Simulation codes for misperceptions 9

We consider the following regression models to generate (Y, X_A, X_B).

$$
\begin{aligned}
Y &= \beta_0 + \beta_A X_A + \beta_B X_B + \epsilon, \\
X_A &= \lambda_A Z + \eta, \\
X_B &= \lambda_B Z + \gamma,
\end{aligned}
$$

variables Z, ϵ, η, and γ were generated from independent normal distributions with zero means. The variances for Z, ϵ, η, and γ are σ_Z^2, σ_ϵ^2, σ_η^2, and σ_γ^2, respectively.

The following function is used to generate data (Y, X_A, X_B) with β_0.

```
> simDat <- function(n, ba, bb, la, lb, sa, sb, sz, se = 1) {
+   Z <- rnorm(n, 0, sz)
+   Xa <- la * Z + rnorm(n, 0, sa)
+   Xb <- lb * Z + rnorm(n, 0, sb)
+   Y <- ba * Xa + bb * Xb + rnorm(n, 0, se)
+   data.frame(Y = Y, Xa = Xa, Xb = Xb)
+ }
```

The arguments to be supplied in `simDat()` are:

- `n` is an integer value to specify the sample size
- `ba` is a numerical value to specify β_A
- `bb` is a numerical value to specify β_B
- `la` is a numerical value to specify λ_A
- `lb` is a numerical value to specify λ_B
- `sa` is a numerical value to specify σ_η
- `sb` is a numerical value to specify σ_γ
- `sz` is a numerical value to specify σ_Z
- `se` is a numerical value to specify σ_ϵ

The `simDat()` function returns a `data.frame` with `n` rows and three columns. The three columns are `Y`, `Xa`, and `Xb`, representing Y, X_A, and X_B. The following is an example code that generates a simulated data from Scenario I with parameters stated in Table 1 of the main document.

```
> set.seed(0); dat <- simDat(500, .4, -.3, sqrt(.75), sqrt(.75), .5, .5, 1, .75)
> dim(dat)
[1] 500   3
> head(dat)
            Y          Xa          Xb
1  0.08620043  0.78056638  0.9503247
2 -0.24671706 -0.04185872  0.6380271
3  0.28215226  1.99927549  1.0732578
4  1.14112189  0.22134297  0.4070548
5  0.01746587  0.45809652 -0.3774620
6  1.26299240 -1.13496131 -1.3683953
```

The following fits the full model and the reduced model based on `dat`.

```
> ## Full model
> lm(Y ~ Xa + Xb, data = dat)

Call:
lm(formula = Y ~ Xa + Xb, data = dat)

Coefficients:
(Intercept)           Xa           Xb
   0.002032     0.403388    -0.318948
> ## Reduced model
> lm(Y ~ Xa, data = dat)

Call:
lm(formula = Y ~ Xa, data = dat)

Coefficients:
(Intercept)           Xa
   0.006821     0.170744
```

It is frequently said that a confounder must be associated with the independent variable or "exposure" of interest and the outcome of interest. However, there is no universally agreed-upon simple definition of a confound or confounding variable. One definition might be a variable that, if not properly controlled for via experimental design or statistical adjustment, leads to a biased estimate of the causal effect of an independent variable on an outcome as inferred from its association. That is, we might define a confounder as a variable which, if not properly controlled for, leads the association between the independent variable and the dependent variable to not represent the causal effect of the independent variable on the dependent variable. Although not all definitions of confounders would embrace this set of conditions, one could simply say a covariate whose exclusion leads to bias in the association as an indicator of a causal effect also termed "omitted variable bias". Regardless of terminology, it is clear that there can be variables which, if not included in the analysis, will lead to a bias in the observed association between a plausible cause and its postulated outcome as an indicator of the causal effect, even when that covariate has zero correlation with the outcome. One way in which this can occur is through a phenomenon sometimes referred to as suppression [74].

Consider a simple causal model depicted in Figure 3. At the left side of Figure 3, we have a variable of the belief that dietary fat consumption is not dangerous or, conceived alternatively, one minus the strength of belief that dietary fat consumption is dangerous or should be avoided. Suppose this variable relates to dietary consumption of two kinds of dietary fats, A and B, where dietary fat of type A decreases some health outcome of interest. In contrast, dietary fat of type B decreases the negative health outcome. We can use following linear model to describe the causal effects in Figure 3;

$$M_F : Y = \beta_0 + \beta_A X_A + \beta_B X_B + \epsilon, \tag{1}$$

where Y is the response variable, X_A and X_B are independent variables representing the fat consumptions of dietary fat types A and B, respectively, and ϵ is an independent error term with the variance σ^2_ϵ. Of the two covariates, we suppose X_A is the exposure of interest that is correlated with Y, and X_B is a confounding variable that is not correlated with Y. Following the causal diagram in Figure 3, we generate X_A and X_B from a latent variable Z, where

$$X_A = \lambda_A Z + \eta,$$
$$X_B = \lambda_B Z + \gamma,$$

and η and γ are independent error terms with variances σ^2_η and σ^2_γ, respectively. Without loss of generality, we assume that variables X_A, X_B, and Z have been standardized to unit variance, and

the additional regression parameters are chosen so that the Y also has a unit variance. This then implies the causal effects $\rho(Y, X_A) = \beta_A + \beta_B\lambda_A\lambda_B$ and $\rho(Y, X_B) = \beta_B + \beta_A\lambda_A\lambda_B$.

To demonstrate the concept of omitted variable bias, we considere scenarios when X_A and Y are correlated, i.e., $\rho(Y, X_A) \neq 0$ while X_B and Y are uncorrelated, i.e., $\rho(Y, X_B) = 0$. Consider the following reduced model where the confounding variable, X_B, is excluded from the full model (1),

$$M_R : Y = \beta_0 + \beta_A X_A + \epsilon^*, \tag{2}$$

and $\epsilon^* \equiv \beta_B X_B + \epsilon$. The least-squares estimate for β_A under the reduced model (2) is

$$\begin{aligned}
\widehat{\beta}_A &= \frac{\text{Cov}(Y, X_A)}{\text{Var}(X_A)} \\
&= \text{Cov}(\beta_0 + \beta_A X_A + \epsilon^*, X_A) \\
&= \beta_A + \text{Cov}(\epsilon^*, X_A) \\
&= \beta_A + \text{Cov}(\beta_B X_B + \epsilon, X_A) \\
&= \beta_A + \beta_B \cdot \text{Cov}(X_B, X_A) \\
&= \beta_A + \beta_B \cdot \lambda_A \cdot \lambda_B.
\end{aligned}$$

The above derivation shows that $\widehat{\beta}_A$ is unbiased for β_A if X_B and Y have no linear relationship (i.e., $\beta_B = 0$), or X_A and X_B are independent (i.e., $\lambda_A = 0$ or $\lambda_B = 0$). However, those requirements contradict with the imposed assumption in the causal model of Figure 3 indicating that the omitted variable bias cannot be avoided.

In spite of the theoretical justification, we conducted simulation studies to illustrate our points. In order to generate simulated data under the imposed assumptions, we select regression parameters following the restrictions:

$$\beta_A + \beta_B\lambda_A\lambda_B \neq 0; \tag{3}$$

$$\beta_B + \beta_A\lambda_A\lambda_B = 0; \tag{4}$$

$$\lambda_A^2 + \sigma_\eta^2 = 1; \tag{5}$$

$$\lambda_B^2 + \sigma_\gamma^2 = 1; \tag{6}$$

$$\beta_A^2 + \beta_B^2 + \sigma_\epsilon^2 + 2\beta_A\beta_B\lambda_A\lambda_B = 1, \tag{7}$$

where restrictions (5), (6), and (7) are required to have $\text{Var}(X_A) = 1$, $\text{Var}(X_B) = 1$, and $\text{Var}(Y) = 1$, respectively. Plugging (5) and (6) into (3) yields $\sigma_\gamma^2 + \sigma_\eta^2 - \sigma_\gamma^2 \cdot \sigma_\eta^2 \neq 0$. We consider simulation settings based on the parameter specifications presented in Table 1, where variables Z, ϵ, η, and γ were generated from independent normal distributions with zero means. For all scenarios considered, the empirical Pearson's correlations between Y and X_B are close to zero. With the simulated data, we examined the bias of least-squares estimator for β_A under the full model of (1) and the reduced model (2). With 10,000 replications and three levels of sample sizes $n \in \{500, 1000, 2000\}$, the summary of bias is presented in Table 2. As expected, the bias of $\widehat{\beta}_A$ is virtually zero when controlling for X_B in the full model. On the contrary, failing to control for X_B in the model, one would mistakenly estimate the causal effect between X_A and Y resulting in a bias that agrees closely to $\beta_B\lambda_A\lambda_B$. Our simulation results confirms that excluding confounding variables from the model could bias the coefficient estimates, hence introducing omitted variable biases. In addition, our result dispute the premise that a covariate that is uncorrelated with the outcome cannot be biasing the results of an association test between another variable and the outcome as an indicator of a causal effect and disputes the premise we began with. Whereas in the psychometrics literature such patterns have usually been termed suppressor effects, in a nutrition epidemiology paper they were referred to as negative confounders [CITATION Cho08 ł2052].

Table 1: Paramters for Misperception 9

Scenario	β_A	β_B	λ_A	λ_B	σ^2_ϵ	σ^2_η	σ^2_γ
I	-0.4	0.3	$\sqrt{3}/2$	$\sqrt{3}/2$	0.93	0.25	0.25
II	0.4	-0.3	$\sqrt{3}/2$	$\sqrt{3}/2$	0.93	0.25	0.25
III	-0.24	0.5	0.8	0.6	0.8076	0.36	0.64
IV	0.24	-0.5	0.8	0.6	0.8076	0.36	0.64

Table 2: Summary of bias when fitting the full model (M_F) and the reduced model (M_R).

Scenario	$n = 500$		$n = 1000$		$n = 2000$	
	M_F	M_R	M_F	M_R	M_F	M_R
I	-0.0007	0.2248	0.0001	0.2251	-0.0001	0.2249
II	0.0005	-0.2249	0.0003	-0.2249	0.0002	-0.2248
III	-0.0001	0.2400	0.0003	0.2405	-0.0003	0.2399
IV	0.0004	-0.2396	-0.0002	-0.2400	-0.0005	-0.2402