Small Is Beautiful: The Use and Interpretation of R2 in Social Research

Article *in* Szociologiai Szemle · January 1999

Ferenc Moksony
Corvinus University of Budapest

Ferenc Moksony: Small is beautiful. The use and interpretation of R^2 in social research. Szociológiai Szemle, 1999. Special issue. 130-138.

SMALL IS BEAUTIFUL. THE USE AND INTERPRETATION OF R^2 IN SOCIAL RESEARCH*

Ferenc Moksony

* I am indebted to Rita Hegedüs, György Lengyel and Ákos Róna-Tas for their helpful comments on an earlier draft of this paper.

ABSTRACT. Few statistical measures are as highly respected by social scientists as is the coefficient of determination. R^2 is an indispensable part of any serious research report and its sheer magnitude is often regarded as the most important indicator of the quality of a study. In this paper, I challenge this view and argue that in research aimed at the test of a theory, R^2, whether big or small, is, in general, completely irrelevant. I maintain, moreover, that the common interpretation of R^2 as a measure of "explanatory power" is misleading, as is the belief that a high value of R^2 testifies that the "true" or "best" or "complete" model has been found. I also discuss the implications for research practice of the effect that the spread of the independent variable exerts on the coefficient of determination.

> "These measures of goodness of fit have a fatal attraction. Although it is generally conceded among insiders that they do not mean a thing, high values are still a source of pride and satisfaction to their authors, however hard they may try to conceal these feelings." (Cramer, 1987: 253)

Few statistical measures are as highly respected by social scientists as is the coefficient of determination. R^2 is an indispensable part of any serious research report and its sheer magnitude is regarded by many as the most important indicator of the quality of a study.[1] The coefficient of determination, however, is not only among the most popular statistical measures; it is also among the ones that are most often used inappropriately. My aim in this paper is to discuss some of these misuses and to point out the limitations of R^2.

R^2 and the purpose of research

Researchers compute and report R^2 almost automatically, without pondering if doing so is at all necessary or sensible. Whether the coefficient of determination conveys any useful information is largely dependent on the purpose of the study. If the aim is **prediction**, then it is obviously important to know how accurately the dependent variable can be estimated from the explanatory variables. In this case, using R^2 may make some sense, since a high value generally indicates small prediction error. As we shall see shortly, however, the coefficient of determination actually is a rather poor measure of how close the estimated values come to the observed ones.

The situation is entirely different when the purpose of research is the **test of a theory.** In this case, we derive empirical implications from the theory and then check to see if they are true. Although these implications are fairly different in their substantive content, their formal structure is almost always the same - they describe the effect of one variable on another. What we need in order to assess our theory, then, is some measure of effect such as a regression coefficient; **the coefficient of**

[1] Statisticians sometimes distinguish r^2 and R^2, the "simple" and the multiple coefficient of determination. Given that what I am about to say pertains to both, I use the terms "R^2 " and "coefficient of determination" interchangeably.

determination is completely irrelevant.[2] A low value of R^2 indicates merely that the dependent variable is affected by a host of other factors in addition to the ones considered in the analysis; this, however, is quite immaterial given that our intention is to establish a particular causal relationship, not to prepare a full list of the various causes of a phenomenon.

R^2 and explanatory power

One often hears that the coefficient of determination measures the explanatory power of the variables included in a regression model. This view, although very popular, is rather misleading, as it **fails to distinguish between substantive and statistical explanations**. In a purely statistical sense, R^2 does in fact indicate the proportion of variance explained by the independent variables; this, however, has nothing to do with substantive explanation. If we used the dependent variable itself as the independent variable, we would be guaranteed to get an R^2 of 1.00, suggesting that a perfect explanation has been provided. Still, nobody would seriously say that we really explained anything; that we came any closer to a real understanding of the phenomenon we are interested in (see Lewis-Beck, 1993: 16; King, 1986: 677).

A classic example may serve to illustrate the basic difference between substantive and statistical explanations. The number of births in a given locality can be estimated reasonably well from the number of storks in the same area; if we ran a regression with the number of births as the dependent and the number of storks as the independent variable, we would probably get a fairly large R^2. But does this mean the number of storks explain, in a substantive sense, the level of fertility? Obviously not; the statistical explanatory power of this variable derives entirely from the fact that it is correlated with the real determinant of the number of births - namely, the degree of urbanization. Rural areas have more storks and they also have a higher birth rate.

Notice that the problem would not arise if we used the number of storks to **predict**, rather than to explain, fertility. If all we want to do is to forecast, with sufficient precision, the future course of a phenomenon, then essentially any variable would do, quite irrespective of whether its effect is real or spurious (Cook & Campbell, 1979: 296-297; Elster, 1990: 10). In fact, since variables with no real influence are often easier to capture empirically than are the true causal forces, from a purely practical standpoint, the former may well be more useful than the latter. If, however, the aim is explanation, then the issue of spuriousness is obviously critical and in this case we have to be very much aware that **a high value of R^2 does not at all necessarily imply real causal impact**.

The coefficient of determination is frequently employed to measure the **relative** explanatory power of independent variables. This application is typically found in the context of stepwise regression, a procedure no less questionable than the indiscriminate use of R^2. Stepwise regression generally ranks explanatory variables on the basis of their contribution to the increase of the coefficient of determination. This would present no problem if the variables were uncorrelated with each other, as in this case R^2 could be partitioned unambiguously and each variable would get the explanatory power that it rightly deserves. Independent variables are rarely uncorrelated, however, and, consequently, explanatory power cannot usually be divided unequivocally among them. In addition to the explanatory power that is unique to each variable, there is also a portion they share with one another and that, therefore, cannot legitimately be assigned to any of them. Which variable eventually gets this portion depends on which variable is first included in the regression and this, in turn, depends, in general, on which variable is more correlated with the dependent variable, irrespective of whether this correlation is causal or spurious. The result is that **true causal forces may look unimportant and may therefore be dropped from the model, while variables with no real impact may appear important and may be kept in the equation** (see Lewis-Beck, 1978; Pedhazur, 1982: 167-171; Kennedy, 1992: 63-64).

The previous example on storks and fertility may serve to demonstrate this important point. Imagine we have two correlated variables - the number of storks and the degree of urbanization - and wish to

[2] This is acknowledged even by those who otherwise defend the use of the coefficient of determination. Lewis-Beck and Skalaban (1991: 169), for instance, write, “when the researcher wants to know ‘the effect of X’, the R^2 has little utility. In that case, he or she should consult the relevant slope estimate”.

establish which of them is more important in explaining the level of fertility. Imagine further that, for whatever reason, the number of storks is slightly more correlated with the number of births than is the level of urbanization. What would happen if we ran stepwise regression in this case? Most likely, the number of storks would be included in the model, since at the first step, the criterion for inclusion is usually the size of the simple, zero-order correlation with the dependent variable. And the degree of urbanization would probably be excluded because its explanatory power has already been "soaked up", so to speak, by the other variable, the number of storks. We see, then, that by adhering blindly to the stepwise procedure, we end up dropping what should be kept and keeping what actually should be dropped.

R^2 and goodness of fit

Perhaps the most common use of the coefficient of determination is as a measure of goodness of fit. Applied in this way, the size of R^2 is generally taken to indicate how well the regression model fits the data; that is, how close the estimated values of the dependent variable come to the observed ones. Although in studies aimed at the test of a theory, this issue typically is of minor significance, in studies aimed at prediction, the precision of estimates is obviously of vital importance. In these latter cases, then, we do in fact need some measure of goodness of fit; the only question is if R^2 really is the most suitable for this purpose.

Despite the widely held belief, the coefficient of determination is of rather limited usefulness as an indicator of fit. This is because it is affected not only by how tightly the data points cluster around the regression line, but also by **the variance of the explanatory variable**. The same degree of fit will produce a higher R^2 if there is a greater spread in the explanatory variable. The effect of spread is clearly seen from equation 1, in which $\hat{Y}$ is the estimated value of the dependent variable; $\overline{Y}$ and $\overline{X}$ are the means of the dependent and independent variables, respectively; and b_1 is the unstandardized regression coefficient expressing the impact of the explanatory variable:

$$\sum(\hat{Y}-\overline{Y})^2 = b_1{}^2 * \sum(X-\overline{X})^2 \qquad (1)$$

The sum of squares due to regression - which is to the left of the equal sign and which is the numerator of R^2 - is, as can be seen, a function of the degree of variation in the explanatory variable, which is on the right hand side of the equation. Provided all other factors are constant, the larger the spread of the independent variable, the larger the regression sum of squares and the larger the coefficient of determination.

What does all this imply for research practice? In sociological studies, we typically do not have experimental control over the values of the explanatory variables; what we usually do, instead, is passively recording whatever values we happen to observe. Still, sampling sometimes allows us to manipulate the distribution of independent variables. This occurs, for instance, when we select cases so that they represent extreme values of the explanatory variable, or when we choose the same number of observations from the two categories of a dichotomous variable. All these "tricks" tend to increase the variation of the independent variable, thereby raising the magnitude of R^2.

This effect of sampling on the size of the coefficient of determination is aptly illustrated by Blalock (1964: 114-124) and Weisberg (1985: 74-76), who introduce artificial changes into the variance of the explanatory variable and see what happens as a result to several common statistical measures, including R^2. This sort of methodological experiment is useful also because it shows that while R^2 varies rather widely as the variance of the explanatory variable changes, another indicator of goodness of fit, the standard error of the regression estimate, remains largely constant.[3] The standard error, then, seems to be more robust and thus it could be preferred over the coefficient of determination as a

[3] This, however, presupposes homoscedasticity; if the size of the error differs greatly across the range of the independent variable, then truncating or artificially widening this range obviouy affects the standard error.

measure of goodness of fit.[4] Another advantage of the standard error is that it expresses the degree of fit in terms of the original metric of the dependent variable, whereas R^2 is a scale-free indicator and is, therefore, usually more difficult to interpret substantively (Achen, 1982: 61-64).

So far I talked about the dangers of too much variance, showing that if we are able to widen the range of the independent variable, then the coefficient of determination can be inflated almost at will and should therefore be interpreted very carefully. But caution is also needed when the variance of the explanatory variable is just too small and cannot be increased. This may occur when the variable refers to a **rare event** such as suicide. People who attempted to take their life usually comprise but a minor portion of the sample, resulting in a highly skewed distribution, with the great bulk of the cases falling in one of the two categories. And this, in turn, implies small variance, since the variance of a dichotomy is simply the product of the two relative frequencies. If we are interested in the effects of rare events, then, we should be aware that a low value of R^2 does not necessarily indicates that the impact is small and negligible. (For a more detailed discussion of this issue, along with empirical examples, see Glenn & Shelton, 1983.)

R^2 and the perfect model

People often think that the coefficient of determination indicates the correctness or completeness of the regression model. The higher the value of R^2, it is believed, the better the model, the more faithfully it represents reality. In fact, many researchers treat the coefficient of determination as a sort of quality assurance that automatically guarantees the worth of the work done. This view, however, is fundamentally wrong and efforts to increase R^2 in the hope of eventually finding the perfect model are totally senseless. To begin with, there is just no such thing as a perfect model; not because perfection is unattainable, but because models are, by their very definition, simplifications of reality (King, 1991: 1048). They purposely stress and magnify some aspects of the world, while paying little or no attention to others. Every model rests on some theory and reflects the particular emphases of this theory. And every model can be criticized only on the basis of another, competing theory; not on the basis of the size of the coefficient of determination. When we add new variables to the regression equation, the goal is never to boost R^2, to reach the perfect model; it is, instead, to rule out alternative explanations (Achen, 1982: 52). Whether a model is good or bad can only be decided by theoretical reasoning; the coefficient of determination has no say in this debate.

That R^2 really has nothing to do with the quality of our model can be illustrated by the following example. Suppose we are assessing the effect of an educational program designed to help unemployed individuals return to the labor market. Suppose, further, that participation is **voluntary**, with people themselves deciding whether or not to use the opportunity that is made available to them. After the program is over, we find that participants are, on average, more successful in getting a new job than are those who chose not to participate. We know, of course, that just because participation was voluntary, this result does not, in itself, unambiguously prove that the program was effective. It may well be that participants had stronger motivation from the very beginning and tried the program precisely because they were already more eager to find a job. If this were really the case, then participation would be a consequence rather than a cause and participants would have better chances anyway, without any supplementary education. It is also possible that they were younger or more educated and were, therefore, more likely to become employed, again without any program whatsoever. In order to establish the true effect of our program, all these characteristics - motivation, age, education - have to be included as control variables in the analysis. What does this imply for the coefficient of determination? By adding these variables to the regression equation, R^2 will, in all probability, increase greatly, making the model look very impressive.

[4] However, if prediction error has immediate practical relevance, then the standard error might still not be quite ideal. This is because it rests on the squared difference between the estimated and the observed value of the dependent variable and thus it overemphasizes large prediction errors. It is possible that the practical costs of making large and small errors are the same; in that case, absolute deviations might be preferable to squared ones. (For a detailed treatment of this issue, see Berk, 1986; for a more general discussion of the costs of prediction errors, see Goodman, 1966.)

Now suppose participation is not voluntary but, instead, we use **randomization** to decide who will get the program and who will not. In this case, the two groups are, on average, identical in all respects: they are roughly of the same age and education, and they are also quite similar to each other in prior motivation. What follows from this? In order to assess the true effect of the program, this time we do not need the control variables used before, since none of them is now correlated with participation.[5] However, excluding these variables also implies that the coefficient of determination will probably be much lower than it was previously, when there was no randomization and people themselves decided whether or not to try the program. But does this mean the second model is inferior to the first? Not at all; in fact, it is much better, since from the point of view of establishing causality, randomized investigations are the very best research tools possible.

Using the coefficient of determination as an indicator of the perfection of the model is seriously misguided for another reason as well. In their strong desire to increase R^2 beyond any limit, researchers often **fit their model to the random fluctuations** that are present in the data (Kennedy, 1992: 70), disregarding that that any set of observation is but a sample, one of the many data sets possible. Had we drawn another sample, the distribution of the data points would be slightly different and, consequently, the fit of our model would not be as good as it was before. What should we do in this case? Should we try another equation that produces a somewhat better fit to this new set of observation? But even this second model would probably fail in a third sample - and the process would continue infinitely. There is obviously not much point in having a high R^2 that only applies to a single data set. In fitting our models, we should strive for perfection only to the extent that the results reflect regular patterns in the data - patterns that are fairly stable across samples -, and not the random perturbations that are unique to each set of observation.

It is just this principle that is neglected by those who pursue what is commonly labeled as **curve-fitting**. They are not content with fitting a line but try, instead, a second degree polynomial, which they soon replace with a third degree one, and so on until they arrive at an n-1 degree curve that goes through each and every data point, produces a marvelous R^2 - but is quite useless as it only describes the particular n observation at hand and is thus of no help whatsoever in uncovering the more general pattern that is underlying the data (see Lieberson, 1985: 93).

Earlier I stated that in research aimed at prediction, a large value of the coefficient of determination is usually good news and most textbooks do indeed regard a high R^2 as a prerequisite of successful prediction (e.g., Lewis-Beck, 1993: 16). In view of what has just been said about the danger of fitting the model to the chance fluctuations in the data, some qualification now seems to be in order. If the high value of R^2 is merely the result of fitting the equation to the particularities of the sample that we happen to have, then the coefficient of determination, no matter how large, does not at all guarantee that the same close fit will be achieved outside that sample. In fact, as Mayer (1975) notes, “if we are interested in hypotheses that are valid beyond the sample period, goodness of fit statistics are a very poor guide” (p. 882).

How poor a guide these statistics might be can be illustrated by Lieberson’s (1985: 97-99) example. Suppose we toss a large number of fair coins, ten times each. If we count the number of “heads”, the result will be different for different coins. In some cases, we only get two or three “heads” - instead of five, the theoretically expected value -, but in other cases we get as many as eight or nine or even ten. Now suppose we try to explain these variations. If we are patient enough, we can identify some characteristics of the coins that are associated with the number of heads. One such characteristic might be the year or decade in which the coin was minted; another might be where it was produced; still another might be when, in the sequence of coin tossing, it was thrown. But no matter how hard we try, how many characteristics we consider, our effort is doomed to failure. It is because “the subset of ... coins that generated a large number of heads in the first round will be no more likely to generate an unusually large number of heads on a second sequence of ten tossings than will the set of ... coins that came up with only two or three heads in the first sequence. Thus there is an excellent fit of the coins being tossed, but a very poor result with the characteristics of the individual coins.” (Lieberson, 1985: 98).

[5] Of course, these variables might still be included in the model for other purposes, such as to reduce the residual variance.

Conclusions[6]

Having discussed what seem to me the most important misconceptions about R^2, and having pointed out the limitations of this measure, it is useful, in conclusion, to present a formula that provides a summary picture of the factors affecting the coefficient of determination. As a first step, recall equation 1, which expressed the numerator of R^2 as a function of the variation in the explanatory variable, as well as of the regression coefficient:

$$\sum(\hat{Y}-\overline{Y})^2 = b_1^2 * \sum(X-\overline{X})^2 . \qquad (1)$$

The total sum of squares - which is but the denominator of R^2 - consists, as is well known, of two parts, the regression sum of squares and the residual sum of squares:

$$\sum(Y-\overline{Y})^2 = \sum(\hat{Y}-\overline{Y})^2 + \sum(Y-\hat{Y})^2 . \qquad (2)$$

Let us substitute equation 1 into equation 2:

$$\sum(Y-\overline{Y})^2 = b_1^2 * \sum(X-\overline{X})^2 + \sum(Y-\hat{Y})^2 .$$

Using these expressions, the coefficient of determination can now be described as:

$$R^2 = \frac{b_1^2 * \sum(X-\overline{X})^2}{b_1^2 * \sum(X-\overline{X})^2 + \sum(Y-\hat{Y})^2} .$$

Or, in words:

$$R^2 = \frac{\text{magnitude of effect}^2 * \text{variation of X}}{\text{magnitude of effect}^2 * \text{variation of X} + \text{goodness of fit}}$$

As this formula demonstrates, the coefficient of determination is, in essence, a mixture of three factors: the impact of the explanatory variable, the degree of variation in this variable, and, finally, the size of the spread around the regression line. Precisely because it is affected by so many factors, R^2 is unable to reflect any of them accurately. Neither is it a good measure of the strength of the effect of the independent variable; nor is it an appropriate indicator of how well the model fits the data. For both these purposes, we have better measures available: the impact of the explanatory variable can be captured by the unstandardized regression coefficient or its analogues, whereas goodness of fit can be judged by the standard error of the regression estimate. In view of all this, the high respect social researchers commonly have for the coefficient of determination does not seem to be warranted; the popularity of R^2 derives, it appears, much more from its rhetorical value than from its actual accomplishments.

References

Achen, Ch. 1982. Interpreting and using regression. Beverly Hills - London: Sage Publications.

Berk, R.A. 1986. How applied sociology can save basic sociology. Unpublished manuscript.

[6] This part of the paper relies heavily on Achen (1982), especially page 63.

Blalock, H. 1964. Causal inferences in nonexperimental research. Durham, N.C.: University of North Carolina Press.

Cook, Th. & Campbell, D.T. 1979. Quasi-experimentation. Design and analysis issues for field settings. Boston etc.: Houghton Mifflin Co.

Cramer, J.S. 1987. Mean and variance of R^2 in small and moderate samples. Journal of Econometrics, 35: 253-266.

Darlington, R. 1990. Regression and linear models. New York etc.: McGraw - Hill Publishing Co.

Elster, J. 1990. Nuts and bolts for the social sciences. Cambridge etc.: Cambridge University Press

Glenn, N. D. & Shelton, B. A. 1983. Pre-adult background variables and divorce: a note of caution about overreliance on explained variance. Journal of Marriage and the Family, 45: 405-410.

Goodman, L. 1966. Generalizing the problem of prediction. 277-281 in: Lazarsfeld, P.F. & Rosenberg, M., eds.: The language of social research. 5th ed., Toronto.

Kennedy, P. 1992. A guide to econometrics. Oxford, UK. - Cambridge, USA: Blackwell Publishers

King, G. 1986. How not to lie with statistics: Avoiding common mistakes in quantitative political science. American Journal of Political Science, 30: 666-687.

King, G. 1991. “Truth” is stranger than prediction, more questionable than causal inference. American Journal of Political Science, 35: 1047-1053.

Lewis-Beck, M. 1993. Applied regression: an introduction. In: Lewis-Beck, M., ed.: Regression analysis. International Handbooks of Quantitative Applications in the Social Sciences. Vol. 2. London - Thousand Oaks, CA - New Delhi: Sage Publications

Lewis-Beck, M. & Skalaban, A. 1991. The R-squared: Some straight talk. Political Analysis, 2: 153-171.

Lewis-Beck, M. 1978. Stepwise regression: a caution. Political Methodology, 5: 213-240.

Lieberson, S. 1985. Making it count. The improvement of social research and theory. Berkeley - Los Angeles - London: University of California Press

Mayer, T. 1975. Selecting economic hypothesis by goodness of fit. Economic Journal, 85: 877-883.

Pedhazur, E. 1982. Multiple regression in behavioral research. 2nd ed. Forth Worth etc.: Harcourt Brace Jovanovich College Publishers

Weisberg, S. 1985. Applied linear regression. 2nd ed. New York etc.: John Wiley & Sons.