The DMREG Procedure

Overview

Procedure Syntax

PROC DMREG Statement
CLASS Statement
CODE Statement
DECISION Statement
FREQ Statement
MODEL Statement
NLOPTIONS Statement
REMOTE Statement
SCORE Statement

Details

Examples

Example 1: Linear and Quadratic Logistic Regression with an Ordinal Target (Rings Data)
Example 2: Performing a Stepwise OLS Regression (DMREG Baseball Data)
Example 3: Comparison of the DMREG and LOGISTIC Procedures when Using a Categorical Input Variable

References

Copyright 2000 by SAS Institute Inc., Cary, NC, USA. All rights reserved.
Overview

DMREG enables you to fit both linear and logistic regression models. Linear regression attempts to predict the value of a continuous target as a linear function of one or more independent inputs. Logistic regression attempts to predict the probability that a categorical (binary, ordinal, or nominal) target will acquire the event of interest as a function of one or more independent inputs. The procedure supports forward, backward, and stepwise selection methods. It also allows you to score data sets or generate SAS DATA step code to score a data set.

Copyright 2000 by SAS Institute Inc., Cary, NC, USA. All rights reserved.
The DMREG Procedure

Procedure Syntax

PROC DMREG < option(s)>;
   MODEL dependent=independent(s) </ model-option(s)>;
   CLASS variable(s);
   CODE code-option(s);
   DECISION DECDATA=<libref:SAS-data-set<DECVARS=decision-variable(s)> <option(s)>;
   FREQ variable;
   NLOPTIONS nonlinear-option(s);
   REMOTE remote-option(s);
   SCORE scoring-option(s);

Copyright 2000 by SAS Institute Inc., Cary, NC, USA. All rights reserved.
The DMREG Procedure

PROC DMREG Statement

Invokes the DMREG procedure.

PROC DMREG <option(s)>;

Required Arguments

DATA=<libref.> SAS-data-set
   Identifies the training data set.

DMDBCAT=<libref.> SAS-catalog
   Identifies the training data catalog.

Options

COVOUT
   Specifies that the OUTEST= data set is to include the variance-covariance matrix of the parameter estimates.

DESCENDING
   Specifies that the order of categorical target is to be reversed.

ESTITER=n
   Specifies that the OUTEST= data set contains parameter estimates and fit statistics (for training, test, and validation data) for every n-th iteration.

   Default: 0. Only the parameter estimates of the final iteration are output.

INEST=<libref.> SAS-data-set
   Identifies the data set that contains initial estimates.

MINIMAL
   Specifies the use of minimal resources to fit a logistic regression model. Memory for the Hessian matrix is not needed. The optimization defaults to the conjugate gradient technique and standard errors of the regression parameters are not computed. Model selection is disabled when this option is specified. This option does not apply to the normal error regression models.

NAMELEN=n
   Specifies the length of effect names in the printed output to be n characters, where n is a value between 20 and 200. The default length is 20 characters.

OUTEST=<libref.> SAS-data-set
IDENTIFIES THE OUTPUT DATA SET CONTAINING ESTIMATES AND FIT STATISTICS. SEE FOR MORE INFORMATION.

NOPRINT
  Suppresses all printed output.

SIMPLE
  Prints simple descriptive statistics of the input variables.

TESTDATA=<libref.> SAS-data-set
  Identifies the data set containing test data.

VALIDATA=<libref.> SAS-data-set
  Identifies the data set containing validation data.

Copyright 2000 by SAS Institute Inc., Cary, NC, USA. All rights reserved.
CLASS Statement

Specifies one or more categorical variables to be used in the analysis.

CLASS variable(s); 

Required Argument

variable(s)

Specifies a list of categorical variables to be used in the analysis. You must specify the target variable if it has a categorical (binary, ordinal, or nominal) measurement level.
The DMREG Procedure

CODE Statement

Controls the creation of SAS code that can be used to score data sets.

Tip: If neither FILE= nor METABASE= is specified, then the SAS code is written to the SAS log.

```
CODE <code-option(s)>;
```

CODE Options

ERROR
   Specifies that the error function is to be computed.

FILE=
   Specifies the path for writing the code to an external file. For example, FILE="c:\mydir\scorecode.sas".

FORMAT=
   Specifies numeric formats for printing the estimated parameters.

GROUP=
   Specifies the group identifier (up to four characters) for group processing.

METABASE= mylib.mycat.myentry
   Specifies the code catalog entry to which the results are written.

RESIDUAL
   Specifies that residuals are to be computed.
DECISION Statement

Specifies information used for decision processing in the DECIDE, DMREG, NEURAL, and SPLIT procedures. This documentation applies to all four procedures.

Tip: The DECISION statement is required for the DMREG and NEURAL procedures. It is optional for PROC SPLIT.

DECISION DECDATA=<libref.> SAS-data-set <DECVARS=decision-variable(s)><option(s)>;

DECDATA= <libref.> SAS-data-set

Specifies the input data set that contains the decision matrix. The DECDATA= data set must contain the target variable.

Note: The DECDATA= data set may also contain decision variables specified by means of the DECVARS= option, and prior probability variable(s) specified by means of the PRIORVAR= option.

The target variable is specified by means of the TARGET statement in the DECIDE, NEURAL, and SPLIT procedures or by using the MODEL statement in the DMREG procedure. If the target variable in the DATA= data set is categorical then the target variable of the DECDATA= data set should contain the category values, and the decision variables will contain the common consequences of making those decisions for the corresponding target level. If the target variable is interval, then each decision variable will contain the value of the consequence for that decision at a point specified in the target variable. The unspecified regions of the decision function are interpolated by a piecewise linear spline.

Tip: The DECDATA= data set may be of TYPE=LOSS, PROFIT, or REVENUE. If unspecified, TYPE=PROFIT is assumed by default. TYPE= is a data set option that should be specified when the data set is created.

DECVARS=decision-variable(s)

Specifies the decision variables in the DECDATA= data set that contain the target-specific consequences for each decision.

Default: None

COST=cost-option(s)

Specifies numeric constants that gives the cost of a decision, or variables in the DATA= data set that contain the case-specific costs, or any combination of constants and variables. There must be the same number of cost constants and variables as there are decision variables in the DECVARS=
option. In the COST= option, you may not use abbreviated variable lists such as D1-D3, ABC--XYZ, or PQR:

| Default:            | All costs are assumed to be 0. |

**CAUTION:**

The COST= option may only be specified when the DECDATA= data set is of TYPE=REVENUE.

**PRIORVAR=variable**

Specifies the variable in the DECDATA= data set that contains the prior probabilities to use for making decisions.

| Default:            | None |

Copyright 2000 by SAS Institute Inc., Cary, NC, USA. All rights reserved.
FREQ Statement

Specifies the variable that contains frequencies for training data.

FREQ <variable> ;

variable

Specifies the frequency variable. If specified, the FREQ variable overrides whatever is in the DMDB metadata. If the FREQ statement contains no name, then a FREQ variable is not used.

CAUTION:

If there is a frequency variable in the DMDB, it is not advisable to use another variable as a frequency variable because the training data does not contain observations with invalid values in the FREQ variable specified in the DMDB. For example, if the frequency variable specified in the DMDB contains a 0 or negative value, then that observation is discarded even if the FREQ variable that you specified in the FREQ statement of the DMREG procedure contains valid frequency values.

Default: If the FREQ statement is not specified, the frequency variable in the DMDB is used. If the FREQ statement is specified without a variable, a frequency of 1 is used for all observations.

Range: The frequency variable can contain integer or non-integer values.
MODEL Statement

Specifies modeling options.

Requirements: Model statement is required.

MODEL dependent=independent(s) / model-option(s);

Required Argument

dependent=independent(s)

where the arguments are defined as follows:

dependent
    Specifies the response variable (target).

independents
    Specifies the explanatory variables or effects (inputs). The syntax of effects is described in.

Options

model-options(s)

Specifies options that affect the fit, confidence intervals, variable selection, and specification of the model as follows:

MODEL Options - Fitting Options

MISCCONV=n

Specifies the critical misclassification rate at which to stop iterations.

Default: \( n = 0 \)

Range: \( 0 - 1 \)

STARTMISC=n

Specifies the number of iterations to be processed before checking misclassification rate.
<table>
<thead>
<tr>
<th>Default:</th>
<th>Depends on the optimization technique:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( n = 3 )</td>
</tr>
<tr>
<td></td>
<td>TECHNIQUE = NEWRAP, NRRIDG, TRUREG</td>
</tr>
<tr>
<td></td>
<td>( n = 5 )</td>
</tr>
<tr>
<td></td>
<td>TECHNIQUE = QUANEW, DBLDOG</td>
</tr>
<tr>
<td></td>
<td>( n = 10 )</td>
</tr>
<tr>
<td></td>
<td>TECHNIQUE = CONGRA</td>
</tr>
<tr>
<td>Alias:</td>
<td>STATMISC</td>
</tr>
</tbody>
</table>

**MODEL Options - Miscellaneous Options**

**ALPHA=\( n \)**

Specifies the significance level of confidence intervals for regression parameters.

**Default:** 0.05

**CLPARM**

Specifies the computation of confidence intervals for parameters.

**CORRB**

Specifies that the correlation matrix is to be printed.

**COVB**

Specifies that the covariance matrix is to be printed.

**MODEL Options - Selection Options**

**CHOOSE=AIC | NONE | SBC | TDECDATA | VDECDATA | VERROR | VMISC | XDECDATA | XERROR | XMISC**

Specifies the criterion for the selection of the model.

**AIC**

Represents the Akaike Information Criterion. The model with the smallest criterion value is chosen.

**NONE**

Chooses standard variable selection based on the entry and/or stay \( P \)-values.

**SBC**

Represents the Schwarz Bayesian Criterion. The model with the smallest criterion value is chosen.

**TDECDATA**

Represents the total profit/loss for the training data. The model with the largest profit or the smallest loss is chosen.
VDECDATA
Represents the total profit/loss for the VALIDATA= data set. The model with the largest profit or the smallest loss is chosen.

VERROR
Represents the error rate for the VALIDATA= data set. The error is the sum of square errors for least-square regression and negative log-likelihood for logistic regression. The model with the smallest error rate is chosen.

VMISC
Represents the misclassification rate for the VALIDATA= data set. The model with the smallest misclassification rate is chosen.

XDECDATA
Represents the total profit/loss for cross-validation of the training data. The model with the largest profit or the smallest loss is chosen.

XERROR
Represents the error rate for cross validation. The error is the sum of square errors for least-square regression and negative log-likelihood for logistic regression. The model with the smallest error rate is chosen.

XMISC
Represents the misclassification rate for cross validation. The model with the smallest misclassification rate is chosen.

Default: If decision processing is specified, the default is CHOOSE=TDECDATA; if the VALIDATA= data set is also specified, the default is CHOOSE=VDECDATA.

DETAILS
Prints details at each model selection step.

HIERARCHY=ALL | CLASS
Specifies how containment is to be applied.

ALL
Specifies that all independent variables that meet hierarchical requirements are included in the model.

CLASS
Specifies that only CLASS variables that meet hierarchical requirements are included in the model.

Default: ALL

INCLUDE=n
Specifies that the first $n$ effects in the model are to be included in each model.

Default: 0
**MAXSTEP=n**

Specifies the maximum number of steps for the STEPWISE variable selection method.

**Default:** Two times the number of effects specified in the MODEL statement.

**NODESIGNPRINT**

Suppresses the display of the coding of the CLASS inputs.

** Alias:** NODP

**RULE=MULTIPLE | SINGLE | NONE**

Specifies the rule for inclusion of effects for SELECTION=FORWARD, BACKWARD, or STEPWISE.

**MULTIPLE**

One or more effects can be considered for entry or removal at the same time provided the hierarchical rule is observed. For example, if main effects A and B and interactions A*B are not in the model, effects that can be considered for entry in a single step are A alone, or B alone, or A, B, and A*B together.

**SINGLE**

A single effect is considered for entry into the model only if its lower order effects are already in the model; a single effect is considered for removal from the model only if its higher order effects are not in the model.

**NONE**

Effects are included or excluded one at a time without preservation of any hierarchical order.

**Default:** RULE=NONE


**SELECTION= FORWARD | BACKWARD | STEPWISE | NONE**

Specifies the variable selection methods.

**FORWARD**

Begins with no inputs in the model and then, systematically, adds inputs that are related to the target.

**BACKWARD**

Begins with all inputs in the model and then, systematically, removes inputs that are not related to the target.

**STEPWISE**

Systematically adds and deletes inputs from the model. Stepwise selection is similar to forward selection except that stepwise may remove an input after it has entered the model.
and replace it with another input.

NONE

All inputs are used to fit the model.

**Default:** NONE

**SEQUENTIAL**

Specifies the addition or deletion of variables in sequential order, as specified in the MODEL statement.

**SLENTRY=**\(n\)

Specifies the significance level for addition of variables.

**Default:** .05

**SLSTAY=**\(n\)

Specifies the significance level for removal of variables.

**Default:** .05

**START=**\(n\)

Specifies that the first \(n\) effects be included in the starting model.

**Default:**

<table>
<thead>
<tr>
<th>0 - for the FORWARD or the STEPWISE method</th>
</tr>
</thead>
<tbody>
<tr>
<td>(s) (the total number of effects in the MODEL statement)- for the BACKWARD method</td>
</tr>
</tbody>
</table>

**Range:** The value of \(n\) ranges from 0 to \(s\), where \(s\) is the total number of effects in the MODEL statement.

**STOP=**\(n\)

Specifies the maximum (FORWARD method) or minimum (BACKWARD method) number of effects to be included in the final model. The variable selection process is stopped when \(n\) effects are added or deleted. The STOP= option has no effect when SELECTION=NONE or STEPWISE.

**Range:** The value of \(n\) ranges from 0 to \(s\), where \(s\) is the total number of effects in the MODEL statement.

**Default:**

<table>
<thead>
<tr>
<th>(s) - for the FORWARD method</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 - for the BACKWARD method</td>
</tr>
</tbody>
</table>

**MODEL Options - Specification Options**
CODING= DEVIATION | GLM

Specifies design variable coding for CLASS inputs.

DEVIATION

Deviation from mean coding, which is also known as effect coding.

GLM

Non-full rank GLM coding as used in the GLM procedure.

**Default:** CODING=DEVIATION

LEVEL=INTERVAL | NOMINAL | ORDINAL

Specifies the measurement level of the target variable.

INTERVAL

Interval variable.

NOMINAL

Nominal variable.

ORDINAL

Ordinal variable.

**Default:** ORDINAL for a categorical target; INTERVAL for a numerical target.

ERROR=MBERNOULLI | NORMAL

Specifies the error distribution.

MBERNOULLI

Multinomial distribution with on trial. This includes the binomial distribution with on trial. MBERNOULLI is not available if the target measurement level is interval.

**Alias:** BINOMAIL or MULTINOMIAL

NORMAL

Normal distribution. NORMAL is not allowed e if the target measurement level is nominal.

**Default:** ERROR=NORMAL (for LEVEL=INTERVAL), ERROR=MBERNOULLI (otherwise).

LINK= CLOGLOG | IDENTITY | LOGIT | PROBIT

Specifies the link function that represents the expected values of the target to the linear predictors.

CLOGLOG

Specifies the complementary log-log function, which is the inverse of the extreme value distribution function. The CLOGLOG function is available for ordinal or binary targets.

IDENTITY

Specifies the identity function. The IDENTITY function can only be used for the linear
regression analysis (ERROR=NORMAL).

LOGIT

Specifies the logit function, which is the inverse of the logistic distribution function. The LOGIT function is available for nominal, ordinal, or binary targets.

PROBIT

Specifies the probit function, which is the inverse of the standard normal distribution function. The PROBIT function is available for ordinal or binary targets.

<table>
<thead>
<tr>
<th>Default:</th>
<th>LOGIT (for ERROR=MBERNOULLI), IDENTITY (for ERROR=NORMAL).</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>IDENTITY (for ERROR=NORMAL)</td>
</tr>
</tbody>
</table>

Tip: The CLOGLOG, LOGIT, and PROBIT link functions are used for a logistic regression analysis. The IDENTITY link function is used for a linear regression analysis.

NOINT

Suppresses the intercept for the binary target model or the normal error linear regression model.

SINGULAR= n

Specifies the tolerance for testing singularity.

Default: $1e^{-6}$
The DMREG Procedure

NLOPTIONS Statement

Specifies options for nonlinear optimizations. These options only apply to logistic regression models.

NLOPTIONS nonlinear-option(s);

Nonlinear-Options

**ABSCONV= number**

Specifies an absolute function convergence criterion. ABSCONV= is a function of the log-likelihood for the intercept-only model. The optimization is to maximize the log-likelihood.

<table>
<thead>
<tr>
<th>Default:</th>
<th>The default value is 1e-3 times the log-likelihood of the null model (intercept-only model).</th>
</tr>
</thead>
<tbody>
<tr>
<td>Range:</td>
<td>number &gt; 0</td>
</tr>
</tbody>
</table>

**ABSFCONV= number**

Specifies an absolute function convergence criterion.

<table>
<thead>
<tr>
<th>Default:</th>
<th>$10^{-3}$ times the log-likelihood of the intercept-only model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Range:</td>
<td>number &gt; 0</td>
</tr>
</tbody>
</table>

**ABSGCONV= number**

Specifies the absolute gradient convergence criterion.

<table>
<thead>
<tr>
<th>Default:</th>
<th>1E-5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Range:</td>
<td>number &gt; 0</td>
</tr>
</tbody>
</table>

**ABSXCONV= number**

Specifies the absolute parameter convergence criterion.

<table>
<thead>
<tr>
<th>Default:</th>
<th>1E-8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Range:</td>
<td>number &gt; 0</td>
</tr>
</tbody>
</table>

**DAMPSTEP= number**

Specifies that the initial step size value for each line search used by the QUANEW, CONGRA, or NEWRAP techniques cannot be larger than the product of number and the step size value used in the previous iteration.
**DIAHES**

Forces the optimization algorithm (TRUREG, NEWRAP, or NRRIDG) to take advantage of the diagonality.

**FCONV= number**

Specifies a function convergence criterion.

<table>
<thead>
<tr>
<th>Default</th>
<th>$10^{-FDIGITS}$, where FDIGITS is the value of the FDIGITS= option.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Range</td>
<td>$number &gt; 0$</td>
</tr>
</tbody>
</table>

**FCONV2= number**

Specifies another function convergence criterion.

<table>
<thead>
<tr>
<th>Default</th>
<th>$10^{-4}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Range</td>
<td>$number &gt; 0$</td>
</tr>
</tbody>
</table>

**FDIGITS= number**

Specifies the number of accurate digits in evaluations of the objective function.

<table>
<thead>
<tr>
<th>Default</th>
<th>$-\log_{10}(\varepsilon)$, where $\varepsilon$ is the machine precision.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Range</td>
<td>$number &gt; 0$</td>
</tr>
</tbody>
</table>

**FSIZE= number**

Specifies the parameter of the relative function and relative gradient termination criteria.

<table>
<thead>
<tr>
<th>Default</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Range</td>
<td>$number \geq 0$</td>
</tr>
</tbody>
</table>

**GCONV= number**

Specifies the relative gradient convergence criterion.

<table>
<thead>
<tr>
<th>Default</th>
<th>$10^{-6}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Range</td>
<td>$number &gt; 0$</td>
</tr>
</tbody>
</table>

**GCONV2= number**

Specifies another relative gradient convergence criterion.

<table>
<thead>
<tr>
<th>Default</th>
<th>0</th>
</tr>
</thead>
</table>
HESCAL= 0 | 1 | 2 | 3  
Specifies the scaling version of the Hessian or cross-product Jacobian matrix used in NRRIDG, TRUREG, LEVMAR, NEWRAP, or DBLDOG optimization.

Default: 1 - for LEVMAR minimization technique  
0 - for all others

INHESIAN= number  
Specifies how to define the initial estimate of the approximate Hessian for the quasi-Newton techniques QUANEW and DBLDOG.

Range: number ≥ 0  
Default: The default is to use a Hessian based on the initial estimates as the initial estimate of the approximate Hessian. When r=0, the initial estimate of the approximate Hessian is computed from the magnitude of the initial gradient.

INSTEP= number  
Specifies a larger or smaller radius of the trust region used in the TRUREG, DBLDOG, and LEVMAR algorithms.

Default: 1  
Range: number > 0

LINESEARCH= number  
Specifies the line-search method for the CONGRA, QUANEW, and NEWRAP optimization techniques.

Default: 2  
Range: 1 ≤ number ≤ 8

LSPRECISION= number  
Specifies the degree of accuracy that should be obtained by the second and third line-search algorithms.
### Table of Line-Search Precision Values

<table>
<thead>
<tr>
<th>TECHNIQUE=</th>
<th>UPDATE=</th>
<th>LSPRECISION VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>QUANEW</td>
<td>DBFGS, BFGS</td>
<td>0.4</td>
</tr>
<tr>
<td>QUANEW</td>
<td>DDFP, DFP</td>
<td>0.06</td>
</tr>
<tr>
<td>CONGRA</td>
<td>all</td>
<td>0.1</td>
</tr>
<tr>
<td>NEWRAP</td>
<td>no update</td>
<td>0.9</td>
</tr>
</tbody>
</table>

**Range:** \( number > 0 \)

**MAXFUNC= number**

Specifies the maximum number of function calls in the optimization process. The objective function that is minimized is the negative log-likelihood.

**Default:**

- 125 for TRUREG, NRRIDG, and NEWRAP.
- 500 for QUANEW and DBLDOG.
- 1000 for CONGRA.

**Range:** \( number > 0 \)

**MAXITER= number**

Specifies the maximum number of iterations in the optimization process.

**Default:**

- 50 for TRUREG, NRRIDG and NEWRAP
- 200 for QUANEW and DBLDOG
- 400 for CONGRA

**Range:** \( number > 0 \)

**MAXSTEP= number**

Specifies the upper bound for the step length of the line-search algorithms.

**Default:** The largest double precision value

**Range:** \( number > 0 \)
MAXTIME= *number*

Specifies the upper limit of CPU time for the optimization process. It is measured in seconds.

| Default: | 7 days, that is, MAXTIME=604800 seconds |
| Range:   | *number* > 0 |

NOPRINT

Suppresses all output printed and only ERRORs, WARNINGs, and NOTEs are printed on the log file.

PALL

Prints all optional output except the output generated by the PSTDERR, LIST, or LISTCODE options.

PHISTORY

Prints the optimization history. If PSUMMARY or NOPRINT are not specified, then the PHISTORY option is set automatically. The iteration history is printed by default.

PSUMMARY

Restricts the amount of default printed output to a short form of iteration history and NOTEs, WARNINGs, and ERRORs.

RESTART= *number*

Specifies that the QUANEW or CONGRA algorithm is restarted with a steepest descent/ascent search direction after the *number* of iterations has been completed.

| Default: | For TECHNIQUE=CONGRA, and UPDATE= PB, restart is done automatically, so *number* is not used; For TECHNIQUE=CONGRA, and UPDATE not = PB, *number* is the number of parameters. For TECHNIQUE=QUANEW, *number* is the largest integer available. |
| Range:   | *number* > 1 |

SINGULAR= *number*

Specifies an absolute singularity criterion for the computation of the inertia of Hessian and cross-product Jacobian and their projected forms.

| Default: | 1E-8 |
| Range:   | *number* > 0 |

TECHNIQUE= *method*

where *method* is one of the following:
NONE
Specifies no method; no optimization is performed.

TRUREG
   Specifies the Trust-Region optimization technique.

NEWRAP
   Specifies the Newton-Raphson with Line Search optimization technique.

NRRIDG
   Specifies the Newton-Raphson with Ridging optimization technique. This is the default when the number of parameters to be estimated is $n \leq 40$.

DBLDOG
   Specifies the Double-Dogleg optimization technique.

QUANEW
   Specifies the quasi-Newton optimization technique. This is the default when the number of convergence parameters to be estimated is in the range: $40 < n \leq 400$.

CONGRA
   Specifies the Conjugate Gradient optimization technique. This is the default when the number of convergence parameters to be estimated is $n \geq 400$.

**Default:** The default technique is either NRRIDG, QUANEW, or CONGRA, depending on the value of the number of convergence parameters to be estimated.

See for more information.

**UPDATE=update-type**

where *update-type* is one of the following:

**BFGS**
   For TECHNIQUE=QUANEW, performs the BFGS (Broyden-Fletcher-Goldfarb-Shanno) update of the Cholesky factor of the Hessian matrix.

**CD**
   For TECHNIQUE=CONGRA, performs a conjugate descent update of Fletcher.

**DBFGS**
   For TECHNIQUE=DBLDOG or QUANEW, performs the dual BFGS (Broyden-Fletcher-Goldfarb-Shanno) update of the Cholesky factor of the Hessian matrix. This is the default for TECHNIQUE=QUANEW and DBLDOG.

**DDFP**
   For TECHNIQUE=DBLDOG or QUANEW, performs the dual DFP (Davidson-Fletcher-Powell) update of the Cholesky factor of the Hessian matrix.

**DFP**
   For TECHNIQUE=QUANEW, performs the original DFP (Davidson-Fletcher-Powell) update of the inverse Hessian matrix.
For TECHNIQUE=CONGRA, performs the Fletcher-Reeves update.

For TECHNIQUE=CONGRA, performs the automatic restart update method of Powell and Beale. This is the default for TECHNIQUE= CONGRA.

For TECHNIQUE=CONGRA, performs the Polak-Ribiere update.

**VERSION= 1 | 2| 3**

Specifies the version of the hybrid quasi-Newton optimization technique or the version of the quasi-Newton optimization technique with nonlinear constraints.

| Default: | 2 |

**XCONV= number**

Specifies the relative parameter convergence criterion.

| Default: | 1E-8 |
| Range: | $number > 0$ |

**XSIZE= number**

Specifies the number of successive iterations for which the criterion must be satisfied before the optimization process can be terminated.

| Default: | 0 |
| Range: | $number \geq 0$ |
The REMOTE statement is implemented in the NEURAL, DMREG, and DMVQ procedures in Enterprise Miner 4.1. You can use it to communicate with an MFC monitor (an external process on a Window client) to observe the progress of the iterative algorithm or to interrupt the iterative process. The monitor has a Graph tab and a Status tab as shown below:
The Graph tab displays the iteration history: objection function versus iteration number and maximum absolute gradient versus iteration number. Click [Stop Current] or [Stop All] to stop the current or all optimization process. The Status tab displays the objective function and the maximum absolute element of the gradient vectors for each iteration.

REMOTE remote-option(s);

Options

remote-options can be the following:

SOCKET=socket-reference

establishes a TCP/IP socket connection to an MFC monitor on the Window client to receive the report of the ongoing optimization. The socket reference contains the IP address and the port number and can be defined by using the following FILENAME statement:

FILENAME <socket-reference> SOCKET '<ip_address:portnum>';

where ip_address is the IP address of the Window client and portnum is the socket port number. The socket port number os any number that you use to invoke the MFC monitor.

PLOTFILE=fileref | ' external-file'

Specifies the external file that contains the iterative history (for example, the iteration number, the objective function, and the maximum absolute gradient). You can specify the path of the external file in quotes or you can use the FILENAME statement to specify a file reference. This option is obsolete if you can take advantage of the SOCKET= option.

STOPFILE

Specifies an external file that the iterative process will be terminated if this file exists. This is useful when you run a project with a large data set. To stop the process, you must create the external file. The DMREG procedure stops the iterative process when it detects this file. The file does not have to have any content. You can specify the path of an external file in quotes or use the LIBNAME statement to specify the file reference. This option is obsolete if you can take advantage of the SOCKET= option.

Example:

FILENAME abc SOCKET 'd6026.us.sas.com:12234';
PROC DMREG DATA=SAMPSIO.DMDCENS DMDBCAT=SAMPSIO.DMDCENS;
  REMOTE SOCKET=abs;
  CLASS CLASS WORKCLAS MARTAL OCCUPATN RELATION RACE SEX COUNTRY;
  MODEL CLASS=AGE FNLWGT EDUC_NUM CAP_GAIN CAP_LOSS HOURWEEK
  WORKCLAS MARITAL OCCUPATN RELATION RACE SEX COUNTRY
  / SELECTION=F CHOOSE=AIC;
RUN;

You can invoke the monitor any time by using the port number (1234) that you choose. After the socket connect is made you can see the display of the iteration history of the ongoing optimization.
The DMREG Procedure

SCORE Statement

Specifies options for scoring data.

SCORE scoring-option(s);

Options

Scoring-options can be the following:

ADDITIONALRESIDUALS

Specifies that the OUT= data set contains additional residuals such as: RS_ for logistic regressions and RS_, RT_, RD_, RDS_, RDT_ for normal error regression. See for more detail.

Alias: ADDRES | AR

ALPHA=number

Specifies the significance level \( p \) for the construction of 100(1-p)\% confidence interval for the posterior probabilities. This number must be between 0 and 1.

Default: 0.05

CLP

Specifies that the OUT= data set contains the confidence limits for the posterior probabilities. The significance level is controlled by the ALPHA= option.

DATA=<libref.> SAS-data-set

Specifies the input data set that contains inputs and optionally targets.

Default: The default is the same as the DATA= data set in the PROC statement.

DMDB | NODMDB

Specifies whether an explicit DATA= data set has been DMDB-encoded or if the data set contains raw data.

Default: If the DATA= option is not specified in the SCORE statement, the training data is used and the NODMDB option is invalid.

Caution: If the DATA= in the SCORE statement specifies a data set other than the training data, either DMDB or NODMDB must be specified in the SCORE statement.

OUT=<libref.> SAS-data-set
Specifies the output data set with outputs.

Default: DATAn

Names for computed variables are normally taken from the data dictionary. If necessary, names for these variables can be generated by concatenating a prefix to the name of the corresponding target variable according to the rules in the following tables:

**Statistics Generated in the OUT=SAS-data-set for Normal ErrorRegression**

<table>
<thead>
<tr>
<th>NAME</th>
<th>LABEL</th>
</tr>
</thead>
<tbody>
<tr>
<td>P_targetname</td>
<td>Predicted: targetname</td>
</tr>
<tr>
<td>E_targetname</td>
<td>Error Function: targetname</td>
</tr>
<tr>
<td>R_targetname</td>
<td>Residuals: targetname</td>
</tr>
<tr>
<td>RD_targetname</td>
<td>Deviance Residuals: targetname</td>
</tr>
</tbody>
</table>

If the target is declared as a categorical variable, the OUT=SAS-data-set also includes:

<table>
<thead>
<tr>
<th>NAME</th>
<th>LABEL</th>
</tr>
</thead>
<tbody>
<tr>
<td>F_targetname</td>
<td>From: targetname</td>
</tr>
<tr>
<td>I_targetname</td>
<td>Into: targetname</td>
</tr>
</tbody>
</table>

If the ADDITIONALRESIDUALS option is also specified, the OUT=SAS-data-set includes:
RS_targetname  Standardized
Residuals:

targetname

RT_targetname  Studentized
Residuals:

targetname

RDS_targetname  Standardized
Deviance
Residuals:

targetname

RDT_targetname  Studentized
Deviance
Residuals:

targetname

**Note:** In the table above, targetname is the name of the target variable. For example, if PURCHASE is the targetname, the predicted value statistic is named P_PURCHA and the studentized deviance residual is named RDT_PURC. [If the constructed names are longer than the maximum of eight characters allowed for SAS variable names, they are truncated to eight characters.] ■

*Statistics Generated in the OUT= SAS-data-set for
Binomial or Multinomial Regression*

<table>
<thead>
<tr>
<th>NAME</th>
<th>LABEL</th>
</tr>
</thead>
<tbody>
<tr>
<td>P_targetname&amp;value</td>
<td>Predicted: targetname=targetvalue</td>
</tr>
<tr>
<td>F_targetname</td>
<td>From: targetname</td>
</tr>
<tr>
<td>I_targetname</td>
<td>Into: targetname</td>
</tr>
<tr>
<td>E_targetname</td>
<td>Error Function:</td>
</tr>
<tr>
<td>R_targetname&amp;value</td>
<td>Residual: targetname=targetvalue</td>
</tr>
</tbody>
</table>
If the ADDITIONALRESIDUALS option is specified, the OUT=SAS-data-set includes:

\[
\text{RS\_targetname\&value Standardized Residual: targetname=targetvalue}
\]

**Note:** In the table above, targetname\&value is a combination of the target name (targetname) and target value (targetvalue). For example, if PURCHASE is the targetname and "YES" and "NO" are the two values possible for targetvalue, the predicted value statistics are named P_PURYES and P_PURNOS.

**OUTFIT= <libref.>SAS-data-set**

Specifies the output data set with fit statistics. For more information, see .

**OUTSTEP**

Scores the data for each model selection step.

**ROLE=role-value**

Specifies the role of the DATA= data set. The ROLE= option primarily affects which fit statistics are computed and what their names and labels are.

**Role-value** can be:

**TRAIN**

This value is the default when the same data set name is used in the DATA= option in both the PROC and SCORE statements. Specifying TRAIN with any data set other than the actual training set is an error.

**VALID | VALIDATION**

This value is the default when the DATA= data set name in the SCORE statement is the same as the data set in the VALIDDATA= in the PROC statement.

**TEST**

This value is the default when the DATA= data set name in the SCORE statement is the same as the data set name in the TESTDATA= option of the PROC statement.

**SCORE**

Predicted values are produced but residuals, error functions, and other fit statistics are not produced.
### Details

#### Input

The input to the DMREG procedure can be assigned one of these roles:

**Training**
- The DATA= data set is used to fit the initial model.

**Validation**
- The VALIDATA= data set is used to compute assessment statistics and to fine-tune the model during stepwise selection.

**Test**
- The TESTDATA= data set is an additional "hold out" data set that you can use to compute assessment statistics.

**Score**
- The DATA= data set in the SCORE statement is used for predicting target values for a new data set that may not contain the target.

#### Specification of Effects

Different types of effects can be used in the DMREG procedure. In the following list, assume that A, B, and C are class variables and that X1, X2, and Y are continuous variables:

1. **Regressor effects** are specified by writing continuous variables individually:
   - X1 X2
2. **Polynomial effects** are specified by joining two or more continuous variables with asterisks:
   - X1*X1 X1*X2
3. **Main effects** are specified by writing class variables individually:
   - AC
4. **Crossed effects** (interactions) are specified by joining class variables with asterisks:
   - A*BB*CA*B*C
5. **Continuous-by-class effects** are written by joining continuous variables and class variables with asterisks:
   - X1*A.

**Note:** Nested effects are not supported.
Optimization Methods

The following table provides a list of the general nonlinear optimization methods and the default maximum number of iterations and function calls for each method.

*Optimization Methods for the Regression node.*

<table>
<thead>
<tr>
<th>Optimization Method</th>
<th>Maximum Iterations</th>
<th>Maximum Function Calls</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conjugate Gradient</td>
<td>400</td>
<td>1000</td>
</tr>
<tr>
<td>Double Dogleg</td>
<td>200</td>
<td>500</td>
</tr>
<tr>
<td>Newton-Raphson with Line Search</td>
<td>50</td>
<td>125</td>
</tr>
<tr>
<td>Newton-Raphson with Ridging</td>
<td>50</td>
<td>125</td>
</tr>
<tr>
<td>Quasi-Newton</td>
<td>200</td>
<td>500</td>
</tr>
<tr>
<td>Trust-Region</td>
<td>50</td>
<td>125</td>
</tr>
</tbody>
</table>

You should set the optimization method based on the size of the data mining problem, as follows:

1. Small-to-medium problems - The Trust-Region, Newton-Raphson with Ridging, and Newton-Raphson with Line Search methods are appropriate for small and medium sized optimization problems (number of model parameters up to 40) where the Hessian matrix is easy and cheap to compute. Sometimes, Newton-Raphson with Ridging can be faster than Trust-Region, but Trust-Region is numerically more stable. If the Hessian matrix is not singular at the optimum, then the Newton-Raphson with Line Search can be a very competitive method.

2. Medium Problems - The quasi-Newton and Double Dogleg methods are appropriate for medium optimization problems (number of model parameters up to 400) where the objective function and the gradient are must faster to compute than the Hessian. Quasi-Newton and Double Dogleg require more iterations than does the Trust-Region or the Newton-Raphson methods, but each iteration is much faster.

3. Large Problems - The Conjugate Gradient method is appropriate for large data mining problems (number of model parameters greater than 400) where the objective function and the gradient are much faster to compute than the Hessian matrix, and where they need too much memory to store the approximate Hessian matrix.

Note: To learn about these optimization methods, see the *SAS/OR Technical Report: The NLP Procedure* (1997).
If the number of parameters is less than or equal to 40, then the default method is set to Newton-Raphson with Ridging. If the number of parameters is greater than 40 and less than 400, then the default method is set to quasi-Newton. If the number of parameters is greater than 400, then Conjugate Gradient is the default method.

---

**Fit Statistics for OUTFIT= Data Sets**

The OUTFIT= data set in the PROC DMREG statement contains fit statistics for the training, test, and/or validation data. Depending on the ROLE= option in the SCORE statement, the OUTFIT= data set contains fit statistics for either the training, test, or validation data.

---

**Fit Statistics for the Training Data**

<table>
<thead>
<tr>
<th>Fit Statistic</th>
<th>Training Data</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>AIC</em></td>
<td>Train: Akaike's Information Criterion</td>
</tr>
<tr>
<td><em>ASE</em></td>
<td>Train: Average Squared Error</td>
</tr>
<tr>
<td><em>AVERR</em></td>
<td>Train: Average Error Function</td>
</tr>
<tr>
<td><em>DFE</em></td>
<td>Train: Degrees of Freedom for Error</td>
</tr>
<tr>
<td><em>DFM</em></td>
<td>Train: Model Degrees of Freedom</td>
</tr>
<tr>
<td><em>DFT</em></td>
<td>Train: Total Degrees of Freedom</td>
</tr>
<tr>
<td><em>DIV</em></td>
<td>Train: Divisor for ASE</td>
</tr>
<tr>
<td><em>ERR</em></td>
<td>Train: Error Function</td>
</tr>
<tr>
<td><em>FPE</em></td>
<td>Train: Final Prediction Error</td>
</tr>
</tbody>
</table>
### Fit Statistics for the Test Data

<table>
<thead>
<tr>
<th>Fit Statistic</th>
<th>Test Data</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>TASE</em></td>
<td>Test: Average Squared Error</td>
</tr>
</tbody>
</table>
_TASEL_  Test: Lower 95% Confidence Limit for TASE

_TASEU_  Test: Upper 95% Confidence Limit for TASE

_TAVERR_ Test: Average Error Function

_TDIV_ Test: Divisor for TASE

_TERR_ Test: Error Function

_TMAX_ Test: Maximum Absolute Error

_TMSE_ Test: Mean Square Error

_TNOBS_ Test: Sum of Frequencies

_TRASE_ Test: Root Average Squared Error

_TRMSE_ Test: Root Mean Square Error

_TSSE_ Test: Sum of Square Errors

_TSUMW_ Test: Sum of Case Weights Times Frequency

_TMISC_ Test: Misclassification Rate
Test: Lower 95% Confidence Limit for TMISC

Test: Upper 95% Confidence Limit for TMISC

### Fit Statistics for the Validation Data

<table>
<thead>
<tr>
<th>Fit Statistic</th>
<th>Validation Data</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>VASE</em></td>
<td>Valid: Average Squared Error</td>
</tr>
<tr>
<td><em>VAVERR</em></td>
<td>Valid: Average Error Function</td>
</tr>
<tr>
<td><em>VDIV</em></td>
<td>Valid: Divisor for VASE</td>
</tr>
<tr>
<td><em>VERR</em></td>
<td>Valid: Error Function</td>
</tr>
<tr>
<td><em>VMAX</em></td>
<td>Valid: Maximum Absolute Error</td>
</tr>
<tr>
<td><em>VMSE</em></td>
<td>Valid: Mean Square Error</td>
</tr>
<tr>
<td><em>VNOBS</em></td>
<td>Valid: Sum of Frequencies</td>
</tr>
<tr>
<td><em>VRASE</em></td>
<td>Valid: Root Average Squared Error</td>
</tr>
<tr>
<td><em>VRMSE</em></td>
<td>Valid: Root Mean Square Error</td>
</tr>
<tr>
<td><em>VSSE</em></td>
<td>Valid: Sum of Square Errors</td>
</tr>
</tbody>
</table>
Examples

The following examples were executed using the HP-UX version 10.20 operating system and the SAS software release 6.12TS045.

**Example 1: Linear and Quadratic Logistic Regression with an Ordinal Target (Rings Data)**

**Example 2: Performing a Stepwise OLS Regression (DMREG Baseball Data)**

**Example 3: Comparison of the DMREG and LOGISTIC Procedures when Using a Categorical Input Variable**

Copyright 2000 by SAS Institute Inc., Cary, NC, USA. All rights reserved.
Example 1: Linear and Quadratic Logistic Regression with an Ordinal Target (Rings Data)

This example demonstrates how to perform both a linear and a quadratic logistic regression with an ordinal target. The example DMDB training data set SAMPSIO.DMDRING contains an ordinal target with 3 levels (C=0, 1, or 2) and two continuous inputs (X and Y). There are 180 observations in the data set. The SAMPSIO.DMSRING data set is scored using the scoring formula from the trained models. Both data sets are stored in the sample library.

### Linear-Logistic Program

```plaintext
proc gplot data=sampsio.dmdring;
    plot y*x=c /haxis=axis1 vaxis=axis2;
    symbol c=black i=none v=dot;
    symbol2 c=red i=none v=square;
    symbol3 c=green i=none v=triangle;
    axis1 c=black width=2.5 order=(0 to 30 by 5);
    axis2 c=black width=2.5 minor=none order=(0 to 20 by 2);
    title 'Plot of the Rings Training Data';
run;
```

```plaintext
proc dmreg data=sampsio.dmdring dmdbcat=sampsio.dmdring;
    class c;
    model c = x y;
    score out=out outfit=fit;
    score data=sampsio.dmsring nodmdb out=gridout;
    title 'Linear-Logistic Regression with Ordinal Target';
run;
```

```plaintext
proc print data=fit noobs label;
    var _aic_ _max_ _rfpe_ _misc_;
    title2 'Fit Statistics for the Training Data Set';
run;
```

```plaintext
proc freq data=out;
    tables f_c*i_c;
    title2 'Misclassification Table: Training Data';
run;
```
proc gplot data=out;
  plot y*x=i_c / haxis=axis1 vaxis=axis2;
  symbol c=black i=none v=dot;
  symbol2 c=red i=none v=square;
  symbol3 c=green i=none v=triangle;
  axis1 c=black width=2.5 order=(0 to 30 by 5);
  axis2 c=black width=2.5 minor=none order=(0 to 20 by 2);
  title2 'Classification Results';
run;

proc gcontour data=gridout;
  plot y*x=p_c1 / pattern ctext=black coutline=gray;
  plot y*x=p_c2 / pattern ctext=black coutline=gray;
  plot y*x=p_c3 / pattern ctext=black coutline=gray;
  title2 'Posterior Probabilities';
  pattern v=msolid;
  legend frame;
run;

---

Linear-Logistic Output

PROC GPlot Plot of the Rings Training Data

DMREG Summary Profile Information

PROC DMREG first lists background information about the fitting of the linear-logistic model. Included are the name of the input data set, the response variable, the number of response levels, the number of observations used, the error distribution, and the link function.
**DMREG Response Profile**

The Response Profile table lists the target categories, their ordered values, and their total frequencies for the given data.

<table>
<thead>
<tr>
<th>Target Profile</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Ordered Value</td>
<td>c</td>
<td>Total Frequency</td>
</tr>
<tr>
<td>1</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>59</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>113</td>
<td></td>
</tr>
</tbody>
</table>
Newton-Raphson Ridge Optimization
Without Parameter Scaling

Parameter Estimates  4

Optimization Start

Active Constraints  0  Objective Function  143.32716812
Max Abs Gradient Element  2.5948717949

<table>
<thead>
<tr>
<th>Iter</th>
<th>Restarts</th>
<th>Function Calls</th>
<th>Active Constraints</th>
<th>Objective Function</th>
<th>Objective Function Change</th>
<th>Max Abs Gradient Element</th>
<th>Ridge</th>
<th>Predicted Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>2</td>
<td>0</td>
<td>142.85145</td>
<td>0.4757</td>
<td>0.1180</td>
<td>0</td>
<td>1.001</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>3</td>
<td>0</td>
<td>142.85125</td>
<td>0.000198</td>
<td>0.000051</td>
<td>0</td>
<td>1.000</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>4</td>
<td>0</td>
<td>142.85125</td>
<td>6.43E-11</td>
<td>1.26E-11</td>
<td>0</td>
<td>1.001</td>
</tr>
</tbody>
</table>

GCONV convergence criterion satisfied.

### DMREG Model Fitting Information and Testing Global Null Hypothesis Beta=0

The Model Fitting Information and Testing Global Null Hypothesis Beta=0 table contains the negative of twice the log likelihood (-2 LOG L) for the fitted model. Results of the likelihood ratio test and the efficient score test for testing the joint significance of the explanatory inputs are also printed in the table.

<table>
<thead>
<tr>
<th></th>
<th>Intercept Only</th>
<th>Intercept and Covariates</th>
<th>Chi-Square for Covariates</th>
</tr>
</thead>
<tbody>
<tr>
<td>-2 LOG L</td>
<td>286.654</td>
<td>285.703</td>
<td>0.952 with 2 DF (p=0.6213)</td>
</tr>
</tbody>
</table>

### DMREG Analysis of Maximum Likelihood Estimates

The Analysis of Maximum Likelihood Estimates table lists the parameter estimates, their standard errors, and the results of the Wald test for the individual parameters. A standardized estimate for each slope parameter and the odds ratio for each estimate is also printed. An odds ratio is obtained by exponentiating the corresponding parameter estimate.
DMREG Odds Ratio Estimates

The Odd Ratio Estimates table lists the odd ratios for the explanatory inputs. The odd ratio estimates provide the change in odds for a unit increase in each input.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>DF</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>Wald Chi-square</th>
<th>Pr &gt; Chi-square</th>
<th>Standardized Estimate</th>
<th>exp(Est)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>-3.0449</td>
<td>0.5918</td>
<td>26.47</td>
<td>&lt;.0001</td>
<td>.</td>
<td>0.048</td>
</tr>
<tr>
<td>Intercept</td>
<td>2</td>
<td>-0.4945</td>
<td>0.4931</td>
<td>1.01</td>
<td>0.3159</td>
<td>.</td>
<td>0.610</td>
</tr>
<tr>
<td>x</td>
<td>1</td>
<td>-0.0156</td>
<td>0.0222</td>
<td>0.56</td>
<td>0.4561</td>
<td>-0.063289</td>
<td>0.984</td>
</tr>
<tr>
<td>y</td>
<td>1</td>
<td>0.0231</td>
<td>0.0367</td>
<td>0.39</td>
<td>0.5298</td>
<td>0.053024</td>
<td>1.023</td>
</tr>
</tbody>
</table>

**DMREG Odds Ratio Estimates**

The Odd Ratio Estimates table lists the odd ratios for the explanatory inputs. The odd ratio estimates provide the change in odds for a unit increase in each input.

**Linear-Logistic Regression with Ordinal Target**

**The DMREG Procedure**

**Odds Ratio Estimates**

<table>
<thead>
<tr>
<th>Input</th>
<th>Odds Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>0.984</td>
</tr>
<tr>
<td>y</td>
<td>1.023</td>
</tr>
</tbody>
</table>

**PROC PRINT Report of Selected Fit Statistics for the Training Data Set**

The misclassification rate for the training data set is only 37.22%.

**Linear-Logistic Regression with Ordinal Target**

**Fit Statistics for the Training Data Set**

<table>
<thead>
<tr>
<th>Train: Akaike's Information Criterion</th>
<th>Train: Maximum Absolute Error</th>
<th>Train: Final Prediction Error</th>
<th>Train: Root Mean Square Error</th>
<th>Train: Misclassification Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>293.703</td>
<td>0.95656</td>
<td>0.41066</td>
<td>0.37222</td>
<td></td>
</tr>
</tbody>
</table>

**PROC FREQ Misclassification Table for the Training Data**

All observations in the training data are classified into the C=3 level. The linear model is not adequate.
**Linear-Logistic Regression with Ordinal Target**

*Misclassification Table: Training Data*

**The FREQ Procedure**

**Table of F_c by l_c**

<table>
<thead>
<tr>
<th>F_c(From: c)</th>
<th>l_c(Into: c)</th>
<th>Frequency</th>
<th>Percent</th>
<th>Row Pct</th>
<th>Col Pct</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3</td>
<td>Total</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>8</td>
<td>4.44</td>
<td>4.44</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td>59</td>
<td>32.78</td>
<td>32.78</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td>113</td>
<td>62.78</td>
<td>62.78</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td>180</td>
<td>100.00</td>
<td>100.00</td>
</tr>
</tbody>
</table>

**PROC GPLOT Plot of the Classification Results**

The target classes are not linearly separable.
PROC GCONTOUR Plots of the Posterior Probabilities
Quadratic-Logistic Program
proc dmreg data=sampsio.dmdring dmdbcat=sampsio.dmdring;
class c;
model c=x|x|y|y @2;
score out=qout outfit=qfit;
score data=sampsio.dmsring nodmdb out=qgridout;
title1 'Quadratic-Logistic Regression with Ordinal Target';
run;

proc print data=qfit noobs label;
  var _aic_ _max_ _rfpe_ _misc_;
title2 'Fit Statistics for the Training Data Set';
run;

proc freq data=qout;
tables f_c*i_c;
title2 'Misclassification Table: Training Data';
run;

proc gplot data=qout;
  plot y*x=i_c / haxis=axis1 vaxis=axis2;
symbol1 c=black i=none v=dot;
symbol2 c=red i=none v=square;
symbol3 c=green i=none v=triangle;
axis1 c=black width=2.5 order=(0 to 30 by 5);
axis2 c=black width=2.5 minor=none order=(0 to 20 by 2);
title2 'Classification Results';
run;

proc gcontour data=qgridout;
  plot y*x=p_c1 / pattern ctext=black coutline=gray;
  plot y*x=p_c2 / pattern ctext=black coutline=gray;
  plot y*x=p_c3 / pattern ctext=black coutline=gray;
title2 'Posterior Probabilities';
  pattern v=msolid;
  legend frame;
run;

---

**Quadratic-Logistic Output**

**DMREG Output**
Quadratic-Logistic Regression with Ordinal Target

The DMREG Procedure

Training Data Set: SAMP1O.DMDRING
DMDB Catalog: SAMP1O.DMDRING
Target Variable: c
Target Measurement Level: Ordinal
Number of Target Categories: 3
Error: MBernoulli
Link Function: Logit
Number of Model Parameters: 7
Number of Observations: 180

Target Profile

<table>
<thead>
<tr>
<th>Ordered Value</th>
<th>c</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>59</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>113</td>
<td></td>
</tr>
</tbody>
</table>

Newton-Raphson Ridge Optimization

Without Parameter Scaling

Parameter Estimates 7

Optimization Start

Active Constraints 0  Objective Function 143.32716812
Max Abs Gradient Element 8.4022792023

<table>
<thead>
<tr>
<th>Iter</th>
<th>Restarts</th>
<th>Function Calls</th>
<th>Active Constraints</th>
<th>Objective Function</th>
<th>Objective Function Change</th>
<th>Max Abs Gradient Element</th>
<th>Ridge</th>
<th>Predicted Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>50.31687</td>
<td>33.0103</td>
<td>8.4596</td>
<td>0</td>
<td>1.095</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>26.02018</td>
<td>24.2957</td>
<td>2.9573</td>
<td>0</td>
<td>1.292</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>13.98860</td>
<td>12.0316</td>
<td>1.1136</td>
<td>0</td>
<td>1.306</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>7.43545</td>
<td>6.5532</td>
<td>0.5352</td>
<td>0</td>
<td>1.312</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>6</td>
<td>0</td>
<td>3.75004</td>
<td>3.6854</td>
<td>0.3031</td>
<td>0</td>
<td>1.310</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>7</td>
<td>0</td>
<td>1.75436</td>
<td>1.9957</td>
<td>0.1649</td>
<td>0</td>
<td>1.301</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>8</td>
<td>0</td>
<td>0.75995</td>
<td>0.9944</td>
<td>0.0776</td>
<td>0</td>
<td>1.290</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>9</td>
<td>0</td>
<td>0.30934</td>
<td>0.4506</td>
<td>0.0323</td>
<td>0</td>
<td>1.280</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>10</td>
<td>0</td>
<td>0.12182</td>
<td>0.1875</td>
<td>0.0125</td>
<td>0</td>
<td>1.274</td>
</tr>
</tbody>
</table>

Optimization Results

Iterations 9 Function Calls 12
Hessian Calls 11 Active Constraints 0
Objective Function 0.121823985 Max Abs Gradient Element 0.0125231148
Ridge 0 Actual Over Pred Change 1.2737233683

ABSCONV convergence criterion satisfied.

NOTE: At least one element of the (projected) gradient is greater than 1e-3.
PROC PRINT Report of Selected Fit Statistics for the Training Data

Note that the training misclassification rate is 0. All cases are correctly classified by the quadratic-logistic model.

-2 LOG L  286.654  0.244  286.411 with 5 DF (p<.0001)

Analysis of Maximum Likelihood Estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>DF</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>Wald Chi-square</th>
<th>Pr &gt; Chi-square</th>
<th>Standardized Estimate</th>
<th>exp(Est)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>-232.2</td>
<td>236.4</td>
<td>0.96</td>
<td>0.3260</td>
<td>.</td>
<td>0.000</td>
</tr>
<tr>
<td>Intercept</td>
<td>2</td>
<td>-197.4</td>
<td>188.0</td>
<td>1.10</td>
<td>0.2939</td>
<td>.</td>
<td>0.000</td>
</tr>
<tr>
<td>x</td>
<td>1</td>
<td>18.4037</td>
<td>16.3812</td>
<td>1.27</td>
<td>0.2590</td>
<td>70.701785</td>
<td>999.000</td>
</tr>
<tr>
<td>x*x</td>
<td>1</td>
<td>-0.6584</td>
<td>0.4552</td>
<td>2.09</td>
<td>0.1481</td>
<td>.</td>
<td>0.518</td>
</tr>
<tr>
<td>y</td>
<td>1</td>
<td>22.5413</td>
<td>31.5097</td>
<td>0.51</td>
<td>0.4744</td>
<td>51.803360</td>
<td>999.000</td>
</tr>
<tr>
<td>x*y</td>
<td>1</td>
<td>0.0469</td>
<td>0.7657</td>
<td>0.00</td>
<td>0.9512</td>
<td>.</td>
<td>1.048</td>
</tr>
<tr>
<td>y*y</td>
<td>1</td>
<td>-1.2320</td>
<td>1.3963</td>
<td>0.78</td>
<td>0.3776</td>
<td>.</td>
<td>0.292</td>
</tr>
</tbody>
</table>

PROC PRINT Report of Selected Fit Statistics for the Training Data

Note that the training misclassification rate is 0. All cases are correctly classified by the quadratic-logistic model.

Quadratic-Logistic Regression with Ordinal Target
Fits Statistics for the Training Data Set

<table>
<thead>
<tr>
<th>Train: Akaike’s Information Criterion</th>
<th>Train: Maximum Absolute Error</th>
<th>Train: Final Prediction Error</th>
<th>Train: Misclassification Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>14.2436</td>
<td>0.023582</td>
<td>.002403151</td>
<td>0</td>
</tr>
</tbody>
</table>

PROC FREQ Misclassification Table for the Training Data
### Quadratic-Logistic Regression with Ordinal Target

**Misclassification Table: Training Data**

The FREQ Procedure

**Table of F_c by l_c**

<table>
<thead>
<tr>
<th>F_c</th>
<th>l_c</th>
<th>Frequency</th>
<th>Percent</th>
<th>Row Pct</th>
<th>Col Pct</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>8</td>
<td>4.44</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>8</td>
<td>4.44</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>59</td>
<td>32.78</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>59</td>
<td>32.78</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>113</td>
<td>62.78</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>113</td>
<td>62.78</td>
<td>100.00</td>
<td>100.00</td>
</tr>
</tbody>
</table>

Total | 180 | 100.00 |

**PROC GPLOT Plot of the Classification Results**
PROC GCONTOUR Plots of the Posterior Probabilities

Quadratic—Logistic Regression with Ordinal Target
Posterior Probabilities

Predicted: c=1

Predicted: c=2
PROC GPLOT creates a scatter plot of the Rings training data.

``` Sas 
proc gplot data=sampsio.dmdring;
   plot y*x=c /haxis=axis1 vaxis=axis2;
   symbol c=black i=none v=dot;
   symbol2 c=red i=none v=square;
   symbol3 c=green i=none v=triangle;
   axis1 c=black width=2.5 order=(0 to 30 by 5);
   axis2 c=black width=2.5 minor=none order=(0 to 20 by 2);
   title 'Plot of the Rings Training Data';
run;
```

The PROC DMREG statement invokes the procedure. The DATA= option identifies the DMDB encoded training data set that is used to fit the model. The DMDBCAT= option identifies the DMDB training data catalog. You can create DMDB encoded data sets and catalogs with the DMDB procedure.

```
proc dmreg data=sampsio.dmdring dmdbcat=sampsio.dmdring;
```
The CLASS statement identifies the target C as a categorical variable.

class c;
The MODEL statement specifies the linear-logistic model.

model c = x y;
The SCORE statement scores the training data set and outputs fit statistics to the OUTFIT= data set. A note is printed in the log that indicates the training data set is scored when the DATA= option is omitted.

```
score out=out outfit=fit;
```
The second SCORE statement scores the SAMPSIO.DMSRING data set. The NODMDB option specifies that the score data set contains raw values instead of DMDB encoded data.

```
score data=sampsio.dmsring nodmdb out=gridout;
   title 'Linear-Logistic Regression with Ordinal Target';
run;
```
PROC PRINT report of selected fit statistics for the training data.

proc print data=fit noobs label;
    var _aic_ _max_ _rfpe_ _misc_ ;
    title2 'Fit Statistics for the Training Data Set';
run;
PROC FREQ report of the misclassification rate for the training data set. The F_C variable is the actual target value for each case and the I_C variable is the target value into which the case is classified.

```
proc freq data=out;
  tables f_c*i_c;
  title2 'Misclassification Table: Training Data';
run;
```
PROC GPLOT produces a plot of the classification results for the training data.

proc gplot data=out;
  plot y*x=i_c / haxis=axis1 vaxis=axis2;
  symbol  c=black i=none v=dot;
  symbol2 c=red i=none v=square;
  symbol3 c=green i=none v=triangle;
  axis1 c=black width=2.5 order=(0 to 30 by 5);
  axis2 c=black width=2.5 minor=none order=(0 to 20 by 2);
  title2 'Classification Results';
run;
PROC GCONTOUR produces plots of the posterior probabilities.

```sas
proc gcontour data=gridout;
  plot y*x=p_c1 / pattern ctext=black coutline=gray;
  plot y*x=p_c2 / pattern ctext=black coutline=gray;
  plot y*x=p_c3 / pattern ctext=black coutline=gray;
  title2 'Posterior Probabilities';
  pattern v=msolid;
  legend frame;
run;
```
The model statement specifies the quadratic-logistic model. The vertical bars indicate that interactions of the specified inputs should be generated. "@2" indicates that only interactions up to the second order should be used.

```sas
proc dmreg data=sampsio.dmdring dmdbcat=sampsio.dmdring;
  class c;
  model c=x|x|y|y @2;
  score out=qout outfit=qfit;
  score data=sampsio.dmsring nodmdb out=qgridout;
  title1 'Quadratic-Logistic Regression with Ordinal Target';
run;
```
PROC PRINT produces a report of selected fit statistics for the training data.

proc print data=qfit noobs label;
  var _aic_ _max_ _rfpe_ _misc_;  
  title2 'Fit Statistics for the Training Data Set';
run;
PROC FREQ creates a report of the misclassification matrix for the training data set.

```sas
proc freq data=qout;
   tables f_c*i_c;
   title2 'Misclassification Table: Training Data';
run;
```
PROC GPLOT plots the classification results for the training data set.

```
proc gplot data=qout;
    plot y*x=i_c / haxis=axis1 vaxis=axis2;
    symbol  c=black i=none v=dot;
    symbol2 c=red   i=none v=square;
    symbol3 c=green i=none v=triangle;
    axis1 c=black width=2.5 order=(0 to 30 by 5);
    axis2 c=black width=2.5 minor=none order=(0 to 20 by 2);
    title2 'Classification Results';
run;
```
PROC GCONTOUR plots the posterior probabilities.

```
proc gcontour data=qgridout;
    plot y*x=p_c1 / pattern ctext=black coutline=gray;
    plot y*x=p_c2 / pattern ctext=black coutline=gray;;
    plot y*x=p_c3 / pattern ctext=black coutline=gray;;
    title2 'Posterior Probabilities';
    pattern v=msolid;
    legend frame;
run;
```
Example 2: Performing a Stepwise OLS Regression (DMREG Baseball Data)

This example demonstrates how to perform a stepwise OLS regression using the DMREG procedure. The example DMDB training data set SAMSPIO.DMBASE (baseball data set) contains performance measures and salary levels for regular hitters and leading substitute hitters in major league baseball for the year 1986 (Collier 1987). There is one observation per hitter. The continuous response variable is the log of the players salary (logsalar). The SAMSPIO.DMTBASE data set is a test data set which is scored using the scoring formula from the trained model. The SAMSPIO.DMBASE and SAMSPIO.DMTBASE data sets and the SAMSPIO.DMDBASE data mining catalog are stored in the sample library.

Program

```
proc dmreg data=sampsio.dmdbase dmdbcat=sampsio.dmdbase
   testdata=sampsio.dmtbase outest=regest;
   class league division position;
   model logsalar = no_atbat no_hits no_home no_runs no_rbi no_bb
                  yr_major cr_atbat cr_hits cr_home cr_runs
                  cr_rbi cr_bb league division position no_outs
                  no_assts no_error
                  / error=normal
                  choose=sbc
                  selection=stepwise
                  slentry=0.25 slstay=0.25;
   score data=sampsio.dmtbase nodmdb
      out=regout(rename=(p_logsal=predict r_logsal=residual));
   title 'Output from the DMREG Procedure';
run;
```
**Design Matrix For Classification Effects**

The DMREG procedure uses a deviation from the means method to generate the design matrix for the classification inputs. Each row of the design matrix is generated by a unique combination of the nominal input values. Each column of the design matrix corresponds to a model parameter.

If a nominal variable SWING has \( k \) levels (3), then its main effect has \( k-1 \) (2) degrees of freedom, and the design matrix has \( k-1 \) (2) columns that correspond to the first \( k-1 \) levels. The ith column contains a 1 in the ith row, a -1 in the last row, and 0 everywhere else. If \( \alpha_i \) denotes the parameter that corresponds to the ith level of variable SWING, then \( k-1 \) columns yield estimates of the independent parameter \( \alpha_1, \alpha_2, \ldots, \alpha_{k-1} \). The last parameter is not needed because DMREG constrains the \( k \) parameters to sum to 0. Crossed effects, such as SWING*LEAGUE, are formed by the horizontal direct product of main effects.

---

**Output**

### Summary Profile Information

The first section of the output lists the two-level data set name, the response variable, the number of observations, the error distribution, and the link function.

Output from the DMREG Procedure

<table>
<thead>
<tr>
<th>The DMREG Procedure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training Data Set:</td>
</tr>
<tr>
<td>DMDB Catalog:</td>
</tr>
<tr>
<td>Target Variable:</td>
</tr>
<tr>
<td>Target Measurement Level:</td>
</tr>
<tr>
<td>Error:</td>
</tr>
<tr>
<td>Link Function:</td>
</tr>
<tr>
<td>Number of Model Parameters:</td>
</tr>
<tr>
<td>Number of Observations:</td>
</tr>
</tbody>
</table>

---
Data
Levels
for
SWING

<table>
<thead>
<tr>
<th></th>
<th>Design</th>
<th>Columns</th>
</tr>
</thead>
<tbody>
<tr>
<td>Left</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Right</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Switch</td>
<td>-1</td>
<td>-1</td>
</tr>
</tbody>
</table>

The printing of the design matrix can be suppressed by using the MODEL statement option NODESIGNPRINT.

\[
\text{Input Class Level Information}
\]

\[
\text{Design Variables}
\]

<table>
<thead>
<tr>
<th>Class</th>
<th>Value</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
</tr>
</thead>
<tbody>
<tr>
<td>league</td>
<td>American</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>National</td>
<td>-1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>division</td>
<td>East</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>West</td>
<td>-1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>position</td>
<td>1B</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2B</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2S</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3B</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3D</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3S</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>C</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>CD</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>CF</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>DH</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>DO</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>LF</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>O1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>OF</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>OS</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>RF</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>S3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SS</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>UT</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
</tr>
</tbody>
</table>
Model Fitting Information for Each Subset Model of the Stepwise Selection Process

For brevity, only steps number 5 and 8 from the stepwise selection process are listed in the following output. Step number 5 contains the model that has the smallest SBC statistic. This model is used to score the test data set. Because no other inputs met the condition for removal from the model and no other variables met the criterion for addition to the model, the stepwise algorithm terminates after step number 8.

For each model subset of the stepwise modeling process, DMREG provides:

1. An analysis of variance table which lists degrees of freedom, sums of squares, mean squares, the Model F, and its associated p-value.
2. Model fitting information which contains the following statistics that enable you to assess the fit of each stepwise model:
   - **R-square** - which is calculated as $1 - \frac{\text{SSE}}{\text{SST}}$, where SSE is the error sums of squares and SST is the total sums of squares. The $R^2$ statistic ranges from 0 to 1. Models that have large values of $R^2$ are preferred. For step number 8, the regression equation explains 60.17% of the variability in the target.
   - **Adj R-sq** - the Adj-$R^2$ is an alternative criterion to the $R^2$ statistic that is adjusted for the number of parameters in the model. This statistic is calculated as $1 - \left( \frac{(n - i) \left(1 - R^2\right)}{(n - p)} \right)$, where n is the number cases, and i is an indicator variable that is 1 if the model includes an intercept and 0, otherwise. Large differences between the $R^2$ and the Adj-$R^2$ values for a given model can indicate that you have used too many inputs in the model.
   - **AIC** - Akaike's Information Criterion, which is a goodness-of-fit statistic that you can use to compare one model to another. Lower values indicate a more desirable model. It is calculated as $(n) \ln \left(\frac{\text{SSE}}{n}\right) + 2p$, where n is the number of cases, SSE is the error sums of squares, and p is the number of model parameters.
   - **BIC** - Bayesian Information Criterion is another goodness-of-fit statistic that is calculated as $(n) \ln \left(\frac{\text{SSE}}{n}\right) + (p + 2) q - 2q^2$, where $q = \frac{\text{MSE}}{(\text{SSE}/n)}$ (MSE is obtained from the full model). Smaller BIC values are preferred.
   - **SBC** -Schwarz's Bayesian Criterion is another goodness-of-fit statistic that is calculated as $(n) \ln \left(\frac{\text{SSE}}{n}\right) + (p) \ln (n)$ Models that have small SBC values are preferred. Because the CHOOSE=SBC option was specified, DMREG selects the model that has the smallest SBC value.
   - **C(p)** - Mallow's Cp Statistic enables you to determine if your model is under or overspecified. This statistic is calculated as $\left( \frac{\text{SSE}(p)}{\text{MSE}} \right) - (n - 2p)$, where SSE(p) is the error sums of squares for the subset model with p parameters including the intercept if any, MSE is the error mean square for the full model, and n is the number of cases. For any subset model $C(p) > p$, there is evidence of bias due to an incompletely specified model (your model may not contain enough inputs). However, if there are values of $C(p) < p$, the full model is said to be overspecified. When the right model is chosen, the parameter estimates are unbiased, and this is reflected in $Cp < p$ or at least near p.

3. Analysis of effects and parameter estimates that contains the effect, degrees of freedom, parameter estimate, standard error, type II sums of squares, F-value and the corresponding p-value.
Step 5. Effect no_bb entered:

### Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>5</td>
<td>72.833938</td>
<td>14.566768</td>
<td>43.41</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>157</td>
<td>52.678742</td>
<td>0.335533</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>Corrected Total</td>
<td>162</td>
<td>125.512680</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
</tbody>
</table>

### Model Fitting Information

- R-square: 0.5803
- Adj R-sq: 0.5669
- AIC: -172.1147
- BIC: -169.6060
- SBC: -153.5522
- C(p): 5.3257

### Type III Analysis of Effects

<table>
<thead>
<tr>
<th>Effect</th>
<th>DF</th>
<th>Type III SS</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>no_hits</td>
<td>1</td>
<td>5.6348</td>
<td>16.7935</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>no_bb</td>
<td>1</td>
<td>1.7104</td>
<td>5.0976</td>
<td>0.0253</td>
</tr>
<tr>
<td>cr_hits</td>
<td>1</td>
<td>27.5518</td>
<td>82.1135</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>no_outs</td>
<td>1</td>
<td>2.5542</td>
<td>7.6123</td>
<td>0.0065</td>
</tr>
<tr>
<td>no_error</td>
<td>1</td>
<td>2.0405</td>
<td>6.0812</td>
<td>0.0147</td>
</tr>
</tbody>
</table>

### Analysis of Parameter Estimates

| Parameter   | DF | Estimate | Standard Error | t Value | Pr > |t| |
|-------------|----|----------|----------------|---------|------|---|
| Intercept   | 1  | 4.5835   | 0.1290         | 35.52   | <.0001 |
| no_hits     | 1  | 0.00562  | 0.00137        | 4.10    | <.0001 |
| no_bb       | 1  | 0.00602  | 0.00267        | 2.26    | 0.0253 |
| cr_hits     | 1  | 0.000701 | 0.000077       | 9.06    | <.0001 |
| no_outs     | 1  | 0.000453 | 0.000164       | 2.76    | 0.0065 |
| no_error    | 1  | -0.0180  | 0.00729        | -2.47   | 0.0147 |
**Summary of the Stepwise Selection Process**

The Summary of Stepwise Procedure section provides the step number, the explanatory input or inputs entered or removed at each step, the $F$ statistic, and the corresponding $p$-value in which the entry or removal of the input is based. For this example, 8 of the 19 original inputs met the 0.25 entry and stay probability values.

---

### Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>$F$ Value</th>
<th>Pr &gt; $F$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>8</td>
<td>75.525299</td>
<td>9.440662</td>
<td>29.08</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>154</td>
<td>49.987381</td>
<td>0.324593</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>Corrected Total</td>
<td>162</td>
<td>125.512680</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
</tbody>
</table>

### Model Fitting Information

- $R$-square: 0.6017
- Adj $R$-sq: 0.5810
- AIC: -174.6627
- BIC: -170.9032
- SBC: -146.8189
- C(p): 3.3390

### Type III Analysis of Effects

<table>
<thead>
<tr>
<th>Effect</th>
<th>DF</th>
<th>Type III SS</th>
<th>$F$ Value</th>
<th>Pr &gt; $F$</th>
</tr>
</thead>
<tbody>
<tr>
<td>no_hits</td>
<td>1</td>
<td>1.3988</td>
<td>4.3095</td>
<td>0.0396</td>
</tr>
<tr>
<td>no_rbi</td>
<td>1</td>
<td>0.4606</td>
<td>1.4191</td>
<td>0.2354</td>
</tr>
<tr>
<td>no_bb</td>
<td>1</td>
<td>2.1274</td>
<td>6.5541</td>
<td>0.0114</td>
</tr>
<tr>
<td>yr_major</td>
<td>1</td>
<td>1.6517</td>
<td>5.0885</td>
<td>0.0255</td>
</tr>
<tr>
<td>cr_hits</td>
<td>1</td>
<td>1.6766</td>
<td>5.1651</td>
<td>0.0244</td>
</tr>
<tr>
<td>cr_bb</td>
<td>1</td>
<td>0.7491</td>
<td>2.3077</td>
<td>0.1308</td>
</tr>
<tr>
<td>no_outs</td>
<td>1</td>
<td>2.3525</td>
<td>7.2475</td>
<td>0.0079</td>
</tr>
<tr>
<td>no_error</td>
<td>1</td>
<td>1.3915</td>
<td>4.2869</td>
<td>0.0401</td>
</tr>
</tbody>
</table>

### Analysis of Parameter Estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>DF</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>t Value</th>
<th>Pr &gt;</th>
<th>t</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>4.2938</td>
<td>0.1834</td>
<td>23.41</td>
<td>&lt;.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>no_hits</td>
<td>1</td>
<td>0.00424</td>
<td>0.00204</td>
<td>2.08</td>
<td>0.0396</td>
<td></td>
<td></td>
</tr>
<tr>
<td>no_rbi</td>
<td>1</td>
<td>0.00362</td>
<td>0.00304</td>
<td>1.19</td>
<td>0.2354</td>
<td></td>
<td></td>
</tr>
<tr>
<td>no_bb</td>
<td>1</td>
<td>0.00880</td>
<td>0.00344</td>
<td>2.56</td>
<td>0.0114</td>
<td></td>
<td></td>
</tr>
<tr>
<td>yr_major</td>
<td>1</td>
<td>0.0580</td>
<td>0.0257</td>
<td>2.26</td>
<td>0.0255</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cr_hits</td>
<td>1</td>
<td>0.00571</td>
<td>0.000251</td>
<td>2.27</td>
<td>0.0244</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cr_bb</td>
<td>1</td>
<td>-0.00084</td>
<td>0.000551</td>
<td>-1.52</td>
<td>0.1308</td>
<td></td>
<td></td>
</tr>
<tr>
<td>no_outs</td>
<td>1</td>
<td>0.000439</td>
<td>0.000163</td>
<td>2.69</td>
<td>0.0079</td>
<td></td>
<td></td>
</tr>
<tr>
<td>no_error</td>
<td>1</td>
<td>-0.0152</td>
<td>0.00733</td>
<td>-2.07</td>
<td>0.0401</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

NOTE: No (additional) effects met the 0.25 significance level for entry into the model.

---

### Step 8. Effect no_rbi entered:

**Analysis of Variance**

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>$F$ Value</th>
<th>Pr &gt; $F$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>8</td>
<td>75.525299</td>
<td>9.440662</td>
<td>29.08</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>154</td>
<td>49.987381</td>
<td>0.324593</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>Corrected Total</td>
<td>162</td>
<td>125.512680</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
</tbody>
</table>

### Model Fitting Information

- $R$-square: 0.6017
- Adj $R$-sq: 0.5810
- AIC: -174.6627
- BIC: -170.9032
- SBC: -146.8189
- C(p): 3.3390

### Type III Analysis of Effects

<table>
<thead>
<tr>
<th>Effect</th>
<th>DF</th>
<th>Type III SS</th>
<th>$F$ Value</th>
<th>Pr &gt; $F$</th>
</tr>
</thead>
<tbody>
<tr>
<td>no_hits</td>
<td>1</td>
<td>1.3988</td>
<td>4.3095</td>
<td>0.0396</td>
</tr>
<tr>
<td>no_rbi</td>
<td>1</td>
<td>0.4606</td>
<td>1.4191</td>
<td>0.2354</td>
</tr>
<tr>
<td>no_bb</td>
<td>1</td>
<td>2.1274</td>
<td>6.5541</td>
<td>0.0114</td>
</tr>
<tr>
<td>yr_major</td>
<td>1</td>
<td>1.6517</td>
<td>5.0885</td>
<td>0.0255</td>
</tr>
<tr>
<td>cr_hits</td>
<td>1</td>
<td>1.6766</td>
<td>5.1651</td>
<td>0.0244</td>
</tr>
<tr>
<td>cr_bb</td>
<td>1</td>
<td>0.7491</td>
<td>2.3077</td>
<td>0.1308</td>
</tr>
<tr>
<td>no_outs</td>
<td>1</td>
<td>2.3525</td>
<td>7.2475</td>
<td>0.0079</td>
</tr>
<tr>
<td>no_error</td>
<td>1</td>
<td>1.3915</td>
<td>4.2869</td>
<td>0.0401</td>
</tr>
</tbody>
</table>

### Analysis of Parameter Estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>DF</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>t Value</th>
<th>Pr &gt;</th>
<th>t</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>4.2938</td>
<td>0.1834</td>
<td>23.41</td>
<td>&lt;.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>no_hits</td>
<td>1</td>
<td>0.00424</td>
<td>0.00204</td>
<td>2.08</td>
<td>0.0396</td>
<td></td>
<td></td>
</tr>
<tr>
<td>no_rbi</td>
<td>1</td>
<td>0.00362</td>
<td>0.00304</td>
<td>1.19</td>
<td>0.2354</td>
<td></td>
<td></td>
</tr>
<tr>
<td>no_bb</td>
<td>1</td>
<td>0.00880</td>
<td>0.00344</td>
<td>2.56</td>
<td>0.0114</td>
<td></td>
<td></td>
</tr>
<tr>
<td>yr_major</td>
<td>1</td>
<td>0.0580</td>
<td>0.0257</td>
<td>2.26</td>
<td>0.0255</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cr_hits</td>
<td>1</td>
<td>0.00571</td>
<td>0.000251</td>
<td>2.27</td>
<td>0.0244</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cr_bb</td>
<td>1</td>
<td>-0.00084</td>
<td>0.000551</td>
<td>-1.52</td>
<td>0.1308</td>
<td></td>
<td></td>
</tr>
<tr>
<td>no_outs</td>
<td>1</td>
<td>0.000439</td>
<td>0.000163</td>
<td>2.69</td>
<td>0.0079</td>
<td></td>
<td></td>
</tr>
<tr>
<td>no_error</td>
<td>1</td>
<td>-0.0152</td>
<td>0.00733</td>
<td>-2.07</td>
<td>0.0401</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

NOTE: No (additional) effects met the 0.25 significance level for entry into the model.
List Report of Selected Variables in the OUTEST= data set

The example PROC PRINT report of the OUTEST= data set lists selected fit statistics for the training and test data sets. The default OUTEST data set contains two observations for each step number. These observations are distinguished by value of the _TYPE_ variable:

- _TYPE_ = "PARMS" - contains parameter estimate statistics
- _TYPE_ = "T" - contains the t-value for the parameter estimate

Because a WHERE statement was used to select only values of TYPE = "PARMS", this report contains one observation per step number. An additional observation is displayed that identifies the model chosen based on the SBC criterion (CHOOSE="SBC").

Partial Listing of the OUTEST= Data Set

<table>
<thead>
<tr>
<th>Model Selection Step Number</th>
<th>Model Chosen Criterion</th>
<th>Train: Schwarz's Bayesian Square Error Function</th>
<th>Train: Mean Average Error Function</th>
<th>Train: Square Error Function</th>
<th>Test: Mean Average Error Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>-37.505</td>
<td>0.77477</td>
<td>0.77002</td>
<td>0.81858</td>
</tr>
<tr>
<td>1</td>
<td>109.377</td>
<td>-148.441</td>
<td>0.37312</td>
<td>0.36625</td>
<td>0.34632</td>
</tr>
<tr>
<td>2</td>
<td>152.041</td>
<td>-153.552</td>
<td>0.35597</td>
<td>0.34732</td>
<td>0.37438</td>
</tr>
<tr>
<td>3</td>
<td>153.438</td>
<td>-153.552</td>
<td>0.34424</td>
<td>0.33368</td>
<td>0.38553</td>
</tr>
<tr>
<td>5</td>
<td>153.552</td>
<td>-153.552</td>
<td>0.35553</td>
<td>0.32318</td>
<td>0.37563</td>
</tr>
<tr>
<td>6</td>
<td>152.974</td>
<td>-150.418</td>
<td>0.32459</td>
<td>0.30667</td>
<td>0.35978</td>
</tr>
<tr>
<td>7</td>
<td>150.418</td>
<td>-146.819</td>
<td>0.32459</td>
<td>0.30667</td>
<td>0.35978</td>
</tr>
<tr>
<td>8</td>
<td>153.552</td>
<td>1</td>
<td>0.35553</td>
<td>0.32318</td>
<td>0.37563</td>
</tr>
</tbody>
</table>

GPLOT Diagnostic Plots for the Scored Baseball Test Data

Plot of the log of salary versus the predicted log of salary.
Plot of the residual values versus the predicted log of salary.
Diagnostic Plots for the Scored Baseball Data

Residual: lgsolar

Predicted: lgsolar
The PROC DMREG statement invokes the procedure. The DATA= option identifies the training data set that is used to fit the model. The DMDBCAT= option identifies the training data catalog.

```plaintext
proc dmreg data=sampsio.dmdbase dmdbcat=sampsio.dmdbase
```
The TESTDATA= option identifies the test data set. The OUTEST= option creates the output data set containing estimates and fit statistics.

testdata=sampsio.dmtbase outest=regest;
The CLASS statement specifies the categorical variables to be used in the regression analysis.

```plaintext
class league division position;
```
The MODEL statement specifies the linear model. The ERROR=normal model option specifies to use the normal error distribution. The CHOOSE=SBC model option specifies to choose the model subset with the smallest Schwarz Bayesian criterion.

\[
\text{model } \log\text{salar} = \text{no\_atbat} \text{ no\_hits} \text{ no\_home} \text{ no\_runs} \text{ no\_rbi} \text{ no\_bb} \\
\text{yr\_major} \text{ cr\_atbat} \text{ cr\_hits} \text{ cr\_home} \text{ cr\_runs} \\
\text{cr\_rbi} \text{ cr\_bb} \text{ league} \text{ division} \text{ position} \text{ no\_outs} \\
\text{no\_assts} \text{ no\_error}
\]

/ error=normal
choose=sbc
The MODEL option SELECTION=STEPWISE specifies to use the stepwise variable selection method. Stepwise selection systematically adds and deletes inputs from the model based on the SLENTRY= and SLSTAY= significance levels. The subset models are created based on the SLENTRY and SLSTAY significance levels, but the model that is chosen is based solely on the subset model that has the smallest SBC criterion.

```
selection=stepwise
slentry=0.25 slstay=0.25;
```
The SCORE statement specifies the data set that you want to score in conjunction with training. The DATA= option identifies the score data set (for this example, the test data set).

```
score data=sampsio.dmtbase nodmdb
```
The OUT=option identifies the output data set that contains estimates and fit statistics for the scored data set. The RENAME=option enables you to rename variables in the OUT= data set.

```plaintext
out=regout(rename=(p_logsal=predict r_logsal=residual));
title 'Output from the DMREG Procedure';
run;
```
PROC PRINT produces a report of selected variables from the OUTEST= data set.

```plaintext
proc print data=regest noobs label;
  var _step_ _chosen_ _sbc_ _mse_ _averr_ _tmse_ _taverr_
  where _type_ = 'PARMS';
  title 'Partial Listing of the OUTEST= Data Set';
run;
```
PROC GPLOT produces diagnostic plots of the scored test data. The first PLOT statement plots the response versus the predicted values.

```
proc gplot data=regout;
pplot logsalar*predict / haxis=axis1 vaxis=axis2 frame;
symbol c=black i=none v=dot h=3 pct;
axis1 c=black width=2.5;
axis2 c=black width=2.5;
title 'Diagnostic Plots for the Scored Baseball Data';
```
The second PLOT statement plots the residuals versus the predicted values.

    plot residual*predict / haxis=axis1 vaxis=axis2;
run;
quit;
This example provides a comparison of the DMREG and LOGISTIC procedures when using a categorical input to model a binary target. The example data set SAMPSIO.HMEQ contains fictitious mortgage data where each case represents an applicant for a home equity loan. All applicants have an existing mortgage.

The binary target BAD represents whether or not an applicant eventually defaulted or was ever seriously delinquent. There are nine continuous inputs available for modeling. JOB is the only categorical input used to predict the target BAD.

When you compare the output from the DMREG and LOGISTIC procedures code, you must take into consideration how each procedure handles the categorical variables. By default, DMREG uses a deviations from the means coding to code the classification variables. The design matrix for the class effects has values of 0, 1, and -1 for the reference levels. This coding is sometimes referred to as "effects", "center-point", and "full-rank" coding. The parameters for these categorical indicators measure the difference from each level to the average across levels.

Because the LOGISTIC procedure does not enable you to specify class inputs directly in the MODEL statement, you must first create an input data set that contains the design matrix for the class variables. To create the design matrix data set for input to the LOGISTIC procedure, you can use a SAS DATA step, a TRANSREG procedure, or a GENMOD procedure. If you use the deviations from the means coding method to code the class variables, then the LOGISTIC output will automatically match the output generated from the DMREG run. If you use the GLM non-full rank coding (0, 1) to code the class variables, you must set the DMREG CODE= MODEL statement option in GLM. In this case, both procedures will generate the same output.

### Program: Deviations from the Mean Coding

```sas
proc freq data=sampsio.hmeq;
   tables job / missing;
   title 'JOB Classification Table';
run;

data hmeq;
   set sampsio.hmeq;
   if job = '.' then job='Other';
run;

proc transreg data=hmeq design;
   model class (job/deviations);
   id bad loan mortdue value yoj derog clage ninq clno debtinc;
   output;
run;
```
/*
data dumyhmeq;
  set hmeq;
  select (job);
  when ('Mgr')
    do;
      j_mgr=1;
      j_off=0;
      j_other=0;
      j_prof=0;
      j_sales=0;
      j_self=-1;
    end;
  when ('Office')
    do;
      j_mgr=0;
      j_off=1;
      j_other=0;
      j_prof=0;
      j_sales=0;
      j_self=-1;
    end;
  when ('Other')
    do;
      j_mgr=0;
      j_off=0;
      j_other=1;
      j_prof=0;
      j_sales=0;
      j_self=-1;
    end;
  when ('ProfExe')
    do;
      j_mgr=0;
      j_off=0;
      j_other=0;
      j_prof=1;
      j_sales=0;
      j_self=-1;
    end;
  when ('Sales')
    do;
      j_mgr=0;
      j_off=0;
      j_other=0;
      j_prof=0;
      j_sales=1;
      j_self=-1;
    end;
  when ('Self')
    do;
      j_mgr=-1;
      j_off=-1;
      j_other=-1;
      j_prof=-1;
    end;
*/
*\j_sales=-1;  
\j_self=-1;  
end;  
\otherwise;  
end;  
run;  
*/  

\proc logistic descending;  
\model bad = \&_trgind loan mortdue value yoj  
\derog clage ninq clno debtinc;  
\title 'LOGISTIC Home Equity Data: Deviations from the Mean Coding';  
run;  

\proc dmdb batch data=hmeq  
\out=dm_data dmdbcat=dm_cat;  
\var loan mortdue value yoj derog  
\clage ninq clno debtinc;  
\class bad(desc)  
\job(asc);  
\target bad;  
run;  

\proc dmreg data=dm_data  
\dmdbcat=dm_cat;  
\class bad job;  
\model bad = job loan mortdue value yoj derog  
\clage ninq clno debtinc;  
\title1 'DMREG Home Equity Data:  
\Default Deviations from the Mean Coding';  
run;  

---  

Output: Deviations from the Mean Coding  

FREQ Classification Table for JOB.  

The categorical input JOB contains 7 levels. Notice that 279 cases have missing values. Both the DMREG and LOGISTIC procedures omit observations that have missing values from the analysis. For this example, the missing values are imputed using the mode of JOB.  

<table>
<thead>
<tr>
<th>JOB</th>
<th>Frequency</th>
<th>Percent</th>
<th>Cumulative Frequency</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>\Mgr</td>
<td>279</td>
<td>4.68</td>
<td>279</td>
<td>4.68</td>
</tr>
<tr>
<td>Office</td>
<td>767</td>
<td>12.87</td>
<td>1046</td>
<td>17.55</td>
</tr>
<tr>
<td>Other</td>
<td>948</td>
<td>15.91</td>
<td>1994</td>
<td>33.46</td>
</tr>
<tr>
<td>ProfEx</td>
<td>2388</td>
<td>40.07</td>
<td>4382</td>
<td>73.52</td>
</tr>
<tr>
<td>Sales</td>
<td>1276</td>
<td>21.41</td>
<td>5658</td>
<td>94.93</td>
</tr>
<tr>
<td>Self</td>
<td>109</td>
<td>1.83</td>
<td>5767</td>
<td>96.76</td>
</tr>
<tr>
<td></td>
<td>193</td>
<td>3.24</td>
<td>5960</td>
<td>100.00</td>
</tr>
</tbody>
</table>
LOGISTIC Output

LOGISTIC Home Equity Data: Deviations from the Mean Coding

The LOGISTIC Procedure

Model Information

Data Set WORK.DATA1
Response Variable BAD
Number of Response Levels 2
Number of Observations 3527
Model binary, logit
Optimization Technique Fisher's scoring

Response Profile

<table>
<thead>
<tr>
<th>Ordered Value</th>
<th>Total BAD</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>313</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>3214</td>
<td></td>
</tr>
</tbody>
</table>

Probability modeled is BAD=1.

NOTE: 2433 observations were deleted due to missing values for the response or explanatory variables.

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Intercept Only</th>
<th>Intercept and Covariates</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIC</td>
<td>2115.535</td>
<td>1765.011</td>
</tr>
<tr>
<td>SC</td>
<td>2121.703</td>
<td>1857.534</td>
</tr>
<tr>
<td>-2 Log L</td>
<td>2113.535</td>
<td>1795.011</td>
</tr>
</tbody>
</table>

Testing Global Null Hypothesis: BETA=0

<table>
<thead>
<tr>
<th>Test</th>
<th>Chi-Square</th>
<th>DF</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Likelihood Ratio</td>
<td>378.5235</td>
<td>14</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Score</td>
<td>488.0227</td>
<td>14</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Wald</td>
<td>250.3468</td>
<td>14</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

Analysis of Maximum Likelihood Estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>DF</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>Chi-Square</th>
<th>Wald</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td></td>
<td>-4.7354</td>
<td>0.4913</td>
<td>120.5087</td>
<td>&lt;.0001</td>
<td></td>
</tr>
<tr>
<td>JOBMgr</td>
<td>1</td>
<td>-0.1672</td>
<td>0.1755</td>
<td>0.3073</td>
<td>0.5408</td>
<td></td>
</tr>
<tr>
<td>JOBOffice</td>
<td>1</td>
<td>-0.5672</td>
<td>0.1745</td>
<td>10.5634</td>
<td>0.0012</td>
<td></td>
</tr>
<tr>
<td>JOBOther</td>
<td>1</td>
<td>-0.2373</td>
<td>0.1337</td>
<td>3.1513</td>
<td>0.0759</td>
<td></td>
</tr>
<tr>
<td>JOBProfExe</td>
<td>1</td>
<td>-0.2414</td>
<td>0.1551</td>
<td>2.4231</td>
<td>0.1196</td>
<td></td>
</tr>
<tr>
<td>JOBSales</td>
<td>1</td>
<td>0.7526</td>
<td>0.3348</td>
<td>5.0514</td>
<td>0.0246</td>
<td></td>
</tr>
<tr>
<td>LOAN</td>
<td>1</td>
<td>-0.00003</td>
<td>7.786E-6</td>
<td>10.4819</td>
<td>0.0012</td>
<td></td>
</tr>
<tr>
<td>MORTDUE</td>
<td>1</td>
<td>-5.98E-6</td>
<td>3.723E-6</td>
<td>2.5780</td>
<td>0.1084</td>
<td></td>
</tr>
<tr>
<td>VALUE</td>
<td>1</td>
<td>4.651E-6</td>
<td>3.184E-6</td>
<td>2.1339</td>
<td>0.1441</td>
<td></td>
</tr>
<tr>
<td>YQJ</td>
<td>1</td>
<td>-0.00818</td>
<td>0.00971</td>
<td>0.7099</td>
<td>0.3995</td>
<td></td>
</tr>
<tr>
<td>DEROG</td>
<td>1</td>
<td>0.8898</td>
<td>0.0984</td>
<td>81.7161</td>
<td>&lt;.0001</td>
<td></td>
</tr>
</tbody>
</table>
DMREG Output

Notice that the DMREG output matches the output generated from the LOGISTIC run.

DMREG Home Equity Data: Default Deviations from the Mean Coding

The DMREG Procedure

Training Data Set: WORK.DM_DATA
DMDB Catalog: WORK.DM_CAT
Target Variable: BAD
Target Measurement Level: Ordinal
Number of Target Categories: 2
Error: MBernoulli
Link Function: Logit
Number of Model Parameters: 15
Number of Observations: 3527

Target Profile

<table>
<thead>
<tr>
<th>Ordered Value</th>
<th>Total BAD Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>313</td>
</tr>
<tr>
<td>2</td>
<td>3214</td>
</tr>
</tbody>
</table>

Input Class Level Information

Design Variables

<table>
<thead>
<tr>
<th>Class</th>
<th>Value</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>JOB</td>
<td>Mgr</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Office</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Other</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>ProfExe</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Sales</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Self</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
</tr>
</tbody>
</table>

DMREG Home Equity Data: Default Deviations from the Mean Coding

The DMREG Procedure

Newton-Raphson Ridge Optimization

Without Parameter Scaling

Parameter Estimates

Optimization Start

Active Constraints 0
Objective Function 1056.7673865
Max Abs Gradient Element 29.970626595
Optimization Results

<table>
<thead>
<tr>
<th>Iter</th>
<th>Restarts</th>
<th>Function Calls</th>
<th>Active Constraints</th>
<th>Objective Function</th>
<th>Objective Function Change</th>
<th>Max Abs Gradient Element</th>
<th>Ridge</th>
<th>Ratio Between Actual and Predicted Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>912.15841</td>
<td>144.6</td>
<td>116.9</td>
<td>0</td>
<td>0.593</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>868.27555</td>
<td>48.8829</td>
<td>9.4267</td>
<td>0</td>
<td>1.006</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>867.50765</td>
<td>0.7679</td>
<td>0.4844</td>
<td>0</td>
<td>1.033</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>867.50562</td>
<td>0.00203</td>
<td>0.00140</td>
<td>0</td>
<td>1.002</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>6</td>
<td>0</td>
<td>867.50562</td>
<td>1.67E-8</td>
<td>1.184E-8</td>
<td>0</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Testing Global Null Hypothesis BETA=0

<table>
<thead>
<tr>
<th>Intercept Only</th>
<th>Intercept and Covariates</th>
<th>Chi-Square for Covariates</th>
</tr>
</thead>
<tbody>
<tr>
<td>-2 LOG L</td>
<td>2113.535</td>
<td>1735.011</td>
</tr>
<tr>
<td></td>
<td>378.524 with 14 DF (p&lt;.0001)</td>
<td></td>
</tr>
</tbody>
</table>

Type III Analysis of Effects

<table>
<thead>
<tr>
<th>Effect</th>
<th>DF</th>
<th>Wald</th>
<th>Pr &gt;</th>
<th>Chi-Square</th>
<th>Chi-Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>JOB</td>
<td>5</td>
<td>14.3409</td>
<td>0.0136</td>
<td>129.54</td>
<td></td>
</tr>
<tr>
<td>LOAN</td>
<td>1</td>
<td>10.4820</td>
<td>0.0012</td>
<td>11.31</td>
<td></td>
</tr>
<tr>
<td>MORTDUE</td>
<td>1</td>
<td>2.5780</td>
<td>0.1084</td>
<td>2.5780</td>
<td></td>
</tr>
<tr>
<td>VALUE</td>
<td>1</td>
<td>2.1339</td>
<td>0.1441</td>
<td>2.1339</td>
<td></td>
</tr>
<tr>
<td>YOJ</td>
<td>1</td>
<td>0.7099</td>
<td>0.3995</td>
<td>0.7099</td>
<td></td>
</tr>
<tr>
<td>DEROG</td>
<td>1</td>
<td>81.7165</td>
<td>&lt;.0001</td>
<td>81.7165</td>
<td></td>
</tr>
<tr>
<td>CLAGE</td>
<td>1</td>
<td>26.7300</td>
<td>&lt;.0001</td>
<td>26.7300</td>
<td></td>
</tr>
<tr>
<td>NINQ</td>
<td>1</td>
<td>7.3653</td>
<td>0.0066</td>
<td>7.3653</td>
<td></td>
</tr>
<tr>
<td>CLNO</td>
<td>1</td>
<td>0.1735</td>
<td>0.6770</td>
<td>0.1735</td>
<td></td>
</tr>
<tr>
<td>DEBTINC</td>
<td>1</td>
<td>105.6561</td>
<td>&lt;.0001</td>
<td>105.6561</td>
<td></td>
</tr>
</tbody>
</table>

Analysis of Maximum Likelihood Estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>DF</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>Wald</th>
<th>Pr &gt;</th>
<th>Standardized Estimate</th>
<th>exp(Est)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>-4.7355</td>
<td>0.4313</td>
<td>129.54</td>
<td>&lt;.0001</td>
<td>.</td>
<td>0.009</td>
</tr>
<tr>
<td>NGR</td>
<td>1</td>
<td>-0.1672</td>
<td>0.1755</td>
<td>0.91</td>
<td>0.3408</td>
<td>.</td>
<td>0.846</td>
</tr>
<tr>
<td>JOB</td>
<td>1</td>
<td>-0.5672</td>
<td>0.1745</td>
<td>10.56</td>
<td>0.0012</td>
<td>.</td>
<td>0.567</td>
</tr>
<tr>
<td>VALUE</td>
<td>1</td>
<td>-0.2373</td>
<td>0.1337</td>
<td>3.15</td>
<td>0.0759</td>
<td>.</td>
<td>0.789</td>
</tr>
<tr>
<td>YOJ</td>
<td>1</td>
<td>-0.2415</td>
<td>0.1551</td>
<td>2.42</td>
<td>0.1196</td>
<td>.</td>
<td>0.785</td>
</tr>
<tr>
<td>LOAN</td>
<td>1</td>
<td>-0.00003</td>
<td>7.78E-6</td>
<td>10.48</td>
<td>0.0012</td>
<td>-0.158376</td>
<td>1.000</td>
</tr>
<tr>
<td>MORTDUE</td>
<td>1</td>
<td>-5.38E-6</td>
<td>3.72E-6</td>
<td>2.58</td>
<td>0.1084</td>
<td>-0.147394</td>
<td>1.000</td>
</tr>
<tr>
<td>VALUE</td>
<td>1</td>
<td>4.651E-6</td>
<td>3.18E-6</td>
<td>2.13</td>
<td>0.1441</td>
<td>0.138844</td>
<td>1.000</td>
</tr>
<tr>
<td>YOJ</td>
<td>1</td>
<td>-0.00818</td>
<td>0.00971</td>
<td>0.71</td>
<td>0.3995</td>
<td>-0.034557</td>
<td>0.992</td>
</tr>
<tr>
<td>DEROG</td>
<td>1</td>
<td>0.8898</td>
<td>0.0984</td>
<td>81.72</td>
<td>&lt;.0001</td>
<td>0.282757</td>
<td>2.435</td>
</tr>
<tr>
<td>CLAGE</td>
<td>1</td>
<td>-0.00520</td>
<td>0.00101</td>
<td>26.74</td>
<td>&lt;.0001</td>
<td>-0.235922</td>
<td>0.995</td>
</tr>
<tr>
<td>NINQ</td>
<td>1</td>
<td>0.0070</td>
<td>0.0360</td>
<td>7.37</td>
<td>0.0066</td>
<td>0.082609</td>
<td>1.103</td>
</tr>
<tr>
<td>CLNO</td>
<td>1</td>
<td>-0.00314</td>
<td>0.00753</td>
<td>0.17</td>
<td>0.6770</td>
<td>-0.016175</td>
<td>0.997</td>
</tr>
<tr>
<td>DEBTINC</td>
<td>1</td>
<td>0.1033</td>
<td>0.0101</td>
<td>105.6561</td>
<td>&lt;.0001</td>
<td>0.488171</td>
<td>1.109</td>
</tr>
</tbody>
</table>

Odds Ratio Estimates

Input Odds Ratio

Odds Ratio
Program: GLM Non-Full Rank (0, 1) Coding

data dumyhmeq;
    set hmeq;
    j_mgr=(job='Mgr');
    j_off=(job='Office');
    j_other=(job='Other');
    j_prof=(job='ProfExe');
    j_sales=(job='Sales');
    j_self=(job='Self');
run;

proc logistic data=dumyhmeq descending noprint;
    model bad = j_mgr j_off j_other j_prof j_sales j_self
           loan mortdue value yoj derog
           clage ninq clno debtinc;
    output out=logfit(keep=bad p_bad1) p=p_bad1;
    title 'LOGISTIC Home Equity Data: GLM coding';
    run;

proc dmdb batch data=hmeq
    out=dm_data dmdbcat=dm_cat;
    var loan mortdue value yoj derog
        clage ninq clno debtinc;
    class bad(desc)
        reason(asc)
        job(asc);
    target bad;
run;

proc dmreg data=dm_data
    dmdbcat=dm_cat
    noprint;
    class bad job;
    model bad = job loan mortdue value yoj derog
               clage ninq clno debtinc / coding=glm;
    score out=dmscore;
    title1 'DMREG Home Equity Data: GLM coding';
    run;

proc compare data=dmscore compare=logfit note
    method=absolute
criterion=1e-7;
var p_bad1;

run;

Output: GLM Non-Full Rank (0, 1) Coding
PROC COMPARE results.

DMREG Home Equity Data: GLM coding

The COMPARE Procedure
Comparison of WORK.DMSCORE with WORK.LOGFIT
(Method=ABSOLUTE, Criterion=0.0000001)

Data Set Summary

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Created</th>
<th>Modified</th>
<th>NVar</th>
<th>NObs</th>
</tr>
</thead>
<tbody>
<tr>
<td>WORK.DMSCORE</td>
<td>01NOV00:14:19:52</td>
<td>01NOV00:14:19:52</td>
<td>20</td>
<td>5960</td>
</tr>
<tr>
<td>WORK.LOGFIT</td>
<td>01NOV00:14:19:50</td>
<td>01NOV00:14:19:50</td>
<td>2</td>
<td>5960</td>
</tr>
</tbody>
</table>

Variables Summary
Number of Variables in Common: 2.
Number of Variables in WORK.DMSCORE but not in WORK.LOGFIT: 18.
Number of Variables with Differing Attributes: 2.
Number of VAR Statement Variables: 1.

Listing of Common Variables with Differing Attributes

<table>
<thead>
<tr>
<th>Variable</th>
<th>Dataset</th>
<th>Type</th>
<th>Length</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>P_BAD1</td>
<td>WORK.DMSCORE</td>
<td>Num</td>
<td>8</td>
<td>Predicted: BAD=1</td>
</tr>
<tr>
<td></td>
<td>WORK.LOGFIT</td>
<td>Num</td>
<td>8</td>
<td>Estimated Probability</td>
</tr>
</tbody>
</table>

Observation Summary

<table>
<thead>
<tr>
<th>Observation</th>
<th>Base</th>
<th>Compare</th>
</tr>
</thead>
<tbody>
<tr>
<td>First Obs</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>First Unequal</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Last Unequal</td>
<td>5960</td>
<td>5960</td>
</tr>
<tr>
<td>Last Obs</td>
<td>5960</td>
<td>5960</td>
</tr>
</tbody>
</table>

Number of Observations in Common: 5960.
Total Number of Observations Read from WORK.DMSCORE: 5960.
Total Number of Observations Read from WORK.LOGFIT: 5960.

Number of Observations with Some Compared Variables Unequal: 5822.
Number of Observations with All Compared Variables Equal: 138.

Values Comparison Summary

Number of Variables Compared with All Observations Equal: 0.
Number of Variables Compared with Some Observations Unequal: 1.
Number of Variables with Missing Value Differences: 1.
Total Number of Values which Compare Unequal: 5822.
Total Number of Values not EXACTLY Equal: 5960.
Maximum Difference: 0.0000064507.

Copyright 2000 by SAS Institute Inc., Cary, NC, USA. All rights reserved.
PROC FREQ step to create a classification table for the categorical input JOB.

```plaintext
proc freq data=sampsio.hmeq;
  tables job / missing;
  title 'JOB Classification Table';
run;
```
SAS DATA step to replace the missing JOB values with the variable's mode. It does not matter whether or not you perform data imputation prior to modeling - DMREG and LOGISTIC will produce the same results if you use the same method to code the class variables. Some of the continuous inputs have missing values. DMREG and LOGISTIC do not use observations that have missing values in the analysis. You can impute the missing values for the continuous inputs by using the STDIZE procedure.

```sas
data hmeq;
  set sampsio.hmeq;
  if job = '' then job='Other';
run;
```
PROC TRANSREG step to create the design matrix for the classification input JOB. The DESIGN option specifies that the goal is design matrix creation, not analysis.

proc transreg data=hmeq design;
The MODEL statement specifies the class variable JOB. The DEVIATIONS (or EFFECTS) t-option requests a deviations from the means coding.

model class (job/deviations);
The ID statement also specifies to output the target and the continuous inputs to the temporary design matrix data set. PROC TRANSREG automatically creates the macro variable &_TRGIND that contains the list of independent variables. This macro variable is used in the MODEL statement in PROC LOGISTIC.

```plaintext
id bad loan mortdue value yoj derog clage ninq clno debtinc;
output;
run;
```
You can also create the design matrix for the classification variable(s) in a SAS DATA step although this task is too time consuming for databases that contain several class variables. The DATA step is commented out, but it does demonstrate how to manually code a categorical variable using the deviations from the MEANS method.

```sas
/*
data dumyhmeq;
    set hmeq;
    select (job);
    when ('Mgr')
        do;
            j_mgr=1;
            j_off=0;
            j_other=0;
            j_prof=0;
            j_sales=0;
            j_self=-1;
        end;
    when ('Office')
        do;
            j_mgr=0;
            j_off=1;
            j_other=0;
            j_prof=0;
            j_sales=0;
            j_self=-1;
        end;
    when ('Other')
        do;
            j_mgr=0;
            j_off=0;
            j_other=1;
            j_prof=0;
            j_sales=0;
            j_self=-1;
        end;
    when ('ProfExe')
        do;
            j_mgr=0;
            j_off=0;
            j_other=0;
            j_prof=1;
            j_sales=0;
            j_self=-1;
        end;
end;
```
when ('Sales') do;
    j_mgr=0;
    j_off=0;
    j_other=0;
    j_prof=0;
    j_sales=1;
    j_self=-1;
end;

when ('Self') do;
    j_mgr=-1;
    j_off=-1;
    j_other=-1;
    j_prof=-1;
    j_sales=-1;
    j_self=-1;
end;
otherwise;
end;
run;

*/
PROC LOGISTIC step to model the binary target BAD. The macro variable \&_TRGIND obtains the classification design matrix from the subsequent PROC TRANSREG run. The DESCENDING option causes the procedure to model the probability that BAD = 1 (bad applicants).

```
proc logistic descending;
   model bad = &_trgind loan mortdue value yoj
defrog clage ninq clno debtinc;
   title 'LOGISTIC Home Equity Data: Deviations from the Mean Coding';
run;
```
PROC DMDB step to create the DMDB data set and catalog that are required as input to DMREG.

```plaintext
proc dmdb batch data=hmeq
   out=dm_data dmdbcat=dm_cat;
   var loan mortdue value yoj derog
       clage ninq clno debtinc;
   class bad(desc)
       job(asc);
   target bad;
run;
```
Because the order of the target BAD was set to descending in the DMDB data set, DMREG also models the probability that BAD=1 (bad applicants). By default, DMREG using deviation from the means coding to create the design matrix for the class variables.

```sas
proc dmreg data=dm_data
dmdbcat=dm_cat;
  class bad job;
  model bad = job loan mortdue value yoj derog clage ninq clno debtinc;
  title1 'DMREG Home Equity Data: Default Deviations from the Mean Coding';
run;
```
DATA step program to code the class variable JOB using GLM non-full rank (0, 1) coding.

```sql
data dumyhmeq;
  set hmeq;
  j_mgr=(job='Mgr');
  j_off=(job='Office');
  j_other=(job='Other');
  j_prof=(job='ProfExe');
  j_sales=(job='Sales');
  j_self=(job='Self');
run;
```
PROC LOGISTIC step to model the binary target BAD.

proc logistic data=dumyhmeq descending noprint;
   model bad = j_mgr j_off j_other j_prof j_sales j_self
      loan mortdue value yoj derog
      clage ning clno debtinc;
   output out=logfit(keep=bad p_bad1) p=p_bad1;
title 'LOGISTIC Home Equity Data: GLM coding';
run;
The NOPRINT option suppresses the printing of the DMREG output. PROC COMPARE is used to compare the predicted values from the LOGISTIC and DMREG models. The CODING=GLM option creates the design matrix for the class variables using GLM non-full rank coding.

```plaintext
proc dmreg data=dm_data
dmdbcat=dm_cat
   noprint;
   class bad job;
   model bad = job loan mortdue value yoj derog
clage ninq clno debtinc / coding=glm;
   score out=dmscore;
   title1 'DMREG Home Equity Data: GLM coding';
run;
```

References


