The Collinearity Problem in Regression-Discontinuity Models

Frank E. Eaton
Senior Research Consultant
Abt Associates of Canada
and
Douglas A. Smith
Department of Economics
Carleton University
and
Associate, Abt Associates of Canada

ABSTRACT

The article addresses the issue of collinearity in regression-discontinuity models, as raised by a recent book review in this Journal. It examines the causes of collinearity between the constant term in the model and a dichotomous explanatory variable that indicates participation in a program being evaluated. It shows that removal of the constant does not solve the collinearity probelm but in fact may bias the estimated program impact. Collinearity can truly be overcome only by collecting more data (of an appropriate kind).

RÉSUMÉ

Cet article s'adresse au problème de collinéarité dans des modèles de régression discrets, comme présenté dans un récent compte rendu de livre dans cette Revue. On examine les causes de collinéarité entre le terme constant du modèle et une variable explicative booléenne qui indique la participation à un programme à évaluer. On montre que la suppression du term constant ne résout pas le problème de collinéarité, mais qu'elle peut en fait biaiser l'impact estimé du programme. On ne peut pas vraiment surmonter la collinéarité sauf en recueillant plus de données (d'une type approprié).

Introduction

A recent book review in this Journal (Mason, 1986, reviewing Trochim, 1984) discusses the problem of collinearity of a program variable and the intercept term in a regression-discontinuity model. The review asserts that including a constant term in this type of regression model may produce a biased estimate of the discontinuity effect being tested for. According to the review, this is the case if either participants or controls constitute a large fraction of the total number of observations. This note provides an alternative view of the best approach to this problem.

In the context of program evaluation, consider a design for evaluating a program with the objective of assessing the impact of the program on the post-program income and employability levels of participants. The data for the regression model consist of pre-program and post-program labour market variables for a group of program participants and for a comparison group.

The potential multicollinearity problem depends on the relationship between the intercept or constant term and the program variable. Typically

in a regression-disconinuity model, the program variable is dichotomous, taking a value of one for participants and zero for comparison group members. If most of the observations are on participants, then most of the values for the program variable are equal to one.

The intercept term is a measure of the size of the dependent variable if all independent variables are zero in value. In the matrix arithmetic required to produce least-squares estimates of the regression coefficients, the constant term enters as a column of ones. In other words, it has a constant value (of "1") for all observations. If the regression model includes such a term, collinearity may occur between the constant and the program variable. In the extreme case of a database with no comparison group, both the participant variable and the constant would consist of a column of ones. (No least-squares estimates would be possible, because the matrix would be singular.)

If collinearity is a problem, it will cause an increase in the standard error of the estimated program coefficient. The more the participant and comparison groups differ in size (number of members), the greater the collinearity and potential for problems arising in the analysis. In other words, the model becomes less efficient at concluding that an effect due to the program is statistically significant. Mason therefore considers the inclusion of an intercept term in such models to be an "elementary statistical error" at least when collinearity is a problem.

This note first considers the problem of collinearity and alternative ways of dealing with it. We show that some intuitively appealing solutions may confound the problem. We then discuss the issue of obtaining samples of appropriate sizes for the two groups. In many cases, the successful use of regression-discontinuity methods will depend critically on this issue.

The Problem of Collinearity

Collinearity problems arise in a regression model when the degree of correlation among two or more explanatory variables becomes sufficiently high to make an assessment of their separate effects on an outcome variable difficult. Collinearity is a problem usually arising from the data. There is no generally accepted analytical solution other than acquiring more or different data. If two highly related independent variables both affect the dependent variable, estimation of the effects attributable solely to each variable will be imprecise. In this note, however, we are concerned particularly with the impact of research design on collinearity.

It is important to understand the precise impact of collinearity and its implications for the interpretation of regression results. Collinearity leads to *imprecise* parameter estimates because it inflates the standard errors of the collinear parameters. Collinearity is a problem of precision in measurement not of bias in the size of parameter estimates. It is incorrect to assert as Mason does in his review (p. 95) that collinearity will lead to *biased* parameter estimates.

As we have already indicated, there are few remedies for most collinearity problems. Generally, acquiring more data will be ineffective because these data will also be collinear. In the case of a dichotomous variable, however,

more data will help if such data help to balance the sample. Intuitively, it seems clear that the best design is one in which the sizes of participant and comparison groups are equal. This design minimizes the correlation between the constant and the dichotomous variable (Conlisk, 1979).

Alternative Approches

For a number of reasons, it may not always be possible to have an equal number of participants and comparison group members. Cost factors and other design issues mean that the sizes of the two groups will be equal only rarely in practice. Are there solutions to the problem other than equalizing the sizes of the two groups?

Mason (p. 95) suggests dropping the intercept term from the regression to overcome the collinearity problem. This is an intuitively appealing solution. The program variable is collinear with the intercept and we are concerned primarily with the program variable. Therefore, why not simply exclude the intercept?

The answer to this question is that in a regression model, the intercept has a crucial role to play. It represents the value of the outcome variable when all the explanatory variables are set equal to zero. Dropping the intercept term means that the included dichotomous program variable then measures the intercept plus the program effect. This produces a biased estimate of the most important parameter in the model. This proposed solution replaces an imprecise but unbiased estimate with a biased estimate. This point is reinforced by considering in more detail the possible consequences of omitting the intercept term from the regression.

First, suppose removing the intercept has no effect on the coefficient of the program variable (i.e., the intercept coefficient is near zero). Then one may conclude that the collinearity between the two terms was not a serious problem given the overall covariance structure of the data used in the model. The model would leave us with the same conclusion as to whether the program had a significant effect on the outcome variable in question.

On the other hand, suppose removing the intercept results in a significant change in the coefficient of the program variable. One still must consider the problem of incorrectly attributing all of the significant effect to the program variable when the model fails to include another highly related variable. Excluding the constant term may artificially inflate or deflate the estimated coefficient for the program variable.

We further illustrate the above problem with a geometric example. In a typical regression model, each coefficient represents the slope of a line in multi-dimensional space. To address the model in two dimensions, we consider the following, simpler model:

$$Y = a + bX + cP$$

where Y is the outcome variable (such as wage)
X represents another explanatory variable (such as age, sex, education, or number of dependents)
P is the dichotomous program variable
a is the estimated coefficient of the constant term
b is the estimated coefficient of X
c is the estimated coefficient of P

In this model, "a" represents the value of Y when X and P equal zero. The value of "b" measures the amount that Y will change if X increases by one unit, and thus represents the slope of the fitted regression line relating Y to X. Because P may equal either zero or one, the model actually can be thought of as representing two lines that will be "c" units apart. (These lines are represented by the equations $Y = a + bX + c$ for participants and $Y = a + bX$ for members of the comparison group). Part A of the exhibit below (based on fictional data) shows these relationships graphically. It shows the effect of removing the constant term from the model. The result is a model which forces "a" to be zero: $Y = bX + cP$. Again, we may think of this as two fitted lines: $Y = bX + c$ for participants and $Y = bX$ for comparison group members.

We interpret the effect shown in Part A of the exhibit as follows. With the constant term in the model, we see two lines that represent the data very well. The estimated value of "c" is 1.81, which is statistically very significant in this case ($t = 6.37$). The lines with no constant term offer a visibly poor fit and produce an estimate of "c" that is biased strongly upward at 4.06 and has less significance ($t = 4.84$). While both t-statistics are highly significant, the removal of the constant term has actually reduced the significance of the observed influence of the program variable, P, on the outcome variable Y.

As stated earlier, if "a" is near zero anyway, removal of the constant term will have little effect. Part B of the exhibit shows the small effect in this case.

We conclude that, unless the estimated intercept is close to zero, removal of the intercept term could significantly alter the slope of the fitted line, which would clearly affect the estimated coefficients. This approach may thus introduce considerable bias to the model and to the measure of program effectiveness.

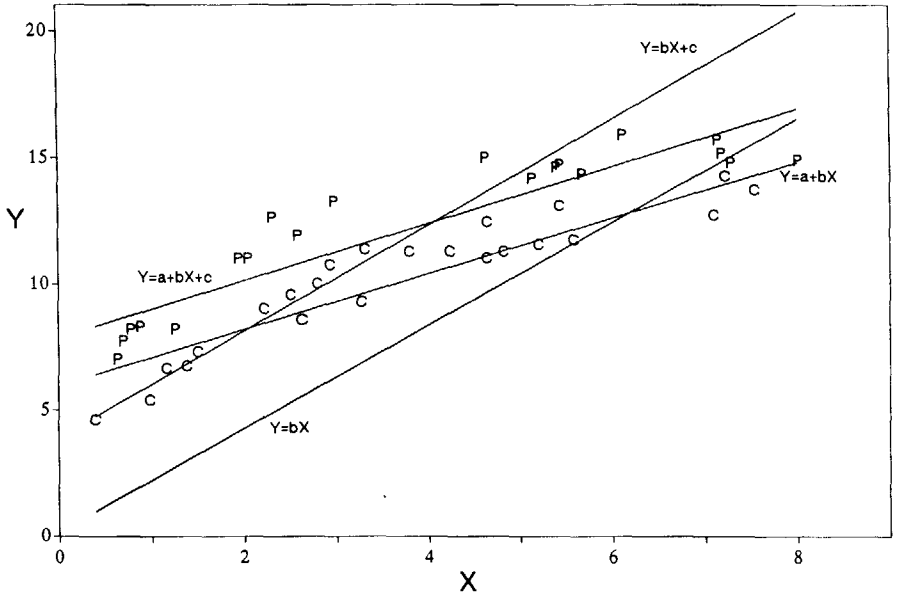
The above argument does not provide a solution to the problem of collinearity. We hope that it shows that the problem arises from the data, not from the analytical approach used. It cannot be overcome through analysis. It can be prevented by designing the evaluation study to avoid the problem of imbalance among sample sizes. The ideal arrangement from a statistical point of view is groups of equal size.

Problem of Achieving Sample Sizes

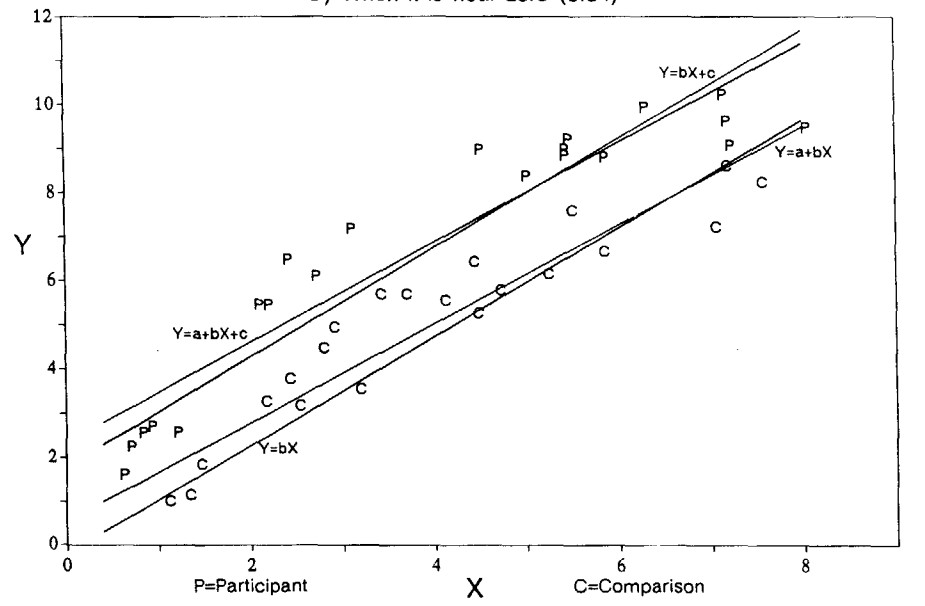
Governments seldom administer public programs in a way that permits true statistical experimentation. Such an experiment would mean withholding

Effect of Removing the Intercept Term

A) When it is far from zero (6.04)



B) When it is near zero (0.54)



services or benefits, on a random basis, from some eligible applicants while others received them.

In many cases, the best one can hope for, in the sense of the purity of the experiment, is the existence of a group of people somehow comparable to those participating in the program. But most programs keep data only on participants, or, perhaps, on rejected applicants as well. Consequently, adequate comparison group samples are hard to achieve. The members of such groups are expensive to contact and to interview because little information about them is available.

Evaluators are thus loathe to demand large comparison samples, especially when data from participants can provide so much information pertaining to evaluation issues other than effectiveness. Evaluators should be aware, however, of the effect that a small sample in the comparison group can have on the analysis of program effectiveness. This awareness could well affect their decisions either about sample sizes or about the rigour of the methodology to be used to measure effectiveness.

Conclusion

The problem of collinearity is best solved by ensuring adequate and balanced samples. Of course, this solution must be taken at the design stage of an analytical study. Once the data have been gathered, the only way to improve the balance among samples and to overcome collinearity is to discard data from the larger group. If this wasteful approach must be taken, the cases to be discarded should be selected at random to avoid introducing bias.

We further point out that our comment is not restricted to the use of a dichotomous variable only to indicate participation or non-participation in a program. A constant term is required in all models unless we have *a priori* knowledge that the true relation runs through the origin. Any dichotomous variable (e.g. sex or regional "dummy" variables) is potentially collinear with the constant term. Researchers should be aware of this infrequently considered source of collinearity in a potentially wide variety of situations.

References

Conlisk, J. (1979) "Choice of Sample Size in Evaluating Manpower Programs: Comments on Pitcher and Stafford" in F. Block (ed.) *Evaluating Manpower Training Programs*, JAI Press, pp. 79-96.

Mason, G. (1986) "Review of *Research Design for Program Evaluation: The Regression-Discontinuity Approach*, by W.M.K. Trochim," in *Canadian Journal of Program Evaluation*, Vol. 1, pp. 93-95.

Trochim, W.M.K. (1984) *Research Design for Program Evaluation: The Regression-Discontinuity Approach*, Sage Publications.