http://faculty.washington.edu/tlumley/gee/geejss/node1.html

**Next:** [An introduction to](http://faculty.washington.edu/tlumley/gee/geejss/node2.html) **Up:** [XLISP-Stat tools for building](http://faculty.washington.edu/tlumley/gee/geejss/geejss.html) **Previous:** [XLISP-Stat tools for building](http://faculty.washington.edu/tlumley/gee/geejss/geejss.html)

**Introduction**

Generalised Estimating Equations models, proposed by Liang and Zeger in 1986, are probably the simplest method for analysing data collected in groups where observations within a group may be correlated but observations in separate groups are independent. A complete description of the method is given in their two 1986 papers. The basic principle of the method is a generalisation of the fact that weighted least squares analyses give unbiased parameter estimates no matter what weights are used. Generalised linear models, such as logistic regression, have similar robustness properties, giving asymptotically correct parameter estimates even when the data are correlated. This means that it is possible to estimate regression parameters using any convenient or plausible assumptions about the true correlation between observations and get the right answer even when the assumptions are not correct. It is only necessary to use a ``model-robust'' or ``agnostic'' estimate of the standard errors. It would be unreasonable to expect this freedom of choice to be without cost and it turns out that there is a moderate gain in efficiency resulting from choosing a working correlation structure close to the true one.

Useful references include the two original papers (Zeger & Liang 1986, Liang & Zeger 1986) and two recent books: Diggle, Liang & Zeger (1993) and Fahrmeir & Tutz (1995). As far as I know the most elementary treatment anywhere in the literature is still Zeger & Liang (1986).

Section [2](http://faculty.washington.edu/tlumley/gee/geejss/node2.html#geeinfo) gives an overview of the theory and use of Generalised Estimating Equations. Section [3](http://faculty.washington.edu/tlumley/gee/geejss/node5.html#use) describes how to use the Lisp-Stat code, including diagnostics. Finally there is a brief discussion of missing data handling and of other software for fitting GEE models. Appendix [A](http://faculty.washington.edu/tlumley/gee/geejss/node13.html#details) describes some aspects of the implementation, including the global variables (Table [3](http://faculty.washington.edu/tlumley/gee/geejss/node19.html#globals)) that control many program options.

*Thomas Lumley   
Sun Dec 8 16:10:41 PST 1996*

**Next:** [GEE models in](http://faculty.washington.edu/tlumley/gee/geejss/node3.html) **Up:** [XLISP-Stat tools for building](http://faculty.washington.edu/tlumley/gee/geejss/geejss.html) **Previous:** [Introduction](http://faculty.washington.edu/tlumley/gee/geejss/node1.html)

# An introduction to GEE models

* [GEE models in theory](http://faculty.washington.edu/tlumley/gee/geejss/node3.html" \l "SECTION00021000000000000000)

* [GEE models in practice](http://faculty.washington.edu/tlumley/gee/geejss/node4.html" \l "SECTION00022000000000000000)

Thomas Lumley   
Sun Dec 8 16:10:41 PST 1996

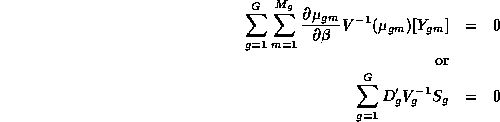
**Next:** [GEE models in](http://faculty.washington.edu/tlumley/gee/geejss/node4.html) **Up:** [An introduction to](http://faculty.washington.edu/tlumley/gee/geejss/node2.html) **Previous:** [An introduction to](http://faculty.washington.edu/tlumley/gee/geejss/node2.html)

## GEE models in theory

Suppose observations and **p** covariates http://faculty.washington.edu/tlumley/gee/geejss/img1.gifare available in http://faculty.washington.edu/tlumley/gee/geejss/img2.gifgroups with at most http://faculty.washington.edu/tlumley/gee/geejss/img3.gifin each group, and that the mean and variance of each observation is thought to satisfy a generalised linear model:

http://faculty.washington.edu/tlumley/gee/geejss/img4.gif

but that the observations within a group are correlated. If the variables were uncorrelated the (quasi)score equations would be:



where http://faculty.washington.edu/tlumley/gee/geejss/img6.gifis http://faculty.washington.edu/tlumley/gee/geejss/img7.gifmatrix of derivatives, http://faculty.washington.edu/tlumley/gee/geejss/img8.gifis a diagonal http://faculty.washington.edu/tlumley/gee/geejss/img9.gifmatrix of variances and http://faculty.washington.edu/tlumley/gee/geejss/img10.gifis the vector of the http://faculty.washington.edu/tlumley/gee/geejss/img11.gifresiduals http://faculty.washington.edu/tlumley/gee/geejss/img12.gif.

The left-hand side of this equation has zero expectation at the true value of http://faculty.washington.edu/tlumley/gee/geejss/img13.gifeven if there is correlation within groups, so the estimates of http://faculty.washington.edu/tlumley/gee/geejss/img14.gifwill be consistent even for correlated data. In this equation the changing variance of http://faculty.washington.edu/tlumley/gee/geejss/img15.gifis accounted for in the weighting by http://faculty.washington.edu/tlumley/gee/geejss/img16.gifbut no account is made for the correlations. The consistency results still hold if is replaced by a covariance matrix for group **g**, either prespecified or depending on a finite set of parameters. If this covariance matrix is close to the true covariance of then http://faculty.washington.edu/tlumley/gee/geejss/img19.gifwill be estimated more efficiently. The assumed form of http://faculty.washington.edu/tlumley/gee/geejss/img20.gifis called the ``working covariance model''.

In order for this to be useful it is necessary to have reliable standard error estimates for the estimates http://faculty.washington.edu/tlumley/gee/geejss/img21.gif. An argument based on the delta method (or a first-order Taylor series expansion) shows that

http://faculty.washington.edu/tlumley/gee/geejss/img22.gif

the so-called ``sandwich estimator'' gives asymptotically correct standard errors and various simulation studies have shown that its properties are good when **M** is small (5 or fewer). Less is known for large **M**.

Thomas Lumley   
Sun Dec 8 16:10:41 PST 1996

**Next:** [How to use](http://faculty.washington.edu/tlumley/gee/geejss/node5.html) **Up:** [An introduction to](http://faculty.washington.edu/tlumley/gee/geejss/node2.html) **Previous:** [GEE models in](http://faculty.washington.edu/tlumley/gee/geejss/node3.html)

## GEE models in practice

The models are used in the same way as standard generalised linear models, and the coefficients have the same interpretation. They measure differences in the response for a unit change in the predictor, averaged over the whole sample. GEE models are thus particularly suitable when the correlation is of no substantive interest and is merely a nuisance parameter.

The data used for the example in section [3](http://faculty.washington.edu/tlumley/gee/geejss/node5.html#use) illustrate a common, but very simple situation. The number of hospital visits for 72 children were recorded, together with the sex of the child and the age and smoking status of the mother. We are interested in whether maternal smoking is related to the number of hospital visits and the obvious analysis is a Poisson regression. The complication is that the number of hospital visits was recorded over four separate time periods. Using these as separate observations violates the independence assumption for Poisson regression; adding them together to get a single outcome is likely to result in overdispersion and loses any information about variation over time.

With a working covariance model of independence the parameter estimates will be the same as from a Poisson regression, but the standard errors will be valid. This is often sufficient, especially when the correlations are not too high and the number of observations on each individual is the same.

Another popular model is the exchangeable correlation model, in which all pairwise correlations between different times are the same. This is analogous to the popular but unreliable method of analysing repeated Normal measurements using split-plot ANOVA. An exchangeable correlation GEE fits the same working modelbut gets asymptotically valid standard error estimates even when the correlations are not truly exchangeable.

It is probably the case that observations close together in time are more similar than those far apart in time. A working covariance model that incorporated this structure might give better estimates. One possible correlation structure for these data would be stationary 3-dependence. This working model estimates three correlation parameters: the correlations at lag 1, lag 2 and lag 3. If there were sufficient data it would even be possible to estimate all 6 correlation parameters. This is known as the ``saturated'' working model. In addition to possibly greater efficiency these alternative working models provide estimates of the correlations between different time points. It should be noted, however, that these estimates have low efficiency and may be biased if the working model is not close to the truth.

Little is known about good ways of choosing correlation structures. The options provided in this implementation allow reasonable flexibility to approximate a wide range of structures. In the next section these data are analysed in Lisp-Stat using multiple correlation structures and diagnostic methods.

**Next:** [How to use](http://faculty.washington.edu/tlumley/gee/geejss/node5.html) **Up:** [An introduction to](http://faculty.washington.edu/tlumley/gee/geejss/node2.html) **Previous:** [GEE models in](http://faculty.washington.edu/tlumley/gee/geejss/node3.html)

Thomas Lumley   
Sun Dec 8 16:10:41 PST 1996