USING DUMMY VARIABLE REGRESSION FOR LONGITUDINAL ANALYSIS

Alan Fox and Bennie A. Clemmer
Office of Research and Statistics
Social Security Administration

The objective of our research was to investigate predictors of income change between 1968 and 1970, based on a longitudinal survey of persons nearing retirement age. Most of our predictor variables were categorical, and therefore a form of dummy variable regression was used. The dependent variable is a ratio of two numbers, and several predictor variables encompass both level and change, based on answers from two waves of the survey. Nonresponse and errors in measurement are much more serious than with cross-sectional analysis.

We concentrated on the largest group in the survey, married men and their wives. The possible correlates of income change that we used were levels and change in work status of the respondent and his wife, type and duration of his previous employment, and his education, health, and perception of retirement.

This work was based on data from the Longitudinal Retirement History Study being conducted by the Social Security Administration. This is a ten-year study of men and nonmarried women who were 58 to 63 years old in 1969, with reinterviews every two years until 1979, when they will be 68 to 73, and of course mostly retired. Since we only had the 1969 and 1971 interview data, we were somewhat restricted in the number of persons doing what we were most interested in, namely retiring. With later waves of data we will have larger numbers to work with.

In section 1 of this paper we describe the dependent and predictor variables. Section 2 describes the dummy variable regression package used, and briefly raises the question of interaction. Section 3 describes our results to date. And in Section 4 we briefly examine sampling errors.

1. The Dependent and Predictor Variables

The dependent variable was the change in total money income of the couple, from 1968 to 1970. This change can be measured in several ways: we concentrated on the ratio of 1970 income to 1968 income. Since we are dealing with two distinct points in time, changing underlying conditions must be accounted for so that income in the two years is comparable. Among these underlying conditions is the cost of living; therefore we have reduced 1970 income by the 11.6 percent rise in the Consumer Price Index. 1/ Also, as people retire the taxes on their income components change--notably from taxable earnings to nontaxable Social Security benefits--and their disposable income should reflect that fact. Some preliminary estimates are given later in this paper.

Nonresponse and incorrect response are major problems in longitudinal analysis, and caused a substantial decrease in usable sample size. With very conservative editing, letting a nonresponse on a single income item cause a nonresponse on the total, only about 60 percent of the couples gave usable income responses in both years. 2/

We postulated that many couples with extreme TMI ratios had reported incorrectly in one or the other survey year. Therefore, we eliminated the small number whose TMI ratios were less than 20 percent or more than 200 percent (the latter represents a doubling of income as the couple ages.) 3/ In so doing, we achieved a close-to-normal distribution of the dependent variable, and considerably reduced the variation to be explained by the predictors.

The level of TMI can be important, and has been dealt with by including the 1968-70 average as a predictor variable. This is better than using either 1968 or 1970 TMI to describe level, because of the possibility of incorrect responses: for instance, if a couple underreports its 1968 income but correctly reports its 1970 income, the TMI ratio will be high. Thus there will be a spurious correlation between low initial level and high increase. 4/

Certain predictor variables, such as occupation of the longest job and education, can be expected to remain unchanged from one survey wave to the next. Other variables may change, but only rarely; home ownership and urban-rural residence are examples. These variables are therefore entered only as absolute values. Other predictor variables encompass both level and change: for instance, respondent's and wife's employment, retirement status, health, and happiness. When using such composite variables the number of possible combinations can become unwieldy, and simplifications must be made. For instance, retirement status has four levels in each year (not retired, partly retired, completely retired, and nonresponse); a detailed transition matrix enabled us to choose the six most frequent combinations out of a possible 16, with a residual class including rare combinations or nonresponse in either year. 5/

Employment status was asked as of the week before the interview (1969 and 1971) and was a poor predictor of income in the calendar year preceding the interview (1968 and 1970). Therefore a proxy was used for employment status in the income year: if any earnings were reported in the income year, the respondent was considered employed in that year. Thus a person who quit work during the income year is counted as employed in that year.

Retirement status was also asked as of the survey week, and therefore is an imperfect

measure of retirement in the income year. However, it was included in the model because it might correlate with the amount of work in the income year.

2. The Programs Used

Most of the predictor variables were categorical. Therefore, a form of dummy-variable regression, Multiple Classification Analysis (MCA; see Andrews, et al.), was used. This program, developed at the University of Michigan Survey Research Center, differs from conventional dummy-variable packages mainly in that all categories of a predictor variable are represented, without omitting one of each set. The coefficients indicate the difference between each category and the grand mean, not the omitted category.6/ Furthermore, since it is not necessary to construct a large set of dummy variables, computer programming is simplified.

For each category of a predictor, two coefficients are printed. The unadjusted coefficient indicates the difference between the mean of the dependent variable for that category and the grand mean. The adjusted coefficient indicates the deviation from the grand mean for persons in that category, after accounting for the effects of all other predictors in the model.

Two summary statistics are presented for each predictor. Eta-square indicates the proportion of the total sum of squares explained by the predictor, without accounting for any other predictors. It is analogous to a simple correlation coefficient (squared). Beta-square, on the other hand, measures the importance of a predictor after accounting for the effects of all other predictors (Andrews, et al., p. 7).

The differences between beta-square and eta-square, and between the unadjusted and adjusted coefficients, indicate the degree of intercorrelation between a particular predictor and other predictor variables. Regression analysis attempts to sort out the separate influences of the various predictors. For instance, age might be expected to be an important predictor of income change when taken in isolation, but regression analysis reveals that this was due to the intercorrelation between age and work stoppage.

Additivity and Interaction.--In the additive linear model, to predict the value of the dependent variable one simply sums the coefficients associated with each category of every variable that describes the person. With interaction present, the effect on the dependent variable of being in one predictor category is not the same for all persons.7/ For instance, one might expect attitude toward retirement to have a different effect for retired workers than the nonretired. If substantial interaction is present, the coefficients on attitude for the entire sample would be misleading. Therefore, interaction should be accounted for in the model, either by adding interaction terms (such as attitude toward retirement crossed by retirement status), or by running separate regressions (for the retired and nonretired). Running separate regressions is especially useful if many interactions are present, or if the interactions involve more than two predictors.

An attractive way to locate interaction is with another program from Michigan, the Automatic Interaction Detector (AID). That program makes a series of splits on the dependent variable, such that at each stage the predictor chosen for the split accounts for the greatest decrease in unexplained variance (see Sonquist, et al.).

We have been somewhat disappointed in the results of AID trees. Often it is difficult to interpret the results of a complex tree. Furthermore, with slightly different samples the program may construct a substantially different tree if there were close competitors and if the first split is different. The variables that almost cause a split (that is, the near misses) often tell as much as the variables that do split. This is evident from a careful analysis of the printout, but is not clear from the actual tree of splits. Finally, although it gives a rough impression of which variables are important, either over the entire sample or for some specific subgroup, this impression must be confirmed in other ways. For instance, in the regression analysis reported later in this paper, wife's employment was quite important. However, that variable did not show up in an AID tree run on the entire sample, although a careful analysis of the printout would have given some indication of its importance.

Thus, AID can sometimes help to identify interactions, but our experience has shown that intuition and preliminary cross-tabulations can often be just as productive. For instance, we expect retired people to have different characteristics than nonretired persons. To test this, we will either run separate analyses for the two groups, or insert interaction terms into a single regression analysis.

3. Results

Results of the MCA analysis of the entire sample of married men and their wives are shown in Table 1. Predictors are listed in decreasing order of beta-square. It should come as no surprise that the three most important predictors relate to employment status of the respondent (V3 and V14) and of his wife (V6). Respondent's Employment 1968-70 (V3) and Respondent's Retirement Status 1969-71 (V14) are highly intercorrelated, as seen by the fact that beta-square is about one-half of eta-square for both predictors (this is also evident from a cross-tabulation of one predictor against the other).

The wife's employment (V6) was practically uncorrelated with any other variable, as seen by the close correspondence between beta-square and eta-square, and between the adjusted and unadjusted coefficients. Tabular analysis revealed

no tendency for the wife to quit work in response to her husband's retirement.

Nothing mattered beyond the three employment-related variables. This is clear from the small adjusted coefficients, the precipitous decline of beta-square, and is even clearer when a set of regressions is run, entering predictors in decreasing order of beta-square. This table gives the marginal contribution of the last predictor entered:

Predictor(s)	Total R^2	R^2 added
Husband's employment only (V3)	8.4%	8.4%
Adding retirement status (V3, V14)	11.0	2.6
Adding wife's employment (V3, V6, V14)	13.4	2.4
Adding average TMI (V3, V6, V10, V14)	13.6	.2
Adding 12 remaining predictors	13.8	.2

To predict the TMI ratio for a given individual or group, one theoretically must add the adjusted coefficients for all predictors in the model. In this particular model this is not particularly important, as only the three labor force variables had coefficients significantly different from zero (see discussion of sampling errors, below). Furthermore, the respondent's labor force variables (V3 employment status, and V14 retirement status) were practically uncorrelated with the wife's employment status (V6), and therefore in this model it is possible to consider the effect of the husband's work status without simultaneously considering the effect of the wife's work status. However, for the sake of precision we have rerun the regression with only the three labor force variables, and will discuss the three predictors simultaneously. The adjusted coefficients in the restricted model (Table 2) are practically the same as in the full model (Table 1); this is further indication that only these three variables were important.

The additive nature of the restricted model can be illustrated as follows. For couples where the respondent worked in 1968 but not in 1970 (V3, category 2) and was not retired in 1969 but completely retired in 1971 (V14, category 3) and where the wife was not employed in either year (V6, category 3), estimated 1970 TMI was about 65 percent of 1968 TMI, after adjusting for price increases (94.4 - 20.8 - 10.3 - .2 = 63.1%). As expected, where the husband only partly retired in 1971 the couple did somewhat better (94.4 - 20.8 - 7.2 - .2 = 66.2%).

If the wife worked in 1968 but not in 1970 (V6, category 2), the TMI ratio decreased by 11 percent. Thus, if she retired at the same time as her husband, their 1970 TMI decreased to about one-half of their 1968 TMI (94.4 - 20.8 - 10.3 - 11.3 = 52.0%).

Among couples where there was no change in employment status, total money income rose about as fast as prices. Where both husband and wife were employed in both years, and the husband was not retired in both years, the estimated TMI ratio was 101 percent (94.4 + 1.0 + 4.4 + 1.2 = 101.0%). Where the wife was not working in either year the TMI ratio was about 100 percent (94.4 + 1.0 + 4.4 - .2 = 99.6%).

The restricted model has been rerun using total money income net of Federal income and Social Security taxes. A description of the tax estimation procedure is available from the authors. For couples where both husband and wife quit work, and the husband retired completely, the estimated TMI ratio was 58.5 percent, 6.5 percentage points higher than when income was adjusted only for changes in the cost of living (Table 3). This reflects the fact that much retirement income is tax-free. On the other hand, for those who continued working the estimated TMI ratio was virtually the same as when gross income was used.

One obvious problem is that rather few RHS sample persons had quit work by 1970, and so our sample of "retirees" is small. The 1973 data, which have only recently arrived, should give us far more retirees to work with. In particular, we will want to focus on the determinants of TMI change among the retirees, and will probably do so by running a separate regression for that group, since we expect interaction to be present.

4. A Note on Sampling Errors

Sampling errors are not calculated by the MCA program, since it is assumed to be run on complex samples where sampling error is generally greater than in simple random samples. A useful approximation suggested by Michigan researchers is (Morgan, et al, Vol. I, p. 356):

For unadjusted category coefficients,

$$SE = \frac{\text{Standard deviation } \sqrt{DEFF}}{\sqrt{n_i}}$$

For adjusted category coefficients,

$$SE = \frac{\text{Standard deviation } \sqrt{1 - R^2}\sqrt{DEFF}}{\sqrt{n_i}}$$

Where SE = standard error of the predictor category,

Standard deviation = standard deviation of the dependent variable,

n_i = number of sample cases in the predictor category,

R^2 = coefficient of multiple determination for the run, and

DEFF = estimate of the design effect of the survey.8/

From previous work the design effect for RHS has been estimated to be about 1.2. Using these

approximations, all coefficients discussed above which involve a change in work status are significantly different from zero at the .01 level (3SE).

Note that some of the unadjusted coefficients in the full model (Table 1) were significant, while the corresponding adjusted coefficients were not. This reflects the adjustment process of regression analysis. For instance, while being 64 or 65 years old in 1971 (V11, category 3) might appear significant from a tabular (univariate) analysis, it is so only because of intercorrelation with other variables, notably change in employment status.

We are currently working on a Balanced Repeated Replication design to get more precise standard errors for MCA runs from this survey.

Footnotes

1/ Without the price adjustment, the results were as expected: the grand mean TMI ratio shifted upward by about 12 percentage points, while all coefficients remained the same.

2/ Nonresponse on asset income was the main problem. Slightly over 70 percent answered all other items in both years. Since asset income is a minor source for most couples, we might eliminate it entirely in future analysis, or substitute predicted values.

3/ Two percent had TMI ratios less than 20 percent, while 3 percent were over 200 percent (of the latter, fully one-half reported 1968 TMI below $2,000).

4/ This is known as "regression toward the mean." See Morgan et al, Vol. III, p. 47, footnote 1.

5/ There is usually no need to eliminate a case with a nonresponse in some predictor variable.

6/ Dummy-variable regression coefficients can be made equivalent to MCA coefficients. See Andrews et al, p. 46-47, and Sweeney and Ulveling.

7/ Interaction involves a combined effect of two or more predictor variables on the dependent variable that is greater or less than the sum of the individual effects. On the other hand, intercorrelation (often termed collinearity or multicollinearity) involves a systematic relationship between two or more predictor variables.

8/ The design effect (DEFF) is the ratio of the actual variance of a complex sample to the variance of a simple random sample of the same size.

References

Andrews, Frank M., James N. Morgan, John A. Sonquist, and Laura Klem, Multiple Classification Analysis, second edition, University of Michigan, 1973.

Kish, Leslie, and Martin R. Frankel, "Balanced Repeated Replications for Standard Errors," Journal of the American Statistical Association, September, 1970.

Morgan, James N., et al., Five Thousand American Families--Patterns of Economic Progress, (in 3 volumes, 1974-75), Institute for Social Research, The University of Michigan.

Morgan, James N., "Using Survey Data from the University of Michigan's Survey Research Center," American Economic Review, May, 1975, p. 250 - 256.

Sonquist, John A., Elizabeth Lauh Baker, and James N. Morgan, Searching for Structure, Institute for Social Research, The University of Michigan, 1971.

Sweeney, Robert E. and Edwin F. Ulveling, "A Transformation for Simplifying the Interpretation of Coefficients of Binary Variables in Regression Analysis," The American Statistician, December, 1972, p. 30.

Table 1.--Distributions and changes in total money income, married men and their wives, 1969-71 Retirement History Study 1/

General statistics

Grand mean (1970 TMI/1968 TMI) 1/	94.4%
Coefficient of variation 2/	32%
Multiple R^2	.138
Number of cases	3412

Predictor and class	Number of cases	Percent of cases	Unadjusted coefficient	Adjusted coefficient	Eta^2	$Beta^2$	Beta rank
V3: Respondent's employment 1968-70 5/							
Employed both years	2872	84%	2.5**	1.1			
Employed 1968, not in 1970	266	8	-29.3**	-20.7**	.084	.047	1
Not employed either year	255	8	- .1	7.6**			
Not employed 1968, employed 1970	19	1	4/	4/			
V14: Respondent's retirement status 1969-71							
Not retired 1969 or 1971	2150	63%	5.4**	4.0**			
Not 1969, partly 1971	268	8	- 7.4**	- 7.4**			
Not 1969, retired 1971	393	12	-14.1**	- 9.9**	.067	.035	2
Partly retired 1969 and 1971	104	3	- 4.5	- 4.2			
Partly 1969, retired 1971	88	3	-14.7**	- 6.4			
Retired 1969 and 1971	334	10	- 9.0**	- 6.6**			
Nonresponse; other	75	2	8.7*	5.9			
V6: Wife's employment 1968-70							
Employed both years	1218	36%	1.8	1.0			
Employed 1968, not in 1970	312	9	-12.6**	-11.3**	.027	.024	3
Not employed either year	1717	50	- .3	- .1			
Not employed 1968, employed 1970	165	5	13.9**	14.6**			
V10: Average TMI 1968-70							
$1-1999	146	4%	- 5.9*	- 2.0			
2000-3999	404	12	- 1.0	1.9			
4000-5999	563	17	- 3.0*	- 1.3			
6000-7999	668	20	- .6	- .7	.010	.003	4
8000-9999	559	16	- 1.8	- 2.0			
10000-14999	678	20	3.7*	1.9			
15000 and over	394	12	4.6*	1.4			
V18: Occupation of longest job							
Professional and technical	286	8%	2.1	- .5			
Farmers	332	10	- .7	3/			
Managers and proprietors	474	14	2.3	1.3			
Clerical and sales	277	8	3.3	1.5			
Crafts	818	24	- 1.4	- 1.4	.004	.002	5
Operatives	710	21	- 2.2	- 1.1			
Service: Private household	2	3/	4/	4/			
Service other than household	189	6	.7	1.1			
Laborers and farm foremen	295	9	.5	2.7			
Nonresponse	29	1	4/	4/			
V16: Happiness in 1969 and 1971							
Happy both years	2618	77%	.9	3/			
Not happy both years	215	6	- 6.6*	- .8			
Happy in 1969, not in 1971	254	7	- 5.5*	- 2.7	.008	.002	6
Not happy 1969, happy 1971	299	9	.5	2.4			
Nonresponse	26	1	4/	4/			
V17: Attitude toward retirement in 1971							
Very positive	403	12%	- 3.5*	- .6			
Positive	2072	61	.1	.1			
Negative	840	25	1.7	.5	.003	.002	7
Very negative	40	1	4/	4/			
Nonresponse	57	2	- 5.9	- 9.0*			

Table 1.(continued)--Distributions and changes in total money income, married men and their wives, 1969-71 Retirement History Study 1/

Predictor and class	Number of cases	Percent of cases	Unadjusted coefficient	Adjusted coefficient	Eta^2	$Beta^2$	Beta rank
V15: Work limitations in 1969 and 1971							
No limitations either year	1910	56%	2.9**	.8			
Limited in both years	775	23	- 5.7**	- 1.8			
No limitation 1969, limited in 1971	450	13	- 1.6	.5	.014	.002	8
Limited in 1969, not in 1971	273	8	- 1.6	- 1.8			
Nonresponse	4	3/	4/	4/			
V12: Home ownership in 1969							
Own home	2383	70%	- .5	- .8	.001	.002	9
Rent home	1029	30	- 1.2	1.8			
V7: Education of respondent							
0-4 grades	250	7%	- 3.3	- .8			
5-8 grades	1244	37	- 1.7	- .9			
9-11 grades	656	19	- .1	.5	.005	.001	10
12 grades	734	22	2.0	.7			
College	518	15	2.7	.7			
Nonresponse	10	3/	4/	4/			
V13: Father's education							
0-4 years	568	17%	.2	.8			
5-8 years	1156	34	.3	- .2			
9-12 years	328	10	3.6	2.2	.002	.001	11
College	169	5	- 1.9	- 2.4			
Nonresponse	1191	35	- 1.1	- .5			
V9: Duration of longest job							
0-10 years	615	18%	- .4	- .8			
11-20 years	1001	29	.7	.3			
21-30 years	957	28	.8	.5	.001	.001	12
31-40 years	603	18	- 2.2	- 1.3			
41 years and over	236	7	.3	2.0			
V11: Age of respondent in 1971							
60-61 years	1274	37%	1.8	- 1.0			
62-63 years	1132	33	.2	.6	.003	.001	13
64-65 years	1006	30	- 2.5*	.6			
V5: Race							
White	3121	92%	3/	- .2	3/	3/	14
Black, other	291	9	.2	1.6			
V8: Area of residence in 1969							
Urban	2227	65%	3/	- .3			
Rural farm	305	9	- .4	- .2	3/	3/	15
Rural nonfarm	880	26	.1	.7			
V4: Second pension receipt in 1970							
Receiving pension	594	17%	- 7.6**	- .6	.013	3/	16
Not receiving pension	2818	83	1.6*	.1			

1/ Restricted to married men and their wives reporting fully on all sources of income in both 1968 and 1970. The 1970 TMI has been adjusted for 11.6 percent rise in the cost of living (CPI). Cases with net TMI ratio under 20 percent or over 200 percent omitted. Marital status is as of 1971.
2/ C.V. = Standard deviation as percent of grand mean.
3/ Less than 0.5 (for eta-square and beta-square, less than .001).
4/ Not shown where base under 50 sample cases.
5/ Based on receipt of earned income in 1968 or 1970.
*Significantly different from zero at .05 level. } See discussion in text
**Significantly different from zero at .01 level. }

Source: MCA run 16 7/2/75

Table 2.--Restricted model with the three most important predictors only

General statistics

Grand mean (1970 TMI/1968 TMI) 1/	94.4%
Coefficient of variation 2/	32%
Multiple R^2	.134
Number of cases	3412

Predictor and class	Number of cases	Percent of cases	Unadjusted coefficient	Adjusted coefficient	Eta^2	$Beta^2$	Beta rank
V3: Respondent's employment 1968-70 5/							
Employed both years	2872	84%	2.5**	1.0			
Employed 1968, not in 1970	266	8	-29.3**	-20.8**	.084	.048	1
Not employed either year	255	8	- .1	7.9**			
Not employed 1968, employed 1970	19	1	4/	4/			
V14: Respondent's retirement status 1969-71							
Not retired 1969 or 1971	2150	63%	5.4**	4.4**			
Not 1969, partly 1971	268	8	- 7.4**	- 7.2**			
Not 1969, retired 1971	393	12	-14.1**	-10.3**	.067	.041	2
Partly retired 1969 and 1971	104	3	- 4.5	- 4.3			
Partly 1969, retired 1971	88	3	-14.7**	- 7.0*			
Retired 1969 and 1971	334	10	- 9.0**	- 8.4**			
Nonresponse; other	75	2	8.7*	5.5			
V6: Wife's employment 1968-70							
Employed both years	1218	36%	1.8	1.2			
Employed 1968, not in 1970	312	9	-12.6**	-11.3**	.027	.025	3
Not employed either year	1717	50	- .3	- .2			
Not employed 1968, employed 1970	165	5	13.9**	14.6**			

Footnotes: Same as Table 1.

Source: MCA run 16E

Table 3.--Restricted model with income net of Federal income and Social Security taxes.

General statistics

Grand mean (1970 TMI/1968 TMI) 1/	96.2%
Coefficient of variation 2/	30%
Multiple R^2	.116
Number of cases	3424

Predictor and class	Number of cases	Percent of cases	Unadjusted coefficient	Adjusted coefficient	Eta^2	$Beta^2$	Beta rank
V3: Respondent's employment 1968-70 5/							
Employed both years	2881	84%	2.3**	1.1			
Employed 1968, not in 1970	268	8	-25.9**	-18.9**	.071	.041	1
Not employed either year	256	8	- .6	5.9**			
Not employed 1968, employed 1970	19	1	4/	4/			
V14: Respondent's retirement status 1969-71							
Not retired 1969 or 1971	2157	63%	4.6**	3.6**			
Not 1969, partly 1971	269	8	- 6.3**	- 6.4**			
Not 1969, retired 1971	395	12	- 11.3**	- 7.9**	.052	.029	2
Partly retired 1969 and 1971	104	3	- 3.9	- 3.8			
Partly 1969, retired 1971	88	3	- 13.1**	- 6.2			
Retired 1969 and 1971	336	10	- 8.4**	- 7.0**			
Nonresponse: other	75	2	7.7 8	5.2			
V6: Wife's employment 1968-70							
Employed both years	1219	36%	1.4	.8			
Employed 1968, not in 1970	314	9	-12.1**	-10.9**	.031	.028	3
Not employed either year	1722	50	- .3	- .2			
Not employed 1968, employed 1970	169	5	15.9**	16.4**			

Footnotes: Same as Table 1.

Source: MCA Run 17A, 7/10/75

Number of cases differs slightly from other runs because screen on extreme cases is applied to TMI ratio net of taxes.