DAGs and confounding

ISTA 410 / INFO 510: Bayesian Modeling and Inference

U. of Arizona School of Information

March 24, 2021

Outline

Last time:

- Multiple regression
- Total vs. direct causal effect

Today:

- Hazards of regression: multicollinearity, confounding, and collider bias
- Causal DAGs

Aside: categorical variables

Categorical variables in regression

Ways to handle a categorical variable c:

- indicator variable
- if c is binary, assign value 0 to one category, 1 to another

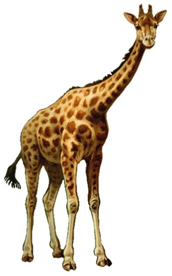
$$\mu_j = \alpha + \beta_c c$$

Potential problem: more uncertainty in one category than another

Example: human height

Simple example: modeling giraffe height stratified by sex. Assign $s = 0$ for female, $s = 1$ for male.

$$
\begin{aligned}
h &\sim \mathrm{Normal}(\mu, \sigma) \\
\mu &= \alpha + \beta s \\
\alpha &\sim \mathrm{Normal}(5, 1.5) \\
\beta &\sim \mathrm{Normal}(0, 0.4)
\end{aligned}
$$



- $\mathrm{Var}\,\mu$ for males $= \mathrm{Var}\,\alpha + \mathrm{Var}\,\beta$
- $\mathrm{Var}\,\mu$ for females $= \mathrm{Var}\,\alpha$

Alternatives: one-hot encoding or index variables

One alternative: one-hot encoding

- create an indicator variable for every category

$$\mu = \beta_f f + \beta_m m$$

Why drop α? Not enough constraints:

- Suppose female giraffes average 4.8 m, males 5.1; then which is correct?

$$\mu = 5 + 0.1m - 0.2f$$

$$\mu = 4 + 1.1m + 0.8f$$

Alternatives: one-hot encoding or index variables

Another: index variables

- create a vector of intercepts and use the value of the categorical variable to index out the right one

$$\mu = \beta_s \qquad s \in \{0, 1\}$$

  Requires encoding values of the variable as ordinal values $\{0, 1, 2, \ldots, n\}$

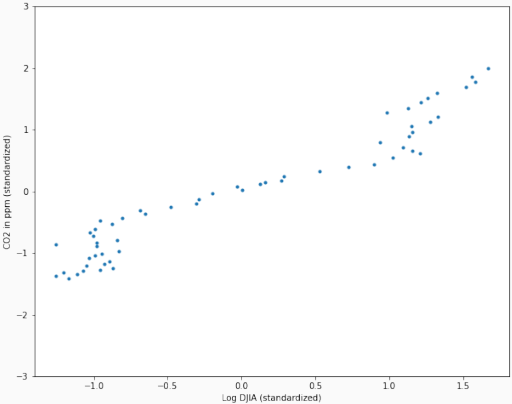
Same as what I suggest for the multilevel model in the midterm: create a vector of θs and use the `statecode` as a vector index

Multicollinearity in regression

Multicollinearity

Multicollinearity: when several predictors are tightly correlated with one another

- Example: CO_2 and DJIA from last time



Multicollinearity

Summary tables from three models:

- CO2 only:

	mean	sd	hdi_3%	hdi_97%
alpha	0.000	0.044	-0.079	0.084
beta_c	0.940	0.043	0.859	1.021
sigma	0.332	0.032	0.275	0.394

- DJIA only:

	mean	sd	hdi_3%	hdi_97%
alpha	0.001	0.052	-0.088	0.105
beta	0.910	0.054	0.815	1.015
sigma	0.404	0.041	0.333	0.481

- Both:

	mean	sd	hdi_3%	hdi_97%
alpha	-0.000	0.043	-0.077	0.085
beta_c	0.783	0.155	0.499	1.082
beta_d	0.164	0.155	-0.111	0.475
sigma	0.335	0.032	0.277	0.396

Identifiability

The problem with multicollinearity is a problem of *identifiability*

- A model is *identifiable* if, given an infinite amount of data, the model parameters could be inferred exactly
- The height model with the extra intercept is non-identifiable – even if you had perfect estimates, there is a 1D space of equivalent parameter vectors
- Multicollinearity usually doesn't imply true non-identifiability (unless predictors are perfectly correlated), but "weak identifiability"

DAGs as probabilistic models

What is a DAG?

What is a DAG?

- Directed acyclic graph
- Nodes are variables
- Directed arrows are causal associations

What are we using DAGs for? Probabilistic models, on two levels:

- probabilistic model for causal associations between variables
- metadata that guides choice of variables for inference

References

BDA mentions causal inference and gives some details, but doesn't use DAGs

Statistical Rethinking chapters 5 and 6 (my main source for this)

Core book in the field: Judea Pearl, *Causality* (available online through UofA library)
Chapter/section references:

- DAGs as probabilistic models: Chapter 1
- The backdoor criterion: Section 3.3
- Simpson's paradox and confounding: Chapter 6

Three technical slides

Probabilistic model of a DAG

The probabilistic nature of a DAG is implied *conditional independence*.

Say we have n variables $X_1, \ldots, X_n$. We can always write

$$p(x_1, \ldots, x_n) = \prod_i p(x_i | x_1, x_2, \ldots, x_{i-1})$$

(the chain rule). We are interested in the case where each x_j is dependent on only some of the other variables:

$$p(x_i | x_1, \ldots, x_{i-1}) = p(x_i | pa_i)$$

where PA_i is a subset of the remaining variables, called the "parents" of X_i.

Graphical example

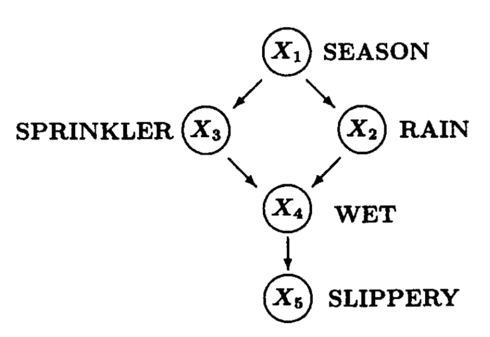


Figure from *Causality*

Graphical example

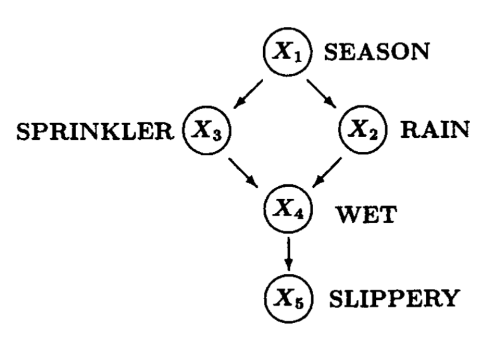


Figure from *Causality*

$$P(x_1, \ldots, x_5) = P(x_1)P(x_2|x_1)P(x_3|x_1)P(x_4|x_2, x_3)P(x_5|x_4)$$

Controlling "flows"

When we're trying to estimate the effect of one variable on another:

- Control "flow" of information along paths
- Information flows along or against arrows
- Including a variable in the regression can either "block" or "open" paths

Three basic paths

Three basic paths

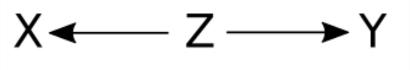
In a DAG, information flows along paths (both with and against the arrows).

A path from X to Y can be a direct path – an arrow between X and Y. Or it can be an indirect path $X \leftrightarrow Z \leftrightarrow Y$ (or a concatenation of several of these).

Indirect paths can lead to confounding / spurious associations; to deal with this, we need to classify the different types of indirect paths.

The "fork" path

The *fork* is the form most students learn as the sole definition of "confounding" in introductory classes: X and Y are confounded by their common cause, Z:

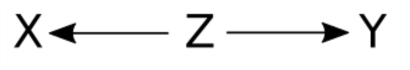


A statistical association exists between X and Y because they are both influenced by Z.

Example: X is ice cream sales; Y is drowning deaths; Z is temperature

The “fork” path

The *fork* is the form most students learn as the sole definition of “confounding” in introductory classes: X and Y are confounded by their common cause, Z:

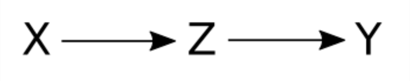


Conditional independence:

- DAG property means: conditional on Z, X and Y are independent.
- So, condition/stratify/control on Z to block the path and estimate effect of X on Y

The “chain” path

The *chain* is a similar-looking form, where Z sits in the middle of a causal path:



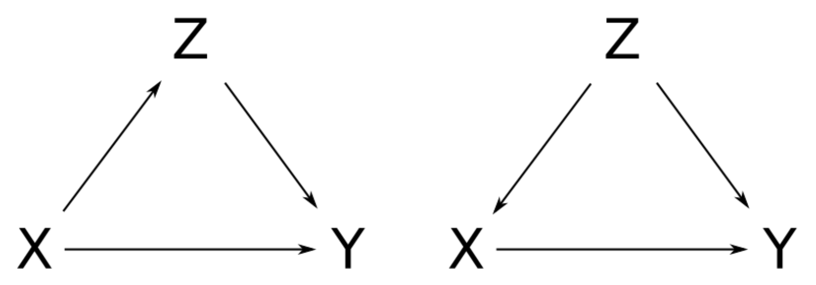
Typical case: Z is an effect of X that mediates the effect on Y

Example: X is pesticide application; Z is the pest population; Y is crop yield.

Controlling for Z blocks information flow along the path.

When the data can't tell you

Multiple paths: should you include the variable Z or not?



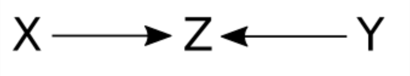
The data/model cannot tell you the difference between these, because they imply the same set of conditional independences

The “collider” path

The third form is the *collider* or inverted fork, and it behaves quite differently:

In contrast to the fork or chain, information flows through the collider only when it *is* observed / controlled; controlling *unblocks* the path.

Heuristic example



X: switch state on/off Z: light bulb on/off Y: power working/not working

The presence of power and the state of the switch are independent; but,

- turn on the switch and observe the light: it's off
- is the power working?

The explaining-away effect

This property of colliders is responsible for a sometimes counterintuitive effect:

- "explaining away": observing one of the common causes
- Berkson's paradox: conditioning on a variable can introduce a spurious association

They're really the same effect; explaining away common in AI/ML; Berkson's paradox in statistics

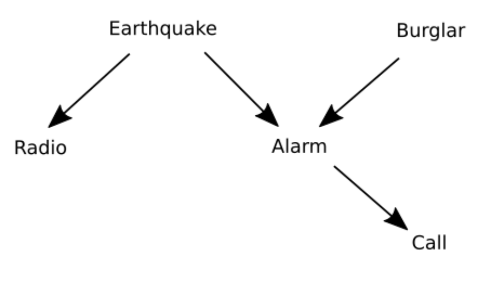
Explaining away: the burglar alarm

From Pearl by way of Mackay:

> *Fred lives in Los Angeles and commutes 60 miles to work. Whilst at work, he receives a phone-call from his neighbour saying that Fred's burglar alarm is ringing. What is the probability that there was a burglar in his house today? While driving home to investigate, Fred hears on the radio that there was a small earthquake that day near his home. 'Oh', he says, feeling relieved, 'it was probably the earthquake that set off the alarm'. What is the probability that there was a burglar in his house?*

Explaining away: the burglar alarm

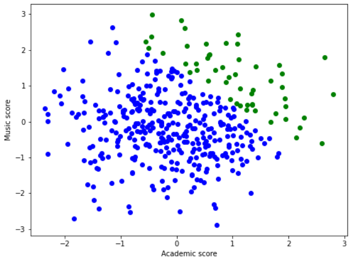
A DAG for the burglar alarm problem, showing the collider:



The alarm sits at a collider.

Conditioning on colliders creates confounding

The spurious-association effect of conditioning on a collider:



Berkson's paradox a.k.a. *selection bias*

Recent example

Recent example: risk factors for COVID-19

- Early studies of COVID-19 were based on observational studies
- Testing availability was low, so the population whose status could be confirmed was subject to selection bias

Examining the effect of smoking

Example study: does smoking protect against severe disease?

- early observational data suggested a negative association between smoking and probability of severe COVID-19
- this is a surprising finding!

Examining the effect of smoking

Example study: does smoking protect against severe disease?

- early observational data suggested a negative association between smoking and probability of severe COVID-19
- this is a surprising finding!

Implicit collider: COVID-19 testing

Examining the effect of smoking

In the early stages of the pandemic, two groups of people were tested most commonly:

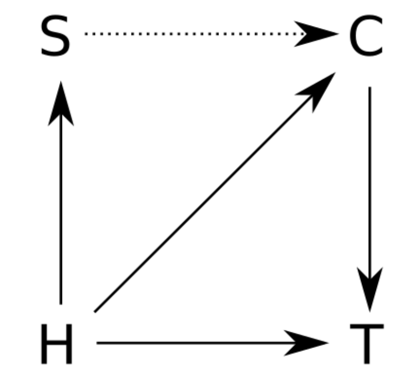
- people with severe disease
- healthcare workers

Conditioning on testing introduces an association between these two traits

Griffith et al., "Collider bias undermines our understanding of COVID-19 risk and severity" (Nature, 12 Nov 2020)

A DAG for the smoking confound

Here is a DAG:



The backdoor criterion

d-separation

A (possibly undirected) path p through a DAG G is said to be *d-separated* or *blocked* by a set of nodes Z if:

1. p contains a chain $X_i \to M \to X_j$ or fork $X_i \leftarrow M \to X_j$ such that $M \in Z$; or,
2. p contains a collider $X_i \to M \leftarrow X_j$ such that $M \notin Z$ and no descendent of M is in Z.

(Why the descendant property? Look back at the burglar alarm.)

The d-separation (blocking) definition for paths leads to another definition, for sets of variables.

The backdoor criterion

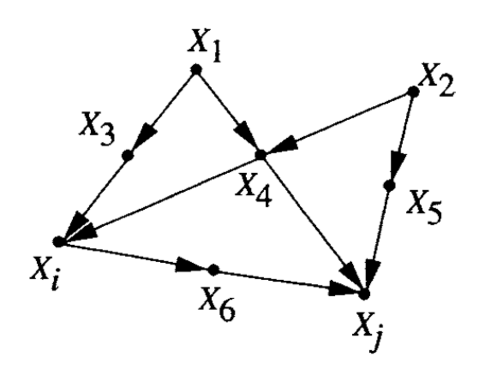
A related definition:

Definiton
A set of variables Z satisfies the backdoor criterion with respect to an ordered pair of variables (X_i, X_j) in G if:

1. *no node in Z is a descendent of X_i; and,*
2. *Z blocks every path from X_i to X_j that contains an arrow into X.*

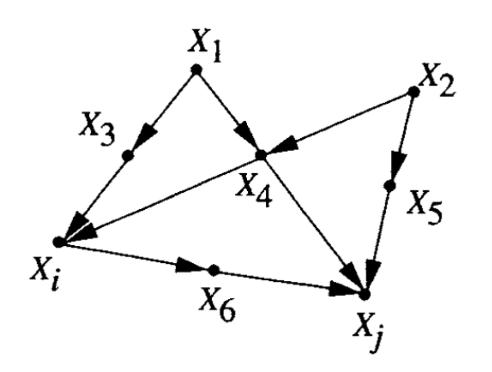
To estimate the causal effect of X on Y, condition on a set of variables satisfying the backdoor criterion with respect to (X, Y).

Example



Which variables satisfy the backdoor criterion?

Example

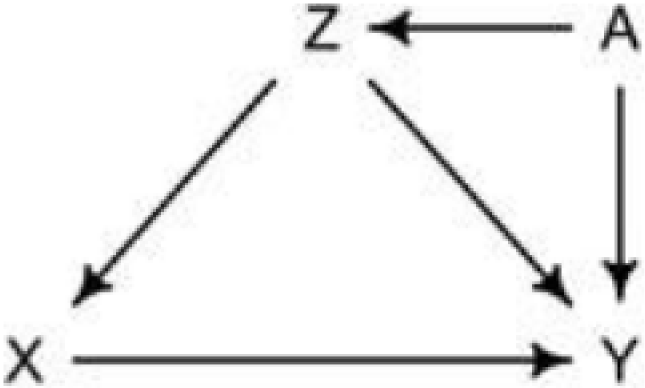


Which variables satisfy the backdoor criterion?

- $\{X_3, X_4\}$ or $\{X_4, X_5\}$
- Not $\{X_4\}$ (doesn't block every backdoor path), nor $\{X_6\}$ (descendent of X_i)

Group exercise

For each DAG, which variable should be conditioned on to estimate total causal influence of X on Y?



Summary

Today:

- DAGs
- Three types of confounder
- Collider bias

Next week:

- interactions between variables
- correlated parameters