PROC Logistic: The CLASS and CONTRAST statements, and the LACKFIT option

PROC Logistic is versatile and fun to use! Here are some options for this procedure that were not covered in the tutorial.

The CLASS Statement

When do you use it?

The following story illustrates an instance when the CLASS statement can be handy. Maria Euglena is interested in answering the following research question: For first year college men, does pledging a fraternity differ by birth order? In particular, do men who were the third or fourth child pledge fraternities at higher rates? Her outcome variable is PLEDGE where 1 represents men who pledged a fraternity in their first year of college. Her main question predictor is called BIRTH and is a categorical variable with a range from 1-9 that denotes where in their series of siblings surveyed men fall. Because Maria Euglena would like to test the impact of BIRTH on the probability of PLEDGE=1 to determine whether certain placements in the birth order are more important than others, she should probably recode BIRTH into a series of dummies: 1stborn, 2ndborn, 3rdborn, etc. Instead, because Maria Euglena is sharp and knows her SAS code short cuts she uses the CLASS statement in her PROC LOGISTIC paragraph as a short cut for creating these dummies. [??]

What does it do?

In fact, the CLASS statement will take any categorical variable and create a series of dummies from it that may be output to the dataset used for the analysis or a new dataset. In the example shown below, these new dummies are not output to a dataset.

ADD SOMETHING ABOUT IT ASSIGNING THE COMPARISON CATEGORY AMG THE SERIES OF DUMMIES.

SAS will create dummy variables for a categorical variable on-the-fly. There are various coding schemes from which to choose. The default coding for all the categorical variables in proc logistic is the effect coding. Here we changed it to dummy coding by using the param = ref option. We can specify the comparison group by using ref = option after the variable name. There are other coding schemes available, such as orthogonal polynomial coding scheme and reference cell coding. We can double check what coding scheme is used and which group is the reference group by looking at the Class Level Information part of the output.

What does it look like?

```
proc logistic data = dataset ;
  class categoricalvar (ref='1') /param = ref;
```

```
  model outcome (event='1') = continuousvar categoricalvar;
run;
```

The CONTRAST Statement

When do you use it?

Maria Euglena has done a bit more research before coming back to her analysis and has decided that what she is really interested in is whether being 1st born is more important to predicting PLEDGE=1 than being 3rd born. SAS provides a simple way of conducting this test called the CONTRAST statement.

What does it do?

In the parameter estimates, we only see the comparison of level 2 vs. 1 and level 3 vs. 1 for variable prog. If we want to compare level 2 vs. level 3, we can use the contrast statement. Usually, contrast is done using less than full rank, reference cell coding as used in proc glm. We chose this type of coding by using param = glm option in the class statement. We also used estimate option at the end of contrast statement to get the estimate of the difference between group 1 and group 2. It is always a good idea to check the Class Level Information to see how the variable is coded so we know that the contrast statement gives us the expected contrast among groups.

What does it look like?

```
proc logistic data = dataset ;
  class categoricalvar /param = glm;
  model outcome (event='1') = continuousvar categoricalvar;
contrast '1 vs 2 of categoricalvar' categoricalvar 1 -1 0 / estimate;
run;
```

The LACKFIT Option

When do you use it?

What does it do?

The Hosmer-Lemeshow test of goodness-of-fit can be performed by using the lackfit option after the model statement. This test divides subjects into deciles based on predicted probabilities, then computes a chi-square from observed and expected frequencies. It tests the null hypothesis that there is no difference between the observed and predicted values of the response variable. Therefore, when the test is not significant, as in this example, we cannot reject the null hypothesis and say that the model fits the data well.

What does it look like?

```
proc logistic data = dataset ;
  class categoricalvar /param = glm;
  model outcome (event='1') = continuousvar categoricalvar/ lackfit;
```

```
contrast '1 vs 2 of categoricalvar' categoricalvar 1 -1 0 / estimate;
run;
```

Adapted From:
http://www.ats.ucla.edu/stat/sas/seminars/sas_logistic/logistic1.htm