METHODOLOGICAL
PERSPECTIVES

A series edited by
RICHARD J. HILL
Purdue University

Consulting Editor
HOWARD S. BECKER
Northwestern University

CAUSAL MODELS IN THE SOCIAL SCIENCES

edited by H. M. BLALOCK, JR.

University of Washington

ALDINE • ATHERTON/*Chicago* • *New York*

ABOUT THE EDITOR

H. M. BLALOCK, JR., is Professor of Sociology in the Department of Sociology at the University of Washington, Seattle. He was educated at Dartmouth College, Brown University, and the University of North Carolina, where he received his Ph.D. in 1954. Professor Blalock has written or edited six previous books, dealing with general methodology, applied statistics, theory building, race relations, and social power; he has contributed more than 40 articles to academic and professional journals. He has served on a number of professional councils and as associate editor of several journals.

First published 1971 by
Aldine · Atherton, Inc.
529 South Wabash Avenue
Chicago, Illinois 60605

Library of Congress Catalog Card Number 70-133304
ISBN 202-30076-5, cloth; 202-30228-5, paper
Printed in the United States of America

To Paul F. Lazarsfeld

leadership role. Influentials in this study, compared to those less influential, adopt significant innovations which contribute, in turn, to increased production. Decisions leading to the adoption of innovations may require considerable information from many sources, including the mass media. While influentials may (or may not) pass on information to those who are less influential, there is considerable evidence which suggests that this information is used for the pursuit of purely economic goals. Thus, influentials appear to play a key role in the agricultural production process, a role distinct—but not independent—from their prestige and other attributes of social position. The fact that they are more productive, compared to those less influential, enhances their influence position and undoubtedly has implications for their capital position as well.

The importance of the influential can be noted when the model is used to predict the consequences of changing the level of a variable. For example, if an increase in agricultural production is a desired goal, how does one go about increasing production in an area comparable to one in which this study was conducted? That is, which variables must be manipulated? An examination of the important variables in equation (24) suggests that individuals who adopt innovations and who also control resources are likely to be high producers. Those who adopt innovations, according to equation (23), are influentials who utilize technological sources of information. And influentials (equation (21)) are those who are well educated, who expose themselves to relevant mass media channels, and who are prestigious in their community. A starting point for increasing production would appear to be knowledge of the influentials.

Conclusions

The purpose of this study was to describe the logic of, and to utilize, the Theil-Basmann two-stage estimation procedure for systems of simultaneous equations. Using behavioral data, the methodology appears to have promise for explaining or accounting for the effects of interdependent systems involving sociological and economic variables.

Specifically, variables relating to the diffusion of technical innovations were found to represent an interdependent system—a system in which the relationships among endogenous variables were examined and the effects of a set of exogenous variables were tested. The methodology is recommended for analysis of other kinds of interdependent systems in behavioral areas which heretofore have yielded less with less powerful techniques.

Chapter 13

PEER INFLUENCES ON ASPIRATIONS: A REINTERPRETATION

OTIS DUDLEY DUNCAN
University of Michigan
ARCHIBALD O. HALLER
University of Wisconsin
ALEJANDRO PORTES[1]
University of Wisconsin

The hypothesis that interaction with peers influences levels of educational and occupational aspirations of adolescents, enunciated by Haller and Butterworth in 1960,[2] has proved to be an intriguing one, to judge by the attention given it subsequently. Without mentioning the several studies of "school climates," wherein the hypothesis is more or less assumed to be correct in order to interpret ostensible "school effects," we may refer to studies specifically directed to tests of the hypothesis itself or one of several closely related hypotheses which have also been subjected to scrutiny.[3]

To summarize our present knowledge: (1) The supposition that homophily with respect to socioeconomic characteristics is generated by socioeconomic segregation of school populations was considered by Rhodes,

Reprinted by permission of the authors and publisher from the *American Journal of Sociology*, Vol. 74, pp. 119–137, and Vol. 75, pp. 1042–1046. Copyright 1968 and 1970, The University of Chicago Press.

[1] This is a report from Project 5-0074 (EO-191), "Socioeconomic Background and Occupational Achievement: Extensions of a Basic Model," carried out under Contract OE-5-85-072 with the U.S. Office of Education. The assistance in computation of Ruthe C. Sweet is gratefully acknowledged.

[2] A. O. Haller and C. E. Butterworth, "Peer Influences on Levels of Occupational and Educational Aspiration," *Social Forces*, XXXVIII (May, 1960), 289–295.

[3] C. Norman Alexander, Jr., and Ernest Q. Campbell, "Peer Influences on Adolescent Educational Aspirations and Attainments," *American Sociological Review*, XXIX (August, 1964), 568–575; Ernest Q. Campbell and C. Norman Alexander, "'Structural Effects and Interpersonal Relationships," *American Journal of Sociology*, LXXI (November, 1965), 284–289; M. Richard Cramer, "The Relationship between Educational and Occupational Plans of High School Students" (presented at the 1967 meeting of the Southern Sociological Society).

Reiss, and Duncan,[4] who were able to show that a minor part of the observed homophily is due to school segregation, the greater part to assortative choices of friends within schools. (2) In their study, Haller and Butterworth were concerned to eliminate socioeconomic homophily as a complete explanation of similarity in aspirations of friends. Hence, they considered peer-pairs whose members were alike in social class background and with respect to level of parental aspiration for their sons, as well as in measured intelligence. They showed that a positive intraclass correlation of close friends' aspirations held within such homogeneous pairs. (3) Invoking balance theory, Alexander and Campbell observed that agreement between friends' plans and desires to attend college was greater when the friendship choice was reciprocated than when it was not.[5] (4) In a second analysis, the same writers showed that "structural effects" of school socioeconomic composition were mediated by individual status effects, in that the former disappeared when the latter were taken into account: "Given knowledge of an individual's immediate interpersonal influences, the characteristics of the total collectivity provide no additional contribution to the prediction of his [college plans]."[6] (5) While Cramer notes an appreciable frequency of extreme incongruity between educational plans and occupational aspirations, the correlation between the two variables is relatively high, and too high to be explained fully by the operation of background factors as common causes.[7]

That "structural effects" are no large part of the explanation of similarity in friends' aspirations is easily demonstrated by an even more straightforward method than that used by Campbell and Alexander. For illustration, consider the correlation of .4986 between educational plans and friend's plans obtained by William H. Sewell (unpublished) for an all-Wisconsin sample of 4,386 boys who were high school seniors in 1957. Sewell found a correlation of .3364 between respondent's plans and the percentage of students in his class planning to go to college and similarly a correlation of .3327 between friend's plans and percentage going. (This is formally the same as the correlation ratio of individual plans on school.) Hence, if "structural effects" explained the correlation between friends, the latter would have been (.3364) (.3327) = .1119, which falls short of the observed correlation by .3867. Stated otherwise, the average within-school correlation between plans and friend's plans was .4354 (closely comparable to figures cited below for boys in a single school district).

[4] Albert Lewis Rhodes, Albert J. Reiss, Jr., and Otis Dudley Duncan, "Occupational Segregation in a Metropolitan School System," *American Journal of Sociology*, LXX (May, 1965), 682–694, and LXXI (July, 1965), 131.

[5] Alexander and Campbell, *op. cit.*

[6] Campbell and Alexander, *op. cit.*, p. 288.

[7] Cramer, *op. cit.*

The evidence is clear, therefore, that neither status homophily nor similarity in aspirations of friends is close to being adequately explained by school "structural effects" or school segregation. Yet the latter factors are important enough that they must be taken into account in estimating the impact of peer interaction on the formation of aspirations. By limiting the inquiry to a single school (or, alternatively, by looking at average within-school relationships), one can eliminate these factors from the analysis.

There remains the question of how to estimate the magnitude of peer influence, a problem not tackled directly in the studies cited, where the authors were content to stop with the detection of significant relationships indicating that a non-spurious correlation between aspirations (or plans) and friend's aspirations actually exists. The estimation (as distinct from detection) of such effects appears to be a particular instance of a generic problem for which explicit explanatory models have not yet been proposed. The purpose of this report is to suggest the possibilities in one kind of model. The purpose is realized with a presentation of some possible reinterpretations of the data originally collected and analyzed by Haller and Butterworth in the paper already cited.

Data

The original sample consisted of all seventeen-year-old boys in school in Lenawee County (Michigan) during the spring of 1957, interviews and test data being secured for 442 persons. For 329 of these boys, data were included in the same sample for at least one person listed as a best friend. Whether the friendship was reciprocated is not considered in this analysis. Some of the friends are, of course, included among the 329 respondents identified by this procedure, but others are not. It is important to note that the data on social and psychological characteristics as well as aspirations were obtained directly from the friend and not via the respondent's report on his friend.

The five variables to be considered are briefly enumerated (for lengthier descriptions, see the original publication):[8] *Level of occupational aspiration* is the score on Haller and Miller's Occupational Aspiration Scale. *Level of educational aspiration* is a composite score based on several questions about the number of years of college or university training the respondent planned to complete. *Socioeconomic status (SES)* is measured by Sewell's socioeconomic status scale, which includes items of parental educational attainment as well as material possessions in the home. *Intelligence* refers to raw scores on Cattell's Test of G Culture Free. *Parental aspirations* are indexed by a composite score

[8] See also Archibald O. Haller and Irwin W. Miller, *The Occupational Aspiration Scale: Theory, Structure and Correlates* (Technical Bulletin 288 [East Lansing: Agricultural Experiment Station, Michigan State University, 1963]).

from the answers to questions asked about the degree to which parents "encouraged" the respondent to have high levels of achievement. The distribution of each of these variables was normalized as a preliminary step to all further calculations.

The intercorrelations of these variables are presented in Table 13.1. For comparison, the same table includes the correlations derived from the complete data on all 442 boys. It appears that the correlations either for friends or for respondents are fairly representative of those obtaining within the whole population.

Models

It is easy to demonstrate that neither socioeconomic homophily nor friendship assortment by intelligence, nor the combination of the two, suffices to account for the correlation between respondent's and friend's aspirations. For example, the correlation between the two occupational aspirations is given in Table 13.1 as .42. In a multiple regression of respondent's occupational aspiration on friend's occupational aspiration, with the intelligence, parental aspiration, and SES of both boys included as additional independent variables, the standardized net regression coefficient (β coefficient) is a highly significant .26. If the regression is turned around, with friend's aspiration as the dependent variable and respondent's aspiration as one of the seven independent variables, its coefficient is .24.

While these results demonstrate that there is a net relationship between respondent's and friend's aspirations with homophily taken into account, the method by which these regression coefficients are estimated is not acceptable if we postulate both a causal influence of friend's on respondent's aspiration and vice versa. If friend's aspiration influences respondent's, it is illogical to use a model in which the latter is an explanatory variable in accounting for the former without reckoning with the reciprocal influence of the former upon the latter. Given that only one observation is made upon each aspiration, presumably at a stage when both have become relatively crystallized, we must think of the two dependent variables as being simultaneously determined, each being influenced by the other as well as by the remaining variables in the model.

Table 13.1. Observed correlations for 329 respondents and their best friends (above diagonal), and "synthetic" correlations, including observed correlations for 442 respondents (below diagonal)

Variable	*Symbol*	X_a	X_b	X_c	Y_1	Y_2	X_d	X_e	X_f	Y_3	Y_4
Respondent:											
Intelligence	X_a	....	.1839	.2220	.4105	.4043	.3355	.1021	.1861	.2598	.2903
Parental aspiration	X_b	.16*	....	.0489	.2137	.2742	.0782	.1147	.0186	.0839	.1124
Family SES	X_c	.23*	.09*	....	.3240	.4047	.2302	.0931	.2707	.2786	.3054
Occupational aspiration	Y_1	.45*	.22*	.37*	....	.6247	.2995	.0760	.2930	.4216	.3269
Educational aspiration..	Y_2	.41*	.29*	.41*	.64*	....	.2863	.0702	.2407	.3275	.3669
Best friend:											
Intelligence	X_d	.34	.09	.21	.28	.29	....	.2087	.2950	.5007	.5191
Parental aspiration	X_e	.09	.11	.06	.08	.09	.16*	....	—.0438	.1988	.2784
Family SES	X_f	.21	.06	.27	.29	.27	.23*	.09*	....	.3607	.4105
Occupational aspiration	Y_3	.28	.08	.29	.42	.33	.45*	.22*	.37*	....	.6404
Educational aspiration.	Y_4	.29	.09	.27	.33	.37	.41*	.29*	.41*	.64*	....

* Observed correlations for 442 respondents; remaining correlations below diagonal are obtained by averaging correlations above diagonal, as explained in the text.

Some exposition of simultaneous models, or models incorporating reciprocal influences, appears in the literature on path analysis.[9] Models of this type have been extensively considered in econometrics.[10] It is clear from the discussion in both these contexts that straightforward regression of one dependent variable upon a set of predictor variables does not yield the proper estimate of the effect of a variable which is simultaneously being determined within the model. Our interpretation here is presented in the framework of path analysis, but takes advantage of some of the approaches developed in econometrics. The reader is referred to a previous discussion of path analysis in sociological research,[11] with the warning that some of the simplified algorithms for path diagrams stated there apply only to simple recursive systems and not to simultaneous systems. This paper, then, affords an introduction to and illustration of the treatment of such systems in a path framework. Emphasis is placed on explicit statement of the equations and specifications of the model and the derivations that may be made therefrom to secure equations from which estimates may be calculated.

A just-identified model

Several models will be presented for didactic purposes before we arrive at one that represents a reasonably firm though tentative interpretation. Model I

[9] Sewall Wright, "The Treatment of Reciprocal Interaction, with or without Lag, in Path Analysis," *Biometrics*, XVI (September, 1960), 423–445.

[10] J. Johnston, *Econometric Methods* (New York: McGraw-Hill Book Co., 1963), chap. ix; Arthur S. Goldberger, *Econometric Theory* (New York: John Wiley & Sons, 1964), chap. vii; E. Malinvaud, *Statistical Methods of Econometrics* (Chicago: Rand McNally & Co., 1966), Part V; R. L. Basmann, "An Expository Note on Estimation of Simultaneous Structural Equations," *Biometrics*, XVI (September, 1960), 464–480. For a more elementary presentation, Mordecai Ezekiel and Karl A. Fox, *Methods of Correlation and Regression Analysis* (3d ed.; New York: John Wiley & Sons, 1959), chap. xxiv. See also Everett R. Dempster, "The Question of Stability with Positive Feedback," *Biometrics*, XVI (September, 1960), 481–483.

[11] Otis Dudley Duncan, "Path Analysis: Sociological Examples," *American Journal of Sociology*, LXXII (July, 1966), 1–16.

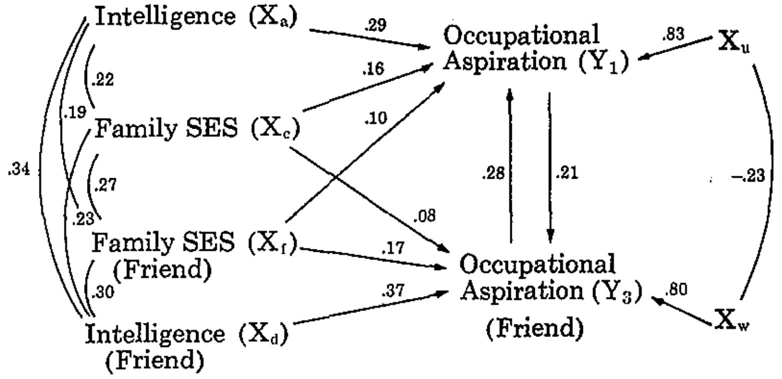

Figure 13.1. *Model I.*

(Fig. 13.1) represents occupational aspiration as depending directly on intelligence, family SES, friend's family SES, friend's occupational aspiration, and unspecified residual factors (or "disturbance," in the econometrician's parlance). The crucial assumption of the model is that occupational aspiration does *not* depend *directly* on friend's intelligence. Occupational aspiration does, however, depend directly on friend's family SES, an assumption that seems plausible enough if we consider the possibility that members of the friend's family as well as one's own may afford role models.

Three classes of variables are distinguished. The variables to the left, denoted by X's with letter subscripts taken from the first part of the alphabet, are "predetermined variables" with respect to this model. (The same variables might, of course, occur in a different role in some other model.) The Y's, with numerical subscripts, are the "endogenous variables," determined within the model, sometimes also called "jointly dependent variables." The X's with letter subscripts drawn from the end of the alphabet are "disturbances," standing for the residue of factors influencing aspiration but not explicitly measured.

The essential specification for all models of this type is that the disturbances are uncorrelated with the predetermined variables, though not (of course) with the endogenous variables and not necessarily with each other. Thus $r_{au}=r_{cu}=r_{fu}=r_{du}=r_{aw}=r_{cw}=r_{fw}=r_{dw}=0$. (We use r_{au} as a more economical notation for $r_{X_aX_u}$, etc.)

The intercorrelations of the predetermined variables are represented by the curved lines connecting them; similarly, there is a correlation between the two disturbances represented by a curved line. (The arrowheads at both ends of such curved lines, which are conventional in path diagrams, are omitted here to suggest that one must not rely on the usual algorithm for reading compound paths from the diagram when the system is non-recursive.)

While the diagram, on the conventions understood for reading it, specifies the model adequately, it is essential that the model also be stated explicitly in equations. Model I comprises two equations:

$$Y_1=p_{1a}X_a+p_{1c}X_c+p_{1f}X_f+p_{13}Y_3+p_{1u}X_u$$

and (Model I)

$$Y_3=p_{3d}X_d+p_{3f}X_f+p_{3c}X_c+p_{31}Y_1+p_{3w}X_w.$$

To these equations must be attached the earlier stipulation that disturbances are uncorrelated with predetermined variables.

To calculate estimates of the path coefficients, we must derive equations that yield a solution for them. To do this, we take advantage of the condition concerning the zero correlation of the predetermined variables and disturbances. Consider the first equation of the Model I. Multiply both sides of the equation by X_a, sum both sides over sample observations, and divide both sides by N, the number of sample observations. This yields

$$\frac{\Sigma Y_1X_a}{N}=p_{1a}\frac{\Sigma X_aX_a}{N}+p_{1c}\frac{\Sigma X_cX_a}{N}+p_{1f}\frac{\Sigma X_fX_a}{N}+p_{13}\frac{\Sigma Y_3X_a}{N}+p_{1u}\frac{\Sigma X_uX_a}{N}.$$

Since we are dealing with standardized variables (zero mean, unit variance), an expression like $\Sigma Y_1X_a/N$ simplifies immediately to r_{1a} while $\Sigma X_aX_a/N=1$. Note especially that $\Sigma X_uX_a/N=r_{ua}=0$, on the specification already stated. We proceed to multiply the first equation through by each of the four predetermined variables and to simplify the results as just indicated. This work yields the following set of four equations:

$$r_{1a}=p_{1a}+p_{1c}r_{ac}+p_{1f}r_{af}+p_{13}r_{3a},$$
$$r_{1c}=p_{1a}r_{ac}+p_{1c}+p_{1f}r_{cf}+p_{13}r_{3c},$$
$$r_{1f}=p_{1a}r_{af}+p_{1c}r_{cf}+p_{1f}+p_{13}r_{3f},$$
$$r_{1d}=p_{1a}r_{ad}+p_{1c}r_{cd}+p_{1f}r_{df}+p_{13}r_{3d}.$$

Or, in matrix form,

$$\begin{pmatrix} 1 & r_{ac} & r_{af} & r_{3a} \\ r_{ac} & 1 & r_{cf} & r_{3c} \\ r_{af} & r_{cf} & 1 & r_{3f} \\ r_{ad} & r_{cd} & r_{df} & r_{3d} \end{pmatrix} \begin{pmatrix} p_{1a} \\ p_{1c} \\ p_{1f} \\ p_{13} \end{pmatrix} = \begin{pmatrix} r_{1a} \\ r_{1c} \\ r_{1f} \\ r_{1d} \end{pmatrix}.$$

It will be noted that the square matrix on the left is not symmetric, unlike the case of the normal equations in conventional regression. Nevertheless, it will almost always be possible to invert this matrix (barring excessive collinearity among the predetermined variables) and thus to solve the four equations for the four unknown path coefficients, all the correlations on both the left- and right-hand sides of the equations being given in the data.

That we have exactly four equations to solve for four unknowns is essentially what is meant by saying that the model, or, more particularly, the relation in which Y_1 figures as the dependent variable, is "just identified." (The same holds, in this model, for the relation in which Y_3 is the dependent variable, but it need not always be the case that the situation in regard to identifiability is the same with respect to all relations in the model.) The econometricians have much more elegant ways of analyzing identifiability, but what it comes down to in cases like the one at hand is this; Y_1 depends on four variables explicitly (leaving aside the disturbance term), and there is just the same number of *predetermined* variables in the model. If there were four explanatory variables for Y_1 but fewer predetermined variables in the model, then the relation for Y_1 would be "underidentified"; that is, we could not derive a sufficient number of equations to yield a unique solution for the estimated paths leading to Y_1 (at least, not without some alteration in the specification of the model). On the other hand, if there are more predetermined variables than there are explanatory variables occurring in the model's equation for Y_1, this relation would be "overidentified" (examples of overidentification are given below).

The same procedure can be followed to secure equations from which we may solve for the paths leading to Y_3; that is, the second equation of the model is multiplied through, in turn, by each of the four predetermined variables, and the results are simplified to yield four equations in known correlations and four unknown path coefficients.

The method just presented yields exactly the same results as does the method of "indirect least squares," which is applicable to just-identified relations in a model. Computationally, however, it is a considerable simplification over indirect least squares, as has been noted in at least one textbook of econometrics.[12]

Some tedious but straightforward work remains if we are to calculate residual paths and the correlation between disturbances. We now multiply through each of the model equations by each of the endogenous variables and by each of the disturbance variables. This yields eight equations, which take the following form upon simplification:

$$r_{11}=1=p_{1a}r_{1a}+p_{1c}r_{1c}+p_{1f}r_{1f}+p_{13}r_{13}+p_{1u}r_{1u};$$
$$r_{13}=p_{1a}r_{3a}+p_{1c}r_{3c}+p_{1f}r_{3f}+p_{13}+p_{1u}r_{3u};$$
$$r_{33}=1=p_{3d}r_{3d}+p_{3f}r_{3f}+p_{3c}r_{3c}+p_{31}r_{13}+p_{3w}r_{3w};$$

[12] Goldberger, *op. cit.*, p. 348.

$$r_{13}=p_{3d}r_{1d}+p_{3f}r_{1f}+p_{3c}r_{1c}+p_{31}+p_{3w}r_{1w};$$
$$r_{1u}=p_{13}r_{3u}+p_{1u};\text{ hence }p_{1u}^2=p_{1u}r_{1u}-p_{13}p_{1u}r_{3u};$$
$$r_{3w}=p_{31}r_{1w}+p_{3w};\text{ hence }p_{3w}^2=p_{3w}r_{3w}-p_{31}p_{3w}r_{1w};$$
$$r_{3u}=p_{31}r_{1u}+p_{3w}r_{uw};$$
$$r_{1w}=p_{13}r_{3w}+p_{1u}r_{uw}.$$

A convenient solution routine is to compute $p_{1u}r_{1u}$ from the first equation (all other terms in it now being known), $p_{1u}r_{3u}$ from the second, $p_{3w}r_{3w}$ from the third, and $p_{3w}r_{1w}$ from the fourth. The values thus obtained can be substituted into the fifth and sixth equations to compute p_{1u} and p_{3w}, and it is then possible to obtain r_{1u}, r_{3w}, r_{3u}, and r_{1w} by a back solution. Either of the last two equations can then be used to obtain r_{uw}; if both are used, one has a partial check on the arithmetic.

Although it is advisable to retain a generous number of decimal places in the initial correlations and all intermediate calculations, the results are not likely to be meaningful for more than two places; the rounded estimates are shown in Figure 13.1.

Model I was also used with educational aspiration in place of occupational aspiration; that is, Y_2 replaced Y_1 and Y_4 replaced Y_3, with the predetermined variables and specifications remaining unchanged. For comparison, the results are as follows:

$$\begin{array}{lll} p_{2a}=.27 & p_{4c}=.04, & p_{2v}=.84, \\ p_{2c}=.26, & p_{4f}=.22, & p_{4z}=.79, \\ p_{2f}=.02, & p_{4d}=.36, & r_{vz}=-.38, \\ p_{24}=.25, & p_{42}=.29, & \end{array}$$

where X_v and X_z are the disturbances for Y_2 and Y_4, respectively.

The two sets of results are similar in most respects. Intelligence and family SES are appreciable influences on aspiration, while the influence of friend's family SES is barely detectable. The reciprocal paths of influence of friend's aspiration on respondent's aspiration, and vice versa, are around .2 or .3, by no means a negligible value. In both sets of results, consistent estimates are obtained only by acknowledging a rather substantial negative correlation between the disturbances of the two aspiration variables. If the analyst feels uncomfortable with the size of this correlation, as one may well feel in the absence of any evident rationalization of it, he may be inclined to reject the model even though the remaining estimates are reasonable.

Several things could be awry. Perhaps it is mistaken to argue that friend's intelligence can have no direct influence on one's aspiration (even though insertion of an additional path for this variable in Model I will render it underidentified). Perhaps there is a variable (such as "ambition," introduced into Model IV) omitted from the model with respect to which there is pronounced homophily and which is a significant cause of aspiration. In this event, the paths between the aspiration variables are probably overestimated, and the residual correlation is forced to compensate for this. In the present state of theory in social psychology, we are not likely to have firm grounds for asserting the validity of a model on grounds completely independent of the data for a given problem. Hence, it would seem that the best we can do is propose reasonable models, consider their plausibility, and, where indicated, undertake the construction of alternative ones (or await the work of a critic who may do so). Several alternatives to Model I were in fact attempted which yielded even less satisfactory results, but this does not prove that still another alternative could not be plausibly proposed.

An overidentified model

One such alternative is instructive, both as an example of estimation procedure when one is confronted with overidentification and as an indication of the sensitivity of the results to what may appear to be minor modifications of the model. In Model II (Fig. 13.2), we delete the path from friend's family SES to aspiration (and vice versa), suspecting on the basis of results with Model I that this variable is not very consequential. With this deletion, the equations are:

$$Y_1 = p_{1a}X_a + p_{1c}X_c + p_{13}Y_3 + p_{1u}X_u$$

and (Model II)

$$Y_3 = p_{3d}X_d + p_{3f}X_f + p_{31}Y_1 + p_{3w}X_w.$$

In the presence of overidentification, we cannot proceed as before to translate the foregoing specification about the model directly into a set of equations for estimating the model's parameters from a set of sample data. If we were to require that all four predetermined variables be uncorrelated with each residual variable *in the sample data*, we would be led to an inconsistency. From the first equation of the model, for example, requiring $r_{au} = r_{cu} = r_{fu} = r_{du} = 0$ would imply that there are four equations (one each for r_{1a}, r_{1c}, r_{1f}, and r_{1d}) involving just three unknowns (p_{1a}, p_{1c}, and p_{13}). The solution obtained from any three of these equations will not, in general, satisfy the fourth. Indirect least squares, or the equivalent procedure described above, is not available as a method of estimation.

In this situation we derive a set of estimating equations that are implied by the econometrician's method of "two-stage least squares." (The proof of the equivalence of our procedure to 2SLS, though elementary, is omitted.) We begin by computing the "first stage regression" of each dependent variable on all the predetermined variables, obtaining the regression equations:

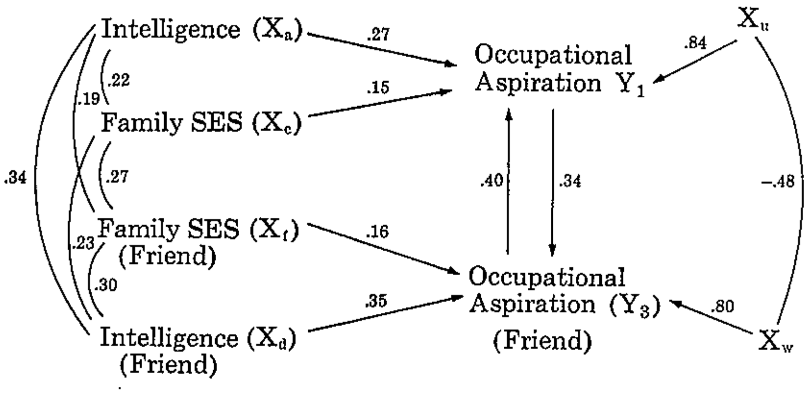

Figure 13.2. *Model II.*

$$\hat{Y}_1 = \beta_{1a.cfd}X_a + \beta_{1c.afd}X_c + \beta_{1f.acd}X_f + \beta_{1d.acf}X_d$$

and

$$\hat{Y}_3 = \beta_{3a.cfd}X_a + \beta_{3c.afd}X_c + \beta_{3f.acd}X_f + \beta_{3d.acf}X_d$$

where the β-coefficients are the ordinary regression coefficients in standard form computed from the sample data. These β-coefficients are of interest in connection with the "reduced form" of the model, as is pointed out in a subsequent section of the paper; but their immediate use is the one noted in the next paragraph.

The 2SLS method entails the following restrictions on the correlations between predetermined variables and residuals *in the sample*: $r_{au} = r_{cu} = r_{fw} = r_{dw} = 0$; $\beta_{3f.acd}r_{fu} + \beta_{3d.acf}r_{du} = 0$; and $\beta_{1a.cfd}r_{aw} + \beta_{1c.afd}r_{cw} = 0$. Note that neither r_{fu} nor r_{du}, individually, is zero, but their weighted sum is forced to be zero, the weights being those that obtain as a consequence of the 2SLS method. A parallel statement holds for r_{aw} and r_{cw}.

Let us multiply through the first equation of Model II, in turn, by each of the predetermined variables. We obtain:

$$r_{1a} = p_{1a} + p_{1c}r_{ac} + p_{13}r_{3a}$$
$$r_{1c} = p_{1a}r_{ac} + p_{1c} + p_{13}r_{3c}$$
$$r_{1f} = p_{1a}r_{af} + p_{1c}r_{cf} + p_{13}r_{3f} + p_{1u}r_{fu}$$
$$r_{1d} = p_{1a}r_{ad} + p_{1c}r_{cd} + p_{13}r_{3d} \times p_{1u}r_{du}$$

Now, multiply the third of these equations by $\beta_{3f.acd}$ and the fourth by $\beta_{3d.acf}$ and add the two together. This yields (omitting the secondary subscripts of the β's)

$$\beta_{3f}r_{1f}+\beta_{3d}r_{1d}=p_{1a}(\beta_{3f}r_{af}+\beta_{3d}r_{ad})+p_{1c}(\beta_{3f}r_{cf}+\beta_{3d}r_{cd}) \\ +p_{13}(\beta_{3f}r_{3f}+\beta_{3d}r_{3d}).$$

We need not show the term $p_{1u}(\beta_{3f}r_{fu}+\beta_{3d}r_{du})$ on the right hand side, since it is zero, by the requirement stated above. We can similarly write the set of three estimating equations from the second equation of the model:

$$r_{3f}=p_{3d}r_{df}+p_{3f}+p_{31}r_{1f}$$
$$r_{3d}=p_{3d}+p_{3f}r_{df}+p_{31}r_{1d}$$
$$\beta_{1a}r_{3a}+\beta_{1c}r_{3c}=p_{3d}(\beta_{1a}r_{ad}+\beta_{1c}r_{cd})+p_{3f}(\beta_{1a}r_{af}+\beta_{1c}r_{cf}) \\ +p_{31}(\beta_{1a}r_{1a}+\beta_{1c}r_{1c})$$

The solutions of the two sets of estimating equations are shown as the path coefficients in Figure 13.2. The solutions for the residual paths (p_{1u} and p_{3w}) and for the correlation between residuals (r_{uw}) are obtained by a routine like that described in the previous section. The equations to be solved are the following:

$$\begin{aligned}
r_{11}&=1=p_{1a}r_{1a}+p_{1c}r_{1c}+p_{13}r_{13}+p_{1u}r_{1u}\\
r_{13}&=p_{1a}r_{3a}+p_{1c}r_{3c}+p_{13}+p_{1u}r_{3u}\\
r_{33}&=1=p_{3d}r_{3d}+p_{3f}r_{3f}+p_{31}r_{13}+p_{3w}r_{3w}\\
r_{13}&=p_{3d}r_{1d}+p_{3f}r_{1f}+p_{31}+p_{3w}r_{1w}\\
r_{1u}&=p_{13}r_{3u}+p_{1u}\\
r_{3w}&=p_{31}r_{1w}+p_{3w}\\
r_{1w}&=p_{1a}r_{aw}+p_{1c}r_{cw}+p_{13}r_{3w}+p_{1u}r_{uw}\\
r_{3u}&=p_{3d}r_{du}+p_{3f}r_{fu}+p_{31}r_{1u}+p_{3w}r_{uw}
\end{aligned}$$

Note that the last two of these equations include the four non-zero correlations of predetermined variables with residuals. Their values are obtained from the respective estimating equations that contain them. For example, from the equation for r_{1f} given earlier, we obtain $r_{fu}=(r_{1f}-p_{1a}r_{af}-p_{1c}r_{cf}-p_{13}r_{3f})/p_{1u}$. From equations like these, we find

$$\begin{aligned}
r_{fu}&=.0665\\
r_{du}&=-.0340\\
r_{aw}&=-.0346\\
r_{cu}&=.0551
\end{aligned}$$

These small values do not seriously call into question the assumption of the model that all the correlations between predetermined variables and disturbances *in the universe* are zero. No such assumption is made in regard to the correlation between disturbances, so that we are prepared to find that r_{uw} differs from zero. Nevertheless, the substantial negative correlation, $r_{uw}=-.48$, seems difficult to interpret in substantive terms.

If we neglect the correlations of predetermined variables with residuals, we can compute a set of "implied correlations" that depend only on the path coefficients and the intercorrelations of the predetermined variables. Deviations of the implied correlations from the corresponding observed correlations provide another perspective on the "goodness of fit" of the model. For the implied correlation, r'_{1f}, for example, we have $r'_{1f}=p_{1a}r_{af}+p_{1c}r_{cf}+p_{13}r_{3f}$. Shown below are the implied correlations that are not constrained to equal their observed counterparts.

$$\begin{array}{ll}
r'_{1f}=.2371\ (-.0559) & r'_{3a}=.2876\ (.0278)\\
r'_{1d}=.3281\ (.0286) & r'_{3c}=.2342\ (-.0444)
\end{array}$$

Deviations from observed values are in parentheses. These deviations appear (with signs reversed) in the numerators of formulas of the type already illustrated for calculating the correlations of predetermined variables with sample residuals. Were these deviations to be sizable, we should be inclined to call the model into question.

A block-recursive model

Both models discussed thus far are "simultaneous" models in that the endogenous variables are jointly determined by the model within a single period of observation. In simple recursive models, by contrast, we assume that the endogenous variables are successively determined, and the specification is altered to exclude correlation of the disturbances among themselves. The latter specification (zero intercorrelation of disturbances) is common to the Simon-Blalock procedure of causal analysis and to the examples of stratification models given in a previous publication on path analysis.[13] There is no apparent reason, however, why features of both types of model cannot be combined in a single construction. (The term "construction" might well be preferred to "model" in contexts like the present one, to emphasize that we are indeed "construing" the data to mean what our interpretation via a diagram or system of equations represents them to mean.)

[13] Hubert M. Blalock, Jr., *Causal Inferences in Non-experimental Research* (Chapel Hill: University of North Carolina Press, 1964); Herbert A. Simon, *Models of Man* (New York: John Wiley & Sons, 1957), chap. ii; Duncan, *op. cit.*

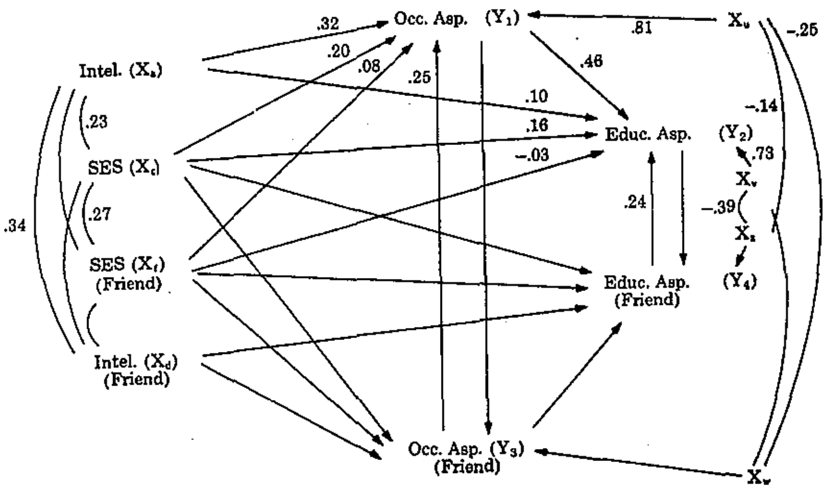

Figure 13.3. *Model III.*

In Model III (Fig. 13.3), we reason as though respondents and their friends "first" make up their minds what occupations they would like to pursue, and "then," on the basis of the occupational choice and other considerations, "decide" what educational preparation they may need. That such a construction is at best an oversimplification may be evident from introspection. This model, like the preceding ones, serves primarily a didactic purpose.

In the first sector of the model, occupational aspirations Y_1 and Y_3 are simultaneous endogenous variables and the specification is exactly the same as for Model I. We then assume that with respect to educational aspiration (Y_2 and Y_4) the predetermined variables include not only X_a, X_c, X_f, and X_d but also Y_1 and Y_3. Hence the specification $r_{1v}=r_{3z}=r_{uv}=r_{wz}=0$; but r_{uz}, r_{vw}, r_{uw}, r_{vz}, r_{3v}, and r_{1z} are not necessarily zero. The equations of the model are

$$
\begin{aligned}
Y_1&=p_{1a}X_a+p_{1c}X_c+p_{1f}X_f+p_{13}Y_3+p_{1u}X_u,\\
Y_3&=p_{3d}X_d+p_{3f}X_f+p_{3c}X_c+p_{31}Y_1+p_{3w}X_w, \qquad \text{(Model III)}\\
Y_2&=p_{2a}X_a+p_{2c}X_c+p_{2f}X_f+p_{21}Y_1+p_{24}Y_4+p_{2v}X_v,\\
Y_4&=p_{4d}X_d+p_{4f}X_f+p_{4c}X_c+p_{43}Y_3+p_{42}Y_2+p_{4z}X_z.
\end{aligned}
$$

In estimating the parameters of this model, we have engaged in a preliminary manipulation of the data which has nothing to do with the properties of the model but which is suggested by the somewhat artificial design of the data matrix. From the original correlation matrix in Table 13.1, we constructed a "synthetic" correlation matrix which was forced to be symmetrical in variables pertaining to respondent and friend. The intercorrelations of educational aspiration, occupational aspiration, family SES, and intelligence were assumed to be the same for respondents and friends, and the values thereof were assumed to be those observed in the entire original sample of 442 boys. The correlation between friend and respondent on each of these variables was retained from the data on the 329 pairs. The "cross-correlations" between friend and respondents were averaged; thus, for example, in the synthetic correlation matrix $r_{1a}=r_{3d}$ and each is the average of r_{1a} and r_{3d} as initially computed. The synthetic matrix is shown below the diagonal in Table 13.1. Given the symmetry of the data and the model, it is necessary to carry out computations for estimating only the path coefficients in the first and third equations.

Estimates for the first equation are obtained in the same way as described for Model I, and, indeed, the results differ from those in Model I only slightly. The third equation (for Y_2) is just identified, in virtue of the specification $r_{1v}=0$. Thus we can multiply this equation through by Y_1, X_a, X_c, X_f, and X_d in turn and simplify to obtain five equations in the five unknown path coefficients and known correlations. The calculation of residual paths for all equations proceeds along the lines already illustrated for Model I.

From the perspective of the hypothesis that orients this study, the most interesting estimates are those for the reciprocal paths for aspirations: $p_{13}=p_{31}=.25$ and $p_{24}=p_{42}=.24$. The paths from predetermined to endogenous variables seem reasonable, except perhaps for the negative though small value of $p_{2f}(=p_{4c})$, $-.03$. That family SES has an apparently larger influence on educational aspiration than does intelligence may be explained by two considerations: intelligence has a sizable influence on occupational aspiration, which, in this model, intervenes between the latter and educational aspiration; and the family SES scale is heavily weighted by parental educational attainment.

In short, the model gives satisfactory estimates on the whole, if we are prepared to accept the excessively rationalistic assumption that occupational decisions precede educational decisions (both being measured well before the actual decision point), and if we are prepared to overlook the substantial negative correlations between disturbances, $r_{uw}=-.25$ and $r_{vz}=-.39$. Dissatisfaction with our ability to rationalize the latter motivated one more alternative construction on these data.

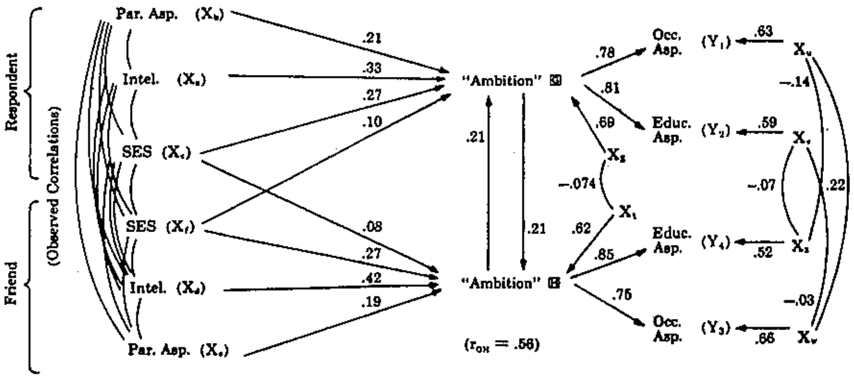

Figure 13.4. *Model IV.*

A composite model

The approach taken in Model IV (Fig. 13.4) was suggested by a remark of Turner and Stevens, who called attention to the possibility of constructions including aspects of both factor analysis and causal modeling.[14] The analogy with factor analysis here consists in the postulation of an unobserved variable, called "ambition" for lack of a better term, that underlies both educational and occupational aspiration. There are perhaps two justifications for such a postulate. First, both aspiration variables are probably rather unreliable relative to the other variables in the model, and it might be advisable to follow a procedure that reduces the attenuation introduced by measurement error before proceeding to a causal interpretation. Second, on a purely introspective basis it seems likely that many boys do not make a neat conceptual separation of educational and occupational aspirations, nor do they make plans for schooling and job choices in any fixed order. The argument here is just the opposite of the one that would have to be used to justify the precedence of occupational aspiration with respect to educational aspiration in Model III. If the reader likes, he may identify "ambition" with general level of aspiration, as the latter was conceptualized in the classic literature on the subject.[15] Our procedure would then suggest that level of aspiration can only be manifested in and recognized by orientations toward particular

[14] Malcolm E. Turner and Charles D. Stevens, "The Regression Analysis of Causal Paths," *Biometrics*, XV (June, 1959), 236–258.

[15] K. Lewin, Tamara Dembo, L. Festinger, and Pauline S. Sears, "Level of Aspiration," in J. McV. Hunt (ed.), *Personality and the Behavior Disorders* (New York: Ronald Press, 1944), I, 333–378.

goals, such as educational attainment or occupational achievement. We are, in effect, using verbalized educational and occupational aspirations as *indicators* of the construct, level of aspiration or "ambition."

Although the approach taken to the construction of hypothetical "ambition" variables is suggested by factor analysis, the actual procedure is an ad hoc one that depends on various heuristic considerations rather than on one of the standard factor models. Lest this arouse undue anxiety, we observe that the classic factor models, in their time, were similarly motivated by heuristic concerns. That they have become frozen into a canonical procedure taken over by investigators of problems whose structures are quite different from those of the pioneers is perhaps a commentary on the relative potence of propensities to imitate and to innovate in research. Certainly, after Wright's illuminating contrast between the interpretive and the purely mathematical approaches to factor analysis,[16] we need no longer feel constrained to follow a single routine that someone has ventured to dub "objective" or "optimal."

We begin, as usual, by writing the entire set of equations which are graphically represented in the path diagram of Model IV:

$$
\begin{aligned}
Y_1 &= p_{1G}G + p_{1u}X_u, \\
Y_2 &= p_{2G}G + p_{2v}X_v, \\
Y_3 &= p_{3H}H + p_{3w}X_w, \\
Y_4 &= p_{4H}H + p_{4z}X_z, && \text{(Model IV)} \\
G &= p_{Gb}X_b + p_{Ga}X_a + p_{Gc}X_c + p_{Gf}X_f + p_{GH}H + p_{Gs}X_s, \\
H &= p_{He}X_e + p_{Hd}X_d + p_{Hf}X_f + p_{Hc}X_c + p_{HG}G + p_{Ht}X_t.
\end{aligned}
$$

The last two equations contain the substance of the model. G and H are the simultaneous endogenous variables, $X_a, \ldots, X_f$ are the predetermined variables, and X_s and X_t are the disturbances, which need not be uncorrelated with each other although both are uncorrelated with all predetermined variables. Each relation is overidentified, since each includes only five explanatory variables while there are six predetermined variables in the model. We shall again employ the version of two-stage least squares illustrated for Model II when we come to the estimation of the path coefficients in these two equations.

Before proceeding to this task, we must first construct the two hypothetical endogenous variables by deriving their correlations with the predetermined variables, making use of the intercorrelations of the aspiration scores. Hence, attention will be focused first on the right-hand portion of the diagram.

[16] Sewall Wright, "The Interpretation of Multivariate Systems," in O. Kempthorne *et al.* (eds.), *Statistics and Mathematics in Biology* (Ames: Iowa State College Press, 1954), chap. ii, especially pp. 27–32.

The factor model employed here says that the correlation between a boy's educational and occupational aspirations is completely accounted for by his "ambition." The correlation between the aspirations of friends is substantially accounted for by the correlation of their respective values on "ambition," but the model allows for some between-friend correlation of specific elements in the two aspiration scores. Hence, the specifications $r_{uv}=r_{wz}=0$; but r_{uz}, r_{vw}, r_{vz}, and r_{uw} are not necessarily zero. There are six known correlations among the aspiration variables, $Y_1, \ldots, Y_4$, and we have four equations of complete determination, stipulating that each aspiration variable is completely determined by the appropriate "ambition" variable and the residual. These ten conditions, however, do not suffice to estimate the paths in the first four equations of Model IV, inasmuch as we must estimate the four paths from "ambition" to aspiration, the four residual paths, four intercorrelations of residuals, and the unknown correlation between the two "ambition" variables, to wit, r_{GH}, for a total of thirteen unknowns. This portion of the model, by itself, is underidentified. Underidentification is, of course, the typical situation in factor analysis, which the analyst gets around by imposing various mathematical constraints on the solution. Here we invoke, instead, the "external" information provided by the predetermined variables as a constraint on the solution.

We begin by noting that the six correlations among the aspiration variables can be written as follows (obtained by "multiplying through" each of the first four equations in the model by the other aspiration variables):

$$
\begin{aligned}
r_{12} &= p_{1G}p_{2G} = r_{1G}r_{2G},\\
r_{34} &= p_{3H}p_{4H} = r_{3H}r_{4H},\\
r_{13} &= p_{1G}p_{3H}r_{GH} + p_{1u}p_{3w}r_{uw},\\
r_{24} &= p_{2G}p_{4H}r_{GH} + p_{2v}p_{4z}r_{vz},\\
r_{14} &= p_{1G}p_{4H}r_{GH} + p_{1u}p_{4z}r_{uz},\\
r_{23} &= p_{2G}p_{3H}r_{GH} + p_{2v}p_{3w}r_{vw}.
\end{aligned}
$$

Bringing the predetermined variables into the picture, we note that $r_{1a}=p_{1G}r_{aG}$ and $r_{2a}=p_{2G}r_{aG}$. From these it would follow that $p_{1G}/p_{2G}=r_{1a}/r_{2a}$; but we can equally well compute $p_{1G}/p_{2G}=r_{1b}/r_{2b}$, and so on. As a compromise among the six possible estimates of this ratio, we take

$$p_{1G}/p_{2G}=\sum_i r_{1i}/\sum_i r_{2i},$$

where $i=a, \ldots, f$. Given this ratio p_{1G}/p_{2G} and the previously noted $r_{12}=p_{1G}p_{2G}$, we can solve at once for p_{1G} and p_{2G}. A similar procedure yields p_{3H} and p_{4H}.

Now, if we disregard correlations among the residual factors of the Y's, we may compute the implied correlations,

$$
\begin{aligned}
r'_{13} &= p_{1G}p_{3H}r_{GH},\\
r'_{24} &= p_{2G}p_{4H}r_{GH},\\
r'_{14} &= p_{1G}p_{4H}r_{GH},\\
r'_{23} &= p_{2G}p_{3H}r_{GH}.
\end{aligned}
$$

Let $\delta_1=r_{13}-r'_{13}$, $\delta_2=r_{24}-r'_{24}$, $\delta_3=r_{14}-r'_{14}$, $\delta_4=r_{23}-r'_{23}$, and $S=\Sigma\delta^2$. We wish to select r_{GH} so that S is at its minimum. Finding dS/dr_{GH}, setting it equal to zero, and rearranging, we obtain

$$r_{GH}=\frac{r_{13}p_{1G}p_{3H}+r_{24}p_{2G}p_{4H}+r_{14}p_{1G}p_{4H}+r_{23}p_{2G}p_{3H}}{(p_{1G}p_{3H})^2+(p_{2G}p_{4H})^2+(p_{1G}p_{4H})^2+(p_{2G}p_{3H})^2}.$$

The path coefficients in this expression have already been estimated and the correlations are known; hence r_{GH} is easily calculated as .56, which is the hypothetical correlation between "ambition" of respondent and "ambition" of friend. It is a materially higher value than any of the observed correlations between aspiration scores. This is the most important result of the work done to this point. We complete this phase of the calculations by computing residual paths for the aspiration variables, using formulas like $p_{1u}{}^2=1-p_{1G}{}^2$, which are obvious from the model equations. Finally, we may return to the identities stated for the intercorrelations among the Y's with enough information in hand to solve the last four equations for r_{uw}, r_{vz}, r_{uz}, and r_{vw}, respectively.

One more step is necessary before we can begin the estimation of the last two model equations. As already noted, $r_{1a}=p_{1G}r_{aG}$ and $r_{2a}=p_{2G}r_{aG}$. Since r_{1a} and r_{2a} are known and we now have estimates of p_{1G} and p_{2G}, we obtain two solutions for r_{aG}. These are not the same, so we strike a simple average of the two and follow an analogous procedure to secure correlations of G and H with all the predetermined variables.

By this sequence of estimates and approximations, we have arrived at a complete correlation matrix for variables G, H, and $X_a, \ldots, X_f$. Coefficients in the last two equations of Model IV are now estimated by the two-stage least-squares procedure already illustrated for Model II. The estimating equations are derived by multiplying through the last two equations of Model IV (those involving G and H) by each of the six predetermined variables, imposing the following restrictions on the sample correlations between residuals and predetermined variables, in conformity with the 2SLS principle of estimation: $r_{bs}=r_{as}=r_{cs}=r_{fs}=r_{ct}=r_{ft}=r_{dt}=r_{et}=\beta_{Hd}r_{ds}+\beta_{He}r_{es}=\beta_{Gb}r_{bt}+\beta_{Ga}r_{at}=0$.

The solutions for the residual paths and for the non-zero correlations of predetermined variables and residuals are accomplished by the routine that was illustrated for Model II. We find

$$r_{ds}=.0291$$
$$r_{es}=-.0683$$
$$r_{bt}=-.0122$$
$$r_{at}=.0079$$

Neglecting these residual correlations we have the following implied values for correlations that are not forced to equal their observed values (deviations from the latter are shown in parentheses):

$$r'_{Gd}=.3508\ (-.0201)$$
$$r'_{Ge}=.1397\ (.0471)$$
$$r'_{Hb}=.1293\ (.0076)$$
$$r'_{Ha}=.3383\ (-.0049)$$

None of these results casts doubt on the credibility of the model, which seems acceptable, moreover, in regard to the reasonable magnitudes of the path coefficients and the small size of the correlation between residuals ($r_{st}=-.074$, as shown in Figure 13.4). We may, as a final indication of goodness of fit, compute all the correlations involving the aspiration variables that are implied by the path coefficients, neglecting correlations involving residuals (except for r_{st}). Typical formulas for this purpose are:

$$r'_{1a}=p_{1G}r_{Ga}$$
$$r'_{1d}=p_{1G}r'_{Gd}$$
$$r'_{14}=p_{1G}r_{GH}p_{4H}$$

The implied correlations and their deviations from observed correlations are shown in Table 13.2. The deviations of the implied from the observed correlations are analogous to the "residual correlations" obtained after extracting the "meaningful" factors in a factor analysis. They may be attributed to sampling error, if this seems reasonable. If not, they afford material for an investigator who wishes to essay a more convincing interpretation than that afforded by Model IV.

Concerning the reduced form

In all the discussion thus far we have considered only what the econometricians call the "structural form" of the models. For some purposes, it is instructive also to pay attention to the "reduced form." Returning to the last two equations of Model IV, which constitute the substance of that model,

Table 13.2. Correlations implied by Model IV between aspiration variables and other variables in the model

Variable	Variable (see stub)			
	Y_1	Y_2	Y_4	Y_5
Respondent:				
Parental aspiration (X_b)	.2388(.0251)	.2482(–.0260)	.1104(–.0020)	.0969(.0130)
Intelligence (X_a)	.3997(–.0108)	.4155(.0112)	.2889(–.0014)	.2537(–.0061)
Family SES (X_c)	.3567(.0327)	.3707(–.0340)	.3114(.0060)	.2734(–.0052)
Friend:				
Family SES (X_f)	.2623(–.0307)	.2726(.0319)	.4106(.0001)	.3605(–.0002)
Intelligence (X_d)	.2719(–.0276)	.2826(–.0037)	.5447(.0256)	.4783(–.0214)
Parental aspiration (X_e)	.1083(.0323)	.1126(.0424)	.2524(–.0260)	.2216(.0228)
Respondent:				
Occupational aspiration (Y_1) .	. . . .	.6247*	.3738(.0469)	.3283(–.0933)
Educational aspiration (Y_2) . . .	. . . .	. . . .	.3885(.0216)	.3412(.0137)
Friend:				
Educational aspiration (Y_4) . .	. . . .	. . . .	. . . .	.6404*
Occupational aspiration (Y_5) . .	. . . .	. . . .	. . . .	. . . .

NOTE. Deviations from observed correlations are in parentheses.

* Model permits no deviation of implied from observed correlation. (Deviations from the other observed intercorrelations among the Y's occur if residual intercorrelations are neglected.)

let us substitute the expression for H given in the last equation into the next-to-last equation. We obtain a result that can be put into the form

$$G=\alpha_aX_a+\alpha_bX_b+\alpha_cX_c+\alpha_dX_d+\alpha_eX_e+\alpha_fX_f+\alpha_sX_s+\alpha_tX_t.$$

Similarly, substituting the fifth equation into the sixth, we obtain

$$H=\beta_aX_a+\beta_bX_b+\beta_cX_c+\beta_dX_d+\beta_eX_e+\beta_fX_f+\beta_sX_s+\beta_tX_t.$$

Let $\gamma=1/(1-p_{GH}p_{HG})$. Then the constants in the two foregoing reduced-form equations are defined as follows:

$$\alpha_a=p_{Ga}\gamma,\quad \alpha_b=p_{Gb}\gamma,\quad \alpha_c=(p_{Gc}+p_{GH}p_{Hc})\gamma,\quad \alpha_d=p_{GH}p_{Hd}\gamma,\quad \alpha_e=p_{GH}p_{He}\gamma,$$
$$\alpha_f=(p_{Gf}+p_{GH}p_{Hf})\gamma,\quad \alpha_s=p_{Gs}\gamma,\quad \alpha_t=p_{GH}p_{Ht}\gamma,$$
$$\beta_a=p_{HG}p_{Ga}\gamma,\quad \beta_b=p_{HG}p_{Gb}\gamma,\quad \beta_c=(p_{Hc}+p_{HG}p_{Gc})\gamma,\quad \beta_d=p_{Hd}\gamma,\quad \beta_e=p_{He}\gamma,$$
$$\beta_f=(p_{Hf}+p_{HG}p_{Gf})\gamma,\quad \beta_s=p_{HG}p_{Gs}\gamma,\quad \beta_t=p_{Ht}\gamma.$$

Each reduced-form equation represents one endogenous variable as a linear combination of all the predetermined variables and the disturbances. The coefficients in this linear combination are not, however, linear combinations of the coefficients of the structural form.

Since X_s and X_t are specified to be uncorrelated with the predetermined variables, the coefficients $\alpha_a, \ldots, \alpha_f$ can be estimated by the ordinary regression of G on $X_a, \ldots, X_f$ and the β's from the regression of H on the same six variables. (These regressions have already been computed for the first stage of the two-stage estimation procedure.)

It may be observed that $\alpha_d/\beta_d = p_{GH}$; but $\alpha_e/\beta_e = p_{GH}$, also. If we use our first-stage regression estimates of the α's and β's, the former implies $p_{GH} = .2891$ while the second yields $-.0838$. Similarly, estimating p_{HG} from β_a/α_a and β_b/α_b yields two inconsistent answers, .2338 and .1722, respectively. These inconsistencies present the problem of overidentification in an especially striking fashion. Had the equations in Model IV both been just identified, such inconsistencies would not have arisen. Indeed, the method of "indirect least squares," alluded to earlier as a technique for estimating coefficients in a just-identified system, consists precisely in estimating the reduced-form coefficients first and then deriving therefrom the estimates of coefficients in the structural equations of the model.

Although the overidentification means that the reduced-form coefficients do not yield unique estimates of the structural coefficients, we may still be interested in the reduced form of the model. To begin with, having obtained the structural coefficients by the two-stage procedure, we may compute the implied values of the reduced-form coefficients from them. Table 13.3 shows the reduced-form coefficients of Model IV, both as estimated from the first-stage regressions and as computed from the structural coefficients. For the most part, the discrepancies appear small, although it is these very discrepancies that preclude the indirect least-squares approach.

Table 13.3. Reduced-form coefficients for Model IV

Independent Variable	*Symbol*	*Dependent variable*			
		Set (A)		*Set (B)*	
		G	*H*	*G*	*H*
Respondent:					
Intelligence	X_a	.3378	.0790	.3418	.0725
Parental aspiration	X_b	.2199	.0379	.2169	.0460
Family SES.	X_c	.3056	.1426	.2919	.1396
Best Friend:					
Intelligence.	X_d	.1291	.4464	.0913	.4399
Parental aspiration	X_e	$-.0159$	.1900	.0418	.2016
Family SES.	X_f	.1499	.3034	.1585	.2942

NOTE. Set (A), as estimated in first-stage regression; set (B), as computed from structural coefficients estimated in two-stage regression.

Apart from computation and estimation, the reduced-form coefficients, as defined in terms of the structural coefficients, have some conceptual or interpretive significance, for their definitions indicate something of the "mechanisms" through which the predetermined variables influence the endogenous variables. Consider

$$\alpha_a = \frac{p_{Ga}}{1 - p_{GH}p_{HG}} = \frac{.3267}{.9560} = .34.$$

The form of this expression indicates that X_a influences G directly, via p_{Ga}, and that this influence is slightly amplified, to the extent of $1/(.956) = 1.046$, by the reciprocal action of respondent's and friend's "ambition." A more complex mechanism is suggested by

$$\alpha_c = \frac{p_{Gc} + p_{GH}p_{Hc}}{1 - p_{GH}p_{HG}} = \frac{.2749 + (.2074)(.0785)}{.956} = .29.$$

In this case, there is not only the direct effect, p_{Gc}, but also a compound or indirect effect of X_c on G, via H, as represented by $p_{GH}p_{Hc} = .0163$, while the sum of the two is amplified by the reciprocation. The third type of mechanism is exemplified by

$$\alpha_d = \frac{p_{GH}p_{Hd}}{1 - p_{GH}p_{HG}} = \frac{(.2074)(.4205)}{.956} = .09.$$

Here the influence of X_d on G is entirely indirect, via H, with the same factor of amplification as in the previous examples. Indeed, it is the assumption that some of the predetermined variables influence the endogenous variables only indirectly that permits the model to be identified in the first place. (Econometricians somewhat confusingly call both just-identified and overidentified models or equations "identified.")

The formulas for the reduced-form coefficients, of course, merely reflect the assumptions built into the model. But if the assumptions are accepted, it is of interest to include in the interpretation some evaluation of the relative importance of the various mechanisms that the model implicitly postulates.

Evaluation and discussion

Closing remarks are confined to comments on the rationale and results of Model IV. Some deficiencies of the earlier models have already been mentioned, and others that merit emphasis are shared by them with Model IV.

To recapitulate, the study was concerned with the hypothesis that adolescent boys influence each other in forming their occupational and educational aspirations. The earlier analysis of Haller and Butterworth had demonstrated that some correlation between the aspirations of boys in a peer group remained even if the groups were confined to those homogeneous on background factors presumed to give rise to aspirations. The present analysis accepts a different task: not that of hypothesis testing, but that of *estimation* in the context of an *explicit causal interpretation* of the influences on aspiration. The estimates are meaningful only to the extent that the initial study design is adequate to the purposes of identifying determinants of aspiration and of ascertaining the patterns of homophily operative in a relevant population.

The first question, then, is posed by a limitation on the design noted by the authors of the original study: with these data we cannot rule out the possibility that friendships are formed partially on the basis of common interests in educational and occupational goals. If this be the case, then all the estimates attempted here are beside the point, because we have treated aspirations as outcomes of the background characteristics of the respondent and his friend (treated as "predetermined variables") and of their respective influences on each other. As was also noted in the earlier paper, a longitudinal design would be required to eliminate the possibility of assortment on the basis of aspirations (although it is not entirely clear how the requisite causal inferences would be made, even if the design were longitudinal). The results here are, therefore, in the same provisional status as those of the predecessor study. If assortment on the basis of aspirations proved to be important, our estimates of the mutual influence of friends on each other's aspirations are not merely wrong; they become irrelevant.

Supposing, however, that friendship assortment occurs primarily on the basis of social and personal characteristics other than aspirations (though possibly, as the models suggest, on the basis of factors affecting aspirations); then we must reckon with the further question of whether the effects of such characteristics are adequately accounted for. Percentages of "explained" variance in the jointly dependent variables of simultaneous models are not readily computed as they are for ordinary multiple regressions. From the size of the path coefficients for the disturbances, however, we might be prone to assume that some relevant background characteristics are omitted. Still, it is not obvious what they might be, since most studies of status achievement and aspirations have focused on variables much like those used here. Strictly speaking, the disturbance in a model represents all variables that operate "accidentally" or randomly with respect to the influence of predetermined variables. Retrospective introspection certainly suggests that many accidental experiences, not necessarily shared with one's best friend, may have an impact on the formation of aspirations. Again, the results must be left in provisional form: if further investigation discloses major background factors inducing high or low aspirations, and if there is significant homophily with respect to these factors, incorporation of such factors into constructions like Model IV may well result in drastic reduction of the paths representing reciprocal influence of respondent's and friend's aspirations.

Looking more specifically at the results with Model IV, we may note that some of the asymmetry between respondent and friend that seemed implausible in Models I and II has disappeared in the more elaborate model. Except that friend's "ambition" seems to be more heavily influenced by his intelligence than is the case for respondent's "ambition," the results for the two boys are much alike. It is only a coincidence that $p_{GH}=p_{HG}$ when the results are rounded to two decimal places. But either of these coefficients at a value of roughly .2 seems like a reasonable estimate of the influence of a friend's aspirations upon one's own. This estimate does not differ greatly from those for Models I and III, but is on the conservative side by comparison with them. Recalling that the correlation between H and G, the two constructed "ambition" variables, came out as .56, the .21 value for p_{GH} and p_{HG} suggests that a significant part of the explanation for resemblance between aspirations of peers is due to mutual influence, but a goodly part of it is also due to the way in which peers come to associate (assortatively with respect to background characteristics) in the first place—bearing in mind the reservation already stated concerning the cogency of this interpretation.

The result that p_{GH} is very nearly the same as p_{HG} may be regarded as somewhat anomalous. In Model III the two reciprocal paths were forced to be equal, but in Models I and II, where this was not the case, the influence of friend on ego appeared to be somewhat stronger than that of ego on friend. This is perhaps what we should expect, given that the friendship pairs analyzed here are not defined by mutual choices but by the unilateral choice of the respondent. We know (or can presume) that friend is a significant other for ego, but we cannot be sure that the converse is true. Clearly, the whole matter of the extent to which an individual's dispositions are influenced by significant others should be further explored in research designed to include measures of degree of significance of those others, estimated independently of the dependent variable under study.

Of the discrepancies between observed correlations and those implied by the model, only the one of $-.09$ for r_{13}, between the two friends' occupational aspirations, seems interesting (this discrepancy appears in a different guise as $r_{uw}=.22$). The model may seem to fail to represent quite adequately some specific aspect of similarity of friend's occupational choices. It need not be argued, however, that the model underestimates mutual influence of friends' occupational aspirations. If friends encounter the same role models, apart from their families, this could induce some similarity in their cognitive and affective orientations to the world of work.

A final reservation will be stated, although others may well occur to the reader. The parental aspiration variable is based on the respondents' reports of their parents' attitudes. Hence this variable may well be contaminated to some degree by the dependent variables which it supposedly helps to explain. Fortunately, the study design precludes a similar contamination of the data on friend's aspirations.

The reader may well be appalled at all the apparatus brought into play in an attempt to demonstrate the reasonableness of what he already knew—even though he has been privileged to witness only a small part of the trial and error going into the construction of Model IV. The rejoinder to such a possible criticism would surely be that if a hypothesis is worth considering at all, it should be worthwhile to do some hard work to estimate its significance. On the purely conceptual level, it may suffice to recognize peer-group influence on aspiration as an actual process and to reason from that in a qualitative way to some of its consequences. Ultimately, however, we shall want to know, of the factors and processes that operate in the real world, which ones do how much of the work. Constructions like those exhibited in this paper not only offer one approach to the rendering of relevant estimates, but also present interpretations in such a form that their weaknesses—and those of the theories giving rise to them—are fairly evident. If the results of more of our research could be cast into this form, we would begin to understand better how much we do and do not know.

Chapter 14

THE CHOICE OF INSTRUMENTAL VARIABLES IN THE ESTIMATION OF ECONOMY-WIDE ECONOMETRIC MODELS

Franklin M. Fisher[1]
Massachusetts Institute of Technology

1. Introduction

This paper is concerned with an important class of problems encountered in the estimation of economy-wide econometric models. The essential characteristics of such models for our purposes are three. They are dynamic, including lagged endogenous variables as essential parts of the system. They are large and nearly self-contained so that they include relatively few truly exogenous variables. Finally, they are essentially interdependent in that their dynamic structure is indecomposable.

Because an economy-wide model tends to be large, it is frequently impossible to estimate its reduced form by unrestricted least squares, since the number of exogenous and lagged endogenous variables is greater than the number of available observations. In addition, as will be brought out below,

Reprinted by permission of the author and publisher from the *International Economic Review*, Vol. 6, pp. 245–274. Copyright 1965, Kansai Economic Federation.

[1] This paper which forms part of a longer contribution to *The Brookings-SSRC Quarterly Econometric Model of the United States*, J. Duesenberry, G. Fromm, E. Kuh, and L. R. Klein, eds., to be published by North-Holland Publishing Co., 1965, a volume describing the Social Science Research Council model of the United States economy, was largely written during my tenure of a National Science Foundation Postdoctoral Fellowship at the Econometric Institute of the Netherlands School of Economics. I am indebted to T. J. Rothenberg for helpful conversations and to L. R. Klein and E. Kuh for criticism of earlier drafts, but retain responsibility for any errors.