Abstract
Logistic regression analysis can be used to determine whether demographic or other factors are associated with the occurrence of adverse events (AE) in clinical trials. In this paper, we begin by introducing the SAS procedures GENMOD and LOGISTIC for performing logistic regression analyses. We then present a SAS macro which generates multivariate and univariate models dynamically for adverse events satisfying prespecified conditions. In these models, the outcome variable is defined as whether the subject had an AE of interest, and the covariates include age, sex, race, and weight. Both univariate and multivariate models were fit for each AE under consideration. Scenarios are presented if an AE did not occur in the reference category for a categorical covariate. If any given AE did not occur in the reference category for a given categorical covariate, the univariate model was not run for that covariate and the multivariate model did not include that covariate. This is because the odds ratio will become inestimable for that covariate. In multivariate analysis, a single model is fit for each AE that includes all the covariates (unless a categorical covariate needs to be excluded as described above). The estimate and the corresponding p-values for continuous variables and for individual levels of the categorical covariates are stored in the ODS dataset created by GENMOD procedure. The odds ratio can be computed as exp(ESTIMATE).

Introduction
Regression methods are being used in data analysis concerned with describing the relationship between a response variable and one or more explanatory variables. Binary responses (i.e., success and failure), ordinal responses (i.e., normal, mild, and severe), and nominal responses (i.e., gender, age group, side effects associated with the use of therapeutic agents) arise in many fields of study. Logistic regression analysis is often used to investigate the relationship between these discrete responses and a set of explanatory variables. Logistic regression involves producing model coefficients with significance tests.

Important assumptions that need to be kept in mind before choosing logistic regression are:

- the response of a "success" occurs with a constant probability \( p \) \( (0 < p < 1) \) for every trial in each group
- the responses are independent across the trials, i.e., a "success" on any given trial should not influence a "success" on any another trial
Response variables can be interpreted using odds ratio (number of "success" responses to number of "failures" over the independent trials).

Logistic regression analysis in SAS can be done using PROC LOGISTIC as well as PROC GENMOD. A brief description of LOGISTIC and GENMOD procedures and comparison between the two are discussed in this paper. Other procedures available in SAS for performing logistic regression analysis include PROC NLMIXED, CATMOD, SURVEYLOGISTIC.

**PROC GENMOD**

The GENMOD procedure fits generalized linear models. Its function and application is similar to PROC GLM (general linear model) for ANOVA models. However, GENMOD allows the response variable to assume distributions such as binomial (for dichotomous data) and Poisson (for counts).

Using GENMOD procedure, you can construct a generalized linear model by deciding on response and explanatory variables for your data and choosing an appropriate link function and response probability distribution. Some examples of generalized linear models follow. Explanatory variables can be any combination of continuous variables, classification variables, and interactions.

Syntax for logistic regression using PROC GENMOD is:

```
PROC GENMOD DATA=ae descending;
CLASS gender;
MODEL response = gender / dist=binomial link=logit / <options> ;
RUN;
```

**PROC LOGISTIC**

The LOGISTIC procedure can be used to fit linear logistic regression models for discrete response data. The model estimation is achieved by using the method of maximum likelihood. It can also perform conditional logistic regression for binary response data and exact conditional logistic regression for binary and nominal response data. The Fisher-scoring algorithm or the Newton-Raphson algorithm is used to carry out maximum likelihood estimation, and starting values for the parameter estimates can be specified. The logit link function in the logistic regression models can be replaced by the probit function; the complementary log-log function; or the generalized logit function.

The LOGISTIC procedure enables one to specify categorical variables and/or continuous variables as independent variables.

Syntax for logistic regression using PROC LOGISTIC is:

```
PROC LOGISTIC DATA=ae descending;
CLASS gender/PARAM=ref;
MODEL response = gender / <options>;
RUN;
```

PROC GENMOD presents a unified approach to the analysis of categorical data including Poisson and Negative Binomial (for counts), gamma, and normally distributed
data (though for this distribution, GLM, REG, or MIXED will likely work better). It also handles repeated measures for count data in much the same way as PROC MIXED works with repeated measures for continuous data.

PROC LOGISTIC is designed for regression applications with one response (0/1) collected from each subject or several independent responses aggregated over subjects; it is the procedure designed to compute ROC curves. It can also perform exact logistic regression when you have small sample sizes or 0 counts in some of the cells (a technique that may be of value when given the warning "quasi-complete separation" in the log file).

**Data Overview**

The adverse event dataset used for analysis was coded using The Medical Dictionary for Regulatory Activities (MedDRA) terminologies. The Dataset complied with CDISC SDTM standards. The dataset was vertical in structure (i.e. one observation per subject per adverse event experienced). Two categorical variables: race, sex; and two continuous variables: age, weight were the independent variables used for analysis. The response variable was whether the subject had an adverse event or not.

**SAS Programming Overview**

The sections below will discuss the steps involved in the proposed macro from SAS programming perspective in terms of processing considerations, creation of analysis datasets, and statistical analysis.

**Processing Considerations**

Multivariate and univariate models are generated dynamically in such a way that the models converge. Convergence of the models defined in GENMOD procedure depends on data obtained from clinical trials. The models do not converge in some scenarios and the odds ratio becomes inestimable making it difficult to conclude the association of covariates to the occurrence of adverse events. Different scenarios are presented where the models do not converge with reasonable solutions. Multivariate and univariate logistic regression analyses were carried out. For each adverse event under consideration, whether univariate model needs to be run and the selection covariates used for multivariate model need to be decided. This is done dynamically depending on the nature of data so that the odds ratio becomes estimable. This processing is different for continuous and categorical covariates.

**Categorical covariates:**

If any given AE did not occur in the reference category for a given categorical covariate, the univariate model was not run for that covariate and multivariate model did not include that covariate because the odds ratio becomes inestimable for that covariate. In multivariate analysis, a single model is fit for each AE that includes all covariates (unless a categorical covariate needs to be excluded, as described above).

Also, if the only occurrence(s) of the event occur in a single category of a categorical variable, the model will not converge; however, group size seems to play a factor. If all groups are sufficiently large, it is not an issue if all the events occur in one category.
**Continuous covariates:**
For continuous covariates, if the only occurrence(s) of the event happen(s) on the boundary of the covariate, the model usually does not converge. One solution is to tweak the data so the covariate value is no longer completely on the boundary for the people with adverse events. For example, if the covariate is age, and the only event(s) happen(s) in the youngest or oldest person (or people) in the dataset, we slightly change the age of that person (or people) so that it is no longer the youngest or oldest.

**Analysis Dataset Creation**
The analysis dataset contained one-record-per-patient for each of the relevant adverse events. Steps involved in creating analysis dataset are:

1. For each relevant adverse event, create a one-record-per-patient dataset and eliminate all records corresponding to multiple occurrences of an adverse event for same-subject. Every subject involved in the study should be represented in each of these data sets.

2. Create the outcome variable: This variable is assigned a value of "1" if the patient had incidence of the adverse event and "0" if otherwise.

3. Exclude subjects from the dataset if they have any missing covariates.

**Statistical Analysis**
This section will focus on the steps involved to generate multivariate and univariate models dynamically for adverse events.

Univariate and multivariate logistic regression analyses were carried out on adverse event dataset described above using the **GENMOD** procedure. Each of them will be discussed in detail here.

**Univariate Analysis**
For each adverse event, a separate univariate model is run for each covariate. The following code will perform the univariate analyses for a single adverse event and for a single covariate and send the relevant statistics to the output dataset "PARAMS."

```
Sample1: SAS Code for Full Model For Univariate Analysis

ODS OUTPUT parameterestimates=parms modelfit=fit_full;
PROC GENMOD DATA=<analysis data set name goes here> descending;
CLASS xxx ; <indicate SEX or RACE, or do not include this line, as needed>
MODEL had_ae=<covariate name goes here> / dist=binomial link=logit;
WHERE ae_term=<name of the outcome variable goes here>;
RUN;
ODS OUTPUT CLOSE;
```
The model shown in Sample 1 provides all relevant statistics except the p-value corresponding to the general race category. The method to calculate p-value is addressed later. In the dataset "PARMS," the relevant parameters for the continuous covariates are contained in the records where PARAMETER = <covariate variable name>. The relevant parameters for the categorical covariate SEX are contained in the row where LEVEL1='F'. The relevant parameters for the individual levels of the categorical covariate (RACE) are contained in the row where LEVEL1 is not ='White'. The odds ratio (OR) can be computed as \(\exp(\text{ESTIMATE})\). "PROBCHI" is the p-value. Note that LEVEL1, ESTIMATE, PROBCHI and PARAMETER are the names of the variables in the output dataset "PARMS".

NOTE: If a categorical covariate was excluded from the model, the corresponding odds ratio and p-value are not estimable.

Getting the overall p-value for RACE

Running the model shown in Sample 1 for all covariates will provide all statistics except for the p-value for the overall race category. To get the p-value for the overall race category involves performing the likelihood ratio test (LRT) for race. The LRT requires running a model with race (full model – shown in Sample 1); and then running the same model except without race (called the reduced model – shown in Sample 2); and using the log likelihood statistics and degrees of freedom for each model. The following is the reduced model (Sample 2).

Sample 2: SAS Code for Reduced Model for Univariate Analysis

```
ODS OUTPUT modelfit=fit_reduced;
PROC GENMOD DATA=<analysis data set name goes here> descending;
MODEL had_ae=<leave this blank> / dist=binomial link=logit;
WHERE ae_term=< name of relevant adverse event goes here>;
RUN;
ODS OUTPUT CLOSE;
```

The output data set FIT_FULL from the full model (Sample 1) contains the full model degrees of freedom in each of the first four rows of the variable DF (choose any row to get the DF), and the full model log likelihood is in the variable VALUE where the variable CRITERION=Log likelihood. Similarly, the output data set FIT_REduced contains the relevant information from the reduced model. The relevant data from FIT_FULL and FIT_REduced should be combined into a single data set, and the LRT chi squared statistic should be computed as \(2.0*[\text{Loglikelihood(full)} - \text{loglikelihood(reduced)}]\). The degrees of freedom for the LRT is the DF from the reduced model minus DF from the full model. The overall race p-value (for the table output) is computed as PROBCHI(LRT chi squared statistic, degrees of freedom). A sample listing of FIT_FULL dataset is as shown in Table 1.
<table>
<thead>
<tr>
<th>Obs</th>
<th>Criterion</th>
<th>DF</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Deviance</td>
<td>xxx</td>
<td>xx.xxxx</td>
</tr>
<tr>
<td>2</td>
<td>Scaled Deviance</td>
<td>xxx</td>
<td>xx.xxxx</td>
</tr>
<tr>
<td>3</td>
<td>Pearson Chi-Square</td>
<td>xxx</td>
<td>xx.xxxx</td>
</tr>
<tr>
<td>4</td>
<td>Scaled Pearson X2</td>
<td>xxx</td>
<td>xx.xxxx</td>
</tr>
<tr>
<td>5</td>
<td>Log Likelihood</td>
<td>_</td>
<td>xx.xxxx</td>
</tr>
</tbody>
</table>

**Table 1**: Listing Of FIT_FULL Dataset.

**Multivariate Analysis**

For each AE_TERM, a single model is needed that includes all covariates (unless a categorical covariate needs to be excluded as described in the processing considerations section). The code shown in Sample 3 will perform the analysis for a single AETERM and send the relevant output to the dataset PARMS. The likelihood ratio test information for the full model will be in the output dataset FIT_FULL.

<table>
<thead>
<tr>
<th>Sample 3: SAS Code for Full Model for Multivariate Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>ODS OUTPUT parameterestimates=parms modelfit=fit_full;</td>
</tr>
<tr>
<td>PROC GENMOD DATA=&lt;analysis data set name&gt; descending;</td>
</tr>
<tr>
<td>CLASS SEX RACE;</td>
</tr>
<tr>
<td>MODEL had_ae=race† sex† age weight /dist=binomial link=logit;</td>
</tr>
<tr>
<td>WHERE ae_term=&lt;name of relevant adverse event goes here&gt;;</td>
</tr>
<tr>
<td>RUN;</td>
</tr>
<tr>
<td>ODS OUTPUT CLOSE;</td>
</tr>
</tbody>
</table>

†This categorical variable should be excluded from the model if the respective AE did not occur for the reference category.

To get the reduced model information for the likelihood ratio test for race, it will be necessary to run another model which is identical to the above except it will not include race, as follows. The relevant information will be contained in the dataset FIT_REDUCED.

<table>
<thead>
<tr>
<th>Sample 4: SAS Code for Reduced Model for Multivariate Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>ODS OUTPUT parameterestimates=parms modelfit=fit_reduced;</td>
</tr>
<tr>
<td>PROC GENMOD DATA=&lt;analysis data set name&gt; descending;</td>
</tr>
<tr>
<td>CLASS SEX;</td>
</tr>
<tr>
<td>MODEL had_ae=sex† age weight /dist=binomial link=logit;</td>
</tr>
<tr>
<td>WHERE ae_term=&lt;name of relevant adverse event goes here&gt;;;</td>
</tr>
<tr>
<td>RUN;</td>
</tr>
<tr>
<td>ODS OUTPUT CLOSE;</td>
</tr>
</tbody>
</table>
†This categorical variable should be excluded from the model if the respective AE did not occur for the reference category.

To get the p-value for the overall race category requires performing the LRT. Use the datasets FIT_FULL and FIT_REduced in an identical fashion to that described for the univariate race LRT.

Notes on the model statements:

- The 'NOPRINT' option will suppress the ODS output and thus should not be used.
- The option DESCENDING makes HAD_AE=1 the outcome whose probability is being modeled.

The CLASS statement in the univariate model is needed when the covariate is categorical. If this statement is included when the covariate is not categorical, it will be ignored.

As described in the previous sections, different models are used based on the nature of adverse event data. A macro was written to dynamically generate the model statement of GENMOD procedure. If any given AE did not occur in the reference category for a given categorical covariate, the univariate model was not run for that covariate, and the multivariate model did not include that covariate.

**Conclusion and Summary**
Convergence of the models defined in GENMOD procedure depends on the nature of data. Some of the scenarios where the models do not converge and odds ratio becomes inestimable were presented in this paper. It is difficult to conclude the association of covariates to the occurrence of adverse events in these scenarios. Reasonable solutions were presented to estimate the odds ratio in these scenarios. The proposed approach was successful in carrying out adverse event logistic regression analysis by dynamically varying the model parameters depending on the nature of the data. The model covariates are selected dynamically so that the model would converge and the odds ratio is estimable.

**References**

3. Data coding for logistic regression in SAS: [http://cc.uoregon.edu/cnews/spring2005/saslogistic.htm](http://cc.uoregon.edu/cnews/spring2005/saslogistic.htm)
4. Articles on data analysis issues: [http://darkwing.uoregon.edu/~robinh/analysis.html](http://darkwing.uoregon.edu/~robinh/analysis.html)
Acknowledgments
We would like to thank our manager Amy Gillespie and management of MaxisIT Inc. for their kind support and cooperation. We would also like to thank Greg Golm, Associate Director, Merck & Co., Inc. for his valuable inputs.

Contact Information
Your comments and questions are valued and encouraged. Contact the authors at:

Author Name: Suhas R. Sanjee
Company : MaxisIT Inc.
Address : 1199 Amboy Ave
City: Edison  State: NJ  ZIP: 08830
Email: sanjee@maxisit.com

Author Name: Sheng Zhang
Company : Merck & Co., Inc.
Address : 351 Sumneytown Pike
City: Upper Gwynedd State: PA  ZIP: 19454
Email: sheng_zhang@merck.com

Trademark Information
SAS and SAS/MACRO software are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are registered trademarks or trademarks of their respective companies.