Identifying direct effects

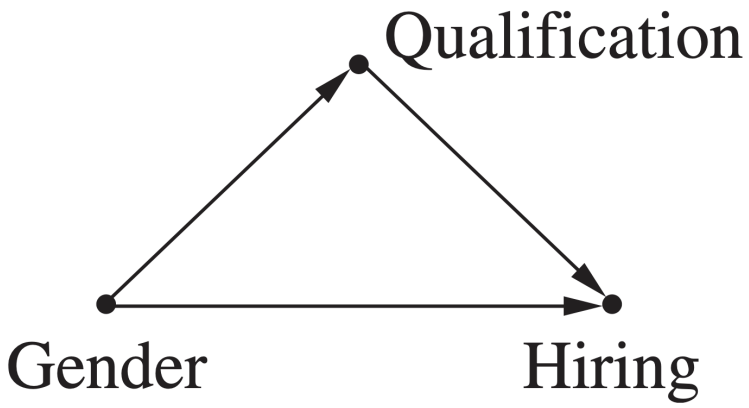
The techniques we've seen so far were about estimating the *total* causal effect of X on Y

- This corresponds to *all* directed paths from X to Y in the causal graph

We may also be interested in measuring just the direct effect

Example: A direct effect of gender on hiring is discrimination, but we want to distinguish it from an indirect effect via differences in qualification

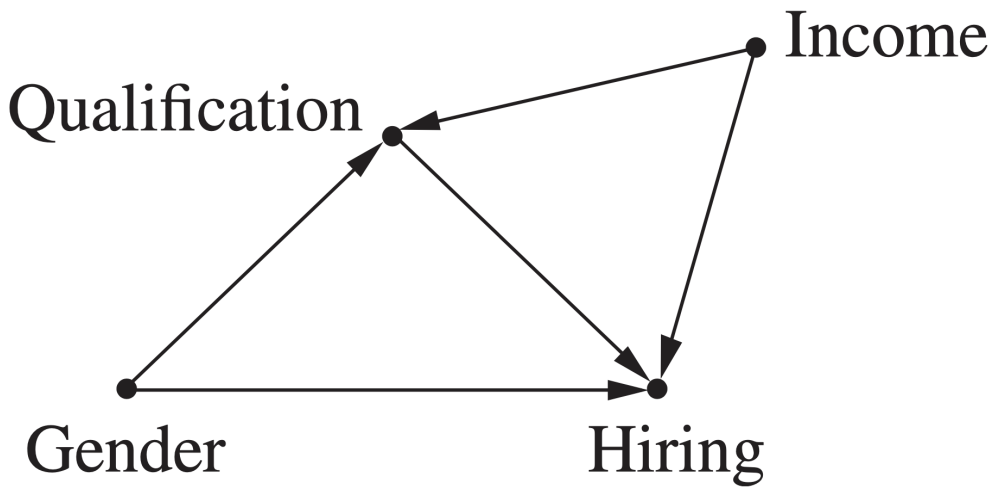
Example: direct effect of gender on hiring (1/3)



A mediator is a variable on a directed path between a specified cause and effect

To find the direct causal effect in this graph, we can condition on the mediator: find the *covariate-specific effect* $P(H \mid \text{do}(G), Q = q)$ for each q

Example: direct effect of gender on hiring (2/3)

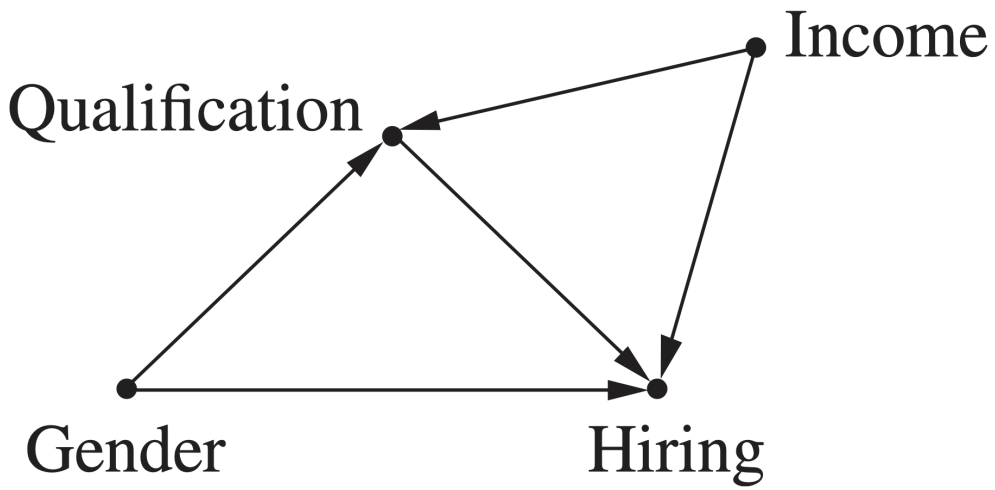


Conditioning on Q (as for covariate-specific effect) now opens the path $G \rightarrow Q \leftarrow I \rightarrow H$

Solution: *intervene* on all the mediators

What we want to know: $P(H \mid \text{do}(G = g, Q = q))$

Example: direct effect of gender on hiring (2/3)



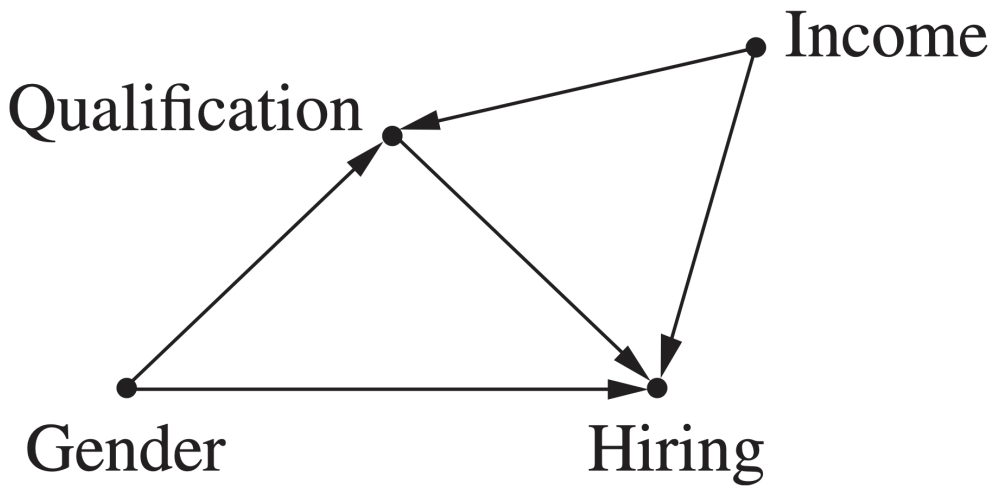
Conditioning on Q (as for covariate-specific effect) now opens the path $G \to Q \leftarrow I \to H$

Solution: *intervene* on all the mediators

What we want to know: $P(H \mid \text{do}(G = g, Q = q))$
...from which we can compute the controlled direct effect on H of changing G from g' to g *(← wrong way around in book)*:

$$P(H = h \mid \text{do}(G = g, Q = q)) - P(H = h \mid \text{do}(G = g', Q = q))$$

Example: direct effect of gender on hiring (3/3)



$$
\begin{aligned}
&P(H = h \,|\, \mathrm{do}(G = g, Q = q)) \\
&\quad = P(H = h \,|\, G = g, \mathrm{do}(Q = q)) \\
&\quad = \sum_i P(H = h \,|\, G = g, Q = q, I = i) P(I = i)
\end{aligned}
$$

Indirect effects

What if we want to know the indirect effect only?

In theory, we can find this as the difference between the total and the direct effect

Might not give a clear answer if in the structural equation of the effect, the cause and the mediator *interact* in a complex way (more about this in chapter 4)

Forms of structural equations (1/2)

It is common to make some assumptions about the type of the structural equations

- *Nonparametric*: No restriction on the form. Could be approximated by e.g. fitting a sufficiently large neural network.
- *Discrete case*: If a variable X and its parents are categorical, a finite number of parameters suffices to specify the distribution of X for each value of the parents

Forms of structural equations (2/2)

- *Linear*: functional form is $X_i = \alpha_0 + \alpha_1 X_j + \alpha_2 X_k + \ldots + U_i$
- *Linear Gaussian*: additionally assume that U_i follows a normal distribution
- *Additive noise*: functional form is $X_i = f(\text{parents}) + U_i$
  Much weaker assumption than linearity

Linear Gaussian case

Some special properties hold in the linear Gaussian case:

- Any (joint/marginal/conditional) distribution of the variables is a multivariate Gaussian: completely described by means, variances and covariances
- As a consequence, conditional expectations are linear functions
  - Both the above also hold for interventional distributions/expectations, because the intervened model is also linear
- So if we are interested in, say, $E(Y \mid \mathrm{do}\, X = x + 1) - E(Y \mid \mathrm{do}\, X = x)$, this is the same number regardless of x
  - This also means that the indirect effect now makes sense to talk about

Structural vs. regression coefficients

In the linear Gaussian case, the difference between structural equations and regression equations becomes very clear:

- For any set of variables, we can find a regression (i.e. do supervised learning) to predict Y in terms of X, but not all of these correspond to the model's structural equations
- Very easy to tell the difference now that they are just single numbers (assuming we're interested in the slope)

Other applications of causal inference in ML

We have been looking at how to predict causal effects from observational distributions, using causal models

But there are other, less obvious ways that causal inference may play a role in machine learning!

- Fairness
- Domain adaptation
- Many others I won't say more about, e.g. explainability

Fairness

(see e.g. Kilbertus et al (2017). Avoiding Discrimination through Causal Reasoning)

As we saw, we can describe discrimination in terms of causal models

- Without causal models, we can't always tell the difference between 'allowed' and 'forbidden' influences
- Important in machine learning applications
- Still hard due to e.g. proxy variables

Domain adaptation

(see e.g. Subbaswamy, Suchi and Saria (2019). Preventing failures due to dataset shift: Learning predictive models that transport)

Even if we're not doing an 'intervention' exactly, we still may want to relate different distributions using causal models

- E.g. a machine learning model for diagnosis trained in one hospital, but to be used in another where some aspect of the patient distribution is different
- Causal models allow us to make precise what will change and what will stay the same