Collider Effect in the Analysis of Observational Data: Epidemiological Insight and Interactive Visualization

Correlation between a variable with the exposure and the outcome in the analysis of epidemiological data is common but correlation does not mean causation

Miguel Angel Luque-Fernandez[1], Daniel Rodondo-Sanchez[2], Michael Schomaker[3]

1 Introduction

Classical epidemiology during the last 30 years has focussed on the control of confounding [1]. However, it is just recently that epidemiologists have started to describe the bias produced by other source of structures such as colliders and mediators [2, 3]. In the epidemiological literature different explanations have been proposed to describe the paradoxical protective effect of established risk factors on an outcome such as the birth weight and the pre-eclampsia smoking paradoxes [4, 5]. The use of direct acyclic graphs (DAGs) help to visualize these new structures and distinguishes between biases resulting from (inappropriate) conditioning on common effects (collider bias) and lack of conditioning on common causes of exposure and outcome (confounding) [6, 7](Figure 1). A collider for a certain pair of variables such as an exposure and an outcome is third variable that is causally influenced by both of them. Controlling for, or conditioning analysis on, such a variable can introduce a spurious association between its causes (exposure and outcome) explaining why medical literature is plenty of paradoxical findings [8]. In DAG terminology, a collider is the variable in the middle of an inverted fork (i.e., variable Z in A $\rightarrow$ Z $\leftarrow$ Y) [9]. This methodological note is structured as follows: i) we unveil the statistical structure of the collider bias through a simulated data generation, ii) we illustrate the effect of conditioning on a collider based on a noncommunicable disease epidemiology example, iii) we provide R-code in easy-to-read boxes throughout the manuscript and in a GitHub repository: https://github.com/migariane/ColliderApp for replicability and iv) we provide readers with a Shiny application allowing to dynamically visualize the effect of a collider.

2 Collider statistical structure and simulation

```
# Define Intervention: start ART if CD4 count < 750 or CD4% < 25%
cd4_750.1  <- (mydata_wide$cd4a_cf.1 < sqrt(750)  | mydata_wide$cd4p_cf.1 < 25)
cd4_750.3  <- (mydata_wide$cd4a_cf.3 < sqrt(750)  | mydata_wide$cd4p_cf.3 < 25)   | cd4_750.1
cd4_750.6  <- (mydata_wide$cd4a_cf.6 < sqrt(750)  | mydata_wide$cd4p_cf.6 < 25)   | cd4_750.3
abar750 <- as.matrix(cbind(cd4_750.1,cd4_750.3,cd4_750.6))
abar750[is.na(abar750)] <- 0
# Define collection of estimation methods to be used by SuperLearner
mylibrary <- list(Q=c("SL.glm", "SL.stepAIC", "SL.step.interaction", "SL.gam"),
                  g=c("SL.glm", "SL.stepAIC", "SL.gam"))
```

[1]Biomedical Research Institute. Non-Communicable and Cancer Epidemiology Group (ibs.Granada), Andalusian School of Public Health, University of Granada, Granada, Spain, miguel.luque.easp@juntadeandalucia.es

[2]Biomedical Research Institute. Non-Communicable and Cancer Epidemiology Group (ibs.Granada), Andalusian School of Public Health, University of Granada, Granada, Spain, daniel.redondo.easp@juntadeandalucia.es

[3]University of Cape Town, Centre for Infectious Disease Epidemiology and Research, Observatory, 7925; Cape Town, South Africa, michael.schomaker@uct.ac.za

3 Motivating Example

4 Collider effect

5 Conclusion

Making causal inferences on the basis of correlational data is very hard.

Contributors and sources

6 Key message

-
-
-
-

Acknowledgements

Miguel Angel Luque Fernandez is supported by the Spanish National Institute of Health, Carlos III Miguel Servet I Investigator Award (CP17/00206).

7 Figures and Tables

Table 1: Univariate, bivariate and multivariate adjusted models for the association between systolic blood pressure and age, n = 1,000

	Dependent variable:		
	Systolic Blood Pressure in mmhg		
	(Univarite)	(Bivariate)	(Multivariate collider)
Age in years	1.972***	1.255***	−0.319***
	(0.062)	(0.007)	(0.030)
Proteinuria in mg			0.419***
			(0.008)
Sodium intake in g		9.989***	2.825***
		(0.032)	(0.136)
Constant	−3.570	−0.311	−0.105
	(4.039)	(0.407)	(0.208)
Observations	1,000	1,000	1,000
R^2	0.504	0.995	0.999
Adjusted R^2	0.503	0.995	0.999
Residual Std. Error	9.757 (df = 998)	0.983 (df = 997)	0.502 (df = 996)
F Statistic	1,013.503*** (df = 1; 998)	98,648.120*** (df = 2; 997)	252,906.500*** (df = 3; 9
Note:			*p<0.1; **p<0.05; ***p<0

8 Annex

```
library(visreg)
library(broom)
library(stargazer)

## Data Generation based on Noncommunicable Disease Epidemiology
generateData <- function(n){
    age <- rnorm(n, 65, 5)
    sodium  <- age/15 + rnorm(n)
    sbp <- 10*sodium + 1.25*age + rnorm(n)
    proteinurie  <- 1.5*age + 1.8*sbp - 0.9*sodium + rnorm(n)
    data.frame(sbp,age,sodium,proteinurie)
}

set.seed(777)
ObsData <- generateData(n=1000)

## Models Fit
fit0 <- lm(sbp ~ age, data = ObsData);tidy(fit0)
fit1 <- lm(sbp ~ age + sodium , data = ObsData);tidy(fit1)
fit2 <- lm(sbp ~ age + proteinurie + sodium, data = ObsData);tidy(fit2)

## Models visualization
par(mfrow=c(1,3))
visreg(fit0, ylab = "SBP in mmhg", line=list(col="blue"),  points = list(cex = 1.5, pch = 1), ji
visreg(fit1, ylab = "SBP in mmhg", line=list(col="blue"),  points = list(cex = 1.5, pch = 1), ji
visreg(fit2, ylab = "SBP in mmhg", line=list(col="red"),   points = list(cex = 1.5, pch = 1), ji

## Comparing Models
stargazer(fit0, fit1, fit2, type = "latex", multicolumn = FALSE)
```