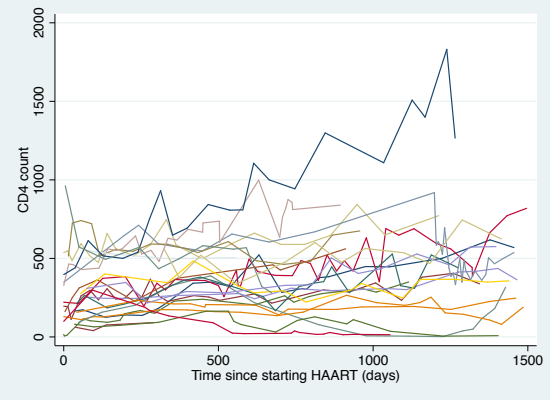


Longitudinal Data
Fall 2014

Chapter 5-6

Approaches to Repeated Measures
Estimation of Marginal (GEE) Models

Instructors

Nick Jewell (*jewell@berkeley.edu*)

GSI

Robin Mejia (*mejia@nasw.org*)

Quick Summary of Repeated Measures Strategies

- Transition Models Relies on some assumption of conditional independence of repeated outcomes, by adjusting for previous outcome value(s): $E(Y_{ij}| X_{ij}, Y_{i(j-1)})$.

- Mixed Effects Models – explicit model of sources of random variability at cluster level, $E(Y_{ij}|X_{ij},\alpha_i)$, $\alpha_i \sim N(0,\sigma^2_\alpha),\ldots.$

- Generalized Estimating Equation (GEE) approach – only specify parameter of interest ($E(Y_{ij}|X_{ij})$) (and perhaps exploiting a guess at correlation structure) and then adjust inference for correlated repeated measures.

Example: observations within subjects: The Effect of Drug and Alcohol Use on Teenage Sexual Activity

- Minnis & Padian (2001) conducted a longitudinal study of teenagers in San Rafael, California to investigate the association between drug and alcohol use and sexual activity on the same day.

- Participants were asked to keep track of their activities over approximately one month and binary indicator variables were created to show whether drug/alcohol use and/or sexual activity were reported for each 24 hour period.

Example of Binary Outcome: Sex, Drugs and Teenagers

- A longitudinal study of the effects of drug-use on sexual activity.
- Let X_{ij}, the only explanatory variable of interest for now, indicate whether or not subject i reported drug-use (1=yes, 0=no) on day j.
- Let Y_{ij} denote whether subject had sex (1=yes, 0=no), i.e., Y_{ij} is a binary outcome and thus its expectation can be modeled via the logit transform.

Data

	i	T_{ij}	X_{ij}	Y_{ij}
	eid	today	drgalcoh	sx24hrs
1.	10122	03 Jun 98	yes	no
2.	10123	04 Jun 98	no	no
3.	10123	05 Jun 98	no	no
4.	10123	06 Jun 98	yes	no
5.	10123	07 Jun 98	no	no
6.	10123	08 Jun 98	no	no
7.	10123	09 Jun 98	no	no
8.	10123	12 Jun 98	no	no
9.	10123	14 Jun 98	yes	no
10.	10123	16 Jun 98	no	no
11.	10123	17 Jun 98	no	no
12.	10123	18 Jun 98	no	yes
13.	10123	19 Jun 98	no	no
14.	10123	20 Jun 98	no	no
15.	10123	21 Jun 98	no	no
16.	10123	23 Jun 98	no	no
17.	10123	25 Jun 98	no	yes
18.	10123	28 Jun 98	no	no
19.	10123	29 Jun 98	no	yes
20.	10123	01 Jul 98	no	yes
21.	10123	02 Jul 98	no	no
22.	10123	03 Jul 98	no	no
23.	10123	04 Jul 98	no	no
24.	10123	05 Jul 98	no	no
25.	10124	04 Jun 98	no	no
26.	10124	07 Jun 98	no	no
27.	10124	08 Jun 98	no	no

Transition Model for Teenage Sex and Drug-Use

- For time-sequenced repeated measures, build the joint distribution by specifying a sequence of distributions that are conditioned on previous measurements on the individual. These are called transition (Markov) models.

- For the study of teenage sex:

$$\text{logit}[P(Y_{ij}=1 \mid \text{X}_{\text{ij}}, Y_{ij-1}, Y_{ij-2}, ..., Y_{i1})] = \beta_0^{\text{TM}} + \beta_1^{\text{TM}} X_{ij} + \delta Y_{ij-1}$$

- where Y_{i1} is outcome at time T_{i1}, Y_{i2} at T_{i2}, ..., and $T_{i1} < T_{i2} < ... < T_{in_i}$.

Transition Model for Teenage Sex and Drug-Use

- This approach constructs the likelihood, for the case of this model

$$\text{logit}[P(Y_{ij}=1 \mid X_{ij}, Y_{ij-1}, Y_{ij-2}, \ldots, Y_{i1})] = \beta_0^{\text{TM}} + \beta_1^{\text{TM}} X_{ij} + \delta Y_{ij-1}$$

- assuming some conditional independence, e.g.,

$$Y_{ij} \perp (Y_{ij-2}, Y_{\text{ij-3}}, \ldots, Y_{\text{i1}}) \mid X_{ij}, Y_{ij-1}$$

Interpretation of Parameters in This Transition Models

- *exp(δ)* = odds ratio (OR) of among subjects who did versus did not have sex during the prior day, keeping drug status fixed.

- exp(β_1^{TM}) = OR of drug use vs. not for either subjects who reported having sex or did not have sex the previous day.

- use generalized linear models (glm) software (e.g., linear, logistic, poisson regression).

- Easiest to use for “clean”, time-structured data.

Sexual Activity and drug/alcohol use among teenagers revisted

Main Variables

sex24hrs - sex in last 24 hrs. (0=no, 1=yes)

drgalcoh - drug or alcohol use in last 24 hrs.

tues-sun - dummy variables designating day of week

Results in STATA

```
.sort eid today
* This is how one puts Y_ij-1 onto same line as Y_ij to be used as covariate
.by eid: gen sxyest = sx24hrs[_n-1]
.by eid: replace sxyest = . if _n==1

.logistic sx24hrs drgalcoh sxyest

. Logit estimates                                 Number of obs   =       1607
                                                  LR chi2(2)      =      55.39
                                                  Prob > chi2     =     0.0000
Log likelihood = -942.60915                       Pseudo R2       =     0.0285

------------------------------------------------------------------------------
     sx24hrs | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
exp(β1™)drgalcoh 1.63798    .1986677     4.07   0.000     1.291421     2.07754
exp(δ)    sxyest 2.051903   .2478562     5.95   0.000     1.619338     2.600018
------------------------------------------------------------------------------
```

- sxyest - sex in 24-48 hrs ago (0=no, 1=yes)
- drgalcoh - drug or alcohol use in last 24 hrs.

Random Effects Models

- Uses a random effect to model the sources of unmeasured variability.

- Assumes model can be written such that outcomes Y_{ij} and Y_{ik}, $j \neq k$ are independent given some realized value of a random effect (β_{i0}) and the covariates, e.g., $Y_{ij} \perp Y_{ik} \mid X_{ij}, \beta_{0i}$

- These *latent* random effects (variables) are assumed to come from a particular parametric model, typically normal.

Random Effects Model for Teenage Sex and Drug-Use

$$\log it[P(Y_{ij}=1\mid\beta_{0i},X_{ij}=x_{ij})]=\log\left(\frac{P(Y_{ij}=1\mid\beta_{0i},X_{ij}=x_{ij})}{P(Y_{ij}=0\mid\beta_{0i},X_{ij}=x_{ij})}\right)=\beta_0^{RE}+\beta_{0i}+\beta_1^{RE}x_{ij}$$

- In this case, build likelihood assuming the repeated observations within *i*th teenager are independent of one another given β_{i0} and X_{ij}: $\text{for } j\neq k: Y_{ij}\perp Y_{ik}\mid X_{ij},\beta_{0i}$
- Assumes parametric distribution for random effects: $\beta_{i0} \sim N(0,\tau^2)$.
- $exp(\beta_1^{RE})$ has interpretation as "individual-level" odds ratio, that is, OR for having sex when subject *i* reports drug-use relative to when same subject does not report drug-use.

Motivation for This Approach

- Intuitive for modeling heterogeneity across individuals.
- This heterogeneity can be represented by a probability distribution
- Model suggests one can make inferences about individuals, not just comparison of populations of individuals.

Motivation for Mixed Models

- Can be extended to hierarchy of units (multi-level modeling), such as repeated longitudinal measures of a person, within a household, within a community.
- Can include non-nested sources of variation (so-called crossed random effects).
- Useful for estimating the contributions to variability from different sources (e.g., within and among individuals).

Some available software for random effects models

- Linear Models
  - Proc Mixed in SAS
  - xtreg in STATA (only simple random effects models)
  - xtmixed in STATA 10
  - lme in R
- Logistic and Poisson Models
  - xtlogit and xtpoisson in STATA for simple random effects, xtmelogit and xtmepoisson for general mixed models in STATA version 10
  - gllamm – for general mixed models is STATA add-on

Random effects using xtlogit in STATA

```
. xtlogit sx24hrs drgalcoh, or i(eid) re

Random-effects logit                            Number of obs      =      1708
Group variable (i) : eid                        Number of groups   =       109

Random effects u_i ~ Gaussian                   Obs per group: min =         1
                                                               avg =      15.7
                                                               max =        33

                                                Wald chi2(1)       =      5.48
Log likelihood  = -921.39213                    Prob > chi2        =    0.0192

------------------------------------------------------------------------------
     sx24hrs |         OR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
exp(β1^RE)   |   1.447266   .2284893     2.34   0.019     1.062096    1.972119
-------------+----------------------------------------------------------------
    /lnsig2u |   .5483488   .2428238                      .0724228    1.024275
-------------+----------------------------------------------------------------
  τ  sigma_u |   1.315444   .1597106                      1.036875    1.668854
         rho |   .3446819   .0166718                      .2463036    .4584528
------------------------------------------------------------------------------
Likelihood ratio test of rho=0: chibar2(01) =   184.17 Prob >= chibar2 = 0.000
```

Estimation of Marginal Models (GEE)

- Focus not on modeling sources of variability, but just estimating so-called “marginal” regression model.

- Marginal model describes how means change in *populations* of individuals as covarites change (sometimes called population average models).

- Thus, the parameter of interest, $E[Y_{ij} \mid X_{ij} = x_{ij}]$ is defined as the mean value of an observation Y_{ij} in the theoretical experiment where one randomly draws an observation from a population where everyone has $X_{ij} = x_{ij}$.

Marginal Models (GEE)

- For instance, if Y_{ij} is the cholesterol and X_{ij} = *yes* if one smokes, *no* otherwise. In a marginal model, $E[Y_{ij} \mid X_{ij} = yes]$ will be the mean of a randomly drawn Y_{ij} from the sub-population where everyone smokes.

- Marginal is odd term, and depends on assuming the underlying model is a random effects model, e.g., $E[Y_{ij} \mid X_{ij} = x_{ij}] = \int E[Y_{ij} \mid X_{ij} = x_{ij}, \beta_{0i}] f(\beta_{0i}) d\beta_{0i}$

- The parameters (coefficients) of these models are estimated by *generalized estimating equations(GEE).*

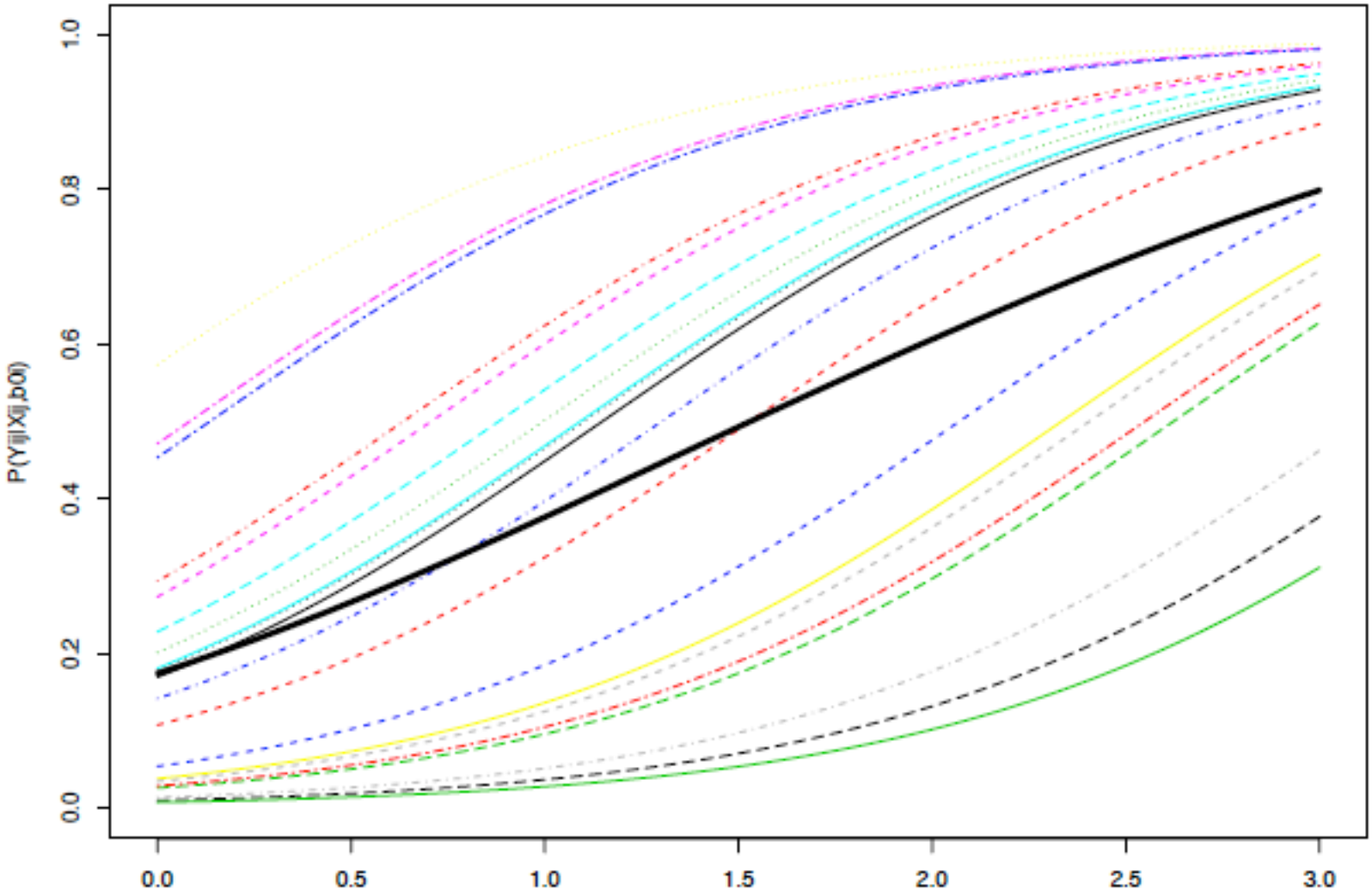
Parameter Interpretation in a marginal model

- Parameters in an equivalent random effects and GEE model have subtly different interpretations.

- Coefficients in a random effects model can be interpreted as expected “differences” (odds ratios, relative risks, etc) *within an individual*, given a change in their *X* from one value to another

- Coefficients in a marginal model represent expected differences (odds ratios, relative risks, etc) comparing populations defined by differences in *X.*

Figure 5.1: Difference Between Subject-Specific Logistic Regressions and Marginal Version: the individual thin lines represent curves for different individuals - the thick black line represents the marginal probability of disease, averaged over the individual curves



Marginal Models (GEE)

- GEE software typically allows several different working correlation models (e.g., exchangeable, auto-regressive, unstructured, etc.) that are used in the estimating phase of GEE.
- These correlation models are used to build weight matrices, which are used in a weighted regression.
- When deriving inferences for the coefficients, though, it calculates "robust" standard errors, which does not assume one guessed the right correlation model.

Examples of Correlation Models

$$V = \sigma^2 \begin{bmatrix} R_{01} & 0 & 0 & \cdots & 0 \\ 0 & R_{02} & 0 & \cdots & 0 \\ 0 & 0 & R_{03} & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & R_{0n} \end{bmatrix}$$

- *R01* is correlation matrix for individual *i=1*.
- Implies each individual is independent of all others.
- Correlation within individuals across longitudinal observations has the same form.

Structure for R_0

- General structure:

$$R_0 = \begin{bmatrix} 1 & \rho_{12} & \rho_{13} & \cdots & \rho_{1n} \\ \rho_{12} & 1 & \rho_{23} & \cdots & \rho_{2n} \\ \rho_{13} & \rho_{23} & 1 & \cdots & \rho_{3n} \\ \vdots & \vdots & \vdots & 1 & \vdots \\ \rho_{1n} & \rho_{2n} & \rho_{3n} & \cdots & 1 \end{bmatrix}$$

- A lot of unknown parameters

Uniform correlation (compound symmetry or exchangeable)

$$R_0 = \begin{bmatrix} 1 & \rho & \rho & \cdots & \rho \\ \rho & 1 & \rho & \cdots & \rho \\ \rho & \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \vdots & 1 & \vdots \\ \rho & \rho & \rho & \cdots & 1 \end{bmatrix}$$

- Can be inspired from random effects model

$$Y_{ij} = \alpha + \alpha_i + \beta x_{ij} + e_{ij}$$

Errors e_{ij} uncorrelated, and independent of x_{ij} and α_i

$$\rho = \frac{Var(\alpha_i)}{Var(\alpha_i) + Var(e_{ij})}$$

Correlation Models (contd):Time-Decaying Correlations (Auto-regressive)

$$R_0 = \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{n-3} \\ \vdots & \vdots & \vdots & 1 & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \cdots & 1 \end{bmatrix}$$

Auto-regressive: $e_{ij} = \rho e_{ij-1} + \eta_{ij}$

Not great for unequally spaced longitudinal data

Exponential correlation model generalizes this to $\text{corr}(y_{ij}, y_{ik}) = \rho^{|t_j - t_k|}$ rather than $\rho^{|j-k|}$

Examples of var-cov. models

Description	Abbrev.	Var-Cov. Matrix			
Compound Symmetry	CS	$\sigma^2+\sigma_0^2$	σ_0^2	σ_0^2	σ_0^2
		σ_0^2	$\sigma^2+\sigma_0^2$	σ_0^2	σ_0^2
		σ_0^2	σ_0^2	$\sigma^2+\sigma_0^2$	σ_0^2
		σ_0^2	σ_0^2	σ_0^2	$\sigma^2+\sigma_0^2$
Unstructured	UN	σ_1^2	σ_{12}	σ_{13}	σ_{14}
		σ_{12}	σ_2^2	σ_{23}	σ_{24}
		σ_{13}	σ_{23}	σ_3^2	σ_{34}
		σ_{14}	σ_{24}	σ_{34}	σ_4^2
Autoregressive	AR(1)	σ^2	$\rho\sigma^2$	$\rho^2\sigma^2$	$\rho^3\sigma^2$
		$\rho\sigma^2$	σ^2	$\rho\sigma^2$	$\rho^2\sigma^2$
		$\rho^2\sigma^2$	$\rho\sigma^2$	σ^2	$\rho\sigma^2$
		$\rho^3\sigma^2$	$\rho^2\sigma^2$	$\rho\sigma^2$	σ^2
Banded Diagnonal	UN(1)	σ_1^2	0	0	0
		0	σ_2^2	0	0
		0	0	σ_3^2	0
		0	0	0	σ_4^2
Spatial Power	SP(POW)(c)	σ^2	$\rho^{d12}\sigma^2$	$\rho^{d13}\sigma^2$	$\rho^{d14}\sigma^2$
		$\rho^{d12}\sigma^2$	σ^2	$\rho^{d23}\sigma^2$	$\rho^{d24}\sigma^2$
		$\rho^{d13}\sigma^2$	$\rho^{d23}\sigma^2$	σ^2	$\rho^{d34}\sigma^2$
		$\rho^{d14}\sigma^2$	$\rho^{d24}\sigma^2$	$\rho^{d34}\sigma^2$	σ^2

The GEE Algorithm

- Algorithm is similar to the one used for the non-repeated measures problems (e.g., OLS for continuous data, logistic regression for binary and Poisson regression for counts).

- Let $R(\alpha)$ be a $n_i \times n_i$ "working" correlation matrix that is fully characterized by a vector of parameters, α.

- V_i is again the variance-covariance of the observations which will be a function of the mean ($E(Y_i|X_i)$), a scale parameter, ϕ and $R(\alpha)$.

Standard Errors of Coefficients

- GEE will normally return two estimates of the variance of the coefficient estimates, 1) naive and 2) robust.
- Naive assumes that the chosen model for R(α), such as compound symmetry, is correct.
- Robust is a more nonparametric estimate that does not assume your guess for *R(α)* is correct. However, its variance estimates can be more variable.

GEE Marginal Model for Teenage Sex and Drug-Use

$$\log it[P(Y_{ij}=1 \mid X_{ij}=x_{ij})] = \log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) = \log\left(\frac{P(Y_{ij}=1 \mid X_{ij}=x_{ij})}{P(Y_{ij}=0 \mid X_{ij}=x_{ij})}\right) = \beta_0^M + \beta_1^M x_{ij}$$

- $var(Y_{ij}) = \mu_{ij}(1-\mu_{ij})^*$, $corr(Y_{ij}, Y_{ik}) = \rho$ (i.e., assume compound symmetry).
- $exp(\beta_1^M)$ is a ratio of population frequencies, i.e., it is a population averaged parameter. It is the odds ratio of the probabilities (proportions) of teenagers who would engage in sexual activity in populations reporting drug use vs. populations not reporting drug-use.
- * Semi-robust inference (relies on variance model)

Sexual Activity and drug/alcohol use among teenagers revisted

Main Variables

sex24hrs - sex in last 24 hrs. (0=no, 1=yes)

drgalcoh - drug or alcohol use in last 24 hrs.

tues-sun - dummy variables designating day of week

Results using xtgee in STATA

robust SE

```
. xtgee sx24hrs drgalcoh, eform i(id) family(binomial) cor(ind) robust

GEE population-averaged model                   Number of obs      =      1708
Group variable:                           id    Number of groups   =       109
Link:                                  logit    Obs per group: min =         1
Family:                             binomial                   avg =      15.7
Correlation:                     independent                   max =        33

                                (standard errors adjusted for clustering on id)
------------------------------------------------------------------------------
             |               Semi-robust
     sx24hrs | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
exp(β1^M) drgalcoh 1.739521   .3149874     3.06   0.002     1.219823    2.480635
------------------------------------------------------------------------------
```

non-robust (naive) SE

```
. xtgee sx24hrs drgalcoh, eform i(eid) family(binomial) cor(ind)

------------------------------------------------------------------------------
     sx24hrs | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    drgalcoh |   1.739521     .20244     4.76   0.000     1.384744    2.185194
------------------------------------------------------------------------------
```

xtgee Options

- family(?), link(?) -- identify that we wish linear regression with continuous outcome (as compared to, say, binary outcomes – more later)
- corr(ind) -- identify that we will assume independence for our correlation structure (some other possibilities include exchangeability and autoregressive structures)
- i(?)--identify which variable indentifies the individual (or cluster)
- ro -- identifies that we wish robust estimates of variability

Model 2 – same marginal model, different working correlation.

$$\log it[P(Y_{ij}=1 \mid X_{ij}=x_{ij})] = \log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) = \log\left(\frac{P(Y_{ij}=1 \mid X_{ij}=x_{ij})}{P(Y_{ij}=0 \mid X_{ij}=x_{ij})}\right) = \beta_0^M + \beta_1^M x_{ij}$$

x_{ij} = 0 if drug/alcohol use is no, 1 if yes

y_{ij} = 0 if no sex in last 24 hours, 1 if yes

$cor(Y_{ij}, Y_{ij'})=\rho$ (compound symmetry or exchangeable correlation structure)

Results of Model 2 using STATA

robust SE

```
. xtgee sx24hrs drgalcoh, eform i(id) family(binomial) cor(exc) robust

GEE population-averaged model                   Number of obs      =      1708
Group variable:                            id   Number of groups   =       109
Link:                                   logit   Obs per group: min =         1
Family:                              binomial                   avg =      15.7
Correlation:                     exchangeable                   max =        33
                                 (standard errors adjusted for clustering on id)
------------------------------------------------------------------------------
             |             Semi-robust
     sx24hrs | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    drgalcoh |   1.393705   .1919735     2.41   0.016     1.063956    1.825653
------------------------------------------------------------------------------
```

non-robust (naive) SE

```
. xtgee sx24hrs drgalcoh, eform i(eid) family(binomial) cor(exc)
------------------------------------------------------------------------------
     sx24hrs | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    drgalcoh |   1.393705   .1701631     2.72   0.007     1.097095    1.770507
------------------------------------------------------------------------------
```

Estimated Working Correlation

```
. xtcorr
             c1          c2          c3          c4          c5          c6          c7          c8          c9
 r1      1.0000
 r2      0.1614      1.0000
 r3      0.1614      0.1614      1.0000
 r4      0.1614      0.1614      0.1614      1.0000
 r5      0.1614      0.1614      0.1614      0.1614      1.0000
 r6      0.1614      0.1614      0.1614      0.1614      0.1614      1.0000
 r7      0.1614      0.1614      0.1614      0.1614      0.1614      0.1614      1.0000
 r8      0.1614      0.1614      0.1614      0.1614      0.1614      0.1614      0.1614      1.0000
 r9      0.1614      0.1614      0.1614      0.1614      0.1614      0.1614      0.1614      0.1614      1.0000
r10      0.1614      0.1614      0.1614      0.1614      0.1614      0.1614      0.1614      0.1614      0.1614
r11      0.1614      0.1614      0.1614      0.1614      0.1614      0.1614      0.1614      0.1614      0.1614
r12      0.1614      0.1614      0.1614      0.1614      0.1614      0.1614      0.1614      0.1614      0.1614
r13      0.1614      0.1614      0.1614      0.1614      0.1614      0.1614      0.1614      0.1614      0.1614
r14      0.1614      0.1614      0.1614      0.1614      0.1614      0.1614      0.1614      0.1614      0.1614
r15      0.1614      0.1614      0.1614      0.1614      0.1614      0.1614      0.1614      0.1614      0.1614
r16      0.1614      0.1614      0.1614      0.1614      0.1614      0.1614      0.1614      0.1614      0.1614
r17      0.1614      0.1614      0.1614      0.1614      0.1614      0.1614      0.1614      0.1614      0.1614
r18      0.1614      0.1614      0.1614      0.1614      0.1614      0.1614      0.1614      0.1614      0.1614
r19      0.1614      0.1614      0.1614      0.1614      0.1614      0.1614      0.1614      0.1614      0.1614
```

Model 3 – adjusting for day of week

$$\log it[P(Y_{ij}=1 \mid x_{ij}, day_{ij})] = \beta_0 + \beta_1 x_{ij} + \gamma_1 z_{1ij} + \gamma_2 z_{2ij} + \cdots + \gamma_6 z_{6ij}$$

x_{ij} = 1 if drug/alcohol use is yes, 0 if no

z_{1ij} = 1 if interview day is Tuesday, 0 if not

z_{2ij} = 1 if interview day is Wed., 0 if not.....

z_{6ij} = 1 if interview day is Sunday, 0 if not

y_{ij} = 1 if sex in last 24 hours, 0 if no

$cor(Y_{ij}, Y_{ij'}) = \rho$ (compound symmetry or exchangeable correlation structure)

Results of Model 3 using STATA

```
. xtgee sx24hrs drgalcoh tues wed thur fri sat sun, eform i(id) family(binomial
> ) cor(exc) robust


GEE population-averaged model                   Number of obs      =      1708
Group variable:                        id       Number of groups   =       109
Link:                               logit       Obs per group: min =         1
Family:                          binomial                      avg =      15.7
Correlation:                 exchangeable                      max =        33
                                                Wald chi2(7)       =     11.40
Scale parameter:                        1       Prob > chi2        =    0.1220

                            (standard errors adjusted for clustering on id)
------------------------------------------------------------------------------
             |               Semi-robust
     sx24hrs | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    drgalcoh |   1.373029   .1845197     2.36   0.018     1.055086    1.786782
        tues |   1.239246   .2320747     1.15   0.252     .8585234    1.788804
         wed |   1.234437   .2523307     1.03   0.303      .826942    1.842734
        thur |   1.099757    .233122     0.45   0.654     .7258761    1.666215
         fri |   .9833647   .1933837    -0.09   0.932     .6688388    1.445799
         sat |   1.277403   .2490991     1.26   0.209     .8716457    1.872043
         sun |   1.577958    .306514     2.35   0.019     1.078331     2.30908
------------------------------------------------------------------------------
```

Model for drug/alcohol use vs. day of week

$$\log it[P(X_{ij}=1 \mid day_{ij})] = \gamma^*_0 + \gamma^*_1 z_{1ij} + \gamma^*_2 z_{2ij} + \cdots + \gamma_6{}^* z_{6ij}$$

X_{ij} = 1 if drug/alcohol use is yes, 0 if no

z_{1ij} = 1 if interview day is Tuesday, 0 if not

z_{2ij} = 1 if interview day is Wed., 0 if not.....

z_{6ij} = 1 if interview day is Sunday, 0 if not

$cor(Y_{ij}, Y_{ij'}) = \rho$ (compound symmetry or exchangeable correlation structure)

Results of drug/alcohol use Model using STATA

```
. xtgee drgalcoh tues wed thur fri sat sun, eform i(id) family(binomial) cor(ex
> c) robust

GEE population-averaged model                   Number of obs      =      1708
Group variable:                        id       Number of groups   =       109
Link:                               logit       Obs per group: min =         1
Family:                          binomial                      avg =      15.7
Correlation:                 exchangeable                      max =        33
                                                Wald chi2(6)       =     28.91
Scale parameter:                        1       Prob > chi2        =    0.0001

                            (standard errors adjusted for clustering on id)
------------------------------------------------------------------------------
             |             Semi-robust
    drgalcoh | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        tues |   .7484218   .1301296    -1.67   0.096     .5322875    1.052317
         wed |   .7043399   .1440654    -1.71   0.087     .4717131    1.051687
        thur |   .9226514    .171617    -0.43   0.665     .6407825    1.328509
         fri |   1.197263   .2206008     0.98   0.329      .834357    1.718015
         sat |   1.666645   .3147173     2.71   0.007     1.151088    2.413115
         sun |   1.371219    .205994     2.10   0.036     1.021488    1.840688
------------------------------------------------------------------------------
```

Continuous Outcome Example (Linear Model): Respiratory Function

- Random sample of 300 girls from Topeka
- Measurements of fev_1, height, age (fev_1 is forced expired volume in first second after spirometry in ml)

OLS -- ignores correlation (no robust variabilty estimates)

```
. xtgee  lnfev age, family(gaussian) link(id) corr(ind) i( childid)

GEE population-averaged model                   Number of obs      =      1994
Group variable:                      childid    Number of groups   =       300
Link:                                identity   Obs per group: min =         1
Family:                              Gaussian                  avg =       6.6
Correlation:                       independent                 max =        12
                                                Wald chi2(1)       =   6299.69
Scale parameter:                     .0262556   Prob > chi2        =    0.0000

Pearson chi2(1994):                    52.35   Deviance           =     52.35
Dispersion (Pearson):                .0262556  Dispersion         =  .0262556
```

lnfev	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	.0866927	.0010923	79.37	0.000	.084552	.0888335
_cons	-.2741518	.014197	-19.31	0.000	-.3019775	-.2463261

(Same as OLS on entire data set)

```
. regress  lnfev age

------------------------------------------------------------------------------
       lnfev |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0866927   .0010928    79.33   0.000     .0845496    .0888359
       _cons |  -.2741518   .0142042   -19.30   0.000    -.3020084   -.2462953
------------------------------------------------------------------------------
```

OLS with Robust Variability Estimates

```
. xtgee  lnfev age, family(gaussian) link(id) corr(ind) i( childid) ro

GEE population-averaged model                   Number of obs      =      1994
Group variable:                      childid    Number of groups   =       300
Link:                                identity   Obs per group: min =         1
Family:                              Gaussian                  avg =       6.6
Correlation:                         independent               max =        12

                                                (standard errors adjusted for clustering on childid)
------------------------------------------------------------------------------
             |               Semi-robust
       lnfev |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0866927   .0011288    76.80   0.000     .0844804    .0889051
       _cons |  -.2741518   .0158196   -17.33   0.000    -.3051577   -.2431459
------------------------------------------------------------------------------
```

.0011288 as compared to non-robust .0010923 (and .0158 vs .0142)

Review of Modeling Longitudinal vs. X-sectional Associations

- Consider the model:

$$E[Y_{ij} \mid X_{i1} = x_{i1}, X_{ij} = x_{ij}] =$$

$$\beta_0 + \beta_C x_{i1} + \beta_L (x_{ij} - x_{i1})$$

- β_L represents the expected change in Y given a change in X_{ij} relative to the baseline value (X_{i1}) - longitudinal effect.

- β_C represents the expected difference in average Y across two sub-populations that differ by their baseline values, X_{i1} - cross-sectional effect.

Alternative Parameterization

- An identical fit to the data would be:

$$E[Y_{ij} \mid X_{i1} = x_{i1}, X_{ij} = x_{ij}] =$$

$$\beta_0 + \beta_C^* x_{i1} + \beta_L x_{ij}$$

- β_L still represents the expected change in Y given a change in X_{ij} relative to the baseline value (X_{i1}) - longitudinal effect.

- β^*_C represents the difference in the x-sectional vs. longitudinal (or $\beta^*_C = \beta_C - \beta_L$).

Model for Lung Function

- Consider the model:

$$E[Y_{ij} \mid X_{i11} = x_{i11}, X_{ij1} = x_{ij1}, X_{i12} = x_{i12}, X_{ij2} = x_{ij2}] =$$

$$\beta_0 + \beta_1 x_{i11} + \beta_2 x_{ij1} + \beta_3 x_{i12} + \beta_3 x_{ij2}$$

with X_{ij1} height for subject i, time j, and X_{ij2} is the corresponding age.

More complicated Model -- still OLS

```
xtgee  lnfev  lnheight age initlnheight initage, family(gaussian) link(id)
corr(ind) i( childid)

GEE population-averaged model                   Number of obs      =      1994
Group variable:                       childid   Number of groups   =       300
Link:                                identity   Obs per group: min =         1
Family:                              Gaussian                  avg =       6.6
Correlation:                      independent                  max =        12
                                                Wald chi2(4)       =  14199.25
Scale parameter:                     .0134473   Prob > chi2        =    0.0000

------------------------------------------------------------------------------
       lnfev |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   lnheight |   2.056183   .0699129    29.41   0.000     1.919156     2.19321
        age |   .0284979   .0021109    13.50   0.000     .0243606    .0326352
initlnheight |   .4074967   .0839699     4.85   0.000     .2429187    .5720746
    initage |   -.016087   .0040224    -4.00   0.000    -.0239708   -.0082032
      _cons |  -.3309375     .02105   -15.72   0.000    -.3721947   -.2896803
------------------------------------------------------------------------------
```

More complicated Model, different parameterization

```
 xtgee  lnfev  lnheightchange agechange initlnheight initage, family(gaussian) link(id) corr(ind)
           i(childid)

Iteration 1: tolerance = 1.427e-13

GEE population-averaged model                 Number of obs      =      1994
Group variable:                     childid   Number of groups   =       300
Link:                              identity   Obs per group: min =         1
Family:                            Gaussian                 avg =       6.6
Correlation:                    independent                 max =        12
                                               Wald chi2(4)       =  14199.25
Scale parameter:                   .0134473   Prob > chi2        =    0.0000

Pearson chi2(1994):                  26.81   Deviance           =     26.81
Dispersion (Pearson):              .0134473   Dispersion         =  .0134473

------------------------------------------------------------------------------
       lnfev |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
lnheightchange   2.056183   .0699129    29.41   0.000     1.919156     2.19321
   agechange |   .0284979   .0021109    13.50   0.000     .0243606    .0326352
initlnheight |    2.46368   .0649965    37.90   0.000     2.336289    2.591071
     initage |   .0124109    .003436     3.61   0.000     .0056765    .0191453
       _cons |  -.3309375     .02105   -15.72   0.000    -.3721947   -.2896803
------------------------------------------------------------------------------
```

More complicated Model -- still OLS + Robust

```
xtgee  lnfev  lnheight age initlnheight initage, family(gaussian) link(id) corr(ind) i( childid) ro

GEE population-averaged model                   Number of obs      =      1994
Group variable:                       childid   Number of groups   =       300
Link:                                identity   Obs per group: min =         1
Family:                              Gaussian                  avg =       6.6
Correlation:                      independent                  max =        12

                                              (standard errors adjusted for clustering on childid)
------------------------------------------------------------------------------
             |               Semi-robust
       lnfev |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    lnheight |   2.056183   .0792847    25.93   0.000     1.900788    2.211578
         age |   .0284979   .0022755    12.52   0.000      .024038    .0329578
initlnheight |   .4074967   .1828943     2.23   0.026     .0490305    .7659628
     initage |   -.016087    .008835    -1.82   0.069    -.0334034    .0012293
       _cons |  -.3309375   .0432665    -7.65   0.000    -.4157383   -.2461367
------------------------------------------------------------------------------
```

More complicated Model, different parameterization

```
 . xtgee  lnfev  lnheightchange agechange initlnheight initage, family(gaussian) link(id) corr(ind)
i( childid) ro

Iteration 1: tolerance = 1.427e-13

GEE population-averaged model                   Number of obs      =      1994
Group variable:                      childid    Number of groups   =       300
Link:                               identity    Obs per group: min =         1
Family:                             Gaussian                   avg =       6.6
Correlation:                     independent                   max =        12
                                                Wald chi2(4)       =  11417.58
Scale parameter:                    .0134473    Prob > chi2        =    0.0000

Pearson chi2(1994):                    26.81    Deviance           =     26.81
Dispersion (Pearson):               .0134473    Dispersion         =  .0134473

                          (standard errors adjusted for clustering on childid)
------------------------------------------------------------------------------
             |             Semi-robust
       lnfev |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
lnheightch~e |   2.056183   .0792847    25.93   0.000     1.900788    2.211578
   agechange |   .0284979   .0022755    12.52   0.000      .024038    .0329578
initlnheight |    2.46368   .1775394    13.88   0.000     2.115709    2.811651
     initage |   .0124109   .0087532     1.42   0.156    -.0047451    .0295668
       _cons |  -.3309375   .0432665    -7.65   0.000    -.4157383   -.2461367
------------------------------------------------------------------------------
```

Comparison of Standard Errors

Variable	Naïve SE	Robust SE	Naïve z	Robust z
lnheight	.0699	.0793	29.4	25.9
age	.0021	.0023	13.5	12.5
initlnheight	.0840	.1829	4.8	2.2
initage	.0040	.0088	-4.0	-1.8
_cons	.0211	.0433	-15.7	-7.6