

Fitting generalized estimating equation (GEE) regression models in Stata

Nicholas Horton
horton@bu.edu
Dept of Epidemiology and Biostatistics
Boston University
School of Public Health

3/16/2001

Nicholas Horton, BU SPH

1

Outline

- Regression models for clustered or longitudinal data
- Brief review of GEEs
 - mean model
 - working correlation matrix
- Stata GEE implementation
- Example: Mental health service utilization
- Summary and conclusions

3/16/2001

Nicholas Horton, BU SPH

2

Regression models for clustered or longitudinal data

- Longitudinal, repeated measures, or clustered data commonly encountered
- Correlations between observations on a given subject may exist, and need to be accounted for
- If outcomes are multivariate normal, then established methods of analysis are available (Laird and Ware, Biometrics, 1982)
- If outcomes are binary or counts, likelihood based inference less tractable

Generalized estimating equations

- Described by Liang and Zeger (Biometrika, 1986) and Zeger and Liang (Biometrics, 1986) to extend the generalized linear model to allow for correlated observations
- Characterize the marginal expectation (average response for observations sharing the same covariates) as a function of covariates
- Method accounts for the correlation between observations in generalized linear regression models by use of empirical (sandwich/robust) variance estimator
- Posits model for the working correlation matrix

The marginal mean model

- We assume the marginal regression model:

$$g(E[Y_{ij} | x_{ij}]) = x'_{ij} \beta$$

- Where x_{ij} is a p times 1 vector of covariates, β consists of the p regression parameters of interest, $g(\cdot)$ is the link function, and Y_{ij} denotes the j th outcome (for $j=1, \dots, J$) for the i th subject (for $i=1, \dots, N$)
- Common choices for the link function include:
 - $g(a)=a$ (identity link)
 - $g(a)=\log(a)$ [for count data]
 - $g(a)=\log(a/(1-a))$ [logit link for binary data]

Model for the correlation

- Assuming no missing data, the $J \times J$ covariance matrix for Y is modeled as:

$$V_i = \phi A_i^{1/2} R(\alpha) A_i^{1/2}$$

- Where ϕ is a glm dispersion parameter, A is a diagonal matrix of variance functions, and $R(\alpha)$ is the working correlation matrix of Y

Model for the correlation (cont.)

- If mean model is correct, correlation structure may be misspecified, but parameter estimates remain consistent
- Liang and Zeger showed that modeling correlation may boost efficiency
- But this is a large sample result; there must be enough clusters to estimate these parameters
- Variety of models that are supported in Stata

Model for the correlation (cont.)

- Independence

$$R(\alpha) = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

- Number of parameters: 0

Model for the correlation (cont.)

- Exchangeable (compound symmetry)

$$R(\alpha) = \begin{pmatrix} 1 & \alpha & \cdots & \alpha \\ \alpha & 1 & \cdots & \alpha \\ \vdots & \vdots & \ddots & \vdots \\ \alpha & \alpha & \cdots & 1 \end{pmatrix}$$

- Number of parameters: 1

Model for the correlation (cont.)

- Unstructured

$$R(\alpha) = \begin{pmatrix} 1 & \alpha_{12} & \cdots & \alpha_{1J} \\ \alpha_{12} & 1 & \cdots & \alpha_{2J} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{1J} & \alpha_{2J} & \cdots & 1 \end{pmatrix}$$

- Number of parameters: $J(J-1)/2$

Model for the correlation (cont.)

- Auto-regressive

$$R(\alpha) = \begin{pmatrix} 1 & \alpha & \cdots & \alpha^{J-1} \\ \alpha & 1 & \cdots & \alpha^{J-2} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha^{J-1} & \alpha^{J-2} & \cdots & 1 \end{pmatrix}$$

- Number of parameters: 1

Model for the correlation (cont.)

- Stationary (g-dependent)

$$R(\alpha) = \begin{pmatrix} 1 & \alpha_1 & \cdots & \alpha_{J-1} \\ \alpha_1 & 1 & \cdots & \alpha_{J-2} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{J-1} & \alpha_{J-2} & \cdots & 1 \end{pmatrix}$$

- Number of parameters: $0 < g \leq J-1$

Model for the correlation (cont.)

- Fixed

$$R(\alpha) = \begin{pmatrix} 1 & c_{12} & \cdots & c_{1J} \\ c_{12} & 1 & \cdots & c_{2J} \\ \vdots & \vdots & \ddots & \vdots \\ c_{1J} & c_{2J} & \cdots & 1 \end{pmatrix}$$

- Number of parameters: 0 (user specified)

Model for the correlation (cont.)

- If J is small and data are balanced and complete, then an unstructured matrix is recommended
- If observations are mistimed, then use a structure that accounts for correlation as function of time (stationary, or auto-regressive)
- If observations are clustered (i.e. no logical ordering) then exchangeable may be appropriate
- If number of clusters small, independent may be best
- Issues discussed further in Diggle, Liang and Zeger (1994, book)

Missing data

- Standard GEE models assume that missing observations are Missing Completely at Random (MCAR) in the sense of Little and Rubin (book, 1987)
- Robins, Rotnitzky and Zhao (JASA, 1995) proposed methods to allow for data that is missing at random (MAR)
- These methods not yet implemented in standard software (requires estimation of weights and more complicated variance formula)

Variance estimators

- Empirical (aka *sandwich* or *robust/semi-robust*)
consistent when the mean model is correctly specified
(if no missing data)
- Model-based (aka *naïve*) [default in Stata]
consistent when both the mean model and the
covariance model are correctly specified

Syntax for xtgee

**xtgee depvar varlist, family(family) link(link) corr(corr)
i(idvar) t(timevar) robust**

Family: binomial, gaussian, gamma, igaussian, nbinomial,
poisson

Link: identity, cloglog, log, logit, nbinomial, opwer, power,
probit, reciprocal

Correlation: independent, exchangeable, ar#, stationary#,
nonstationary#,unstructured, fixed

Also options to change the scale parameter, use weighted
equations, specify offsets

Example: Mental Health Service Utilization

- Connecticut child studies (Zahner et al, AJPH, 1997)
- Outcome: use of general health, school, or mental health services (dichotomous report)
- Sample: 2,519 children
- Other dichotomous predictors: age, gender, academic problems

Data format and variables

```

                                A S           G
                                C E S M E
                                A T C E N
                                D T H N E S
O                                B O P I O T R E
B   I                            O L R N O A A R
S   D                            Y D O G L L L V

1  90111502  0  0  0  0  1  0  0  0
2  90111502  0  0  0  1  0  1  0  0
3  90111502  0  0  0  2  0  0  1  0
4  80111206  0  0  0  0  1  0  0  0
5  80111206  0  0  0  1  0  1  0  0
6  80111206  0  0  0  2  0  0  1  0
7  40111608  1  0  0  0  1  0  0  0
8  40111608  1  0  0  1  0  1  0  0
9  40111608  1  0  0  2  0  0  1  0
```

3/16/2001

Nicholas Horton, BU SPH

21

Stata code to fit model

```
iis id
tis setting
xtdes
xi: xtgee serv i.old*mental i.old*school
    i.boy*mental i.boy*school
    i.acadpro*mental i.acadpro*school,
    link(logit) corr(unst) family(binomial)
    robust
xtcorr
```

3/16/2001

Nicholas Horton, BU SPH

23

Describe cross-sectional data (xtdes)

```

id: 1, 2, ..., 2519          n =      2519
setting: 0, 1, ..., 2       T =        3
      Delta(type) = 1; (2-0)+1 = 3
      (id*setting uniquely identifies each observation)
  
```

```

Distribution of T_i: min   5%   25%   50%   75%   95%   max
                      3     3     3     3     3     3
  
```

```

-----+-----
Freq.  Percent   Cum. | Pattern
-----+-----
2519   100.00  100.00 | 111
-----+-----
2519   100.00           | XXX
  
```

(No missing data!)

3/16/2001

Nicholas Horton, BU SPH

24

```

GEE population-averaged model          Number of obs   =   7557
Group and time vars:      id setting   Number of groups =   2519
Link:                      logit       Obs per group: min =    3
Family:                    binomial     avg =             3.0
Correlation:              unstructured  max =             3
                                           Wald chi2(11)    =   605.12
Scale parameter:          1             Prob > chi2      =   0.0000
                                (standard errors adjusted for clustering on id)
  
```

```

-----+-----
serv |          Coef.   Semi-robust          z   P>|z|   [95% Conf. Interval]
-----+-----
  _Iold_1 |   .1233576   .1441123   0.86   0.392   -.1590973   .4058124
  mental |  -.3520988   .1933698  -1.82   0.069   -.7310967   .0268992
  _IoldXment~1 | .2905076   .189558   1.53   0.125   -.0810192   .6620344
  school |   .1850487   .1734874   1.07   0.286   -.1549804   .5250778
  _IoldXscho~1 | .330549   .162133   2.04   0.041   .0127742   .6483239
  _Iboy_1 |   .3652564   .1464068   2.49   0.013   .0783043   .6522084
  _IboyXment~1 | -.2779134   .1894824  -1.47   0.142   -.6492921   .0934654
  _IboyXscho~1 | -.1538587   .1650033  -0.93   0.351   -.4772592   .1695418
  _Iacadpro_1 |   .7239641   .1445971   5.01   0.000   .440559   1.007369
  _IacaXment~1 |   .1843236   .1911094   0.96   0.335   -.1902441   .5588912
  _IacaXscho~1 |   1.136088   .1669423   6.81   0.000   .8088873   1.463289
  _cons |  -2.944382   .1489399  -19.77   0.000   -3.236298  -2.652465
  
```

3/16/2001

Nicholas Horton, BU SPH

25

Estimates of working correlation (xtcorr)

Estimated within-id corr matrix R

```
      school mental general
      c1      c2      c3
r1  1.0000
r2  0.1646  1.0000
r3  0.1977  0.2270  1.0000
```

Multidimensional test of OLD effect

```
test _IoldXmenta_1=0
( 1) _IoldXmenta_1 = 0.0
     chi2( 1) = 2.35
     Prob > chi2 = 0.1254
test _IoldXschoo_1=0,accumulate
( 1) _IoldXschoo_1 = 0.0
( 2) _IoldXmenta_1 = 0.0
     chi2( 2) = 4.55
     Prob > chi2 = 0.1029 ←
test _Iold_1=0,accumulate
( 1) _IoldXschoo_1 = 0.0
( 2) _IoldXmenta_1 = 0.0
( 3) _Iold_1 = 0.0
     chi2( 3) = 20.61
     Prob > chi2 = 0.0001 ←
```

Results from Example

- There is a significant interaction between service setting and academic problems ($df=2, p<0.0001$), but not for age and setting ($df=2, p=0.10$) or gender and setting ($df=2, p=0.33$)
- Overall, a higher proportion of boys use services ($df=3, p=0.04$) and older children use them more than younger children ($df=3, p=0.0001$)

More resources

- Generalized estimating equations: an annotated bibliography (Ziegler, Kastner and Blettner, Biometrical Journal, 1998)
- Review of software to fit Generalized Estimating Equation regression models (Horton and Lipsitz, The American Statistician, 1999, article online at <http://www.biostat.harvard.edu/~horton/geereview.pdf>)