Causal inference is not a statistical problem

Lucy D'Agostino McGowan

Introduction

Anscombe's quartet is a set of four datasets that have the same statistical properties in terms of summary statistics (means, variances, correlations, and linear regression fits), but exhibit different distributions and relationships when plotted on a graph. The quartet, often used to teach introductory statistics courses, was created to illustrate the importance of visualizing data before drawing conclusions based on statistical analyses alone. Here, we propose a different quartet, where again statistical summaries do not provide insight into the true underlying mechanism, but even visualizations do not solve the issue. Here, an understanding or assumption of the data generating mechanism is required to correctly capture the relationship between the available factors. This proposed quartet is meant to help readers better understand the assumptions underlying causal inference methods, further driving home the point that in order to accurately estimate causal effects we require more information than can be gleaned from statistical tools alone.

The data generated to create the figures displayed here are included in an R package titled `causalquartet`.

Methods

We propose the following four data generation mechanisms, summarized by the equations below as well as the directed acyclic graphs displayed in Figure 1. Here, X is presumed to be some exposure of interest, Y an outcome, and Z a known, measured factor. The M-Bias equation includes two additional unmeasured factors, U_1 and U_2.

(1) Collider:

$$X \sim N(0, 1)$$

$$Y = X + \varepsilon_y,\ \varepsilon_y \sim N(0, 1)$$

$$Z = 0.45X + 0.77Y + \varepsilon_z$$

,

$$\varepsilon_z \sim N(0,1)$$

(2) Confounder:

$$Z \sim N(0,1)$$

$$X = Z + \varepsilon_x,\ \varepsilon_x \sim N(0,1)$$

$$Y = 0.5Z + \varepsilon_y,\ \varepsilon_y \sim N(0,1)$$

(3) Mediator:

$$X \sim N(0,1)$$

$$Z = X + \varepsilon_z,\ \varepsilon_z \sim N(0,1)$$

$$Y = Z + \varepsilon_y,\ \varepsilon_y \sim N(0,1)$$

(4) M-Bias:

$$U_1 \sim N(0,1)$$

$$U_2 \sim N(0,1)$$

$$Z = 8U_1 + U_2 + \varepsilon_z,\ \varepsilon_z \sim N(0,1)$$

$$X = U_1 + \varepsilon_x,\ \varepsilon_x \sim N(0,1)$$

$$Y = X + U_2 + \varepsilon_y,\ \varepsilon_y \sim N(0,1)$$

In each of these scenarios, a linear model fit to estimate the relationship between X and Y with no further adjustment will result in a $\hat{\beta}$ coefficient of 1. Additionally, the correlation between X and the additional known factor Z is 0.70.

We have simulated 100 data points from each of the four mechanisms, each is displayed in Figure 2. This set of figures demonstrates that despite the very different data generating mechanisms, there is not a clear way to determine the "appropriate" way to model the effect of the exposure X and the outcome Y without additional information. For example, the unadjusted models are displayed in Figure 2, showing a relationship between X and Y of 1. This is the correct causal model for data generating mechanisms (1) and (4), however it overstates the effect of X for data generating mechanism (2), and describes the total effect of X on Y for data generating mechanism (3), but not the direct effect (Table 1). Indeed, even examining the correlation between X and the known factor Z does not help us determine whether adjusting for Z is appropriate, as it is 0.7 in all cases (Table 2).

Table 1: Correct causal models and causal effects for each data generating mechanism.

Data generating mechanism	Correct causal model	Correct causal effect
(1) Collider	Y ~ X	1
(2) Confounder	Y ~ X + Z	0.5
(3) Mediator	Direct effect: Y ~ X + Z	Direct effect: 0
	Total Effect: Y ~ X	Total effect: 1
(4) M-Bias	Y ~ X	1

Table 2: Coefficients for the exposure under each data generating mechanism depending on the model fit as well as the correlation between X and Z.

Data generating mechanism	Y ~ X	Y ~ X + Z	Correlation of X and Z
(1) Collider	1	0.55	0.7
(2) Confounder	1	0.50	0.7
(3) Mediator	1	0.00	0.7
(4) M-Bias	1	0.88	0.7

Discussion

Here we have demonstrated that when presented with an exposure, outcome, and some measured factors, statistics alone, whether summary statistics or data visualizations, are not sufficient to determine the appropriate causal estimate. Additional information about the data generating mechanism is needed in order to draw the correct conclusions.

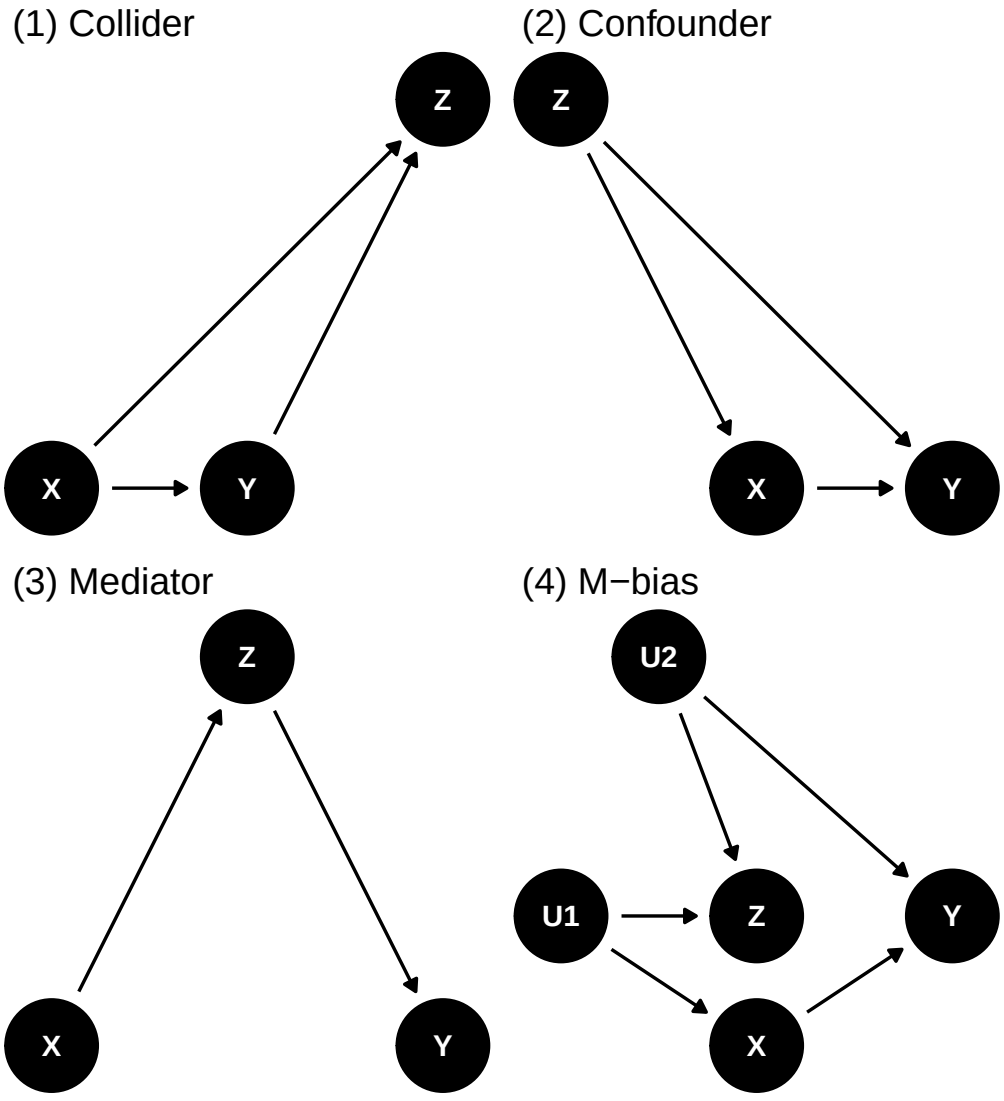

Figure 1: Directed Acyclic Graphs describing the four data generating mechanisms: (1) Collider (2) Confounder (3) Mediator (4) M-Bias.

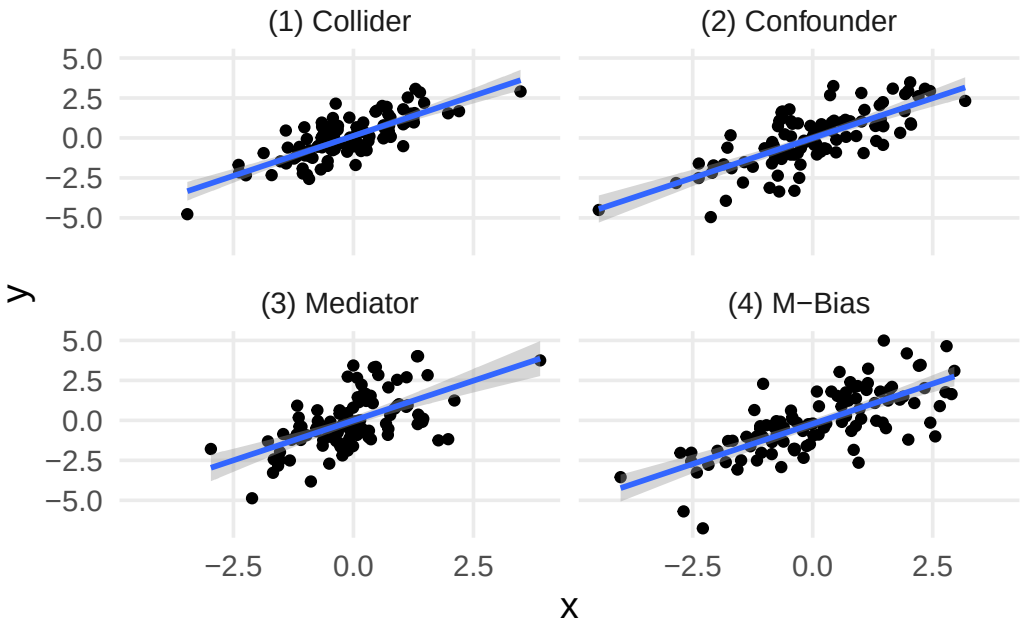

Figure 2: 100 points generated using the data generating mechanisms specified (1) Collider (2) Confounder (3) Mediator (4) M-Bias The blue line displays a linear regression fit estimating the relationship between X and Y, in each case the slope is 1.

Appendix

R code to generate the tables and figures:

```
library(tidyverse)
# devtools::install_github("LucyMcGowan/causalquartet")
library(causalquartet)

## Figure 2

ggplot(causalquartet, aes(x = x, y = y)) +
  geom_point() +
  geom_smooth(method = "lm", formula = "y ~ x") +
  facet_wrap(~type)

## Table 2

causalquartet %>%
  nest_by(type) %>%
  mutate(`Y ~ X` = round(coef(lm(y ~ x, data = data))[2], 2),
         `Y ~ X + Z` = round(coef(lm(y ~ x + z, data = data))[2], 2),
         `Correlation of X and Z` = round(cor(data$x, data$z), 2)) %>%
  select(-data, `Data generating mechanism` = type) %>%
  knitr::kable()
```