Open access • Journal Article • DOI:10.1086/225554

Recovering Individual Data in the Presence of Group and Individual Effects

Gudmund R. Iversen

Published on: 01 Sep 1973 - American Journal of Sociology (University of Chicago Press)

Topics: Ecological fallacy and Grouped data

9-1-1973

Recovering Individual Data In The Presence Of Group And Individual Effects

Gudmund R. Iversen
Swarthmore College, iversen@swarthmore.edu

Recommended Citation
Gudmund R. Iversen. (1973). "Recovering Individual Data In The Presence Of Group And Individual Effects". *American Journal Of Sociology*. Volume 79, Issue 2. 420-434. DOI: 10.1086/225554
https://works.swarthmore.edu/fac-math-stat/214

Recovering Individual Data in the Presence of Group and Individual Effects
Author(s): Gudmund R. Iversen
Source: *American Journal of Sociology*, Vol. 79, No. 2 (Sep., 1973), pp. 420-434
Published by: The University of Chicago Press
Stable URL: http://www.jstor.org/stable/2776466
Accessed: 07-02-2018 18:15 UTC

Recovering Individual Data in the Presence of Group and Individual Effects[1]

Gudmund R. Iversen
Swarthmore College

The ecological fallacy of relating variables on the group level, when the individual-level relationship is desired, can only be avoided by using individual-level data. This paper gives some conditions for occasions when individual-level data can successfully be recovered from grouped data. Such a recovery is illustrated using data on urban or rural residence and participation or not in the labor force as an example. The conditions are given in terms of the distinction between individual- and group-level effects of one variable on another. Recovering individual data, on the one hand, and the study of individual- and group-level effects, on the other hand, represent two separate areas of thought that have received considerable attention. Here a link is made between the two lines of development to facilitate the recovery of individual-level data. Some consequences of the models for research design and recovery of historical data are explored.

INTRODUCTION

Since Robinson's (1950) now famous attack on the indiscriminate use of ecological correlations when individual-level data are not available, sociologists have given considerable attention to the problem of estimating individual-level relationships from grouped data. In the same period, sociologists have also been concerned with the distinction between individual-level and group-level effects. In spite of the interest in these two flourishing lines of related thought, few attempts have been made to link the two together. One example of such a link is the paper by Goodman (1959), the richness of which does not seem to have been fully appreciated in the literature. Another and more explicit link is made in this paper. In particular this paper shows how a classification of individual-level and group-level effects helps to clarify the problem of estimating individual relationships from grouped data. Most of the resulting models are special cases of those presented by Goodman (1959).

As an example of the problem pursued here consider table 1. The table classifies the adult U.S. population (in thousands) by the 1950 census with respect to whether people lived in an urban or a rural area and whether or not they were in the labor force. In addition to this table, there are similar tables (not shown here) available for each of the 48 states. The joint

[1] Valuable comments on an earlier draft were provided by James A. Davis, Otis Dudley Duncan, Leo A. Goodman, Leslie Kish, Frederick Mosteller, and Donald E. Stokes.

TABLE 1

PEOPLE CLASSIFIED BY URBAN OR RURAL RESIDENCE AND PARTICIPATION IN LABOR FORCE WITH THE JOINT CLASSIFICATION MISSING: NATIONAL 1950 U.S. CENSUS (IN THOUSANDS)

Residence	In the Labor Force	Not in the Labor Force	Total
Urban	...	...	73,589
Rural	...	...	37,212
Total	59,314	51,487	110,801

distribution of the two variables that define this 2×2-contingency table is, for the moment, not known. Information is only available for the two marginal distributions. Without the joint distribution, it is not possible to study the relationship between the two variables on the level of the individual. If people migrate from rural to urban areas because it would then be possible to join the labor force, there should be a correlation between the two variables in table 1. If there is such a correlation, one could perhaps lessen the migration by increasing the rural labor force and thereby lessen some of the many pressures in our urban areas. But without the cell entries in table 1 one cannot discover whether the two variables are correlated.

Most of the discussion here is limited to the study of the effects of one variable X (residence) on a variable Y (labor force), where both variables are dichotomies. The data, in the case of K ($=48$) groups, can be arranged in K contingency tables, one for each state, in addition to the table for the country as a whole. The relationship between X and Y on the individual level is seen from the interior cell entries of the tables, and the relationship on the group level is seen from the margins. It was Robinson (1950) who pointed out to the sociological profession the danger of correlating the means of the marginal distributions across the K tables when one really wants to study the relationship between X and Y on the level of the individual. Correlating the marginal means gives the ecological, and not the individual, correlation.

The sociological literature, as exemplified by Davis, Spaeth, and Huson (1961), points out that the variable Y may be determined by two aspects of the variable X. First, the individual's own score on X influences his score on Y. This is the effect of X on Y on the level of the individual. Second, by belonging to a group, the level of X in the group influences his score on Y. This is the effect of X on the group level. In addition to individual and group effect, there may exist an effect which can be seen as an interaction effect of the two. Davis et al. introduce a classification scheme that enables one to decide whether such individual, group, and

interaction effects are present in a set of data when group- and individual-level data are available. This classification scheme is used below.

But complete data on both the group and individual level are not always available. In sociological research, univariate distributions are more common than bivariate distributions. Thus, we are often left with only the margins and not the cell entries in a set of tables. Even though, in general, it is impossible to recover the cell entries from the margins, Miller (1952), Goodman (1953*a*, 1953*b*, 1959), Madansky (1959), Telser (1963), and Lee, Judge, and Zellner (1968), among others, have shown that under certain circumstances it is possible to recover the cell entries. One may want to estimate the cell entries because they are of interest in their own right. More commonly the cell entries are needed in order to find the individual correlation between the two variables.

However, the methods of recovery have not fully benefited from the distinction between individual and group effects. Insights obtained from such a distinction between these types of effects can often clarify, in critical respects, the problem of estimating the missing cell entries. This paper examines the estimation of missing cell entries in the light of the notion of individual and group effects. It also suggests that for some models it may be possible to make use of additional, and incomplete, data to improve the estimates.

Many statistical estimation problems result from the fact that not all the necessary information is available, since only a sample of observations is taken. Such problems can be solved with more observations. Other statistical problems occur because the measurements contain errors, and these problems can be solved by better measuring devices. The estimation problem discussed here may contain sampling and measurement errors, but we tend to ignore these because our problem is solved neither by more nor better observations. Instead, the problem would have been solved if we had the right kind of data. The parameters in our models below are further removed from the observed data than is usually the case, and it is this increased latency of our parameters that makes the estimation more complicated.

INDIVIDUAL AND GROUP EFFECTS

The concepts of individual and group effects for 2×2 tables are formally introduced in this section, with the notation being developed as it is needed. The discussion is limited to 2×2 tables, even though some generalizations to larger tables are possible.

In a set of K contingency tables of size 2×2 let the kth table have marginal proportions (p) and conditional row proportions (r) for the two variables X and Y as shown in table 2. The proportions in table 2, in our

TABLE 2

MARGINAL (p) AND CONDITIONAL ROW PROPORTIONS (r) FOR THE kTH 2×2-CONTINGENCY TABLE, $k = 1, \ldots, K$

	Y		
	Labor Force	Not Labor Force	Total
X:			
Urban	r_{11k}	$1 - r_{11k}$	$p_{1\cdot k}$
Rural	r_{21k}	$1 - r_{21k}$	$1 - p_{1\cdot k}$
Total	$p_{\cdot 1k}$	$1 - p_{\cdot 1k}$	1

case with the p's known and the r's unknown, can be taken to be generated by some stochastic process with probabilities as shown in table 3.

TABLE 3

MARGINAL (π) AND CONDITIONAL ROW PROBABILITIES (ρ) FOR THE kTH TABLE

	Y		
	Labor Force	Not Labor Force	Total
X:			
Urban	ρ_{11k}	$1 - \rho_{11k}$	$\pi_{1\cdot k}$
Rural	ρ_{21k}	$1 - \rho_{21k}$	$1 - \pi_{1\cdot k}$
Total	$\pi_{\cdot 1k}$	$1 - \pi_{\cdot 1k}$	1

Davis et al. (1961) consider, in this notation, the relationships of the two conditional row probabilities ρ_{11} and ρ_{21} to the marginal probability $\pi_{1\cdot}$. The subscript k is dropped in order to consider the general relationships between the conditional and marginal probabilities.

Figure 1 shows how these probabilities are related in the case of individual effect only. The group composition, as measured by $\pi_{1\cdot k}$ in the kth table, has no effect on ρ_{11} and ρ_{21}, since the conditional probabilities are the same for all tables. This would be the case if the probability of belonging in the labor force depended only on residence and not on the proportion of urban population in the particular state.

The relationship in figure 1 can be expressed:

$$\rho_{11k} = a + (0)\pi_{1\cdot k} \quad \text{and} \quad \rho_{21k} = c + (0)\pi_{1\cdot k}. \tag{1}$$

Thus, the case of individual effects only is represented as linear relationships with zero slopes and different intercepts for ρ_{11} and ρ_{21}. The difference $\rho_{11} - \rho_{21}$ can be taken as a measure of the degree of individual

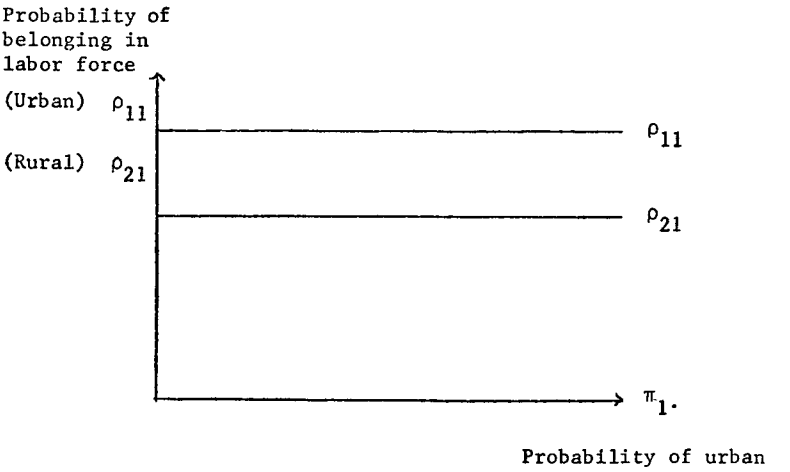

FIG. 1.—Individual effect only

effects; the further away from zero the difference, the larger the individual effect.

The case of group effect only is shown in figure 2. The lines for the two

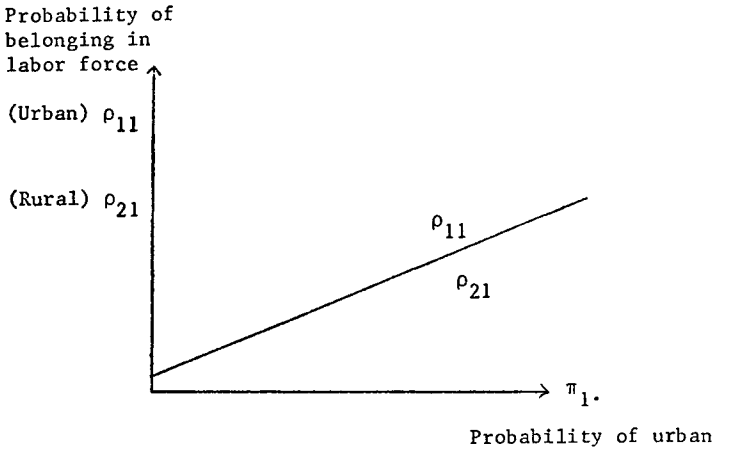

FIG. 2.—Group effect only

conditional row probabilities coincide. Mathematically, this can be expressed in the equations

$$\rho_{11k} = a + b\pi_{1\cdot k} \quad \text{and} \quad \rho_{21k} = a + b\pi_{1\cdot k}. \tag{2}$$

For given model parameters a and b, $\rho_{11k} = \rho_{21k}$, and the actual value of the conditional probabilities is determined by the group composition $\pi_{1\cdot k}$.

Here the probability of a person belonging to the labor force is the same whether the person has an urban or a rural residence. The probability depends only on the proportion of the population in the state that is urban. This means we have the case of group effects only when the two characteristics—residence and membership in the labor force—are independent.

Equations (1) and (2) above are examples of the more general case where ρ_{11} and ρ_{21} are assumed to depend linearly on $\pi_{1\cdot}$ as expressed in the equations

$$\rho_{11k} = a + b\pi_{1\cdot k} \quad \text{and} \quad \rho_{21k} = c + d\pi_{1\cdot k}. \tag{3}$$

The case $a \neq c$ and $b = d = 0$ gives individual effect only, while $a = c$ and $b = d$ gives group effect only. When $a \neq c$ and $b = d$, we have both individual and group effects but no interaction between the two effects. Interaction is present when $b \neq d$. Implications of the three different cases for the estimation of cell entries from margins are discussed below. These cases are contained in Goodman (1959), as well as in Boudon (1963) and other places.

ESTIMATION OF CONSTANT CONDITIONAL PROBABILITIES

The conditional row proportions and the marginal proportions in the kth table are related according to the basic row equation

$$p_{\cdot 1k} = r_{11k}p_{1\cdot k} + r_{21k}(1 - p_{1\cdot k}). \tag{4}$$

Miller (1952) seems to be the first to have suggested that if the r's do not vary much from table to table, equation (4) can be used to estimate the unknown row proportions.

Small variations in the r's across tables can be treated as the case of the presence of only individual effect. If the sociological problem under study is one in which there are substantive reasons to believe that only the individual effect is present, it is possible to estimate the missing cell entries.

The important condition for the estimation to give valid results is that the unknown r's do not vary much from table to table. Assuming only individual effect, as expressed in equation (1), we can write

$$r_{11k} = a + e_{1k} \quad \text{and} \quad r_{21k} = c + e_{2k}. \tag{5}$$

This substituted in equation (4) results in the equation

$$p_{\cdot 1k} = ap_{1\cdot k} + c(1 - p_{1\cdot k}) + e_k, \tag{6}$$

where, for the residual term, $e_k = e_{1k}p_{1\cdot k} + e_{2k}(1 - {}_{1\cdot k})$. By minimizing the sum Σe_k^2 with respect to a and c, we get the least-squares estimates $\hat{a}$ and $\hat{c}$. These estimates are used to estimate the conditional row proportions, that is, $\hat{r}_{11k} = \hat{a}$ and $\hat{r}_{21k} = \hat{c}$. Since the residual term depends upon the

independent variables, a weighted-estimated procedure has been suggested by Madansky (1959). As shown by Iversen (1969), the differences $|a - \hat{a}|$ and $|c - \hat{c}|$ depend on the values of e_{1k} and e_{2k}. If all the terms e_{1k} and e_{2k} ($k = 1, 2, \ldots, K$) are small, one is guaranteed that the estimates $\hat{a}$ and $\hat{c}$ are close to a and c. This, again, means that the corresponding estimated cell entries are close to the true cell entries.

Essential for this estimating procedure is the assumption of the presence of only individual effect. This assumption should ideally follow from the sociological theory underlying the research being performed. One sign of the assumptions being inappropriate is that the estimated proportions fall outside the admissible range from zero to one. To force the estimates to be in the admissible range by methods like quadratic programming, as done by Irwin and Meeter (1969), is to neglect strong evidence in the data that it is the model and not the original estimation that is wrong. If the true values are close to zero or one, sampling variations could give estimates outside the range, but one should have good substantive reasons for accepting estimates near zero or one. Partial checks on whether the assumption of only individual effect is present have been developed by Goodman (1959) and include that there should be a linear relationship between $p_{1\cdot}$ and $p_{\cdot 1}$, the estimated $p_{\cdot 1}$ should be close to the observed $p_{\cdot 1}$, and the estimated correlation on the individual level should lie between the bounds discussed by Duncan and Davis (1953).

ZERO-, FIRST-, AND SECOND-ORDER MARGINAL RELATIONSHIPS

Having discussed the case of individual effect only above, we consider here various configurations of group and individual effects and how they influence the estimation of missing cell entries.

Similar to the relationship between the conditional and marginal proportions in equation (4), the conditional and marginal probabilities are related as shown in the equation

$$\pi_{\cdot 1k} = \rho_{11k}\pi_{1\cdot k} + \rho_{21k}(1 - \pi_{1\cdot k}). \tag{7}$$

If the conditional probabilities ρ_{11k} and ρ_{21k} are taken to depend linearly on the marginal probability $\pi_{1\cdot k}$, as expressed in equation (3), we can substitute from equation (3) into equation (7). That substitution results in the equation

$$\pi_{\cdot 1k} = c + (a - c + d)\pi_{1\cdot k} + (b - d)\pi_{1\cdot k}^2, \tag{8}$$

which specifies the relationship between the marginal row and column probabilities in terms of the model parameters a, b, c, and d.

Equation (8) gives the relationship between the marginal probabilities in the face of the full configuration of individual and group effects as well

as interaction between the two effects. The relationship is of second order and contains four parameters. Boudon (1963) arrives at the same relationship, even though, due to what must be some minor error, his parameter corresponding to d above is missing in the factor for π_1. Goodman (1959) discusses several aspects of equation (8), mainly in terms of an example relating illiteracy rates and a black-white dichotomy. A more general version of this equation is given by Goodman (1959, p. 624), and he gets this equation by setting $z = x$, in his notation.

Equation (8) contains the four parameters a, b, c, and d, and, using the observed margins, we will only be able to estimate the three linear combinations c, $a - c + d$, and $b - d$. In order to estimate all four parameters, we therefore are going to need information beyond what is contained in the margins. One such type of information is whatever substantive sociological theory we have that can guide us as to the presence of individual, group, and interaction effects. Another type of information consists of data on the cell entries for one of the K tables or data on the cell entries for what can be called the sum table. As an example of this last kind of data, we may have a sample survey giving cell entries for the whole country, while tables for each of the 50 states give only the marginal proportions.

Before returning to the case where additional data are available, equation (8) is examined in some detail with respect to individual, group, and interaction effects. Several different cases are discussed below.

Case 0.—In this case, $a = c$, $b = d = 0$. This is the case of no individual or group effect. Equation (8) reduces to

$$\pi_{\cdot 1k} = a, \tag{9}$$

which is a zero-order relationship between the margins. This is the case where the two conditional row probabilities in each table have the same value, and this value does not differ from table to table. Since $\rho_{11k} = \rho_{21k}$, there is independence between the two characteristics in each table. The observed points $(p_{1\cdot k}, p_{\cdot 1k})$ should, according to equation (9), scatter around a line with intercept a and slope zero. If we have such a scatterplot and can, in addition, specify the absence of both individual and group effects, equation (9) can be used for estimation of the missing cell entries. This case is also obtained from Goodman (1959, pp. 623–24) by setting, in his notation, $z = x$ and $B = F = 0$, as well as $C = G$ and $F = H = 0$.

Case 1A.—Here, $b = d = 0$. Such a case corresponds to the illustration in figure 1, where the conditional row probabilities do not vary from table to table. With these restrictions, equation (8) becomes

$$\pi_{\cdot 1k} = c + (a - c)\pi_{1\cdot k}, \tag{10}$$

which gives a first-order relationship between the margins. Here there is an individual effect but no group effect present. This is the case that is

more extensively discussed above. Equation (10) forms the basis for much of the work done by earlier authors in this field. The observed points ($p_{1\cdot k}$, $p_{\cdot 1k}$) lie scattered around a line with intercept c and slope $a - c$. Because a and c are restricted to the interval from zero to one, the line intersects the east and west and not the north and south sides of the unit square. With such a scatterplot and substantive reasons for believing that individual, but not group, effect is present, equation (10) can be used for the estimation of the missing cell entries. Goodman (1959, pp. 623–24) gets this case, in his notation by setting $z = x$ and $F = 0$, as well as $F = H = 0$. Least-squares estimation of constant conditional probabilities has also been considered by Madansky (1959), Telser (1963), and others. Lee et al. (1968) have considered maximum likelihood and Bayesian estimation as well.

Case 1B.—In this case, $a = c$ and $b = d$. Here we have no individual, only group, effect present, as shown in figure 2. Such restrictions on the model parameters reduce equation (8) to

$$\pi_{\cdot 1k} = a + b\pi_{1\cdot k}. \tag{11}$$

As in case 1A there is a first-order (linear) relationship between the marginal probabilities. In order to identify this case we therefore have to be able to justify, on substantive grounds, the presence of a group, and not an individual, effect. In Goodman's example, this occurs if "the average difference between the illiteracy rate for Negroes and the rate for whites is zero in states having the same proportion x of Negroes" (1959, p. 623). He discussed this as the case where $B = 0$, in his notation.

Case 1C.—Here, $b = d$, which corresponds to the presence of individual and group effects, but without any interaction between the two effects. Equation (8) becomes

$$\pi_{\cdot 1k} = c + (a + b - c)\pi_{1\cdot k}, \tag{12}$$

which is the third example of a first-order relationship between the marginal probabilities. This equation is also contained in Goodman (1959, p. 623). As an additional complication, there are three parameters to estimate in this case, and we cannot hope to estimate more than the two quantities c and $a + b$. If one therefore decides on substantive grounds that there are individual and group effects without interaction, additional data from survey work or other sources are necessary in this case in order to obtain estimates of the cell entries.

Three cases have been presented above where there is a linear relationship between the marginal probabilities. Data showing a linear trend in the scatterplot of the points ($p_{1\cdot k}$, $p_{\cdot 1k}$) could come from any of the three cases, and there is therefore no simple way of estimating the missing cell entries. Considerable attention has to be given to the question of whether on sub-

stantive grounds one can decide on the presence of individual and group effects. The more usual case would be that both effects are present, and additional data therefore become a necessity.

Case 2.—The full, second-order relationship between the marginal probabilities has been introduced in equation (8). When the points $(p_{1\cdot k}, p_{\cdot 1k})$ show a nonlinear trend that can be accounted for by the second-order term $\pi_1.^2$, there is good reason to believe that the tables have been generated by some variant of the model in equation (3). It should be kept in mind, however, that a nonlinear marginal relationship can be obtained from other models as well.

Here a second-order marginal relationship is accounted for by the presence of interaction between the individual and group effects. Regressing the proportion in the labor force on the proportion urban and the square of the proportion urban results in the equation

$$\hat{p}_{\cdot 1} = 0.54 - 0.13p_{1\cdot} + 0.16p_{1\cdot}^2, \qquad R = 0.32 \; s_e = 0.03. \quad (13)$$

Since the two regression slopes are about of the same magnitude and have opposite signs, the proportion in the labor force does not vary much from state to state. The coefficients suggest the following estimates:

$$\begin{aligned} \hat{c} &= 0.54, \\ \widehat{a - c + d} &= -0.13, \\ \widehat{b - d} &= 0.16. \end{aligned} \quad (14)$$

In this model $a - c$ is a measure of the individual-level effect; b and d, measures of the group effect; and $b - d$, measure of the interaction effect on the participation in the labor force. The estimated coefficients indicate a slight interaction effect, which would imply the presence of both an individual- and group-level effect from residence.

Estimating the four parameters, and thereby the cell entries, from the three equations in equation (14) is not possible without additional data. Such additional data can occur in many ways, and as an example we consider the case when the proportion r_{11k} is available for the kth table. That means we know the proportion in the labor force among the urban residents for one state. Such a proportion could be available from a sample survey, the state employment agency, etc. Using least-squares methods we can get the estimates $\hat{c}$, $(\widehat{a - c + d})$, and $(\widehat{b - d})$. Adding the three estimates, we get the estimate $(\widehat{a + b}) = 0.57$. From equation (3) we have $r_{11k} = \hat{a} + \hat{b}p_{1\cdot k}$. We assume that $(\widehat{a + b}) = \hat{a} + \hat{b}$, where we know the left side but not the two separate components $\hat{a}$ and $\hat{b}$. Solving the last two equations in $\hat{a}$ and $\hat{b}$, we get

$$\hat{a} = \frac{r_{11k} - (\widehat{a + b})p_{1\cdot k}}{p_{1\cdot k}} \quad \text{and} \quad \hat{b} = (\widehat{a + b}) - \hat{a}. \quad (15)$$

That gives $\hat{d} = \hat{b} - \widehat{(b - d)}$, and we have thereby been able to estimate all four model parameters.

With these estimates, we can estimate the row proportions in all the remaining tables according to

$$\hat{p}_{11i} = \hat{a} + \hat{b}p_{1\cdot i} \quad \text{and} \quad \hat{p}_{21i} = \hat{c} + \hat{d}p_{1\cdot i}. \tag{16}$$

In this case we have

$$\hat{a} = \frac{r_{11k} - 0.57p_{1\cdot k}}{1 - p_{1\cdot k}}, \qquad \hat{b} = 0.57 - \hat{a} \tag{17}$$
$$\hat{d} = \hat{b} - 0.16.$$

Suppose we knew that in a particular state $r_{11} = 0.54$ and $p_{1\cdot} = 0.68$. That results in

$$\hat{r}_{11} = 0.48 + 0.09p_{1\cdot} \quad \text{and} \quad \hat{r}_{21} = 0.54 - 0.07p_{1\cdot}. \tag{18}$$

Since there is only one degree of freedom in each table, one need only estimate one of the four missing cell entries. Using $\hat{r}_{11}$ from equation (18) to estimate the cell entries in the 48 tables leads to an estimated frequency of 39,999,200 people who are in the labor force and have an urban residence. The Bureau of the Census has published these cell entries, and the reported number equals 40,674,000 people. The difference between the true and the estimated frequencies is very small in this case. The difference depends, in a crucial way, on the values of r_{11} and $p_{1\cdot}$ in the kth group used as the missing piece of information in order to estimate a, b, c, and d.

It may also be possible to discover the past with methods like the one outlined above. Historical data are abundantly available from sources like past censuses and elections, but many of those data only permit analysis on the group, or ecological, level, since the proper individual data were not observed. For example, relating the votes in pairs of elections we have only the marginal proportions. But in recent years there has been a growth of sample surveys that do give the cell entries as well as the margins. Such cell entries can give us the r_{11k} needed in equation (13) above to estimate the cell entries in the remaining tables that refer to past elections.

The use of additional data, as outlined above, also has implications for the design of new research. Assume that data on the margins are available for a series of units, say states, and cell entries are needed for the whole country. A survey may be conducted across the whole country in order to obtain estimates of the country cell entries. The method above suggests the possibilities of surveying only a single state, with a considerable saving in time and money, and then combining such survey results with the existing information available about the margins.

TRANSPOSED TABLES

By arranging the variables as shown in table 1, the discussion here has been developed in terms of conditional row probabilities. It may, however, be possible to estimate the entries using column instead of row probabilities. That amounts to transposing all the tables.

The column probabilities may not be very meaningful in terms of the substantive content of the variables X and Y, but that is not a sufficient reason for not considering using these probabilities for the estimation. The structure of the tables may be such that these probabilities can be successfully used for the recovery of the missing cell entries. In general, the recovery is possible when there exists some simple relationship between the conditional column probabilities and the corresponding marginal probabilities.

OTHER MODELS AND LARGER TABLES

It may be that the relationships between the conditional probabilities and the corresponding marginal probabilities are not linear. Various functional forms can be specified, and one possible model would consist of setting

$$\rho_{ijk} = a + b\pi_{i\cdot k} + c\pi_{i\cdot k}^2. \tag{19}$$

No further attention is given to such a model.

Turning to tables with more than two rows and two columns complicates matters considerably. With R rows there are $R - 1$ degrees of freedom from the marginal proportions, and we therefore ought to consider a model of the form

$$\rho_{ijk} = b_0 + b_1\pi_{1\cdot k} + b_2\pi_{2\cdot k} + \ldots + b_{R-1}\pi_{(R-1)\cdot k} \tag{20}$$

for the conditional probability in the (i, j) cell. The presence of only an individual effect is characterized by $b_1 = b_2 = \ldots = b_{R-1} = 0$, that is, constant probabilities across all the tables. An array of group-effect patterns is possible with some b's equal to zero and others not. But it is unlikely that we have any substantive theory that can identify what parameters can be set equal to zero.

The relationships between the marginal probabilities for a set of tables of size $R \times C$ are obtained from the C equations of the type

$$\pi_{\cdot jk} = \rho_{ij}\pi_{1\cdot k} + \ldots + \rho_{Rj}\pi_{R\cdot k} \qquad (j = 1, \ldots, c) \tag{21}$$

by substitution from equation (20). With the full model in equation (20), the resulting marginal relationship contains many more parameters than can be estimated. But we get terms of the type $\pi_{i\cdot k}$, $\pi_{i\cdot k}^2$, and $\pi_{m\cdot k}\pi_{n\cdot k}$ $(m \neq n)$, so that with R rows we can estimate as many as $R(R + 1)/2$

parameters. With 2×2 tables we can estimate three parameters, as seen above. With 3×3 tables, we can estimate six parameters in each equation. That means that each of the three conditional probabilities in a particular column can be written as a function of two parameters. For instance, we can have

$$\begin{aligned} \rho_{11k} &= a + b\pi_{1\cdot k}, \\ \rho_{21k} &= c + d\pi_{2\cdot k}, \\ \rho_{31k} &= e + f(1 - \pi_{1\cdot k} - \pi_{2\cdot k}), \end{aligned} \tag{22}$$

and all these six model parameters a, b, c, d, e, and f can be estimated using equation (21). Thus, larger tables offer more opportunities for model building, even though it becomes even more difficult to distinguish between competing models.

CONCLUSIONS

This paper attempts to bring together two fairly divergent developments. On the one hand, sociologists have been concerned with the distinction between effects on the individual level and effects on the group level. The other concern has been with how to estimate individual-level data from group-level data and thereby avoid the ecological fallacy of relating variables on the group level when what is desired is the relationship between the variables on the individual level.

By borrowing the notions of group- and individual-level effects we can gain a better understanding of the issues involved in estimating the missing cell entries. The models arrived at here are not new; indeed, they can be seen as special cases of models proposed by Goodman (1959), but if we derive the models explicitly from group- and individual-level effects as I have defined these effects, they can possibly be used with greater success. The problem of estimating missing cell entries is still beset by difficulties, but I have presented some circumstances under which it is possible to estimate the cell entries.

It cannot be determined from the margins alone whether the conditions for the models discussed here are satisfied, since the conditions are expressed in terms of the missing cell entries. Care must therefore be exercised in the use of these and other models. Severe biases may appear in the estimates if the wrong model is used, and these biases are not always easily detected. Because of the high risk of bias, we can place less emphasis on trying to construct estimates that have small sampling errors and other desirable properties. Instead, the effort should be spent on trying to reduce possible bias. For example, if a model leads to inadmissible estimates, for instance, if the estimate of a proportion falls outside the range from zero to one, it is

better to change the model than to force the estimating process through quadratic programming to give admissible estimates.

All the work that has been done on the problem of estimating missing cell entries has resulted in providing cell entries that add up correctly to the given margins. The margins provide restrictions on the range of possible cell entries. Additional data in the form of known cell entries in some partial way provide further restrictions on the range of possible cell entries. We construct models that will give estimates that are consistent with the available information. The more information that is available, the tighter the restrictions that can be specified, and therefore the closer the estimates will be to the true values.

We have seen here that an important source of additional information is contained in the distinction between individual- and group-level effects. It may be possible to bring in substantive considerations to help decide whether group- or individual-level effects are present. Based on such considerations, it may be possible to specify the values of some of the parameters in the model and to use the available group data to estimate the remaining parameters.

We have also seen that the parameters can be estimated if cell entries are available in at least one of the tables. This opens up some unexplored possibilities, and it was pointed out that certain research design implications follow from this.

Bayesian statistics may have something to offer for the estimation of missing cell entries. Because of the latency of the parameters we want to estimate, any past knowledge of the parameters we can bring to the analysis should be included. Bayesian statistics seems ideally suited for this purpose. This may be a case where the prior distribution could contribute significantly to the determination of the posterior distribution of the parameters and the resulting cell entries.

Statistical theory can only present necessary conditions for when a particular model holds. The sufficient conditions will have to come from the substantive theory underlying the variables X and Y. With the interplay of these two sources of information, the latency of the parameters can possibly be overcome and the missing cell entries successfully recovered.

REFERENCES

Boudon, Reymond. 1963. "Propriétés individuelles et propriétés collectives: un problème d'analyse écologique." *Revue française de sociologie* 4:275–99.

Davis, J. A., J. L. Spaeth, and C. Huson. 1961. "A Technique for Analyzing the Effects of Group Composition." *American Sociological Review* 26, no. 2 (April): 215–25.

Duncan, O. Dudley, and Beverly Davis. 1953. "An Alternative to Ecological Correlation." *American Sociological Review* 18, no. 6 (December): 665–66.

Goodman, Leo A. 1953*a*. "A Further Note on Miller's 'Finite Markov Processes in Psychology.'" *Psychometrika* 18, no. 3 (September): 245–48.

———. 1953*b*. "Ecological Regressions and the Behavior of Individuals." *American Sociological Review* 18, no. 6 (December): 663–64.

———. 1959. "Some Alternatives to Ecological Correlation." *American Journal of Sociology* 64, no. 6 (May): 610–25.

Irwin, G. A., and Duane A. Meeter. 1969. "Building Voter Transition Models from Aggregate Data." *Midwest Journal of Political Science* 13, no. 4 (November): 545–66.

Iversen, Gudmund R. 1969. "Estimation of Cell Entries in Contingency Tables When Only Margins Are Observed." Ph.D. dissertation, Department of Statistics, Harvard University.

Lee, T. C., G. G. Judge, and Arnold Zellner. 1968. "Maximum Likelihood and Bayesian Estimation of Transition Probabilities." *Journal of the American Statistical Association* 63, no. 324 (December): 1162–79.

Madansky, A. 1959. "Least Squares Estimation in Finite Markov Processes." *Psychometrika* 24, no. 2 (June): 137–44.

Miller, George A. 1952. "Finite Markov Processes in Psychology." *Psychometrika* 17, no. 2 (June): 49–167.

Robinson, W. S. 1950. "Ecological Correlations and the Behavior of Individuals." *American Sociological Review* 15, no. 3 (June): 351–57.

Telser, L. G. 1963. "Least Squares Estimation of Transition Probabilities." In *Measurement in Economics,* edited by C. Christ. Stanford, Calif.: University Press.