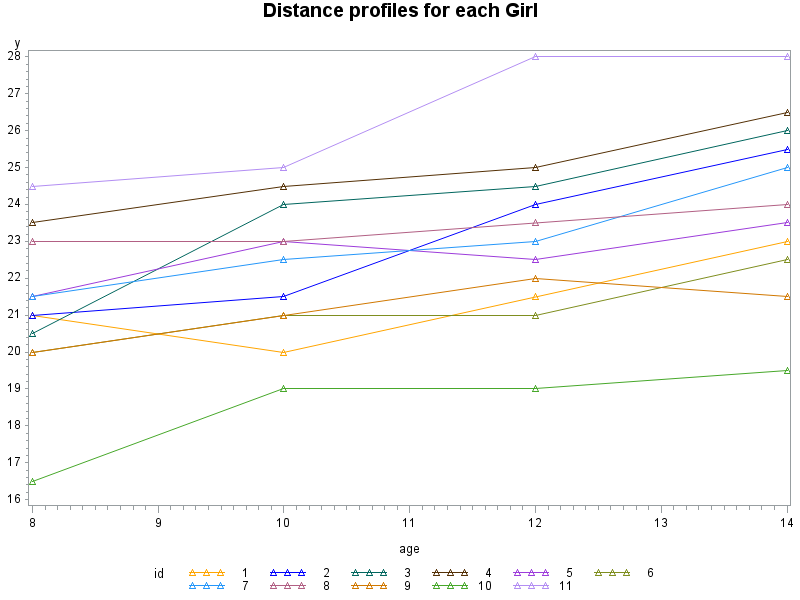
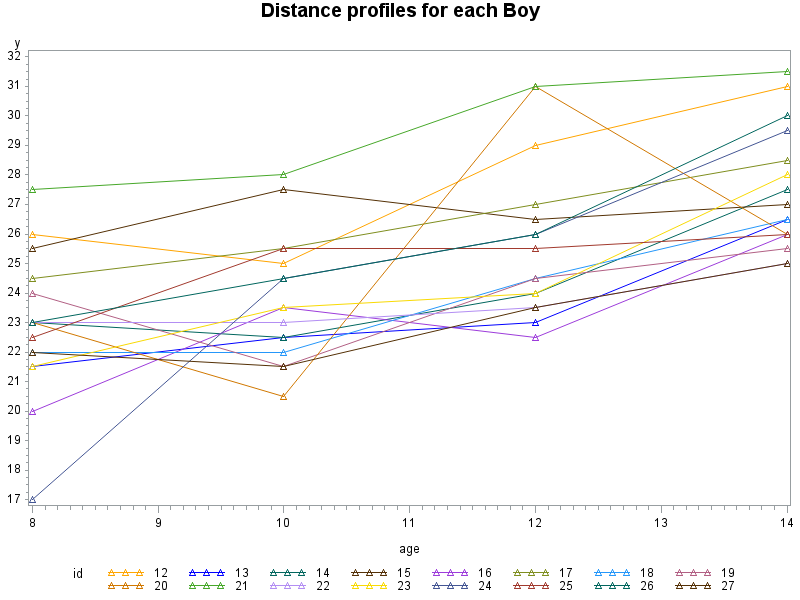
1. **Descriptive Analysis**

**a.Describe key features of the study’s design and the completeness of data.**

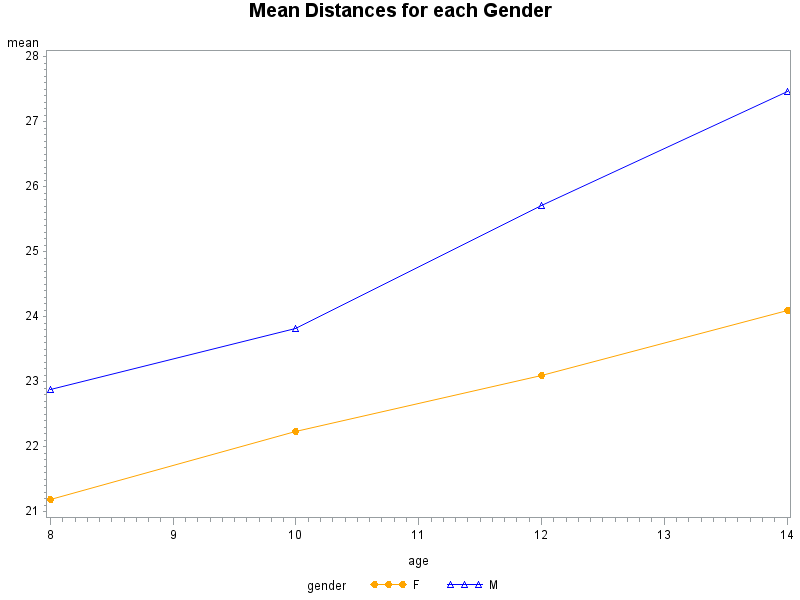
It’s an observational study with 2 cohorts (girls and boys). It’s a balanced study design over 4 equal-spaced measurement occasions and without missing data (or complete data).

**b.Create two plots, one for each gender, showing the trajectories of each subject’s distance over time. Comment on the notable features in these trajectories.**

****



The trajectories for girls and boys all go up over the study period, almost in a linear fashion. But for boys the rate of change seems larger than girls, on general. That is, most of girls grew at a modest increase rate without much variation between individuals; while for boys, most of them grew at a relative faster increase rate than girls and also with larger variation between individuals.

**c.Create a single plot of the distance over time for each gender. Comment on the notable features in this plot.**

|  |  |  |
| --- | --- | --- |
| gender | age | mean |
| F | 8 | 21.1818 |
| F | 10 | 22.2273 |
| F | 12 | 23.0909 |
| F | 14 | 24.0909 |
| M | 8 | 22.8750 |
| M | 10 | 23.8125 |
| M | 12 | 25.7188 |
| M | 14 | 27.46 |

At baseline, boys have higher mean distances than girls which would be expected in an observational study that the baseline values are different between groups. Throughout the study period, girls have lower distance levels than boys at all occasions.

For the rate of change, both groups increase in a linear fashion, if look at entire study period, boys have a little sharper increase rate than girls. If look at separate intervals, between age 8 and 10, boys and girls have similar rate of change; while between age 10 and 14, boys start to grow faster than before and girls remain the same as before, as a result, boys have higher rate of growth than girls between age 8 and 14, on average.

1. **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.State the model being fitted including any distributional assumptions. As much as possible, use the same notation as in the course notes.**

The distance for ith subject on jth occasion:

* The Fixed Part:
* The Random Part:

* Where is the time from baseline for ith subject on jth occasion (continuous); if ith subject is a girl, and 0 otherwise.
* is the subject-specific random intercept, is the subject-specific random slope.

Assuming ) , ) , and . is between-subject intercepts variation, is between-subject slopes variation, is between-subject co-variation of intercepts and slopes.

Or we can re-write as, ( where

* And assuming within-subject erroe . is the within-subject error variation.

**b. Compare the within-subject “error” variance to the between subject variance of the intercepts. What do you conclude?**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Estimated G Matrix | | | | |
| Row | **Effect** | **id** | **Col1** | **Col2** |
| 1 | Intercept | 1 | **3.2340** | -0.05886 |
| 2 | t | 1 | -0.05886 | 0.1301 |

|  |  |  |
| --- | --- | --- |
| Covariance Parameter Estimates | | |
| Cov Parm | **Subject** | **Estimate** |
| UN(1,1) | **id** | **3.2340** |
| UN(2,1) | **id** | -0.05886 |
| UN(2,2) | **id** | 0.1301 |
| Residual |  | **1.7162** |

The within-subject error variance is 1.7162, and the between-subject variance of the intercepts is

Thus, between-subject heterogeneity/variance of level of distance is about 1.88 times the within-subject measurement error variance.

The total variance at jth occasion is:

So the proportion of variation explained by between-subject variation at jth occasion is:

)

For example, at baseline, 35% of total variation can be explained by within-subject error, and the rest 65% of total variation at baseline can be explained by between-subject variation of level of distance.

We can conclude that the majority of total variation is due to between-subject variability of natural propensity to response (here is distance).

**c. 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** | **Pr > |t|** | **Lower** | **Upper** |
| Intercept |  | 22.6156 | 0.5265 | <.0001 | 21.5313 | 23.7000 |
| t |  | 1.5687 | 0.1720 | <.0001 | 1.2145 | 1.9230 |
| gender | **F** | -1.4065 | 0.8249 | 0.0939 | -3.0603 | 0.2472 |
| t\*gender | **F** | -0.6097 | 0.2695 | 0.0277 | -1.1499 | -0.06940 |

The estimate for the difference in mean intercept is -1.4065 (SE=0.8249, p-value=0.0939). We can conclude that, at baseline (Age=8), on average, girls have 1.4065mm smaller distance than that of boys, but it is not statistical significant, and there is a 95% chance that the true difference of mean distance at baseline would be between -3.0603 and 0.2472.

The estimate for the difference in mean slope is -0.6097 (SE=0.2695, p-value=0.0277). We can conclude that, on average, the rate of increase/change for girls is 0.6097 smaller (statistical significant) than that of boys (which would be expected from looking at the mean profiles plot), and there is a 95% chance that the true difference of rate of change would be between -1.1499 and -0.06940.

**d. Obtain and provide an interpretation for the correlation of the intercept and slope among subjects.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Estimated G Correlation Matrix | | | | |
| Row | **Effect** | **id** | **Col1** | **Col2** |
| 1 | Intercept | 1 | 1.0000 | **-0.09075** |
| 2 | t | 1 | **-0.09075** | 1.0000 |

The correlation of the random intercepts and slopes among subjects is -0.09075.

It means subjects(both girls and boys) who start at a higher baseline value than his/her group mean baseline value(boys and girls have different group mean baseline value), would be expected to have lower rate of change over time than his/her group mean rate of change/increase (boys and girls have different group mean rate of change/increase; subjects who start at a lower baseline value than his/her group mean baseline value would be expected to have larger rate of change over time than his/her group mean rate of change.

This can be explained by the phenomenon “regression to/towards the mean”. Subjects start with extreme values at 1st observation will tend to be closer to average. In this example, subjects with higher deviation from group average at baseline will tend to have smaller rate of increase than group average, and subjects with lower deviation from group average at baseline will tend to have larger rate of increase than group average.

**e. From your model, conditional upon gender, obtain the standard deviation of the slopes among subjects. Provide a clinically meaningful interpretation of this standard deviation.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Estimated G Matrix | | | | |
| Row | **Effect** | **id** | **Col1** | **Col2** |
| 1 | Intercept | 1 | 3.2340 | -0.05886 |
| 2 | t | 1 | -0.05886 | **0.1301** |

Conditional on gender, the standard deviation of the slopes among subjects is sqrt(0.1301)=0.3606938,

indicating that there is some variability from girl to girl (or from boy to boy) in rate of change/distance increase. For example, approximately 95% of girls have rate of change between 0.2520402 and 1.66596 (i.e., 1.5687-0.6097+/-1.96\* 0.3606938); and approximately 95% of boys have rate of change between 0.8617402 and 2.27566 (i.e., 1.5687+/-1.96\* 0.3606938).

**f. Obtain the marginal correlation matrix for the vector of responses, Yi =(Yi0, Yi1, Yi2, Yi3), for the subject with ID=1. Considering your findings in response to question 2b, why are the correlations relatively large?**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 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 |

From question 2b we know that, between-subject heterogeneity/variance of level of distance is about 1.88 times the within-subject measurement error variance and concluded that the majority of total variation is due to between-subject variability of natural propensity to response (here is distance). The natural propensity to response means, there’re some individuals who are always higher than average, some who are always lower than average. Whether it be “high”, “medium”, or “low”, this natural propensity to response is shared by all of the repeated measures obtained on the same subject, which leads to correlation among repeated measures. Here, the between-subject heterogeneity/variance of natural propensity(e.g., level of distance mm) is relatively large compared to within-subject measurement error, as such, leads to the marginal correlations relatively large (but not for conditional correlations which is zero’s).

**g. 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?**

To choose among random effects covariance models: Random Intercept Model versus Random Intercept & Slope Model, we can use (1) Likelihood Ratio Test using REML (since random intercept model is nested within the random intercept and slope model) (2) AIC using REML

1. Likelihood Ratio Test (REML)

i.e., the slope of time is not random

Table: LRT for random effects covariance models

|  |  |
| --- | --- |
| Model | -2 Maximized REML log-likelihood |
| Random Intercept Model | 431.0 |
| Random Intercept & Slope Model | 429.8 |
| Difference | 1.2 |

Since the null hypothesis is on the boundary of the parameter space (e.g., the variance of a random effect equal to zero), the distribution of likelihood ratio test is a 50:50 mixture of chi-squared distributions with 1 and 2 degrees of freedom. The critical value from Appendix C of FLW at 0.05 significance level is 5.14.

LRT yields <5.14, so we can’t reject the null hypothesis and conclude that the model with only random intercept is an adequate fit to the data compared to the random intercept and slope model.

1. AIC (REML)

|  |  |
| --- | --- |
| Model | AIC |
| Random Intercept Model | 435.0 |
| Random Intercept & Slope Model | 437.8 |

AIC also indicates that the random intercept model is preferred.

**h. The data for this study could also be analyzed using the repeated measures approach discussed extensively earlier in the course.**

**i) Define a repeated measures model with an unstructured variance-covariance matrix and the same form of fixed effects model as in question 2a which could be used.**

The distance for ith subject on jth occasion:

* Where is the time from baseline for ith subject on jth occasion (continuous); if ith subject is a girl, and 0 otherwise.
* Assuming , and is unstructured variance-covariance matrix

**ii) 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?**

Since we have concluded that the random intercept model is an adequate fit to the data compared to the random intercept and slope model, we now compare the random intercept covariance model to the repeated measures with unstructured variance-covariance model. The two covariance models have the same model for the mean, we can compare AIC’s from REML estimate.

|  |  |
| --- | --- |
| Model | AIC(REML) |
| Unstructured covariate pattern Model | 441.8 |
| Random Intercept Model | 435.0 |

AIC indicates that the random intercept model is preferred. The random intercept model induces a compound symmetry variance-covariance structure, thus the compound symmetry structure might be a better fit to the data than the repeated measures model with unstructured variance-covariance matrix.

**iii) 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.**

Since these two models are not nested within each other, we can only use AIC to compare them. Since they have the same number of parameters (4 mean parameters+2 covariance parameters), comparing AIC is the same as comparing REML log-likelihood:

|  |  |
| --- | --- |
| Model | AIC(REML) |
| Compound Symmetry covariate pattern Model | 435.0 |
| Random Intercept Model | 435.0 |

|  |  |
| --- | --- |
| Model | -2 Maximized REML log-likelihood |
| Compound Symmetry covariate pattern Model | 431.0 |
| Random Intercept Model | 431.0 |

They yield the exactly same AIC=435.0 and -2 Maximized REML log-likelihood=431.0. So the two models fit the data equivalently well. The reason behind this similarity is, although the random intercept model has different modeling approach compared to the compound symmetry covariance pattern model, marginally (in terms of population average), these two models have the same model for the mean response and the same variance-covariance structure (compound symmetry), thus any estimation and inference about the population(marginally) are the same for the two models.

We can see that their parameter estimation and inference for the marginal mean and marginal covariance are exactly the same:

**Random Intercept Model:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Solution for Fixed Effects | | | | | |
| Effect | **gender** | **Estimate** | **Standard Error** | **t Value** | **Pr > |t|** |
| Intercept |  | 22.6156 | 0.5388 | 41.98 | <.0001 |
| t |  | 1.5687 | 0.1550 | 10.12 | <.0001 |
| gender | **F** | -1.4065 | 0.8441 | -1.67 | 0.0996 |
| t\*gender | **F** | -0.6097 | 0.2428 | -2.51 | 0.0141 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Estimated V Matrix for id 1 | | | | |
| Row | **Col1** | **Col2** | **Col3** | **Col4** |
| 1 | 5.2207 | 3.2986 | 3.2986 | 3.2986 |
| 2 | 3.2986 | 5.2207 | 3.2986 | 3.2986 |
| 3 | 3.2986 | 3.2986 | 5.2207 | 3.2986 |
| 4 | 3.2986 | 3.2986 | 3.2986 | 5.2207 |

**Compound Symmetry covariate pattern Model**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Solution for Fixed Effects | | | | | |
| Effect | **gender** | **Estimate** | **Standard Error** | **t Value** | **Pr > |t|** |
| Intercept |  | 22.6156 | 0.5388 | 41.98 | <.0001 |
| t |  | 1.5688 | 0.1550 | 10.12 | <.0001 |
| gender | **F** | -1.4065 | 0.8441 | -1.67 | 0.1081 |
| t\*gender | **F** | -0.6097 | 0.2428 | -2.51 | 0.0141 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Estimated R Matrix for id 1 | | | | |
| Row | **Col1** | **Col2** | **Col3** | **Col4** |
| 1 | 5.2207 | 3.2986 | 3.2986 | 3.2986 |
| 2 | 3.2986 | 5.2207 | 3.2986 | 3.2986 |
| 3 | 3.2986 | 3.2986 | 5.2207 | 3.2986 |
| 4 | 3.2986 | 3.2986 | 3.2986 | 5.2207 |

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Appendix : SAS Codes

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

\*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 ;

**proc** **print**;**run**;

\* transpose from wide format to long format;

**data** ldental;

set dental;

y=y\_Age8;age=**8**;t=**0**;tc=**0**;output;

y=y\_Age10;age=**10**;t=**1**;tc=**1**;output;

y=y\_Age12;age=**12**;t=**2**;tc=**2**;output;

y=y\_Age14;age=**14**;t=**3**;tc=**3**;output;

drop Y\_Age8 Y\_Age10 Y\_Age12 Y\_Age14 ;

**proc** **print**;**run**;

\*Q1b: Plot each subject's response trajectories;

**proc** **gplot** data=ldental;

symbol1 interpol=join value=triangle;

symbol2 interpol=join value=triangle;

symbol3 interpol=join value=triangle;

symbol4 interpol=join value=triangle;

symbol5 interpol=join value=triangle;

plot y\*age=id;

where gender='F';

title 'Distance profiles for each Girl';

**run**;

**proc** **gplot** data=ldental;

symbol1 interpol=join value=triangle;

symbol2 interpol=join value=triangle;

symbol3 interpol=join value=triangle;

symbol4 interpol=join value=triangle;

symbol5 interpol=join value=triangle;

plot y\*age=id;

where gender='M';

title 'Distance profiles for each Boy';

**run**;

\*Q1c: mean response profiles;

**proc** **means** data=ldental n mean std nway;

var y;

class gender age;

output out=meandata mean=mean;

**proc** **print** data=meandata;

**run**;

**proc** **gplot** data=meandata;

symbol1 color=orange interpol=join value=dot;

symbol2 color=blue interpol=join value=triangle;

plot mean\*age=gender;

title 'Mean Distances for each Gender';

**run**;

\*Q2: Random intercept and slope;

**proc** **mixed** data=ldental ;

class id gender;

model y=t gender t\*gender/s chisq cl;

random intercept t/type=un subject=id G Gcorr vcorr=**1**;

**run**;

\*Random intercept;

**proc** **mixed** data=ldental ;

class id gender;

model y=t gender t\*gender/s chisq cl;

random intercept /type=un subject=id G Gcorr v=**1** vcorr=**1**;

**run**;

\*Unstructured covariate pattern Model;

**proc** **mixed** data=ldental;

class id gender tc;

model y=t gender t\*gender/s chisq cl;

repeated tc/type=un subject=id r rcorr;

**run**;

\*Compound Symmetry covariate pattern Model;

**proc** **mixed** data=ldental;

class id gender tc;

model y=t gender t\*gender/s chisq cl;

repeated tc/type=cs subject=id r rcorr;

**run**;