LDA 140.665 Midterm

Due: 02/21/2005

Question 1:

The randomized intervention trial is designed to answer the scientific questions: whether social network method is effective in retaining drug users in treatment programs, thereby reducing the drug consumption. Therefore, one possible outcome that we can define for the analysis is *the amount of drug consumption each person consumes per week* (i.e. the amount of drug can be calculated as the number of pills times the dosage of drug). For each individual, the amount of drug consumption is measured at the beginning of the trial (week 0), and then is repeatedly measured every week after entering the trial (week 1, 2…n). This outcome is a *continuous* variable. The drug consumption across individuals at a particular time can be used for cross-sectional analysis. The repeated measurements of drug consumption over time can be used in the longitudinal data analysis.

Question 2:

To estimate β_c: $y_{i1} = \beta_c x_{i1} + \varepsilon_{i1}$, i=1…m

(i.e. regress baseline response (y_i at time 1) on baseline predictor (x_i at time 1) for each person i)

To estimate β_L: $y_{ij} - y_{i1} = \beta_L (x_{ij} - x_{i1}) + (\varepsilon_{ij} - \varepsilon_{i1})$, j=2…n

(i.e. regress the *difference* of response between time j and time 1 on the *difference* of predictor between time j and time 1 for each person i)

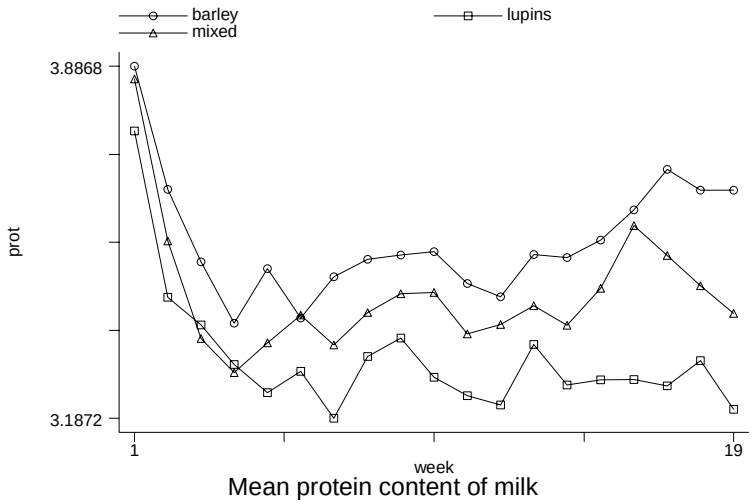
Question 3: Use the Cow dataset:

Scientific Question: we want to determine whether diet will affect the protein content in the milk

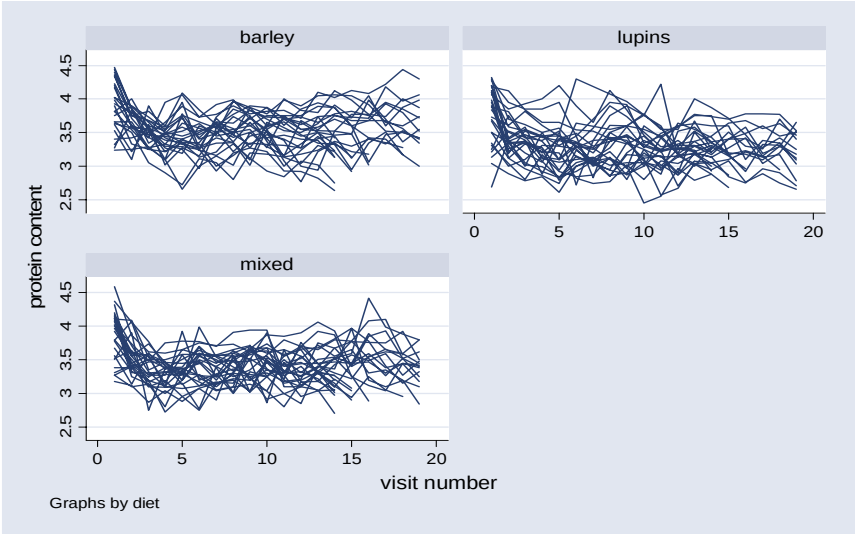
Study design: 79 cows were maintained on one of the 3 diets (barley, mixture of barley and lupins, or lupins alone). The milk was collected weekly (i.e. up to 19 weeks). The outcome of interest is the protein content in the milk. The time variable is weeks. There are 25 cows on barley, 27 on mixed, and 27 on lupins diet.

Compare the protein content of milk by the 3 diet groups:

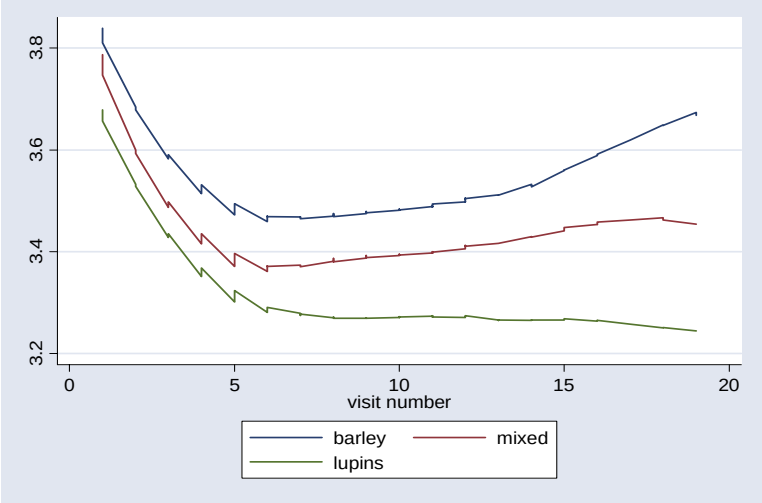
Mean Trend Plot (by diet type):



Spaghetti Plots (by diet type):



Lowess Smooth Curve (by diet type):



Overall average (i.e. mean of all data points) of protein content (by diet):

Diet (no. of observations)	Average of Protein Content
Barley (425)	3.531929
Mixed (459)	3.429695
Lupins (453)	3.312384

Summary: Overall, cows in the barley and mixed diet groups have higher protein content in milk than the cows in the lupins diet. When we further look at the time trend of the protein content change among the three groups, it appeared that the protein content in the three groups dropped at the first 5-6 weeks. However, after 5-6 weeks, the protein appeared to slightly increase in the barley and mixed diet groups, but remain to decrease in the lupin diet group over time. It is worth noticing that at the beginning of the study, the difference of mean protein content between the three groups differed only slightly, but as the time increased, the difference became greater.

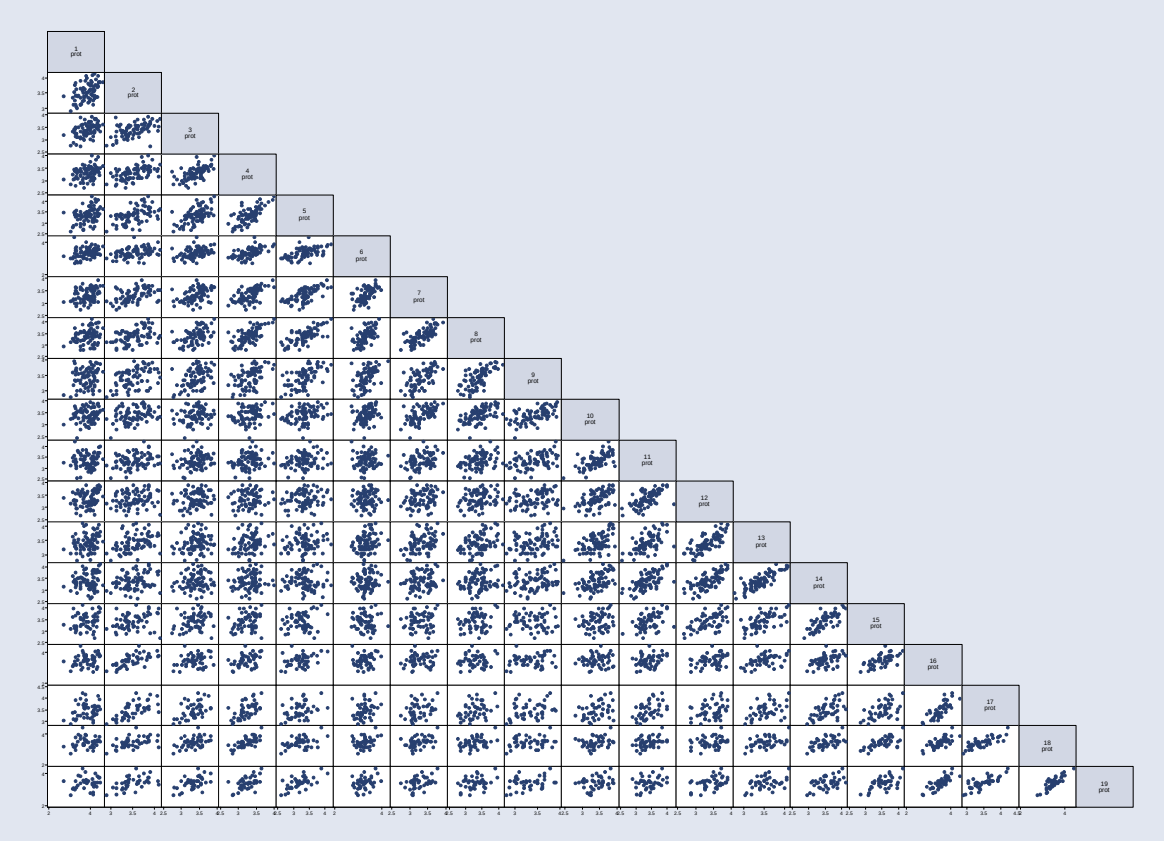
This suggested that the diet would affect the protein content in the milk.

Scatterplot Matrix:

The graph shows each of the 19 choose 2 scatterplots of responses from a person at different times.

(see next page)

```
graph matrix prot1-prot19, half
```

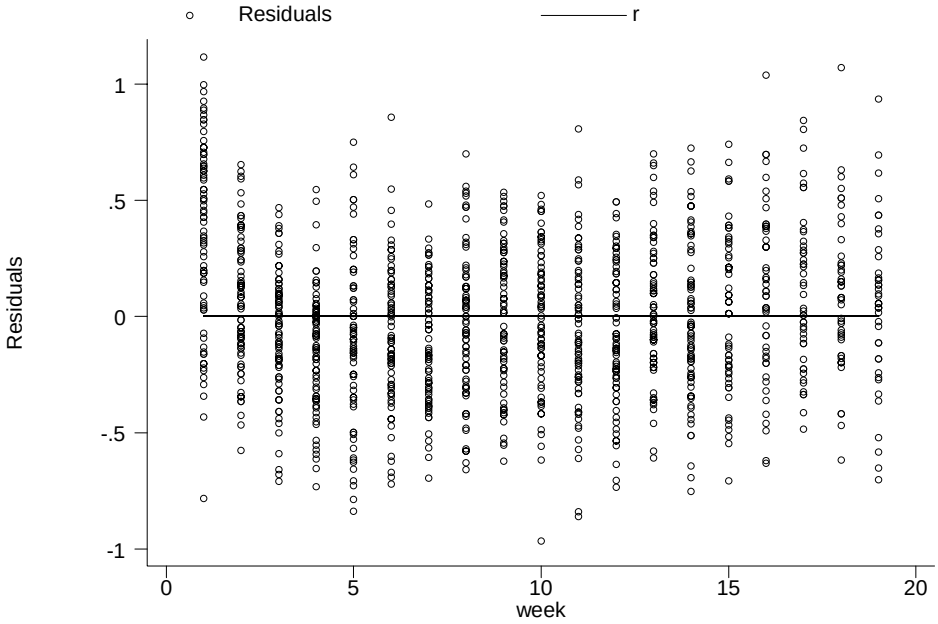


There is a positive liner trend for pair wise scatter plots that are one week apart, and the positive

association seems to be less obvious as the time lag increased.

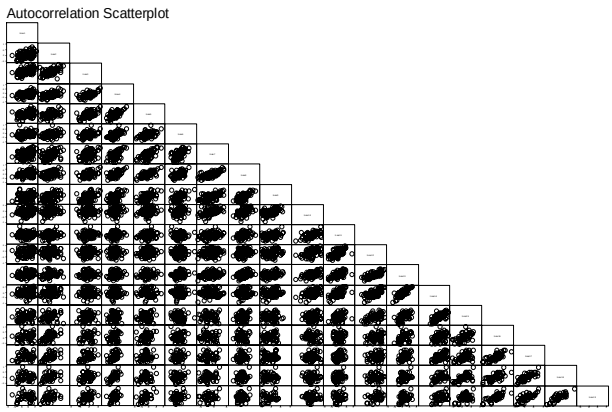
To assess the assumption of stationarity, we first regressed the protein content on the time variable (i.e. week) to get the residuals. If the data is stationary, then we would expect that the residuals of protein content after adjusting for the time variable (i.e. week) remain constant. Based on the graph below (residuals vs. week), we can see that the variance seems to remain constant and does not fluctuate much, except at the week 1. Therefore, the assumption of stationarity is reasonable.

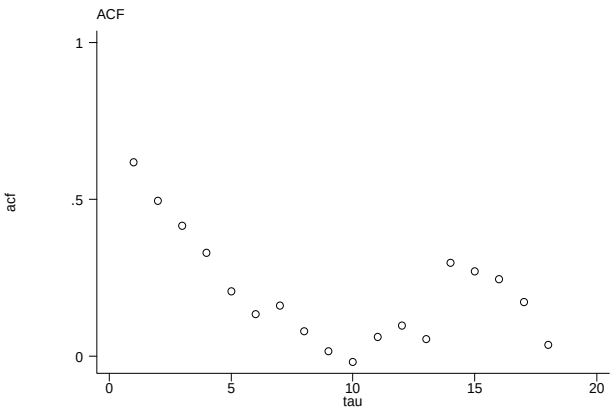
```
regress prot week
predict prot_res, residuals
gen r=0
graph7 prot_res r week, c(.l) s(oi) xlab ylab
(see next page for graph)
```



Because the stationarity assumption is reasonable, we can estimate the autocorrelation function. First, we remove the effect of covariates by adjusting for diet and week to get the residuals, then we use the residuals to plot the scatterplot matrix and autocorrelation function.

```
xi: regress prot week i.diet
predict prot_res, residuals
autocor prot_res week id
```





Comment: Based on the autocorrelation scatterplot, the correlation is reasonably constant along the diagonal in the matrix, but the correlation decreases as the observations are moved away from the diagonals. The means that the correlation between y_{ij} and y_{ik} depends on the time lag (t_{ij}-t_{ik}). After week 10, the correlation fluctuated and does not follow a particular pattern.

<u>Question 4:</u> Dental Study

I. Suppose I don't know the data represent observations on different subjects and assume all responses are independent

Stata command: **by sex: regress dist age**

For Type 0: $\beta_{00} = 17.37$; $\beta_{01} = 0.48$; Var(y_j)= $\sigma_0^2 = 4.683$

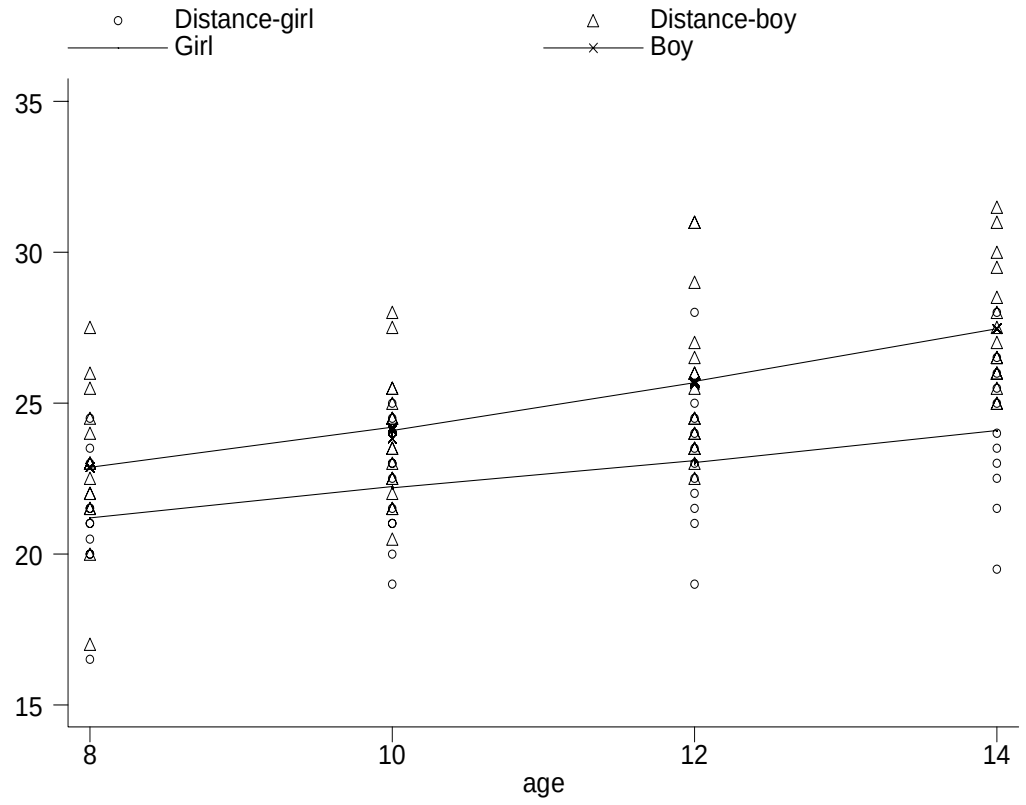
→ Model: $y_j = 17.37 + 0.48x_j + \varepsilon_j$, $\varepsilon_j \sim N(0, \sigma_0^2 = 4.683)$

For Type 1: $\beta_{10} = 16.34$; $\beta_{11} = 0.78$; Var(y_j)= $\sigma_1^2 = 5.372$

→ Model: $y_j = 16.34 + 0.78x_j + \varepsilon_j$, $\varepsilon_j \sim N(0, \sigma_1^2 = 5.372)$

II. The data represent repeated observations on each of 27 children.

1. Plot the data (using Lowess Smooth Curve)



2. Fit Model with Ordinary Least Square (OLS)

Stata command: **by sex: regress dist age** (i.e. dist: distance)

The result is the same as the previous analysis, which assumes cross-section data)

	Intercept	**SE(intercept)**	**Slope**	**SE(slope)**	
Girls	$\beta_{0G} = 17.3727$	1.637755	$\beta_{0G} = 0.4795$	.1459028	$\sigma_0^2 = 4.683$
Boys	$\beta_{0B} = 16.3406$	1.454371	$\beta_{1G} = 0.7844$	.1295657	$\sigma_1^2 = 5.372$

3. Fit the model with Generalized Least Square (GLS) assuming different models for the covariance matrix:

(a) Independent Correlation Structure (same as OLS)

Stata command: **by sex: xtgls dist age, igls**

	Intercept	**SE(intercept)**	**Slope**	**SE(slope)**
Girls	$\beta_{0G} = 17.3727$	1.600101	$\beta_{0G} = 0.4795$	.1425483
Boys	$\beta_{0B} = 16.3406$	1.431466	$\beta_{1G} = 0.7844$	.1275252

(b) Uniform Correlation Structure (This is random effect GLS regression)

Stata command: **by sex: xtreg dist age, re i(id)**

	Intercept	**SE(intercept)**	**Slope**	**SE(slope)**	**rho**
Girls	$\beta_{0G} = 17.3727$	.8587419	$\beta_{0G} = 0.4795$	.0525898	0.875
Boys	$\beta_{0B} = 16.3406$	1.12872	$\beta_{1G} = 0.7844$	.0938154	0.484

(c) Exponential Correlation Structure

Stata command: **by sex: xtgls dist age, igls corr(ar1) i(id) force**

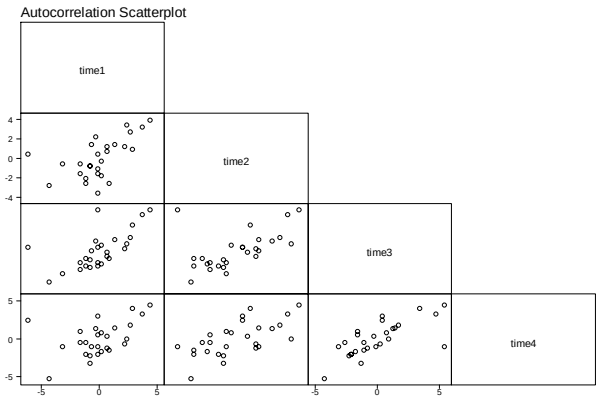
	Intercept	**SE(intercept)**	**Slope**	**SE(slope)**
Girls	$\beta_{0G} = 17.3163$	1.133636	$\beta_{0G} = 0.4841$	.0963581
Boys	$\beta_{0B} = 16.5965$	1.528134	$\beta_{1G} = 0.7694$	.1316101

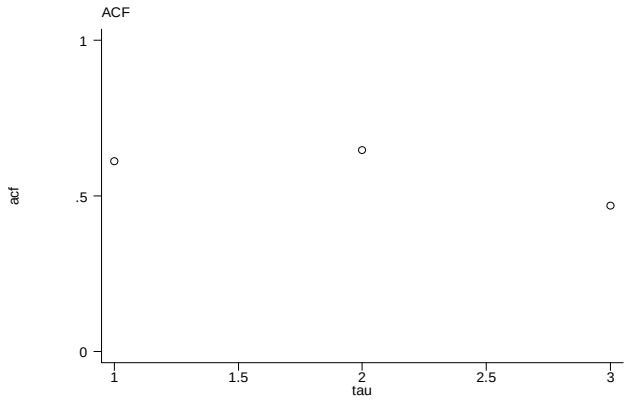
4. To determine which correlation model is the most appropriate, I first regress distance (response) on the two covariates, age and gender to get the residuals. The residuals of distance are then used to get the autocorrelation function and variogram.

Based on the autocorrelation function and the variogram, I think that **uniform correlation model** seemed to be the most appropriate one for the data.

```
xi: regress dist i.age sex
predict dist_res, residuals
```

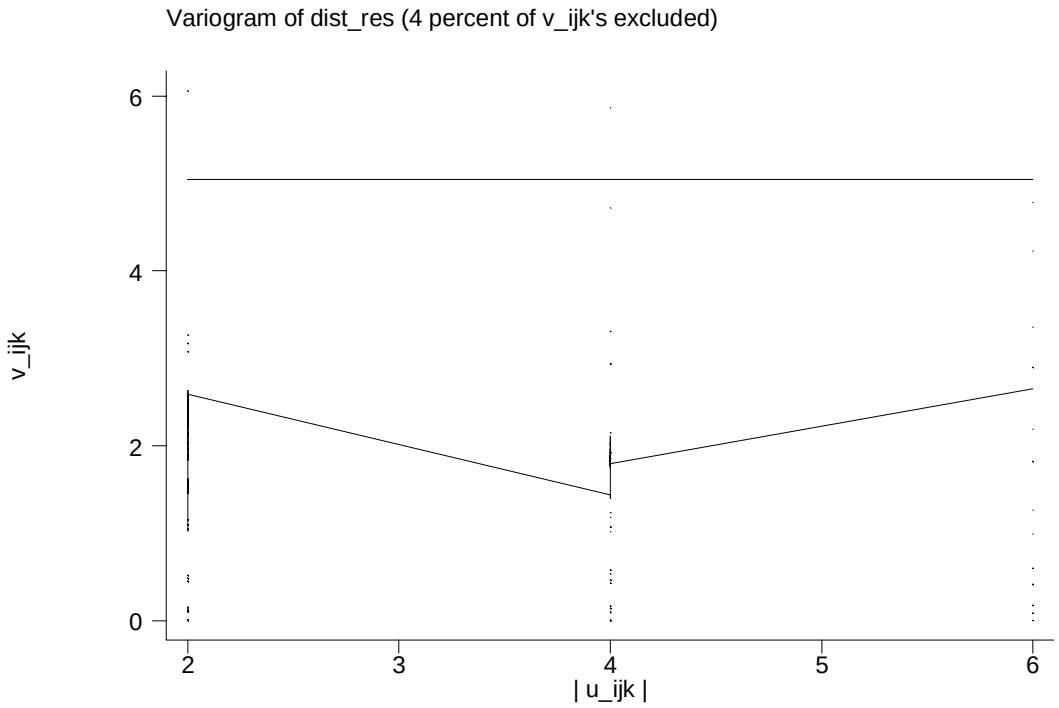
autocor dist_res age id





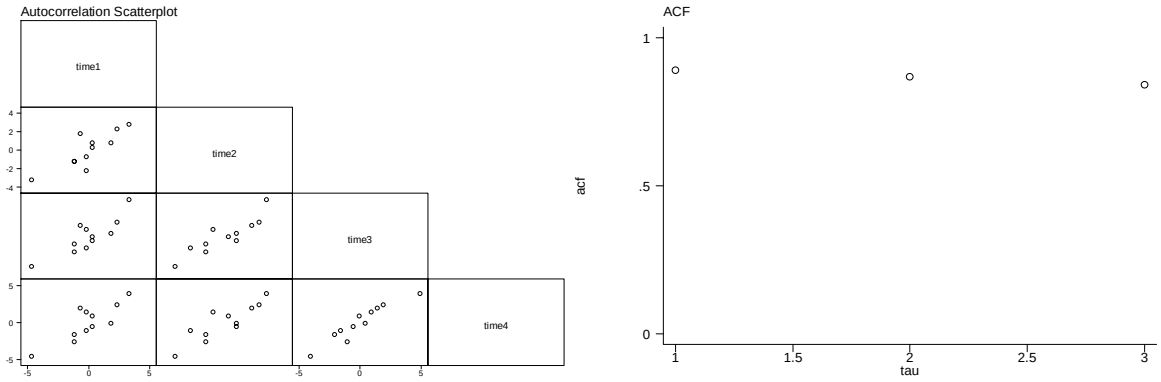
Variogram:

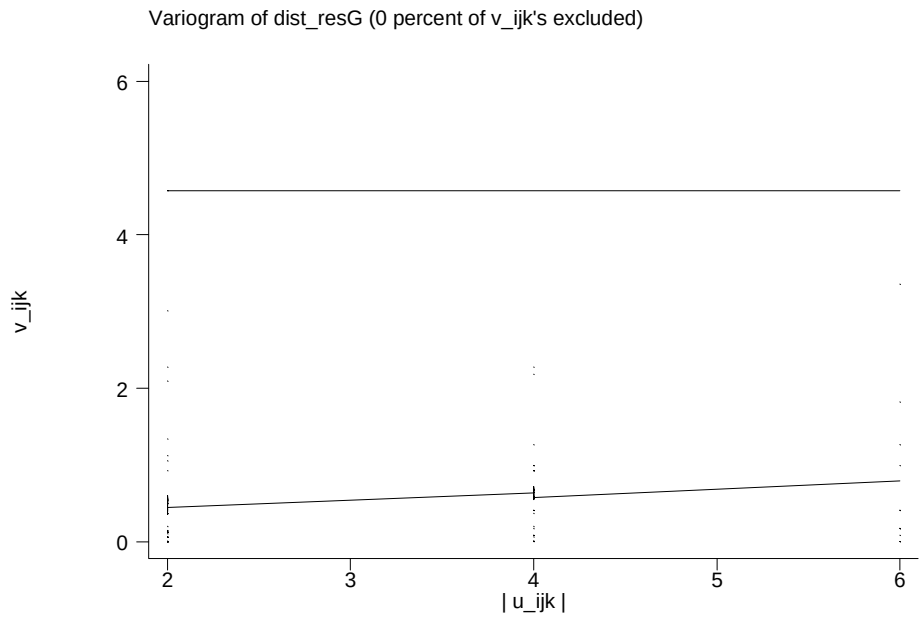
Stata command: `variogram dist_res`



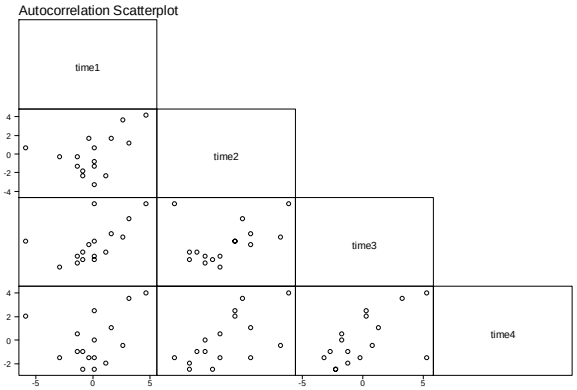
Alternatively, we look at the autocorrelation and variogram for girls and boys separately (using residuals after adjusting for age). This also suggested that uniform model seems to fit the data best for both boys and girls as well.

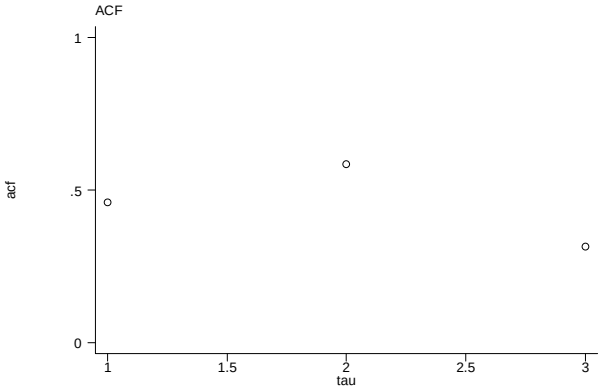
Girls only:

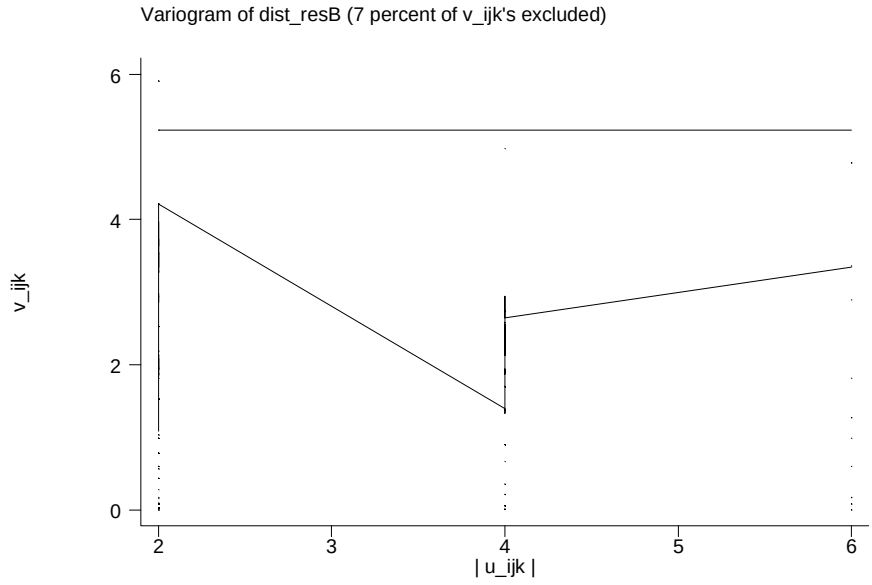



Variogram of dist_resG (0 percent of v_ijk's excluded)
v_ijk
| u_ijk |

Boys Only:


Autocorrelation Scatterplot
time1
time2
time3
time4


ACF
acf
tau


Variogram of dist_resB (7 percent of v_ijk's excluded)
v_ijk
| u_ijk |

5.

(1) WLE under uniform correlation:

Stata command: **by sex: xtreg dist age, re i(id)**

	Intercept	**SE(intercept)**	**Slope**	**SE(slope)**	**rho**
Girls	$\beta_{0G} = 17.3727$	.8587419	$\beta_{0G} = 0.4795$	.0525898	0.875
Boys	$\beta_{0B} = 16.3406$	1.12872	$\beta_{1G} = 0.7844$	.0938154	0.484

(2) MLE under uniform correlation

Stata command: **by sex: xtreg dist age, mle i(id)**

	Intercept	**SE(intercept)**	**Slope**	**SE(slope)**	**rho**
Girls	$\beta_{0G} = 17.3727$	.8310691	$\beta_{0G} = 0.4795$	.0517866	0.868
Boys	$\beta_{0B} = 16.3406$	1.112954	$\beta_{1G} = 0.7844$	.0928289	0.470

6. Compare the estimate and inference of OLS and WLS, we found that both approach got the same estimates of coefficient. However, OLS (the incorrect method) has *larger* standard error of coefficient estimates than WLS. Therefore, the confidence intervals of the coefficients are wider in the OLS as compared to WLS. In terms of inference, both OLS and WLS analysis showed that there is a significant increase in the distance per unit increase in age among both boys and girls in this example. However, with WLS approach (the correct method), we were able to obtain a more precise estimate (i.e. a narrower confidence interval) than the OLS approach.

7. Summary:

Based on the analysis using the longitudinal data, we found that the average distance increases as the age of the children increases. In addition, the rate of distance change differed between boys and girls. Based on the WLS approach under uniform correlation model, the average distance increased by 0.48 (95% CI: 0.376 to 0.583) per year increase in age among girls; the average distance increased by 0.78 (95% CI: 0.601 to 0.968) per year increase in age among boys. This suggested that boys had higher rate of increase than girls did.