Comparison of estimation procedures for linear and non-linear outcome

I simulated a simple dataset from a cross-over trial comparing treatment to placebo (trt = 1 for treatment, 0 for placebo). The period variable indicates the ordering of the treatments. There are two outcomes of interest: alcohol consumption (Y, a continuous measure) and alcohol dependence (AD, a binary measure: 1 yes, 0 no).

These hypothetical data are available at the course website:
http://www.biostat.jhsph.edu/~ejohnson/multilevel.htm

Linear model of Y as a function of period and treatment

1. First regress Y on period and treatment ignoring the correlation in the data (i.e. ordinary least squares.

```
reg Y period trt
```

```
      Source |       SS       df       MS              Number of obs =      30
-------------+------------------------------           F(  2,    27) =    1.83
       Model |  52.0714286     2  26.0357143           Prob > F      =  0.1798
    Residual |  384.228571    27  14.2306878           R-squared     =  0.1193
-------------+------------------------------           Adj R-squared =  0.0541
       Total |       436.3    29  15.0448276           Root MSE      =  3.7724

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      period |   2.571429   1.380542     1.86   0.073    -.2612092    5.404066
         trt |  -.4285714   1.380542    -0.31   0.759    -3.261209    2.404066
       _cons |   15.22857   1.220997    12.47   0.000     12.72329    17.73385
------------------------------------------------------------------------------
```

2. Now fit a marginal model using GEE where we specify an independence working correlation structure. This will provide us with a "robust" estimate of variance.

```
xtgee Y period trt, i(id) corr(ind)

Iteration 1: tolerance = 1.477e-15

GEE population-averaged model                   Number of obs      =        30
Group variable:                         id      Number of groups   =        15
Link:                             identity      Obs per group: min =         2
Family:                           Gaussian                     avg =       2.0
Correlation:                  independent                      max =         2
                                                Wald chi2(2)       =      4.07
Scale parameter:                  12.80762      Prob > chi2        =    0.1310

Pearson chi2(30):                   384.23      Deviance           =    384.23
Dispersion (Pearson):             12.80762      Dispersion         =  12.80762

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      period |   2.571429   1.309697     1.96   0.050     .0044697    5.138387
         trt |  -.4285714   1.309697    -0.33   0.743     -2.99553    2.138387
       _cons |   15.22857    1.15834    13.15   0.000     12.95827    17.49888
------------------------------------------------------------------------------
```

3. Fit another GEE model but assume an exchangeable correlation structure. In this case, since we only have two observations per person, this is just like estimating the correlation between the two observations and using this information in the model. In general cases where you have more than two observations per person over time, this correlation structure assumes that there is no meaning to the "times" or that all the observations from the same person are exchangeable over time.

```
xtgee Y period trt, i(id) corr(exch)

Iteration 1: tolerance = 2.114e-15

GEE population-averaged model                   Number of obs      =        30
Group variable:                         id      Number of groups   =        15
Link:                             identity      Obs per group: min =         2
Family:                           Gaussian                     avg =       2.0
Correlation:                  exchangeable                     max =         2
                                                Wald chi2(2)       =      6.89
Scale parameter:                  12.80762      Prob > chi2        =    0.0320

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      period |   2.571429   1.006357     2.56   0.011     .5990044    4.543853
         trt |  -.4285714   1.006357    -0.43   0.670    -2.400996    1.543853
       _cons |   15.22857   1.068604    14.25   0.000     13.13415      17.323
------------------------------------------------------------------------------
```

After running xtgee, you can obtain the estimate of the correlation using the "xtcorr" command. Below we see that the estimate of the correlation in Y among subjects (within subject correlation) is approximately 0.41.

```
xtcorr

Estimated within-id correlation matrix R:

        c1      c2
r1  1.0000
r2  0.4096  1.0000
```

4. Now fit a subject specific random effects model. Here we used the xtreg command, you could also use xtmixed and gllamm. Sometimes the estimation is difficult and the xtreg and xtmixed commands will not provide you with a solution. In those cases, try gllamm. This command uses a different estimation procedure and generally works in cases where xtreg and xtmixed don't.

```
xtreg Y period trt, re i(id)

Random-effects GLS regression                   Number of obs      =        30
Group variable (i): id                          Number of groups   =        15

R-sq:  within  = 0.3146                         Obs per group: min =         2
       between = 0.0000                                        avg =       2.0
       overall = 0.1193                                        max =         2

Random effects u_i ~ Gaussian                   Wald chi2(2)       =      5.97
corr(u_i, X)       = 0 (assumed)                Prob > chi2        =    0.0506

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      period |   2.571429   1.081001     2.38   0.017     .4527052    4.690152
         trt |  -.4285714   1.081001    -0.40   0.692    -2.547295    1.690152
       _cons |   15.22857   1.126053    13.52   0.000     13.02155     17.4356
-------------+----------------------------------------------------------------
     sigma_u |  2.3040814
     sigma_e |  2.9538576
         rho |  .37827891   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

Sigma_u is the estimate of the standard deviation in Y across subjects; represents subject to subject variability in the outcome Y. Sigma_e is the within subject statistical standard deviation. Rho provides the fraction of the total variance that is explained by subject to subject variability: 2.30^2 / (2.95^2+2.30^2) = 0.38.

Logistic model of Y as a function of period and treatment

5. Fit the logistic regression model of odds of Y as a function of period and treatment. Assumes independence of observations within subject. The constant or intercept of the model is the log odds of AD among persons not receiving the treatment during period 0. The period and treatment coefficients are log odds ratios!

```
logit AD period trt

Iteration 0:   log likelihood = -20.727699
Iteration 1:   log likelihood = -17.809803
Iteration 2:   log likelihood = -17.777181
Iteration 3:   log likelihood = -17.777121

Logistic regression                                     Number of obs   =         30
                                                        LR chi2(2)      =       5.90
                                                        Prob > chi2     =     0.0523
Log likelihood = -17.777121                             Pseudo R2       =     0.1423

------------------------------------------------------------------------------
          AD |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      period |   .7927895   .8336501     0.95   0.342    -.8411347    2.426714
         trt |   1.821201   .8351845     2.18   0.029     .1842697    3.458133
       _cons |  -1.142702   .7496396    -1.52   0.127    -2.611969    .3265641
------------------------------------------------------------------------------
```

6. Fit the logistic regression model using GEE specifying an independence working correlation structure. Again, this provides a robust variance estimate.

```
xtgee AD period trt, family(binomial) i(id) corr(ind)

Iteration 1: tolerance = 3.306e-11

GEE population-averaged model                   Number of obs      =        30
Group variable:                         id      Number of groups   =        15
Link:                                logit      Obs per group: min =         2
Family:                           binomial                     avg =       2.0
Correlation:                   independent                     max =         2
                                                Wald chi2(2)       =      4.97
Scale parameter:                         1      Prob > chi2        =    0.0833

Pearson chi2(30):                    29.83      Deviance           =     35.55
Dispersion (Pearson):              .9942692     Dispersion         =  1.185141

------------------------------------------------------------------------------
          AD |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      period |   .7927895   .8336523     0.95   0.342    -.8411389    2.426718
         trt |   1.821201   .8351866     2.18   0.029     .1842654    3.458137
       _cons |  -1.142702   .7496417    -1.52   0.127    -2.611973    .3265683
------------------------------------------------------------------------------
```

7. Now fit the GEE model with an exchangeable correlation structure

```
xtgee AD period trt, family(binomial) i(id) corr(exch)

Iteration 1: tolerance = .01978125
Iteration 2: tolerance = .00041008
Iteration 3: tolerance = .00001051
Iteration 4: tolerance = 1.605e-07

GEE population-averaged model                   Number of obs      =        30
Group variable:                         id      Number of groups   =        15
Link:                                logit      Obs per group: min =         2
Family:                           binomial                     avg =       2.0
Correlation:                  exchangeable                     max =         2
                                               Wald chi2(2)       =      3.21
Scale parameter:                         1     Prob > chi2        =    0.2012

------------------------------------------------------------------------------
          AD |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      period |   .7580283   1.025071     0.74   0.460    -1.251074    2.767131
         trt |   1.801878   1.027685     1.75   0.080    -.2123477    3.816104
       _cons |  -1.116056   .8302357    -1.34   0.179    -2.743288    .5111761
------------------------------------------------------------------------------
```

8. Fit the subject specific random effects model using gllamm. Remember, you have to specify that the data is binomial; otherwise you get estimates assuming your outcome is continuous.

```
gllamm AD period trt, i(id) f(binomial) l(logit) adapt

number of level 1 units = 30
number of level 2 units = 15

Condition Number = 3.4900529

gllamm model

log likelihood = -17.777121

------------------------------------------------------------------------------
          AD |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      period |   .7927886   .8336527     0.95   0.342    -.8411407    2.426718
         trt |   1.821207   .8351871     2.18   0.029     .1842706    3.458144
       _cons |  -1.142706   .7496422    -1.52   0.127    -2.611977    .3265661
------------------------------------------------------------------------------


Variances and covariances of random effects
------------------------------------------------------------------------------

***level 2 (id)

    var(1): 3.322e-11 (6.531e-06)
------------------------------------------------------------------------------
```

Here we estimate that the variance of the random subject effects is 3.322e-11, very close to zero. So that it is not strange that our estimates are very similar to the ordinary logistic regression estimates.

Same modeling procedures using SAS:

Linear model

```
proc genmod;
class id;
model Y period trt;
run;

proc genmod;
class id;
model Y period trt;
repeated id / type=ind;
run;

proc genmod;
class id;
model Y period trt;
repeated id / type=exch;
run;

proc mixed;
class id;
model Y period trt;
random intercept / subject=id;
run;
```

Logistic model

```
proc genmod;
class id;
model AD period trt / dist=binomial;
run;

proc genmod;
class id;
model AD period trt / dist=binomial;
repeated id / type=ind;
run;

proc genmod;
class id;
model AD period trt / dist=binomial;
repeated id / type=exch;
```

```
run;

proc glimmix;
class id;
model AD period trt / dist=binomial solution;
random intercept / subject=id;
run;
```