P are determined within the model and called the endogenous variables. The structural parameter δ can be identified by linear regression provided $\text{Var}(\epsilon_P) > 0$, and the structural parameter ν can be identified by linear regression provided $\text{Var}(X) > 0$.

Under the conditions stated above the parameters of these structural equations coincide with the projection parameters.

What do we mean by the model being structural? The term structural means that each of the equations is *assumed* to provide comparative statics and answers to counterfactual questions. Setting the right-hand-side variables to their potential values, we have

$$Y(p, x) := \delta p + x'\beta + \epsilon_Y,$$
$$P(x) := x'\nu + \epsilon_P.$$

The conceptual operation of "setting" or "fixing" the variables is supposed to leave the structure invariant. More generally, the structural parameters are supposed to be invariant to changes in the distribution of exogenous variables – X, ϵ_Y, ϵ_P – that have been generated outside of the model. Therefore, we can use these structural parameters to generate counterfactual predictions.

The jargon *comparative statics* refers to the determination of how endogenous variables change in response to changes in exogenous variables. Similarly, *counterfactual questions* coincide with asking how outcomes or endogenous variables change when variables are set to new values with other features of the model remaining fixed; e.g. asking how demand changes when price is set to some new value by a firm with household characteristics, price shocks, and demand shocks unaffected.

6.2 Drawing the Model: Causal Diagrams, aka DAGs

Sewall and Philip Wright [2], [3] would have depicted system of equations (6.1.3) graphically as a causal (path) diagram as in Figure 6.1. Observed variables are shown as nodes, causal paths are shown by directed arrows, and the structural (causal) parameters are given by the symbols placed next to the arrows.

The graph represents a structural economic model that can answer causal (comparative statics) questions. For example, the elasticity parameter δ tells us how household demand will respond to a firm *setting* a new price. Note that a firm setting a new price will not alter household characteristics or the other exogenous features of the model, and thus only the parameter δ is relevant for answering this question within the model.

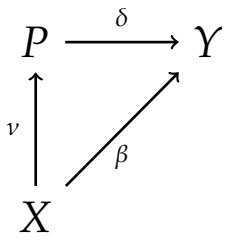


Figure 6.1: A simple causal diagram representation of the TSEM for the household gasoline demand example.

We could have expanded the previous graph to include unobserved shocks ϵ_P and ϵ_Y as follows:

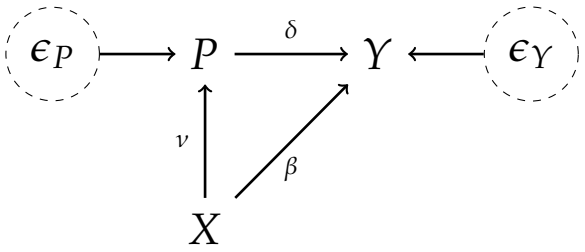


Figure 6.2: An expanded causal diagram representation of the TSEM that shows the unobserved shocks ϵ_P and ϵ_Y as root nodes.

The graph initiates with the *root nodes* ϵ_P, X, and ϵ_Y. The absence of links between the root nodes signifies the orthogonality between the nodes: namely, the absence of correlation. Understanding the orthogonality structure between nodes is an important input into identification of structural parameters via projection. The nodes X and ϵ_P are *parents* of P; the nodes P, X, and ϵ_Y are *parents* of Y. The node Y is a *collider* on all paths, because it contains only incoming arrows.

The main effect of interest is δ, which we call the structural causal effect of P on Y. This effect is identified after adjusting for X. In terms of the graph above, there are two paths connecting P and Y:

$$P \to Y \text{ and } P \leftarrow X \to Y.$$

The second path is called a *backdoor path* because there is an arrow pointing back to P from X. This connection indicates that there is a common cause for P and Y. Figuratively speaking, controlling or adjusting for X is said to be like "closing the backdoor path," shutting down the non-causal sources of statistical dependence between Y and P.

This visual characterization of the adjustment for X is due to J. Pearl [8] and generalizes to much more complicated graphs. We revisit these ideas throughout subsequent chapters.

How do household characteristics impact our model? X affects Y through two paths:

- the direct effect β via $X \to Y$,
- and the indirect effect $v\delta$ via $X \to P \to Y$.

The indirect effect is said to be "mediated" by P. We saw in Section 6.1 that we can identify δ and β from projection of Y on P and X, and we can identify ν by projection of P on X. Therefore both the direct and indirect effects are identified.

Mediation structures appeared right at the outset in the Wrights' work [2], [3].

The total effect of X on Y is

$$\nu\delta + \beta,$$

which can be identified in this case by projection of Y on X. To verify this, we plug the first equation from the TSEM in (6.1.3) into the second equation producing

$$Y = (\nu\delta + \beta)'X + V; \quad V = \epsilon_Y + \delta\epsilon_P.$$

We see that the composite disturbance V is orthogonal to X,

$$V \perp X,$$

and, therefore, $(\nu\delta + \beta)$ coincides with the projection coefficient in the projection of Y on X. The latter point can be seen graphically: There are no "backdoor" paths from X to Y, so it is not necessary to adjust or control for anything to identify the total effect of X on Y.

In fact, while conditioning on P would allow us to identify the direct effect of X, β, it would prevent us from retrieving the total effect $\nu\delta + \beta$. In empirical practice, we may think of conditioning on P as "conditioning on the outcome," as P is determined by its parents, including X, so may be thought of as an outcome relative to X.

Remark 6.2.1 (Statistical Identification) Statistical identification typically relies on a combination of orthogonality or conditional independence restrictions and additional conditions – referred to as "rank conditions" in some settings – that ensure there is variation available for learning parameters of interest. For example, we need that $\text{Var}(\epsilon_P) > 0$ if we wish to learn δ in the TSEM in (6.1.3), and we need overlap for learning ATE as discussed in Chapter 5. Graphical methods provide a tool for representing orthogonality and conditional

independence relationships. They typically do not immediately reveal the additional rank-type conditions one would use in establishing statistical point identification. Examining the graphical structure does reveal what causal effects are potentially learnable within the structure, and additional restrictions, such as $\text{Var}(\epsilon_P) > 0$ in the TSEM, can then be deduced. Throughout the remainder of this book, we abstract away from rank-type conditions when discussing graphical models and talk about identifying parameters from the implied orthogonality or conditional independence structure.

To summarize, to learn a causal parameter, we must first define the causal parameter of interest and then carefully consider the choice of what to condition on to learn this effect. These choices are particularly important given the existence of *collider bias*.

The Notebooks 6.6.1 provide a simple simulated example of collider bias based on the SEM (6.3.1).

6.3 When Conditioning Can Go Wrong: Collider Bias, aka Heckman Selection Bias

Consider the following SEM:

$$\begin{aligned} T &:= \epsilon_T \\ B &:= \epsilon_B \\ C &:= T + B + \epsilon_C \end{aligned} \tag{6.3.1}$$

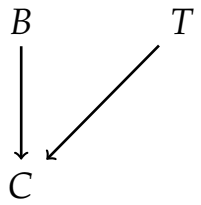


Figure 6.3: DAG with a collider representing SEM (6.3.1).

where ϵ_T, ϵ_B, and ϵ_C are independent $N(0,1)$ shocks. Here the average structural function for T, which does not depend on what values B might take, is zero,

$$\text{E}[T] = 0.$$

Regression without conditioning on C correctly identifies that T is not causally impacted by B:

$$\text{E}[T \mid B = b] = 0.$$

However, further conditioning on C removes the causal interpretation of the projection coefficient:[4]

$$\text{E}[T \mid B, C] = (C - B)/2; \implies \text{E}[\text{E}[T \mid B = b, C]] = -b/2 < 0.$$

This regression suggests that, controlling for C, the predictive effect of B on T is $-1/2$. This predictive effect is not a causal effect.

4: Dividing by 2 may seem counterintuitive, but it is correct. See the Collider Bias Notebooks 6.6.1 for detail.

Collider bias illustrates that conditioning on outcomes may produce the wrong conclusions about causality, so conditioning on outcomes should be always approached with care. In econometrics, collider bias is known as a form of sample selection bias[5] ("conditioning on endogenous variables" or Heckman selection bias [9]).

5: J. Heckman was awarded the Nobel Memorial prize "for his development of theory and methods for analyzing selective samples." Source: Nobelprize.org

A Serious Digression on Colliders. Within our toy SEM framework, regression on a collider is clearly the wrong thing to do if one wants to identify the causal effect of B on T. However, we do note that regression on a collider can be *very useful* for other predictive tasks.

The following example draws on the discussion given in the "Book of Why" [10] to illustrate collider bias.

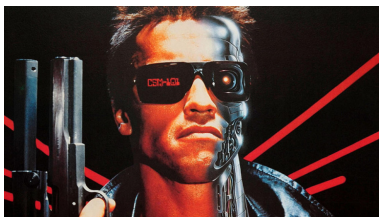


Figure 6.4: Our SEM predicts that this actor, A. Terminator, is (essentially) the most talented actor in Hollywood.

Example 6.3.1 (Structural Model of Hollywood) Suppose that the preceding SEM provides a cartoon depiction of people in Hollywood where T denotes acting talent, C denotes celebrity (i.e. success or popularity), and B denotes bonhomie (i.e. approachability or friendliness). Note that the SEM indicates that more talent and approachability cause more success. Further, for a person to remain in Hollywood, we would expect $C > 0$. As shown above, the causal effect of B on T in this SEM is 0. However, the best linear predictor of T given B conditional on $C > 0$ is

$$\approx .6 - B/4.$$

That is, bonhomie and talent are negatively correlated in Hollywood despite the fact that approachability does not causally impact talent. This correlation is useful for making predictions. For example, the individual depicted in the margin appears quite imposing and not approachable, perhaps with $B = -20$. We would then predict the expected value of his talent to be $t \in [+5.6 \pm 2]$, which is at least 3.6 standard deviations above the average talent of zero in the overall population within our model. From that, we should *predict* that this person is an incredibly talented actor but should not draw any conclusions about causality between B and T.

The example illustrates how simple theoretical models are often used in economics. Causal reasoning is made within a simple model, such as the SEM (6.3.1). This reasoning then leads to some testable restrictions, such as negative correlation between T and B conditional on $C > 0$. Even though we may not believe that the stylized model provides a complete model of reality, the

implications of the simple model provide some insight into how observed phenomena, such as a negative correlation between T and B conditional on $C > 0$, may arise. Such reversion of the correlation between two variables has been observed empirically in several cases, a prominent one being the birth-weight paradox [11] described below.

Example 6.3.2 (Birth-weight "paradox" [11]) In a study conducted in 1991 in the US, it was found that infants born to smokers had higher risk of low birth-weight (LBW) and higher risk of infant mortality than infants born to non-smokers. However, when looking at the sub-group of infants with LBW, the comparison is reversed and the risk of infant mortality is lower for infants born to smokers, than for infants born to non-smokers. How is that possible? Does smoking have a positive causal effect on infant mortality conditional on LBW?

A more plausible alternative explanation can be uncovered through the lens of SEMs and Causal Diagrams if one starts to think of competing risks and collider bias. Let's denote with S the smoking indicator, Y the infant death outcome, and B the low birth-weight indicator. We will also denote with U an abstract variable corresponding to the multitude of competing risks that can cause LBW. It is highly plausible that smoking is a risk factor for LBW and also has a direct effect on mortality. Moreover, LBW and the competing risk factors can also have a direct effect on mortality. Putting these factors together leads to the Causal Diagram depicted in Figure 6.5. In this setting, an infant with a smoking parent may be highly likely to have LBW caused by smoking. At the same time, LBW can be much less frequent for non-smoking parents. When we further focus in on the group of infants of non-smoking parents with LBW, it is highly probable that LBW was caused by some other competing risk which can adversely affect mortality. Thus, conditioning on LBW, we could essentially be comparing infants of smoking parents without competing risks to infants of non-smoking parents with competing risks.

To illustrate how the unconditional association between Y and S uncovers the true causal effect, while conditioning on B introduces bias and can even reverse the sign of the true effect, let's look at a simple linear SEM that corresponds to

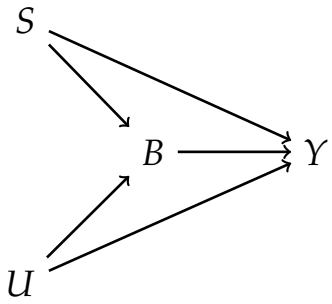


Figure 6.5: DAG with a collider representing low birth-weight "paradox" Example 6.3.2.

the causal diagram depicted in Figure 6.5:

$$\begin{aligned} Y &:= S + B + \kappa U + \epsilon_Y \\ B &:= S + U + \epsilon_B \\ S &:= \epsilon_S \\ U &:= \epsilon_U \end{aligned} \tag{6.3.2}$$

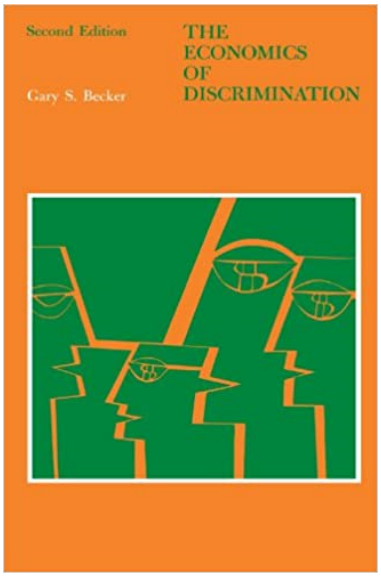
where $\epsilon_Y, \epsilon_B, \epsilon_S$ and ϵ_U are independent $N(0,1)$ shocks. Note that if we simply project Y on S, then we recover the correct positive causal effect of 2, since conditional exogeneity is satisfied. However, when we project Y on S and B, we learn a CEF of the form:

$$\begin{aligned} \mathrm{E}[Y \mid S, B] &= S + B + \kappa \mathrm{E}[U \mid S, B] \\ &= S + B + \kappa(B - S)/2 = (1 - \kappa/2)S + (1 + \kappa/2)B. \end{aligned}$$

If the competing risks increase infant mortality a lot, i.e. $\kappa \gg 1$, then this projection recovers an erroneous large negative(!) effect $1 - \kappa/2$ of smoking on mortality.

6.4 Wage Gap Analysis and Discrimination

> "The central question in any employment-discrimination case is whether the employer would have taken the same action had the employee been of a different race (age, sex, religion, national origin etc.) and everything else had remained the same." (In Carson versus Bethlehem Steel Corp., 70 FEP Cases 921, 7th Cir. (1996) [12]).



Wage regressions are widely used by labor economists to characterize the wage gap between men and women and to link the wage gap to discrimination; see, e.g., [13] and [14]. Some economists have asserted that it is wrong to study discrimination by doing wage gap regressions, e.g. [15], and that we should instead look at the unconditional difference in outcomes across groups. Their reasoning is based on the argument that key job characteristics – e.g., education and occupation – are determined in response to both a group identity and discrimination and are therefore (intermediate) outcomes. Controlling for these characteristics may then introduce a form of selection bias. Which of these two sets of economists is right?

In what follows, we present a simple SEM in (6.4.1), which postulates that different groups receive equal wages if there

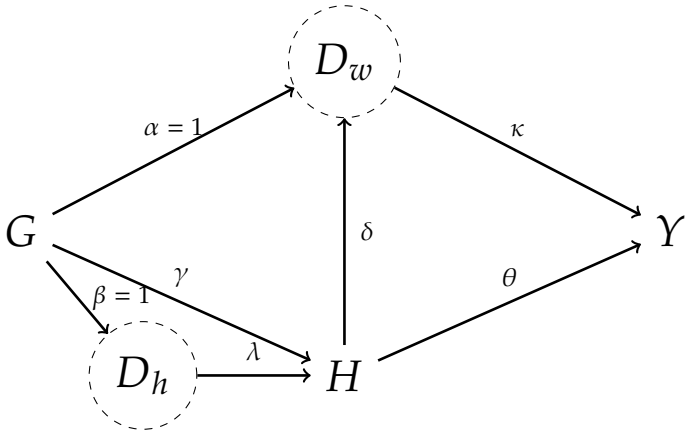


Figure 6.6: A Simple Model of Discrimination. Here G denotes a group (e.g., sex), H is human capital, and Y is the wage. D_w denotes unobserved wage discrimination occuring in the work place, and D_h denotes unobserved discrimination that occurs in the accumulation of human capital.

are no conditional productivity differences between the groups. We will see that, in this SEM, wage gap regressions do uncover well-defined discrimination effects that occur in wage-setting mechanisms. In contrast, the unconditonal average wage gap uncovers a more complicated causal object, which absorbs discrimination in wage setting, discrimination in human capital and occupational acquisitions, as well as group specific preferences for occupations.

Here we begin with the linear SEM and the equivalent DAG shown in Figure 6.6:

$$\begin{aligned} Y &:= \kappa D_w + \theta H + \epsilon_Y, \\ D_w &:= \alpha G + \delta H + \epsilon_{D_w}, \\ H &:= \gamma G + \lambda D_h + \epsilon_H, \\ D_h &:= \beta G + \epsilon_{D_h}, \\ G, & \end{aligned} \tag{6.4.1}$$

where the shocks ϵ_Y, ϵ_{D_w}, ϵ_H, ϵ_{D_h}, and G are all mean zero and uncorrelated.

The outcome Y is wage, G is group (e.g., sex), H is human capital (a scalar index that includes labor-relevant characteristics such as education, occupation, etc.),[6] D_w is latent wage discrimination arising in the work-place, and D_h is latent discrimination arising in acquisition of human capital. There could be other observed confounders that we don't show for the sake of simplicity.

6: H can be easily made a vector with a slightly more complicated notation.

The discrimination variables D_w and D_h are latent variables that are important for our model but cannot be directly observed. We maintain throughout that these variables are non-degenerate and related to group identity G. Under these assumptions,

the scale of these latent variables is non-zero but arbitrary, so we normalize the effect $G \to D_w$ to unity, $\alpha = 1$, and the effect $G \to D_h$ to unity as well, $\beta = 1$. There is no edge from G to Y, reflecting our assumption that there is no systematic group difference in productivity conditional on H and D_w. In the absence of productivity differences between workers, economic reasoning suggests that they would be assigned the same wage in a discrimination-free economy [16]. Thus, we would expect $\kappa = 0$ in a discrimination-free economy in the case that H captures all sources of productivity differences between workers.

Within this model, the parameter of interest is then the causal or structural effect of discrimination on wages given by

$$\kappa.$$

If $\kappa \neq 0$, we can conclude that wages are assigned unfairly within the framework of this SEM.

If we observed D_w directly, we could learn the effect of discrimination on wages, κ, by regression of Y on D_w and H. Identification of κ from this regression follows from the back-door criterion discussed in Section 6.2. We don't observe D_w directly, but we postulate that this variable is determined only by G, H, and a stochastic shock. Dependence on H captures the idea that discrimination may be larger or smaller depending on education level, profession, etc. We return to using this additional structure to learn about κ below.

Discrimination may operate through channels other than simple wage differences. For example, in the 1960s, there were relatively few women or African American lawyers, a highly paid occupation. Discrimination that operates through occupational choice or human capital formation is captured by latent variable D_h. In our model, H, which captures productivity differences between individuals, can be determined as a result of both discrimination and group preferences.[7] The parameter γ then captures the effect of group preferences on the formation of H, while the effect of discrimination on H is captured by λ. Since D_h is not observed, there is no way to separately identify these two effects.

7: For example, 90% of firefighters in the US are men, which may reflect a genuine preference for this occupation among men. At the same time, even preference for occupation may be a result of cultural institutions that could themselves be interpreted as discriminatory in broader, cross-cultural, contexts.

It is easy to show, within the model, that the population linear regression of Y on G and H recovers the wage dis-

crimination effect,

$$\kappa,$$

and that the linear regression of H on G recovers

$$\gamma + \lambda,$$

the sum of the group preference effect and the human capital discrimination effect; see Appendix 6.A for details. If a further strong assumption is made that there is no group preference effect, $\gamma = 0$, the linear regression of Y on G recovers the total discrimination effect:

$$\kappa + \lambda(\kappa\delta + \theta).$$

Endogenous Sample Selection. There is an important issue with our empirical example. We are only able to look at earnings of people who are employed. Thus, we are conditioning on

$$Y > R,$$

where R is the reservation wage. In other words, we are conditioning on the outcome which may cause major selectivity issues: People get employed, and end up in our data, only if the offered wage is higher than some reservation wage. This sample selection on the basis of the outcome can cause major biases in the analysis. The potential for large biases was recognized by James J. Heckman [9] in the 70s and led to the development of the celebrated Heckman selection correction and related methods.

An alternate approach to applying a selection correction in our example is to select a subset S of people who are employed with probability one (or very close to one). For example, one could look at highly educated, unmarried people. Within this subset, we would then have

$$P(Y > R|S) \approx 1.$$

That is, the value of the wage offer, Y, is approximately unrelated to whether we observe individual wages for this subset of people. This type of strategy has been employed by Casey Mulligan and Yona Rubinstein [17]. Mulligan and Rubinstein continue to find evidence in favor of the existence of wage gaps in their analysis of a subsample where selection effects are likely small. This finding then

suggests that the broad conclusion of the existence of wage gaps is not driven entirely by sample selection issues.

In summary, we have the following observations:

- In general, wage gap regressions just estimate predictive effects or associations.
- When we assume a SEM like the one above holds and there are no endogenous sample selection effects, wage gap regressions estimate wage discrimination effects.
- Unconditional wage gaps generally reflect a combination of different types of discrimination and group preferences and thus do not isolate solely the effects of discrimination.

Figure 6.7: Early 20th century: The work of Sewall and Philip Wright made it possible for humans to begin to "fly" in the space of causal models. Another family of Wrights made it possible for humans to begin to fly in the air.

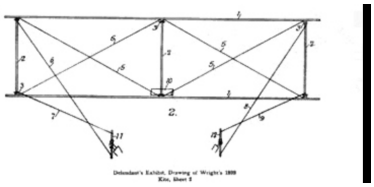


Figure 6.8: An early drawing for an airplane appears very much like an early drawing of a DAG.

6.5 Notes

This chapter presented an approach to causal inference that goes back to the works of Sewall and Philip Wright [2], [3], Tinbergen [4], Haavelmo [5], and others. This tradition lives in modern structural causal models used in econometrics (especially, industrial organization) and in the artificial intelligence community. The latter community, inspired by the foundational work of J. Pearl [8], strongly adopted the use of causal diagrams, known as directed acyclical graphs (DAGs). We continue exploring this approach throughout the remainder of our treatment on causal inference.

6.6 Notebooks

Notebook 6.6.1 (Collider Bias) Collider Bias R Notebook and Collider Bias Python Notebook provide a simple simulated example of collider bias, informing our discussion of conditioning on Celebrity in our Structural Model of Hollywood.

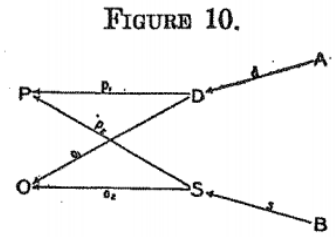


Figure 6.9: DAG for Supply-Demand Systems in P. Wright's work in 1928 [2].

6.7 Exercise

Exercise 6.7.1 (Collider Bias) Explain collider bias to a friend in simple terms. Use no more than two paragraphs. Illustrate

your explanation using a simulation experiment.

Exercise 6.7.2 (Wage Gap Revisited) Empirical: Revisit the group wage gap analysis from Chapter 4, focusing on college-educated workers. Is there a structural/causal interpretation for the estimated wage gap? Is there a group gap in education achievement? Does this group gap in education have a structural/causal interpretation? Some of these questions are open ended and have no simple answers, but it is useful to think about them. (If you have other data sets that might illuminate discrimination in other settings, please use them in place of the wage data set).

Exercise 6.7.3 (Mechanisms for Wage Gap) Free-style exercise: The model for wage discrimination presented in our notes is very stylized and subject to multiple criticisms. For example, it does not deal with promotion and hiring decisions. There are several interesting models of discrimination in hiring, college admissions, and pay. For example, see "The Book of Why"[10] and the Bickel et al. 1975 paper [18] for an analysis of Berkeley undergraduate admissions decisions. Nina Roussile's (2020) [19] paper isolates the ask gap as the central mechanism for the subsequent wage gap. Referring to one such analysis, draw or write down a linear structural causal model that captures the structural idea of the analysis and discuss identification in the model.

6.A Details of the Wage Discrimination Analysis

We write out some of the structural equations corresponding to our stylized DAG for discrimination (Figure 6.6):

$$\begin{aligned} Y &:= \kappa D_w + \theta H + \epsilon_Y, \quad \epsilon_Y \perp D_w, H, G \\ D_w &:= G + \delta H + \epsilon_{D_w}, \quad \epsilon_{D_w} \perp G, H \end{aligned}$$

where the orthogonality relations are implied by the model.

Linear regression analysis would use observable variables only, so we substitute the model for the unobserved D_w in terms of G and H into the equation for Y to obtain

$$Y = \kappa G + (\kappa\delta + \theta)H + U, \quad U := \kappa\epsilon_{D_w} + \epsilon_Y \perp (G, H).$$

The composite error term U is orthogonal to G and H. Therefore, regression of Y on G and H learns κ and $(\kappa\delta + \theta)$, with our main target being κ. We can also see that by partialling out H,

"This is elementary, my dear Watson," said Sherlock Holmes after seeing this.

$$\tilde{Y} = \kappa\tilde{G} + U, \;\; U \perp \tilde{G}.$$

Thus, κ is retrievable only if there is non-zero variation in $\tilde{G}$ after taking out the linear effect of H.

Now suppose we want to study discrimination effects in occupational choices, captured by H in our model. We write out the relevant structural equations:

$$\begin{aligned} H &:= \gamma G + \lambda D_h + \epsilon_H, \quad \epsilon_H \perp (G, D_h), \\ D_h &:= G + \epsilon_{D_h}, \qquad \epsilon_{D_h} \perp G. \end{aligned}$$

Recall that γ is the group preference effect and λ is the discrimination effect. Since D_h is not directly observed, we substitute it out to arrive at

$$H = (\gamma + \lambda)G + V; \quad V := \gamma\epsilon_{D_h} + \epsilon_H \perp G.$$

Therefore, $\gamma + \lambda$ is the projection coefficient in the projection of H on G. Hence, we can identify $\gamma + \lambda$, but we can't identify γ and λ separately.

Going further, suppose that the group preference effect is zero, so $\gamma = 0$. Then, the previous argument would identify λ and we could identify the total discrimination effect arising from two different channels:

$$\kappa + \lambda(\kappa\delta + \theta).$$

from the regression of Y on G.

We can assert that the unconditional difference in wages measures discrimination only if the group preference effect in determining H is zero ($\gamma = 0$). Of course, most economists would probably not agree with the assumption that $\gamma = 0$. Empirically, there are large differences in group composition among different professions. These differences likely reflect both discrimination and genuine preferences.

Bibliography

[1] R. Frisch. 'A Dynamic Approach to Economic Theory: Lectures by Ragnar Frisch at Yale University'. Frisch Archives, Department of Economics, University of Oslo. 1930 (cited on page 143).

[2] Philip G. Wright. *The Tariff on Animal and Vegetable Oils*. New York: The Macmillan company, 1928 (cited on pages 144, 147, 149, 157).

[3] Sewall Wright. 'Correlation and Causation'. In: *Journal of Agricultural Research* 20.7 (Jan. 1921), pp. 557–585 (cited on pages 144, 147, 149, 157).

[4] Jan Tinbergen. 'Bestimmung und Deutung von Angebotskurven Ein Beispiel'. In: *Zeitschrift für Nationalökonomie* 1.5 (1930), pp. 669–679 (cited on pages 144, 157).

[5] Trygve Haavelmo. 'The probability approach in econometrics'. In: *Econometrica* 12 (1944), pp. iii–vi+1–115 (cited on pages 144, 157).

[6] Jerry A. Hausman and Whitney K. Newey. 'Nonparametric estimation of exact consumers surplus and deadweight loss'. In: *Econometrica* 63.6 (1995), pp. 1445–1476 (cited on page 144).

[7] Charles W. Cobb and Paul H. Douglas. 'A Theory of Production'. In: *The American Economic Review* 18.1 (1928), pp. 139–165 (cited on page 144).

[8] Judea Pearl. *Causality*. Cambridge University Press, 2009 (cited on pages 148, 157).

[9] James J. Heckman. 'Sample selection bias as a specification error'. In: *Econometrica* 47.1 (1979), pp. 153–161 (cited on pages 151, 156).

[10] Judea Pearl and Dana Mackenzie. *The Book of Why*. Penguin Books, 2019 (cited on pages 151, 158).

[11] Sonia Hernández-Díaz, Enrique F Schisterman, and Miguel A Hernán. 'The birth weight "paradox" uncovered?' In: *American Journal of Epidemiology* 164.11 (2006), pp. 1115–1120 (cited on page 152).

[12] 'Carson v. Bethlehem Steel Corp.' In: 82 F.3d 157, 158, 7th Cir. (1996) (cited on page 153).

[13] Francine D. Blau and Lawrence M. Kahn. 'The gender wage gap: Extent, trends, and explanations'. In: *Journal of Economic Literature* 55.3 (2017), pp. 789–865 (cited on page 153).

[14] Sonja C. Kassenboehmer and Mathias G. Sinning. 'Distributional changes in the gender wage gap'. In: *ILR Review* 67.2 (2014), pp. 335–361 (cited on page 153).

[15] Elise Gould, Jessica Schieder, and Kathleen Geier. 'What is the gender pay gap and is it real'. In: *Economic Policy Institute* (2016) (cited on page 153).

[16] Gary S. Becker. *The Economics of Discrimination*. University of Chicago Press, 2010 (cited on page 155).

[17] Casey B. Mulligan and Yona Rubinstein. 'Selection, Investment, and Women's Relative Wages Over Time'. In: *Quarterly Journal of Economics* 123.3 (2008), pp. 1061–1110. DOI: 10.1162/qjec.2008.123.3.1061 (cited on page 156).

[18] Peter J. Bickel, Eugene A. Hammel, and J. William O'Connell. 'Sex Bias in Graduate Admissions: Data from Berkeley: Measuring bias is harder than is usually assumed, and the evidence is sometimes contrary to expectation.' In: *Science* 187.4175 (1975), pp. 398–404 (cited on page 158).

[19] Nina Roussille. 'The central role of the ask gap in gender pay inequality'. In: *URL: https://ninaroussille. github. io/files/Roussille_ askgap. pdf* 34 (2020), p. 35 (cited on page 158).