Multiple regression and causal DAGs

ISTA 410 / INFO 510 - Bayesian Modeling and Inference

University of Arizona School of Information

September 13, 2021

Outline

Last week:

- Model specification in PyMC3
- Inference with quadratic approximation via `quap`
- Posterior predictive sampling
- Linear regression

Today:

- Multiple regression
- Causal DAGs

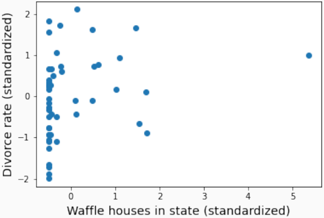
Starting chapter 5 of *Rethinking*

Goals for this week

- Introduce multiple regression and explore its properties
  - Used correctly, these models can uncover masked associations or eliminate spurious correlations
  - Used carelessly, they can introduce spurious associations and lead to confusion
- Basics of causal inference
  - Directed acyclic graphs (DAGs)
  - DAG structures and causal relationships
  - Criteria for including/excluding variables

Waffle House and divorce

Presence of Waffle Houses statistically correlated with divorce



- Are Waffle Houses dens of iniquity?

It's not the waffles

It seems most plausible that:

- Waffle Houses don't cause divorce
- Some other property influences the presence of WH and also the rate of divorce

It's not the waffles

It seems most plausible that:

- Waffle Houses don't cause divorce
- Some other property influences the presence of WH and also the rate of divorce

- The US South

Anything idiosyncratic to the South will be associated with Waffle House.

Correlation is not causation

- Truism from intro stats: correlation does not imply causation
- Causation does not imply correlation
- Causation implies *conditional correlation*
- Example (due to Rachael Meager):
  - Pressure on the gas pedal has a causal effect on your car's speed
  - If you go from a flat to a steep hill, you might increase pressure to keep speed constant

Other predictors

What other features of the South might be responsible for the divorce rate?

- poverty
- religiosity → seems to suggest opposite association

Other predictors

What other features of the South might be responsible for the divorce rate?

- The South has generally higher religiosity
- The South has generally higher poverty rates

what we said

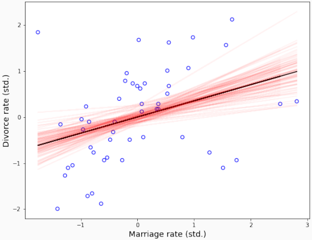
- Two predictors that we'll use
  - Marriage rate
  - Median age at (first) marriage

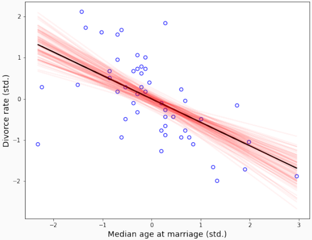
in the data set

Marriage rate & age at marriage

- Marriage rate
    - You have to get married to get divorced (+)
    - Society values marriage, so opposes divorce (-)
- Median age at (first) marriage
    - Younger people make worse decisions
    - People change

Marriage rate and median age at marriage





- Does higher marriage rate cause higher divorce rate?
- Does higher age at marriage cause lower divorce rate?

DAGs

What's a DAG?

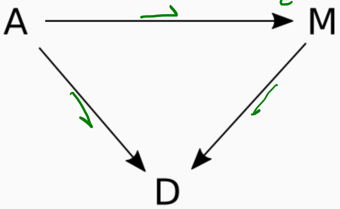
- Directed (edges are arrows)
- Acyclic (No directed loops)
- Graph (nodes and edges)

Use as a heuristic model for causal relationships

- Not a mechanical model – does not include explanation of how or why the causal relationship exists
- Not even a statistical model – does not specify probability distributions, only conditional dependence/independence

Our first DAG

Here is a DAG representing a possible model for the relationships between age at marriage, marriage rate, and divorce rate:



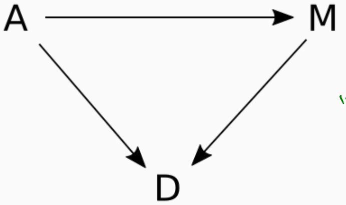
people getting married younger have more people to marry

What this DAG says:

1. A directly influences D
2. M directly influences D
3. A directly influences M

Our first DAG

Here is a DAG representing a possible model for the relationships between age at marriage, marriage rate, and divorce rate:



"paths" - undirected connections from one var to another.

In this model, total causal effect of A on D:

1. $A \to D$ – direct causal effect
2. $A \to M \to D$ – indirect effect

Multiple regression and control

Multiple regression provides statistical "control." This means *conditioning on the information in one variable*, not *setting the value of one variable*

- Multiple regression answers: once we know all other predictors, how is each predictor associated with the outcome?
- To interpret the effect of statistical control, we need a clear model of what the causal relationships might be – this is what the DAG offers us

Two competing DAGs

Here are two DAGs:

- Both DAGs are consistent with the inferences of our previous models
- Statistical association between M and D appears in both

Multiple regression

Multiple regression estimates conditional associations:

- What is the value of a predictor, once we know the other predictors?
    - If we already know the marriage rate in a state, what do we learn from the age at marriage?
    - If we already know the age at marriage in a state, what do we learn from the marriage rate?

$$D_i \sim \mathrm{Normal}(\mu_i, \sigma) \quad \leftarrow \text{likelihood}$$

$$\mu_i = \alpha + \underbrace{\beta_M}_{\text{conditional on } A} M + \underbrace{\beta_A}_{\text{conditional on } M} A \quad \leftarrow \text{model equation}$$

Multiple regression model

$$D_i \sim \mathrm{Normal}(\mu_i, \sigma)$$

$$\mu_i = \alpha + \beta_M M + \beta_A A$$

$$\alpha \sim \mathrm{Normal}(0, 0.2) \leftarrow \text{prior on intercepts}$$

$$\beta_M \sim \mathrm{Normal}(0, 0.5)$$

$$\beta_A \sim \mathrm{Normal}(0, 0.5)$$

priors on slopes

$$\sigma \sim \mathrm{Exponential}(1) \leftarrow \text{prior on std. dev.}$$

Priors

What are these priors about?

- Working with standardized data

$$z_i = \frac{x_i - \bar{x}}{s}$$

- Each variable has mean 0, SD 1
- Intercept parameter should be 0:

$$\alpha \sim \mathrm{Normal}(0, 0.2)$$

← tighter prior

- Priors on slopes fairly vague

$$\beta_M \sim \mathrm{Normal}(0, 0.5)$$
$$\beta_A \sim \mathrm{Normal}(0, 0.5)$$

Fitting the model

context

```
# Model with marriage rate and median age at marriage
with pm.Model() as linear_model:
    # Priors for model parameters
    alpha = pm.Normal('alpha', mu = 0, sigma = 0.2)
    betaA = pm.Normal('betaA', mu = 0, sigma = 0.5)
    betaM = pm.Normal('betaM', mu = 0, sigma = 0.5)
    sigma = pm.Exponential('sigma', 1)

    # Model equation
    mu = alpha + betaA * A + betaM * M

    # Observed variable and inference
    D_obs = pm.Normal('D_obs', mu = mu, sigma = sigma, observed = D)
    qp = quap()
```

names

likelihood

quadratic approx

A, M, D are arrays with values for these variables

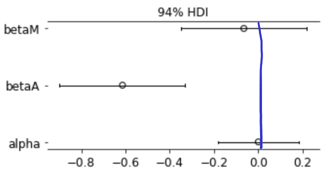
Fitting the model

```
# Model with marriage rate and median age at marriage
with pm.Model() as linear_model:
    # Priors for model parameters
    alpha = pm.Normal('alpha', mu = 0, sigma = 0.2)
    betaA = pm.Normal('betaA', mu = 0, sigma = 0.5)
    betaM = pm.Normal('betaM', mu = 0, sigma = 0.5)
    sigma = pm.Exponential('sigma', 1)

    # Model equation
    mu = alpha + betaA * A + betaM * M

    # Observed variable and inference
    D_obs = pm.Normal('D_obs', mu = mu, sigma = sigma, observed = D)
    qp = quap()
```

	mean	sd	hdi_3%	hdi_97%
alpha	0.000	0.098	-0.184	0.184
betaA	-0.614	0.151	-0.897	-0.330
betaM	-0.065	0.151	-0.349	0.218
sigma	0.793	0.079	0.645	0.941

slopes

conditional
on other parameters



Comparing three model results

- Age only:

	mean	sd	hdi_3%	hdi_97%
alpha	0.000	0.098	-0.185	0.185
betaA	-0.568	0.110	-0.775	-0.361
sigma	0.796	0.079	0.648	0.945

neg association
between A, D

- Marriage rate only:

	mean	sd	hdi_3%	hdi_97%
alpha	0.000	0.109	-0.205	0.205
betaM	0.350	0.126	0.113	0.587
sigma	0.919	0.091	0.749	1.090

pos association
between M, D

- Both:

	mean	sd	hdi_3%	hdi_97%
alpha	0.000	0.098	-0.184	0.184
betaA	-0.614	0.151	-0.897	-0.330
betaM	-0.065	0.151	-0.349	0.218
sigma	0.793	0.079	0.645	0.941

neg association
between A, D
none between M, D.

Exploring the result graphically

Two approaches to exploring our results graphically

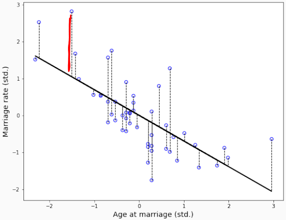
- Residual analysis – bad idea for analysis but useful for understanding the mechanics
  - Do a regression of marriage rate on age at marriage
  - Extract residuals and use them as inputs in a second regression
- Posterior predictive plots
  - Compare predicted divorce rate to actual divorce rate
  - Check overall model fit
  - Identify places where the model fails – can it be improved?

Residual analysis

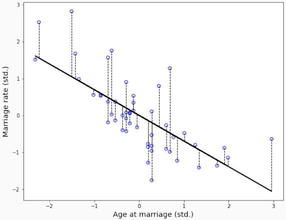
Do it both ways:

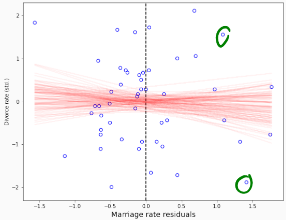
- Regress marriage rate on age at marriage
- Use the residuals as predictors for divorce
    - Idea: residuals represent "leftover" variability in marriage rate once age at marriage is known
    - If marriage rate has a direct effect on divorce, it will be visible here
- Do the same starting from a regression of age on marriage rate

Residual analysis

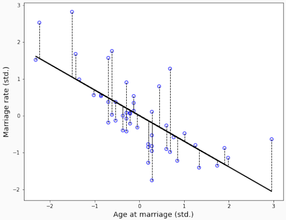


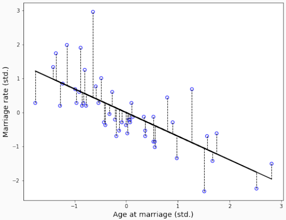
Residual analysis

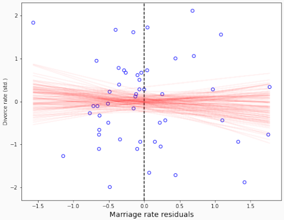




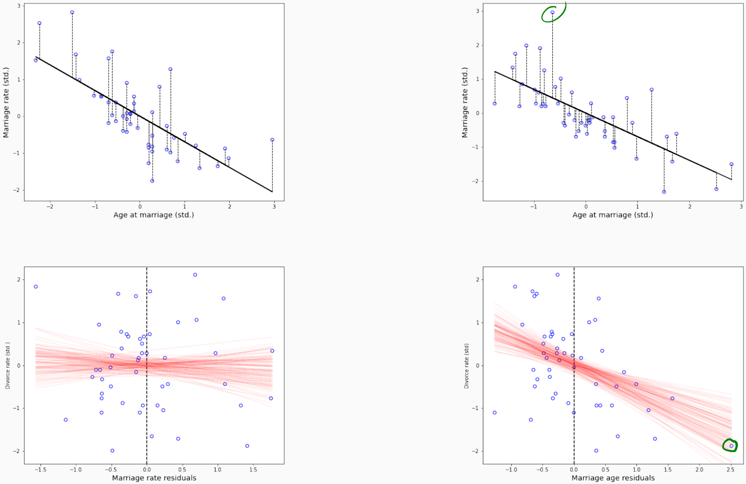
Residual analysis







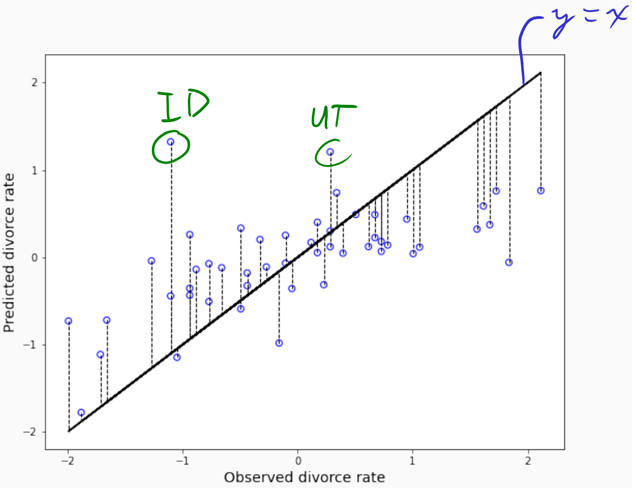
Residual analysis



Statistical "control"

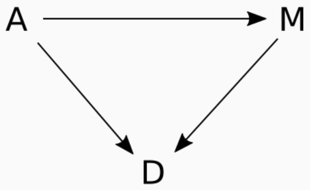
- Multiple regression often described as "controlling" for each variable
  - Not really control – we're not fixing input values in an experiment
  - Probabilistically: we are computing *conditional* distributions
  - Better way to think about this: "stratify"
- How is each predictor associated with the outcome, once all other predictors are known?

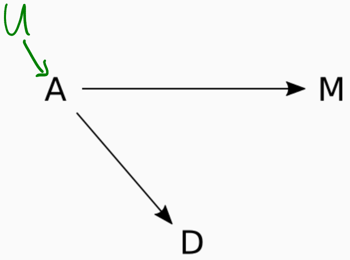
Posterior predictive plot



Summary

Recall our initial DAGs:

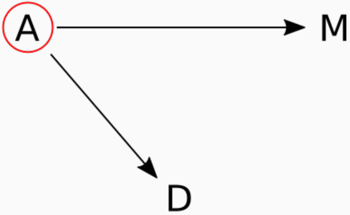




- Multiple regression reveals that the direct effect of M on D is weak/zero
- Note: if A is unknown, M is still useful; still carries *predictive* power
- If you wanted to stage interventions to reduce divorce, though, M wouldn't be a good target

Conditioning on A

Here, adding variables to the regression helped us determine the effect:



Conditioning on A eliminated the spurious association of M with D

In some cases, conditioning on a variable can *introduce* a spurious association – depends on the structure of the DAG

Masked associations

- Previously: used multiple regression to suppress a spurious correlation
- Masked association:
  - There is really a direct causal effect of predictor on outcome
  - A simple regression shows no (or suppressed) association
- Typical situation:
  - Have two predictors that are associated with the outcome in the opposite direction
  - Two predictors are positively associated with one another
  - Effects cancel

Urban foxes in London

Data set: urban foxes living in groups

- We want to monitor the health of groups of foxes living in the city
- Easy health metric: fox weight
    - Foxes with adequate food, no disease, etc. are heavier
    - Low weight is a marker for various health conditions (malnutrition, disease, age)
- Is the presence of food a good predictor for weight?
- Is the presence of food causally related to weight?

What happens if you feed the foxes?

If food is added to an area, will the foxes get bigger?

- This is a causal question, not just a statistical question
- Difference: talks about an intervention
- Simplest thing to try: regress fox weight on average food

Simple linear regression

Linear regression for fox weight:

$$\begin{aligned}
w_i &\sim \mathrm{Normal}(\mu_i, \sigma) \\
\mu_i &= \alpha + \beta_f \mathrm{food} \\
\alpha &\sim \mathrm{Normal}(0, 0.2) \\
\beta_f &\sim \mathrm{Normal}(0, 0.5) \\
\sigma &\sim \mathrm{Exponential}(1)
\end{aligned}$$

What does the fox model say?

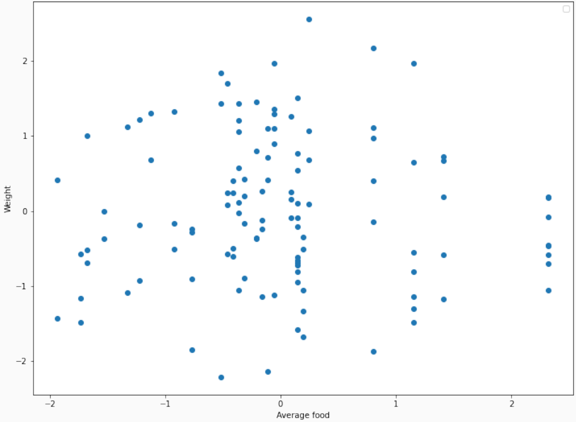
Here are the estimates from the fox model:

	mean	sd	hdi_3%	hdi_97%
bF	-0.024	0.092	-0.191	0.150
alpha	-0.003	0.099	-0.183	0.185
sigma	1.013	0.068	0.884	1.131

This is about as close to zero as we can get. Does this check out with the data?

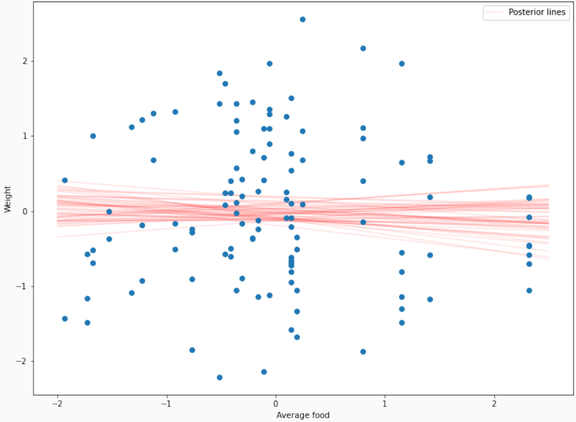
What does the fox model say?

Scatterplot of weight vs. average food:



What does the fox model say?

Scatterplot of weight vs. average food:



What does the fox model say?

The fox model tells us:

- No apparent association between food availability and fox weight
- But intuition tells us: if we provide more food, it must go somewhere!

What does the fox model say?

The fox model tells us:

- No apparent association between food availability and fox weight
- But intuition tells us: if we provide more food, it must go somewhere!
- More foxes

How can we check this? Include both variables.

A multiple regression

$$
\begin{aligned}
w_i &\sim \mathrm{Normal}(\mu_i, \sigma) \\
\mu_i &= \alpha + \beta_f \mathrm{food} + \beta_g \mathrm{groupsize} \\
\alpha &\sim \mathrm{Normal}(0, 0.2) \\
\beta_f &\sim \mathrm{Normal}(0, 0.5) \\
\beta_g &\sim \mathrm{Normal}(0, 0.5) \\
\sigma &\sim \mathrm{HalfCauchy}(1)
\end{aligned}
$$

Multiple regression results

Here are the results from the multiple regression:

	mean	sd	hdi_3%	hdi_97%
bG	-0.568	0.189	-0.937	-0.224
bF	0.475	0.188	0.153	0.859
alpha	0.001	0.090	-0.176	0.160
sigma	0.967	0.068	0.854	1.103

neg. association
$G \leftrightarrow W$

pos. association
$F \leftrightarrow W$

- Conditioned on group size, more food means bigger foxes
- Conditioned on food, bigger group means smaller foxes.

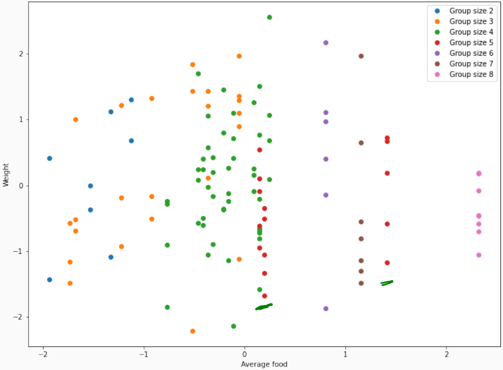
Multiple regression results

Here are the results from the multiple regression:

	mean	sd	hdi_3%	hdi_97%
bG	-0.568	0.189	-0.937	-0.224
bF	0.475	0.188	0.153	0.859
alpha	0.001	0.090	-0.176	0.160
sigma	0.967	0.068	0.854	1.103

Can we see this in the scatter plot?

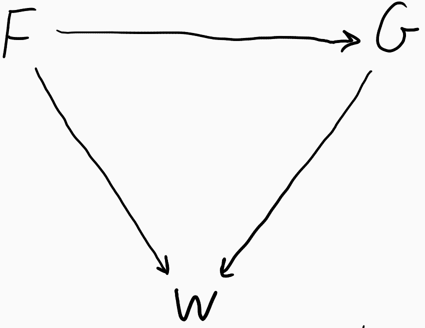
Statistical control as stratification



The association between food and weight appears when the data is stratified by group size, but not before

DAG for the fox model

Let's draw a DAG for the fox model:



Total effect of F on W
- Direct effect F -> W
- Indirect effect F -> G -> W

Direct vs. total effect

The graph for the fox model gives us a distinction between two effects of food on weight:

- direct effect: associated with the arrow from F to W; effect of food on weight at fixed group size
  - estimated by the multiple regression, but not the simple regression
- total effect: associated with all directed paths from F to W; effect of food on weight, including those mediated by changes in group size
  - estimated by the simple regression, but not the multiple regression

Revealing and eliminating associations

In both cases we gained something by including the extra variable:

- In the marriage example, a spurious association is eliminated
- In the fox example, a masked association is revealed

This is the power of multiple regression: it thinks "hypothetically" about the variables

Beware: including the wrong variables can introduce spurious associations!

Structure of DAGs

What is a DAG?

What is a DAG?

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

What can we use DAGs for? Probabilistic models, on two levels:

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

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

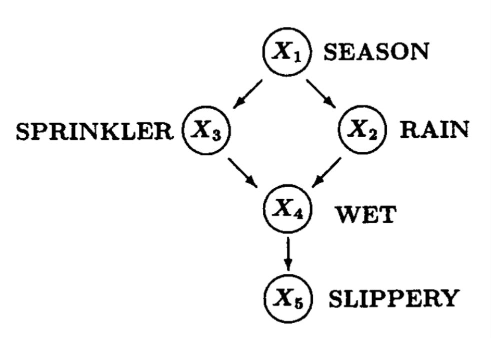


Figure from *Causality* by Judea Pearl

Graphical example

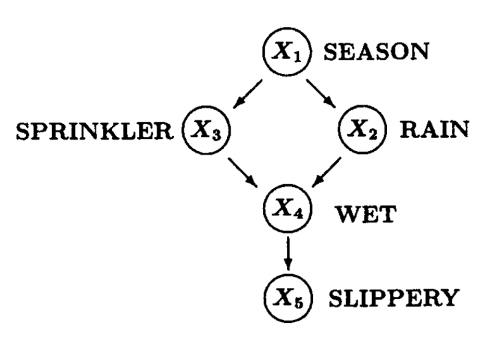


Figure from *Causality* by Judea Pearl

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

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

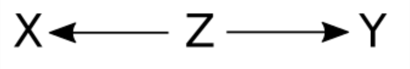
In a DAG, information flows along paths (both with and against the arrows).

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:

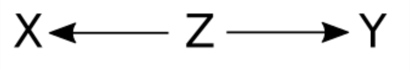


A statistical association exists between X and Y because they are both influenced by Z.

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:

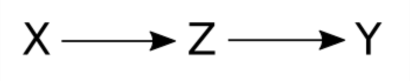


Conditional independence:

- 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:



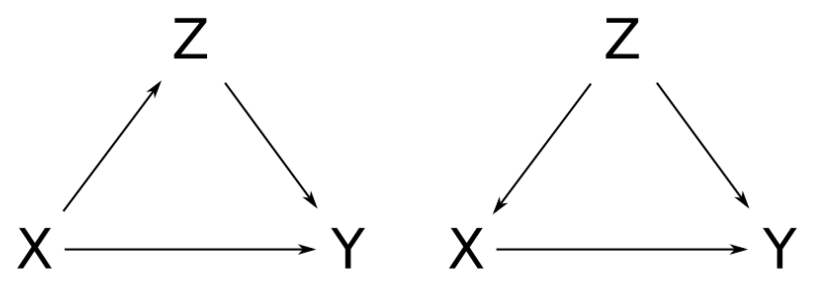
Typical case: Z is an effect of X that mediates the effect on Y

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?



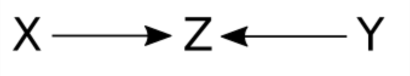
The data/model cannot tell you the difference between these, because they imply the same set of conditional independences

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

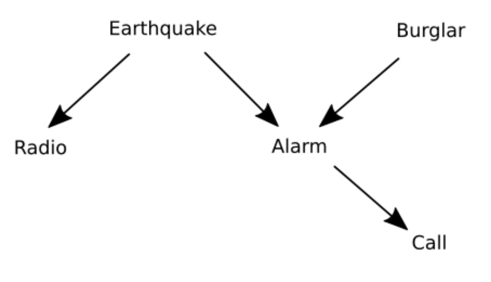
Explaining away: the burglar alarm

From Judea Pearl by way of David 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

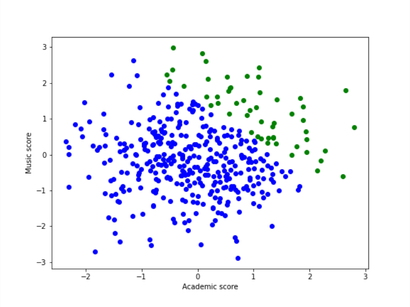
A DAG for the burglar alarm problem, showing the collider:



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

This led to the potential for collider bias. Why?

- Any study of confirmed COVID-19 cases can only be applied to people who are tested (still true!)
- Data sets are implicitly conditional on having been tested

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: who is getting tested in early 2020?

Examining the effect of smoking

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

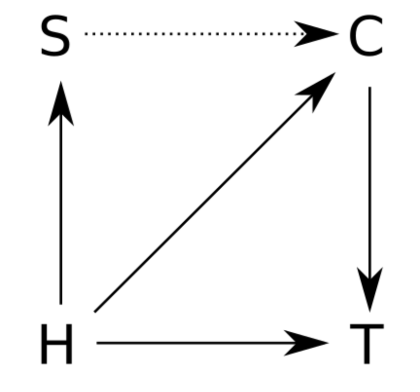
- people with severe disease
- healthcare workers

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 $X_i \rightarrow M \rightarrow X_j$ or fork $X_i \leftarrow M \rightarrow X_j$ such that $M \in Z$; or,
2. p contains a collider $X_i \rightarrow M \leftarrow X_j$ 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.

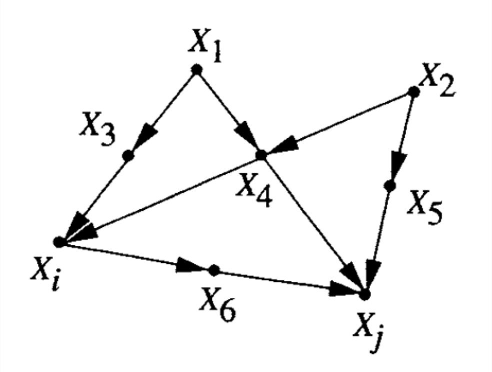
A related definition:

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.*

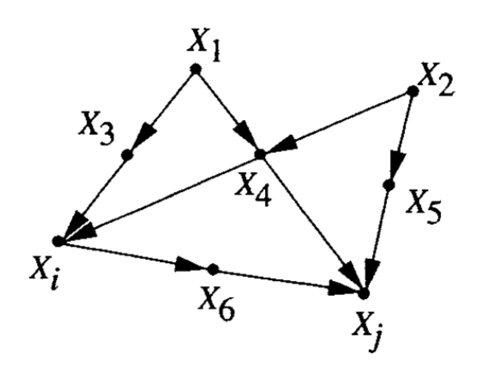
To estimate the causal effect of X on Y, condition on a set of variables satisfying the backdoor criterion with respect to (X, Y).

Example



Which variables satisfy the backdoor criterion?

Example



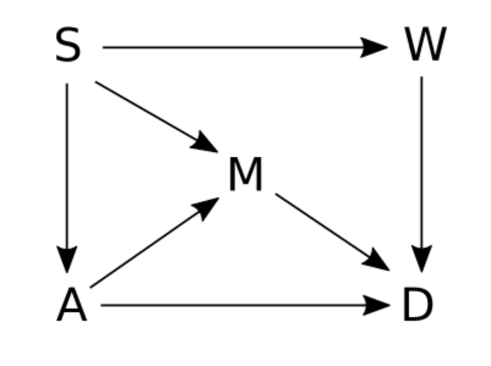
Which variables satisfy the backdoor criterion?

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

Return to the Waffle House

A bigger DAG from the Waffle House example, including:

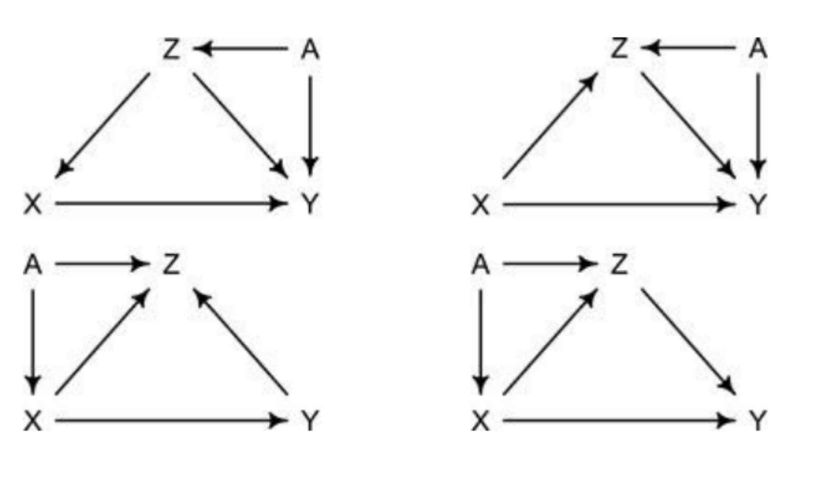
- W: number of Waffle Houses in the state
- S: Indicator variable for South



To estimate the direct effect of W on D, what do we condition on?

Group exercise

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



Summary

Today:

- Multiple regression
- Causal DAGs

Next time:

- Structure of DAGs and the backdoor criterion

Pearl