** this file provides example SAS code to analyze simulated data that was generated to mimic data arising from the AUTISM SMART study (PI: Connie Kasari). the simulated data set file is "AUTISM_Simulated_Data.sas7bdat". an accompanying hand out "Autism Handout.pdf" describes the variables in the data.;

** the practicum exercise is to fill in all of the ???'s;

***************
*** READ IN DATA ***
***************

* create a SAS library that points to the folder that contains AUTISM_Simulated_Data.sas7bdat;
libname libdat  "C:\Users\dalmiral\Dropbox\workshop_SummerInstitute_2014\workshop preparation instructors\workshop preparation instructions for participants for the *

data autismdat;
    set libdat.autism_simulated_data;
    s = 0; if a1=1 and r=0 then s=1;
    label a1 = "initial option: A1=-1=JASP+EMT+SGD; A1=1=JASP+EMT"
    a2 = "second option: A2=-1=Add SGD; A2=1=INTSFY"
    r = "R=0=early non-response; R=1=early response"
    s = "S=1=re-randomized; S=0=not re-randomized"
    o11 = "# of spontaneous words spoken at baseline: hi is good"
    o12 = "# of unintelligible utterances at baseline: low is good"
    o21 = "# of communicative functions during initial txt: hi is good"
    o22 = "# of spontaneous words during initial treatment: hi is good"
    y = "# of spontaneous words 6 months post: hi is good"
    id = "study identification #"
    run;

** brief description of the data;
proc means data=autismdat; run;
** take a look at a frequency table of the initial treatment, early response by week 8, and second stage treatments ;
proc sort data=autismdat; by a1 a2 r s;
proc freq data=autismdat;
    table a1;
    table r;
    table a1*r;
    table a2;
run;

* center baseline covariates at the means. this will facilitate interpretation of the regression parameters. ;
proc means data=autismdat; run;
data dat1;
    set autismdat;
    o11c = o11 - 33.501626;
$o_{12c} = o_{12} - 17.013570$

$o_{21c} = o_{21} - 5.5849851$

$o_{22c} = o_{22} - 48.368327$

run;

proc means data=dat1;
  var o11c o12c o21c o22c;
run;

********************
*** MODULE 4 ***
********************

** primary data ;
** analyses     ;
** part I       ;

********************
*** PI(a) ***
********************

*********************************************************
** code to examine the main effect of initial treatment. ;
*********************************************************

* regression to compare first-line treatments on end of study outcome $Y = \#$ of spontaneous words 6 months

proc genmod data = dat1;
  model $y = a_1 \, o_{11c} \, o_{12c}$;
  estimate 'Mean $Y$ under JASP+EMT' intercept 1 $a_1$ 1;
  estimate 'Mean $Y$ under JASP+EMT+SGD' intercept 1 $a_1$ -1;
  estimate 'Between groups diff' $a_1$ 2;
run;

** the following implements a t-test which is an alternative to the regression model; 
* first, assign labels;

data dat2;
  set dat1;
  if $a_1 = 1$ then $a1tmp =$"noSGD";
  if $a_1 = -1$ then $a1tmp =$"yesSGD";
run;

* run the ttest;

proc ttest data=dat2;
  class a1tmp; var $y$;
run;

* the analysis above is for illustrative purposes only. 
* in an actual SMART, there will be longitudinal measures for $y$, 
* therefore, the analysis above can be carried out using a longitudinal 
* data analysis method such as a growth curve model (linear mixed model).;

**********************************************************************************
** code to examine impact of JASP+EMT and JASP+EMT+SGD on short-term responses rate.             
* this is not an useful analysis in terms of building an AI. 
* because this outcome does not incorporate the effects of future treatment decisions. 
* nonetheless, is still interesting in its own right 
* and we analyze it here for completeness 

proc freq data=dat1;
* again for simplicity, here we used a simple frequency table with a chi-squared test ;
* to compare the proportions responding to initial treatment.
* in practice, we could also use a logistic regression to adjust for initial covariates;

** code to compare the best second-line tactic among non-responders to EMT.

** this code allows investigators to examine whether, among non-responders to JASP+EMT, 
** it is better to INTENSIFY JASP+EMT (A2=1) or add the device JASP+EMT+SGD (A2=-1).

* this analysis is only among those who were re-randomized,
* which were those initially on JASP+EMT who did not respond.

* run the regression;
proc genmod data = dat1;
model y = a2 ;
estimate 'Mean Y: INTENSIFY JASP+EMT ' intercept 1 a2 1 ;
estimate 'Mean Y: Add SGD' intercept 1 a2 -1 ;
estimate 'Between groups difference' a2 2 ;
run;

* the analysis above is for illustrative purposes only.
* in an actual SMART, there will be longitudinal measures for y,
* therefore, the analysis above can be carried out using a longitudinal 
* data analysis method such as a growth curve model (linear mixed model).;

** code to estimate the mean outcome had the entire population followed
** (JASP+EMT, INTENSFY) AI;

* create indicator for the (JASP+EMT, INTENSFY) AI;
* and assign weights;
data dat5; set dat2;
  * create indicator Z1 = -1/1 = didn't/did follow (JASP+EMT, INTENSFY);
  Z1=-1;
  if A1=1 and R=1 then Z1=1;
  if A1=1 and R=0 and A2=1 then Z1=1;
  * an easier way is just "if A2=1 then Z1=1 because only those with A1=1 had A2"

  * define weights;
  W=2;
  if A1=1 and R=0 then W=4;
  * an easier way is just "if S=1 then W=4"
run;
run the W-weighted regression:
proc genmod data = dat5;
class id;
model y = z1 ;
scwgt w;
repeated subject = id / type = ind; * this lines requests robust standard errors;
estimate 'Mean Y under (JASP+EMT, INTENSFY) AI' intercept 1 z1 1;
run;

* an alternative is to just restrict the analysis to the Z1=1 kids;
* ie, those following (JASP+EMT, INTENSFY), as defined above;
* and then obtain a weighted mean just for them;
data dat5tmp; set dat5; if Z1=1; run;
proc genmod data = dat5tmp;
class id;
model y = ;
scwgt w;
repeated subject = id / type = ind; * this lines requests robust standard errors;
run;
* we get identical estimates and standard errors!;

*************************************************************************************;
** code to estimate the mean outcome had the entire population followed
** (JASP+EMT, Add SGD) AI;

* create indicator and assign weights;
data dat6; set dat2;
   * create indicator Z2 = -1/1 = didn't/did follow (JASP+EMT, Add SGD) AI;
     Z2=-1;
     if A1=1 and R=1           then Z2=1;
     if A1=1 and R=0 and A2=-1 then Z2=1;
   * define weights;
     W=2;
     if A1=1 and R=0 then W=4;
run;
* run the W-weighted regression;
proc genmod data = dat6;
class id;
model y = z2 ;
scwgt w;
repeated subject = id / type = ind; * this lines requests robust standard errors;
estimate 'Mean Y under (JASP+EMT, Add SGD) AI' intercept 1 z2 1;
run;

** Estimating the mean outcome under the (JASP+EMT+SGD,INTENSFY) is trivial ;
** Indeed, it is the mean outcome for all who started on A1=-1 ;
** No need to use weights in this case, because both responders ;
** and non-responders are equally represented in the sample mean ;
proc genmod data = dat6;
model y = ;
   where A1= -1 /*restrict to participant who got SGD initially*/;
run;

****************;
** code to compare the mean under the 2 AIs beginning with different treatments. 
** let's compare (JASP+EMT, Add SGD) vs (JASP+EMT+SGD, INTENSFY).
* first, create indicator variables and define the weights;
data dat7; set dat2;
  * create indicator Z2 = -1/1 = didn't/did follow (JASP+EMT, Add SGD) AI;
  Z2=-1;
  if A1=1 and R=1 then Z2=1;
  if A1=1 and R=0 and A2=-1 then Z2=1;
* second, create indicator Z3 to identify those following (JASP+EMT+SGD, INTENSFY);
  Z3=-1;
  if A1=-1 then Z3=1;
* define weights;
  W=2;
  if A1=1 and R=0 then W=4;
run;
** next, limit the data set only to those following these two AIs;
data dat8; set dat7; if Z2=1 or Z3=1; run;
** finally, conduct a regression analysis to compare mean outcomes under the two AIs;
proc genmod data = dat8;
  class id;
  model y = z2;
  scwgt w;
  repeated subject = id / type = ind;
  estimate 'Mean Y under (JASP+EMT, Add SGD) AI' intercept 1 z2 1;
  estimate 'Mean Y under (JASP+EMT+SGD, INTENSFY) AI' intercept 1 z2 -1;
  estimate '     Diff:           ' z2 2;
run;

** code to estimate and compare the mean outcome under all 3 of the embedded AIs
** using 1 regression analysis.
** this analysis differs from the above in that it requires both weighting and replication. see the slides for intuition concerning replication.
* first, replicate responders and define weights;
data dat9; set dat2;
  * only responders to EMT get replicated;
  * because they are the only ones consistent with more than 1 AI;
  if R=1 and A1=1 then do;
ob = 1; A2 = -1; w = 2; output;       * replication happens here;
ob = 2; A2 = 1; w = 2; output;       end;
else if A1=1 and R=0 then do;       * replication does not happen here;
ob = 1; w = 4; output;              end;
else if A1=-1 then do;              * replication does not happen here;
ob = 1; w = 2; output;              end;
run;
data dat9;
set dat9;
* create indicators to differentiate
*    (JASP+EMT,    INTENSFY) vs
*    (JASP+EMT,    Add SGD) vs
*    (JASP+EMT+SGD,INTENSFY)    ;
* A1 helps differentiate A1= 1=[(JASP+EMT,INTENSFY) or (JASP+EMT, Add SGD)]
*                   from A1=-1=(AAC,AAC) ;
* but within A1=1 we need to be able to differentiate between the first two ;
A1A2=A1*A2;
* but A2 is missing for people following (AAC,AAC) strategy, so we reset missing to zero;
if A1A2=. then A1A2=0;
* the above code made it so that:
* (JASP+EMT,INTENSFY) is A1= 1 and A1A2= 1 ;
* (JASP+EMT,Add SGD) is A1= 1 and A1A2=-1 ;
* (JASP+EMT+SGD,...) is A1=-1 and A1A2= 0 ;
run;
proc genmod data = dat9;
class id;
model y = a1 a1a2 ;
sccwgt w;
repeated subject = id / type = ind;
* these statements request the mean under each AI;
estimate 'Mean Y under (JASP+EMT,INTENSFY) AI'  int 1 a1 1 a1a2  1 ;
estimate 'Mean Y under (JASP+EMT, Add SGD) AI'  int 1 a1 1 a1a2 -1 ;
estimate 'Mean Y under (JASP+EMT+SGD,...) AI'       int 1 a1 -1 a1a2  0 ;
* these statements are to get all pairwise comparisons;
estimate 'Diff: (JASP+EMT,Add SGD) - (JASP+EMT+SGD,...)' int 0 a1 2 a1a2  -1 ;
run;
** a statistical advantage about estimating the means under all 3 AI      ;
** simultaneously is that we can increase statistical efficiency       ;
** (lower standard errors) in the estimation of the differences in means. ;
** we can do this by adjusting for baseline covariates that might      ;
** explain variability in Y. however, we must be careful not to adjust for ;
** post-baseline/time-varying covariates or intermediate outcomes.     ;
proc genmod data = dat9;
class id;
model y = a1 a1a2               o11c;
sccwgt w;
repeated subject = id / type = ind;
* these statements request the mean under each AI;
estimate 'Mean Y under (JASP+EMT,INTENSFY) AI'  int 1 a1 1 a1a2  1 ;
estimate 'Mean Y under (JASP+EMT, Add SGD) AI'  int 1 a1 1 a1a2 -1 ;
estimate 'Mean Y under (JASP+EMT+SGD,...) AI'       int 1 a1 -1 a1a2  0 ;
* these statements are to get all pairwise comparisons;
**as you can see, after we adjust for baseline covariates that are predictive**
**we increase the power and now the CI for (EMT,EMT)-(JASP+EMT, JASP+EMT+SGD) is much tighter.**

** Secondary **
** data analyses; **
** Q-learning **

** Secondary data analyses;**
** Q-learning**

** the first step in q-learning is to examine intermediate moderators of the effect of intensify vs switch **
* so we only use non-responders to EMT answer this question;*
* that is, we limit analysis to s=1.  

* first, we fit a full model;*
proc genmod data = dat10;
  model y = o11c o12c o21 o22    a2 a2*o21 a2*o22;
run;
* only o21 appears to moderate the effect   *
* second, fit a reduced model and make appropriate comparisons;*
proc genmod data = dat10;
  model y = o11c  o21     a2 a2*o21;
* the effect at a higher level of utterances;*
estimate 'INTENSIFY, o21=5'     intercept 1 o11c 0 o21 5 a2  1 a2*o21  5;
estimate 'ADD   , o21=5'        intercept 1 o11c 0 o21 5 a2 -1 a2*o21 -5;
estimate 'INTSFY-ADD, o21=5'    intercept 0 o11c 0 o21 0 a2  2 a2*o21 10;
* the effect at a lower level of utterances;*
estimate 'INTENSIFY, o21=2'     intercept 1 o11c 0 o21 2 a2  1 a2*o21  2;
estimate 'ADD   , o21=2'        intercept 1 o11c 0 o21 2 a2 -1 a2*o21 -2;
estimate 'INTSFY-ADD, o21=2'    intercept 0 o11c 0 o21 0 a2  2 a2*o21  4;
run;
* we find that among non-responders to JASP+EMT at the end of 3 months:             
* at higher level of communicative functions=O21, it is better to A2=INTENSIFY.;
* at lower level of communicative functions=O21, it is better to A2=ADD.   ;

* how do we know where the cut-off point is on O21?*
* class discussion;*

** the code above helps understand how to use intermediate outcomes to make stage 2 decisions about intensifying vs augmenting. the above is the first **
step in the Q-learning algorithm to develop a more richly-individualized adaptive intervention. The second step is, assuming that future treatment decisions are optimal, how do we use baseline covariates to make stage 1 decisions about medication vs behavioral modification. The code below implements the second step. It relies on the model used above for the stage 2 decision. The code proceeds in 3 steps and fits these 2 regressions:

\[
\text{stage 2: } y = b_7 + b_8 o_{11c} + b_9 o_{21} + b_{10} a_{1} o_{21} + b_{11} a_{2} + e
\]

\[
\text{stage 1: } y_{opt2} = b_1 + b_2 o_{11} + b_3 o_{12} + b_4 a_{1} o_{11} + b_5 a_{1} o_{12} + b_6 a_1 + e
\]

\(y_{opt2}\) is the \(y\) adjusted for having followed the best stage 2 text.

```latex
proc qlearn data=dat2 deriveci;
    tailor1 o11 o12;
    main2 o11c;
    tailor2 o21;
    response y;
    stg2sample s;
    stg1trt a1;
    stg2trt a2;
run;
```

* Based on the results above, we remove \(o_{12} o_{12} a_1\)

* The reduced stage 1 model is: \(y_{opt2} = b_1 + b_2 o_{11} + b_3 a_{1} o_{11} + b_4 a_1 + e\)

* Before fitting it, however, we first specify contrasts of interest

```latex
data contrasts1;
    input C1 C2 C3 C4; *no. of matrix columns = no. of stage 1 parameters = 4;
    datalines;
    1 10 10 1
    1 10 -10 -1
    0 0 20 2
    1 70 70 1
    1 70 -70 -1
    0 0 140 2
;
* Taking into account optimal future decisions (to intensify vs augment) *
* Contrast 1 = mean future-optimal outcome under JASP+EMT for children w/o11 = 10 = low;
* Contrast 2 = mean future-optimal outcome under JASP+EMT+SGD for children w/o11 = 10 = low;
* Contrast 3 = mean difference no SGD - SGD for children w/o11 = 10 = low;
* Contrast 4 = mean future-optimal outcome under JASP+EMT for children w/o11 = 70 = hi;
* Contrast 5 = mean future-optimal outcome under JASP+EMT+SGD for children w/o11 = 70 = hi;
* Contrast 6 = mean difference no SGD - SGD for children w/o11 = 70 = hi;
run;
```

* Finally, we fit q-learning

```latex
proc qlearn data=dat2 contrast1=contrasts1 deriveci;
    tailor1 o11;
    main2 o11c;
    tailor2 o21;
    response y;
    stg2sample s;
    stg1trt a1;
    stg2trt a2;
run;
```

* How do we know where the cut-off point is on O11?

* Class discussion;
the following code is used to estimate the mean outcome under the more richly-developed adaptive treatment strategy identified using q-learning.

** that is, given the optimal AI identified using q-learning, this code is to obtain the estimated mean outcome had the entire population followed this optimal AI (the one found by qlearning).

data dat11;
  set dat2;
  yopt = 51.5675 + 0.2872*o11 + abs( 0.2309*o11*a1 + -5.3776*a1 );
run;
proc means data=dat11 mean ;
  var yopt;
run;
** the mean of yopt is the estimated mean under the optimal AI found via q-learning ;
** in the future, PROC QLEARN will automatically report this value with an estimated CI;
** the estimated mean is 64.6051 words under the optimal AI found under Q-learning ;
** note this is larger than the Mean Y under best the "design-embedded" AI (JASP+EMT, JASP+EMT+SGD), ;
** which was 62.2434;

*** eof;