STATISTICS IN MEDICINE
Statist. Med. 2000; **19**:715–722

GEE Analysis of negatively correlated binary responses: a caution

James A. Hanley[1,2,*,†], Abdissa Negassa[3] and Michael D. deB. Edwardes[2]

[1]*Department of Epidemiology and Biostatistics, McGill University, Montreal, Canada*
[2]*Division of Clinical Epidemiology, Royal Victoria Hospital, Montreal, Canada*
[3]*Division of Epidemiology, Department of Oncology, McGill University, Montreal, Canada*

SUMMARY

The method of generalized estimating equations has become almost standard for analysing longitudinal and other correlated response data. However, we have found that if binary responses have less than binomial variation over clusters, and are modelled using exchangeable correlations, prevailing software implementations may give unreliable results. Bounding the negative correlation away from its theoretical minimum may not always be a satisfactory solution. In such instances, using the independence working correlation structure and robust SEs is a more trustworthy alternative. Copyright © 2000 John Wiley & Sons, Ltd.

1. INTRODUCTION

The generalized estimating equations (GEE) approach[1,2] has become the method of choice for analysing longitudinal and other correlated response data. It is now available in most statistical packages,[3–5] but some users use their own implementations, or rely on older macros.[6,7]

In this note we relate our experience when we used a cluster sample involving binary responses[8] to explain the essence of the GEE approach to non-statisticians. We chose the example to allow them to compare the GEE estimate of a proportion, and its standard error (SE), with those calculated by classical methods. Table I shows the raw data. Clusters (households) range in size from 1 to 7. Individuals in each household were classified as to (i) whether they had consulted a physician in the past 12 months and (ii) gender.

We obtained a GEE estimate, and associated SE, of the proportion who had visited a physician in the past year. The results from several software implementations were virtually identical to each other, and to those from the classical analysis.[8] The SE, larger than the one that might be (naively) calculated from the binomial model, reflects the considerable similarity (positive correlation) of responses within the same household.

* Correspondence to: James A. Hanley, Department of Epidemiology and Biostatistics, McGill University, 1020 Pine Avenue West, Montreal, Que., H3A 1A2, Canada.
† E-mail: Jimh@epid.lan.mcgill.ca

Contract/grant sponsor: Natural Sciences and Engineering Research Council of Canada
Contract/grant sponsor: NIH-CA; Contract/grant number: 70269
Contract/grant sponsor: Fonds de la recherche en santé du Québec

CCC 0277–6715/2000/050715–08$17.50
Copyright © 2000 John Wiley & Sons, Ltd.
Received February 1998
Accepted April 1999

Table I. Data on (v) physician visits and (g) gender for a cluster sample of 30 households

	Households																												Total
h	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	
n_h	5	6	3	3	2	3	3	3	4	4	3	2	7	4	3	5	4	4	3	3	4	3	3	1	2	4	3	4	2	4	104
(v) y_h^+	5	0	2	3	0	0	0	0	0	0	0	0	0	4	1	2	0	0	1	3	2	0	0	0	2	2	0	2	0	1	30
(g) y_h^+	1	3	1	1	1	1	1	1	2	3	2	1	3	3	2	3	3	3	2	1	1	2	2	0	1	3	1	2	1	2	53

h household
n_h number of persons in household h
(v): y_h^+ number who had visited a physician in previous year
(g): y_h^+ number who were of male gender
Source: Cochran[8]

Table II. Point estimates of percentage of males, with associated standard error, using various models/software. Data are from Table I. An asterisk (*) denotes non-convergence

Model/software	Point estimate (%)	Standard error (%)		$\hat{\rho}$
		Model-based	Robust	
SRS	50·96 (53/104)	4·90		
Cluster	50·96	3.40		
GEE:				
as in text, r unrestricted	53·66	3·47	5·60	− 0·1852
SAS Genmod[3]	48·20	2·88	3·07	− 0·1567
S-plus[4]	*	*	*	*
STATA[5]	47·57	2·84	4·08	− 0·1811
SAS Macro v1.25[6]	*	*	*	*
SAS Macro v2.03[7]	*	*	*	*
independence	50·96	4·90	3·33&	0·0

SRS: simple random sample

Responses as to gender showed *less* than binomial variation across households,[8] but we again expected each software package to produce virtually the same GEE estimates for the proportion of males. However, as is shown in Table II, several implementations were not able to estimate the proportion of males, and associated SE; if estimates were produced, they were somewhat different from each other. This note explores why, and suggests a strategy for dealing with such negatively correlated data.

2. GEE ANALYSIS

We denote the binary response of the jth person in the hth household as y_{hj}. The regression model for the binary responses contains just one parameter, $E[y_{hj}] = \pi$, the intercept in an intercept-only regression model with a single predictor variable $x_0 \equiv 1$ for each y. If the available GEE

Copyright © 2000 John Wiley & Sons, Ltd.

software allows it, one can estimate π directly by using the identity function to link $E[y_{hj}]$ to the linear predictor πx_0, and by specifying binomial (Bernoulli) variation; otherwise, one can estimate it indirectly from the estimate of logit$[\pi]$. When we 'unpacked' the data in Table I for the GEE analysis, there was no information on the occurrence of the outcome among individual family members, that is, parents, siblings etc., and so we treated the responses as exchangeable. Thus, in the $n_h \times n_h$ correlation matrix R_h for y's from a household containing n_h persons, we assumed that all expected pairwise correlations are of the same magnitude ρ.

The GEE estimate of π then satisfies the single estimating equation

$$\frac{1}{\pi(1-\pi)} \sum \mathbf{1}' R_h^{-1} (\mathbf{y}_h - \pi \mathbf{1}) = 0$$

where $\mathbf{y}_h$ is the vector of binary responses in household h, $\mathbf{1}$ is a vector of n_h ones, and the summation is over households.[1] Since the on- and off-diagonal elements of R_h^{-1} are given by

$$-\{1 + (n_h - 2)\rho\}/[\{\rho - 1\}\{1 + (n_h - 1)\rho\}] \quad \text{and} \quad \rho/[\{\rho - 1\}\{1 + (n_h - 1)\rho\}],$$

respectively, the equation can be rewritten as a summation over individuals and households

$$\frac{1}{\pi(1-\pi)} \sum\sum w_{hj}(y_{hj} - \pi) = 0.$$

Thus, the solution $\hat{\pi}$ can be written as the weighted average of individual responses

$$\hat{\pi} = \frac{\sum\sum w_{hj} y_{hj}}{\sum\sum w_{hj}} \tag{1}$$

with the weight w_{hj} for the response from person j in household h given by

$$w_{hj} = 1/\{1 + (n_h - 1)\hat{\rho}\}. \tag{2}$$

The model-based variance estimator is

$$\text{var}(\hat{\pi}) \cong \frac{\hat{\pi}(1-\hat{\pi})}{\sum\sum w_{hj}}. \tag{3}$$

A second, robust variance estimator for $\hat{\pi}$ is available

$$\text{var}(\hat{\pi}) \cong \frac{\sum W_h^2 (\hat{\pi}_h - \hat{\pi})^2}{\{\sum W_h\}^2} \tag{4}$$

where $W_h = \sum w_{hj}$ is the sum of the weights for individuals in household h and $\hat{\pi}_h = \sum_j y_{hj}/n_h$.

The moment-based estimator of ρ takes the form

$$\hat{\rho} = \frac{\sum_h \sum_{j<k} (y_{hj} - \hat{\pi})(y_{hk} - \hat{\pi})}{\hat{\pi}(1-\hat{\pi}) \sum_h \{n_h(n_h - 1)/2\}} \tag{5}$$

based on all available unique pairs of responses within households.

The iterative 'cycling' between successive values of $\hat{\pi}$ and $\hat{\rho}$ is shown for the physician visit data in the top half of Figure 1. As expected,[1] $\hat{\pi}$ is a very weak function of $\hat{\rho}$. Indeed, in all of the software implementations examined, convergence occurs in two or fewer steps.

Copyright © 2000 John Wiley & Sons, Ltd.

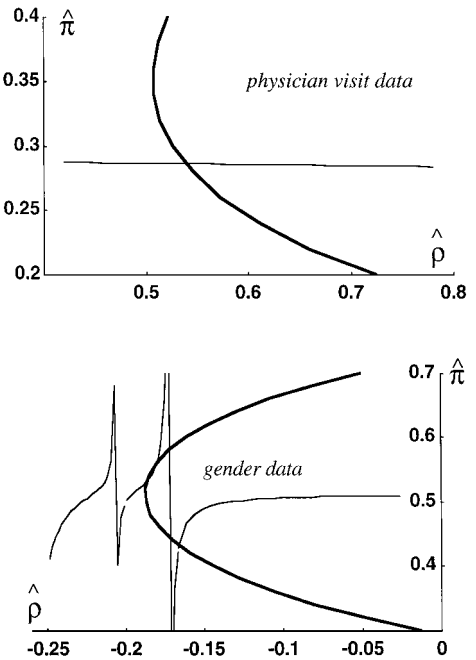

Figure 1. $\hat{\pi}$ as a function of $\hat{\rho}$ (thinner line, obtained from equation (1)) and of $\hat{\rho}$ as a function of $\hat{\pi}$ (thicker line, obtained from equation (5)). A point of intersection of the two functions represents an estimate of (π, ρ)

When the GEE approach just described is applied to the responses on gender, the [2]-[1]-[5] cycle described above ends at an estimate of $\hat{\pi} = 53{\cdot}8$ per cent, based on $\hat{\rho} = -0{\cdot}1852$; there is a large discrepancy between the model-based and robust SEs (3·5 per cent and 5·6 per cent). The GENMOD procedure in SAS yields $\hat{\pi} = 48{\cdot}2$ per cent, a correlation of $-0{\cdot}1567$, and model-based and robust SEs that are both close to 3 per cent. STATA produces an estimate of 47·57 per cent and model-based and robust SEs of approximately 3 per cent and 4 per cent, respectively. The sequence of estimates produced by the GEE solver used in S-plus fails to converge. The estimates produced by the algorithm used in the second version of the SAS macro[7] also fail to converge; estimates of π cycle between 53 per cent and 73 per cent, while those for ρ cycle between $-0{\cdot}1866$ and 0. The algorithm in the first version of the now less-used SAS macro[6] does not allow 'singletons', but when the one singleton (the 24th household in Table I) is deleted, the estimates still fail to converge.

We first discovered this problem of convergence before the GEE procedure became available in SAS PROC GENMOD, while we were still using the SAS macros. We first suspected that the starting value of $\hat{\rho} = 0$ was a poor one; however, different starting values did not lead to convergence. The reason for the non-converging sequence of estimates only became obvious to us when we reconstructed the 'trail' in the bottom half of Figure 1, which shows the two functions $\hat{\pi}[\hat{\rho}]$ and $\hat{\rho}[\hat{\pi}]$. The w's involve singularities at $\hat{\rho} = -1/6,\ -1/5,\ -1/4$, etc. All statisticians are taught at some point that an $n \times n$ correlation matrix with on- and off-diagonal elements consisting of 1's and $-1/(n-1)$'s respectively, is not invertible, but only three of the many colleagues we consulted recognized this as the reason why the matrix inversions (our prime suspect) were unstable.

Copyright © 2000 John Wiley & Sons, Ltd.

3. PREVENTIVE ACTION

To avoid these singularies, one can use an estimator of ρ which respects the limits on ρ, namely

$$\hat{\rho} \geqslant \max[-1/\{\max_h(n_h) - 1\} + \varepsilon, \hat{\rho}_{\text{equation (5)}}]$$

where ε is a small positive quantity. This is how SAS PROC GENMOD obtained the correlation of $-0{\cdot}1567$; the largest household was $n_{13} = 7$ and the procedure bounded $\hat{\rho}$ away from $-1/6(= -0{\cdot}1667)$ by an ε of 0·01. None of the other software put a bound on ρ.

To investigate the wisdom of this approach, and the sensitivity of the results to the choice of ε, we fitted estimates by specifying working correlations that were successively closer to $\rho = -1/6$. Results are shown in Table III. Because of the highly accurate numerical methods available in SAS, one can obtain GEE estimates from correlations *very* close to this 'black hole'; only when the specified correlation is within 0·0000001 of $r = -1/6$ does the procedure report that 'the working correlation has been ridged with a maximum value of 4·2072747E-6 to avoid singularity'. With this ridged value, it produces an estimate of $3/7 = 42{\cdot}86$ per cent and a SE – by both methods – of virtually zero. The 42·8 per cent is the fraction of males in the (one!) household with $n = 7$ members. The SE is zero because the estimator gives no weight to households of other sizes, but infinite weights to those in the house of seven. As might be gauged from the steep slopes in Figure 2, changing $\hat{\rho}$ just slightly has dramatic effects. Already, at $\hat{\rho} = -0{\cdot}15$ the weights for individual responses from 1- to 7-member households range from 1 to 10; at the $\hat{\rho} = -0{\cdot}1567$ used to produce the estimate of 48·2 per cent, they range from 1 to 16·7, so that almost 2/5ths of the estimate derives from the seven individuals in the household of seven. At $\hat{\rho} = -0{\cdot}16$ the weights range from 1 to 25; at $\hat{\rho} = -0{\cdot}166$ from 1 to 250; and the range increases tenfold for every additional '6' added.

A safer approach is to avoid altogether the constraints placed on $\hat{\rho}$, and instead adopt a GEE model with a working uncorrelated structure, that is, with $\rho = 0$. This independence model yields the binomial point estimate

$$\hat{\pi} = \frac{\sum\sum y_{hj}}{\sum n_h} \tag{6}$$

and robust variance estimator

$$\text{var}(\hat{\pi}) \cong \frac{\sum n_h^2(\hat{\pi}_h - \hat{\pi})^2}{\{\sum n_h\}^2}. \tag{7}$$

Applied to this example, this estimator yields an SE of 3·33 per cent, a value which – apart from small variations arising from differing values of the scale parameter used – was reproduced by all of the statistical packages examined.

4. DISCUSSION

While our example shows that we need to be careful in the use of negative intracluster correlations in GEE models, we do not wish to raise a general panic. Such examples of negative correlation are uncommon in biostatistics, and may be difficult to interpret biologically.[9] Apart from Cochran, none of the other texts or papers we consulted gives a real example. We have however encountered examples ourselves. They involved the birthweights of human twins, and

Copyright © 2000 John Wiley & Sons, Ltd.

Table III. Sensitivity of GEE estimates to values of $\hat{\rho}$ close to $-1/6$. Point estimates* and associated standard errors* for proportion of males from data in Table I

Working correlation†	Point estimate (%)	Standard error (%)	
		Empirical	Model-based
− 0·1533	48·74	3·14	3·05
− 0·1567†	48·20	3·07	2·88
− 0·16	47·39	2·91	2·62
− 0·163	45·97	2·42	2·15
− 0·165	44·77	1·68	1·68
− 0·16666	42·87	0·01	0·12
− 0·166666	42·86	0·00	0·04
− 0·1666666‡	42·86	0·00	0·02

* Using SAS GENMOD procedure[3]
† $\hat{\rho} = -0{\cdot}1567$ produced automatically with the CORR = EXCHangeable option; all other correlations user-supplied with CORR = USER option
‡ 'The working correlation has been ridged with a maximum value of 4·2072747E-6 to avoid singularity' – SAS note

the lung sizes of animal litter-mates, where nature, faced with limited space or nutrition, in an attempt to maximize survival of fewer offspring, allows considerable inequality among the individual 'competitors'. Some evidence of a somewhat weaker 'constant-sum' tendency is also evident in published animal data,[10] where pups from larger litters tend to weigh less than those from litters with fewer pups. One might wonder what the intra-household correlation in the household survey would have been had Cochran inquired about the amount of house*work* performed by each family member!

Alternatives to GEE based on sampling theory have been available for many years. For example, the computer package PC CARP[11] for multi-stage survey designs performs regression analyses by the method of weighted least squares and multinomial logistic regression by maximum likelihood.[12,13] An advantage of using PC CARP is that the regression model does not include design factors. Since additive design factors are not required, PC CARP models are less restrictive than GEE models. For the data in Table I, the results obtained from the two PC CARP methods are identical to those shown by Cochran for cluster samples.

The above example raises the question, as have other authors, of the appropriateness of using the usual Pearson correlation for binary responses. One suggestion is to use other types of correlations, such as conditional and marginal odds ratios.[14,15] These strategies are an attempt to deal with the fact that correlations for binary data are 'constrained in complicated ways by the marginal means'. However, all deal with positive correlations. Only two authors discuss negative correlations. Kupper and Haseman[16] gives as an advantage of their correlated binomial model the fact that the correlation 'can be positive or negative' whereas the intraclass correlation coefficient in the beta-binomial model of Williams[17] 'must be positive'. Prentice[18] showed that 'the correlation coefficient delta need not be positive in the beta binomial model as previously thought' but that its lower bound is

$$\text{delta}_{\min} = \max(-\pi/(n-\pi-1),\ -(1-\pi)/(n+\pi-2)).$$

Copyright © 2000 John Wiley & Sons, Ltd.

Prentice, while remarking that underdispersion is rare in application, shows how his extended beta-binomial model can accommodate such underdispersion, enabling inference in the vicinity of zero pairwise correlation. In our example, the largest n is 7 so with π and $1-\pi$ both at approximately 0.5, a lower bound is $-0{\cdot}5/(7-0{\cdot}5-1) = -1/11$ or $-0{\cdot}09$. The estimates of ρ produced by the unconstrained GEE approach used above were double that.

Equally interesting is the finding by Hendricks *et al.*[19] while carrying out a simulation study to calculate power in a GEE model for a proposed study of an intervention which will be allocated to clusters of individuals rather than individuals. They modelled between-cluster variation using a beta distribution. Understandably, failure to account for intracluster correlation led to overestimates of power as well as inflation of type I error. Somewhat surprising though was the finding that although the GEE method accounted for the intracluster correlation when present, estimates of the intracluster correlation were *negatively* biased when no intracluster correlation was present. In addition, they also found inflated type I error estimates from the GEE method and suspect that it may be related to the negatively biased estimates of intracluster correlation. We wonder if it might be due to the failure of their GEE approach to respect the theoretical parameter boundaries pointed out by Prentice.

What should one do if confronted with negative correlation? As our calculations show, it may be difficult to decide by how much one should bound ρ away from its theoretical lower bound. Instead, we suggest that in such cases, if some numerical investigation (such as we performed in Table III) shows that point estimates and SEs are not stable, one should revert to the independence working correlation and associated robust SE.

ACKNOWLEDGEMENTS

This research was supported by the Natural Sciences and Engineering Research Council of Canada (JH), NIH-CA grant 70269 (AN) and Fonds de la recherche en santé du Québec (M DeB. E).

REFERENCES

1. Liang, K. Y. and Zeger, S. L. 'Longitudinal data analysis using generalized linear models', *Biometrika*, **73**, 13–22 (1986).
2. Zeger, S. L. and Liang, K. Y. 'The analysis of discrete and continuous longitudinal data', *Biometrics*, **42**, 121–130 (1986).
3. SAS Institute. *Statistical Analysis System, release 6.12*. Release T5051 of GENMOD procedure, modified 1998 to deal correctly with clusters containing one observation (singletons), SAS Institute Inc., Cary, NC, U.S.A., 1998.
4. *gee S-function, version 4.13*, modified 98/01/27 (1998) to deal correctly with clusters containing one observation (singletons), available from http://lib.stat.cmu.edu/S/gee.
5. Stata Corporation, *Stata version* 5, *program xtgee*, updated 14 Oct 1997 to deal correctly with clusters containing one observation (singletons), Stata Corporation, College Station, Texas, U.S.A., 1997.
6. Karim, R. M. and Zeger, S. L. 'GEE: A SAS Macro for Longitudinal Data Analysis (Version - 1.25)', Department of Biostatistics, The Johns Hopkins University, Baltimore, MD, Technical Report #674, June, 1988.
7. Groemping, U. *A SAS macro for Generalized Estimating Equation - Version 2.03*, based on the SAS macro by Karim and Zeger, 1988. Fachbereich Statistik, Universitaet Dortmund, Germany, November, 1994, available from ftp://statlab.uni-heidelberg.de/pub/statlib/GEE/GEE1.
8. Cochran, W. G. *Sampling Techniques*, Wiley, New York, 1953, pp. 124–127.
9. Sokal, R. R. and Rohlf, F. J. *Introduction to Biostatistics*, 2nd edn, Freeman, New York, 1973.

Copyright © 2000 John Wiley & Sons, Ltd.

10. Dempster, A. P., Selwyn, M. R., Patel, C. M. and Roth, A. J. 'Statistical and computational aspects of mixed model analysis', *Applied Statistics*, **33**, 203–214 (1984).
11. *PC CARP*, Statistical Laboratory, Iowa State University, Ames, Iowa, U.S.A., 1989.
12. Fuller, W. A. 'Regression analysis for sample surveys', *Sankhya* C, **37**, 117–132 (1975).
13. Fuller, W. A. 'Least squares and related analyses for complex survey design', *Survey Methodology*, **10**, 97–112 (1984).
14. Diggle, P. J., Liang, K-Y. and Zeger, S. L. *Analysis of Longitudinal Data*, Clarendon Press, Oxford, 1994.
15. Carey, V., Zeger, S. L. and Diggle, P. 'Modelling multivariate binary data with alternating logistic regressions', *Biometrika*, **80**, 517–526 (1993).
16. Kupper, L. L. and Haseman, J. K. 'The use of a correlated binomial model for the analysis of toxicological experiments', *Biometrics*, **34**, 69–76 (1978).
17. Williams, D. A. 'The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity', *Biometrics*, **31**, 949–952 (1975).
18. Prentice, R. L. 'Binary regression using an extended beta-binomial distribution, with discussion of correlation induced by covariate measurement errors', *Journal of the American Statistical Association*, **81**, 321–327 (1996).
19. Hendricks, S. A., Wassell, J. T., Collins, J. W. and Sedlak, S. L. 'Power determination for geographically clustered data using generalized estimating equations', *Statistics in Medicine*, **15**(17–18), 1951–1960 (1996).

Copyright © 2000 John Wiley & Sons, Ltd.