

140.655: Sitka Spruce Lab

Scientific problem: Assess the effect of ozone pollution on tree growth.

Lab goals

1. Review exploratory analysis techniques for longitudinal data
2. Decide which correlation structure to use in our model
3. Compare GEE models using QIC
4. Discuss robust estimation of standard errors

Data

The raw **sitka.data** file available on the course data website contains data on 79 trees over two growing seasons. The columns in the raw dataset contain the following information:

- logsize: our outcome is log tree size, $y_{ij} = \log(\text{height}_{ij} \times \text{diameter}_{ij}^2)$
- days: time in days since 1 January 1988
- chamber (1,2 = ozone; 3,4 = normal): The first two chambers, containing 27 trees each, have an ozone enriched atmosphere, the remaining two, containing 12 and 13 trees respectively, have a normal (control) atmosphere.
- ozone (0 = normal, 1 = ozone): indicator for ozone-enriched environment
- year: identifier for year, either 1988 or 1989
- tree (1,...,79): identifier for tree

Step 0: load the data, produce analytic dataset

```
* read in the data using the infile command
* change the location of the data to where you've saved it

. cd "C:\Documents and Settings\Sandrah Eckel\Desktop\LDA lab7"
C:\Documents and Settings\Sandrah Eckel\Desktop\LDA lab7

. infile logsize days chamber ozone year tree using "sitka.data"
'logsize' cannot be read as a number for logsize[1]
'days' cannot be read as a number for days[1]
'chamber' cannot be read as a number for chamber[1]
'ozone' cannot be read as a number for ozone[1]
'year' cannot be read as a number for year[1]
'tree' cannot be read as a number for tree[1]
(1028 observations read)

* the first observation is blank because the 'sitka.data' file
* had the column names listed in the file, so drop the first observation

. drop in 1
(1 observation deleted)

* change the tree identifier from 'tree' to 'id'

. rename tree id
* you may want to save the data as .dta file for later convenience
. save "sitka.dta", replace
file sitka.dta saved
```

```
** limit our analysis to the 1988 growing season
. keep if year==88
(632 observations deleted)
```

Step 1: ALWAYS explore your data

```
** explore missing data patterns
. xtset id days
    panel variable:  id (strongly balanced)
    time variable:  days, 152 to 258, but with gaps
```

```
. xtodes
```

```
id: 1, 2, ..., 79          n =          79
days: 152, 174, ..., 258  T =           5
Delta(days) = 22; (258-152)/22 + 1 = 5.8181818
(id*days uniquely identifies each observation)
```

```
Distribution of T_i:  min      5%      25%      50%      75%      95%      max
                   5         5         5         5         5         5
```

```
      Freq.  Percent   Cum. | Pattern
-----+-----
      79    100.00  100.00 | 11111
-----+-----
      79    100.00      | XXXXX
```

We observe all 79 trees at each time point. The data is balanced (same number of trees at each time point) but not equally spaced (the minimum time between observations is 22 days and the maximum is 31 days).

```
** reshape wide for some data exploration commands
. reshape wide logsize, i(id) j(days)
(note: j = 152 174 201 227 258)
```

```
Data                                long  ->  wide
-----+-----
Number of obs.                       395  ->   79
Number of variables                     6  ->   9
j variable (5 values)                   days  -> (dropped)
xij variables:
                                         logsize  ->  logsize152 logsize174 ... logsize258
```

Data summaries

What does the continuous variable logsize look like over time?

```
. summ logsize152-logsize258
```

Variable	Obs	Mean	Std. Dev.	Min	Max
logsize152	79	4.093291	.6578612	2.23	5.46
logsize174	79	4.518481	.6237667	2.89	5.79
logsize201	79	4.908608	.6052796	3.16	6.12
logsize227	79	5.262405	.637771	3.38	6.42
logsize258	79	5.421139	.6530107	3.52	6.63

The mean logsize increases over time and the standard deviation of logsize fluctuates slightly, with no apparent increasing or decreasing trend over time.

For the (not time-varying) categorical variables:

```
. tab chamber
```

chamber	Freq.	Percent	Cum.
1	27	34.18	34.18
2	27	34.18	68.35
3	12	15.19	83.54
4	13	16.46	100.00
Total	79	100.00	

```
. tab ozone
```

ozone	Freq.	Percent	Cum.
0	25	31.65	31.65
1	54	68.35	100.00
Total	79	100.00	

** exploratory data analysis in long format

```
. reshape long logsize, i(id) j(days)
(note: j = 152 174 201 227 258)
```

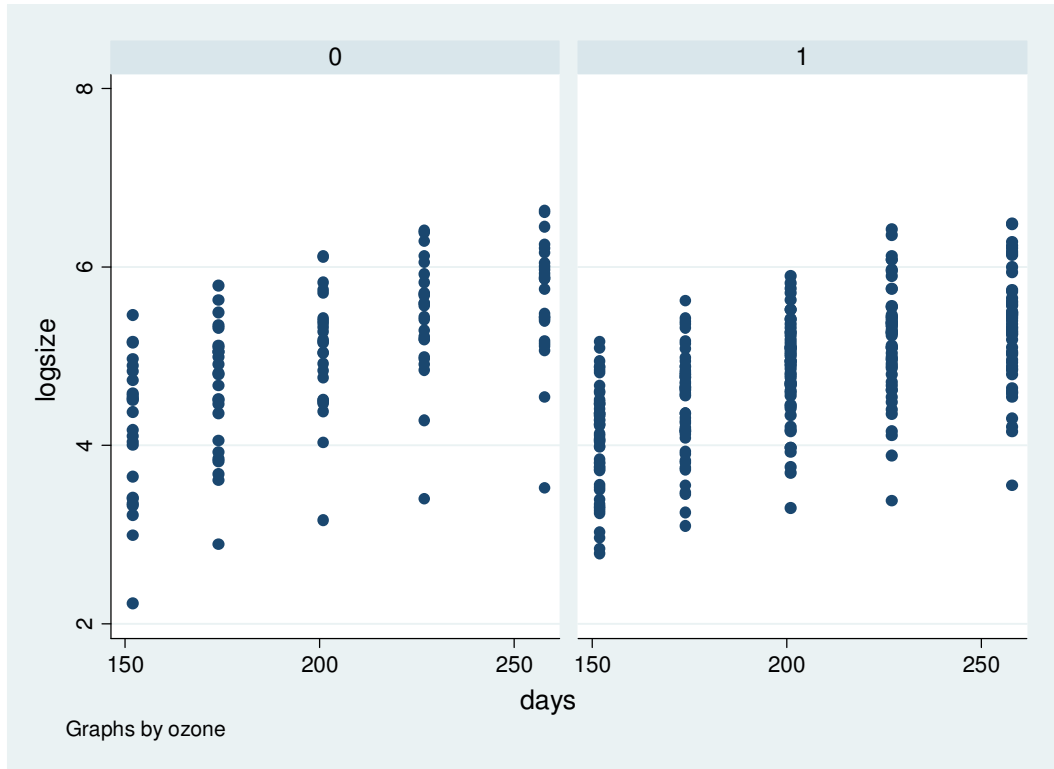
```
Data
```

	wide	->	long
Number of obs.	79	->	395
Number of variables	9	->	6
j variable (5 values)		->	days
xij variables:			
logsize152 logsize174 ... logsize258		->	logsize

Visualize the change in logsize over time, differentiating between normal and ozone environments (ozone = 0 (normal), ozone = 1 (ozone))

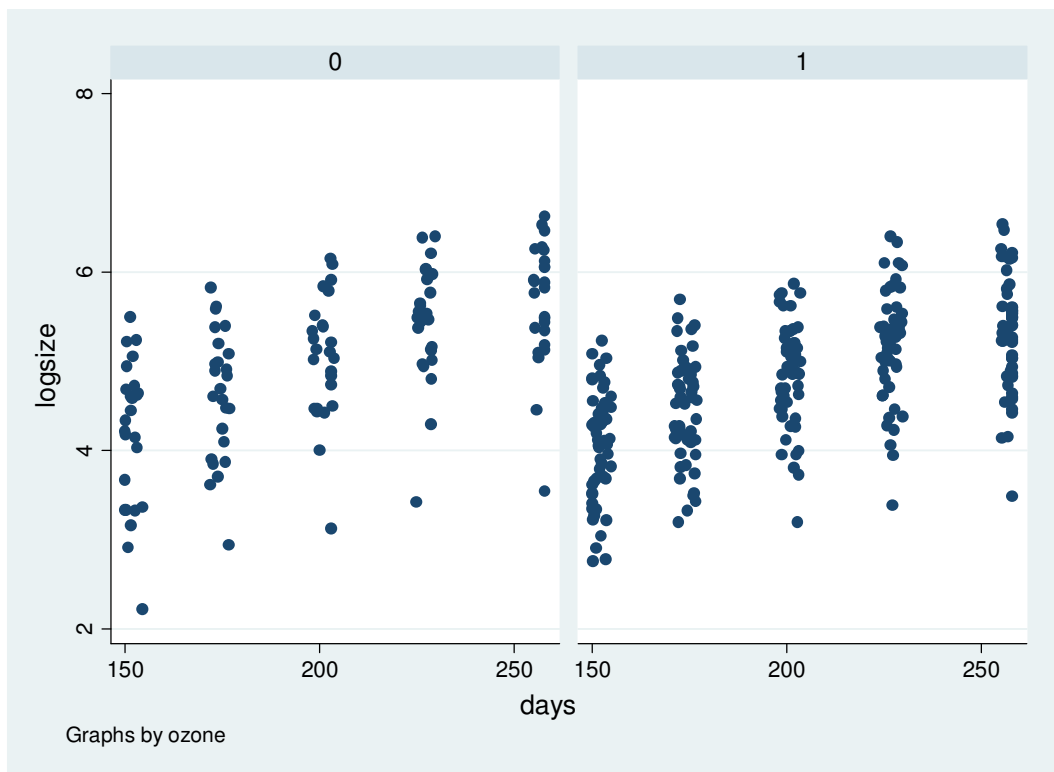
(1) Scatterplots

```
. twoway scatter logsize days, by(ozone) ylab(2 (2) 8)
```



Always jitter your plots when you have points potentially on top of one another!

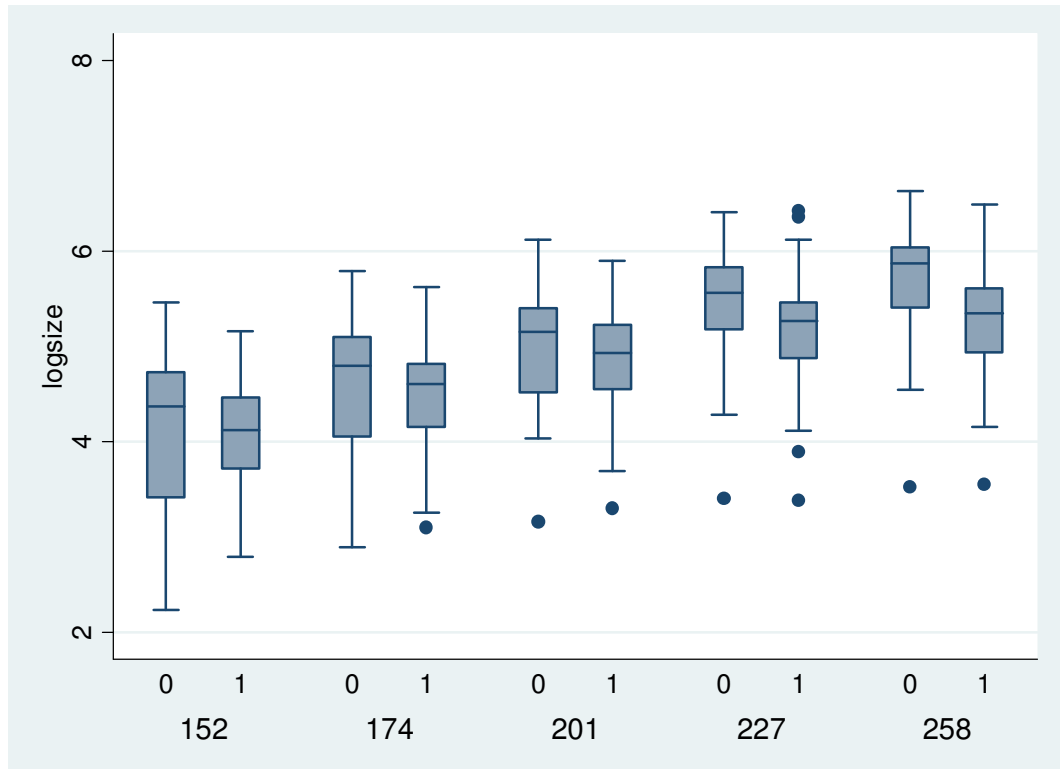
```
. twoway scatter logsize days, by(ozone) jitter(3) ylab(2 (2) 8)
```



This plot is still not all that informative...

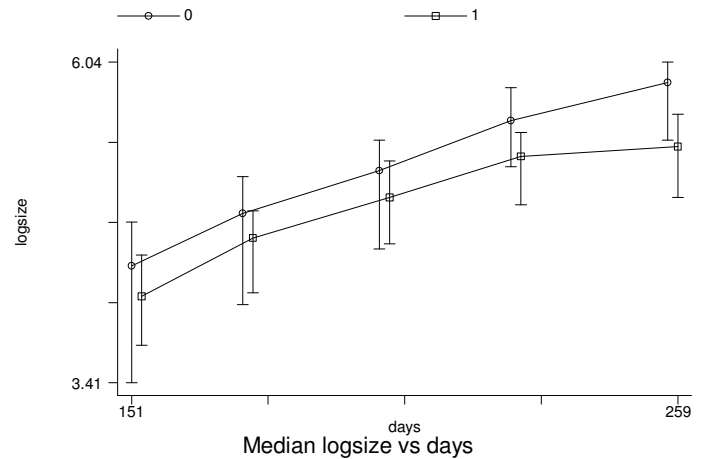
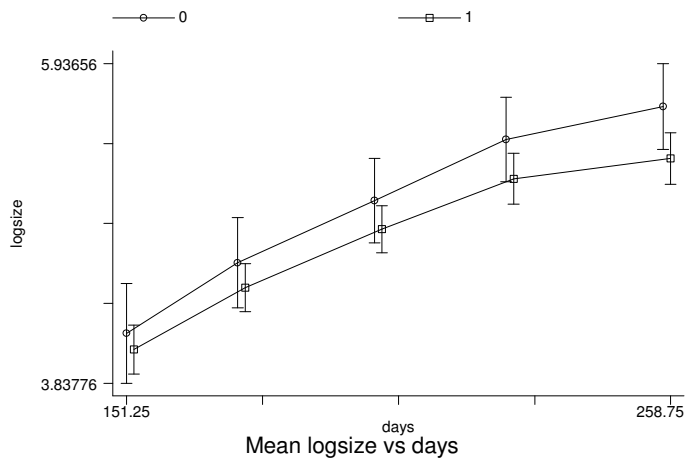
(2) Box plots

```
. graph box logsize, over(ozone) over(days) ylab(2 (2) 8)
```



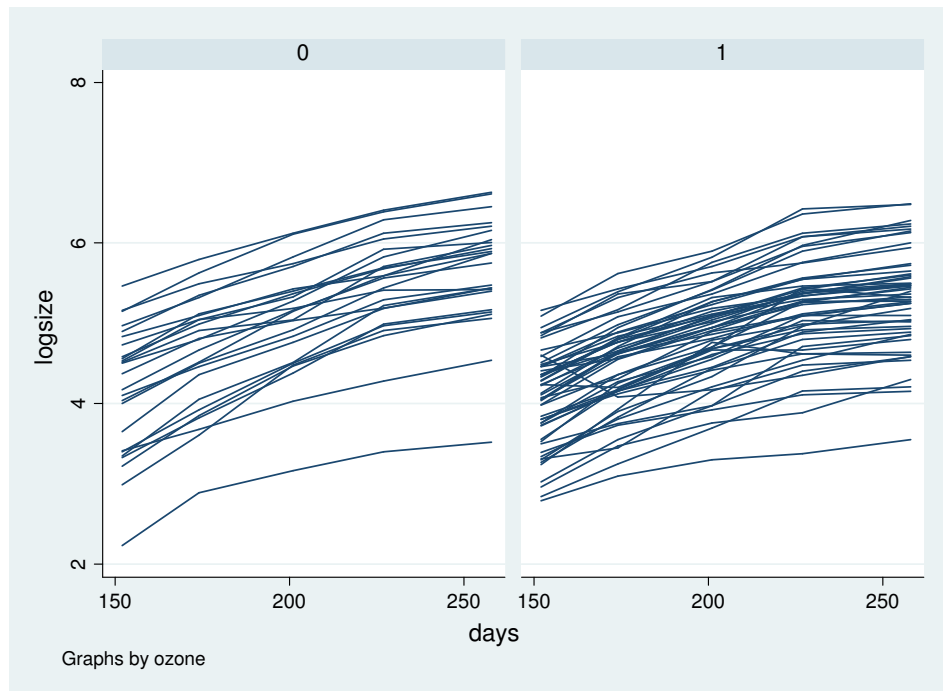
(3) Mean/median trend plots

```
. xtgraph logsize, group(ozone) ti("Mean logsize vs days") bar(ci) offset(1.5)  
. xtgraph logsize, group(ozone) ti("Median logsize vs days") av(median) bar(iqr)  
offset(2)
```



(4) Individual trajectories: Spaghetti plots

```
. sort ozone id days  
. twoway line logsize days, by(ozone) c(L)
```



Model based data analysis – OLS and WLS

The overall growth pattern is clearly non-linear. The focus of today's lab is not to model the overall growth pattern parametrically, so we will simply model the growth curve using dummy variables for each time point. We will concentrate our modeling efforts on the control (normal environment) versus treatment (ozone environment) contrast, in the first growing season (1988).

(1) Create the variables we will use to model the mean structure

In order to reproduce the results on p.76 of the Diggle, Heagerty, Liang and Zeger textbook, we generate an indicator variable for 'control' atmosphere.

```
. gen control=1-ozone
```

Next we generate an interaction between day and atmosphere type using following notation

```
. gen day_control=days/100*control
```

(2) "Remove" mean structure and explore the correlation structure by calculating the autocorrelation function of the residuals from an OLS

```
. xi, noomit: reg logsize i.days control day_control, noconstant
```

Source	SS	df	MS	Number of obs =	395
Model	9353.83562	7	1336.26223	F(7, 388) =	3381.85
Residual	153.309291	388	.395127038	Prob > F =	0.0000
				R-squared =	0.9839
				Adj R-squared =	0.9836
Total	9507.14491	395	24.0687213	Root MSE =	.62859

logsize	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
days_152	4.060577	.07937	51.16	0.000	3.904528 4.216626
days_174	4.470879	.0756955	59.06	0.000	4.322054 4.619703
days_201	4.842733	.0739281	65.51	0.000	4.697383 4.988083
days_227	5.178935	.075257	68.82	0.000	5.030973 5.326898
days_258	5.31669	.0805032	66.04	0.000	5.158413 5.474968
control	-.2216775	.3729256	-0.59	0.553	-.9548854 .5115304
day_control	.213851	.1811625	1.18	0.239	-.1423321 .5700341

```
. estimates store model_ind
. predict residforcorr, resid
```

Autocorrelation function of the residuals:

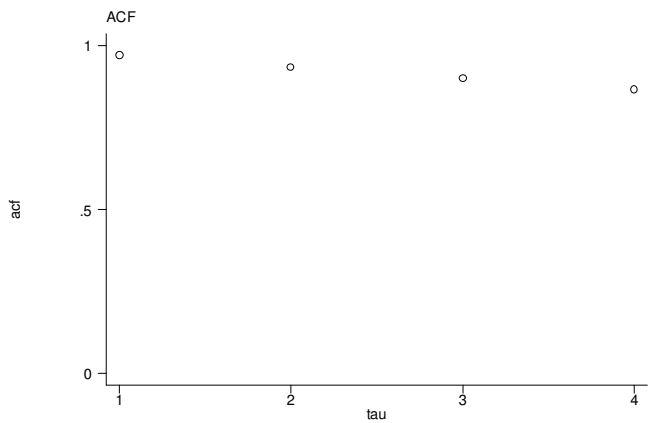
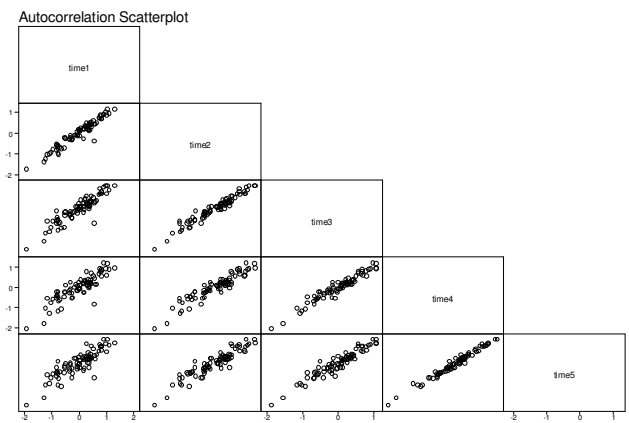
```
. autocor residforcorr days id
```

	time1	time2	time3	time4	time5
time1	1.0000				
time2	0.9621	1.0000			
time3	0.9189	0.9716	1.0000		
time4	0.8750	0.9377	0.9652	1.0000	
time5	0.8668	0.9299	0.9524	0.9876	1.0000

```

+-----+
|       acf       |
+-----+
1. | .9709176 |
2. | .9343636 |
3. | .901251  |
4. | .8667914 |
+-----+

```



logsize	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
daysb152	4.060577	.0843937	48.11	0.000	3.895169	4.225986
daysb174	4.470879	.0841771	53.11	0.000	4.305895	4.635863
daysb201	4.842733	.0840764	57.60	0.000	4.677947	5.00752
daysb227	5.178935	.0841519	61.54	0.000	5.014001	5.34387
daysb258	5.31669	.0844624	62.95	0.000	5.151147	5.482234
control	-.2216775	.1736163	-1.28	0.202	-.5619592	.1186043
day_control	.213851	.0458595	4.66	0.000	.123968	.303734

. xtcorr

Estimated within-id correlation matrix R:

	c1	c2	c3	c4	c5
r1	1.0000				
r2	0.9348	1.0000			
r3	0.9348	0.9348	1.0000		
r4	0.9348	0.9348	0.9348	1.0000	
r5	0.9348	0.9348	0.9348	0.9348	1.0000

Correlation structure: AR1

```
. xi, noomit: xtgee logsize i.days control day_control,
noconstant i(id) corr(ar1)
```

note: observations not equally spaced
modal spacing is delta days = 22
spacing declared with tsset is 1
79 groups omitted from estimation
no observations
r(2000);

The xtgee command cannot model an AR1 correlation structure with unequally spaced data. (See below)

Excerpt from the Stata help file on xtgee:

Correlation structures and the allowed spacing of observations within panel

Correlation	--characteristics allowed--		
	Unbalanced	Unequal spacing	Gaps
independent	yes	yes	yes
exchangeable	yes	yes	yes
ar k	yes (*)	no	no
stationary k	yes (*)	no	no
nonstationary k	yes (*)	no	no
unstructured	yes	yes	yes
fixed	yes	yes	yes

(*) All panels must have at least k+1 obs.

Definitions:

1. Panels are balanced if each has the same number of observations.
2. Panels are equally spaced if the interval between observations is constant.
3. Panels have gaps if some observations are missing.

The number of days between observations in our data is:

```
. display 258-227
    31
. display 227-201
    26
. display 201-174
    27
. display 174-152
    22
```

We can force xtgee to treat the observations as equally spaced using the 'force' command.

```
. xi, noomit: xtgee logsize i.days control day_control, noconstant i(id) corr(ar1)
force
```

```
GEE population-averaged model
Group and time vars:      id days
Link:                     identity
Family:                   Gaussian
Correlation:              AR(1)
Scale parameter:         .3881351
Number of obs            =      395
Number of groups         =      79
Obs per group: min      =       5
                        avg      =      5.0
                        max      =       5
Wald chi2(6)            =     5892.16
Prob > chi2              =      0.0000
```

logsize	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
daysd152	4.059906	.0847529	47.90	0.000	3.893794 4.226019
daysd174	4.469646	.084052	53.18	0.000	4.304907 4.634385
daysd201	4.840811	.0836945	57.84	0.000	4.676772 5.004849
daysd227	5.176348	.0838797	61.71	0.000	5.011947 5.34075
daysd258	5.313312	.0847741	62.68	0.000	5.147157 5.479466
control	-.2318289	.2187354	-1.06	0.289	-.6605424 .1968845
day_control	.2219242	.0802091	2.77	0.006	.0647173 .3791312

```
. xtcorr
```

Estimated within-id correlation matrix R:

```

      c1      c2      c3      c4      c5
r1  1.0000
r2  0.9570  1.0000
r3  0.9158  0.9570  1.0000
r4  0.8764  0.9158  0.9570  1.0000
r5  0.8387  0.8764  0.9158  0.9570  1.0000
```

Correlation structure: Unstructured

```
. xi, noomit: xtgee logsize i.days control day_control, noconstant i(id) corr(uns)
```

```
GEE population-averaged model
Group and time vars:          id days
Link:                          identity
Family:                         Gaussian
Correlation:                    unstructured
Scale parameter:                .3881764
Number of obs                  =      395
Number of groups               =       79
Obs per group: min             =        5
                               avg      =       5.0
                               max      =        5
Wald chi2(6)                   =     8714.87
Prob > chi2                     =      0.0000
```

logsize	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
dayse152	4.068008	.0841851	48.32	0.000	3.903009 4.233008
dayse174	4.475508	.0840788	53.23	0.000	4.310717 4.6403
dayse201	4.843925	.0840398	57.64	0.000	4.67921 5.00864
dayse227	5.176816	.0840976	61.56	0.000	5.011988 5.341644
dayse258	5.310624	.0842883	63.01	0.000	5.145422 5.475826
control	-.3063249	.1612672	-1.90	0.058	-.6224027 .0097529
day_control	.2540909	.0340712	7.46	0.000	.1873126 .3208692

```
. xtcorr
```

Estimated within-id correlation matrix R:

	c1	c2	c3	c4	c5
r1	1.0000				
r2	0.9943	1.0000			
r3	0.9180	0.9162	1.0000		
r4	0.9141	0.9247	0.9210	1.0000	
r5	0.9156	0.9270	0.9189	0.9963	1.0000

The standard errors from the uniform correlation model are ‘closest’ to those from the unstructured correlation model, which means uniform correlation is a relatively efficient parametric model in this case.

Is there a better way to decide on a correlation structure?

QIC - an extension of AIC for GEE models

“The generalized estimating equation (GEE) approach is a widely used statistical method in the analysis of longitudinal data in clinical and epidemiological studies. It is an extension of the generalized linear model (GLM) method to correlated data such that valid standard errors of the parameter estimates can be drawn. Unlike the GLM method, which is based on the maximum likelihood theory for independent observations, the GEE method is based on the quasiliikelihood theory and no assumption is made about the distribution of response observations. Therefore, **Akaike’s information criterion, a widely used method for model selection in GLM, is not applicable to GEE directly.** However, Pan (Biometrics 2001; 57:120-125) proposed a model-selection method for GEE and termed it quasiliikelihood under the independence model criterion. This criterion can also be used to select the best-working correlation structure.” – James Cui, *QIC program and model selection in GEE analyses*

Make sure you're connected to the internet and then download the function for QIC in Stata:

```
** install QIC, the extension of AIC for xtgee models **  
ssc install qic
```

QIC for GEE models with the 4 different correlation structures

. * independent

```
. xi, noomit: qic logsize i.days control day_control, noconstant fam(gaussian)  
corr(indep)  
QIC and QIC_u
```

Corr =	indep
Family =	gaussian
Link =	iden
p =	7
Trace =	10.720
QIC =	174.750
QIC_u =	167.309

. * uniform

```
. xi, noomit: qic logsize i.days control day_control, noconstant fam(gaussian)  
corr(exc)  
QIC and QIC_u
```

Corr =	exc
Family =	gaussian
Link =	iden
p =	7
Trace =	10.720
QIC =	174.750
QIC_u =	167.309

. * ar1

```
. xi, noomit: qic logsize i.days control day_control, noconstant fam(gaussian)  
corr(ar1) force  
QIC and QIC_u
```

Corr =	ar1
Family =	gaussian
Link =	iden
p =	7
Trace =	11.034
QIC =	175.382
QIC_u =	167.313

. * unstructured

```
. xi, noomit: qic logsize i.days control day_control, noconstant fam(gaussian)  
corr(unst)  
QIC and QIC_u
```

Corr =	unst
Family =	gaussian
Link =	iden
p =	7
Trace =	10.798
QIC =	174.926
QIC_u =	167.330

Robust estimation for the standard error

Using the robust option specifies that the Huber/White/sandwich estimator of variance is to be used in place of the traditional calculation

Robust estimation does not affect the estimation of the working correlation matrix that is used in the process of fitting the model. **All that robust estimation does is to change the standard error of the coefficients.** If the structure that we have assumed for the correlation is incorrect, using the Huber/White/sandwich estimator will ‘fix’ this. However, the robustness of our standard error estimates is still dependent on an adequate mean function, so the standard errors are only ‘semi-robust’.

Robust SE estimate with Independence correlation assumption

```
. xi, noomit: xtgee logsize i.days control day_control, noconstant i(id) corr(ind)
robust nolog
```

logsize	Coef.	Semi-robust Std. Err.	z	P> z	[95% Conf. Interval]	
daysp152	4.060577	.0790991	51.34	0.000	3.905546	4.215609
daysp174	4.470879	.0776508	57.58	0.000	4.318686	4.623071
daysp201	4.842733	.0773668	62.59	0.000	4.691097	4.994369
daysp227	5.178935	.0820593	63.11	0.000	5.018102	5.339769
daysp258	5.31669	.0838617	63.40	0.000	5.152325	5.481056
control	-.2216775	.2429765	-0.91	0.362	-.6979027	.2545478
day_control	.213851	.0789394	2.71	0.007	.0591327	.3685694

```
. xtcorr
```

Estimated within-id correlation matrix R:

	c1	c2	c3	c4	c5
r1	1.0000				
r2	0.0000	1.0000			
r3	0.0000	0.0000	1.0000		
r4	0.0000	0.0000	0.0000	1.0000	
r5	0.0000	0.0000	0.0000	0.0000	1.0000

The working correlation matrix is still the same!

Robust SE estimate with Exchangeable correlation assumption

```
. xi, noomit: xtgee logsize i.days control day_control, noconstant i(id) corr(exc)
robust nolog
```

logsize	Coef.	Semi-robust Std. Err.	z	P> z	[95% Conf. Interval]	
daysq152	4.060577	.0790991	51.34	0.000	3.905546	4.215609
daysq174	4.470879	.0776508	57.58	0.000	4.318686	4.623071
daysq201	4.842733	.0773668	62.59	0.000	4.691097	4.994369
daysq227	5.178935	.0820593	63.11	0.000	5.018102	5.339769
daysq258	5.31669	.0838617	63.40	0.000	5.152325	5.481056
control	-.2216775	.2429765	-0.91	0.362	-.6979027	.2545478
day_control	.213851	.0789394	2.71	0.007	.0591327	.3685694

SE estimate with Exchangeable correlation assumption (not robust)

```
. xi, noomit: xtgee logsize i.days control day_control, noconstant i(id) corr(exc)
nolog
```

logsize	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
dayss152	4.060577	.0843937	48.11	0.000	3.895169 4.225986
dayss174	4.470879	.0841771	53.11	0.000	4.305895 4.635863
dayss201	4.842733	.0840764	57.60	0.000	4.677947 5.00752
dayss227	5.178935	.0841519	61.54	0.000	5.014001 5.34387
dayss258	5.31669	.0844624	62.95	0.000	5.151147 5.482234
control	-.2216775	.1736163	-1.28	0.202	-.5619592 .1186043
day_control	.213851	.0458595	4.66	0.000	.123968 .303734

Robust SE estimate with Unstructured correlation assumption

```
. xi, noomit: xtgee logsize i.days control day_control, noconstant i(id) corr(uns)
robust nolog
```

```
GEE population-averaged model
Group and time vars:      id days
Link:                     identity
Family:                   Gaussian
Correlation:              unstructured
Scale parameter:         .3881764
Number of obs            =      395
Number of groups         =       79
Obs per group: min      =        5
                        avg      =       5.0
                        max      =        5
Wald chi2(6)             =     6578.66
Prob > chi2              =      0.0000
```

(Std. Err. adjusted for clustering on id)

logsize	Coef.	Semi-robust Std. Err.	z	P> z	[95% Conf. Interval]
daysr152	4.068008	.079685	51.05	0.000	3.911829 4.224188
daysr174	4.475508	.0778079	57.52	0.000	4.323008 4.628009
daysr201	4.843925	.0771505	62.79	0.000	4.692713 4.995137
daysr227	5.176816	.0814303	63.57	0.000	5.017216 5.336416
daysr258	5.310624	.0827203	64.20	0.000	5.148495 5.472752
control	-.3063249	.2146579	-1.43	0.154	-.7270467 .1143969
day_control	.2540909	.0620627	4.09	0.000	.1324502 .3757316

Use this model specification if you want to be as ‘conservative’ as possible for your standard errors.