I. Introduction

In a medical or epidemiological setting, frequently information is collected for a large number of categorical variables. The relationship between these variables and one response (or dependent) variable is often of primary interest. When the number of independent variables is large, it becomes necessary to select a subset of "manageable" size for further analysis. This paper briefly reviews criteria and procedures based on these criteria for selecting an important subset of independent variables. SAS macros have been developed for implementing a categorical selection procedure for the situation where all variables are dichotomous; thus the following discussion is presented with the dichotomous case in mind. The use of the macros is discussed and an example based on the selection procedure proposed by Higgins and Koch (1977) is reviewed.

II. Measuring association across two-way tables

Initially, when considering the association between the dependent variable and each independent variable, a chi-square test of independence is often used. If we have observed frequencies $x_{ij}$ and expected frequencies $e_{ij}$, then the test statistic $X^2 = \sum \frac{(x_{ij} - e_{ij})^2}{e_{ij}}$ is asymptotically distributed as a $\chi^2$ with one degree of freedom under the hypothesis of independence. Another measure asymptotically distributed as $\chi^2$ under the hypothesis of independence is the likelihood ratio statistic:

$$G^2 = 2 \sum x_{ij} \log \frac{x_{ij}}{e_{ij}}.$$  

One desirable property of this statistic, which the chi-square statistic does not possess, is that it is minimized by maximum likelihood estimates. However, when both $X^2$ and $G^2$ are calculated, results rarely vary (Bishop, Fienberg, Holland (1975)).

Since it is usually not practical to consider all possible combinations of variables, generally a forward stepwise selection approach is used, with selection of the first important covariate based on one of the above measures. After this is done, it becomes necessary to assess the relationship between the dependent variable and remaining independent or candidate variables, taking into account the selected variable. Various methods, based on stratification by the selected factors, have been suggested. After step $s$, we will have selected $s$ variables, and stratifying on the levels of these factors yields $k = 2^s - 1$ two-by-two tables between the dependent and independent variable, for each remaining independent variable.

One procedure for estimating the association between the dependent variable and a candidate variable involves calculating the chi-square test statistic for each of the $k$ tables, and then summing them. This method does not take into account the direction of the differences $p_1 - p_2$ (where $p_1$ and $p_2$ are the observed proportions of a given level of the dependent variable within each of the two levels of the independent variable) in the various tables; it will detect departures from independence when the direction differs from that expected and vice versa, as well as when the directions are consistently the same.

Cochran (1954) suggested calculating the value

$$\sum \frac{(x_{ij} - e_{ij})^2}{e_{ij}}$$

and assigning a sign based on the direction of $p_1 - p_2$ in each table, summing these values then dividing by the square root of the number of tables. This has approximately a standard normal distribution and is valid if $p_1$ and $p_2$ are in the 20-80 percent range and the sample sizes of the tables are similar. Note that this will primarily detect departures from independence when the differences are consistently in the same direction. Fleiss (1973) points out serious drawbacks to this method, particularly when the sample sizes among the tables vary greatly.

Another method proposed by Cochran to handle more general situations involves the weighted mean difference. For the $i$th table, let $x_{ii}$ and $e_{ii}$ be defined as above, $n_{i1}$ and $n_{i2}$ are the number of observations at the two levels of the independent variable, $n_i$ is the total number of observations, $\hat{p}_i = \frac{n_{i1}}{n_i}$, and $d_i = \hat{p}_i^1 - \hat{p}_i^2$, then the test statistic

$$\sum \frac{w_i d_i^2}{\frac{1}{n_i} \sum w_i}$$

is compared to standard normal tables. Since this statistic is based on asymptotic binomial model results, sample sizes of at least 20 for each table are desirable (Landis, Leyman, Koch (1978)).

Mantel and Haenszel (1958) developed a slightly different test of association based on the hypergeometric distribution that is applicable even when individual table sizes are small. We have $k$ frequency tables, one for each stratum, of the following form:
stratum i

candidate 1
0 A_i B_i N_{ii}
dependent 1
C_i D_i N_{i1}
M_{i1} M_{i2} T_i

The test statistic is

\[ \chi^2 = \frac{1}{\sqrt{\sum_{i=1}^{k} \hat{E}(A_i) - \frac{k}{E(A_i)} | - 1/2)^2}} \]

where \( \hat{E}(A_i) = N_i M_{i1} M_{i2} / T_i \) and \( \hat{V}(A_i) = N_i N_{i2} M_{i1} M_{i2} / T_i^2 (T_i - 1) \).

Birch (1964) has shown that this statistic has an approximate chi-square distribution with one degree of freedom.

III. Variable selection

Higgins and Koch developed a categorical variable selection procedure and demonstrated its use with data from a study of the respiratory ailment byssinosis and possibly related covariates. The first variable selected was determined by the value of the chi-square test statistic (of byssinosis and each covariate) divided by the degrees of freedom. Subsequent variable selection was based on the chi-square statistic, divided by its degrees of freedom for the two-way table of the dependent variable, and each level of the previously selected variables combined with the levels of the candidate variables. The basis for terminating selection of variables involves two test statistics: (1) the sum of the individual chi-square statistics calculated from each level of selected variables, and (2) a chi-square statistic with one degree of freedom developed by Cochran, and Mantel and Haenszel; modified by Campbell, and Koch and Rienfurt. The first termination statistic is suggested for use in the early steps when sample sizes are more likely to allow the detection of both main effects and interaction between the candidate and previously selected variables. In the later steps, when the individual tables become sparse, and the degrees of freedom large, the second termination statistic is suggested as it is less affected by reductions in sample size. Selection of important covariates is concluded when the appropriate termination statistic is not significant at some predetermined level.

Freeman (1979) describes two methods of variable selection based on likelihood ratio test statistics. These techniques were developed for the situation where there is not a dependent variable designated in advance. The first method, conditional table selection, is analogous to Higgins' and Koch's termination criteria. The second method, compound table selection, is somewhat comparable to Higgins' and Koch's selection criteria. The two sets of variables identified by these methods are then suggested for further analysis.

Standish and Allred (1981) select variables based on reduction of the entropy function of the dependent variable. If, after selection of the first few variables, the sample size of any table is not large enough (less than 20), the Mantel Haenszel statistic is then used as the selection criteria.

IV. Use of categorical variable selection macros to implement Higgins-Koch selection procedure

Before this series of macros was developed, using the Higgins-Koch selection procedure in SAS was a tedious and time-consuming chore. SAS was useful in generating the appropriate tables and chi-square statistics, but much output was generated and the Mantel-Haenszel statistic could not be obtained. These macros were developed particularly for the situation where there are a large number of independent variables, and have simplified the necessary programming immensely.

One step of the selection procedure is generated each time a program with this series of macros is run. It is possible to run up to six steps, but this could easily be extended. The variables must all be dichotomous, coded 0 or 1, and the dependent variable must be named Y. The following statements need to be specified at the beginning of the program:

MACRO STEP number % - number must be 1, 2, 3, 4, 5, or 6
MACRO V SELECT list % - list variables selected by previous steps; last variable listed must be Y, the dependent variable
MACRO V REMAIN list % - list variables that have not been previously selected
MACRO SORT VAR variable % - variable is test statistic by which final listing is sorted; possible choices are described further in the following section.

The macro GEN MAC must be called in the user's program, and the statement "% INCLUDE MACROS;" (note that a file MACROS must be defined in JCL) has to appear in a data step following GEN MAC. Both of these must occur before the main macro VARSEL is called. The macros are listed in the appendix and the following example is presented to demonstrate and explain the results of their use.

V. Example

The data used are from the byssinosis study as reported by Higgins and Koch. Since dichotomous variables were necessary, two variables were included for years employment (<10, >10 and <20, >20) and workplace (1, 2 or 3 and 1 or 2, 3). Although these adjustments were made, the same set of important variables was selected. The details of the analysis and conclusions may be found in Higgins' and Koch's paper.
Four steps were necessary; the results of steps 1 and 2 are listed in Tables A and B. Table B (step 3) will be used to explain the output. Note that the variables WORK1 and YRS10 were selected in steps 1 and 2, respectively. There are 2 (step 1) lines printed for each remaining independent variable, with each line corresponding to a level of the previously selected variables. The order is determined by the order of variables in macro V_SELECT, and is as follows in this example:

<table>
<thead>
<tr>
<th>STRATUM</th>
<th>YRS10</th>
<th>WORK1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

The frequency tables are arranged in the following manner:

<table>
<thead>
<tr>
<th>smoke(candidate variable)</th>
</tr>
</thead>
<tbody>
<tr>
<td>stratum 1</td>
</tr>
<tr>
<td>1 0</td>
</tr>
<tr>
<td>byssinosis (Y)</td>
</tr>
<tr>
<td>0  A-986  B-1360</td>
</tr>
<tr>
<td>1  C-10  D-24</td>
</tr>
<tr>
<td>T-2380</td>
</tr>
</tbody>
</table>

CHI SQ is the uncorrected chi-square test statistic for each stratum and ODD is the odds ratio AD/BC. The variable OVER CHI is the overall chi-square test statistic, the variable that determines selection. (Note that the degrees of freedom is the same for each table, thus OVER CHI is not divided by this value.) SUM CHI is the termination statistic (1.), is the sum of each stratum's chi-square value, and PR SUM is the probability of a random variable with a chi-square distribution being larger than the SUM CHI value. MH CHI is the Mantel-Haenszel test statistic; PR MH is the probability of obtaining a larger value. RR ME is the associated overall relative risk estimate \( \left( \frac{\sum A_iD_i/T_i}{\sum B_iC_i/T_i} \right) \). The listings may be sorted by any of these variables, with DESCENDING OVER CHI seeming most appropriate for this selection procedure.

VI. Discussion

Higgins and Koch note that their selection procedure yields an important set of explanatory variables, but not necessarily the best set. Also, the order in which the variables are selected may not reflect their relative importance. The process of selecting a variable does not follow strictly defined criteria, but may be based on the selection and/or the termination statistics in combination with the user's insight regarding the data. Hence, the macros presented generate only one step of the procedure at a time so that the decision to select a variable can be carefully made.

The macros, as presented, allow the selection of six independent variables. This could be expanded, but we have found that after four or five steps many of the expected cell counts within the individual strata become prohibitively small. In addition to sample size, the within-strata chi-square statistic and odds ratio should be examined for location of possible associations and consistency (or lack of it) in the direction of associations.

Use of the statistics printed in the output and the SORT VAR option allow some flexibility in modifying the selection procedure from that given by Higgins and Koch. For example, when many of the independent variables have missing information, it may be preferable to divide the selection criterion by the number of observations with complete information for that factor and then make the selection based on this adjusted statistic. This would require only a simple programming statement in the _VARSEL_ macro.

References


### Table A
CATEGORICAL VARIABLE SELECTION USING SAS
HIGGINS KOCJ EXAMPLE

#### STEP NUMBER 1
VARIABLES PREVIOUSLY SELECTED: Y

<table>
<thead>
<tr>
<th>VARIABLE</th>
<th>STRATA</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>T</th>
<th>CHI_SQ</th>
<th>ODD</th>
<th>OVER_CHI</th>
<th>SUM_CHI</th>
<th>PR_SUM</th>
<th>MH_CHI</th>
<th>PR_MH</th>
<th>PR_MH</th>
</tr>
</thead>
<tbody>
<tr>
<td>WORK1</td>
<td>1</td>
<td>4690</td>
<td>564</td>
<td>60</td>
<td>105</td>
<td>5419</td>
<td>413.7257</td>
<td>14.4579</td>
<td>413.7257</td>
<td>413.7257</td>
<td>0.0000</td>
<td>410.7760</td>
<td>0.0000</td>
<td>14.5523</td>
</tr>
<tr>
<td>WORK12</td>
<td>1</td>
<td>3408</td>
<td>1846</td>
<td>42</td>
<td>123</td>
<td>5419</td>
<td>1074.109</td>
<td>5.4056</td>
<td>1074.109</td>
<td>1074.109</td>
<td>0.0000</td>
<td>195.6483</td>
<td>0.0000</td>
<td>5.4066</td>
</tr>
<tr>
<td>SEX</td>
<td>1</td>
<td>2788</td>
<td>2466</td>
<td>128</td>
<td>37</td>
<td>5419</td>
<td>38.6707</td>
<td>0.3268</td>
<td>38.6707</td>
<td>38.6707</td>
<td>0.0000</td>
<td>37.6838</td>
<td>0.0000</td>
<td>0.3286</td>
</tr>
<tr>
<td>SMOKER</td>
<td>1</td>
<td>2190</td>
<td>3064</td>
<td>40</td>
<td>125</td>
<td>5419</td>
<td>20.0921</td>
<td>1.8363</td>
<td>20.0921</td>
<td>20.0921</td>
<td>0.0000</td>
<td>19.5095</td>
<td>0.0000</td>
<td>1.8363</td>
</tr>
<tr>
<td>HH10</td>
<td>1</td>
<td>2788</td>
<td>4466</td>
<td>102</td>
<td>63</td>
<td>5419</td>
<td>18.0969</td>
<td>0.5996</td>
<td>18.0969</td>
<td>18.0969</td>
<td>0.0000</td>
<td>17.5080</td>
<td>0.0000</td>
<td>0.5996</td>
</tr>
<tr>
<td>YRS20</td>
<td>1</td>
<td>1902</td>
<td>3252</td>
<td>76</td>
<td>9</td>
<td>5419</td>
<td>6.7097</td>
<td>0.6645</td>
<td>6.7097</td>
<td>6.7097</td>
<td>0.0000</td>
<td>6.2899</td>
<td>0.0000</td>
<td>0.6645</td>
</tr>
<tr>
<td>RACE</td>
<td>1</td>
<td>3424</td>
<td>1830</td>
<td>92</td>
<td>73</td>
<td>5419</td>
<td>6.2130</td>
<td>1.4835</td>
<td>6.2130</td>
<td>6.2130</td>
<td>0.0000</td>
<td>5.8122</td>
<td>0.0000</td>
<td>1.4835</td>
</tr>
</tbody>
</table>

### Table B
CATEGORICAL VARIABLE SELECTION USING SAS
HIGGINS KOCJ EXAMPLE

#### STEP NUMBER 2
VARIABLES PREVIOUSLY SELECTED: YRS10, WORK1 Y

<table>
<thead>
<tr>
<th>VARIABLE</th>
<th>STRATA</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>T</th>
<th>CHI_SQ</th>
<th>ODD</th>
<th>OVER_CHI</th>
<th>SUM_CHI</th>
<th>PR_SUM</th>
<th>MH_CHI</th>
<th>PR_MH</th>
<th>PR_MH</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMOKE</td>
<td>1</td>
<td>986</td>
<td>1260</td>
<td>10</td>
<td>24</td>
<td>2380</td>
<td>2.1924</td>
<td>1.7400</td>
<td>2.1924</td>
<td>2.1924</td>
<td>0.0055</td>
<td>0.9817</td>
<td>0.017</td>
<td>1.8363</td>
</tr>
<tr>
<td>SMOKE</td>
<td>2</td>
<td>81</td>
<td>101</td>
<td>11</td>
<td>57</td>
<td>310</td>
<td>7.6070</td>
<td>2.6070</td>
<td>7.6070</td>
<td>7.6070</td>
<td>0.0000</td>
<td>1.0000</td>
<td>0.0000</td>
<td>2.6070</td>
</tr>
<tr>
<td>SMOKE</td>
<td>3</td>
<td>1004</td>
<td>1340</td>
<td>12</td>
<td>14</td>
<td>2370</td>
<td>0.1558</td>
<td>0.8741</td>
<td>0.1558</td>
<td>0.1558</td>
<td>0.0000</td>
<td>1.0000</td>
<td>0.0000</td>
<td>0.8741</td>
</tr>
<tr>
<td>SEX</td>
<td>1</td>
<td>119</td>
<td>263</td>
<td>7</td>
<td>30</td>
<td>359</td>
<td>4.7931</td>
<td>2.5123</td>
<td>4.7931</td>
<td>4.7931</td>
<td>0.0000</td>
<td>1.0000</td>
<td>0.0000</td>
<td>2.5123</td>
</tr>
<tr>
<td>SEX</td>
<td>2</td>
<td>538</td>
<td>68</td>
<td>0</td>
<td>310</td>
<td>310</td>
<td>1.1287</td>
<td>0.0000</td>
<td>1.1287</td>
<td>1.1287</td>
<td>0.0000</td>
<td>1.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>SEX</td>
<td>3</td>
<td>1009</td>
<td>1344</td>
<td>7</td>
<td>19</td>
<td>2370</td>
<td>2.0583</td>
<td>2.0583</td>
<td>2.0583</td>
<td>2.0583</td>
<td>0.0000</td>
<td>1.0000</td>
<td>0.0000</td>
<td>2.0583</td>
</tr>
<tr>
<td>SEX</td>
<td>4</td>
<td>267</td>
<td>55</td>
<td>3</td>
<td>359</td>
<td>359</td>
<td>1.9724</td>
<td>0.4238</td>
<td>1.9724</td>
<td>1.9724</td>
<td>0.0000</td>
<td>1.0000</td>
<td>0.0000</td>
<td>0.4238</td>
</tr>
<tr>
<td>RACE</td>
<td>1</td>
<td>2209</td>
<td>137</td>
<td>34</td>
<td>0</td>
<td>2380</td>
<td>1.0688</td>
<td>0.0000</td>
<td>1.0688</td>
<td>1.0688</td>
<td>0.0000</td>
<td>1.0000</td>
<td>0.0000</td>
<td>1.0688</td>
</tr>
<tr>
<td>RACE</td>
<td>2</td>
<td>156</td>
<td>86</td>
<td>46</td>
<td>22</td>
<td>310</td>
<td>0.2371</td>
<td>0.8675</td>
<td>0.2371</td>
<td>0.2371</td>
<td>0.0000</td>
<td>1.0000</td>
<td>0.0000</td>
<td>0.8675</td>
</tr>
<tr>
<td>RACE</td>
<td>3</td>
<td>997</td>
<td>1347</td>
<td>9</td>
<td>17</td>
<td>2370</td>
<td>1.3961</td>
<td>0.7252</td>
<td>1.3961</td>
<td>1.3961</td>
<td>0.0000</td>
<td>1.0000</td>
<td>0.0000</td>
<td>0.7252</td>
</tr>
<tr>
<td>RACE</td>
<td>4</td>
<td>62</td>
<td>260</td>
<td>3</td>
<td>34</td>
<td>359</td>
<td>2.7026</td>
<td>0.7252</td>
<td>2.7026</td>
<td>2.7026</td>
<td>0.0000</td>
<td>1.0000</td>
<td>0.0000</td>
<td>0.7252</td>
</tr>
<tr>
<td>YRS20</td>
<td>1</td>
<td>1720</td>
<td>618</td>
<td>27</td>
<td>7</td>
<td>2380</td>
<td>0.5731</td>
<td>0.7249</td>
<td>0.5731</td>
<td>0.5731</td>
<td>0.0000</td>
<td>0.9922</td>
<td>0.0000</td>
<td>0.7249</td>
</tr>
<tr>
<td>YRS20</td>
<td>2</td>
<td>174</td>
<td>68</td>
<td>49</td>
<td>19</td>
<td>310</td>
<td>0.0007</td>
<td>0.9922</td>
<td>0.0007</td>
<td>0.0007</td>
<td>0.0000</td>
<td>1.0000</td>
<td>0.0000</td>
<td>0.9922</td>
</tr>
<tr>
<td>YRS20</td>
<td>3</td>
<td>2344</td>
<td>0</td>
<td>25</td>
<td>2370</td>
<td>1.1515</td>
<td>0.7252</td>
<td>1.1515</td>
<td>0.7252</td>
<td>1.1515</td>
<td>0.0000</td>
<td>1.0000</td>
<td>0.0000</td>
<td>0.7252</td>
</tr>
<tr>
<td>YRS20</td>
<td>4</td>
<td>322</td>
<td>0</td>
<td>37</td>
<td>359</td>
<td>359</td>
<td>0.1236</td>
<td>0.7252</td>
<td>0.1236</td>
<td>0.1236</td>
<td>0.0000</td>
<td>1.0000</td>
<td>0.0000</td>
<td>0.7252</td>
</tr>
<tr>
<td>WORK12</td>
<td>1</td>
<td>1742</td>
<td>644</td>
<td>24</td>
<td>10</td>
<td>2380</td>
<td>0.2535</td>
<td>1.2017</td>
<td>0.2535</td>
<td>0.2535</td>
<td>0.0000</td>
<td>0.9912</td>
<td>0.0126</td>
<td>1.1515</td>
</tr>
<tr>
<td>WORK12</td>
<td>2</td>
<td>242</td>
<td>0</td>
<td>68</td>
<td>310</td>
<td>310</td>
<td>0.0000</td>
<td>0.9912</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>1.0000</td>
<td>0.0000</td>
<td>0.9912</td>
</tr>
<tr>
<td>WORK12</td>
<td>3</td>
<td>1666</td>
<td>678</td>
<td>18</td>
<td>8</td>
<td>2370</td>
<td>0.1425</td>
<td>1.0921</td>
<td>0.1425</td>
<td>0.1425</td>
<td>0.0000</td>
<td>1.0000</td>
<td>0.0000</td>
<td>1.0921</td>
</tr>
<tr>
<td>WORK12</td>
<td>4</td>
<td>322</td>
<td>0</td>
<td>37</td>
<td>359</td>
<td>359</td>
<td>0.0000</td>
<td>1.0921</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>1.0000</td>
<td>0.0000</td>
<td>1.0921</td>
</tr>
</tbody>
</table>
\* VARIABLES MUST ALL BE CODED 0 OR 1
\* DEPENDENT VARIABLE MUST BE NAMED Y
\* EXAMPLE OF JCL

 Appendices

**EXAMPLE OF JCL**

```
DNAME=MACHOS
UN1T=SPM.
P=

SPACE=(TRK,1,10,5,0,0)
CH7=(RECFS=F,RECFM=BLK,LSIZE=1600)
```

**USER SPECIFIES STFP NUMBER (1 THROUGH 6)**

MACRO 3

**USER SPECIFIES SELECTED AND DEPENDENT VARIABLES ENDING WITH DEPENDENT**

MACRO V_SELECT Y=1 WORK 1

**USER SPECIFIES REMAINING INDEPENDENT VARIABLES**

MACRO V_REMAIN SEX MACK WORK 1? Y=0 20

**USER SPECIFIES CRITERIA FOR SORING FINAL LISTING**

MACRO SORT_VAR DEPENDING OVER CHI

**THIS MACRO GENERATES MACROS NEEDED LATER IN PROGRAM**

MACRO GEN_MAC DATA _NULL_

```
ARRAY X V_REMAIN; DO OVER X; NVAR=_1_; EN1;
LASTN='NX1'; SUBSTR(LASTN,3)=NVAR;
LASTSUM='SX1'; SUBSTR(LASTSUM,3)=NVAR;
LASTLAG='LS1'; SUBSTR(LASTLAG,3)=NVAR;
ARRAY SEL V_SELECT; DO OVER SEL; NSEL=_1_; EN1;
LASTMAIN='MX1'; SUBSTR(LASTMAIN,3)=NSEL;
```

**MACROS**

```
FILE MACROS:

PUT 'MACRO NX NX1-LASTNX3% MACHOS SX SX1-LASTSUM3% MACHOS LLAG_S
LS1-LASTLAG3% MACRO V_MAINT MX1-LASTMAIN3% 3:
```

**THIS MACRO GENERATES THE OVERALL CHI SQUARE VALUE**

MACRO OVERCHI PROC MATRI

```
FETCH MAKI DATA=NEW; MATRI=MAT1: OUTPUT MAI1 OUT=MAI2:
DATA CALCHI1 CALCHI2 SET MAI2:
DO i=1 TO 32: DROP 1:
 IF _N_ EQ 4*I-3 OR _N_ EQ 4*I-2 THEN OUTPUT CALCHI1:
 IF _N_ EQ 4*I-1 OR _N_ EQ 4*I THEN OUTPUT CALCHI2: EN1:
```

**MACRO GEN_MAC DATA _NULL_**

```
ARRAY X V_REMAIN; DO OVER X; NVAR=_1_; EN1;
LASTN='NX1'; SUBSTR(LASTN,3)=NVAR;
LASTSUM='SX1'; SUBSTR(LASTSUM,3)=NVAR;
LASTLAG='LS1'; SUBSTR(LASTLAG,3)=NVAR;
ARRAY SEL V_SELECT; DO OVER SEL; NSEL=_1_; EN1;
LASTMAIN='MX1'; SUBSTR(LASTMAIN,3)=NSEL;
```

**MACRO _VARSEL_**

PROC NOSORT:

```
PROC SORT:
```

**THIS MACRO GENERATES NECESSARY STRATA, CALCULATES AND PRINTS RESULTS**

MACRO _VARSEL_

PROC SORT: BY V_SELECT:

```
PROC MEANS NOPRINT N SUM HY V_REMAIN: VAR V_REMAIN:
OUTPUT OUT=NEW N=NX SUM=SX1:
```

**DATA NEW2**

```
ARRAY NUM NX;
ARRAY V_S V_SELECT;
ARRAY A V_REMA1N;
ARRAY LAG_SEL LAG_S:
DO OVER SUM; IF SUM=_1_ THEN SUM=0 END1;
```

**KEEP V_REMAIN V_SELECT CELL PREV RETAIN PREV:**

```
IF _N_=1 THEN PREV='C11:
IF PREV='C11 AND Y=1 THEN LINK ZERO1:
IF PREV='A11 AND Y=0 THEN LINK ZEH0:
IF Y=1 THEN GO TO L2:
CELL='A11 DO OVER NUM: X=NUM-SUM: END1 OUTPUT:
CELL='B11 DO OVER NUM: X=NUM END1 OUTPUT:
PREV='A11 RETURN:
```

**L21**

```
CELL='A11 DO OVER NUM: X=NUM-SUM: END1 OUTPUT:
CELL='F11 DO OVER NUM: X=NUM END1 OUTPUT:
PREV='C11 RETURN:
```

716