Paper 1521-2015

Sums of Squares: The Basics and a Surprise

Sheila Barron, The University of Iowa
Michelle A. Mengeling, Iowa City Veterans Affairs Health Care System & The University of Iowa

ABSTRACT

Most Design of Experiments textbooks cover Type I, Type II, and Type III sums of squares, but many researchers and statisticians fall into the habit of using one type mindlessly. This Breakout Session will review sum of squares basics, illustrate the importance of choosing the appropriate sum of squares type, and highlight differences in PROC GLM and PROC REG parameterization coding.

INTRODUCTION

This paper began with a mystery. A researcher contacted the Statistics Outreach Center at the University of Iowa with a problem. He had conducted a 2 x 2 factorial analysis of variance (ANOVA) and found significant main effects and a non-significant interaction. He then analyzed the same data using multiple regression expecting to get the same results. However, in the regression, nothing was statistically significant.

I asked a series of questions:

- Were the data balanced? Yes, 28 observations in each of the four cells.
- Were the two independent variables dichotomous? Yes, they were both dummy coded (0 and 1 were the only possible values).
- Was the interaction term formed by multiplying the two independent variables? Yes, for the regression analysis. In ANOVA, the researcher included the code for the interaction in the model statement.

It may be that you know the cause of the difference. Or, it may be that you learned why this would happen in a statistics class once and have since forgotten. In discussing the issue with a number of colleagues, we came to the conclusion that a review of the basics would be helpful for many researchers.

BACKGROUND

It is common to see ANOVA referred to as a special case of regression. In fact, there are numerous texts that would lead the reader to believe that the results will be the same regardless of whether the data is analyzed as an ANOVA or as a regression (e.g., Field & Miles, 2010). This is misleading as indicated by the results of the current analysis.

ANOVA methods were developed in the context of experimental data and as such, commonly involved categorical explanatory variables. They are included in the general linear model employed in GLM procedure, not through their values, but though levelization. This is the process by which a categorical variable is transformed into a set of indicator variables which identify the level of the categorical variable. There is one indicator variable per level of the explanatory variable. For the interaction effect, levelization will create one indicator variable for each combination of levels of the main effects involved in the interaction.

Regression methods were developed in the context of observational studies collecting naturally-occurring data which were often continuous. As such, the process of levelization is not traditionally employed in regression analysis. A researcher wanting to use the REG procedure with a categorical variable may create the indicator variables through a preliminary data step. Such dummy coding is commonly used. The primary difference between dummy coding and levelization is that when a researcher includes dummy variables in a regression analysis, the number of dummy variables included is always one less

than the number of categories represented in the data. However, in levelization the number of indicator variables included is equal to the number of categories included in the data.[1]

To understand the implications of this difference, it is necessary to understand sums of squares.

SUMS OF SQUARES

There are numerous SAS procedures that can report sums of squares. This paper focuses on two: 1) general linear model (PROC GLM); and 2) regression (PROC REG). Both procedures fit under the umbrella of linear model analysis, and in fact, a common example routinely employed by statistics classes illustrates how, under certain circumstances, the two procedures produce identical results.

In factorial ANOVA and least-squares multiple regression, the inferential statistics can be determined by the ratio of effect and error sums of squares, each divided by the appropriate degrees of freedom. Effect sums of squares are a decomposition of the total sum of squared deviations from the overall mean (SST). How the SST is decomposed depends on characteristics of the data as well as the hypotheses of interest to the researcher.

DATA LAYOUT

Data layouts for analysis of variance are categorized as balanced, proportional, non-systematic or chaotic based on the relationship among the cell sample sizes (Driscoll & Borror, 2000). Balanced indicates equal cell sample sizes. Proportional indicates a constant ratio exists among conditional sample sizes (e.g., if in row 1, column 1 has twice as many observations as column 2, the same ratio will apply in the other rows in the design). Non-systematic is used to represent the situation where proportionality does not hold but there is data in all cells. Finally, chaotic applies to the situation where one or more cells in the design are empty. Analysis of chaotic data will not be discussed in this paper.

As there are four types of data layouts, there are four types of sums of squares. We will review Types I, II, and III. Note, Type IV sums of squares are identical to Type III sums of squares except when there are empty cells (chaotic data) and that is beyond the scope of this paper.

TYPE I SUMS OF SQUARES – SEQUENTIAL

Type I sums of squares (SS) are based on a sequential decomposition. For example, if your ANOVA model statement is “MODEL Y = A|B” the sum of squares are considered in effect order A, B, A*B, with each effect adjusted for all preceding effects in the model. The sums of squares for A will be obtained without consideration of the B or A*B effects. The sums of squares for B will be obtained based on a model already containing A, and the sums of squares for the interaction will be obtained based on a model already containing A and B. Thus, any variance that is shared between the various effects will be subsumed by the variable entered earlier.

Factor A	SS(A)
Factor B	SS(B|A)
Interaction	SS(AB|A,B)

[1] If you remember your linear algebra you will know that this results in a design matrix that is singular, and this method of parameterization is sometimes called singular parameterization.

TYPE II SUMS OF SQUARES – ADDED AFTER OTHER MAIN EFFECTS

With Type II SS, each main effect is considered as though it were added after the other main effects but before the interaction. Then any interaction effects are calculated based on a model already containing the main effects. Thus, any variance that is shared between A and B is not considered part of A or B. Thus interaction variance that is shared with A or with B will be counted as part of the main effect, and not as part of the interaction effect.

Factor A	SS(A|B)
Factor B	SS(B|A)
Interaction	SS(AB|A,B)

TYPE III SUMS OF SQUARES – ADDED LAST

Type III SS considers all effects as though they are added last. Thus any shared variance ends up not being counted in any of the effects.

Factor A	SS(A|B, AB)
Factor B	SS(B|A, AB)
Interaction	SS(AB|A,B)

In ANOVA, when the data are balanced (equal cell sizes) the factors are orthogonal and all three types of sums of squares are identical (). Orthogonal, or independent, indicates that there is no variance shared across the various effects, and the separate sums of squares can be added to obtain the model sums of squares.

This is a key point in the mystery of the differing results in PROC GLM and PROC REG. Recall that the design was balanced – the cell sizes were all 28, and in PROC GLM the Type I and Type III sums of squares were identical as one would expect with an orthogonal design.

Now let us present examples from both PROC GLM and PROC REG in order to illustrate the results that sparked this discussion.

ANOVA

With balanced data (equal cell sizes), the results from the ANOVA show that the Type I and Type III sums of squares (SS), printed by default in PROC GLM, are identical. This occurs because all three effects (A, B, and A*B) are orthogonal (independent).

Below is a simplified version of the data from the original study for illustration along with the output (Output 1):

```
data work.one;
input A B Y @@;
  ID = _N_;
  output;

datalines;
0 0 50 0 0 54 0 0 56 0 0 60
0 1 46 0 1 50 0 1 52 0 1 56
1 0 47 1 0 51 1 0 52 1 0 55
1 1 38 1 1 42 1 1 43 1 1 46
;

proc glm;
  class A B;
  model Y=A|B;
run;
```

```
                                The GLM Procedure

                                Class Level Information

                              Class         Levels    Values
                              A                  2    0 1
                              B                  2    0 1

Dependent Variable: Y

                                          Sum of
      Source                      DF         Squares     Mean Square    F Value    Pr > F
      Model                        3     350.2500000     116.7500000       8.27    0.0030
      Error                       12     169.5000000      14.1250000
      Corrected Total             15     519.7500000

                      R-Square     Coeff Var      Root MSE        Y Mean
                      0.673882      7.535487      3.758324      49.87500


      Source                      DF       Type I SS     Mean Square    F Value    Pr > F
      A                            1     156.2500000     156.2500000      11.06    0.0060
      B                            1     169.0000000     169.0000000      11.96    0.0047
      A*B                          1      25.0000000      25.0000000       1.77    0.2081


      Source                      DF     Type III SS     Mean Square    F Value    Pr > F
      A                            1     156.2500000     156.2500000      11.06    0.0060
      B                            1     169.0000000     169.0000000      11.96    0.0047
      A*B                          1      25.0000000      25.0000000       1.77    0.2081
```

Output 1. PROC GLM Output

Another way to consider the statistical comparisons is to use contrasts. A quick review: a contrast is a comparison involving two or more factor level means and is defined by a linear combination of the factor level means μ_I where the coefficients c_i sum to zero ($\sum_{i=1}^{r} c_i = 0$). Of importance to this discussion, contrasts are not necessarily orthogonal (independent) from one another. To ensure orthogonality, the multiplicative sum of the contrast coefficients must be equal to zero. For example, if a_i represents the coefficients for Contrast 'Main A' and b_i represents the coefficients for Contrast 'Main B', then for contrasts 'Main A' and 'Main B' to be orthogonal, the following would be true $\sum_{i=1}^{r} c_i d_i = 0$).

For example, Table 1 shows the orthogonal contrasts for the GLM example:

	A1B1	A1B2	A2B1	A2B2
Contrast 'Main A'	1	1	-1	-1
Contrast 'Main B'	1	-1	1	-1
Contrast 'Interaction'	1	-1	-1	1

Table 1 Orthogonal Contrasts that are equivalent to PROC GLM output

The SAS coded orthogonal contrasts, shown below, produce identical results to those provided by PROC GLM:

```
contrast 'Main A' A 1 -1;
contrast 'Main B' B 1 -1;
contrast 'Interaction' A*B 1 -1 -1 1;
```

MULTIPLE REGRESSION

Using the same data from the PROC GLM example, the only additional code needed for the multiple regression analysis involves creating the interaction term (AxB) in a data step:

```
data work.one;
set work.one;
AxB = A*B;
run;
```

In PROC REG, we need to explicitly request the sums of squares be included in the output. SAS documentation outlines options to request Type I (SS1) and Type II (SS2) sums of squares on the MODEL statement:

```
proc reg;
model Y = A B AxB/ss1 ss2;
run;
```

This was another clue to understanding why the PROC GLM and PROC REG results differed. Type III sums of squares aren't an option in PROC REG. The only difference between Type II and Type III SS is whether or not variance shared with the interaction effect is removed when testing the main effects. The epiphany was that in regression analysis, the interaction term is created before the procedure is run, and it is entered as another variable. Thus, SAS doesn't "know" that the variable AxB is an interaction term—it is treated as though it is a just a third variable in the analysis. Thus in PROC REG, Type II and Type III sums of squares would always be identical, and there is no need to include options for both types.

The next clue came from examining the regression results, including the SS1 and SS2 output (Output 2). The Type I and Type II SS were not the same. The design was balanced—in the original data there were 28 observations per cell (for this example, we have used only 4 observations per cell). But once again, the key was how the interaction was computed for the regression analysis. As is standard practice, we multiplied the A and B variables to obtain the interaction. This meant that when either (or both) of the main effects had a value of 0, AxB equaled 0, and only when both A and B had a value of 1 did AxB have a value of 1.

A look at the contrast weights (see Table 2) and a little math demonstrate that an interaction calculated from 0/1 dummy coded variables is not orthogonal to the two main effects. The sum of the product of the weights for A and B equaled 0 indicating A and B are orthogonal. But the sum of the product of the weights for A and A*B do not equal 0 (nor does the sum of the product of the weights B and AxB), thus A and AxB are not independent (and B and AxB are not independent).

A1B1	A1B2	A2B1	A2B2
1	1	-1	-1
1	-1	1	-1
1	1	1	-3

Table 2. Non-Orthogonal Contrasts that are equivalent to PROC REG output when 0/1 Dummy Coding is Used

The REG Procedure

Model: MODEL1
Dependent Variable: Y

Number of Observations Read 16
Number of Observations Used 16

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	350.25000	116.75000	8.27	0.0030
Error	12	169.50000	14.12500		
Corrected Total	15	519.75000			

Root MSE	3.75832	R-Square	0.6739
Dependent Mean	49.87500	Adj R-Sq	0.5924
Coeff Var	7.53549		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS	Type II SS
Intercept	1	55.00000	1.87916	29.27	<.0001	39800	12100
A	1	-3.75000	2.65754	-1.41	0.1836	156.25000	28.12500
B	1	-4.00000	2.65754	-1.51	0.1581	169.00000	32.00000
AxB	1	-5.00000	3.75832	-1.33	0.2081	25.00000	25.00000

Output 2. PROC REG Output with Dummy Coding

The problem was the use of 0/1 dummy coding. In order to produce results in PROC REG equivalent to the PROC GLM (i.e., orthogonal contrasts), we should have used effect coding (e.g., -1/1), which would create independence among the three factors (i.e., A, B, AxB). Since the design is balanced, -1/1 effect coding is the same as centering the variable (see Table 3). Rerunning the regression analysis with the revised -1/1 effect coding resulted in sums of squares (and p-values) equivalent to the GLM results (Output 3). In addition, the Type I and Type II sums of squares in PROC REG are equal once the variables are centered.

```
data work.two;
set work.one;
if A = 1 then newA = 1;
else          newA =-1;
if B = 1 then newB = 1;
else          newB =-1;
newAxB = newA*newB;
```

A1B1	A1B2	A2B1	A2B2
1	1	-1	-1
1	-1	1	-1
1	-1	-1	1

Table 3. Orthogonal Contrasts that are equivalent to PROC REG output when -1/1 Effect Coding is Used

The REG Procedure

Model: MODEL1
Dependent Variable: Y

Number of Observations Read 16
Number of Observations Used 16

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	350.25000	116.75000	8.27	0.0030
Error	12	169.50000	14.12500		
Corrected Total	15	519.75000			

Root MSE	3.75832	R-Square	0.6739
Dependent Mean	49.87500	Adj R-Sq	0.5924
Coeff Var	7.53549		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS	Type II SS
Intercept	1	49.87500	0.93958	53.08	<.0001	39800	39800
newA	1	-3.12500	0.93958	-3.33	0.0060	156.25000	156.25000
newB	1	-3.25000	0.93958	-3.46	0.0047	169.00000	169.00000
newAxB	1	-1.25000	0.93958	-1.33	0.2081	25.00000	25.00000

Output 3. PROC REG Output with Effect Coding

CONCLUSION

It is true that analysis of variance and multiple regression are at the heart the same type of analysis. However, due to differences in the contexts of their development, there are lingering differences in the conventional usage of these procedures. Therefore, although it is true that you can produce identical output using the two procedures; it is unlikely to happen mindlessly. Even in a simple 2x2 balanced factorial design, we encountered differences where we did not expect to find them. Working through the basics to identify the cause for the difference, helped us develop a deeper appreciation for the role of history and convention in the details of the methodologies we use.

Reviewing the literature on sums of squares, interaction effects, and contrasts, clarified the underlying issue, a difference in the coding of the interaction effect for PROC REG. This paper unpacks why differences were found when a researcher decided to compare PROC GLM and PROC REG results. The choice of coding scheme had no impact in PROC GLM and thus, it had no unintended effects, however, the same was not true for PROC REG. The dummy coding initially used produced non-orthogonal comparisons in PROC REG even though the design was balanced, and thus, the results were not comparable to PROC GLM. This paper provides students, researchers, and SAS programmers with several avenues of investigation (i.e., SSs, contrasts) to understand the potential source of differences between PROC GLM and PROC REG results.

REFERENCES

Driscoll, M. F. and Borror, C. M. 2000. "Sums of Squares and Expected mean Squares in SAS", Qual. Reliab. Engng. Int., 16: 423–433 (http://verde.esalq.usp.br/~jorge/cursos/cesar/Semin%E1rios/SumsSquares_Expected(SQ)_SAS.pdf).

Field A. and Miles J. 2010. Discovering Statistics using SAS. Thousand Oaks, CA. Sage Publications Inc.

Scholer, F. 2012. "ANOVA (and R)" Accessed March 20, 2015. http://goanna.cs.rmit.edu.au/~fscholer/anova.php

ACKNOWLEDGMENTS

Thank you to Scott Asay for initially coming to the Statistics Outreach Center at the University of Iowa with the data and question that led to this paper. It was your curiosity that led us to realize there was a gap in our understanding of the relationship between ANOVA and multiple regression.

RECOMMENDED READING

Usage Note 22910: Why are PROC GLM and PROC REG giving different results?
http://support.sas.com/kb/22/910.html

Parameterization of PROC GLM Models
http://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#statug_glm_sect030.htm

The Four Types of Estimable Functions
https://support.sas.com/documentation/cdl/en/statugestimable/61763/PDF/default/statugestimable.pdf

Parameterizing Models to Test the Hypotheses You Want: Coding Indicator Variables and Modified Continuous Variables http://www2.sas.com/proceedings/sugi30/212-30.pdf

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Sheila Barron
University of Iowa
sheila-barron@uiowa.edu
https://www.linkedin.com/in/sheilabarron