This PDF is a selection from an out-of-print volume from the National Bureau of Economic Research

Volume Title: Annals of Economic and Social Measurement, Volume 2, number 2

Volume Author/Editor: NBER

Volume Publisher:

Volume URL: http://www.nber.org/books/aesm73-2

Publication Date: April 1973

Chapter Title: Ransacking CPS Tabulations' Applications of The Log Linear Model To Poverty Statistics

Chapter Author: Frederick J. Scheuren

Chapter URL: http://www.nber.org/chapters/c9888

Chapter pages in book: (p. 158 - 181)

Annals of Economic and Social Measurement, 2/2, 1973

RANSACKING CPS TABULATIONS: APPLICATIONS OF THE LOG LINEAR MODEL TO POVERTY STATISTICS

BY FREDERICK J. SCHEUREN

The log-linear model as developed by Goodman, Kullback, and others affords researchers a powerful tool for analyzing tabulations of survey data. Presented are some applications of the model to counts of the poor published by the Census Bureau from the annual income supplement to the Current Population Survey (CPS).

In keeping with the use of the word "ransacking" in the title, the approach is exploratory and descriptive. Formal hypothesis testing and other confirmatory techniques are dealt with only peripherally. Some attention is paid, though, to the statistical problems posed by the complex (multi-stage) nature of the CPS sample.

1. INTRODUCTION

The annual income and poverty reports, published by the Census Bureau, from the Current Population Survey (CPS) are one of the most important sources of information on the economic status of Americans. This paper takes some of the well-known techniques for fitting log-linear models to tabular material, and applies them to the CPS poverty figures. In the cases examined, the relationship between a family's poverty status and the demographic characteristics of the family head can be described quite simply and succinctly. Nearly all the information in several long and involved cross-tabulations can be summarized by the models studied.

1.1. *Formulating the Model*

To introduce the notation we will need, consider the following data taken from the March 1971 CPS.

TABLE A

NUMBER OF U.S. FAMILIES BY POVERTY STATUS, AGE, SEX, AND RACE OF HEAD

(In Thousands)

Age and Sex of Head	Poor		Nonpoor	
	White	Nonwhite	White	Nonwhite
Male-headed:				
Under 65 years old	1,821	495	34,649	2,873
65 years or older	783	181	4,896	300
Female-headed:				
Under 65 years old	959	773	2,552	651
65 years or older	138	64	737	76

Source: U.S. Bureau of the Census, *Current Population Reports*, Series P-60, No. 81, "Characteristics of the Low-Income Population, 1970" U.S. Government Printing Office, Washington, D.C., 1971 (page 67).

Table A has four dimensions: Poverty, race, sex, and age. To refer to an individual cell of the table let $N(ijkm)$ denote the total number of families having the ith poverty status ($i = 1$ if the family is poor, $i = 2$ if nonpoor), jth race ($j = 1$ nonwhite, $j = 2$ white), kth sex ($k = 1$ if family is headed by a male, $k = 2$ if head is a female) and mth age ($m = 1$ if the head is under 65, $m = 2$ if the head is 65 years or older). The true proportions of families in any cell will be denoted by

$$P(ijkm) = \frac{N(ijkm)}{N} \tag{1.1}$$

where N is the total number of families in the U.S. noninstitutional population. Estimates $\{\hat{p}(ijkm)\}$ are formed from Table A by substituting sample values for both $N(ijkm)$ and N in (1.1).

Depending on the head's age, race and sex, the odds that a given family will be poor vary considerably. For example the odds that a white male-headed family will be poor are 34,649 to 1,821 or about 19 to 1 if the head is under 65 but grow to 4,896 to 783 or about 6 to 1 if the head is 65 or more. For nonwhite male-headed families the odds are not as favorable as for whites: 2,873 to 495 if the head is under 65 and 300 to 181 if the head is 65 or more. An interesting result emerges if one looks at the relative odds' ratios for whites and nonwhites at each age level. For male heads under 65 this relative poverty ratio is

$$\frac{N(2,2,1,1)/N(1,2,1,1)}{N(2,1,1,1)/N(1,1,1,1)} = \frac{34,649/1,821}{2,873/495} = 3.28 \tag{1.2}$$

which is not too different from the ratio for families with male heads 65 or more, i.e.,

$$\frac{N(2,2,1,2)/N(1,2,1,2)}{N(2,1,1,2)/N(1,1,1,2)} = \frac{4,896/783}{300/181} = 3.77. \tag{1.3}$$

It turns out, in fact, that for any given combination of age and sex of the family head the odds of being *non*poor are about $3\frac{1}{3}$ times better for whites than for nonwhites.

1.2. *General Model Equations*

To pursue this type of analysis rigorously for Table A, the natural logorithms of the cell proportions will be fit to a model with coefficients which are functions of the relative odds' ratios considered above. In its full generality the model equation is

$$\begin{aligned} \ln p(ijkm) = {} & \beta_0 + \beta_i^{\mathrm{P}} + \beta_j^{\mathrm{R}} + \beta_k^{\mathrm{S}} + \beta_m^{\mathrm{A}} \\ & + \beta_{ij}^{\mathrm{PR}} + \beta_{ik}^{\mathrm{PS}} + \beta_{im}^{\mathrm{PA}} + \beta_{jk}^{\mathrm{RS}} + \beta_{jm}^{\mathrm{RA}} + \beta_{km}^{\mathrm{SA}} \\ & + \beta_{ijk}^{\mathrm{PRS}} + \beta_{ijm}^{\mathrm{PRA}} + \beta_{ikm}^{\mathrm{PSA}} + \beta_{jkm}^{\mathrm{RSA}} + \beta_{ijkm}^{\mathrm{PRSA}}. \end{aligned} \tag{1.4}$$

The superscripts P, R, S, and A stand for poverty, race, sex, and age respectively. The four β's having only one superscript reflect the contribution of each of the factors taken by itself. There are six β's needed to account for the factors acting in pairs; the β's with three subscripts absorb the interaction of sets of three dimensions simultaneously; $\beta_{ijkm}^{\text{PRSA}}$ is the four-way interaction.

Expression (1.4) is the usual dummy variable regression model except that the independent variables have been suppressed for the sake of brevity. Readers who find the notation troublesome should consult the footnote.[1] To have a defined system some of the coefficients must be dropped. The convention will therefore be adopted of setting to zero all β's having a "2" as any part of their subscript.

From (1.4) it can be shown that the log of the poverty odds ratio for a given age, race or sex group is

$$
\begin{aligned}
\Psi_{jkm} &= \ln\{p(1jkm)/(2jkm)\} \\
&= \ln p(1jkm) - \ln p(2jkm) \\
&= \beta_1^{\text{P}} + \beta_{1j}^{\text{PR}} + \beta_{1k}^{\text{PS}} + \beta_{1m}^{\text{PA}} \\
&\quad + \beta_{1jk}^{\text{PRS}} + \beta_{1jm}^{\text{PRA}} + \beta_{1km}^{\text{PSA}} + \beta_{1jkm}^{\text{PRSA}}
\end{aligned}
\tag{1.5}
$$

The coefficients of the logit model (1.5) are factors which taken together give the odds of a family's being poor. The overall odds are a function of β_1^{P} while the relative odds by race, sex and age are determined from β_{1j}^{PR}, β_{1k}^{PS}, and β_{1m}^{PA} respectively. The remaining four terms are corrections to these relative odds made necessary by the fact that sometimes two or more dimensions act jointly. More will be said about the interpretation of the model parameters in Section 3 where the actual numerical values for Table A are discussed.

2. Fitting the Log Linear Model

Models such as (1.4) or (1.5) can be fit in regression by (weighted) least squares [2; 26]. We will, however, employ another estimation procedure here [9; 18], one based on the theory of minimum discrimination information. While to some extent the choice between these two possible procedures is a matter of taste, there are often computational advantages to the use of information-theoretic techniques. They also can allow one to visualize in an intuitively satisfying way the implications of a particular model for the table being examined. Readers not interested in the mathematical details of the fitting algorithms can safely skip the rest of this section provided they are willing to accept our measure of fit, I^2, and use it as one could use R^2 in ordinary regression.

[1] Let

$$
X_i = \begin{cases} 1 & i = 1 \\ 0 & i = 2, \end{cases} \qquad X_j = \begin{cases} 1 & j = 1 \\ 0 & j = 2, \end{cases} \qquad X_k = \begin{cases} 1 & k = 1 \\ 0 & k = 2, \end{cases} \qquad X_m = \begin{cases} 1 & m = 1 \\ 0 & m = 2 \end{cases}
$$

then there is an exact correspondence between (1.4) and the more familiar model

$$
\begin{aligned}
\ln p(ijkm) = \beta_0 &+ \beta^{\text{P}} X_i + \beta^{\text{R}} X_j + \beta^{\text{S}} X_k + \beta^{\text{A}} X_m + \beta^{\text{PR}} X_i X_j \\
&+ \beta^{\text{PS}} X_i X_k + \beta^{\text{PA}} X_i X_m + \beta^{\text{RS}} X_j X_k + \beta^{\text{RA}} X_j X_m + \beta^{\text{SA}} X_k X_m \\
&+ \beta^{\text{PRS}} X_i X_j X_k + \beta^{\text{PRA}} X_i X_j X_m + \beta^{\text{PSA}} X_i X_k X_m + \beta^{\text{RSA}} X_j X_k X_m \\
&+ \beta^{\text{PRSA}} X_i X_j X_k X_m.
\end{aligned}
$$

2.1. *Minimum Discrimination Information*

As applied to tabulated data the Minimum Discrimination approach involves consideration of the quantity

(2.1) $$I(\hat{p}:\tilde{p}) = \sum n\hat{p}(ijkm) \ln \frac{\hat{p}(ijkm)}{\tilde{p}(ijkm)}$$

where n is the sample size, the $\{\hat{p}(ijkm)\}$ are the survey estimates of the cell proportions, and the $\{\tilde{p}(ijkm)\}$ are selected to minimize $I(\hat{p}:\tilde{p})$ subject to the restrictions imposed by the model chosen, including the requirements that

(2.2) $$\sum \tilde{p}(ijkm) = 1 \quad \text{and} \quad \tilde{p}(ijkm) > 0 \qquad \text{for all } i, j, k, \text{ and } m.$$

To see how the $\{\hat{p}(ijkm)\}$ are used to obtain the model parameters we will write (1.4) in matrix form. Let $\mathbf{y}$ be the column vector of natural logarithms of the estimated cell proportions, e.g. in Table A

(2.3) $$\mathbf{y} = (\ln \hat{p}(1, 1, 1, 1), \ln \hat{p}(1, 1, 1, 2), \ldots, \ln \hat{p}(2, 2, 2, 2))'$$

then the mathematical models to be studied can be expressed succinctly in the form

(2.4) $$\mathbf{y} = X\boldsymbol{\beta} + \mathbf{e}$$

where X is a matrix of exogenous variables (assumed to be of full rank), $\boldsymbol{\beta}$ is a vector of unknown parameters and $\mathbf{e}$ is a random variable with zero mean and variance–covariance matrix V.

Using the Minimum Discrimination approach, the estimated value of $\boldsymbol{\beta}$ is obtained from

(2.5) $$\tilde{\boldsymbol{\beta}} = (X'X)^{-1}X'\tilde{\mathbf{y}}$$

where in Table A

(2.6) $$\tilde{\mathbf{y}} = (\ln \tilde{p}(1, 1, 1, 1), \ln \tilde{p}(1, 1, 1, 2), \ldots, \ln \tilde{p}(2, 2, 2, 2))'.$$

This way of proceeding is just backwards from that in ordinary regression (with $V = \sigma^2 I$). In regression one first gets $\tilde{\boldsymbol{\beta}}$ from

(2.7) $$\tilde{\boldsymbol{\beta}} = (X'X)^{-1}X'\mathbf{y}$$

and then the "predicted" values $\tilde{\mathbf{y}}$ are given by

(2.8) $$\tilde{\mathbf{y}} = X\tilde{\boldsymbol{\beta}}.$$

2.2. *Iterative Scaling Procedure*

For the types of models we will mainly consider in this paper, a direct relationship exists between the equation one assumes and the marginal totals of the table. Broadly speaking, once one has specified what rim totals the table is to have, the model has also been determined.

The marginals needed to fit a particular model are found by examining the parameters assumed to be nonzero. For instance if

(2.9) $$\ln p(ijkm) = \beta_0 + \beta_i^{\mathrm{P}} + \beta_j^{\mathrm{R}} + \beta_k^{\mathrm{S}} + \beta_m^{\mathrm{A}} + \beta_{ij}^{\mathrm{PR}}$$

then the Poverty–Race marginal is needed since β_{ij}^{PR} is hypothesized to be nonzero.

Because this two-way marginal determines the one-way Poverty and Race marginals, estimating β_i^{P} or β_j^{R} creates no new problems. But to obtain β_k^{S} and β_m^{A} the one-way sex and age marginals must also be used.[2]

The estimated cell entries implied by the model are found by an iterative process. Commonly the initial step in a computer program is to enter "1's" in all the cells. These values are then scaled so that the table will agree with the first marginal one has specified. The resulting array is used as input to the next step where the entries are fitted to a second specified marginal. In subsequent steps the other marginals are introduced in turn. The iterative cycle may need to be repeated a number of times, each stage beginning with the cell values taken from the previous stage until the desired degree of accuracy has been achieved. Convergence is generally quite rapid.

One can also use the iterative scaling procedure to "standardize" a table's values by fitting it to a marginal or marginals taken from another table. When engaged in standardization the iteration does not begin with "1's" in all the cells, but with the original entries. For an illustration of this technique, see Table D.

2.3. *Fitting Criterion*

Considerations of parsimony make it desirable to reduce the number of estimated β's as far as possible without leaving out something "essential." To do this, reliance will be placed on a criterion [9:246] similar to R^2. Expressed in the notation of Table A, the relative information statistic I^2 is obtained as follows:

Let $\{\tilde{\tilde{p}}\}$ be the set of cell proportions estimated when fitting the model

$$\Psi_{jkm} = \ln \frac{p(1jkm)}{p(2jkm)} = \beta_1^{\mathrm{P}}. \tag{2.10}$$

Further, let $\{\tilde{p}\}$ be the set of cell proportions estimated for some other variant of (1.5), including the parameter β_1^{P}, such as

$$\Psi_{jkm} = \beta_1^{\mathrm{P}} + \beta_{1j}^{\mathrm{PR}} + \beta_{1k}^{\mathrm{PS}}. \tag{2.11}$$

It can then be shown [17] that

$$I(\hat{p}:\tilde{\tilde{p}}) = I(\tilde{p}:\tilde{\tilde{p}}) + I(\hat{p}:\tilde{p}) \tag{2.12}$$

where the $\{\hat{p}\}$ are the original estimated cell proportions. $I(\hat{p}:\tilde{\tilde{p}})$ is the *total* amount of variation in the cell frequencies which remains unexplained when we assume that the odds of being poor are constant for all groups. $I(\tilde{p}:\tilde{\tilde{p}})$ is a measure of the variation *explained* by allowing for the association (regression) between poverty, race and sex. $I(\hat{p}:\tilde{p})$ is the variation which continues to remain unexplained under model (2.11). Thus (2.12) is of the form

Total variation = Explained + Unexplained.

[2] It should be noted for future reference (page 163) that in fitting (3.4) by assumption the race-poverty effect was taken to be independent of age and sex; hence all the information about the association between them is found in the race-poverty marginal totals. Similarly the information about the age-sex-poverty effect is contained entirely in the age-sex-poverty marginal. Since from (1.4) we must also deal with relationships between age, race and sex which do not involve poverty, the age-race-sex marginal totals must be preserved. Thus to fit (3.4) a table was constructed which conformed to the marginals: poverty crossed with race, poverty crossed with age-sex, and race crossed with age-sex.

Dividing both sides of (2.12) by $I(\hat{p}:\bar{\hat{p}})$ and rearranging terms we define I^2 as

(2.13) $$I^2 = \frac{I(\tilde{p}:\bar{\hat{p}})}{I(\hat{p}:\bar{\hat{p}})} - 1 = \frac{I(\hat{p}:\tilde{p})}{I(\hat{p}:\bar{\hat{p}})}.$$

Since [17]

(2.14) $$I(\hat{p}:\bar{\hat{p}}) \geq I(\hat{p}:\tilde{p}) \geq 0,$$

then, except for the trivial case when $I(\hat{p}:\bar{\hat{p}}) = 0$

(2.15) $$0 \leq I^2 \leq 1.$$

This definition allows us to interpret I^2 in much the same way as the R^2 of standard regression. Of course, R^2 itself could have been used in assessing relative fit. However, to do so would be to introduce an extraneous element. We prefer I^2 because it is directly linked to the estimation process.

2.4. *Descriptive Use of Log Linear Model*

The approach taken to the CPS data in this paper is frankly exploratory and descriptive [e.g., 5: 23]. The use of the word "Ransacking" in the title was meant to imply this. We have not resorted to formal hypothesis testing as such. As a matter of fact, given a belief in the inherent granularity of large finite populations (like the universe of all U.S. families), one would not expect that any of the β's in a model such as (1.5) could actually be left out and still have an exact fit to data collected in a complete census. Often enough though, some of the higher-order interactions, whose meaning can be hard to get hold of intuitively, may be so close to zero that to assume that they are does not seriously impair the model's descriptive power.

With large-scale surveys, like the CPS, a subjective measure of fit such as I^2 may be a better guide for the researcher than considerations of statistical significance. For one thing when the sample size is large relative to the number of cells then substantively insignificant effects can become statistically significant. It also turns out to be quite difficult to make even approximate significance statements when the data come from complex multi-stage samples, designs which seem to be so common in practical work.

3. The Odds of Being Poor Given Age, Race, and Sex

One of the problems inherent in using the relative information, I^2, as a guide in choosing a model is deciding how large it must be for the fit to be "satisfactory." Considerations such as descriptive simplicity, the size of the table, and still other concerns all play a part in addressing what is inherently a subjective question. For situations like Table A where only a small number of cells are involved we propose to use a rather stringent criterion requiring that $I^2 \geq 95$ percent. Since poverty is relatively greater among nonwhites, among families headed by a woman or by someone 65 years or older it is natural to begin with a model which brings in all of these factors in some way. The simplest form for doing this is

(3.1) $$\Psi_{jkm} = \beta_1^{\mathrm{P}} + \beta_{1j}^{\mathrm{PR}} + \beta_{1k}^{\mathrm{PS}} + \beta_{1m}^{\mathrm{PA}}.$$

In (3.1) we posit that there is only a pairwise association between poverty and each of the other three dimensions, i.e. that the relationship between poverty and any one "independent" variable is the same no matter what values are taken on by the other two variables. To see what is meant, consider again the relative odds ratio for whites and nonwhites, as was done in (1.2) and (1.3). From (1.4) with some algebra the ratio

$$\frac{N(2,2,k,m)/N(1,2,k,m)}{N(2,1,k,m)/N(1,1,k,m)} = \frac{p(1,1,k,m)/p(2,1,k,m)}{p(1,2,k,m)/p(2,2,k,m)} = \exp\{\beta^{PR}_{11} + \beta^{PRS}_{11k} + \beta^{PRA}_{11m} + \beta^{PRSA}_{11km}\} \tag{3.2}$$

In the special case of pairwise association this ratio becomes

$$\exp\{\beta^{PR}_{11}\} \tag{3.3}$$

that is, a constant which does not vary from one age–sex combination to another.

When the pairwise associative model (3.1) was fit to Table A the relative information accounted for was 91.3 percent. At the cost of including just one more coefficient (the poverty–age–sex interaction, β^{PSA}_{1km}) a very good fit ($I^2 = 99.9$ percent) was obtained. In what follows we will discuss the latter model in some detail.

First, the fact that the poverty–age–sex interaction is nonzero indicates that it might be better to treat age and sex as just one dimension in looking at poverty since they do not act separately but jointly. Thinking of age and sex as one factor the model can be rewritten as

$$\Psi_{jr} = \beta^{P}_{1} + \beta^{PR}_{1j} + \beta^{PSA}_{1r} \tag{3.4}$$

where the $\{\beta^{PSA}_{1r}\}r = 1, \ldots, 4$ are the quantities required to account for the impact of sex and age on poverty. The actual numerical values of the β's were:

$\hat{\beta}^{P}_{1}$	$= -2.950$	(Overall poverty coefficient)
$\hat{\beta}^{PR}_{11}$	$= +1.206$	(Poverty coefficient for nonwhites)
$\hat{\beta}^{PSA}_{11}$	$= +1.952$	(Poverty coefficient for female heads under 65)
$\hat{\beta}^{PSA}_{12}$	$= +1.341$	(Poverty coefficient for female heads 65 or older)
$\hat{\beta}^{PSA}_{14}$	$= +1.134$	(Poverty coefficient for male heads 65 or older)

where we set $\beta^{PR}_{12} = \beta^{PSA}_{13} = 0$ (because of the restrictions required when using dummy variables).

The sign and size of the parameters are of course indicative of the direction and strength of the interrelationships we are studying. For example the poverty coefficient for nonwhites is $+1.206$. The positive sign means that poverty is more likely to be found among nonwhites than whites—in fact, $\exp\{1.206\} = 3.34$ more likely.

The age–sex coefficients show that the incidence of poverty is greatest among families headed by a female under 65 with families headed by a female 65 or older in second place. Not only are male-headed families less poor than female-headed ones but the pattern is also different with poverty being at its lowest for families with a male head under 65. This difference in pattern incidentally is why the effects of age and sex could not be treated additively but had to be combined.

To readers familiar with the literature on poverty none of the relationships we have been discussing are at all new. The example was in fact chosen with this in mind. It allowed us to put the emphasis on the methodology rather than on the findings.

3.1. *Interrelationships Over Time*

An example in which the results are less obvious can be constructed by looking at how stable the relationships between poverty and race, age, and sex have been over the period 1959–1970. To do this the logit model

$$\Psi_{jrt} = \beta^{P}_{1t} + \beta^{PR}_{1jt} + \beta^{PSA}_{1rt} \tag{3.5}$$

can be fit using each year's figures $t = 1959, \ldots, 1970$. All that is required for the analysis is to introduce "time" as an additional dimension of the table.

The fits obtained using (3.5) were remarkably good in each year (the average value of I^2 was 99.7 percent). However there have been considerable changes in the coefficients as can be seen from Table B. Poverty itself, of course, has declined fairly steadily from 1959 to 1969 with only a small increase in 1970.

The impact of race on poverty has also been substantially reduced as the table shows. Most of the decline in the relative incidence of poverty between whites and nonwhites occurred between 1965 and 1968, a period of quite low unemployment. Even so, except for the 1964 figure (which appears to be an anomaly) there has been some improvement from year to year in reducing the disproportionate burden of poverty borne by nonwhites.

The relative incidence of poverty by age and sex of head changed over the period we are examining but the pattern was not nearly as regular as for race. The most important movement seems to be in the growing disparity between families headed by a male under 65 and all other families. This is made evident by the fact that the coefficients for female-headed families and families headed by a male 65 or older tend to get larger and larger as time goes on. The high unemployment in 1970 reversed this trend somewhat but there are reasons to suspect it will continue over the long run due in part at least to the poverty definition itself. This definition is based on a set minimum standard, updated annually using the Consumer Price Index. Thus, as has been pointed out elsewhere [25 (81)], those dependent on fixed incomes (such as the aged) or in jobs with limited upward mobility (often women) necessarily will become a proportionately larger share of the poverty population, all other things being equal.

To summarize then, three trends have been isolated in Table B: An overall decline in the incidence of poverty, and tendencies for the declines to be relatively greater among nonwhite families and families headed by a male under 65. We will now try to assess the relative importance of each of these phenomenon. As part of this assessment the model

$$\Psi_{jrt} = \beta^{P}_{1t} + \beta^{PR}_{1j} + \beta^{PSA}_{1r} \tag{3.6}$$

was estimated. The difference between the minimum discrimination information for (3.5) and that obtained for (3.6) is, of course, a measure of the loss of fit incurred

TABLE B

RACE AND AGE SEX COEFFICIENTS FOR POVERTY MODEL (3.9)

Year	Overall Poverty Coefficient	Race and Poverty	Age, Sex and Poverty: Male 65+	Female Under 65	Female 65+
1970[r]	−2.950	1.206	1.134	1.952	1.341
1969[r]	−3.028	1.243	1.287	2.018	1.625
1968[r]	−2.922	1.256	1.142	1.900	1.457
1967[r]	−2.798	1.385	1.320	1.766	1.506
1966[r]	−2.738	1.487	1.262	1.771	1.156
1966	−2.650	1.448	1.298	1.810	1.167
1965	−2.473	1.550	1.027	1.686	1.416
1964	−2.302	1.461	0.925	1.466	1.079
1963	−2.272	1.591	0.953	1.579	1.299
1962	−2.150	1.637	0.891	1.593	1.064
1961	−2.070	1.638	0.963	1.428	1.214
1960	−2.060	1.658	0.915	1.525	1.016
1959	−2.060	1.689	1.073	1.514	1.051

[r] Based on revised methodology for processing income data as explained in Series P-60: No. 81, pp. 23–25.

Source Data for Coefficients: U.S. Bureau of the Census, *Current Population Reports*, Series P-60: No. 81, p. 67; No. 76, p. 52, No. 68, pp. 33–37.

by assuming that the relative incidence of poverty was not changing by age, race or sex. Similarly comparing the minimum discrimination information for (3.6) and

$$\Psi_{jrt} = \beta_1^{P} + \beta_{1j}^{PR} + \beta_{1r}^{PSA} \tag{3.7}$$

provides an indication of the importance over time of the change in the incidence of poverty. The difference between the minimum discrimination information for (3.5) and (3.7) provides an overall measure of the total lack of fit from all causes. When one examines this total, 90.1 percent is due to uniform shifts in the general incidence of poverty in the population. Only 9.9 percent is the result of changes in the relative incidence of poverty among age–race–sex groups. Of this remainder about one-third of the lack of fit is due to changes in the race effect and two-thirds to changes by age and sex of head.[3]

At first glance there would seem to be some problem in squaring the above analysis with the figures in Table C which show that all of the decline in the number of poor families has occurred among those with male heads; in fact the number of poor female-headed families has actually increased slightly.

The logit model and its corresponding coefficient estimates depend on the relative number of poor families within each age, race and sex class. They are only indirectly affected by the counts in the individual cells being examined. On the other hand, Table C summarizes the net result of both an altered pattern in the incidence of poverty and also changes in the relative sizes of various demographic groups and of the overall total number of families.

[3] It should be mentioned that the relative importance of each of these causes is *not* independent of the order in which they are examined. The sequence followed makes a difference as it does in regression.

TABLE C

NUMBER OF POOR FAMILIES BY SEX OF HEAD, 1970 AND 1959

(In Thousands)

Sex of Head	1970	1959	Change 1959 to 1970
Total	5,214	8,320	−3,106
Male	3,280	6,404	3,124
Female	1,934	1,916	+18

Source: U.S. Bureau of the Census, *Current Population Reports*, Series P-60, No. 81, p. 29.

Table D below was created in an attempt to sort out all the factors acting on the poverty totals.[4] However, the partialing out of the importance of any one change cannot be done independently of the others. Thus the adjustments shown in Table D are conditional in nature. Each represents the net additional change made by a factor given the other factors whose effects have already been taken account of. Despite this limitation it may be useful to compare the differential impact of

TABLE D

ELEMENTS OF THE 1959 TO 1970 SHIFT IN THE NUMBER OF POOR FAMILIES

(In Thousands)

Item	Total	Male-Headed Families	Female-Headed Families
Poor Families in 1959	8,320	6,404	1,916
Population Composition Changes:			
Growth overall	+1,282	+987	+295
Race	+186	+113	+73
Sex	+181	−198	+379
Age	+85	+55	+30
Poverty Incidence Changes:			
Decline overall	−4,874	−3,832	−1,042
Race	−485	−298	−187
Age and sex	+519	+49	+470
Poor families in 1970	5,214	3,280	1,934
Net Changes, 1959 to 1970	−3,106	−3,124	+18

Note: The adjustments are not independent of the order in which they were made. Rather each line represents the net change obtained by altering an additional factor. The population composition changes were derived by a sequential standardization process. First the overall 1959 table's total was increased to agree with that for 1970 then the marginal totals by race were made to agree with those for 1970. The increase in the number of poor families caused by this change was then derived. The next step was to force the 1959 table to agree with the 1970 race sex marginals and finally with the 1970 age–race sex marginal table.

[4] Methodological improvements in the collection and processing of the CPS also had an effect on the poverty totals. Adjustments for this have not been made separately.

population composition and poverty incidence changes on male and female-headed families.

Since 1959 there has been an overall 15 percent growth in the number of U.S. families. The increase has been somewhat faster for nonwhites than for whites. The most important change though is the quite rapid growth of female-headed families relative to those headed by a male. There were also changes in the proportion of male and female-headed families by age of head with male heads being older and female heads younger in 1970 than 1959. If one does not allow for the lowering in the incidence of poverty over the period then these changes have the cumulative effect of increasing the number of poor male-headed families by 15 percent and the number of poor female-headed families by 41 percent.

However there has been, as the table shows, an overall decline in the incidence of poverty for both male and female-headed families. This is not apparent in the overall 1959–1970 differences because population composition changes swamp the relative decline for female-headed families.

4. The Odds of Being Poor Given Education and Work Experience

In this section we will examine the relationship between family poverty and the educational attainment of the head. Two 5-way tables will be looked at: The classifiers for the first are race (Black, Nonblack), Poverty, Sex, Age (25 to 34 years, 35 to 44, 45 to 54, 55 to 64, 65 or more) and highest grade completed (Less than 8 grades, 8 grades, 9 to 11, High school graduate, some college). The second table is exactly the same as the first except that in place of race the family head's work experience (Year-round full-time, other) is used as a classifier. (These tabulations, like Table A, are from the 1970 CPS Poverty Report, Series P-60, No. 81.)

Several purposes are served by introducing these additional examples. Both are tables of moderate size (200 cells) and differ in other ways from the small (16 cells) table just studied. For one thing, two of the dimensions (age and education) can be treated as quantitative rather than strictly qualitative variables if so desired. Perhaps the most important topic we will take up is how one can combine the results of the separate analyses into one overall model.

4.1. *Model Notation*

The two tables to be studied can be dealt with in a unified way. Each is a (5-way) marginal of the 6-way table formed by the factors; age, sex, race, education, work experience and poverty status. Even though the more detailed tabulation is not available to us it is convenient to set up our definitions as if it were. Therefore let $\hat{p}(ijkmr)$ be the estimated cell proportions of the overall table where $i = 1, 2$ is used to designate a family's poverty status, $j = 1, \ldots, 10$ is a combined index identifying the family head's age and sex:[5] $k = 1, \ldots, 5$ denotes the educational attainment of the head: and $m = 1, 2$ and $r = 1, 2$ are used to identify the head's race and work experience respectively.

[5] In effect, combining age and sex reduces the 6-way table we started with to simply 5 distinct dimensions. Age and sex are treated as one dimension since, as we saw in Table A, they act jointly in determining a family's poverty status.

The cell proportions in the published tables can be defined as

$$
\hat{p}(ijkm\cdot) = \sum_{r=1}^{2} \hat{p}(ijkmr)
$$

(4.1)

$$
\hat{p}(ijk\cdot r) = \sum_{m=1}^{2} \hat{p}(ijkmr)
$$

Let us now consider two dummy variable logit models with the odds of being poor as the "dependent" variable—one based on the table having race as a classifier, the other based on the table separating families by the work experience of the head. Adhering to the notation established earlier in this paper these models can be expressed by

(4.2)
$$
\begin{aligned}
\Psi_{jkm} &= \ln\{p(1jkm\cdot)/p(2jkm\cdot)\} \\
&= \beta_1^{P} + \beta_{1j}^{PSA} + \beta_{1k}^{PE} + \beta_{1m}^{PR}
\end{aligned}
$$

and

(4.3)
$$
\begin{aligned}
\Psi_{jkr} &= \ln\{p(1jk\cdot r)/p(2jk\cdot r)\} \\
&= \beta_1^{P} + \beta_{1j}^{PSA} + \beta_{1k}^{PE} + \beta_{1r}^{PW}
\end{aligned}
$$

(The dimensions not in our first example are identified by the super-scripts "E," education, and "W," work experience.)

4.2. *Goodness of Fit*

Despite the fact that the above equations do not include any high-order interaction terms, they seem to represent an adequate summary of the relationship between poverty incidence and the other variables. The relative amounts of explained variation were $I^2 = 96.2$ percent for (4.2) and $I^2 = 95.8$ percent for (4.3).

The reader might find the I^2 value for (4.2) inconsistent with the much better fit (99.9 percent) obtained earlier in (3.4). After all both models include age, race and sex and (4.2) also includes education. Arguing from the similarity we said exists between R^2 and I^2 one's expectation would be that the fit for (4.2) would be better, not worse.

The apparent anomaly is explainable chiefly by taking account of the differences in the sizes of the tables being used.[6] In fitting (3.4) to Table A there are only 16 cells involved and five (poverty) parameters were needed for the model. With (4.2) we have a 200 cell table to describe and do so quite well with just 15 parameters. To properly compare models (3.4) and (4.2) the fitting should be done using the same table for both. When this was tried age, sex and race taken together had an I^2 value of 68.7 percent as compared to the 96.2 percent fit obtained with education added.

The situation we are discussing is an instance of what happens when one goes from one level of aggregation to another. Commonly the amount of "noise" in our figures grows relatively faster as we disaggregate than does the amount of

[6] Differences between the two tables in the classifications used for the race and age variables also play a minor role.

additional information obtained. A well-known example of this phenomenon can arise with R^2 itself when one looks at the same relationship in a cross-section or over time. The R^2 value is typically smaller with the cross-section data. Disaggregation tends to raise the importance of "accidental" factors and thus lower R^2 (or I^2).

4.3. *Coefficient Estimates*

Rather than display all the coefficients for models (4.2) and (4.3) we will look only at education and age to see to what extent these dimensions can be treated as quantitative.

The education coefficients are shown in Table E below. Both sets of coefficients are in reasonably close agreement and exhibit the expected pattern of getting

TABLE E

EDUCATION COEFFICIENTS FOR MODELS (4.2) AND (4.3)

Notation	Equation (4.2)	Equation (4.3)	Interpretation
$\hat{\beta}^{PE}_{11}$	+1.026	+1.038	Poverty coefficient for heads with less than 8th grade education.
$\hat{\beta}^{PE}_{12}$	+0.327	+0.274	Poverty coefficient for heads who completed the 8th grade.
$\hat{\beta}^{PE}_{13}$	0	0	Coefficient for those with some high school (set to zero by definition).
$\hat{\beta}^{PE}_{14}$	−0.706	−0.634	Coefficient for High School Graduates.
$\hat{\beta}^{PE}_{15}$	−1.123	−1.018	Coefficient for heads who completed one or more years of college.

smaller (algebraically) as the head's education increases. What is not clear is how we can incorporate the actual values for highest grade completed in explaining the relationship to poverty. However, if attention is confined to the rank order of the classifications then a fairly satisfactory model for the poverty–education interaction is given by

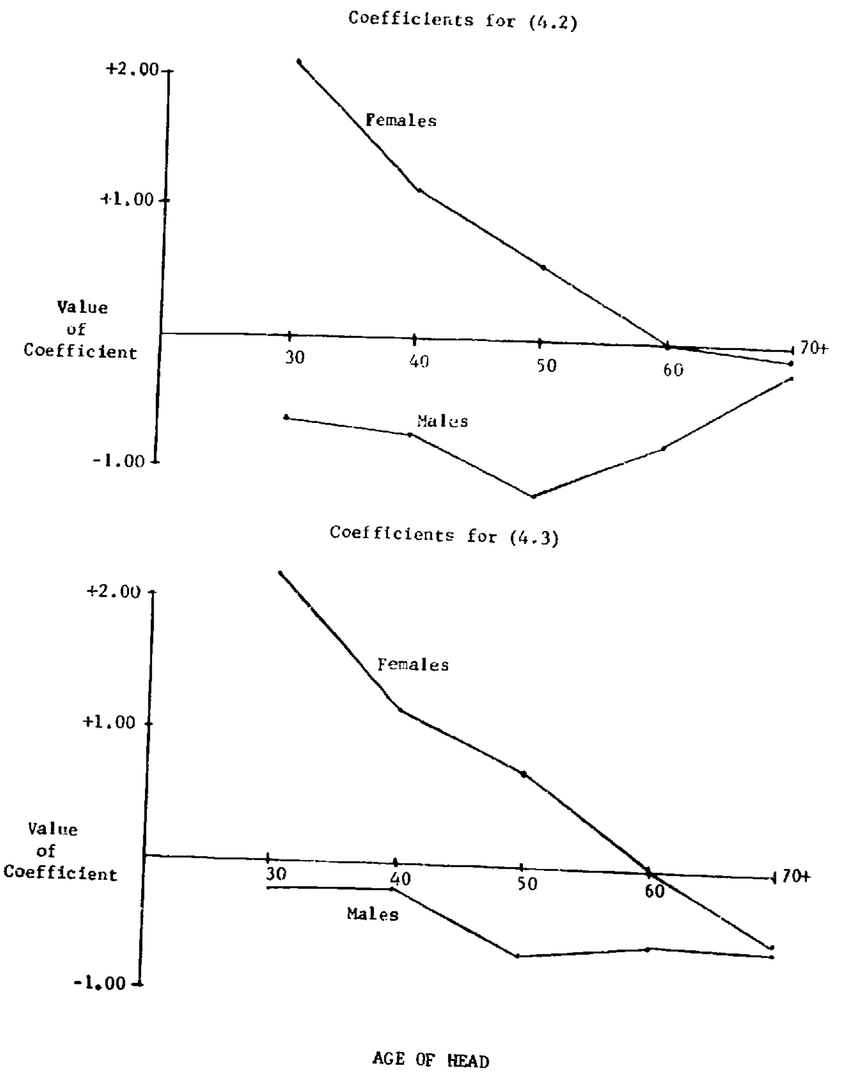
$$\beta^{PE}_{1k} = \beta(k-3) \qquad \text{for } k = 1, \ldots, 5. \tag{4.4}$$

Whether one would actually resort to (4.4) as a summarization device is open to question but it does point up the fact that education is an ordinal rather than an interval-scaled variable. (After all it is simply not true that the difference between an eleventh and twelfth grade education is the same as the difference between completing the tenth and eleventh grades.)

Chart A displays the age–sex–poverty coefficients graphed against the middle of the age bracket to which they apply. In every case the coefficients for female-headed families are larger than those for families headed by a male. The (log) odds of being poor seem to decline with age in a regular (almost linear) fashion for female-headed families. This pattern is strikingly similar for (4.2) and (4.3), perhaps due to the infrequency with which female heads work year-round full-time.

CHART A

AGE SEX COEFFICIENTS FOR (4.2) AND (4.3)



For male-headed families, the age–poverty coefficients are affected not only by the head's labor force participation and earnings which tend to grow until middle life but also by contributions to the family income of working wives.

4.4. *Combining Tables*

In order to incorporate race and work experience together in a logit model with poverty, age–sex and education, all six dimensions must be cross-classified. As we have already mentioned, such a 6-way table is not available. However there is an option short of rerunning the survey data tapes which can be employed to create the needed tabulation. What will be done is to use the published marginals to obtain a *fitted* version of the table sought. Obviously such a procedure will be satisfactory only under certain assumptions.

For the particular example at hand three 5-way marginals were available the two we have been discussing and a table crossing age, sex, race, and work experience of the head with the family's poverty status [25 (81)]. These three tables were then incorporated as marginals in the usual iterative fitting process to produce the needed overall table.

The model

$$\begin{aligned} \Psi_{jkmr} &= \ln \{p(1jkmr)/p(2jkmr)\} \\ &= \beta_1^{P} + \beta_{1j}^{PSA} + \beta_{1k}^{PE} + \beta_{1m}^{PR} + \beta_{1r}^{PW} \end{aligned} \tag{4.5}$$

was then derived from the constructed table with the value of the relative information being $I^2 = 94.3$ percent.

Implicit in the way we created the overall table is the assumption that the relationship between poverty and the other factors is simple enough to be adequately mirrored in the three marginals we possess when taken together. While the estimates of (4.5) are not themselves affected by the validity of this assumption, we may be mislead as to how good a summary the model represents. After all in the overall fitting process some smoothing takes place which necessarily reduces the amount of residual error. Thus I^2 as computed above should be considered only an upper bound, although in this case one may guess that it does not overestimate the true value by very much.

A second assumption is made by the procedure just outlined. Not only are some poverty relationships disregarded but there are also interrelationships among the other factors which are ignored. In particular, the race-work experience-education interaction is treated as if it were zero. Table F illustrates the effect on the poverty coefficients of different assumptions about how the nonpoverty factors vary. The first column provides the greatest possible interaction given the way the overall table was constructed. Column two was derived by letting the nonpoverty factors interact in sets of three (with the exception already noted). The third column allows the nonpoverty factors to interact only in pairs and the last column treats the nonpoverty factors as if they were conditionally independent.

The agreement between the first two methods (columns one and two) is extremely good. Even when the fit is confined just to two-way relationships the coefficients are not badly off. In this instance, there does not seem to be much sensitivity in our estimates to relationships of order higher than two. As the last column of the table demonstrates, however, we cannot ignore interrelationships among the nonpoverty factors altogether.

It might be noted in passing that the coefficients obtained under the assumption of conditional independence are the same values one would obtain if looking at each dimension's contribution to poverty without regard to how much of the association is explained by the joint action of several factors.[7] To be specific, consider the poverty parameter for blacks in the tables we have examined. The net

[7] The distinction being made here is the same as that between the coefficient of an independent variable in a simple or a multiple regression.

TABLE F

POVERTY COEFFICIENTS FOR AGE, RACE, SEX, WORK EXPERIENCE AND EDUCATION OF HEAD COMPUTED USING ALTERNATIVE STANDARDIZATION TECHNIQUE

Coefficients	Type of Fit (Marginals Employed)			
	Three 5-Way Marginals	All Possible 3-Way Marginals*	All Possible 2-Way Marginals†	Two-Way Poverty Marginals Only†
Overall Poverty Coefficient	−1.518	−1.517	−1.446	−0.594
Male Heads:				
25 to 34 years	−0.189	−0.190	−0.266	−1.432
35 to 44 years	−0.129	−0.129	−0.184	−1.300
45 to 54 years	−0.574	−0.576	−0.608	−1.600
55 to 64 years	−0.491	−0.491	−0.551	−1.114
65 years or older	−0.441	−0.441	−0.490	−0.215
Female Heads:				
25 to 34 years	+2.068	+2.067	+2.021	+1.572
35 to 44 years	+1.253	+1.252	+1.215	+0.774
45 to 54 years	+0.665	+0.664	+0.644	+0.303
55 to 64 years	0.000	0.000	0.000	0.000
65 years or older	−0.380	−0.380	−0.427	+0.100
Education completed				
Less than 8 grades	+0.939	+0.938	+0.911	+0.933
8th grade	+0.319	+0.319	+0.295	+0.109
9 to 11 grades	0.000	0.000	0.000	0.000
12th grade	−0.545	−0.544	−0.562	−0.567
Some college	−0.879	−0.880	−0.903	−1.392
Poverty Coefficient for Negroes	+0.855	+0.855	+0.856	+1.397
Poverty Coefficient for year-round workers	−1.638	−1.638	−1.646	−2.039

* Except the work experience–Education–Race marginal.
† Age and sex are treated as one dimension.

overall disadvantage of being black is summarized by the value $\beta_{11}^{PR} = +1.397$; when the contributions to this differential due to age, sex, education and work experience are taken out, the poverty–race relationship declines to $\beta_{11}^{PR} = +0.855$.

4.5. *Some Analytic Issues*

The subject of combining tables is an important one especially when consideration is given to the nature of the CPS figures we have been using. In government-conducted surveys, like the CPS, traditionally results have been displayed only in tabular form with the information on individual schedules not being subjected to further examination. For example, published CPS data on the distribution of personal income (in Series P-60) exists from 1947 on but only in recent years, beginning with 1964, has there been any release by the Census Bureau of the complete survey files.[8] Thus researchers interested in looking at relatively long-

[8] Computer files with some information on families (but not individuals) exist from 1959 income year on. For both families and persons identifying items have been removed to protect the confidentiality of the interview.

term shifts in income patterns must employ techniques like those in this paper for dealing with grouped data.

For the earlier years the published tabulations are not extensive enough to look at more than two or three variables at a time. Even using the 1970 CPS poverty tabulations. which were quite voluminous. one cannot study relationships of order higher than that already dealt with above. Without at least two-way tables relating all the variables it would seem that the only course open to us is to prepare a number of separate (incomplete) analyses. An alternative exists however which we can only just mention for reasons of space. This is to standardize the published historical material with data taken from more recent surveys. There are interpretative issues which must be faced in adopting such a procedure but useful results can emerge. In biological and medical settings and in demography, standardization techniques are widely accepted; perhaps they have a role to play with CPS income data as well. A paper on this subject with some empirical findings is in preparation.

5. Bias and Mean Square Error of Model Coefficients

Fitting log linear models. as we have tried to show through some examples. provides the researcher with a powerful data analysis tool for describing a surveyed population. What have not been dealt with are the statistical properties of the figures obtained. This section will investigate such properties—in particular. the bias and variance, or more precisely mean square error. of the logit model coefficients.

5.1. *Bias in Coefficient Estimates*

In regression analysis. bias in the coefficient estimates is often discussed in terms of errors made in specifying the model. Such a context is inappropriate here because we are just using the logit fitting process as a device for summarizing interrelationships among factors in the finite population from which the observations were drawn. Ignoring some of the more complicated interactions. as we have said, does not necessarily imply acceptance of the hypothesis that they do not exist but rather that a "satisfactory" parsimonious description (as measured by I^2) can be achieved without them.

However, even with misspecification error ruled out. the coefficient estimates $\{\hat{\beta}$'s$\}$ are biased. Nonetheless under quite general conditions it can be shown that the expected value of $\hat{\beta}$. denoted $E\hat{\beta}$. is

$$E\hat{\beta} = \beta + 0\left(\frac{1}{n}\right) \tag{5.1}$$

where the term $0(1/n)$ goes to zero as the sample size "n" gets large.

Some situations for which (5.1) does not hold may be worth mentioning. If the sample elements were not selected with equal probability. then preparing the cell proportions using the *unweighted* counts will lead to a bias which may not disappear with increasing sample size. In a stratified cluster design, like the CPS, (5.1) may not apply to small subpopulations concentrated in parts of the country (e.g. outside the big cities) which are not included with certainty. The difficulty is that the number of sampled areas or PSU's must be "large," not just the number of families or individuals in the survey. A final note of caution should be sounded in cases where the marginals being used to obtain the model coefficients contain one or more cell entries which are close to zero. Two methods for alleviating this last type of bias, which is of $O(1/n)$, will be discussed below.

Bias Reduction

One method of bias reduction which is often advocated [e.g. 9: 229-230] involves adding a small amount, usually $1/2n$, to the original cell proportions before fitting the table. Only in one very special case can such a technique be shown to be beneficial, namely when all the β's are assumed nonzero. (The assumption of simple random sampling is also required.) In point of fact, adding a fixed amount to every cell can actually be harmful when fitting models in which some of the coefficients are set to zero.[9]

A far more general bias reducing procedure is a method called the "Jackknife," by Tukey [19: 134], "to suggest the broad usefulness of the technique as a substitute for specialized tools.... just as the Boy Scout's trusty tool serves so variedly." To see how the Jackknife can be applied to survey data let us assume that the overall sample can be divided into "r" independent subsamples or replicates each identical in design and of size "$\bar{n}$."

The Jackknifed coefficients are defined by

$$\hat{\boldsymbol{\beta}} = \frac{1}{r} \sum_{k=1}^{r} \hat{\boldsymbol{\beta}}_k \tag{5.2}$$

with

$$\hat{\boldsymbol{\beta}}_k = r\tilde{\boldsymbol{\beta}} - (r-1)\tilde{\tilde{\boldsymbol{\beta}}}_k \tag{5.3}$$

where $\tilde{\boldsymbol{\beta}}$ is the estimator we have been discussing all along and the $\{\tilde{\tilde{\boldsymbol{\beta}}}_k\}$ are constructed just like $\tilde{\boldsymbol{\beta}}$ except instead of adding together all "r" replicates the fit is obtained with only $r-1$ of them, i.e. by leaving out the kth, $k = 1, \ldots, r$. Now if the

$$\text{Bias}\,\{\tilde{\boldsymbol{\beta}}\} = \frac{r-1}{r}\,\text{Bias}\,\{\tilde{\tilde{\boldsymbol{\beta}}}_k\} + 0\left(\frac{1}{n^2}\right) \tag{5.4}$$

[9] Adding small amounts to cells is also suggested in the literature on contingency tables for dealing with *zero* cells (e.g. [6]). Zeroes can be a serious problem in applied work when they are found in the marginals one wishes to fit. For example in creating the 6-way table of the previous section there were a few zeroes in the 5-way marginals. Arbitrarily a small amount was added to each cell. The analyses of the coefficients in Table F shows that in this case the zeroes made very little difference; however, that will not always be true, particularly when there are a great many. It should be recognized that when the marginal cell proportions are very small the coefficient estimates can be quite unstable and very large samples will be needed to obtain satisfactory results.

then

$$(5.5)\qquad \begin{aligned} E\hat{\beta} &= \frac{1}{r}\sum_{k=1}^{r} E\hat{\beta}_k \\ &= r[\beta + \text{Bias}\,\{\tilde{\beta}\}] \\ &\quad -(r-1)\left[\beta + \frac{r}{r-1}\,\text{Bias}\,\{\tilde{\beta}\}\right] + 0\left(\frac{1}{n^2}\right) \\ &= \beta + 0\left(\frac{1}{n^2}\right). \end{aligned}$$

The CPS is not made up of independent identically-designed subsamples [24], so if the Jackknife is to be applied at all certain practical compromises are necessary. One way of Jackknifing in the CPS is to divide the overall sample into "replicates" on the same lines that are used to create the eight rotation groups which make up each month's survey. Such subsamples, while identical in design, would not be independent.

Dependence among the replicates makes it impossible for (5.4) to be satisfied: nonetheless, given the nature of the CPS, it can be shown that appreciable reductions in the absolute value of the expected bias may still be achieved by Jackknifing, making the extra trouble taken worthwhile (particularly for large tables where the average cell size is small).

A numerical illustration of the Jackknife appears in Table G below. For purposes of the example the CPS rotation panels for March, 1971, were considered

TABLE G

ILLUSTRATIVE JACKKNIFED RACE AND AGE-SEX COEFFICIENTS FOR POVERTY MODEL (3.4) USING EIGHT "REPLICATES"

Item	Overall Poverty Coefficient	Race and Poverty	Age, Sex and Poverty		
			Males 65 +	Females Under 65	Females 65 +
Original coefficients, $\hat{\beta}$	−2.9496	1.2062	1.1337	1.9521	1.3407
Jackknife Average, $\hat{\beta}$	−2.9488	1.2077	1.1333	1.9504	1.3358
Individual values:					
$\hat{\beta}_1$	−3.1135	1.4468	1.2519	1.9730	1.2826
$\hat{\beta}_2$	−2.8936	1.0256	1.1639	1.9491	1.7346
$\hat{\beta}_3$	−2.9454	1.1881	0.9913	2.0197	1.3157
$\hat{\beta}_4$	−2.9314	1.2061	1.0792	2.0339	1.4311
$\hat{\beta}_5$	−2.9554	1.2116	0.9402	1.9378	1.4389
$\hat{\beta}_6$	−2.9054	1.0705	1.1705	1.8225	1.3611
$\hat{\beta}_7$	−2.8881	1.2555	1.2168	1.8703	1.0255
$\hat{\beta}_8$	−2.9579	1.2574	1.2529	1.9965	1.0968

Note: For the sake of convenience the coefficients $\{\hat{\beta}_k\}$ were constructed using the CPS rotation panels rather than subsamples selected to be identical. Although all the panels start out the same in terms of the way they are drawn, at any one survey point each rotation group will have been interviewed a different number of times. Since re-interviewing has some effect on response patterns, using the panels as "replicates" would not be desirable in general. Technically (see, for example [22]), each replicate should be weighted using the same scheme that is applied to the overall sample. This refinement was also skipped since the figures are only meant to be illustrative. Instead the estimates were prepared simply using the already existing weights.

to be identically designed (dependent) replicates and Jackknifed poverty coefficients for Table A were derived. Although some of the fine points have been ignored (as the note to Table G makes clear), the figures shown may be of interest.

There are only slight differences between our original estimates and the Jackknife average, something one could almost have predicted ahead of time given the smallness of the table and the size of the sample. The differences also exhibit the expected pattern of being *larger* for coefficients based on marginals which are *smaller*.

5.3. *Variance of Coefficient Estimates*

A convenient way of dealing with any study's variances $\{v_i^2\}$ is to relate them to the variances $\{\sigma_i^2\}$ one would have obtained from a sample random sample (with replacement) of exactly the same size. This can be done using the expression

$$v_i^2 = \delta_i \sigma_i^2 \tag{5.6}$$

where, following Kish [15:258], the $\{\delta_i\}$ are called "design effects."

Typically in a cluster sample the $\{\delta_i\}$ are larger than one. For example, in the CPS when looking at proportions the estimated simple random sampling standard errors sometimes understate the actual standard errors by as much as 50 percent or more. The variances of logit coefficients are related to the variances of the table's cell proportions. Thus, unless some adjustment is made to the sample random sampling estimates normally computed, confidence interval statements will be off. (For the 1970 poverty tabulations analyzed in this paper the square root of the design effect for proportions averaged about $\sqrt{\delta} = 1.23$.)

5.4. *Calculating Variances*

The standard survey approach to the variance of a nonlinear function, like $\tilde{\boldsymbol{\beta}}$, involves the use of a Taylor expansion. One either implicitly or explicitly depends on being able to express the statistic, to a close approximation, as a linear combination of sample means and totals. Variance calculations based on replication or jackknifing are comparatively easy since they only *implicitly* rely on the Taylor Series results. Procedures which require that the expansion be exhibited *explicitly* will not be discussed in this paper since they are too difficult to apply routinely as part of the analysis of a contingency table. Instead we will briefly deal with three "short-cut" techniques which, as applied to the CPS, yield approximations good enough for most purposes.

The first and best known "short-cut" method of estimating variances involves replication. If the overall sample is made up of "r" independent identically designed subsamples, one can obtain an estimate of the variance-covariance matrix of $\tilde{\boldsymbol{\beta}}$ by deriving the coefficients $\tilde{\boldsymbol{\beta}}_k$ for each replicate and using

$$\tilde{V}(\tilde{\boldsymbol{\beta}}) = \frac{1}{r(r-1)} \sum_{k=1}^{r} (\tilde{\boldsymbol{\beta}}_k - \tilde{\boldsymbol{\beta}}.)(\tilde{\boldsymbol{\beta}}_k - \tilde{\boldsymbol{\beta}}.)' \tag{5.7}$$

where $\tilde{\boldsymbol{\beta}}.$ is the average of the replicate values, i.e.

$$\tilde{\boldsymbol{\beta}}. = \frac{1}{r} \sum_{k=1}^{r} \tilde{\boldsymbol{\beta}}_k. \tag{5.8}$$

A related method which also produces an asymptotically unbiased variance estimator of $V(\hat{\beta})$ is to use the Jackknife values $\bar{\bar{\beta}}_k$ in the calculating formula

(5.9) $$\tilde{V}(\bar{\beta}) = \frac{r-1}{r}\left[\sum_{k=1}^{r} \bar{\bar{\beta}}_k \bar{\bar{\beta}}_k' - r\bar{\bar{\beta}}.\bar{\bar{\beta}}'.\right].$$

where

(5.10) $$\bar{\bar{\beta}}. = \frac{1}{r}\sum_{k=1}^{r} \bar{\bar{\beta}}_k.$$

Both of these methods suffer from the disadvantage that the variance of the variance estimator can be large. This. of course. is the price one pays for ease of computation. Of the two. the Jackknife is to be preferred because it will be less sensitive to the problem of zero cells which can arise when looking at the sample replicate by replicate.

As we have seen. since the CPS cannot be divided into independent identically designed subsamples the replicate and Jackknife variance estimators are not strictly appropriate. However. if the eight rotation panels are treated as independent. the resulting standard errors calculated are underestimates. For most statistics. except those based heavily on persons living outside metropolitan areas. an upward adjustment in the standard deviation on the order of 6 percent is required. For nonmetropolitan area statistics somewhat larger correction factors should be used.[10]

For researchers using only the published CPS tables. perhaps the best that can be done is to calculate the simple random sampling variance $\tilde{\Sigma}$ and then correct it with an adjustment factor derived from the standard error tables which accompany all CPS reports. $\tilde{\Sigma}$ is obtained [18] by first calculating the quantity $(X'TX)^{-1}$ where T is a diagonal matrix of the table's weighted cell counts as fitted under the model and X is the array of independent factors in equation (2.4). Dropping the first row and column of $(X'TX)^{-1}$. one then obtains W times $\tilde{\Sigma}$ where "W" is the average sampling weight.

For proportions. the published CPS standard error tables are calculated using the expression

(5.11) $$\text{Standard Error of } \hat{p} = \left\{\frac{b}{Y}\hat{p}(1-\hat{p})\right\}^{1/2}$$

where Y is the estimated total number of persons or families in the subpopulation (e.g. black males) to which the proportion applies. "b" plays a role similar to the design effect and in fact

(5.12) $$\frac{b}{Y} = \frac{(b/W)}{(Y/W)} = \frac{\delta}{n}.$$

[10] CPS tapes can be bought from the Census Bureau that allow one to calculate variances based on the collapsed stratum technique. Collapsing strata, however. often leads to an *overestimate* of the variance. See [1]. [11] and [21] for details and a discussion of still other methods.

For example, the value of $b = 2{,}074$ was used to create generalized standard error estimates for proportions of families in the 1970 CPS report [25 (81)]. Since the average weight for families was 1.372, the overall design effect for proportions is $\delta \doteq 1.5$.

The work of Kish and Frankel [16] suggests that it would be unwise to simply apply the "δ" appropriate for proportions to $\hat{\sum}$. For the usual regression parameters. Kish found that, on the average, the increase in the standard error for a complex design was 6 percent or about one-third of that for sample means (17 percent). Using this result as a guide, the effect for proportions ($\sqrt{\delta} = 1.23$) in the 1970 CPS report was reduced to $1.00 + (0.23)(\frac{6}{17}) = 1.08$ when calculating the standard errors of the β's in Table H.

Table H compares standard error estimates for the CPS poverty coefficients obtained as part of our analysis of Table A. All three approaches are in quite close agreement, considering the rough nature of the approximations employed. Further work on the validity of these methods is needed however, and the reader is cautioned to take the results in Table H only as illustrative.

TABLE H

ILLUSTRATIVE STANDARD ERROR ESTIMATES: 1970 RACE AND AGE-SEX COEFFICIENTS FOR POVERTY MODEL (3.4)

Type of Standard Error Estimate	Overall Poverty Coefficient	Race and Poverty	Age, Sex, and Poverty		
			Males 65+	Females Under 65	Females 65+
Replicate	0.0285	0.0490	0.0488	0.0489	0.0894
Jackknife	0.0269	0.0509	0.0449	0.0501	0.0807
Adjusting Simple Random Sampling	0.0288	0.0478	0.0526	0.0482	0.1052

Note: Replicate and Jackknife estimators were calculated by treating the 8 CPS rotation panels as independent. A correction factor was then applied as is explained in the text. The simple random sampling errors were adjusted by 1.08 before being shown. See the note to Table G for further limitations on these results.

6. COMPUTER PROGRAMS AND BIBLIOGRAPHICAL NOTES

The models fit in this paper have a simple dummy variable structure. However the computer programs employed are applicable to more complicated parameterizations [4]. There is also no necessity, for instance, to look only at logit models where the "dependent" dimension (in our case poverty) is dichotomous; polychotomous dependent variables present no new problems [9:238].

6.1. *Computer Programs*

At the Office of Economic Opportunity (OEO) three contingency table programs for fitting log linear models are in use. Two of these are for batch processing on an IBM 360/50 and the third is an APL program. All were developed at the George Washington University Statistics Department. C. Terence Ireland wrote the first of these programs—CONTAB II [12]. A main feature of this algorithm is that there is practically *no* limit (except CORE) as to the size of the table which can be analyzed. Marian Fisher modified CONTAB II to increase its flexibility still further. Her program CONTAB MOD [7] allows the researcher to fit general models, not just dummy variable ones. Also marginal totals can be introduced from outside the sample. In addition to these, Ireland prepared an APL contingency table package which has since been augmented at the Office of Economic Opportunity by H. Lock Oh. As yet the APL program is restricted to tables of less than 500 cells.

Future refinements in some or all of these programs are anticipated. In particular, we are looking at the possibility of modifying the iteration scheme so that it can deal efficiently with stratified designs where the probabilities of selection vary considerably from stratum to stratum. So long as the sampling weights are used, the present iterative procedure gives asymptotically unbiased coefficients; but, if the weights differ widely from cell to cell, competitive techniques exist which can yield estimates having smaller variances [14]. Since the CPS begins as a "self-weighting" sample no modification of the standard fitting procedure was deemed necessary for the work presented in this paper.

6.2. *Bibliographical Notes and Acknowledgements*

Lack of space has lead us to slight many aspects of log-linear model fitting. For example much more could be said about methods for hypothesis testing with survey data, e.g. [20], and their implications. We have only dealt with this indirectly by looking at the variances of a model's coefficients. The implicit assumption has been made that approximate normal theory confidence intervals for the coefficients can be constructed using the estimated standard errors (once corrected for design effects). Another important part of the theory which needs to be considered is the examination of residuals and the suppression of outliers [13].

The title of this paper comes in part from a 1969 article by Goodman [8], "How to ransack social mobility tables and other kinds of cross-classification tables." Ransacking seemed just too good a word not to use again, especially since it so aptly conjures up the kind of hunting for relationships that researchers must engage in if they hope to tap the riches of data like that obtained from the Current Population Survey. There are, of course, elements of subjectivity in such a search. It was because of this subjectivity that the statistic I^2 was used. Unlike R^2, it is linked closely with the fitting process and for this reason to be preferred. A full discussion of the development and properties of the class of measures of which I^2 is a member can be found in Goodman [e.g., 9: 246; 10: 42–44].

The nature of an applied paper is to take many results for granted. Such is the case here. Heavy reliance has been placed on ideas to be found in Goodman [9] and Kullback [18]. The writer has also profited at various points from conversations with Dr. Ireland and Dr. Kullback. Editorial and other assistance were provided by Wray Smith, Gary Liberson and Lock Oh of OEO and Easley Hoy of the Census Bureau.

Policy Research Division,
Office of Economic Opportunity

BIBLIOGRAPHY

1. Banks, M. J. and Shapiro, G. M. (1971). Variance of the Current Population Survey, including within and between-PSU components and the effect of the different stages of estimation. *Proc. Soc. Stat. Sec.*, 40–49.
2. Berkson, J. (1955). Maximum likelihood and minimum X^2 estimates of the logistic function. *J. Amer. Stat. Assn.* 50, 130–162.
3. Brillinger, P. R. (1966). The application of the Jackknife to the analysis of sample surveys. *Commentary*, 8: 74–80.
4. Darroch, J. N. and Ratcliff, D. (1972). Generalized iterative scaling for log-linear models. *Annals Math. Stat.*, 43, 1470–1480.
5. Dempster, A. P. (1971). An overview of multivariate data analysis. *J. Mult. Analysis*, 1: 316–346.
6. Feinberg, S. E. and Holland, P. W. (1970). Methods for eliminating zero counts in contingency tables. *Random Counts in models and structures I*. Penn. State Univ. Press, University Park, 233–260.
7. Fisher, Marian (1972). *CONTAB MOD*. George Washington University, Washington, Unpublished.
8. Goodman, L. A. (1969). How to ransack social mobility tables and other kinds of cross-classification tables. *American Journal of Sociology*, 75: 1–40.
9. Goodman, L. A. (1970). The multivariate analysis of qualitative data: interactions among multiple classifications. *J. Amer. Stat. Assn.*, 65: 226–256.
10. Goodman, L. A. (1972). A modified multiple regression approach to the analysis of dichotomous variables. *Amer. Soc. Rev.*, 37: 28–46.
11. Hanson, R. H. (1970). Variance estimates for unbiased estimates: 449 CPS PSU design (Unpublished).
12. Ireland, C. T. (1971). *CONTAB II: A computer program for analyzing contingency tables*. George Washington University, Washington.
13. Ireland, C. T. (1972). Sequential cell deletion in contingency tables (Unpublished).
14. Johnson, W. D. and Koch, G. G. (1970). Analysis of qualitative data: linear functions. *Approaches and techniques*, 358–369.
15. Kish, L. (1965). *Survey sampling*. Wiley, New York.
16. Kish, L. and Frankel, M. R. (1970). Balanced repeated replications for standard errors. *J. Amer. Stat. Assn.*, 65: 1071–1094.
17. Kullback, S. (1959). *Information theory and statistics*. Wiley, New York, 1968 ed. Dover, New York.
18. Kullback, S. (1970). Minimum discrimination information estimation and application. Invited paper presented to Sixteenth Conference on the Design of Experiments in Army Research, Development and Testing, U.S.A. Logistics Management Center, Fort Lee, Va., 21 October, 1970. *ARO-D*.
19. Mosteller, F. and Tukey, J. W. (1968). Data analysis, including statistics. *The Handbook of Social Psychology* (2nd edition). Addison-Wesley, New York.
20. Nathan, G. (1969). Tests of independence in contingency tables from stratified samples. *New Developments in Survey Sampling*. Wiley, New York.
21. Scheuren, F. J. (1972). Unbiased replicate variance estimates for differentiable functions of CPS totals (Unpublished).
22. Simmons, W. R. (1968). Pseudo-replication in the N.C.H.S. Health Examination Survey. *Proc. Amer. Stat. Assn. Soc. Stat. Sec.*, 19–30.
23. Tukey, J. W. and Wilk, M. B. (1965). Data analysis and statistics: techniques and approaches. *Proc. symposium on information processing in sight sensory systems*. California Institute of Technology, Pasadena.
24. U.S. Bureau of the Census (1967). Concepts and methods used in manpower statistics from the current population survey. *Current population reports*, Series P-23, No. 22.
25. U.S. Bureau of the Census. *Current population reports*. Series P-60, Nos. 68, 76, and 81.
26. Zellner, A. and Lee, T. H. (1965). Joint estimation of relationships involving discrete random variables. *Econometrica*, 33: 382–394.