**Lab 2: Plotting Longitudinal Data**

**Overview**

1. **Profile plots**

**Overall and by covariate value**

1. **Repeated box plots over time**
2. **Empirical summary plot and empirical prediction plot (sample mean +/- 2se and +/- 2sd error bars over time)**

**And by covariate**

1. **Correlation matrix**
2. **Scatterplot matrix**

Need to download 4 macros, bigmice, smallmice, and pain datasets from web.

**Lab Tasks**

Use the Big Mice data for this part. As usual, change libnames and file locations as appropriate for your computing setup.

1. **Profile plot of Big Mice**

/\*This Macro will be a shortcut so that we don't have to type

in the full directory name every time.

It's useful because if you want to run the lab code but

you change the directory of your files (i.e. running in at home)

then all you have to do is change the directory name ONCE right HERE\*/

\*%let workdir=C:\Documents and Settings\Leannica\Desktop\SCHOOL\B236 Winter 2008;

%let workdir=C:\Documents and Settings\sph\Desktop\;

\*Use Macro to Assign Library where downloaded data sets

and macros will be found;

libname b236 "&workdir";

\*SET GRAPH OPTIONS;

goptions cback=white;

axis1 minor=none label=( angle=**90** 'Weight(mg)')

order=(**100** to **1200** by **100**);

axis2 minor=none label=('Days') order=(**0** to **20** by **5**) offset=(**5**,**5**);

 \*offset puts space before the first and after the last tickmarks;

\*USE BIGMICE data;

\*PROFILE PLOTS;

**proc** **gplot** data=b236.bigmice;

 /\*This graph is on page 34 of the text\*/

 plot weight\*day=id /vaxis=axis1 haxis=axis2 nolegend;

 symbol1 i=join v=circle r=**35** color=black line=**1**;

 \*r = repeat (use number of subjects);

 title 'Profile Plot of BIG MICE Data';

 **run**;

1. **Repeated box plots of Big Mice data**

\*BOXPLOTS;

 /\*This graph is on page 32\*/

**proc** **gplot** data=b236.bigmice;

 plot weight\*day/ vaxis=axis1 haxis=axis2;

 symbol i=boxt00 bwidth=**3**;

 \*"00" high/low extremes for whisker length;

 \*"t"=add tops and bottoms to each line;

 \* bwidth= box width;

 title 'Box plots of MICE Data';

 **run**;

1. **Empirical prediction plot (mean +/- 2sd error bars)**

**and empirical summary plot (mean +/- 2se error bars)**

\*EMPIRICAL PREDICTION PLOT: MEAN +/- 2 SD;

 /\*This graph is on page 68\*/

**proc** **gplot** data=b236.bigmice;

 plot weight\*day / vaxis=axis1 haxis=axis2;

 symbol1 i=std2jt v=none color=black r=**1**;

 \*std=standard deviations;

 \*2=#of SD's;

 \*J=provide line throught means;

 \*t=add tops and bottoms to each line;

 title 'Average Weight(mg) +/- 2 SD';

 **run**;

\*EMPIRICAL SUMMARY PLOT: MEAN +/- 2 SE;

 /\*This graph is on page 68\*/

**proc** **gplot** data=b236.bigmice;

 plot weight\*day / vaxis=axis1 haxis=axis2;

 symbol1 i=std2mjt v=none color=black r=**1**;

 \*NOTE: USE of "m" makes this STANDARD ERRORS;

 \*std=standard errors;

 \*2=#of SE's;

 \*J=provide line throught means;

 \*t=add tops and bottoms to each line;

 title 'Average Weight(mg) +/- 2 SE';

**run**;

**quit**;

**3b. Alternative code for empirical prediction plot and empirical summary plot**

The following approach to drawing means +/- 2SE has some advantages over the previous code. In the previous code, the Y axis range will be for the entire data set. If you restrict the range to allow you to see the plot better, SAS will also erroneously restrict the data that is used in calculating means and sds, causing you to get an incorrect plot. The following approach of saving the means and sds to a file and working with the saved file allows you (1) to have a more sensible y axis range and (2) this code can be re-used later when we fit models to our data sets. We will use our fitted models to create a data set of estimated means and SEs and reuse the code to create an equivalent “inference plot” as opposed to these “empirical summary plots”.

\*Create a file with summary statistics for weight by day;

**proc** **means** data=b236.bigmice noprint;

 class day;

 var weight;

 output out=micem mean=mean std=std stderr=stderr n=n;

**run**;

\*Create a file with mean and mean+-2\*SD for weight by day;

**data** empred (keep=Y day); set micem;

 Y=mean; output;

 Y=mean - **2**\*std; output;

 Y=mean + **2**\*std; output;

**run**;

\*EMPIRICAL PREDICTION PLOT: MEAN +/- 2 SD;

**proc** **gplot** data=empred;

 plot Y\*day /vaxis=axis1 haxis=axis2;

 symbol1 i=hilocjt v=none color=black r=**1**;

 \*hilo=a solid vertical line connect the minimum and maximum Y values for each X value;

 \*c=assumes that there are three values of Y (HIGH, LOW, and CLOSE) for every value of X;

 \*j=joins the mean values or the close values (if HILOC is specified) with a line;

 \*t=add tops and bottoms to each line;

 title 'Average Weight(mg) +/- 2 SD';

**run**;

**quit**;

\*Create a file with mean and mean+-2\*SE for weight by day;

**data** emsum (keep=Y day); set micem;

 Y=mean; output;

 Y=mean - **2**\*stderr; output;

 Y=mean + **2**\*stderr; output;

**run**;

\* change the range of Y-axis;

axis3 minor=none label=( angle=**90** 'Weight(mg)') order=(**100** to **1100** by **100**);

\*EMPIRICAL SUMMARY PLOT: MEAN +/- 2 SE;

**proc** **gplot** data=emsum;

 plot Y\*day /vaxis=axis3 haxis=axis2;

 symbol1 i=hilocjt v=none color=black r=**1**;

 title 'Average Weight(mg) +/- 2 SE';

**run**;

**quit**;

1. **Empirical standard deviations over time**

\*PLOT STANDARD DEVIATIONS OVER TIME;

\*Change verical axis for SD PLOT;

axis1 minor=none label=( angle=**90** 'SD of Weight(mg)') order=(**0** to **140** by **20**);

/\*This plot is on page 41\*/

**proc** **gplot** data=micem;

 plot std\*day/vaxis=axis1 haxis=axis2;

 symbol1 i=none v=circle color=black height=**2** r=**1** ;

 title 'SD over time, BIG MICE DATA';

**run**;

1. **Correlation matrix and correlogram with the small mice data**

\*CREATING A CORRELOGRAM IS A MULTISTEP PROCESS;

\*For a better explanation of correlogram, see pgs 68-70 in the text.;

\*Use small mice dataset;

\*TRANSPOSE DATASET FROM LONG FORMAT TO WIDE FORMAT;

**proc** **transpose** data=b236.smallmice out=micet

 prefix=weight;

 by id;

 id day;

 var weight;

**run**;

\*CREATE DATA SET OF CORRELATIONS;

**proc** **corr** data=micet;

 var weight2--weight20;

 ods select pearsoncorr;

 \*only outputs the correlation table;

 ods output pearsoncorr=pearsoncorr;

 \*creates an output dataset for the correlation table;

 title 'Correlation Matrix for SMALL MICE';

**run**;

\*CREATE DATASET WITH ONLY THE UPPER TRIANGULAR VALUES;

 \*i.e. delete all values on and below the diagonal;

**data** tri (drop=i obsno variable);

 set pearsoncorr (drop=pweight2--pweight20);

 obsno=\_N\_;

 \*\_N\_ specifies the observation numbers;

 day=input(substr(variable,**7**,**2**),best.);

 \*This is used to create a day variable...

 We want the number after "weight"

 so we go to the 7th position of the string and extract

 the last 2 positions (the number). Then using input,

 we convert the number (which is a character string)

 into a number.;

 \*best refers to the format for the number.;

 array w{\*}weight2--weight20;

 do i=**1** to **7**;

 if i<=obsno then w{i}=**.**;

 end;

 \*Array and do loop sets all values on and below

 the diagonal to missing;

**run**;

\*CREATE THE DATASET IN LONG FORM;

**data** trilong (keep = day wt);

 set tri;

 array w{\*} weight2--weight20;

 do i=**1** to **7**;

 wt=w{i};

 output;

 end;

**run**;

\*DELETE MISSING OBSERVATIONS;

**data** trilong;

 set trilong;

 if wt=**.** then delete;

**run**;

\*CREATE LAG VARIABLE;

**data** trilag;

 set trilong;

 by day;

 if first.day then lags=**3**;

 else lags+**3**;

**run**;

\*SET UP AXES;

axis1 minor=none label=( angle=**90** 'Correlation' );

axis2 minor=none label=('Lag(days)');

\*PLOT CORRELOGRAM;

/\*Similar plot on page 67\*/

**proc** **gplot** data=trilag;

 plot wt\*lags=day/vaxis=axis1 haxis=axis2 nolegend;

 title 'Correlogram of SMALL MICE DATA';

 symbol i=join l=**1** v=circle color=black r=**10**;

**run**;

1. **Scatterplot matrix using the Pain data.**

/\*NEED TO DOWNLOAD THE 4 MACROS [boxanno.sas](http://rem.ph.ucla.edu/rob/rm/new/sascodes/boxanno.sas), [gdispla.sas](http://rem.ph.ucla.edu/rob/rm/new/sascodes/gdispla.sas). [gensym.sas](http://rem.ph.ucla.edu/rob/rm/new/sascodes/gensym.sas), [scatmat.sas](http://rem.ph.ucla.edu/rob/rm/new/sascodes/scatmat.sas);

These are available from course web site under lab 2.

<http://rem.ph.ucla.edu/rob/rm/new/index.html>\*/

\*SCATTERPLOT MATRIX USING PAIN DATA;

\*TRANSPOSE DATASET TO WIDE FORMAT;

**proc** **sort** data=b236.pain; by id;**run**;

**proc** **transpose** data=b236.pain out=painwide1(drop= \_NAME\_) prefix=l2pain;

 by id;

 id trial;

 copy cs treatment;

 var l2paintol;

 **run**;

\*SINCE WE COPIED OVER CS AND TREATMENT WE HAVE EXTRA OBSERVATIONS... SO NEED TO DELETE;

**proc** **sort** data=painwide1; by id; **run**;

**data** painwide;

 set painwide1;

 by id;

 if first.id;

 **run**;

\*USE MACROS TO CREATE SCATTERPLOT;

/\*This plot is on page 62\*/

%include "&workdir\boxanno.sas";

%include "&workdir\gdispla.sas";

%include "&workdir\gensym.sas";

%include "&workdir\scatmat.sas";

options device=WIN;

%***scatmat***(data=painwide, var=l2pain1-l2pain4, anno=BOX);

**TO DO TASKS**

Use the pain data for this part. When appropriate, plot the full data set and also separately by coping style.

1. Draw a profile plot for the pain data.

\*WE NEED TO CHANGE THE r=64;

* 1. See the skewness in the original data, so transform pain tolerance by taking a log.

\*THIS variable already exists in the dataset called l2paintol.

* 1. Draw the profile plot separately for attenders and distracters. How do the two coping styles differ?

DON’T FORGET TO SORT FIRST by coping style;

1. Draw repeated box plots over time, and separately by coping style.
2. Draw the empirical summary plot for the pain data. Draw separate curves by coping style but draw them on the same plot. Again, how do the two curves differ?

\*ADD A SMALL NUMBER TO TRIAL SO THAT PLOTS AREN'T ON TOP OF EACH OTHER;

**data** pain2;

 set b236.pain;

 if cs='attender' then trial=trial+**0.05**;

**run**;

\*TO GET THE GRAPH TO USE THE SECOND SYMBOL COMMAND FOR ONE OF THE TWO CS GROUPS WE HAVE TO EXPICITLY SAY R=1 IN THE SYMBOL1 STATEMENT BECAUSE WE PREVIOUSLY HAD A SYMBOL1 COMMAND WITH R=64 WHICH TELLS SAS TO USE SYMBOL1 IN A GRAPH 64 TIMES BEFORE SWITCHING TO SYMBOL2;

symbol1 i=std2mjt v=none color=red r=1;

symbol2 i=std2mjt v=none color=black l=2; /\*l=2 GIVES THE LINE TYPE\*/

1. Draw the scatterplot matrix (a) separately by coping style, and (b) in one plot, with different plotting symbols for attenders and distracters.

\*CREATE TWO DATA SETS BY COPING STYLE FROM PAINWIDE DATASET;

\*REMEMBER WE NEED TO USE WIDE DATA SETS FOR SCATTER MATRIX;

**proc** **sort** data=painwide; by cs;**run**;

**data** attenders distracters;

set painwide;

by cs;

if cs='attender' then output attenders;

else output distracters;

**run**;

\*FOR (a), include “names=A1 A2 A3 A4” in %scatmat so that we know it is attenders data … ditto for distracters;

\*FOR (b), include “group=cs” in %scatmat;

1. Calculate the variance over time and correlation matrix separately for each coping style.

\*VARIANCE OVER TIME REQUIRES DATA IN LONG FORMAT, BUT CORRELATION MATRIX REQUIRES DATA IN WIDE FORMAT;

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**TO DO TASK SOLUTIONS**

\*TO DO TASKS;

\*Profile Plot for Pain Data Set;

\*PROFILE PLOTS;

\*SET GRAPH OPTIONS;

goptions cback=white;

axis1 minor=none label=( angle=**90** 'Time(sec)' );

axis2 minor=none label=('Trial');

\*IT IS VERY IMPORTANT FOR CORRECTLY MAKING THE PROFILE PLOT TO SET r= (WHICH STANDS FOR REPEAT)

TO A NUMBER AT LEAST AS LARGE AS THE NUMBER OF SUBJECTS;

symbol1 i=join v=circle r=**64** color=black;

title 'Profile Plot of PAIN Data Set';

**proc** **gplot** data=b236.pain;

 plot paintol\*trial=id /vaxis=axis1 haxis=axis2 nolegend;

**run**;

quit;

\*PROFILE PLOT OF LOG TRANSFORMED PAIN VARIABLE;

symbol1 i=join v=circle r=**64** color=black;

title 'Profile Plot of Log2(PAINTOL)';

**proc** **gplot** data=b236.pain;

 plot l2paintol\*trial=id /vaxis=axis1 haxis=axis2 nolegend;

**run**;

quit;

\*PROFILE PLOT OF LOG TRANSFORMED PAIN VARIABLE BY COPING STYLE;

\*SORT FIRST;

**proc** **sort** data=b236.pain; by cs;**run**;

symbol1 i=join v=circle r=**64** color=black;

title 'Profile Plot of Log2(PAINTOL)';

**proc** **gplot** data=b236.pain;

by cs;

 plot l2paintol\*trial=id /vaxis=axis1 haxis=axis2 nolegend;

**run**;

quit;

\*DIFFERENCE IN COPING STYLE? - ans: distracters appear to have higher log2paintol values.;

\*REPEATED BOX PLOTS OVER TIME BY COPING STYLE;

\*BOXPLOTS;

**proc** **sort** data=b236.pain; by cs trial;**run**;

**proc** **boxplot** data=b236.pain;

by cs;

 plot l2paintol\*trial;

 title 'Box plots of Log2(Paintol)';

**run**;

\*Draw the empirical summary plot for the pain data.

Draw separate curves by coping style but draw them on the same plot.

Again, how do the two curves differ?

\*ADD A SMALL NUMBER TO TRIAL SO THAT PLOTS AREN'T ON TOP OF EACH OTHER;

**data** pain2;

 set b236.pain;

 if cs='attender' then trial=trial+**0.05**;

**run**;

\*TO GET THE GRAPH TO USE THE SECOND SYMBOL COMMAND FOR ONE OF THE TWO CS GROUPS

WE HAVE TO EXPICITLY SAY R=1 IN THE SYMBOL1 STATEMENT BECAUSE WE PREVIOUSLY HAD A SYMBOL1

COMMAND WITH R=64 WHICH TELLS SAS TO USE SYMBOL1 IN A GRAPH 64 TIMES BEFORE SWITCHING TO SYMBOL2;

axis1 minor=none label="Trial";

axis2 minor=none label=("Average Log2(Paintol)");

symbol1 i=std2jt v=none color=red r=**1**;

symbol2 i=std2jt v=none color=black l=**2**; /\*l=2 GIVES THE LINE TYPE\*/

\*SUMMARY PLOTS;

title 'Average Log2(Paintol) +/- 2 SD';

**proc** **gplot** data=pain2;

 plot l2paintol\*trial=cs/ vaxis=axis2 haxis=axis1;

**run**;

quit;

\*WE MODIFLY THE SYMBOL STATEMENT BY INSERTING M, SPECIFYING SE of MEAN, rather than SD;

symbol1 i=std2mjt v=none color=red r=**1**;

symbol2 i=std2mjt v=none color=black l=**2**; /\*l=2 GIVES THE LINE TYPE\*/

\*SUMMARY PLOTS;

title 'Average Log2(Paintol) +/- 2 SE';

**proc** **gplot** data=pain2;

 plot l2paintol\*trial=cs/ vaxis=axis2 haxis=axis1;

**run**;

quit;

\*CREATE TWO DATA SETS BY COPING STYLE FROM PAINWIDE DATASET;

\*REMEMBER WE NEED TO USE WIDE DATA SETS FOR SCATTER MATRIX;

**proc** **sort** data=painwide; by cs;**run**;

**data** attenders distracters;

set painwide;

by cs;

if cs='attender' then output attenders;

else output distracters;

**run**;

**proc** **print** data=painwide;**run**;

%***scatmat***(data=attenders, var=l2pain1-l2pain4,names=Al2pain1 Al2pain2 Al2pain3 Al2pain4, anno=box);

%***scatmat***(data=distracters, var=l2pain1-l2pain4, names=Dl2pain1 Dl2pain2 Dl2pain3 Dl2pain4, anno=box);

\*(b) in one plot, with different plotting symbols for attenders and distracters;

%***scatmat***(data=painwide, var=l2pain1-l2pain4, group=cs, anno=box);

\*5. Calculate the variance over time and correlation matrix separately for each coping style.;

\*PLOT STANDARD DEVIATIONS OVER TIME;

**proc** **sort** data=b236.pain; by cs;**run**;

**proc** **means** data=b236.pain noprint;

by cs;

class trial;

var l2paintol;

output out=painm mean=mean var=var std=std;

**run**;

symbol1 i=none v=circle color=red r=**1**;

axis1 minor=none label=( angle=**90** 'SD of Log2(Paintol)' );

axis2 minor=none label=('Trial');

**proc** **gplot** data=painm;

by cs;

plot std\*trial/vaxis=axis1 haxis=axis2;

title 'SD over time, PAIN DATA';

**run**;

\*CORRELATION MATRIX;

\*USE WIDE DATA SET;

\*ods trace on <-this is used is you want to find out in the log what output the proc gives you;

ods select pearsoncorr; \*only outputs the correlation table;

ods output pearsoncorr=pearsoncorr; \*creates an output dataset for the correlation table;

title 'Correlation Matrix for PAIN DATA';

**proc** **corr** data=painwide; by cs;var l2pain1-l2pain4;**run**;

\*ods trace off <-turn off the "trace" from above;