



ERIC A. HANUSHEK

JOHN E. JACKSON

Statistical Methods  
for Social Scientists

# Statistical Methods for Social Scientists

*ERIC A. HANUSHEK*

*Department of Economics  
Yale University  
New Haven, Connecticut*

*JOHN E. JACKSON*

*Department of Government  
Harvard University  
Cambridge, Massachusetts*



**ACADEMIC PRESS, INC.**

Harcourt Brace Jovanovich, Publishers  
Orlando San Diego New York  
Austin London Montreal Sydney  
Tokyo Toronto

COPYRIGHT © 1977, BY ACADEMIC PRESS, INC.  
ALL RIGHTS RESERVED.  
NO PART OF THIS PUBLICATION MAY BE REPRODUCED OR  
TRANSMITTED IN ANY FORM OR BY ANY MEANS, ELECTRONIC  
OR MECHANICAL, INCLUDING PHOTOCOPY, RECORDING, OR ANY  
INFORMATION STORAGE AND RETRIEVAL SYSTEM, WITHOUT  
PERMISSION IN WRITING FROM THE PUBLISHER.

ACADEMIC PRESS, INC.  
Orlando, Florida 32887

*United Kingdom Edition published by*  
ACADEMIC PRESS, INC. (LONDON) LTD.  
24/28 Oval Road, London NW1 7DX

**Library of Congress Cataloging in Publication Data**

Hanushek, Eric Alan,           Date  
Statistical methods for social scientists.

(Quantitative studies in social relations series)

Bibliography:           p.

1. Estimation theory.           2. Least squares.           3. Social  
sciences—Statistical methods.   I. Jackson, John Edgar,  
joint author.           II. Title.

QA276.8.H35           519.5'4           76-9158

ISBN 0-12-324350-5

PRINTED IN THE UNITED STATES OF AMERICA

sample; a lack of information is not solved by throwing out information. A given observation or set of observations may contain redundant information about the effects of two variables, i.e., the variables in those observations may exhibit the same intercorrelations as observed in other observations, but that is no reason to eliminate the observation.

#### 4.5 Model Specification and Multicollinearity in Practice

Multicollinearity can have a powerful effect upon model specification and, particularly, on statistical tests of model specification. When we discussed model specification in terms of an omitted variable, there were two possible situations: data on the omitted variable exist but were ignored, and data on the omitted variable do not exist. In the first case, the standard  $t$ -test of the null hypothesis  $\beta=0$  is a test of the specification that includes  $X$ , and performance of that statistical test provides information about appropriate model specification. But multicollinearity confounds this test and weakens the ability to judge among model specifications. Since multicollinearity reduces the precision of the estimates (increases their variance), it becomes difficult to develop tests that are good at distinguishing between alternative values of a parameter and alternative specifications of the model.

Likewise we must be very careful about the model specification. If minor changes in the model specification or the definition of variables yield large changes in the estimated coefficients, the model should be treated with some caution. In particular, the precise estimates of any given specification may represent an artifact more of the sample than of the true underlying structure. They may rely heavily upon one or two data points that exhibit a slightly different pattern of intercorrelations but which are not necessarily representative of the population. In other words, multicollinearity reduces our confidence in any particular point estimates of parameters. This lowered confidence is usually, but not always, reflected in the estimated coefficient variances. Moreover, since the estimates become very sensitive to sample and specification, the results that are obtained from experimentation with a variety of specifications and variable definitions are quite suspect by themselves. They require more than the usual amount of verification from other samples of data.

We have concentrated our discussion on the problems associated with omitting an important variable from the analysis. The reverse case is also true: there are costs associated with including irrelevant variables. The expected value of the estimated coefficient for such a variable will of course be zero, and the other coefficients will remain unbiased, so that faulty conclusions are not expected. However, the effects on the single estimate of each coefficient obtained from one data set may not be trivial because the

variance of all estimates will increase. The formula for the variance of each estimated coefficient does not take into account that one of the included variables has no influence on  $Y$ ; it includes only terms involving the variances and covariances of the explanatory variables. For example, in the model examined in Chapter 2, the variance of  $b_2$  is a function of the variance of  $X_2$  and of the correlation between  $X_2$  and  $X_3$ . In that case if  $X_3$  does not influence  $Y$  and should be omitted from the model, including it only increases the variance of the estimate of  $b_2$ . Thus the cost of including extraneous variables in the estimation is reflected in higher variances for the estimates of the coefficients for the variables that belong in the equation. In equations with more than two explanatory variables, irrelevant variables have the same effect because they generally increase the collinearity within the set of included variables, which reduces the size of the determinant and increases the variance of each estimated coefficient. The implication of this discussion is that you do not want to include unnecessary variables, particularly if they are collinear with other variables, just as you do not want to omit necessary ones. The question is how to tell the two apart. The only certain answer is with the theory used to construct the model in the first place. This, and possibly previous empirical findings, are the only sure way to make such decisions.

In some instances researchers will use the statistical tests described in Chapter 3 to decide whether a variable is extraneous or not. If the  $t$ -statistic falls below the critical value for a specified confidence level, say 0.1, the researcher will decide to accept the null hypothesis that the true coefficient for that variable is zero, and reestimate the equation with that variable omitted. This process runs the very considerable risk of biasing the remaining coefficients because the statistical tests used are not set up to test the null hypothesis implicit in this decision process. The null hypothesis the researcher is actually using is that the true coefficient is not zero,  $H_0: \beta \neq 0$ ; but the  $t$ -statistic is testing the null hypothesis that  $\beta$  equals zero. Not being able to reject the null hypothesis that  $\beta = 0$  is not equivalent to rejecting the null hypothesis that  $\beta \neq 0$ .

Nature, as the designer of social scientists' experiments, is particularly perverse on this problem. Low  $t$ -statistics can result either from the true coefficient being close to zero or because the estimated coefficient has a high variance, possibly caused by multicollinearity. If one could be sure that the true influence of a variable is small and that the low  $t$ -statistic is the result of the true coefficient being close to zero, the amount of bias from omitting this variable would be relatively small. The gain in precision by reducing the variance of the remaining coefficients could offset this small bias. However, large gains in precision are possible only when highly collinear variables are omitted. Unfortunately, this collinearity increases the variance of the estimated coefficients and implies that one cannot be confident that the true value of the coefficient is close to zero. The result of using  $t$ -tests to justify

omitting variables may lead to the exclusion of collinear but substantively important variables. Thus there is considerable risk of biased coefficients if one adopts this strategy. We return to our previous comment that in the face of multicollinearity the researcher must be more cautious in evaluating and interpreting the results and must provide much more information about the behavior being modeled. This information can come only from theoretical considerations and previous empirical work.

A common estimation procedure known as stepwise regression is particularly vulnerable to the problems of specification and multicollinearity just described. In stepwise regression, the researcher specifies only the dependent variable and a list of possible explanatory variables rather than the exact model to be estimated. The program doing the regression<sup>5</sup> then successively selects variables for inclusion in the equation on the basis of which one will yield the greatest increase in  $R^2$ . In some cases cutoffs are established in terms of the number of variables to be included or the minimum change in  $R^2$  required for inclusion of the next variable.

Stepwise regression represents a series of ordinary least squares estimates where the number of variables is progressively increased. At any stage, the coefficient estimates, estimated standard errors,  $R^2$ , etc. arrived at through a stepwise procedure will be identical to the estimates obtained from a simple OLS regression that includes the same variables. Thus, the issue is not the numerical coefficient estimates. Instead, it is whether the additional information generated in intermediate stages of the stepwise process is useful in interpreting (or constructing) the model itself or in ascertaining anything about relationships between individual independent variables and the dependent variable.

Two common justifications for the use of stepwise regression are that such a procedure is useful in determining the "most important" variables in explaining the behavior in question and that, because there is uncertainty about just which variables should be in the equation, this procedure allows the data "to tell the best model." Let us consider these in order.

$R^2$  was interpreted as the amount of dependent variable variation explained by the exogenous variables. Thus, it seems logical that the variables that "explain the most" are the "most important" in determining the behavior. However,  $R^2$  is a sample specific statistic. As such it is determined not only by the strength of the relationship (the  $\beta_k$ ) but also by the intercorrelations among the exogenous variables and by the variance in each of the exogenous variables. These last two terms are dependent upon the specific characteristics of the sample and, thus, cannot be easily generalized to the entire population under consideration. Further, since the procedure operates on increments to  $R^2$ , or changes in explained variation, the amount of

<sup>5</sup>Some stepwise programs start with a full set of variables identified in the equation and eliminate variables on the basis of the smallest reduction in  $R^2$ . This is referred to as stepwise deletion. Other programs combine these two methods.

variation attributable to any variable is dependent upon the order in which it is entered, i.e., on the set of other variables that are already in the equation (entered in an earlier step) and on the set not yet entered.

Consider a simple example where two exogenous variables each have a strong influence on the dependent variable (i.e., a large value of  $\beta_k$ ) but which are highly correlated with each other in the sample. A stepwise procedure would select one of the variables for inclusion but might neglect the second because it would add little to  $R^2$ . The individual parameter estimate for the included variable would be biased (as discussed previously), and the procedure would be misleading if we interpreted the stepwise regression as indicating that the included variable was important and the excluded variable was unimportant. The individual coefficients at any stage in the procedure are biased in just the way discussed under the heading of model specification.

Further, there is little assurance that the final model—the model selected at the end of the entire stepwise procedure—bears any relationship to the underlying population model. First, a variable entered at an early stage may have no influence on the dependent variable ( $\beta_k = 0$ ) but may be correlated in the sample with several other variables that do influence the dependent variable. The stepwise procedure may include this variable because it “proxies” several other variables—variables that do have a significant relationship with the dependent variable. Because of the level of intercorrelation, the true variables may never be included. Second, it is possible that some important variables are not included (and neither are any proxies for them). A set of variables might be skipped in the search process if their effects are “offsetting”; i.e., in the case of two variables, if both have similar effects on  $Y_i$  (in terms of  $\beta_k$ ) but are negatively correlated in the sample, or if each has an opposing effect on  $Y_i$  but they are positively correlated in the sample, the estimated importance of either one taken separately will be understated.

The point of this discussion is simple. Stepwise regression appears to promise something that it cannot deliver. It is not possible to use stepwise regression to give both the model and the parameter estimates. Nor is it possible to use either the order of entry into a stepwise procedure or the parameter estimates of intermediate stages to make inferences about the importance of particular variables (except in the context of one specific sample). To the extent that the purpose of estimation is to make inferences about population relationships on the basis of sample information, a stepwise procedure can be very misleading.

#### 4.6 Functional Forms

The previous discussion of model specification has centered exclusively on which variables should be included. Specifying the relationship among them is also very important. Choosing the correct *functional form* of the model is often done with even less guidance than choosing the variables of the model.