Explaining, and Explaining Away, Inequality

Contents

Orientation

Previously, we've looked at using regression models to adjust for covariates. We have an outcome Y, a main variable of interest X, and a bunch of other variables or covariates. Regression models, linear or otherwise, always try to estimate $\mathbb{E}[Y|X, Z, W, \ldots]$. That is, regression models are ways of estimating the expected outcome for particular sub-populations. (We use the model rather than just the average for that sub-population because we often need to interpolate and/or reduce noise.) Hopefully this sounds like a reasonable starting point for comparisons: we want to compare otherwise-similar groups or individuals, not totally disparate ones.

The crucial question, though, is usually what covariates, if any, should be included in the regression as controls.

As a concrete example: women have lower average incomes than men. That's a well-established and inarguable fact[1]. Here are some other facts[2]:

- Women are more likely than men to be part-time workers, and to work fewer hours, and both of these lead to lower incomes.
- Women tend to have fewer years of job experience.
- Women are more likely to be in occupations with lower average salaries.

Should we adjust the comparison of women to men to control for hours worked, for job experience, and for occupation?

Doing so, making those adjustments, definitely narrows the scope of the comparison to make it more similar, and to come closer to what one might call bare or naked sexism: "Here are two surgeons, both working full

[1] It shows up very clearly in the kind of CPS data we've been working with, but the particular data file we've used only has *household* and not *individual* income. There will be another data set soon.

[2] Notice, by the way, how hard it is to use ordinary language to state any of these things with precision. There are of course plenty of women who work for more hours than plenty of men, *even though* the hours-worked distribution curve for the sub-population of women is shifted to the left compared to the hours-worked distribution curve for the sub-population men. Ordinary language likes to turn differences between population distributions into categorical contrasts between ideal types or abstractions.

time and equally experienced, why is one paid more than another?" (Or maybe we find that the differences are negligible once we've controlled for everything, and congratulate ourselves on the equality[3].)

Doing so *also* puts aside lots of places where discrimination due to sex is, very plausibly, at play. *Why* do women work fewer hours, do more part-time work, and have fewer years of experience on the job? Because they're taking care of kids (and, increasingly, aging parents) much more than men are. *Some* of that is biology, but at least a *lot* of that is changeable social custom. (We know it's changeable because men have, on average, begun to spend significantly more time on child care just during my lifetime.) Why are certain jobs predominantly taken by women and others by men? Again, maybe *some* of it is due to biological differences, but a *lot* of it is social custom, and plain sexism[4]. Controlling for these variables is, plausibly, not so much explaining variation in the outcome, but rather explaining away inequality.

If I have been doing my job, by this point you should be feeling confused.

- Not controlling for anything at all is good, because it gives us clear information about inequalities.
- Not controlling for anything at all is bad, because it's attributing differences which *aren't* due to the categorical division to the categorical division.
- Controlling for everything we can is good, because it's ensuring fair, apples-to-apples comparisons.
- Controlling for everything we can is bad, because it means explaining away inequality and discrimination.
- ontrolling for everything we can is good, because it lets us measure the impact of direct discrimination.

Clarity comes from thinking about goals of the analysis, and from being explicit about causal structure

1. One goal is often to understand the *mechanisms* by which inequality happens. This is scientifically interesting, and also important for policy and for interventions (e.g., *how much* would it help to equalize hours worked?[5])
2. Another is to measure the sheer magnitude of the effects, i.e., to assess how big the inequality is.

Let me expand on the "causal structure" bit.

Causal structure

There are a couple of competing formalisms for representing causal structure; the clearest one, I think, is the one sometimes called "graphical causal models", a.k.a. "structural equation models", a.k.a. "path modeling". I am going to briefly lay out the essentials here. (It's actually been shown that the graphical-models formalism and its main competitor, the "potential outcome" formalism, are equivalent, though the translation between the two is not always easy (Richardson and Robins 2013).)

Graphical causal models in a nutshell

Every variable gets to be a **node**, a box (or circle) in the diagram. We draw an **edge**, or arrow, between variable A and variable B when we think A is a direct cause of B; the pointy end of the arrow points into B.[6]

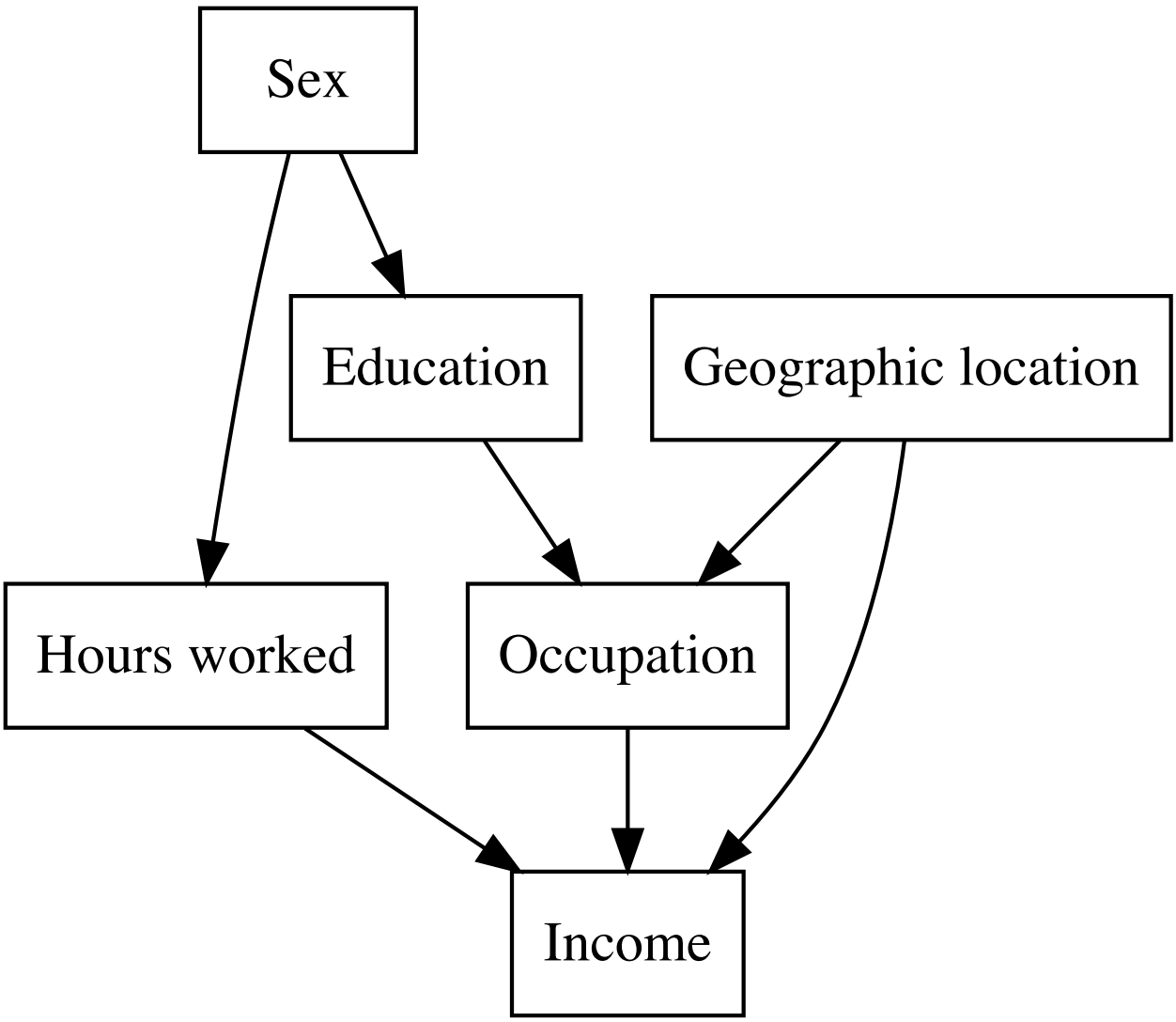
Here's one way to diagram out the boxes and arrows for our running example of sex inequality in incomes.

[3]Or, if you've got the right sort of values, you bemoan the lack of inequality which you think should exist. (Very few people want complete equality across every possbile social division.)

[4]Every human society has *some* form of gendered division of labor, and there are *some* cross-cultural commonalties to those divisions, but also a lot of variation. O'Connor (2019) is an interesting theory about why this is the case, which we'll look at later in the course, and gives good references to the anthropological literature.

[5]There could well be side-effects of some interventions which are unacceptable — if we reduce income inequality by sex through somehow compelling women to spend more time working for other people, when they would in fact rather *not* be doing so, have we really improved the situation? Might it be better to reduce the hours worked by *men*, by inducing them to spend more time taking care of their kids? etc., etc. But before we start assessing such trade-offs, it'd help to know what the effects of a policy would be!

[6]Mathematicians long ago seized on the name "graph theory" for studying the properties of collections of nodes linked by edges, even though they're not "graphs" in the plot-a-function or scatter-plot sense. That's what makes these models "graphical". Because our edges or arrows have directions, we have "directed graphs", sometimes abbreviated to "digraphs".



(This isn't meant to be a fully fleshed-out model of this important problem, just something simple enough to follow and get the points across. I will come back, later, to the issue of where the graph comes from, and/or possible rival graphs.)

We say that B is a **child** of A, and that A is a **parent** of B. **Ancestor** and **descendant** are defined transitively. If a variable has no parents, it is **exogenous** ("born outside"), otherwise it is **endogenous** ("born inside").

Every variable has some distribution given its parents. The joint distribution of all variables is the product of these conditional distributions.

If we can go from one node to another along arrows, that's a **path** connecting the nodes. If all the arrows point in the same direction, that's a **directed path**. X is a causal ancestor of Y if, and only if, there is a directed path from X to Y. A path which starts and ends at the same node is called a **cycle**. We usually insist that no directed path is also a cycle, so these are sometimes called **directed acyclic graph** (**DAG**) models. This means that no variable can be its own ancestor[7].

In a DAG, every variable is statistically independent of its non-descendants, *conditional on* its parents. (This is called the **(graphical) Markov property**.) One consequence is that an exogenous variable is independent

[7] "What about feedback?" you ask. "What low income leads to bad education which leads to bad jobs which leads to low income?" "Distinguish between low income now and low income later", I reply.

of any variable that isn't its descendant. In particular, any two exogenous variables are independent.

We can work out more independence relations by thinking about paths. Paths are ways of linking variables together, so they let us draw inferences about one variable from knowledge of another. We can think of them as channels along which information can flow[8]. The question is whether a path or channel is **open** (or **unblocked**), and information can flow along it, or **closed** (or **blocked**), and something keeps us from continuing a sequence of deductions.

- A **chain step**, or just **chain**, in a path is two consecutive arrows pointing the same way, $A \to B \to C$, or $A \leftarrow B \leftarrow C$. Either way, this step is by default open, but we block it by conditioning on the middle variable B. In words, if we know A causes B, and B causes C, in general we can draw some inferences about C from knowing A. But if we already know B, knowing A doesn't let us draw any *additional* inferences about C (at least not by this route). (If we know women work fewer hours, and working fewer hours leads to lower income, we can deduce that a woman probably earn less than average. But if we already know how many hours someone works, the additional knowledge that they're a woman doesn't help us predict their income, *by this channel*.)
- A **fork** is when two arrows diverge from a common cause, $A \leftarrow B \to C$. Similarly to a chain, a fork is open by default, but closed by conditioning on the middle variable. (If we see that someone works few hours, we can deduce that they're more likely to be a woman, and from that we can deduce that they're more likely to be in certain occupations than others. But if we already know they're a woman, we don't get more information, *by this channel*, from their hours worked.)
- Finally, a **collider** is when two arrows meet at a common effect, $A \to B \leftarrow C$. A collider is, by default, *closed*. However, a collider becomes *opened* if we condition on it, or any descendant of it. Say A is "hours worked", B is "income" and C is "occupation". If we know someone's occupation, that lets us draw some inferences about their income, but if we don't *know* their income, the sequence of inferences stops there, it doesn't let us go backwards to their hours worked. If we condition on their income, however, then we *can* go backwards to hours. Since getting food stamps depends on income, if we conditioned on food stamps we know *something* about income, and that's enough to let us draw inferences about hours worked (they'll be less precise than if we conditioned on income, but there'll still be *some* information)[9]. Creating dependencies by conditioning on colliders might seem like a weird "edge case" of little importance, but they often arise due to "selection effects", where being included in the sample at all is a consequence of variables included in the analysis.[10]

A path is blocked or closed if it's blocked or closed at any step; the path is open if it's open at every step.

Here is the key fact:

> If conditioning on the variable or variables S blocks every path between X and Y, then X and Y are independent given S.

The converse statement — if there is an open path between X and Y, then they are dependent — *can* fail, but failure requires weird coincidences where the contributions of different paths cancel each other out exactly. The assumption that these weird coincidences don't happen is called **faithfulness**, or the distribution being **faithful** to the graph. Assuming faithfulness, X and Y are dependent given S if, and only if, there is at least

[8] That analogy can be made very precise using information theory (Raginsky 2011, and see also @Venkatesh-Dutta-Grover-information-flow)

[9] A lot of the tricks Sherlock Holmes pulls in the stories (or the TV shows) are about conditioning on colliders. Think of when Holmes meets Watson for the first time, and says "I see you have been in Afghanistan". The reasoning is that Watson is a British man of a certain age who holds himself very straight, walks with a limp and has a deep tan. You can draw the DAG here for yourself.

[10] Claims that standardized test scores are bad predictors of success in college or graduate school is a famous example (at least among academics...). We only see academic success for *admitted* students. If test scores are one of the inputs into admissions decisions, then "being in the sample" is a descendant of test scores. If test scores aren't the *only* input into admissions decisions, there are paths linking test scores to those other inputs, and all of those paths have colliders at the "is admitted" variable. Even if test scores were independent of other admissions inputs, this would tend to create *negative* correlations among the inputs when we just look at the admitees. (If X and Y are independent and $Z = X + Y$, then, *conditional on* $Z = z$, $X = z - Y$ and X and Y are perfectly negatively correlated given Z.) Students who get admitted despite low test scores will tend to be strong on other factors like grades, letters of recommendation, etc., and conversely. If all those inputs really do predict academic success, the relationship will be attenuated in the sub-population of admitted students.

one path between X and Y which is open conditional on S.

What's "causal" about these models anyway?

We imagine **intervening on**, **manipulating**, or otherwise altering a variable, say X. The most basic sort of manipulation would be to set X to a particular value x. We show this in symbols as conditioning not on $X = x$, but on $do(X = x)$. In terms of the graph, this would correspond to deleting all the incoming arrows into x, and creating a new probability distribution in this "surgically altered" graph where $X = x$ with probability 1. All the *consequences* from this follow as usual, but any *inferences* we might ordinarily draw from x to its parents (and so to other variables) are blocked.

(If we want to do more complex manipulations, let randomly assigning $X = x_{\text{treatment}}$ or $X = x_{\text{control}}$, those follow from being able to calculate the consequences of these "atomic" interventions.)

You might well ask what the difference is between $\mathbb{P}(Y|X = x)$ and $\mathbb{P}(Y|do(X = x))$ (or similarly for conditional expectations, etc.). An example from Rubin and Waterman (2006) may help make this vivid. The more often sales reps from a given company visit a particular doctor, the more likely that doctor is to prescribe drugs from that company. That's the kind of information which lets us say things about $\mathbb{P}(Y|X = x)$, with Y = number of prescriptions written and X = number of sales visits. But maybe the sales reps visit doctors who they know are very likely to prescribe *anyway*, so they'll get credit for "making a sale" which would have happened even without their visit. The pharmaceutical company really wants to know whether the sales reps made a *difference*, i.e., $\mathbb{P}(Y|do(X = x))$. The worst case, from the point of view of the drug company[11], would be if lots of expensive sales visits were a *sign* that a doctor was going to prescribe, but not a *cause* of those prescriptions.

Similarly, we often want to know whether membership in some group is a *cause* of some inequality of interest to us, or merely a *sign* which points, indirectly, to some other cause. In the notes to lecture 9, I alluded to the dispute about whether racial inequality is more important (in contemporary America) than class inequality, or vice versa. Put crudely: we know race is inherited, and we know class is inherited. This means that *current* race and *current* class are correlated, because they share causal ancestry. To what extent are black people disadvantaged, compared to whites and Asians, *because* they are black, and to what extent *because* they are disproportionately lower-class? One can at least *imagine* a situation where all the causal paths from race to outcomes like income, risk of violence, etc., go through class, so that lower-class people of all races have the same disadvantages. (Few scholars, I think, believe that to be the case, but it's an imaginable world.)

Estimating causal effects[12]

If we want to estimate the effect of X on Y, a good set S of controls to include is one where (i) no variable in S is a descendant of X, and (ii) conditioning on S blocks every path between X and Y with an arrow *into* X (the "back-door paths"). The point of blocking the back door paths is to eliminate any associations between X and Y which are due to their sharing common causal ancestry. The point of not including descendants of X is that allowing them risks blocking one of the directed paths through which X influences Y, and/or activating a collider and inducing weird, non-causal dependencies between X and Y.

[11] You might well wonder whether selling drugs more effectively is good for doctors, patients, or society as a whole, but that's a separate issue, and it's easy to come up with examples which would still apply after the Revolution.

[12] You might well ask whether the phrase "causal effect" isn't redundant — doesn't the term "effect" imply a cause? But this arises because it's extremely ommon to talk about coefficients in a linear regression model, or contrasts in any regression model, as the "effects" of the relevant variable, *whether or not* one really regards the model as causal. (Thus in the pharmaceutical-marketing example, the slope coefficient in simple regression of prescriptions on sales visits might be called the "effect" of sales visits.) I try to avoid using "effects" in this non-causal way, because I'm a pedant, but even I slip up from time to time.

If we have a set of variables which meet this **back-door criterion**, then

$$\mathbb{P}\left(Y|do(X=x)\right) = \sum_s \mathbb{P}\left(Y|X=x, S=s\right)\mathbb{P}\left(S=s\right) \quad (1)$$

$$\mathbb{E}\left[Y|do(X=x)\right] = \sum_s \mathbb{E}\left[Y|X=x, S=s\right]\mathbb{P}\left(S=s\right) \quad (2)$$

$$\approx \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[Y|X=x, S=s_i\right] \quad (3)$$

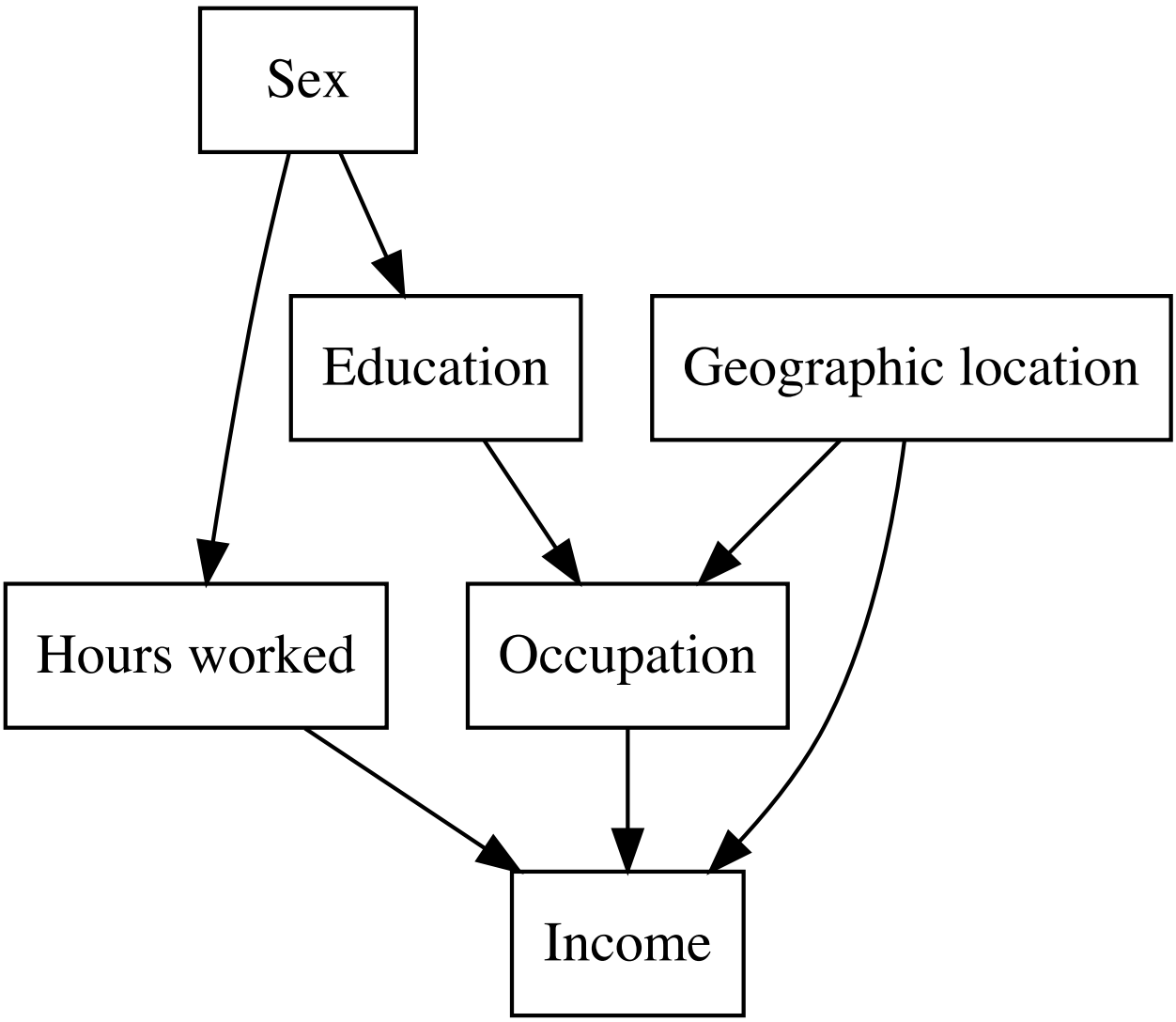
(In the last equation, the $\approx$ comes from the law of large numbers, and is meant to suggest an obvious estimation strategy. It's not always the most *efficient* way of doing the estimation, however.) If we can't find a set of variables which meets this criterion, there can still be other ways of estimating the effect of X on Y, though they generally won't be as simple as just doing a regression. (I will point you to Part III of Shalizi (n.d.), and/or the further reading at the end of these notes, and/or Prof. Branson's introduction to causal inference course next semester.)

If we're interested in how much of the effect is mediated by a particular directed path, we should block the *other* directed paths.

If we're interested in a direct, unmediated effect, we should block all the longer directed paths (leaving only the one-step $X \to Y$ path, if it exists).

An illustration relating to the gap in income by sex

Here's our running example again, since it's been a few pages.



Since, in this model, sex is exogenous, if we're interested in the *total* effect of sex on income, we should in fact control for nothing whatsoever. If we're interested in how much of the inequality is driven by hours worked, we should block the *other* directed path from sex to income, by controlling for education or occupation. If we're interested in how much of this is driven by occupational sorting, we should control for hours worked. If we want to see whether there is a direct effect of sex on income, or, more cautiously, one *not* mediated by the occupation and hours-worked channel, we should control for both of those. Controlling for geography is pointless. (Controlling for occupation would induce dependence between sex and geography.)

To elaborate, as I said in class, if we're running a hospital and want to make sure we're not engaging in sex discrimination in what we pay our workers, it might well make sense for us to control for hours worked and even for occupation. This wouldn't be because we think sex discrimination *isn't* why relatively more men than women who become surgeons, and relatively more women than men who become nurses, but because that is, to a large extent, beyond our control — whereas direct sex discrimination in pay *is* something we can control. This is not an iron-clad argument — we could do things about recruitment, we could adjust our family-leave or flex-time policies, etc. — but it's at least a sensible line of inquiry.

I hope this graph seems at least superficially plausible, but people could draw many different ones. Maybe occupation drives location. (People move for jobs!) Maybe sex is a direct parent of occupation, not just a grand-parent via education. Etc., etc. It is often the case that different people, studying the same variables, will have different ideas about the causal structure, which will lead them to control for different things, which

will lead to different conclusions. One advantage of actually drawing the graph is that it forces people to be *explicit* about their causal assumptions, so that they can be examined, debated, and, where necessary, improved (it can even lead you to revise your own assumptions — "Wait, does that really make sense?") *Some* of this is legitimate differences in opinion, and/or scientific uncertainty.

Arguments about causal structure do not take place purely at the level of "seems plausible to me". Possible graphs are constrained by background knowledge about the world, *and* by their testable implications. Let me expand on the second point. Remember that if S blocks all paths between X and Y, then X and Y are independent conditional on S. *We can test for conditional independence.* Statistical tests aren't infallible, but they do let us get evidence for or against different causal models, and in particular can weed out bad models. There can be multiple graphs which all imply the same independence relations, and then the data don't distinguish between these "Markov equivalent" graphs. Ideally, we'd repeat our analyses for every equivalent graph, but I admit that few people do (an exception: Maathuis et al. (2010), Maathuis, Kalisch, and Bühlmann (2009)). There is even an interesting sub-field of **causal discovery** which tries to learn the true graph (or class of possible graphs) from observations, essentially by weeding out the ones which are *in*-compatible with the data (Shalizi, n.d., ch. 22). It has not, to date, been much used in the study of social inequality, but some of us think it ought to be (Glymour 1998).

Summary

Which covariates we should control for are determined by the interplay between two things:

1. What is the goal of our analysis? What is it, exactly, we're trying to learn?
2. How do we think the variables fit together? That is, what's the causal structure at work?

Further reading

If you are interested in the issue of gendered inequality in income, a good place to start is Bach, Chernozhukov, and Spindler (2018), which is a careful look at the predictors of differences in *hourly wages* (not total income) using modern regression methods. This documents that the magnitude of the gap differs a lot between different social groups.

Morgan and Winship (2015) is a good introduction to causal inference for social scientists with some knowledge of statistics. (Nothing in it goes, mathematically or methodologically, beyond our 36-401.) Pearl (2009) is a concise summary of the contributions of one of the acknowledged masters of the field. Spirtes, Glymour, and Scheines (1993); Spirtes, Glymour, and Scheines (2001) is another great work from pioneers, especially on causal discovery.

References

Bach, Philipp, Victor Chernozhukov, and Martin Spindler. 2018. "Closing the U.S. Gender Wage Gap Requires Understanding Its Heterogeneity." arxiv:1812.04345. https://arxiv.org/abs/1812.04345.

Glymour, Clark. 1998. "What Went Wrong? Reflections on Science by Observation and *The Bell Curve*." *Philosophy of Science* 65:1–32. https://doi.org/10.1086/392624.

Maathuis, Marloes H., Diego Colombo, Markus Kalisch, and Peter Bühlmann. 2010. "Predicting Causal Effects in Large-Scale Systems from Observational Data." *Nature Methods* 7:247–48. https://doi.org/10.1038/nmeth0410-247.

Maathuis, Marloes H., Markus Kalisch, and Peter Bühlmann. 2009. "Estimating High-Dimensional Intervention Effects from Observational Data." *Annals of Statistics* 37:3133–64. https://doi.org/10.1214/09-AOS685.

Morgan, Stephen L., and Christopher Winship. 2015. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Second. Cambridge, England: Cambridge University Press.

O'Connor, Cailin. 2019. *The Origins of Unfairness: Social Categories and Cultural Evolution*. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780198789970.001.0001.

Pearl, Judea. 2009. "Causal Inference in Statistics: An Overview." *Statistics Surveys* 3:96–146. https://doi.org/10.1214/09-SS057.

Raginsky, Maxim. 2011. "Directed Information and Pearl's Causal Calculus." In *Proceedings of the 49th Annual Allerton Conference on Communication, Control and Computing*, edited by S. Meyn and B. Hajek, 958–65. IEEE. https://doi.org/10.1109/Allerton.2011.6120270.

Richardson, Thomas S., and James M. Robins. 2013. "Single World Intervention Graphs (SWIGs): A Unification of the Counterfactual and Graphical Approaches to Causality." 128. Center for Statistics; the Social Sciences, University of Washington. http://www.csss.washington.edu/Papers/wp128.pdf.

Rubin, Donald B., and Richard P. Waterman. 2006. "Estimating the Causal Effects of Marketing Interventions Using Propensity Score Methodology." *Statistical Science* 21:206–22. https://doi.org/10.1214/088342306000000259.

Shalizi, Cosma Rohilla. n.d. *Advanced Data Analysis from an Elementary Point of View*. Cambridge, England: Cambridge University Press. http://www.stat.cmu.edu/~cshalizi/ADAfaEPoV.

Spirtes, Peter, Clark Glymour, and Richard Scheines. 1993. *Causation, Prediction, and Search*. 1st ed. Berlin: Springer-Verlag. https://doi.org/10.1007/978-1-4612-2748-9.

———. 2001. *Causation, Prediction, and Search*. 2nd ed. Cambridge, Massachusetts: MIT Press.

Venkatesh, Praveen, Sanghamitra Dutta, and Pulkit Grover. 2020. "Information Flow in Computational Systems." *IEEE Transactions on Information Theory* 66:5456–91. https://doi.org/10.1109/TIT.2020.2987806.