The Stata Journal (2007)
7, Number 2, pp. 209–220

QIC program and model selection in GEE analyses

James Cui
Department of Epidemiology and Preventive Medicine
Monash University
Melbourne, Australia
james.cui@med.monash.edu.au

Abstract. The generalized estimating equation (GEE) approach is a widely used statistical method in the analysis of longitudinal data in clinical and epidemiological studies. It is an extension of the generalized linear model (GLM) method to correlated data such that valid standard errors of the parameter estimates can be drawn. Unlike the GLM method, which is based on the maximum likelihood theory for independent observations, the GEE method is based on the quasilikelihood theory and no assumption is made about the distribution of response observations. Therefore, Akaike's information criterion, a widely used method for model selection in GLM, is not applicable to GEE directly. However, Pan (Biometrics 2001; 57: 120–125) proposed a model-selection method for GEE and termed it quasilikelihood under the independence model criterion. This criterion can also be used to select the best-working correlation structure. From Pan's methods, I developed a general Stata program, `qic`, that accommodates all the distribution and link functions and correlation structures available in Stata version 9. In this paper, I introduce this program and demonstrate how to use it to select the best working correlation structure and the best subset of covariates through two examples in longitudinal studies.

Keywords: st0126, qic, Akaike's information criterion, GEE, likelihood, model, quasilikelihood under the independence model criterion

1 Introduction

The generalized estimating equation (GEE) approach (Liang and Zeger 1986) is a widely used statistical method in the analysis of longitudinal data in clinical and epidemiological studies (Diggle et al. 2002; Fitzmaurice et al. 2004). It specifies how the average of a response variable of a subject changes with covariates while allowing for the correlation between repeated measurements on the same subject over time. The focus of this method is the estimation of regression parameters that have a population-average interpretation and the correlation structure is treated as a nuisance parameter (Hardin and Hilbe 2003; Singer and Willett 2003; Weiss 2005). An attractive feature of this method is that, when the mean response is correctly specified, consistent parameter estimates will be derived even if the correlation structure is misspecified. The mean and variance of the response variable are usually specified by one of the distribution functions in the exponential family (McCullagh and Nelder 1989).

Essentially, GEE is an extension of the generalized linear model (GLM) (Nelder and Wedderburn 1972) to correlated data such that valid standard errors of the parameter estimates can be drawn. Correlated data are often encountered in clinical and epidemiological studies. For example, common cancers are known to be clustered within a family because of possible underlying genetic risks (Cui et al. 2001a,b). Also, repeated measures of the same subject in follow-up studies are likely to be correlated because of the continuity of the measurement over time (Rabe-Hesketh and Skrondal 2005). To take account of the correlation, a specification of a working correlation structure is required in GEE, which can be independence, exchangeable, autoregressive, stationary, nonstationary, or unstructured specification in Stata version 9 (StataCorp 2005).

Unlike GLM, which is based on the maximum likelihood theory for independent observations (McCullagh and Nelder 1989), the GEE method is based on the quasilikelihood theory (Wedderburn 1974), and no assumption is made about the distribution of response observations. Therefore, some of the statistics derived under the likelihood theory cannot be applied to GEE directly. For instance, Akaike's information criterion (AIC; Akaike 1974), a widely used method for model selection in GLM, is not applicable to GEE. However, under appropriate modification of the AIC method, Pan (2001) proposed a model-selection method for GEE and termed it quasilikelihood under the independence model criterion (QIC). This criterion can also be used to select the best working correlation structure in GEE analyses.

Although the QIC method was published in 2001 and included in Hardin and Hilbe (2003), the application of this method in practice is relatively slow. One possible reason is that as of this writing this method has not been included in any popular statistical software. From Pan's theory, I developed a general Stata program, `qic`, that accommodates all the distribution and link functions and correlation structures available in Stata version 9. The aim of this paper is to introduce this program and demonstrate how to use it to select the best working correlation structure and the best subset of covariates through two examples in longitudinal studies.

2 Methods

Denote y as the response variable and x a vector of covariates. Under GLM $g(\mu) = \beta' x$, where $g()$ is the link function and $\mu = E(y)$, the AIC is given by

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

where LL is the log likelihood and p is the number of parameters in the model. Pan (2001) modified the above formula and made an adjustment for the penalty term $2p$ for GEE, deriving the QIC as

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

where I represents the independent covariance structure used to calculate the quasi-likelihood. Here $\widehat{\mu} = g^{-1}(x\widehat{\beta})$ and $g^{-1}()$ is the inverse link function. The coefficient estimates $\widehat{\beta}$ and robust variance estimator $\widehat{V}_R$ are obtained from a general working covariance structure R. Another variance estimator $\widehat{\Omega}_I$ is obtained under the assumption of an independence correlation structure.

The QIC value in (1) can be used to select the best correlation structure and the best fitting model in GEE analyses (Pan 2001). A correlation matrix with the smallest QIC value is chosen as the preferred correlation structure. A subset of covariates with the smallest QIC value is the preferred model. When $\text{trace}(\widehat{\Omega}_I^{-1}\widehat{V}_R) \approx \text{trace}(I) = p$, there is a simplified version of QIC, called QIC_u (Pan 2001),

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

However, this simplified QIC_u cannot be applied to select the optimal working covariance structure because of the assumption of asymptotic equivalence of $\widehat{\Omega}_I$ and $\widehat{V}_R$.

The quasilikelihood in model (1) and (2) is of the general form (Wedderburn 1974)

$$Q(\mu) = \int_y^\mu \frac{y - t}{\phi V(t)} dt \tag{3}$$

where ϕ is a dispersion parameter. The variance of the response observations is a function of the mean μ and denoted as $V(\mu)$. The value of $V(\mu)$ is given in table 1 for some of the commonly used distributions in the exponential family. Substituting $V(\mu)$ in (3) with the corresponding value in table 1, we can compute the value of the quasilikelihood $Q(\mu)$, which is also listed in table 1.

Table 1: Variance and quasilikelihood functions for commonly used distributions in the exponential family

Distribution	$V(\mu)$	$Q(\mu)$
Bernoulli	$\mu(1-\mu)$	$y\ln(\frac{\mu}{1-\mu}) + \ln(1-\mu)$
Normal	1	$-\frac{1}{2}\sum(y-\mu)^2$
Poisson	μ	$y\ln(\mu) - \mu$
Gamma	μ^2	$-(y/\mu + \ln(\mu))$
Negative binomial	$\mu + \mu^2$	$y(\ln(\mu) - 2\ln(\mu+1))$
Inverse Gaussian	μ^3	$-\frac{y}{2\mu^2} + \frac{1}{\mu}$

From the above statistical theory, I developed a general Stata program, `qic`, to calculate the QIC and QIC_u values in GEE analyses. This program was implemented in Stata version 9.

3 The qic program

3.1 Syntax

`qic` *depvar* [*indepvars*] [*if*] [*in*] [*weight*] [`,` `i(`*varname*$_i$`)` `t(`*varname*$_t$`)`
`family(`*familyname*`)` `link(`*linkname*`)` `exposure(`*varname*`)` `offset(`*varname*`)`
`noconstant` `force` `corr(`*correlation*`)` `robust` `nmp` `rgf` `scale(x2|dev|phi|`*#*`)`
`level(`*#*`)` `eform` `iterate(`*#*`)` `tolerance(`*#*`)` `nolog` `trace` `nodisplay`]

3.2 Description

The `qic` program calculates and displays the QIC and QIC$_u$ values for a GEE model. Here *depvar* is the name of the response variable and it must be specified after `qic`. Other items are optional, including a list of independent variable names (denoted by *indepvars*) and options related to GEE methods (such as `i(`*varname*$_i$`)` and `t(`*varname*$_t$`)` [see [XT] **xtgee**]). Specification of *if* and *in* statements is allowed.

Similar to the `xtgee` command, the panel ID variable can be specified using the `iis` *varname*$_i$ command before running `qic` or using the option `i(`*varname*$_i$`)` within `qic`. Also, the time variable can be specified using command `tis` *varname*$_t$ or option `t(`*varname*$_t$`)`. The `qic` command includes nearly all the options available in `xtgee` except for the `robust` options because the variance Ω_I has to be estimated without the `robust` option and the variance V_R estimated with the `robust` option. An underlined part of an option may be used instead of the whole word.

In addition to calculating the QIC and QIC$_u$ values, the `qic` program also calculates and displays the value of the trace of $\widehat{\Omega}_I^{-1}\widehat{V}_R$ to compare how close the value of QIC$_u$ approximates the value of QIC. When comparing two or more different correlation structures for a specified distribution and link function, a correlation structure with the smallest QIC is the preferred correlation structure. Usually the full model with all available covariates is used in selecting the best correlation structure (Hardin and Hilbe 2003). Under the preferred correlation structure, a subset of covariates with the smallest QIC will be the preferred model. Details of GEE-related options follow.

3.3 Options

`i(`*varname*$_i$`)` specifies *varname*$_i$ as the panel ID variable.

`t(`*varname*$_t$`)` specifies *varname*$_t$ as the time variable.

`family(`*familyname*`)` specifies the distribution of *depvar*. `family(gaussian)` is the default, which corresponds to the Gaussian distribution.

`link(`*linkname*`)` specifies the link function. The default is the canonical link function for the specified `family()`.

`exposure(`*varname*`)` requests that ln(*varname*) be included in the model with its coefficient being constrained to 1.

`offset(`*varname*`)` requests that *varname* be included in the model with its coefficient being constrained to 1. The `exposure()` and `offset()` options can be used in Poisson regression models to reflect the different amount of exposure for which the *depvar* events were observed.

`noconstant` specifies that the linear predicator has no intercept term and thus forces it to pass through the origin on the scale defined by the link function.

`force` requests that estimation be forced even though `t()` is not equally spaced. It is relevant only for correlation structures that require knowledge of `t()` and observations to be equally spaced.

`corr(`*correlation*`)` specifies the within-subject correlation structure. The default is the exchangeable correlation structure `corr(exchangeable)`.

`robust` specifies that the Huber/White/sandwich estimator of variance be used in place of the default GLS variance estimator. This produces valid standard errors even if the correlations within group are not as hypothesized by the specified correlation structure. It does, however, require that the model correctly specify the mean. The resulting standard errors are thus labeled "Semi-Robust" instead of "Robust". Although there is no `cluster()` option, results are as if there were a `cluster()` option and you specified clustering on `i()`.

`nmp` specifies that the divisor $N - P$ be used instead of the default N, where N is the total number of observations and P is the number of coefficients estimated.

`rgf` specifies that the robust variance estimate be multiplied by $(N-1)/(N-P)$, where N = # of observations and P = # of coefficients estimated. This option can be used only with `family(gaussian)` when `robust` is either specified or implied by the use of `pweights`. Using this option implies that the robust variance estimate is not invariant to the scale of any weights used.

`scale(x2|dev|#|phi)` overrides the default scale parameter of `scale(1)`; see [R] **estimation options**.

`level(#)` specifies the confidence level, as a percentage, for confidence intervals. The default is `level(95)` or as set by `set level`; see [U] **20.6 Specifying the width of confidence intervals**.

`eform` displays the exponentiated coefficients and associated standard errors and confidence intervals.

`iterate(#)` specifies the maximum number of iterations allowed in the optimization. It must be a positive integer. `iterate(100)` is the default.

`tolerance(#)` specifies the convergence criterion for the coefficient vector. When the relative change in the coefficient vector between two consecutive iterations is less than or equal to #, the optimization process is stopped. `tolerance(1e-6)` is the default.

`nolog` suppresses display of the iteration log.

`trace` specifies that the current estimates are printed at each iteration.

`nodisplay` suppresses display of the GEE tables during the calculation of $\widehat{\Omega}_I$ and $\widehat{V}_R$.

3.4 Saved results

`qic` saves the following in `r()`:

Scalars

`r(p)`	number of parameters	`r(trace)`	value of trace
`r(qicu)`	value of QIC_u	`r(qic)`	value of QIC

Macros

`r(family)`	probability distribution	`r(corr)`	correlation structure
`r(link)`	link function		

A list of standard saved `e()` results associated with execution of the `xtgee` command with the specified correlation structure and the `robust` option can also be obtained by using `ereturn list` (see [XT] **xtgee**).

4 Example 1

I demonstrate how to use the `qic` program to select the best fitting model through two examples. The first example has a normal distribution and the second has a binomial distribution. For simplicity, we do not give examples for other distributions here. Similar procedures of applying `qic` can easily be extended to other distributions.

The first example comes from the National Longitudinal Survey of Labor Market Experience (Center for Human Resource Research 1989). A subsample of 3,913 women aged between 14 and 26 years who have completed their education with wages in excess of $1/hour but less than $700/hour is used in this analysis (see [XT] **xtgee**). The response variable in this example is a continuous variable `ln_wage`, representing the logarithm of the wage of each woman. Each individual has an identification number denoted as `idcode`. The year of survey is denoted by variable `year`. Explanatory variables that are used in this paper include `grade`, representing the current grade that has been completed (ranging from 0–18); `age`, representing the age at the survey; and `south`, representing whether a person comes from the south (1 if a person comes from the south and 0 otherwise).

The first step is to choose the best working correlation structure. Because the response outcome `ln_wage` is a continuous variable, we use the normal distribution and the identity link function, which are also the default choices of the `qic` program. Therefore, we do not need to specify them explicitly in the `qic` command.

We first calculate the QIC and QIC_u values for the exchangeable correlation structure, based on the full model with all covariates of interest `age`, `grade`, and `south`. The output is displayed below, in which the first part corresponds to a GEE model with independence correlation matrix to calculate the $\widehat{\Omega}_I^{-1}$ and the second part to a GEE model with the specified correlation matrix to calculate the robust variance $\widehat{V}_R$. The last part gives the QIC and QIC$_u$ values and displays the distribution and link function and the correlation structure used in the program, plus the number of parameters in the regression model. We can suppress display of the GEE tables by using the `nodisplay` option (see *Example 2*).

```
. use nlswork2, clear
(National Longitudinal Survey.  Young Women 14-26 years of age in 1968)
. qic ln_wage age grade south, i(idcode) t(year) corr(exchangeable)
Iteration 1: tolerance = 5.706e-14
GEE population-averaged model                   Number of obs      =     16077
Group variable:                     idcode      Number of groups   =      3911
Link:                             identity      Obs per group: min =         1
Family:                           Gaussian                     avg =       4.1
Correlation:                   independent                     max =         9
                                                Wald chi2(3)       =   4957.31
Scale parameter:                  .1360754      Prob > chi2        =    0.0000
Pearson chi2(16077):               2187.68      Deviance           =   2187.68
Dispersion (Pearson):             .1360754      Dispersion         =  .1360754

     ln_wage       Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]

         age     .0229159   .0007681    29.84   0.000     .0214105    .0244213
       grade      .068458   .0014044    48.75   0.000     .0657055    .0712105
       south    -.1563942   .0060314   -25.93   0.000    -.1682155    -.144573
       _cons     .2491559   .0234644    10.62   0.000     .2031664    .2951453

Iteration 1: tolerance = .09574304
Iteration 2: tolerance = .00291689
Iteration 3: tolerance = .00006597
Iteration 4: tolerance = 1.487e-06
Iteration 5: tolerance = 3.350e-08
GEE population-averaged model                   Number of obs      =     16077
Group variable:                     idcode      Number of groups   =      3911
Link:                             identity      Obs per group: min =         1
Family:                           Gaussian                     avg =       4.1
Correlation:                  exchangeable                     max =         9
                                                Wald chi2(3)       =   2098.45
Scale parameter:                  .1369477      Prob > chi2        =    0.0000
                                  (Std. Err. adjusted for clustering on idcode)

                          Semi-robust
     ln_wage       Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]

         age     .0257849   .0010161    25.38   0.000     .0237934    .0277763
       grade     .0700327   .0022904    30.58   0.000     .0655435    .0745218
       south    -.1316293    .009427   -13.96   0.000    -.1501058   -.1131527
       _cons     .1261871   .0347916     3.63   0.000     .0579968    .1943773
```

```
                QIC and QIC_u
  ___________________________________________
  Corr =        exchangeable
  Family =               gau
  Link =                iden
  p =                      4
  Trace =              9.687
  QIC =             2221.083
  QIC_u =           2209.709
  ___________________________________________
```

Similarly, the QIC values for other correlation structures can be calculated; they are summarized in table 2. The `force` option is used in the autoregressive, stationary, and nonstationary correlation matrix because observations are not equally spaced. The second-order stationary and third-order nonstationary correlation matrices are used because the first-order correlation model does not converge. The independence correlation structure has the smallest QIC and thus is chosen as the preferred correlation matrix, and the corresponding QIC_u value is also listed in table 2.

Under the independence correlation structure, we fit different models with different subsets of covariates and find that the full model has the smallest QIC (marked in boldface font) and thus is chosen as the best-fitting model to the data (see table 2). On the basis of the approximate criterion QIC_u, the full model still has the smallest QIC_u value and thus is chosen as the preferred model.

Table 2: QIC for model selection under normal distribution for the National Longitudinal Survey data

Correlation	Variable	p	Trace	QIC	QIC_u
Independent	`age, grade, south`	4	11.05	**2,209.78**	**2,195.68**
Exchangeable	`age, grade, south`	4	9.69	2,221.08	
Autoregressive	`age, grade, south`	4	10.74	2,217.03	
Stationary3	`age, grade, south`	4	18.45	2,232.26	
Nonstationary4	`age, grade, south`	4	26.55	2,252.88	
Unstructured	`age, grade, south`	4	9.46	2,227.71	
Independent	`age`	2	5.64	2,686.42	2,679.14
Independent	`grade`	2	6.22	2,399.30	2,390.86
Independent	`south`	2	6.09	2,732.60	2,724.42
Independent	`age, grade`	3	8.36	2,297.22	2,286.50
Independent	`age, south`	3	8.47	2,527.97	2,517.03
Independent	`grade, south`	3	8.89	2,327.86	2,316.08

NOTE: Values in boldface indicate smallest QIC value.

5 Example 2

This example also comes from the National Longitudinal Survey of Labor Market Experience (Center for Human Resource Research 1989). A part of 4,434 women with union membership information from 1970 to 1988 is used in this analysis (see [XT] **xtlogit**). The response variable in this example is a binary indicator, `union`, which equals 1 if a person is a union member and 0 otherwise. For illustrative purpose, the same explanatory variables `age`, `grade`, and `south` as outlined in example 1 are used here again.

Because the response outcome `union` is a binary variable, we use the binomial (Bernoulli) distribution and the probit link function in this analysis. To shorten the `qic` command, `iis idcode` and `tis year` are used to declare the panel ID variable and time variable before using the `qic` command. The `nolog` and `nodisplay` options are used to suppress display of the iteration log and the GEE-related tables. The output of the calculation is shown below.

```
. use union, clear
(NLS Women 14-24 in 1968)
. iis idcode
. tis year
. qic union age grade south, family(bin) link(probit) corr(exc) nolog nodisplay

              QIC and QIC_u
___________________________________________
Corr =                    exc
Family =                  bin
Link =                 probit
p =                         4
Trace =                12.626
QIC =               27193.900
QIC_u =             27176.648
___________________________________________
```

Similar calculations are conducted using other correlation structures, and details of the analysis results are shown in table 3. The `force` option is used in the autoregressive, stationary, and nonstationary correlation matrices because observations are not equally spaced. The fifth-order nonstationary correlation matrix is used because the first four orders' correlation structures do not converge. The independence correlation matrix has the smallest QIC and thus is chosen as the preferred correlation structure, and the corresponding QIC_u value is also listed in table 3.

(*Continued on next page*)

Table 3: QIC for model selection under binomial (Bernoulli) distribution probit link function for the National Longitudinal Survey data

Correlation	Variable	p	Trace	QIC	QIC$_u$
Independent	`age`, `grade`, `south`	4	14.49	**27,166.63**	**27,145.65**
Exchangeable	`age`, `grade`, `south`	4	12.63	27,193.90	
Autoregressive	`age`, `grade`, `south`	4	12.91	27,171.92	
Stationary4	`age`, `grade`, `south`	4	24.77	27,201.77	
Nonstationary5	`age`, `grade`, `south`	4	29.20	27,279.97	
Unstructured	`age`, `grade`, `south`	4	11.65	27,223.97	
Independent	`age`	2	6.15	27,714.61	27,706.31
Independent	`grade`	2	8.96	27,625.40	27,611.47
Independent	`south`	2	7.81	27,249.15	27,237.51
Independent	`age, grade`	3	10.93	27,614.53	27,598.67
Independent	`age, south`	3	9.77	27,219.26	27,205.72
Independent	`grade, south`	3	12.57	27,185.96	27,166.82

NOTE: Values in boldface indicate smallest QIC value.

Under the independence correlation structure, we find that the full model has the smallest QIC and thus is chosen as the preferred model (see table 3). The QIC$_u$ criterion also indicates that the full model is the most parsimonious.

6 Conclusion

In this paper, I introduced a new program, `qic`, for calculating the QIC and QIC$_u$ values for selecting the best correlation structure and the most parsimonious models in GEE analyses. All the distribution and link functions and all the correlation structures available in Stata version 9 can be specified in this program.

Although both QIC and QIC$_u$ select the same model in the two examples presented here, sometimes they select different models because QIC$_u$ is just an approximation to QIC. Furthermore, QIC$_u$ cannot be used to select the best correlation structure. Therefore, I recommend using QIC in practice especially when they select different models.

Other statistics, such as the Wald χ^2 and deviance, are also produced by Stata in fitting a GEE model. However, these statistics cannot be used for comparing nonnested GEE models because they do not take into account the number of parameters in the model. Therefore, they may give misleading conclusions for model selection in GEE analyses. With the availability of this new `qic` program, I hope that more applications of the QIC method can be seen in practice in the future.

7 Acknowledgment

I thank Dr. Guoqi Qian, University of Melbourne, for help during development of this program.

8 References

Akaike, H. 1974. A new look at statistical model identification. *IEEE Transactions on Automatic Control* AC-19: 716–723.

Center for Human Resource Research. 1989. National Longitudinal Survey of Labor Market Experience, Young Women 14–26 years of age in 1968. Ohio State University.

Cui, J., A. C. Antoniou, G. S. Dite, M. C. Southey, D. J. Venter, D. F. Easton, G. G. Giles, M. R. E. McCredie, and J. L. Hopper. 2001a. After BRCA1 and BRCA2—what next? Multifactorial segregation analyses of three-generational, population-based Australian female breast cancer families. *American Journal of Human Genetics* 68: 420–431.

Cui, J., M. P. Staples, J. L. Hopper, D. R. English, M. R. E. McCredie, and G. G. Giles. 2001b. Segregation analyses of 1476 population-based Australian families affected by prostate cancer. *American Journal of Human Genetics* 68: 1207–1218.

Diggle, P., P. Heagerty, K.-Y. Liang, and S. Zeger. 2002. *Analysis of Longitudinal Data*. 2nd ed. Oxford: Oxford University Press.

Fitzmaurice, G. M., N. M. Laird, and J. W. Ware. 2004. *Applied Longitudinal Data*. 2nd ed. Hoboken, NJ: Wiley.

Hardin, J. W., and J. M. Hilbe. 2003. *Generalized Estimating Equations*. Boca Raton, FL: Chapman & Hall/CRC.

Liang, K.-Y., and S. L. Zeger. 1986. Longitudinal data analysis using generalized linear models. *Biometrika* 73: 13–22.

McCullagh, P., and J. A. Nelder. 1989. *Generalized Linear Models*. 2nd ed. London: Chapman & Hall.

Nelder, J. A., and R. W. M. Wedderburn. 1972. Generalized linear models. *Journal of the Royal Statistical Society, Series A* 135: 370–384.

Pan, W. 2001. Akaike's information criterion in generalized estimating equations. *Biometrics* 57: 120–125.

Rabe-Hesketh, S., and A. Skrondal. 2005. *Multilevel and Longitudinal Modeling Using Stata*. College Station, TX: Stata Press.

Singer, J., and J. Willett. 2003. *Applied Longitudinal Data Analysis*. Oxford: Oxford University Press.

StataCorp. 2005. *Stata Statistical Software: Release 9*. College Station, TX: StataCorp.

Wedderburn, R. W. M. 1974. Quasilikelihood functions, generalized linear models, and the Gauss–Newton method. *Biometrika* 61: 439–447.

Weiss, R. E. 2005. *Modeling Longitudinal Data*. New York: Springer.

About the author

James Cui is a statistician at the Department of Epidemiology and Preventive Medicine, Monash University, Melbourne, Australia. He has had 23 years of working experience in teaching biostatistics and conducting research in epidemiology.