Bias amplification for non-Gaussian data in environmental epidemiology

Suyeon Kang*, Minsoo Kim, Seungpil Jung, Woojoo Lee

Department of Statistics, INHA University

Introduction

- Recent environmental epidemiology studies have shown importance of multiple pollutant exposure on health and diseases development.
- A common aim in epidemiology is to identify and accurately measure causal relationship between exposures and outcomes, controlling the confounders.
- Researchers have employed a number of regression models to measure an effect of exposure to mixtures.
- In practice, the whole set of exposure variables are used as covariates in the regression model because information for the causal effect of exposures is not sufficient.
- However, it is almost impossible to identify the confounders accurately corresponding to the whole set of exposure variables. In addition, some confounders may not be observed. Therefore, the causal effect estimate is often biased.
- When exposure variables are positively correlated, the biases of causal effect estimates for exposures can be amplified if the whole set of exposure variables are used as covariates in Gaussian linear regression models without controlling the confounders[1].
- In this study, we illustrate the problem of bias amplification using Directed Acyclic Graphs(DAGs) and provide new bias formulae for Poisson and Bernoulli data.

DAGs and Bias Analysis

- DAG is a graphical tool to represent our qualitative knowledge and a priori assumptions about the causal structure of interest.
- Causal effects are represented by directed arrows, for example, in Figure 1, $X_1 \rightarrow Y$ indicates X_1 is a cause of Y.
- The absence of arrow between two variables indicates there is no causal relationship between them.

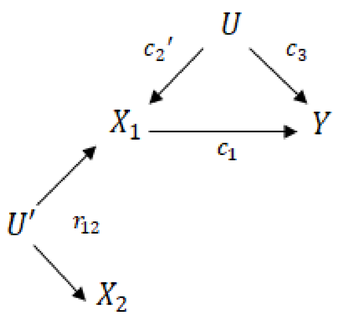

Figure 1. Causal Diagram for scenario 1

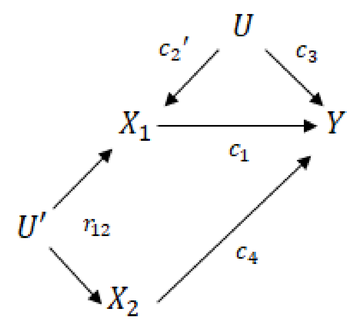

Figure 2. Causal Diagram for scenario 2

X_1, X_2: exposures such that $corr(X_1, X_2) = r_{12}$, Y: outcome, U: unmeasured confounder, U': common source between X_1 and X_2, c_i, c_2': causal effect (i = 1,3,4)

- In Figure 1 and 2, X_1, X_2, U and Y denote exposure 1, exposure 2, unmeasured confounder and outcome, respectively.
- The pathway $X_2 \leftarrow U' \rightarrow X_1$ is a backdoor pathway, and X_1 is a collider in the pathway $X_2 \leftarrow U' \rightarrow X_1 \leftarrow U \rightarrow Y$.
- The backdoor pathway $X_2 \leftarrow U' \rightarrow X_1 \leftarrow U \rightarrow Y$ is blocked if X_1 is not controlled, but opened if it is.
- For the bias analysis, we will consider two scenarios represented in Figure 1 and 2. For each scenario, we perform a simulation study for different distributions of Y and different sample sizes.
- We assume Gaussian, Poisson, Bernoulli for Y. Sample size is 100 or 500.
- In Figure1, $U \sim N(0,1)$, $U' \sim N(0,1)$, $X_1' = r_1 * U' + c_2 * U + \varepsilon$, $\varepsilon \sim N(0,1)$, $X_2' = r_2 * U' + \varepsilon$, $\varepsilon \sim N(0,1)$ and X_1 and X_2 are the centered and scaled X_1' and X_2', respectively. And $g(E(Y)) = c_1 * X_1 + c_3 * U$, the link function $g(\cdot)$ is identified as identity, log and logit function for Gaussian, Poisson and Bernoulli data, respectively.
- In Figure2, $U \sim N(0,1)$, $U' \sim N(0,1)$, $X_1' = r_1 * U' + c_2 * U + \varepsilon$, $\varepsilon \sim N(0,1)$, $X_2' = r_2 * U' + \varepsilon$, $\varepsilon \sim N(0,1)$ and X_1 and X_2 are the centered and scaled X_1' and X_2', respectively. And $g(E(Y)) = c_1 * X_1 + c_3 * U + c_4 * X_2$.
- Analytic bias formulae when both X_1 and X_2 are used as covariates in the generalized linear models (GLMs) are derived and provided in Table 1.

Table 1. Analytic bias formulae for the regression coefficients of X_1 and X_2 for each scenario

Data	Exposure variable	True value	GLMs using X_1 and X_2
Gaussian /Poisson	Figure 1.		
	X_1	c_1	$c_1 + [c_2'c_3/(1-r_{12}^2)]$ [a]
	X_2	0	$-r_{12}\,c_2'c_3/(1-r_{12}^2)$
	Figure 2.		
	X_1	c_1	$c_1 + [c_2'c_3/(1-r_{12}^2)]$
	X_2	c_4	$c_4 - [r_{12}\,c_2'c_3/(1-r_{12}^2)]$
Bernoulli	Figure 1.		
	X_1	c_1	$[c_1 + [c_2'c_3/(1-r_{12}^2)]/\gamma$ [b]
	X_2	0	$[-r_{12}\,c_2'c_3/(1-r_{12}^2)]/\gamma$
	Figure 2.		
	X_1	c_1	$[c_1 + [c_2'c_3/(1-r_{12}^2)]/\gamma$
	X_2	c_4	$[c_4 - [r_{12}\,c_2'c_3/(1-r_{12}^2)]]/\gamma$

[a] $c_2' = \frac{c_2}{\sqrt{r_1^2+c_2^2+1}}$

[b] $\gamma = \sqrt{1+\frac{\pi c_3^2}{8(1-r_{12}^2)}(1-r_{12}^2-\frac{c_2^2}{1+r_1^2+c_2^2})}$

- In Table 1, **"True value"** means the original causal effect that would be identified when the confounder U is controlled. And **"GLMs using X_1 and X_2"** means the asymptotic limit of the maximum likelihood estimator for the regression coefficient associated with the two exposures when the confounder is not controlled.
- For example, in Figure 1 for Gaussian and Poisson data, the original causal effects of X_1 and X_2 are c_1 and 0, respectively. But when the confounder U is not controlled, the asymptotic limits of the corresponding maximum likelihood estimators are $c_1 + [c_2'c_3/(1-r_{12}^2)]$ and $-r_{12}\,c_2'c_3/(1-r_{12}^2)$, respectively.

Numerical Study

- We set $c_1 = 1$, $c_2 = c_3 = c_4 = 2$, $r_1 = r_2 = 5$ (i.e, r_{12}=0.90). We assume that the confounder U is unobserved. Here, two exposures are highly and positively correlated, which reflects the property of real data often observed in environmental epidemiology.
- We report the averages of the regression coefficients when either X_1 or X_2 is used as a covariate in GLMs. In Table 2 and 3, we call this "**Single**". **"Multiple"** denotes the corresponding results when both X_1 and X_2 are used as covariates in GLMs.
- For 1000 independent simulation data, we also report the empirical coverage rates (the proportion that the 95% bootstrap confidence interval includes the true causal effect) in Table 2 and 3.

Simulation result for Figure 1 : bias and empirical coverage rate

Table 2.

Data	Sample size(n)	Exposure variable	True value	Single	Multiple			Bias	
				Estimate	Theoretical	Estimate	Coverage	Single	Multiple
Gaussian	n=100	X_1	1	1.723	**4.675**	4.667	95.10%	0.723	3.675
		X_2	0	0.888	**-3.290**	-3.286	96.00%	0.888	-3.290
	n=500	X_1	1	1.730	**4.675**	4.673	95.50%	0.730	3.675
		X_2	0	0.896	**-3.290**	-3.288	95.00%	0.896	-3.290
Poisson	n=100	X_1	1	1.571	**4.675**	4.533	96.80%	0.571	3.675
		X_2	0	0.750	**-3.290**	-3.200	96.60%	0.750	-3.290
	n=500	X_1	1	1.614	**4.675**	4.575	95.50%	0.614	3.675
		X_2	0	0.810	**-3.290**	-3.225	95.00%	0.810	-3.290
Bernoulli	n=100	X_1	1	1.163	**3.796**	4.071	93.50%	0.163	2.796
		X_2	0	0.502	**-2.671**	-2.879	93.50%	0.502	-2.671
	n=500	X_1	1	1.140	**3.796**	3.879	94.20%	0.140	2.796
		X_2	0	0.497	**-2.671**	-2.725	94.90%	0.497	-2.671

Simulation result for Figure 2 : bias and empirical coverage rate

Table 3.

Data	Sample size(n)	Exposure variable	True value	Single	Multiple			Bias	
				Estimate	Theoretical	Estimate	Coverage	Single	Multiple
Gaussian	n=100	X_1	1	3.519	**4.675**	4.668	94.60%	2.519	3.675
		X_2	2	2.887	**-1.290**	-1.284	95.60%	0.887	-3.290
	n=500	X_1	1	3.522	**4.675**	4.672	95.00%	2.522	3.675
		X_2	2	2.899	**-1.290**	-1.285	94.80%	0.899	-3.290
Poisson	n=100	X_1	1	3.304	**4.675**	4.501	96.70%	2.304	3.675
		X_2	2	2.474	**-1.290**	-1.239	95.50%	0.474	-3.290
	n=500	X_1	1	3.315	**4.675**	4.536	95.60%	2.315	3.675
		X_2	2	2.459	**-1.290**	-1.270	94.00%	0.459	-3.290
Bernoulli	n=100	X_1	1	2.971	**3.796**	4.143	94.10%	1.971	2.796
		X_2	2	1.729	**-1.047**	-1.118	95.00%	-0.271	-3.047
	n=500	X_1	1	2.845	**3.796**	3.934	93.80%	1.845	2.796
		X_2	2	1.672	**-1.047**	-1.081	94.60%	-0.328	-3.047

- From this numerical study, we confirm that the analytic bias formulae are valid.
- We also confirm that under practical situations, when both X_1 and X_2 are used as covariates in the GLMs, the biases are uniformly larger than their counterparts where either X_1 or X_2 is used as a covariate.

Conclusion

- It is important to recognize that the bias can be amplified when several exposure variables are used as covariates in a regression model. In particular, the extent of the bias can be large when exposures are highly correlated.
- We strongly advise researchers not to blindly include multiple exposure variables in a regression model.

Acknowledgement

- Woojoo Lee was supported by a Grant from the Next-Generation BioGreen 21 program (Project No. PJ01337701), Rural Development Administration, Republic of Korea

Reference

[1] Marc G. Weisskopf, Ryan M. Seals, and Thomas F. Webster. (2018). "Bias Amplification in Epidemiologic Analysis of Exposure to Mixtures"