The GLM procedure implements dummy-variable regression to handle analysis-of-variance or analysis-of-covariance models. Because the procedure can use excessive computer resources, particularly for large problems, PROC GLM should not be used where classical textbook formulas provide correct results.*

3.2 THE DUMMY-VARIABLE MODEL

This section presents the analysis-of-variance model using dummy variables, methods for solving the resulting overspecification problem, and the analysis for the model with PROC GLM. For simplicity, an analysis-of-variance model with one-way structure, such as would result from a completely randomized design, illustrates the discussion. In application, however, such a structure is adequately (and more economically) analyzed by using PROC ANOVA (section 2.3.4).

3.2.1 The Simplest Case: One-Way Structure

Data for the one-way structure consist of measurements classified according to a one-dimensional criterion. An example of this kind of structure is the set of student exam scores, where each student is taught by one of three teachers. The exam scores are thus grouped or classified according to TEACHER. The most straightforward model for data of this type is

$$y_{ij} = \mu_i + \varepsilon_{ij}$$

where

y_{ij}	represents the jth measurement in the ith group.
μ_i	represents the population mean for the ith group.
ε_{ij}	represents the random error with mean=0 and variance=σ_2.
$i=1, \ldots, t$	where t equals the number of groups.
$j=1, \ldots, n_i$	where n_i equals the number of observations in the ith group.

This is called the means or μ-model because it uses the means $\mu_1, \ldots, \mu_t$ as the basic parameters in the mathematical expression to model the data (Hocking and Speed 1975). The corresponding estimates of these parameters are

$$\hat{\mu}_1 = \bar{y}_{1.}$$
$$\vdots$$
$$\hat{\mu}_t = \bar{y}_{t.}$$

where $\bar{y}_{i.}=(\Sigma_j y_{ij})/n_i$ is the mean of n_i observations in group i.

* For nested or heirarchical analyses, use PROC NESTED. For balanced factorial structures, use PROC ANOVA.

In these situations, the statistical inference of interest is often about differences between the means of the form $(\mu_i - \mu_{i'})$ or between the means and some reference or baseline value μ. Therefore, many statistical textbooks present a model for the one-way structure that employs these differences as basic parameters. This is the familiar analysis-of-variance model illustrated in section **2.3.4**:

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

where μ equals the reference value and

$$\tau_i = \mu_i - \mu \quad .$$

Thus, the means can be expressed as

$$\mu_i = \mu + \tau_i \quad .$$

It is important to understand that the definition or interpretation of the parameter μ is arbitrary. The definition may depend on the situation at hand, and, in fact, it may not be necessary to define μ at all.

For the implementation of the dummy-variable model, the analysis-of-variance model

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

is rewritten as a regression model

$$y_{ij} = \beta_0 x_0 + \beta_1 x_1 + \ldots + \beta_t x_t + \varepsilon_{ij}$$

where $\beta_0 = \mu$, $\beta_i = \tau_i$, and the dummy variables $x_0, \ldots, x_t$ are defined as follows:

x_0	always equals 1.
x_1	equals 1 for an observation in group 1 and 0 otherwise.
x_2	equals 1 for an observation in group 2 and 0 otherwise.
.	
.	
.	
x_t	equals 1 for an observation in group t and 0 otherwise.

In matrix notation, the model equations for the data become

$$\mathbf{Y} = \begin{bmatrix} y_{11} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ \vdots \\ y_{2n_2} \\ \vdots \\ y_{t1} \\ \vdots \\ y_{tn_t} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_t \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{1n_1} \\ \varepsilon_{21} \\ \vdots \\ \varepsilon_{2n_2} \\ \vdots \\ \varepsilon_{t1} \\ \vdots \\ \varepsilon_{tn_t} \end{bmatrix} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Thus, the matrices of the normal equations are

$$\boldsymbol{X}'\boldsymbol{X} = \begin{bmatrix} n_{\cdot} & n_1 & n_2 & \cdots & n_t \\ n_1 & n_1 & 0 & \cdots & 0 \\ n_2 & 0 & n_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ n_t & 0 & 0 & \cdots & n_t \end{bmatrix}, \quad \boldsymbol{X}'\boldsymbol{Y} = \begin{bmatrix} Y_{\cdot\cdot} \\ Y_{1\cdot} \\ Y_{2\cdot} \\ \vdots \\ Y_{t\cdot} \end{bmatrix}$$

where $Y_{i.}$ and $Y_{..}$ are totals corresponding to $\bar{y}_{i.}$ and $\bar{y}_{..}$. The normal equations $(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{Y}$ are equivalent to the set

$$\begin{aligned}
\hat{\beta}_0 + \hat{\beta}_1 &= \bar{y}_{1.} \\
\hat{\beta}_0 + \hat{\beta}_2 &= \bar{y}_{2.} \\
&\cdot \\
&\cdot \\
&\cdot \\
\hat{\beta}_0 + \hat{\beta}_t &= \bar{y}_t \quad .
\end{aligned}$$

Because there are only t equations, there is obviously no unique solution for the $(t+1)$ estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_t$. This follows from the fact that the reference parameter β_0 is arbitrary and is the reason that the model is said to be overspecified. Similarly, the $\mathbf{X}'\mathbf{X}$ matrix describing the set of normal equations is of less-than-full rank, or singular. Clearly, in this model, the first row of $\mathbf{X}'\mathbf{X}$ is equal to the sum of the other t rows. The same relationship exists among the columns of $\mathbf{X}'\mathbf{X}$.

3.2.2 Obtaining Useful Estimates

There are two popular methods for obtaining solutions in spite of the overspecification problem: 1) the restrictions method and 2) the method of the generalized inverse. The latter is used by PROC GLM. This section reviews both methods in order to put the approach used by PROC GLM into perspective.

The restrictions method is based on the fact that any definition of one of the parameters in the model (say the reference parameter) causes the other parameters to be uniquely defined. The definition can be restated in the form of a restriction. Another view of the term restriction in the statistical literature is that the parameters are defined to have a unique interpretation, and the corresponding estimates are then required to coincide with the definition of the parameters.

One approach is to define one of the τ_i equal to zero, say $\tau_t=0$. In this case, μ becomes the mean of the tth group $\mu_\tau=\mu+\tau_t=\mu$, and τ_i becomes the difference between the mean for the ith group and the mean for the tth group, $\tau_i=\mu_i-\mu=\mu_i-\mu_t$.

The corresponding restriction on the solution to the normal equations is to require $\hat{\tau}_t=0$. Requiring $\hat{\tau}_t=0$ leads automatically to a unique set of values for the remaining set of estimates $\hat{\mu}, \hat{\tau}_1, \ldots, \hat{\tau}_{t-1}$. This occurs because τ_t is dropped from the linear model, and, consequently, the column corresponding to τ_t is dropped from the $\mathbf{X}$ matrix, producing the following model equation:

To illustrate the effects of unbalanced data on the estimation of differences between means and computation of sums of squares, consider the data in this two-way table:

		B	
		1	2
A	1	7,9	5
	2	8	4,6

Within level 1 of B, the cell mean for each level of A is 8; that is, $\bar{y}_{11.}=(7+9)/2=8$ and $\bar{y}_{21.}=8$; hence, there is no evidence of a difference between the levels of A within level 1 of B. Similarly, there is no evidence of a difference between levels of A within level 2 of B because $\bar{y}_{12.}=5$ and $\bar{y}_{22.}=(4+6)/2=5$. Hence, you may conclude that there is no evidence in the table of a difference between the levels of A. However, the marginal means for A are

$$\bar{y}_{1..} = (7 + 9 + 5)/3 = 7 \quad \text{and} \quad \bar{y}_{2..} = (8 + 4 + 6)/3 = 6 \quad .$$

The difference of $7-6=1$ between these marginal means may be erroneously interpreted as measuring an overall effect of the factor A. Actually, the observed difference between the marginal means for the two levels of A measures the effect of factor B in addition to the effect of factor B. This can be verified by expressing the observations in terms of the analysis-of-variance model $y_{ijk}=\mu+\alpha_i+\beta_j$. (For simplicity, the interaction and error terms have been left out of the model.)

		B	
		1	2
A	1	$7 = \mu + \alpha_1 + \beta_1$ $9 = \mu + \alpha_1 + \beta_1$	$5 = \mu + \alpha_1 + \beta_2$
	2	$8 = \mu + \alpha_2 + \beta_1$	$4 = \mu + \alpha_2 + \beta_2$ $6 = \mu + \alpha_2 + \beta_2$

The difference between marginal means for A_1 and A_2 is shown to be

$$\begin{aligned}\bar{y}_{1..} - \bar{y}_{2..} &= (1/3)[(\alpha_1 + \beta_1) + (\alpha_1 + \beta_1) + (\alpha_1 + \beta_2)] \\ &\quad - (1/3)[(\alpha_2 + \beta_1) + (\alpha_2 + \beta_2) + (\alpha_2 + \beta_2)] \\ &= (\alpha_1 - \alpha_2) + (1/3)(\beta_1 - \beta_2) \quad .\end{aligned}$$

Thus, instead of estimating $(\alpha_1 - \alpha_2)$, the difference between the marginal means of A estimates $(\alpha_1-\alpha_2)$ plus a function of the factor B parameters $(\beta_1-\beta_2)/3$. In other words, the difference between the A marginal means is biased by factor B effects.

The null hypothesis about A that would normally be tested is

$$H_0: \alpha_1 - \alpha_2 = 0 \quad .$$

However, for this example, the sum of squares for A computed by PROC ANOVA can be shown to equal $3(\bar{y}_{1..} - \bar{y}_{2..})^2/2$; hence, the PROC ANOVA F test for A actually tests the hypothesis

$$H_0: (\alpha_1 - \alpha_2) + (\beta_1 - \beta_2)/3 = 0$$

which involves the factor B difference $(\beta_1 - \beta_2)$ in addition to the factor A difference $(\alpha_1 - \alpha_2)$.

In terms of the μ model $y_{ijk} = \mu_{ij} + \varepsilon_{ijk}$, you usually want to estimate $(\mu_{11} + \mu_{12})/2$ and $(\mu_{21} + \mu_{22})/2$ or the difference between these quantities. However, the A marginal means for the example are

$$\bar{y}_{1..} = (2\mu_{11} + \mu_{12})/3 + \bar{\varepsilon}_{1..}$$

and

$$\bar{y}_{2..} = (\mu_{21} + 2\mu_{22})/3 + \bar{\varepsilon}_{2..}$$

which estimate $(2\mu_{11} + \mu_{22})/3$ and $(\mu_{21} + 2\mu_{22})/3$, respectively. These estimates are functions of the usually irrelevant cell frequencies and may be useless.

In summary, a major problem in the analysis of unbalanced data is the contamination of differences between factor means by effects of other factors. The solution to this problem is to adjust the means to remove the contaminating effects.

This phenomenon is related to multicollinearity in regression (Freund and Littell 1986) where appropriate estimates are obtained by defining partial regression coefficients and associated partial sums of squares. Using this analogy, it is logical to expect that the partial coefficients and associated statistics of a dummy-variable regression may provide appropriate statistics for the analysis of unbalanced data. Although this expectation is essentially correct, the overspecification of the dummy-variable model and the use of estimable functions opens additional possibilities for defining estimates.

3.3.2 Sums of Squares Computed by PROC GLM

PROC GLM recognizes different theoretical approaches to analysis of variance by providing four types of sums of squares and associated statistics. In general, these approaches relate to

1. the orthogonality of effects, that is, tests on one factor being independent of (often called adjusted for or partial of) tests of other effects.
2. the involvement of the cell sample sizes in the linear function of the parameters being tested.

The four types of sums of squares in PROC GLM are called Type I, Type II, Type III, and Type IV sums of squares (Goodnight 1978).

Type I functions retain the properties discussed in Chapter 1, "Regression." They correspond to adding each source (factor) sequentially to the model in the order listed. For example, the Type I sum of squares for the first factor listed is the same as PROC ANOVA would compute for that effect. It reflects differences between unadjusted means of that factor as if the data consisted of a one-way structure. The TYPE I SS may not be particularly useful for analysis of unbalanced multiway structures but may be useful for nested models (section 3.5), polynomial

The four sums of squares are requested in PROC GLM as options in the MODEL statement. For example, the following SAS statement specifies the printing of Type I and Type IV sums of squares:

```
MODEL . . . / SS1 SS4;
```

Any or all types can be requested. If no sums of squares are specified, PROC GLM computes the Type I and Type III sums of squares by default.

The next two sections interpret the different sums of squares in terms of reduction notation and the μ-model.

3.3.3 Interpreting Sums of Squares in Reduction Notation

Using reduction notation (section **1.1.2**) is difficult in the context of dummy variables because of the singularity of the **X'X** matrix before restrictions are imposed on the parameters of the model. The term *restriction on parameters* can be understood to mean that the parameters are uniquely defined. The interpretation of the parameters and associated sums of squares may be dependent on the restriction imposed.

As an example, consider a 2 x 3 factorial structure with n_{ij} observations in the cell in row i, column j. The equation for the model is

$$y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \varepsilon_{ijk}$$

where

i equals 1, 2
j equals 1, 2, 3
k equals 1, . . . , n_{ij}

and $n_{ij}>0$ for all i, j is assumed. An expression of the form $R(\alpha \mid \mu, \beta)$ means the same as $R(\alpha_1, \alpha_2 \mid \mu, \beta_1, \beta_2, \beta_3)$. The sums of squares printed by PROC GLM can be interpreted in reduction notation most easily under the restrictions

$$\Sigma_i \alpha_i = \Sigma_j \beta_j = \Sigma_i \alpha\beta_{ij} = \Sigma_j \alpha\beta_{ij} = 0 \tag{3.1}$$

that is, by taking a full-rank X matrix given by

$$
\mathbf{X} = \begin{bmatrix}
\mu & \alpha_1 & \beta_1 & \beta_2 & \alpha\beta_{11} & \alpha\beta_{12} \\
1 & 1 & 1 & 0 & 1 & 0 \\
. & . & . & . & . & . \\
. & . & . & . & . & . \\
. & . & . & . & . & . \\
1 & 1 & 1 & 0 & 1 & 0 \\
1 & 1 & 0 & 1 & 0 & 1 \\
. & . & . & . & . & . \\
. & . & . & . & . & . \\
. & . & . & . & . & . \\
1 & 1 & 0 & 1 & 0 & 1 \\
1 & 1 & -1 & -1 & -1 & -1 \\
. & . & . & . & . & . \\
. & . & . & . & . & . \\
. & . & . & . & . & . \\
1 & 1 & -1 & -1 & -1 & -1 \\
1 & -1 & 1 & 0 & -1 & 0 \\
. & . & . & . & . & . \\
. & . & . & . & . & . \\
. & . & . & . & . & . \\
1 & -1 & 1 & 0 & -1 & 0 \\
1 & -1 & 0 & 1 & 0 & -1 \\
. & . & . & . & . & . \\
. & . & . & . & . & . \\
. & . & . & . & . & . \\
1 & -1 & 0 & 1 & 0 & -1 \\
1 & -1 & -1 & -1 & 1 & 1 \\
. & . & . & . & . & . \\
. & . & . & . & . & . \\
. & . & . & . & . & . \\
1 & -1 & -1 & -1 & 1 & 1
\end{bmatrix}
\begin{matrix}
\\ \\
\left.\vphantom{\begin{matrix}a\\b\\c\\d\\e\end{matrix}}\right\} n_{11} \text{ rows for obs. in 11 cell} \\
\left.\vphantom{\begin{matrix}a\\b\\c\\d\\e\end{matrix}}\right\} n_{12} \text{ rows for obs. in 12 cell} \\
\left.\vphantom{\begin{matrix}a\\b\\c\\d\\e\end{matrix}}\right\} n_{13} \text{ rows for obs. in 13 cell} \\
\left.\vphantom{\begin{matrix}a\\b\\c\\d\\e\end{matrix}}\right\} n_{21} \text{ rows for obs. in 21 cell} \\
\left.\vphantom{\begin{matrix}a\\b\\c\\d\\e\end{matrix}}\right\} n_{22} \text{ rows for obs. in 22 cell} \\
\left.\vphantom{\begin{matrix}a\\b\\c\\d\\e\end{matrix}}\right\} n_{23} \text{ rows for obs. in 23 cell}
\end{matrix}
$$

With this set of restrictions or definitions of the parameters, the sums of squares that result from the MODEL statement

```
MODEL Y=A B A*B / SS1 SS2 SS3 SS4;
```

are summarized below.

Sums of Squares Printed by PROC GLM in Reduction Notation

Effect	Type I	Type II	Type III = Type IV
A	$R(\alpha \mid \mu)$	$R(\alpha \mid \mu, \beta)$	$R(\alpha \mid \mu, \beta, \alpha\beta)$
B	$R(\beta \mid \mu, \alpha)$	$R(\beta \mid \mu, \alpha)$	$R(\beta \mid \mu, \alpha, \alpha\beta)$
A*B	$R(\alpha\beta \mid \mu, \alpha, \beta)$	$R(\alpha\beta \mid \mu, \alpha, \beta)$	$R(\alpha\beta \mid \mu, \alpha, \beta)$

You should be careful when using reduction notation with less-than-full-rank models. If no restrictions had been specified on the model for the two-way structure above, then $R(\alpha \mid \mu, \beta, \alpha\beta)=0$ because the columns of the X matrix corresponding to the α_i would be linearly dependent on the columns corresponding to μ and the $\alpha\beta_{ij}$.

In addition, the dependence of reduction notation on the restrictions imposed cannot be overemphasized. For example, imposing the restriction

$$\alpha_2 = \beta_3 = \alpha\beta_{21} = \alpha\beta_{22} = \alpha\beta_{13} = \alpha\beta_{23} = 0 \tag{3.2}$$

results in a different value for $R(\alpha \mid \mu, \beta, \alpha\beta)$. Although the restrictions of equation 3.1 are those that correspond to the sums of squares computed by PROC GLM, the restrictions of equation 3.2 are those that correspond to the (biased) parameter estimates computed by PROC GLM. This is illustrated in section **3.3.4**.

There is a relationship between the four types of sums of squares and four types of data structures in a two-way classification. The relationship derives from the principles of adjustment that the sums of squares types obey. Letting n_{ij} denote the number of observations in level i of factor A and level j of factor B, the four types of data structures are listed below:

1. equal cell frequencies: n_{ij}=common value for all i, j.
2. proportionate cell frequencies: $n_{ij}/n_{il}=n_{kj}/n_{kl}$ for all i, j, k, l.
3. disproportionate, nonzero, cell frequencies: $n_{ij}/n_{il} \neq n_{kj}/n_{kl}$ for some i, j, k, l, but $n_{ij}>0$ for all i, j.
4. empty cell(s): $n_{ij}=0$ for some i, j.

The display below shows the relationship between sums of squares types and data structure types pertaining to the following MODEL statement:

```
MODEL Y=A B A*B / SS1 SS2 SS3 SS4;
```

For example, writing III=IV indicates that Type III is equal to Type IV.