Paper 263

Draft: Simplification on Learning Model by Using PROC GENMOD

Kuolung Hu
Karen Nowak
David McKinzie
Institute of Psychiatric Research, Indiana University School of Medicine
Indianapolis, IN

Abstract

Longitudinal data sets are comprised of repeated observations of an outcome and a set of covariates for each of many subjects. One objective of the statistical analysis of such data is to describe the marginal expectation of the outcome variable as a function of the covariates, while accounting for the correlation among the repeated observations for a given subject. In this paper, we apply the generalized estimating equation (GEEs) approach to fit one of the class of models, learning model, with repeated binary outcomes which is common in the biological and social sciences. Our intention was to discuss what we consider to be a recently widely applicable methodology for longitudinal data, discuss its advantages and disadvantages, and illustrated by SAS® system. The method is demonstrated with an analysis of data obtained from a dog-shock experiment and a delay of reward experiment between alcohol preferring and non-preferring rats.

Introduction

One objective in the analysis of longitudinal data sets is to model the marginal expectations of the outcome as a function of the predictor variables. Because repeated observations are made on each single subject, the correlation is anticipated among subject’s measurements. It must be properly accounted for between the observations to obtain a correct statistical analysis.

To construct a marginal model for multivariate binary data and explore the relationship of the response with explanatory variables, Zeger and Liang (1986) and Liang and Zeger (1986) propose a moment methods methodology, generalized estimating equations (GEEs), for discrete longitudinal data that used the quasi-likelihood (Wedderburn, 1974) approach. They recommend the quasi-likelihood method because of the sparseness of multivariate distributions for non-Gaussian data. They specify that a known function of the marginal expectation of the dependent variate, followed a generalized linear model (McCullagh and Nelder, 1989), is a linear function of the covariates, and assume that the variance is a known function of the mean. In addition, they specify a “working” correlation matrix for the observations for each subject. This configuration leads to generalized estimating equations which give consistent estimators of the regression coefficients and of their variances under weak assumptions about the actual correlation among a subject’s observations.

Furthermore, Lipsitz *et al.* (1994) proposed one-step generalized estimating equations and compared the performance to that of a fully iterated estimator in small sample simulations. They found out that GEEs are more efficient than ordinary logistic regression with variance correction for estimating the effect of a time-varying covariate. Therefore, the learning model, the presentation of an event is changing stochastically, depending on what type of events happened at one or more immediately preceding times, would be considered a practical example to be utilized by GEEs methodology.

Learning Model

Learning model, which constitutes the accumulation of events of different types, are commonly used in the behavioral sciences. The occurrence of an event to be modeled depends on the total number of previous events of different types. For instance, it can determine which types of event experiences most influence being recorded, and its probability depends on accumulated previous such events.

In previous methodology, a transitional model (Lindsey, 1993) was proposed. For example, considering the Solomon-Wynne experiment in which a dog was required to learn to avoid a shock (Kalbfleisch, 1985, pp.83-88). The dog was in a compartment with a floor through which shock could be applied. When the lights were turned out and a barrier raised; ten second later, the shock occurred. Thus, the dog had ten seconds, after the lights go out, to jump the barrier and avoid the shock.

Suppose that the subject learns from previous trials. The probability of avoidance will depend on the number of previous shocks ($y_{ij}=0$) and on the number of previous avoidances. ($y_{ij}=1$). Let P_j be the probability of a shock at trial j (j=1,....,25), given performance on previous trials, and X_{ij} be the number of previous avoidances before trial j. Then j-X_{ij} is the number of previous shocks. We use the model

$$p_j = p_1^{Xij} p_2^{j-Xij}$$

or

$$\log(Pj) = \alpha Xij + \beta(j - Xij)$$

where

$$\alpha = \log(p_1)\ and\ \beta = \log(p_2).$$

Because

$$\frac{P_j}{P_{j-1}} = \begin{cases} P_1\ if\ \ X_{i,j-1} = 1 \\ P_2\ if\ \ X_{i,j-1} = 0 \end{cases}$$

, the probability of shock changes by a factor of P1 if there was an avoidance at that previous trial or P2 if there was a shock. Furthermore, with the proper link function, we can obtain these parameters estimated by the logistic regression.

Generalized Linears Models

The class of generalized linear models is an extension of the traditional linear models with three distinct components that can be specified by the user:

1. *The response distribution* Yi(i=1,....,N) are independent random variables with means u_i and observed values $y_i=u_i+e_i$, where the e_i are the residuals or 'errors'.

2. *The linear predictor* A set of T unknown parameters, B=(B_1,....,B_T)^T, and the corresponding set of known explanatory variables $X_{nxT}=(X_1,....,X_n)^T$, the design of the model matrix, are such that:

$$\eta_i = \sum_{t=1}^{T} x_{it}\beta_t$$

define the linear predictor. This describes how the location of the response distribution changes with the explanatory variables. In the general case, the shape of the distribution changes as a function of location, because the variance depends on the mean in the distribution. It is not observed in a simple linear normal regression. This leads to the third point.

3. *The link function* the relationship between the mean of the ith response and its linear predictor is given by the link function $g_i(.)$:

$$\eta_i = g_i(u_i) = x_i^T \beta$$

Notice that it must be monotonic and differentiable.

These link functions lead to an important property. With the canonical link function, all unknown parameters of linear structure have sufficient statistics if the response distribution is a member of the exponential family and the scale parameter is known. Many useful statistical models can be formulated as generalized linear models by selection of an appropriate link function and response probability distribution.

GEEs Estimate

Let Y_{it} have the likelihood

$$f(y_{it}) = \exp\{[y_{it}\theta_{it} - a(\theta_{it}) + b(y_{it})]/\phi\}$$

Corresponding to the generalized linear model, where $\theta_{it} = h(\eta_{it})$, and $\eta_{it}=X'_{it}\beta$. Liang and Zeger (1986) consider 'Score-like' estimating equations of the form:

$$\sum_{i=1}^{n} \mathbf{D}_i^{`} \mathbf{V}_i^{-1} \mathbf{S}_i \qquad (1)$$

where S_i is the $Tx1$ vector of deviations, Y_{it} - E(Y_{it}), for the ith individual (where E(Y_{it})=d$a(\theta_{it})$/dθ_{it}), Vi is the T x T 'working' covariance matrix for Yi, and Di= d{E(Yi)}/dβ.

Vi is represented as

$$\mathbf{V}i = \mathbf{A}_i^{1/2}\,\mathrm{Corr}(\mathbf{Y}_i)\,\mathbf{A}_i^{1/2}$$

Where Ai (=diag[$var(Y_{it})$]) is specified by the marginal distributions. The estimation algorithm consists of specifying a model for Corr(Y*i*) and cycling between of β via (1) and Corr(Y*i*) by the method of moments. Under the mild conditions, the estimator of β is asymptotically unbiased and normal for any choice of Corr(Y*i*), with asymptotic variance depending on both the assumed and true covariance pattern. Moreover, a consistent variance estimate is also available under equally weak conditions. The method allows time-varying covariates and could be adapted to allow missing observations. By specifying restricted forms for Corr(Y*i*) such as constant diagnosis or first-order autoregressive structure, one can limit the number of parameters when *T* is large.

Progabide Example

The GENMOD procedure in SAS Version 6.12 system fits generalized linear models to the data by maximum likelihood estimation of the parameters. This class of generalized linear models is an extension of traditional linear models that allow the population mean to depend on a linear predictor through a nonlinear link function and the response probability distribution to be any member of an exponential family. GENMOD also provided the GEEs model fitting by quasi-likelihood approach, controlled with new REPEATED statement. Furthermore, there are exchangeable, unstructured, auto regressive(1), independent ,m-dependent and self-specified correlation structures available. Therefore, we could fit the model by GEEs with several different correlation structures with TYPE=option in REPEATED statement. For more details, refer to Johnston and Stokes (1997) and SAS/STAT references.

1. Dog-Shock Experiment

Back to the Solomon-Wynne experiment on dog shock. The following statements input the data, which are arranged as one outcome per trial:

```
data temp;
do id=1 to 30;
do trial=1 to 25;
input outcome @;
output;
end;
end;
cards;
0 0 1 0 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
0 0 0 0 0 1 1 0 1 1 0 0 1 1 0 1 0 1 1 1 1 1 1 1 1
..............;
```

In this analysis, the marginal relationship between the response variable and the covariates was modeled. We simply applying the PROC GENMOD with GEEs to fitting the marginal model:

```
*/ For GEES calculation /*;
proc genmod;
where trial notin (1);
class id ;
model outcome=trial/dist=bin;
repeated subject=id/covb type=ind;
run;
```

The degrees of freedom and deviance are listed below. Without the mass effort to count those accumulated success events and probabilities like the transitional model, we could easily fit this model to access the effect of the covariates. We may also notice that there is not much of a difference between these two models in the deviance/DF ratio:

Model Structure	DF	Deviance	Deviance/ DF
Transitional Model	718	552.2	0.7690
GEEs	718	567.31	0.7901

Furthermore, the output for the parameter estimate in GEEs are listed below:

```
             Analysis Of GEE Parameter Estimates
             Empirical Standard Error Estimates

               Empirical  95% Confidence Limits
Parameter  Estimate  Std Err  Lower   Upper   Z    Pr>|Z|

INTERCEPT -2.1578  0.2561 -2.6598 -1.6558 -8.424 0.0000
TRAIL      0.2774  0.0275  0.2235  0.3313 10.090 0.0000
Scale      1.0093     .       .       .      .      .
```

The following table displays the estimated regression coefficient, empirical standard errors for intercept and trial effects with four different kinds of working correlation structures; independent, unstructured, exchangeable and auto regressive (1), specified.

Model	INTERCEPT	TRIAL	SCALE
GEEs with IND	-2.1578 (0.2561)	0.2774 (0.0275)	1.0093
GEEs with UN	-1.5962 (0.2543)	0.2263 (0.0232)	0.9291
GEEs with EXCH	-2.1454 (0.2645)	0.2792 (0.0272)	1.0218
GEEs with AR	-2.1612 (0.2559)	0.2775 (0.0276)	1.0093

Note: (.) inside is the empirical std error

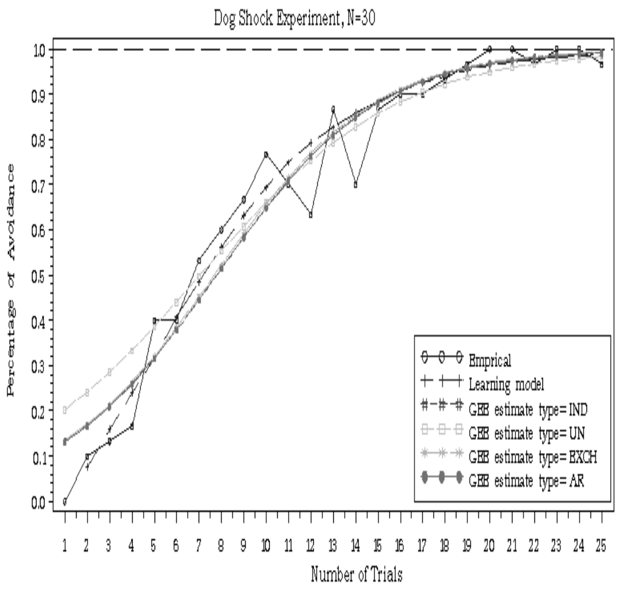

Figure 1. Dog Shock Experiment

With the parameter estimate, we can plot the estimated response curves (see figure 1) to identify which one could perform similarly to the empirical percentage of avoidance. Therefore, we may say that with the exception of the unstructured working correlation model, the rest of the GEES estimates and the transitional estimates are all sensitive to the empirical curve.

By specifying the COVB option, we can produce the model-based and the empirical covariance matrix in order to examine the adequacy of the correlation model we built.

```
                Covariance Matrix (Model-Based)
    Covariances are Above the Diagonal and Correlations
                          are Below

   Parameter
  Number                PRM1        PRM2
  PRM1               0.05359   -0.004517
  PRM2              -0.88981   0.0004809

                 Covariance Matrix (Empirical)
    Covariances are Above the Diagonal and Correlations
                          are Below

   Parameter
  Number                PRM1        PRM2
  PRM1               0.06561   -0.005840
  PRM2              -0.82931   0.0007558
```

2. Impulsivity between Alcohol Preferring (P) and Non-preferring (NP) rats

This example is derived from an experimental condition reported by Nowak *et al.*(1998). Male P and NP rats (n=10/group) were tested for their responses to delay-of-reward in a T-maze apparatus. Rats were first trained to traverse a runway and choose a side (L or R) which contained either a small (1 Kix) or large (5 Kix) reward. The side for the large reward was kept constant for each animal; but counterbalanced across animals. When the large reward was chosen >90% of the time, a 5 seconds delay was then imposed, in which, the rat was prevented from reaching the goal box when choosing the large reward. There was no delay for access to the small reward. Rats received 5 trials/day for 5 consecutive days. The outcomes (L or R) were recorded. The objective was to determine whether group differences existed in the percentage choice of large reward with repeated exposure to a 5 second delay.

This model was fit in the SAS system using the GENMOD procedure, which has been well-illustrated from the previous example. The following lines transferred the outcome to a binary data set (0=small vs. 1=large):

```
data temp1;
set temp;
array c[*] c1-c25;
do i=1 to 25;
if c[i]=side then c[i]=1;
else c[i]=0;
end;
do i=1 to 25;
outcome=c[i];
trial=i;id=rat_id;output;
end;
```

Furthermore, looking at the group and trial factors, we apply the REPEATED statement in the GENMOD procedure with the correlation structure specified. We would obtain the GEEs solutions:

```
*/ For GEES calculation /*;

proc genmod data=temp1;
class id group;
model outcome= group trial /dist=bin type3;
repeated subject=id/covb type=ind;
run;
```

The information generated from GEE option is as follows:

```
                Covariance Matrix (Model-Based)
    Covariances are Above the Diagonal and Correlations
                          are Below
Parameter
Number        PRM1        PRM2        PRM4

PRM1       0.05270    -0.01185   -0.002732
PRM2      -0.24643     0.04388   -0.000373
PRM4      -0.82580    -0.12364   0.0002077
```

```
              Covariance Matrix (Empirical)
     Covariances are Above the Diagonal and Correlations
                         are Below

 Parameter
 Number        PRM1        PRM2        PRM4

 PRM1       0.02365    -0.03730   -0.000338
 PRM2      -0.59041     0.16875   -0.001988
 PRM4      -0.13864    -0.30568   0.0002506
```

The deviance of this model is 570.9031 with 497 *d.f.* and the ratio is 1.1487. Furthermore, the following table displays the estimated regression coefficient, empirical standard errors and standardized statistics (estimates/s.e.) for intercept, group and trial effects by four correlation structures and acquisition curve plots (see figure 2) are also listed below:

Covariate	Correlation	Regression Coefficients		
	Structure	Est.	Std. Err	Est./S.E.
Intercept	UN	0.7589	0.0909	8.3480
	IND	0.9143	0.1538	5.9453
	EXCH	0.8978	0.1526	5.8854
	AR(1)	0.9163	0.1539	5.9557
Group	UN	1.4609	0.2324	6.2855
	IND	1.3980	0.4108	3.4032
	EXCH	1.4241	0.4569	3.1166
	AR(1)	1.3996	0.4111	3.4049
Trial	UN	-0.0474	0.0110	-4.323
	IND	-0.0626	0.0158	-3.952
	EXCH	-0.0626	0.0160	-3.905
	AR(1)	-0.0627	0.0158	-3.958

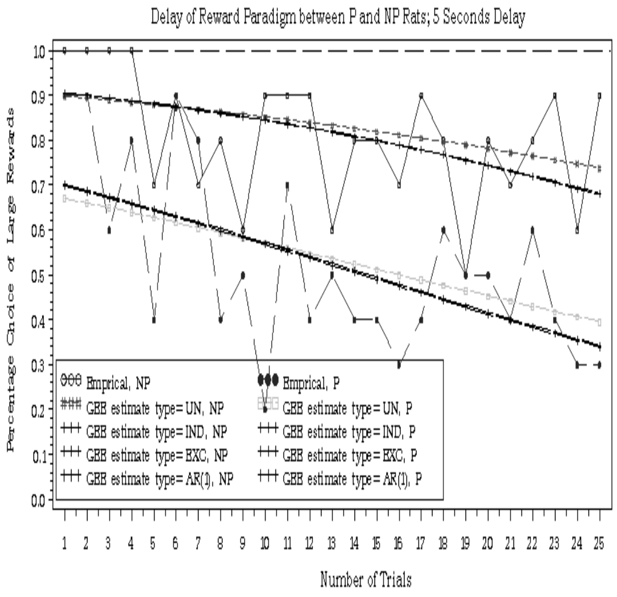

Figure 2 Delay of Reward Experiment

However, we can tell that the estimate lines for the two groups are most likely parallel, indicating a possible group effect and ,furthermore, the estimates of the odds of choosing a large reward versus a small reward are approximately $e^{1.398}$= 4.047 times higher for the NP rats than for the P rats.

Conclusion

We applied the marginal models in the repeated multivariate binary response, instead of a transitional (conditional) one, in the learning model. While modelling the relationship of the response with explanatory variables is the objective, a GEEs method is easy to implement and gives efficient estimates of regression coefficients under weak correlation assumptions. However, estimates of the association among the binary outcomes could be inefficient.

References

Johnston, G and Stokes, M (1997), “Repeated Measures Analysis with Discrete Data using the SAS System” Proceeding of twenty-two SAS User Group Conference International, 1300-1305.

Kalbfleisch, J.g. (1985), “Probability and Statistical Inference”, vol2, Springer Verlag, Berlin.

Liang, K-Y, and Zeger, S.L. (1986), “Longitudinal Data analysis Using Generalized Linear Models”, Biometrika, 73,1,13-22.

Lindsey, J.K.(1993), “Models for Repeated Measurements”, Oxford : Clarendon Press; New York.

Lipsitz, S.R., Fitzmaurice, G.M., Orav, E.J., and Laird N.M. (1994), “Performance of Generalized Estimating Equations in Practical Situations”, Biometrics 50, 270-278.

Mccullagh, P. and Nelder, J.A.(1989), “Generalized Linear Models”, 2nd ed., Chapman and Hall, London.

Nowak, K.L., McKinzie, D.L., Dagon, C.L., Murphy, J.M., Mcbride, W.J., Lumeng, L. and Li T.-K.(1998). Differences in an Operant Measurement of Impulsivity between Alcohol-preferring P and non-preferring NP Rats. Research Society on Alcoholism, Hilton Head, S. Carolina.

SAS Institute Inc. (1996), “SAS/STAT Software Changes and Enhancement”, Cary , N.C., SAS Institute Inc.

Wedderburn, R.W.M. (1974), “Quasi-likelihood functions, Generalized Linear Models, and the Gauss-Newton Method”, Biometrika, 61, 439-447.

Zeger, S.L. and Liang K-Y (1986), “Longitudinal Data Analysis for Discrete and Continuous Outcomes”, Biometrics, 42, 121-130.

Acknowledgments

SAS and SAS/STAT are registered trademarks of SAS institute Inc. in the USA and in other counties. ® indicates USA registration.

Contact Information

Your comments and questions are values and encouraged. Contact the author at:

Kuolung Hu
Institute of Psychiatric Research
Indiana University, School of Medicine
791 Union Drive
Indianapolis, IN 46202
Work Phone: (317) 278-1381
Fax: (317) 274-1365
e-mail: khu@iupui.edu