CAUSAL THINKING

Linking scientific and statistical models

Diogo Melo
Lewis-Sigler Institute of Integrative Genomics
damelo@princeton.edu

WHAT ARE MODELS FOR?

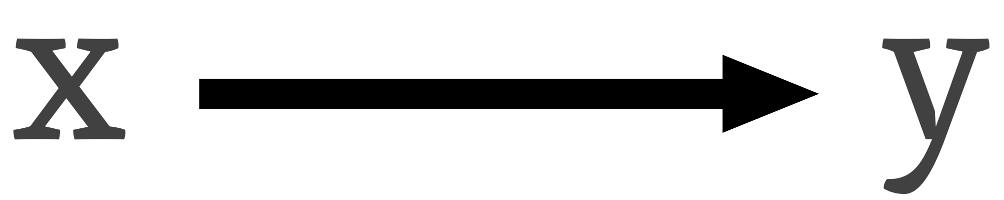
PREDICTION VS CAUSAL INFERENCE

CORRELATION AND CAUSATION

Why does correlation not imply causation?

GRAPH MODEL REPRESENTATION

- We can use graphs to represent our putative causal model.
- An arrow between variables represents a potential causal effect.



This is a Directed Acyclic Graph, a **DAG**

ELEMENTAL TRIADS

- All DAGs can be decomposed into a set of 3 elemental motifs:
    - The pipe, the fork and the collider
- We can use these to structure our thinking about our models and decide what variable to include or exclude

The pipe:

$$x \longrightarrow z \longrightarrow y$$

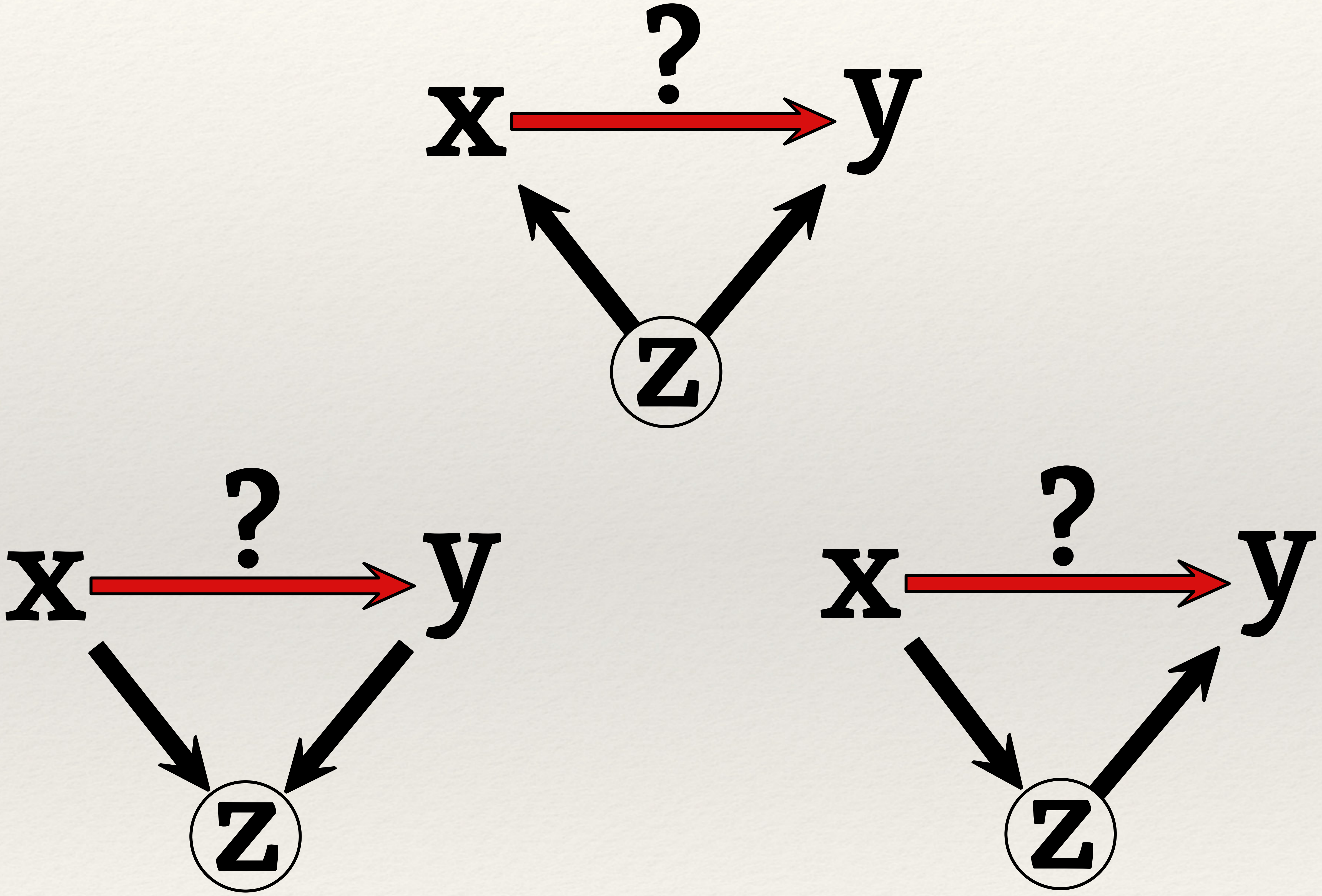
The fork:

$$x \longleftarrow z \longrightarrow y$$

The collider:

$$x \longrightarrow z \longleftarrow y$$

HOW DOES A CONFOUNDER AFFECT OUR ESTIMATE OF THE EFFECT OF X ON Y?

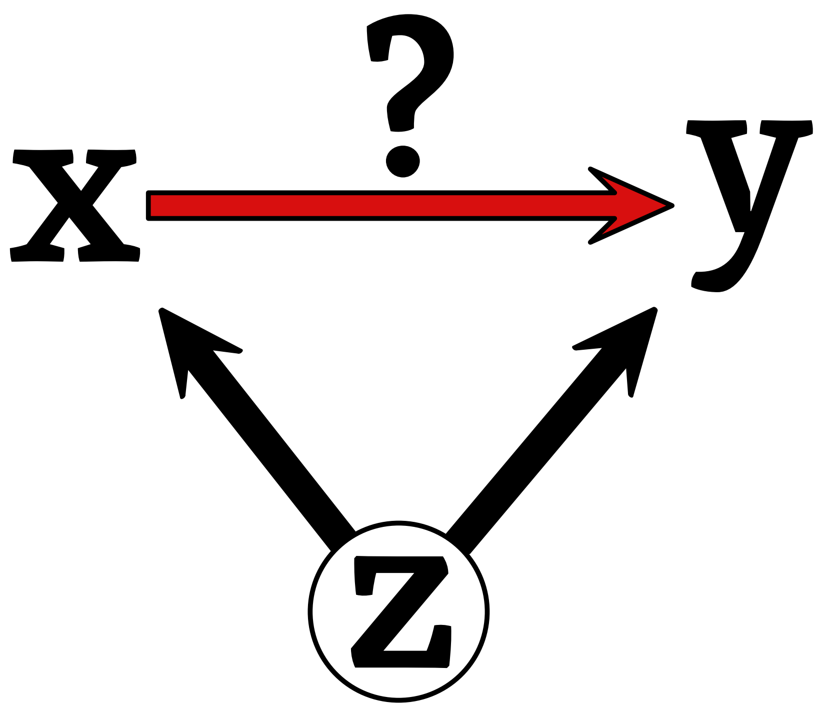


THE FORK

SIMULATING THE EFFECT OF A FORK

- Every DAG implies a causal relation between variables.
- We can use distributions to simulate the generative model implied by this DAG:

$$y \sim Normal(\alpha_y + \beta_{yx}x + \beta_{yz}z, \sigma_y)$$
$$x \sim Normal(\alpha_x + \beta_{xz}z, \sigma_x)$$
$$z \sim Normal(\alpha_z, \sigma_z)$$



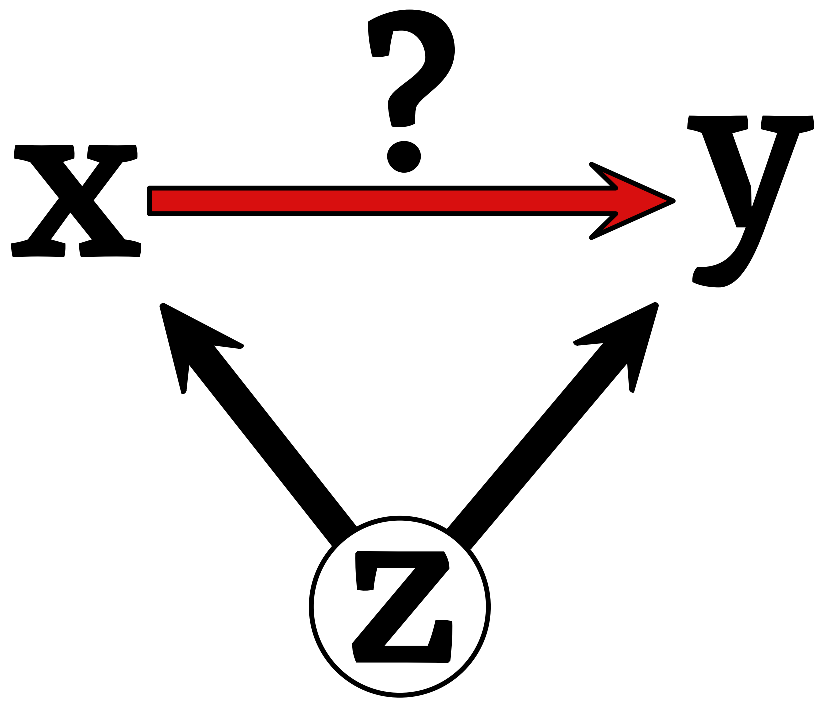
SIMULATING THE EFFECT OF A FORK

Math

$$y \sim Normal(\alpha_y + \beta_{yx}x + \beta_{yz}z, \sigma_y)$$

$$x \sim Normal(\alpha_x + \beta_{xz}z, \sigma_x)$$

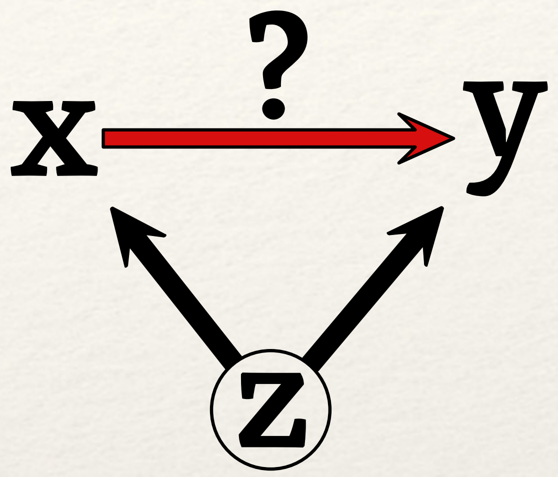
$$z \sim Normal(\alpha_z, \sigma_z)$$



R Code

```
N = 100
z = rnorm(N)                # z ~ normal(0, 1)
x = rnorm(N, 1 + z)         # x ~ normal(1 + z, 1)
y = rnorm(N, 1 + x + z) # y ~ normal(1 + x + z, 1)
```

STATISTICAL MODEL WITHOUT THE CONFOUNDER Z



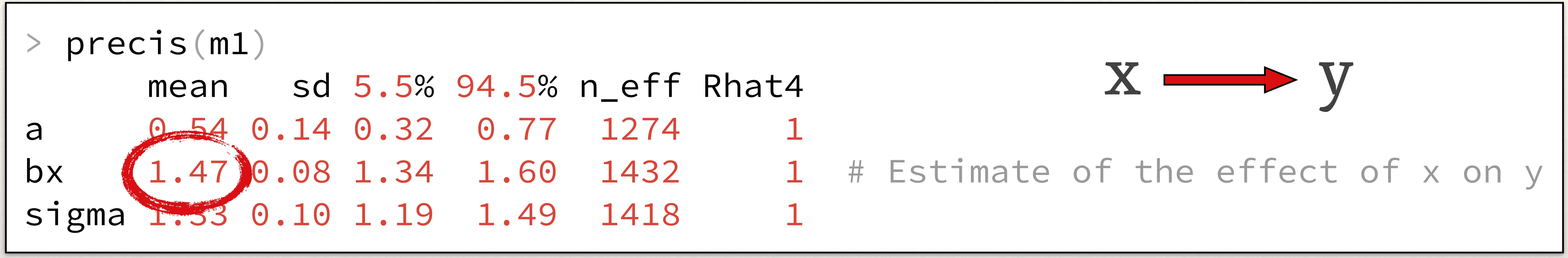
Math $\quad x \rightarrow y$

$$y \sim Normal(\mu, \sigma)$$
$$\mu = a + bx$$
$$a \sim Normal(0,0.3)$$
$$b \sim Normal(0,0.3)$$
$$\sigma \sim Exponential(1)$$

Rethinking Code $\quad x \rightarrow y$

```
m1 = ulam(alist(
    y ~ normal(mu, sigma),
    mu = a + bx*x,
    a ~ normal(0, 0.3),
    bx ~ normal(0, 0.3),
    sigma ~ exponential(1)
), data = list(y = y, x = x))
```

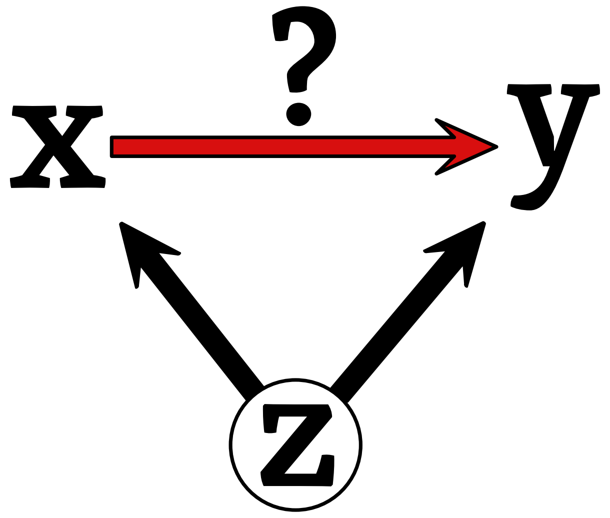
MODEL ESTIMATES WITHOUT THE CONFOUNDER



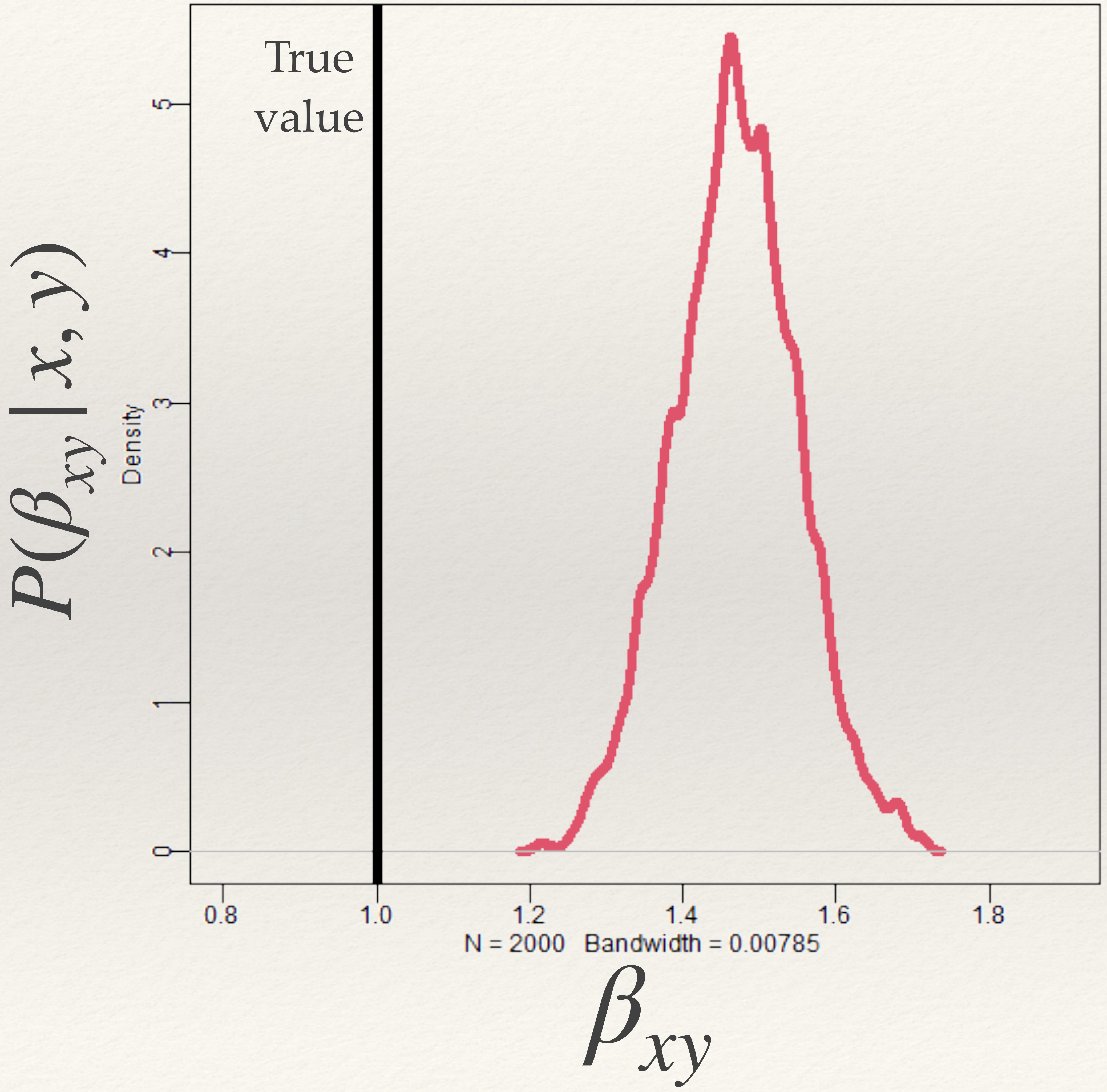
```
> precis(m1)
       mean   sd 5.5% 94.5% n_eff Rhat4
a      0.54 0.14 0.32  0.77  1274     1
bx     1.47 0.08 1.34  1.60  1432     1  # Estimate of the effect of x on y
sigma  1.33 0.10 1.19  1.49  1418     1
```

Simulation R code

```
N = 100
z = rnorm(N)            # z ~ normal(0, 1)
x = rnorm(N, 1 + z)     # x ~ normal(1 + z, 1)
y = rnorm(N, 1 + x + z) # y ~ normal(1 + x + z, 1)
```

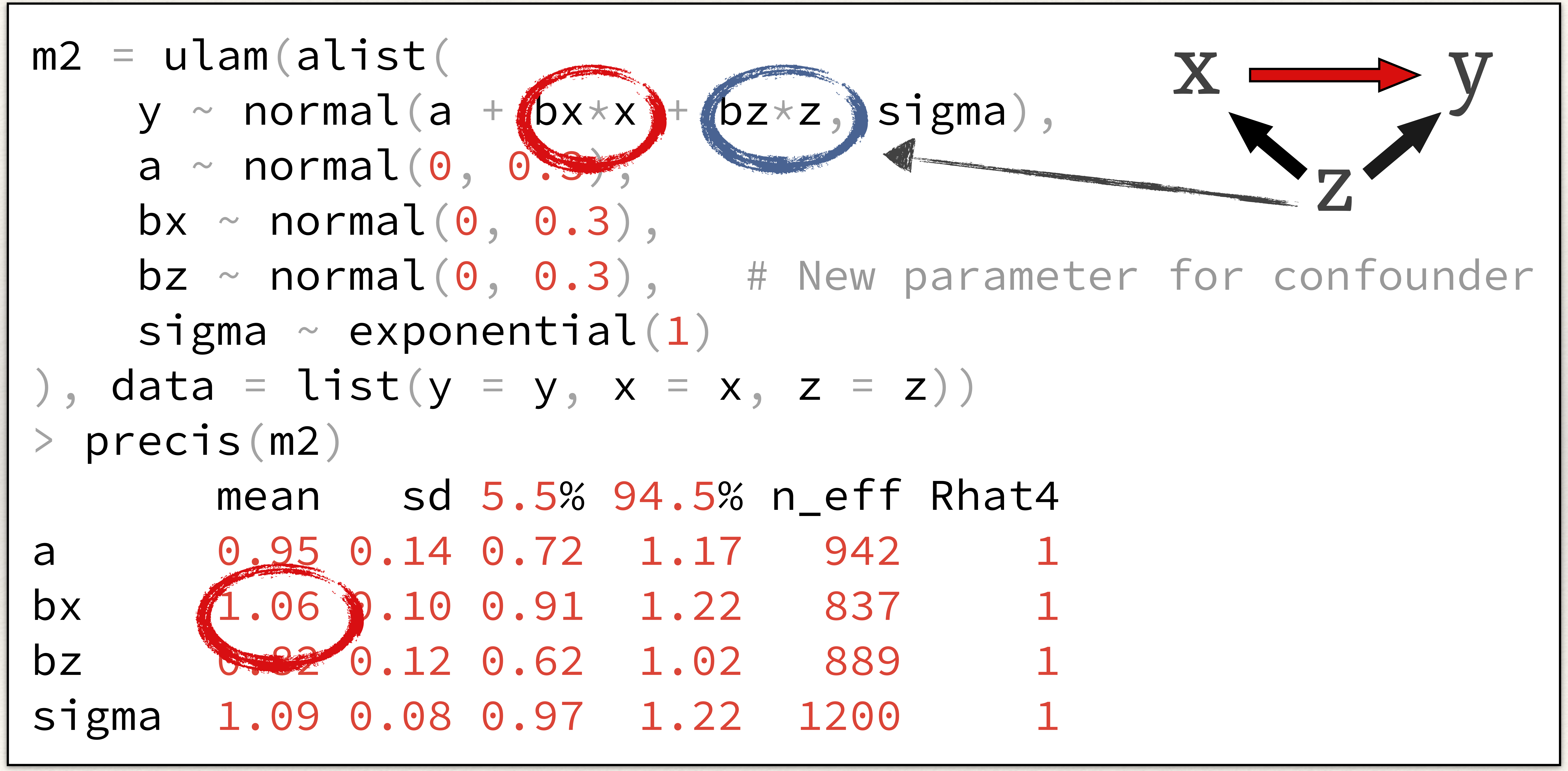


POSTERIOR DISTRIBUTION OF β_{xy} WITHOUT THE CONFOUNDER



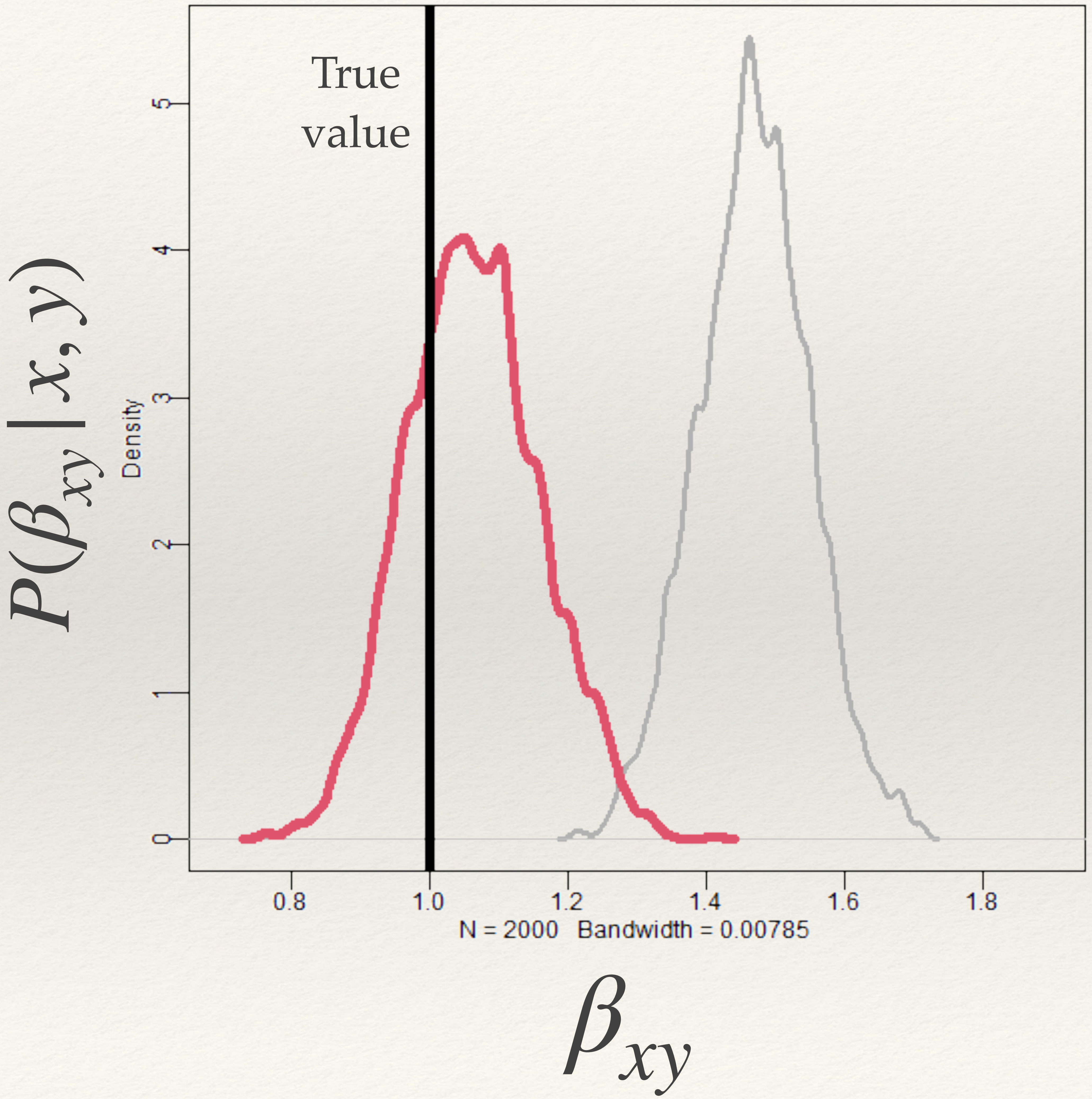
x → y

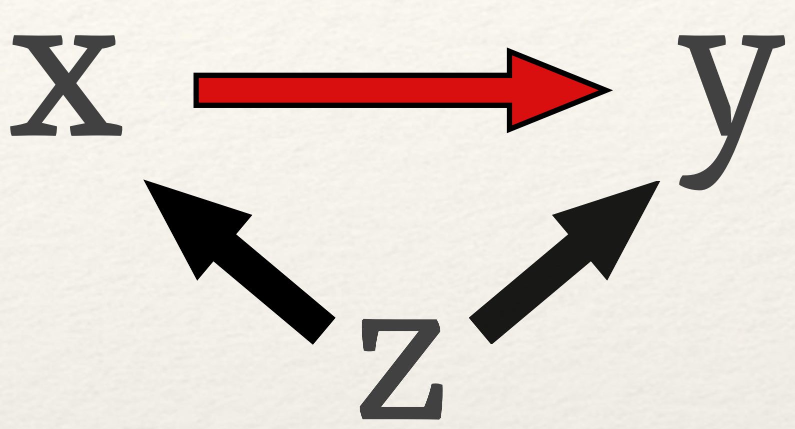
INCLUDING THE CONFOUNDER



```
m2 = ulam(alist(
    y ~ normal(a + bx*x + bz*z, sigma),
    a ~ normal(0, 0.3),
    bx ~ normal(0, 0.3),
    bz ~ normal(0, 0.3),    # New parameter for confounder
    sigma ~ exponential(1)
), data = list(y = y, x = x, z = z))
> precis(m2)
       mean   sd 5.5% 94.5% n_eff Rhat4
a      0.95 0.14 0.72  1.17   942     1
bx     1.06 0.10 0.91  1.22   837     1
bz     0.82 0.12 0.62  1.02   889     1
sigma  1.09 0.08 0.97  1.22  1200     1
```

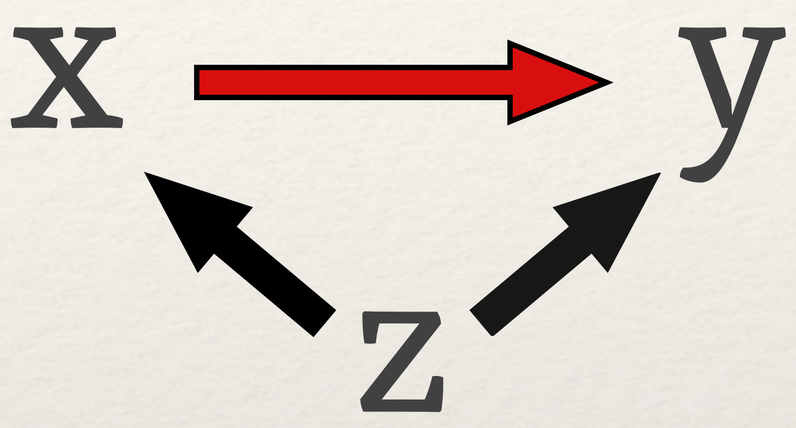
POSTERIOR DISTRIBUTION OF β_{xy} WITH THE CONFOUNDER





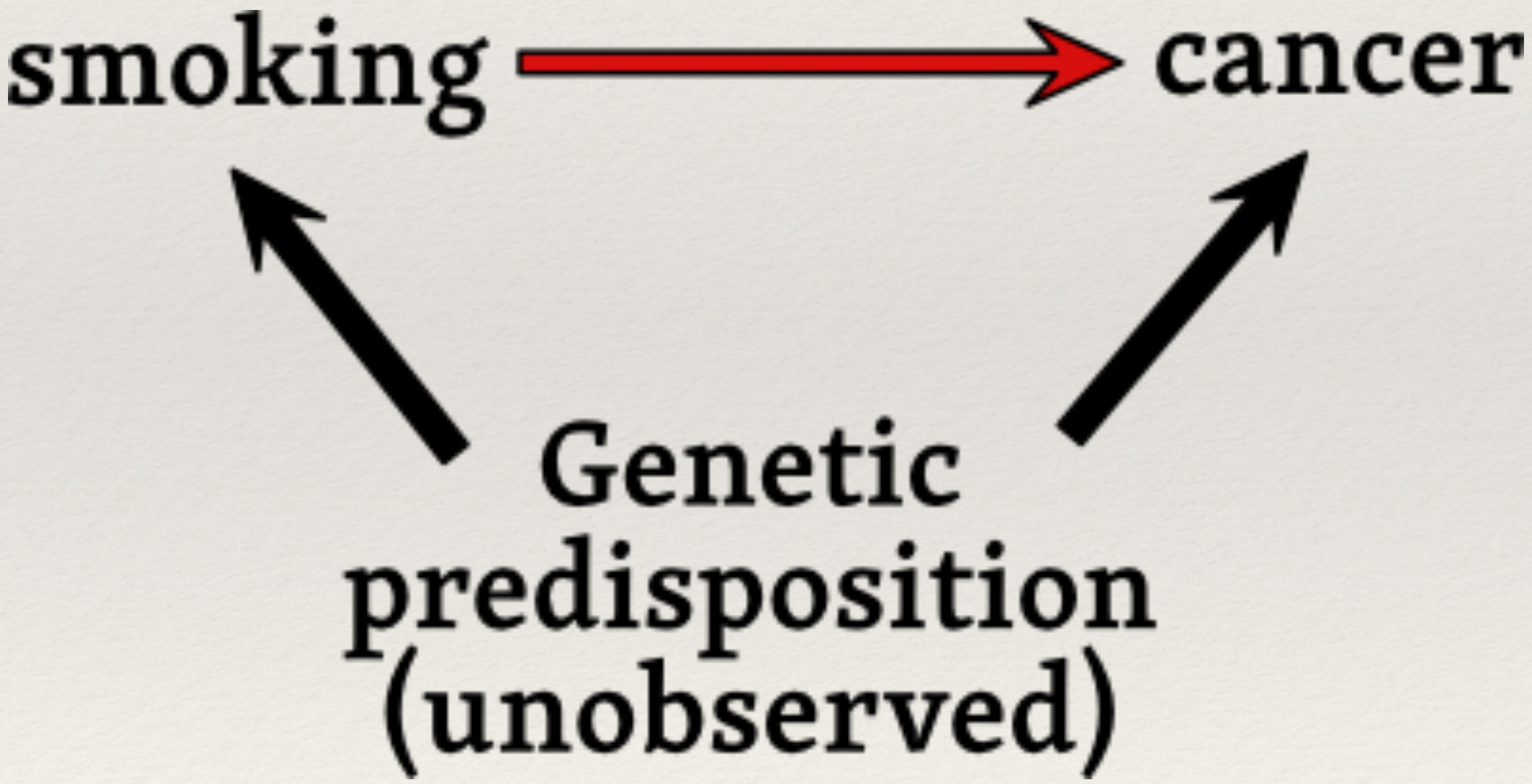
EXAMPLES OF FORKS OR CONFOUNDERS

- This is the quintessential “control” variable.
- Most variables are included in the model under the assumption that they are confounders and need have their effects taken into consideration.

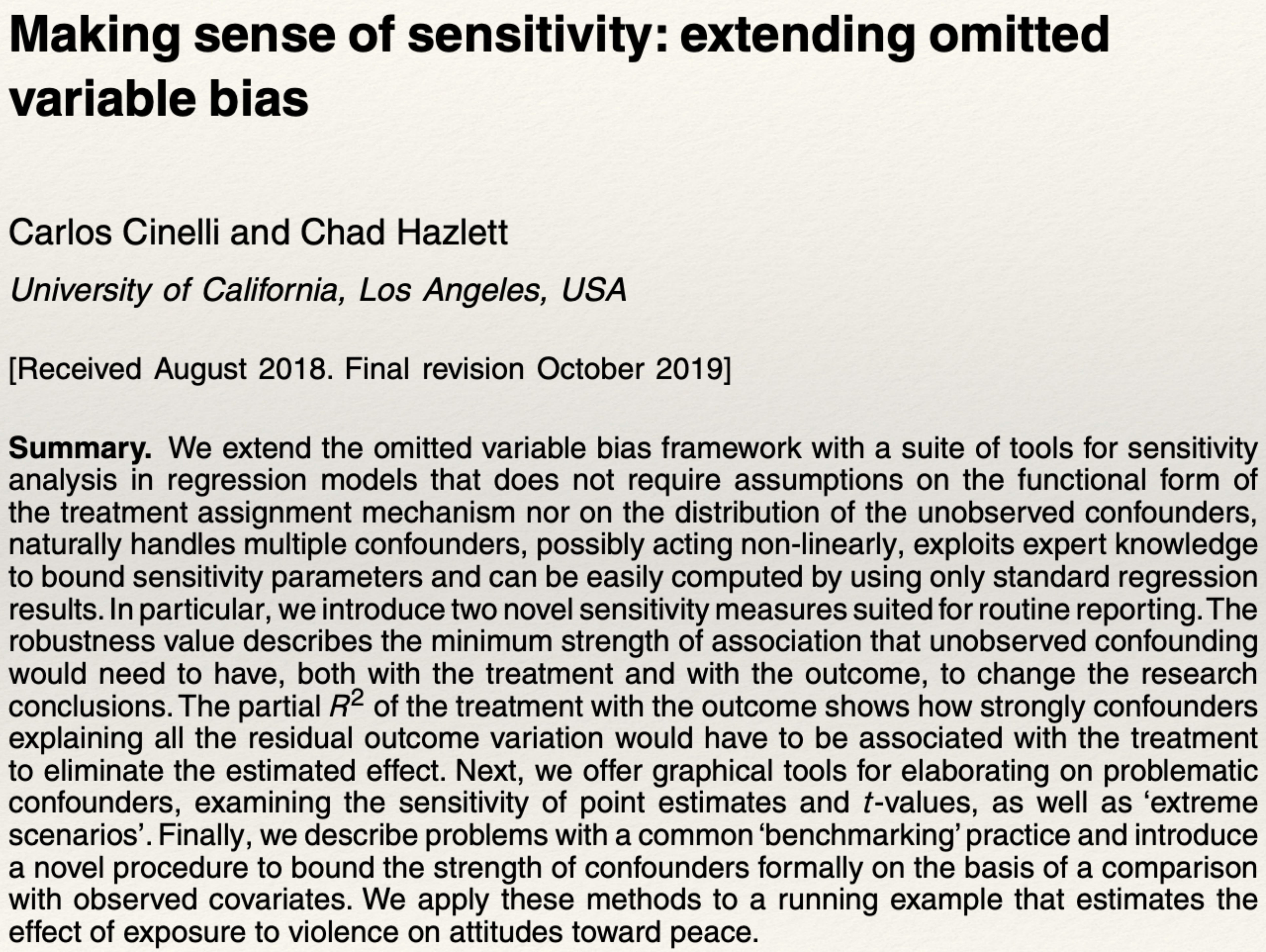


SMOKING AND CANCER

Famously, R. A. Fisher was not convinced that smoking caused cancer, and proposed that an unobserved propensity variable caused both cancer and smoking



OMITTED VARIABLE BIAS



Making sense of sensitivity: extending omitted variable bias

Carlos Cinelli and Chad Hazlett

University of California, Los Angeles, USA

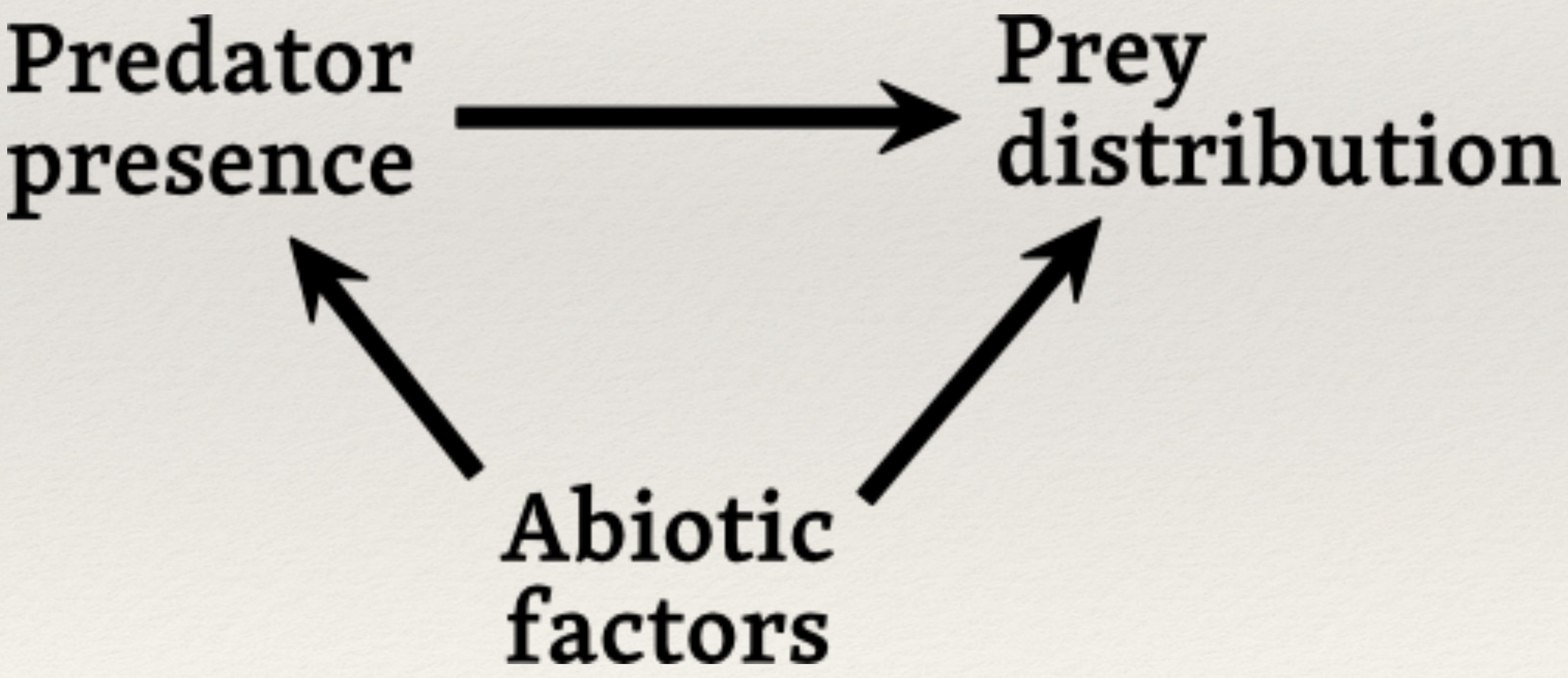
[Received August 2018. Final revision October 2019]

Summary. We extend the omitted variable bias framework with a suite of tools for sensitivity analysis in regression models that does not require assumptions on the functional form of the treatment assignment mechanism nor on the distribution of the unobserved confounders, naturally handles multiple confounders, possibly acting non-linearly, exploits expert knowledge to bound sensitivity parameters and can be easily computed by using only standard regression results. In particular, we introduce two novel sensitivity measures suited for routine reporting. The robustness value describes the minimum strength of association that unobserved confounding would need to have, both with the treatment and with the outcome, to change the research conclusions. The partial R^2 of the treatment with the outcome shows how strongly confounders explaining all the residual outcome variation would have to be associated with the treatment to eliminate the estimated effect. Next, we offer graphical tools for elaborating on problematic confounders, examining the sensitivity of point estimates and t-values, as well as 'extreme scenarios'. Finally, we describe problems with a common 'benchmarking' practice and introduce a novel procedure to bound the strength of confounders formally on the basis of a comparison with observed covariates. We apply these methods to a running example that estimates the effect of exposure to violence on attitudes toward peace.

Cinelli, C. & Hazlett, C. Making sense of sensitivity: extending omitted variable bias. J. R. Stat. Soc. Series B Stat. Methodol. ***82****, 39–67 (2020)*

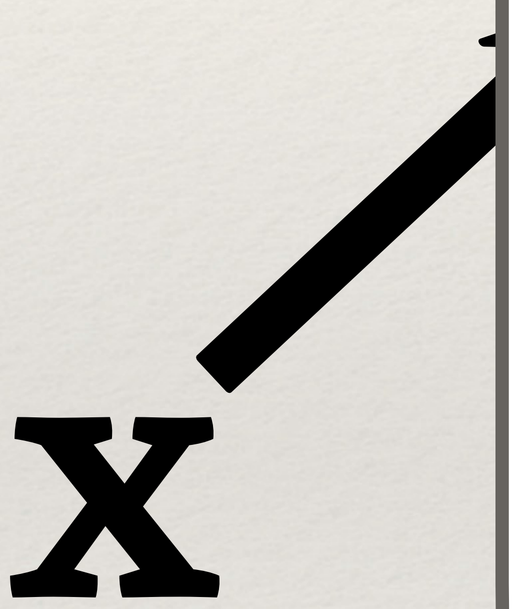
SPECIES DISTRIBUTION

Maybe want to evaluate the effect of some predator on the distribution of a prey. But the spacial distribution of both the predator and the prey are affected by some abiotic factor:



PIPE

ALL THE EFFECT OF X ON Y IS MEDIATED BY Z



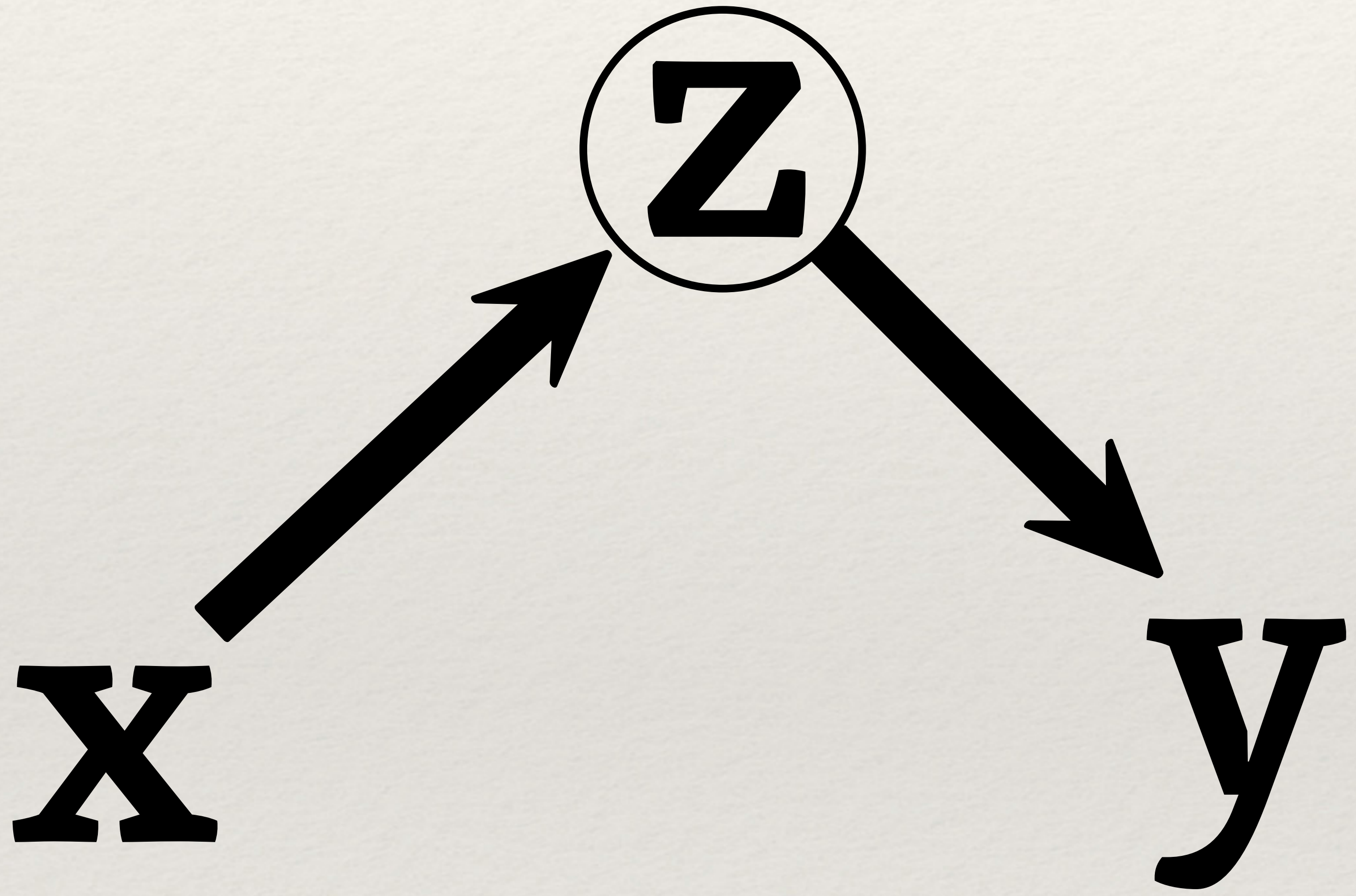
Math

$$y \sim Normal(\alpha_y + \beta_{yz} z, \sigma_y)$$

$$z \sim Normal(\alpha_z + \beta_{zx} x, \sigma_z)$$

$$x \sim Normal(\alpha_x, \sigma_x)$$

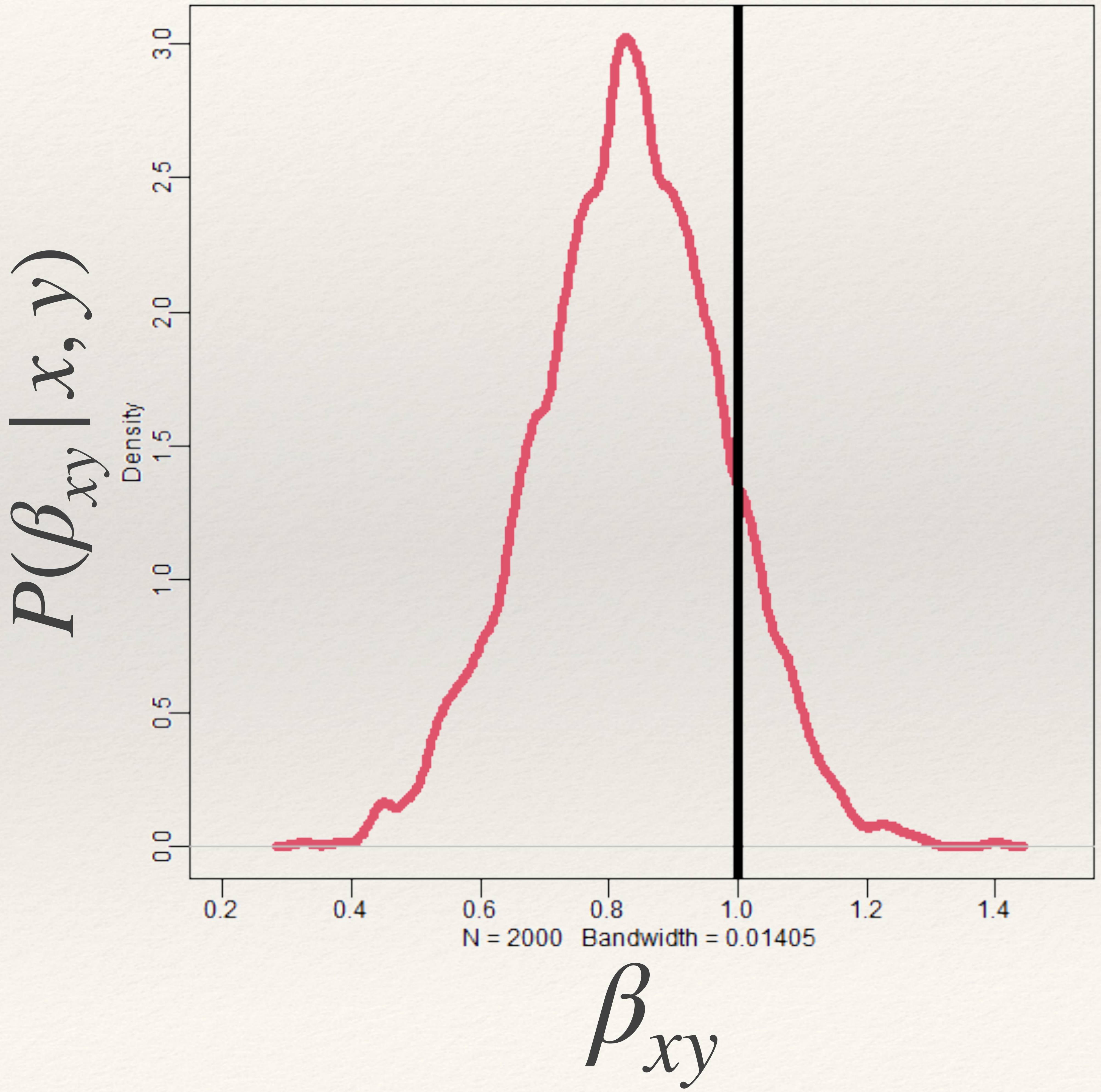
MODEL WITHOUT THE MEDIATOR



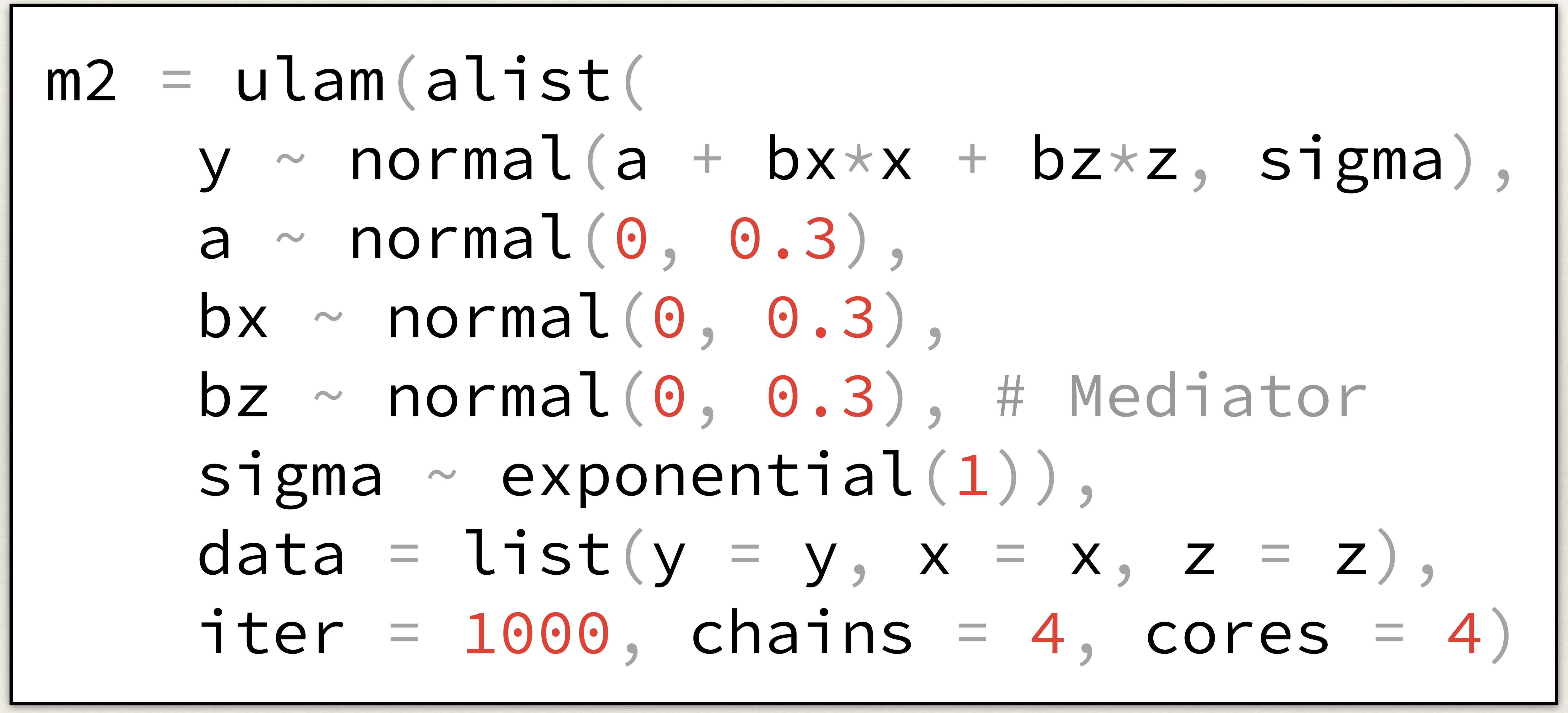
```
set.seed(1)
N = 100
x = rnorm(N)        # x ~ normal(0, 1)
z = rnorm(N, 1 + x) # z ~ normal(1 + x, 1)
y = rnorm(N, 1 + z) # y ~ normal(1 + z, 1)

m1 = ulam(alist(
    y ~ normal(a + bx*x, sigma),
    a ~ normal(0, 0.3),
    bx ~ normal(0, 0.3),
    sigma ~ exponential(1)),
    data = list(y = y, x = x),
    iter = 1000, chains = 4, cores = 4)
```

MODEL WITHOUT THE MEDIATOR

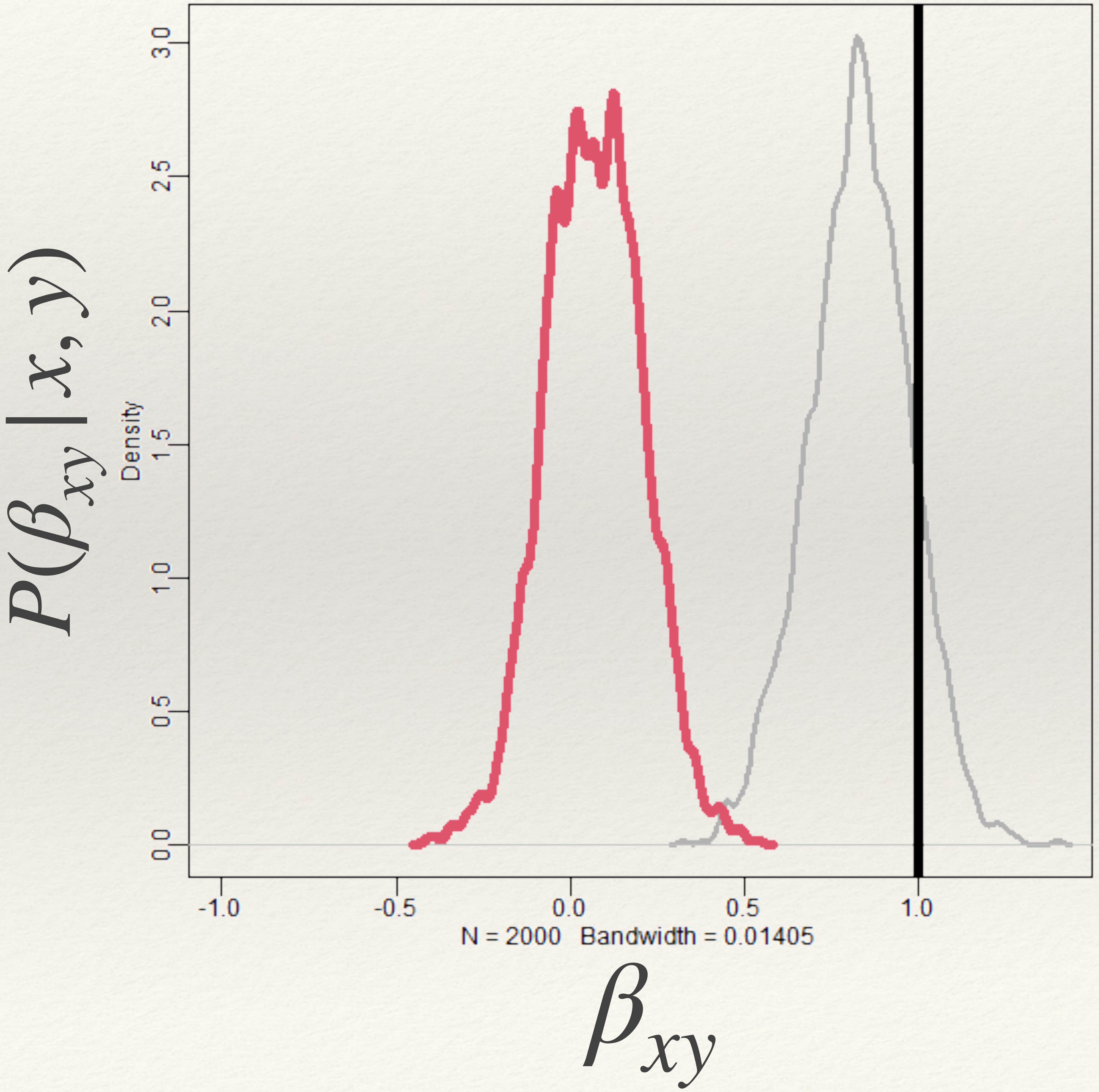
```
m1 = ulam(alist(
    y ~ normal(a + bx*x, sigma),
    a ~ normal(0, 0.3),
    bx ~ normal(0, 0.3),
    sigma ~ exponential(1)),
    data = list(y = y, x = x),
    iter = 1000, chains = 4, cores = 4)
```



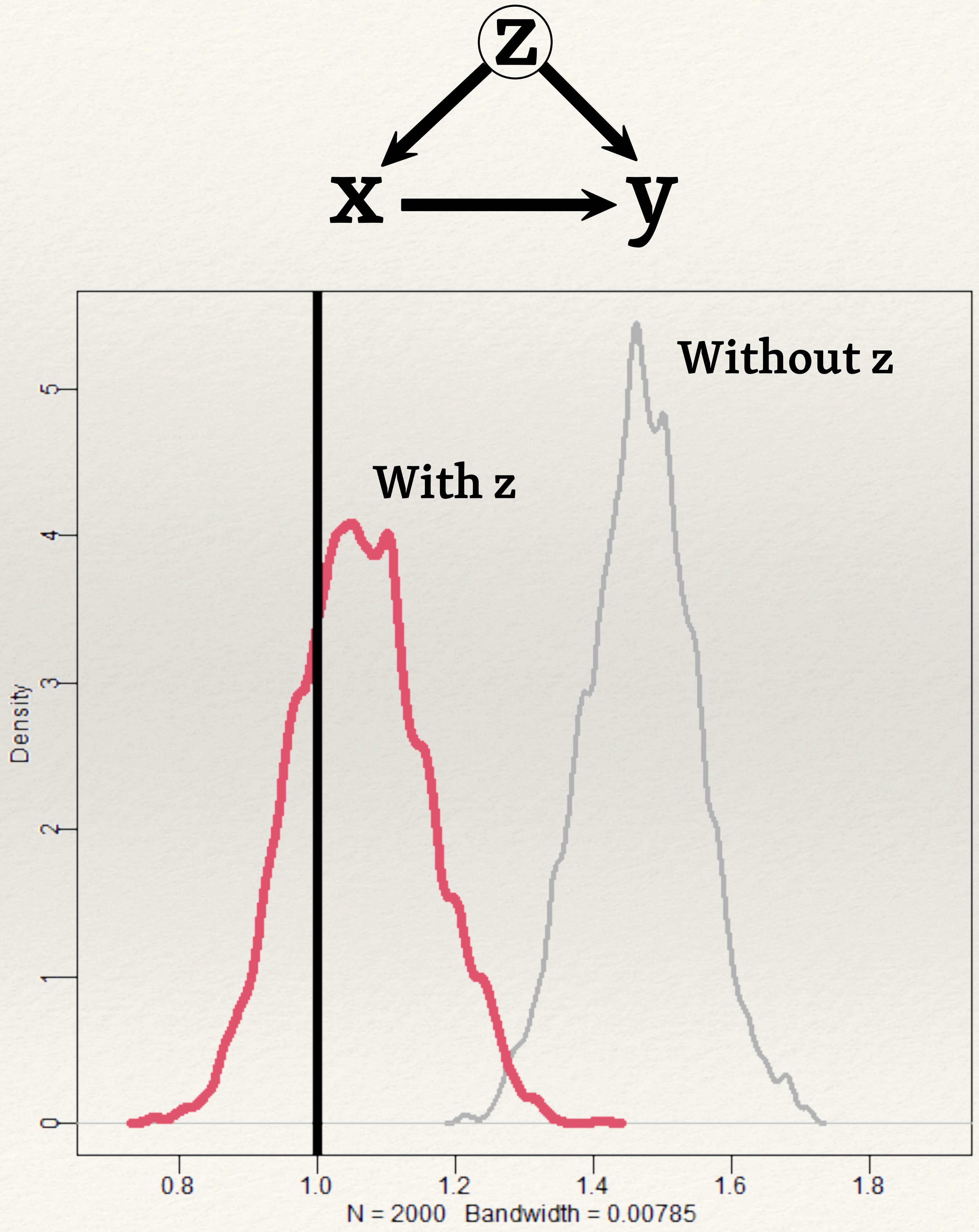
INCLUDING THE MEDIATOR



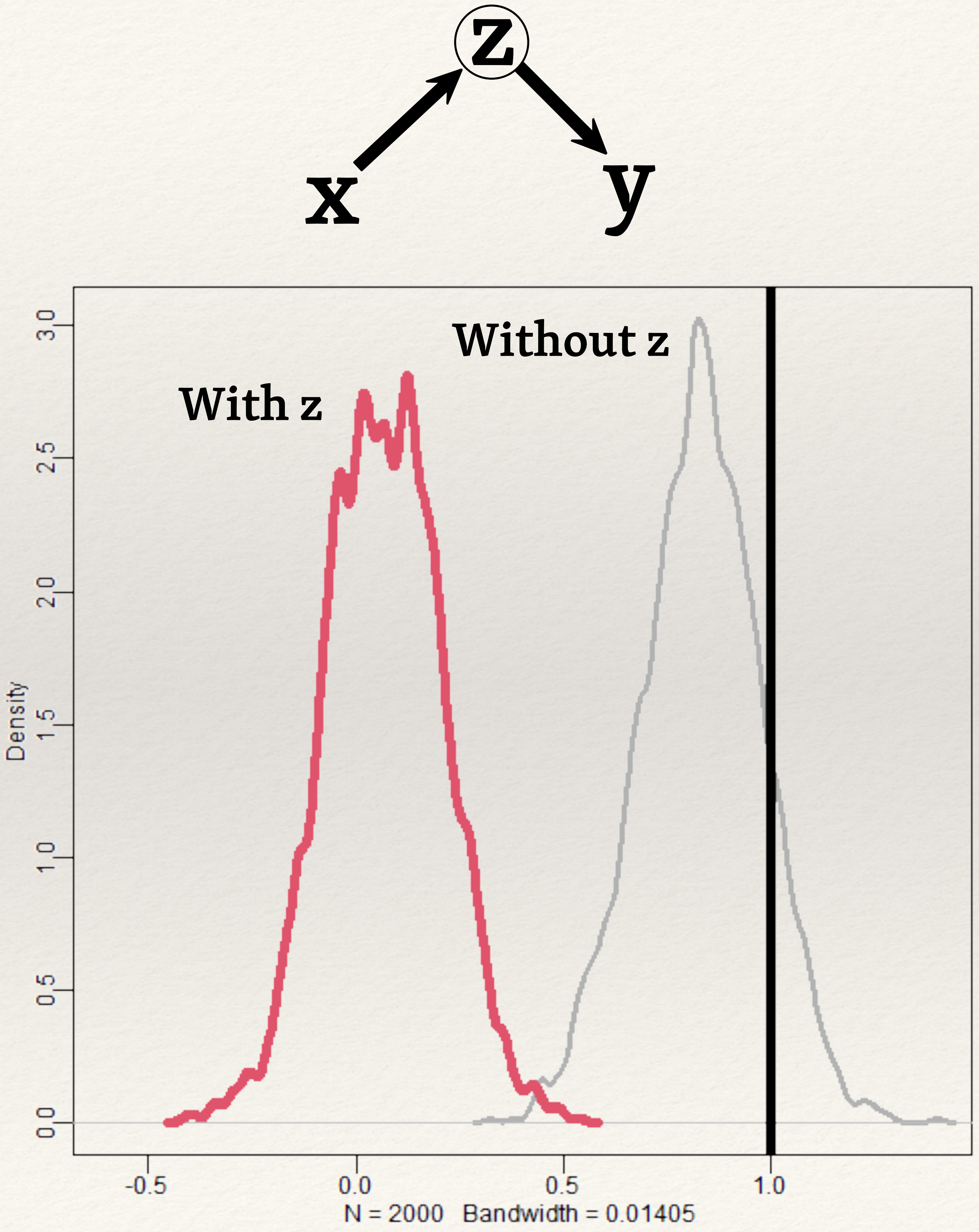
```
m2 = ulam(alist(
    y ~ normal(a + bx*x + bz*z, sigma),
    a ~ normal(0, 0.3),
    bx ~ normal(0, 0.3),
    bz ~ normal(0, 0.3), # Mediator
    sigma ~ exponential(1)),
    data = list(y = y, x = x, z = z),
    iter = 1000, chains = 4, cores = 4)
```



PIPE VS. FORK



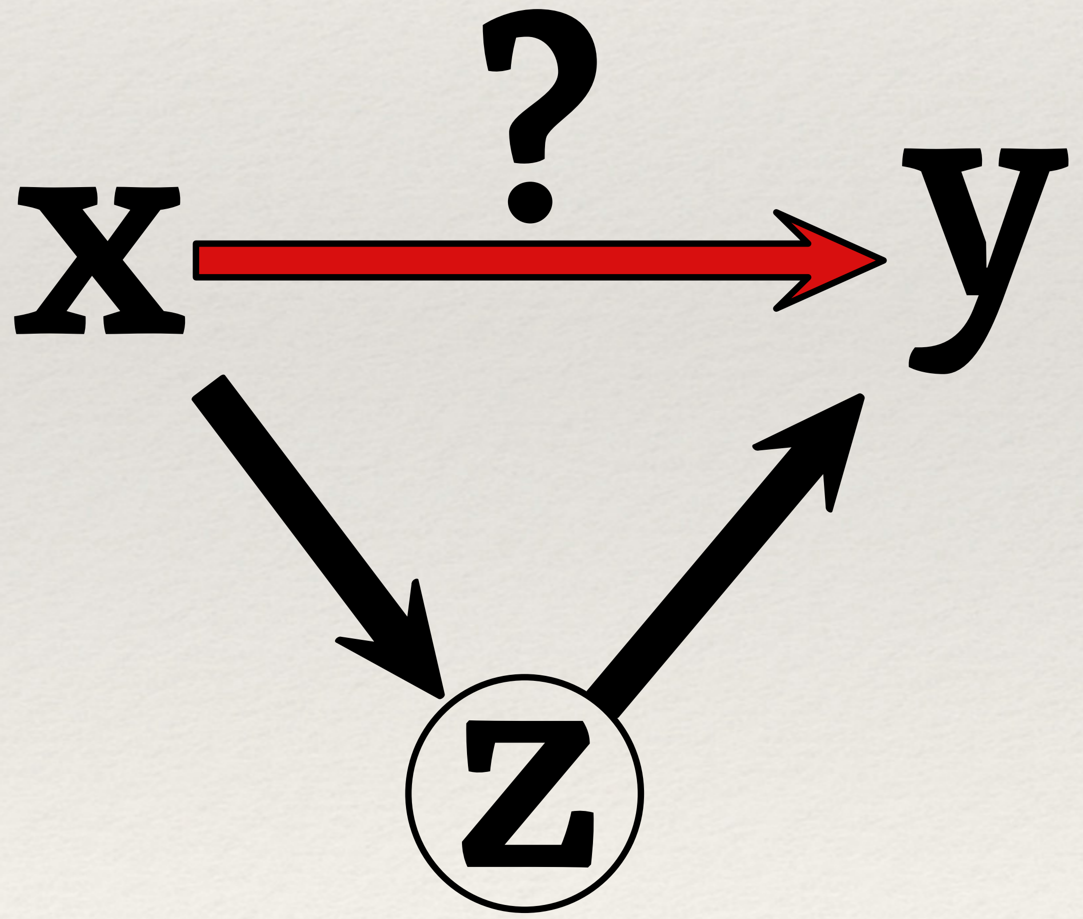
z
x
y
Without z
With z
Density
0
1
2
3
4
5
0.8
1.0
1.2
1.4
1.6
1.8
N = 2000 Bandwidth = 0.00785



z
x
y
Without z
With z
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
-0.5
0.0
0.5
1.0
N = 2000 Bandwidth = 0.01405

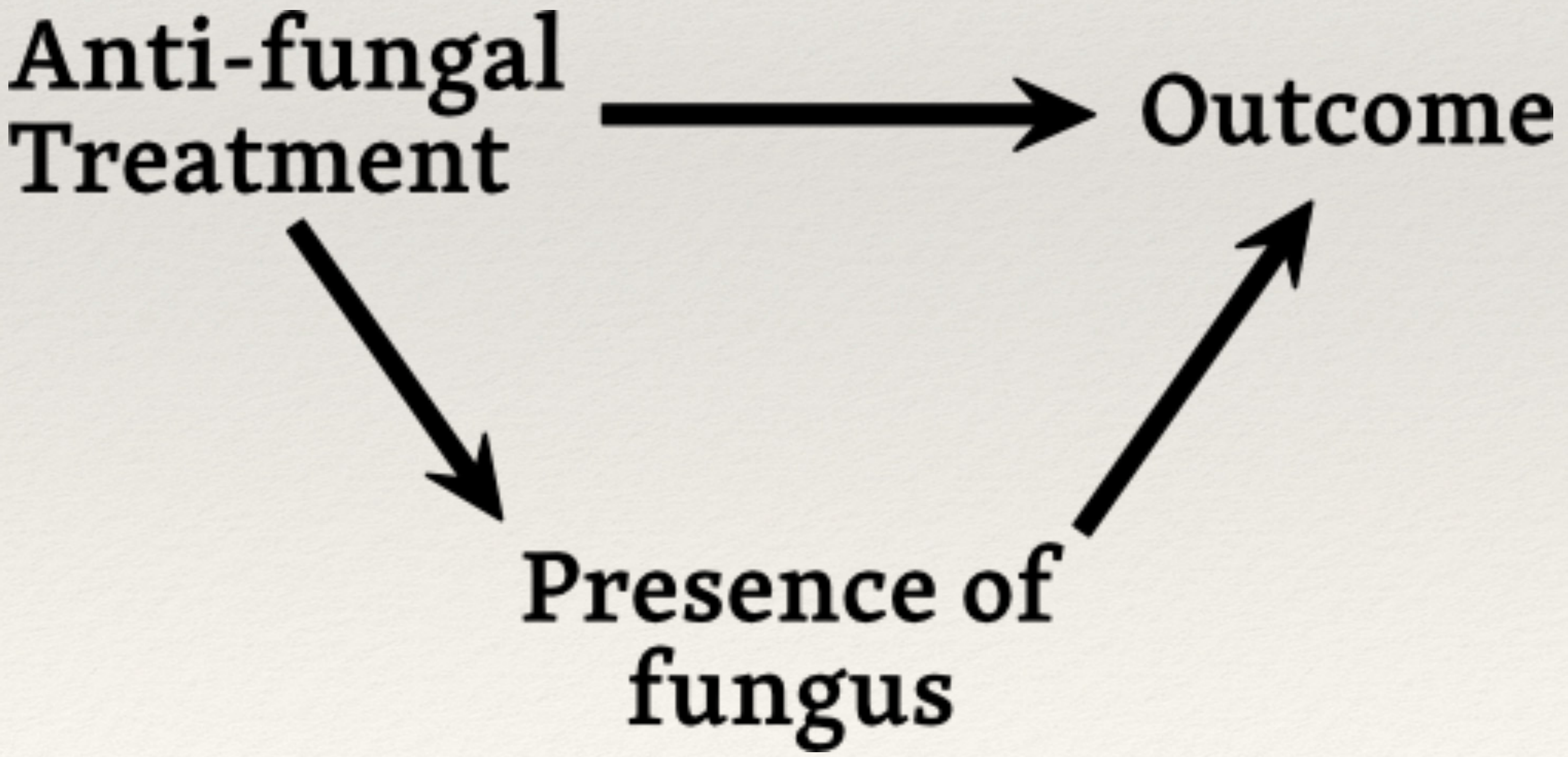
EXAMPLES OF PIPES OR MEDIATORS

Including a mediator in our models can have catastrophic effects. A common mistake is to include post-treatment variables in the model.



POST-TREATMENT VARIABLES

If we are evaluating the effectiveness of a fungal treatment, most of the effect of the treatment could be mediated by the presence of fungus. So, using presence of fungus in our model would mask the effect of the treatment.

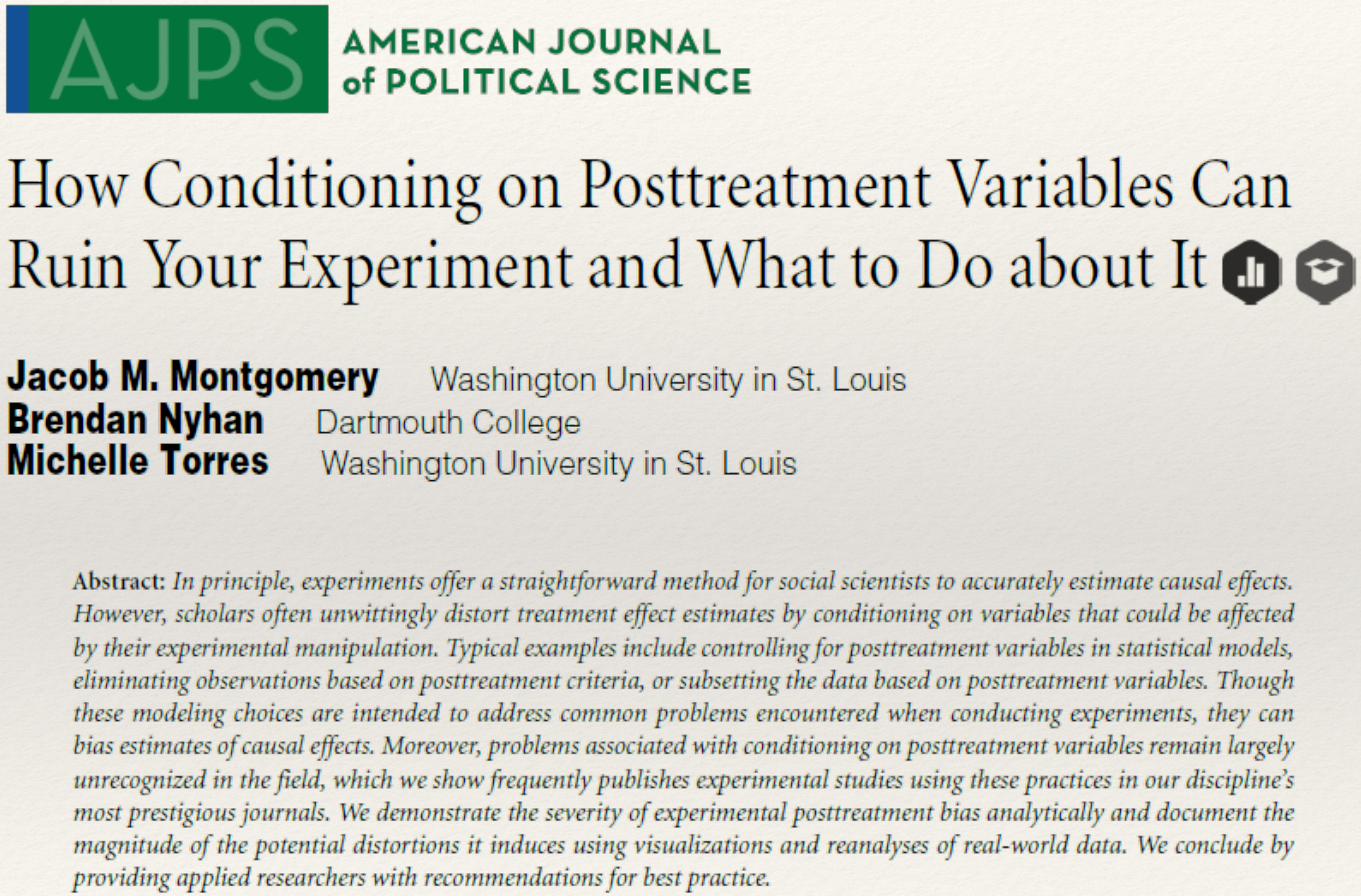


Anti-fungal Treatment

Outcome

Presence of fungus

CONDITIONING ON POSTTREATMENT VARIABLES



AJPS AMERICAN JOURNAL of POLITICAL SCIENCE

How Conditioning on Posttreatment Variables Can Ruin Your Experiment and What to Do about It

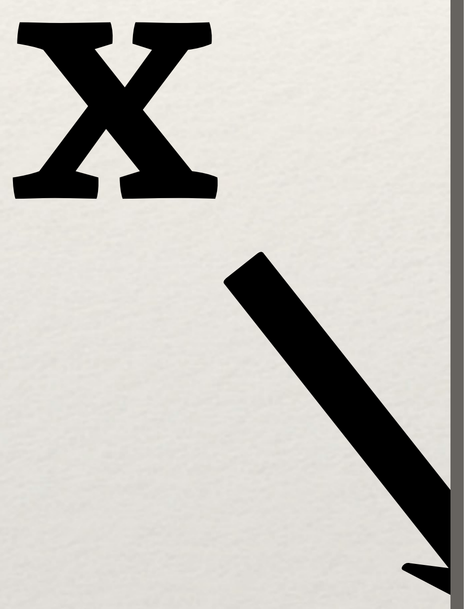
Jacob M. Montgomery Washington University in St. Louis
Brendan Nyhan Dartmouth College
Michelle Torres Washington University in St. Louis

Abstract: *In principle, experiments offer a straightforward method for social scientists to accurately estimate causal effects. However, scholars often unwittingly distort treatment effect estimates by conditioning on variables that could be affected by their experimental manipulation. Typical examples include controlling for posttreatment variables in statistical models, eliminating observations based on posttreatment criteria, or subsetting the data based on posttreatment variables. Though these modeling choices are intended to address common problems encountered when conducting experiments, they can bias estimates of causal effects. Moreover, problems associated with conditioning on posttreatment variables remain largely unrecognized in the field, which we show frequently publishes experimental studies using these practices in our discipline's most prestigious journals. We demonstrate the severity of experimental posttreatment bias analytically and document the magnitude of the potential distortions it induces using visualizations and reanalyses of real-world data. We conclude by providing applied researchers with recommendations for best practice.*

Montgomery, J. M., Nyhan, B. & Torres, M. How conditioning on posttreatment variables can ruin your experiment and what to do about it. Am. J. Pol. Sci. ***62****, 760–775 (2018)*

COLLIDER

NO EFFECT OF X ON Y, BUT BOTH AFFECT Z



Math

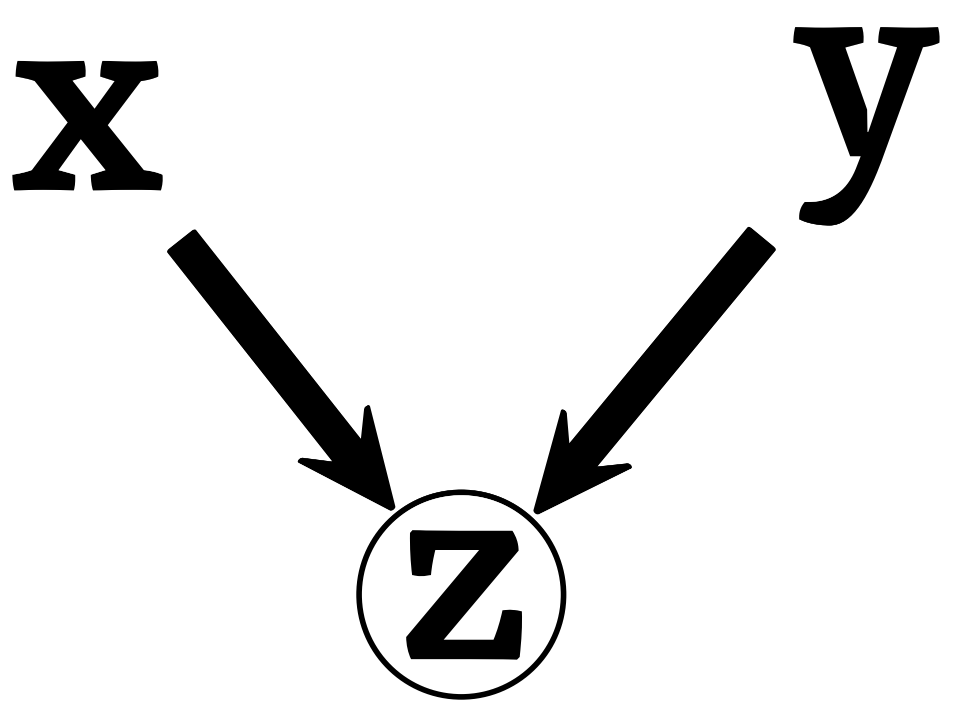
$$y \sim Normal(\alpha_y, \sigma_y)$$

$$x \sim Normal(\alpha_x, \sigma_x)$$

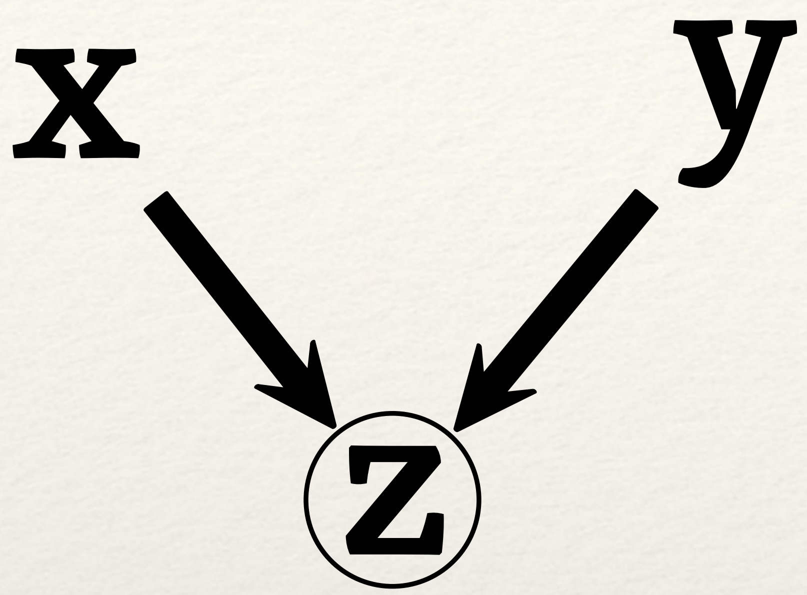
$$z \sim Normal(\alpha_z + \beta_{zx}x + \beta_{zy}y, \sigma_z)$$

NO EFFECT OF X ON Y, BUT BOTH AFFECT Z

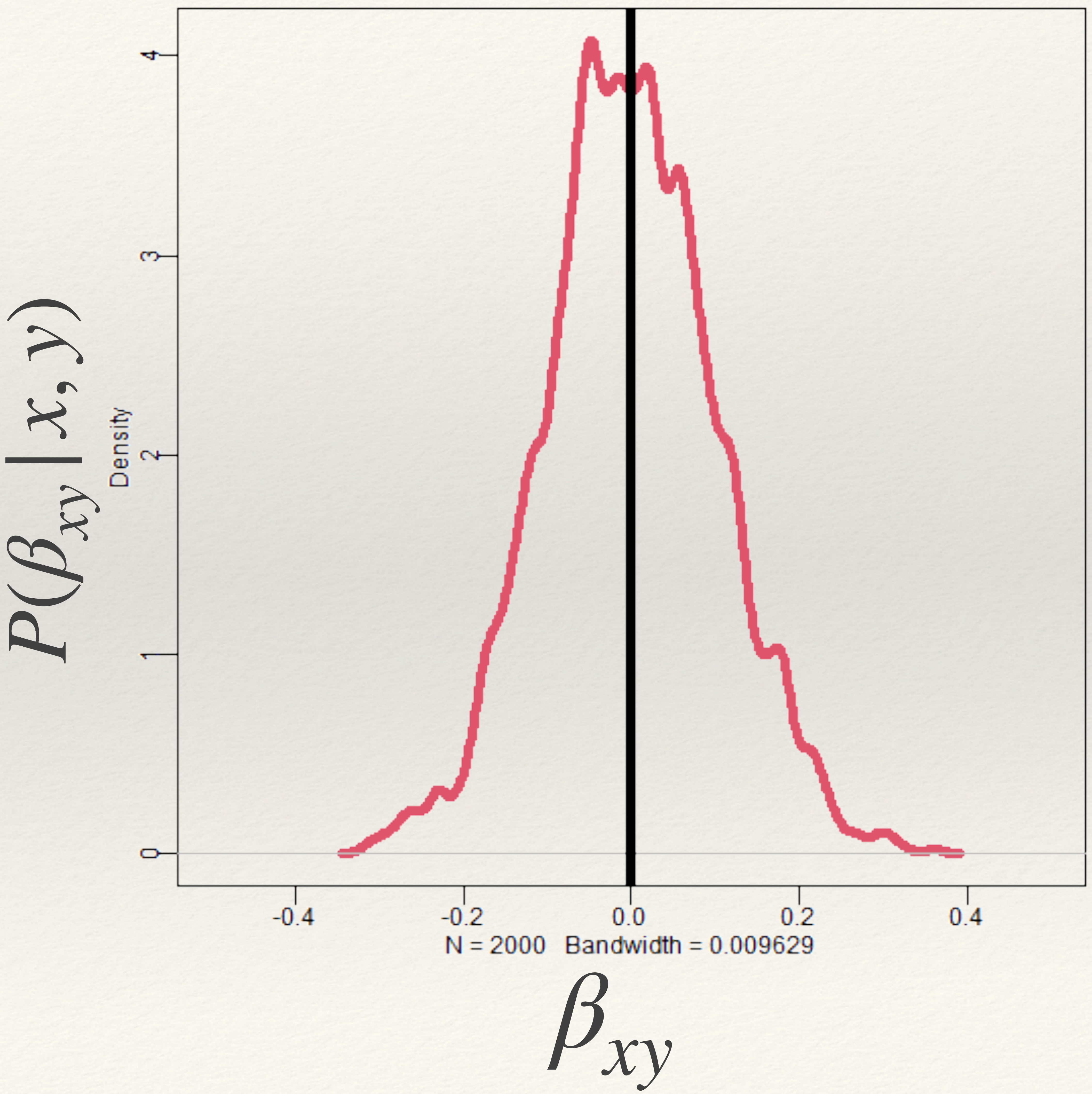
```r
set.seed(1)
N = 100
x = rnorm(N)             # x ~ normal(0, 1)
y = rnorm(N)             # y ~ normal(0, 1)
z = rnorm(N, 1 + x + y) # z ~ normal(1 + x + y, 1) -> collider

m1 = ulam(alist(
    y ~ normal(a + bx*x, sigma),
    a ~ normal(0, 0.3),
    bx ~ normal(0, 0.3),
    sigma ~ exponential(1)),
    data = list(y = y, x = x),
    iter = 1000, chains = 4, cores = 4)
```



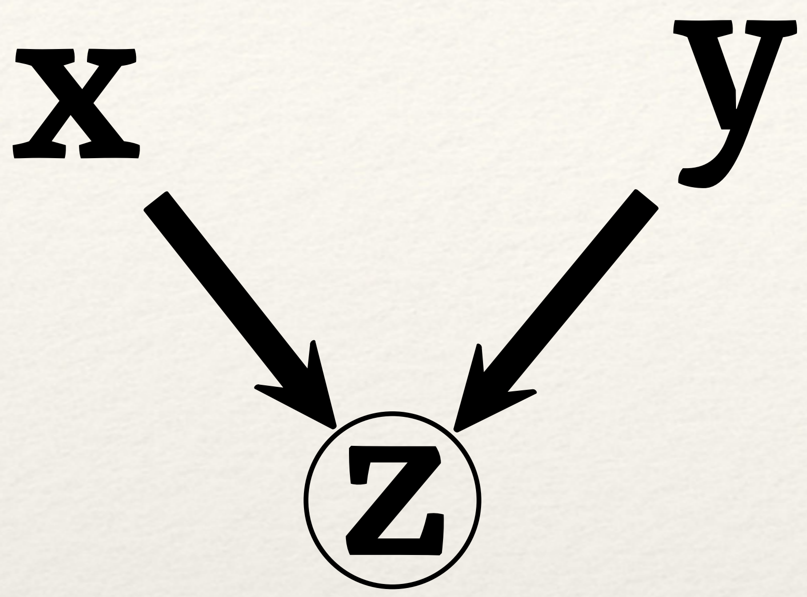
EFFECT OF X ON Y WITHOUT THE COLLIDER



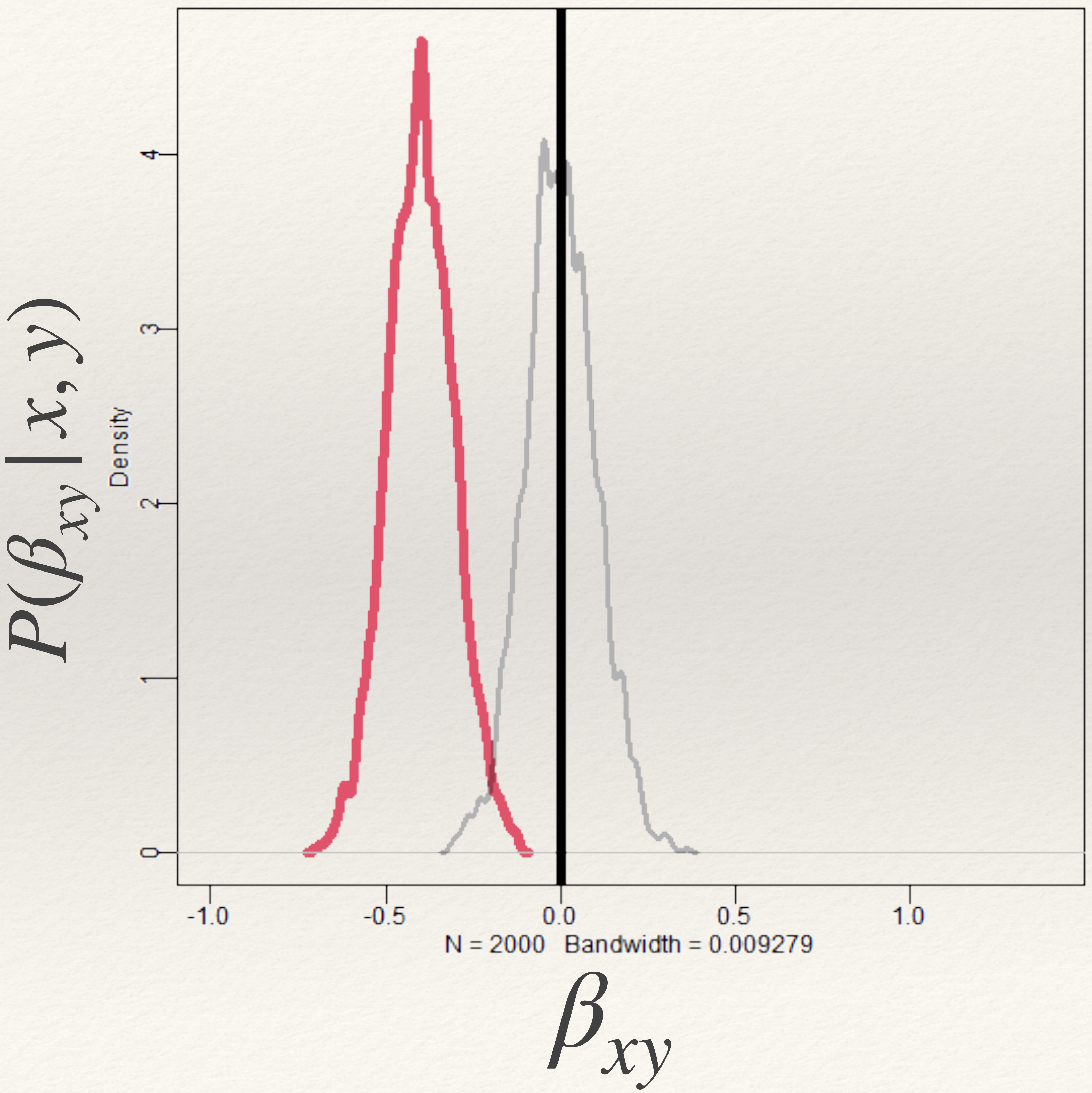
```
m1 = ulam(alist(
    y ~ normal(a + bx*x, sigma),
    a ~ normal(0, 0.3),
    bx ~ normal(0, 0.3),
    sigma ~ exponential(1)),
    data = list(y = y, x = x),
    iter = 1000, chains = 4, cores = 4)
```



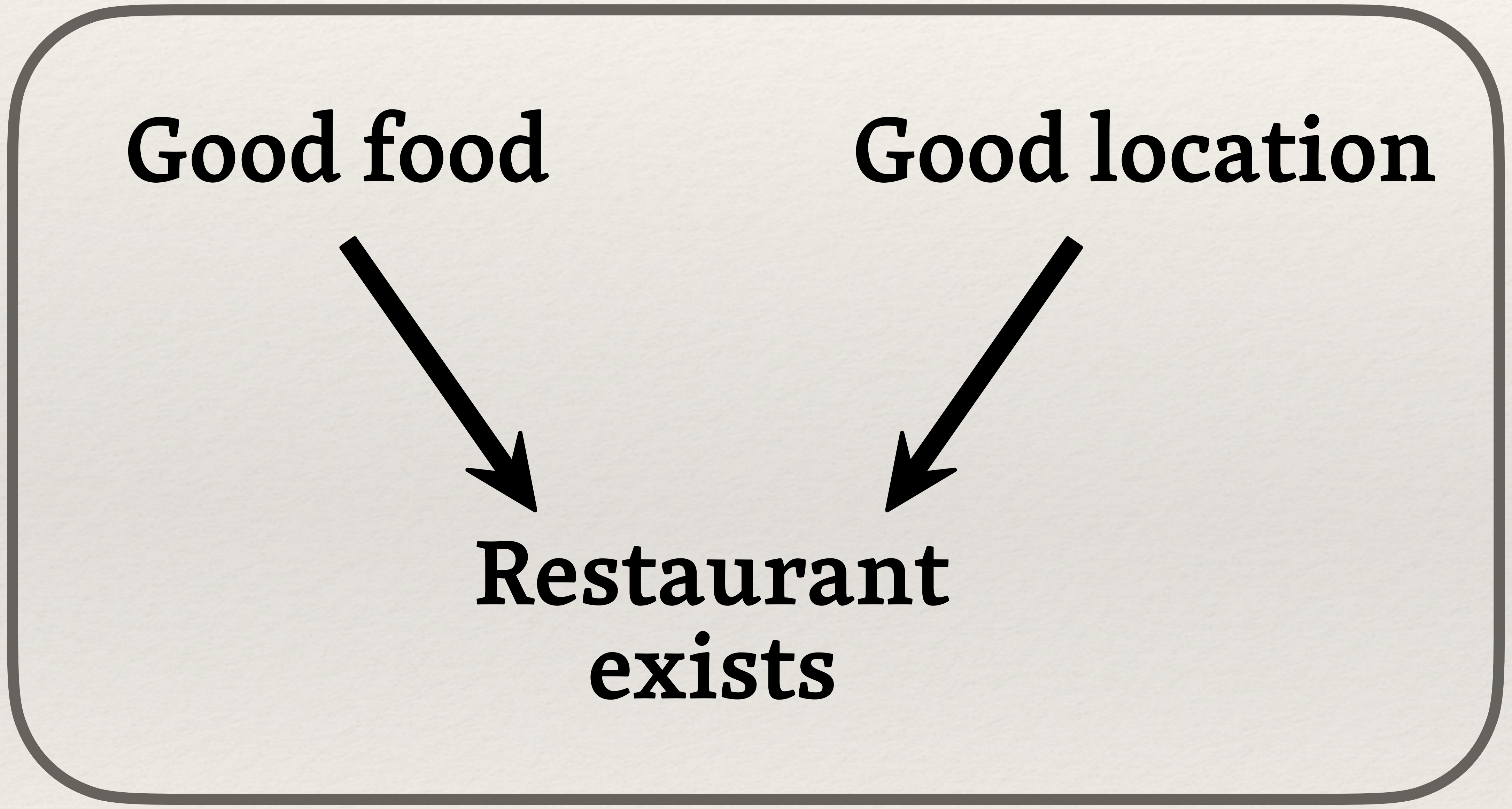
EFFECT OF X ON Y WITH THE COLLIDER



```
m2 = ulam(alist(
    y ~ normal(a + bx*x + bz*z, sigma),
    a ~ normal(0, 0.3),
    bx ~ normal(0, 0.3),
    bz ~ normal(0, 0.3), # Collider
    sigma ~ exponential(1)),
    data = list(y = y, x = x, z = z),
    iter = 1000, chains = 4, cores = 4)
```



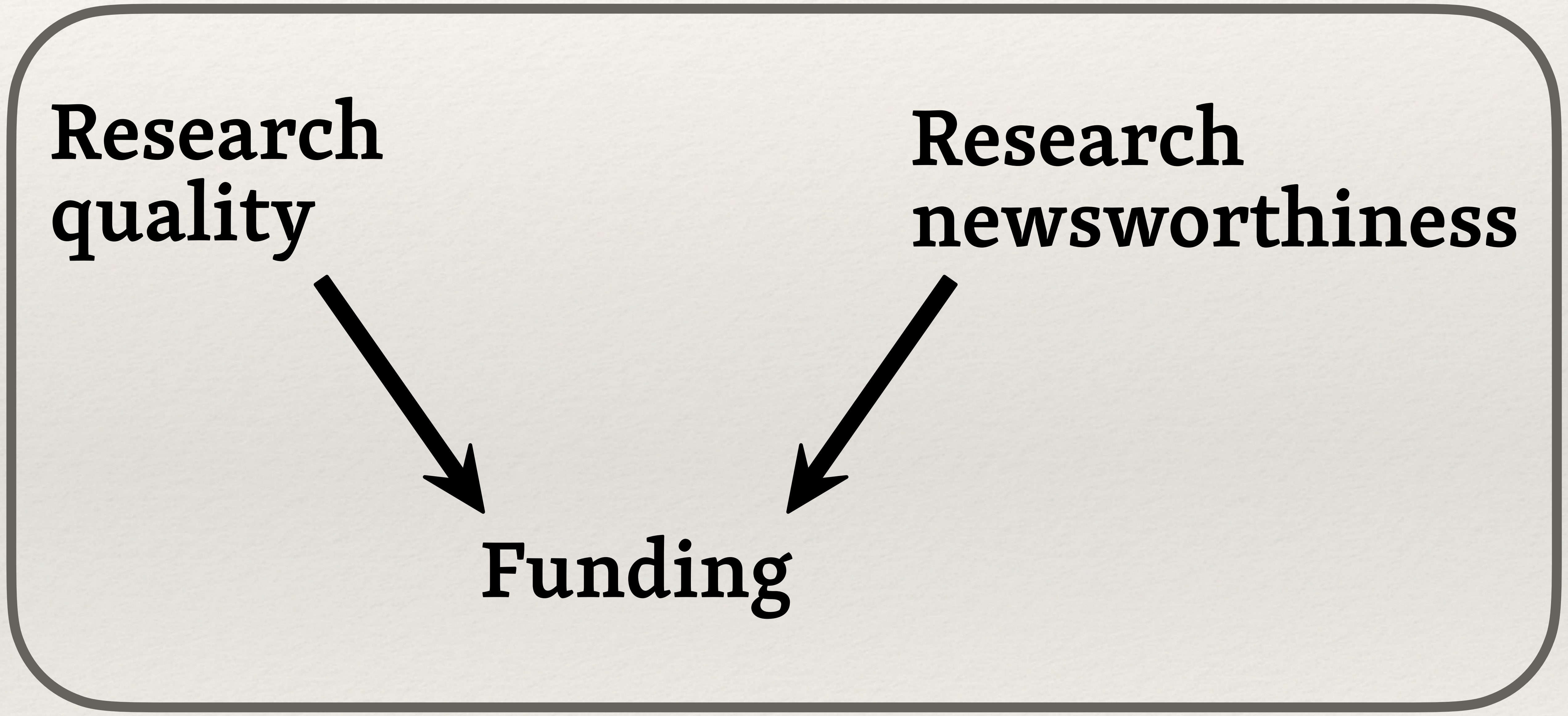
COLLIDERS CAN CAUSE OUR SAMPLES TO BE BIASED



Good food

Good location

Restaurant exists



Research quality

Research newsworthiness

Funding

WHAT NOW?!

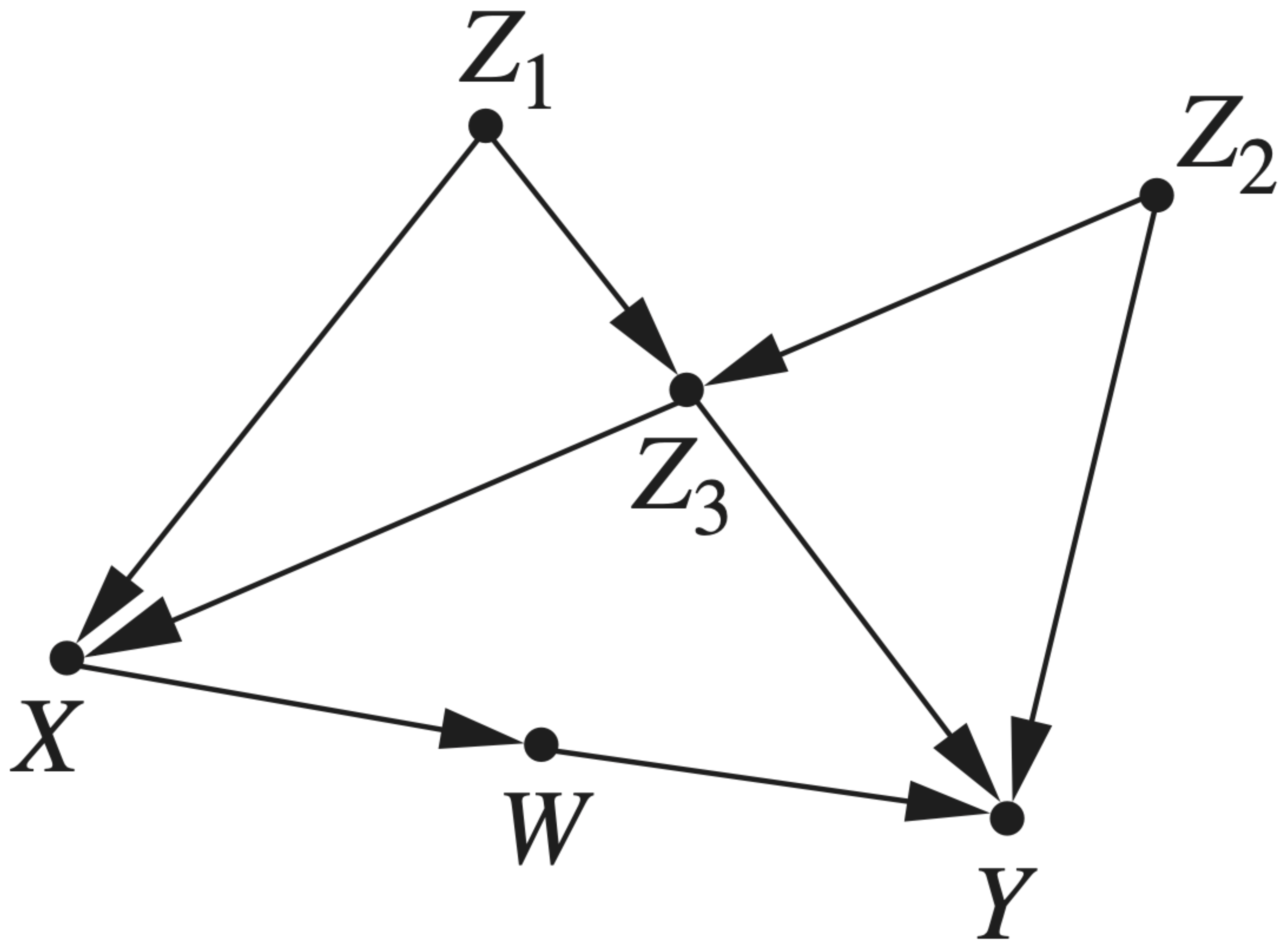
USING DAGS TO BUILD MODELS

If we represent our putative causal relations using DAGs, we have a set of rules that tells us what variables we need to include in the model in order to calculate a particular effect.

OPEN AND CLOSED PATHS

- Paths containing uncontrolled **pipes and forks** are open
- Paths containing **colliders are closed by default**, but open if we condition on the collider
- To estimate the true causal effect of x on y, we need **all non-causal paths from x to y to be closed** in our model

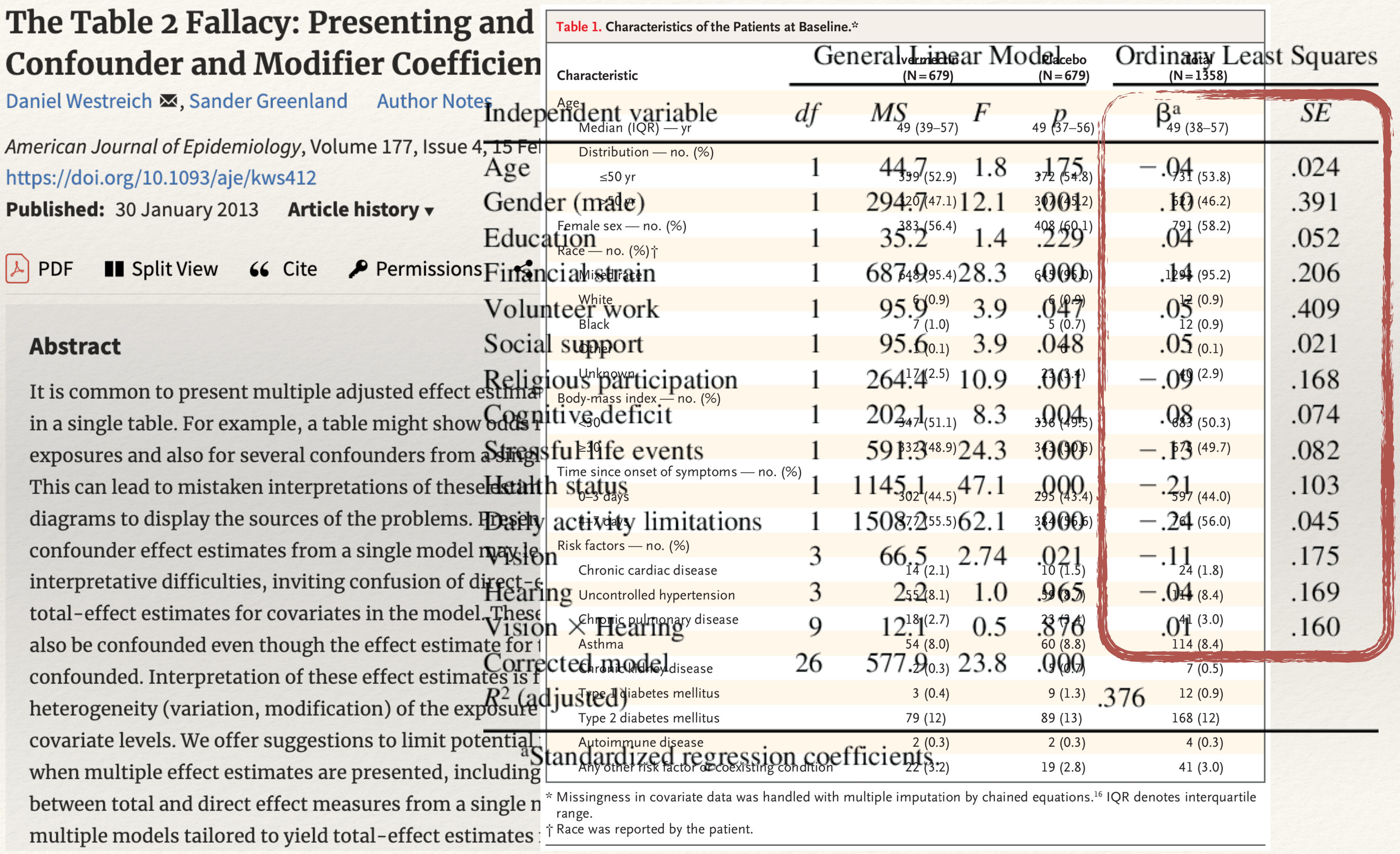
Identify all the open paths from X to Y



BACK DOOR CRITERION

To estimate the causal effect of X on Y, identify a set of control variables such that no descendants of X are in the control set, and all paths between X and Y that contain an arrow into X are blocked.

NOT ALL ESTIMATES ARE CAUSAL



The Table 2 Fallacy: Presenting and
Confounder and Modifier Coefficien

Daniel Westreich ✉, Sander Greenland   Author Notes

American Journal of Epidemiology, Volume 177, Issue 4, 15 Fe
https://doi.org/10.1093/aje/kws412

Published: 30 January 2013   **Article history** ▾

Abstract

It is common to present multiple adjusted effect estima
in a single table. For example, a table might show odds r
exposures and also for several confounders from a singl
This can lead to mistaken interpretations of these estim
diagrams to display the sources of the problems. Presen
confounder effect estimates from a single model may le
interpretative difficulties, inviting confusion of direct-e
total-effect estimates for covariates in the model. These
also be confounded even though the effect estimate for t
confounded. Interpretation of these effect estimates is f
heterogeneity (variation, modification) of the exposure
covariate levels. We offer suggestions to limit potential
when multiple effect estimates are presented, including
between total and direct effect measures from a single m
multiple models tailored to yield total-effect estimates

Table 1. Characteristics of the Patients at Baseline.*

Characteristic	Ivermectin (N = 679)	Placebo (N = 679)	Total (N = 1358)
Age			
Median (IQR) — yr	49 (39–57)	49 (37–56)	49 (38–57)
Distribution — no. (%)			
≤50 yr	359 (52.9)	372 (54.8)	731 (53.8)
>50 yr	320 (47.1)	307 (45.2)	627 (46.2)
Female sex — no. (%)	383 (56.4)	408 (60.1)	791 (58.2)
Race — no. (%)†			
Mixed race	648 (95.4)	645 (95.0)	1293 (95.2)
White	6 (0.9)	6 (0.9)	12 (0.9)
Black	7 (1.0)	5 (0.7)	12 (0.9)
Other	1 (0.1)	0	1 (0.1)
Unknown	17 (2.5)	23 (3.4)	40 (2.9)
Body-mass index — no. (%)			
<30	347 (51.1)	336 (49.5)	683 (50.3)
≥30	332 (48.9)	343 (50.5)	675 (49.7)
Time since onset of symptoms — no. (%)			
0–3 days	302 (44.5)	295 (43.4)	597 (44.0)
4–7 days	377 (55.5)	384 (56.6)	761 (56.0)
Risk factors — no. (%)			
Chronic cardiac disease	14 (2.1)	10 (1.5)	24 (1.8)
Uncontrolled hypertension	55 (8.1)	59 (8.7)	114 (8.4)
Chronic pulmonary disease	18 (2.7)	23 (3.4)	41 (3.0)
Asthma	54 (8.0)	60 (8.8)	114 (8.4)
Chronic kidney disease	2 (0.3)	5 (0.7)	7 (0.5)
Type 1 diabetes mellitus	3 (0.4)	9 (1.3)	12 (0.9)
Type 2 diabetes mellitus	79 (12)	89 (13)	168 (12)
Autoimmune disease	2 (0.3)	2 (0.3)	4 (0.3)
Any other risk factor or coexisting condition	22 (3.2)	19 (2.8)	41 (3.0)

* Missingness in covariate data was handled with multiple imputation by chained equations.[16] IQR denotes interquartile range.

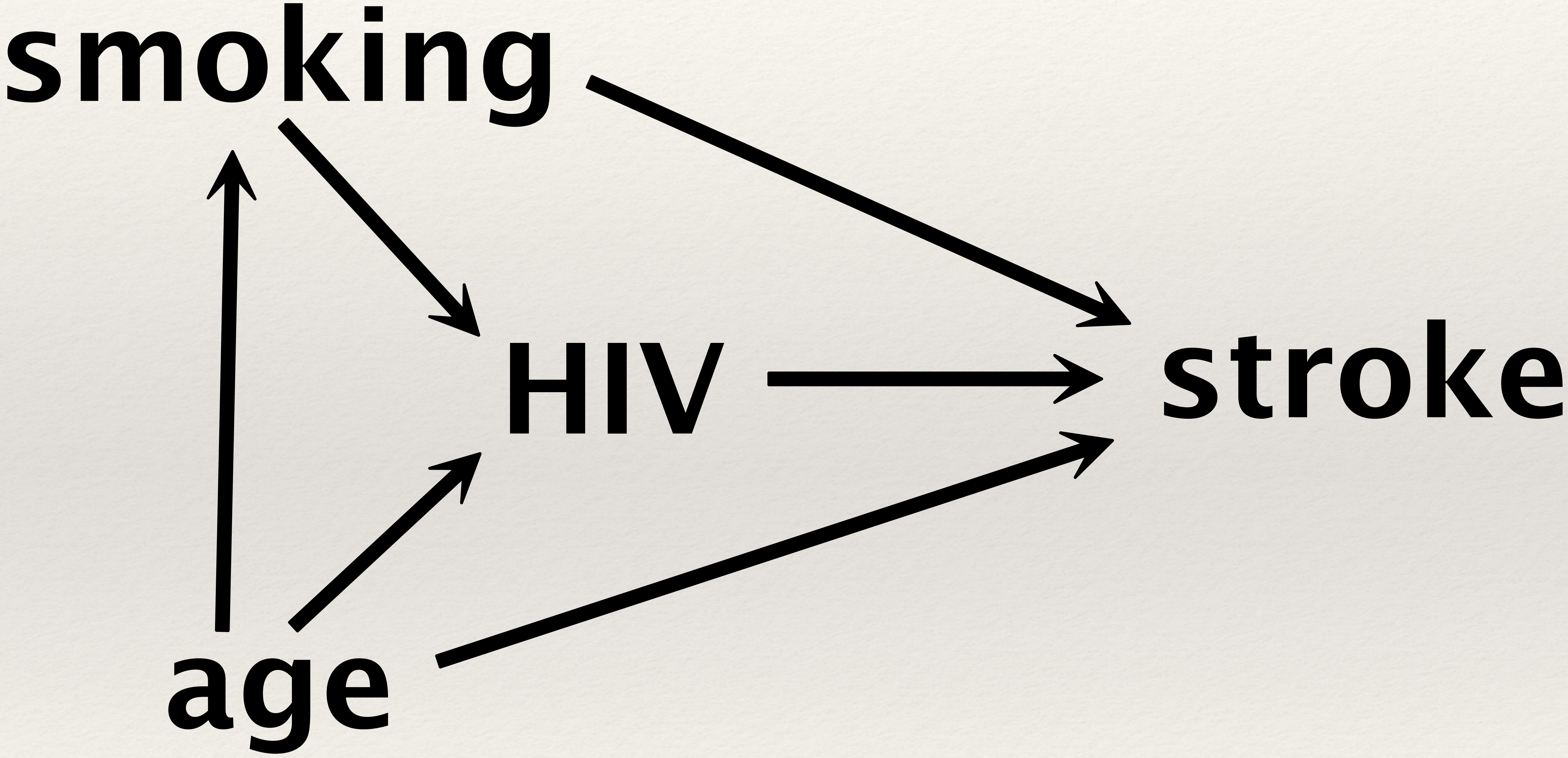
† Race was reported by the patient.

	General Linear Model				Ordinary Least Squares	
Independent variable	*df*	*MS*	*F*	*p*	β[a]	*SE*
Age	1	44.7	1.8	.175	−.04	.024
Gender (male)	1	294.7	12.1	.001	.10	.391
Education	1	35.2	1.4	.229	.04	.052
Financial strain	1	687.9	28.3	.000	.14	.206
Volunteer work	1	95.9	3.9	.047	.05	.409
Social support	1	95.6	3.9	.048	.05	.021
Religious participation	1	264.4	10.9	.001	−.09	.168
Cognitive deficit	1	202.1	8.3	.004	.08	.074
Stressful life events	1	591.3	24.3	.000	−.13	.082
Health status	1	1145.1	47.1	.000	−.21	.103
Daily activity limitations	1	1508.2	62.1	.000	−.24	.045
Vision	3	66.5	2.74	.021	−.11	.175
Hearing	3	2.2	1.0	.965	−.04	.169
Vision × Hearing	9	12.1	0.5	.876	.01	.160
Corrected model	26	577.9	23.8	.000		
R^2 (adjusted)					.376	

[a] Standardized regression coefficients.

Westreich, D. & Greenland, S. The table 2 fallacy: presenting and interpreting confounder and modifier coefficients. Am. J. Epidemiol. ***177****, 292–298 (2013)*

HIV AND STROKE



GOOD AND BAD CONTROLS

Good controls

- Block non-causal paths
- Improve precision
- Allow inference of causal effects

Bad controls

- Block causal paths (blocking pipes)
- Open non-causal paths (opening colliders)
- Reduce precision
- Prevents causal inference

Cinelli, C., Forney, A. & Pearl, J. A crash course in good and bad controls. (2020) doi:10.2139/ssrn.3689437