Multi-Dimensional Tables

Categorical data with more than two factors are referred to as multi-dimensional distributions. Procedure CATMOD will be used for analyses concerning such data. PROC CATMOD may also be used to analyze one-and two-way data structures (as will be seen later), however it is an effective means to approach more complex data structures.

Response Functions

PROC CATMOD utilizes a different technique to do categorical analysis than the 'Pearson type' chi-square. The analysis is based on a transformation of the cell probabilities. This transformation is called the response function. The exact form of the response function depends on the data type and it is normally motivated by certain theoretical considerations. SAS offers many different forms of response functions and even allows the user to specify their own, however, the most common (default) is the Generalized Logit. This function is defined as:

$$\text{Generalized Logit} = \text{LOG}(p_i/p_k),$$

where p_i is the ith cell probability and p_k is the last cell probability. The ratio of p_i/p_k is called an **odds ratio** and the log of the odds ratio is just a comparison of the ith category to the last, on a log scale. The logit can be rewritten as:

$$\text{Generalized Logit} = \text{LOG}(p_i) - \text{LOG}(p_k).$$

It should be noted that if there are k categories, then there will be only k-1 response functions since the kth one will be zero.

Model Types

PROC CATMOD will analyze two different model types. The difference between them is somewhat philosophical, but in most cases the bottom line is that they reach common conclusions. The two types are referred to as Linear models and Log-Linear models. Linear models have well defined distinctions between dependent and independent factors. This is most analogous to the usual analysis of

variance. However, many categorical situations are not so clearly defined. For example, in the survey data introduced earlier the distinction between the dependent and independent variables may not be obvious. The Log-Linear model makes no such distinction, and instead, looks for association among all the variables. Even when clear definitions exist for the variables (such as a designed experiment) the Log-Linear approach may be more informative. The exact specification for each type will be demonstrated later.

Estimation of either model type in PROC CATMOD can be done by least squares procedures or maximum likelihood techniques. The maximum likelihood method, an iterative procedure, is designated as the default. It is considered to be a better method for this type of data especially when the cell counts are small.

Statements

MODEL

The MODEL statement in PROC CATMOD can be very similar to that of other Procedures such as ANOVA and GLM. The basic form is:

MODEL ***response-effects* = *design-effects* / *options*;**

Depending on the model type, the response and design effects may look different, but the general form is the same. Design-effects are those variables which may explain potential variability, such as treatments in the case of a Linear model. All notational conventions supported by PROC GLM will also be accepted by PROC CATMOD. Some examples for the Linear model are:

```
MODEL R = A B A*B;
```

or equivalently

```
MODEL R = A|B;     .
```

Nesting of terms is also allowed.

```
MODEL R = A B(A);     .
```

Log-Linear models require additional statements and key words which will be addressed later.

As with most MODEL statements in SAS, there are several options available. PROC CATMOD is very generous in its output. Since it uses an iterative estimation method (ML), it will print extensive information concerning every iteration. It will also print the levels of

the response-effects, design-effects, and the actual value of the response functions. In many cases the additional information may be suppressed with the options NOITER, NORESPONSE, NODESIGN, and NOPROFILE. Other important options include ML, and GLS to explicitly request the maximum likelihood or least squares estimation methods. Two-way tables of cross classifications can be obtained with FREQ, and the predicted values for cell frequencies and probabilities are printed with PRED=FREQ and PRED=PROB, respectively.

WEIGHT

The WEIGHT statement in PROC CATMOD works in the same manner as it did in PROC FREQ (handout #4). It lists a variable which represents the number of times an observation occurs. This statement is only required if the data is in the count form.

DIRECT

A DIRECT statement is used when a continuous variable is to be included in the analysis. It must always precede the MODEL statement. The effect of a continuous variable is analogous to an

analysis of covariance for continuous data. When the response function is a logit, the resulting analysis is called logistic regression and is simply a regression done on logit transformed proportions. Other procedures such as PROC LOGISTIC and PROC REG specifically address this type of analysis however, and would probably provide a more informative analysis.

LOGLIN

In order to run a log-linear model in PROC CATMOD, a new statement must be used in conjunction with special key words in the MODEL statement. Since the log-linear philosophy does not distinguish between independent and dependent variables, the user is left with how to put the same variables on both sides of the equation. In the survey data for example, gender, degree, and field were responses to the survey, but at the same time they are the factors of interest for hypothesis testing. To deal with this SAS has split the expression of the MODEL statement into two parts. A general log-linear model would be stated as:

```
MODEL A*B*C = _RESPONSE_/ML;
LOGLIN A B A*B;
```

The MODEL statement lists the response categories on the left side and leaves the right hand side to a general term _RESPONSE_. This is a key word and tells SAS to expect further definition later. The LOGLIN statement provides that definition. This looks familiar to what would appear on the right hand side of a PROC GLM MODEL statement, that is the main effects of A & B and their interaction.

Examples

In the survey data a linear model may be proposed to measure the influence of DEGREE and FIELD on the GENDER response. SAS codes to examine this model would be:

```
PROC CATMOD DATA=SURVEY;
     WEIGHT COUNTS;
     MODEL GENDER = DEGREE FIELD DEGREE*FIELD/ML NOGLS;
```

Which produces the following output:

MAXIMUM-LIKELIHOOD ANALYSIS-OF-VARIANCE TABLE

Source	DF	Chi-Square	Prob
INTERCEPT	1	1.90	0.1682
DEGREE	1	2.92	0.0877
FIELD	5	3.65	0.6015
DEGREE*FIELD	5	1.24	0.9406
LIKELIHOOD RATIO	0	.	.

The title terminology for the above table is **not** a correct one. This is **not** an analysis of variance, rather a table of chi-square statistics for **various effects**. The chi-square statistics are yet another version of the $P^2$and are called **Wald** chi-squares. They should be very close in value to the corresponding Pearson's chi-square from PROC FREQ. The likelihood ratio at the bottom of the table is also a chi-square statistic which measures the combined effect of terms left out of the model. Since this model specified all possible terms (e.g. a saturated model), there is nothing left to test and the line can be ignored. From the table it can be concluded that DEGREE and FIELD have no influence with regard to gender in this survey. Note that this examines the effect of DEGREE and FIELD on the response of GENDER and **does not** directly test for differences in the response to DEGREE or FIELD. This is the distinction between the linear and log-linear model. The next example examines the log-linear approach.

```
PROC CATMOD DATA=SURVEY;
     WEIGHT COUNTS;
     MODEL GENDER*DEGREE*FIELD = _RESPONSE_/ML NOGLS;
     LOGLIN GENDER|DEGREE|FIELD;
```

Notice that the expression of variables in the LOGLIN statement follow the same syntax rules as the MODEL statement. It is allowable to

give only main effects or certain interaction terms in any combination appropriate for the statistical model.

MAXIMUM-LIKELIHOOD ANALYSIS-OF-VARIANCE TABLE

Source	DF	Chi-Square	Prob
GENDER	1	1.90	0.1682
DEGREE	1	24.03	0.0000
GENDER*DEGREE	1	2.92	0.0877
FIELD	5	30.07	0.0000
GENDER*FIELD	5	3.65	0.6015
DEGREE*FIELD	5	3.19	0.6713
GENDER*DEGREE*FIELD	5	1.24	0.9406
LIKELIHOOD RATIO	0	.	.

The form of the output is similar to that of the linear model, but this time has the appearance of a full factorial design. Of interest are the terms GENDER*DEGREE, GENDER*FIELD, and GENDER*DEGREE*FIELD. These are exactly the same as the terms from the linear model. When the linear model listed the term DEGREE and its chi-square, it is actually looking at the **association** between GENDER and DEGREE. In contrast, the DEGREE term in

the log-linear model directly measures the **effect** of DEGREE as a response regardless of gender, which in this case is highly significant. The likelihood ratio chi-square again measures the effect of terms left out of the model.

PROC CATMOD also has the ability to handle the one-way and two-way data structures. The following example reproduces the one-way test of equal proportions for REGION using the log-linear model.

```
PROC CATMOD DATA=SURVEY;
     WEIGHT COUNTS;
     MODEL REGION = _RESPONSE_/ML NOGLS;
     LOGLIN REGION;
```

Although the chi-square values will not be **exactly** the same, they should be close to each other. Similar models could be derived to examine desired two-way data structures.

There is an obvious advantage to doing analyses using PROC CATMOD. The preceding output indicated what factors are significant in the model, but PROC CATMOD can go further by examining which levels within a factor are causing the significance. The DATA STEP or PROC FREQ can not easily do this. PROC CATMOD uses a

technique similar to PROC GLM , i.e.: the CONTRAST statement, to address this.

Contrasts

In order to understand contrasts in PROC CATMOD, the earlier discussion of response functions must be considered. PROC CATMOD uses a linear combination of response functions to do its estimation. However, the number of response functions is always one less than the number of categories. The problem arises when the user needs to compare K categories with only K-1 response function parameters. SAS handles this by having all K parameters sum to zero, by definition, so that the Kth parameter is equal to the negative sum of all other parameters. To illustrate this, the one-way example will be used.

The one-way example had 5 regions defined, thus K=5. The response functions would then be:

$LOG(p_1/p_5)$, $LOG(p_2/p_5)$, $LOG(p_3/p_5)$, and $LOG(p_4/p_5)$.

Therefore, only four parameters will be estimated, say α_1, α_2, α_3, and

α_4. The fifth parameter, α_5, will be defined as:

$$\alpha_5 = -(\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4)$$

This information will then have to be used to determine the coefficients for the desired contrasts. If the contrast only involves the first four categories, then the coefficients are easily calculated. For example, to contrast region 1 (CA) with region 2 (EA) the contrast would be:

$$1*\alpha_1 + -1*\alpha_2 + 0*\alpha_3 + 0*\alpha_4 + 0*\alpha_5$$

leading to the CONTRAST statement:

```
CONTRAST 'CA vs EA' _RESPONSE_ 1 -1 0 0;
```

Notice that in the CONTRAST statement, the fifth parameter coefficient is **omitted**. Otherwise the contrast is similar to PROC GLM. The difference will be apparent when the fifth category is included as part of the contrast. For example, suppose the contrast between region 2 (EA) and region 5 (PC) is to be made. The desired contrast is:

$$0*\alpha_1 + 1*\alpha_2 + 0*\alpha_3 + 0*\alpha_4 + -1*\alpha_5.$$

The definition for α_5 must be used to reduce the contrast to four coefficients. The contrast now becomes:

$$0*\alpha_1 + 1*\alpha_2 + 0*\alpha_3 + 0*\alpha_4 + -1*(-\alpha_1 - \alpha_2 - \alpha_3 - \alpha_4).$$

Grouping like terms yields:

$$1*\alpha_1 + 0*\alpha_2 + 1*\alpha_3 + 1*\alpha_4,$$

thus, the CONTRAST statement will be:

```
CONTRAST 'EA vs PC' _RESPONSE_ 1 0 1 1;
```

Although this looks strange, it will perform the desired test! Because the parameterization of PROC CATMOD differs from PROC GLM, users should be careful to work out the desired contrasts. The results of the specified contrasts follow:

Contrast	DF	Chi-Square	Prob
CA vs EA	1	26.75	0.0000
EA vs PC	1	10.19	0.0014

The tests show that the Eastern region had a different response rate than either the Canadian or Pacific regions. Notice that the relative sizes of the chi-square statistics also agree with the actual data. In the CA vs EA contrast, the comparison is between frequencies of 11 and 60 respectively, and as would be expected, results in a large P^2. On the other hand, the EA vs PC contrast compares frequencies which are closer to each other, i.e.: 60 and 39, respectively, and consequentially results in a smaller P^2 (and hence a larger p-value).

The log-linear form of the model can lead to a large number of parameters. In the example above with GENDER*DEGREE*FIELD, there are 2 x 2 x 6 = 24 categories and therefore 23 estimated parameters. A CONTRAST statement would have to have 23 coefficients. The order of the coefficients can be determined from the response profile printed in the output. A contrast to compare the first FIELD with the second would be written as:

```
CONTRAST 'Fld1 vs Fld2'
         _RESPONSE_  0 0 0 1 -1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0;
```

This ordering results from the way the data is read-in and the output (not shown) indicated that the parameters 4 - 8 coresponded to FIELD. Since all parameters must be accounted for, 21 zeros must also be included. The following result is for the above contrast:

CONTRASTS OF MAXIMUM-LIKELIHOOD ESTIMATES

Contrast	DF	Chi-Square	Prob
Fld1 vs Fld2	1	4.64	0.0313

This indicates that the fields were marginally (p=.03) different in the number of responses.

With the linear model, each factor would be considered separately and would have its own set of parameter estimates, thus making contrast statements a bit more manageable. From the same example, there would be 1 estimate for DEGREE, 5 for FIELD and 5 for DEGREE*FIELD. A contrast on FIELD categories 1 vs 2 would be:

```
CONTRAST 'Fld1 vs Fld2' FIELD 1 -1 0 0 0;   ,
```

which results in the following output:

Contrast	DF	Chi-Square	Prob
Fld1 vs Fld2	1	0.05	0.8193

This contrast is not significant. The result is not conflicting with the previous contrast, however ot reflects the difference between the log-linear and linear models. For the linear model, there is no difference in the way FIELD 1 and FIELD 2 influence GENDER, whereas the log-linear model found a significant difference in the number of responses between FIELD 1 and FIELD 2.