GLM | SAS ANNOTATED OUTPUT

This page shows an example of analysis of variance run through a general linear model (glm) with footnotes explaining the output. The data were collected on 200 high school students, with measurements on various tests, including science, math, reading and social studies. The response variable is writing test score (**write**), from which we explore its relationship with gender (**female**) and academic program (**prog**).

The syntax for the page is provided below. The **class** statement defines which variables are to be treated as categorical variables in the **model** statement. The **model** statement has the main effects of **female** and **prog**, as well as their interaction; the interaction is specified by taking the product of the two main effect terms. The option **ss3** tells SAS we want type 3 sums of squares; an explanation of type 3 sums of squares is provided below.

**proc glm data = "c:\temp\hsb2";**

**class female prog;**

**model write = female prog female\*prog /ss3;**

**run; quit;**

The GLM Procedure

Class Level Information

Class Levels Values

female 2 0 1

prog 3 1 2 3

Number of Observations Read 200

Number of Observations Used 200

Dependent Variable: write

Sum of

Source DF Squares Mean Square F Value Pr > F

Model 5 4630.36091 926.07218 13.56 F

female 1 1261.853291 1261.853291 18.48

### Class Level Information

Class Level Information

Classa Levelsb Valuesc

female 2 0 1

prog 3 1 2 3

Number of Observations Readd 200

Number of Observations Usedd 200

a. **Class** – Underneath are the categorical (factor) variables, which were defined as such in the **class**statement. Had the categorical variables not been defined in the **class** statement and just entered in the **model** statement, the respective variables would be treated as continuous variables, which would be inappropriate.

b.**Levels** – Underneath are the respective number of levels (categories) of the factor variables defined in the **class** statement.

c.**Values** – Underneath are the respective values of the levels for the factor variables defined in the **class**statement.

d.**Number of Observations Read** and **Number of Observations Used** – This is the number of observations read and the number of observation used in the analysis. The **Number of Observations Used** may be less than the **Number of Observations Read** if there are missing values for any variables in the equation. By default, SAS does a listwise deletion of incomplete cases.

### Model Information

Dependent Variablee: write

Sum of

Sourcef DFg Squaresh Mean Squarei F Valuej Pr > Fj

Model 5 4630.36091 926.07218 13.56 k Coeff Varl Root MSEm write Meann

0.258985 15.65866 8.263856 52.77500

Sourceo DFp Type III SSq Mean Squarer F Values Pr > Fs

female 1 1261.853291 1261.853291 18.48

e. **Dependent Variable** – This is the dependent variable in our glm model.

f. **Source** – Underneath are the sources of variation of the dependent variable. There are three parts, Model, Error, and Corrected Total. With glm, you must think in terms of the variation of the response variable (sums of squares), and partitioning this variation. The variation in the response variable, denoted by Corrected Total, can be partitioned into two unique parts. The first partition, Model, is the variance in the response accounted by our model (**female prog female\*prog**). The second source, Error, is the variation not explained by the Model. These two sources, the explained (Model), and unexplained (Error), add up to the Corrected Total, SSCorrected Total= SSModel+ SSError.

The term “Corrected Total” is called such, as compared to “Total”, or more correctly, “Uncorrected Total,” because the “Corrected Total” adjusts the sums of squares to incorporate information on the intercept. Specifically, the Corrected Total is the sum of the squared difference between the response variable and the mean of the response variable, whereas the Uncorrected Total is the sum of the squared values of just the response variable.

g. **DF** – These are the degrees of freedom associated with the respective sources of variance. As with the additive nature of the sums of squares, the degrees of freedom are also additve, DFCorrected Source= DFModel+ DFError. The DFCorrected Totalhas N-1 degrees of freedom, where N is the total sample size. See DF, superscript p, for the calculation of the DF for each individual predictor variable, which add up to DFModel. Hence, DFError=DFCorrected Total– DFModel. The DFModeland DFError define the parameters of the F-distribution used to test F Value, superscript j.

h. **Sum of Squares** – These are the sums of squares that correspond to the three sources of variation. SSModel – The Model sum of squares is the squared difference of the predicted value and the grand mean summed over all observations. Suppose our model did not explain a significant proportion of variance, then the predicted value would be near the grand mean, which would result with a small SSModel, and SSError would nearly be equal to SSCorrected Total. SSError – The Error sum of squares is the squared difference of the observed value from the predicted value summed over all observations. SSCorrected Total – The Corrected Total sum of squares is the squared difference of the observed value from the grand mean summed over all observations.

i. **Mean Square** – These are the Mean Squares (MS) that correspond to the partitions of the total variance. The MS is defined as SS/DF.

j. **F Value** and **Pr > F** – These are the F Value and p-value, respectively, testing the null hypothesis that the Model does not explain the variance of our response variable. F Value is computed as MSModel/ MSError, and under the null hypothesis, F Value follows a central F-distribution with numerator DF = DFModeland denominator DF =DFError. The probability of observing an F Value as large as, or larger, than 13.56 under the null hypothesis is < 0.0001. If we set our alpha level at 0.05, our willingness to accept a Type I error, we’d reject the null hypothesis and conclude that our model explains a statistically significant proportion of the variance.

k. **R-Square** – This is the R-Square value for the model. R-Square defines the proportion of the total variance explained by the Model and is calculated as R-Square = SSModel/SSCorrected Total = 4630.36/17878.88=0.259.

l. **Coeff Var** – This is the Coefficient of Variation (CV). The coefficient of variation is defined as the 100 times root MSE divided by the mean of response variable; CV = 100\*8.26/52.775 = 15.659. The CV is a dimensionless quantity and allows the comparison of the variation of populations.

m. **Root MSE** – This is the root mean square error. It is the square root of the MSErrorand defines the standard deviation of an observation about the predicted value.

n. **write Mean** – This is the grand mean of the response variable.

o. **Source** – Underneath are the variables in the model. Our model has **female**, **prog**, and the interaction of **female** and **prog**. The interaction disallows the effect of, say, **prog,** over the levels of **female** to be additive. Also, our model follows the hierarchical principal, i.e., if an interaction term is in the model (**female\*prog**), the lower order terms (**female** and **prog**) must be included. Further, when there is a significant interaction in the model, the main effects (the lower order terms) are difficult to interpret. If the interaction term is not statistically significant, some would advise dropping the term and rerunning the model with just the main effects, so that the main effects would have an unambiguous meaning. The traditional anova approach would leave the nonsignificant interaction in the model and interpret the main effects in the normal manner. If the interaction term is found statistically significant, one would leave the model as is and evaluate the simple main effects.

p. **DF** – These are the degrees of freedom for the individual predictor variables in the model. From the class level information section, the lower order term DF is given by the number of levels minus one. For example, **female** as two levels, therefore DFfemale = 2-1=1. Also, **prog** has three levels and DFprog = 3-1=2. For the interaction term, DFfemale\*prog = DFprog\* DFfemale = 1\*2 =2. The DF of the predictor variables, along with the DFError, define the parameters of the F-distribution used to test the significance of F Value, superscript s.

q. **Type III SS** – These are the type III sum of squares, which are referred to as partial sum of squares. For a particular variable, say **female**, SSfemale is calculated with respect  to the other variables in the model, **prog**and **female\*prog**. Also, we showed earlier that SSCorrected Total= SSModel+ SSError, and we might expect that SSModel= SSfemale+ SSprog+ SSprog\*female; however, this is generally not the case (this is only true for a balanced design).

r. **Mean Square** – These are the mean squares for the individual predictor variables in the model. They are calculated as SS/DF, and along MSError, they are used to calculate F Value, superscript s.

s. **F Value** and **Pr > F** – These are the F Value and p-value, respectively, testing the null hypothesis that an individual predictor in the model does not explain a significant proportion of the variance, given the other variables are in the model. F Value is computed as MSSource Var/ MSError. Under the null hypothesis, F Value follows a central F-distribution with numerator DF = DFSource Var, where Source Var is the predictor variable of interest, and denominator DF =DFError. Following the point made in Source, superscript o, we focus only on the interaction term. **female\*prog** – This is the F Value and p-value testing the interaction of **female** and **prog** on the response variable, given the other variables are in the model. The probability of observing an F Value, as large as, or larger, than 2.39 under the null hypothesis that there is not an interaction of **female**and **prog**, given the other variables are in the model, is 0.0946. If we set our alpha level at 0.05, the probability of a Type I error, we would fail to reject the null hypothesis that **female** and **prog**do not interact. Based on this finding, some would advise rerunning the model without the interaction term, including only the main effects in the model (and the intercept). This would in turn permit a valid interpretation of the main effects of **female** and **prog**.