INTERNATIONAL JOURNAL OF NATURAL AND APPLIED SCIENCES (IJNAS), VOL.4, NOS.1& 2 (2009); P. 77 – 86, 10 TABLES, 1 FIG.

Fitting drug data using generalized estimating equations as regression model

C . E. Onwukwe[*1], E. Eteng[1], J. A. Ugboh[1] and T. A. Ugbe[1]

ABSTRACT

Researchers are always interested in analyzing data that arise from a longitudinal or clustered design. Although there are a variety of standard likelihood-based approaches to analysis when the outcome variables are approximately multivariate normal, models for discrete-type outcomes generally require a different approach. Liang and Zeger (1986) formalized an approach to this problem using generalized estimating equations(GEEs) to extend generalized linear models(GLMs) to a regression setting with correlated observations within subjects. SAS Proc Genmod was used to fit a model in a drug data . The model fitted is y=0.3291–1.1553*period*–1.4994*older*+1.2542*A*+0.3404*B* .

INTRODUCTION

Generalized linear models (GLMs) (McCullagh and Nelder1989) are a standard method used to fit regression models for univariate data that are presumed to follow an exponential family distribution. Frequently researchers are interested in analyzing data that arise from a longitudinal, repeated measures or clustered design , and there exists correlation between observations on a given subject. If the outcomes are approximately multivariate normal, then there well established methods of analysis (Laird and Ware 1982). But if the outcomes are binary or counts, general likelihood based approaches are less tractable. For clustered binary outcomes, several approaches have been suggested (e.g., Fitzmaurice and Laird 1993) . Generalized estimating equations (GEEs) were developed to extend the GLM to accommodate correlated data, and are widely used by researchers in a number of fields. In this paper we will fit GEE model using statistical package SAS

BRIEF REVIEW OF GLM'S AND GEE'S

McCullagh and Nelder (1989) introduced the GLM for the exponential family data with the form $f_y(y,\theta,\phi)=\exp\left\{(y\theta-b(\theta))/a(\varphi)+c(y,\phi)\right\}$, where $a(.),b(.)\ and\ c(.)$ are given, θ is the canonical parameter, and ϕ is the dispersion parameter. The GLM is then given by $g(\mu_i) = g\left(E\left[Y_i\right]\right) = x_i'\beta,$ where x_i is a $p\times1$ vector of covariates for the i^{th} and β is a $p\times1$ vector of regression parameters. One of the attractive properties of the GLM is that it allows for linear as well as nonlinear models under a single framework .

It is possible to fit models where the underlying data are normal, inverse Guassian, gamma, Poisson, binomial , geometric and negative binomial by suitable choice of the link function $g(.)$ (Hilbe, 1994).

Liang and Zeger (1986) and Zeger and Liang(1986) introduced generalized estimating equations (GEEs) to account for the correlation between observations in generalized linear regression models. One aspect of their approach builds upon previous methods of variance estimation developed to protect against inappropriate assumptions about the variance (Huber 1967; White 1980, 1982). GEE's are used to characterize the marginal expectation of a set of outcomes as a function of a set of study variables. In a marginal model , the analyst is interested in modeling the marginal expectation (average response for observations sharing the same covariates) as a function of explanatory variables. Diggle, et al. (1994) provided a detailed review of marginal models as well as other approaches (including random effects models and transition (markov) models).

Let $Y_{ij}, i=1,---,n, j=1,---,t$ be the j^{th} outcome for the i^{th} subject, where we assume that observations on different subjects are independent, though we allow for association between outcomes observed on the same subject. In the GEE setting we are not assuming that Y_{ij} is a member of the exponential family, but we are assuming that the mean and variance are characterized as in the GLM.

*Corresponding author.
Manuscript received by the Editor January 3, 2007; revised manuscript accepted November 24, 2008.
[1]Department of Mathematics /Statistics & Computer Science ,University of Calabar, Calabar, Nigeria

We assumed the marginal regression model

$$g\left(E\left[Y_{ij}\right]\right) = x'_{ij}\beta \qquad (1)$$

where x_{ij} is a $p\times 1$ vector of study variables (covariates) for the i^{th} subject at the j^{th} outcome, β consists of the p regression parameters of interest and $g(.)$ is the link function. Common choices for the link function might be $g(a)=a$ for measured data (the identity link) $g(a)=\log(a)$ for count data (log link) or $g(a)=\log\left(\frac{a}{1-a}\right)$ for binary data (logit link). Since likelihood methods for binary do not commonly exist in general purpose statistical software, GEE's have been popular approach to regression model fitting for this type of data. For binary data with the logit link, we have that

$\log\left(\frac{E\left[Y_{ij}\right]}{1-E\left[Y_{ij}\right]}\right)=x'_{ij}\beta$, which implies that $E\left[Y_{ij}\right]=\mu_{ij}=\frac{\exp(x'_{ij}\beta)}{1+\exp(x'_{ij}\beta)}$, and since the outcomes are binary, we have that

$$\text{var}(Y_{ij})=V_{ij}=\frac{\exp(x'_{ij}\beta)}{1+eap(x'_{ij}\beta)^2}. \qquad (2)$$

In addition to this marginal mean model, we need to model the covariance structure of the correlated observations on a given subject. Assuming no missing data, the $t\times t$ covariance matrix of Y_{ij} is modeled as $V_i=\phi A_i^{1/2}R(\alpha)A_i^{1/2}$. where A_i is diagonal matrix of variance function $V(u_{ij})$ and $R(\alpha)$ is the working correlation matrix of Y_i indexed by a vector of parameters α. We will now describe specifications for R.

Specification of working correlation matrix

There are a variety of common structures that may be appropriate to use to model the working correlation matrix. *Table 1* displays a number of such matrices. Issues guiding the choice of correlation structures are beyond the scope of this paper (see Diggle et al. 1994 for a readable discussion), but in general if the number of observations per cluster is small in a balanced and complete design, then an unstructured matrix is recommended. For datasets with mistimed measurements, it may be reasonable to consider a model where the correlation is a function of the time between observations (i.e., M-dependent or auto-regressive). For datasets with clustered observations, there may be no logical ordering for observations within a cluster and an exchangeable structure may be most appropriate. Comparisons of estimates and standard errors from several different correlation structures may indicate sensitivity to misspecification of the variance structure. For both the independence working structure and the fixed working structure, no estimation of α is performed. We note that use of the exchangeable (also referred to as compound symmetry) working correlation matrix with measured data and identity link function is equivalent to a random effects model with a random intercept per cluster. Fixed working correlation in *Fig.1* is symmetric with 1's on the diagonal, specifies a banded structure with a fixed correlation and linear decline as the distance between observation increases.

$$\begin{pmatrix} 1.0 & 0.9 & 0.8 & 0.7 \\ 0.9 & 1.0 & 0.9 & 0.8 \\ 0.8 & 0.9 & 1.0 & 0.9 \\ 0.7 & 0.8 & 0.9 & 1.0 \end{pmatrix}$$

Fig.1 Example of Fixed Working Correlation Matrix

Empirical and model based variance estimators

Zeger and Liang(1986) referred to V_{ij} as a "working " matrix because it is not required to be correctly specified for the parameter estimates in model (1) to be consistent (as long as the mean model itself is correct and there is no missing data). However, Liang and Zeger (1986) showed that there can be important gains in efficiency realized by correctly specifying the working correlation matrix.

A set of estimating equations are solved (through an iterative process) to find the value of the estimator $\hat{\beta}$. An empirical variance estimator can be used to estimate $\text{var}(\hat{\beta})$. This variance estimator is also referred to as a "sandwich" or "robust" estimator. Another variance estimate available from GEE models is the model-based (or "naive") estimate, which is consistent when both the mean model and the covariance model are correctly specified.

Table 1. Fixed Working Correlation

Common Working Correlation Models

Structure	*Definition*	*Example*	*No. Parameters*
Independent	$R_{u,v}=\begin{cases}1, & if \quad u=v\\ 0, & otherwise\end{cases}$	$\begin{pmatrix}1 & 0 & \cdots & 0\\ 0 & 1 & \cdots & 0\\ & & & \\ 0 & 0 & \cdots & 0\end{pmatrix}$	0
Exchangeable	$R_{u,v}=\begin{cases}1, & if \quad u=v\\ \rho, & otherwise\end{cases}$	$\begin{pmatrix}1 & \alpha & \cdots & \alpha\\ \alpha & 1 & \cdots & \alpha\\ & & & \\ \alpha & \alpha & \cdots & \alpha\end{pmatrix}$	1
Unstructured	$R_{u,v}=\begin{cases}1, & if \quad u=v\\ \rho_{u,v}, & otherwise\end{cases}$	$\begin{pmatrix}1 & \rho_{1,2} & \cdots & \rho_{1,t}\\ \rho_{1,2} & 1 & \cdots & \rho_{2,t}\\ & & & \\ \rho_{1,t} & \rho_{2,t} & \cdots & 1\end{pmatrix}$	$\frac{t(t-1)}{2}$
Auto-regressive	$R_{u,v}=\begin{cases}1, & if \quad u=v\\ \rho^{\lvert u-v\rvert}, & otherwise\end{cases}$	$\begin{pmatrix}1 & \rho & \cdots & \rho^{t-1}\\ \rho & 1 & \cdots & \rho^{t-2}\\ & & & \\ \rho^{t-1} & \rho^{t-2} & \cdots & 1\end{pmatrix}$	1
1dependent	$R_{u,v}=\begin{cases}1, & if \quad u=v\\ \rho_{\lvert u-v\rvert}, & otherwise\end{cases}$	$\begin{pmatrix}1 & \rho_1 & \cdots & \rho_{t-1}\\ \rho_1 & 1 & \cdots & \rho_{t-2}\\ & & & \\ \rho_{t-1} & \rho_{t-2} & \cdots & 1\end{pmatrix}$	$0<M\le t-1$
Fixed	$R_{u,v}=\begin{cases}1, & if \quad u=v\\ r_{u,v}, & otherwise\end{cases}$	$\begin{pmatrix}1 & r_{1,2} & \cdots & r_{1,t}\\ r_{1,2} & 1 & \cdots & r_{2,t}\\ & & & \\ r_{1,t} & r_{2,t} & \cdots & 1\end{pmatrix}$	0(user specified

Since in general the analyst will not know the correct covariance structure, the empirical variance estimate will be preferred when the number of cluster is large. When the number of clusters is small, say < 20, the model based variance estimator may have better properties (Prentice 1988) even if the "working variance" is wrong. This is because the robust variance estimator is asymptotically unbiased , but could be highly biased when the number of clusters is small.

APPLICATION

Longitudinal or clustered studies often have missing data, either by design or happenstance. If a litter in a teratology study is the level of clustering, litter size may vary between litters. Patients in an observational study may miss appointments or drop out of the study. The protocol for a clinical trial may call for patients to be observed at specified intervals, but their actual observations may take place at varying times. Such unbalanced and/or incomplete data can complicate GEE analyses. If the missingness can be thought of as being missing completely at random (MCAR) in the sense of Little and Rubin(1987), then the consistency results established by Liang and Zeger (1986) hold. However, the notation and calculations for arbitrary missing data patterns are more complicated than in the balanced and complete case. Robins et al (1995) proposed methods to allow for data that is missing at random (MAR). Their inverse probability censoring weight (IPCW) approach requires that the missingness law be modeled and that weights corresponding to the inverse probability of missingness be included in the GEE. This will yield consistent parameter estimates, but the variance will tend to be incorrect (since the weights are being estimated but are treated as constants by default) . Unfortunately, the method of Robins et al (1995) only works well when there is dropout-that is, once a subject misses a time, that subject is not seen again. Often subjects miss a single observation, and then are seen at the net time. The probability of missingness pattern over time is not estimable with a simple logistic regression in this case, so the Robins et al (1995) method is more difficult to implement. Laird proposed a modification to the GEE approach that combines restricted maximum likelihood (REML) estimating equations for the parameters in the variance-covariance matrix.The variable CARRY takes the value N(no) if the observation is from the first period, it takes the value A or B if it comes from the second period and the treatment in the first periods A or B, respectively.

If the subject received the placebo in the first period, the value of CARRY is also set to N for the observations in the second period. The following PROC GENMOD statement fit the GEE model. Since there are 300 subjects in the crossover study there are 300 clusters or experimental units in the GEE analysis. With responses for both periods, the cluster size is two. There are no missing values, so both the minimum and maximum cluster size two. A logistic regression analysis is appropriate for these data so DIST=BIN is specified in the MODEL statement. The logit link is used by default. Both SUBJECT and AGE are specified in the CLASS statement, since AGE reflects a classification into groups. The model includes main effects for period, age, drug and carryover effects and interactions for period and age and drug and age. The option TYPE=UNSTR specifies the unsaturated correlation structure. Since there are only two measurements per subject, this is the same as exchangeable structure.

RESULTS

From Table 2, since there are 300 subjects in the crossover study, there are 300 clusters or experimental units in the GEE analysis. With responses for both periods, the cluster size is two. There are no missing values, so both the minimum and maximum cluster size is two. *Tables* 3 and 4 showed that the score statistic for the two – level carry variable is 1.15 with p - value equal to 0.5626. In addition, the age × drug interaction appears to be unimportant, with a score chi – square statistic of 0.72 for 2 df (p = 0.6981) see *Table* 7. The joint test is definitely nonsignificant, with a chi – square value of 1.31 for 4 df and a p – value 0.8595. The Typ3 tests indicate that period, age and drug are highly significant with a p – value of 0.0240, the period×age interaction cannot be dismissed. The unstructured correlation structure is the same as the exchangeable correlation structure when you have two responses per cluster. The correlation is estimated to be 0.1959. Next test for differences whether two parameters for drugs A and B is equal to zero was carried out. The single degree of freedom test and the chi – square value of 19.15 for the score test is highly significant (see *Table*1)

Table 2. Working correlation matrix

Obs	age	sequence	time1	time2	i	subject	period	drug	carry	response
1	older	AB	F	F	1	1	1	A	N	F
2	older	AB	F	F	1	1	0	B	A	F
3	older	AB	F	F	2	2	1	A	N	F
4	older	AB	F	F	2	2	0	B	A	F
5	older	AB	F	F	3	3	1	A	N	F
6	older	AB	F	F	3	3	0	B	A	F
7	older	AB	F	F	4	4	1	A	N	F
8	older	AB	F	F	4	4	0	B	A	F
9	older	AB	F	F	5	5	1	A	N	F
10	older	AB	F	F	5	5	0	B	A	F
11	older	AB	F	F	6	6	1	A	N	F
12	older	AB	F	F	6	6	0	B	A	F
13	older	AB	F	F	7	7	1	A	N	F
14	older	AB	F	F	7	7	0	B	A	F
15	older	AB	F	F	8	8	1	A	N	F

.

Table 3. Parameter Information

Parameter	Effect	age	drug	carry
Prm1	Intercept			
Prm2	period			
Prm3	age	older		
Prm4	age	younger		
Prm5	drug		A	
Prm6	drug		B	
Prm7	drug		P	
Prm8	period*age	older		
Prm9	period*age	younger		
Prm10	carry			A
Prm11	carry			B
Prm12	carry			N
Prm13	age*drug	older	A	
Prm14	age*drug	older	B	
Prm15	age*drug	older	P	
Prm16	age*drug	younger	A	
Prm17	age*drug	younger	B	
Prm18	age*drug	younger	P	

Table 4. Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	590	730.8056	1.2387
Scaled Deviance	590	730.8056	1.2387
Pearson Chi-Square	590	597.8187	1.0133
Scaled Pearson X2	590	597.8187	1.0133
Log Likelihood		-365.4028	

Table 5. Score Statistics For Type 3 GEE Analysis

Source	DF	Chi-Square	Pr>Chisq
Period	1	4.61	0.0318
age	1	36.03	< .0001
drug	2	27.66	< .0001
Period*age	1	4.69	0.0303
Carry	2	1.15	0.5626
age*drug	2	0.72	0.6981

Table 6. Analysis Of GEE Parameter Estimates Empirical Standard Error Estimates

Parameter		Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept		0.3291	0.4115	-0.4774	1.1356	0.80	0.4239
period		-1.1553	0.3406	-1.6424	-0.3072	-2.86	0.0042
age	older	-1.4994	0.3345	-2.2620	-0.9509	-4.80	<.0001
drug	A	1.2542	0.3623	-0.3129	1.1074	1.10	0.2729
drug	B	0.3404	0.3600	0.5001	1.9113	3.35	0.0008

y = 0.3291 - 1.1553 period - 1.4994 older + 1.2542 A + 0. 3404 B.

Table7. Contrast Results for GEE Analysis

Contrast	DF	Chi-Square	Pr > ChiSq	Type
carry	2	1.15	0.5626	Score
inter	2	0.72	0.6981	Score
joint	4	1.31	0.8595	Score

Table 8. Working Correlation Matrix

	Col1	Col2
Row1	1.0000	0.1959
Row2	0.1959	1.0000

Table 9. Score Statistics For Type 3 GEE Analysis

Source	DF	Chi-Square	Pr>Chisq
Period	**1**	**24 . 98**	**<0.0001**
age	**1**	**35 . 53**	**< .0001**
drug	**2**	**39 . 31**	**< .0001**
Period*age	**1**	**5 . 10**	**0.0240**

Table 10. Contrast Results for GEE Analysis

Contrast	DF	Chi-Square	Pr > ChiSq	Type
A versus B	1	19.15	< 0.0001	Score

CONCLUSION

The joint test is definitely nonsignificant, with a chi-square of 1.31 for 4 df and a *p – value* of 0.8595 in *Table 7* . The unstructured correlation structure is the same as exchangeable correlation structure when you have two responses per cluster. The correlation is estimated to be 0.2274. This is a single degree of freedom test and the chi-square value of 19.15 *Table 10* for the score test is highly significant. The type 3 test indicates that period, age and drug are highly significant. With a *p-value* of 0.0240. *Table 9*, the period*age interaction cannot be dismiss

y=0.3291–1.1553*period*–1.4994*older*+1.2542*A*+0.3404*B*

REFERENCES

Diggle, D. A., Liang, K. Y. and Zeger, S. L. (1994). *Analysis of Longitudinal Data*, Clarendon Press, London.

Fitzmaurice, G. M. and Laird, N. M. (1993). A likelihood-based method for analysing longitudinal binary responses, *Biometrika*, 80: 141-151.

Hilbe, JU. M. (1994). Generalized linear models, *The American Statistician*, 48:255-265.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data, *Biometrics*, 38:963-974.

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

Little, R. J. A., and Rubbin, D. B. (1987). *Statistical Analysis Missing Data* , Wiley, New York.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, Chapman and Hall, New York: 266pp

Prentice, R. L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics*, 44: 1033-1048.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1995). Analysis of semi-parametric regression models for repeated outcomes in the presence of missing data, *Journal of the American Statistical Association*, 90: 106-121.

White, H (1980).A Heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity, *Econometrica*, 48: 817-838.

White, H (1982). Maximum likelihood estimation of misspecified models *Econometrica*, 48:50, 1-25.

Zeger, S. L. and Liang, K. Y. (1986). Longitudinal data analysis for discrete and continuous outcomes, *Biometrics*, 42: 121-130.