Prof. Brent Coull

BIO 226: Applied Longitudinal Analysis

Homework 2 Solutions
Due Thursday, February 26, 2015
[100 points]

Purpose:
To provide an introduction to the use of PROC MIXED for analyzing data from a single-group repeated measures design.

Instructions:
1. For each question requiring data analysis, support your conclusions by including the relevant SAS output in your answer.
2. Include your SAS program as an Appendix to your solutions (PLEASE dont include the output from your program!).
3. Homework should be turned by class on the due date.

Joint Modeling of Mean and Covariance in the Dental Growth Study:
In a study of dental growth, measurements of the distance (mm) from the center of the pituitary gland to the pteryomaxillary fissure were obtained on 11 girls and 16 boys at ages 8, 10, 12 and 14 years (Potthoff and Roy, 1964). The data are in the file "dental.txt" on the course web page. Each row of the data set contains the following six variables: subject ID, gender ("F" or "M"), and the measurements at ages 8, 10, 12 and 14 years, respectively.

For all analyses in this homework, subset the dataset to contain only the measurements for the girls.

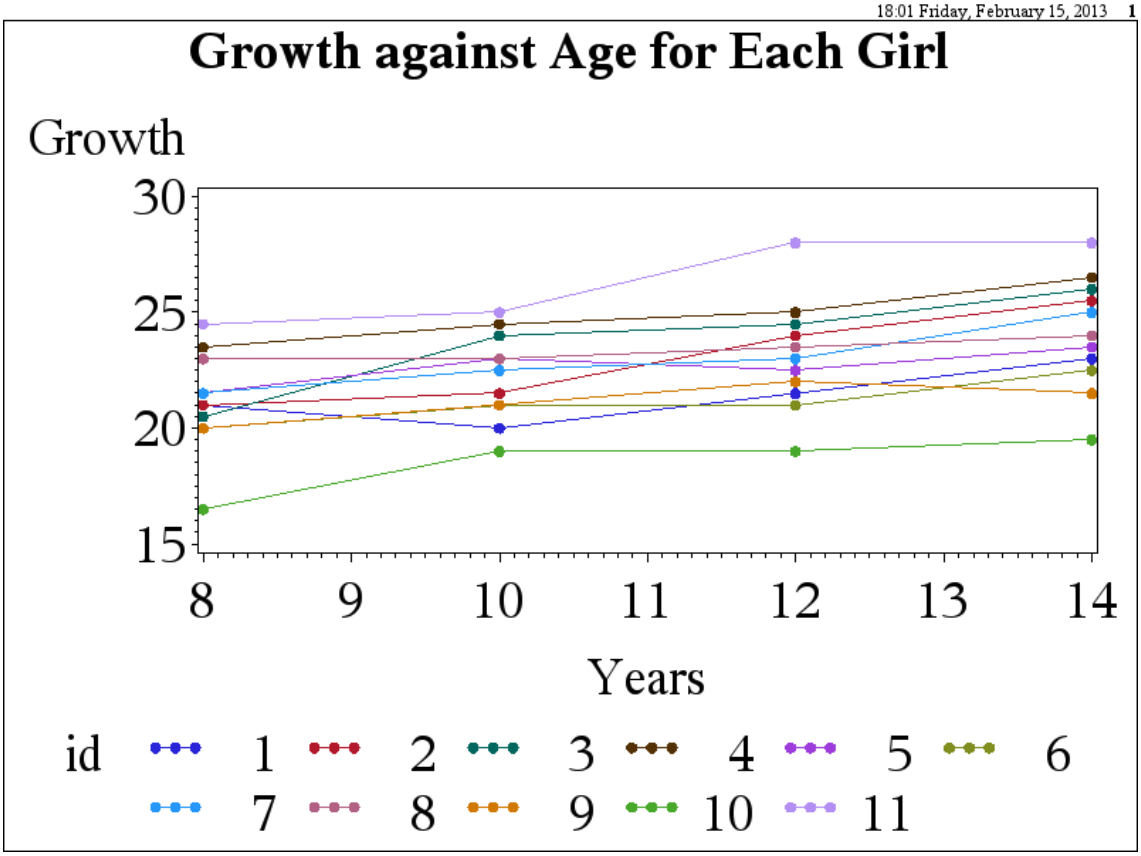
PART A: Descriptive Analyses

Problem 1
[15 points: 5 for plot, 10 for a quick description that includes mention of increasing trend.] Plot the observed trajectories of distance for the 11 girls (all on one plot). Briefly, describe the important patterns that are apparent in the plot.

Measurements seems to increase over time for all eleven girls in the study at approximately the same rate.

```
data dental1;
infile 'dental.txt';
input id gender $ y8 y10 y12 y14;
run;

data girls1;
set dental1;
if gender="F";
run;
```

```
data dental2;
set dental1;
age=8; age8 = 0; agecat=8; y=y8; output;
age=10; age8 = 2; agecat=10; y=y10; output;
age=12; age8 = 4; agecat=12; y=y12; output;
age=14; age8 = 6; agecat=14; y=y14; output;
keep id gender age agecat age8 y;
run;

data girls2;
set dental2;
if gender="F";
run;

proc gplot data=girls2;
symbol1 interpol=join value=dot;
symbol2 interpol=join value=triangle;
plot y*age=id /haxis=axis1 vaxis=axis2;
axis1 label=("Years");
axis2 label=("Growth");
title 'Growth against Age for Each Girl';
run;
```


Growth against Age for Each Girl
Growth
30
25
20
15
8
9
10
11
12
13
14
Years
id
1
2
3
4
5
6
7
8
9
10
11

Problem 2

[15 points: 5 for statistics, 5 for plot, 5 for comment.] Obtain descriptive statistics for the distances in girls including means and standard deviations at each measurement occasion. Plot the means against time and comment on the shape of their trajectory. Also, comment on how the standard deviations are changing with time.

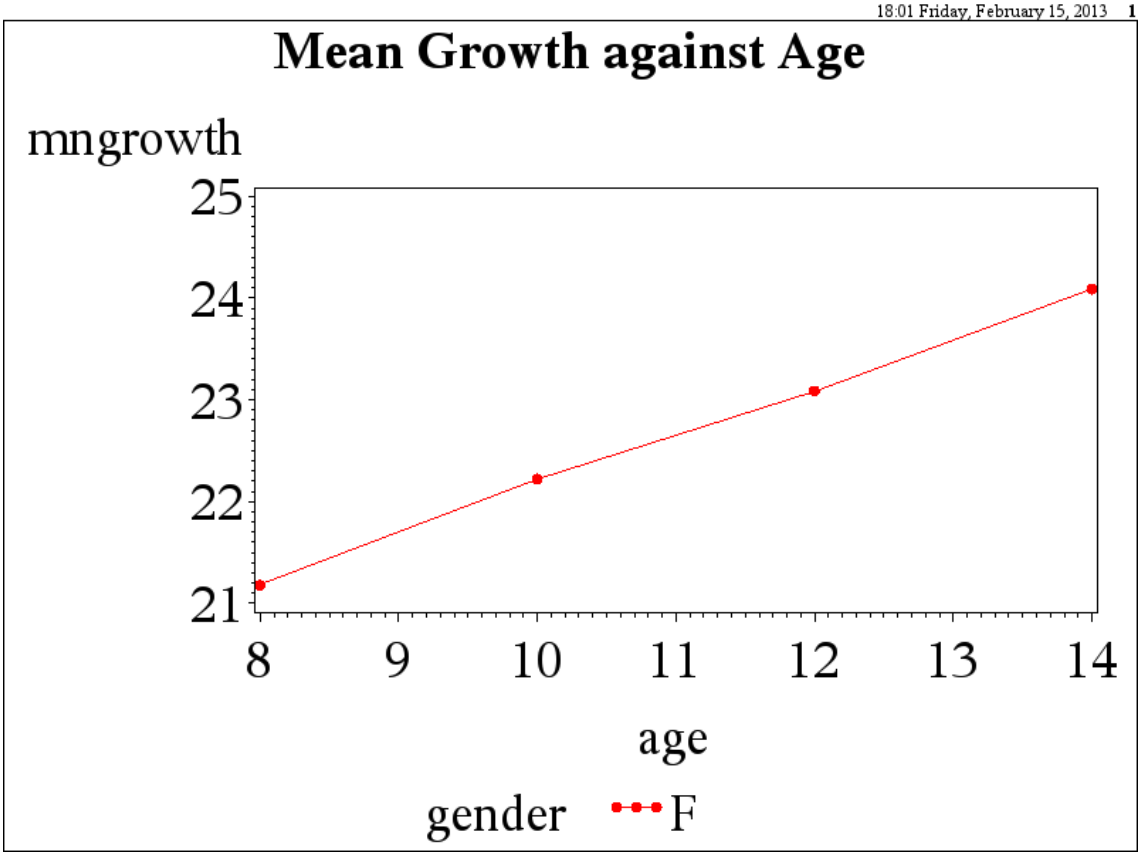
```
proc means data=girls2 n mean std var nway;
var y;
class gender age;
output out=meandata mean=mngrowth;
run;
```

```
---------------------------------------------------------------------------
```

```
                            The MEANS Procedure

                          Analysis Variable : y

                             N
gender               age    Obs     N           Mean        Std Dev       Variance
---------------------------------------------------------------------------------
F                      8     11    11     21.1818182      2.1245320      4.5136364

                      10     11    11     22.2272727      1.9021519      3.6181818

                      12     11    11     23.0909091      2.3645103      5.5909091

                      14     11    11     24.0909091      2.4373980      5.9409091
---------------------------------------------------------------------------------
```

```
---------------------------------------------------------------------------
proc gplot data=meandata;
symbol1 color=red interpol=join value=dot;
symbol2 interpol=join value=triangle;
plot mngrowth*age=gender;
title 'Mean Growth against Age';
run;
```

The mean growth increases over time in a roughly linear way. The standard deviations appear relatively constant over time, with only a slight increase.


Mean Growth against Age
mngrowth
25
24
23
22
21
8
9
10
11
12
13
14
age
gender
F

Problem 3

[15 points: 5 for covariances, 5 for correlations, 5 for comments on correlations.] Obtain the variance-covariance and correlation matrices for the repeated measurements over time. Briefly, describe the important patterns in the correlations.

```
proc corr data=girls1 cov;
by gender;
var y8 y10 y12 y14;
run;
```

Covariance Matrix, DF = 10

	y8	y10	y12	y14
y8	4.513636364	3.354545455	4.331818182	4.356818182
y10	3.354545455	3.618181818	4.027272727	4.077272727
y12	4.331818182	4.027272727	5.590909091	5.465909091
y14	4.356818182	4.077272727	5.465909091	5.940909091

Pearson Correlation Coefficients, N = 11
Prob > |r| under H0: Rho=0

	y8	y10	y12	y14
y8	1.00000	0.83009 0.0016	0.86231 0.0006	0.84136 0.0012
y10	0.83009 0.0016	1.00000	0.89542 0.0002	0.87942 0.0004
y12	0.86231 0.0006	0.89542 0.0002	1.00000	0.94841 <.0001
y14	0.84136 0.0012	0.87942 0.0004	0.94841 <.0001	1.00000

The correlations are roughly constant across time. They are quite high and positive, at above 0.8.

Problem 4

[10 points] Identify one unexpected feature of this correlation structure generated in question 3. Provide a possible reason for this unexpected feature.

The correlation does not really seem to decrease over time. Note that the y8 and y10 correlation is lower than the y8 and y12 correlation. Part of the reason may be small sample sizes, which would mean that the certainty about these correlations is low.

PART B: Repeated Measures Models

Problem 5

[20 points: 5 for fitting correct models, 5 for each part a,b,c] Use PROC MIXED to fit a repeated measures model in which there is separate parameter for the mean distance at each of the four ages [HINT: Use the NOINT option on the MODEL statement]. Use an unstructured variance-covariance structure. For estimating parameters, fit the model using (i) the METHOD=ML option and (ii) the METHOD=REML option on the PROC MIXED statement.

```
title 'ML method';
proc mixed data=girls2 method=ml;
class id agecat age;
model y=agecat / noint chisq s;
/* NOINT - excludes fixed-effect intercept from model */
repeated age / type=un subject=id r rcorr;
run;
title;
```

```
                          Estimated R Matrix for id 1

               Row        Col1        Col2        Col3        Col4

                 1      4.1033      3.0496      3.9380      3.9607
                 2      3.0496      3.2893      3.6612      3.7066
                 3      3.9380      3.6612      5.0826      4.9690
                 4      3.9607      3.7066      4.9690      5.4008

                     Estimated R Correlation Matrix for id 1

               Row        Col1        Col2        Col3        Col4

                 1      1.0000      0.8301      0.8623      0.8414
                 2      0.8301      1.0000      0.8954      0.8794
                 3      0.8623      0.8954      1.0000      0.9484
                 4      0.8414      0.8794      0.9484      1.0000

                                    Standard
    Effect    agecat    Estimate       Error      DF    t Value    Pr > |t|

    agecat     8         21.1818      0.6108      11      34.68      <.0001
    agecat    10         22.2273      0.5468      11      40.65      <.0001
    agecat    12         23.0909      0.6797      11      33.97      <.0001
    agecat    14         24.0909      0.7007      11      34.38      <.0001
```

```
title 'REML method';
proc mixed data=girls2;
class id agecat age;
```

```
model y=agecat / noint chisq s;
repeated age / type=un subject=id r rcorr;
run;
title;
```

Estimated R Matrix for id 1

Row	Col1	Col2	Col3	Col4
1	4.5136	3.3545	4.3318	4.3568
2	3.3545	3.6182	4.0273	4.0773
3	4.3318	4.0273	5.5909	5.4659
4	4.3568	4.0773	5.4659	5.9409

Estimated R Correlation Matrix for id 1

Row	Col1	Col2	Col3	Col4
1	1.0000	0.8301	0.8623	0.8414
2	0.8301	1.0000	0.8954	0.8794
3	0.8623	0.8954	1.0000	0.9484
4	0.8414	0.8794	0.9484	1.0000

Effect	agecat	Estimate	Standard Error	DF	t Value	Pr > |t|
agecat	8	21.1818	0.6406	11	33.07	<.0001
agecat	10	22.2273	0.5735	11	38.76	<.0001
agecat	12	23.0909	0.7129	11	32.39	<.0001
agecat	14	24.0909	0.7349	11	32.78	<.0001

(a) Does either method of estimation give the same estimates of the variances and covariances as in your answer to question 3? If so, which one or ones?

REML gives the same estimates of the variances and covariances as in question 3. ML does not.

(b) Compare the variances and covariances for the two methods of estimation (i.e. ML versus REML). If they are not the same, comment on why they are different and state which method should be preferred and why?

REML gives a less biased estimate of the variances and covariances than ML because it takes into account the fact that β is also estimated. Thus, REML gives a preferred estimate for the variances and covariances.

(c) Compare the estimated betas in the regression model to the sample means calculated in question 2. Provide a reason for any relationship you notice between these two sets of estimates.

The estimated βs are the same for both ML and REML, and both are identical to the sample means because we fit the saturated model to the data.

Problem 6

[25 points: 10 for part a, 10 for part b, and 5 for part c] Using METHOD=ML and an unstructured variance-covariance structure, fit a model which assumes that the mean distance changes linearly with time.

```
proc mixed data=girls2 method=ml;
class id agecat age;
model y=agecat / noint chisq s;
/* NOINT - excludes fixed-effect intercept from model */
repeated age / type=un subject=id r rcorr;
run;
```

```
                    Fit Statistics

-2 Log Likelihood                130.5
```

```
proc mixed data=girls2 method=ml;
class id age;
model y=agecat / chisq s;
repeated age / type=un subject=id r rcorr;
run;
```

```
                    Fit Statistics

-2 Log Likelihood                130.6
```

(a) Why can this model be considered to be nested within the model fitted in question 5?

As in slide 28 Lab 2,

$E[growth_{ij}] = \beta_0^c I(age_{ij} = 8) + \beta_1^c I(age_{ij} = 10) + \beta_2^c I(age_{ij} = 12) + \beta_3^c I(age_{ij} = 14)$

$E[growth_{ij}] = \beta_0 + \beta_1 age_{ij}$

at age 8 $E[growth] = \beta_0^c = \beta_0 + 8\beta_1$

at age 10 $E[growth] = \beta_1^c = \beta_0 + 10\beta_1$

at age 12 $E[growth] = \beta_2^c = \beta_0 + 12\beta_1$

at age 14 $E[growth] = \beta_3^c = \beta_0 + 14\beta_1$

so we see that model 2 is a special case of model 1. Model 2 is nested in model 1.

(b) By undertaking a likelihood ratio test, evaluate the goodness of fit of this model compared with the model fitted (with METHOD=ML) in question 5. Which model do you prefer and why?

As in slide 29 Lab 2,

$$\chi_2^2 = 2(\log \hat{L}_{full} - \log \hat{L}_{reduced}) = (-2 \log \hat{L}_{reduced}) - (-2 \log \hat{L}_{full}) = 130.6 - 130.5 = 0.1$$

```
DATA pvalues;
chsq = SDF('chisquare',0.1,2);
RUN;
PROC PRINT DATA=pvalues;
RUN;
```

```
                                Obs      chsq

                                 1     0.95123
```

Thus, we do not reject the linear model. We prefer the simpler linear model because it fits well.

(c) Why should you not use the REML log likelihood to compare models in this question?

REML should not be used to compare different mean functions because the penalty term in REML depends on the regression mean model specification. See Lecture 5, slide 25.