Sins of Commission vs. Sins of Omission: How Confounding Can be Induced by Including 'Irrelevant' Covariates in Regression

Thomas Richardson
10 April 2002

CSSS Seminar
Overview of Talk

• Q: which variables should be regressed on if the causal structure is known, and I want to be able to interpret the coefficients causally?

• Two fallacious ‘Rules of thumb’ in common use.

• Examination of the fallacies.

• Valid solution based on graphical methods

• Quantification of the error in the rules of thumb

• Conclusion: distinction between endogenous and exogenous variables is too crude.
Hypothesized Model:

\[ Y_i = a_0 + a_X X_i + a_Z Z_i + \epsilon_{Y_i} \]

This is a *structural equation*, i.e. it describes the way in which the variable \( Y \) is causally determined by the variables \( X \) and \( Z \).

This ‘equation’ might be better written as:

\[ Y_i \leftarrow a_0 + a_X X_i + a_Z Z_i + \epsilon_{Y_i} \]

since it is better viewed as an assignment, rather than an equation.

*Throughout the talk* \( a_0, a_X \) and \( a_Z \) *will denote these structural coefficients.*
Suppose that we have a sample of data on $X$, $Y$ and $Z$ and we are interested in the coefficient $a_X$.

Suppose further that

$$E(Y \mid X, Z) = \beta_0 + \beta_X X + \beta_Z Z$$

$$E(Y \mid X) = \gamma_0 + \gamma_X X$$

The following questions arise:

1. Under what conditions will $\beta_X = a_X$, so that the coefficient of $X$ in the regression of $Y$ on $X$ and $Z$ will be an unbiased estimator of $a_X$?

2. Under what conditions will $\gamma_X = a_X$, so that the coefficient of $X$ in the regression of $Y$ on $X$ alone will be an unbiased estimator of $a_X$?
Fallacious Rule of Thumb (I)

*Mirer’s “Rule”:*
Whenever, $\gamma_X = a_X$ then $\beta_X = a_X$.

Equivalently, if the coefficient from the regression of $Y$ on $X$ alone is an unbiased estimator of $a_X$, then so is the coefficient from the regression of $Y$ on $X$ and $Z$.

Put briefly: adding a new variable into a regression equation never introduces bias in any of the other coefficients, viewed as estimators of structural coefficients.

‘*Spurious’ dependence may be removed but will not be introduced by adding variables to a regression.*

(It is usually implicit that $Z$ precedes $X$ and $Y$.)
Fallacy in Mirer’s “Rule”

We assume that

\[ a_X = \gamma_X = \frac{\text{cov}(X, Y)}{V(X)} \]

while

\[ \beta_X = \frac{\text{cov}(X, Y \mid Z)}{V(X \mid Z)} \]

where

\[ \text{cov}(X, Y \mid Z) = \text{cov}(X, Y) - \frac{\text{cov}(X, Z) \text{cov}(Y, Z)}{V(Z)} \]
Consider the following structure:

\[ l_1 \quad l_2 \]
\[ \quad \quad z \quad \]
\[ \quad \quad \quad \quad y \]
\[ x \quad \]

Where \( L_1 \) and \( L_2 \) can be thought of as independent unobserved variables.

Here \( \text{cov}(X, Y) = 0 \) hence \( \alpha_X = \gamma_X = 0 \).

However, \( \text{cov}(X, Z) \neq 0 \neq \text{cov}(Y, Z) \), hence \( \beta_X \neq 0 \), so \( \beta_X \neq \gamma_X \).
Example from WWII Pilot Selection

\[ \text{test taking ability} (l_1) \quad \text{mechanical aptitude} (l_2) \]

\[ \text{vocabulary score} (x) \quad \text{mechanical comprehension score} (z) \quad \text{flight simulator score} (y) \]

(apologies to Bob Abbott)
Possible source of the error

Compare Mirer equations (13.15) and (13.17)

\[ Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i \]
\[ Y_i = \gamma_0 + \gamma_1 X_{1i} + \nu_i \]

Mixer states (p.293)

... when (13.17) is the correct model, (13.15) could be considered “correct” if the restriction that \( \beta_2 = 0 \) is added and somehow taken into account during estimation.

but for (13.15) to be correctly specified (w.r.t. both \( X_1 \) and \( X_2 \)) we also require,

\[ E(u \mid X_1, X_2) = 0 \]

This is certainly not implied by the condition

\[ E(v \mid X_1) = 0 \]

which is all that we require in order for (13.17) to be correctly specified.
Valid conditions

Answer to Q.1: if, in the population,
\[ E(\varepsilon_Y \mid X, Z) = \delta_0 + \delta_Z Z, \]
i.e. it is not a function of \( X \) then
\[
E(Y \mid X, Z) = E(a_0 + a_X X + a_Z Z + \varepsilon_Y \mid X, Z) \\
= E(a_0 + a_X X + a_Z Z \mid X, Z) + \delta_0 + \delta_Z Z \\
= (a_0 + \delta_0) + a_X X + (a_Z + \delta_Z) Z
\]
so in this case \( \beta_X = a_X \).

Answer to Q.2: if in the population
\[ E(a_Z Z + \varepsilon_Y \mid X) = \phi_0 \]
i.e. it is not a function of \( X \), then
\[
E(Y \mid X) = E(a_0 + a_X X + a_Z Z + \varepsilon_Y \mid X) \\
= E(a_0 + a_X X \mid X) + \phi_0 \\
= (a_0 + \phi_0) + a_X X
\]
so in this case \( \gamma_X = a_X \)
Simple case (I)

Consider the following graph, describing causal relations:

We have

\[ E(\epsilon_Y \mid X, Z) = \text{constant}_1, \]

since there are no unmeasured variables confounding \( Y \) and \( Z \), or \( Y \) and \( X \).

Note that

\[ E(\alpha_Z Z + \epsilon_Y \mid X) = \text{constant}_2 + \alpha_Z E(Z \mid X) \]

which in general will depend on \( X \). Hence \( Z \) must be included.
Simple case (2)

\[ E(\epsilon_Y \mid X, Z) = \text{constant}_1, \]

since there are no unmeasured variables confounding \( Y \) and \( Z \), or \( Y \) and \( X \).

Note that

\[
E(a_Z Z + \epsilon_Y \mid X) = \text{constant}_2 + a_Z E(Z \mid X)
\]
\[
= \text{constant}_2 + a_Z E(Z)
\]
\[
= \text{constant}_3
\]

which does not depend on \( X \).

Hence regressing \( Y \) on \( X \) alone also yields an unbiased estimate of \( a_X \).
Case where Mirer’s rule fails

In this case, $E(\epsilon_Y \mid X, Z)$ will be a function of $X$ (and $Z$), hence the coefficient of $X$ in the regression of $Y$ and on $X$ and $Z$ will not yield an unbiased estimator of $a_X$.

However, $E(\epsilon_Y \mid X) = \text{constant}_1$. Hence the coefficient of $X$ in the regression of $Y$ on $X$ alone will yield an unbiased estimator of $a_X$. 

13
Question: In what general circumstances will a given set of covariates in a regression be such that the coefficient of \( X \) is an unbiased estimate of \( a_X \)?

In particular, can we describe these assumptions in causal terms?

i.e. given a hypothesized causal structure, can we identify a regression which will allow us to estimate a given causal coefficient?
Fallacious Rule of Thumb (II)

‘Confounding path’ rule

According to this rule, a set of variables is sufficient to control confounding so long as it contains at least one variable on any path of the form:

- $X \leftarrow \cdots \leftarrow \cdots \rightarrow Y$, or

- $X \rightarrow \cdots \rightarrow Y$.

(We assume that $X$ occurs prior to $Y$ so $Y \rightarrow \cdots \rightarrow X$ does not occur in the graph.)
Example showing the rule is fallacious

According to the ‘rule’ conditioning on $Z$ is sufficient to control confounding between $X$ and $Y$. However, as we have seen in the case of Mirer’s rule, this is not correct.
Source of the fallacy

The fallacy arises as follows:

*If there were no ‘confounding paths’ then regressing $Y$ on $X$ alone would give an unbiased estimate of $a_X$.*

*Conditioning on a variable on a confounding variable does block that path.*

However, as we saw with Mirer such conditioning, can induce additional confounding paths, that do not take the form described in the rule.

Consequently the rule describes a *necessary*, but *not sufficient* condition for a set to control confounding.
Consequences for inferences about ‘mediators’
Correct rule (1)

A non-endpoint vertex $W$ on a path is a *collider on the path* if the edges preceding and succeeding $W$ on the path have an arrowhead at $W$, i.e. $\rightarrow W \leftarrow$,

A non-endpoint vertex $W$ on a path which is not a collider is a *non-collider on the path*, i.e. $\leftarrow W \rightarrow$, $\leftarrow W \leftarrow$, $\rightarrow W \rightarrow$. 
Correct rule (2)

A path between vertices $X$ and $Y$ in a graph $G$ is said to be \textit{d-connecting given a set $Z$} (possibly empty) if

(i) every non-collider $V$ on the path is not in $Z$, and

(ii) every collider $V$ on the path is either in $Z$ or there is a directed path $V \to \cdots \to Z_i$ where $Z_i$ is in $Z$.

If there is no path d-connecting $X$ and $Y$ given $Z$, then $X$ and $Y$ are said to be \textit{d-separated given $Z$}.
Correct rule (3)

The coefficient of $X$ when regressing $Y$ on $X$ together with the variables in a set $Z$ ($X, Y \notin Z$) will yield an unbiased estimate of the causal effect $\alpha_X$ if:

(a) $X$ and $Y$ are d-separated by $Z$ in the graph $G'$ formed by removing the $X \rightarrow Y$ edge from $G$ (if such an edge is present in $G$).

(b) There is no vertex $Z$ in $Z$ such that $Y \rightarrow \cdots \rightarrow Z$. 
Sins of Omission

In each of these graphs failure to include $Z$ in the regression, will make the coefficient of $X$, a biased estimator of $a_X$. Including $Z$ will remove the bias (or in the last case including either of $Z_1$ or $Z_2$).
Sins of Commission

In each of these graphs including $Z$ in the regression, will make the coefficient of $X$, a biased estimator of $a_X$. Regressing $Y$ on $X$ alone will yield an unbiased estimator.
Time for a new idea

\[ l_1 \rightarrow z_1 \rightarrow z_2 \rightarrow y \]

\[ x \rightarrow y \]

There is no regression model which will estimate \( a_X \) in these two cases. However, \( a_X \) may still be estimated by other methods.
Abandon Hope!

In each of these graphs the causal effect $a_X$ cannot be identified without either making additional assumptions, or measuring additional variables.