# Parameterizing Causal Marginal Models

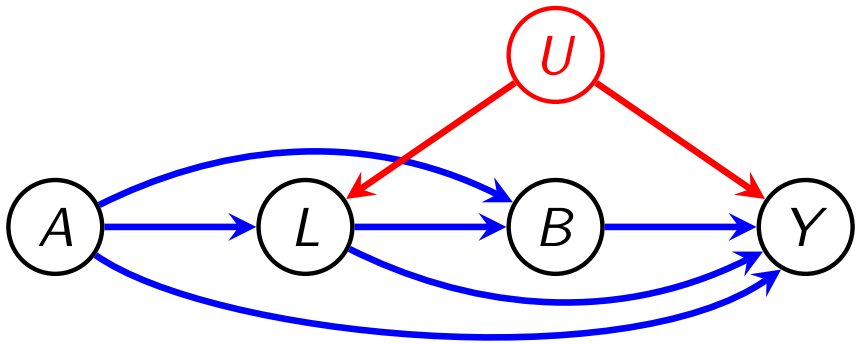
Robin Evans, University of Oxford
Bohao Yao, University of Oxford
Vanessa Didelez, Leibniz Institute, Bremen

MRC-BSU Seminar
University of Cambridge
16th June 2020

# Outline

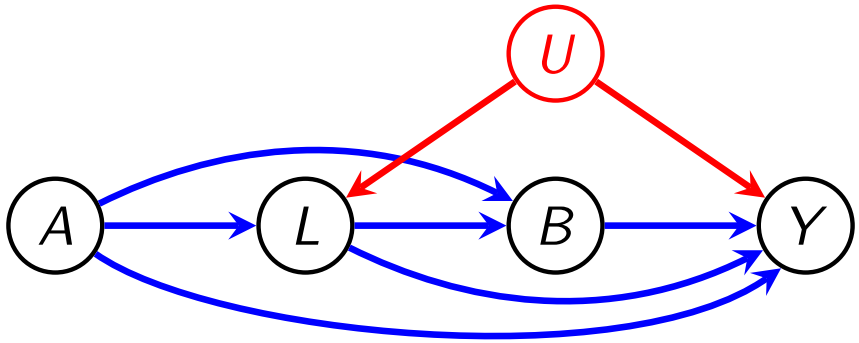
# Causal Models

Take a simple two-step dynamic treatment model.




- $A, B$ treatments (randomised);
- $L$ intermediate outcome;
- $Y$ final outcome;
- $U$ unobserved confounders.

# Identification




**Question:** how do the treatments causally affect the final outcome? Or, if we treated everyone with $(a, b)$, what would happen?

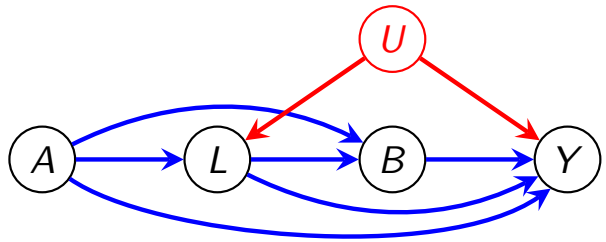
We want $P(Y \mid do(A = a, B = b))$. How do we identify this?

- $P(Y \mid A = a, B = b)$: ignoring/marginalizing $L$;
- $P(Y \mid A = a, B = b, L = \ell)$: conditioning on $L$.

**Neither** has the desired causal interpretation!

# Identification

We can 'reweight' a sample or distribution to pretend that $B$ was assigned independently of $A$ and $L$:




$$P^*(a, \ell, b, y) = P(a, \ell, b, y)\frac{P(b)}{P(b \mid a, \ell)} = P(y \mid a, \ell, b)P(b)P(\ell \mid a)P(a).$$

In this new 'world', $L$ is post-treatment, so just ignore it!

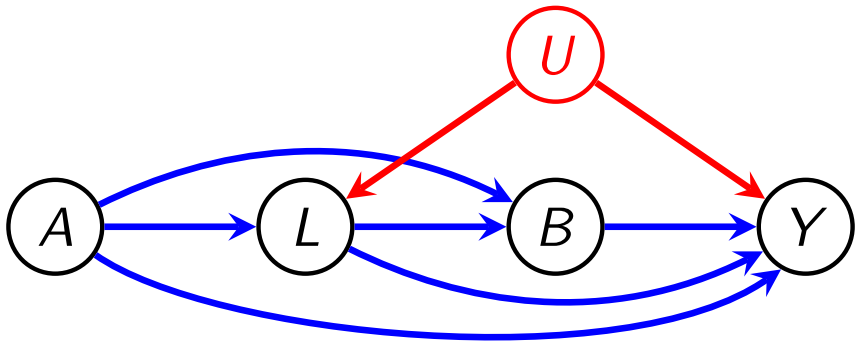
Then $P^*(y \mid a, b)$ **does** have the desired causal interpretation:

$$P^*(y \mid a, b) = \sum_{\ell} P^*(\ell, y \mid a, b) = \sum_{\ell} P(y \mid a, \ell, b) \cdot P(\ell \mid a).$$

This example is due to Robins (1986); more general results are available (see, e.g., Shpitser and Pearl, 2006).

# Parameterizing Causal Models

For likelihood-based inference and simulation, need a parameterization.




Standard parameterizations lead to the **g-null paradox**.

For example, with

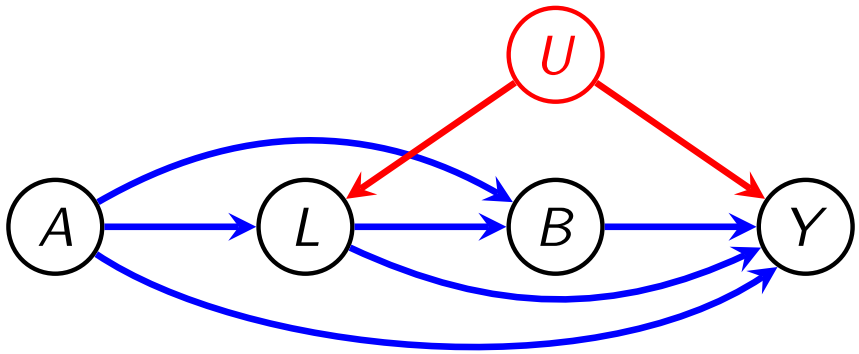
- linear model for $Y$ given $A, B, L$;
- any model for binary $L$ given continuous, unbounded $A$;

then it is almost impossible for $P(Y \mid do(A = a, B = b))$ **not** to depend upon $A$ except in trivial cases (Robins and Wasserman, 1997).

Naturally, this is disastrous for hypothesis testing.

# Simulation

Havercroft and Didelez (2012) note that even simulating data from model so that $P(y \mid do(a, b))$ independent of $A$ is sometimes difficult.




Why?

Simulation requires $P(a, \ell, b, y)$;
relationship to $P(y \mid do(a, b))$ seems complicated.

g-null paradox shows we can't just specify a nice parametric model for $P$ and then fix parameters until independence holds.

# Recast the Problem

Define

$$\begin{aligned} P^*(y, \ell \mid a, b) &\equiv P(y, \ell \mid do(a, b)) \\ &= P(y \mid a, \ell, b) \cdot P(\ell \mid a). \end{aligned}$$

**Message:** $P^*$ is *just* a (conditional) probability distribution.

## Desired Properties of $P^*$

- nice model for $P^*(y \mid a, \ell, b) = P(y \mid a, \ell, b)$ for simulation.
- nice model for $P^*(y \mid a, b)$ for statistical inference;
- nice model for $P^*(\ell \mid a, b) = P(\ell \mid a)$ to ensure $L \perp\!\!\!\perp B \mid A\,[P^*]$.

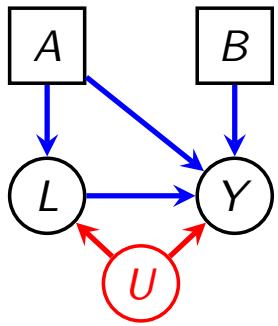
So how do we get this?

Short answer: **we can't**! It doesn't make sense to try to specify $P^*(y \mid a, \ell, b)$ and $P^*(y \mid a, b)$ separately.

# Margins

A better way to think about this: given interventional distribution $P^*$ suppose we have:

- a model for $P^*(y \mid a, b)$;
- a model for $P^*(\ell \mid a, b) = P(\ell \mid a)$;




These do not fully specify $P^*(y, \ell \mid a, b)$ so what else do we need?

**Answer:** some sort of dependence measure:

$$\phi_{LY|AB}(\ell, y \mid a, b);$$

e.g. a copula or the odds ratio.

Any additional information given by $P(y \mid a, \ell, b)$ is then **redundant**.

# A Principled Approach

For our problem, separately specify (nice, parametric) models for:

- $P(a, \ell, b)$;
- $P(y \mid do(a, b))$;
- $\phi_{LY|AB}$ (some dependence measure, e.g. the conditional odds ratio).

This is variation independent, and has no redundancy.

Modelling $\phi_{LY|AB}$ is data-dependent, but:

- discrete case: use **odds ratios** (Bergsma and Rudas, 2002);
- Gaussian case: **partial correlation** $\rho_{LY \cdot AB}$;
- general $A, B$, continuous $L, Y$: **copula** models.

# Marginal Tension

We will see that there is generally a tension between:

- simple specification of the **joint distribution**, in order to facilitate simulation and likelihood-based inference;
- simple specification of the **target of inference** (i.e. some marginal quantity) in order that it is interpretable;
- enforcing marginal **constraints** implied by the causal model.

# Weight Function

Of course, the 'margins' we are interested in are non-standard.

## Definition

Let $w(z \mid x)$ be a smooth function of $P(x, z)$, with:

K1. $w(z \mid x) \geq 0$;

K2. $\int w(z \mid x)\, dz = 1$ for each $x$;

K3. $w(z \mid x) \cdot P(x)$ equivalent to $P(x, z)$.

Note that K1, K2 imply that $w$ is a **kernel**.

# Cognate Probabilities

## Definition

We say $P^*(y \mid x)$ is **cognate** to $P(y \mid x)$ (within $P(z, x, y)$) if

$$P^*(y \mid x) \equiv \int P(y \mid x, z) \cdot w(z \mid x)\, dz.$$

for some $w$ satisfying K1–K3.

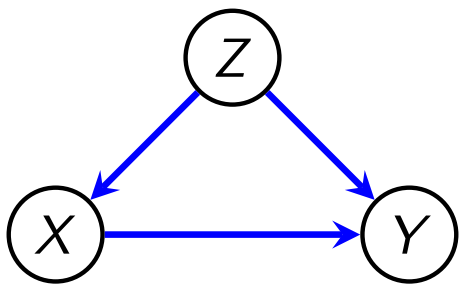
# Cognate Probabilities: Examples

**Examples**

$$P(y \mid x) = \sum_z P(y \mid x, z) \cdot P(z \mid x)$$

$$P(y \mid x, c_0) = \sum_z P(y \mid x, z, c_0) \cdot P(z \mid x, c_0)$$

$$P(y \mid do(x)) = \sum_z P(y \mid x, z) \cdot P(z).$$



# Results

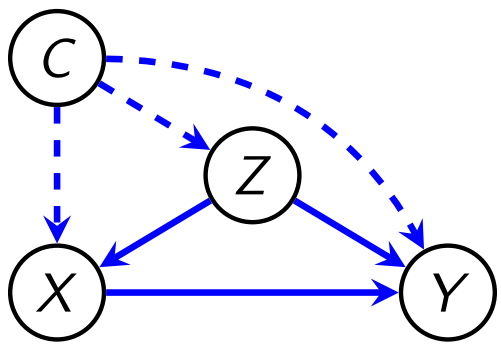
## Theorem

Consider an outcome $Y$, and causally prior variables $C, X, Z$. Then we can smoothly parameterize the joint distribution $P(c, z, x, y)$ with:

$$P(c, z, x) \qquad \underbrace{P^*(y \mid c, x)}_{\text{cognate to } P(y \mid c,x)} \qquad \underbrace{\phi_{ZY|CX}(z, y \mid c, x)}_{P^*\text{-dependence measure}}$$

and these three pieces are variation independent of one another.

Any of $C, X, Z, Y$ can be vector valued.




This gives us the **best of both worlds**: a coherent joint distribution and a marginal specification of our choice.

# Sketch Proof

We have $P(c, z, x)$, from which we can compute $w(z \mid c, x)$.

Note also, that

$$P^*(c, z, x, y) = P(c, z, x, y) \frac{w(z \mid c, x)}{P(z \mid c, x)}.$$

We can (smoothly) recover the left hand side from $w(z \mid c, x)$, $P^*(y \mid x, c)$ and $\phi_{ZY|CX}$ just by using the inverse map (e.g. IPF will work with the odds ratios).

Now, since we know $P(c, z, x)$ and $w(z \mid c, x)$, we can recover $P$.

Variation independence follows from results of Csiszár (1975) with odds ratios.

# Results

A corollary of this is that we can always parameterize a different cognate probability for each new variable.

Let $X_1, \ldots, X_m$ be the different variables, and $\boldsymbol{K}_i \subseteq \{X_1, \ldots, X_{i-1}\}$.

## Corollary

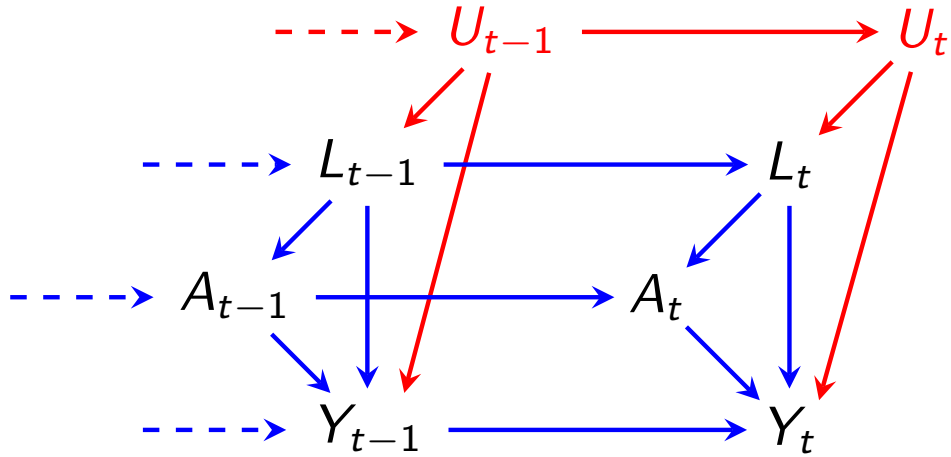
We can obtain a variation independent parameterization which includes

$$P^*(x_i \mid \boldsymbol{k}_i) \qquad \forall i = 1, \ldots, m,$$

provided each $P^*(x_i \mid \boldsymbol{k}_i)$ is cognate to $P(x_i \mid \boldsymbol{k}_i)$.

# Example: Survival Models

Young and Tchetgen Tchetgen (2014) consider survival models:




What is probability of survival ($Y = 1$) to next time point, given treatment?

$$P(Y_t = 1 \mid Y_{t-1} = 1, do(a_1, \ldots, a_t)).$$

No problem! What remains is the dependence structure between $L$'s and $Y_t$ given $A_1, \ldots, A_t$.

# Example: Survival Models

Hence simulation in some cases becomes relatively easy under a null; e.g.:

$$P(Y_t \mid Y_{t-1} = 1, do(a_1, \ldots, a_t)) = P(Y_t \mid Y_{t-1} = 1).$$

Young and Tchetgen Tchetgen note that this is not at all trivial.

Can also easily incorporate, for e.g., a **stationarity assumption**:

$$P(Y_t \mid Y_{t-1} = 1, do(A_t = a)) = g(a).$$

# Variation Independence and Covariates

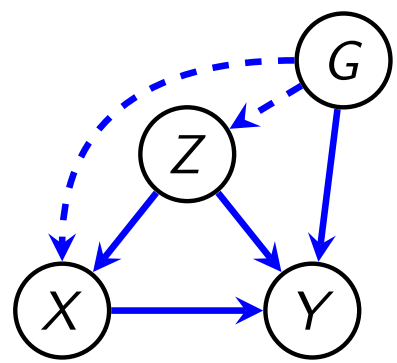
The variation independence is useful:

- easy to incorporate covariates in GLM form;
- no danger of choosing impossible interaction parameters (so no g-null paradox!);
- means independent priors are valid.

Example, suppose want to model:

$$\text{logit}\, P(Y = 1 \mid do(x), g) = f(x, g);$$

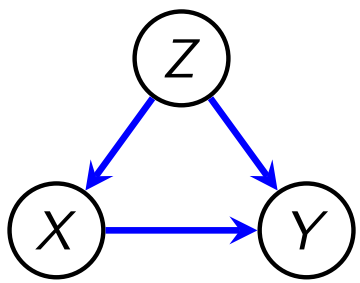
i.e. how is causal effect of $X$ on $Y$ modulated by $G$?
We can do this with a logistic regression.

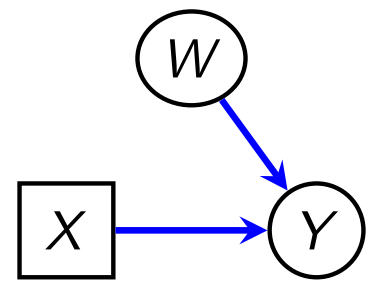


# Multiple Experiments and Transportability

The parameterization approach is also important if we want to combine information from different experimental settings with some (but not all) parameters in common.

For example, observational and randomized trials on $X$:




Might assume that $P(y \mid do(x))$ common to both settings, fit the model and do a likelihood ratio test.

# Results

## Proposition

Let $\phi_{ZY|CX}$ be the odds ratio parameters.

Then these are the same in $P$ as in $P^*$.

**Proof sketch.** Take the binary case. We have

$$\begin{aligned}
\log \phi_{ZY|CX} &= \sum_{(c,z,x,y)\in\{0,1\}^4} (-1)^{|(c,z,x,y)|} \log P^*(c,z,x,y) \\
&= \sum_{(c,z,x,y)\in\{0,1\}^4} (-1)^{|y|+|(c,z,x)|} \log P^*(c,z,x) P^*(y \mid c,z,x) \\
&= \sum_{(c,z,x,y)\in\{0,1\}^4} (-1)^{(c,z,x,y)|} \log P\,(y \mid c,z,x),
\end{aligned}$$

since the $\log P^*(c,z,x)$ terms all cancel one another, and $P^*(y \mid c,z,x) = P(y \mid c,z,x)$.

# How Do We Simulate?

In practice, if $X$ or $Y$ is continuous we need to use rejection sampling.

1. First, simulate data from $P^*(z, x, y)$.
2. Then determine

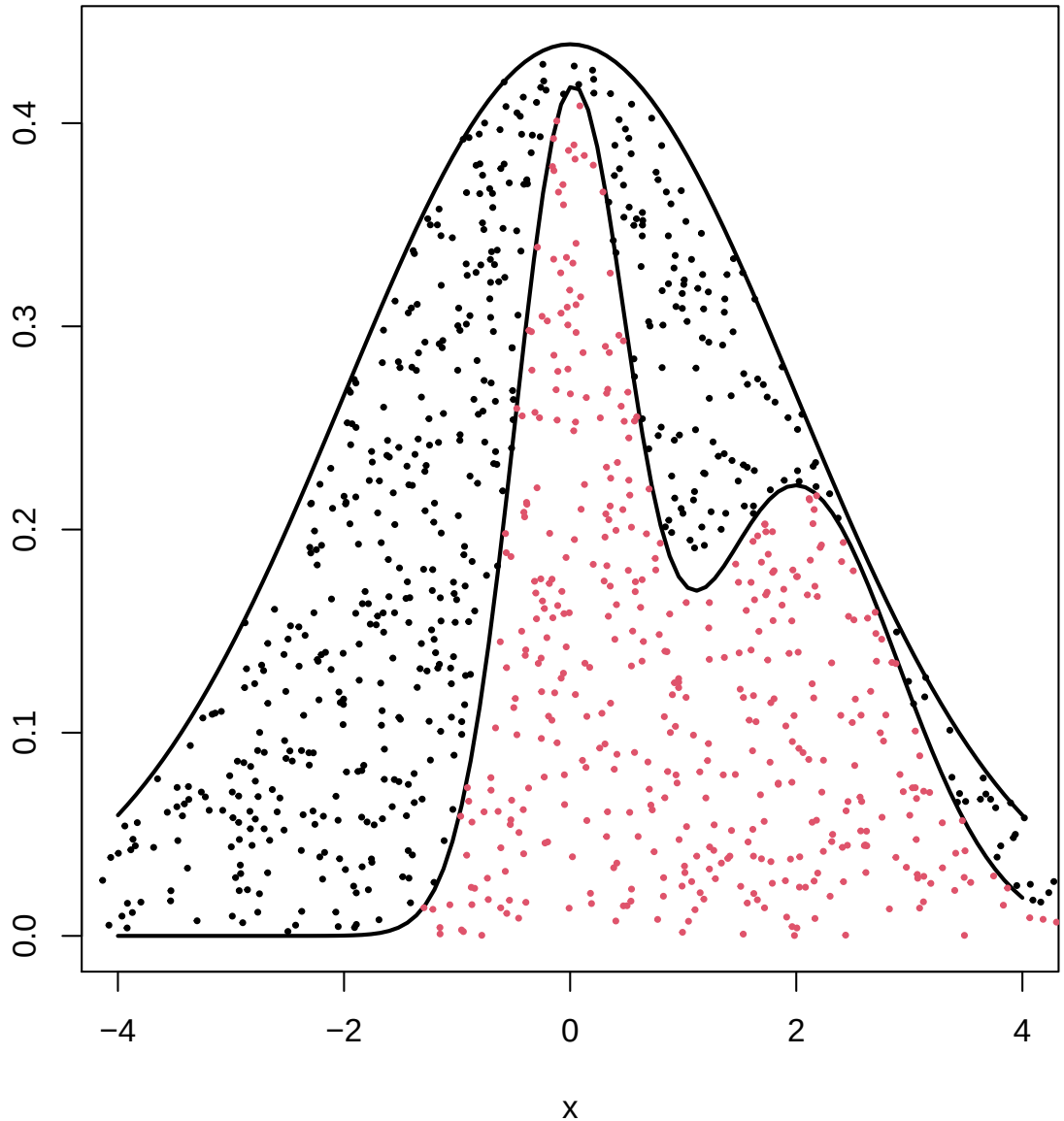
$$M \equiv \sup_{z,x} \frac{P(z, x)}{P^*(z, x)}.$$

3. For each sample $(z_i, x_i, y_i)$, simulate an independent $U_i \sim \text{Unif}(0, 1)$ and reject if

$$U_i > M^{-1} \frac{P(z_i, x_i)}{P^*(z_i, x_i)}.$$

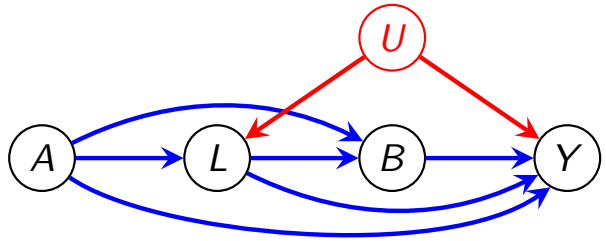
Notice that this doesn't involve $Y$, so the causal distribution is preserved.

# Rejection Sampling



# Copula Model Example

Take the two-step dynamic model from Havercroft and Didelez (2012).




We choose:

- $A, B \sim \text{Bernoulli}(\frac{1}{2})$;
- Gaussian copula model:

$$\begin{pmatrix} \Phi^{-1}(U) \\ \Phi^{-1}(L') \\ \Phi^{-1}(Y') \end{pmatrix} \sim N\left(0, \begin{pmatrix} 1 & 0.4 & 0.5 \\ & 1 & 0.3 \\ & & 1 \end{pmatrix}\right);$$

- $L \mid A = a \sim \text{Exp}(\exp(-0.3 + 0.75a))$;
- $Y \mid do(A = a, B = b) \sim \text{Exp}(\exp(-0.5 + 0.2a + 0.3b))$.
- $B \mid A = a, L = \ell \sim \text{Bernoulli}(\text{expit}(-1.5 + 0.4a + \ell))$;

# Copula Model Example

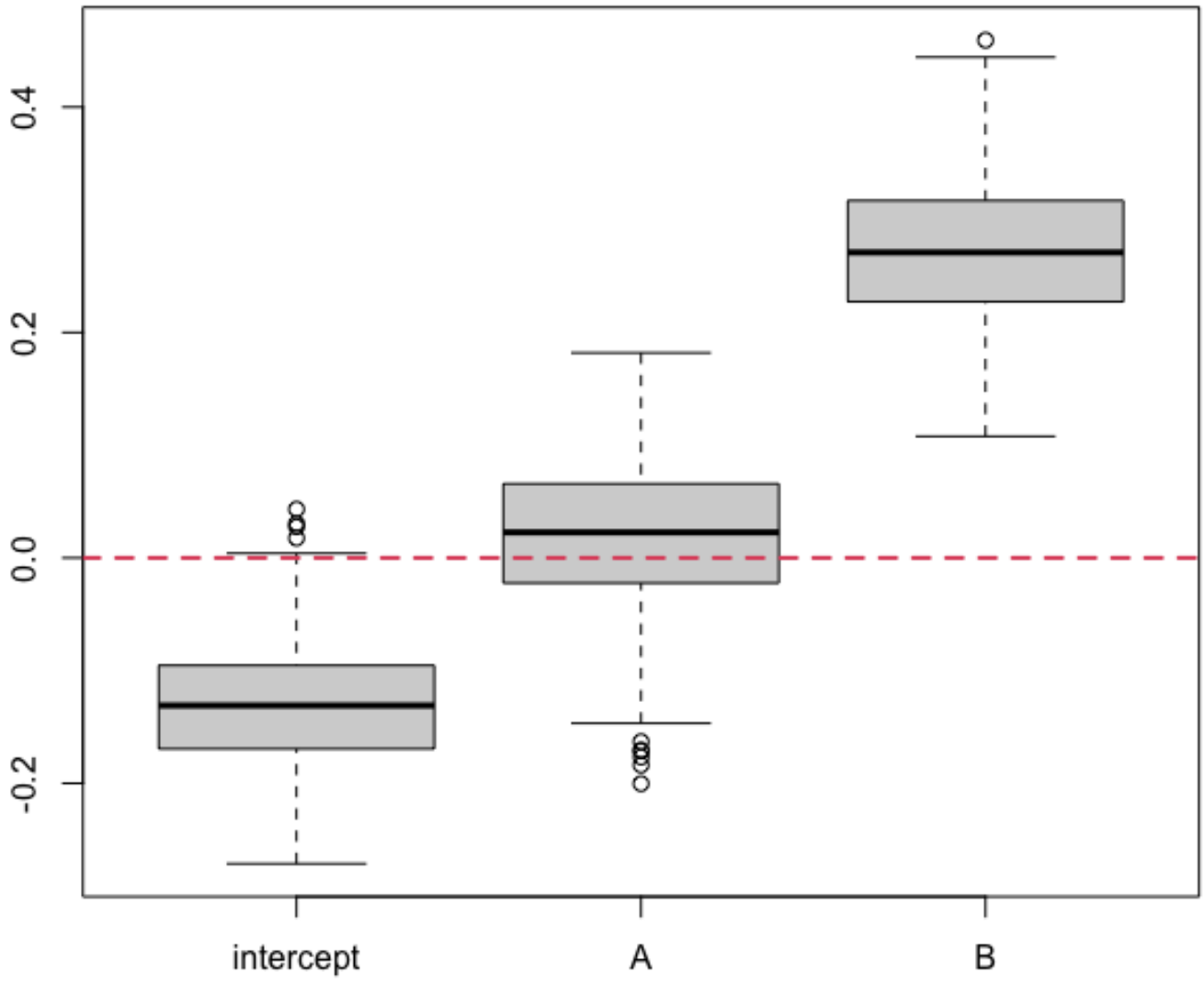
Suppose we simulate $n = 10^4$ observations this way.

If we fit an ordinary gamma GLM with $\log \mathbb{E} Y = \beta_0 + \beta_1 a + \beta_2 b$, then the results are wrong:

| coefficient | truth | estimate | std err. | p-value |
|---|---|---|---|---|
| intercept | 0.5 | 0.357 | 0.017 | $< 10^{-16}$ |
| $A$ | -0.2 | -0.315 | 0.020 | $1.30 \times 10^{-3}$ |
| $B$ | -0.3 | -0.018 | 0.020 | $< 10^{-16}$ |

# Copula Model Example

Bias over 1,000 fits to simulated data ($n = 10^3$).



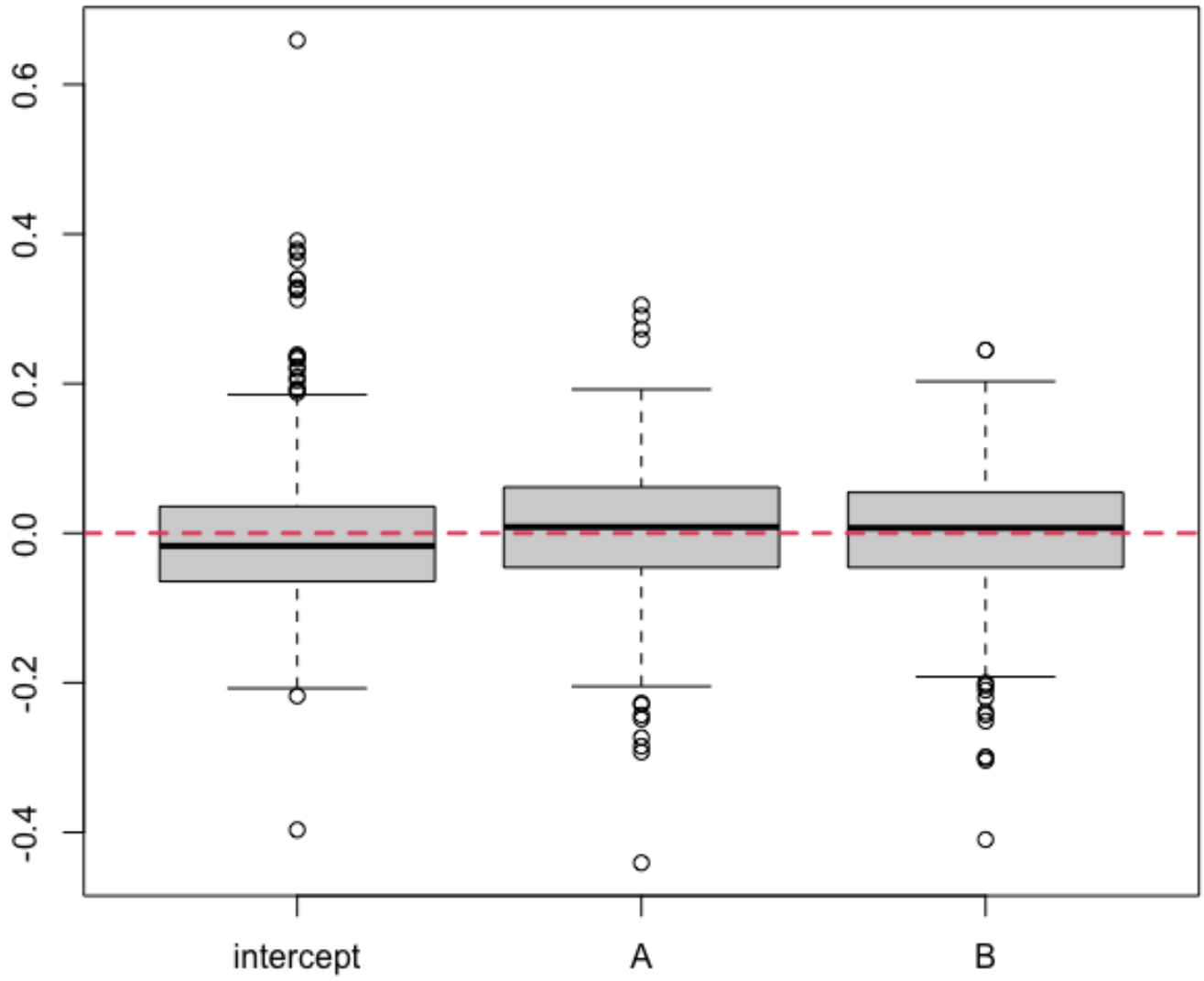
# Copula Model Example

Alternatively, if we fit a reweighted GLM with bootstrapped standard errors to the $n = 10^4$ data, the results are fine!

| coefficient | truth | estimate | std err. | p-value |
|---|---|---|---|---|
| intercept | 0.5 | 0.559 | 0.054 | 0.277 |
| $A$ | -0.2 | -0.209 | 0.041 | 0.828 |
| $B$ | -0.3 | -0.342 | 0.040 | 0.290 |

# Copula Model Example

Bias over 1,000 fits to simulated data ($n = 10^3$).



# Survival Model Example

Consider a survival model with binary treatment $X_i$, and measured covariate $Z_i$. We take $Y_i = 1$ to mean the patient has survived.

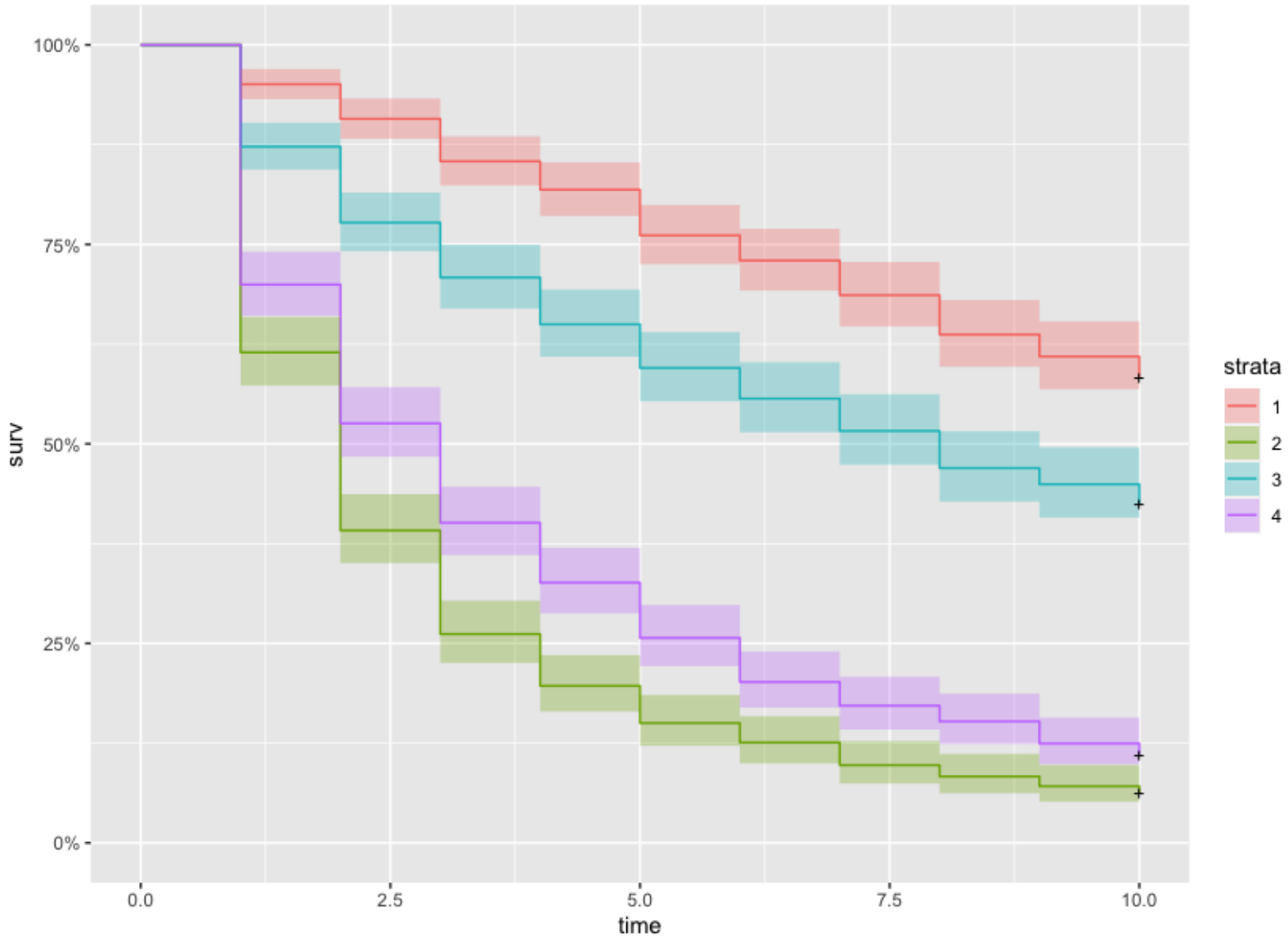
We pick:

- $Z_i$ is a stationary Gaussian process (mean 0, variance 1);
- $X_1 \sim \text{Bernoulli}(\text{expit}(2Z_1))$;
- $X_i \sim \text{Bernoulli}(\text{expit}(-10 + 5X_{i-1}))$;
- $Y_i \mid X_i = 0 \sim \text{Bernoulli}(\text{expit}(1 + \frac{1}{2}X_{i-1} + Z_i))$;
- $Y_i \mid X_i = 1 \sim \text{Bernoulli}(\text{expit}(3 + \frac{1}{2}X_{i-1} + 2Z_i))$.

We can compare this to:

- $X_1 \sim \text{Bernoulli}(\frac{1}{2})$.

# Survival Plot



# Summary

- **Causal models are marginal models** (most of the time!);
- there is a large literature on marginal models to look at for other cases.
- This has applications to marginal structural models, survival models, dynamic treatment regimes, structural nested mean models, stationarity, transportability...;
- simulation becomes much easier in Gaussian, discrete cases, some copula models.

- Limitation: with continuous outcomes this method (generally) relies on rejection sampling, which may be inefficient in higher dimensions.

# Thank you!

# References


Bergsma and Rudas. Marginal log-linear parameters, *Ann. Statist.*, 2002.

Bergsma, Croon, Hagenaars. Advancements in Marginal Modeling for Categorical Data, *Sociological Methodology*, 2013.

Csiszár. *I*-Divergence Geometry of Probability Distributions and Minimization Problems, *Ann. Prob.*, 1975.

Evans, Yao, Didelez. Parameterizing Causal Marginal Models, 2020. (check arXiv soon!)

Havercroft and Didelez. Simulating from marginal structural models with time-dependent confounding, *Stat. Med.*, 2012.

Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect, *Math. Modelling*, 1986.

Robins and Wasserman. Estimation of Effects of Sequential Treatments by Reparameterizing DAGs, *UAI*, 1997.

Shpitser and Pearl, Identification of Joint Interventional Distributions in Recursive Semi-Markovian Causal Models, *AAAI*, 2006.

Young and Tchetgen Tchetgen. Simulation from a known Cox MSM using standard parametric models for the g-formula, *Stat. Med.*, 2014.

# Generalising Odds Ratios

Let $p$ be a density for $X, Y$.

The **odds ratio** for $X, Y$ is the equivalence class of functions $\phi_{XY}$ such that

$$\phi_{XY}(x, y) = p(x, y) \cdot u(x) \cdot v(y).$$

some functions $u, v > 0$.

Some points to note:

- defined for any distribution with a density;
- $p$ is a member of the equivalence class;
- there's no requirement for $p$ to be positive;
- iterative proportional fitting recovers the joint distribution.

# Specifying Margins

Let $r_{XY}(x, y)$ be a joint distribution with odds ratio $\phi_{XY}$.

**Theorem**

Let $p_X$ and $p_Y$ be densities such that $p_X \ll r_X$ and $p_Y \ll r_Y$. Then there exists a unique joint distribution with margins $p_X$, $p_Y$ and odds ratio $\phi_{XY}$.

This follows from Csiszár (1975).

This is a form of **variation independence**: we can paste together essentially any dependence structure with any margins and get a distribution.

# Examples

- For discrete variables this reduces to the 'usual' odds ratio;
- for Gaussian variables:

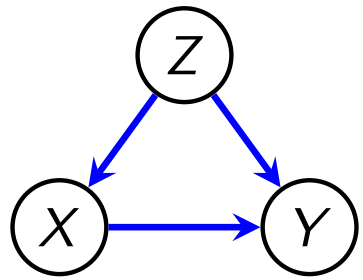
$$\phi_{XY} \sim \exp\left(\frac{\rho xy}{\sigma_x \sigma_y (1-\rho^2)}\right)$$

- multivariate $t$-distribution ($\boldsymbol{x} = (x, y)^T$):

$$\phi_{XY} \sim \left(1 + \nu^{-1} \boldsymbol{x}^T \Sigma^{-1} \boldsymbol{x}\right)^{-\nu/2-1}$$

# Margins

Let's think about the simplest example of this kind.




$$P(y \mid do(x)) = \sum_z P(z)P(y \mid x, z).$$

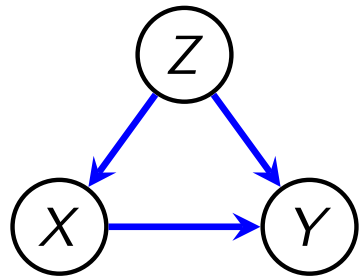
This is a 'margin' of the joint distribution

$$P^*(z, y \mid x) \equiv P(z)P(y \mid x, z).$$

To work with $P^*$ we need to model the $XY$-margin (because that's the quantity of interest) and the $XZ$-margin (to enforce the independence).

So what's left to know?

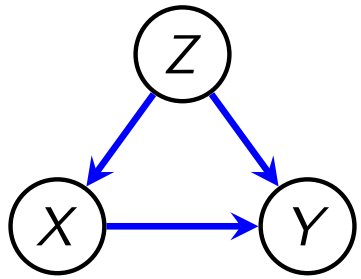
# Odds Ratios




Bergsma and Rudas' results show that the remaining information is precisely the odds ratio between $Y$ and $Z$ conditional upon $X$.

Attempting to specify any additional information given this, $P(y \mid do(x))$ and $P(x, z)$ doesn't really make any sense.

# Odds Ratios




But there's nothing to stop us specifying that the parameters $\beta$ and $\gamma$ are from this model:

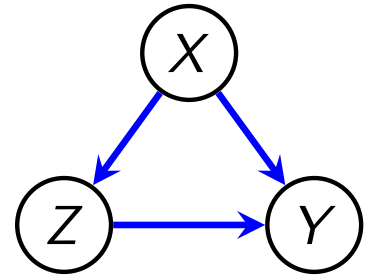
$$\text{logit}\, P(y \mid x, z) = \mu + \alpha x + \beta z + \gamma xz.$$

But $\mu$ and $\alpha$ are **not free**.

Take home - you can have part of a nice model on $X, Y, Z$ just don't expect all of it!

# g-null Paradox Illustration

Suppose that we have continuous $X$ and $Y$, but binary $Z$.




An innocuous seeming model would be:

$$\mathbb{E}[Y \mid X = x, Z = z] = \mu + \beta x + \gamma z.$$

But:

$$\begin{aligned}\mathbb{E}[Y \mid X = x] &= \sum_z \mathbb{E}[Y \mid X = x, Z = z] \cdot P(Z = z \mid X = x) \\ &= \mu + \beta x + \gamma P(Z = 1 \mid X = x).\end{aligned}$$

Now $P(Z = 1 \mid X = x)$ can't be a linear function of $x$ (unless it's constant). So $\mathbb{E}[Y \mid X = x]$ is only a linear function if either:

- $Z \perp\!\!\!\perp X$; or
- $\gamma = 0$ (so $Y \perp\!\!\!\perp Z \mid X$).