Regression Analysis and Causal Inference: Cause for Concern?

By Norman A. Constantine

Norman A. Constantine is senior scientist and director, Center for Research on Adolescent Health and Development, Public Health Institute, Oakland, CA, and clinical professor of community health and human development, School of Public Health, University of California, Berkeley.

Regular readers of *Perspectives on Sexual and Reproductive Health* may have noticed a preference among published authors for regression analysis as their primary methodology. In fact, last year, a substantial majority of this journal's articles employed some form of regression analysis, predominantly logistic regression. Such methodological dominance—which is by no means unique to this journal—supports the need for a critical review of the common uses and misuses of these types of analyses, and a careful examination of the validity threats and issues associated with the claims, conclusions and recommendations that typically result.

When used appropriately, regression analysis can be a powerful tool: It allows one to statistically model the relationship between a dependent (outcome) variable and a set of independent (predictor) variables. Linear regression is used with continuous dependent variables, such as number of sex partners or infant birth weight, while logistic regression is used with dichotomous dependent variables, such as a history of pregnancy or STD infection. Both forms of regression allow for the assessment of whether an independent variable (such as age, attitudes, protective behaviors or services received) is associated with an outcome variable while controlling for (statistically removing) the outcome's overlapping associations with other variables. These types of analyses are generally applied to correlational data, such as survey, census or administrative data, in so-called observational studies.

Of course, regression analysis is not needed in every study employing quantitative data. For example, results of abortion surveillance over time and across locations can be usefully presented employing just percentage or rate distributions.[1] The potential power and added complexity of regression analysis are best reserved for either predicting outcomes or explaining relationships.[2] The prediction of outcomes on the basis of current characteristics is possible without regard to the causal relationships among variables. For instance, regression analyses have shown that Asian and college-educated parents are the least likely among all social and demographic subgroups to support human papillomavirus vaccination for their daughters;[3] these findings can be used to identify the need for educational campaigns for these two subgroups even without understanding why they are the least supportive of vaccination. However, to develop effective educational campaigns, it is also necessary to understand the factors that influence parents' support for having their daughters vaccinated. When the goal is to understand (i.e., explain) the causal influences on a population outcome—a prerequisite for the design and development of any sexual health intervention—regression analysis can be a powerful tool, but it has some fundamental limitations. Its appropriate use requires substantial care and skill, as well as sufficient inferential humility.

DO VIRGINITY PLEDGES CAUSE VIRGINITY?

Tabachnick and Fidell provide an excellent introduction to the mechanics and approaches of regression analysis, together with a frank discussion of its limitations. This includes a review and assessment of the important and often neglected statistical and logical assumptions required in its use, and a blunt reminder that correlation does not imply causation:

"Regression analyses reveal relationships among variables but do not imply that the relationships are causal. Demonstration of causality is a logical and experimental, rather than statistical, problem. An apparently strong relationship between variables could stem from many causes, including the influence of other currently unmeasured variables."[4(p. 122)]

For example, if adolescents who voluntarily sign a virginity pledge are found to be more likely than nonpledgers to remain virgins, one must ask whether this difference is due to the pledging itself or instead to a preexisting inclination to abstain from sex.[5] This fundamental limitation applies even when correlational data are subject to regression analysis.

The editors of the *International Journal of Epidemiology* have warned against such unjustified causal inferences in epidemiological research, noting that observational studies have revealed "apparently protective effects" of a number of substances (e.g., beta-carotene and vitamin C on cardiovascular health), but that randomized controlled trials have not borne out these results. What these studies have in common, the editors write, "is that the groups of people who were apparently receiving protection … in the observational studies were very different from the groups not using them, on a whole host of characteristics."[6(p. 5)] Furthermore, they say, "belief that these differences could be summed up in measures of a few 'potential confounders' and adequately adjusted for in statistical analyses fails to recognize the complexity of the reasons why people differ."[6(p. 5)]

Other researchers have provided similar critiques.[7–10] From the perspective of behavioral science, Rutter also lamented this problem:

"From an early point in their training, all behavioral scientists are taught that statistically significant correlations do not necessarily mean any kind of causative effect. Nevertheless, the literature is full of studies [containing direct or implied causal conclusions] that are exclusively based on correlational evidence."[11(p. 377)]

Rutter classified most researchers into two camps—"those who are careful to use language that avoids any direct claim for causation, and yet, in the discussion section of their papers … imply that the findings do indeed mean causation," and those who "take refuge in the claim that they are studying only associations and not causation."[11(p. 377)] He viewed this second approach as disingenuous, arguing that in most cases, the reported findings would be of little interest without at least implied causal conclusions.[11] So in this view, most researchers working with correlational data either euphemize their implied causal conclusions through use of noncausal language (first camp) or disavow causal claims while counting on readers to draw causal implications on their own (second camp). These variations of denial are used in essence to diminish a researcher's responsibility for sufficiently justifying intended causal conclusions and implications.

CAUSE FOR CONCERN

In sexual and reproductive health research, as in many other applied research areas, a typical article using regression analysis employs data from a national survey data set or from a state, community or institutional survey. After descriptive statistics and percentage distributions are presented, regression results are reported, often for a large number of potential predictor variables. Consistent with Rutter's observations, article titles, as well as their main bodies, generally avoid causal claims, employing instead correlational terms—"association," "predictor," "risk," "correlate" and the like. Yet, in the Discussion sections of these articles, regression results are often transmuted directly into causal claims (e.g., identified risk characteristics lead to such and such consequences) or causal implications (e.g., interventions should focus on changing the identified risk characteristics, implying that this should lead to outcome changes). And typically, one or more of these causal claims or implications appear in the abstract—the part of the article most likely to be read and remembered, and to be used in policy discussions. Furthermore, these direct or implied causal conclusions regularly appear in research summaries and syntheses, frequently without sufficient acknowledgment of the limited evidence of causality.

One high-profile example is the narrative review "Factors That Affect Teens' Sexual Behaviors," in which more than 500 individual and environmental characteristics were claimed to affect one or more adolescent sexual risk behaviors and outcomes.[12] The review included a table of the 71 risk and protective "factors" deemed most important, on the basis of findings from a large collection of primarily observational studies employing some form of regression analysis. A fundamental limitation in this review was the failure to sufficiently distinguish between characteristics that were merely associated with and occurred before the sexual behaviors (risk or protective markers) and those for which evidence of causality had been found (risk or protective factors)—for example, when an experimental manipulation of the characteristic had been found to contribute to a change in the outcome. For most characteristics listed, the evidence of causality was weak or completely absent, yet causal claims were made throughout the review.

PLAUSIBLE ALTERNATIVE EXPLANATIONS AND CAUSAL ARGUMENTS

An important first step toward improving the application and interpretation of regression analyses is to better understand and differentiate between causal and correlational language.[13,14] This would help clarify authors' intended interpretations, making explicit those claims and implications that require a higher standard of causal evidence. An article title that modestly promises associations, predictors or risk characteristics should not yield a focus on causal claims and implications. And a title that boldly proclaims "effects of" or "influences on" or "factors that affect" should signal the presentation of compelling evidence in support of specific causal claims.

But what kind of evidence is compelling? One research method that can yield strong evidence to support a claim of causality is the randomized controlled trial, which involves random assignment of units (e.g., persons, schools, clinics or communities) to intervention or nonintervention (control) conditions. The putative power of randomization is its potential "to control an infinite number of rival hypotheses without specifying what any of them are."[15(p. viii)] But as with any method, the devil is in the details, and such trials, even when feasible, do not always yield compelling evidence.[5,16,17] As Shadish and colleagues noted, "validity is a property of inferences [and not] of designs or methods, as the same design may contribute to more or less valid inferences under different circumstances."[18(p. 35)] Building an argument for the validity of research inferences involves identifying and ruling out plausible alternative explanations, or rival hypotheses, for research findings. In fact, this activity represents "the core of the scientific method."[15(p. viii)]

The British epidemiologist Austin Bradford Hill presented nine now classic criteria for evaluating a causal relationship: strength of association, consistency, specificity, temporality, dose-response relationship, plausibility, coherence, reversibility and analogy.[19] According to Hill, it all comes down to ruling out plausible alternative explanations:

"None of my nine [criteria] can bring indisputable evidence for or against the cause-and-effect hypothesis and none can be required as a sine qua non. What they can do, with greater or less strength, is to help us to make up our minds on the fundamental question—is there any other way of explaining the set of facts before us, is there any other answer equally, or more, likely than cause and effect?"[19(p. 299)]

By the same token, regression analysis on its own does not bring indisputable evidence for or against a cause-and-effect hypothesis. Rather, regression analysis—and the same is true for any statistical methodology—is a tool for developing and integrating evidence into arguments to support conclusions. And like any tool, it can be used more or less skillfully. Including a large number of variables as potential predictors in a regression analysis to see which are statistically significant is a crude and generally inappropriate use of this tool, even if, as is commonly done, the variables are prescreened for bivariate associations with the outcome. But more sophisticated uses of regression analysis are possible.

APPROPRIATE USE

Observational studies employing regression analysis sometimes provide evidence relevant to the understanding of causality. This generally occurs not through the routine use of complex statistical methods, but instead through careful analysis and understanding of potential alternative explanations and threats to validity. For example, employing 2002 National Survey of Family Growth data, Kohler and colleagues found that U.S. adolescents who had received comprehensive sex education were significantly less likely to report teenage pregnancies than were adolescents who had received no or abstinence-only sex education.[20] These findings were derived from a strong study design and analysis that statistically controlled for plausible alternative explanations involving preexisting group differences (e.g., in family income). This approach was appropriate because the type of sex education received was unlikely to be due to motivated self-selection (in contrast, for example, to virginity pledging or dietary habits). Absent self-selection, preexisting group differences that might have influenced both the sex education received and sexual behavior outcomes could be more successfully removed through statistical adjustment. One strength of this study is that it provided evidence of convergent validity—its results were consistent with those from other studies of the same issue, employing different research designs and exhibiting different strengths and weaknesses.[21] Nevertheless, cautious interpretation and further study are still warranted.

Theory-based regression analysis strategies also can be used to help develop causal evidence from correlational data. One such strategy is referred to as the hierarchical, or sequential, approach.[4] Kerlinger and Pedhazur[22] popularized this approach in the behavioral sciences, and others further elaborated it;[2,23] Victora adapted it for use in epidemiology.[24] The hierarchical approach involves comparisons across a series of theory-informed regression models in which independent variables are sequentially added in small subsets. The ordering of these models is critical, as all variables are statistically adjusted for all other variables in the same model and previous models, but not for variables in subsequent models. Jerman and Constantine employed Victora's hierarchical conceptual framework to study the potential influences of parent demographic and psychological characteristics on parent-adolescent communication about sex.[25] One key hypothesis tested in the initial model was the interaction between parent and adolescent gender; because gender was considered a basic (distal) characteristic, statistical tests of gender variables and interactions were not adjusted for more proximate characteristics, such as parental education and comfort with communication, which were added in higher level models. This hypothesis was supported by the statistical significance of the gender interaction in the initial model.

Another theory-based strategy involves the elaboration method, developed by Lazarsfeld and colleagues in the 1940s for use with contingency tables,[26] and adapted to regression analysis by Aneshensel.[27] This method begins with the specification of a hypothesized causal relationship between a pair of variables (referred to as the focal relationship), and employs a series of regression models to rule out alternative explanations and "to bolster causal interpretation by demonstrating that the focal relationship fits within an encompassing system of relationships."[27(p. 58)] Aneshensel illustrated this approach with a series of theory-driven analyses from her Los Angeles Survey of Adolescent Experience study. One such analysis provided evidence in support of the hypothesized mediating role of parental demands on the focal relationship between family income and adolescent psychological mastery (i.e., the perceived ability to achieve desired outcomes and avoid undesirable ones).[27(p. 170)]

Other, more complex theory-based approaches include path analysis, structural equation modeling and those based on directed acyclic graphs.[28] The added complexity of these approaches requires access to a specialized statistician to design, implement and interpret the analyses. Yet they all suffer from the same fundamental limitations as the less esoteric regression approaches: "[A specified] model does not 'confirm' causal relationships. Rather it assumes causal links and then tests how strong they would be if the model were a correct representation of reality."[18(p. 398)]

All of these theory-based strategies are tools to support the development and comparison of explanatory regression models under the guidance of an explicit theoretical framework. They can provide useful direction and standardized approaches, but are neither necessary nor sufficient to the process of causal inquiry. Researchers can build and compare regression models to evaluate theory-driven hypotheses in many ways. Ultimately, it is the investigator's responsibility to make compelling arguments involving carefully developed evidence in support of the study's conclusions and proposed implications. When skillfully applied, theory-based regression analysis can help develop this evidence.

CONCLUDING THOUGHTS

Perspectives has long been known for its applied focus and emphasis on real-world problems and solutions, and has filled a critical niche in our field. Certainly, there is value in publishing good noncausal studies—for example, exploratory studies can yield hypotheses to be tested in future research, while descriptive and correlational studies of sexual health surveillance data can help identify the need for interventions among specific populations. But a real-world focus thrives on causal conclusions and implications; without addressing causality, this journal's contributions to the promotion of sexual and reproductive health (like the contributions of similar journals in other applied fields) would be unnecessarily limited. However, the demonstration of meaningful causality with real-world implications is rarely straightforward; in fact, it is usually difficult and often quite messy. Statistical methods such as regression analysis can help guide us through the swamp. But methods do not in themselves provide valid inferences—they are just tools to be skillfully used in the quest for such inferences.

As Abelson noted, "the purpose of statistics is to organize a useful argument from quantitative evidence, using a form of principled rhetoric."[29(p. xiii)] To make valid causal arguments and effectively address plausible alternative explanations requires deep understanding of the research topic, respect for the assumptions and limitations of the analytical tools employed, and perhaps most importantly, a strong theoretical foundation.

REFERENCES

1. Sedgh G et al., Legal abortion worldwide in 2008: levels and recent trends, *Perspectives on Sexual and Reproductive Health,* 2011, 43(3):188–198.

2. Pedhazur EJ, *Multiple Regression in Behavioral Research: Explanation and Prediction,* third ed., New York: Harcourt Brace, 1997.

3. Constantine NA and Jerman P, Acceptance of human papillomavirus vaccination among Californian parents of daughters: a representative statewide analysis, *Journal of Adolescent Health,* 2007, 40(2):108–114.

4. Tabachnick BG and Fidell LS, *Using Multivariate Statistics,* fifth ed., Boston: Allyn & Bacon, 2007.

5. Constantine NA and Braverman MT, Appraising evidence on program effectiveness, in: Braverman MT, Constantine NA and Slater JK, eds., *Foundations and Evaluation: Contexts and Practices for Effective Philanthropy,* San Francisco: Jossey-Bass, 2004, pp. 236–258.

6. Davey Smith G and Ebrahim S, Epidemiology—Is it time to call it a day? *International Journal of Epidemiology,* 2001, 30(1):1–11.

7. Freedman DA, From association to causation: some remarks on the history of statistics, *Statistical Science,* 1999, 14(3):243–258.

8. Freedman DA, Linear statistical models for causation: a critical review, in: Everitt B and Howell D, eds., *Encyclopedia of Statistics in Behavioral Science,* New York: Wiley, 2005, pp. 1061–1073.

9. Freedman DA, *Statistical Models and Causal Inference: A Dialogue with the Social Sciences,* Cambridge, UK: Cambridge University Press, 2010.

10. Sterne JAC and Davey Smith G, Sifting the evidence—What's wrong with significance tests? *BMJ (Clinical Research Ed.),* 2001, 322(7280):226–231.

11. Rutter M, Proceeding from observed correlation to causal inference: the use of natural experiments, *Perspectives on Psychological Science,* 2007, 2(4):377–395.

12. Kirby D, Factors that affect teens' sexual behaviors, in: Kirby D, ed., *Emerging Answers 2007: Research Findings on Programs to Reduce Teen Pregnancy and Sexually Transmitted Diseases,* Washington, DC: National Campaign to Prevent Teen and Unplanned Pregnancy, 2007, pp. 51–80.

13. Kraemer HC et al., Coming to terms with the terms of risk, *Archives of General Psychiatry,* 1997, 54(4):337–343.

14. Kraemer HC, Lowe KK and Kupfer DJ, *To Your Health: How to Understand What Research Tells Us About Risk,* Oxford, UK: Oxford University Press, 2005.

15. Campbell DT, Foreword, in: Yin RK, ed., *Case Study Research: Design and Methods,* fourth ed., Thousand Oaks, CA: Sage, 2009, pp. vii–viii.

16. Constantine NA, Intervention effectiveness research in adolescent health psychology: methodological issues and strategies, in: O'Donohue W, Benuto L and Woodward L, eds., *Handbook of Adolescent Health Psychology,* New York: Springer-Verlag, 2012 (in press).

17. Scriven M, A summative evaluation of RCT methodology: and an alternative approach to causal research, *Journal of Multidisciplinary Evaluation,* 2008, 5(9):11–24.

18. Shadish WR, Cook TD and Campbell DT, *Experimental and Quasi-Experimental Designs for Generalized Causal Inference,* Boston: Houghton Mifflin, 2002.

19. Hill AB, The environment and disease: association or causation? *Proceedings of the Royal Society of Medicine,* 1965, 58(5):295–300.

20. Kohler PK, Manhart LE and Lafferty WE, Abstinence-only and comprehensive sex education and the initiation of sexual activity and teen pregnancy, *Journal of Adolescent Health,* 2008, 42(4):344–351.

21. Constantine NA, Converging evidence leaves policy behind: sex education in the United States, *Journal of Adolescent Health,* 2008, 42(4):324–326.

22. Kerlinger FN and Pedhazur EJ, *Multiple Regression in Behavioral Research,* New York: Holt, Rinehart & Winston, 1973.

23. Cohen J et al., *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences,* third ed., Mahwah, NJ: Erlbaum, 2003.

24. Victora CG et al., The role of conceptual frameworks in epidemiological analysis: a hierarchical approach, *International Journal of Epidemiology,* 1997, 26(1):224–227.

25. Jerman P and Constantine NA, Demographic and psychological predictors of parent-adolescent communication about sex: a representative statewide analysis, *Journal of Youth and Adolescence,* 2010, 39(10):1164–1174.

26. Lazarsfeld PF, Pasanella AK and Rosenberg M, eds., *Continuities in the Language of Social Research,* New York: Free Press, 1972.

27. Aneshensel CS, *Theory-Based Data Analysis for the Social Sciences,* Thousand Oaks, CA: Sage, 2002.

28. Pearl J, *Causality: Models, Reasoning, and Inference,* Cambridge, UK: Cambridge University Press, 2000.

29. Abelson RP, *Statistics as Principled Argument,* Hillsdale, NJ: Erlbaum, 1995.

Acknowledgments

The author thanks Marc Braverman, Wendy Constantine and Petra Jerman for critical review and discussion. Preparation of this viewpoint was facilitated by grants from the Ford Foundation and The William and Flora Hewlett Foundation.

Author contact: *nconstantine@berkeley.edu*

doi: *10.1363/4413412*