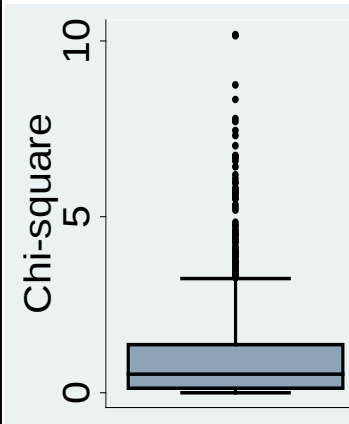


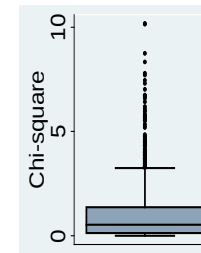
Repeated Measures, Part 3

*Charles E. McCulloch,
Division of Biostatistics,
Dept of Epidemiology and Biostatistics,
UCSF*

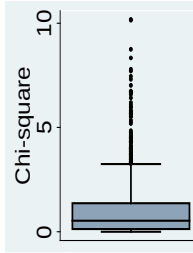


Outline

1. More on **mixed** and **xtgee**
2. Examples
3. Robust standard errors
4. Binary outcomes
5. Changing the link function
6. Modeling practice
7. Summary



MIXED for continuous outcomes

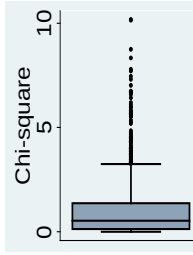


There can be multiple random effects specified by adding additional ||'s and random effects at the end. In this way you can handle multiple hierarchies or levels of clustering. Order is from highest to lowest level of clustering.

Example: backpain data

```
mixed logcost i.pracstyl i.educ ||  
  doctor: || patient:, reml
```

MIXED for continuous outcomes

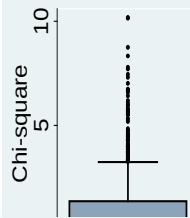


Specifying a random factor with a colon tells MIXED to allow different intercepts for each level of the factor, e.g., different intercepts for each patient.

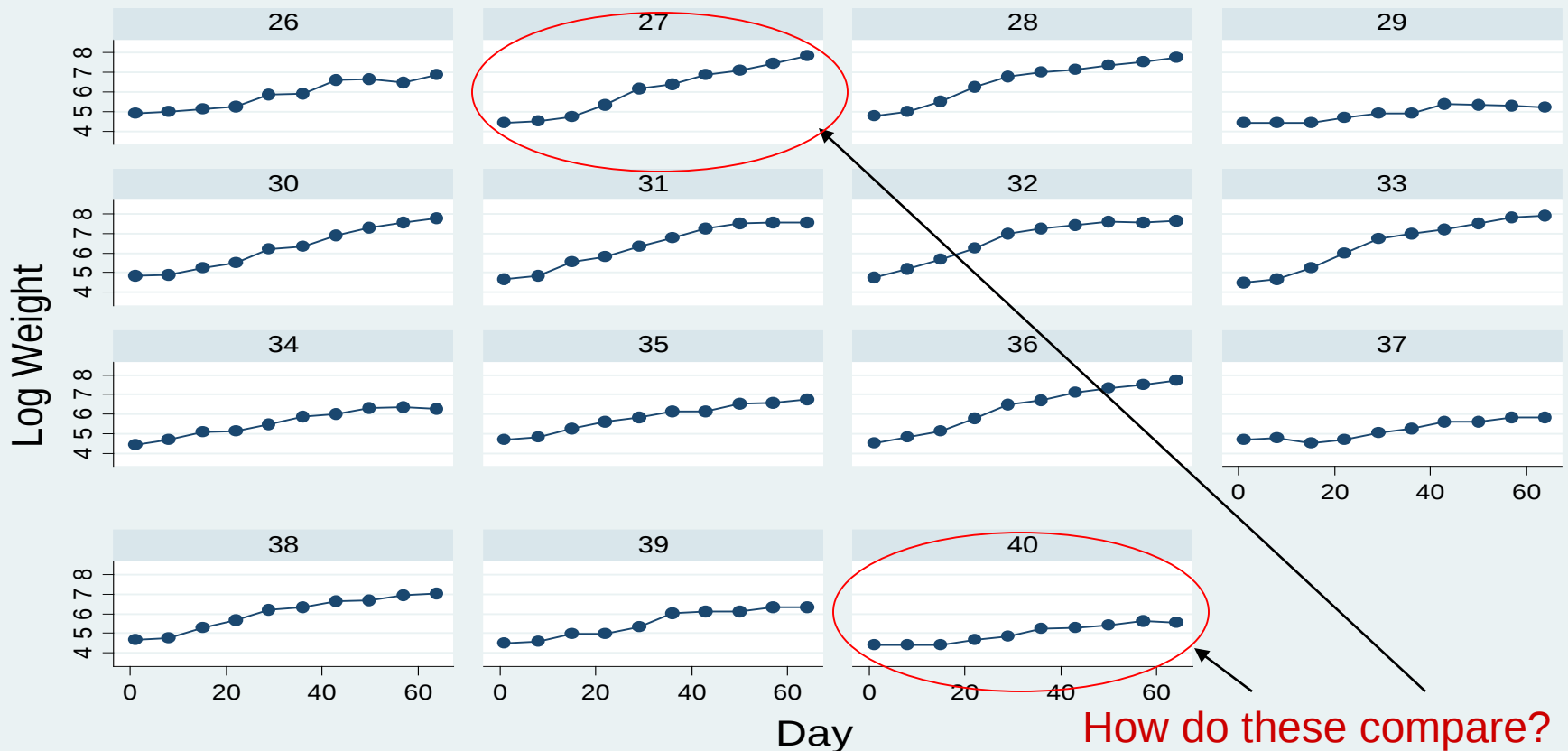
This is often sufficient, but there are cases in which more features of the model need to be included, for example, patient, animal or physician specific terms.

If so, these are added after the colon.

Mouse tumor/weight data

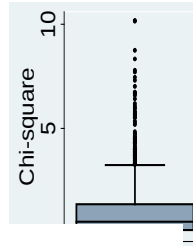


Mouse-specific trajectories of $\log(\text{weight})$
For a single treatment group



Graphs by mouse

Mouse tumor/weight data



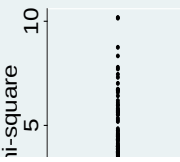
- So - an adequate model would need animal specific slopes and intercepts.

Stata code

- `mixed logw i.group day group#c.day
|| mouseid: day, cov(uns) reml`

Mouse specific
intercepts

Do the treatment groups
change differently over time
(assuming linearity)?
Mouse specific day
coefficients (slopes
with day).

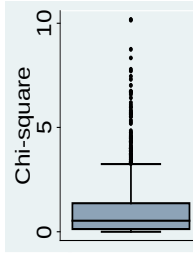


Mouse tumor/weight data

More details in lab, but the “|| mouse:” portion specifies that mouse is a random effect (i.e., allows a mouse specific intercept or constant term) and the “day” term specifies that the slopes over days that are associated with a mouse are also random effects. That is, it allows each mouse to have its own growth trend.

The “cov(un)” specifies that the variance and covariances of the random intercepts and slopes are unstructured. That is, the variances of the random intercepts and slopes are not assumed the same and they may be correlated.

Using xtgee: 5 things to specify



1. What is the distributional *family* (for fixed values of the covariates) that is appropriate to use for the outcome data? Normal, binary, binomial (other?).
2. Which predictors are we going to include in the model? (Nothing new here).
3. In what way are we going to *link* the predictors to the data? (Through the mean? Through the logit of the risk?)
4. What *correlation* structure will be used or assumed temporarily in order to form the estimates?
5. Which variable indicates how the data are clustered?

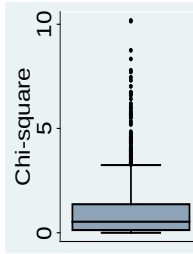
xtgee: syntax

Example: Fecal Fat

```
xtgee fecfat i.pilltype i.sex, i(patid)  
      t(pilltype) family(gaussian) corr(exch)
```

which can be shortened using defaults to

```
xtgee fecfat i.pilltype i.sex, i(patid)
```



More on xtgee

Here are the commonly used options:

Family

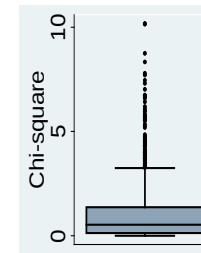
binomial

Gaussian (i.e., normal) default

gamma

nbinomial

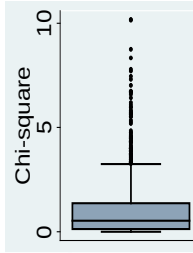
Poisson



More on xtgee

Link

identity (model mean directly)	<u>default for Gaussian</u>
log	<u>default for Poisson</u>
logit	<u>default for binomial</u>
power	
probit	



More on xtgee: main menu

xtgee - Fit population-averaged panel-data models by using GEE

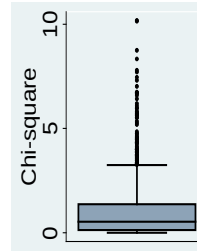
Model | Model 2 | Correlation | by/if/in | Weights | SE/Robust | Reporting | Optimization

Dependent variable:

Independent variables:

Family and link choices:

	Gaussian	Inverse Gaussian	Binomial	Poisson	Negative binomial	Gamma
Identity	☒	☐	☐	☐	☐	☐
Log	☐	☐	☐	☐	☐	☐
Logit			☐			
Probit			☐			
C. log-log			☐			
Power	☐	☐	☐	☐	☐	☐
Odds power			☐			
Neg. binom.					☐	
Reciprocal	☐		☐	☐		☐



More on xtgee

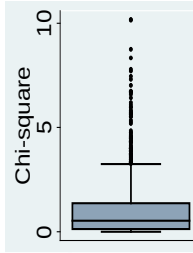
Correlation structure

independent

exchangeable default

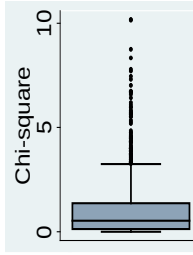
ar #

unstructured



Examples

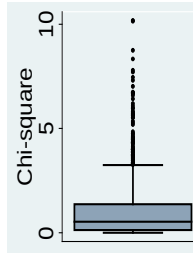
Let's try this on the birthweight data.



- `. xtgee bweight birthord initage, family(gaussian)
link(identity) corr(exchangeable) i(momid)`
- `. xtgee bweight birthord initage, fam(gau) link(i)
corr(exch) i(momid)`
- `. xtgee bweight birthord initage, i(momid)`

all give the same output:

Examples



Iteration 1: tolerance = 7.180e-13

GEE population-averaged model		Number of obs	=	1000
Group variable:	momid	Number of groups	=	200
Link:	identity	Obs per group: min	=	5
Family:	Gaussian	avg	=	5.0
Correlation:	exchangeable	max	=	5
		Wald chi2(2)	=	30.87
Scale parameter:	324458.3	Prob > chi2	=	0.0000

	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
birthord	46.608	9.944792	4.687	0.000	27.11657	66.09943
initage	26.73226	8.957553	2.984	0.003	9.175783	44.28874
_cons	2526.622	162.544	15.544	0.000	2208.042	2845.203

Examples

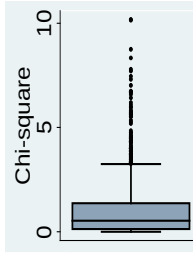
The command

```
. xtcorr
```

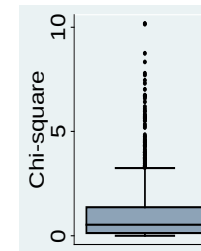
Estimated within-momid correlation matrix R:

	c1	c2	c3	c4	c5
r1	1.0000				
r2	0.3904	1.0000			
r3	0.3904	0.3904	1.0000		
r4	0.3904	0.3904	0.3904	1.0000	
r5	0.3904	0.3904	0.3904	0.3904	1.0000

gives the estimated correlation structure.



Examples: variation I



```
. xtgee bweight birthord initage, i(momid) corr(uns)
```

```
GEE population-averaged model
Group and time vars:      momid birthord
Link:                     identity
Family:                   Gaussian
Correlation:              unstructured

Number of obs      =      1000
Number of groups   =       200
Obs per group: min =         5
                  avg =        5.0
                  max =         5

Wald chi2(2)       =      30.43
Prob > chi2        =      0.0000

Scale parameter:      324495.1
```

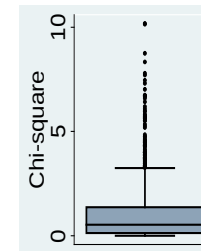
bweight	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
birthord	44.70366	9.935604	4.499	0.000	25.23023 64.17708
initage	28.07164	8.79559	3.192	0.001	10.83261 45.31068
_cons	2505.539	159.0359	15.755	0.000	2193.834 2817.243

```
. xtcorr
Estimated within-momid correlation matrix R:

      c1      c2      c3      c4      c5
r1  1.0000
r2  0.1905  1.0000
r3  0.2977  0.4714  1.0000
r4  0.2475  0.4326  0.6417  1.0000
r5  0.3641  0.3814  0.4286  0.4487  1.0000
```

The estimated correlations are a bit different, but the coefficients and SEs have remained almost unchanged

Examples: variation II



```
. xtgee bweight birthord initage, i(momid) corr(uns) robust
```

```
Iteration 1: tolerance = .04763573
Iteration 2: tolerance = .00062083
Iteration 3: tolerance = .00001004
Iteration 4: tolerance = 1.668e-07
```

```
GEE population-averaged model
Group and time vars:      momid birthord
Link:                     identity
Family:                   Gaussian
Correlation:              unstructured

Number of obs      =      1000
Number of groups   =       200
Obs per group: min =         5
                  avg =        5.0
                  max =         5

Wald chi2(2)       =      29.05
Prob > chi2        =      0.0000

Scale parameter:    324495.1
```

(standard errors adjusted for clustering on momid)

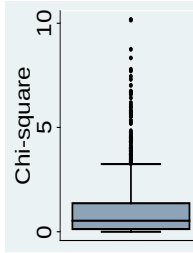
bweight	Coef.	Semi-robust Std. Err.	z	P> z	[95% Conf. Interval]	
birthord	44.70366	9.84746	4.540	0.000	25.40299	64.00432
initage	28.07164	9.141285	3.071	0.002	10.15506	45.98823
_cons	2505.539	161.11	15.552	0.000	2189.769	2821.308

Robust doesn't change
the coefficients, but
recalculates the SEs

Robust standard errors

The “robust” option asks Stata to estimate the standard errors empirically from the data. This has the significant advantage that it gives valid standard errors *even when the assumed correlation structure is wrong*. It is also better than assuming an unstructured variance-covariance structure, because it bypasses the estimation of the correlations over “time” to directly get an estimate of the standard errors.

The robust option works well when there are many “subjects” and not too much data per subject and not much missing data. So, for example, it would work very well when there are 2,000 subjects most measured yearly for four years. It would not work well if the “subjects” were 8 centers in a multi-center trial, each with 1,000 patients enrolled.



Examples: variation III (mixed)

```
mixed bweight birthord initage || momid:, reml
```

Mixed-effects REML regression
Group variable: momid

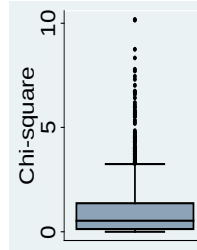
Number of obs = 1000
Number of groups = 200
Obs per group: min = 5
 avg = 5.0
 max = 5

Log restricted-likelihood = -7649.3763
Wald chi2(2) = 30.75
Prob > chi2 = 0.0000

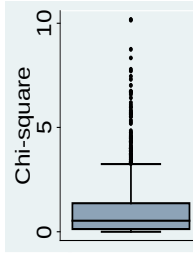
bweight	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
birthord	46.608	9.951014	4.68	0.000	27.10437	66.11163
initage	26.73226	9.002678	2.97	0.003	9.08734	44.37719
_cons	2526.622	163.3387	15.47	0.000	2206.484	2846.76

Random-effects Parameters		Estimate	Std. Err.	[95% Conf. Interval]	
momid: Identity					
	sd(_cons)	358.1759	23.71804	314.5797	407.8139
	sd(Residual)	445.0229	11.13253	423.7298	467.386

LR test vs. linear regression: chibar2(01) = 209.20 Prob >= chibar2 = 0.0000



Binary outcomes/logistic regression



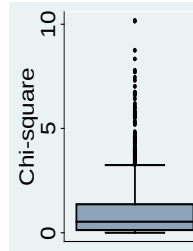
- Now let's take a look at the use of xtgee for clustered logistic regression. I took the Georgia babies data set and artificially dichotomized it as to whether birthweight was above or below 3000 grams.
- What options do we use now?

family =

link =

corr =

Binary outcomes: low birthweight



```
xtgee lowbirth birthord initage, i(momid) family(binomial)
```

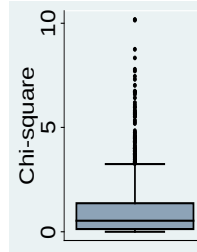
GEE population-averaged model		Number of obs	=	1000
Group variable:	momid	Number of groups	=	200
Link:	logit	Obs per group: min	=	5
Family:	binomial	avg	=	5.0
Correlation:	exchangeable	max	=	5
		Wald chi2(2)	=	11.30
Scale parameter:	1	Prob > chi2	=	0.0035

lowbirth	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
birthord	-.0829363	.0390214	-2.125	0.034	-.159417	-.0064557
initage	-.089028	.0337755	-2.636	0.008	-.1552267	-.0228293
_cons	1.267884	.6036077	2.101	0.036	.0848346	2.450933

How does this compare to the continuous data fit?

OK to assume exchangeable?

Binary outcomes: robust option



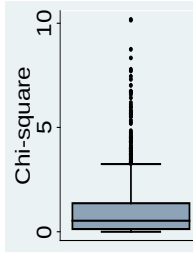
```
xtgee lowbirth birthord initag, i(momid) family(bino) robust
```

```
GEE population-averaged model
Group variable:          momid      Number of obs      =      1000
Link:                   logit        Number of groups   =      200
Family:                 binomial     Obs per group: min =      5
Correlation:            exchangeable avg =      5.0
                                   max =      5
                                   Wald chi2(2)      =      10.64
Scale parameter:        1           Prob > chi2        =      0.0049
```

(Std. Err. adjusted for clustering on momid)

	Semi-robust					
lowbirth	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
-----+-----						
birthord	-.0829363	.0384829	-2.16	0.031	-.1583614	-.0075113
initage	-.089028	.0341776	-2.60	0.009	-.1560149	-.0220412
_cons	1.267884	.6099252	2.08	0.038	.0724524	2.463315

Binary outcomes: low birthweight



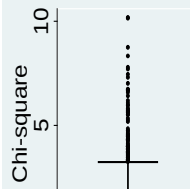
```
. xtgee lowbirth initage birthord, i(momid) link(logit) family(binomial) robust ef
```

```
GEE population-averaged model
Group variable:          momid      Number of obs      =      1000
Link:                   logit       Number of groups   =      200
Family:                 binomial    Obs per group: min =      5
Correlation:            exchangeable avg =      5.0
                                   max =      5
                                   Wald chi2(2)      =      10.64
Scale parameter:        1          Prob > chi2        =      0.0049
```

(Std. Err. adjusted for clustering on momid)

lowbirth		Robust				
		Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
-----+-----						
initage		.9148199	.0312663	-2.60	0.009	.8555464 .9781999
birthord		.9204098	.03542	-2.16	0.031	.8535413 .9925168
_cons		3.553325	2.167263	2.08	0.038	1.075142 11.74368

Low birthweight: mixed effects model



```
. meglm lowbirth initage birthord ||momid:, link(logit) family(binomial) or
```

```
Mixed-effects GLM                                Number of obs      =       1000
Family:                                binomial
Link:                                  logit
Group variable:                            momid

Number of groups      =       200
Obs per group: min    =         5
                  avg  =       5.0
                  max  =         5
```

```
Integration method: mvaghermite                    Integration points =         7
Wald chi2(2)              =       11.90
Prob > chi2                =       0.0026

Log likelihood = -588.05261
```

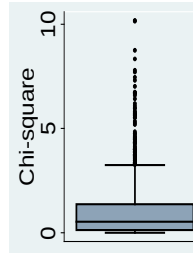
lowbirth	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
initage	.8808697	.0404965	-2.76	0.006	.804969 .9639271
birthord	.8872647	.0500711	-2.12	0.034	.7943597 .9910354
_cons	6.00727	4.979094	2.16	0.031	1.183486 30.49236

```
-----+-----
momid      |
var(_cons)|      2.595213      .5419826      3.907845
-----+-----
```

```
LR test vs. logistic regression: chi2 = 0.0000
                                         25
```

ORs are further from 1 for mixed effects models, though p-values are usually similar. Slightly different interpretation (subject specific).

Changing the link function



What kind of model would this command fit?

```
xtgee bweight birthord initage, i(momid) link(log)
```

GEE population-averaged model		Number of obs	=	1000
Group variable:	momid	Number of groups	=	200
Link:	log	Obs per group: min	=	5
Family:	Gaussian	avg	=	5.0
Correlation:	exchangeable	max	=	5
		Wald chi2(2)	=	30.56
Scale parameter:	324595.5	Prob > chi2	=	0.0000

bweight	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
birthord	.0147553	.0031742	4.648	0.000	.0085339	.0209767
initage	.008179	.0027336	2.992	0.003	.0028212	.0135368
_cons	7.862211	.0503139	156.263	0.000	7.763598	7.960825

Interpretations:

Changing the link function

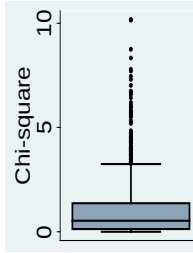
The log link implies we are modeling the log of the mean birthweight. The distribution has defaulted to normal. So our model is

$$\ln(\text{mean birthweight}) = \beta_0 + \beta_1 \text{birthord} + \beta_2 \text{initage}$$

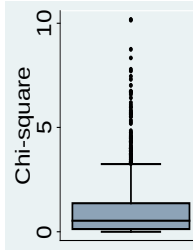
or

$$\text{mean birthweight} = \exp(\beta_0 + \beta_1 \text{birthord} + \beta_2 \text{initage})$$

So with each increase of birth order by one the mean birthweight is predicted to change by $\exp(\beta_1) \times 100\%$, or in our case by $\exp(0.0148) = 1.015 \times 100\% = 101.5\%$ or an increase of 1.5%.



Changing the link function



You can ask Stata to exponentiate for you using the `eform` option:

```
xtgee bweight birthord initage, i(momid) link(log) eform
```

```
GEE population-averaged model
Group variable:                momid
Link:                          log
Family:                        Gaussian
Correlation:                   exchangeable

Number of obs      =      1,000
Number of groups   =       200
Obs per group:
    min            =          5
    avg            =         5.0
    max            =          5

Wald chi2(2)       =      30.56
Prob > chi2        =      0.0000

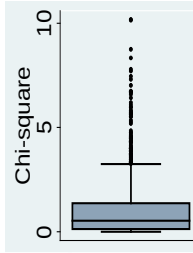
Scale parameter:      324595.5
```

bweight	exp(b)	Std. Err.	z	P> z	[95% Conf. Interval]	
birthord	1.014865	.0032214	4.65	0.000	1.00857	1.021198
initage	1.008213	.0027561	2.99	0.003	1.002825	1.013629
_cons	2597.258	130.6783	156.26	0.000	2353.357	2866.437

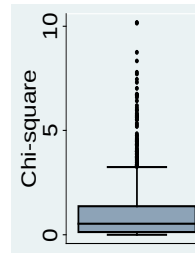
Changing the link function

What kind of model would this command fit?

```
xtgee lowbirth birthord initage,  
      i(momid) family(bino) link(log)
```



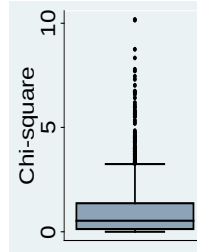
Changing the link function (cont.)



To fit a relative risk model for binary data *with or without clustering* (think of the same question using cross-sectional data) some people recommend the following model. Why should this work?

```
xtgee lowbirth birthord initage,  
      i(momid) family(poisson) robust
```

Changing the link function



Output for binary outcome, log link:

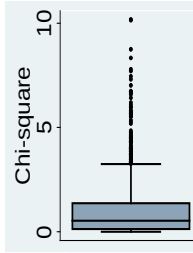
```
xtgee lowbirth birthord initage, i(momid) family(poisson) robust eform
```

```
GEE population-averaged model
Group variable:          momid      Number of obs      =      1,000
Link:                   log         Number of groups   =      200
Family:                 Poisson     Obs per group:
Correlation:            exchangeable      min =          5
                                           avg  =          5.0
                                           max  =          5
                                           Wald chi2(2)    =      10.44
Scale parameter:        1          Prob > chi2         =      0.0054
```

(Std. Err. adjusted for clustering on momid)

		Robust					
lowbirth		IRR	Std. Err.	z	P> z	[95% Conf. Interval]	
-----+-----							
birthord		.9500198	.0227377	-2.14	0.032	.9064839	.9956465
initage		.9426217	.0219578	-2.54	0.011	.900553	.9866557
_cons		1.197149	.4797687	0.45	0.653	.5457846	2.625882

Modeling practice 1



Epileptics were randomly allocated to a placebo or an anti-seizure drug (Progabide) group. The number of seizures was recorded during a baseline period and for four periods after beginning treatment.

Is the drug effective at reducing the number of seizures?

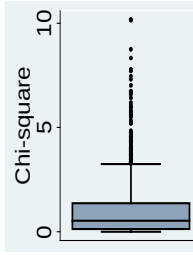
family =

link =

corr =

predictors =

Modeling practice 2



A quality improvement program was designed to reduce prescription of antibiotics for antibiotic resistant infections in emergency departments. 16 hospitals were randomized to the program or control group. Is the program effective?

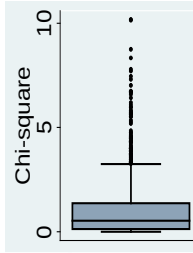
family =

link =

corr =

predictors =

Modeling practice 3



The OAI is a longitudinal cohort study of osteoarthritis. ~5000 participants were studied yearly for 5 years. Is gender, presence of a lesion on the MRI, or race associated with the change in pain over time? Pain is assessed using a numerical scale called the WOMAC.

family =

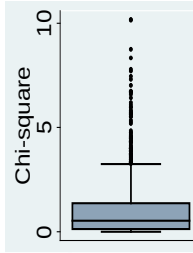
link =

corr =

predictors =

Notes on xtgee

- xtgee is a flexible regression command
- Handles a single level of clustering
- Handles a wide variety of distributions, links and correlation structures
- Five questions: distribution, predictors, link, correlation structure, cluster variable
- Not designed for inferences about the correlation structure
- Doesn't give predicted values for each cluster



Summary

- Hierarchical data structures are common.
- Lead to correlated data.
- Ignoring the correlation can be a serious error.
- `xtgee` can handle single level of clustering and a variety of outcome types.
- `mixed` can handle multiple levels for normally distributed data.
- Not discussed in class: `meologit` and `mepoisson` can handle multiple levels of clustering for binary and Poisson outcomes. Random effects models and robust SEs are also available for time-to-event data.
- GEE methods have the advantage of robust SEs.
- Mixed effects models have the advantage of being able to handle more general types of designs and partition variability.

