Other Linear Models

Recall: One-way ANOVA model equation:

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

SLR model equation:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

● These seem quite different and are used in different data analysis situations.

● But these and other models can be unified. They are each examples of the general linear model.

Dummy Variables

● The one-way ANOVA model may be represented as a regression model by using dummy variables.

Dummy variables (indicator variables): Take only the values 0 and 1 (sometimes -1 in certain contexts).

● One-way ANOVA model (above) is equivalent to:

$$Y = \mu X_0 + \tau_1 X_1 + \tau_2 X_2 + \cdots + \tau_t X_t + \varepsilon$$

where we define these dummy variables:

X_0 =

X_1 =

X_2 =

.

.

X_t =

Example: Suppose we have a one-way analysis with two observations from level 1, two observations from level 2, and three observations from level 3. The X matrix of the "regression" would look like:

● The Y-vector of response values and the vector of parameter estimates would be:

<u>Problem</u>: It turns out that $X^T X$ is not invertible in this case.

● There are t = 3 non-redundant equations and t + 1= 4 unknown parameters here.

● We fix this by adding one extra restriction to the parameters.

● Most common (we used this before): Force $\sum_{i=1}^{t} \tau_i = 0$ by defining $\tau_t = -\tau_1 - \ldots - \tau_{t-1}$.

● Using this approach, we need t – 1 dummy variables to represent t levels.

● If an observation comes from the <u>last</u> level, it gets a value of –1 for <u>all</u> dummy variables X_1, X_2, …, X_{t-1}.

X matrix from previous data set using this approach:

● **Another option: Force the <u>last</u> $\tau_i = 0$.**

● **These options give different numerical estimates for the parameters, but all conclusions about <u>effects and contrasts</u> will be <u>the same</u>.**

<u>Unbalanced Data</u>

● **Using the standard ANOVA formulas is easy, but it will give wrong results when data are unbalanced (different numbers of observations across cells).**

● **Dummy variable approach always gives correct answers.**

<u>Illustration</u>: A unbalanced 2-factor factorial study. (Table 11.3 data, p. 585)

● **<u>Question</u>: Does factor A have a significant effect on the response? (For simplicity, <u>ignore</u> any interaction between A and C for this example).**

Recall: Our F-statistic formula for this type of test was:

F* =

and SSA =

● **This formula is based on the variation between the marginal means $\bar{Y}_{1\bullet\bullet}$ and $\bar{Y}_{2\bullet\bullet}$**

● For the Table 11.3 data:

$\bar{Y}_{1\bullet\bullet} =$

$\bar{Y}_{2\bullet\bullet} =$

→ Based on this, there is <u>some</u> sample variation between the means for levels 1 and 2 of factor A.

● However, let's look at the sample means for levels 1 and 2 of A, <u>separately at each level of C</u>:

For level 1 of C:

$\bar{Y}_{11\bullet} =$

$\bar{Y}_{21\bullet} =$

For level 2 of C:

$\bar{Y}_{12\bullet} =$

$\bar{Y}_{22\bullet} =$

● These results imply that (at each level of C) there is <u>no</u> sample variation between the means for levels 1 and 2 of factor A.

• Which conclusion is correct?

• Our model is (recall there is no interaction term):

<u>Note</u>: $\bar{Y}_{11\bullet} - \bar{Y}_{21\bullet}$ **is an estimate of:**

Also, $\bar{Y}_{12\bullet} - \bar{Y}_{22\bullet}$ **is an estimate of:**

• So these <u>do</u> estimate the true difference in the means for levels 1 and 2 of factor A.

But … $\bar{Y}_{1\bullet\bullet} - \bar{Y}_{2\bullet\bullet}$**, for these data, is:**

which estimates:

● This is <u>not</u> the true difference in factor A's level means that we wanted to estimate.

● For balanced data, the magnitudes of all the coefficients would be the same and everything would cancel out properly.

● With unbalanced data, we need to adjust for the fact that the various <u>cell means</u> are based on <u>different numbers of observations</u> per cell.

● Using a dummy variable regression model implies the effect of factor A is estimated <u>holding factor C constant</u> → produces correct results.

● Analysis for unbalanced data involves the <u>least squares means</u>, not the ordinary <u>factor level means</u>.

● The least squares mean (for, say, level 1 of factor A) is the <u>unweighted average</u> of the cell sample means corresponding to level 1 of factor A. With unbalanced data, this is different than simply averaging all response values for level 1 of factor A. (see example)

● With unbalanced data in the two-way ANOVA, our F-tests about the factors use the Type III sums of squares, rather than the ordinary (Type I) ANOVA SS.

● See example for calculating these F-statistics correctly.

<u>Example</u>: (Table 11.3 data)

● **Least squares means:**

● **Correct F-tests about factor effects:**

● **More complicated example: Suppose A has 3 levels and C has 2 levels.**

● **Now we need to use 3 – 1 = 2 dummy variables for A and 2 – 1 = 1 dummy variable for C.**