

PROC GLM

PROC GLM is set up mainly for testing of statistical hypotheses. It uses a less than full rank coding for the indicator variables in the design matrix. For instance, its design matrix for two factors (3 levels & 2 levels, 2 obs. per cell) is

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 \end{bmatrix}$$

Interactions would be coded as six extra columns (products of cols 2-4 with 5-6). The matrix $\mathbf{X}'\mathbf{X}$ is singular, and the normal equations do not have a solution.

GLM uses a **generalized inverse** which allows a partial solution to the normal equations.

The ordinary inverse of \mathbf{A} produces

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$$

A property of the ordinary inverse is

$$\mathbf{A}^{-1}\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}$$

A **generalized inverse** of the matrix \mathbf{A} , denoted \mathbf{A}^- , is any matrix such that

$$\mathbf{A}^- \mathbf{A} \mathbf{A}^- = \mathbf{A}^-$$

One such matrix is found by finding a smaller matrix \mathbf{A} which can be inverted. Write

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}$$

where \mathbf{A}_{11} is an invertible matrix (say, $m \times m$). Then

$$\mathbf{A}^- = \begin{bmatrix} \mathbf{A}_{11}^{-1} & \mathbf{0}_{12} \\ \mathbf{0}_{21} & \mathbf{0}_{22} \end{bmatrix}$$

is a generalized inverse, where the $\mathbf{0}$'s are matrices of zeros. For example, recall the matrix

$$\mathbf{A} = \begin{bmatrix} 3 & 1 \\ 6 & 2 \end{bmatrix}$$

A generalized inverse is

$$\mathbf{A}^- = \begin{bmatrix} \frac{1}{3} & 0 \\ 0 & 0 \end{bmatrix}$$

A partial solution to a system of equations

$$\mathbf{A}\boldsymbol{\beta} = \mathbf{c}$$

is given by

$$\mathbf{b} = \mathbf{A}^- \mathbf{c}$$

This amounts to “zeroing out” as many equations and variables necessary to get a solvable system of equations. For instance, our system of equations given by

$$3\beta_0 + 1\beta_1 = 12$$

$$6\beta_0 + 2\beta_1 = 8$$

becomes, with the above generalized inverse,

$$\mathbf{b} = \mathbf{A}^- \mathbf{c} = \begin{bmatrix} \frac{1}{3} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 12 \\ 8 \end{bmatrix} = \begin{bmatrix} 4 \\ 0 \end{bmatrix}$$

In other words, drop the second equation, set $\beta_1 = 0$, and solve for β_0 .

The normal equations are

$$(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$$

GLM calculates a partial solution (\mathbf{b} , an $m \times 1$ vector) to the normal equations in the form

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^- \mathbf{X}'\mathbf{y}$$

Why? Statistical hypotheses for the general linear model can be written in matrix form as

$$\mathbf{L}\boldsymbol{\beta} = \mathbf{0}$$

where \mathbf{L} is a row vector of constants. For instance, in a linear regression model, the hypothesis of zero slope results from

$$\mathbf{L} = [0 \quad 1]$$

It turns out that for certain forms of \mathbf{L} (linear combinations of the β_j s called **estimable functions**) the linear function

$$\mathbf{Lb}$$

is an unbiased estimate of $\mathbf{L}\boldsymbol{\beta}$. Actually, \mathbf{L} can even be a matrix, with $\mathbf{L}\boldsymbol{\beta} = \mathbf{0}$ giving a whole set of simultaneous hypotheses on estimable functions. The sums of squares for the hypotheses are

$$SS(\mathbf{L}\boldsymbol{\beta} = \mathbf{0}) = (\mathbf{Lb})'[\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}'](\mathbf{Lb})$$

A test of the hypotheses (as H_0) is provided by the F statistic given by

$$F = \frac{SS(\mathbf{L}\boldsymbol{\beta} = \mathbf{0})/m}{SS(\text{error})/(\text{df for unrestricted model})}$$

where $SS(\text{error})$ is calculated from the generalized inverse:

$$SS(\text{error}) = \mathbf{y}'[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{y}$$

The partial solutions in **b** can be obtained with the SOLUTION option in the MODEL statement, for instance:

```
MODEL Y=A B A*B / SOLUTION;
```

However, those values are not of much interest (unless you have advanced interests).

If one wants to report a model form, for use in prediction say, that contains categorical predictor variables, one might consider the following process: (1) develop the model (i.e. what variables to include) in PROC GLM, then (2) code full rank indicator variables corresponding to the model for use in PROC REG. Use the coefficients reported by PROC REG.