ANOVA Based Designs for the Analysis of Survey Data

Robert W. Hodge

NUPRI Research Paper Series No. 37

March 1987

Robert W. Hodge
Professor
Department of Sociology
University of Southern California

Acknowledgements

This essay was initially prepared in conjunction with research workshops, sponsored by the Rockefeller and Ford Foundations, which were held in Bangkok and Istanbul. The support of the Foundations is acknowledged, but they bear no responsibility for any errors contained herein. I am grateful to Sun Hee Lee for carrying out the computations reported herein and to my colleagues and friends--Fred Arnold, Naohiro Ogawa, and John von Briesen Raz--for numerous discussions of this material. Whatever clarity I bring to the matters discussed herein is largely due to their counsel.

C O N T E N T S

Tables	vi
Abstract	viii
I. Introduction	1
II. Multiple Classification Analysis and Dummy Variables	2
III. Dummy Variables and Multiple Classification: An Empirical Example	8
IV. Alternative Approaches to the Derivation of MCA Coefficients	17
V. Effect Coding of Categorical Variables	19
VI. Summary and Conclusions	24
Notes	26
References	27

T A B L E S

1. Multiple Classification Analysis of Number of Living Children, for Married Women of Childbearing Age, 1967 12

2. Constrained Estimates of the Impact of Husband's and Wife's Education on Number of Living Children, for Married Women of Childbearing Age, Korea, 1967 16

A B S T R A C T

There is a long and relatively extensive literature upon the use of ANOVA type designs for the analysis of classifications with unequal numbers of cases in the various cells. This type of data is typically encountered in surveys and censuses where there is no way that one can meaningfully constrain the predictor variables to be orthogonal to one another. Analyses of this type of data have variously employed the analysis of variance with unequal cell frequencies, dummy variable regressions, multiple classification analysis, and effect coded predictors. While all of these methods are closely related to one another, there are some subtle differences between them. Furthermore, some of the literature relating these methods to one another is almost as hard to find as a fugitive. In this essay, we work out the relationship between dummy variable and multiple classification analysis, showing how one can transform the results obtained from either of these methods to the other. In addition we exhibit computational formula, provide a numerical example derived from a Korean fertility survey, and discuss the relative merits of dummy variable as opposed to effect coding of polytomies.

I. Introduction

In experiments, treatments are typically orthogonalized by design. Randomization of the selection of subjects to treatments underwrites the assumption that the treatments are orthogonal to any other variable whatsoever. These two features of experiments--the design by which the treatments are administered and randomization of the selection of subjects to treatments--jointly imply that the typical ANOVA or ANCOVA design for the estimation of treatment effects is correctly specified. Such is not the case with the analysis of survey data, where the predictor variables ("treatments") are typically correlated and the absence of random selection of subjects to levels of the predictor variables implies confounding with uncontrolled factors. Nonetheless, ANOVA type designs (Yates, 1934) are frequently employed in the analysis of survey data, the most familiar of these being (1) dummy variable regression analysis (Suits, 1957; Draper and Smith, 1966) and (2) multiple classification analysis (Andrews, Morgan, and Sonquist, 1967; Andrews, Morgan, Sonquist, and Klem, 1973; Blau and Duncan, 1967; Duncan, 1964; Hill, 1959).

In a recent article, Suits (1984) has indicated the desirability for <u>interpretative</u> purposes of expressing the estimated effects in ANOVA type designs as deviations from some kind of typical case, as opposed to deviations from some arbitrarily contrived configuration of the levels of the predictor variables. Unfortunately, the particular examples which Suits uses to illustrate his point are rather special ones which, while based on survey type data, have the very atypical property of having equal numbers of observations in each level of the predictor variables. Furthermore, Suits chooses to make no reference to multiple classification analysis which provides a more general solution to the problem he addressed. In particular, the equations developed by Suits to transform dummy variable coefficients into multiple classification coefficients (MCA) are appropriate <u>only</u> for the special cases in which the number of sample cases in each level of a predictor variable are equal. Different predictor variables may, of course, have different numbers of cases in their respective levels, so long as those numbers are equivalent.

In this essay, we do five things: (1) we develop explicitly the relationship between multiple classification analysis and dummy

variable regression analysis, (2) we discuss alternative strategies for deriving MCA coefficients, (3) we indicate the relationship between effect-coded and dummy-coded variables, (4) we provide numerical and, hopefully, substantively interesting illustrations of these relationships, and (5) we briefly overview some nuances in the analysis of survey and experimental data which highlight some differences between three essentially similar methods: ANOVA, MCA, and dummy variable regression analysis. Nothing developed in this essay is, to our knowledge, intellectually novel. We survey that which is already known, but we attempt to synthesize such knowledge in what we believe to be a somewhat novel way.

II. Multiple Classification Analysis and Dummy Variables

If you are confronted with an array of numerical observations on a set of objects, like the hourly earnings of a set of employed persons, and are asked to guess the value of the observation for a randomly chosen subject on the basis of a single summary statistic of the entire distribution, you can control your errors in prediction in several alternative ways. If you want to maximize your probability of being exactly correct, you choose the mode of the distribution as your estimate for the randomly selected case from it. If you want to equate your chances of guessing too high with guessing too low, you choose the median and, if you want to minimize your squared error of prediction in the numerical value, you select the mean. Multiple classification analysis, like ordinary least squares regression analysis, rests on this third criterion. Multiple classification analysis, *therefore, postulates a best guess for the observed value of a particular case as the mean of all observations and adjusts that initial guess upwards or downwards according to the level or category of the predictor (or treatment) variables into which a case falls. In this sense, multiple classification analysis is equivalent to doing ordinary least squares regression analysis with all of the predictor variables centered about their respective means.*

The substantive model for multiple classification analysis is given by the following, in the case of two predictor classifications:

$$Y_i = \overline{Y} + \sum_{j=1}^{r} \alpha_j X_{ji} + \sum_{k=1}^{s} \beta_k Z_{ki} + e_i, \qquad \text{(Eq. 1)}$$

where the X_{ji}'s and Z_{ki}'s are dummy coded 1 or 0 according to whether the ith object is or is not in the jth category of the classification given by the X_j's and the kth category of the classification given by the Z_k's; Y_i represents the numerical value of some dependent variable for the ith respondent; $\overline{Y}$ is the mean of the Y_i's, and the α_j's and β_k's and the coefficients we wish to estimate. Furthermore, we impose the following condition upon the X_j's and Z_k's:

$$\sum_{j=1}^{r} X_{ji} = \sum_{k=1}^{s} Z_{ki} = 1, \text{ for all } i, \qquad \text{(Eq. 2)}$$

which secures that the classifications given by the X_j's and Z_k's are mutually exclusive and exhaustive. Finally, we impose the following constraint upon the α_j's and β_k's:

$$\sum_{j=1}^{r} \alpha_j \overline{X}_j = \sum_{k=1}^{s} \beta_k \overline{Z}_k = 0, \qquad \text{(Eq. 3)}$$

which can viewed in several ways. First, the model given by Eq. 1 and Eq. 2 alone is not identified, owing to the strict or exact multicollinearity between the X_j's and Z_k's. Thus, Eq. 3 can be regarded as an identifying restriction, but it is only one of any number of identifying restrictions which might be postulated. For example, in what follows below, it will be clear that the identifying restrictions imposed in the case of dummy variable regression analysis are twofold: (1) free the intercept from the constraint of equality to the grand mean and (2) set one of the α_j's and one of the β_k's equal to zero. Alternatively, one can propose, as Suits (1984) does, to set the unweighted sum of the α_j's and β_k's equal to zero.[1/] Second, however, the identifying constraint imposed by Eq. 3 can be seen in a different light: It is the only one of an infinitely large possible (though not all plausible) identifying constraints which guarantees that, in the general case, the sum of the squared errors of estimate will be minimized and the coefficients will also be centered about the grand mean. The proof of this assertion goes beyond this review essay, but the reader with scant statistical and mathematical expertise will surely gain insight into why this is so by considering the case in which there is only one, rather than multiple predictor classifications.

Although it is not necessary to our exposition, it is perhaps useful at this juncture to recognize that the X_j's (and the Z_k's) are counter variables, which increment a running sum over all respondents by the value one if the ith respondent happens to fall in the jth category of the classification given by the X_j's. Thus, the sum of X_{ji} over all respondents is just n_j, the number of respondents in the jth category of the classification given by the X_j's and the mean of X_{ji} is just P_j, the proportion of respondents in the jth category of the classification given by the X_j's. We must also note, before continuing, that the model given above is only a substantive model applicable to a particular data set. To turn it into a full mathematical model, from which we could make inferences from samples to populations, we would have to augment it by (1) drawing a distinction between population parameters and the sample parameters (estimated by $\overline{Y}$ and the α_j's and β_k's), as well as (2) imposing a variety of constraints on the e_i's which correspond to those in ANOVA designs, such as normality, independence, and homoscedasticity. Here we ignore these considerations, not because they are unimportant (they are), but because our focus is upon the estimation of parameters rather than their statistical properties.

Owing to the strict (exact) multicollinearity between the X_j's and Z_k's (see Eq. 2), there is no way that Eq. 1 can be estimated directly by ordinary least squares. We can, however, invoke Eq. 2 to eliminate the linear dependencies inherent between the predictors in Eq. 1. For example, we may rewrite Eq. 2 as:

$$X_{ri} = 1 - \sum_{j=1}^{r-1} X_{ji}, \text{ for all } i, \qquad \text{(Eq. 4a)}$$

and

$$Z_{si} = 1 - \sum_{k=1}^{s-1} Z_{ki}, \text{ for all } i, \qquad \text{(Eq. 4b)}$$

which simply express the dummy variables which record membership in the last categories of the classifications given by the X_j's and Z_k's as a function of membership or non-membership in the remaining categories.[2/]

Since Eqs. 4a and 4b are exact identities, like Eq. 2, they can be substituted into Eq. 1 to yield:

$$Y_i = \bar{Y} + \sum_{j=1}^{r-1} \alpha_j X_{ji} + \alpha_r (1 - \sum_{j=1}^{r-1} X_{ji}) + \sum_{k=1}^{s-1} \beta_k Z_{ki} + \beta_s (1 - \sum_{k=1}^{s-1} Z_{ki}) + e_i. \qquad \text{(Eq. 5)}$$

Algebraically rearranging this result, we obtain:

$$Y_i = \bar{Y} + \alpha_r + \beta_s + \sum_{j=1}^{r-1} (\alpha_j - \alpha_r) X_{ji} + \sum_{k=1}^{s-1} (\beta_k - \beta_s) Z_{ki} + e_i, \qquad \text{(Eq. 6)}$$

which can be written as:

$$Y_i = k + \sum_{j=1}^{r-1} a_j X_{ji} + \sum_{k=1}^{s-1} b_k Z_{ki} + e_i, \qquad \text{(Eq. 7)}$$

where $k = \bar{Y} + \alpha_r + \beta_s$, (Eq. 8a)

$a_j = \alpha_j - \alpha_r$, for all $j < r$, (Eq. 8b)

and $b_k = \beta_k - \beta_s$, for all $k < s$. (Eq. 8c)

Formally, Eq. 7 is equivalent to Eq. 1, since it has been deduced from the latter only via the substitution of algebraic identities. However, Eq. 7, unlike Eq. 1, contains no strict multicollinearities between the variables on the right hand side. In fact, the form of Eq. 7 is identical to that of an ordinary dummy variable regression and its coefficients--k, the a_j's, and b_k's may be estimated directly by ordinary least squares (see, e.g., Suits, 1957; Draper and Smith, 1966).

Since Eq. 7 can be estimated, the problem now becomes one of retrieving the α_j's and β_k's of the multiple classification analysis from the numerical estimates of the corresponding a_j's and b_k's in the dummy variable regression analysis. This is where the identifying restriction in Eq. 3 comes into play. Since it follows from Eq. 2, that $\sum_{j=1}^{r} \bar{X}_j = 1$, we may rewrite one of the two identities given by Eq. 3 as follows:

$$\sum_{j=1}^{r-1} \alpha_j \bar{X}_j + \alpha_r (1 - \sum_{j=1}^{r-1} \bar{X}_j) = 0. \qquad \text{(Eq. 9)}$$

Rearranging this equation, we find that

$$\sum_{j=1}^{r-1}(\alpha_j - \alpha_r)\bar{X}_j + \alpha_r = 0, \qquad \text{(Eq. 10)}$$

or, on transposing and substituting Eq. 8b,

$$\alpha_r = -\sum_{j=1}^{r-1}(\alpha_j - \alpha_r)\bar{X}_j = -\sum_{j=1}^{r-1} a_j \bar{X}_j. \qquad \text{(Eq. 11)}$$

We have now an explicit formula for α_r in terms of the estimated values of the a_j's, which enables us to derive the remaining multiple classification coefficients by simply adding α_r to both sides of Eq. 8b to obtain:

$$\alpha_j = a_j + \alpha_r, \text{ for all } j < r. \qquad \text{(Eq. 12)}$$

A similar derivation shows that:

$$\beta_s = -\sum_{k=1}^{s-1} b_k \bar{Z}_k, \qquad \text{(Eq. 13)}$$

and

$$\beta_K = b_k + \beta_s, \text{ for all } k < s. \qquad \text{(Eq. 14)}$$

We have thus shown how the coefficients in a multiple classification analysis may be derived from the coefficients of a dummy variable regression analysis.

The advantage of multiple classification analysis is clear; it represents the effects of the predictor categories as deviations from the grand mean, just as the effects of treatments are represented in ANOVA analyses of experimental designs. There is only one way of doing this, so the solution is unique. By way of contrast, there are as many different ways of doing a dummy variable regression analysis as the product of the number of categories in all of the predictor classifications. For example, in the present case of two classifications, with r and s categories, respectively, there are rs possible ways of doing a dummy variable regression analysis, any one of which may be used to derive the same multiple classification analysis.

The present derivation of the algebraic identities relating multiple classification analysis to dummy variable regression analysis reveals, surprising as it seems, that the coefficients obtained in dummy variable regressions are misinterpreted by a number of authors. Suits, for example, studies a gasoline consumption function in which weekly gasoline expenditures are related to annual household income and its square, as well as to dummy variables reflecting the age of the household head and region of residence. In this consumption function, the omitted regional variable is southern location and the omitted age category is for household heads aged over 64. Suits (1984, pp. 179) states, "Since each of the coefficients of the dummy variables [in the equation] indicates deviation from gasoline consumption of people over 64 residing in the South, they are especially awkward to interpret . . . " This is quite simply misleading. As can be seen by recourse to Eqs. 8b and 8c above, the dummy variable coefficients reflect only deviations from the category implicitly omitted from the classification to which they refer, rather than a deviation from the joint category defined by the combination of categories implicitly omitted from all of the classifications. The deviations so measured have, however, been adjusted for the other classifications. Interpreting dummy variable coefficients in the manner of Suits is like interpreting the coefficient of a variable in an ordinary least squares regression equation as the difference between anyone who has exactly 1 unit of the variable in question and someone who has nothing of anything. That interpretation of an ordinary regression coefficient would lead one to believe that everything is assessed relative to the intercept. But that is not the case; the regression coefficient reflects the difference between any pair of individuals who differ by exactly one unit on the variable in question and share a common set of values on all the remaining variables. Similarly, dummy variable coefficients reflect differences between those in the included categories and those in the excluded category of the same classification, given that they share similar locations in the remaining categories. Other misleading interpretations of dummy variable coefficients as reflecting deviations from the combination of categories implicitly omitted from all of the classifications can be found in substantive articles using the method and even occasionally in textbooks on research methodology. For

example, Lansing and Morgan (1971) say without sufficient explanation that dummy variable coefficients are given "in terms of differences from the excluded group"(ibid., p.317). Appropriate interpretations, parallel to that offered here, can be found in many standard econometric texts, including those of Goldberger (1964, pp. 224ff), Johnston (1972, pp. 176ff), and Kmenta (1971, pp. 415-418).

The present exposition makes clear that the proper interpretation of a dummy variable regression coefficient is the net or adjusted difference between the category reflected in the dummy variable and the implicitly omitted category from the same classification. Correspondingly, differences between pairs of MCA coefficients or, what is the same thing, differences between the dummy variable coefficients associated with the categories to which they refer have parallel interpretations. From Eq. 8b, we have $a_j = \alpha_j - \alpha_r$ and $a_{j+1} = \alpha_{j+1} - \alpha_r$, which implies that

$$a_j - a_{j+1} = (\alpha_j - \alpha_r) - (\alpha_{j+1} - \alpha_r) = \alpha_j - \alpha_{j+1}. \qquad \text{(Eq. 15)}$$

Thus, we find that differences between pairs of MCA coefficients are equivalent to differences between pairs of corresponding dummy variable coefficients. Tests of significance of the difference between any pair of MCA coefficients can, therefore, be constructed from a knowlege of the variance-covariance matrix of the coefficients in the dummy variable regression analysis from which the MCA coefficients were derived. For example, the variance of the difference $(\alpha_j - \alpha_{j+1})$ is given by $\mathrm{Var}[a_j] - 2\mathrm{Cov}[a_j, a_{j+1}] + \mathrm{Var}[a_{j+1}]$. The ratio of the former commodity to the square root of the latter forms a t-test for the significance of the difference between any pair of groups, adjusted for whatever other variables have been included in the analysis.

III. Dummy Variables and Multiple Classification: *An Empirical Example*

Although the computational formulas given by Eq. 11 through Eq. 14 for transforming dummy variable coefficients into the relevant MCA coefficients are straightforward, numerical illustrations of the

calculations are often useful. Here we present an illustration based on a 1967 KAP survey conducted in the Republic of Korea. We consider herein the following variables derived from this survey: (1) wife's education clustered into four categories from high to low, (2) husband's education clustered into four categories exactly comparable to those in which wife's education was clustered, (3) urban vs. rural residence, (4) wife's age, measured as a continuous variable, and (5) number of living children. For purposes of exposition, we denote these variables, respectively, by W_{ji}, j = 1(low), 2, 3, and 4(high); H_{ki}, k = 1(low), 2, 3, and 4(high); R_i = 1(rural) or 0(urban); U_i = 1(urban) or 0(rural); A_{Wi} = wife's age in years, and F_i = number of living children. Evidently, with these definitions;

$$\sum_{j=1}^{4} W_{ji} = \sum_{k=1}^{4} H_{ki} = U_i + R_i = 1, \text{ for all } i. \qquad \text{(Eq. 15a)}$$

we also center A_{Wi} independently about its respective means among rural ($=\overline{A}_{Wr}$) and among urban ($=\overline{A}_{Wu}$) women, which yields the new variable:

$$A_{Wi} = (A_{Wi} - \overline{A}_{Wr})R_i + (A_{Wi} - \overline{A}_{Wu})U_i. \qquad \text{(Eq. 15b)}$$

The model of these data which we entertain is considerably more complicated than the substantively vacuous one we introduced for purposes of exposition. Specifically, we estimate the following model:

$$F_i = \overline{F} + a_u U_i + a_r R_i + b_u A_{Wi} U_i + b_r A_{Wi} R_i + \sum_{j=1}^{4} c_{uj} W_{ji} U_i + \sum_{j=1}^{4} c_{rj} W_{ji} R_i + \sum_{k=1}^{4} d_{uk} H_{ki} U_i + \sum_{k=1}^{4} d_{rk} H_{ki} R_i + e_i, \qquad \text{(Eq. 16)}$$

where the following identities hold, for all respondents, i,

$$U_i + R_i = 1, \qquad \text{(Eq. 17a)}$$

$$\sum_{j=1}^{4} W_{ji} U_i = \sum_{k=1}^{4} H_{ki} U_i = U_i, \qquad \text{(Eq. 17b)}$$

and

$$\sum_{j=1}^{4} W_{ji} R_i = \sum_{k=1}^{4} H_{ki} R_i = R_i. \qquad \text{(Eq. 17c)}$$

In addition, the following restrictions are imposed on the coefficients:

$$\sum_{j=1}^{4} c_{uj} \Pr(W_{ji} = 1 \mid U_i = 1) = \sum_{k=1}^{4} c_{uk} \Pr(H_{ki} = 1 \mid U_i = 1)$$

$$= \sum_{j=1}^{4} c_{rj} \Pr(W_{ji} = 1 \mid R_i = 1)$$

$$= \sum_{k=1}^{4} c_{rk} \Pr(H_{ki} = 1 \mid R_i = 1)$$

$$= a_u \bar{U} + a_r \bar{R} = 0, \qquad \text{(Eq. 18a)}$$

where, for example, $\Pr(W_{ji} = 1 \mid U_i = 1)$ is the probability that a women falls in the jth educational category given that she lives in an urban place.

What does the proposed model postulate? It says, simply (despite all the room it takes up to make it explicit), that the number of living children a woman has depends upon (1) whether she lives in an urban or rural place, (2) her own educational level, (3) her husband's education, and (4) her own age, but that (5) the impacts of (a) her own educational level, (b) her husband's educational level, and (c) her own age are themselves contingent upon whether she resides in an urban or rural place. Since, in this model, the effects of the categorical variables reflecting wife's and husband's education are contingent upon where they reside, their effects are constrained to equal zero (by Eq. 18) within both urban and rural areas. Eq. 18 likewise constrains the impact of urban vs. rural residence to cancel out when properly weighted.

Although it is possible to estimate the model given by Eq. 16 in one fell swoop by judicious invocation of the identities given in Eqs. 17a through 17c, we have chosen to estimate the model in halves, by running one model for rural women and another model for urban women. This makes sense only because the postulated model implies that residence interacts with everything. Johnston (1972, pp. 177) considers, for example, a very simple consumption function in which the marginal propensity to consume is identical in wartime and peacetime, but consumption is higher in peacetime than in wartime. Such a function can be best estimated by combining all the observations, both in peacetime and in wartime, to obtain a single, overall estimate of the

marginal propensity to consume and a constant difference between wartime and peacetime. If, however, the marginal propensity to consume were itself different in wartime and peacetime, then nothing would be gained by estimating a single coefficient for the marginal propensity to consume across both wartime and peacetime years. As Johnston observes (ibid.), one can merely fit two separate equations: one to the peacetime data and one to the wartime data. That is just the strategy we have used here, since we have postulated that the impact of rural residence (call it peace) vs. urban residence (call it war) is not simply additive, but instead influences the impacts of all of the remaining variables in our equation for number of living children.

Since we are fitting separate equations for those in the rough and those in the action, the model given by Eq. 16 is grossly simplified within groups. For $U_i = 1$, those at the front, it just becomes:

$$F_i = \bar{F} + a_u + b_u A_{Wi} + \sum_{j=1}^{4} c_{uj} W_{ji} + \sum_{k=1}^{4} d_{uk} H_{ki} + e_i, \quad \text{(Eq. 19)}$$

and the same constraints of the c_{uj}'s and d_{uk}'s hold. A similar equation applies for the case $R_i = 1$, but it is not exhibited here owing to redundancy. Although the proof is not offered here, it is easy to show that the values of b_u, the c_{uj}'s and d_{uk}'s in Eq. 19 are identical to those of the corresponding coefficients in Eq. 16. Furthermore, it can be shown that $\bar{F} + a_u = \bar{F}_u$, the mean number of living children among urban women. Using this identity and deleting the linear dependencies between the categories of husband's and wife's education, we obtain the following estimable version of Eq. 19:

$$F_i = \bar{F}_u + c_{u4} + d_{u4} + b_u A_{Wi} + \sum_{j=1}^{3} (c_{uj} - c_{u4}) W_{ji} + \sum_{k=1}^{3} (d_{uk} - d_{u4}) H_{ki} + e_i. \quad \text{(Eq. 20)}$$

Estimates of the dummy variable regression shown in Eq. 20 and the corresponding model for rural women are given in Table 1, which also reports the statistical significance of the coefficients in the dummy variable regressions as well as the derived multiple classification coefficients. The means of all of the variables are also shown. The means of the dummy variables for the levels of husband's and wife's

Table 1. Multiple Classification Analysis of Number of Living Children, for Married Women of Childbearing Age, Korea, 1967

Variables and Statistics	Urban			Rural		
	Means	Dummy Variable Coefficients	MCA Coefficients	Means	Dummy Variable Coefficients	MCA Coefficients
Age	37.77	.1880	.1880	32.36	.2041[c]	.2041
Husband's education						
H_1 (low)	.0751	.2527	.1866	.3044	.0583	-.0140
H_2	.2669	.2354[c]	.1693	.4633	.1011	.0288
H_3	.2102	-.0763	-.1424	.1157	.0667	-.0058
H_4 (high)	.4478	--	-.0661	.1166	--	-.0723
Wife's education						
W_1 (low)	.2068	.4582[c]	.0885	.5521	.3057	.1066
W_2	.4896	.4770[c]	.1073	.3982	.0874	-.1117
W_3	.1501	.2756	-.0941	.0315	-.1427	-.3418
W_4 (high)	.1535	--	-.3697	.0182	--	-.1991
Number of living children	3.1018	-3.4941[a]	1799[b]	3.4720	-3.4038[a]	2411[b]
	Coefficients of determination					
R^2 (Adjusted R^2)		.4509 (.4476)[c]			.5126 (.5112)[c]	

[a]Intercept

[b]Number of cases

[c]Significantly different from zero at .05 level.

education are, of course, just the proportions of wives and husbands falling in the various categories of educational attainment.

To illustrate the computation of the multiple classification coefficients, we may consider the categories of husband's education among urban residents. We may compute the value of d_{u4}, the MCA coefficient for husbands with high educational levels, by applying the analogue to Eq. 11 or Eq. 12. This gives

$$d_{u4} = - [(.0751)(.2527) + (.2669)(.2354) + (.2102)(-.0763)]$$

$$= -.0661.$$

The remaining MCA coefficients are obtained by simply adding the estimate of d_{u4} to the remaining dummy variable coefficients. For example, $d_{u3} = (-.0763) + (-.0661) = -.1424$. The remaining MCA coefficients are derived analogously.

The results themselves reveal that, after adjustment for wife's age, there remains a very modest educational differential in fertility among both urban and rural Korean women. These gradients appear among urban women with respect to both husband's and wife's education, but neither gradient is monotonic. Among rural woman, there is virtually no gradient with respect to husband's education, but there are differentials with respect to wife's education. The rural educational differential, like its urban counterpart, is not monotonic. In urban areas, the primary difference appears to rest between the lowest two and the highest two educational categories while in rural areas, there are three tiers to the gradient with respect to wife's education, with the lowest category having about .2 of a child more than the next to lowest category, which in turn has about .1 of a child more than the two highest categories combined.

Some difficulties in the interpretation of dummy variable coefficients are also revealed by the illustration in Table 1. For example, we observe in the results for rural women that none of the coefficients for the categories of husband's and wife's education is twice as large as its standard error. One cannot, however, infer from this that there are no educational differentials in fertility among rural women. The reason is clear; the dummy variable coefficients contrast each of the included categories with the highest educational

group, which among rural women is a very small category, amounting to less than two percent of the total sample. There are somewhat more husbands of rural women in the highest educational category, but it is still relatively small. Thus, even though none of the dummy variable coefficients is significant, there can still be educational differentials overall which could be revealed by contrasting the included categories with one another.

An overall test of the presence of educational differentials among rural women can be made by contrasting the sum of squares explained by the model with the sum of squares explained by wife's age alone. The zero order correlation between age of wife and number of living children is .71303 among rural women. We may calculate:

$$\frac{.5126 - (.71303)^2}{1 - .5126} \cdot \frac{2403}{6} = 3.4415,$$

which is distributed as F with 6 degrees of freedom in the numerator and 2,403 in the denominator. Recourse to tabled values of the F-distribution reveals that the probability of obtaining a value this large or larger is less than .01. Thus, even though none of the dummy variable coefficients is significant by conventional standards, one cannot reject the hypothesis that there are no educational differentials in fertility among rural Korean women.

We may also observe in Table 1 that the educational differentials in fertility with respect to husband's and wife's education are roughly on the same order of magnitude, though the differential with respect to wife's education appears somewhat more marked. We can test whether these differentials are statistically different by contrasting the sums of squares explained in a model in which these differentials are constrained to be identical with the sums of squares explained by the model in Table 1, which allows the impact of husband's education to differ from that of wife's education. To estimate the constrained model, we construct the following new variables:

$$E_{ji} = W_{ji} + H_{ji}, \; j = 1, 2, 3, 4. \qquad \text{(Eq. 21)}$$

It should be noted that the E_j's, although constructed from dummy variables, are not themselves dummy variables. They take on the value

2 when both husband and wife are in the jth educational category; the value 1 when either the husband or the wife but not both are in the jth educational category, and the value 0 when neither the husband nor the wife is in the jth educational category. The E_j's are related by the following identity:

$$\sum_{j=1}^{4} E_{ji} = 2, \text{ for all } i. \qquad \text{(Eq. 22)}$$

Thus, only three of the E_j's can be entered in an ordinary least squares regression. Among urban Korean women, we find that a model including the E_j's and wife's age explains 44.93 percent of the variance in number of living children. We may contrast this model with that given in Table 1 by calculating:

$$\frac{.4509 - .4493}{1 - .4509} \cdot \frac{1191}{3} = 1.3842,$$

which is distributed as F with three degrees of freedom in the numerator and 1191 in the denominator. The probability of obtaining a value this larger or larger is greater than .05, so we cannot reject the hypothesis that fertility differentials with respect to husband's and wife's education are identical among urban Korean women.

The same exercise can be performed among rural women, where we find that a model including the E_j's and wife's age explains 51.14 percent of the variance in number of living children. We can calculate:

$$\frac{.5126 - .5114}{1 - .5126} \cdot \frac{2403}{3} = 1.9721,$$

which does not allow us to reject the hypothesis that the impacts of husband's and wife's education are identical among rural women. Since we cannot reject the hypothesis that the effects of husband's and wife's education differ among either rural or urban women, we may proceed to examine the constrained coefficients. They are reported in Table 2, along with other relevant features of the constrained regressions.

As one can see by comparison of the constrained coefficients in

Table 2. Constrained Estimates of the Impact of Husband's and Wife's Education on Number of Living Children, for Married Women of Childbearing Age, Korea, 1967

Variables and Statistics	Urban			Rural		
	Means	OLS Estimates	MCA Coefficients	Means	OLS Estimates	MCA Coefficients
Age	37.77	.1870[c]	.1870	32.36	.2046[c]	.2046
Education of either spouse						
E_1 (low)	.2819	.3058[c]	.1323	.8565	.2287[c]	.0658
E_2	.7565	.3139[c]	.1404	.8615	.1404	-.0225
E_3	.3603	.0645	-.1090	.1472	.0609	-.1020
E_4 (high)	.6013	--	-.1735	.1348	--	-.1629
Number of living children	3.1018	-3.3734[a]	1199[b]	3.4720	-3.4743[a]	2411[b]
		Coefficients of determination				
R^2 (Adjusted R^2)		.4493 (.4474)[c]			.5114 (.5106)[c]	

[a]Intercept

[b]Number of cases

[c]Significantly different from zero at .05 level.

Table 2 with the unconstrained coefficients in Table 1, the result of constraining the effects of husband's and wife's education to be equal tends to smooth out the educational differentials in fertility. The gradients are now monotonic, save for a very modest reversal of the lowest two educational groups among urban women. In both urban and rural areas, there are educational gradients in fertility. Within both areas, the impacts of husband's and wife's education are approximately identical.

IV. Alternative Approaches to the Derivation of MCA Coefficients

In the preceding two sections we have exhibited the relationship between multiple classification analysis and dummy variable regression analysis and provided an empirical illustration of how one may derive the MCA coefficients from the dummy variable coefficients. We now show how the MCA coefficients may be derived directly by the consideration of ordinary least squares regressions which employ different variables. To derive this alternative approach, we return to the model for multiple classification analysis set forth in Eqs. 1-3. We observe that Eq. 3 may be solved to find a formula for any one of the MCA coefficients. In particular, we have

$$\alpha_r = -\sum_{j=1}^{r-1} \alpha_j(\bar{X}_j/\bar{X}_r), \qquad \text{(Eq. 23a)}$$

and

$$\beta_s = -\sum_{k=1}^{s-1} \beta_k(\bar{Z}_k/\bar{Z}_s). \qquad \text{(Eq. 23b)}$$

These identities may be used to delete α_r and β_s from Eq. 1. This leaves us with:

$$Y_i = \bar{Y} + \sum_{j=1}^{r-1} \alpha_j X_{ji} - \sum_{j=1}^{r-1} \alpha_j(\bar{X}_j/\bar{X}_r)X_{ri} + \sum_{k=1}^{s-1} \beta_k Z_{ki} - \sum_{k=1}^{s-1} \beta_k(\bar{Z}_k/\bar{Z}_s)Z_{si} + e_i. \qquad \text{(Eq. 24)}$$

Combining the terms on the right one obtains:

$$Y_i = \bar{Y} + \sum_{j=1}^{r-1} \alpha_j [X_{ji} - (\bar{X}_j/\bar{X}_r)X_{ri}] + \sum_{k=1}^{s-1} \beta_k [Z_{ki} - (\bar{Z}_k/\bar{Z}_s)Z_{si}] + e_i. \qquad \text{(Eq. 25)}$$

This equation may be estimated by ordinary least squares, but the predictor variables are no longer simple dummy variables. Instead, each predictor variable takes the form $X_{ji} - (\bar{X}_j/\bar{X}_r)X_{ri}$, which takes on the value 1 if the $\underline{i}$th respondent is in the $\underline{j}$th category of the classification given by the X_j's, the value 0 if the $\underline{i}$th respondent is neither in the $\underline{j}$th nor the $\underline{r}$th category of that classification and the generally non-integer value $-\bar{X}_j/\bar{X}_r$, which is just the negative of the ratio of the number of cases in the $\underline{j}$th and $\underline{r}$th categories, if the $\underline{i}$th respondent falls in the $\underline{r}$th category of the classification. These somewhat peculiar variables are defined for any $\underline{r} - 1$ of the categories; here we have omitted the $\underline{r}$th category for algebraic convenience. The use of variables defined in this way has the advantage that one obtains the MCA coefficients directly as part of the output from any standard OLS regression package. Also, the associated t-tests for the significance of the estimated coefficients are direct tests of whether the means for the included categories, adjusted for whatever control classifications and variables are introduced, differ significantly from the grand mean. However, this method, exposited in greater detail by Melichar (1966), does not yield the MCA coefficient for the implicitly omitted category, which can either be obtained by (1) solving Eq. 3 for the missing coefficient or by (2) re-estimating the regression equation after deleting an alternative coefficient from it and creating a new set of variables.

One might reasonably inquire if there is not an approach to multiple classification analysis which yields all the coefficients in one fell swoop. The answer to that question is positive and, in fact, probably the most favored computing algorithm for deriving MCA coefficients employs this alternative, which essentially combines the information in Eq. 1 and Eq. 2 to eliminate one of the "normal" type of equations which could be written on the basis of Eq. 1 and then replaces this missing element with Eq. 3 to obtain a full matrix of rank $\underline{rs}$ which can be inverted to find all of the MCA coefficients. This method is embodied, for example, in the MCA program distributed through the Institute for Social Research at the University of

Michigan. We have treated this popular approach to multiple classification analysis somewhat sketchily here because (1) a knowledge of it does nothing to illuminate the differences between MCA, dummy variable analysis, and ordinary least squares regression; (2) its development and illustration would require appreciable additional space, and (3) it has been clearly exposited in a number of readily available sources (Andrews, Morgan, and Sonquist, 1967; Andrews, Morgan, Sonquist, and Klem, 1973; Blau and Duncan, 1967).

V. Effect Coding of Categorical Variables

An alternative to the dummy coding of categorical variables is to assign the values 1 and -1 rather than those of 1 and 0. This is known as effect coding (Kerlinger and Pedhazur, 1973; see Keyfitz, 1953, for an example). Effect coding is the equivalent to dummy coding in the sense that the sums of squares explained by an effect coded or a set of effect coded variables is identical to the sums of squares explained by the corresponding dummy coded or set of dummy coded variables. The coefficients of the effect coded variables are not, however, equivalent to those of the dummy coded variables and their interpretation is somewhat different.

Let us consider G_{ji} as an effect coded variable which takes on the value 1 if the $\underline{i}$th respondent (or case) belongs to the $\underline{j}$th category of a polytomy containing exactly $\underline{r}$ mutually exclusive and exhaustive categories and the value -1 if the $\underline{i}$th respondent does not belong to the $\underline{j}$ category. Let X_{ji} be, as above, the corresponding set of dummy coded variables. Each of the effect coded variables is related to the corresponding dummy coded variable by

$$G_{ji} = 2X_{ji} - 1, \text{ for all } \underline{i},\ j = 1, 2, \ldots, r. \qquad \text{(Eq. 26)}$$

Since the X_j's are constrained by Eq. 2 to sum to unity, we may easily derive that the G_j's are constrained by

$$\sum_{j=1}^{r} G_{ji} = \sum_{j=1}^{r}(2X_{ji} - 1) = 2 - r, \text{ for all } \underline{i}. \qquad \text{(Eq. 27)}$$

If we solve Eq. 26 for the X_j's, we find that:

$$X_{ji} = (1/2)(G_{ji} + 1), \text{ for all } i, j = 1, 2, \ldots, r. \qquad \text{(Eq. 28)}$$

Since this is an exact identity, we may substitute it into the dummy variable regression given by Eq. 7 above. Leaving the Z_k's in dummy variable format, this gives us

$$Y_i = h + \sum_{j=1}^{r-1} g_j G_{ji} + \sum_{k=1}^{s-1} b_k Z_{ki} + e_i, \qquad \text{(Eq. 29)}$$

where

$$h = k + (1/2)(r - 1), \qquad \text{(Eq. 30a)}$$

and

$$g_j = (1/2)a_j. \qquad \text{(Eq. 30b)}$$

The relationships given by Eqs. 30a and 30b establish two things. First, the intercept obtained in a regression involving effect coded variables for the categories of a polytomy is meaningless: it is equal to the expected value (k) for the combination of categories omitted from the corresponding dummy variable regression plus an increment which is solely a function of the number of categories in the polytomy. Second, the coefficients of the effect coded variables are exactly one-half the value of the coefficients of the corresponding dummy coded variables.

Regressions based on effect coded variables have an inviting interpretation. Since each effect coded variable takes on the value 1 for the group it is designed to represent and the value -1 for everyone else, one is tempted to interpret the coefficient of the effect coded variable as the gain or loss relative to the residue of the population associated with membership in the category coded 1 on the effect coded variable. That seems reasonable because, taken in isolation, the coefficient of G_j in the effect coded regression given by Eq. 29 implies that the members of the jth category receive an average increment or decrement in Y equal to g_j, while the rest of the population loses or gains, respectively, an equivalent amount. Such an

interpretation is inviting because that is precisely the way regression coefficients are ordinarily interpreted. However, in ordinary regressions, like, for example, the regression of total number of different sexual partners on age, height, weight, and social status, a knowledge of the value of a respondent on any one of the predictors provides no absolutely certain clue to his value on any of the remaining predictors. That, however, is not the case in regressions involving either dummy-coded or effect-coded variables representing one's membership in the categories of a polytomy. In the case of dummy coded variables, we know that if a respondent has the value 1 on any one of them, he/she must necessarily have the value 0 on all the rest. In the case of effect coded variables, we know that if a respondent has the value 1 on any one of them, he/she must necessarily have the value -1 on all the rest. Thus, it is not the coefficients of the effect coded variables--the g_j's--which contrast the members of a category with the non-members of that category, but rather functions of the g_j's, given by

$$F_j(g_w) = g_j - \sum_{\substack{w=1 \\ w \neq j}}^{r-1} g_w, \qquad \text{(Eq. 31)}$$

which make the relevant assessment of the gains or losses of those in and those not in the jth category of a classification. There is no way the F_j's can be conveniently read from the coefficients observed in a regression on effect-coded variables.[3/] Consequently, we can see no reason whatsoever why one should introduce effect-coded variables in the analysis of survey-type data. Such an endeavor only guarantees that one will obtain a set of coefficients whose values are inherently without substantive meaning.

Another way of seeing why the apparent interpretation of the coefficients of effect-coded variables is erroneous can be gleaned by exhibiting their relationship to the coefficients obtained from a multiple classification analysis. We already know from Eq. 30b how the coefficients of effect coded variables are related to the coefficients of the corresponding dummy coded variables. Eq. 8b gives the relationship of the coefficients of the dummy coded variables to the corresponding MCA coefficients. Substituting Eq. 8b into Eq. 30b gives the following result:

$$g_j = (1/2)(\alpha_j - \alpha_r). \qquad \text{(Eq. 32)}$$

From this relationship it is self evident that, far from stating the gain or loss attributable to membership in a group vis a vis non-membership in that group, the coefficients of effect coded variables are just equal to one-half the mean difference (adjusted for controlled factors) between the group in hand and the group arbitrarily deleted to secure identification of either the effect coded or dummy coded regression. Evidently, the coefficients of effect coded variables to represent polytomies are as arbitrary as the coefficients of the corresponding dummy coded variables: they depend upon which category is omitted.

The disadvantage of effect coding the categories of a polytomy appears relatively obvious, given the above considerations. In the limiting case of a dichotomy, this disadvantage is less apparent. Since there are only two categories, one receiving the value 1 and the other the value -1, the observed coefficient of the single effect coded variable has the obvious interpretation that what is gained by membership in one of the two categories is lost by membership in the other. However, when the groups contrasted are unequal in size, a situation typical in survey analysis, such an interpretation can be substantively very misleading.

To illustrate this point, let us consider the specification of the function for individual annual earnings which is often encountered in socioeconomic research. For employed men, this function typically has a form which is linearly additive and includes measures of educational attainment, occupational SES, and region of residence, plus a dummy variable for race contrasting nonblacks (= 1) vs. blacks (= 0) or whites (= 1) vs. nonwhites (= 0). In analyses of this kind, the coefficient associated with the dummy variable for race comes to about $1,000.[4/] This coefficient can be interpreted as the "cost" of being black or the "force" of discrimination, although the imputation of such substance to the coefficient's value is not logically justified by its statistical meaning, to wit, the average difference between the earnings of employed whites and nonwhites which would still exist if their regional distributions and levels of education and occupation were equivalent, assuming, of course, the way in which these variables have been related to earnings is correctly specified.

Alternatively, one might have represented the racial variable with an effect coded rather than a dummy coded variable. Had this been done, we know from the above results that the coefficient of the effect coded variable would have been about $500, rather than the approximately $1,000 associated with the dummy coded variable. The interpretation of this coefficient would roughly be that, insofar as earnings are concerned, being a white nets you about $500 a year (adjusted for the control variables), while being nonwhite decrements your earnings by an equivalent amount. The error in this assessment is rooted in the facts that the effect coded variable does not assess differences relative to the grand mean (see Eq. 30a) and that the two racial groups are not balanced. Roughly speaking, about 12 percent of the employed are nonwhite. Using this division of the population, the approximate dummy variable coefficient of $1,000, and the relationship given by Eq. 11, we can calculate that the MCA coefficient for nonwhites is roughly given by -($1,000)(.88) = -$880. The MCA coefficient for whites becomes $1,000 + (-$880) = $120, upon applying Eq. 12. Now this is a very different picture than the one implied by interpreting the coefficient of the effect coded variable. It implies that the average white gains (adjusted for the control variables) about $120, while the average black loses ($880) owing to what might be substantively interpreted as "discrimination." The zero sum quality of effect coding of dichotomies creates the illusion that whites gain, on the average, as much as nonwhites lose of the average, something which could only be the case if the two groups were approximately balanced with respect to their relative sizes. To the contrary, if you took the results of this illustration literally (and it is not far from the actual situation), if every white gave up $10 a month--about a candy bar a day--white and nonwhite incomes could be equated <u>within</u> regions and educational and occupational strata. There is no way one could see a prospect like this from a dummy variable analysis, since it effectively sums the absolute values of the two deviations from the mean. An effect coded analysis only confuses the matter hopelessly by splitting the difference in adjusted means between groups of unequal size. The defect in the method proposed by Suits (1984) is quite simply that it does not allow for the relative sizes of contrasted groups. In the degenerate case of a dichotomy, his method is exactly equivalent to that obtained by the effect coding

of variables.

Although the above paragraphs indicate that effect coded variables have no place in the analysis of non-experimental data, that does not mean that they have no place in statistical enterprises. Effect coded variables are basically designed to compare treatment and control groups in balanced, randomized experimental designs. Modified effect coded variables, taking on the value 1 for a treatment group of interest, 0 for the other treatment groups, and -1 for the common control group are particularly useful in this context. However, in such experimental situations, the marginals are essentially meaningless and are fixed by the investigator. The only things of interest are the differences between the control group and the treatment groups. In particular, the way MCA coefficients split total differences between groups of unequal size would be as misleading in experimental situations as effect coded variables are in non-experimental work. Effect coded variables are also related to the topic of orthogonal polynomials, which are sometimes useful in the analysis of experimental data (see, for example, Hope, 1971, for an application to nonexperimental data of social and demographic relevance). While effect coded variables have a proper place in experimental designs, importing them to nonexperimental situations engenders confusion rather than clarity.

VI. Summary and Conclusions

The purposes of this paper have been primarily the didactic ones of sketching out the relationships between dummy variable and multiple classification analysis, illustrating these relationships with empirical data, indicating the various ways in which one can approach the derivation and estimation of MCA coefficients, and exhibiting the algebraic identities which enable one to pass from dummy variable coding to effect coding of categorical variables. We have also had occasion to take exception to the general applicability of a procedure recently suggested by Suits (1984) for transforming dummy variable coefficients into MCA-like coefficients which are unweighted for the relative sizes of categories.

The present paper also paves the way for the exposition of a

number of related topics. These include the analysis of statistical interactions in the framework of dummy variable and multiple classification analysis, the use of certain types of aggregate data to derive individual level MCA results, and the generalization of MCA-type analyses to situations in which it is unreasonable to assume the underlying functional relationship is additive in the various classifications.

Notes

1/ In fairness to Suits, it must be stressed that all of his examples are cases in which the weighted and unweighted sums of the MCA coefficients are equal, but he fails to distinguish his special case of equal numbers of cases at each level from the general situation in which they are unequal.

2/ These expressions for X_r and Z_s are arbitrary; the dummy variable for any one of the categories of either classification could have been expressed as a function of membership or non-membership in the remaining categories; the particular choice of X_r and X_s was dictated only by algebraic convenience.

3/ In this context, it is useful to note that the summation term on the right hand side of Eq. 31 disappears in the case of a dichotomy.

4/ For articles employing this basic framework, although not necessarily using regression methods, see Duncan (1968) and Siegel (1965).

References

Andrews, Frank, James Morgan, and John Sonquist. 1967. Multiple Classification Analysis. Ann Arbor, Michigan: University of Michigan, Survey Research Center.

Andrews, Frank, James Morgan, John Sonquist, and Laura Klem. 1973. Multiple Classification Analysis: A Report on a Computer Program for Multiple Regression Using Categorical Predictors. 2nd Edition. Ann Arbor, Michigan: University of Michigan, Institute for Social Research.

Blau, Peter M., and Otis Dudley Duncan. 1967. The American Occupational Structure. New York: John Wiley.

Draper, N.R., and H. Smith. 1966. Applied Regression Analysis. New York: John Wiley.

Duncan, Otis Dudley. 1964. "Residential areas and differential fertility," Eugenics Quarterly, 11: 82-89.

_____. 1968. "Inheritance of poverty or inheritance of race?" pp. 85-110 in Daniel P. Moynihan (ed.), On Understanding Poverty. New York: Basic Books.

Goldberger, Arthur S. 1964. Econometric Theory. New York: John Wiley.

Hill, T.P. 1959. "Analysis of the distribution of wages and salaries in Great Britain," Econometrica, 27: 355-381.

Hope, Keith. 1971. "Social mobility and fertility," American Sociological Review, 36: 1019-1032.

Johnston, J. 1972. Econometric Methods. 2nd Edition. New York: McGraw-Hill.

Kerlinger, F.N., and E.J. Pedhazur. 1973. Multiple Regression in Behavioral Research. New York: Holt, Rinehart, and Winston.

Keyfitz, Nathan. 1953. "A factorial arrangement of comparison of family size," American Journal of Sociology, 58: 470-480.

Kmenta, Jan. 1971. Elements of Econometrics. New York: Macmillan.

Lansing, John B., and James N. Morgan. 1971. Economic Survey Methods. Ann Arbor, Michigan: University of Michigan, Institute for Social Research, Survey Research Center.

Melichar, Emanuel. 1966. "Least-squares analysis of economic survey data," pp. 373-385 in 1965 Proceedings of the Business and Economic Statistics Section, American Statistical Association. Washington, D.C.: American Statistical Association.

Siegel, Paul M. 1965. "On the cost of being a Negro," Sociological Inquiry, 35: 41-57.

Suits, Daniel B. 1957. "Use of dummy variables in regression equations," Journal of the American Statistical Association, 52: 548-551.

_____. 1984. "Dummy variables: mechanics v. interpretation," The Review of Economics and Statistics, 66: 177-180.

Yates, Frank. 1934. "The analysis of multiple classifications with unequal numbers in the different classes," Journal of the American Statistical Association, 29: 51-66.