Confounding by Indication, Confounding Variables, Covariates, and Independent Variables: Knowing What These Terms Mean and When to Use Which Term

Chittaranjan Andrade[1] (iD)

ABSTRACT

The terms *independent variables, covariates, confounding variables,* and *confounding by indication* are often imprecisely used in the context of regression. Independent variables are the full set of variables whose influence on the outcome is studied. Covariates are the independent variables that are included not because they are of interest but because their influence on the outcome can be adjusted for, leaving a more precise understanding of how the single remaining independent variable influences the outcome. Confounding variables are variables that are associated with both independent variables and outcomes; so, the relationship identified between independent variables and outcomes may be due to the confounding variable rather than to the independent variable. Potential confounders should be identified, measured, and adjusted for in regression, just as other covariates are. Confounding by indication occurs when the presence of the independent variable is driven by the confounding variable. Confounding by indication is a special kind of confounding; a confounding variable is a special kind of covariate; and a covariate is a special kind of independent variable in regression analysis. These terms and concepts are explained with the help of examples.

Keywords: Independent variables, predictor variable, covariates, confounding variables, confounding by indication, regression

Common terms in research are sometimes imprecisely used because shades of difference in meaning are not recognized. Examples of such terms, in the context of regression, are *independent variables, covariates, confounding variables,* and *confounding by indication*. These four terms are explained in the present article.

Independent Variables

In the simplest form of regression analysis, we examine how a single independent variable predicts a single dependent variable (independent and dependent variables are explained in a previous article[1]). For example, in regression, we can test the hypothesis that (female) sex predicts (greater) vocabulary in preschool children. Sex is the independent or predictor variable, and vocabulary is the dependent or outcome variable.

Independent Variables and Covariates

We know that older children and more intelligent children will have larger vocabularies. So, in the regression analysis referred to above, because the boys and girls in our sample may vary in age and intelligence, we should "adjust" or "control" for age and intelligence when studying the predictive or explanatory

[1]Dept. of Clinical Psychopharmacology and Neurotoxicology, National Institute of Mental Health and Neurosciences, Bangalore, Karnataka, India.

effect of sex on vocabulary. The regression now expands. Sex, age, and intelligence are the independent or predictor (or explanatory) variables; vocabulary remains the dependent or outcome variable. Because there are many independent variables, the regression is called multivariable regression.

As a digression, whether we call sex *predictive* or *explanatory* depends on whether we're using sex to predict vocabulary scores or to merely explain the extent to which it influences vocabulary scores. In this article, prediction and explanation are used interchangeably. The distinction is semantic, not mathematical.

In this regression, sex is the independent variable that is specifically of interest, and age and intelligence are the independent variables that are included as *covariates*. The regression examines how sex predicts vocabulary after adjusting or controlling for the covariates (age and intelligence). Expressed differently, the regression "removes" the explanatory effects of the covariates so that we can more precisely see to what extent sex predicts vocabulary scores.

It should now be apparent that covariates are a special category of independent variable; we're not interested in their effects on the dependent variable; we're only interested in adjusting for their influence so that we can more precisely determine how the independent variable, in which we are interested, predicts the dependent variable.

As a side note, what if we want to look at how all three variables—age, sex, and intelligence—separately predict or explain vocabulary? In that case, we call all of them independent variables. The multivariable regression is run in exactly the same way. However, when we look at the regression coefficient for age, we regard sex and intelligence as the covariates; when we look at the regression coefficient for sex, age and intelligence are the covariates; and when we look at the regression coefficient for intelligence, age and sex are the covariates.

Confounding Variables

Imagine that our study is being conducted in a traditional society in which, unfairly, boys are allowed to go out and play but girls are expected to stay at home and study.

Children who read and study more can be expected to have a larger vocabulary. So, in this situation, time spent reading may explain the relationship between being a girl and having a larger vocabulary. Thus, reading exposure is a *confounding variable*.

A confounding variable is one that is associated with the predictor variable as well as with the outcome variable; through this association, the confounding variable creates a potentially spurious association between the predictor and the outcome.[2] In our hypothetical study, reading is associated with girls as well as with a larger vocabulary. Thus, greater reading exposure in girls, resulting from social factors, can lead us to mistakenly conclude that girls have an *innately* larger vocabulary.

Our vocabulary study already has age and intelligence as covariates. We now realize that we must also include reading exposure as a covariate. So, just as a covariate is a special kind of independent variable, a confounding variable is a special kind of covariate.

Readers may note that, in our study, age and intelligence are not confounding variables. Whereas age and intelligence are both associated with vocabulary, neither is associated with subjects being girls as opposed to being boys. In contrast, whereas reading exposure is associated with vocabulary, in our hypothetical society, it is also associated with girls.

Confounding by Indication

Imagine that we're studying whether gestational exposure to antidepressant drugs (predictor variable) is associated with an increased risk of autism spectrum disorder (ASD) in offspring (outcome variable). The covariates that we include are age, socioeconomic status, family history of ASD, medical comorbidities, and others, because we believe that these variables may influence the ASD risk. Importantly, we include depression as a covariate.

Here, depression is a special covariate; it influences maternal behaviors and the maternal internal environment in many ways, and these effects, separately or together, may predispose to ASD risk in offspring. But depression is also the reason why an antidepressant drug is prescribed. So, depression fulfills the criteria

for a confounding variable: it is associated with both the predictor variable (antidepressant exposure) and, potentially, the outcome (ASD). More specially, depression is the reason or the indication for which the pregnant woman received the antidepressant drug. So, confounding occurred for a special reason: by indication. The confounding variable was the indication for the exposure to the predictor variable.

Confounding by indication was not the case in the example in which reading was a confounding variable; reading was not the reason or the indication why some subjects in the sample were girls. Expressed otherwise, confounding by indication is a special kind of confounding where the confounding variable (e.g., depression) is responsible for the presence or value of the independent variable (e.g., being exposed to an antidepressant).

Summary

Confounding by indication is a special kind of confounding; a confounding variable is a special kind of covariate; and a covariate is a special kind of independent variable in regression analysis.

The Importance of Recognizing Confounding

The importance of covariates is obvious; their influence needs to be adjusted for so that the effect of the predictor variable on the outcome can be better understood. But if a confounding variable is merely a special kind of covariate, why is so much of fuss being made over it?

One answer is that covariates are independent variables that are generally well known to be associated with the outcome. So, identifying and including these in the study design and analysis is straightforward. In contrast, confounding variables are important, specifically in the context of the predictor variable. If the possibility of confounding does not occur to us, and if we do not measure the confound (e.g., reading exposure or depression) and include it as a covariate in the regression, we may draw wrong conclusions about the relationship between the predictor variable (e.g., sex or antidepressant exposure) and the outcome (e.g., vocabulary or ASD in offspring).

In short, we need to understand the concept of confounding so that, at the time of study design itself, we can ask ourselves, "What are the biases associated with our predictor variable that can influence the outcome that we are studying?" It is important to address this question before the study starts because, once data collection is complete, it is too late to go back and measure the confounding variables.

Parting Notes

Sometimes, the confounding variable may be too tightly associated with the independent variable for adjustment for confounding to be possible. For example, in our hypothetical ASD study, we can adjust for depression by including it as a covariate because many women with depression may have preferred to suffer untreated depression during pregnancy rather than expose the pregnancy to an antidepressant drug. That is, depression and antidepressant exposure are not tightly bound. However, in such a situation, confounding by indication shifts to confounding by *severity* of indication because women with severe depression would not have been able to avoid antidepressant use. That is, severe depression and antidepressant exposure are tightly bound, so we would not be able to conclude whether an increased risk (if any) of ASD in exposed offspring is a result of the antidepressant exposure or a result of exposure to the behavioral and physiological manifestations of severe depression.

Similar considerations apply to gestational exposure to antipsychotic drugs and pregnancy outcomes because women with schizophrenia are usually discouraged from discontinuing antipsychotic treatment during pregnancy. So, it is hard to know whether an adverse gestational outcome has resulted from schizophrenia-related behaviors or from the use of the antipsychotic drug during pregnancy.

Finally, consider a study in which we examine the effects of benzodiazepines on cognition. We recognize that sleepiness is associated with poorer cognitive test performance as also with benzodiazepine use. So, is sleepiness a confounding variable that needs to be controlled? No. Sleepiness lies in the path between benzodiazepine use and poorer cognitive test performance; it is a mechanism through which benzodiazepines impair cognition. When a variable lies in the explanatory path between the predictor variable and the outcome variable, it is not a confounding variable, and we do not adjust for it.[2]

As a final digression for the perceptive reader: doesn't reading also lie in the path between being a girl and having a larger vocabulary? The answer is yes, but boys also read; in contrast, non-use of benzodiazepines is much less likely to be associated with sleepiness than use of benzodiazepines. When we adjust for reading in the vocabulary study, the results more accurately tell us whether being a girl predicts a higher vocabulary; the mechanism driven by social influences is removed. *So, when we interpret the results of a regression analysis, we need to interpret it in the context of influences that we have adjusted for and influences that we have left untouched.*

Declaration of Conflicting Interest

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

ORCID iD

Chittaranjan Andrade [iD] https://orcid.org/0000-0003-1526-567X

References

1. Andrade C. A student's guide to the classification and operationalization of variables in the conceptualization and design of a clinical study: Part 1. *Indian J Psychol Med* 2021;43(2):177–179.
2. Rhodes AE, Lin E and Streiner DL. Confronting the confounders: The meaning, detection, and treatment of confounders in research. *Can J Psychiatry* 1999;44(2):175–179.
3. McNamee R. Confounding and confounders. *Occup Environ Med* 2003;60(3):227–234
4. Andrade C. Confounding. *Indian J Psychiatry* 2007;49(2):129–131.