Prof. Brent Coull

BIO 226: Applied Longitudinal Analysis

Homework 5 Solutions

Due Thursday, April 2, 2015

[100 points]

Purpose:
To review concepts of linear mixed effects models.

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 (but not your SAS output) as an appendix to your solutions. In general, this will only be reviewed during grading to help identify a major problem affecting your answers to questions so please do not cross-reference the appendix in your answers to questions.
Late homework will not be graded unless you make arrangements with the Instructor prior to the due date/time.

Analysis of Data from 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.
Note: In HW2 you considered this study, but analyzed data only from the girls. In this homework please analyze ALL the data.

Problems:

Problem 1
[15 points: 5 for each part] **Descriptive Analyses:**

(a) [5 points] Describe key features of the studys design and the completeness of data.

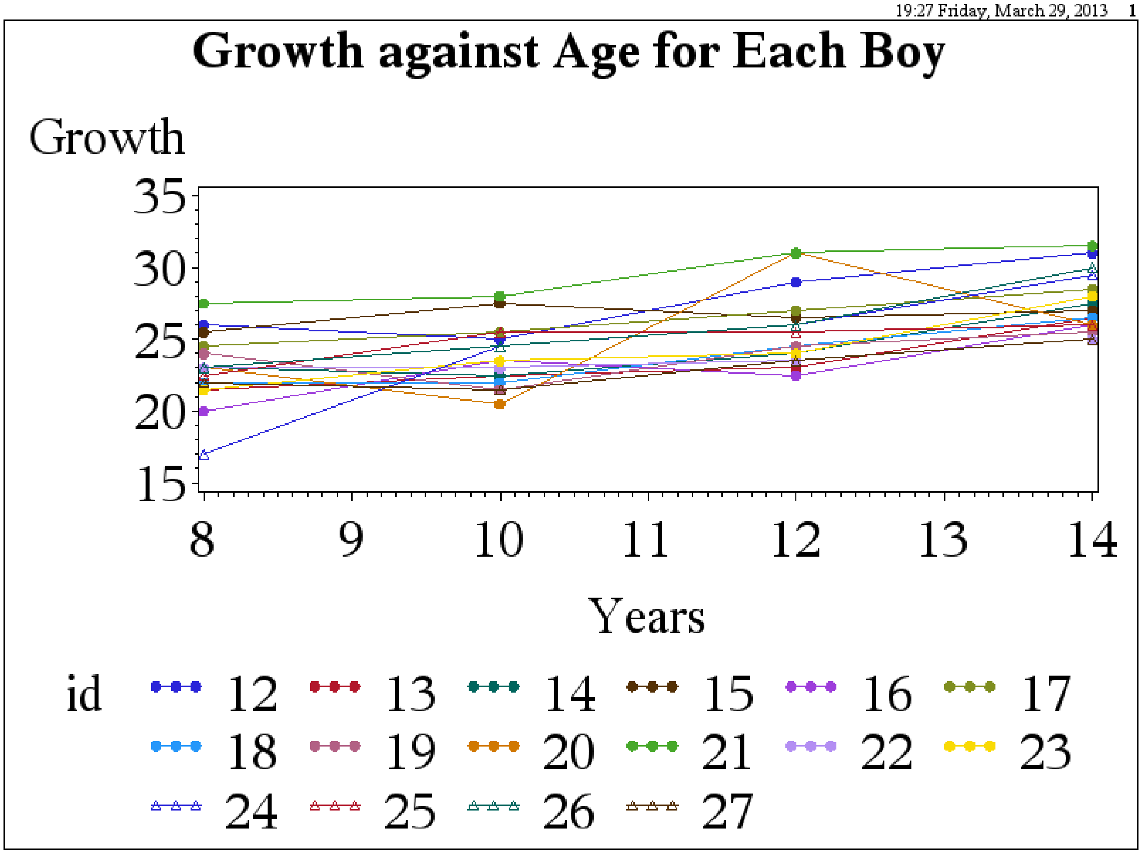
```
data dental1;
infile 'dental.txt';
input id gender $ y8 y10 y12 y14;
run;

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

data boys1;
```

```
set dental1;
if gender="M";
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;

data boys2;
set dental2;
if gender="M";
run;

proc means data=dental2 n mean std var nway;
var y;
class gender age;
output out=meandata mean=mngrowth;
run;
```

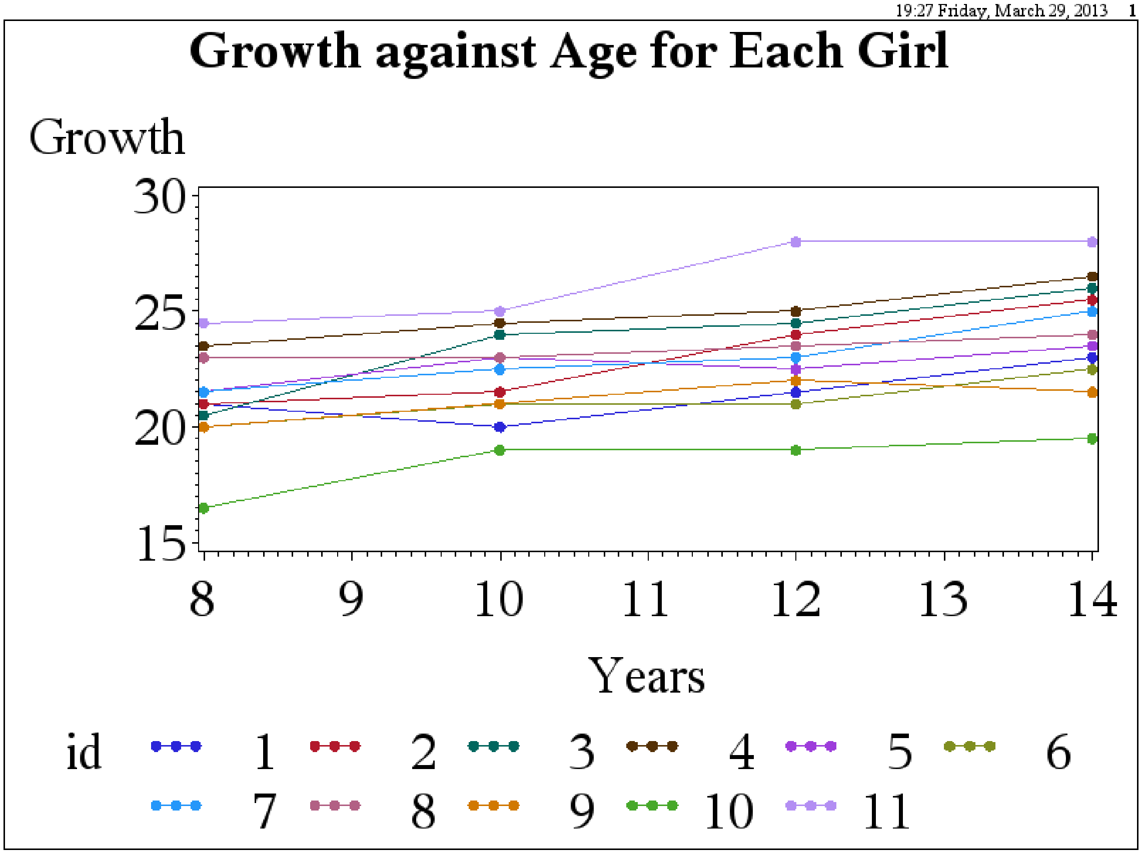
```
------------------------------------------------------------------------------------------
                         N
gender             age  Obs     N          Mean       Std Dev      Variance
--------------------------------------------------------------------------------
F                    8   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

M                    8   16    16    22.8750000     2.4528895     6.0166667

                    10   16    16    23.8125000     2.1360009     4.5625000

                    12   16    16    25.7187500     2.6518468     7.0322917
```

```
                    14      16      16       27.4687500        2.0854156        4.3489583
--------------------------------------------------------------------------------
```

(b) [5 points] Create two plots, one for each gender, showing the trajectories of each subjects distance over time. Comment on the notable features in these trajectories.

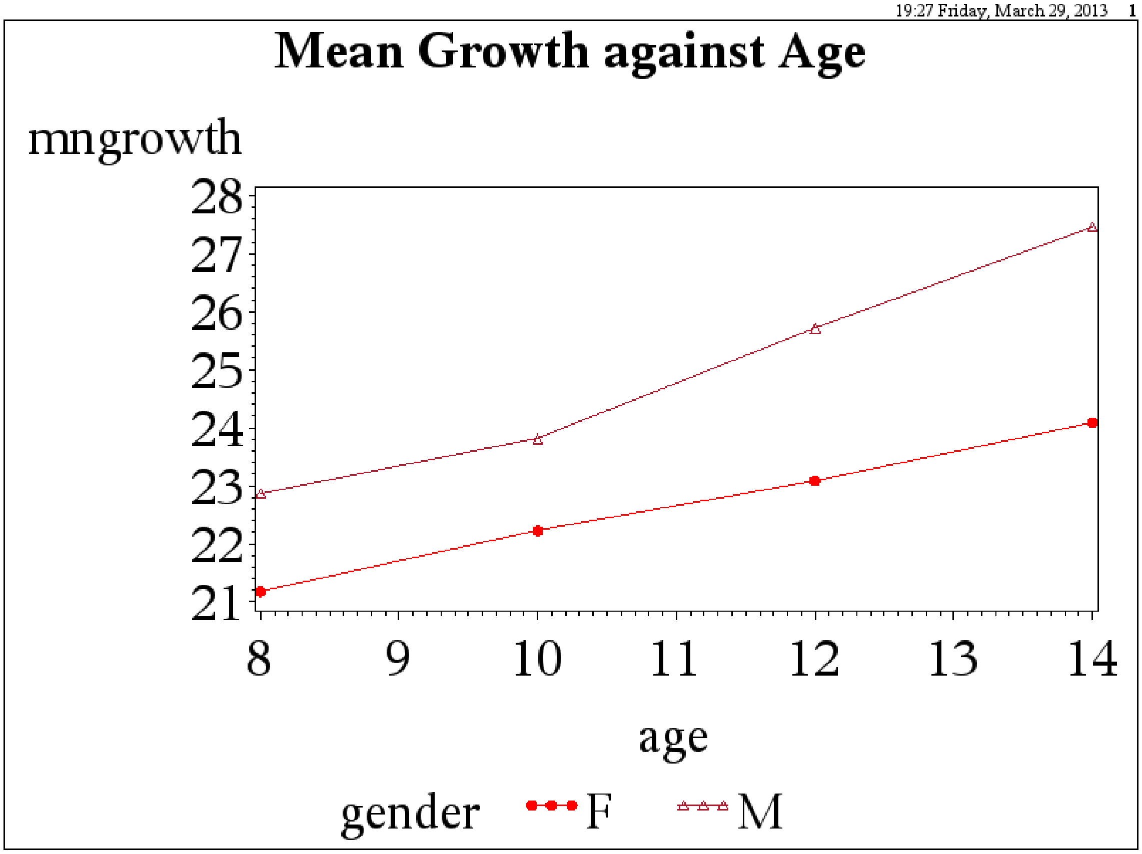


For both boys and girls, there appears on average to be a positive trend between dental growth and age. This is also seen for individual subjects, though there are some declines in dental growth between successive measurements (perhaps due, in part at least, to measurement error?). At any given age, the variability in dental growth measurements appears to be similar for boys as for girls. However, there appears to be less tracking (i.e. maintain similar rank at each successive time) among boys than among girls as suggested by the greater number of crossing trajectories for boys than girls (with the caveat that there are more boys than girls so somewhat difficult to make this assessment).


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

(c) [5 points] Create a single plot of the distance over time for each gender. Comment on the notable features in this plot.

Create a single plot of the mean grip strength over time for each treatment. Comment on the notable features in this plot. 2 points



We notice that the time trends for girls and boys look fairly linear and are increasing with time, and that boys have longer mean growth measurements than girls for all times. It also seems that the rate of increase is higher for boys than girls after age 10.

Problem 2

[85 points:]**(Joint) Mixed Effects Analysis** Using PROC MIXED, fit a model for distance over (continuous) time which includes subject-specific intercepts and slopes as random effects and allows both the mean intercept and the mean slope (fixed effects) to differ by gender.

(a) [10 points] State the model being fitted including any distributional assumptions. As much as possible, use the same notation as in the course notes.

The mixed model can be stated as

$Y_{ij} = \beta_1 + \beta_2 gender_i + \beta_3 age_{ij} + \beta_4 gender_i * age_{ij} + b_{1i} + b_{2i} age_{ij} + \epsilon_{ij}$

where

Y_{ij} is the distance for subject i at age j

age_{ij} is the age of measurement of Y_{ij} (takes values 8, 10, 12 and 14 years)

$\beta_1, \beta_2, \beta_3, \beta_4$ are the fixed effect parameters: intercept, gender main effect, age main effect, and interaction between age and gender.

b_{1i} is the random effect of intercept for subject i, and b_{2i} is the random effect of slope for subject i.

We assume that $(b_{1i}, b_{2i})' \sim N_2(\mathbf{0}, \mathbf{G})$ where $\mathbf{G}$ is the variance-covariance matrix, specifically $Var(b_{1i}) = \mathbf{G}_{11}$, $Var(b_{2i}) = \mathbf{G}_2$, $Cov(b_{1i}, b_{2i}) = \mathbf{G}_{12} = \mathbf{G}_{21}$.

(b) [10 points] Compare the within-subject error variance to the between subject variance of the intercepts. What do you conclude?

From fitting the mixed model, we obtain the following parameter estimates for the marginal covariance:

Covariance Parameter Estimates

Cov Parm	Subject	Estimate
UN(1,1)	id	5.7864
UN(2,1)	id	-0.2896
UN(2,2)	id	0.03252
Residual		1.7162

From the table we find that the within-subject error variance (`Residual`, 1.7162) is small relative to the between-subject variance of the intercepts (the `UN(1,1)` term, which is $Var(b_{1i}) = \mathbf{G}_{11} = 5.7864$).

(c) [10 points] Obtain estimates for the differences in mean intercept and mean slope by gender and their associated standard errors. What do you conclude about these differences?

Solution for Fixed Effects

Effect	gender	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept		16.3406	1.0185	25	16.04	<.0001
age		0.7844	0.08600	25	9.12	<.0001
gender	F	1.0321	1.5957	54	0.65	0.5205
gender	M	0	.	.	.	.
age*gender	F	-0.3048	0.1347	54	-2.26	0.0277
age*gender	M	0	.	.	.	.

The difference in mean intercept is $\widehat{\beta}_2 = 1.0321$, with standard error 1.5957. It is not significantly different from zero (p=0.5205). The difference in mean slope is $\widehat{\beta}_4 = -0.3048$ per year with standard error 0.1347. This is difference is significantly different from zero (p=0.0277) with greater increases in distance for males compared with females.

(d) [10 points] Obtain and provide an interpretation for the correlation of the intercept and slope among subjects.

The G correlation matrix is:

Estimated G Correlation Matrix

Row	Effect	id	Col1	Col2
1	Intercept	1	1.0000	-0.6676
2	age	1	-0.6676	1.0000

The estimated correlation (-0.6676) between the intercept and the slope among subjects is quite negative, indicating a negative association between subjects rate of change in distance and their baseline level within each gender.

(e) [10 points] From your model, conditional upon gender, obtain the standard deviation of the slopes among subjects. Provide a clinically meaningful interpretation of this standard deviation.

From the fitting the mixed model, we obtain the following estimated G matrix:

Estimated G Matrix

Row	Effect	id	Col1	Col2
1	Intercept	1	5.7864	-0.2896
2	age	1	-0.2896	0.03252

The standard deviation of the slopes among subjects is $\sqrt{0.03252} = 0.18$. This standard deviation measures the variability among individuals in change in distance per year, conditional on gender. If the normality assumption is reasonable, then a 95% normal range for change of distance per year would be $0.7844 \pm 1.96 * 0.18 = (0.4316, 1.1372)$ for males and $0.7844 - 0.3048 \pm 1.96 * 0.18 = (0.1268, 0.8324)$ for females.

(f) [10 points] Obtain the marginal correlation matrix for the vector of responses, $Y_i = (Y_{i0}, Y_{i1}, Y_{i2}, Y_{i3})$, for the subject with $ID = 1$. Considering your findings in response to question 3b, why are the correlations relatively large but not extremely strong?

From fitting the mixed to model, we obtain the following marginal covariance and correlation matrices:

Estimated V Matrix for id 1

Row	Col1	Col2	Col3	Col4
1	4.9502	3.1751	3.1162	3.0574
2	3.1751	4.9625	3.3176	3.3888
3	3.1162	3.3176	5.2351	3.7202
4	3.0574	3.3888	3.7202	5.7679

```
                    Estimated V Correlation Matrix for id 1

          Row         Col1          Col2          Col3          Col4

            1       1.0000        0.6406        0.6122        0.5722
            2       0.6406        1.0000        0.6509        0.6334
            3       0.6122        0.6509        1.0000        0.6770
            4       0.5722        0.6334        0.6770        1.0000
```

Since the within-subject variance is smaller than the between-subject variance, particularly the variance of the random intercept, the correlations are somewhat high.

(g) [10 points] Are the random slopes necessary for these data? Answer this question by fitting the model with the same fixed effects but now with just random intercepts. Which model do you prefer for these data and why?

Although one can compare the random intercepts - random slopes model to random intercepts using likelihood ratio, the statistic does not have a χ^2 distribution (see Lab 5) and one has to use specialized cutoff values from the back of the FLW book. Alternatively one can just compare the AIC values from the two models. The random intercepts model has a lower AIC (437.8) than the random intercepts and slopes model (440.6). Therefore, there is no evidence that we need the random slopes for these data.

(h) [15 points: 5 points each] The data for this study could also be analyzed using the repeated measures approach discussed extensively earlier in the course.

i) [5 points] Define a repeated measures model with an unstructured variance-covariance matrix and the same form of fixed effects model as in question 3a which could be used.

$Y_{ij} = \gamma_1 + \gamma_2 age_{ij} + \gamma_3 gender_{ij} + \gamma_4 age_{ij} * gender_{ij} + \tau_{ij}$

where $gender_{ij} = M$ for males and $= F$ for females.

Let $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, Y_{i3}, Y_{i4})'$ and $\gamma = (\gamma_1, \gamma_2, \gamma_3, \gamma_4)'$. The model assumes $\boldsymbol{\tau}_i \sim N(\mathbf{0}, \Sigma)$ where Σ is a 4×4 unstructured matrix (requiring 10 parameters).

ii) [5 points] Fit this model and use the AIC statistic to assess the goodness of fit of the variance-covariance structure induced by the mixed effects model. What do you conclude?

```
proc mixed data=dental2 method=reml;
  class id gender agecat;
  model y=age gender age*gender/s;
  repeated agecat/subject=id type=un r rcorr;
run;
```

The AIC for the repeated measure model (444.5) is larger than the one for the mixed model (440.6). Hence the variance-covariance structure induced by the mixed model is preferred.

iii) [5 points] Now fit the model with a compound symmetry covariate pattern instead of an unstructured variance covariance matrix. Compare the results you get from this model to those obtained from the random intercept model you fit in 2g, and explain the reason behind any similarities or differences between these two sets of results.

The results are identical, because the random intercepts model induces a compound symmetric covariance pattern for the marginal variance covariance matrix (Σ). Therefore these two models are equivalent.