Model and Working Correlation Structure Selection in GEE Analyses of Longitudinal Data

Dr Jisheng Cui

Public Health Research Cluster

Deakin University

1. Introduction

- **GEE method - background information**
- **QIC - comparison with AIC**
- **Stata program - description of options**
- **Examples - Longitudinal data analyses**
- **Summary**

2. GEE method

- **Generalized Estimating Equations (GEE)**
  Liang and Zeger 1986; Biometrika
  Zeger and Liang 1986; Biometrics
  Zeger, Liang, Albert 1988; Biometrics

- **Extension of GLM**
  Nelder and Wedderburn 1972; JRSS A
  McCullagh and Nelder 1989; GLM (2nd Ed)

$$S(\beta; R, D) \equiv \sum_{i=1}^{n} D_i' V_i^{-1} (Y_i - \mu_i) = 0$$

- **Part I: Working correlation structure R**

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

Consistent estimates even if R mis-specified

Independent
Exchangeable
Autoregressive
Stationary
Non-stationary
Unstructured

- **Do we still need to specify R correctly?**

- **Under Ind. model, $\widehat{\beta}$ efficient if R not large**
  Zeger 1988; Biometrics

- **For time-varying covariates $\widehat{\beta}$ may be inefficient**
  Fitzmaurce 1995; Biometrics

- **Population-average model**

- **Part II: Mean & Variance of mean of outcome**
  First two orders of moments
  Different family variance differs

Table 1: Variance functions

Distribution	$V(\mu)$
Bernoulli	$\mu(1-\mu)$
Normal	1
Poisson	μ
Gamma	μ^2
Negative binomial	$\mu+\alpha\mu^2$
Inverse Gaussian	μ^3

- **Part III: Link function between outcome & variables**

  Different model link function differs

  Identity

  Log

  Logit

  Probit

  Power

  Reciprocal

- **Analysis of Longitudinal data**

  – *Subject ID*

  – *Time variable*

- **Model section in GEE**
- **No criterion available between 1986 & 2001**
- **Akaike Information Criterion (AIC) not applicable**

  No distribution was assumed for GEE
  No likelihood defined

- **Quasi-likelihood**

$$Q(\mu) = \int_y^\mu \frac{y-t}{\phi V(t)} dt$$

- **Under Independence correlation matrix**

$$Q(\beta, \phi; I, D) = \sum_{i=1}^{n} \sum_{j=1}^{n_i} Q(\beta, \phi; (Y_{ij}, X_{ij}))$$

- **GEE** $S(\beta; I, D)$ **equivalent** $\partial Q(\beta, \phi; I, D)/\partial\beta$

- **Under a general correlation matrix**
  - **Quasi-likelihood may not exist**

  McCullagh and Nelder 1989

- **Pan (2001) in Biometrics proposed QIC**
- **Hardin & Hilbe (2003)** *Generalized Estimating Equation*
- **Application was slow**

3. QIC method

Quasi-likelihood under Independence model Criterion

Table 2: Quasi-likelihood functions

Distribution	$\phi Q(\mu)$
Bernoulli	$y\ln(\frac{\mu}{1-\mu}) + \ln(1-\mu)$
Normal	$-\frac{1}{2}(y-\mu)^2$
Poisson	$y\ln(\mu) - \mu$
Gamma	$-(y/\mu + \ln(\mu))$
Negative binomial	$y\ln(\frac{\alpha\mu}{1+\alpha\mu}) - \frac{1}{\alpha}\ln(1+\alpha\mu)$
Inverse Gaussian	$-\frac{y}{2\mu^2} + \frac{1}{\mu}$

- **QIC formula**

$$\text{QIC} = -2Q(\widehat{\mu}; I) + 2\text{trace}(\widehat{\Omega}_I^{-1}\widehat{V}_R)$$

  - $\widehat{\Omega}_I$ **variance estimator under** *independence* **correlation structure**
  - $\widehat{\Omega}_R$ *robust* **variance estimator under** *specified working* **correlation structure**

- **Select** *correlation* **structure &** *most parsimonious* **model**

$$\text{QIC}_\text{u} = -2Q(\widehat{\mu}; I) + 2p$$

$$\text{AIC} = -2LL + 2p$$

4. Stata program

`qic` *depvar* [*indepvars*] [*if*] [*in*] [, **i**(*varname*$_i$) **t**(*varname*$_t$)
family(*familyname*) **link**(*linkname*) **corr**(*correlation*)
exposure(*varname*) **offset**(*varname*) **noconstant**
force eform nodisplay nolog iterate(#) **tolerance**(#)
trace]

5. An example

- **Longitudinal study of lung function decline**
  - **Follow-up from 1995 until 2003**
  - **322 males participants used for demonstration)**
- **Variables of interest**
  - FEV_1 **- Forced Expiratory Volume in one second**
  - **Age at each interview**
  - **Smoking status - current, former, nonsmoker**
  - **Age at baseline**
  - **Height at baseline**

- **Two individuals**

id	wave	fev	age	smoking	agebase	htbase
1150024	0	4.58	26.0	0	26.0	1.806
1150024	1	4.68	26.6	0	26.0	1.806
1150024	2	4.35	27.5	0	26.0	1.806
1150024	3	4.58	28.5	0	26.0	1.806
1150024	4	4.74	29.4	0	26.0	1.806
1150024	5	4.64	30.4	0	26.0	1.806
1150024	6	4.68	31.3	0	26.0	1.806
1150032	0	4.49	17.3	0	17.3	1.72
1150032	1	4.31	17.6	0	17.3	1.72
1150032	2	4.36	18.6	1	17.3	1.72

• Stata Command

```
. qic fev age smoking2 smoking3 agebase htbase, i(id) t(wave) corr(exchangeable)
```

```
GEE population-averaged model                   Number of obs      =      1141
Group variable:                        id       Number of groups   =       322
Link:                            identity       Obs per group: min =         1
Family:                          Gaussian                      avg =       3.5
Correlation:                 independent                       max =         9
                                                Wald chi2(5)       =    701.01
Scale parameter:                .2505905        Prob > chi2        =    0.0000

Pearson chi2(1141):              285.92         Deviance           =    285.92
Dispersion (Pearson):           .2505905        Dispersion         =  .2505905
```

fev	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	-.0344535	.0080028	-4.31	0.000	-.0501387	-.0187682
smoking2	-.1772179	.0342687	-5.17	0.000	-.2443834	-.1100524
smoking3	.0346093	.0540126	0.64	0.522	-.0712534	.140472
agebase	.0142697	.0082136	1.74	0.082	-.0018286	.030368
htbase	5.091886	.2154737	23.63	0.000	4.669566	5.514207
_cons	-3.888907	.3858109	-10.08	0.000	-4.645083	-3.132732

```
GEE population-averaged model                   Number of obs      =      1141
Group variable:                        id       Number of groups   =       322
Link:                            identity       Obs per group: min =         1
Family:                          Gaussian                      avg =       3.5
Correlation:                 exchangeable                      max =         9
                                                Wald chi2(5)       =    253.13
Scale parameter:                 .2547384       Prob > chi2        =    0.0000

                                          (Std. Err. adjusted for clustering on id)
------------------------------------------------------------------------------
             |               Semi-robust
         fev |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0354697   .0040787    -8.70   0.000    -.0434638   -.0274755
    smoking2 |  -.0483209     .02919    -1.66   0.098    -.1055323    .0088906
    smoking3 |  -.0098896    .022714    -0.44   0.663    -.0544083    .0346291
     agebase |   .0152248   .0054013     2.82   0.005     .0046385    .0258111
      htbase |   4.962204    .404878    12.26   0.000     4.168658     5.75575
       _cons |  -3.702856    .717312    -5.16   0.000    -5.108762    -2.29695
------------------------------------------------------------------------------
```

```
             QIC
------------------------------------------
Corr =         exchangeable
Family =                gau
Link =                 iden
p =                       6
Trace =              11.535
QIC =               313.727
------------------------------------------
```

```
. qic fev age smoking2 smoking3 agebase htbase, i(id) t(wave) nodisplay

             QIC
-------------------------------------------
Corr =                    exc
Family =                  gau
Link =                   iden
p =                         6
Trace =                11.535
QIC =                 313.727
-------------------------------------------
```

Table 2: QIC for selection of correlation structure

Correlation	Variable	p	QIC
Independent	`age`, `smoking`, `agebase`, `htbase`	6	324.19
Exchangeable	`age`, `smoking`, `agebase`, `htbase`	6	**313.73**
Autoregressive	`age`, `smoking`, `agebase`, `htbase`	6	321.16
Stationary	`age`, `smoking`, `agebase`, `htbase`	6	322.66
Non-stationary	`age`, `smoking`, `agebase`, `htbase`	6	322.49
Unstructured	`age`, `smoking`, `agebase`, `htbase`	6	322.83

```
. qic fev age , i(id) t(wave) corr(exchangeable) nolog nodisplay


             QIC
------------------------------------------
Corr =         exchangeable
Family =                gau
Link =                 iden
p =                       2
Trace =               5.260
QIC =               454.081
------------------------------------------
```

Table 3: QIC for selection of the best model

Correlation	Variable	p	QIC
Exchangeable	`age`	2	454.08
Exchangeable	`agebase`	2	455.19
Exchangeable	`htbase`	2	340.47
Exchangeable	`smoking`	3	469.37
Exchangeable	`age`, `agebase`	3	449.41
Exchangeable	`age`, `htbase`	3	314.94
Exchangeable	`agebase`, `htbase`	3	317.44
Exchangeable	`age`, `smoking`	4	452.72
Exchangeable	`smoking`, `agebase`	4	455.58
Exchangeable	`smoking`, `htbase`	4	342.20
Exchangeable	`age`, `agebase`, `htbase`	4	314.81
Exchangeable	`age`, `smoking`, `agebase`	5	447.22
Exchangeable	`age`, `smoking`, `htbase`	5	314.30
Exchangeable	`smoking`, `agebase`, `htbase`	5	318.51
Exchangeable	`age`, `smoking`, `agebase`, `htbase`	6	**313.73**

6. Summary

- **Developed a new program `qic` using Stata software**
- **Flexible & easy to use**
- **Include all** *distribution functions, link functions and correlation structures* **available in Stata**