Paper 251-2009

Using GEE to Model Student's Satisfaction: A SAS ® Macro Approach

Teck Kiang Tan, Trivina Kang, David Hogan
Centre for Research in Pedagogy and Practice, Nanyang Technological University, Singapore

ABSTRACT

This paper illustrates analysis of longitudinal data using GEE (generalize estimation equations) and shows how output from SAS macros can be streamlined and organized to aid interpretation of the analysis. Although using GEE through procedures such as SAS PROC GENMOD is becoming increasingly commonplace, as far as model evaluation is concerned, its widespread use is somewhat limited by the lack of easily accessible measures to evaluate model goodness-of-fit directly from the default SAS output. This paper gives an example of how this can be done by first building three goodness-of-fit indices, namely the marginal R^2, QIC and QICU, in a SAS macro, using various working correlation matrices, for model comparison. The macro outputs these indices in a summarized SAS output as well as outputs the estimated coefficients, SEs and P-values of these models in an organized way that can be easily exported into EXCEL. With minor adjustments in the EXCEL sheet, these results could be presented in publication ready format. Specifically, we illustrate with a longitudinal data set, how four GEE models with different working correlation matrices specification, with a binomial logit link function, were generated using the macro. The results show that estimated coefficients for the four models were largely similar; supporting Zeger and Liang (1986) point that misspecification of working correlation would still give consistency result. This paper also illustrates the procedure of data management and preliminary data analyses work needed before carrying out similar GEE analyses using several simple SAS macros. These include carrying out statistical procedures such as factor analysis for examining constructs reliability, calculating reliability index Cronbach's Alpha and generating principal component score. Suggestions for future analyses are also provided in this paper.

INTRODUCTION

Using the GEE models, this paper purposes to use SAS macros to extract information from PROC GENMOD and output to a format that is well organized which contains only estimated coefficients, SE and p-value and goodness-of-fit statistics for model comparison. Using the dataset from the centre of research in pedagogy and practice, this paper shows the empirical results of the effects of various family, school and self belief on student's life satisfaction after account for the correlated life satisfaction of students' responses over time. Using different working correlation matrices specification to account for correlated responses, it seeks to show how these models affect the estimations using a SAS macro. In order to carry out the data management and various statistical analyses necessarily for GEE analyses, various smaller SAS macros were written as well. These macros help to organize the SAS program in a structured manner and present output model estimation and goodness-of-fit measures in an organized format that facilitates user interpretation. The paper first presents the purpose of the paper and discusses the method used. It then gives a brief background of GEE models and the three measures used for model comparison. The paper concludes with a discussion of the estimation results and its recommendation for future research.

PURPOSE

The main objective of this paper is to demonstrate the application SAS macros in an efficient way to facilitate the use of GEE models in analyzing longitudinal discrete binary data. The macro %SelectGEE incorporates three measures of evaluating GEE model fit, namely marginal R^2, QIC and QICU, using various working correlation

matrices, for model comparison. The output of these models are summarized into one SAS output with their estimated coefficients, SEs and P-values stored in one dataset which allows it to be easily exported to EXCEL and thus be ready for publication purpose. The paper also briefly presents the procedure of data management for carrying out GEE modeling and how several simple SAS macros for preliminary statistical analyses can be used to make the SAS program easily traceable and tractable. The results of GEE modeling were discussed and suggestions for future modeling strategies were recommended.

METHOD

Data Source and Sample

The data used in this paper is obtained from the first three waves of the life pathway project conducted by the Centre for Research in Pedagogy and Research (CRPP), Nanyang Technological University, Singapore. The total sample consisted of 2,289 Singapore students obtained from nationally representative sample of 39 schools. The first wave of the data started in the year 2005 when the students were in secondary one (Grade 7) and continued annually till students were in Grade 9 in 2007. The selected sample is a representative of the Singapore secondary school population. A stratified sampling frame was used in the study. The schools were classified into three strata according to their past academic performance and equal number of schools randomly selected from each of the stratum.

VARIABLES

The study includes nine manifest variables and six constructs in the GEE analyses. The manifest variables include gender, race, current stream, prior stream, number of siblings, number of closed friends, parent's marital status, anxious about examination, and worry about examination. The constructs are social support, family functioning, mental morbidities, existential satisfaction, self esteem, and body image. There are three main ethnic/racial groups in Singapore, namely Chinese, Malay, and Indian and they are included in the analysis as standalone categories. Students from other ethnic/racial groups were grouped into one named as "other ethnic group" since they are very small in terms of percentage and number. Students' streams/tracks in primary/elementary school range from EM1 (highest) to EM2 and EM3 (lowest). Students' current stream/track in secondary school is also included and range from the highest to the lowest (i.e. gifted/special, express, normal academic, and normal technical). Gender, race, current and prior stream are dummy coded. Since most of the students come from families with parents who are married, this was also dummy coded as a dichotomous variable. The number of indicators of the six constructs ranges from 4 to 8 and are measured on a 6 point likert scale

DATA PROCESSING

During the survey period, data were captured in person-level format (Singer and Willet, 2003) and stored in three separate datasets. They were reshaped into long format or person-period format so that it could be inputted to SAS PROC GENMOD for GEE analyses. These three datasets were concatenated to form a dataset with three times the size of the original datasets in the long format. Factor analyses were carried out for the six constructs and their reliability were examined using macro %Factor and %Alpha respectively. These constructs were converted into principal component scores using the macro %FScore. One of the main rationales of using macro in SAS programming is that it provides routine and repetitive procedures in a compact form. This makes the organization of SAS programming statement neat, tidy, short and traceable. The %FScore below showed that a six line of codes could be summarized into one line and by a glance of the macro specification statements [e.g. %FScore(…,

BodyImage) and %FScore(…, SelfEsteem)], we knew straight away the number of constructs have been generated into scores. In contrast, it would be more difficult to count the number of principal scores generated if we have all the PROC statements layouts in the program.

```
%Macro Factor (Dataset,Var);
Proc Freq Data = &Dataset;
     Tables &Var;
Run;
Options NoLabel;
Proc Corr Data = &Dataset COV;
     Var &Var;
Run;
Options Label;

Proc Factor Data = &Dataset Scree
            Rotate = Promax;
     Var &Var;
Run;
Proc Factor Data = &Dataset Scree
            Rotate = Varimax;
     Var &Var;
Run;
%Mend;
%Macro Alpha (Dataset, Var);
Proc Corr Data = &Dataset Alpha;
     Var &Var;
Run;
%Mend;

%Macro FScore (Dataset,Var,Name);
   Proc Factor Data = &Dataset Method=Prin NFact=1 Out=&Dataset NoPrint;
        Var &Var;
   Data &Dataset;
        Set &Dataset;
             Rename Factor1 = &Name;
   Run;
%Mend;
```

```
%Factor(LongFormatLifeSatisfaction, bodyim1 bodyim2 bodyim3 bodyim4, BodyImage);
%Factor(LongFormatLifeSatisfaction, selfesteem1 selfesteem2 selfesteem3 selfesteem4,
SelfEsteem);

%Alpha(LongFormatLifeSatisfaction, bodyim1 bodyim2 bodyim3 bodyim4, BodyImage);
%Alpha(LongFormatLifeSatisfaction, selfesteem1 selfesteem2 selfesteem3 selfesteem4,
SelfEsteem);

%FScore(LongFormatLifeSatisfaction, bodyim1 bodyim2 bodyim3 bodyim4, BodyImage);
%FScore(LongFormatLifeSatisfaction, selfesteem1 selfesteem2 selfesteem3 selfesteem4,
SelfEsteem);
```

MODELS USED FOR ANLAYSIS

We asked students about their satisfaction in life for all the 3 waves. Since their responses were correlated, the assumption under OLS regression was violated. As such these correlations needed to be taken into account in modeling, otherwise the standard errors of the estimates would be underestimated for the between-subject and overestimated for the within-subject effects. Generalized estimating equations (GEE) were introduced by Liang and Zeger (1986) as an extension of generalized linear models (GLM) to analyze discrete and correlated data. Its strength is that it models a known function of the marginal expectation of the dependent variable as a linear function of explanatory variables. The advantage of GEE is especially obvious when the number of observations is large in relative to number of waves within subjects as in the case of our dataset. Although there are other statistical procedures discussed in the literature which also take into account of correlated responses such as Weighted Least Squares (WLS), we will not be discussing these techniques as they are beyond the scope of this paper except to highlight that compared to WLS, GEE allows us to use continuous variables in the model while

WLS does not. This flexibility makes GEE a better choice than WLS when modeling longitudinal data as WLS is restricted to categorical data in modeling. The following subsections further elaborates the GEE modeling.

Working Correlation Matrix

In GEE modeling, one has to specify the working correlation matrix in estimating the covariance of the parameter estimates. The specification of the working correlation matrix accounts for the form of within-subject correlation of responses on dependent variables. One of the aims of this paper is to find out whether using different working correlation matrices for estimation would affect the estimates and SEs substantially.

Let $\mathbf{R}_i(\boldsymbol{\alpha})$ be an n_i x n_i working correlation matrix that is fully specified by the parameter $\boldsymbol{\alpha}$, the covariance matrix of Y is modeled as follows:

$$\mathbf{V}_i = \varphi \mathbf{A}_i^{1/2} \mathbf{R}(\boldsymbol{\alpha}) \mathbf{A}_i^{1/2}$$

Where $\mathbf{A}_i$ an diagonal matrix $n_i \times n_i$ with diagonal matrix with $v(\mu_{ij})$ as the jth diagonal element

φ is a dispersion parameter, $\mathbf{R}$ is the working correlation

Four types of working correlation were examined in this study in order to measure the relationship between the student's life satisfaction over time. They are briefly summarized below and examples given are presented in matrix form for 4 waves. The exchangeable working correlation specification allows for constant correlations between any 2 measurements within a subject for all the time period. As such, only one parameter needs to be estimated. Autoregressive weights the correlation within two waves by their separated distance and hence correlation coefficients diminish for further distance. Similar to exchangeable R, it requires only one estimated parameter. M dependent working correlation assumes more than m lag with correlation zero and estimates the parameters within m waves. Unstructured R assumes different correlations between any two waves. The following illustrates the four wave structures of the four working correlations.

Working Correlation		Example – 4 Waves
Exchangeable	$Corr(y_{ij}, y_{i,j'}) = \begin{cases} 1 & j = j' \\ \rho & j \neq j' \end{cases}$	$\begin{pmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{pmatrix}$
Autoregressive	$Corr(y_{ij}, y_{i,j+s}) = \rho^s$	$\begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{pmatrix}$
M Dependent	$Corr(y_{ij}, y_{i,j+s}) = \begin{cases} 1 & s = 0 \\ \rho_s & s = 1,2,\ldots,m \\ 0 & s > m \end{cases}$	$\begin{pmatrix} 1 & \rho_1 & \rho_2 & 0 \\ \rho_1 & 1 & \rho_1 & \rho_2 \\ \rho_2 & \rho_1 & 1 & \rho_1 \\ 0 & \rho_2 & \rho_1 & 1 \end{pmatrix}$
Unstructured	$Corr(y_{ij}, y_{i,j'}) = \begin{cases} 1 & j = j' \\ \rho_{jk} & j \neq j' \end{cases}$	$\begin{pmatrix} 1 & \rho_{21} & \rho_{31} & \rho_{41} \\ \rho_{21} & 1 & \rho_{32} & \rho_{42} \\ \rho_{31} & \rho_{32} & 1 & \rho_{43} \\ \rho_{41} & \rho_{42} & \rho_{43} & 1 \end{pmatrix}$

Marginal R^2

Unlike cross-sectional regression, repeated measures are correlated over time, and hence they are not independent of each other. Since the residuals are not independent, the cross-sectional OLS R^2 could not be used directly in the context of GEE. Zheng (2000) introduced an extension of R^2 statistics for GEE models which named as marginal R^2 with the formula as shown below.

$$R^2_{\text{Marginal}} = 1 - \frac{\sum_{t=1}^{T}\sum_{i=1}^{n}\left(Y_{it} - \hat{Y}_{it}\right)^2}{\sum_{t=1}^{T}\sum_{i=1}^{n}\left(Y_{it} - \overline{Y}_{it}\right)^2}$$

Currently, general software like SAS PROC GENMOD does not provide the computation of Marginal R^2 for GEE. The macro %SelectGEE calculated this measure. The marginal R^2 could be interpreted as the portion of variance in the response variable explained by the fitted model (Hardin and Hilbem, 2003). Unlike the OLS R^2, this index could be negative as it depends on how good the prediction of the model, the numerator of the formula. The marginal R^2 would not expect to differ substantially with different working capital specifications as the variables used for the models are the same and the estimated coefficients are expected to be consistent.

QIC and QICU

One of the commonly used and well established goodness-of-fit statistics in comparing competitive models is the AIC (Akaike's Information Criterion). As GEE is not a likelihood-based method, Pan (2001) suggested using the QIC (Quasilikelihood under the Independence Model Criterion) which is analogous to the AIC in evaluating competitive models fit. The rationale of the QIC is similar to the AIC. The first term of QIC is the quasilikelihood computed using a specified working correlation which is analogous to the likelihood estimation counterpart of the AIC and similarly the second term is the penalty term which serves similar effect as the AIC second term as well (Hardin and Hilbe, 2003). Since trace($\hat{\Omega}_{\text{I}}$, $\hat{V}_r$) approximates p when GEE specification is correct, Pan (2001) also suggested QICU which could be used to approximate QIC and potentially useful in variable selection. However, he noted that QICU cannot be applied to select the working correlation matrix **R** as the presumption of QICU is that the specification of working correlation is correct. As such, QICU although is computed in the macro, it is not used for model comparison. The evaluation of choosing the best model of QIC and QICU is the one with the smallest value. Similar to the reason given to marginal R^2, we would not expect QIC to differ much for the various models.

$$QIC = -2Q\left(\hat{\beta}(R); I, D\right) + 2\,\text{trace}\left(\hat{\Omega}_{\text{I}}\hat{V}_r\right)$$

$$QICU = -2Q\left(\hat{\beta}(R); I, D\right) + 2p$$

$\hat{\Omega}_{\text{I}}$: Inverse of the variance matrix by fitting an independence model

$\hat{V}_r$: Modified sandwich estimate of variance from the model with the **R**

$-2Q\left(\hat{\beta}(R); I, D\right)$: Quasilikelihood computed using **R**

p : Dimension of β

$AIC = -2L + 2p$

L : log Likelihood

Since neither the QIC indices nor marginal R^2 measures were currently available in the SAS PROC GENMOD Version 9.1, this paper creates these measures in the macro output. As pointed by Pan (2001), as $\hat{\Omega}_I$ and $\hat{V}_r$ are found in the SAS PROC GENMOD, they can be output as matrices to compute QIC, as well as QICU.

SAS Macro

The macro %SelectGEE specified 6 parameters: &dataset to indicate the name of the input dataset; &cluster for the person ID; &workingmatrix as the SAS working matrix indicator; &Y for the dependent variable and &X for the list of explanatory variables and &num was the running number to specify the order of models. &num was also used to name the datasets created within the macro for the purpose of combining them into the final SAS output.

Three macro variables were created in the macro %SelectGEE for the computation of Marginal R^2. The overall mean was calculated and stored in the macro variable %Overallmean. Total sum of squares and error sum of square were named as macro variable %SST and %SSE respectively.

In sequential order, the macro %SelectGEE performed the following:

1) Calculate the overall mean using PROC MEANS and calculate total sum of squares
2) Fit the model with specified working correlation matrix using PROC GENMOD and output $\hat{V}_r$
3) Fit independent model and output $\hat{\Omega}_I$
4) Read $\hat{\Omega}_I$ and $\hat{V}_r$ into matrices using PROC IML to estimate QIC and QICU
5) Read predicted values output dataset from (2) to calculate Marginal R^2
6) Append results of QIC, QICU and Marginal R^2 of the various models into a dataset
7) Print the summarized results of three measures, and estimated coefficients, standard errors and p-values

SAS Output

The macro %SelectGEE generated the three measures in the following SAS output.

Working Correlation	Marginal R2	QIC	QICU
Exch	0.32473	5751.99	5750.44
Unst	0.32474	5752.01	5750.45
AR(1)	0.32485	5751.32	5750.05
MDep(1)	0.32487	5751.19	5750.01

The macro also produced the following summarized output (Table 1) which was exported from SAS output to EXCEL with minor amendments in wordings.

Table 1 Estimated Coefficients, Standard Errors and P-Values : GEE Models

	Exchangeable			AR(1)			M Dependent			Unstructured		
Parameter	Estimate	SE	P-Value	Estimate	SE	P-Value	Estimate	SE	P-Value	Estimate	SE	P-Value
Intercept	0.429	0.190	0.024	0.405	0.191	0.034	0.401	0.191	0.036	0.427	0.190	0.025
Male	-0.076	0.078	0.328	-0.069	0.078	0.379	-0.068	0.078	0.386	-0.075	0.078	0.337
Malay	-0.280	0.108	0.009	-0.277	0.108	0.011	-0.277	0.109	0.011	-0.278	0.108	0.010
Indian	-0.517	0.138	0.000	-0.532	0.139	0.000	-0.535	0.139	0.000	-0.521	0.138	0.000
Others	-0.101	0.231	0.662	-0.099	0.233	0.670	-0.099	0.233	0.671	-0.095	0.231	0.681
EM1	0.223	0.109	0.040	0.219	0.109	0.044	0.219	0.109	0.045	0.223	0.109	0.041
EM3	-0.096	0.202	0.635	-0.113	0.202	0.578	-0.115	0.203	0.571	-0.105	0.202	0.602
Gifted / Special	0.223	0.136	0.100	0.230	0.136	0.092	0.231	0.136	0.091	0.227	0.136	0.095
Normal Academic	-0.457	0.092	0.000	-0.463	0.093	0.000	-0.463	0.093	0.000	-0.459	0.092	0.000
Normal Technical	-0.722	0.173	0.000	-0.714	0.174	0.000	-0.713	0.174	0.000	-0.719	0.173	0.000
No of Siblings	-0.092	0.039	0.018	-0.094	0.039	0.016	-0.094	0.039	0.016	-0.093	0.039	0.016
Married	0.132	0.121	0.273	0.125	0.121	0.304	0.123	0.121	0.309	0.131	0.121	0.278
Social Support	0.292	0.042	0.000	0.293	0.042	0.000	0.293	0.042	0.000	0.292	0.042	0.000
Family Functioning	0.283	0.042	0.000	0.289	0.042	0.000	0.290	0.042	0.000	0.284	0.042	0.000
No of Closed Friends	0.001	0.011	0.954	0.002	0.011	0.837	0.003	0.011	0.817	0.001	0.011	0.926
Mental Morbidities	-0.291	0.041	0.000	-0.294	0.041	0.000	-0.295	0.041	0.000	-0.291	0.041	0.000
Anxious About Exam	0.128	0.020	0.000	0.127	0.020	0.000	0.127	0.020	0.000	0.127	0.020	0.000
Worry About Exam	0.050	0.020	0.013	0.054	0.020	0.008	0.054	0.020	0.007	0.050	0.020	0.012
Existential Satisfaction	0.564	0.045	0.000	0.563	0.046	0.000	0.563	0.046	0.000	0.563	0.045	0.000
Self Esteem	0.278	0.041	0.000	0.278	0.041	0.000	0.279	0.041	0.000	0.278	0.041	0.000
Body Image	0.282	0.041	0.000	0.286	0.041	0.000	0.287	0.041	0.000	0.283	0.040	0.000
R^2 / Marginal R^2	0.32473			0.32485			0.32487			0.32474		
QIC	5751.99			5751.32			5751.19			5752.01		

RESULTS

Table 1 summarized the GEE results of the four working correlation specification. The Marginal R^2 measure, led us to select the Mdep(1) correlation model (0.32487) while QIC led us to choose the AR(1) correlation model (5751.32). Although the two measures gave different selection results, these measures were close in their estimation. Hardin and Hilbe (2003) had highlighted that the result of QIC measure should not be blindly followed. Similar to interpretation of the OLS R^2, Marginal R^2 should also be a guide for comparison rather than strictly adhered to as a rule of decision for selecting competitive models.

The estimated coefficients of the four models were also very close. This reaffirmed the closeness of the measures for comparison and the consistently supported claim of Zeger and Liang (1986) that such results are expected when working correlations are misspecified. For instance, the estimated coefficients of the four models for self-esteem and existential satisfaction only differed at the third decimal digit which is trivial given these two variables metric ranges from 1 to 6. In general, no significant gender effect was found. The race Chinese was significantly higher in

life satisfaction than race Malay and Indian. Some of the prior and current stream effects were statistically significant. Students with more siblings expressed less satisfaction in life whereas parent's marital status was not significant after controlling other factors. However, it was noted that parent's marital status was statistically significant if other factors were not in consideration. As expected, the availability of social support, family functioning, existential satisfaction, self-esteem and body image all have positive effects on student life satisfaction whereas mental morbidities has negative effect. The number of closed friends was not significant after controlling for other factors, however it was significant if it was modeled under single regression. Although initially counterintuitive that anxiety about examination were positively related to life satisfaction, it is possible that in a highstakes examination environment like Singapore, those who express examination anxiety are those who are concerned about academics and academic performance and this could affect their life satisfaction. This argument is consistent with the findings that students in higher elementary and high school tracks report higher levels of life satisfaction than those in lower tracks.

CONCLUSION AND SUGGESTIONS

This paper reaffirmes the consistent estimate of GEE with the various working correlation matrix. Although the measures used in the paper did not show the same results in the model selection process, they nevertheless provided useful guidelines and supported empirically that the specification of different working correlation specification in the current study did not differ much in their interpretations.

Although the current macro %SelectFEE enables us to compare models with different working correlation specification, it can be easily modify to include comparing different complexity of models with the same working correlation matrix as well. The macro could also be modified and extended by including other distributions like gamma, multinomial, and negative binomial, just to name a few. While the current study's focus is on examining the population average of student's life satisfaction, other longitudinal modeling like random effect models which focus on subject specific could also apply the similar idea of using SAS macros to organize SAS program in an organized manner for efficient data analyses.

ACKNOWLEDGEMENTS

This paper acknowledges the Centre for Research in Pedagogy and Practice (CRPP), National Institute of Education, Singapore for making the data available from the research project "Life Pathway Project" (CRP 046DH). We would also like to thank Melvin Chan, Hock Huan Goh, Lazar Stankov and See Shing Yeung who took time to go through the paper and gave their comment.

REFERENCES

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

SAS Institute Inc. (2004). *SAS/STAT User's Guide Volume 3, Version 9.1*. Cary, NC: SAS Institute Inc.

Singer J. D. and Willet, J. B. (2003). *Applied Longitudinal Data Analysis, Modeling Change and Event Occurrence*. Oxford.

Stokes, M., Davis, C., and Koch, G. (2000). *Categorical Data Analysis Using the SAS System*, 2nd Edition. SAS Institute, Cary, N. C.

Zeger, S. L. and Liang, K. Y. (1986). The analysis of discrete and continuous longitudinal data. *Biometrics, 42*, 121-130.

Zheng, B. (2000). Summarizing the goodness of fit on generalized linear models for longitudinal data. *Statistics in Medicine, 19*, 1265-1275.

AUTHOR CONTACT INFORMATION

Teck Kiang Tan
Centre for Research in Pedagogy and Practice
Nanyang Technological University
1 Nanyang Walk, Block 5, Basement 3
Singapore 637616
Work Phone: 065-62196277
Email: teckkiang.tan@nie.edu.sg

Trivia Kang
Centre for Research in Pedagogy and Practice
Nanyang Technological University
1 Nanyang Walk, Block 5, Basement 3
Singapore 637616
Work Phone: 065-62196277
Email: trivia.kang@nie.edu.sg

David Hogan
Office of Education Research
Nanyang Technological University
1 Nanyang Walk, Block 5, Level 3
Singapore 637616
Work Phone: 065-62196277
Email: david.hogan@nie.edu.sg

APPENDIX

SAS Codes

```
* Macro SelectGEE;
* (1) Generate GEE output for Binary Dependent Model;
* (2) Summary Estimates to an output file for EXCEL output;
* (3) Calculate Marginal R2, QIC and QICU;

%Macro SelectGEE (Dataset, Cluster, WorkingMatrix, Y, X, Num);
Data A;
     Set &Dataset;
          Keep &Y &X &Cluster;
Run;
Proc Means Data = A NoPrint;
     Var &Y;
     Output Out = OverallMean Mean = OverallMean;
Run;
Data OverallMean;
     Set OverallMean;
          Call Symput('OverallMean', OverallMean);
Run;
Data TotalVar;
     Set A;
          Keep &Y;
Data TotalVar;
     Set TotalVar;
          Diff = &Y - &OverallMean;
          SST = Diff * Diff;
     Keep Diff SST;
Run;
Proc Means Data = TotalVar NoPrint;
     Var SST;
     Output Out = SST Sum = SST;
Run;
Data _Null_;
     Set SST;
          Call Symput('SST', SST);
Run;
* Fit Model with Working Correlation Specified;
ods trace on;
Proc GenMod Data = A Descending;
     Class &Cluster;
     Model &Y =  &X  /
     Link = Logit Dist = Binomial;
     Repeated Subject = &Cluster / Type = &WorkingMatrix Corrw ECovB;
          Output Out = PredictedGEE Predicted=Predicted;
     ods output GEERCov=VRModel GEEEmpPEst=GEEParameterEstimates;
Run;

* Fit Independent Model;
Proc GenMod Data = A Descending;
     Class &Cluster;
     Model &Y =  &X  /
     Link = Logit Dist = Binomial;
     Repeated Subject = &Cluster / Type = Ind Corrw ECovB;
     ods output GEERCov=OmegaIndependent;
Run;

Data GEEParameterEstimates;
     Set GEEParameterEstimates;
          Estimate&Num = Estimate;
          ProbZ&Num    = ProbZ;
          StdErr&Num   = StdErr;
          Keep Parm Estimate&Num ProbZ&Num StdErr&Num;
          If Parm = 'Scale' then delete;
Run;
%If &Num=1 %then %Do;
    Data GEEParameterEstimatesAll;
         Set GEEParameterEstimates;
%End; %Else;
%Do;
Data GEEParameterEstimatesAll;
     Merge GEEParameterEstimatesAll GEEParameterEstimates;
Run;
%End;

Data PredictedGEE;
     Set PredictedGEE;
```

```
          Keep &Y Predicted;
Data PredictedGEE;
     Set PredictedGEE;
          Diff = &Y - Predicted;
          SSE = Diff * Diff;
     Keep Diff SSE &Y Predicted;
Run;

Proc IML;
     Use OmegaIndependent; read all into OmegaIndependent;
     Use VRModel; read all into VRModel;
     Use PredictedGEE;
     read all var {&Y} into Y;
     read all var {Predicted} into Predicted;
     Q=Y#log(Predicted/(1-Predicted))+log(1-Predicted);
     QIC=-2*Q[+,1]+2*trace(inv(VRModel)*OmegaIndependent);
     QICU=-2*Q[+,1]+2*nrow(OmegaIndependent);
     create QIC from QIC;
     append from QIC;
     create QICU from QICU;
     append from QICU;
quit;
Data QIC;
     Set QIC;
     Rename Col1 = QIC;
Data QICU;
     Set QICU;
     Rename Col1 = QICU;
Data QICIndex;
     Merge QIC QICU;
Run;

Proc Means Data = PredictedGEE NoPrint;
     Var SSE;
     Output Out = SSE Sum = SSE;
Data _Null_;
     Set SSE;
          Call Symput('SSE', SSE);
Run;
Data SSE;
     Set SSE;
          Length WorkingCorrelation $10;
          Length XPrint $450;
          XPrint = compbl("&X");
          SST = &SST;
          SSE = &SSE;
          WorkingCorrelation = "&WorkingMatrix";
          Y   = "&Y";
       MarginalR2 = 1 - (SSE / SST);
          Drop _Type_ _Freq_;
Run;

Proc Append Base=BaseSSE Data = SSE;
Proc Append Base=BaseQICIndex Data = QICIndex;
Run;
Data All;
     Merge BaseSSE BaseQICIndex;
Proc Print Data = All;
Run;

%Mend;



* Model 1:Exchange;
%SelectGEE (LongFormatLifeSatisfaction, nric_1_s3, Exch, LifeSatsfactionBinary,
            Male Malay Indian Others EM1 EM3
            GiftSpec NAcad NTech NoSiblings Married
            SocialSupport FamilyFunction cfriend
            MentalMorbidities AnxiousExam WorryExam
            ExistentialSatisfaction SelfEsteem BodyImage,1);
* Model 2: AR(1);
%SelectGEE (LongFormatLifeSatisfaction, nric_1_s3, AR(1), LifeSatsfactionBinary,
            Male Malay Indian Others EM1 EM3
            GiftSpec NAcad NTech NoSiblings Married
            SocialSupport FamilyFunction cfriend
            MentalMorbidities AnxiousExam WorryExam
            ExistentialSatisfaction SelfEsteem BodyImage,2);
* Model 3: M Dependent;
%SelectGEE (LongFormatLifeSatisfaction, nric_1_s3, MDep(1), LifeSatsfactionBinary,
            Male Malay Indian Others EM1 EM3
```

```
            GiftSpec NAcad NTech NoSiblings Married
            SocialSupport FamilyFunction cfriend
            MentalMorbidities AnxiousExam WorryExam
            ExistentialSatisfaction SelfEsteem BodyImage,3);
* Model 4: Unstructured;
%SelectGEE (LongFormatLifeSatisfaction, nric_1_s3, Unstr, LifeSatsfactionBinary,
            Male Malay Indian Others EM1 EM3
            GiftSpec NAcad NTech NoSiblings Married
            SocialSupport FamilyFunction cfriend
            MentalMorbidities AnxiousExam WorryExam
            ExistentialSatisfaction SelfEsteem BodyImage,4);
Proc Sort Data = All;
     By MarginalR2;
Proc Print Data = All Noobs;
     Var WorkingCorrelation MarginalR2 QIC QICU;
Run;
Proc Print Data = GEEParameterEstimatesAll Noobs;
Run;
```