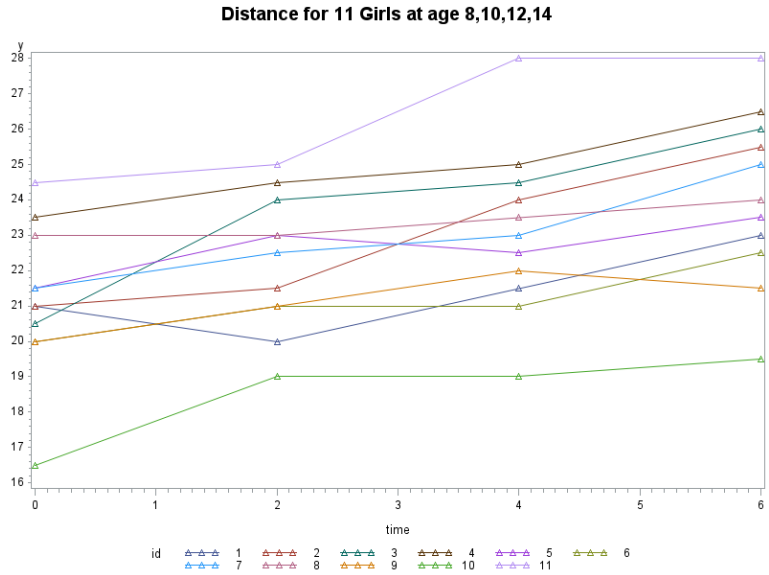
Part A: Descriptive Analyses

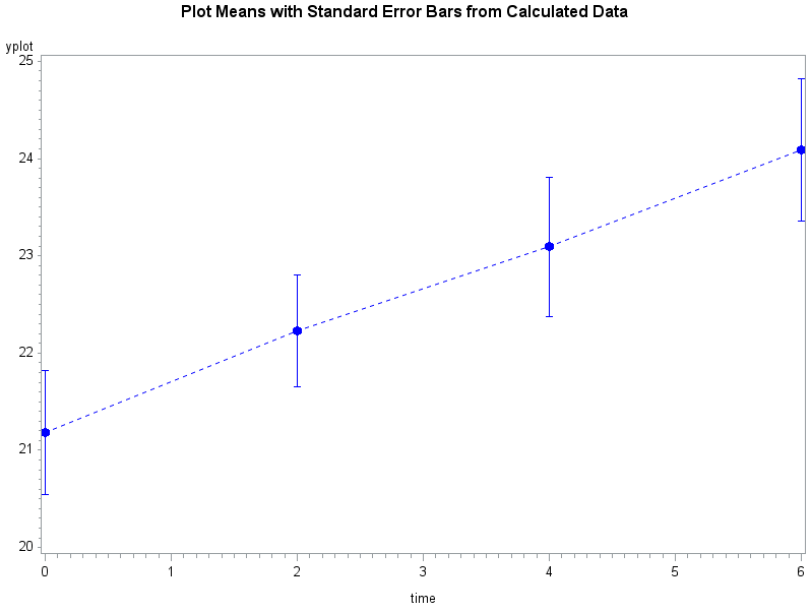
1.



Due to relatively small sample size in this example, the time plot of the raw data, with joined line segments for successive repeated measurements on the same individual, reveals some discernible features. First, the mean response increases approximately linearly over time; Second, within-individual variability appears to be less than between-individual variability, which indicates much heterogeneity of underlying propensity to respond; Third, at baseline, i.e. Age 8, there is discernibly huge variation in the response values, that is, girls entered the study with quite different baseline values; Finally, the variations in each occasion are similar, no discernible increase or decrease over time.

2.

time	mean	stderr
0	21.1818	0.64057
2	22.2273	0.57352
4	23.0909	0.71293
6	24.0909	0.73490



The trajectory of mean response over time increases linearly over time, from 21.2 at baseline to 24.1 at the 4th measurement. And the standard deviation from the sample starts at 0.64, then decreases to 0.57, but afterwards increases to 0.73 at the 4th occasion. Again, although the standard errors from the sample are not exactly the same cross all occasions, they don't differentiate by much, which might be just due to sampling error.

3.

Covariance Matrix, DF = 10				
	Y_Age8	**Y_Age10**	**Y_Age12**	**Y_Age14**
Y_Age8	4.513636364	3.354545455	4.331818182	4.356818182
Y_Age10	3.354545455	3.618181818	4.027272727	4.077272727
Y_Age12	4.331818182	4.027272727	5.590909091	5.465909091
Y_Age14	4.356818182	4.077272727	5.465909091	5.940909091

Pearson Correlation Coefficients, N = 11				
	Y_Age8	**Y_Age10**	**Y_Age12**	**Y_Age14**
Y_Age8	1.00000	0.83009	0.86231	0.84136
Y_Age10	0.83009	1.00000	0.89542	0.87942
Y_Age12	0.86231	0.89542	1.00000	0.94841
Y_Age14	0.84136	0.87942	0.94841	1.00000

Observed patterns in the correlation:

- The correlations are positive
- The correlations between repeated measures never approach zero, even between the baseline measurements and the last measurements which are taken 6 years apart
- The correlations between repeated measures never approaches one, even between occasions taken 2 years apart

4.

Observed unexpected feature in the correlation structure;

- The correlations don't decrease with increasing time separation, mostly

For example, the correlations between measurements at Age 8 (baseline) and Age 12, Age 14, are all larger than more close occasion which Age 10.

This observation is not consistent with most practical experience in longitudinal studies in the biological and health sciences, which has led to an empirical observation that the correlations often decrease with increasing time separation. The rationale behind this particular case might be, from personal point of view, each girl has almost same rate of growth during age 8 to 14, but has different propensity for how long the teeth can growth(i.e., the distance), which is constant over time. The distance at age 8 can be regarded as an approx for this propensity, if a girl has large distance at age 8, that means she will finally have a large distance. Although each girl has different propensity of the distance, their rate of dental growth during age 8 and 14 is the same. This underlying between-individual heterogeneity of propensity and similar growth rate will lead to correlations not decreasing over time, but quite constant.

Part B: Repeated Measures Models

5.

a.

REML Output:

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

ML Output:

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

Question 3 Sample Estimates:

Covariance Matrix, DF = 10				
	Y_Age8	**Y_Age10**	**Y_Age12**	**Y_Age14**
Y_Age8	4.513636364	3.354545455	4.331818182	4.356818182
Y_Age10	3.354545455	3.618181818	4.027272727	4.077272727
Y_Age12	4.331818182	4.027272727	5.590909091	5.465909091
Y_Age14	4.356818182	4.077272727	5.465909091	5.940909091

The REML method gives the exactly same variance and covariance as the sample variance and covariance estimated in question 3.

b.

REML and ML methods give different variance and covariance estimation. REML should be preferred. ML underestimates variances, that is, the diagonal elements in the above ML estimated R Matrix is smaller than what they should be. The bias arises because ML estimate has not taken into account the fact that betas are also estimated from the data, while REML makes a correction or adjustment that leads to a less biased variance-covariance estimation.

c.

REML Output:

Solution for Fixed Effects						
Effect	**time**	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**
Intercept		24.0909	0.7349	10	32.78	<.0001
time	**0**	-2.9091	0.3978	10	-7.31	<.0001
time	**2**	-1.8636	0.3573	10	-5.22	0.0004
time	**4**	-1.0000	0.2335	10	-4.28	0.0016
time	**6**	0	.	.	.	.

ML Output:

Solution for Fixed Effects						
Effect	**time**	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**
Intercept		24.0909	0.7007	10	34.38	<.0001
time	**0**	-2.9091	0.3793	10	-7.67	<.0001
time	**2**	-1.8636	0.3407	10	-5.47	0.0003
time	**4**	-1.0000	0.2227	10	-4.49	0.0012
time	**6**	0	.	.	.	.

Sample Means:

time	**mean**	**stderr**
0	21.1818	0.64057
2	22.2273	0.57352
4	23.0909	0.71293
6	24.0909	0.73490

Observation: REML, ML models both give the exactly same means at each occasion as the sample means calculated earlier.

Because we are fitting a saturated model for both REML and ML models, although generally the estimated variance-covariance matrix will affect the beta's estimates, saturated models estimate the means on each occasion separately (from algebra, variance-covariance matrix cancel out in the estimation equation of beta's)and hence the estimated beta's in both models are exactly same, and they both gives the exactly same means as the sample means. Except that, the estimated variance-covariance matrices in REML and ML are different, they affect the standard errors of beta's and hence p-values. In summary, the saturated model with REML and ML estimation give the precise mean response as sample means, but with slightly different p-values/standard errors (they don't give qualitatively difference results).

6.

a.

Two statistical models are nested if the first model can be transformed into the second model by imposing constraints on the parameters of the first model. In this case, the linear trend model is nested within the saturated model in question 5.

$$Linear\ Trend\ Model: E(Y_{ij}) = \beta_0 + \beta_1 \times Time$$

$$Saturated\ Model: E(Y_{ij}) = b_0 + b_1 \times I\{Time = 0\} + b_2 \times I\{Time = 2\} + b_3 \times I\{Time = 4\}$$

If we put 4 constraints on the parameters of the saturated model and let means on the two models to be equal:

$$Occasion\ 0: b_0 + b_1 = \beta_0$$
$$Occasion\ 2: b_0 + b_2 = \beta_0 + 2\beta_1$$
$$Occasion\ 4: b_0 + b_3 = \beta_0 + 4\beta_1$$
$$Occasion\ 6: b_0 = \beta_0 + 6\beta_1$$

Now the two models have exactly same set of means on all four occasions. Since the two models are indeed only modeling these 4 occasions' mean response, they are the same now. Thus, by definition, the linear trend model is nested within the saturated model. In other words, the linear trend model is a special case of the saturated model, when the above conditions met.

b.

Linear Trend Model (with METHOD=ML)

Fit Statistics	
-2 Log Likelihood	130.6

Saturated Model (with METHOD=ML)

Fit Statistics	
-2 Log Likelihood	130.5

Since the linear trend model is nested within the saturated model, the adequacy of linear trend model cam be assessed by Likelihood Ratio Test. The likelihood ratio test statistics $G^2 = 2(l_{full} - l_{reduced}) = 130.6 - 130.5 = 0.1$, with 2 degree of freedom (p-value=0.95123>0.5). The maximum likelihood of linear trend model is almost the same as the saturated model, but with more parsimonious parameterization. Thus, when compare the linear trend model with the saturated model, the linear trend model is preferred.

c.

Why ML, not REML?

The extra term in REML log-likelihood depends on the regression model specification for the mean. As a result the REML log-likelihood for two nested models for the mean response are based on different transformations of the data (in order to obtain linear combinations of y_{ij} whose distributions don't depend on the mean model or beta's).In short, the REML likelihood for two nested models for the mean are based on two entirely different sets of transformed responses, making comparisons between the mean models meaningless.

Appendix (SAS Program):

```
*********************************
Applied Longitudinal Analysis
Xiner Zhou
2/7/2015
*******************************;

*import dataset;
data dental;
infile  'C:\data\Projects\APCD High Cost\Longitudinal\dental.txt';
input id gender $ Y_Age8 Y_Age10 Y_Age12 Y_Age14 ;
run;

* Subset only girls;
data dental_girl_w;
set dental;
where gender='F';
run;

* transpose from wide format to long format;
data dental_girl_l;
set dental_girl_w;
y=y_Age8;time=0;t=1;output;
y=y_Age10;time=2;t=2;output;
y=y_Age12;time=4;t=3;output;
y=y_Age14;time=6;t=4;output;
drop Y_Age8 Y_Age10 Y_Age12 Y_Age14 ;
run;

*Q1:Plot the observed trajectories of distance for the 11 girls (all on one plot);
proc gplot data=dental_girl_l;
        title 'Distance for 11 Girls at age 8,10,12,14';
        symbol interpol=join value=triangle;
     symbol2 interpol=join value=triangle;
        symbol3 interpol=join value=triangle;
        symbol4 interpol=join value=triangle;
        symbol5 interpol=join value=triangle;
        symbol6 interpol=join value=triangle;
        symbol7 interpol=join value=triangle;
        symbol8 interpol=join value=triangle;
        symbol9 interpol=join value=triangle;
        symbol10 interpol=join value=triangle;
        symbol11 interpol=join value=triangle;
        plot y*time=id;
run;

*Q2: Descriptive stat: Mean & Stderr, then plot mean with error bar over time;
proc sort data=dental_girl_l;by time;run;
proc means data=dental_girl_l noprint;
   by time;
   var y;
   output out=Q2(drop= freq_ _type_) mean=mean stderr=stderr;
        proc print noobs;
run;

* Set the graphics environment;
goptions reset=all cback=white border htext=10pt htitle=12pt;

*Reshape the data to contain three Y values for each occasion: Mean and std error bar;
data reshape;
   set Q2;
   yplot=mean; output;
   yplot=mean-stderr; output;
   yplot=mean + stderr; output;
run;
```

```
/* Plot the error bars using the HILOCTJ interpolation */

/*By specifying the interpol=hiloctj option in the symbol statement, SAS assumes there are three
Y values (i.e. high, low, close) for each X value. These values are required for SAS to draw a
vertical line connecting the high and low values, draw horizontal ticks to the ends (forming the
error bar) and join the close values (forming the mean plot). */
proc gplot data=reshape;
   title1 'Plot Means with Standard Error Bars from Calculated Data';
   symbol1 interpol=hiloctj color=blue line=2;
   symbol2 interpol=none color=blue value=dot height=1.5;
        plot yplot*time mean*time/ overlay ;
run;


*Q3:Obtain the variance-covariance and correlation matrices for the repeated measurements over
time;
proc corr data=dental_girl_w cov nosimple noprob;
var Y_Age8 Y_Age10 Y_Age12 Y_Age14;
run;

*Q5: Saturated Model, Method=REML/ML;
proc mixed data=dental_girl_l method=REML;
        class id time t;
        model y=time/solution chisq;
        repeated t/type=un subject=id R RCORR;
run;
proc mixed data=dental_girl_l method=ML;
        class id time t;
        model y=time/solution chisq;
        repeated t/type=un subject=id R RCORR;
run;

*Q6: Linear trend model;
proc mixed data=dental_girl_l method=ML;
        class id t;
        model y=time /solution chisq;
        repeated t/type=un subject=id R RCORR;
run;

*p value of LRT;
data pvalue;
pvalue=sdf('chisquare',0.1,2);
proc print;
run;
```