ABSTRACT

Probably the most common use for statistics in the research environment is to fit linear models to data obtained experimentally in the lab. The purpose of these models is usually to determine the extent to which some intrinsic variables affect a response of interest. The null hypothesis being tested is usually: if the intrinsic variables have no effect on the response, the coefficients are zero.

The analysis of variance for this problem is straightforward and well established. However, a special class of problems that often occurs, particularly in the chemical industry, requires special handling by SAS. The special treatment required might not be first be obvious.

Mixture problems involve situations where the sum of the ingredients, X's, totals 100 percent or 1. This constraint may be expressed as:

\[ x_1 + x_2 + x_3 + \ldots + x_n = 1 \]  

(1)

The constraint equation reduces the form of the linear model so that no intercept term is required. However the null hypothesis must also change. The null hypothesis then becomes: if none of the variables have any effect on the response, the coefficients are \textit{equal}.

Although the standard SAS GLM output prints the correct coefficients when the NOINT option is used, the ANOVA must be performed in a separate GLM step to account for the changed null hypothesis.

This paper points out the dangers of misinterpreting GLM's ANOVA for MIXTURE problems and how to properly handle the ANOVA for MIXTURE problems using SAS.

1.0 Statistics in the industrial research environment.

Industrial researchers usually use a limited subset of the seemingly boundless body of material called "statistics". One of the main uses of statistical theory in the research environment is to evaluate the EFFECTS of independently controlled variables on one or several responses of interest.

Examples of such research include altering time and temperature in chemical reactions to affect changes in yield, appearance characteristics, cost or other response variables. Any set of controlling variables may be investigated to determine their EFFECT on the response variable(s).

In the research environment one usually has the luxury of designing the experiment before it is carried out, rather than after the fact which too often occurs. In the EXPERIMENTAL design phase of a project, the researcher determines 3 things. First, what his main objective is. Second, what his suspected independent variables are and lastly, what his dependent variables are.

These decisions usually mandate a particular statistical treatment of the data that is to be collected. A design for the experiments may then be chosen which, with minimal experimental effort, will accomplish the desired objectives of the experimenter on a sound statistical basis.

2.0 Background material for FACTORIAL design in "independent" variables.

This section contains a brief description of standard FACTORIAL design and analysis. The material covered will later be compared to the corresponding portions of MIXTURE design and analysis.

Most often in industrial research the goal is to determine the basics: the statistically significant EFFECTS on the responses, the direction of the EFFECT (positive or negative), and the magnitude of the EFFECT.

One of the most commonly used techniques of investigation of the interrelationship between the identified variables is the FACTORIAL design. FACTORIAL designs allow a simplistic, structured, parsimonious framework for data collection which involves very simple statistical interpretations.

2.1 A typical industrial experimental design, the FACTORIAL design.

The typical model chosen by the researcher, which meets his needs as described above, is the linear model. In the linear model, the response is a linear function of the independent variables.

The linear model for a FACTORIAL design in three independent variables is:

\[ E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \]  

(2)

The usual null hypothesis is:

\[ \beta_0 = \beta_1 = \beta_2 = \beta_3 = 0 \]  

(3)

The model null hypothesis may be interpreted as follows: when the variables have no EFFECT on
the response, their coefficients are zero.

When the null hypothesis is satisfied, the model becomes:

\[ E(Y) = \beta_0 \]  

The least squares estimate of the parameter of the null hypothesis model is, of course,

\[ \beta = \bar{Y} \]  

where \( \bar{Y} \) is the mean of the \( Y \)'s.

The interpretation again is: when the independent variables have no effect on the response, the response is at some average, probably non-zero value.

2.2 The space of the independent variables.

The experimental "space" being explored in two level FACTORIAL designs is restricted by the lower and upper limits of each of the independent variables. With three independent variables, this experimental space may be viewed as a rectangular solid in 3 space, i.e. in the coded levels -1 and +1, the experimental space is the volume contained in a cube whose side is length 2.

Any combination of independent variables within the cube is a possible experimental area of interest to the researcher. Although this is the standard visualization of a three term model, it will be shown that for mixtures this visualization is improper.

In the two level FACTORIAL design, the researcher is aided in the interpretation of the parameters by coding his variables into two values.

The experimental settings for 3 variables at two levels for a full factorial design are shown at the right.

<table>
<thead>
<tr>
<th>X</th>
<th>X</th>
<th>X</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>+1</td>
<td>+1</td>
<td>+1</td>
</tr>
<tr>
<td>+1</td>
<td>+1</td>
<td>-1</td>
</tr>
<tr>
<td>+1</td>
<td>-1</td>
<td>+1</td>
</tr>
<tr>
<td>-1</td>
<td>+1</td>
<td>+1</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>+1</td>
</tr>
</tbody>
</table>

Using this coding scheme has two very desirable consequences: orthogonality of the independent variables and correspondence of regression coefficients to EFFECTS.

2.3 Interpretation of the FACTORIAL design. EFFECTS and the NULL HYPOTHESIS

The researcher has no particular interest in the values of the parameters themselves. He only wants to know the EFFECT of his variables on his response and whether those EFFECTS can be distinguished from the null hypothesis.

The full factorial design consists of all the possible combinations of the two levels for all the variables, hence the name FACTORIAL.

These settings of the independent variables have a very desirable mathematical quality, orthogonality.

This orthogonality means that the regression coefficients are completely independent of each other. This independence makes the model easier to interpret since the coefficients will remain the same no matter what subset of the original linear model is proposed.

For instance, if the third term in our linear model is dropped and the \( b \)'s are determined again, the coefficients for the first two terms will be unaffected.

The other desirable consequence of two level FACTORIAL designs is that the coefficients are directly related to an intuitive understanding of what the EFFECTS are.

The EFFECT of a variable on a response may be defined as: (the average response, at the high value of the variable) minus (the average response at the low value, of that variable). The difference in the averages is then assumed to be caused by the changing of the variable from its low to high and is the EFFECT of that variable.

The relationship between the coefficient and the model is straightforward and easy to demonstrate.

When \( X \) is at its high value, the response function has the following form:

\[ \bar{Y} = 8 + 8(1) + 8X + 8X \]  

\[ \text{high} \quad 0 \quad 1 \quad 2 \quad 3 \]

When \( X \) is at its low value, the response function has the following form:

\[ \bar{Y} = 8 + 8(-1) + 8X + 8X \]  

\[ \text{low} \quad 0 \quad 1 \quad 2 \quad 3 \]

The EFFECT for variable 1 is then:

\( \bar{Y} - \bar{Y} = 8(1) - 8(-1) \)  

\[ \text{high} \quad \text{low} \quad 1 \quad 1 \]

\[ = 28 \]

So one may quickly assess the EFFECT of a variable by glancing at the regression coefficient, multiplying by 2, and obtaining the EFFECT of the variables. Because the coefficients are orthogonal, the EFFECTS are
also independent of one another and will not change even if the regression were run again with one of the other terms dropped from the model.

3.0 The MIXTURE constraint and the linear model

The traditional linear model (2) and mixture constraint equation (1) may be combined to produce the following form:

\[ Y = B_1 X_1 + B_2 X_2 + B_3 X_3 + \text{error} \]

This form of the linear model is referred to as the Scheffe canonical form of the linear model for mixtures. In this form the common "intercept" term has disappeared due to the constraint on the sum of the components of the mixture.

One is immediately tempted to use the SAS GLM option, NOINT, of the MODEL statement to obtain the coefficients of the Scheffe model, the B's. And in fact that will give the correct estimates of the coefficients. However, the null hypothesis would not be the correct one for the mixture model. The proper treatment will be described in section 3.2.

3.1 The space of the mixture variables

Once again the researcher may define an experimental "space". With mixtures, the low value of a component is 0 percent and the high value is 100 percent. For three variables, a triangular plane is defined which contains all the possible combinations of linear blending of the three components.

3.2 Interpretation of the mixture design EFFECTS and the NULL HYPOTHESIS

Again the researcher is interested in the EFFECT that each variable has on the response. The definition of EFFECT is more difficult for the mixture case because as one variable goes from high to low, the proportions of the others also change, a consequence of the constraint equation. When variable X is 0 percent, X can be anywhere between 0 percent and 100 percent and the same is true of X.

For mixture models, the EFFECT of a variable is defined as the change in the response as that variable's value goes from 100 percent to 0 percent while the other variables evenly split the remaining percentage. This amounts to comparing the response at a vertex of the mixture space to the response at the midpoint opposite the vertex.

The high value for X is at:

\[ X_1 = 1 \quad X_2 = 0 \quad X_3 = 0 \] (10)

The low value is at:

\[ X_1 = 0 \quad X_2 = .5 \quad X_3 = .5 \] (11)

Relating coefficients to EFFECTS gives:

\[ Y = B(1) + B(0) + B(0) = B \] (12)

\[ Y = B(0) + B(.5) + B(.5) = .5B + .5B \] (13)

and therefore the EFFECT of variable 1:

\[ \text{Effect of } X_1 = \frac{(Y - Y_0)}{X_1 - X_0} = \frac{.5(B + B)}{1 - 0} \] (14)

The intuitive feeling for the EFFECTS of the variables by looking at the model coefficients is not as good as it was for the traditional "independent" variables model. This fact becomes even more apparent when higher order models are evaluated.

It may be shown that the full "second" order model for mixtures is:

\[ Y = B_1 X_1 + B_2 X_2 + B_3 X_3 + B_{12} X_1 X_2 + B_{13} X_1 X_3 + B_{23} X_2 X_3 \] (15)

This equation is the full "second" order equation even though no squared terms appear.

The full "cubic" model for mixtures is:
Y = B X + B X + B X +
1 2 2 3 3
B X X + B X X + B X X +
1 2 1 3 1 3 2 3
B X X X
1 2 3 1 2 3

The Scheffe model null hypothesis for mixtures is:
B = B = B
1 2 3

The null hypothesis for the mixture model may be interpreted as follows: when the variables have no EFFECT on the response, the coefficients are equal. This results in the proper conclusion: if the variables do not affect the response, then the response will be at some average, probably non-zero value.

4.0 Performing the analysis of mixture constraint data using SAS.

This section will discuss the "wrong" way to do the analysis, since it is the easiest one to be trapped into misusing. By "wrong" it is meant that it does not properly reflect the mixture concepts described above.

Then the correct analysis will be shown.

The data used is shown in TABLE 4-1. 10 experimental conditions were run and a dependent variable called ELAST was measured. The order of the experiments was presumably randomized before the experiments were performed. The sums of the X's is 1 which meets the constraint condition of equation (1).

TABLE 4-2 is the output for the "wrong" way of specifying a mixture model and TABLE 4-3 contains the "right" specification.

Seven items on these two outputs will be compared. They are indicated by the circled numbers. On both of these TABLEs the EFFECT of the variables are calculated as in equation (13) using the ESTIMATE statement.

4.1 What to expect for the ANOVA

There were 10 degrees of freedom in the original data. One degree of freedom should be removed for the proper null hypothesis (the values of ELAST are at some average value), reducing the available degrees of freedom to 9 for the total sums of squares.

There are 3 terms in the model, but one degree of freedom is lost to the constraint equation, so the model should have 2 degrees of freedom.

4.2 The "wrong" way

In TABLE 4-2 PROC GLM used a MODEL statement of the Scheffe canonical form with the NOINT option.

1) Because NOINT was specified, the TOTAL degrees of freedom was 10. This reflects, as will other items, the fact that the null hypothesis of GLM with NOINT is:
Y = 0

The model has 3 degrees of freedom. The constraint has not been taken into account.

2) The SUMS OF SQUARES terms for TOTAL is too large because the average value has not been taken into account.

3) The F value is also inflated because the SUMS OF SQUARES is inflated and subsequently the probability value is also inversely affected. The probability value makes this model appear highly significant.

4) R-SQUARE looks pretty good, however, R-SQUARE is too large because SUMS OF SQUARES column is "rang.

5) The model PARAMETER ESTIMATES are correct for the Scheffe form and may be used in equations (14) and (15).

PROC GLM; MODEL: ELAST= X1 X2 X3(NOINT); ESTIMATE 'X1 EFFECT' X1 = 0 1 1 0 1 0 1 0 1 0; ESTIMATE 'X2 EFFECT' X2 = 0 1 1 0 1 0 1 0 1 0; ESTIMATE 'X3 EFFECT' X3 = 0 1 1 0 1 0 1 0 1 0; TITLE "IMPROPER NULL HYPOTHESIS FOR MODEL ANALYSIS OF VARIANCE:
SUMS IMPROPER NULL HYPOTHESIS FOR MODEL ANALYSIS OF VARIANCE:
GENERAL LINEAR MODELS PROCEDURE

DEPENDE MVABLES: ELAST
SOURCE
MODEL 1 3 5.098310.0967158 1.688403 1.2837055 38.19
ERROR 2 13 2464.35419462600 46099.3705587 5 0.006
UNCORRECTED TOTAL 15 54609.200000000 16848.029055 38.19
R-SQUARE 0.55224 21.452 218.33833333 0.73 0.600000000
STDEV ELAST MEAN 69.600000000
PARAMETER ESTIMATE 455.78665311 1.00 0.0199 153.0276744
X1 4 584.84645531 1.00 0.0053 151.0277644
X2 5 507.86665311 1.00 0.0041 151.0277644
X3 6 207.32000000 1.00 0.1879 219.33196323
X1 EFFECT 7 1.00 0.1791 151.0277644
X2 EFFECT 8 -0.00200000 1.00 0.7951 219.33196323
X3 EFFECT 9 1.00 0.1109 219.33196323

TABLE 4-2
"No Intercept" Model Output

236
6) The EFFECTS may properly be estimated from the Scheffe form.

7) None of the EFFECTS are significantly different from zero at the .05 significance level.

4.3 The "right" way

In TABLE 4-3 PROC GLM used a MODEL statement allowing the intercept term to be included. Although the model is not of the Scheffe form, the intercept term will be able to absorb the sum of squares for the null hypothesis.

ESTIMATE statements are the same regardless of the model.

1) 9 total degrees of freedom agrees with what was expected. The model has only 2 degrees of freedom as it should. This happens as a result of the intercept term and the sum of the variables being linearly related.

2) The SUMS OF SQUARES have properly been corrected for the null hypothesis. Both the TOTAL and the MODEL have been greatly reduced.

3) The F value is much lower now that the null hypothesis has been removed from the TOTAL SUMS OF SQUARES. In fact the model is no longer highly significant.

<table>
<thead>
<tr>
<th>PROC GLM MODEL CASE EFFECT X2 X3</th>
<th>ESTIMATE: XII EFFECT X2 X3 = 1.5 2.5 -3.5 1.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESTIMATE: XII EFFECT X2 X3 = -1.5 1.5 -3.5 1.5</td>
<td></td>
</tr>
<tr>
<td>TOT EESTIMATE: XII EFFECT X2 X3 = -1.5 1.5 -3.5 1.5</td>
<td></td>
</tr>
</tbody>
</table>
| TITLE: NULL HYPOTHESIS FOR MODEL ANALYSIS OF VARIANCE; RUN;
<p>| PROPER NULL HYPOTHESIS FOR MODEL, ANALYSIS OF VARIANCE; GENERAL LINEAR MODELS PROCEDURE |
| DEPENDENT VARIABLE: ELAST |</p>
<table>
<thead>
<tr>
<th>SOURCE</th>
<th>COUNTED TOTAL</th>
<th>1.2194</th>
</tr>
</thead>
<tbody>
<tr>
<td>SUMMARY</td>
<td>SUM OF SQUARES</td>
<td>MEAN SQUARE</td>
</tr>
<tr>
<td>---------</td>
<td>----------------</td>
<td>-------------</td>
</tr>
<tr>
<td>MODEL</td>
<td>51509.99598</td>
<td>5099.99598</td>
</tr>
<tr>
<td>ERROR</td>
<td>90900.00000</td>
<td>9090.00000</td>
</tr>
</tbody>
</table>

4) R-SQUARE is also much lower. For mixture models using the Scheffe form, R-SQUARE is usually much larger than it should be. It will almost always be greater than .9 .

5) The model PARAMETER estimates have changed and are biased. A generalized inverse was employed by GLM to find one solution for the PARAMETERS.

6) The EFFECT estimates have not changed with the form of the model even though the model PARAMETERS have.

7) The probability values also remain invariant.

5.0 CONCLUSIONS

Although the mixture constraint equation reduces the forms of the standard polynomial models applied to experimental designs, the Scheffe form of the model should not be used in GLM for ANOVA of the EFFECTS being investigated by the researcher.

If the Scheffe form of the equation is to be used, such as when trying to display the response function graphically, one GLM step will have to be used to obtain the Scheffe coefficients and another to perform the analysis of variance.

6.0 References


TRAINING AND SUPPORT

CHAIRMAN:

John Boling, SAS Institute