THE GLIM PROCEDURE: AN INTERFACE TO THE SAS® SYSTEM

P.M. Grambsch, J.L. Kosanke, T.M. Therneau, D.J. Schaid,
A.R. Zinsmeister, H.S. Wieand, K.P. Offord, J.J. LarsonKeller
Mayo Clinic

INTRODUCTION

The GLIM procedure (PROC GLIM) provides an interface between SAS and the GLIM package (Generalized Linear Models) available from the Numerical Algorithms Group of the Royal Statistical Society (ref. #1). The GLIM package implements a significant subset of the methods found in the book "Generalized Linear Models" by McCullagh and Nelder (ref. #2).

The models fall into a general framework of $E(Y_i) = f(X_i \beta)$, where $X_i \beta$ is the linear predictor found in ordinary linear models, $f$ is a function such as exp or log, and $Y_i$ is assumed to have a probability distribution from the exponential family. Fitting is done by maximum likelihood.

Two methods of interface are provided with this procedure. The first method is appropriate for standard GLIM models, where the distribution is one of: Gaussian, binomial, Poisson or gamma and the function linking the mean and linear predictor ($f^{-1}$) belongs to the power family or, for the binomial, is one of cumulative normal, inverse logit or log log. In this method, the interface appears to the user very similar to any procedure of SAS® software. The user directives consist of ordinary SAS® statements (MODEL, CLASS, etc.) in the usual SAS syntax, described below. These are translated and passed to the GLIM package, the output is captured, and the results are printed. To use capabilities of the GLIM system beyond the SAS-like statements provided by the first method, a PARMCARDS facility can be used to pass "raw" GLIM commands to the package. The GLIM printed output is captured. In either mode, parameter estimates, variances and covariances, and predicted and residual values may be written to SAS® data sets.

Only one MODEL statement, specifying a set of nested models, can be processed by the GLIM procedure in each invocation. You must specify:

- The dependent and independent variables.
- The LINK function (equal to $f^{-1}$ in the above notation).
- The ERROR distribution.
- You may also:
  - Set the tolerances that control convergence of the algorithm and its detection of linear dependencies.
  - Choose the method for estimating the variance or SCALE.
  - Specify weights for the observations.
  - Specify CLASS variables.
  - Fix the value of beta for selected variables in a model.
  - Create output data sets.
  - Specify printout options.

SPECIFICATIONS

The following statements are used with PROC GLIM:

PROC GLIM options:
MODEL dependent - independents/intercept option;
ERROR - error distribution/scale option;
LINK - name of link function;
CLASS = list of classification variables ("factors");
WEIGHT - weighting_variable;
FIXBETA vname - value vname = value ... ;
OUT - dataset_name options;
PARMCARDS; statements for the GLIM program
ID variable list;
BY by variable list;
The MODEL, ERROR, and LINK statements are required, or alternatively, an ID statement with PARMCARDS.

PROC GLIM Statement

The options below can appear in the PROC GLIM statement:
DATA = SASdataset
names the SAS data set containing the data to be analyzed by PROC GLIM. If DATA= is omitted, the most recently created SAS data set is used.
OUTEST = SASdataset
names the SAS data set to contain the parameter estimates produced by PROC GLIM. See Output Data Sets below for more information.
OUTCOV = SASdataset
names the SAS data set to contain the variance/covariance matrix of the parameter estimates. See Output Data Sets below for more information.
POUTTEST = SASdataset
names the SAS data set to contain the parameter estimates produced by PROC GLIM when the PARMCARDS option is used. See Output Data Sets below for more information.
POUTCOV = SASdataset
names the SAS data set to contain the variance/covariance matrix when the PARMCARDS option is used. See Output Data Sets below for more information.
MAXITER - value
sets the maximum number of iterations for the convergence of the fitting algorithm. The default value is 10.
GEPS - value
sets the convergence criteria for the proportional change in the log-likelihood. The default value is 1.0E-4.
MODEL dependent = independents/options;

The MODEL statement defines the model(s) to be fit by GLIM. A set of nested models (including the null model) will be fit, with variables added to the model one at a time from left to right. The full model involves all the independent variables in the model statement. Deviances for each of the nested models are printed, but parameter estimates are available only for the full model.

The MODEL statement can include interaction terms of the form a*b, where a and b must be class variables. The interpretation of these terms is exactly as in GLIM. The more general interaction symbols (the first 8 bytes are used).

There is one option available for the MODEL statement. Specifying NOINT will cause the model to be fit without an intercept term. Without the NOINT option an intercept term is always the first variable of the fit. With NOINT, the null model assumes $f(\mu) = 0$. In order to avoid problems in computing the deviance for the null model, NOINT is allowed only when using the canonical link for Normal (identity), Poisson (log) and Binomial (logit).

NOINT is not allowed for gamma errors. The user can always circumvent this prohibition with the PARMCAVDS option.

ERROR Statement

ERROR distribution-name/options:

The ERROR statement specifies the distribution of the errors, or equivalently, the distribution of $y$. The distribution name must be one of NORMAL, BINOMIAL, POISSON, or GAMMA, where N, B, P, and G are acceptable abbreviations. If the distribution is binomial, then the next item in the ERROR statement must be the name of the variable which contains the denominator for the proportion of successes (while $y$ contains the number of successes).

There is one option, SCALE, which may take one of three settings. The GLIM procedure assumes that the variance of each $y$ satisfies the equation $\text{var}(y) = \text{scale} \times V(\mu)$, where $\mu = E(y)$ is the expectation and the function $V(\mu)$ is

**SCALE** = \text{DEVIANCE or P}

The allowable settings for SCALE are PEARSON or P

PEARSON will cause the scale to be estimated using the sum of squared weighted residuals $(y_i - u_i)$, with the weight of $w_i / V(\mu)$ where $w_i$ is the user-supplied weight (see Weight Statement) and the weighted sum is divided by the degrees of freedom for error. DEVIANCE will cause the scale to be estimated using the deviance divided by the degrees of freedom for error. A numerical value will cause the scale to be set to that value. The variance-covariance matrix of the parameter estimates (and the standard errors) incorporate the scale value as specified or estimated.

The default values for SCALE are SCALE=PEARSON in the normal and gamma case, and SCALE=1 in the binomial and poisson case.

LINK Statement

LINK function-name;

The LINK statement gives the functional relationship between $y$ and the linear predictor ($X \beta$), where $\mu = ((EXy) - f(X \beta)$ and the link function is $f$. There are 8 choices:

<table>
<thead>
<tr>
<th>Function</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>IDENTITY (ID)</td>
<td>Identity function $\mu$</td>
</tr>
<tr>
<td>LOGLOG (11)</td>
<td>$\log(\mu) = \log(1 - \mu/n)$</td>
</tr>
<tr>
<td>LOGIT (LT)</td>
<td>$\log(\mu/(1-\mu))$</td>
</tr>
<tr>
<td>PROBIT (PR)</td>
<td>$\Phi^{-1}(\mu/n)$</td>
</tr>
<tr>
<td>LOG (LG)</td>
<td>$\log(\mu)$</td>
</tr>
<tr>
<td>INVERSE (IN)</td>
<td>$1/\mu$</td>
</tr>
<tr>
<td>SQRT (SQ)</td>
<td>$\mu^{1/2}$</td>
</tr>
<tr>
<td>EXP (E)</td>
<td>$\mu^{*k}$</td>
</tr>
</tbody>
</table>

$\Phi(X)$ is the standard normal cumulative distribution function.

For the EXP function the LINK statement must also include the numeric value for $k$ (e.g. EXP 0.5). The LOGLOG, LOGIT, and PROBIT links are appropriate only for binomial data, and vice versa. The log function is the natural logarithm (base e).

CLASS Statement

CLASS variable list;

The CLASS statement specifies the variables for GLIM to consider as factors. The levels for each factor are assigned based on the sorted values of the class variable.

WEIGHT Statement

WEIGHT variable-name;

The WEIGHT statement assigns case weights
to the observations. The values of the WEIGHT variable are checked and those observations with negative weights eliminated. Case weights are inversely proportional to case variances.

**FIXBETA Statement**

```
FIXBETA variable-name=value variable-name=value ... ;
```

The FIXBETA statement allows the coefficient of one or more of the non-class independent variables to be fixed in advance. Variable names that appear in the FIXBETA statement must also appear as independent variables in the MODEL statement.

**OUT Statement**

```
OUT= SASdataset P= varname DR= varname
PR= varname;
OUT= SASdataset P= varname R= varname;
```

The OUT statement specifies an output data set containing the predicted value (P), deviance residual (DR), Pearson residual (PR), or any combination of the three for each observation for the full model specified by the MODEL statement. The POUT statement specifies an output data set containing the predicted value (P), or Pearson residual (R) for the last model fit by the GLIM commands following the PARMCARDS. Deviance residuals are not available with the PARMCARDS option. The output data set named with the OUT= or with the POUT= contains all the variables in the input data set along with the values resulting from fitting the model. Only observations with non-missing data for the dependent variable are given predicted and/or residual values.

**ID Statement**

```
ID variable-list;
```

The variables in the ID statement are passed to GLIM along with the variables found in the WEIGHT and MODEL statements. The purpose of this statement is to make variables available to GLIM for use with the PARMCARDS facility. (Also see variable names below.)

**PARMCARDS Statement**

```
PARMCARDS; [GLIM commands]
```

The lines following a PARMCARDS statement, up to but not including the next line containing a semicolon, are passed directly to the GLIM package and should contain GLIM commands. The statements contained on these lines will be executed after the statements that are automatically generated by PROC GLIM to handle the MODEL, LINK, etc. statements described above. The variables found in the model statement, along with any in an ID or WEIGHT statement, will be passed to GLIM. The GLIM commands may appear anywhere in columns 1-72. They must follow standard GLIM syntax rules. Any GLIM command may be used except the accuracy command. PROC GLIM operates at accuracy = 9, the maximum accuracy allowed. Channels 20-32 are reserved by PROC GLIM and should not be used in any PARMCARDS statements. Printed output from the GLIM package appears after the GLIM procedure output.

**Output Datasets**

The OUTEST data set contains one observation per estimated unaliased parameter from the full model specified in the model statement and the POUTEST data set contains one observation per estimated unaliased parameter from the last model fit by the GLIM commands following the PARMCARDS. There are two variables on these data sets: ESTIM containing the estimated parameter values and DESCRIP, a character variable containing a brief description of the parameter. The estimate of the scale parameter is not part of the output.

The OUTCOV data set contains the variance/covariance matrix of the unaliased parameter estimates from the full model specified on the MODEL statement. If the scale - Pearson option was used, the covariance matrix will incorporate the Pearson estimate for the scale parameter. The POUTCOV data set contains the variance/covariance matrix of the unaliased parameter estimates fit by the GLIM commands following PARMCARDS. The Pearson scale estimate is not an option when PARMCARDS is used. The variable names are "COL1", "COL2", "COL3", .... etc.

When the BY variable statement is used with the OUTCOV= or POUTCOV= option, a separate SAS data set is created for the variance/covariance matrix of each BY group. The naming convention is that the data set name is the SAS dataset name supplied by the user (in the OUTCOV= SAS dataset option) with a 1 or 2 digit number appended to denote the BY group. The last two letters of the user-supplied name may be truncated if necessary to maintain an 8-character word length. Up to 99 BY groups are supported with this option. (If OUTCOV is not specified, the number of groups is unlimited.)

**Variable Names**

There is a difference between the SAS and GLIM variable names which is visible only to the PARMCARDS user. GLIM names have only 4 characters of which neither the first nor last can be an underscore. PROC GLIM creates GLIM names from SAS names as follows:

1. The dependent variable is called Y.
2. The independent variables are called I1, I2, etc.
3. The denominator for binomial distribution is called N.
4. The weight variable is called W.
5. The variables in the ID list are called D1, D2, etc.

**Printed Output**

The standard printed output consists of four parts.

1. Summary of the variables in the MODEL statement, including the weight variable and ID variables if given. Numeric and class variables are summarized separately. For each numeric variable, the number missing, minimum, maximum and mean are given. For class variables, the number missing, number of classes and values of the class variable are provided. The number of observations, number deleted due to missing data and number used in the analysis are given.
2. Summary of the GLIM directives: The dependent variable, error distribution, option chosen for the scale parameter, link function, weight variable, use of intercept option, use of the fixed beta option and use of parmcards.

3. Deviance table: The deviance for each of the nested models, the difference in deviance and degrees of freedom between successive models, and the difference in deviance divided by the degrees of freedom. The asymptotic test statistic for assessing improvement in fit (chi-square statistic for models with known scale parameter and F-statistic for models with estimated scale) and p-values are given.

4. Estimates of coefficients: The parameter estimates, asymptotic estimated standard errors, test statistic, and p-values for each of the unaliased parameters in the full model. Extrinsically aliased parameters are indicated. Optionally, a matrix with parameter variances on the diagonal, covariances below the diagonal and correlations above it is printed.

GLIM Diagnostics

Warnings and GLIM and system faults that GLIM detects are captured and output.

Warnings: When fitting a model, a condition may occur that indicates that the results are unreliable. If so, one or more warning messages will appear as part of the printed output immediately before the Deviance Table. The warning messages are:

1. NO CONVERGENCE YET
   The numerical algorithm has not converged after the maximum number of iterations.

2. ITERATIONS DIVERGED
   The numerical algorithm has detected divergence, increasing deviance.

3. CHANGE IN DF
   A parameter has become aliased during the fit and results will be incorrect.

4. UNIT I HELD AT LIMIT
   This occurs with the binomial distribution when the linear predictor for unit 1 is tending to ± infinity. The fitted value is set to either zero or the binomial denominator. A given invocation of PROC GLIM may fit multiple nested models. Every warning message appearing on each model fit is printed, but the messages are not identified as to which model causes them. The option PRINT GLIM will provide more details.

Faults: A fault occurs when GLIM cannot execute a statement successfully. GLIM fault messages are appended on the SAS output log.

Examples

The example is taken from reference 2, pp. 162-164. The dataset gives the clotting time of normal plasma diluted to nine different percentage concentrations with prothrombin-free plasma. Data on two different lots of clotting agent were presented; we consider only lot 2. McCullagh and Nelder suggested modeling the clotting time as a function of concentration, using the gamma distribution, the inverse link and the log of the percentage concentration as the linear predictor. We follow their suggestion, using these data to illustrate the estimation of the scale parameter. We then illustrate the PARMCARDS option. We replace the gamma distribution with the inverse Gaussian (Wald) distribution and utilize GLIM macros to specify a user-defined model via PARMCARDS. The first two macros, M1 and M2, specify the link function by giving the inverse of the link function and the derivative of the link function. The last two macros, M3 and M4, specify the distribution by giving the variance as a function of the mean and the deviance for each unit.
The interface was written in PL/I. It is designed for SAS Version 5, GLIM 3.77 and the IBM/MVS environment.

References


SAS and SAS/STAT are registered trademarks of SAS Institute Inc., Cary, NC, USA.

Questions and comments are very welcome. Please contact the Section of Biostatistics, Mayo Clinic, Rochester, MN 55905. (507-284-5567).

1267