**Computer input and output to accompany *Statistical Methods for the Social and Behavioural Sciences: A Model-Based Approach* by David B. Flora**

This document © 2017 by David B. Flora and Joo Ann Lee

SAS syntax and output for Chapter 4: Interactions in Multiple Regression

Much of the material in this document builds on the SAS syntax and output for Chapters 1 to 3. If necessary, refer back to those documents for reminders.

Begin by importing the drink.txt data (for the heavy alcohol use example) into R:

|  |
| --- |
| **data** ch4data;  infile 'c:\drink.txt' FIRSTOBS=**2** dlm=' ';  input id coa alcuse ext numalc;  drop id;  **run**; |

Preliminary descriptive statistics (Table 4.1).

Use the 'means' procedure to get descriptive statistics and the 'corr' procedure to get the correlation matrix:

|  |
| --- |
| **proc** **means** data=ch4data maxdec=**2** n mean std median min max range skew kurt;  **run**;  **proc** **corr** data=ch4data;  **run**; |

The variable 'coa' is the dichotomous parental alcoholism variable, with 1 = child of alcoholic, 0 = control.  
'alcuse' is the heavy alcohol use variable.  
'ext' is externalizing behaviour.  
'numalc' indicates the number of alcoholic parents (0, 1, or 2) for each case.

Scatterplots of alcohol use by externalizing, separately for COAs and controls:

It is important that the x- and y-axes have the same scale in both plots. Specifically, 'min' and 'max' after 'xaxis' is set so that the plot displays an x-axis with values from 0 to 3:

|  |
| --- |
| **data** controls;  set ch4data;  if coa=**0**;  **run**;  title "Controls";  **proc** **sgplot** data=controls;  scatter x=ext y=alcuse;  xaxis label="Externalizing" min=**0** max=**3**;  yaxis label="Alcohol use" min=**0** max=**5**;  **run**;  1.png |

|  |
| --- |
| **data** coas;  set ch4data;  if coa=**1**;  **run**;  title "COAs";  **proc** **sgplot** data=coas;  scatter x=ext y=alcuse;  xaxis label="Externalizing" min=**0** max=**3**;  yaxis label="Alcohol use" min=**0** max=**5**;  **run**;  2.png |

**Modeling an interaction with a dichotomous moderator**

Specify and estimate regression of alcohol use on externalizing, parental alcoholism, and their interaction.

We first need to create a new variable that equals the product of two variables with the data step before using that variable to represent the interaction in ‘proc reg’:

|  |
| --- |
| **data** ch4data;  set ch4data;  extcoa= ext\*coa;  **run**;  **proc** **reg** data=ch4data plots=none;  model alcuse = ext coa extcoa / clb;  **run**;  **quit**; |

The output above provides values matching those in Table 4.2.

Plot of alcohol use by externalizing showing implied simple regression lines for COAs and controls (Figure 4.2):

The 'glm' procedure can be used to plot model-implied slopes of a focal predictor across the levels of a moderator. A temporary dataset is created to rename the variables:

|  |
| --- |
| **data** temp;  set ch4data;  Parental\_alcoholism=coa;  Externalizing=ext;  Alcohol\_use=alcuse;  **run**;  **proc** **glm** data=temp;  class Parental\_alcoholism;  model Alcohol\_use = Parental\_alcoholism|Externalizing;  **run**;  **quit**;  3.png |

**Probing an interaction with simple-slope analysis**

Recode parental alcoholism for the purpose of simple-slope analysis:

|  |
| --- |
| **data** ch4data;  set ch4data;  ncoa=abs(coa-**1**);  extncoa=ext\*ncoa;  **run**; |

Now, specify and estimate the interaction model using re-coded COA variable. Results should match those in Table 4.3:

|  |
| --- |
| **proc** **reg** data=ch4data plots=none;  model alcuse = ext ncoa extncoa / clb;  **run**;  **quit**; |

Get Bonferroni-corrected simultaneous confidence interval for simple slope of externalizing among COAs. The confidence level is set to (1- α) = .975, as explained in Chapter 4, using the ‘alpha’ option:

|  |
| --- |
| **proc** **reg** data=ch4data plots=none;  model alcuse = ext ncoa extncoa / clb alpha=**.025**;  **run**;  **quit**; |

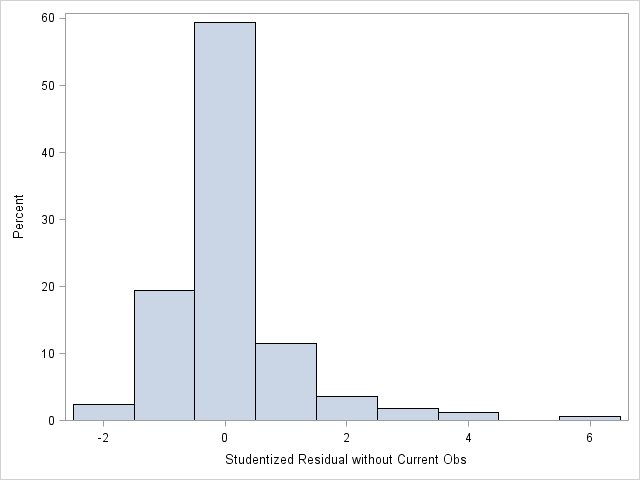
Get Bonferroni-corrected simultaneous confidence interval for simple slope of externalizing among controls:

|  |
| --- |
| **proc** **reg** data=ch4data plots=none;  model alcuse = ext coa extcoa / clb alpha=**.025**;  **run**;  **quit**; |

**Regression diagnostics for the interaction model**

Histogram of Studentized residuals and scatterplot of Studentized residuals against predicted values (Figure 4.3):

|  |
| --- |
| **proc** **reg** data=ch4data plots=none;  model alcuse = ext coa extcoa / clb;  output out=ch4data r=resid h=hatval rstudent=rstudent cookd=cookd p=predict;  **run**;  **quit**;  **proc** **sgplot** data=ch4data;  histogram rstudent;  **run**; |

****

|  |
| --- |
| **proc** **sgplot** data=ch4data;  scatter x=predict y=rstudent;  loess x=predict y=rstudent / lineattrs=(color=blue) smooth=**0.9** nomarkers Name="Lowess" CLM;  **run**;  5.png |

**Modeling an interaction with a three-category moderator**

Create dummy variables for number of alcoholic parents:

|  |
| --- |
| **data** ch4data;  set ch4data;  d1 = **0**;  d2 = **0**;  if numalc = **1** then d1 = **1**;  if numalc = **2** then d2 = **1**;  **run**; |

To follow the hierarchical regression strategy in Chapter 4, first specify the model without interactions (‘mod1’), then specify and estimate the model including interactions with dummy variables (‘mod2’) and test whether the interaction terms significantly increase R2:

|  |
| --- |
| **data** ch4data;  set ch4data;  extd1 = ext\*d1;  extd2 = ext\*d2;  **run**;  **proc** **reg** data=ch4data plots=none;  mod1: model alcuse = ext d1 d2 / clb;  mod2: model alcuse = ext d1 d2 extd1 extd2 / clb;  R2deltaTest: test extd1=**0**, extd2=**0**;  **run**;  **quit**; |

Results above for R2 and significance test for R2 change match those in Chapter 4. Results for 'mod2' match values in Table 4.5.

**Simple-slope analysis with three-category moderator**

Create new dummy variables from 'numalc' so that reference category is 1-alcoholic parent:

|  |
| --- |
| **data** ch4data;  set ch4data;  d12 = **0**;  d22 = **0**;  if numalc = **0** then d12 = **1**;  if numalc = **2** then d22 = **1**;  **run**; |

Specify and estimate the interaction model with new dummy variables.  
Results should match Table 4.6:

|  |
| --- |
| **data** ch4data;  set ch4data;  extd12 = ext\*d12;  extd22 = ext\*d22;  **run**;  **proc** **reg** data=ch4data plots=none;  model alcuse = ext d12 d22 extd12 extd22 / clb;  **run**;  **quit**; |

Create new dummy variables from 'numalc' so that reference category is 2-alcoholic parents:

|  |
| --- |
| **data** ch4data;  set ch4data;  d13 = **0**;  d23 = **0**;  if numalc = **0** then d13 = **1**;  if numalc = **1** then d23 = **1**;  **run**; |

Specify and estimate the interaction model with new dummy variables.  
Results should match Table 4.7:

|  |
| --- |
| **data** ch4data;  set ch4data;  extd13 = ext\*d13;  extd23 = ext\*d23;  **run**;  **proc** **reg** data=ch4data plots=none;  model alcuse = ext d13 d23 extd13 extd23 / clb;  **run**;  **quit**; |

**Modeling an interaction with a continuous moderator**

Import the hasslesNA.txt data (for the psychological symptoms and daily hassles example) into R:

|  |
| --- |
| **data** ch4data2;  infile 'c:\HasslesNA.txt' FIRSTOBS=**2** dlm='09'x;  input hassles support symptoms;  **run**; |

Descriptive stats for hassles example:

|  |
| --- |
| **proc** **corr** data=ch4data2;  **run**; |

|  |
| --- |
| **proc** **means** data=ch4data2 maxdec=**2** n mean std median min max range skew kurt;  **run**; |

The scatterplot matrix in Figure 4.4 can be created using the 'sgscatter' procedure. Most of its options are the same as those in the 'sgplot' procedure. The 'diagonal' option is used to indicate the type of univariate plot for the diagonal:

|  |
| --- |
| **proc** **sgscatter** data=ch4data2;  matrix hassles support symptoms / diagonal=(histogram);  **run**;  **6.png** |

Specify and estimate the interaction model:

|  |
| --- |
| **data** ch4data2;  set ch4data2;  hasslesSupport = hassles\*support;  **run**;  **proc** **reg** data=ch4data2 plots=none;  model symptoms = hassles support hasslesSupport / clb;  **run**;  **quit**; |

Results above match values in Table 4.9.

**Probe the interaction with a continuous moderator**

First, center 'support' at a low value (e.g., 25th percentile = 25 for 'support'):

|  |
| --- |
| **data** ch4data2;  set ch4data2;  suppL=support-**25**;  **run**; |

Then estimate the regression model again:

|  |
| --- |
| **data** ch4data2;  set ch4data2;  hasslesSuppL = hassles\*suppL;  **run**;  **proc** **reg** data=ch4data2 plots=none;  model symptoms = hassles suppL hasslesSuppL / clb;  **run**;  **quit**; |

Results above should match values in Table 4.10.

Next, center 'support' at a medium value (e.g., 50th percentile = 31 for 'support'):

|  |
| --- |
| **data** ch4data2;  set ch4data2;  suppM=support-**31**;  **run**; |

Then estimate the regression model again:

|  |
| --- |
| **data** ch4data2;  set ch4data2;  hasslesSuppM = hassles\*suppM;  **run**;  **proc** **reg** data=ch4data2 plots=none;  model symptoms = hassles suppM hasslesSuppM / clb;  **run**;  **quit**; |

Results above should match values in Table 4.11.

Finally, center 'support' at a high value (e.g., 75th percentile = 34 for 'support'):

|  |
| --- |
| **data** ch4data2;  set ch4data2;  suppH=support-**34**;  **run**; |

Then estimate the regression model again:

|  |
| --- |
| **data** ch4data2;  set ch4data2;  hasslesSuppH = hassles\*suppH;  **run**;  **proc** **reg** data=ch4data2 plots=none;  model symptoms = hassles suppH hasslesSuppH / clb;  **run**;  **quit**; |

Results above should match values in Table 4.12.

But, as reported in Chapter 4, there is a case with a large value for Cook's distance:

|  |
| --- |
| **proc** **reg** data=ch4data2 plots=none;  model symptoms = hassles support hasslesSupport / clb;  output out=ch4data2 r=resid h=hatval rstudent=rstudent cookd=cookd p=predict;  **run**;  **quit**;  **data** ch4data2;  set ch4data2;  id=\_n\_;  **run**;  **proc** **sgplot** data=ch4data2;  scatter x=id y=cookd;  **run**;  7.png |

So we need to figure out which case that is. Here is one approach:

Sort the data by Cook's distance in descending order:

|  |
| --- |
| **proc** **sort** data=ch4data2;  by descending cookd;  **run**; |

The next command prints the first 5 observations in the data set:

|  |
| --- |
| **proc** **print** data=ch4data2(obs=**5**);  **run**; |

In the output above, we can see that the first case is the one with Cook's distance = 1.05. This case has an extremely high value for the 'symptoms' outcome.

The syntax below creates a new data set from which the outlying case has been removed (i.e., only cases with a value of 'symptoms' less than 177 are retained):

|  |
| --- |
| **data** ch4data2;  set ch4data2;  if symptoms < **177** then output;  **run**; |

Now re-estimate the model using this new data set:

|  |
| --- |
| **proc** **reg** data=ch4data2 plots=none;  model symptoms = hassles support hasslesSupport / clb;  **run**;  **quit**; |

In the output above, the interaction is no longer significant, consistent with the results presented in Chapter 4 for the model with the influential case removed.