PROC GENMOD with GEE to Analyze Correlated Outcomes Data Using SAS

Tyler Smith, Department of Defense Center for Deployment Health Research,
Naval Health Research Center, San Diego, CA
Besa Smith, Department of Defense Center for Deployment Health Research,
Naval Health Research Center, San Diego, CA

ABSTRACT

Generalized linear models provide a framework for relating response and predictor variables by extending traditional linear model theory to nonlinear data. This is very important in many areas of epidemiologic research where outcomes are dichotomous or otherwise not normally distributed. To potentially complicate the statistical approach further, some studies, such as prospective cohort studies, follow individuals for a sufficient duration of time to observe multiple disease occurrences. If used as response variables, these disease outcomes are often correlated and should be treated with statistical consideration encompassing repeated measures applications in the regression analyses. Generalized Estimating Equations (GEEs) offer a way to analyze such data with reasonable statistical efficiency.

The GENMOD procedure in SAS® allows the extension of traditional linear model theory to generalized linear models by allowing the mean of a population to depend on a linear predictor through a nonlinear link function. In this paper we investigate a binary outcome modeling approach using PROC LOGISTIC and PROC GENMOD with the link function. Further, we investigate the Generalized Estimating Equation (GEE) capabilities of PROC GENMOD for correlated outcome data to fit models using different correlation structures.

INTRODUCTION

Although often more costly and time consuming to conduct prospective studies, analyses utilizing such a design have gained much attention and grown in use over the past several decades. This is largely due to intense pressure on the research community to better define causal relationships which are demonstrated by several criterion including establishing temporal sequence of exposure and disease. However, when taking repeated measurements on individuals, it is often found that those with an outcome of interest at time point A will have a higher probability of having a like outcome at time point B. The complexity of investigating correlated outcomes in order to make sound statistical inferences becomes a burden that some researchers choose to avoid or ignore. In this paper we will discuss the ease of analyzing correlated dichotomous outcome data using Generalized Linear Models (GLMs) and Generalized Estimating Equations (GEEs).

Longitudinal studies are defined as studies in which the outcome variable is repeatedly measured on two or more occasions over time. However, the models and methods are more broadly applicable to other repeated measure type data. In this paper, we will loosely use longitudinal data to imply those data that are taken repeatedly over time as well as those not taken over time but have otherwise correlated outcome data.

CORRELATED OUTCOMES

Correlated outcomes are collected in many areas of research and occur for a variety of different reasons. Valid scientific inferences are reliant upon properly accounting for the correlation among outcomes within subjects. This type of within subject correlation may be due to a single outcome measured repeatedly over time on the same subject, as in longitudinal studies; or may be due to multiple outcomes measured one or more times each on the same subject, as in clinical trials involving multiple investigative endpoints. Correlation may also be due to a membership relationship among units (families or litters). Examples of studies include disease outcomes in

longitudinal studies of moderate duration resulting in multiple incident cases, twin or sibling studies, self comparison studies, and paired studies.

LINEAR MODELS

Predicting the value for variable Y by changing the values for variable X is simple and intuitive and employed in our every day lives. Linear models embrace the prediction potential of Y by X but also places the distinction that the relationship is in fact linear in nature. We further our inferential potential by suggesting that we may describe the relationship between variables Y and X in a linear fashion and that we may add variables to enhance predictive capabilities, or control for confounding or inadequacies in our data. The distilling of complex information stemming from multiple variables into scientific inferences has become mathematically and practically possible, however interpretation of the often complex results often remain difficult and elusive.

The typical multiple regression model takes the following form where the variable Y is a vector of observations, the X variables are linearly associated covariates, the β's are regression coefficients, and *e* reflects the error variability that cannot be accounted for by the predictors.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + e$$

We assume that the y_i's are normal and independent with standard deviation σ, such that we estimate the vector of β's by minimizing the sum of squares of the differences between observed and model-predicted outcomes.

A researcher may feel that a person's weight can be reasonably predicted by the person's height, gender, how much fast food is eaten in a typical week, and how many hours of exercise the person is able to achieve each week. Using the above linear regression to estimate the respective regression coefficients from a sample of data including height, gender, fast food consumption and exercise yields the regression equation below.

$$\text{Weight} = \beta_0 + \beta_1(\text{height}) + \beta_2(\text{gender}) + \beta_3(\text{fastfood}) + \beta_4(\text{exercise}) + e$$

GENERALIZED LINEAR MODELS (GLMS)

What if the response and predictor variables are not related linearly? Generalized linear models (GLMs) are a generalization of the general linear model described above. GLMs are flexible, providing the researcher tools to work in a wide range of common situations while at the same time allowing most of the familiar ideas of normal linear regression to carry over. If the distribution of the dependent or response variable is not normal, GLMs which assume a *link linear* relationship based on a chosen link function, may be utilized to complete the analysis.

GLMs take on the following standard model form:

$$Y = g(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + e)$$

Where, as in the linear model, the variable Y is a vector of observations, the X variables are linearly associated covariates, the β's are regression coefficients, and *e* reflects the error variability that cannot be accounted for by the predictors. However, there is now an additional function *g(…)* included in the equation. The function *g(…)* is some known monotonic function which acts on E(y) relating the means of the responses to the linear predictors. The estimation of the nonlinear regression equation is then completed by weighting the observations inversely according to the variance functions. This weighting procedure is equivalent to maximum likelihood (ML) estimation when the observations come from an exponential family distribution. Therefore, in GLMs, the values of the β coefficients are obtained by ML estimation requiring iterative computational procedures. There are many iterative methods for ML

estimation in the GLM, of which the Newton-Raphson and Fisher-Scoring methods are among the most efficient and widely used.

LINK FUNCTIONS

The inverse function of *g(…)*, is called the link function. In the general linear model the dependent variable values are expected to follow the normal distribution, the link function is a simple identity function (the linear combination of values for the predictor variables, not transformed). Various link functions (McCullagh and Nelder, 1989) can be chosen, depending on the assumed distribution of the *y* variable (binomial, multinomial, etc):

Traditional Linear Model:
- Continuous response variable
- Normal distribution
- Link function: identity

$$g(\mu) = \mu$$

Logistic Regression
- Proportional response variable
- Binomial distribution
- Link function: logit

$$g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$$

Poisson Regression in Log Linear Model
- Counting response variable
- Poisson distribution
- Link function: log

$$g(\mu) = \log(\mu)$$

Gamma Model with Log Link
- Positive and continuous response variable
- Gamma distribution
- Link function: log

$$g(\mu) = \log(\mu)$$

GENERALIZED ESTIMATING EQUATIONS (GEEs)

Statistical methods for extending linear model theory to repeated measurements of multivariate, normally distributed responses, have been established. These methods may be accomplished using the GLM or MIXED procedures in SAS. The Generalized Estimating Equations (GEEs) approach introduced by Liang and Zeger (1986), is another method for analyzing correlated outcome data, when those data could have been modeled using GLMs if there were no correlated outcomes. This approach for handling *continuous or discrete* responses provides a non-likelihood based or quasi-likelihood approach for modeling correlated responses. In other words, by specifying one of a variety of possible working correlation matrix structures to account for the within-subject correlations, the GEE

method estimates model parameters by iteratively solving a system of equations based on quasi-likelihood distributional assumptions.

The most commonly used within-subject correlation matrices are as follows:

- Independence, repeated observations are uncorrelated
- Unspecified (unstructured), correlations within any two responses are unknown and need to be estimated
- Exchangeable, correlation between any two responses of the i^{th} individual is the same
- Autoregressive of first order [AR(1)], assuming the interval length is the same between any two observations

The researcher may then choose from a variety of model forms by specifying a link function for logistic, log-linear, or linear.

Although this paper will focus on nonlinear discrete correlated outcomes, it should be mentioned that the following types of models may be fit for continuous correlated outcome data as well.

The "marginal model" (also known as population averaged models) is used when the researcher is investigating the population and wishes to model the population averaged response as a function of the covariates. The regression of the response on explanatory variables is modeled separately from within-person correlation. We model the mean of the average response over the sub-population that shares a common value of X and interpret for the population and not the individual.

The "transitional model" (also known as an autoregressive model) is used when the analysis must account for a time dependency. The correlation among the current outcomes exists because the past outcomes explicitly influence the present observation. The past outcomes are treated as additional predictor variables.

The "random effects model" (also known as the mixed effects model) is used when the analysis must account for both fixed and random effects in the model. This occurs when data for a subject are independent observations following a linear model or GLM, but the regression coefficients vary from person to person. Infant growth is a good example where the coefficients represent birth weight and growth rate. Because children are born at different weights and have different growth rates based on genetic and environmental factors, we need to solve for the variability reflecting natural heterogeneity due to unmeasured factors.

NONLINEAR MODELS FOR DISCRETE CORRELATED OUTCOME DATA

The marginal, random effects, and transitional models mentioned previously in linear models may be employed for discrete outcomes. The most commonly used approaches for marginal models when discrete outcomes occur are logistic regression models for dichotomous and poylchotomous outcomes, and Poisson regression models for counts. For the marginal model, regression coefficients have population-averaged interpretation.

ANALYTIC APPROACH

Descriptive statistics using PROC FREQ for categorical variables or PROC UNIVARIATE for continuous variables should be completed to determine possible significant explanatory variables to be included in the model runs. Multicollinearity among potential explanatory variables should be investigated using PROC REG's diagnostic capabilities to ensure no model-burdening correlations exist between variables. After investigation for confounding of variables not independently associated with the outcome, variables with p-values of 0.15 or less are retained in the final model analysis.

The steps involved in the longitudinal model data analysis:

1) Choose the model by specifying the link function, which describes the model form that you wish to use

2) Choose the variance-covariance structure (specifying the working correlation structure for each subject)

3) Choose the distribution of the dependent variable

4) Assess the goodness-of-fit of the model and the variance covariance structure

DATA

In this paper we will focus on a dichotomous outcome variable and use a logit analysis or logistic regression to analyze the binary outcome while controlling for possible confounding. We will apply all analyses to a set of data consisting of 45 observations and 6 variables. The outcome of interest will be myocardial infarction within a window after vaccination with the anthrax vaccine. The data look like this:

input id gender race maritalstat mi anthrax;

```
14 1 1 1 1 1
14 2 1 1 0 1
15 1 1 0 0 1
16 2 3 1 1 0
17 1 2 1 0 0
18 1 1 0 0 0
19 2 2 0 0 0
20 2 3 1 0 1
…
…
…
45 2 3 1 0 1
```

PROC LOGISTIC

Logistic regression is a statistical method used to evaluate many independent variables (X_1, X_2, …, X_p) in order to predict a dichotomous outcome. Generally this outcome is denoted as Y = 1 or Y = 0 for the two possibilities.

In logistic regression the probability of an occurrence of the outcome being investigated is defined as:

$$P(Y=1) = \frac{1}{1+\exp\left[-\beta_0 + \left(\sum_{k=1}^{p} \beta_k X_k\right)\right]}$$

SAS offers several procedures to estimate the binary logit model using ML estimation which include PROC LOGISTIC, PROC GENMOD, PROC PROBIT, and PROC CATMOD. In this paper we will focus on the

comparison of PROC LOGISTIC and PROC GENMOD. PROC LOGISTIC is a procedure for fitting linear regression models for binary or ordinal outcomes. The following is sample code for this procedure:

```
ods html path = 'c:\YourPath' body='Name.html';
proc logistic data=MIdata descending;
    class anthrax (ref = '0') race (ref='1') maritalstat (ref='0')    gender (ref='1')  / param=reference;
    model mi=anthrax race maritalstat gender / clodds=wald  clodds=pl lackfit;
     title1 'Proc Logistic for Log Odds of MI Among those Being Vaccinated Against Anthrax After Controlling
         for Race, Marital Status and Gender';
run;
ods html close;
```

ods html with the close after the run will send your output to an HTML file.

Data=MIdata names the input data set for the logistic regression.

Descending: The default in SAS is to model the probability that the dependent variable outcome of MI is equal to 0. The descending option allows us to model the probability that MI is equal to 1 and compares the probability of outcome to probability of no outcome for the odds ratio.

Class statement allows us to establish the reference category in the categorical variables without first making "dummy" variables in a data step. In this case, we are using reference cell coding.

Param=reference requests that the parameter estimates, odds ratios, and confidence intervals be calculated using reference cell coding. The default parameter estimates would be computed using the effect coding scheme which estimates the difference in the effect of each non-reference level compared to the average effect over the other levels of the variable.

Clodds= requests for each explanatory variable, the 95% (the default alpha level because the ALPHA= option is not invoked) Wald or profile likelihood confidence intervals for the odds ratios.

Lackfit requests the Hosmer-Lemeshow goodness of fit test for the model. The null hypothesis is that there is a good fit of the model to the observed data across the risk groups (we wish to fail to reject the null).

PROC LOGISTIC output:

Note for later GENMOD:

Class Level Information

Class	Value	Design Variables	
anthrax	0	-1	
	1	1	
race	1	-1	-1
	2	1	0
	3	0	1

Class Level Information			
Class	Value	Design Variables	
maritalstat	0	-1	
	1	1	
gender	1	-1	
	2	1	

Output not shown: The AIC (Akaikes information criterion, lower is generally better), SC (Schwartz criterion which penalizes for more parameters then the AIC, lower is generally better), and the -2 log likelihood for the model fit statistics; the likelihood ratio, score, and Wald tests for testing whether all of the parameters taken together in the fitted model are equal to 0 when compared to the model with only the intercept; significance of each variable in it's entirety (not categories of the variable) as well as the different categories.

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
8.1776	7	0.3172

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
anthrax 1 vs 0	1.851	0.403	8.513
race 2 vs 1	2.736	0.414	18.104
race 3 vs 1	7.442	1.054	52.518
maritalstat 1 vs 0	1.526	0.312	7.456
gender 2 vs 1	1.053	0.239	4.653

Interpretation: After controlling for race, marital status and gender, those who were vaccinated against anthrax were 1.85 times more likely to have an outcome of MI when compared to those not vaccinated. This finding was not statistically significant at the alpha=0.05 level (95% CI = 0.40, 8.51). Based on the Hosmer-Lemeshow, there is a good fit of the model to the observed data across the risk groups.

PROC GENMOD

The following is sample code for the GENMOD procedure to attempt to replicate what was done with the LOGISTIC procedure:

```
proc genmod data=MIdata descending;
  class anthrax race maritalstat gender;
  model mi= anthrax race maritalstat gender / dist=binomial link=logit;
  estimate "Anthrax"      anthrax -1 1 / exp;
  estimate "Sex"           gender  -1 1 / exp;
  estimate "Black"         race -1 1 0 / exp;
  estimate "Hispan"        race -1 0 1 / exp;
  estimate "maritalstat"   maritalstat -1 1 / exp;
run;
quit;
```

Descending A very important point since version 8.1 came out is that when fitting a logistic regression using PROC GENMOD, the default now models the probability that the dependent variable MI is equal to 0. Versions prior to 8.1 modeled the higher level of the binary outcome variable (i.e. disease is present).

Class statement in GENMOD is the same as with PROC GLM and PROC ANOVA for determining which variables in the model will define categorical (classification) levels. These should be variables which code for terms such as replication id (for later GEE), exposure level, etc. They can be character or numeric in value.

Dist=binomial option identifies the appropriate distribution for the data, in this case binomial. Other potential choices include Gaussian, Poisson, normal, gamma, inverse Gaussian, negative binomial (negbin), and multinomial (mult). If the DIST = option is omitted, SAS will assume the Gaussian distribution.

Link=logit option refers to a transformation which is carried out on the responses prior to analysis, in this case the logit. Other potential choices include identity, log, power, probit, and complementary log log links. When this option is omitted, SAS will assume the identity link function resulting in no transformation.

Estimate will produce the estimated odds ratio for the exposure effect along with its associated standard error and confidence limits. The syntax for the ESTIMATE statement is exactly the same as that for the CONTRAST statement although the CONTRAST statement tests whether a linear combination of means is significantly different from 0. It should be mentioned that including the statement “lrci” and “waldci” after the link=logit will produce wald and likelihood ratio confidence intervals about the parameter estimates.

The word between the quotes will label the output, and the variable name that comes after the label in quotes will call on the variable you wish to investigate. The contrasts in the input statement (–1 1 for anthrax) are the same as the column of the class level information output from the logistic regression above. Therefore, the output odds ratios will reflect the same comparisons as what was seen in PROC LOGISTIC.

Exp after the backslash requests that the parameter estimates, standard error, and the confidence limits be computed and output.

PROC GENMOD output:

Model Information	
Data Set	WORK.MIDATA
Distri bution	Binomial
Link Function	Logit

Model Information	
Dependent Variable	mi
Observations Used	44

Additionally, the output shows the class level information, the response profile, the parameter information, criteria for assessing goodness of fit (including deviance, and log likelihood), and analysis of parameter estimates.

Contrast Estimate Results							
Label	Estimate	Standard Error	Alpha	Confidence Limits		Chi-Square	Pr > ChiSq
Anthrax	0.6159	0.7785	0.05	-0.9098	2.1417	0.63	0.4288
Exp (Anthrax)	1.8514	1.4413	0.05	0.4026	8.5141		
Sex	0.0521	0.7579	0.05	-1.4333	1.5375	0.00	0.9452
Exp(Sex)	1.0535	0.7984	0.05	0.2385	4.6530		
Black	1.0069	0.9641	0.05	-0.8827	2.8965	1.09	0.2963
Exp(Black)	2.7371	2.6388	0.05	0.4137	18.1107		
Hispan	2.0073	0.9971	0.05	0.0531	3.9615	4.05	0.0441
Exp (Hispan)	7.4434	7.4214	0.05	1.0546	52.5368		
maritalstat	0.4229	0.8093	0.05	-1.1633	2.0091	0.27	0.6013
Exp (maritalstat)	1.5264	1.2353	0.05	0.3124	7.4568		

From this output, it is relieving to see that the odds ratios are consistent with the odds ratios produced by PROC LOGISTIC. Notice also that the confidence limits are consistent to the third digit.

Unlike the LOGISTIC procedure, the GENMOD procedure will not give the global test of the null hypothesis that all of the parameters taken together in the fitted model are equal to 0 when compared to the model with only the intercept. To calculate the likelihood ratio chi-square test, take the deviance (in output) from the reduced model (or null model if you remove all variables) and minus the deviance in the full model. This will give you a chi-square statistic with the degrees of freedom equal to the number of variables removed. PROC GENMOD does include an LSMEANS statement that provides an extension of least squares means to the generalized linear model.

LOGISTIC REGRESSION OF CORRELATED OUTCOME DATA USING PROC GENMOD

The following is sample code to use the GENMOD procedure to model a logistic regression with correlated outcome data:

```
proc genmod data=MIdata descending;
  class id anthrax race maritalstat gender ;
  model mi= anthrax race maritalstat gender / dist=binomial link=logit;
  repeated subject=id / type=cs corrw covb;
```

```
  estimate "Anthrax"      anthrax -1 1 / exp;
  estimate "Sex"           gender  -1 1 / exp;
  estimate "Black"         race -1 1 0 / exp;
  estimate "Hispan"        race -1 0 1 / exp;
  estimate "maritalstat"   maritalstat -1 1 / exp;
run;
quit;
```

Note that the only difference between this PROC GENMOD run and the last run is the repeated line and the variable id in the class statement. The contrast, lsmeans, and estimate statements can be used for the GEE parameter estimates as well.

Repeated statement invokes the GEE method. This line is where we specify the covariance structure of multivariate responses for GEE model fitting, the iterative fitting algorithm used in GEEs, and the optional output.

Subject=id specifies that individual subjects are identified by the variable id. The variable id must also be listed in the class statement. Each distinct value, or level, of the effect identifies a different subject, or cluster. Responses from different subjects are assumed to be statistically independent, and responses within subjects are assumed to be correlated.

Type=cs specifies the structure of the working correlation matrix used to model the correlation of the responses from subjects. The default working correlation type is the independent. Possibilities of type= include autoregressive (AR), exchangeable (EXCH or CS), independent (IND), m dependent (MDEP), unstructured (UN or UNSTR), and user specified correlation matrix (USER or FIXED).

Instead of the Type= statement in this case, if the response is of the single variable type, the distribution is binomial, and the data are binary, alternating least squares can be employed by using the logor option in the repeated statement. This specifies the regression structure of the log odds ratio used to model the association of the responses from subjects for binary data instead of using a working correlation. Possibilities of logor= include exchangeable, fully parameterized clusters, and nested, among others.

Corrw displays the estimated working correlation matrix.

Covb displays the estimated regression parameter covariance matrix. Both model-based and empirical covariances are displayed.

GEE using PROC GENMOD output:

The initial parameter estimates for iterative fitting of the GEE model are computed as in the ordinary generalized linear model. Statistics for the initial model fit such as parameter estimates, standard errors, deviances, and Pearson chi-squares do not apply to the GEE model. Because the GEE model is not estimated by ML, tests such as the likelihood ratio test, AIC, and BIC are not appropriate to use.

GEE model fit information:

GEE Model Information	
Correlation Structure	Exchangeable
Subject Effect	id (27 levels)
Number of Clusters	27
Correlation Matrix Dimension	6

GEE Model Information	
Maximum Cluster Size	6
Minimum Cluster Size	1

The parameter estimated covariance matrices from the COVB option (model-based and empirical):

Covariance Matrix (Model-Based)						
	Prm1	Prm2	Prm4	Prm5	Prm7	Prm9
Prm1	0.72563	-0.21743	-0.49657	-0.41742	-0.10522	-0.24441
Prm2	-0.21743	0.56992	0.05754	-0.09958	-0.15104	-0.01420
Prm4	-0.49657	0.05754	1.07454	0.43673	-0.03347	0.08451
Prm5	-0.41742	-0.09958	0.43673	0.78048	0.001494	0.06920
Prm7	-0.10522	-0.15104	-0.03347	0.001494	0.67126	-0.16292
Prm9	-0.24441	-0.01420	0.08451	0.06920	-0.16292	0.55586

Covariance Matrix (Empirical)						
	Prm1	Prm2	Prm4	Prm5	Prm7	Prm9
Prm1	0.73761	-0.11629	-0.48814	-0.39694	-0.09814	-0.26474
Prm2	-0.11629	0.55015	-0.06621	-0.22658	-0.14959	-0.15470
Prm4	-0.48814	-0.06621	0.91789	0.49595	-0.12467	0.14544
Prm5	-0.39694	-0.22658	0.49595	0.85738	0.002843	0.08172
Prm7	-0.09814	-0.14959	-0.12467	0.002843	0.57484	-0.11494
Prm9	-0.26474	-0.15470	0.14544	0.08172	-0.11494	0.68181

The exchangeable working correlation matrix from the CORRW option:

Working Correlation Matrix						
	Col1	Col2	Col3	Col4	Col5	Col6
Row1	1.0000	0.1750	0.1750	0.1750	0.1750	0.1750
Row2	0.1750	1.0000	0.1750	0.1750	0.1750	0.1750
Row3	0.1750	0.1750	1.0000	0.1750	0.1750	0.1750
Row4	0.1750	0.1750	0.1750	1.0000	0.1750	0.1750

Working Correlation Matrix						
	Col1	Col2	Col3	Col4	Col5	Col6
Row5	0.1750	0.1750	0.1750	0.1750	1.0000	0.1750
Row6	0.1750	0.1750	0.1750	0.1750	0.1750	1.0000

The following output contains the parameter estimates, empirical standard error estimates, confidence intervals, z-scores and p-values. Model based standard errors can be requested by using the "modelse" option in the repeated statement.

Analysis Of GEE Parameter Estimates							
Empirical Standard Error Estimates							
Parameter		Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept		0.7144	0.8588	-0.9689	2.3977	0.83	0.4055
anthrax	0	-0.5529	0.7417	-2.0066	0.9009	-0.75	0.4560
anthrax	1	0.0000	0.0000	0.0000	0.0000	.	.
race	1	-2.0251	0.9581	-3.9029	-0.1473	-2.11	0.0345
race	2	-0.9398	0.9259	-2.7546	0.8750	-1.01	0.3101
race	3	0.0000	0.0000	0.0000	0.0000	.	.
maritalstat	0	-0.4925	0.7582	-1.9785	0.9935	-0.65	0.5160
maritalstat	1	0.0000	0.0000	0.0000	0.0000	.	.
gender	1	-0.0574	0.8257	-1.6758	1.5610	-0.07	0.9446
gender	2	0.0000	0.0000	0.0000	0.0000	.	.

The final table in the output reflects the adjusted odds ratios and the 95% confidence intervals for the contrasts chosen in the estimate statements. Notice the difference in the odds ratios using GEE and those reported in GLM and PROC LOGISTIC.

Contrast Estimate Results							
Label	Estimate	Standard Error	Alpha	Confidence Limits		Chi-Square	Pr > ChiSq
Anthrax	0.5529	0.7417	0.05	-0.9009	2.0066	0.56	0.4560
Exp (Anthrax)	1.7383	1.2893	0.05	0.4062	7.4383		
Sex	0.0574	0.8257	0.05	-1.5610	1.6758	0.00	0.9446
Exp(Sex)	1.0591	0.8745	0.05	0.2099	5.3430		

Contrast Estimate Results							
Label	Estimate	Standard Error	Alpha	Confidence Limits		Chi-Square	Pr > ChiSq
Black	1.0853	0.8851	0.05	-0.6494	2.8201	1.50	0.2201
Exp(Black)	2.9604	2.6202	0.05	0.5223	16.7778		
Hispan	2.0251	0.9581	0.05	0.1473	3.9029	4.47	0.0345
Exp(Hispan)	7.5769	7.2592	0.05	1.1587	49.5450		
maritalstat	0.4925	0.7582	0.05	-0.9935	1.9785	0.42	0.5160
Exp (maritalstat)	1.6364	1.2407	0.05	0.3703	7.2318		

The interpretation of the parameters in the marginal (population averaged) and random (mixed) effects model is analogous to the standard logistic regression model, however there are differences (as noted above) in how we adjust for the correlations. Therefore the sentence would be the typical sentence describing strength, direction, and p-value/confidence limit of the association. In this case, after adjusting for correlated outcome data and controlling for race, marital status and gender, those who were vaccinated against anthrax were 1.74 times more likely to have an outcome of MI when compared to those not vaccinated. This finding was not statistically significant at the alpha=0.05 level (95% CI = 0.41, 7.44).

SUMMARY

The use of PROC LOGISTIC has been well established in the research community for conducting regression of dichotomous or polychotomous endpoints. However, in longitudinal analyses, outcome data may be correlated due to repeated measures on the same subject or comparisons of twins or paired data. The collection of tools offered by SAS allow for the flexibility of analyzing many different data structures that present in longitudinal data analyses. PROC GENMOD is a powerful tool to conduct Generalized Linear Model regressions as well as the extension to General Estimating Equations where correlated outcome data must be taken into account. Future research should be focused on model selection, model diagnostics, and goodness-of-fit statistics and their implication into the SAS system.

REFERENCES

McCullagh P and Nelder J (1989). Generalized Linear Models. Chapman and Hall, London, 2nd edition.

Diggle PJ, Liang KY, Seger SL (1995). Analysis of Longitudinal Data. Oxford University Press, New York.

Zeger SL, Liang KY. Longitudinal data analysis using generalized linear models. *Biometrica.* 1986;73:13-22.

Zeger SL, Liang KY. An overview of methods for the analysis of longitudinal data. *Statistics in Medicine.* Oct-Nov 1992;11(14-15):1825-1839.

SAS Institute Inc., SAS/STAT® *User's Guide, Version 6, Fourth Edition, Volume 1,* Cary, NC: SAS Institute Inc., 1989. 943 pp.

SAS Institute Inc., SAS/STAT® *User's Guide, Version 6, Fourth Edition, Volume 2,* Cary, NC: SAS Institute Inc., 1989. 846 pp.

SAS Institute Inc. SAS/STAT® *Software: Changes and Enhancements through Release 6.11*. Cary, NC: SAS Institute Inc., 1996. 1104 pp.

ACKNOWLEDGMENTS

The authors would like to thank CDR Margaret AK Ryan, Director of the Department of Defense Center for Deployment Health Research at the Naval Health Research Center, San Diego.

Approved for public release: distribution is unlimited.

This research was supported by the Department of Defense, Health Affairs, under work unit no. 60002.

ABOUT THE AUTHORS AND CONTACT INFORMATION

Tyler has used SAS for 14 years including over 10 years as a senior statistician with the DoD Center for Deployment Health Research. Responsibilities include management, analysis, and interpretation of large demographic, health care, and longitudinal data on US military personnel. This work has culminated in many peer-reviewed journal manuscripts in scientific journals as well as multiple SAS user group or technical related publications. Invitations to speak include the International Biometrics Society, the local San Diego SAS group, WUSS, and SUGI conferences.

Tyler C. Smith, MS
Statistician
Department of Defense Center for Deployment Health Research, at the Naval Health Research Center, San Diego
smith@nhrc.navy.mil

Besa has been using SAS for 10 years including her work currently as a senior biostatistician with the DoD Center for Deployment Health Research at the Naval Health Research Center, San Diego. Her responsibilities include management of large military and demographic data sets, mathematical modeling and statistical analysis for longitudinal and health-based studies. She has presented at previous WUSS and local San Diego SAS user group meetings.

Besa Smith, MPH
Biostatistician, Henry Jackson Foundation
Department of Defense Center for Deployment Health Research, at the Naval Health Research Center, San Diego
besa@nhrc.navy.mil