Constructing Indicator Variables with SAS

Last revised: 05-23-96

Some regression procedures in SAS and other statistical software do not automatically generate indicator (dummy) variables for classification variables, their interactions, or polynomial effects. For such procedures, you must specify the model directly in terms of distinct variables. For example, if you want to use SAS's REG procedure to fit a model with a classification variable like sex that is coded M or F, you first need to compute the indicator variable, usually in a DATA step.

SAS introduced a new procedure called GLMMOD in version 6.07 that computes these kinds of effects automatically. The GLMMOD procedure essentially constitutes the model-building front end for the GLM procedure; it constructs and outputs the design matrix for the model you specify. You can take the output from the GLMMOD procedure and use it as input to other SAS procedures or write out the output to an ASCII file for input to other statistical software packages.

To illustrate, consider the following data:

```
OBS    HEIGHT    WEIGHT    AGE    SEX    TYPE

  1      69       112.5     14     M       2
  2      56        84.0     13     F       1
  3      65        98.0     13     F       2
  4      62       102.5     14     F       3
  5      63       102.5     14     M       3
  :       :           :      :     :       :
  :       :           :      :     :       :
```

Assume the intent is to use regression to model the weights of school children as a function of their height, sex, and body type. The interaction of sex with age and body type also is of interest. The SAS code below illustrates how to use the GLMMOD procedure to generate the design matrix:

```
proc glmmod outdesign=xx outparm=parms;
   class sex type;
   model weight=height age sex type sex*type age*sex;
run;
```

If you are familiar with the GLM procedure, you will notice that the syntax is very similar. The CLASS statement is where you identify the classification variables to be used in the analysis. The MODEL statement names the dependent variable(s) and independent effects. The OUTDESIGN= option on the PROC statement names an output data set to contain the columns of the design matrix. OUTPARM= names an output data set to contain information that identifies each of the columns of the design matrix.

The OUTDESIGN= Data Set

The OUTDESIGN= data set contains an observation for each observation in the input data set, with the dependent variable(s) and a variable for each column of the design matrix, with names COL1, COL2, . . .   Following is a listing of the OUTDESIGN= data set generated by GLMMOD in the above example:

```
     W
     E                                 C  C  C  C  C   C   C
     I    C   C   C  C  C  C  C  C  C  O  O  O  O  O   O   O
O    G    O   O   O  O  O  O  O  O  O  L  L  L  L  L   L   L
B    H    L   L   L  L  L  L  L  L  L  1  1  1  1  1   1   1
S    T    1   2   3  4  5  6  7  8  9  0  1  2  3  4   5   6

1  112.5  1  69  14  0  1  0  1  0  0  0  0  0  1  0   0  14
2   84.0  1  56  13  1  0  1  0  0  1  0  0  0  0  0  13   0
3   98.0  1  65  13  1  0  0  1  0  0  1  0  0  0  0  13   0
4  102.5  1  62  14  1  0  0  0  1  0  0  1  0  0  0  14   0
5  102.5  1  63  14  0  1  0  0  1  0  0  0  0  0  1   0  14
:     :   :   :   :  :  :  :  :  :  :  :  :  :  :  :   :   :
:     :   :   :   :  :  :  :  :  :  :  :  :  :  :  :   :   :
```

The OUTPARM= Data Set

The OUTPARM= data set contains an observation for each column of the design matrix with the following variables:

- _COLNUM_ identifying the number of the column of the design matrix corresponding to the observation
- EFFNAME containing the name of the effect that generates the column of the design matrix corresponding to the observation
- the CLASS variables, with the values they have for the columns corresponding to the observation, or blanks if they are not involved with the effect associated with the column.

Following is a listing of the OUTPARM= data set generated by GLMMOD in the above example:

```
OBS    _COLNUM_    EFFNAME     SEX    TYPE

  1        1       INTERCEPT
  2        2       HEIGHT
  3        3       AGE
  4        4       SEX           F
  5        5       SEX           M
  6        6       TYPE                 1
  7        7       TYPE                 2
  8        8       TYPE                 3
  9        9       SEX*TYPE      F      1
 10       10       SEX*TYPE      F      2
 11       11       SEX*TYPE      F      3
 12       12       SEX*TYPE      M      1
 13       13       SEX*TYPE      M      2
 14       14       SEX*TYPE      M      3
 15       15       AGE*SEX       F
 16       16       AGE*SEX       M
```

Use the OUTPARM= data set to determine which effect is contained in each column of the OUTDESIGN= data set. For example, from observation 1 in the above OUTPARM= data set, you determine that the variable, COL1, in the OUTDESIGN= data set corresponds to the intercept effect. Note how the variable SEX (coded M and F) has been recoded into two indicator variables, COL4 and COL5. COL4 is coded as 1 for women and 0 for men. COL5 is coded as 1 for men and 0 for women. In general, if a CLASS variable has *m* levels, GLMMOD will generate *m* variables in the design matrix (OUTDESIGN= data set). This means that the model produced by GLMMOD is overparameterized. There are more columns for these effects than there are degrees of freedom for them. So, you need to drop some of the columns when you fit the model. To illustrate, below is the SAS code you would use to fit the model with the REG procedure:

```
proc reg data=outdesign;
   model weight = col2 col3 col4 col6 col7 col9 col10 col15;
run;
```

Note that the intercept column was not included because the REG procedure adds the intercept by default.

Writing Output to an ASCII File

You may want to use the output from GLMMOD as input to some other statistical package. In this case, you need to write the OUTDESIGN= data set to an ASCII file. This is done with a DATA step as illustrated below:

```
data _NULL_;
   set xx;   /* xx was the name specified for OUTDESIGN= */
   file "~/dissert/design.dat";
   put weight 8.2
       (col2 col3 col4 col6 col7 col9 col10 col15) (4.);
run;
```

Details on writing ASCII data can be found in SSCC Pub. # 3-1, *Using SAS on VMS*, SSCC Pub #7-4, *Using SAS on UNIX*, or *The SAS Language Guide*.

Other Features of the GLMMOD Procedure

WEIGHT, FREQ, and BY statements can also be used with GLMMOD. WEIGHT and FREQ variables are transferred to the OUTDESIGN= data set without change. The ORDER= option is also available which allows you to change the sort order of the classification variables. The sort order is what determines which parameters in the model correspond to each level in the data. Details about these features and others can be found in *SAS Technical Report P-229 SAS/STAT Software: Changes and Enhancements*. This book is available for a short term loan from the CDE Print Library in Social Science 4457.