ETH zürich

To adjust or not to adjust:

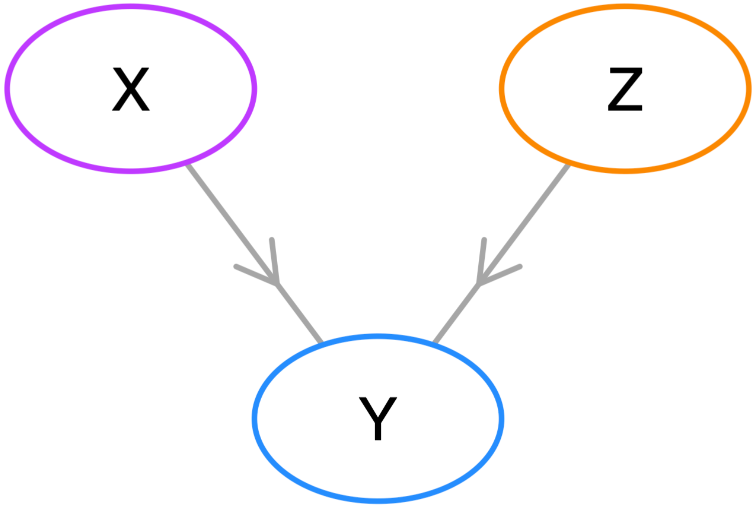
insights from the simplest non-trivial system of discrete variables

Jack Kuipers

25 March 2025

A v-structure

Simplest non-trivial system: a v-structure



- Treatment X, drug/placebo
- Outcome Y, improve/not
- Prognostic factor Z

for binary X, Y and Z

We want to compute the **causal** effect of X on Y

$$p(Y \mid \mathrm{do}(X=1)) - p(Y \mid \mathrm{do}(X=0))$$

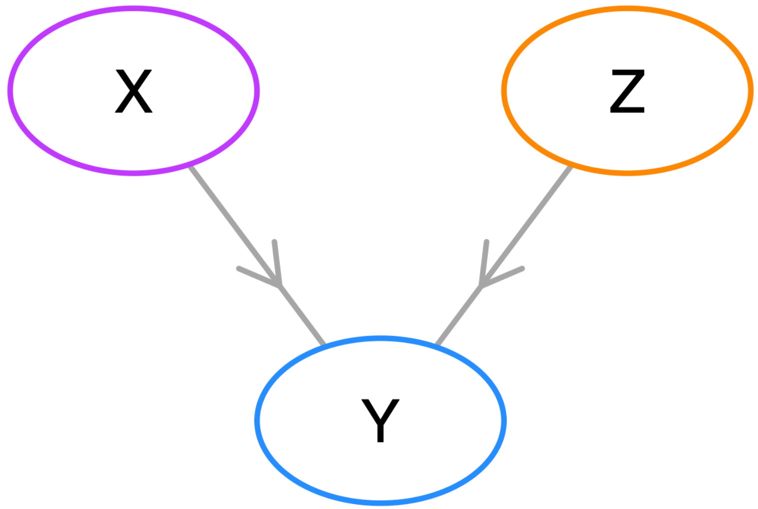
using do-calculus Pearl 1995

Do we adjust for Z?

Two choices

Both choices are valid

- unbiased and targeting the same estimand



No adjustment: use raw conditionals

$$p(Y \mid \mathrm{do}(X)) \overset{R}{=} p(Y \mid X)$$

With adjustment: marginalise

$$p(Y \mid \mathrm{do}(X)) \overset{M}{=} \sum_{Z} p(Y \mid X, Z) p(Z)$$

We want to compare the variance

$$V\left[p(Y \mid \mathrm{do}(X=1)) - p(Y \mid \mathrm{do}(X=0))\right]$$

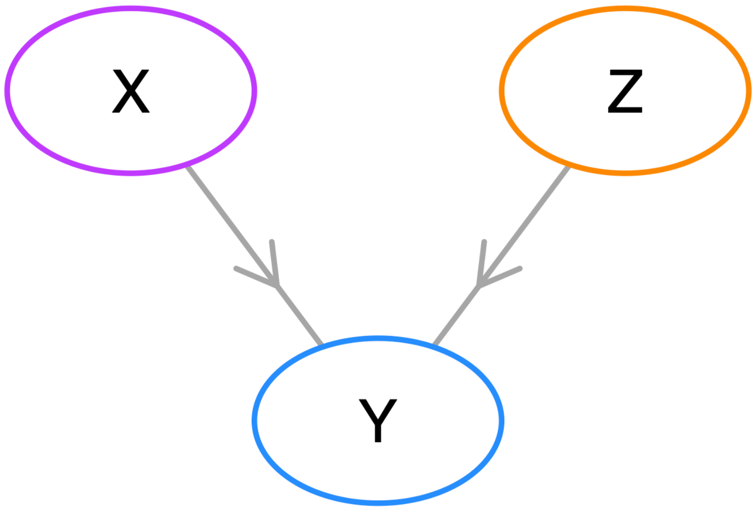
- of the two choices

Probability tables

As data-generating mechanism of the v-structure, we can imagine sampling

- X with probability p_X
- Z with probability p_Z
- Outcome Y depends on both

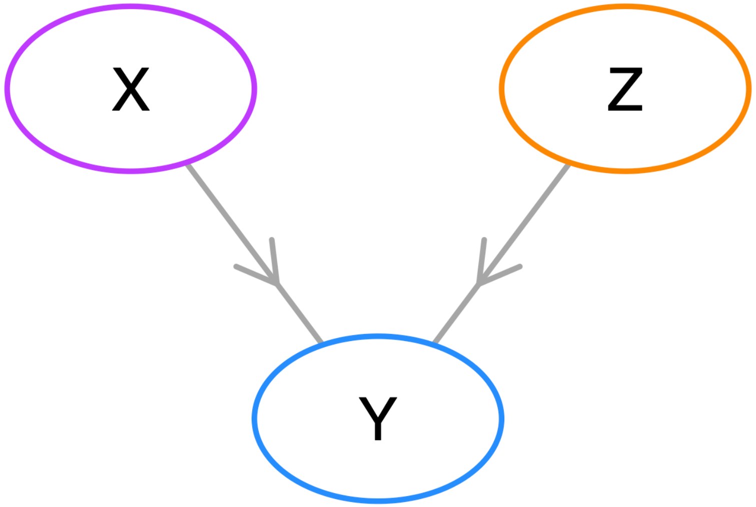
$$
\begin{aligned}
p(Y = 1 \mid X = 0, Z = 0) &= p_{Y,0} \\
p(Y = 1 \mid X = 1, Z = 0) &= p_{Y,2} \\
p(Y = 1 \mid X = 0, Z = 1) &= p_{Y,1} \\
p(Y = 1 \mid X = 1, Z = 1) &= p_{Y,3}
\end{aligned}
$$



Probability tables

As data-generating mechanism of the v-structure, we can imagine sampling

- X with probability p_X
- Z with probability p_Z
- Outcome Y depends on both



Or sampling the combinations of (X, Z, Y) directly from a multinomial distribution

X	Z	Y	p	X	Z	Y	p
0	0	0	$p_0 = (1 - p_X)(1 - p_Z)(1 - p_{Y,0})$	1	0	0	$p_4 = p_X(1 - p_Z)(1 - p_{Y,2})$
0	0	1	$p_1 = (1 - p_X)(1 - p_Z)p_{Y,0}$	1	0	1	$p_5 = p_X(1 - p_Z)p_{Y,2}$
0	1	0	$p_2 = (1 - p_X)p_Z(1 - p_{Y,1})$	1	1	0	$p_6 = p_X p_Z(1 - p_{Y,3})$
0	1	1	$p_3 = (1 - p_X)p_Z p_{Y,1}$	1	1	1	$p_7 = p_X p_Z p_{Y,3}$

Causal effect estimators

Let N_i be the number of sampled cases indexed by $i = 4X + 2Z + Y$

- for total sample size N

Estimate from raw conditionals

$$R = \frac{N_5 + N_7}{N_4 + N_5 + N_6 + N_7} - \frac{N_1 + N_3}{N_0 + N_1 + N_2 + N_3}$$

Estimate from marginalisation

$$M = \frac{N_7}{(N_6+N_7)}\frac{(N_2+N_3+N_6+N_7)}{N} - \frac{N_3}{(N_2+N_3)}\frac{(N_2+N_3+N_6+N_7)}{N} + \frac{N_5}{(N_4+N_5)}\frac{(N_0+N_1+N_4+N_5)}{N} - \frac{N_1}{(N_0+N_1)}\frac{(N_0+N_1+N_4+N_5)}{N}$$

How do we compute their expectations and variances?

The joy of generating functions

Let's take a simpler example

$$R_1 = \frac{N_5 + N_7}{N_4 + N_5 + N_6 + N_7}$$

Introduce the generating function

$$S_N(v, z) = \{[p_0 + p_1 + p_2 + p_3] + [p_4 + p_6 + (p_5 + p_7)v]z\}^N$$

whose multinomial expansion

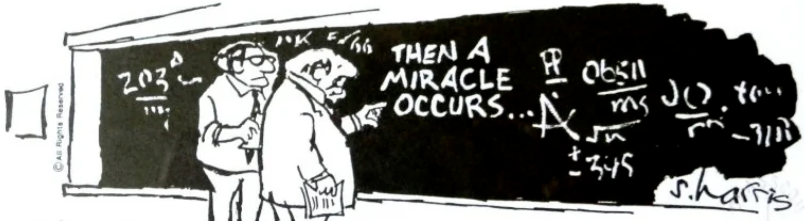
$$S_N = \sum \frac{N!}{N_0! \cdots N_7!} p_0^{N_0} \cdots p_7^{N_7} v^{N_5+N_7} z^{N_4+N_5+N_6+N_7}$$

allows us to keep track of the terms in R_1 through the generating variables (v, z)

And compute expectations through calculus

$$E[R_1] = \sum \frac{N!}{N_0! \cdots N_7!} p_0^{N_0} \cdots p_7^{N_7} \cdot \frac{N_5 + N_7}{N_4 + N_5 + N_6 + N_7} = \int \frac{v}{z} \frac{\partial}{\partial v} S_N \, \mathrm{d}z \bigg|_{\substack{v=1 \\ z=1}}$$

The variance of causal effect estimators



$$V[R] = \frac{(p_5+p_7)(p_4+p_6)}{p_X} N(1-p_X)^{N-1} F\left([1,1,1-N],[2,2],-\frac{p_X}{1-p_X}\right) + \frac{(p_1+p_3)(p_0+p_2)}{(1-p_X)} N p_X^{N-1} F\left([1,1,1-N],[2,2],-\frac{1-p_X}{p_X}\right)$$

no adjustment

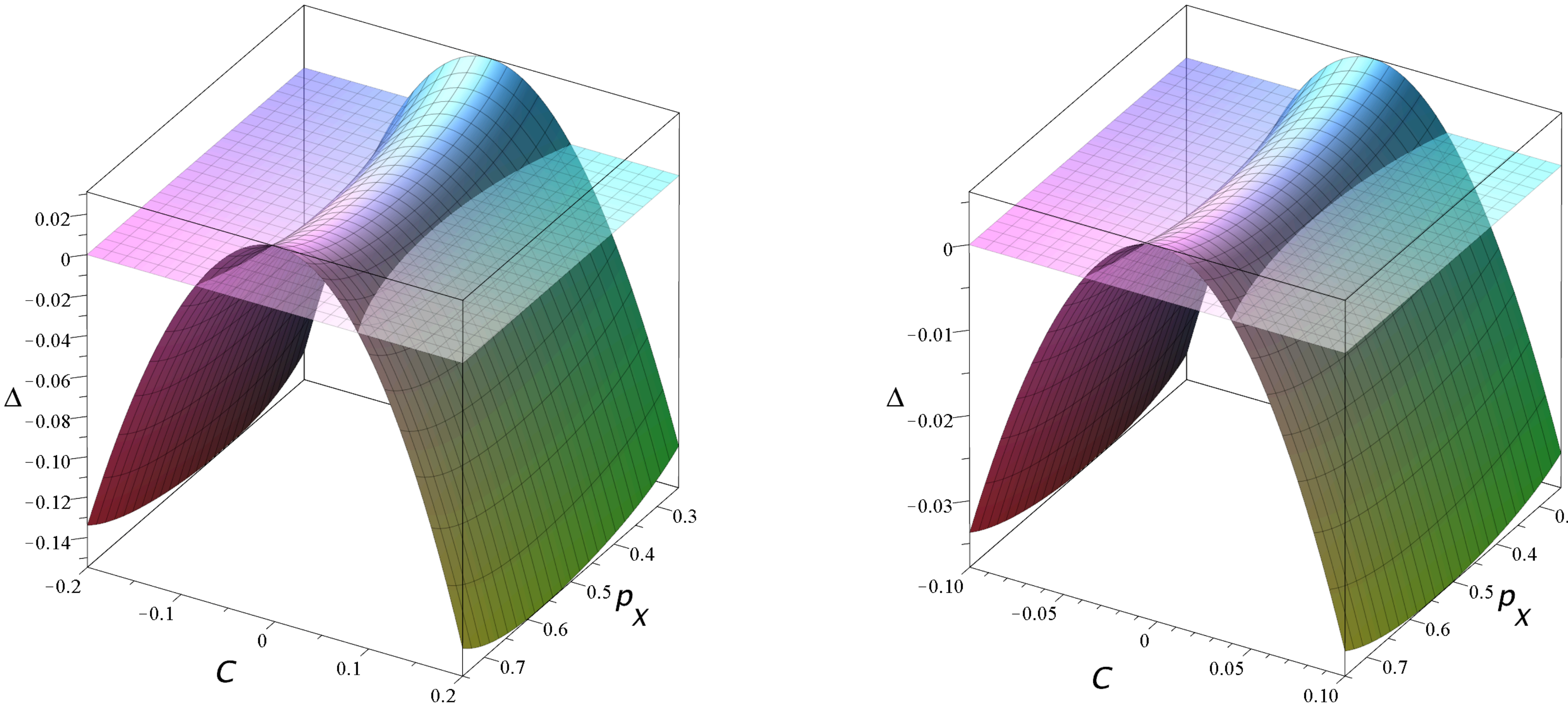
- F are hypergeometric functions

$$\begin{aligned}
V[M]N = & \frac{p_6p_7(p_2+p_3)}{(p_6+p_7)}(N-1)(1-p_6-p_7)^{N-2}F\left([1,1,2-N],[2,2],-\frac{p_6+p_7}{1-p_6-p_7}\right) \\
& + \frac{p_6p_7(p_2+p_3)^2}{(p_6+p_7)}(N-1)(N-2)(1-p_6-p_7)^{N-3}F\left([1,1,3-N],[2,2],-\frac{p_6+p_7}{1-p_6-p_7}\right) \\
& + \frac{p_4p_5(p_0+p_1)}{(p_4+p_5)}(N-1)(1-p_4-p_5)^{N-2}F\left([1,1,2-N],[2,2],-\frac{p_4+p_5}{1-p_4-p_5}\right) \\
& + \frac{p_4p_5(p_0+p_1)^2}{(p_4+p_5)}(N-1)(N-2)(1-p_4-p_5)^{N-3}F\left([1,1,3-N],[2,2],-\frac{p_4+p_5}{1-p_4-p_5}\right) \\
& + \frac{p_2p_3(p_6+p_7)}{(p_2+p_3)}(N-1)(1-p_2-p_3)^{N-2}F\left([1,1,2-N],[2,2],-\frac{p_2+p_3}{1-p_2-p_3}\right) \\
& + \frac{p_2p_3(p_6+p_7)^2}{(p_2+p_3)}(N-1)(N-2)(1-p_2-p_3)^{N-3}F\left([1,1,3-N],[2,2],-\frac{p_2+p_3}{1-p_2-p_3}\right) \\
& + \frac{p_0p_1(p_4+p_5)}{(p_0+p_1)}(N-1)(1-p_0-p_1)^{N-2}F\left([1,1,2-N],[2,2],-\frac{p_0+p_1}{1-p_0-p_1}\right) \\
& + \frac{p_0p_1(p_4+p_5)^2}{(p_0+p_1)}(N-1)(N-2)(1-p_0-p_1)^{N-3}F\left([1,1,3-N],[2,2],-\frac{p_0+p_1}{1-p_0-p_1}\right) \\
& + \frac{p_6p_7}{(p_6+p_7)^2}(p_2+p_3+p_Z) + \frac{p_4p_5}{(p_4+p_5)^2}(p_0+p_1+1-p_Z) \\
& + \frac{p_2p_3}{(p_2+p_3)^2}(p_6+p_7+p_Z) + \frac{p_0p_1}{(p_0+p_1)^2}(p_4+p_5+1-p_Z) \\
& + \left[\frac{p_7}{(p_6+p_7)} - \frac{p_5}{(p_4+p_5)} - \frac{p_3}{(p_2+p_3)} + \frac{p_1}{(p_0+p_1)}\right]^2 p_Z(1-p_Z)
\end{aligned}$$

with adjustment

Full details Kuipers & Moffa, Journal of Causal Inference (2022)

What does the relative variance look like?

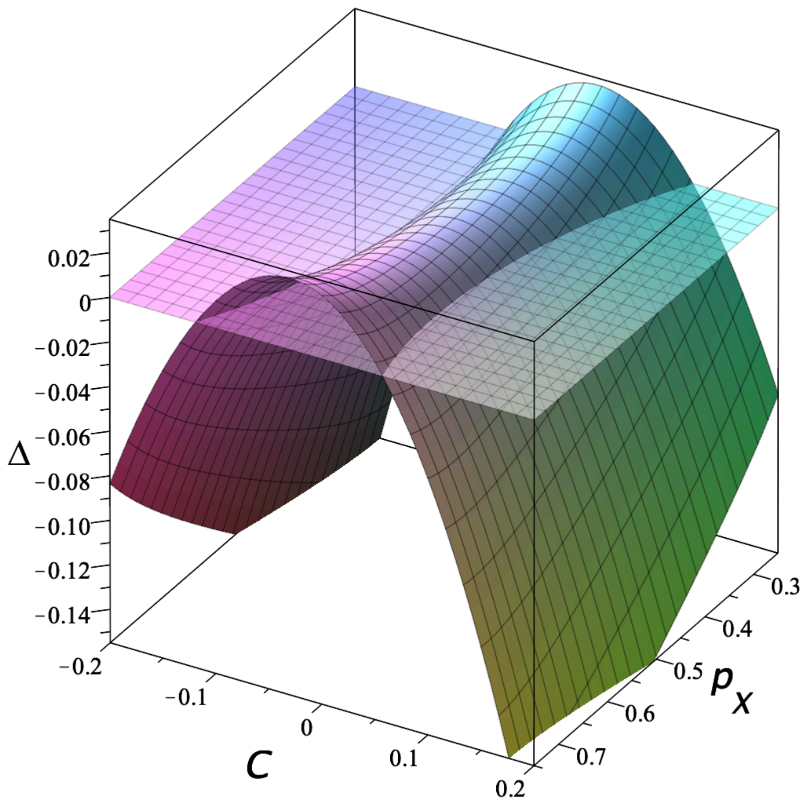


$N = 100$ $N = 400$

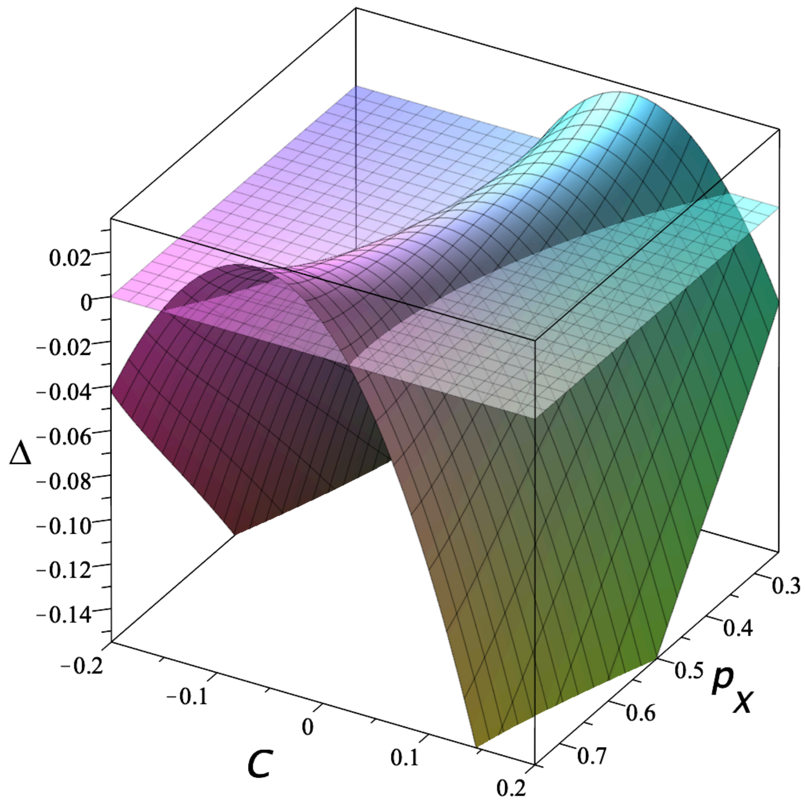
Plot relative difference in variance: $\Delta = \frac{V[M]-V[R]}{V[R]}$

C represents edge strength from $Z \to Y$ (twice the causal effect of Z on Y)

Moderation/Interactions/Product terms



$D = \frac{1}{8}$



$D = \frac{1}{4}$

D is a measure of interaction/moderation of Z and X on Y
(twice the change in causal effect of Z on Y, when changing X)

What does it mean?

There is a parameter regime where it is better not to adjust

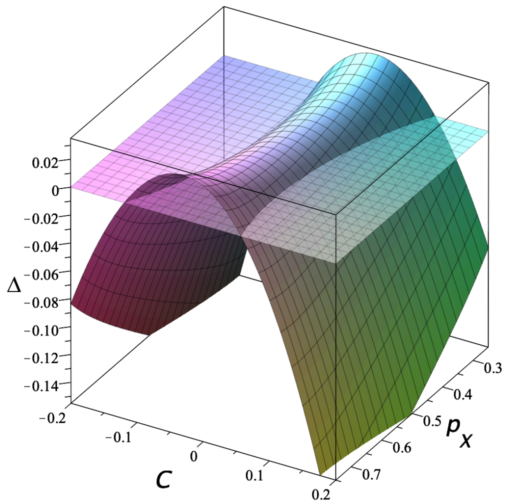
The choice of whether to adjust depends on

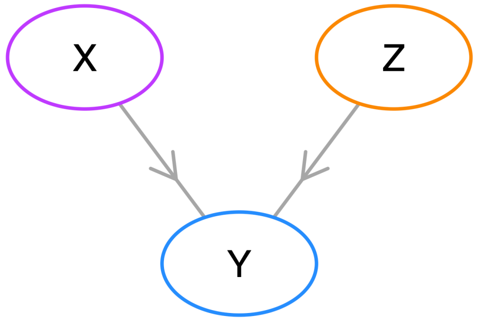
- the strength of the edge $Z \to Y$
- and on the strength of moderation

This is in contrast to leading-order asymptotic results Henckel et al. (2019); Rotnitzky & Smucler (2020)

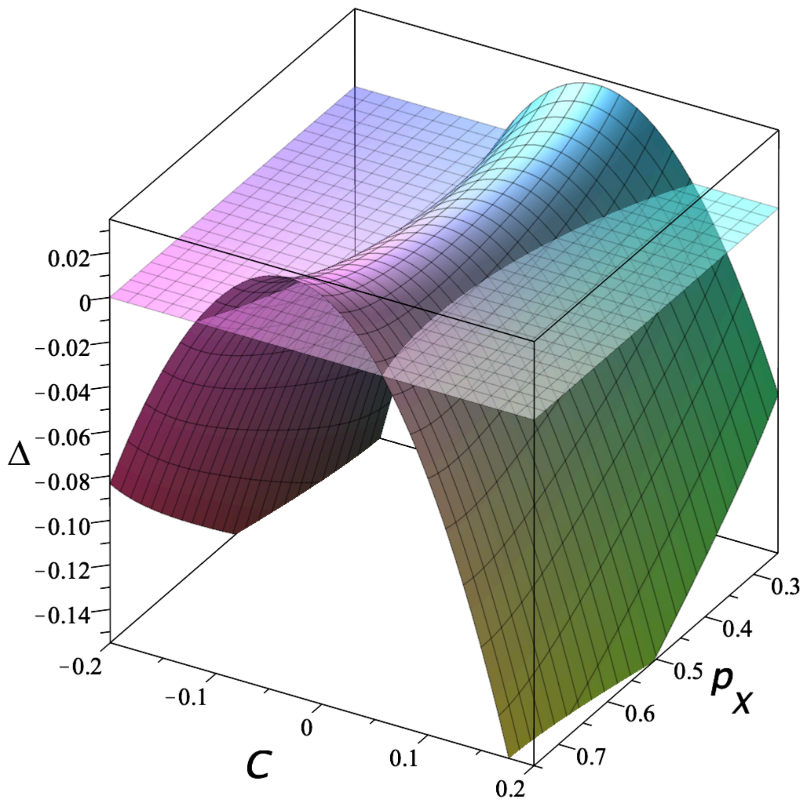
- where the criteria are purely graphical

Can be better not to adjust even when the edge $Z \to Y$ is strong enough to be detectable (through the AIC)

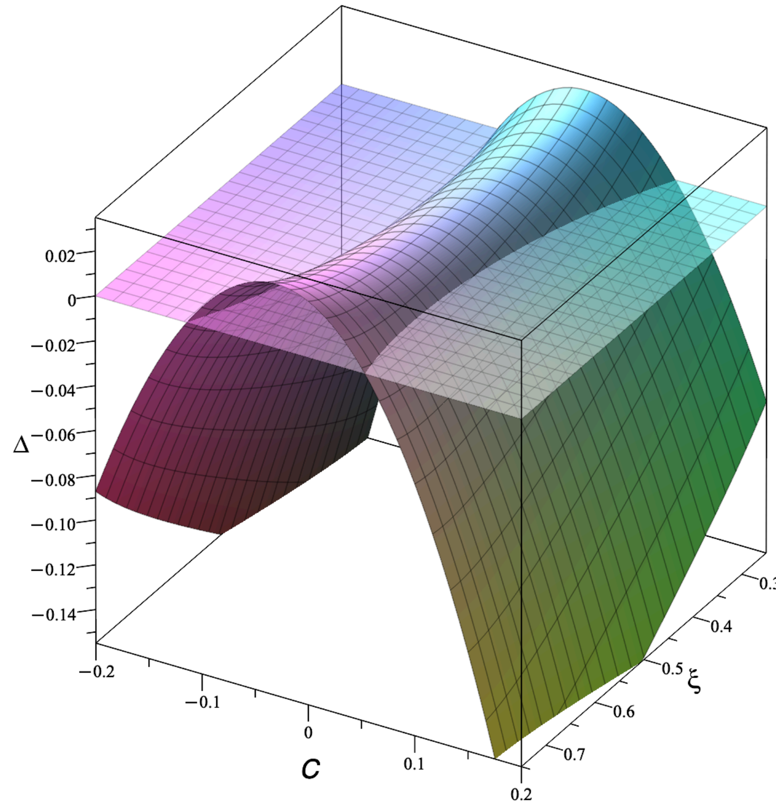




Block randomisation



X random



X fixed

Similar results when block randomising X (predefined number in each category)

Covariance of causal effect estimators

As R and M are estimators of the effect of X on Y

- so is any linear combination of them

$$P = \alpha R + (1 - \alpha) M$$

Variance is

$$V[P] = \alpha^2 V[R] + (1-\alpha)^2 V[M] + 2\alpha(1-\alpha) C[R, M]$$

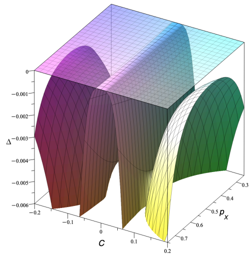
- and lower (at optimal α) than for R and M when

$$C[R, M] < V[R] \quad \wedge \quad C[R, M] < V[M]$$

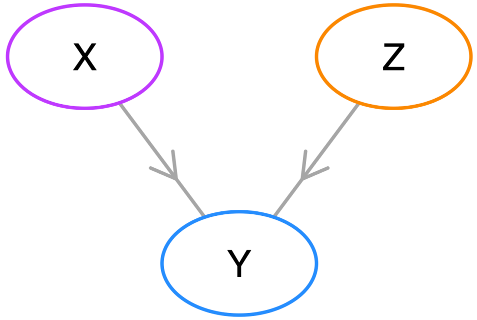
With some asymptotics, this is indeed the case!

$$(C[R, M] - V[R]) \cdot N = -\frac{p_Z(1-p_Z)}{p_X(1-p_X)} \left[2C + (2p_X - 1) D\right]^2 + O(N^{-\frac{3}{2}})$$

$$(C[R, M] - V[M]) \cdot N = -\frac{q_1(1-q_1)(1-p_X)}{N p_X^2} - \frac{q_0(1-q_0)p_X}{N(1-p_X)^2} + O(N^{-\frac{3}{2}})$$



Summary

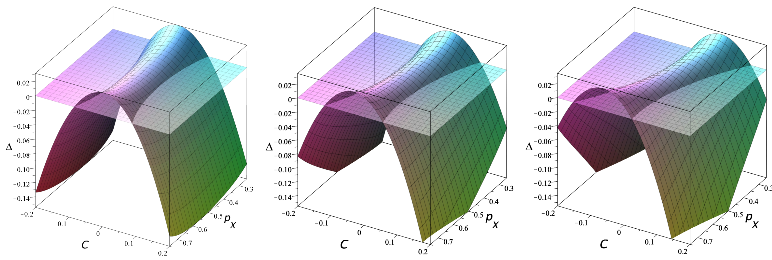


For the simplest non-trivial system

- can analyse analytically
  Kuipers & Moffa, Journal of Causal Inference (2022)

Whether to adjust is **not** purely graphical

- depends on the parameters
- especially the strength of $Z \longrightarrow Y$



Theoretically best result Kuipers & Moffa, arXiv:2503.14242

- combination of estimators

$$P = \alpha R + (1 - \alpha) M$$

All holds whether X is random or blocked

Jack Kuipers
Giusi Moffa

ETH zürich University of Basel Computational Biology Group