**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. This SAS code was used in the MethodsWork workshop on adaptive interventions in Chicago, June 11-12, 2012**

*** READ IN DATA ***

* create a SAS library that points to the folder that contains AUTISM_Simulated_Data.sas7bdat; libname libdat 'C:\Users\dalmiral\Desktop\workshop_MethodsWorks_chicago_13june2012\slides_25april20...'

data autismdat;
    set libdat.autism_simulated_data;
    s = 0; if a1=1 and r=0 then s=1;
    label a1 = "initial txt: A1=-1=AAC; A1=1=EMT"
    a2 = "second txt: A2=-1=SWITCH; 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;
** code to examine the main effect of initial treatment.
** regression to compare first-line treatments on end of study outcome Y = school performance;
proc genmod data = dat1;
  model y = a1 o11c o12c;
  estimate 'Mean Y under EMT'  intercept 1 a1 1;
  estimate 'Mean Y under AAC'  intercept 1 a1 -1;
  estimate 'Between groups diff          ' a1 2;
run;

** the following implements a t-test which is an alternative to the regression model;
* first, assign labels;
data dat2;
  set dat1;
  if a1=1 then a1tmp="EMT";
  if a1=-1 then a1tmp="AAC";
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 EMT and AAC on short-term responses rate.
* this is not an useful analysis in terms of building an adaptive txt strategy.
* 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=dat2;
  table a1*r / chisq nocol nopercent ;
run;
* 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.

** this code allows investigators to examine whether, among non-responders to EMT 
** txt, it is better to INTENSIFY EMT treatment (A2=1) vs SWITCH TO AAC txt (A2=-1).

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

```plaintext
* code to compare the best second-line tactic.
** this code allows investigators to examine whether, among non-responders to EMT txt, it is better to INTENSIFY EMT treatment (A2=1) vs SWITCH TO AAC txt (A2=-1).
** this analysis is only among those who were re-randomized,;
* which were those initially on EMT who did not respond. 

** code to estimate the mean outcome had the entire population followed (EMT,EMT) ATS;

* create indicator for the (EMT,EMT) ATS;
* assign weights;

```plaintext
* code to compare the best second-line tactic.
** this code allows investigators to examine whether, among non-responders to EMT txt, it is better to INTENSIFY EMT treatment (A2=1) vs SWITCH TO AAC txt (A2=-1).
** this analysis is only among those who were re-randomized,;
* which were those initially on EMT who did not respond. 

```plaintext
* code to estimate the mean outcome had the entire population followed (EMT,EMT) ATS;

* create indicator for the (EMT,EMT) ATS;
* assign weights;
```
* an alternative is to just restrict the analysis to the Z1=1 kids;
* ie, those following (EMT, EMT), 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 (EMT,AAC) ATS;
* create indicator and assign weights;
data dat6; set dat2;
  * create indicator Z2 = -1/1 = didn't/did follow (EMT,AAC) ATS;
  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 (EMT,AAC) ATS' intercept 1 z2 1;
run;
** estimating the mean outcome under the (AAC,AAC) is trivial ;
** indeed, it is the mean outcome for all who started on A1=-1;

******************************
*** MODULE 5 ***
******************************
** primary data ;
** analyses     ;
** part II      ;

************************************************************************************;
** code to compare the mean under the 2 ATSs beginning with different treatments. ;
** let's compare (EMT,AAC) vs (AAC,AAC). ;
*
* first, create indicator variables and define the weights;
data dat7; set dat2;
* create indicator $Z_2 = -1/1 = \text{didn't/did follow (EMT,AAC) ATS}$;
$Z_2=-1$;
if $A_1=1$ and $R=1$ then $Z_2=1$;
if $A_1=1$ and $R=0$ and $A_2=-1$ then $Z_2=1$;
* second, create indicator $Z_3$ to identify those following (AAC,AAC);
$Z_3=-1$;
if $A_1=-1$ then $Z_3=1$;
* define weights;
$W=2$;
if $A_1=1$ and $R=0$ then $W=4$;
run;

** next, limit the data set only to those following these two ATSs;**
data dat8; set dat7; if $Z_2=1$ or $Z_3=1$; run;

** finally, conduct a regression analysis to compare mean outcomes under the two ATSs;**
proc genmod data = dat8;
class id;
model y = $z_2$;
scwgt w;
repeated subject = id / type = ind;
estimate 'Mean Y under (EMT,AAC) ATS' intercept 1 $z_2$ 1;
estimate 'Mean Y under (AAC,AAC) ATS' intercept 1 $z_2$ -1;
estimate '     Diff: '                            $z_2$ 2;
run;

*******************************;
*** PII(b) ***;
*******************************;

*******************************************************************************;
** code to estimate and compare the mean outcome under all 3 of the embedded ATSs ;
** 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 ATS;
if $R=1$ and $A_1=1$ then do;
ob = 1; $A_2 = -1$; $w = 2$; output;   * replication happens here;
ob = 2; $A_2 = 1$; $w = 2$; output;
end;
else if $R=0$ and $A_1=1$ then do;
ob = 1; $w = 4$; output;
end;
else if $A_1=-1$ then do;
ob = 1; $w = 2$; output;            * replication does not happen here;
end;
run;
data dat9;
set dat9;
* create indicators to differentiate (EMT,EMT) vs (EMT,AAC) vs (AAC,AAC)    ;
* $A_1$ helps differentiate $A_1=1=[(EMT,EMT) or (EMT,AAC)]$ from $A_1=-1=(AAC,AAC)$ ;
* but within $A_1=1$ we need to be able to differentiate between the first two ;
$A_1A_2=A_1*A_2$;
* but A2 is missing for people following (AAC,AAC) strategy, so...
  if A1A2=0 then A1A2=0;
* the above code made it so that:
* (EMT,EMT) is A1= 1 and A1A2= 1 ;
* (EMT,AAC) is A1= 1 and A1A2=-1 ;
* (AAC,AAC) is A1=-1 and A1A2= 0 ;
run;
proc genmod data = dat9;
  class id;
  model y = a1 a1a2 ;
  scwgt w;
  repeated subject = id / type = ind;
  * these statements request the mean under each ATS;
  estimate 'Mean Y under (EMT,EMT) ATS' int 1 a1 1 a1a2 1 ;
  estimate 'Mean Y under (EMT,AAC) ATS' int 1 a1 1 a1a2 -1 ;
  estimate 'Mean Y under (AAC,AAC) ATS' int 1 a1 -1 a1a2 0 ;
  * these statements are to get all pairwise comparisons;
  estimate '    Diff: (EMT,EMT) - (EMT,AAC)' int 0 a1 0 a1a2 2 ;
  estimate '    Diff: (EMT,EMT) - (AAC,AAC)' int 0 a1 2 a1a2 1 ;
  estimate '    Diff: (EMT,AAC) - (AAC,AAC)' int 0 a1 2 a1a2 -1 ;
run;
** a statistical advantage about estimating the means under all 3 ATS ;
** 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;
  scwgt w;
  repeated subject = id / type = ind;
  * these statements request the mean under each ATS;
  estimate 'Mean Y under (EMT,EMT) ATS' int 1 a1 1 a1a2 1 ;
  estimate 'Mean Y under (EMT,AAC) ATS' int 1 a1 1 a1a2 -1 ;
  estimate 'Mean Y under (AAC,AAC) ATS' int 1 a1 -1 a1a2 0 ;
  * these statements are to get all pairwise comparisons;
  estimate '    Diff: (EMT,EMT) - (EMT,AAC)' int 0 a1 0 a1a2 2 ;
  estimate '    Diff: (EMT,EMT) - (AAC,AAC)' int 0 a1 2 a1a2 1 ;
  estimate '    Diff: (EMT,AAC) - (AAC,AAC)' int 0 a1 2 a1a2 -1 ;
run;
** as you can see, after we adjust for baseline covariates that are predictive ;
** we increase the power and now the CI for (EMT,EMT)-(EMT,AAC) is much tighter.;

***************;
*** MODULE 6 ***;
***************;
** Secondary ;
** data analyses;
** Q-learning **;

***************;
**** S(a) ****;
***************;
** 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.

data dat10;
  set dat1; if s=1;
run;

* 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 'SWITCH , o21=5' intercept 1 o11c 0 o21 5 a2 -1 a2*o21 -5;
  estimate 'INTSFY-SWITCH, 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 'SWITCH , o21=2' intercept 1 o11c 0 o21 2 a2 -1 a2*o21 -2;
  estimate 'INTSFY-SWITCH, o21=2' intercept 0 o11c 0 o21 0 a2 2 a2*o21 4;
run;

* we find that among non-responders to EMT at the end of 3 months:
* at higher level of intermediate utterances=o21, it is better to A2=INTENSIFY.
* at lower level of intermediate utterances=022, it is better to A2=SWITCH.

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

****************;
**** S(b) ****;
****************;

** 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:
* stage 2: y = b7 + b8 o11c + b9 o21 + b10 a1 o21 + b11 a2 + e
* stage 1: yopt2 = b1 + b2 o11 + b3 o12 + b4 a1 o11 + b5 a1 o12 + b6 a1 + e
* yopt2 is the y adjusted for having followed the best stage 2 txt.
proc qlearn data=dat2 deriveci;
tailor1 o11 o12;
main2 o11c;
tailor2 o21;
response y;
stg2sample s;
stg1trt a1;
* based on the results above, we remove o12 o12*a1
* the reduced stage 1 model is: yopt2 = b1 + b2 o11 + b3 a1 o11 + b4 a1 + e
* before fitting it, however, we first specify contrasts of interest

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 EMT for children w/o11 = 10 = low;
* contrast 2 = mean future-optimal outcome under AAC for children w/o11 = 10 = low;
* contrast 3 = mean difference EMT - AAC for children w/o11 = 10 = low
* contrast 4 = mean future-optimal outcome under EMT for children w/o11 = 70 = hi;
* contrast 5 = mean future-optimal outcome under AAC for children w/o11 = 70 = hi;
* contrast 6 = mean difference EMT - AAC for children w/o11 = 70 = hi
run;

* finally, we fit q-learning;
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;

**************
**** S(c) ****
**************;

******************************************************************************;
** 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 ATS identified using q-learning, this code is to
** obtain the estimated mean outcome had the entire population followed this
** optimal ATS (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 ;
** the mean of yopt is the estimated mean under the optimal ATS 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 ATS found under Q-learning ;
** note this is larger than the Mean Y under best the "design-embedded" ATS (EMT,AAC), ;
** which was 62.2434;

*** eof;
*** eof;