A Modified Multiple Regression Approach to the Analysis of Dichotomous Variables

Leo A. Goodman

American Sociological Review, Vol. 37, No. 1 (Feb., 1972), 28-46.

Stable URL:
http://links.jstor.org/sici?sici=0003-1224%28197202%2937%3A1%3C28%3AAMMRAT%3E2.0.CO%3B2-J

American Sociological Review is currently published by American Sociological Association.

A MODIFIED MULTIPLE REGRESSION APPROACH TO THE ANALYSIS OF DICHOTOMOUS VARIABLES *

LEO A. GOODMAN

University of Chicago

American Sociological Review 1972, Vol. 37 (February):28–46

To illustrate the models and methods of the present article, we shall reanalyze those data in the famous study of The American Soldier *by Stouffer et al. (1949), subsequently analyzed by Coleman (1964), Zeisel (1968), and Theil (1970). The methods we present reveal how the odds pertaining to a given dichotomized variable (e.g., the odds that a soldier would prefer a Northern to a Southern Camp assignment) are related to other dichotomized variables (e.g., (a) the soldier's race, (b) his region of origin, (c) his present camp location). The usual regression analysis methods do not suit the case considered here, where the dependent variable is the odds pertaining to a given dichotomous variable. Nor do the usual methods suit the case where the dependent variable is a proportion pertaining to the dichotomous variable. This article presents some relatively elementary models and methods suitable for analyzing the odds (or a proportion) pertaining to the given dichotomous dependent variable. Applying these models and methods to the data referred to above, new insights are obtained.*

LET us begin by describing the data that will be analyzed here for illustrative purposes. These data, which are based on the earlier data first presented by Stouffer et al. (1949), appear in Table 1 below. This four-way table cross-classifies soldiers by the following four dichotomous variables: (A) race (Negro or white); (B) region of origin (North or South); (C) location of present camp (North or South); and (D) preference as to camp location (North or South).

Table 1 shows that for say, a Negro Northerner in a Northern camp the odds are 387 to 36 that he will prefer a Northern camp. This table also shows that for, say, a white Southerner in a Southern camp, the odds are 91 to 869 that he will prefer a Northern camp. In the next section, we present a model that describes quantitatively how these and the other odds in Table 1 are affected by (A) race, (B) region of origin, and (C) present camp location. We show how to test whether the model fits the data, and we measure how well the model fits using an index that is analogous to the usual multiple correlation coefficient of regression analysis. We also show how to assess the statistical significance of the contribution made by certain parameters in the model, and we measure the contribution's magnitude with indices that are analogous to the usual partial and multiple-partial correlation coefficients of regression analysis.

With the model that will be described in the next section, and with the more general model that follows it, we can estimate how the odds for preferring a Northern camp are changed by the "main effects" of race, region or origin, and present camp location, as well as by certain "interaction effects" among these variables. With each model considered here we can also estimate what the expected frequencies in Table 1 would be if the model were true. We can compare these estimated expected frequencies with the corresponding observed frequencies to determine if the model fits the data. For the model that will be described in the next section, Table 2 gives the expected frequencies estimated under the assumption that the model is true. Under this model, the estimated odds are 390.64 to 32.36 that the Negro Northerner in a Northern camp will prefer a Northern camp; and the estimated odds are 91.74 to 868.26 that the white Southerner in a Southern camp will prefer a Northern camp. When we compare the corresponding entries in Tables 1 and 2

* This research was supported in part by Research Contract No. NSF GS 2818 from the Division of the Social Sciences of the National Science Foundation. For helpful comments, the author is indebted to R. D. Bock, S. Haberman, P. F. Lazarsfeld and A. Stinchcombe.

Table 1. Cross-Classification of Soldiers With Respect to Four Dichotomized Variables: (A) Race, (B) Region of Origin, (C) Location of Present Camp, and (D) Preference as to Camp Location

Variable A Race	Variable B Region of Origin	Variable C Location of Present Camp	Variable D Number of Soldiers Preferring Camp* In North	In South
Negro	North	North	387	36
Negro	North	South	876	250
Negro	South	North	383	270
Negro	South	South	381	1712
White	North	North	955	162
White	North	South	874	510
White	South	North	104	176
White	South	South	91	869

*The numbers in this table were recalculated from the percentage table in Stouffer et al. (1949, p. 553). These numbers are consistent with the percentages in the 1949 table, but they may differ somewhat from the actual observed frequencies due to rounding of the percentages. A related percentage table was given also by Coleman (1964, p. 198) and Theil (1970, p. 104). "Preference for camp in North" includes (a) those who prefer to move to a specific camp located in the North, and (b) those whose present camp is in the North and who prefer to stay there. Similarly, "preference for camp in South" includes (a) those who prefer to move to a specific camp located in the South, and (b) those whose present camp is in the South and who prefer to stay there.

(by methods that will be described later herein), we find that the model fits the data well.

Using the expected frequencies estimated under the given model (see Table 2), we can also estimate the expected *proportion* preferring a Northern camp (as well as the *odds* referred to above), for the individuals in each row in Table 2, under the assumption that the model is true. The models considered herein, which describe how the *odds* for preferring a Northern camp are changed by certain specified "main effects" and "interaction effects," can also be used to describe how the *proportion* preferring a Northern camp is changed by these effects.

In order to understand how the dependent variable (preference as to camp location) is related to the other three variables (race, region of origin, and present camp location), we began with the four-way table (Table 1). Is it necessary to use a four-way table to describe how the dependent variable is related to the other three variables, or can this relationship be summarized adequately using the information contained in tables of smaller dimension (e.g., two-way and/or three-way tables)? The methods presented in the present article can be used to answer this question. To estimate the relationship between the dependent variable and the other three variables, the model that will be presented in the next section will actually use only the information contained in (1) the two-way table describing the relationship between the dependent variable and race, and (2) the three-way table describing the relationship between the dependent variable, region of origin and present camp location. Since that model fits the data well, we find that the relationship between the dependent variable and the other three variables can be summarized adequately using only the information contained in the particular two-way and three-way tables noted above. This topic will be discussed more fully later herein when Table 5 is presented.

We propose to analyze the data in Table 1 by methods quite different from those used in earlier analyses of these data. Our model

Table 2. Estimate of the Expected Frequencies in the Four-Way Contingency Table (Table 1), Under the Model in Which the Odds for Preferring a Northern Camp Depend on Race, Region of Origin, Location of Present Camp, and on the Interaction Between Region of Origin and Location of Present Camp

Variable A Race	Variable B Region of Origin	Variable C Location of Present Camp	Variable D Number of Soldiers Preferring Camp In North	In South
Negro	North	North	390.64	32.36
Negro	North	South	879.31	246.69
Negro	South	North	376.79	276.21
Negro	South	South	380.26	1712.74
White	North	North	951.36	165.64
White	North	South	870.69	513.31
White	South	North	110.21	169.79
White	South	South	91.74	868.26

fits the data better than Coleman's (1964) and we present a more parsimonious explanation of these data than he does. With the estimated parameters in our model, we can explain, in a more comprehensive and compact way, various interesting features of these data noted by Zeisel (1968). Some of the models considered in the present article are related to those in Theil (1970), but the methods we use are easier to apply than his. In the final section herein, we shall compare more fully ours with earlier methods.

A MODEL FOR ANALYZING THE ODDS

The symbols A, B, C, and D denote the four dichotomized variables in the four-way table (Table 1): (A) race, (B) region of origin, (C) location of present camp, and (D) preference as to camp location. For variable A, we use numbers 1 and 2 to denote Negro and white. For variables B, C, and D, we use numbers 1 and 2 to denote North and South. Each of Table 1's sixteen cells can be designated (i, j, k, l), where $i = 1$ or 2; $j = 1$ or 2; $k = 1$ or 2; $l = 1$ or 2. For example, entry 387 is in cell (1, 1, 1, 1), and is a case where variables A, B, C, D all take on value 1; entry 36 is in the cell (1, 1, 1, 2), with variable A, B, and C taking on value 1 and variable D value 2; entry 876 is in cell (1, 1, 2, 1) with variables A, B, and D taking on value 1 and variable C value 2; entry 250 is in cell (1, 1, 2, 2) with variables A and B taking on value 1 and variables C and D value 2.

Let f_{ijkl} denote the observed frequency in cell (i, j, k, l) of Table 1. For example, $f_{1111} = 387$, $f_{1112} = 36$, $f_{1121} = 876$, $f_{1122} = 250$, etc. Note that each row of Table 1 can be described by the triplet (i, j, k). For example, the first row is (1, 1, 1); the second (1, 1, 2); etc. Let n_{ijk} denote the total observed frequency in a row (i, j, k). In other words, we can write n_{ijk} as

$$n_{ijk} = f_{ijk1} + f_{ijk2}. \quad (1)$$

For example, $n_{111} = 423$, $n_{112} = 1126$, etc.

For those in row (i, j, k) the observed odds in favor of a preference for a Northern camp (i.e., the odds that variable D will take on value 1) can be written as

$$\omega_{ijk} = f_{ijk1}/f_{ijk2}. \quad (2)$$

For example, $\omega_{111} = 10.75$, $\omega_{112} = 3.50$, etc. In other words, when variables A, B, and C all take on value 1, the odds are 10.75 to 1 that variable D will take on that value. When variables A and B take on value 1 and variable C value 2, the odds are 3.50 to 1 that variable D will take on value 1.

For row (i, j, k) in Table 1, let p_{ijk} denote that row's observed proportion of observations for which variable D takes on value 1. In other words, we can write p_{ijk} as

$$p_{ijk} = f_{ijk1}/n_{ijk}. \quad (3)$$

For example, $p_{111} = .91$, $p_{112} = .78$, etc. We also let q_{ijk} denote the observed proportion

of observations in row (i, j, k) for which variable D takes on value 2. Thus,

$$q_{ijk} = f_{ijk2}/n_{ijk} = 1 - p_{ijk}. \quad (4)$$

From (2)–(4), we see that

$$\omega_{ijk} = p_{ijk}/q_{ijk}. \quad (5)$$

We can also express the p_{ijk} and q_{ijk} in terms of the observed odds ω_{ijk}:

$$p_{ijk} = \omega_{ijk}/(1+\omega_{ijk}),$$
$$q_{ijk} = 1/(1+\omega_{ijk}). \quad (6)$$

From (3) and (6) we see that the observed frequencies f_{ijkl} can be expressed in terms of the observed odds ω_{ijk} and the n_{ijk}:

$$f_{ijk1} = n_{ijk}\omega_{ijk}/(1+\omega_{ijk})$$
$$f_{ijk2} = n_{ijk}/(1+\omega_{ijk}). \quad (7)$$

Let F_{ijkl} denote the expected frequency in cell (i, j, k, l) under some specified model. For example, for the model referred to at the end of the preceding section, we see from Table 2 that F_{1111} and F_{1112} are estimated as 390.64 and 32.36, respectively. (The calculation of the entries in Table 2 will be commented upon later herein after we have presented the material in Table 5.) Letting Ω_{ijk} denote the odds based on the expected frequencies, we see that

$$\Omega_{ijk} = F_{ijk1}/F_{ijk2}. \quad (8)$$

Formula (8) corresponds to (2). In addition, corresponding to (7), we have the following:

$$F_{ijk1} = n_{ijk}\Omega_{ijk}/(1+\Omega_{ijk})$$
$$F_{ijk2} = n_{ijk}/(1+\Omega_{ijk}). \quad (9)$$

(For a related matter, see (44) later herein.) Thus, from the F_{ijkl}, we can calculate the "expected odds" Ω_{ijk}; and from the Ω_{ijk} and n_{ijk}, we can calculate the F_{ijkl}.

Our models will express the Ω_{ijk} in terms of a set of parameters that describe the "main effects" of variables A, B, and C, and certain "interaction effects" among these variables, in a way that is somewhat analogous to the corresponding effects in the usual analysis of variance model. In the present section, we shall present a particular model that fits the data (Table 1) well; and in the next section, we shall present a more general model, namely, a "saturated model" for analyzing the odds, that can help determine the various "unsaturated models" that should be examined further.[1]

Our analysis of this saturated model led us to the particular unsaturated model that we shall present now. For expository reasons we present the unsaturated model first.

Consider the following model:

$$\Omega_{ijk} = \gamma\,\gamma^{A}{}_{i}\gamma^{B}{}_{j}\gamma^{C}{}_{k}\gamma^{BC}{}_{jk} \quad (10)$$

where

$$\gamma^{A}{}_{1} = 1/\gamma^{A}{}_{2},\ \gamma^{B}{}_{1} = 1/\gamma^{B}{}_{2},\ \gamma^{C}{}_{1} = 1/\gamma^{C}{}_{2},$$
$$\gamma^{BC}{}_{11} = \gamma^{BC}{}_{22} = 1/\gamma^{BC}{}_{12} = 1/\gamma^{BC}{}_{21}. \quad (11)$$

Parameters γ, $\gamma^{A}{}_{1}$, $\gamma^{B}{}_{1}$, and $\gamma^{C}{}_{1}$ describe the "main effects" on Ω_{ijk} of the general mean[2] and variables A, B, and C, respectively; and parameter $\gamma^{BC}{}_{11}$ describes the "interaction effect" of variables B and C on Ω_{ijk}.[3]

Formula (10) describes the effects of the parameters on Ω_{ijk}, expressing Ω_{ijk} explicitly in terms of the model's parameters. These parameters can also be explicitly expressed in terms of Ω_{ijk}. From (10)–(11), we obtain the folowing expressions for the parameters in terms of Ω_{ijk}:

$$\gamma = \left[\prod_{i=1}^{2}\prod_{j=1}^{2}\prod_{k=1}^{2}\Omega_{ijk}\right]^{1/8}, \quad (12)$$

$$\gamma^{A}{}_{1} = [\Omega_{1jk}/\Omega_{2jk}]^{1/2} \text{ (for } j = 1, 2;\ k = 1, 2)$$
$$= \left[\prod_{j=1}^{2}\prod_{k=1}^{2}(\Omega_{1jk}/\Omega_{2jk})\right]^{1/8}, \quad (13)$$

$$\gamma^{B}{}_{1} = \left[\prod_{i=1}^{2}\prod_{k=1}^{2}(\Omega_{i1k}/\Omega_{i2k})\right]^{1/8}, \quad (14)$$

$$\gamma^{C}{}_{1} = \left[\prod_{i=1}^{2}\prod_{j=1}^{2}(\Omega_{ij1}/\Omega_{ij2})\right]^{1/8}, \quad (15)$$

$$\gamma^{BC}{}_{11} = [(\Omega_{i11}\Omega_{i22})/(\Omega_{i12}\Omega_{i21})]^{1/4}$$
(for i = 1, 2)
$$= \left[\prod_{i=1}^{2}[(\Omega_{i11}\Omega_{i22})/(\Omega_{i12}\Omega_{i21})]\right]^{1/8}. \quad (16)$$

[1] The saturated model, which we present in the next section, can also be described as a full model or an unrestricted model. The unsaturated models can also be described as restricted models. The various models we consider, which assume that the expected odds Ω_{ijk} are subject to certain multiplicative main and interaction effects (see, e.g., formulas (10) and (29)), are quite different from models of the kind appearing in, for example, Coleman (1964) and Boudon (1968). For further comment, see the final section of the present article.

[2] Since γ is somewhat analogous to the main effect of the general mean in the usual model for the analysis of variance (i.e., the constant term in that model), we refer to γ as the main effect of the general mean on the Ω_{ijk}. γ actually equals the *geometric* mean of the Ω_{ijk} corresponding to the eight possible values of (i, j, k) obtained when i=1, 2; j=1, 2; k=1, 2. For further details, see formula (12) below.

[3] The relationship between the model described above by (10)–(11) and the usual model for the analysis of variance will be clarified when we discuss formulas (20)–(22).

From (12), we see that γ is actually the geometric mean of the eight Ω_{ijk}. From (13), we see that $\gamma^{A}{}_{1}$ is the square-root of the odds-ratio $\Omega_{1jk}/\Omega_{2jk}$. From (12)–(16), we see that all the γ parameters can be expressed in terms of Ω_{ijk}.[4] Since Table 2 presents the estimated values of F_{ijkl} (under model (10)), we can use these to estimate first, Ω_{ijk} (see (8)) and second, the γ parameters (see (12)–(16)). Table 3 gives the estimated values of the γ parameters.

To emphasize the fact that odds Ω_{ijk} pertain to variable D, and that the γ parameters describe the main and interaction effects on these odds, we could replace the symbols Ω_{ijk}, γ, $\gamma^{A}{}_{i}$, $\gamma^{B}{}_{j}$, $\gamma^{C}{}_{k}$, $\gamma^{BC}{}_{jk}$ in (10)–(16) by $\Omega^{\bar{D}}{}_{ijk}$, $\gamma^{\bar{D}}$, $\gamma^{A\bar{D}}{}_{i}$, $\gamma^{B\bar{D}}{}_{j}$, $\gamma^{C\bar{D}}{}_{k}$, $\gamma^{BC\bar{D}}{}_{jk}$, respectively. This notation was used in Table 3 and later in Table 4, where each of the above parameters is identified by its superscript. From Table 3, we see that the estimated main effect of each variable (A, B, C) is positive (i.e., the estimates of $\gamma^{A\bar{D}}{}_{1}$, $\gamma^{B\bar{D}}{}_{1}$, $\gamma^{C\bar{D}}{}_{1}$ are all larger than 1); but the estimated interaction effect between variables B and C is negative (i.e., the estimate of $\gamma^{BC\bar{D}}{}_{11}$ is less than 1). This means, among other things, that the estimated effect on $\Omega^{\bar{D}}{}_{ijk}$ of being a Northerner in a Northern camp is less positive (due to the multiplicative factor of 0.86 pertaining to $\gamma^{BC\bar{D}}{}_{11}$) than might be surmised simply by combining the main effect of being a Northerner with the main effect of his being in a Northern camp. More precisely, after taking account of the model's various main effects, we must multiply the estimate of $\Omega^{\bar{D}}{}_{ijk}$ by the factor 0.86 for a Northerner located in a Northern camp, to account for the interaction effect $\gamma^{BC\bar{D}}{}_{11}$ between variables B and C (i.e., the effect on $\Omega^{\bar{D}}{}_{ijk}$ of the interaction between region of origin and present camp location).[5] By applying the numerical values of Table 3 to formula (10), we see, for example, that

Table 3. Estimate of the Main Effects and Interaction Effects of the Three Variables (A,B,C) on the Odds Ω_{ijk} Pertaining to Variable D in the Four-Way Contingency Table (Table 1), Under Models (10) and (20)

Variable	γ Effects in Model (10)	β Effects in Model (20)
D	1.31	.27
AD	1.45	.37
BD	3.45	1.24
CD	2.14	.76
BCD	0.86	-.15

$$\Omega^{\bar{D}}{}_{111} = (1.31)(1.45)(3.45)(2.14)(0.86) = 12.07. \quad (17a)$$

(All calculations in this paper were carried out to more significant digits than are reported here.)

Further insight into the meaning of the γ parameters can be gained by noting how the estimated value of these parameters affect the estimate of $\Omega^{\bar{D}}{}_{ijk}$, for $i = 1, 2$; $j = 1, 2$; $k = 1, 2$. By applying the numerical values of Table 3 to formulas (10)–(11), we find that the $\Omega^{\bar{D}}{}_{ijk}$ can be estimated by (17a) and as follows:

$$\Omega^{\bar{D}}{}_{211} = (1.31)\left(\frac{1}{1.45}\right)(3.45)(2.14)(0.86) = 5.74, \quad (17b)$$

$$\Omega^{\bar{D}}{}_{121} = (1.31)(1.45)\left(\frac{1}{3.45}\right)(2.14)\left(\frac{1}{0.86}\right) = 1.36, \quad (17c)$$

$$\Omega^{\bar{D}}{}_{112} = (1.31)(1.45)(3.45)\left(\frac{1}{2.14}\right)\left(\frac{1}{0.86}\right) = 3.56, \quad (17d)$$

et cetera. Comparing (17a) with (17b), we

[4] The relationship between formulas (12)–(16) and certain formulas in the analysis of variance will be clarified when we discuss formulas (23)–(27).

[5] From the relationship between $\gamma^{BC}{}_{11}$, $\gamma^{BC}{}_{22}$, $\gamma^{BC}{}_{12}$ and $\gamma^{BC}{}_{21}$ described by formula (11), we see that, after taking account of the various main effects in the model, the estimate of the expected odds $\Omega^{\bar{D}}{}_{1jk}$ favoring a preference for a Northern camp must be multiplied by the factor 0.86 for those whose region of origin is the same as their present camp location (viz., Northerners in a Northern and Southerners in a Southern camp); and it must be divided by the factor 0.86 for those whose region of origin differs from their present camp location (viz., Northerners in a Southern and Southerners in a Northern camp). For further comments, see the final section herein.

see the effect of $\gamma^{A\bar{D}}{}_1$. Comparing (17a) with (17c), we see the effect of $\gamma^{B\bar{D}}{}_1$ and $\gamma^{BC\bar{D}}{}_{11}$. Comparing (17a) with (17d), we see the effect of $\gamma^{C\bar{D}}{}_1$ and $\gamma^{BC\bar{D}}{}_{11}$.

We used the superscript $\bar{D}$ in the preceding two paragraphs to emphasize the fact that the odds Ω_{ijk} pertain to variable D, and that the γ parameters describe the main and interaction effects on these odds. To simplify notation we will delete this superscript hereafter, in all but one section of the paper.

Formula (10) expresses Ω_{ijk} as a product of certain main and interaction effect parameters. This formula can also be expressed in an additive form via logarithms. First, corresponding to Ω_{ijk}, we let Φ_{ijk} denote the natural logarithm of Ω_{ijk}; i.e., we define Φ_{ijk} as

$$\Phi_{ijk} = \log \Omega_{ijk}, \tag{18}$$

where "log" denotes the natural logarithm. Second, corresponding to formula (10)'s set of parameters $(\gamma, \gamma^{A}{}_i, \gamma^{B}{}_j, \gamma^{C}{}_k, \gamma^{BC}{}_{jk})$, we define a new set as follows:

$$\beta = \log \gamma,\ \beta^{A}{}_i = \log \gamma^{A}{}_i,$$
$$\beta^{B}{}_j = \log \gamma^{B}{}_j, \text{ etc.} \tag{19}$$

Then from (10) and (18)–(19) we see that

$$\Phi_{ijk} = \beta + \beta^{A}{}_i + \beta^{B}{}_j + \beta^{C}{}_k + \beta^{BC}{}_{jk}. \tag{20}$$

From (11) and (19) we see that

$$\beta^{A}{}_1 = -\beta^{A}{}_2,\ \beta^{B}{}_1 = -\beta^{B}{}_2,\ \beta^{C}{}_1 = -\beta^{C}{}_2,$$
$$\beta^{BC}{}_{11} = \beta^{BC}{}_{22} = -\beta^{BC}{}_{12} = -\beta^{BC}{}_{21}, \tag{21}$$

which can also be expressed as follows:

$$\sum_{i=1}^{2} \beta^{A}{}_i = 0,\quad \sum_{j=1}^{2} \beta^{B}{}_j = 0,\quad \sum_{k=1}^{2} \beta^{C}{}_k = 0,$$

$$\sum_{j=1}^{2} \beta^{BC}{}_{jk} = 0 \text{ (for } k = 1, 2), \tag{22}$$

$$\sum_{k=1}^{2} \beta^{BC}{}_{jk} = 0 \text{ (for } j = 1, 2).$$

Parameters β, $\beta^{A}{}_1$, $\beta^{B}{}_1$, and $\beta^{C}{}_1$ describe the main effects on Φ_{ijk} of the general mean [6] and variables A, B, and C; and parameter $\beta^{BC}{}_{11}$ describes the interaction effect of variables B and C on Φ_{ijk}. The model described by formula (20), which expresses Φ_{ijk} in terms of five parameters (β, $\beta^{A}{}_1$, $\beta^{B}{}_1$, $\beta^{C}{}_1$, $\beta^{BC}{}_{11}$), is equivalent to that described by formula (10), which expresses the corresponding Ω_{ijk} in terms of the corresponding five parameters (γ, $\gamma^{A}{}_1$, $\gamma^{B}{}_1$, $\gamma^{C}{}_1$, $\gamma^{BC}{}_{11}$).

We noted earlier that model (10)'s parameters could be expressed explicitly in terms of Ω_{ijk} (see formulas (12)–(16)). Similarly, the parameters in model (20) can be expressed explicitly in terms of Φ_{ijk}. From (20)–(22), we obtain the following expressions for these parameters in terms of Φ_{ijk}:

$$\beta = \left[\sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2} \Phi_{ijk}\right]/8, \tag{23}$$

$$\beta^{A}{}_1 = [\Phi_{1jk} - \Phi_{2jk}]/2 \text{ (for } j = 1, 2;\ k = 1, 2) \tag{24}$$
$$= \left[\sum_{j=1}^{2}\sum_{k=1}^{2} (\Phi_{1jk} - \Phi_{2jk})\right]/8,$$

$$\beta^{B}{}_1 = \left[\sum_{j=1}^{2}\sum_{k=1}^{2} (\Phi_{i1k} - \Phi_{i2k})\right]/8, \tag{25}$$

$$\beta^{C}{}_1 = \left[\sum_{i=1}^{2}\sum_{j=1}^{2} (\Phi_{ij1} - \Phi_{ij2})\right]/8, \tag{26}$$

$$\beta^{BC}{}_{11} = [\Phi_{i11} + \Phi_{i22} - \Phi_{i12} - \Phi_{i21}]/4 \tag{27}$$
$$\text{(for } i = 1, 2)$$
$$= \left[\sum_{i=1}^{2} [\Phi_{i11} + \Phi_{i22} - \Phi_{i12} - \Phi_{i21}]\right]/8.$$

Formulas (23)–(27) are equivalent to the corresponding formulas (12)–(16).[7] Formula (23) states that β is the arithmetic mean of Φ_{ijk} (corresponding to the eight possible values of (i, j, k)). Formula (24) states that $\beta^{A}{}_1$ can be expressed both as one-half the difference $\Phi_{1jk} - \Phi_{2jk}$ (for $j = 1, 2$; $k = 1, 2$), and as one-half the arithmetic mean of the differences $\Phi_{1jk} - \Phi_{2jk}$ corresponding to the four possible values of (j, k) obtained when $j = 1, 2$; $k = 1, 2$. Formula (25) states that $\beta^{B}{}_1$ equals one-half the arithmetic mean of the differences $\Phi_{i1k} - \Phi_{i2k}$ corresponding

[6] Since β is somewhat analogous to the main effect of the general mean in the usual model for the analysis of variance (i.e., the constant term in that model), we refer to β as the main effect of the general mean on Φ_{ijk}. β actually equals the *arithmetic* mean of the Φ_{ijk} corresponding to the eight possible values of (i, j, k) obtained when $i=1, 2$; $j=1, 2$; and $k=1, 2$. For further details, see formula (23) below.

[7] Indeed, instead of obtaining (23)–(27) from formulas (20)–(22), we could also have obtained (23)–(27) from formulas (12)–(16) and (18)–(19). Similarly, we could have obtained (12)–(16) from formulas (23)–(27), making use of formula (28) below and the fact that $\Omega_{ijk} = \exp \Phi_{ijk}$.

to the four possible values of (i, k) obtained when i = 1, 2; k = 1, 2. Formula (26) can be similarly expressed, and formula (27) also has a somewhat similar interpretation.

Since Table 2 presents the estimated values of F_{ijkl} (under model (10) or the equivalent model (20)), we can use these values to estimate first, Ω_{ijk} and Φ_{ijk} (see (8) and (18)) and second, the β parameters (see (23)–(27)). Table 3 includes the β parameters' estimated values. We can also use these values to calculate the γ parameters' estimated values since the relationship between the β and γ parameters can be expressed by (19) or by the following equivalent set of formulas: [8]

$$\gamma = \exp \beta,\ \gamma^{A}{}_{i} = \exp \beta^{A}{}_{i},\ \gamma^{B}{}_{j} = \exp \beta^{B}{}_{j}, \quad (28)$$

etc., where "exp" denotes the exponential function.[9]

Earlier we discussed Table 3's γ parameter estimated values. Now let us examine the estimated values of β, $\beta^{A}{}_{1}$, $\beta^{B}{}_{1}$, $\beta^{C}{}_{1}$, $\beta^{BC}{}_{11}$, also given in Table 3. In line with our earlier discussion of Table 3, in examining the estimated β parameters, we note that the estimated main effect of each variable (A, B, C) is positive (i.e., the estimates of the $\beta^{A}{}_{1}$, $\beta^{B}{}_{1}$, $\beta^{C}{}_{1}$, which could have been written as $\beta^{A\bar{D}}{}_{1}$, $\beta^{B\bar{D}}{}_{1}$, $\beta^{C\bar{D}}{}_{1}$, are all positive); but the estimated interaction effect between variables B and C is negative (i.e., the estimate of $\beta^{BC}{}_{11}$, which could have been written as $\beta^{BC\bar{D}}{}_{11}$, is negative).[10]

Later we shall show how to assess the statistical significance of the contribution made by certain parameters (e.g., $\gamma^{BC}{}_{11}$) in model (10), and by certain parameters (e.g., $\beta^{BC}{}_{11}$) in model (20), and we also show how to measure this contribution's magnitude.

Our models express the expected odds Ω_{ijk} in terms of the γ parameters (see (10) and also (29) below); or they express the expected log-odds Φ_{ijk} in terms of the β parameters (see (20) and also (35) below). These two forms of expression are equivalent. Since the expected frequencies F_{ijkl} can be expressed in terms of the Ω_{ijk} (see (9)), our models can also be used to express F_{ijkl} in terms of the γ parameters. In addition, letting P_{ijk} and Q_{ijk} denote the expected proportions F_{ijk1}/n_{ijk} and F_{ijk2}/n_{ijk}, respectively (see (3)–(4)), note that our models can also be used to express P_{ijk} (and Q_{ijk}) in terms of these γ parameters.

Before closing this section, we should note the relationship between model (20) and the usual models for (a) the analysis of variance and (b) the analysis of the "logit" pertaining to variable D.

Model (20) and the usual model for the three-way analysis of variance may be compared in several ways. (Note that the Φ_{ijk} in (20) can be presented in a three-way array, while the F_{ijkl} are presented in a four-way table.) In the usual three-way analysis of variance, one must assume homoscedasticity, i.e., that each observation in the three-way table has the same variance. On the other hand, for our kind of data, the homoscedasticity assumption would be contradicted in a way that could not be ignored.[11] Our data also contradict the assumption in the usual analysis of variance that each observation has a normal distribution.[12]

---

[8] For expository purposes, we discussed the γ before the β parameters. Since Table 3 already provided the γ parameters' estimated values, we could have used them in turn to estimate the β parameters (see (19)). Actually, rather than calculate the estimated β from the estimated γ parameters, calculated earlier from the estimated Ω_{ijk} (see (12)–(16)), it is easier to calculate the estimated γ from the corresponding estimated β parameters (see (28)), which can be calculated from the estimated Φ_{ijk} (see (23)–(27)).

[9] The exponential function is the inverse of the natural logarithm. Comparison of (19) and (28) should make this point clear. For example, for a given γ value, we can calculate β from (19) using a table of natural logarithms; and for a given β value, we can calculate γ either from (19), with the natural-logarithm table used now in so to speak, inverted order, or equivalently from (28) using a table of the exponential function.

[10] The fact that the estimate of $\beta^{BC}{}_{11}$ is negative corresponds to the fact that the estimate of $\gamma^{BC}{}_{11}$ is less than 1 (see (19) and (28)). We interpreted this fact in footnote (5). For further comment, see the final section herein.

[11] In the present context, we note that the variance of the observed proportion p_{ijk} (see (3)) will depend both on the magnitude of n_{ijk} (see (1)) and the expected proportion $P_{ijk}=F_{ijk1}/n_{ijk}$. A similar remark applies to the variance of the observed odds ω_{ijk} (see (2) and (5)) and the variance of the logarithm of ω_{ijk}. (The logarithm of the ω_{ijk} is of interest here since it corresponds to Φ_{ijk} in the same sense that ω_{ijk} corresponds to Ω_{ijk}; see (2), (8), (18).)

[12] On the other hand, when n_{ijk} is large, the

Note also that formulas (20)–(22) and (23)–(27) are similar to formulas appearing in the usual analysis of variance. However, to estimate the β parameters under model (20), we use the estimated values of the expected frequencies F_{ijkl} under the model (see Table 2) to estimate first Ω_{ijk} and Φ_{ijk} (see (8) and (18)); and then we use these estimated values of Φ_{ijk} in (23)–(27) to estimate the β parameters. In contrast, in the usual analysis of variance (assuming homoscedasticity), the quantity corresponding to the estimated Φ_{ijk} in formulas (23)–(27) is replaced by the observation in cell (i, j, k); and formulas (24) and (27) are replaced simply by the corresponding expressions on the second line of these two formulas.

Now let us consider the usual model for analyzing the logit pertaining to variable D. This logit is usually defined as being $\Phi_{ijk}/2$ (see, e.g., Fisher and Yates 1963). Model (20) states that this logit (multiplied by 2) can be expressed as a sum of parameters β, $\beta^{A}{}_{i}$, $\beta^{B}{}_{j}$, $\beta^{C}{}_{k}$, $\beta^{BC}{}_{jk}$ (i.e., the main effects of the general mean and of variables A, B, C, and the interaction effect between variables B and C). We can rewrite this model as a regression model expressing variable D's logit as a linear function of dummy variables pertaining to the main effects of variables A, B, C and the interaction effect between variables B and C; but homoscedasticity can not be assumed in this model. Later we shall test the statistical significance of the contribution made by certain parameters in this model, and we shall measure the contributions' magnitude by applying methods proposed in Goodman (1970, 1971a). For some related material, see also Dyke and Patterson (1952), Bishop (1969), Theil (1970), and the final section below.

A GENERAL MODEL FOR ANALYZING THE ODDS

Model (10) included the main effect on Ω_{ijk} of all three variables (A, B, C), but only one of three possible two-factor interaction effects (viz., $\gamma^{BC}{}_{jk}$); and it did not include the three-factor interaction effect (viz., $\gamma^{ABC}{}_{ijk}$). This model assumed that $\gamma^{AB}{}_{ij}$, $\gamma^{AC}{}_{ik}$, and $\gamma^{ABC}{}_{ijk}$ all equal 1. We shall now consider the model that includes all possible main and interaction effects and that makes no assumptions about which (if any) of these effects equals 1. Instead of model (10), we now have the following "saturated" model:

$$\Omega_{ijk} = \gamma\, \gamma^{A}{}_{i}\gamma^{B}{}_{j}\gamma^{C}{}_{k}\gamma^{AB}{}_{ij}\gamma^{AC}{}_{ik}\gamma^{BC}{}_{jk}\gamma^{ABC}{}_{ijk}, \tag{29}$$

where

$$\gamma^{A}{}_{1} = 1/\gamma^{A}{}_{2}, \ldots, \gamma^{AB}{}_{11} = \gamma^{AB}{}_{22} = 1/\gamma^{AB}{}_{12} = 1/\gamma^{AB}{}_{21}, \ldots,$$

$$\gamma^{ABC}{}_{111} = \gamma^{ABC}{}_{221} = \gamma^{ABC}{}_{212} = \gamma^{ABC}{}_{122} = 1/\gamma^{ABC}{}_{112} = 1/\gamma^{ABC}{}_{121} = 1/\gamma^{ABC}{}_{211} = 1/\gamma^{ABC}{}_{222}. \tag{30}$$

Formula (29) describes the effects of the γ parameters on Ω_{ijk}. It expresses Ω_{ijk} explicitly in terms of the model's γ parameters. These parameters can also be expressed explicitly in terms of Ω_{ijk}. From (29)–(30), we obtain the following expressions for the parameters in terms of the Ω_{ijk}: [13]

$$\gamma = \left[\prod_{i=1}^{2} \prod_{j=1}^{2} \prod_{k=1}^{2} \Omega_{ijk} \right]^{1/8}, \tag{31}$$

$$\gamma^{A}{}_{1} = \left[\prod_{j=1}^{2} \prod_{k=1}^{2} (\Omega_{1jk}/\Omega_{2jk}) \right]^{1/8}, \tag{32}$$

$$\ldots,$$

$$\gamma^{AB}{}_{11} = \left[\prod_{k=1}^{2} (\Omega_{11k}\Omega_{22k})/(\Omega_{12k}\Omega_{21k}) \right]^{1/8}, \tag{33}$$

$$\ldots,$$

$$\gamma^{ABC}{}_{111} = [(\Omega_{111}\Omega_{221}\Omega_{212}\Omega_{122})/(\Omega_{112}\Omega_{121}\Omega_{211}\Omega_{222})]^{1/8}. \tag{34}$$

For the saturated model (29)–(30), we can estimate the γ parameters by formulas (31)–(34), replacing the expected odds Ω_{ijk} in these formulas by the corresponding observed odds ω_{ijk}.[14] With the saturated

observed proportion p_{ijk} will be approximately normally distributed (as long as the expected proportion P_{ijk} differs sufficiently from the extreme values of 0 and 1). A similar remark applies to the observed odds ω_{ijk} and the logarithm of ω_{ijk}.

[13] Formulas (31)–(34) for the saturated model (29)–(30) correspond to formulas (12)–(16) for the unsaturated model (10)–(11).

[14] In contrast to this procedure for the saturated model, note that for an unsaturated model (e.g., model (10)) we use the estimated values of the expected frequencies F_{ijkl} under the model (see Table 2) to estimate Ω_{ijk} (see (8)); then we can use these estimated values of the Ω_{ijk} in (31)–(34) to estimate the γ parameters. (When these estimated values of the Ω_{ijk} are used in (31)–(34), we

Table 4. Estimate of the Main Effects and Interaction Effects of the Three Variables (A,B,C) on the Odds Ω_{ijk} Pertaining to Variable D in the Four-Way Contingency Table (Table 1), Under the Saturated Models (29) and (35)

Variable	γ Effects in Model (29)	β Effects in Model (35)	Standardized Value
D	1.28	.25	6.96
AD	1.44	.37	10.21
BD	3.43	1.23	34.36
CD	2.10	.74	20.65
ABD	0.96	-.04	-1.11
ACD	1.00	.00	0.00
BCD	0.86	-.15	-4.31
ABCD	0.97	-.03	-0.86

model's γ parameters thus estimated, the observed data fit perfectly. (For further comments on this point, see footnote 19 later herein.) Based on Table 1's data, the γ parameters' estimated values are given in Table 4. Note that, for Table 1's data, Table 4's estimated γ's are quite similar to the corresponding quantities of Table 3.

Having replaced the unsaturated (10) with the saturated model (29), we can also replace the unsaturated (20) with the following saturated model:

$$\Phi_{ijk} = \beta + \beta^{A}_{i} + \beta^{B}_{j} + \beta^{C}_{k} + \beta^{AB}_{ij} + \beta^{AC}_{ik} + \beta^{BC}_{jk} + \beta^{ABC}_{ijk}, \quad (35)$$

where

$$\beta^{A}_{1} = -\beta^{B}_{2}, \ \ldots, \ \beta^{AB}_{11} = \beta^{AB}_{22} = -\beta^{AB}_{12} = -\beta^{AB}_{21}, \ \ldots,$$
$$\beta^{ABC}_{111} = \beta^{ABC}_{221} = \beta^{ABC}_{212} = \beta^{ABC}_{122} = -\beta^{ABC}_{112} = -\beta^{ABC}_{121} = -\beta^{ABC}_{211} = -\beta^{ABC}_{222}. \quad (36)$$

obtain the same results as when they are used in (12)–(16).) For an unsaturated model (e.g., model (10)), the entries in Table 2 are the maximum-likelihood estimates of the F_{ijkl} under the model, and they are calculated by an iterative procedure which we shall comment upon later herein after we have presented the material in Table 5. The observed frequencies f_{ijkl} are the maximum-likelihood estimates of the F_{ijkl} under the saturated model, but *not* under an unsaturated model. Similarly, the observed odds ω_{ijk} are the maximum-likelihood estimates of the Ω_{ijk} under the saturated model, but *not* under an unsaturated model.

Model (35)–(36) is, of course, equivalent to model (29)–(30). Similarly, formulas (31)–(34) are equivalent to the following set of formulas:

$$\beta = \left[\sum_{i=1}^{2} \sum_{j=1}^{2} \sum_{k=1}^{2} \Phi_{ijk} \right] /8, \quad (37)$$

$$\beta^{A}_{1} = \left[\sum_{j=1}^{2} \sum_{k=1}^{2} (\Phi_{1jk} - \Phi_{2jk}) \right] /8, \quad (38)$$

$$\ldots,$$

$$\beta^{AB}_{11} = \left[\sum_{k=1}^{2} (\Phi_{11k} + \Phi_{22k} - \Phi_{12k} - \Phi_{21k}) \right] /8, \quad (39)$$

$$\ldots,$$

$$\beta^{ABC}_{111} = [\Phi_{111} + \Phi_{221} + \Phi_{212} + \Phi_{122} - \Phi_{112} - \Phi_{121} - \Phi_{211} - \Phi_{222}]/8. \quad (40)$$

For the saturated model (35)–(36) we can estimate the β parameters by formulas (37)–(40), replacing the "expected log-odds" Φ_{ijk} in these formulas by the corresponding log ω_{ijk}.[15] In addition, the variance of the estimated β parameters can be estimated by the following formula:[16]

$$S^2_{\hat{\beta}} = \left[\sum_{i=1}^{2} \sum_{j=1}^{2} \sum_{k=1}^{2} \sum_{l=1}^{2} (1/f_{ijkl}) \right] /64. \quad (41)$$

By dividing each estimated β parameter by its estimated standard deviation $S_{\hat{\beta}}$, we obtain the corresponding "standardized value" of the estimate. Each standardized value can be used to test whether the corresponding β parameter is nil.[17] Table 4 in-

[15] Remarks similar to those in footnotes 8 and 14 would apply here as well.

[16] Note should be taken of the fact that the estimation method presented herein for the saturated model can be improved upon by replacing f_{ijkl} in (41) by $f_{ijkl}+\frac{1}{2}$, and replacing the ω_{ijk} that are used in (31)–(34) (or in (37)–(40)) by $\omega_{ijk}=(f_{ijk1}+\frac{1}{2})/(f_{ijk2}+\frac{1}{2})$. It should also be noted that formula (41) and some of the other results presented herein are applicable both in the case where the observed four-way table (Table 1) describes results obtained for a random sample of individuals cross-classified with respect to the four variables (A, B, C, D), and also in the case where the f_{ijk1} and f_{ijk2} in row (i, j, k) of Table 1 describe results obtained with respect to variable D for a random sample of n_{ijk} individuals at levels i, j, k on variables A, B, C, respectively. For further details, see Goodman (1970) and Haberman (1970).

[17] The term "standardized value" of a statistic is used here to mean the ratio of the statistic and

cludes the β parameters' estimated values, and their corresponding standardized values.

By examining the magnitudes of Table 4's standardized values, we find that the model in which β^{AB}_{ij}, β^{AC}_{ik}, β^{ABC}_{ijk} are set equal to zero in (35) should merit consideration. (Recall that these three parameters could also have been written as $\beta^{AB\bar{D}}_{ij}$, $\beta^{AC\bar{D}}_{ik}$, $\beta^{ABC\bar{D}}_{ijk}$, respectively.) But the model obtained when these particular parameters are set equal to zero in (35) is equivalent to model (20). Thus, for Table 1's data, examining the saturated models (29) and (35) leads to models (10) and (20).

HOW TO TEST WHETHER A MODEL FOR THE ODDS FITS THE DATA

To test whether the hypothesis H described by model (10) fits Table 1's data, we first estimate the expected frequencies F_{ijkl} under the hypothesis H (see Table 2), and then compare the observed frequency f_{ijkl} in Table 1 with the corresponding estimate of the F_{ijkl} in Table 2, by calculating either the usual chi-square goodness-of-fit statistic

$$\sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2}\sum_{l=1}^{2}(f_{ijkl}-F_{ijkl})^2/F_{ijkl}, \quad (42)$$

or the corresponding chi-square based on the likelihood-ratio statistic; viz.,

$$2\sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2}\sum_{l=1}^{2} f_{ijkl}\log\,[f_{ijkl}/F_{ijkl}]. \quad (43)$$

The chi-square value obtained from (42) or (43) can be assessed by comparing its numerical value with the percentiles of the tabulated chi-square distribution. The degrees of freedom for testing hypothesis H will be $8-5=3$ (since (a) there are eight observed odds in Table 1, and (b) there are five γ parameters estimated in model (10)).

Using (42), we obtain a goodness-of-fit chi-square value of 1.46, and using (43), a likelihood-ratio chi-square value of 1.45. Since there were three degrees of freedom under H, the model fits the data well.

Model (10) is obtained from the saturated model (29) by making a specific set of its γ parameters equal to one.[18] Model (20) is obtained from the saturated model (35) by making a specific set of its β parameters equal to zero. Models obtained this way from saturated models we call "unsaturated."[19] Of course, all unsaturated models obtained from (29) or (35) are models for the odds Ω_{ijk} (or the log-odds Φ_{ijk}) pertaining to variable D. Thus, all these unsaturated models view the four-way table (Table 1) asymmetrically. In the four-way table for variables (A, B, C, D), we treated variable D as the dependent variable; i.e., we viewed the odds (or log odds) pertaining to variable D as depending on the level of variables (A, B, C).

When each individual in a sample is classified by four dichotomous variables (e.g., A, B, C, D), we obtain a four-way table (e.g., Table 1); and, in some contexts, any one of the four variables might be viewed as the dependent variable. For the four-way table, the expected frequencies estimated under a given unsaturated model that treats variable D as the dependent variable (see, e.g., Table 2) will usually differ from the corresponding expected frequencies estimated under a model that treats one of the other variables as the dependent variable.

In some contexts, the research worker will know which variable should be treated as the dependent variable; in others, any one of the four might be treated so. In still others, a different point of view would be appropriate. We could, for example, con-

its estimated standard deviation. The same or similar words have also been used by other writers to denote other things with which the usage here should not be confused.

If a particular β parameter is nil, then the standardized value of the corresponding estimated β will be approximately normally distributed with zero mean and unit variance (when the sample size is large). For comments on related matters, see Goodman (1970, 1971a).

[18] Indeed, the three degrees of freedom used above to test model (10) correspond to the three γ parameters (viz., γ^{AB}_{11}, γ^{AC}_{11}, γ^{ABC}_{111}) in (29) that are set equal to one under model (10).

[19] The number of degrees of freedom used to test a given unsaturated model will equal the number of γ parameters in (29) that are set equal to one under the unsaturated model. Since none of the γ parameters in (29) are set equal to one under that model (i.e., the number of γ parameters set equal to one is zero), there will be zero degrees of freedom under the saturated model. This corresponds to the fact that the observed data fit perfectly under the saturated model, since it includes all possible main and interaction effects (i.e., all possible γ parameters).

sider the case where none of the variables is the dependent variable, but where all are mutually related in some sense (see, e.g., Goodman, 1970). For the four-way table, Goodman (1970) described in his Table 4 a large class of models that would include as special cases models like our (10) and (20) which treat one variable as the dependent variable, as well as "unsaturated" models of a different kind where none of the variables is treated as the dependent variable but where some or all may be mutually related variables. Goodman's Table 4 (1970) contains fifty-three different "unsaturated models" in which one of the four variables is treated as the dependent variable and 113 different "unsaturated" models in which none of the four variables is viewed as the dependent variable. For the case where a given variable (say variable D) is the dependent variable, Goodman's Table 4 (1970) lists nineteen different unsaturated models.

Earlier herein we considered the case where a given variable is the dependent variable. Our models well suit this case (see (10), (20), (29), (35)). Many readers will find our exposition of this case easier to understand than the exposition of the more general case in Goodman (1970). Nevertheless, the more general models and methods of the earlier article also apply to the special case we considered. For each unsaturated model of the kind considered herein, and also for other kinds of "unsaturated" models, Goodman's Table 4 (1970) gave the corresponding degrees of freedom when each variable in the contingency table is dichotomous. He also described ways to calculate the degrees of freedom when some variables are polytomous but not necessarily dichotomous. A single computer program can be used to calculate the estimate of the F_{ijkl}, and the corresponding chi-square values (42) and (43), for any set of "unsaturated" models of the kinds considered herein and in Goodman (1970). For related material dealing with such models see, e.g., Bishop (1969), Goodman (1970, 1971a, 1972).

Let us reconsider model (10), which we obtained from the saturated model (29) by making some of its γ parameters equal to one. We can describe this unsaturated model in any of the following equivalent ways: (1) By listing the γ parameters that are included in model (10); viz., $\gamma^{\bar{D}}$ $\gamma^{A\bar{D}}{}_i$, $\gamma^{B\bar{D}}{}_j$, $\gamma^{C\bar{D}}{}_k$, $\gamma^{BC\bar{D}}{}_{jk}$.[20] (2) By listing the γ parameters in (29) that are set equal to one under the model; viz., $\gamma^{AB\bar{D}}{}_{ij}$, $\gamma^{AC\bar{D}}{}_{ik}$, $\gamma^{ABC\bar{D}}{}_{ijk}$. (3) By listing the particular marginal tables that are fitted under the model—a topic we shall now discuss.

From our Table 1 we can determine n_{ijk} as defined by formula (1). In all unsaturated models obtained from the saturated model (29), the n_{ijk} are considered fixed; thus in these models the expected frequencies F_{ijkl} (under the model) will satisfy the following condition:

$$F_{ijk1} + F_{ijk2} = n_{ijk}. \qquad (44)$$

By comparing the n_{ijk} from Table 1 with the estimated value of $F_{ijk1} + F_{ijk2}$ from Table 2, we see that condition (44) is satisfied. Since the n_{ijk} describe the three-way marginal table pertaining to variables (A, B, C), we shall use the symbol {ABC} to denote this table. Condition (44) states that the marginal table {ABC} is fitted under the model.

In addition to the marginal table {ABC}, two other marginal tables are fitted under model (10); viz., the two-way marginal table {AD} and the three-way marginal table {BCD}. Table 5 gives the three marginal tables fitted under model (10).[21] In the preceding paragraph, we explained why the marginal table {ABC} was fitted. Under model (10), we also fit the marginal tables {AD} and {BCD} because it includes the parameters $\gamma^{A\bar{D}}{}_i$ and $\gamma^{BC\bar{D}}{}_{jk}$, which pertain to the relationship between variables A and D (as displayed in the marginal table {AD})

[20] We return now to the notation used earlier where the letter $\bar{D}$ was included in the superscript of each γ parameter to emphasize the fact that the γ parameters describe the main and interaction effects on the odds pertaining to variable D. This notation will facilitate some of our present exposition. This notation's utility will become clearer two paragraphs below.

[21] The four-way contingency table of observed data (Table 1) can be displayed as a 2 x 2 x 2 x 2 table, or an 8 x 2 table (as in Table 1); similarly the three-way marginal table {ABC} can be displayed as a 2 x 2 x 2 table, or a 4 x 2 table (as in Table 5), or as a 8 x 1 table (as we would obtain if we present it as the marginal of the 8 x 2 table displayed in Table 1).

Table 5. The Three Marginal Tables That are Fitted When Models (10) and (20) are Applied to the Four-Way Contingency Table (Table 1)

I. Table {ABC}

Variable A	Variable B	Variable C	
		North	South
Negro	North	423	1126
Negro	South	653	2093
White	North	1117	1384
White	South	280	960

II. Table {AD}

Variable A	Variable D	
	North	South
Negro	2027	2268
White	2024	1717

III. Table {BCD}

Variable B	Variable C	Variable D	
		North	South
North	North	1263	286
North	South	764	1982
South	North	1829	672
South	South	195	1045

and to the joint relationship among variables B, C, and D (as displayed in the marginal table {BCD}).[22]

The reader will find that the entries in the three marginal tables in Table 5, which were calculated from Table 1's data, equal the corresponding entries in the three marginal tables calculated from Table 2's estimated F_{ijkl}. The computer program, to which we referred in the fourth paragraph preceding this one, calculated the estimated values in Table 2 (viz., the maximum-likelihood estimates of the expected frequencies F_{ikjl} under model (10)) by an iterative procedure which insured that the three marginal tables given in Table 5 would be fitted when Table 2's estimated F_{ijkl} are used. For further details about the computing procedure, see, for example, the literature cited in the paragraph referred to above.

Although the three marginal tables (viz., {ABC}, {AD}, {BCD}) in Table 5 are fitted under model (10), we noted earlier that the reason for fitting {ABC} in the present context is somewhat different from the reason for fitting {AD} and {BCD}. The marginal table {ABC} is considered to be fixed under model (10); i.e., the n_{ijk} in (1) and (44) are viewed as constants. Aside from the n_{ijk} constants, to estimate the F_{ijkl} under model (10), we use only the information contained in the observed marginals tables {AD} and {BCD}.

The above remarks pertain to model (10), but they can be extended in a straightforward way to a wide range of unsaturated models obtained from the saturated model (29) by setting certain specified γ parameters in (29) equal to one. Now let's apply several such unsaturated models to Table 1's data. Table 6 lists the chi-square values (42) and (43) obtained in testing these models. We include both chi-square values (42) and (43) in Table 6; but, in the present context, (43) has some advantages (see, e.g., Goodman 1968, 1970). In the remaining discussion, we shall use only the chi-square value based on (43).

Each model in Table 6 is described there by listing the marginal tables fitted under the model. For the sake of brevity, we actually list in Table 6 the "minimal set" of marginal tables fitted under the model, rather than the entire set of marginal tables that will in fact be fitted (see footnote 22 herein and Goodman 1970). For example, for model (10), which is presented as H_1 in Table 6, we list in Table 6 the following "minimal set" of marginal tables fitted under the model: {ABC}, {AD}, {BCD}. From this "minimal set" of marginal tables,

[22] When the three-way marginal table {BCD} is fitted, then the following two-way marginal tables will fit automatically: {BC}, {BD}, {CD}. Similarly, when a two-way marginal table, say, {AD} is fitted, then the two one-way marginals, {A} and {D}, will fit automatically. Corresponding to the superscript of each γ parameter in model (10) (with the letter $\overline{D}$ added to each superscript), a marginal table pertaining to that superscript will be included in the set of marginal tables fitted under the model. Under model (10), it will suffice to include {AD} and {BCD} (in addition to table {ABC}) in the set of fitted marginal tables, since then all the marginal tables corresponding to the model's γ parameters (viz., {D}, {AD}, {BD}, {CD}, {BCD}) will actually be fitted.

Table 6. Chi-Square Values for Some Models Pertaining to Table 1

Model	Fitted Marginals	Degrees of Freedom	Likelihood-Ratio Chi-Square	Goodness-of-Fit Chi-Square	γ Parameters Included in the Model
H_1	{ABC},{AD},{BCD}	3	1.45	1.46	[D],[AD],[BD],[CD],[BCD]
H_2	{ABC},{AD},{BD},{CD}	4	24.96	25.73	[D],[AD],[BD],[CD]
H_3	{ABC},{BCD}	4	152.65	147.59	[D],[BD],[CD],[BCD]
H_4	{ABC},{BD},{CD}	5	186.36	180.26	[D],[BD],[CD]
H_5	{ABC},{AD},{CD}	5	2286.83	2187.71	[D],[AD],[CD]
H_6	{ABC},{AD},{BD}	5	695.01	727.16	[D],[AD],[BD]
H_7	{ABC},{D}	7	3111.47	2812.64	[D]
H_8	{ABC},{ACD},{BCD}	2	1.32	1.34	[D],[AD],[BD],[CD],[ACD],[BCD]
H_9	{ABC},{ABD},{BCD}	2	0.68	0.69	[D],[AD],[BD],[CD],[ABD],[BCD]
H_{10}	{ABC},{ABD},{ACD}	2	17.29	18.73	[D],[AD],[BD],[CD],[ABD],[ACD]
H_{11}	{ABD},{ACD},{BCD}	2	24.79	25.11	*
H_{12}	None	15	5469.88	5989.11	*

*Models H_{11} and H_{12} cannot be expressed in terms of the γ parameters. See related discussion in the present article.

we find that the following marginal tables will in fact be fitted: {D}, {AD}, {BD}, {CD}, {BCD} as well as {ABC} and all the marginal tables formed from {ABC}. Variable D is included in five of the marginal tables listed above, and the model under consideration (i.e., model H_1 of Table 6) will include the following five γ parameters corresponding to these five marginal tables: $\gamma^{\bar{D}}$, $\gamma^{A\bar{D}}{}_i$, $\gamma^{B\bar{D}}{}_j$, $\gamma^{C\bar{D}}{}_k$, $\gamma^{BC\bar{D}}{}_{jk}$.

Consider now hypothesis H_2 in Table 6. Since this model fits the marginals {ABC}, {AD}, {BD}, {CD}, it will include the following γ parameters: $\gamma^{\bar{D}}$, $\gamma^{A\bar{D}}{}_i$, $\gamma^{B\bar{D}}{}_j$, $\gamma^{C\bar{D}}{}_k$.[23] Similarly, hypothesis H_3 in Table 6 fits the marginals {ABC}, {BCD}; and thus that model includes the following γ parameters: $\gamma^{\bar{D}}$, $\gamma^{B\bar{D}}{}_j$, $\gamma^{C\bar{D}}{}_k$, $\gamma^{BC\bar{D}}{}_{jk}$. Hypothesis H_4 in Table 6 fits the marginals {ABC}, {BD}, {CD}; and thus that model includes the following γ parameters: $\gamma^{\bar{D}}$, $\gamma^{B\bar{D}}{}_j$, $\gamma^{C\bar{D}}{}_k$. Let us discuss these and other models in Table 6 further.

As we have already noted, model (10) is listed as H_1 of Table 6. If we now make $\gamma^{BC\bar{D}}{}_{jk}$ equal to 1 in model (10), we get H_2 of Table 6. In model H_2, the odds pertaining to variable D are expressed in terms of the parameters $\gamma^{\bar{D}}$, $\gamma^{A\bar{D}}{}_i$, $\gamma^{B\bar{D}}{}_j$, $\gamma^{C\bar{D}}{}_k$, i.e., the main effects of the general mean and variables A, B, and C. To test whether the parameter $\gamma^{BC\bar{D}}{}_{jk}$ in model (10) contributes in a statistically significant way, we can use the difference between the corresponding chi-square values for H_2 and H_1 as a chi-square statistic with one degree of freedom. (We get the one degree of freedom by subtracting the corresponding degrees of freedom for H_2 and H_1; i.e., $4 - 3 = 1$.) From Table 6's chi-square values for H_2 and H_1, we see that $\gamma^{BC\bar{D}}{}_{jk}$ does contribute to model (10) in a statistically significant way.

[23] For the reader who has difficulty determining which γ parameters are included in the model from the description of the model in terms of the marginal tables that are fitted, we include this information in Table 6's final column. In that column, we use the symbols [D], [AD], [BD],..., to denote $\gamma^{\bar{D}}$, $\gamma^{A\bar{D}}{}_i$, $\gamma^{B\bar{D}}{}_j$,..., respectively.

If we set $\gamma^{A\bar{D}}_{i}$ equal to 1 in model (10), we get H_3 of Table 6. To test whether the parameter $\gamma^{A\bar{D}}_{i}$ in model (10) contributes in a statistically significant way, we can use the difference between the corresponding chi-square values for H_3 and H_1 as a chi-square statistic with one degree of freedom. From Table 6's chi-square values for H_3 and H_1, we see that $\gamma^{A\bar{D}}_{i}$ does contribute to model (10) in a statistically significant way.

If we set $\gamma^{A\bar{D}}_{i}$ and $\gamma^{BC\bar{D}}_{jk}$ equal to 1 in model (10), we get H_4 of Table 6. If we set $\gamma^{B\bar{D}}_{j}$ and $\gamma^{BC\bar{D}}_{jk}$ equal to 1 in model (10), we get H_5. If we set $\gamma^{C\bar{D}}_{k}$ and $\gamma^{BC\bar{D}}_{jk}$ equal to 1 in model (10), we get H_6. Comparing the magnitudes of Table 6's corresponding three chi-square values, we see that the worst fitting model was H_5, the next worst H_6, and the least worst H_4. In other words, by comparing the three models obtained from H_2 by deleting the main effect of one of the variables (A, B, C), we see that $\gamma^{B\bar{D}}_{j}$ contributes the most.

If we set $\gamma^{A\bar{D}}_{i}$, $\gamma^{B\bar{D}}_{j}$, $\gamma^{C\bar{D}}_{k}$, and $\gamma^{BC\bar{D}}_{jk}$ equal to 1 in model (10), we get H_7 of Table 6. In model H_7, the odds pertaining to variable D depend on $\gamma^{\bar{D}}$ (the main effect of the general mean), but are unaffected by the level of variables A, B, and C. In other words, model H_7 states that variable D is independent of the joint variable A, B, C. From the chi-square value for H_7 in Table 6, we see that the data contradict this model.

If we set $\gamma^{AB\bar{D}}_{ij}$ and $\gamma^{ABC\bar{D}}_{ijk}$ equal to 1 in model (29), we get H_8 of Table 6. If we set $\gamma^{AC\bar{D}}_{ik}$ and $\gamma^{ABC\bar{D}}_{ijk}$ equal to 1 in model (29), we get H_9. If we set $\gamma^{BC\bar{D}}_{jk}$ and $\gamma^{ABC\bar{D}}_{ijk}$ equal to 1 in model (29), we get H_{10}. Table 6 shows that H_8 and H_9 fit the data well, but H_{10} does not.

From the above description of H_8 and H_9, we can express model H_1 as follows: Model H_1 states both that H_8 is true *and* that $\gamma^{AC\bar{D}}_{ik}$ in H_8 equals 1. Model H_1 also states both that H_9 is true *and* that $\gamma^{AB\bar{D}}_{ij}$ in H_9 equals 1. Thus, if H_1 is true, then H_8 and H_9 will also be true; but H_8 and H_9 can be true in cases where H_1 is not. H_1 implies models H_8 and H_9.

Models H_{11} and H_{12} of Table 6 differ from H_1 to H_{10} in an important respect. These last two models do not include the marginal {ABC} among the marginals that are fitted under the model. Therefore, the expected frequencies F_{ijkl} estimated under models H_{11} and H_{12} will *not* satisfy condition (44) (except in some special cases). These two models cannot be expressed as unsaturated models obtained from the saturated model (29), except in cases where condition (44) is satisfied.

Model H_{12} of Table 6 is easier to describe than H_{11}, so I will describe it first. Model H_{12} states that the sixteen cells of Table 1 are equiprobable. From the chi-square value for H_{12} in Table 6, we see that the data contradict this model.

Now let us consider H_{11} of Table 6. As we noted above, this model cannot be expressed as one in which variable D is the dependent variable, since table {ABC} is not fitted under it. However, since the other three-way marginal tables (viz., {ABD}, {ACD}, {BCD}) are fitted under model H_{11}, we see that H_{11} is a model in which any one of the other variables (C, B, or A) can be viewed as the dependent variable. (Note that three of the four possible three-way marginal tables are fitted under H_{11}, and also under H_8, and H_9, and H_{10}.) From the chi-square value for H_{11} in Table 6, we see that the data contradict this model.

We noted earlier that Goodman's Table 4 (1970) included models in which one of the variables is treated as the dependent variable, and it included other kinds of models as well. To test whether any of these other kinds of models might fit the data in our four-way table, we would first consider H_{11} of Table 6; for this model assumes only that the F_{ijkl} are not affected by the three factor "interaction effect" among the three variables A, B, C, nor by the four-factor "interaction effect" among variables A, B, D, C.[24] If this particular model does not fit

[24] Under model H_{11}, the only tables not fitted to the data are the three-way marginal table {ABC} and the four-way table {ABCD}; so the F_{1jkl} (under model H_{11}) are not affected by the three-factor "interaction effect" among variables A, B, C (as displayed in the marginal table {ABC}) nor by the four-factor "interaction effect" among variables A, B, C, D (as displayed in the four-way table). We use the term "interaction effect" in the preceding sentence, and in the sentence to which this footnote applies, in a way

Table 7. Analysis of the Variation in the Odds Pertaining to Variable D in the Four-Way Contingency Table (Table 1)

Source of Variation	Degrees of Freedom	Chi-Square	Numerical Value
1. Total variation due to the "main effects" of variables A,B,C and "interaction effects" among these variables	7	$X^2(H_7)$	3111.47
1a. Due to variation unexplained by model H_1	3	$X^2(H_1)$	1.45
1b. Due to variation explained by model H_1	4	$X^2(H_7)-X^2(H_1)$	3110.02
Partition of (1a)			
1a.1. Due to variation unexplained by model H_9	2	$X^2(H_9)$	0.68
1a.2. Due to variation explained by the $\gamma_{ij}^{AB\bar{D}}$ parameter in model H_9	1	$X^2(H_1)-X^2(H_9)$	0.77

the data, then the data will also contradict any of the other kinds of "unsaturated" models that do not treat variable D as the dependent variable.[25] For examples of data that do not contradict these other kinds of "unsaturated" models, see, Goodman (1970, 1971a).

Before closing this section, we should note that some of the material discussed above could be presented in summary form in tables that are somewhat analogous to the usual analysis of variance tables. Table 7 is an example.

MULTIPLE AND PARTIAL CORRELATION COEFFICIENTS FOR MODELS FOR THE ODDS

In the usual multiple regression analysis for quantitative variables (predicting variable Y from, say, variables X_1 and X_2), the quantity $R^2_{Y\cdot X_1X_2}$, which is the square of the multiple correlation coefficient, can be interpreted as follows: It is the relative decrease in Y's "unexplained variation" obtained when comparing the case where X_1 and X_2 are not used to predict Y with the case where both are used. Similarly, the quantity $r^2_{YX_1\cdot X_2}$, which is the square of the partial correlation coefficient, can be interpreted as follows: It is the relative decrease in Y's unexplained variation obtained when comparing the case where X_2 but not X_1 is used to predict Y with the case where both are used. The quantity $R^2_{Y\cdot X_1X_2}$ is sometimes referred to as the coefficient of multiple determination, and the quantity $r^2_{YX_1\cdot X_2}$ can be called the coefficient of partial determination. Goodman (1970, 1971a) introduced coefficients that are somewhat analogous to the usual coefficients of multiple and partial determination for analyzing the odds pertaining to a given variable in the four-way contingency table. We shall now illustrate their calculation.

For a given model in Table 6 (say model H_i, for $i = 1, 2, \ldots, 12$), we shall use the symbol $X^2(H_i)$ to denote its chi-square value. In the preceding section, we noted, among other things, that the statistic $X^2(H_2) - X^2(H_1)$ could be used to test whether the parameter γ^{BC}_{jk} in H_1 contributed in a statistically significant way.[26] To

related to but different from the way we used it earlier. Earlier the term referred to the interaction effects of certain variables on the expected odds Ω_{ijk} pertaining to variable D; whereas above the term refers to the "interaction effects" *among* certain variables in the four-way table. For further details, see Goodman (1970, 1971a).

[25] Except for model H_{11}, any other unsaturated model that does not treat variable D as the dependent variable can be viewed as a model that states both that H_{11} is true *and* that some additional "interaction effects" (in addition to the particular three and four-factor "interaction effects" noted in sentence one of footnote 24) can be set equal to one. Thus, if any other unsaturated model (of the above kind) is true, then H_{11} will also be true. If H_{11} is not true, then none of the other unsaturated models (of the above kind) can be true. For related matters, see Goodman (1970, 1971a).

[26] To facilitate exposition in the preceding section, we included the letter $\bar{D}$ in the superscript of

measure the contribution's magnitude, we recommend the following coefficient, which we shall call the coefficient of partial determination between the odds ω_{ijk} and the parameter $\gamma^{BC}{}_{jk}$, when the other γ's in model H_1 are taken into account: [27]

$$r^2_{\omega\ \gamma_{BC}\cdot H_1} = [X^2(H_2) - X^2(H_1)]/X^2(H_2). \quad (45)$$

From Table 6, we see that this coefficient equals .94 for Table 1's data.

We also noted in the preceding section that the statistic $X^2(H_3) - X^2(H_1)$ could be used to test whether the parameter $\gamma^A{}_i$ in H_1 contributed in a statistically significant way. As in the preceding paragraph, we shall measure this contribution's magnitude by the following coefficient of partial determination: [28]

$$r^2_{\omega\ \gamma_A\cdot H_1} = [X^2(H_3) - X^2(H_1)]/X^2(H_3). \quad (46)$$

From Table 6, we see that this coefficient equals .99 for Table 1's data.

To measure how well model H_1 fits the data, we consider the following coefficient, which we call the coefficient of multiple determination between ω and the γ parameters in model H_1:

$$R^2_{\omega\cdot H_1} = [X^2(H_7) - X^2(H_1)]/X^2(H_7). \quad (47)$$

From Table 6, we see that this coefficient equals 1.00 (to two decimal places) for Table 1's data.

We might also consider the following coefficient, which we shall call the coefficient of multiple-partial determination between ω and the parameters $\gamma^A{}_i$ and $\gamma^{BC}{}_{jk}$ in model H_1, when H_1's other γ parameters (viz., γ, $\gamma^B{}_j$, $\gamma^C{}_k$) are taken into account.

$$R^2_{\omega\ (\gamma_A,\ \gamma_{BC})\cdot\gamma,\ \gamma_B,\ \gamma_C} = [X^2(H_4) - X^2(H_1)]/X^2(H_4). \quad (48)$$

From Table 6 we see that the coefficient equals .99 (to two decimal places) for Table 1's data. Similarly, we can measure the contribution of $\gamma^B{}_j$ and $\gamma^{BC}{}_{jk}$ (using H_5 rather than H_4 in (48)) or the contribution of $\gamma^C{}_k$ and $\gamma^{BC}{}_{jk}$ (using H_6 rather than H_4 in (48)).

We can also use the above coefficients to measure the magnitude of the contribution made by the parameters in other models in Table 6. For example, to measure the magnitude of $\gamma^{AC}{}_{ik}$ in model H_8, we use the following coefficient of partial determination:

$$r^2_{\omega\ \gamma_{AC}\cdot H_8} = [X^2(H_1) - X^2(H_8)]/X^2(H_1). \quad (49)$$

From Table 6, we see that this coefficient equals .09 for Table 1's data.

All of the r^2 and the R^2 coefficients given by (45)–(49) above took the general form

$$R^2 = [X^2(H'') - X^2(H')]/X^2(H''), \quad (50)$$

where the γ parameters in model H'' are also included among the γ parameters in model H'. We could also write each coefficient as follows:

$$\frac{\sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2}\sum_{l=1}^{2} F'_{ijkl} \log [F'_{ijkl}/F''_{ijkl}]}{\sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2}\sum_{l=1}^{2} f_{ijkl} \log [f_{ijkl}/F''_{ijkl}]} \quad (51)$$

where F'_{ijkl} and F''_{ijkl} denote the expected frequencies estimated under model H' and model H'', respectively. The expression of the coefficient in the form (51) is somewhat analogous to the expression of the coefficients of multiple and partial determination, in the usual multiple regression analysis, as

---

each γ parameter; e.g., $\gamma^{BC}{}_{jk}$ in (10) became $\gamma^{BC\overline{D}}{}_{jk}$. In the present section, we have no need for this more cumbersome notation and will not include the letter $\overline{D}$ in the superscript. The reader should, of course, keep in mind that, say, $\gamma^{BC}{}_{jk}$ here has the same meaning as $\gamma^{BC\overline{D}}{}_{jk}$ earlier, and that the various γ parameters describe the main and interaction effects on the odds pertaining to variable D.

[27] In the subscript of r^2 in (45), we changed the γ^{BC} notation to the γ_{BC} notation because of typographical considerations. This simple notational change should not confuse the reader. Similar notational changes will be made in other formulas in this section.

Since model H_1 includes the γ parameters γ, $\gamma^A{}_i$, $\gamma^B{}_j$, $\gamma^C{}_k$, $\gamma^{BC}{}_{jk}$, we could let

$$r^2_{\omega\gamma_{BC}\cdot\gamma,\ \gamma_A,\ \gamma_B,\ \gamma_C}$$

denote the coefficient defined by (45). To test whether this coefficient differs significantly from zero, we use the statistic $X^2(H_2)$—$X^2(H_1)$ as noted earlier.

[28] Remarks like those in the second paragraph of footnote 27 can be applied to the coefficients defined by (46)–(49). For example, for (46), we could let

$$r^2_{\omega\gamma_A\cdot\gamma,\ \gamma_B,\ \gamma_C,\ \gamma_{BC}}$$

denote this coefficient, and we could assess the statistical significance of this coefficient using the statistic $X^2(H_8)$—$X^2(H_1)$ noted earlier.

a ratio of the "explained variation" (when model H′ is used to "explain" the variation that was not explained by H″) to the "unexplained variation" (when model H″ is used).[29]

COMMENTS ON SOME RELATED WORK

As we noted earlier, Coleman's model and methods differ from ours in several ways. His model does not fit the data as well, and his explanation of the data is less parsimonious. Indeed, he observes in his book (1964) that his model differs from the actual data in certain systematic ways, attributing these deviations to a supposed interaction for Negroes (but not whites) between their region of origin and present camp location.[30] In contrast, we find that (a) model (10) in the present article fits the data very well, (b) it does not require an ostensible interaction for Negroes (but not whites) of the kind considered by Coleman, (c) it includes an interaction effect γ^{BC}_{jk} between region of origin and present camp location, which applies equally to both Negroes and whites, (d) this interaction effect is statistically significant, and (e) it both reduces the expected odds favoring a preference for a Northern camp for those whose region of origin is the same as their present camp location, and increases these expected odds for those whose region of origin differs from their present camp location, after the various main effects in the model have been taken into account.[31]

Coleman's article did not show how to test whether his model fit the actual data, nor was he able to measure how well it fit. Furthermore, he did not show how to test the statistical significance of the contribution made by the various parameters in the model, nor could he measure their contribution's magnitude. In addition, the variance of Coleman's estimates of the main effects in his model was larger than it would have been had he used more efficient estimation methods (e.g., maximum-likelihood estimation methods); and his estimates are biased to the extent that his model excluded relevant interaction effects.[32]

Coleman's model states that the effects on the expected proportions P_{ijk} are linear.[33] Applying Coleman's estimation methods to his model, it is possible to obtain clearly incorrect estimates of the P_{ijk} under the model; e.g., estimates of the expected *proportions* P_{ijk} that are negative or larger than one.[34] Furthermore, his model and methods do not take into account the fact that the variance of the observed proportion p_{ijk} will depend on the magnitude of P_{ijk}.[35]

Some of the limitations of Coleman's approach apply to the usual multiple regression model (and analysis of variance model) if used in the present context. For data of the kind considered in the present article, the assumption of homoscedasticity made in the usual multiple regression model (and in

---

[29] When H″ is taken as H_7 of Table 6, then the denominator in (51) (i.e., the "unexplained variation" when model H_7 is used) corresponds to the "total variation" in the denominator of the usual coefficient of multiple determination in multiple regression analysis. For related matters, see Goodman (1970).

[30] Coleman (1964) did not provide methods for including interaction effects in his models, and so could not measure the magnitude of the ostensible interaction to which he referred, nor could he judge whether introducing the ostensible interaction would improve the fit of his model.

[31] For further details, see footnote 5. In addition to the effect of γ^{BC}_{jk} described there, the effect can also be described as follows: For the estimate of the expected odds $\Omega^{\overline{D}}_{1jk}$ in favor of a Northern camp, the effect on the estimated $\Omega^{\overline{D}}_{1jk}$ of being at present in a Northern rather than a Southern camp is less for those from the North than from the South. Similarly, the effect on the estimated $\Omega^{\overline{D}}_{1jk}$ of being a Northerner rather than a Southerner is less for those presently in a Northern rather than a Southern camp.

[32] The remarks above apply to Coleman's (1964) and Boudon's (1968) articles, except that Boudon's model did allow for interaction effects.

[33] Recall that $P_{1jk}=F_{1jk1}/n_{1jk}$, using our notation. In contrast to Coleman's, our model states that the expected odds Ω_{1jk} can be expressed in terms of multiplicative effects. In many substantive contexts, it will be more useful to consider multiplicative rather than additive effects. Furthermore, from the point of view of statistical theory, there are a number of reasons for preferring multiplicative models of our kind for analyzing data of the kind presented in Table 1. We shall not pursue these matters further here.

[34] Since P_{1jk} denotes an expected *proportion,* it should not be negative nor larger than one. Therefore, it is undesirable to use models and methods that can lead to *estimates* of the P_{1jk} that are negative or that are larger than one.

[35] In other words, Coleman implicitly assumes homoscedasticity when, on the contrary, his data violate this assumption.

the usual analysis of variance model) would be contradicted in a way that could not be ignored. In addition, as with Coleman's analysis, if one applied the usual multiple regression methods to the model in which the effects on the expected proportions P_{ijk} are linear, one could obtain clearly incorrect estimates of the P_{ijk} under the model, in the sense described above.[36]

We noted earlier that our data were also analyzed by Zeisel (1968) and Theil (1970). Zeisel described various interesting features of these data. These features can be explained, in a more comprehensive and compact way, in terms of the estimated parameters in model (10) of the present article. For example, from the estimates for model (10) presented in Table 3 herein, we find that the estimated product of the parameters γ and γ^{BC}_{11} is approximately one (more precisely, this product is 1.13), and this single fact can be used to explain the following features of the data: (a) the preference for a Northern camp location among Negro Northerners in Northern camps is approximately equal to the preference for a Southern camp location among white Southerners in Southern camps; and (b) the preference for a Northern camp location among white Northerners in Northern camps is approximately equal to the preference for a Southern camp location among Negro Southerners in Southern camps. (In order to see that this single fact explains features (a) and (b), insert the estimated values of the parameters in model (10).) The other interesting features noted by Zeisel can also be explained in similar terms, with one exception. This exception pertains to Zeisel's mention of a supposed effect on camp preference due to the interaction between race and region of origin, among those in Northern camps. Applying the methods of the present article, we find that this ostensible interaction effect is not statistically significant, and there is no need to include it in our model (10).

We comment next on the article by Theil (1970). He used the logit model corresponding to (20), but his estimation method and his analysis differed from ours. Theil (1970) used a weighted least-squares procedure, as did Grizzle, Starmer, and Koch (1969) in the same context; whereas, all the estimates presented in the present article are maximum-likelihood estimates. In commenting on the weighted least-squares procedure, the Grizzle-Starmer-Koch article notes that estimates obtained by their procedure have a somewhat larger variance than maximum-likelihood estimates (see also Rao 1965); similarly Theil's estimates have a somewhat larger variance than our maximum-likelihood estimates. We also find that it is harder to use the methods proposed by Theil, and by Grizzle, Starmer, and Koch than the methods proposed in the present article, when studying the kinds of hypotheses we have discussed for the four-way contingency table (or in Goodman 1970, for the five-way table).[37]

Before closing, we remind the reader that our methods were for the case where a given variable (e.g., variable D) can be viewed as the dependent variable which is affected by the other variables under consideration. Where this is not the case, we refer the reader to the more general techniques presented in, for example, Goodman (1970, 1972).

[36] The comments in footnotes 33 and 34 are relevant here.

[37] For further details, see Goodman 1971a.

REFERENCES

Bishop, Y. Y. M.
1969 "Full contingency tables, logits, and split contingency tables." Biometrics 25:383–400.

Boudon, R.
1968 "A new look at correlation analysis." In H. M. Blalock, Jr. and A. Blalock (eds.), Methodology in Social Research. New York: McGraw-Hill.

Coleman, J. S.
1964 Introduction to Mathematical Sociology. New York: Free Press.

Dyke, G. V. and H. D. Patterson
1952 "Analysis of factorial arrangements when the data are proportions." Biometrics 8:1–12.

Fisher, R. A. and F. Yates
1963 Statistical Tables for Biological, Agricultural and Medical Research. Sixth Edition, New York: Hafner Publishing Co., Inc.

Goodman, L. A.
1968 "The analysis of cross-classified data: Independence, quasi-independence and interactions in contingency tables with or with-

out missing entries." Journal of the American Statistical Association 63:1091–1131.
1970 "The multivariate analysis of qualitative data: Interactions among multiple classifications." Journal of the American Statistical Association 65:226–256.
1971a "The analysis of multidimensional contingency tables: Stepwise procedures and direct estimation methods for building models for multiple classifications." Technometrics 13:33–61.
1971b "Partitioning of chi-square, analysis of marginal contingency tables, and estimation of expected frequencies in multidimensional contingency tables." Journal of the American Statistical Association 66: 339–344.
1972 "A general model for the analysis of surveys." American Journal of Sociology 77 (in press).

Grizzle, J. E., C. F. Starmer, and G. G. Koch
1969 "Analysis of categorical data by linear models." Biometrics 25:489–504.

Haberman, S. J.
1970 "The general log-linear model." Ph.D. Thesis. University of Chicago.

Lazarsfeld, P. F.
1971 "Regression analysis with dichotomous attributes." Unpublished manuscript.

Rao, C. R.
1965 "Criteria of estimation in large samples." Pp. 345–362 in Contributions to Statistics. New York: Pergamon Press.

Stouffer, S. A., E. A. Suchman, L. C. Devinney, S. A. Star, and R. M. Williams, Jr.
1949 The American Soldier: Adjustment during Army Life. Studies in Social Psychology in World War II, Vol. 1. Princeton, N.J.: Princeton University Press.

Theil, H.
1970 "On the estimation of relationships involving qualitative variables." American Journal of Sociology 76:103–154.

Zeisel, H.
1968 Say It with Figures. Fifth Edition, Revised. New York: Harper & Row.

INDUSTRIAL CONFLICT AND UNIONIZATION *

DAVID BRITT
Assistant Professor
Florida Atlantic University

AND

OMER R. GALLE
Associate Professor
Vanderbilt University

American Sociological Review 1972, Vol. 37 (February):46–57

People who analyze industrial conflict from strike activity usually base their study on one overall measure of conflict, such as Kerr and Siegel's measure of strike propensity. We suggest it might be easier to identify several components of strike activity rather than one. We identify a composite variable, volume of conflict, which resembles Kerr and Siegel's Strike Propensity, and demonstrate the mathematical relationship to its components: proneness to conflict, extensity of conflict, and intensity of conflict. We analyze the relative contributions of each component to the composite variable, and conclude that extensity of conflict and proneness to conflict exert strong influences on volume of conflict, while intensity of conflict is weaker in impact. We then introduce two union variables—degree of unionization and average union size—and speculate on their effect on the various dimensions of industrial conflict We propose a tentative model which explains the unionization variables' difference in impact in terms of external support, threat potential, and factionalism.

INDUSTRIAL conflict calls to mind a wide variety of phenomena embracing vertical, horizontal, individual and collective forms. Dahrendorf (1959:236–40) treats in dustrial conflict as vertical class conflict between labor, the subject class, and management, the dominant class. Other authors note the relevance of more truncated authority distributions to variations in the degree of industrial conflict. Lopreato (1968), in testing Dahrendorf's thesis, notes great conflict within the dominant class between greater and lesser authorities. Similarly, Michels (1962:33–56) and others describe conflicts between authority levels in unions. Where authority is split among parallel hierarchies, industrial conflict is described as more nearly horizontal: Dalton (1959), focuses on line staff conflicts, Harvey and Mills (1970:181–213) and Lawrence and Lorsch (1967) refer to more general sub-unit conflict over the resource and power allocation. Finally, Udy (1967:678—709) delineates variables which channel anger modes into individual or collective forms.

* The research for this paper was supported in part by the Urban and Regional Development Center, Vanderbilt University. The writers wish to thank John McCarthy, Mayer Zald and especially Leo Rigsby for their comments.