2/21/2005

LDA midterm

1) - Cross-sectional outcome measure: Total number of participants who drops out of the drug treatment program at the end of the trial in comparison to the usual the social work arm.

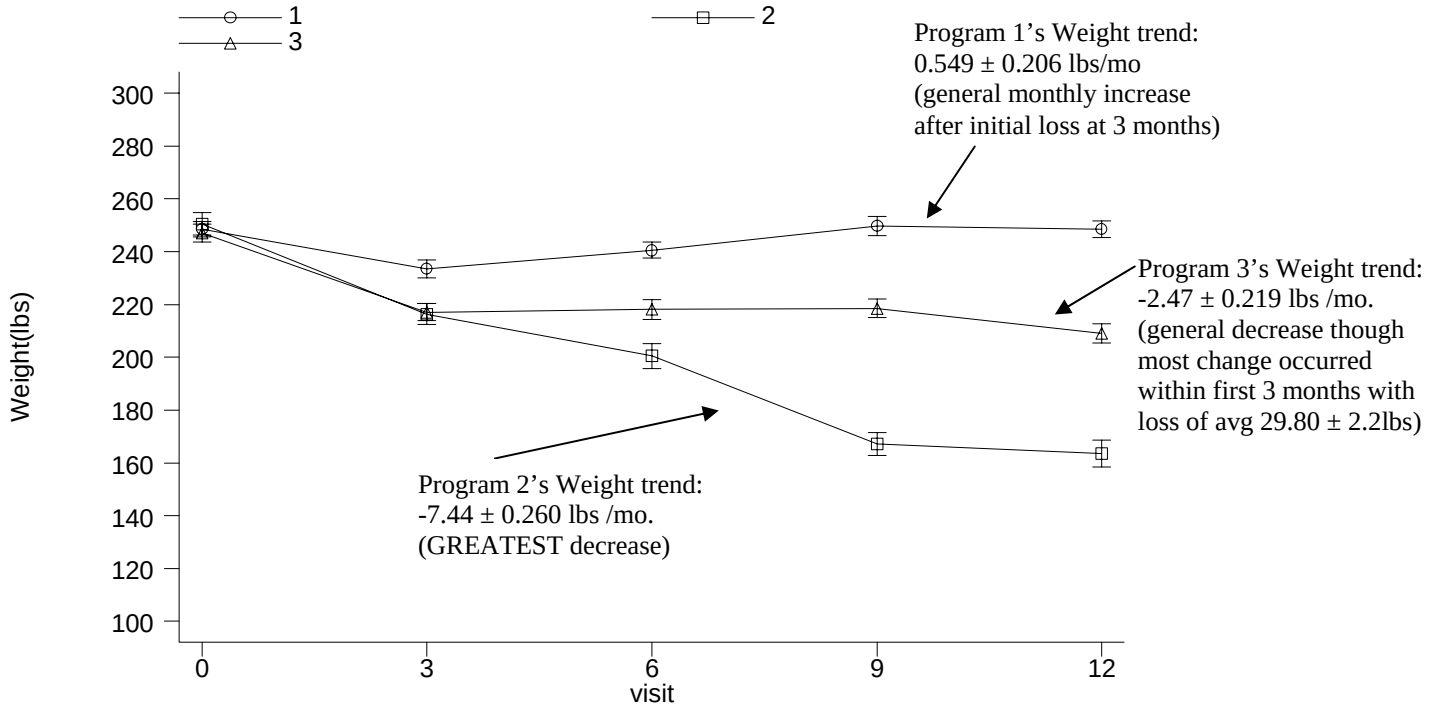
- Longitudinal outcome measure: Number of drug uses per week over the next 5 months, looking at the trends in drug taking behavior in the 2 arms of intervention.

2) To predict βc which is the cross-sectional estimate of the difference in Y where x differs by 1 unit, one can regress y_i with x_1 where x_1 are the measurements of x at baseline. For β_L one needs to include the repeated measurements of x to regress y with x_{ij} for the estimate of the change in y per unit change in x in each individual over time (since $x_{ij} - x_{i1}$ = measurement at j visit minus baseline measurement).

3) Data set – weightloss.raw –

- Goal: To explore the effectiveness of 3 weightloss programs in a total of 100 individuals followed over a year. 34 individuals were on program 1, 28 individuals were on program 2, and 38 individuals were on program 3. Individuals were weighed 5 times, once at the start of the program and then again every 3 months until the end of the year.

- Display of the evidence:

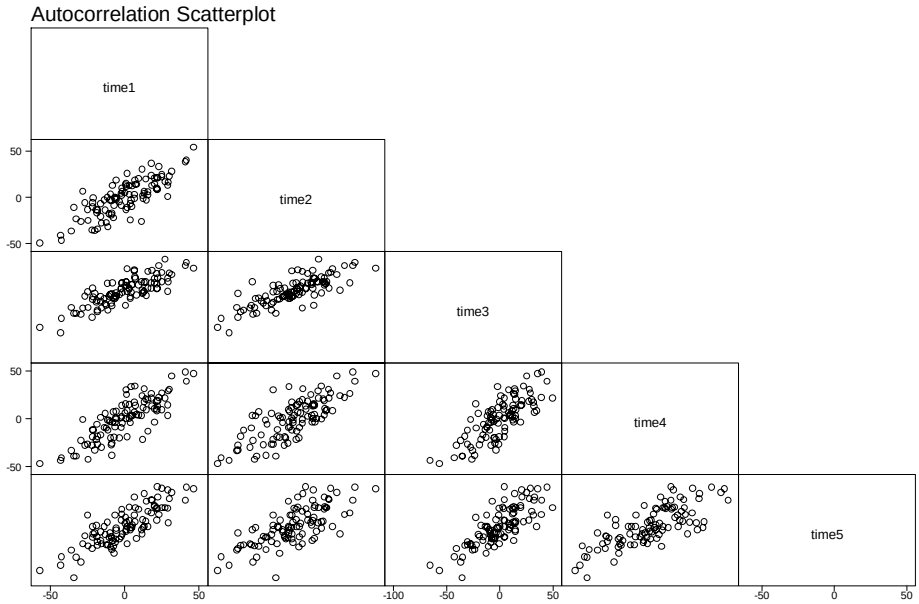


Mean Weight Loss Among 3 Programs Over 1 Year

- Summary of evidence from display:

This display (from xtgraph with some xtreg, re results) summarizes the evidence of the differences in effectiveness between the 3 weight loss programs best. There is clear separation of the weight trend at 3 months for program 1 vs 2 and 3. There are clear differences also at 6 months for program 2 vs 3. Program 2 seems the most effective of the 3, having a relatively linear decrease in mean weights over the whole year. Program 3 is the next best but all the weight loss occurred in the first 3 months. The mean weight for the program 1 individuals decreased slightly after the first 3 months but then they regained their weight (and perhaps gained even more weight than they started) so that the difference from beginning of the year to the end remained relatively unchanged in the 240-260 lbs range. The groups all started at relatively the same mean weight (~250lb). Of note is that program 2 had fewer people in it (n=28) vs program 1 (n=34) and program 3 (n=38). This is somewhat shown in the variance around the estimates (se bars in the display) which are somewhat wider for the program 2 group. However, even taking that into account, the individuals in program 2 lost the most weight over the whole year, about (250-170=) 80 lbs mean loss. Program 3 ended at about a 40lbs mean decrease from the start and the individuals, though they did not lose more weight, were able (unlike program 1 individuals) to maintain their weight loss.

- A scatterplot matrix was created with the residuals after account for the full model (including interaction terms as suspected from the display above). Weight was regressed on visit (treated nominally because of suspected nonlinearity), program, and interaction of i.visit*i.program.



The variance (as already mentioned before) seems to be constant through time and so too for the mean. The correlation between the measurements then is only dependent on the lag in time (tij-tik). This is seen in this autocorrelation scatterplot which shows a uniform correlation structure. This means that there was no sudden, fast decays in the correlation, which is evidence for stationarity. (Per reference for how to use scatterplot matrices to check stationarity: http://www.itl.nist.gov/div898/handbook/pmc/section4/pmc446.htm; accessed 2/21/05) Hence, the **autocorrelation function** (with the residuals accounting for time and treatment and the interaction of the two) was estimated to be: 0.783 for 1st lag, 0.795 for 2nd lag, 0.774 (for 3rd lag), 0.828 (for 4th lag) where each lag spanned 3 months. Again the acf supports a uniform structure.

4) Dental Study:

a) Assuming we didn't know that the data were of repeated measurements:

Model 1 and Parameters for Type 0: $y_i = ß_{00} + ß_{01} + \varepsilon_j$ j=1…44 $var(y_i) = \sigma_0^2$

Model 2 and parameters for Type 1: $y_i = ß_{10} + ß_{11} + \varepsilon_j$ j=1…64 $var(y_i) = \sigma_1^2$

	Model 1 for Type 0				Model 2 for Type 1			
Parameters	$ß_{00}$	SE $ß_{00}$	$ß_{01}$	SE $ß_{01}$	$ß_{10}$	SE $ß_{10}$	$ß_{11}$	SE $ß_{11}$
Estimates	16.34	1.45	0.7844	0.1296	17.37	1.64	0.4795	0.1459

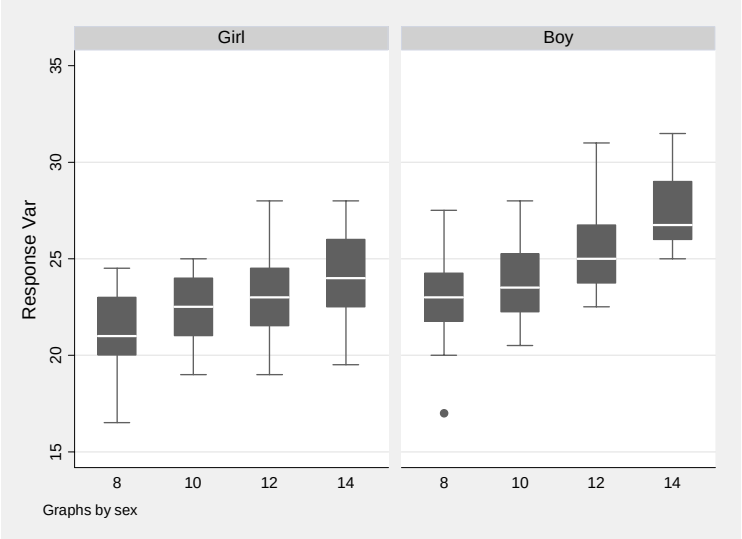
Under OLS, the parameters tells us that the mean Y (distance) for a Type 0, age 0 person is 17.37 and for each increase in X (age) there is a 0.48 in Y. Those with Type 1 were slightly different, starting with a mean of 16.34 at age 0 and increasing at a rate of 0.78 with each increase in age.

b) Assuming we KNOW that the data is of repeated measurements:

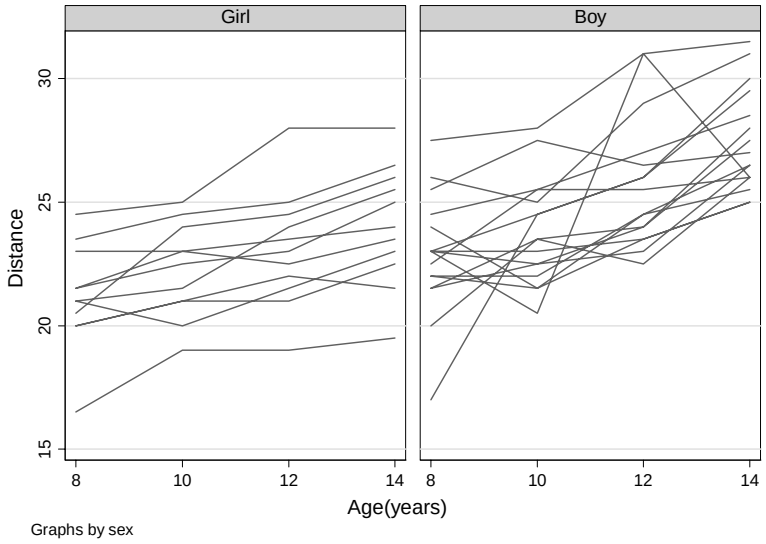
Model 1 for Boys: $y_i = ß_{0B} + ß_{1B}t_{ij} + \varepsilon_j$ w/ i = 1,2,…,16; j=1,2,3,4

Model 2 for Girls: $y_i = ß_{0G} + ß_{1G}t_{ij} + \varepsilon_j$ w/ i = 1,2,…16; j=1,2,3,4

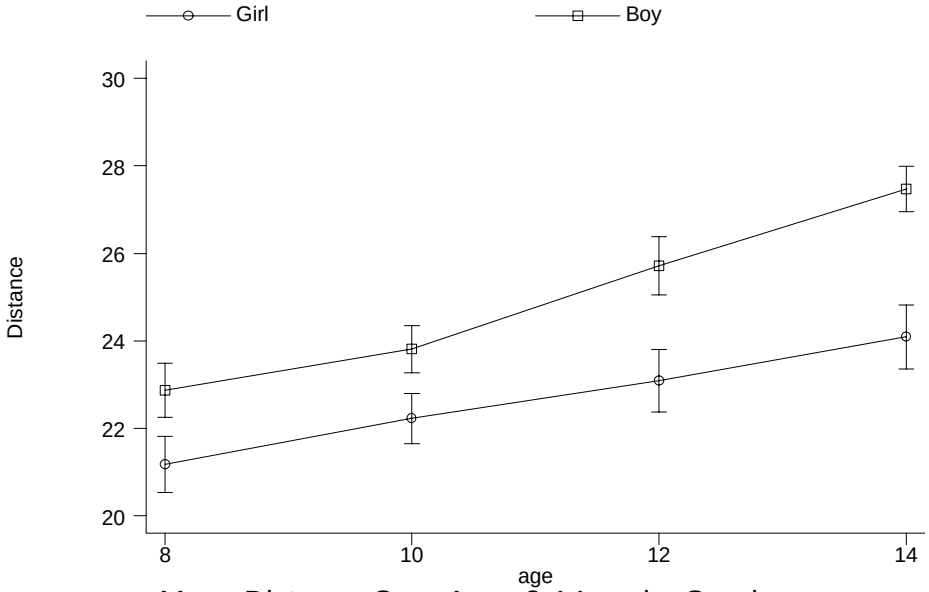
- Plot of the data:



Boxplots show that the distance variable is fairly normally distributed. None of the boxplots look greatly skewed except for maybe the distribution among the distance for the 14 year old boys. The variance around the estimates also seem fairly constant for all 4 measurements.



The spaghetti plot (above) shows that each child is increasing in their response distance (between the pituitary to the pteryomaxillary fissure) as they age. In terms of tracking, it seems as though the girls stayed in their relative order by distance more so than the boys. The boys had more cross-overs where a boy with a low distance compared to the others at age 8 suddenly became a boy with a very great distance compared to others. The sphagetti plot also shows that the variance was fairly constant over time and the spread of the data remained the same. The boys seem to be starting at at a greater distance than the girls and growing slightly faster (greater slope) than the girls which may mean that there's an interaction between gender and time. (More on picking a model in the summary.) This is confirmed in the mean plot of the growth patterns by gender below:



Mean Distance Over Ages 8-14 yrs by Gender

- Fitting OLS:

Ordinary least squares gives the same estimates as above when we introduced the dataset as 108 pairs of singular measurements instead of repeated measurements.

	For Boys				For Girls			
Models	ß0B	SE ß0B	$ß_{1B}t_{ij}$	SE $ß_{1B}t_{ij}$	$ß_{0G}$	SE $ß_{0G}$	$ß_{1G}t_{ij}$	SE $ß_{1G}t_{ij}$
OLS	16.34	1.45	0.7844	0.1295657	17.37	1.64	0.4795	0.1459028

The reason for this is because we are looking at the variable time as the covariate and see if there are differences by sex even in accounting for the non-repeated measurements. Ordinary least squares assumes no correlation between the measurements (taking them as independent) so the standard errors are incorrect here but are hence similar to the ones from the first part of this problem as taking them to be 108 independent pairs.

- GLS with different covariance matrices - The following summarizes the different parameters of the various models fit to the data with stratification by gender to see differences between boys and girls and their changes in distance over time (age). *Note: the coefficients for the constants are centered at age 8 (gen age8 = age-8) to make the data more interpretable without having a constant being an extropolation of what happens at age 0.* Below are also the commands used in STATA to run the models for the estimates for the girls.

		For Boys				For Girls			
	Models	ß0B	SE ß0B	$ß_{1B}t_{ij}$	SE $ß_{1B}t_{ij}$	$ß_{0G}$	SE $ß_{0G}$	$ß_{1G}t_{ij}$	SE $ß_{1G}t_{ij}$
	OLS	22.62	0.4848	0.7844	0.1295657	21.21	0.5459	0.4795	0.1459028
WLS	Independent correlation	22.62	0.4772	0.7844	0.1275252	21.21	0.5334	0.4795	0.1425483
WLS	uniform corr	22.62	0.5230	0.7844	0.092833	21.21	0.6247	0.4795	0.0517869
WLS	uniform w/ random effects	22.62	0.5369	0.7844	0.0938154	21.21	0.6540	0.4795	0.0525898
WLS	exponential	22.75	0.6287	0.7694	0.1316101	21.19	0.4952	0.4841	0.0963581
WLS	MLE (uniform)	22.62	0.5230	0.7844	0.0928289	21.21	0.6247	0.4795	0.0517866
	GEE Robust	22.62	0.5511	0.7844	0.1015729	21.21	0.5878	0.4795	0.066214
	MLE/Unstruc	22.66	0.5166	0.7788	0.1004584	21.24	0.5869	0.4702	0.0703936

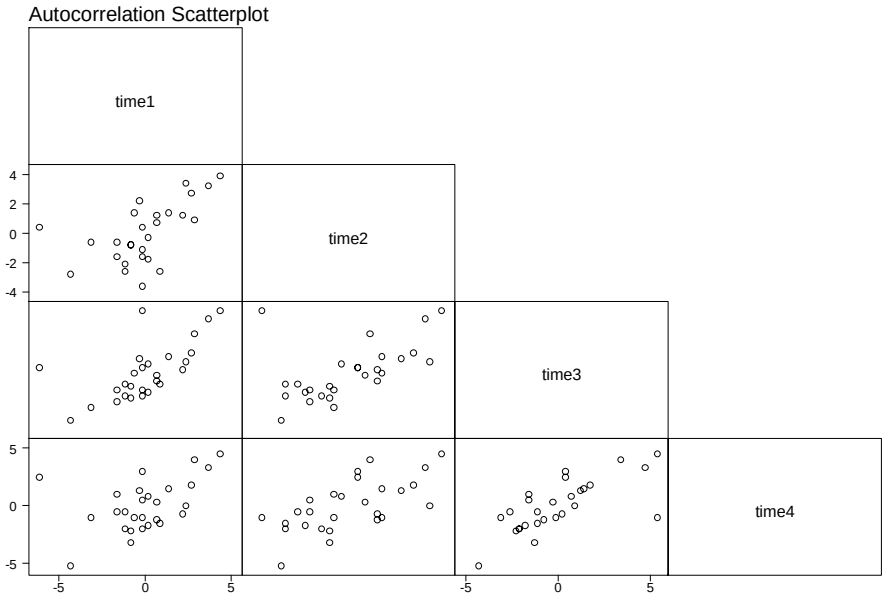
Corresponding STATA commands for the girl models: (boy model was same except w/ “sex==1”)

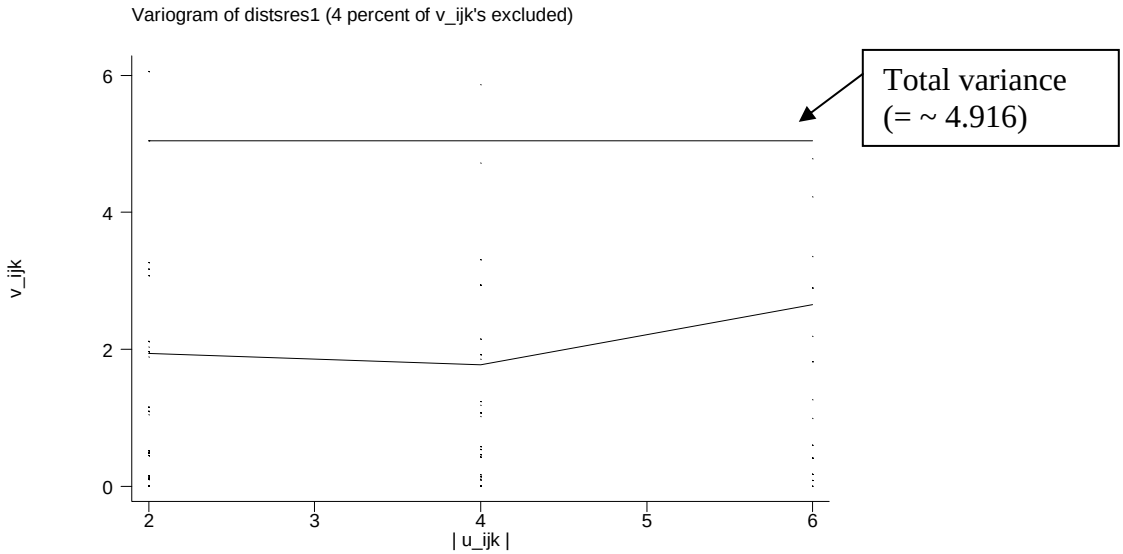
Models	**STATA Commands For Girls**
OLS	regress dist age if sex==0
Independent correlation	xtgls dist age if sex==0, i(id) corr(ind)
uniform corr	xtreg dist age if sex==0, i(id) pa
uniform w/ random effects	xtreg dist age if sex==0, re i(id)
exponential	xtgls dist age if sex==0, igls corr(ar1) i(id) force
WLS w/ MLE	xtreg dist age if sex==0, mle

GEE Robust	xtgee dist age if sex==0, robust
MLE/Unstruc	xtgee dist age if sex==0, i(id) corr(uns)

- Finding the appropriate correlation structure:

From the description of the data (xtdes), the data was found to be balanced (no one was missing any of the 4 measurements taken) and equally spaced (every one was measured at ages 8, 10, 12, and 14). Exploring the scatterplot of the residuals of the distance variable after adjusting for age and sex, one sees that correlation remains between each individual's measurements for all 3 lags in measurement. So distance measurements taken at age 14 are still correlated to those take at age 8. The variogram below confirms this and gives other information as well.





The variogram above was created with the residuals after adjusting distance for time (age), gender, and the interaction between time and gender (i.sex*age). It shows that the most appropriate correlation structure would be a uniform correlation structure (the line is almost straight across) with a random intercept (the line did not reach the total variance line). This total variance is about (2.22^2 by xtsumcorr) 4.916. The estimated between group (or within individuals) variance to total variance (rho) was about 0.628. 0.63 is fair amount of correlation that would make the estimates less efficient by producing incorrect SEs. The variogram also shows that there is measurement error in this model since it does not start at zero. The interaction model was used to best account for any differences in gender but residuals adjusting for just sex and age with no interaction term produced similar variograms and correlations (by xtsumcorr).

- Fitting the WLS with MLE on the uniform correlation model:

This was already explored above but to repeat here:

	For Boys				For Girls			
Models	ß0B	SE ß0B	$ß_{1B}t_{ij}$	SE $ß_{1B}t_{ij}$	$ß_{0G}$	SE $ß_{0G}$	$ß_{1G}t_{ij}$	SE $ß_{1G}t_{ij}$
WLS w/ MLE	22.62	0.5230	0.7844	0.0928	21.21	0.6247	0.4795	0.0518

- Comparing the OLS and WLS estimates. Referring back to the chart above the STATA commands with all the GLS models (and a few GEE models even though those won't be compared here, more in the summary).

The estimates of the coefficients from the different models were similar. All the p-values given for the age variable were 0.001 or less even though the variances differed slightly, so ultimately the inferences from these estimates would not differ no matter the model. Among the girls, each year increase in age corresponded to 0.4795 units ($ß_{1G}t_{ij}$) of increase in distance from the pituitary to the pteryomaxillary fissure. On average, the girls at 8 years old measured about 21.21 in distance ($ß_{0G}$). Among the boys the increase was in 0.7844 units of distance per year and they started at a distance of 22.62 at 8 years old. Though this was stratified, there's evidence (more in summary) that the rates of growth in girls vs boys are not statistically significantly different from each other since their 95% confidence intervals (at 2SEs) overlap.

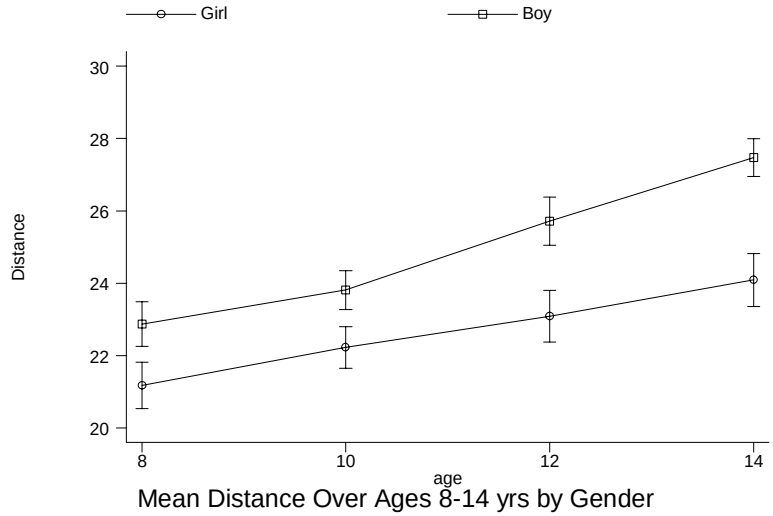
The efficiency of these estimates by the different modelling methods varied from the lowest with the MLE estimate of the uniform correlation structure for the girl models giving a standard error of ±0.52 to the greatest and least efficient OLS estimate of the standard error being 0.146. Since the variogram showed that the correlation structure should be relatively uniform and with random effects, it is not surprising that the exponential model does not fit. From the variogram, a parametric model of the correlation structure that would best fit was inferred to be one with a uniform correlation and a random intercept. The uniform model, the uniform model with random effects and the MLE model, all gave similar variances for the increase by age in distance (~0.052). MLE is best at estimating when the resources are available and since this is a small dataset with small number of covariates, this did not take the computer long to iterate. However, in terms of the parametric models, the uniform correlation model with random effects would fit the variogram best.

- Summary of Findings:

The Change in Distance from Pituitary to Pteryomaxillary Fissure
in Boys and Girls over Ages 8 to 14 years

Introduction:

Twenty seven children, 11 girls and 16 boys, were followed to observe the growth in distance from the pituitary to the pteryomaxillary fissure. Total of four measurements were made, once every two years, starting from age 8 and stopping at age 14. There were no losses to follow up or missed visits so the data were balanced. The goal of the study was to observe rate of growth of this distance over time as well as see if there were differences with respect to the distance and gender. A simple plot (Fig 1.1) of the mean distance over time by gender is very telling. The boys consistently had a greater mean distance than the girls. The slopes of the lines were fairly similar, though the boys may have a greater increase in distance over time. This possible interaction would be assessed in picking out a logical model for this study.



Mean Distance Over Ages 8-14 yrs by Gender

Fig 1.1: Mean Distance over Ages 8-14 in Boys and Girls

Girls and boys both increased in distance. Boys started with larger distances and remained so. Rate of growth were similar in both genders.

Methods:

In modelling the data to best answer the question about the growth rates between boys and girls and in general, a few variations on the models were tested to see if they were different from the simple model of having the distance adjusted for age and sex. To try to test whether the age response variable should be treated as a linear variable (age) or a nominal variable (i.age) allowing for non-linear increases in distance, the regression model with distance, i.age, and sex was tested against the same model but with ordinal age. The likelihood ratio test showed that these were not significantly different from each other ($p= 0.83$) so the age variable was left as linear variable. The plot above also shows that the increase were fairly linear. Next, to truly test for whether the slopes of the lines between the growth in distance for the boys were different from the girls, an interaction model was built including the terms age, sex, and i.sex*age. The likelihood ratio test again showed that this was not different ($p= 0.12$) from a model without the interaction variable meaning that the boys and the girls are growing at the same rate (i.e. slopes were not significantly different). Hence, the original model of distance adjusted with age (linear) and sex was kept and no stratification was needed since the interaction was not significant.

Next, a scatterplot of the data (Fig 1.2) and a variogram (Fig 1.3) was used to determine the correlation structure of the covariance matrix. These showed that a uniform correlation

structure with random effects would be best to account for the correlation. The estimates of the autocorrelation function supported the uniform correlation structure. They were 0.62 for measurements with lag of 2 years, 0.69 for lag of 4 years, and 0.51 for lag of 6 years. The variogram also gave some concern though that there may be some measurement error in the data. This is not surprising since it is probably hard to measure the distance from the "center" from the pituitary to the pteryomaxillary in children that are growing so quickly.

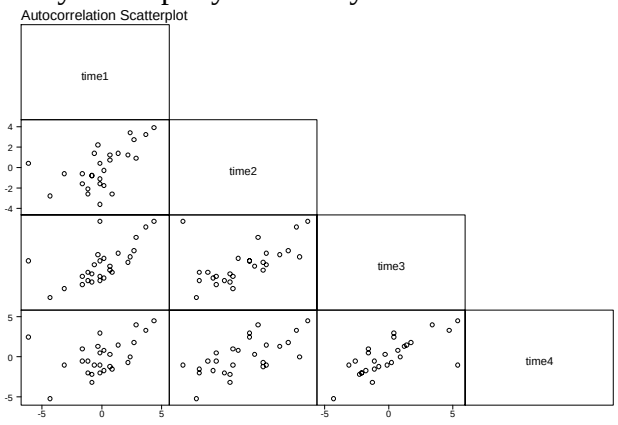


Figure 1.2 Scatterplot matrix of the distance residuals adjusted for age and sex shows that the data is stationary and that the correlation remained strong over longer lags in time between measurements.

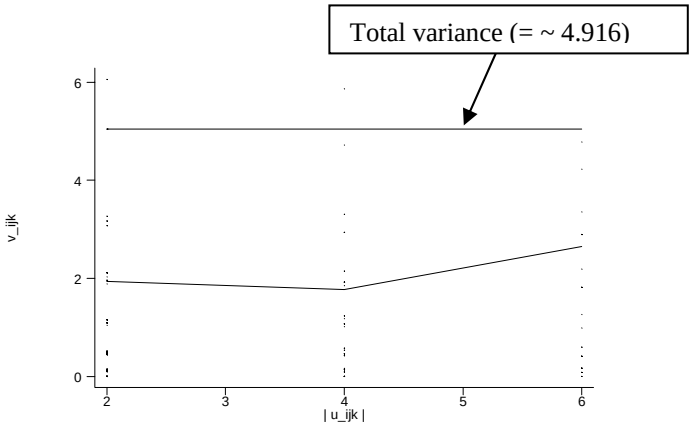


Figure 1.3 Variogram of the distance residuals adjusted for age and sex. Total variance of 4.916 was not reached meaning random effects at work. Uniform correlation structure from fairly straight line across despite time lags . Measurement error in data evidenced by starting of variogram at ~2.

Results:

Knowing that the model for the correlation structure should be uniform with random effects, a parametric model was first fit and the equation for the mean model: $E(y_i)= Ui + ß_0 + ß_1t_{ij} + ß_2t_{ij} + \varepsilon_j$ w/ i = 1,2,…,27; j=1,2,3,4 Where y_i is the distance, Ui is the random intercept (or the difference from the mean intercept) each child can have, $ß_0$ is for the distance centered at age 8, $ß_1$ is the rate of growth in distance over time adjusted for gender, and $ß_2$ is the difference in distance between boys and girls adjusted for age. Two other models were then also investigated to see if the efficiency in the estimates of the variance differed. These other two models were using an MLE estimator for the correlation and using a robust estimate of the correlation. The results below show that they did not differ in their point estimates and the estimates of the variance did not differ greatly. All estimates were significant at $p<0.001$, meaning there were significant growth of distance for all the children and that the distance in boys were slightly greater than girls.

	For all kids					
	Distance at 8 yrs of age	SE	Rate of growth Adjusted for Sex	SE	Difference in Distance btw Boys & Girls	
Models	ß0B	SE ß0B	ß1tij	SE ß1tij	ß2tij	SE ß2tij
uniform w/ random effects	20.667	0.615	0.660	0.062	2.321	0.761
MLE uniform	20.667	0.593	0.660	0.061	2.321	0.733
GEE Robust	20.667	0.623	0.660	0.071	2.321	0.764

. Though the efficiency for the estimates of the variance were probably not significantly different in these 3 models, the MLE estimates were probably best. For this dataset since the data was balanced and equally spaced, and with few covariates (only sex and age), the MLE model was most accurate in estimating even though this was more resource intensive. But it did not take much longer for the statistic program to iteratively estimate the correlation as compared to the other two methods of estimation.

Conclusion:

In summary, the data by the maximum likelihood estimation shows that there is a difference in the mean distance between boys and girls but no differences by gender in the rate of growth over time. Children grow about 0.66 (95% CI: 0.54 - 0.78) units of distance each year between the ages of 8 and 14. There were no differences in this rate of growth by gender. Boys had consistently longer distances than girls, approximately averaging 2.32 units (95% CI: 0.89 - 3.76) greater than girls. Future studies will be done to see if milk consumption and other covariates change these relationships!