Collider Effect in the Analysis of Epidemiological Observational Data: Statistical Insights and Interactive Reactive 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 a 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]. While this methodological note will not close the vexing gap between correlation and causation, it will unveil the statistical structure of the so called colliders. The rest of the methodological note is structured as follows: i) we unveil the statistical structure of the collider bias through a simulated data generation using a linear system of structural equations, ii) we illustrate the effect of conditioning on a collider based on a noncommunicable disease epidemiology simulated and easy to interpret 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 https://watzilei.com/Epicollider.

2 Collider statistical structure

2.1 Reviewing confounding

Regression is the tool that classically epidemiologist use to address causal questions in an observational studies. Let's begin reviewing what we mean with Controlling for factors A (treatment or exposure), W (confounder), and Y (outcome), () via the following causal graph (Figure 1A). In the above diagram, Z is a confounder and will distort the perceived causal relationship between A and Y if unaccounted for. We now review adjustment for confounding via linear regression models. In Box-one we generate data consistent with Figure 1A and run both unadjusted and adjusted regression models to illustrate "confounding" through the interpretation of regression models coefficients and the visualization of the linear fit for the adjusted model.

[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

```
N <- 100000
library(visreg)
W <- rnorm(N)
A <- 0.5 * W + rnorm(N)
Y <- 0.3 * A + 0.4 * W + rnorm(N)
UnadjustedModel <- summary(lm(Y ~ A))
AdjustedModel   <- summary(lm(Y ~ A + W))
visreg(AdjustedModel)
```

Note that our confounder W is the only variable that is exogenous i.e. it is not dependent or generated from the outcome Y or the exposure A and is the only variable that we can set in any way we want. However, both A and Y depend on Z, and they are endogenous to the system of linear equations. This is too the mechanism of randomization in an experimental design where by design Z is exogenous. Thus it does not depends of A or Y. The first regression (that ignores W) incurs upward bias in Ws coefficient due to the confounders positive effects on both W and Y. The second regression including the confounder W recovers zs true coefficient while increasing the R-squared by a few percentage points thus statistically speaking, adjusting for W improves our model.

2.2 Collider simulation

Unlike in Figure 1A where the causal arrows emanate from W, they now point towards W from A and Y. If you sent marbles moving in the direction of the arrows above, two of them might collide at w, earning it the label collider. Causal diagram theory says that if you condition on a collider, you create an artificial situation that appears as if the directions of the arrows pointing toward the collider have flipped. Take a moment to think about that. Do you see the problem? If we condition on w, we will have created a confounder. Lets see it happen in R. First, we generate the data, again using a simple linear data generating mechanism. Notice that x is now exogenous while w is endogenous.

x ¡- rnorm(N) y ¡- 0.7 * X + rnorm(N) w ¡- 1.2 * X + 0.9 * Y + rnorm(N) summary(lm(y x)) fit ¡- lm(y x + w) summary(fit) visreg(fit)

Unlike in previous section, the simpler regression without w recovers the true coefficient of x, while the regression with w has a horribly biased estimate. But the second model is not unequivocally inferior; it has an R-squared thats roughly 20 percentage points higher than the first! The collider w might have ruined our regression coefficients, but it still helps us predict and is an important part of the conditional expectation function E(y—x, z). Unfortunately, you cant in general rely on that function for understanding how the world works.

Summary

We investigated a situation where adding a certain type of variable to a regression, called a collider, will bias coefficients while still increasing predictive power. Whether this is good or bad depends on the research objective. If the goal is to obtain a predictive model that makes accurate predictions of the response, its good. If the goal is to create a model of reality that is useful in making decisions, then the collider bias is almost certainly a bad thing.

Of course, discarding a variable that adds to the predictive power of a model is easier said than done. Models in any organization are evaluated by some metric, typically a predictive one, and trying to convince your peers (and boss) that your model is better because it doesnt have colliders may be a tough sell.

Determining if a variable is a collider involves thinking critically about the way the world works. If an explanatory variable could actually be caused by the response as well as another predictor, then you have a candidate that is perhaps better left out of the regression. Its a complex world out there with many difficult decisions, not all of which can be based on data. This author wishes you a 2016 with more difficult decisions that in 2015, in regression analysis at least!

```
# 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
```

many*difficult decisions, not all of which can be based on data. This author wishes you a 2016 with more difficult decisions that in 2015, in regression analysis at least!

```
# 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"))
```

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)
```

References

[1] Sander Greenland and Hal Morgenstern. Confounding in health research. *Annual Review of Public Health*, 22(1):189–212, May 2001.

[2] Stephen R Cole, Robert W Platt, Enrique F Schisterman, Haitao Chu, Daniel Westreich, David Richardson, and Charles Poole. Illustrating bias due to conditioning on a collider. *International Journal of Epidemiology*, 39(2):417–420, Nov 2009.

[3] Tyler J. Vanderweele and Stijn Vansteelandt. Conceptual issues concerning mediation, interventions and composition. *Statistics and Its Interface*, 2(4):457–468, 2009.

[4] Miguel Angel Luque-Fernandez, Helga Zoega, Unnur Valdimarsdottir, and Michelle A. Williams. Deconstructing the smoking-preeclampsia paradox through a counterfactual framework. *European Journal of Epidemiology*, 31(6):613–623, Jun 2016.

[5] S. Hernandez-Diaz, E. F. Schisterman, and M. A. Hernan. The birth weight "paradox" uncovered? *American Journal of Epidemiology*, 164(11):1115–1120, Sep 2006.

[6] Miguel A. Hernán, Sonia Hernández-Díaz, and James M. Robins. A structural approach to selection bias. *Epidemiology*, 15(5):615–625, Sep 2004.

[7] James M. Robins, Miguel Ángel Hernán, and Babette Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560, Sep 2000.

[8] Julia M Rohrer. Thinking clearly about correlations and causation: Graphical causal models for observational data. 2017.

[9] JUDEA PEARL. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.