Causal Graphs for Basic Epidemiologic Data

Part 3—Measuring causal effects using *do*-expressions and graph surgery

Tomás J. Aragón

2019-11-22

Case studies revisited: stories behind the data

Data are not sufficient to draw causal inferences: we must know how the data was generated. In other words, we must know the "story behind the data."

Case study 1

This was an observational study where 700 patients were given access to a new drug for an ailment. A total of 350 patients chose to take the drug and 350 patients did not. The patients were assessed for clinical recovery. Here are some additional facts:

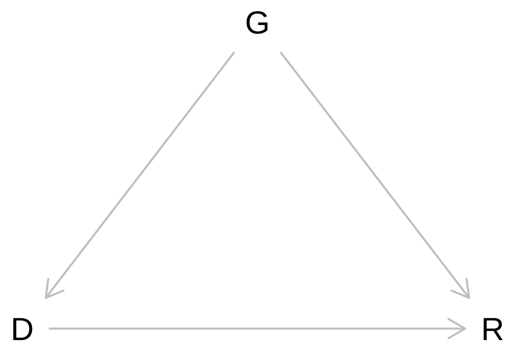
- Estrogen has a negative effect on recovery
- Women are more likely to take the drug compared to men

To analyze this data we had

- Designed a causal graph based on the "story behind the data" (see below)
- Set the conditional probabilities using data or experts

Now we need to evaluate the primary causal question: does consuming the available drug improve recovery after adjusting for potential biases. From the previous work we had designed the following causal graph which we display using the `dagitty` package:

```
library(dagitty)
g1 <- dagitty('dag{ D -> R  G -> D G -> R }')
coordinates(g1) <- list(x = c(D = 0, G = 1, R = 2),
                        y = c(D = 1, G = 0, R = 1))
plot(g1)
```



We want to know what is the causal effect of `D` on `R`? Can we use the causal graph to determine the causal effect *as if* we had "intervened" directly. By *intervene* we mean:

- giving every person in the affected population the drug and assessing recovery, *and*
- not giving every person in the affected population the drug and assessing recovery.

To simulate this intervention we use *do*-expressions. The population causal effect of giving everyone the drug is expressed as

$$P(R = 1 \mid do(D = 1))$$

and the population causal effect of not giving everyone the drug is expressed as follows:

$$P(R = 1 \mid do(D = 0))$$

Therefore, the *average causal effect* (ACE) is the *causal effect difference* using *do*-expressions:

$$ACE = P(R = 1 \mid do(D = 1)) - P(R = 1 \mid do(D = 0))$$

How are *do*-interventions expressed in causal graphs? We design a *manipulated* causal graph using "graph surgery" (Figure @ref(fig:surg1)).

```
g1 <- dagitty('dag{ do -> D -> R  G -> R }')
coordinates(g1) <- list(x = c(do =   0, D = 0, G = 1, R = 2),
                        y = c(do = 0.5, D = 1, G = 0, R = 1))
plot(g1)
```

Next, we want to evaluate the original causal graph in a way that is equivalent to the *do*-intervention causal graph with its new P_m distribution. From Pearl [CITE] we have the following derivation:

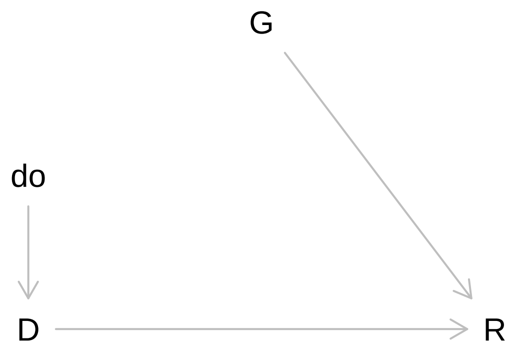


Figure 1: Modified causal graph represents *do*-intervention with new P_m distribution

$$
\begin{aligned}
P(R = r \mid do(D = d)) &= P_m(R = r \mid D = d) && \text{(by definition)} \\
&= P_m(r \mid d) && \text{(for notational convenience)} \\
&= \frac{P_m(r, d)}{P_m(d)} \\
&= \sum_g \frac{P_m(r, d, g)}{P_m(d)} && \text{(by Law of Total Probability)} \\
&= \sum_g \frac{P_m(r, d, g)}{P_m(d)} \frac{P_m(d, g)}{P_m(d, g)} && \text{(multiplying by 1)} \\
&= \sum_g \frac{P_m(r, d, g)}{P_m(d, g)} \frac{P_m(d, g)}{P_m(d)} && \text{(by rearrangement)} \\
&= \sum_g P_m(r \mid d, g) P_m(g \mid d) \\
&= \sum_g P_m(r \mid d, g) P_m(g) && \text{(by independence in manipulated graph)} \\
&= \sum_g P(r \mid d, g) P(g) && \text{(by invariance comparing graphs)} \\
&= \frac{\sum_g P(r, d, g)}{P(d \mid g)} && \text{(by rearrangement)}
\end{aligned}
$$

To summarize, we have derived the *adjustment formula*; in this case "adjusting for gender."

$$
\begin{aligned}
P(R=r \mid do(D=d)) &= \sum_g P(R=r \mid D=d, G=g)P(G=g) \\
&= \sum_g \frac{P(R=r \mid D=d, G=g)}{P(D=d \mid G=g)}
\end{aligned}
$$

This means we must use the contingency table stratified by Gender and the probabilities from the conditional probability table from the previous analysis.

```
bn1.mle$R$prob
```

```
## , , G = Men
##
##      D
## R              No        Yes
##   No  0.13333333 0.06896552
##   Yes 0.86666667 0.93103448
##
## , , G = Women
##
##      D
## R              No        Yes
##   No  0.31250000 0.26996198
##   Yes 0.68750000 0.73003802
```

```
bn1.mle$G$prob
```

```
##   Men Women
##  0.51  0.49
```

Next, we must calculate these *do*-intervention probabilities from the observed conditional probabilities. Here is the adjustment formula for the population without drug treatment:

$$P(R=1 \mid do(D=0)) = P(R=1 \mid D=0, G=1)P(G=1) + P(R=1 \mid D=0, G=2)P(G=2)$$

and here is the adjustment formula for the population with drug treatment:

$$P(R=1 \mid do(D=1)) = P(R=1 \mid D=1, G=1)P(G=1) + P(R=1 \mid D=1, G=2)P(G=2)$$

where $G=1$ for men and $G=2$ for women. Using the conditional probability tables,

```r
Pr.R1_D1.G1 <- bn1.mle$R$prob['Yes','Yes','Men']
Pr.R1_D1.G2 <- bn1.mle$R$prob['Yes','Yes','Women']
Pr.G1 <- bn1.mle$G$prob['Men']
Pr.G2 <- bn1.mle$G$prob['Women']
(Pr.R1_do.D1 <- unname(Pr.R1_D1.G1 * Pr.G1 + Pr.R1_D1.G2 * Pr.G2))
```

```
## [1] 0.8325462
```

```r
Pr.R1_D0.G1 <- bn1.mle$R$prob['Yes','No','Men']
Pr.R1_D0.G2 <- bn1.mle$R$prob['Yes','No','Women']
(Pr.R1_do.D0 <- unname(Pr.R1_D0.G1 * Pr.G1 + Pr.R1_D0.G2 * Pr.G2))
```

```
## [1] 0.778875
```

```r
(ACE1 <- Pr.R1_do.D1 - Pr.R1_do.D0)
```

```
## [1] 0.05367122
```

An alternative calculation approach is to notice the matrix alegebra structure:

$$\begin{bmatrix} P(R=1 \mid D=0, G=1) & P(R=1 \mid D=0, G=2) \\ P(R=1 \mid D=1, G=1) & P(R=1 \mid D=1, G=2) \end{bmatrix} \begin{bmatrix} P(G=1) \\ P(G=2) \end{bmatrix}$$

```r
(R.mtx <- bn1.mle$R$prob['Yes',,])
```

```
##      G
## D           Men     Women
##   No  0.8666667 0.6875000
##   Yes 0.9310345 0.7300380
```

```r
(G.vec <- bn1.mle$G$prob)
```

```
##   Men Women
##  0.51  0.49
```

```r
(Pr.R1_do.D <- R.mtx %*% G.vec)
```

```
##
## D          [,1]
##   No  0.7788750
##   Yes 0.8325462
```

```r
(ACE <- diff(Pr.R1_do.D))
```

```
##
## D           [,1]
##   Yes 0.05367122
```

Case study 2

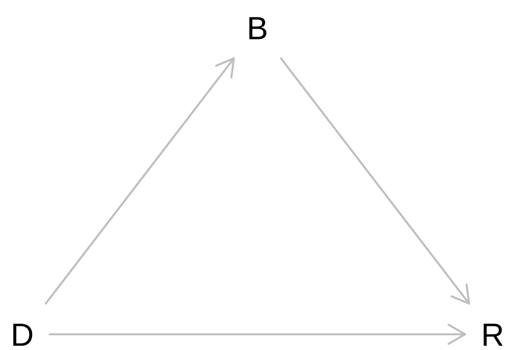
Again data are not sufficient to draw causal inferences: we must know how the data was generated. In other words, we must know the "story behind the data."

This was a treatment study with 700 patients, half of whom were assigned a new drug for their ailment. At the end of the study the patients were assessed for clinical recovery, and their blood pressure was measured. Here are some additional facts:

- Blood pressure was measured at the end of the study
- Drug treatment affects recovery by lowering blood pressure
- Lowering blood pressure also has toxic side effects

Now we need to evaluate the primary causal question: does consuming the available drug improve recovery after adjusting for potential biases. From the previous work we had designed the following causal graph which we display using the `dagitty` package:

```
library(dagitty)
g2 <- dagitty('dag{ D -> R  D -> B  B -> R }')
coordinates(g2) <- list(x = c(D = 0, B = 1, R = 2),
                        y = c(D = 1, B = 0, R = 1))
plot(g2)
```



We want to know what is the causal effect of `D` on `R`? Can we use the causal graph to determine the causal effect *as if* we had "intervened" directly. By *intervene* we mean:

- giving every person in the affected population the drug and assessing recovery, *and*
- not giving every person in the affected population the drug and assessing recovery.

To simulate this intervention we use *do*-expressions. The population causal effect of giving everyone the drug is expressed as

$$P(R = 1 \mid do(D = 1))$$

and the population causal effect of not giving everyone the drug is expressed as follows:

$$P(R = 1 \mid do(D = 0))$$

Therefore, the *average causal effect* (ACE) is the *causal effect difference* using *do*-expressions:

$$ACE = P(R = 1 \mid do(D = 1)) - P(R = 1 \mid do(D = 0))$$

How are *do*-interventions expressed in causal graphs?

Here is the adjustment formula for this causal graph:

$$P(R = r \mid do(D = d)) = P(R = r \mid D = d)$$

And the ACE,

$$ACE = P(R = 1 \mid D = 1) - P(R = 1 \mid D = 0)$$