

## **LDA 140.655**

### **Guidelines to Solutions for the Midterm 2009**

*This midterm is assigned February 16<sup>th</sup>, 2009.*

*Your solution write-up and appendix is due on February 23<sup>rd</sup>, 2009 at 5PM sharp.*

*Two options for turning in:*

*1) A hard copy of the write-up and appendix must be placed in Howard Chang's departmental mailbox on the 3<sup>rd</sup> floor outside of the Biostatistics Departmental office.*

*2) Turn in a hard copy in class to Francesca on Feb 23<sup>rd</sup>.*

*Late submissions are not allowed.*

*Electronic (email) submissions are not allowed.*

*Midterm rules:*

- 1) You are allowed to consult notes and the textbook*
- 2) You are allowed to consult a LDA classmate*
- 3) Your write-up must be independent work and not team work*
- 4) As an appendix, well-documented code will accompany the write-up*

A study was conducted to investigate two treatments for male patients suffering from multiple sclerosis. 150 male suffers of the disease were recruited into the study, and 75 were randomized to receive azathioprine (AZ) alone (group 1), and 75 were randomized to receive azathioprine + methylprenisommne (AZ + MP, group 2). For each participant, a measure of autoimmunity, AFCR, was planned at clinic visits at baseline (time 0, at initiation of the treatment) and at 3,6,9,12,15, and 18 months thereafter. Multiple sclerosis affects the immune system: low values of AFCR (approaching 0) are evidence that immunity is improving, which is hopefully associated with a better prognosis for suffers of MS. Also recorded for each subject was age at entry into the study and an indicator of whether or not the subject had had previous treatment with either of the study agents (0=no, 1=yes). The average age of the men across groups was 50.45, with SD 6.69.

The primary scientific aim of the study are to investigate whether:

- both treatments (AZ or AZ + MP) lower AFCR over the 18 months period;
- treatment with AZ + MP results in different immune system response than does AZ alone, and if so how it is different in terms of response over time.

It was also suspected that a subject's age and prior history might be related to their AFMR level at baseline and to the rate at which AFMR changes during the 18 months period. The square root of AFMR is the response variable of interest (square roots were taken so that the AFMR observations better satisfy the assumption of normality – note that this is sub-optimal but we'll work with it for now).

The data are in the file `afmr.raw`, which you can download from the class data web page. In the file, each record corresponds to a single observation, with columns:

- col1 = subject id
- col2 = time (months)
- col3 = square root AFMR
- col4 = group (1 = AZ alone, 2 = AZ + MP)
- col5 = prior treatment indicator
- col6 = age (years)

Download these data from the course website and read them into Stata using the `infile` command (See Lab 1 documents for examples of using the `infile` command). Make sure you have the data in “long” format.

```
infile id time sqafmr group ptreat age using "afmr.raw "
```

### Part I: Exploratory Data Analysis (EDA)

- a. Describe the structure of the data set, including distribution of observation times and the number of subjects observed at each time point. Understand the distributions of covariates and whether or not they are time varying.

**We have a total of 150 subjects (shown as `n = 150` in `xtdes` table). Measurements are made at 7 different time points. Among the 150 subjects, the number of subjects with observations at time = 0, 3, 6, 9, 12, 15 and 18 are 123, 116, 124, 123, 125, 117 and 119, which is shown by STATA command “`bysort time: sum sqafmr`”. The panel (`id`) variable and time variable (`time`) can uniquely identify each observation, which is critical for longitudinal data analysis. Second, among the subjects, the minimum number of observations is 2, whereas the maximum number of observations is 7. The median number of observations is 6, which means 50% of the subjects have at least 6 observations. Finally, only 19.33% subjects have the observations at all 7 time points. For other subjects, however, some measurements are not obtained at some time points. The observation time patterns are shown in the last part of the `xtdes` table below. The time-varying variables are `time` and `sqafmr`. The baseline variables are `age`, `ptreat` and `group`. The distributions of baseline variables are shown below. None of the baseline variables have missing values.**

```
. xtset id time
      panel variable:  id (unbalanced)
      time variable:  time, 0 to 18, but with gaps
                   delta: 1 unit
```

**Note – if you have Stata 9 you should use tsset above.**

**. xtides**

```

id: 1, 2, ..., 150          n =      150
time: 0, 3, ..., 18        T =        7
Delta(time) = 1 unit
Span(time) = 19 periods
(id*time uniquely identifies each observation)

```

```

Distribution of T_i:  min    5%    25%    50%    75%    95%    max
                   2      4      5      6      6      7      7

```

Freq.	Percent	Cum.	Pattern*
29	19.33	19.33	1111111
13	8.67	28.00	11111.1
12	8.00	36.00	1.11111
10	6.67	42.67	1111.11
10	6.67	49.33	111111.
7	4.67	54.00	11.1111
6	4.00	58.00	.111111
6	4.00	62.00	111.111
4	2.67	64.67	1.111.1
53	35.33	100.00	(other patterns)
150	100.00		XXXXXXXX

\*Each column represents 3 periods.

**bysort time: sum sqafcr**

-> time = 0

Variable	Obs	Mean	Std. Dev.	Min	Max
sqafcr	123	13.36504	2.327983	8	18.7

-> time = 3

Variable	Obs	Mean	Std. Dev.	Min	Max
sqafcr	116	12.44138	2.073437	7.9	17.5

-> time = 6

Variable	Obs	Mean	Std. Dev.	Min	Max
sqafcr	124	11.92581	2.225019	5.5	17.1

-> time = 9

Variable	Obs	Mean	Std. Dev.	Min	Max
sqafcr	123	11.50732	2.427816	4.8	17.3

-> time = 12

Variable	Obs	Mean	Std. Dev.	Min	Max
sqafcr	125	10.836	2.274806	5.1	17.3

-> time = 15

Variable	Obs	Mean	Std. Dev.	Min	Max
sqafcr	117	10.10085	2.131031	3.6	14.9

-> time = 18

Variable	Obs	Mean	Std. Dev.	Min	Max
sqafcr	119	9.578992	2.468647	2	16.1

. codebook group ptreat age

-----  
group  
(unlabeled)  
-----

type: numeric (float)  
range: [1,2] units: 1  
unique values: 2 missing .: 0/847  
tabulation: Freq. Value  
                  429 1  
                  418 2

-----  
ptreat  
(unlabeled)  
-----

type: numeric (float)  
range: [0,1] units: 1  
unique values: 2 missing .: 0/847  
tabulation: Freq. Value  
                  321 0  
                  526 1

-----  
age  
(unlabeled)  
-----

type: numeric (float)  
range: [32,73] units: 1  
unique values: 33 missing .: 0/847  
mean: 50.4109  
std. dev: 6.69489  
percentiles: 10% 25% 50% 75% 90%  
                  42 46 50 54 59

```
. xtsum age
```

Variable		Mean	Std. Dev.	Min	Max	Observations
age	overall	50.41086	6.694889	32	73	N = 847
	between		6.694269	32	73	n = 150
	within		0	50.41086	50.41086	T-bar = 5.64667

```
. xtsum ptreat
```

Variable		Mean	Std. Dev.	Min	Max	Observations
ptreat	overall	.6210153	.4854209	0	1	N = 847
	between		.4886176	0	1	n = 150
	within		0	.6210153	.6210153	T-bar = 5.64667

```
. xttab group
```

group	Overall		Between		Within
	Freq.	Percent	Freq.	Percent	Percent
1	429	50.65	75	50.00	100.00
2	418	49.35	75	50.00	100.00
Total	847	100.00	150	100.00	100.00

(n = 150)

b. Visualize the longitudinal trajectories of the subjects' square root of AFRCR over time. Include a smooth lowess curve on each of the following three plots to show the overall trend in the square root of AFRCR.

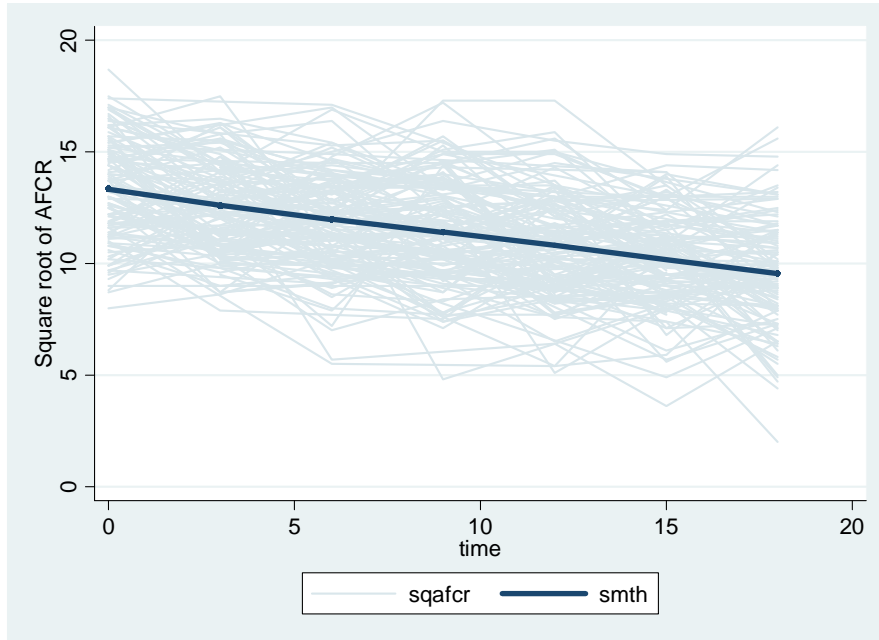
i. Make a spaghetti plot of the square root of AFRCR using information on all the subjects

```
. *generate a smooth lowess curve  
. lowess sqafcr time, gen(smth) nograph
```

```
.
```

```
. *make a spaghetti plot using all the data  
. sort id time
```

```
. twoway line sqafcr time, pstyle(p15) connect(ascending) || line smth  
time, pstyle(p1) clwidth(thick) sort ||, ytitle("Square root of  
> AFRCR")
```



**This spaghetti plot displays so many trajectories that it is hard to pick out individual patterns. We will display a spaghetti plot with fewer trajectories.**

- ii. Make a spaghetti plot of the square root of AFCR using information on a random sample of 10% of the subjects

```
. *make a spaghetti plot using 10% of the subjects (15 subjects)
. ** reshape to wide format to get one row per subject
. reshape wide sqafcr smth, i(id) j(time)
(note: j = 0 3 6 9 12 15 18)
```

Data	long	->	wide
Number of obs.	847	->	150
Number of variables	7	->	18
j variable (7 values)	time	->	(dropped)
xij variables:	sqafcr	->	sqafcr0 sqafcr3 ... sqafcr18
	smth	->	smth0 smth3 ... smth18

```
. ** set seed to 345 (guarantee you can reproduce the same plot again)
. set seed 345
```

```
. ** generate a random variable from uniform(0,1) distribution.
. gen random=uniform()
```

```
. ** reshape data back to long format.
. reshape long
(note: j = 0 3 6 9 12 15 18)
```

Data	wide	->	long
Number of obs.	150	->	1050
Number of variables	19	->	8
j variable (7 values)		->	time
xij variables:			

```

sqafcr0 sqafcr3 ... sqafcr18 -> sqafcr
smth0 smth3 ... smth18 -> smth

```

```

. ** find the 10% quantile of the uniform(0,1) variable.
. centile random, centile(10)

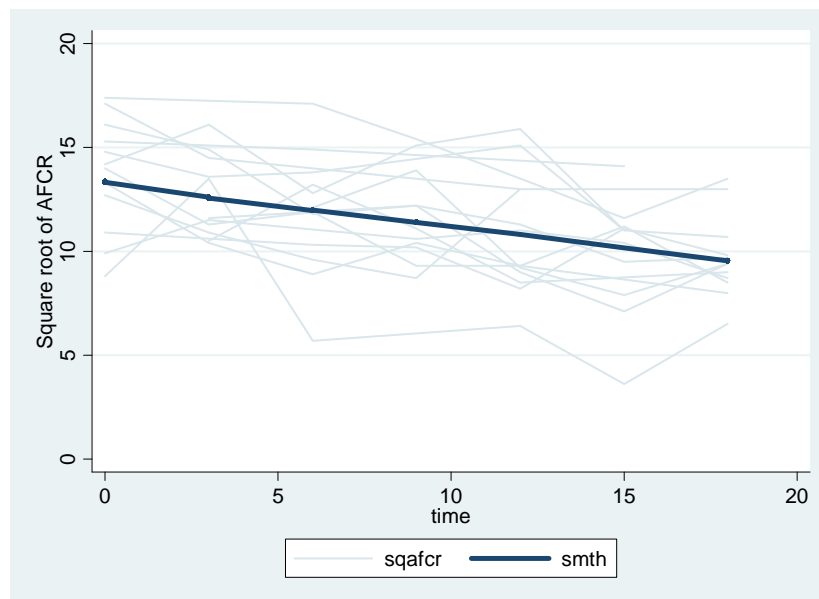
```

Variable	Obs	Percentile	Centile	-- Binom. Interp. -- [95% Conf. Interval]	
random	1050	10	.0498481	.0403368	.0699716

```

. ** plot trajectories only for subjects with value of 'random' < 10th
percentile.
. twoway line sqafcr time if random<.0498491, pstyle(p15) connect(ascending) ||
line smth time, pstyle(p1) clwidth(thick) sort ||, yti
> tle("Square root of AFCR")

```



From this spaghetti plot we can more clearly see how the square root of AFCR seems to be decreasing over time for the majority of the individuals.

- iii. Make a ZAP plot where you plot the trajectories for a subset of individuals selected based on quantiles of the individual-specific median square root of AFCR.

```

*ZAP plot
** use the sqafcr data (instead of residuals - which is another, valid
approach)

```

```

. ** calculate median sqafcr for each subject.
. egen medsqafcr=median(sqafcr), by(id)

```

```

. ** Reshape data to wide format
. reshape wide
(note: j = 0 3 6 9 12 15 18)

```

```

Data                                long  ->  wide

```

```

Number of obs.                1050  ->   150
Number of variables           9      ->   20
j variable (7 values)         time   -> (dropped)
xij variables:
                                sqafcr -> sqafcr0 sqafcr3 ... sqafcr18
                                smth    -> smth0 smth3 ... smth18
-----

. egen maxmedsqafcr=max(medsqafcr)

. egen minmedsqafcr=min(medsqafcr)

. egen medmedsqafcr=median(medsqafcr)

. egen medsqafcr25=pctile(medsqafcr), p(25)

. egen medsqafcr75=pctile(medsqafcr), p(75)

. ** Reshape data to long format
. reshape long
(note: j = 0 3 6 9 12 15 18)

Data                                wide  ->  long
-----
Number of obs.                      150  ->  1050
Number of variables                   25  ->   14
j variable (7 values)                 ->  time
xij variables:
    sqafcr0 sqafcr3 ... sqafcr18     ->  sqafcr
    smth0 smth3 ... smth18           ->  smth
-----

. gen maxsqafcr=sqafcr if medsqafcr==maxmedsqafcr
(1040 missing values generated)

. gen minsqafcr=sqafcr if medsqafcr==minmedsqafcr
(1045 missing values generated)

. gen msqafcr=sqafcr if medsqafcr==medmedsqafcr
(1029 missing values generated)

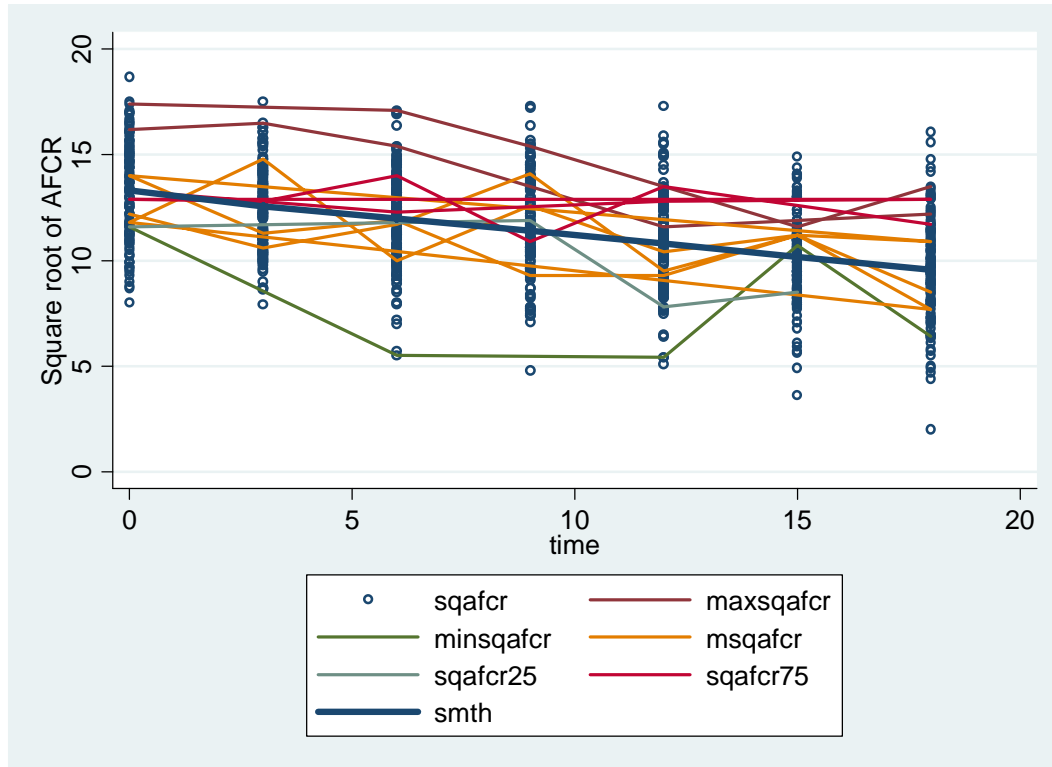
. gen sqafcr25=sqafcr if medsqafcr==medsqafcr25
(1046 missing values generated)

. gen sqafcr75=sqafcr if medsqafcr==medsqafcr75
(1040 missing values generated)

. ** Make a ZAP spaghetti plot using raw data
. sort id time

. scatter sqafcr time, s(oh)|| line maxsqafcr minsqafcr msqafcr sqafcr25
sqafcr75 time, clwidth(medthick medthick medthick medthick me
> dthick) connect(ascending)|| line smth time, pstyle(p1) clwidth(thick) sort
||, legend(lab(1 "sqafcr")) ytitle(Square root of AFCR)

```



Notice that we have a problem with our ZAP plot. We must have multiple individuals at the pre-specified quantiles of the median sqafcr because we are seeing multiple trajectories drawn for some of the quantiles. In particular, the problematic quantiles appear to be the max and the 75<sup>th</sup> and 50<sup>th</sup> percentiles. We will identify the number of individuals at each quantile and the id numbers for these individuals.

```
. * identify the number of individuals at each of our quantiles of interest
. tab id maxsqafcr
```

id	maxsqafcr								Total
	11.6	12.2	13.5	15.4	16.2	16.5	17.1	17.4	
109	1	0	1	1	0	0	1	1	5
139	1	1	0	1	1	1	0	0	5
Total	2	1	1	2	1	1	1	1	10

The trajectories of the two individuals with the same max value look fine – we can see the two distinct trajectories.

```
. tab id minsqafcr
```

id	minsqafcr					Total
	5.4	5.5	6.4	10.7	11.6	
120	1	1	1	1	1	5
Total	1	1	1	1	1	5

```
. tab id msqafcr
```

id	msqafcr								
	7.7	8.5	9.3	9.5	10	10.4	10.6	10.9	11.2
36	1	0	0	1	0	0	1	0	1
117	0	0	0	0	1	1	0	1	1
Total	1	0	0	1	1	1	1	1	2

1	131	7	0	1	2	0	0	0	0	0	1
-----											
1	Total	21	1	1	2	1	1	1	1	1	3

We see approximately 5 trajectories at the median – we need to address this.

id	msqafcr								Total
	11.7	11.8	11.9	12.2	12.6	14	14.1	14.8	
36	1	0	0	1	0	0	1	0	7
117	0	1	0	0	1	0	0	1	7
131	0	0	1	0	0	1	0	0	7
Total	1	1	1	1	1	1	1	1	21

. tab id sqaocr25

id	sqaocr25				Total
	7.8	8.5	11.6	11.9	
88	1	1	1	1	4
Total	1	1	1	1	4

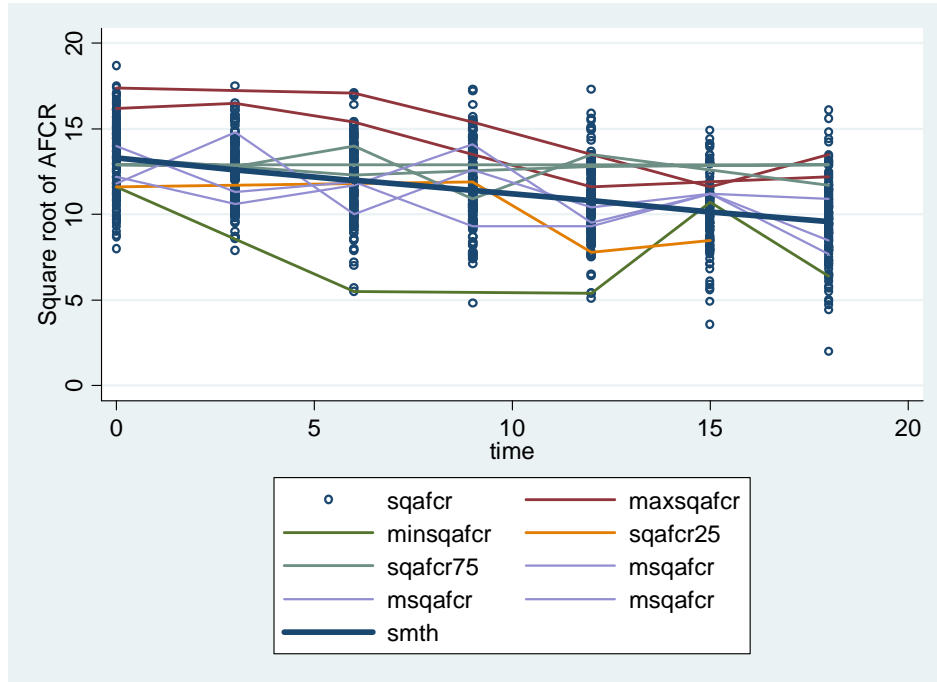
. tab id sqaocr75

id	sqaocr75								Total
	10.9	11.7	12.3	12.8	12.9	13.3	13.5	14	
84	0	0	1	1	1	1	0	0	4
122	1	1	0	1	1	0	1	1	6
Total	1	1	1	2	2	1	1	1	10

We see three trajectories at the 75<sup>th</sup> percentile when we should see only two.

We will fix this problem by plotting the trajectories for each of the problematic individuals directly.

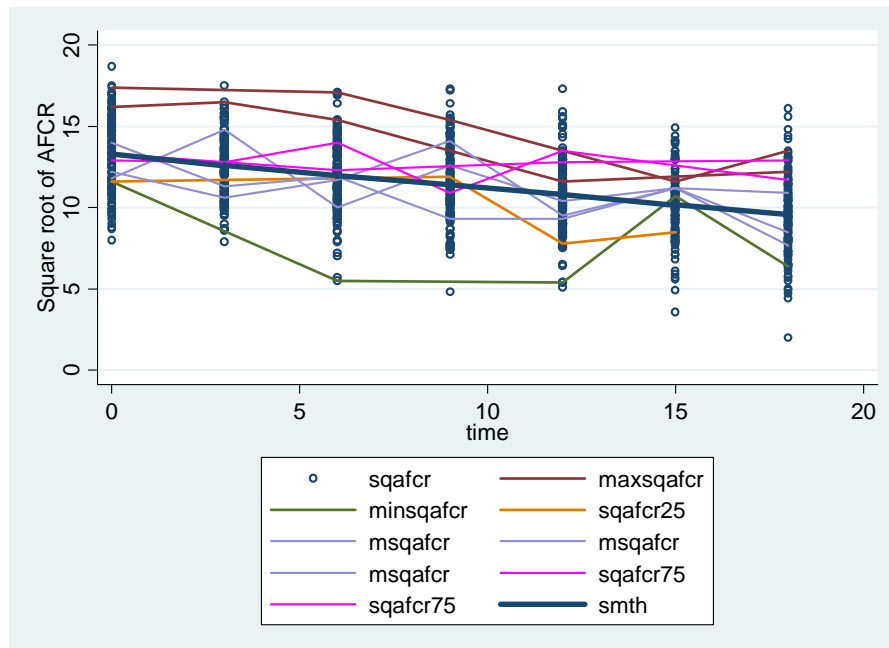
```
**second try making a ZAP spaghetti plot using response data, fix problem with msqafcr
sort id time
scatter sqaocr time, s(oh) || line maxsqaocr minsqaocr sqaocr25 sqaocr75 time,
clwidth(medthick medthick medthick medthick) connect(ascending) || line msqafcr time if
id==36, lcol(lavender) || line msqafcr time if id==117, lcol(lavender) || line msqafcr
time if id==131, lcol(lavender) || line smth time, pstyle(p1) clwidth(thick) sort ||,
legend(lab(1 "sqaocr")) ytitle(Square root of APCR)
```



```

** third try making a ZAP spaghetti plot using response data, fix problem with sqafcr75
** we see three trajectories - there should only be two
sort id time
scatter sqafcr time, s(oh) || line maxsqafcr minsqafcr sqafcr25 time, clwidth(medthick
medthick medthick) connect(ascending) || line msqafcr time if id==36,
lcol(lavender) || line msqafcr time if id==117 , lcol(lavender) || line msqafcr time if
id==131 , lcol(lavender) || line sqafcr75 time if id==84, lcol(magenta) || line sqafcr75
time if id==122, lcol(magenta) || line smth time, pstyle(p1) clwidth(thick) sort ||,
legend(lab(1 "sqafcr")) ytitle(Square root of AFCR)

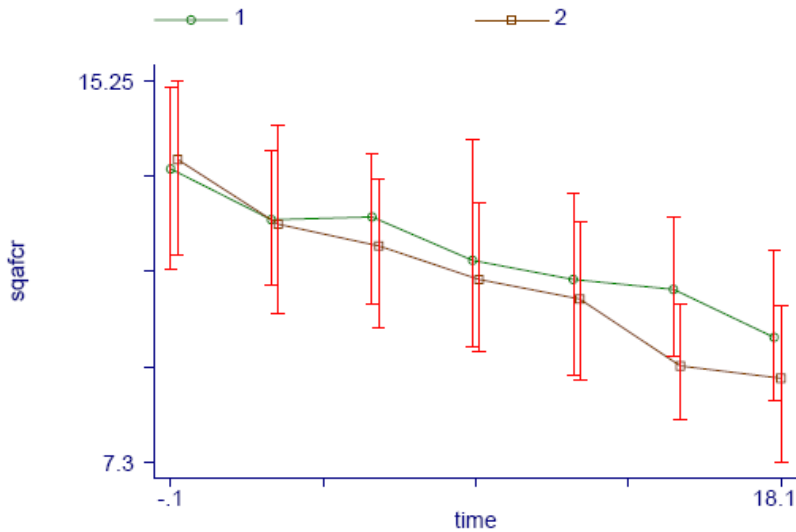
```



The legend isn't that pretty, but now we are seeing the correct number of trajectories!!

- c. Explore overall trends in the data with respect to the primary scientific aim of the study. Present a plot that illustrates the trend in the response (square root of ACFR) over time with respect to treatment group. Ignore age and prior treatment effects for now. Write a **few** sentences summarizing the results.

```
* ignore age and prior treatment effects for now
* grouped by treatment, offset a little bit to make error bars more visible
. xtgraph sqafcr, group(group) av(median) bar(iqr) offset(0.2)
```



The figure above shows that the median values of the square root of the ACFR across time in two treatment groups, AZ alone (Group 1, open circle) and AZ + MP (Group 2, open square). The error bars represent the 25% and 75% quantiles of the square root of the ACFR at each time point. The trend is that the square root of the ACFR decreases over time in both treatment groups. The square root of the ACFR seems to decrease faster over time in Group 2 (AZ + MP).

- d. Explore the correlation structure of the response variable using correlation matrices and the sample autocorrelation function (ACF). Make sure to remove covariates effects as you see appropriate. Include a confidence interval around the ACF estimate. Describe your results.

```
. *remove the effects of covariates, including time, group, ptreat, and age
. xi: reg sqafcr i.time group ptreat age
i.time          _Itime_0-18          (naturally coded; _Itime_0 omitted)
```

Source	SS	df	MS	Number of obs = 847		
Model	2558.5448	9	284.282756	F( 9, 837)	=	77.23
Residual	3080.94176	837	3.680934	Prob > F	=	0.0000
-----				R-squared	=	0.4537
-----				Adj R-squared	=	0.4478
Total	5639.48656	846	6.66605976	Root MSE	=	1.9186

sqafcr	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
_Itime_3	-.8992528	.2483372	-3.62	0.000	-1.38669 - .411816

_Itime_6	-1.481594	.2441822	-6.07	0.000	-1.960875	-1.002312
_Itime_9	-1.824689	.244761	-7.45	0.000	-2.305106	-1.344271
_Itime_12	-2.547711	.2436842	-10.45	0.000	-3.026015	-2.069407
_Itime_15	-3.28142	.2479534	-13.23	0.000	-3.768103	-2.794736
_Itime_18	-3.819111	.2467218	-15.48	0.000	-4.303377	-3.334845
group	-.4615394	.1323548	-3.49	0.001	-.7213257	-.201753
ptreat	.8599866	.1365661	6.30	0.000	.5919343	1.128039
age	.1626793	.0099026	16.43	0.000	.1432423	.1821162
_cons	5.327328	.5767137	9.24	0.000	4.195353	6.459303

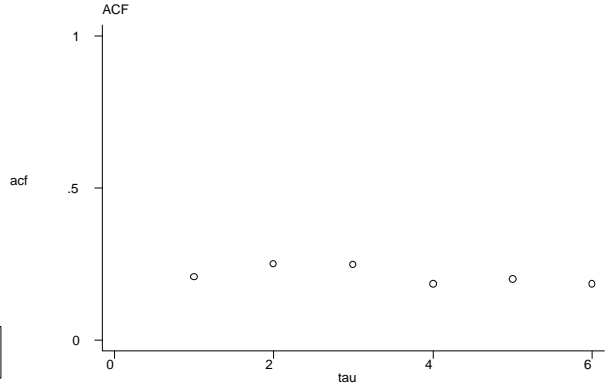
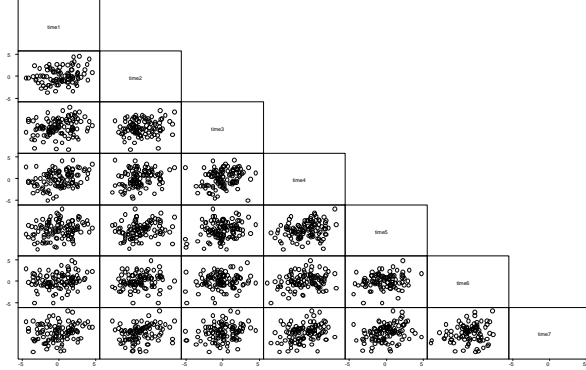
```
. predict afcres1, resid
(203 missing values generated)
```

```
. autocor afcres1 time id
```

	time1	time2	time3	time4	time5	time6	time7
time1	1.0000						
time2	0.2357	1.0000					
time3	0.2253	0.1417	1.0000				
time4	0.3195	0.2357	0.2447	1.0000			
time5	0.2150	0.2017	0.2537	0.2692	1.0000		
time6	0.1833	0.1159	0.1495	0.2368	0.1909	1.0000	
time7	0.1856	0.2269	0.2076	0.3052	0.3141	0.1659	1.0000

	acf
1.	.2095357
2.	.2507799
3.	.2481375
4.	.1853979
5.	.2026924
6.	.1856144

Autocorrelation Scatterplot



**\*\*calculate ACF confidence interval**

**. keep id time afcrrres1**

**. reshape wide afcrrres1, i(id) j(time)**

**(note: j = 0 3 6 9 12 15 18)**

```
Data                                long  ->  wide
-----
Number of obs.                      847  ->   150
Number of variables                   3    ->    8
j variable (7 values)                time  -> (dropped)
xij variables:
                                afcrrres1 -> afcrrres10 afcrrres13 ... afcrrres118
```

**. pwcorr afcrrres10-afcrrres118, obs**

	afcrr~10	afcrr~13	afcrr~16	afcrr~19	afcr~112	afcr~115	afcr~118
afcrrres10	1.0000 123						
afcrrres13	0.2357 93	1.0000 116					
afcrrres16	0.2253 102	0.1417 95	1.0000 124				
afcrrres19	0.3195 100	0.2357 95	0.2447 103	1.0000 123			
afcrrres112	0.2150 103	0.2017 97	0.2537 105	0.2692 104	1.0000 125		
afcrrres115	0.1833 97	0.1159 91	0.1495 96	0.2368 96	0.1909 92	1.0000 117	

afcrres118	0.1856	0.2269	0.2076	0.3052	0.3141	0.1659	1.0000
	99	92	99	98	100	89	119

The number of independent pairs for each lag  $N(u)$  can be calculated by adding up the number of observations along the diagonals from top left to bottom right. The standard error of the ACF for each lag  $u$  is  $1/\sqrt{N(u)}$ .

```

. *number of independent pairs for lag 1
. display 93+95+103+104+92+89
576

. *number of independent pairs for lag 2
. display 102+95+105+96+100
498

. *number of independent pairs for lag 3
. display 100+97+96+98
391

. *number of independent pairs for lag 4
. display 103+91+99
293

. *number of independent pairs for lag 5
. display 97+92
189

. *number of independent pairs for lag 6
. display 99
99

. insheet using acf, names clear
(1 var, 6 obs)
. gen lag=_n
. gen int num = 576 in 1
(5 missing values generated)
. replace num = 498 in 2
(1 real change made)
. replace num = 391 in 3
(1 real change made)
. replace num= 293 in 4
(1 real change made)
. replace num= 189 in 5
(1 real change made)
. replace num= 99 in 6
(1 real change made)

. gen se = 1/sqrt(num)
(1 missing value generated)

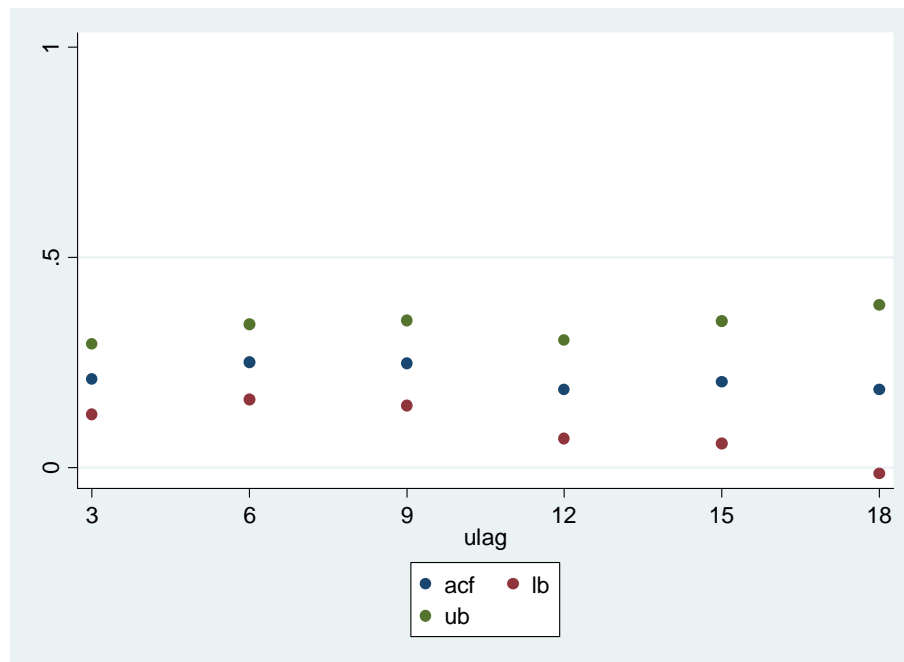
. **plot acf with 95% CI
. gen lb = acf - 2*se
(1 missing value generated)

. gen ub = acf + 2*se
(1 missing value generated)

. replace ub = 1 if ub>1
(1 real change made)

```

```
.gen ulag = lag*3
.scatter acf lb ub ulag, ylabel(0 (0.5) 1) xlabel(3 (3) 18)
```



```
.sort ulag
.save "acf_temp.dta", replace
```

The correlation matrix and the estimated sample correlation function of the square root of the AFCR after removing the covariate effects (time, treatment group, prior treatment and age at entry into the study) are shown above. Generally, all the associations are relatively small ( $\text{corr} < 0.3$ ) with the different time lags. The association appears to become slightly stronger with a time lag of 6 months ( $\text{corr} = 0.25$ ) or 12 months ( $\text{corr} = 0.248$ ) than with a time lag of 3 months. The association becomes slightly weaker after the time lag of 12 months. However, notice that overall, it appears that the estimated correlation may be approximately uniform across the different time lags. The 95% CI for the AFCR is wider at larger time lags because we have less data to estimate the AFCR at larger time intervals.

- e. Create and plot a variogram of the afcr data. Use this variogram to create an ACF plot. Plot the variogram-estimated ACF on the same graph as the ACF estimated above. (Use the same residuals that you used to create the ACF in the previous problem.)

```
. ** Plot Variogram (more useful in the continuous time situation
. ** where it is hard to compute ACF directly)
. ** need to load the afcr data again **
. ** we have already saved the acf data in acf_temp **
```

```

. use "afcr.dta", clear

. ** need longitudinal environment for the variogram **
. xtset id time
      panel variable:  id (unbalanced)
      time variable:  time, 0 to 18, but with gaps
      delta: 1 unit

```

**You need to run variogram.ado to define this command if the .ado file isn't saved in the appropriate place on your computer**

```

. ** need to run xtdiff.ado, ksmapprox.ado before running variogram.ado **

```

```

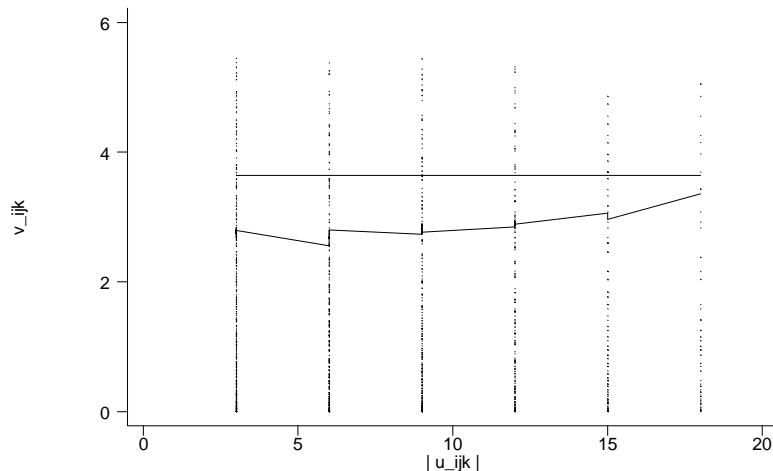
. quietly xi: reg sqaocr i.time group ptreat age
. predict afcrres1, resid
. *plot the variogram

```

```

. variogram afcrres1
Computing smooth lowess model for v in ulag
Variogram of afcrres1 (16 percent of v_ijk's excluded)

```



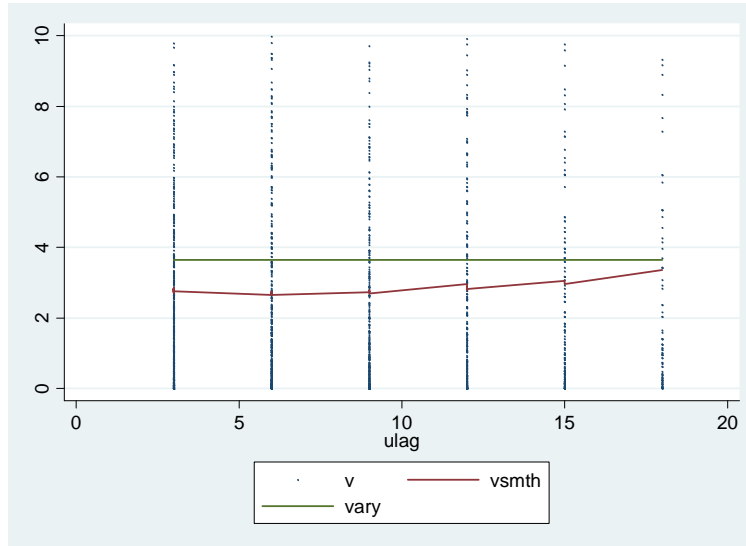
```

. ** Calculate the ACF from variogram **
. ** save current data set before infile the data for variogram from the
. ** "variogram" command above
. save "afcr_temp.dta", replace
file afcr_temp.dta saved

.
. ** read in the variogram data generated by "variogram.ado" **
. ** vario in the following command is the file name in which the variogram
. ** values are stored **
. ** this file is in the working directory of STATA **
. insheet using vario, names clear
(4 vars, 2046 obs)

.
. ** Make our own variogram plot:
. ** v is the sample variogram **
. ** vsmth is the variogram which is the smooth average of sample variogram
. ** vary is the total variance, the horizontal line **
. ** ulag is the time-difference **
. twoway scatter v vsmth vary ulag if v<10, msymbol(p i i) connect(i l l)

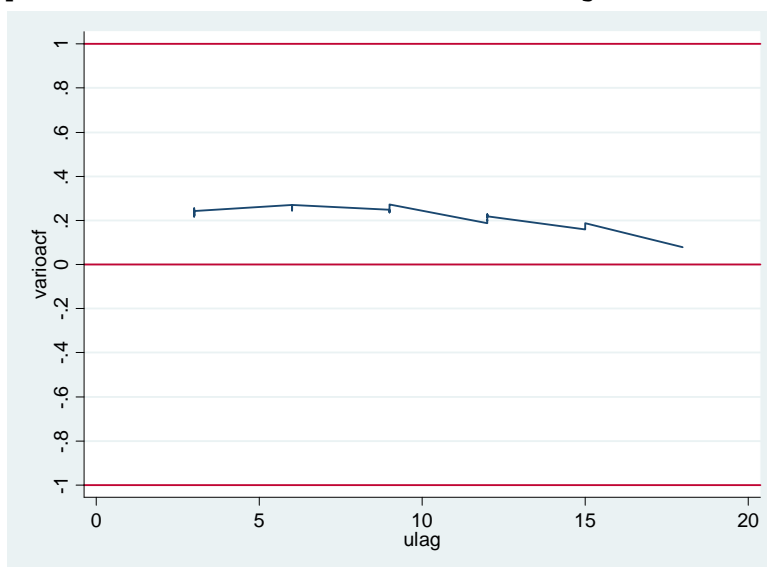
```



```

. ** calculation of the ACF from variogram
. ** This is the way to estimate the ACF when you have continuous time data
. ** and you can't calculate the ACF directly from the data unless you group
the data
.
. gen varioacf = 1 - vsmth/vary
(1846 missing values generated)
.
. twoway line varioacf ulag, ylabel(-1(.2)1) yline(-1 0 1)
. **make the ylabel from -1 to 1 in order to allow negative correlation *

```



```

. sort ulag
. save "vgram_temp.dta", replace
.
. ** merge the variogram data with the acf data
. ** so that we can plot the variogram - based acf on the same plot as

```

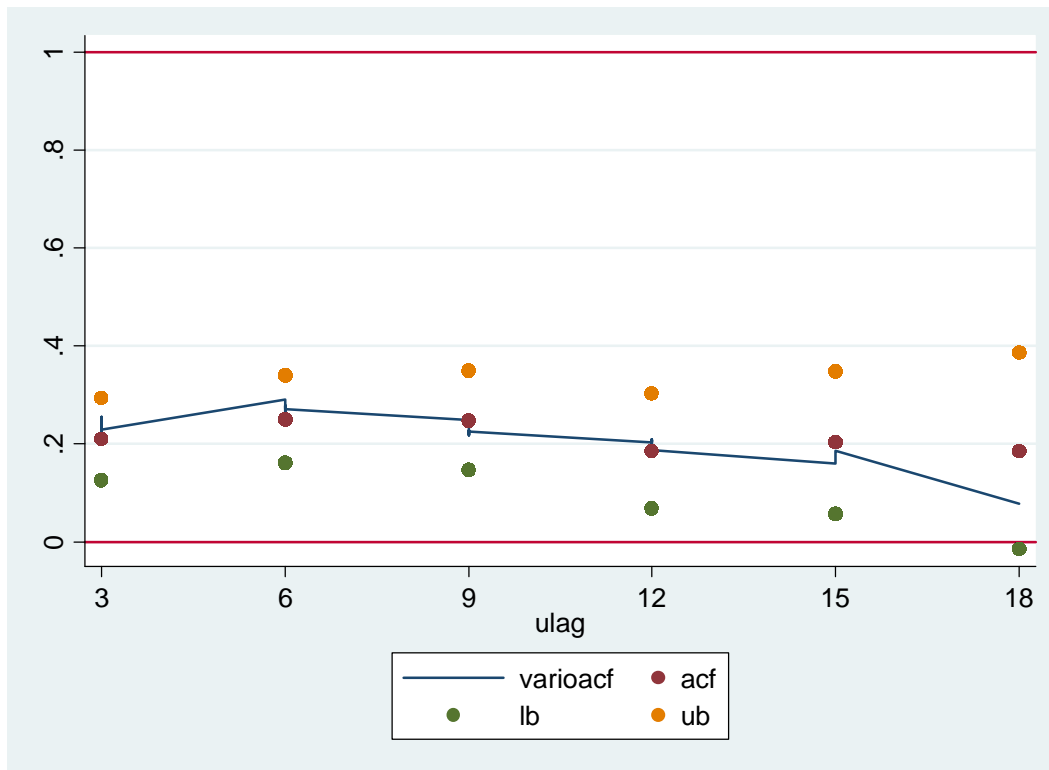
```

. ** the acf that was calculated directly

. merge ulag using "acf_temp.dta"

. ** plot the estimates of the acf using the two different techniques
. ** no negative correlation, so just plot the postive range from 0 to
  **1 of the y-axis
. twoway line varioacf ulag || scatter acf lb ub ulag, ylabel(0(.2)1)
  yline(0 0 1) xlabel(3 (3) 18)

```



We observe that the two techniques produce similar estimates of the ACF. Any differences between the two estimates may be explained by a violation of the “stationary” assumption that is required to derive the relationship between ACF and variogram. See lab 4 for more details. Overall, we see that the two estimates of the ACF result in estimated correlation that is approximately uniform across the time lags.

## Part II: Modeling the AFCR data

Use the AFCR dataset to assess and compare the effects of two treatments for patients suffering from MS.

- Formulate a general mean model that includes, at minimum, an effect for treatment group, for age, prior treatment, for time, and all two-way interactions between time and the other three covariates (formulate a model that assumes independence – i.e., ignore the correlation in the responses). Write down your model. Run the model.

$$y_{ij} = \beta_0 + \beta_1 time_{ij} + \beta_2 group_i + \beta_3 ptreat_i + \beta_4 age_i + \beta_5 (age_i * time_{ij}) + \beta_6 (I_{(group=2)} * time_{ij}) + \beta_7 (I_{(ptreat=1)} * time_{ij}) + \varepsilon_{ij}$$

where  $y_{ij}$  is square root AFCR, and  $t_{ij}$  is 0, 3, 6, 9, 12, 15 and 18.

```
. use "afcr.dta", clear
.
.
.
. xtset id time
      panel variable:  id (unbalanced)
      time variable:  time, 0 to 18, but with gaps
                   delta: 1 unit
.
. gen aget=age*time
.
. gen group2 = group-1
.
. gen gp2t = group2*time
.
. gen ptrtt = ptreat*time
.
. regress sqafcr time group2 ptreat age aget gp2t ptrtt
```

Source	SS	df	MS	Number of obs =	847
Model	2600.19828	7	371.456897	F( 7, 839) =	102.54
Residual	3039.28828	839	3.62251285	Prob > F =	0.0000
				R-squared =	0.4611
				Adj R-squared =	0.4566
Total	5639.48656	846	6.66605976	Root MSE =	1.9033

sqafcr	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
time	-.2079053	.0856172	-2.43	0.015	-.3759543	-.0398563
group2	.2628024	.2369678	1.11	0.268	-.2023169	.7279218
ptreat	.7851258	.244474	3.21	0.001	.3052733	1.264978
age	.1554481	.0177037	8.78	0.000	.1206993	.190197
aget	.0007207	.0016498	0.44	0.662	-.0025176	.003959
gp2t	-.0807612	.021977	-3.67	0.000	-.1238977	-.0376248
ptrtt	.0098912	.0226216	0.44	0.662	-.0345105	.0542928
_cons	4.784678	.9172267	5.22	0.000	2.98435	6.585007

```
. predict afcrrres, resid
.
. autocor afcrrres time id
```

	time1	time2	time3	time4	time5	time6	time7
time1	1.0000						
time2	0.2158	1.0000					
time3	0.2319	0.1256	1.0000				
time4	0.3257	0.2425	0.2452	1.0000			



time	-.1310576	.1586148	-0.83	0.409	-.4419368	.1798217
group2	.3582529	.4317011	0.83	0.407	-.4878656	1.204371
ptreat	.2542262	.4585738	0.55	0.579	-.6445619	1.153014
age	.1659538	.0348391	4.76	0.000	.0976704	.2342372
aget	-.0024018	.0031896	-0.75	0.451	-.0086532	.0038496
gp2t	-.0480013	.0405718	-1.18	0.237	-.1275207	.031518
ptrtt	.0802613	.0429307	1.87	0.062	-.0038813	.1644039
_cons	4.881544	1.725814	2.83	0.005	1.49901	8.264078

(c) Using GEE and your selected correlation structure from (b), test whether treatment has an effect on AFCR, either at baseline and/or on the rate of change of AFCR over time. Use the Huber-White (sandwich) method of robust variance estimation to construct your test. Clearly state null and alternative hypotheses in terms of the model parameters in (a).

**Use the robust variance estimation by specifying “robust”:**

```
. xtgee sqaqcr time group2 ptreat age aget gp2t ptrtt, corr(exch) nmp robust
```

```
GEE population-averaged model
Group variable:          id
Link:                    identity
Family:                  Gaussian
Correlation:             exchangeable
Scale parameter:        3.623731
Number of obs          =      847
Number of groups       =      150
Obs per group: min     =         2
                   avg     =        5.6
                   max     =         7
Wald chi2(7)           =    659.46
Prob > chi2             =     0.0000
```

(Std. Err. adjusted for clustering on id)

sqaqcr	Semi-robust		z	P> z	[95% Conf. Interval]	
	Coef.	Std. Err.				
time	-.1964194	.0729649	-2.69	0.007	-.339428	-.0534108
group2	.2891601	.25823	1.12	0.263	-.2169613	.7952815
ptreat	.7728736	.2576378	3.00	0.003	.2679127	1.277834
age	.1556474	.0171889	9.06	0.000	.1219577	.1893372
aget	.0006288	.0013804	0.46	0.649	-.0020767	.0033343
gp2t	-.0869412	.0194981	-4.46	0.000	-.1251567	-.0487257
ptrtt	.0105459	.0205204	0.51	0.607	-.0296734	.0507651
_cons	4.731701	.9311071	5.08	0.000	2.906765	6.556638

```
. xtcorr, compact
```

```
Error structure: exchangeable
Estimated within-id correlation: 0.2289
```

**Test the null hypothesis that treatment group does not come into play in our model at all:  $H_0: \beta_2 = \beta_6 = 0$  vs.  $H_1$ : at least one of these coefficients  $\neq 0$**

```
. test group2 gp2t
```

```
( 1) group2 = 0
( 2) gp2t = 0
```

```
chi2( 2) = 26.73
Prob > chi2 = 0.0000
```

$\chi^2=26.73$ , p-value<0.05. At the significance level 0.05 we would reject the null and conclude that there is evidence to suggest that treatment has an effect on AFCR either at baseline or on the rate of change of AFCR over time. In other words, either both  $\beta_2$  and  $\beta_6$  are nonzero or one of them is nonzero, while the other is not statistically significantly different from zero.

We now check to see whether treatment has an effect only at baseline by testing whether the coefficient on the interaction between treatment group and rate of change of AFCR over time is not zero in our model, we could test :  $H_0: \beta_6=0$  vs.  $H_1: \beta_6 \neq 0$ .

```
. test gp2t
( 1) gp2t = 0
      chi2( 1) = 19.88
      Prob > chi2 = 0.0000
```

$\chi^2=19.88$  and p-value<0.05. At the significance level 0.05 we would reject the null and conclude that there is evidence to suggest that the treatment has an effect on the rate of change of AFCR over time. Note this is the same p-value you would obtain just by looking at the xtgee results which you can see by remarking that the square of the z-statistic gives the chi-square statistic  $(-4.46)^2 = 19.9$ . The ‘test’ command is useful for looking at whether a set of coefficients are all zero. For a single coefficient, you can just use the model output to test whether this coefficient is zero.

We will not test  $H_0: \beta_2=0$  vs.  $H_1: \beta_2 \neq 0$ . Since we have just shown that we should include the interaction between treatment and time in our model, we should also include the ‘main effects’ of both treatment and time as predictors in our model.

(d) Repeat (c) for the prior treatment variable instead of the treatment variable in the study.

**Test the null hypothesis that prior treatment does not come into play in our model at all:  $H_0: \beta_3=\beta_7=0$  vs.  $H_1: \text{at least one } \neq 0$**

```
. test ptreat ptrtt
( 1) ptreat = 0
( 2) ptrtt = 0
      chi2( 2) = 20.87
      Prob > chi2 = 0.0000
```

$\chi^2=20.87$ , p-value<0.05. At the significance level 0.05 we would reject the null and conclude that there is evidence to suggest that prior treatment has an effect on AFCR either at baseline or on the rate of change of AFCR over time.

Test the null hypothesis that the interaction between prior treatment and rate of change of AFCR over time is not ‘important’ in our model:  $H_0: \beta_7=0$  vs.  $H_1: \beta_7 \neq 0$ .

```
. test ptrtt
( 1) ptrtt = 0

      chi2( 1) =    0.26
    Prob > chi2 =    0.6073
```

$\chi^2=0.26$  and  $p\text{-value}>0.05$ . At the significance level 0.05 we do not reject the null and conclude that there is no evidence to suggest that the prior treatment has an effect on the rate of change of AFCR over time.

- (e) Based on your results in (c) and (d), fit a reduced model, if any, and present key parameter estimates and CIs based on your GEE model fit. Interpret the effects of treatment group and prior treatment on AFCR at baseline and over time. Again, use the robust variance estimator. Make sure your interpretation of parameter estimates is for “population average” or “marginal model” effects.

Since there is no strong evidence to show that the prior treatment has effect on the rate of change of AFCR over time, we take out the interaction of ptreat and time, and refit the model:

$$y_{ij} = \beta_0 + \beta_1 time_{ij} + \beta_2 group_i + \beta_3 ptreat_i + \beta_4 age_i + \beta_5 (age_i * time_{ij}) + \beta_6 (I_{(group=2)} * time_{ij}) + \varepsilon_{ij}$$

where  $y_{ij}$  is square root AFCR, and  $t_{ij}$  is 0, 3, 6, 9, 12, 15 and 18.

Based on the robust variance estimator, we found that treatment group did not have statistically significant effect on AFCR after adjusting the other variables. Prior treatment had statistically significant effect on AFCR after adjusting the other variables. Prior treatment also had an effect on the rate of change of AFCR over time

```
. xtgee sqafcr time group2 ptreat age aget gp2t, corr(exch) nmp robust

GEE population-averaged model
Group variable:          id          Number of obs   =    847
Link:                   identity     Number of groups  =    150
Family:                 Gaussian     Obs per group: min =     2
Correlation:           exchangeable   avg              =    5.6
Scale parameter:       3.620236      max              =     7
Wald chi2(6)           =    662.66
Prob > chi2             =    0.0000
```

(Std. Err. adjusted for clustering on id)

sqafcr	Semi-robust		z	P> z	[95% Conf. Interval]	
	Coef.	Std. Err.				
time	-.1922435	.0721017	-2.67	0.008	-.3335602	-.0509268

group2	.2960125	.2582211	1.15	0.252	-.2100915	.8021165
ptreat	.867898	.1903785	4.56	0.000	.4947631	1.241033
age	.1552254	.0173312	8.96	0.000	.1212568	.1891939
aget	.0006801	.0013885	0.49	0.624	-.0020413	.0034015
gp2t	-.087443	.0194481	-4.50	0.000	-.1255606	-.0493254
_cons	4.690649	.9273086	5.06	0.000	2.873158	6.508141

(f) Refit your model from (e) using:

- your chosen correlation structure with model-based variance estimates;
- independence correlation structure with model-based variance estimates;
- independence correlation structure with robust variance estimates.

Compare the standard errors generated by each of the four model fits and comment on similarities and differences (i.e. try to explain them).

**Robust estimation does not affect the estimation of the working correlation matrix, but only the standard error of the coefficients. As we compare the standard errors generated by each of the four model fits, we can see that standard errors from uniform correlation model is close to those from unstructured correlation whether with robustness or not, which means uniform correlation is a efficient parametric model. Overall, the standard errors generated from the four models are quite similar except the standard error for ptreat. Model 3 had relatively lower standard error for ptreat. Based on robust variance estimation, we got the same inference about the all coefficients no matter which correlation model we assume. With the ROBUST option, XTGEE use a sandwich estimator of variance, which when the mean structure is correctly specified makes your standard error estimates still valid even if you specify the wrong correlation matrix.**

**Model 1: xtgee sqafcr time group2 ptreat age aget gp2t, corr(exch) nmp robust (from part (e))**

**Model 2: xtgee sqafcr time group2 ptreat age aget grp2t, corr(exch) nmp**

**Model 3: xtgee sqafcr time group2 ptreat age aget grp2t, corr(ind) nmp**

**Model 4: xtgee sqafcr time group2 ptreat age aget grp2t, corr(ind) nmp robust**

```
. xtgee sqafcr time group2 ptreat age aget gp2t, corr(exch) nmp
```

```
GEE population-averaged model
Group variable:                id      Number of obs      =      847
Link:                          identity Number of groups   =      150
Family:                         Gaussian Obs per group: min =         2
Correlation:                    exchangeable max           =         7
Scale parameter:                3.620236 Wald chi2(6)       =      619.18
Prob > chi2                     =      0.0000
```

sqafcr	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
time	-.1922435	.0754899	-2.55	0.011	-.3402011 -.0442859
group2	.2960125	.2574557	1.15	0.250	-.2085913 .8006164
ptreat	.867898	.1949355	4.45	0.000	.4858315 1.249965
age	.1552254	.0192157	8.08	0.000	.1175633 .1928874
aget	.0006801	.0014587	0.47	0.641	-.0021789 .0035391

```

gp2t | -.087443 .0194408 -4.50 0.000 -.1255462 -.0493398
_cons | 4.690649 .9932157 4.72 0.000 2.743982 6.637316
-----

```

```

. xtgee sqafcr time group2 ptreat age aget gp2t, corr(ind) nmp

```

```

GEE population-averaged model
Group variable:          id          Number of obs   =   847
Link:                   identity     Number of groups =   150
Family:                 Gaussian      Obs per group: min =    2
Correlation:           independent    avg =           5.6
                                           max =           7
                                           Wald chi2(6)    =   718.29
Scale parameter:       3.619025      Prob > chi2     =   0.0000

Pearson chi2(840):     3039.98      Deviance        =   3039.98
Dispersion (Pearson): 3.619025      Dispersion      =   3.619025

```

```

-----
sqafcr |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----+-----
time   |  -.2040281   .0851157    -2.40  0.017    - .3708519   -.0372044
group2 |   .2688721   .2364469     1.14  0.255    - .1945554    .7322996
ptreat |   .8740927   .1354581     6.45  0.000     .6085998    1.139586
age    |   .155046    .0176713     8.77  0.000     .1204108    .1896812
aget   |   .0007695   .0016453     0.47  0.640    - .0024552    .0039941
gp2t   |  -.0812142   .021942    -3.70  0.000    - .1242198   -.0382086
_cons  |   4.746805   .9126879     5.20  0.000     2.95797    6.53564
-----

```

```

. xtgee sqafcr time group2 ptreat age aget gp2t, corr(ind) nmp robust

```

```

GEE population-averaged model
Group variable:          id          Number of obs   =   847
Link:                   identity     Number of groups =   150
Family:                 Gaussian      Obs per group: min =    2
Correlation:           independent    avg =           5.6
                                           max =           7
                                           Wald chi2(6)    =   677.70
Scale parameter:       3.619025      Prob > chi2     =   0.0000

Pearson chi2(840):     3039.98      Deviance        =   3039.98
Dispersion (Pearson): 3.619025      Dispersion      =   3.619025

```

(Std. Err. adjusted for clustering on id)

```

-----
sqafcr |      Coef.   Semi-robust Std. Err.      z    P>|z|     [95% Conf. Interval]
-----+-----
time   |  -.2040281   .072902    -2.80  0.005    - .3469135   -.0611427
group2 |   .2688721   .2603234     1.03  0.302    - .2413524    .7790966
ptreat |   .8740927   .1921535     4.55  0.000     .4974788    1.250707
age    |   .155046    .017741     8.74  0.000     .1202742    .1898178
aget   |   .0007695   .0014015     0.55  0.583    - .0019774    .0035164
gp2t   |  -.0812142   .0195435    -4.16  0.000    - .1195187   -.0429098
_cons  |   4.746805   .9532822     4.98  0.000     2.878406    6.615204
-----

```

### Saturated model

```

. * saturated model *
. xtgee sqafcr time group2 ptreat age aget gp2t, corr(uns)

```

```

GEE population-averaged model
Group and time vars:    id time   Number of obs   =   847
Link:                   identity     Number of groups =   150
Family:                 Gaussian      Obs per group: min =    2
Correlation:           unstructured  avg =           5.6
                                           max =           7

```

Scale parameter: 3.592024 Wald chi2(6) = 657.38  
 Prob > chi2 = 0.0000

sqafcr	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
time	-.1945504	.0727226	-2.68	0.007	-.3370841	-.0520167
group2	.2954708	.2513709	1.18	0.240	-.1972071	.7881487
ptreat	.8684114	.1930242	4.50	0.000	.490091	1.246732
age	.1501698	.0187462	8.01	0.000	.1134279	.1869116
aget	.0007177	.0014045	0.51	0.609	-.0020351	.0034705
gp2t	-.0890872	.0187239	-4.76	0.000	-.1257853	-.0523891
_cons	4.936241	.9690243	5.09	0.000	3.036988	6.835493

### Part III: Theoretical considerations

Required for Biostatistics PhD, ScM, and MHS Students

a) Derive the variance of the  $\beta_{WLS}$  estimate using matrix notation for a general weight matrix  $W$ .

We assume  $Y \sim MVN(X\beta, V)$ , where  $Var(Y) = V$ . Suppose we use a general weight matrix  $W$ . Then  $\hat{\beta}_{WLS} = (X'WX)^{-1}X'WY$ . Since  $W$  must be symmetric, we have  $W = W'$ .

$$\begin{aligned} Var(\hat{\beta}_{WLS}) &= Var[(X'WX)^{-1}X'WY] = (X'WX)^{-1}X'WVar(Y)[(X'WX)^{-1}X'W]' \\ &= (X'WX)^{-1}X'WVWX(X'WX)^{-1} \end{aligned}$$

b) Derive the variance of the  $\beta_{WLS}$  estimate using matrix notation when the weight matrix  $W$  is  $V^{-1} = Var(Y)^{-1}$ .

We assume  $Y \sim MVN(X\beta, V)$ , where  $Var(Y) = V$ . Since  $V$  is symmetric, we have  $V = V'$ . Plugging into the results from 12. the weight matrix  $W = V^{-1}$ , we have .

$$\begin{aligned} \hat{\beta}_{GLS} &= (X'V^{-1}X)^{-1}X'V^{-1}Y \\ Var(\hat{\beta}_{GLS}) &= (X'V^{-1}X)^{-1}X'V^{-1}VV^{-1}X(X'V^{-1}X)^{-1} = (X'V^{-1}X)^{-1}(X'V^{-1}X)(X'V^{-1}X)^{-1} \\ &= (X'V^{-1}X)^{-1} \end{aligned}$$