**Class Exercise: Parameterization and Multi-Linear**

**Parameterization:**

**1) Determine for LOGISTIC Regression, whether CLASS variable coding makes sense without a reference e.g. Effect (reference=mean) or Reference (reference=last level/value) coding.**

**2) If (1) makes sense, find a way to perform the regression analysis - in SAS or R.**

Answer: For me to use LOGISTIC Regression, it makes sense to use CLASS variable coding with a reference. I checked a paper (web link: <http://www2.sas.com/proceedings/sugi29/194-29.pdf>) and tried to run PROC LOGISTIC with specifying reference and without ref.

For example, I choose BATH to run the PROC LOGISTIC. Because the last level/value for BATH is 2, I choose to use the same value to specify REF=2. However, the result is different.

Here are the codes:

/\*PROC LOGISTIC\*/

ods graphics on;

proc logistic data=census.psam\_h17\_subset1 alpha=.05

plots(only)=(effect oddsratio);

class BATH(ref='2') / param=ref;

model value(event='0') = BATH / clodds=pl;

title 'LOGISTIC MODEL A: PARAM = REF';

run;

quit;

ods graphics on;

proc logistic data=census.psam\_h17\_subset1 alpha=.05

plots(only)=(effect oddsratio);

class BATH;

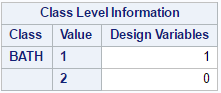
model value(event='0') = BATH / clodds=pl;

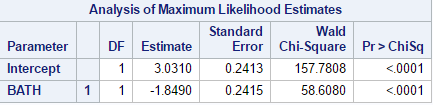
title 'LOGISTIC MODEL B: PARAM = default REF';

run;

quit;

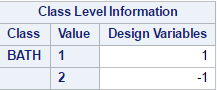
In model A, I use the reference parameterization and specify REF=2. This changes the design matrix and the odds.



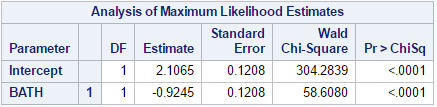


Therefore, the odds of having value = 0 for bath=1 vs. bath=1 is -1.8490.

In Model B, the method of parameterization is not specified, so the default EFFECT parameterization will be used. Also, by default the last ordered category will be used as the reference category. The columns of the design matrix are created based on the EFFECT coding scheme, as follows: For each variable, c-1 columns comprise the design matrix, where c is the number of levels of the classification variable. For the nonreference levels, the columns represent group membership (1=member, 0 =non-member), and for the reference level, the row is populated with a –1. In our Model B, the reference level is the last ordered category, so the design matrix is:



Using EFFECT coding, the beta estimates are estimating the difference in the effect of each nonreference level compared to the average effect over all levels.



Therefore, the odds of having value = 0 for bath=1 vs. bath=1 is -0.9245.

**Multiple Linear Regression:**

**1) Using school data, perform multiple linear regression to find the best fit model. Obtain and compare the adjusted R-squared values.**

**(To create the School dataset and other datasets associated with other sections of the SAS course, reference the SAS codes published. Run  ST200d01.sas to create the dataset)**

|  |  |  |
| --- | --- | --- |
| **response variable** | **predictor variables** | **adjusted R-squared value** |
| fluency2 | phonics2 words2 reading2 | 0.9338 |
| **fluency3** | **phonics3 words3 reading3** | **0.8847** |
| reading3 | letters1 words1 phonics1 phonics2 words2 phonics3 words3 | 0.827 |

Here are the codes:

title 'School Data: Regression and Diagnostics -- fluency2';

proc glmselect data=STAT2.school;

model fluency2=phonics2 words2 reading2 / selection=none;

output out=out r=residuals;

run;

title 'School Data: Regression and Diagnostics -- fluency3';

proc glmselect data=STAT2.school;

model fluency3=phonics3 words3 reading3 / selection=none;

output out=out r=residuals;

run;

title 'School Data: Regression and Diagnostics -- reading3';

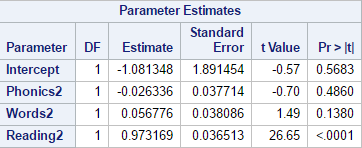
proc glmselect data=STAT2.school;

model reading3=letters1 words1 phonics1 phonics2 words2 phonics3 words3 / selection=none;

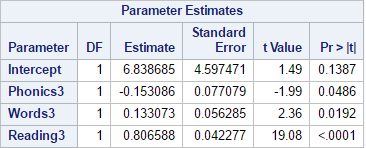
output out=out r=residuals;

run;

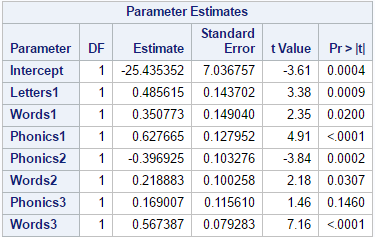
**School Data: Regression and Diagnostics -- fluency2**



**School Data: Regression and Diagnostics -- fluency3**



**School Data: Regression and Diagnostics -- reading3**



Comparing these three models, the second one is more making sense.

**The response variable is fluency3; the predictor variables are phonics3, words3 and reading3; the adjusted R-squared value is 0.8847.**

The estimated regression equation is as follows:

**fluency3 = 6.838685 -0.153086\* phonics3 + 0.133073\*words3 + 0.806588\* reading3**