



THE  
POWER  
TO KNOW.

**SAS<sup>®</sup> Enterprise Miner<sup>™</sup> and  
SAS<sup>®</sup> Text Miner Procedures  
Reference for SAS<sup>®</sup> 9.1.3  
The DMINE Procedure  
(Book Excerpt)**

The correct bibliographic citation for this manual is as follows: SAS Institute Inc., SAS® *Enterprise Miner™* and SAS® *Text Miner Procedures: Reference for SAS® 9.1.3*, Cary, NC: SAS Institute Inc.

**SAS® Enterprise Miner™ and SAS® Text Miner Procedures: Reference for SAS 9.1.3**

Copyright © 2008 by SAS Institute Inc., Cary, NC, USA.

All rights reserved. Produced in the United States of America.

**For a Web download or e-book:** Your use of this publication shall be governed by the terms established by the vendor at the time you acquire this publication.

**U.S. Government Restricted Rights Notice.** Use, duplication, or disclosure of this software and related documentation by the U.S. government is subject to the Agreement with SAS Institute and the restrictions set forth in FAR 52.227-19 Commercial Computer Software-Restricted Rights (June 1987).

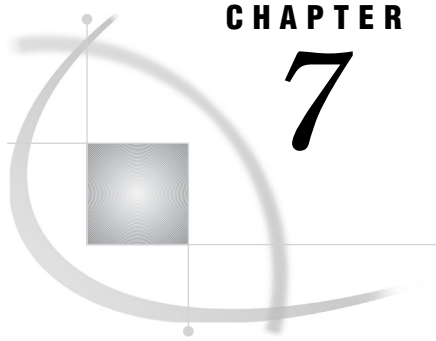
SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513.

1st electronic book, October 2008

SAS® Publishing provides a complete selection of books and electronic products to help customers use SAS software to its fullest potential. For more information about our e-books, e-learning products, CDs, and hard-copy books, visit the SAS Publishing Web site at [support.sas.com/publishing](http://support.sas.com/publishing) or call 1-800-727-3228.

SAS® and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are registered trademarks or trademarks of their respective companies.



## CHAPTER

## 7

## The DMINE Procedure

---

<i>Overview: DMINE Procedure</i>	97
<i>Syntax: DMINE Procedure</i>	97
<i>PROC DMINE Statement</i>	98
<i>CODE Statement</i>	100
<i>FREQ Statement</i>	101
<i>TARGET Statement</i>	102
<i>VARIABLES Statement</i>	102
<i>WEIGHT Statement</i>	102
<i>Details: DMINE Procedure</i>	103
<i>Fit Statistics for OUTEST and OUTFIT Data Sets</i>	103
<i>Missing Values</i>	105
<i>Examples: DMINE Procedure</i>	105
<i>Example 1: Modeling a Continuous Target with the DMINE Procedure (Simple Selection Settings)</i>	105
<i>Example 2: Including the AOV16 and Grouping Variables into the Analysis (Detailed Selection Settings)</i>	112
<i>Example 3: Modeling a Binary Target with the DMINE Procedure</i>	118

---

### Overview: DMINE Procedure

Many data mining databases have hundreds of potential model inputs (independent variables). The DMINE procedure enables you to quickly identify the input variables that are useful for predicting the target variable(s) based on a linear models framework. The procedure facilitates ordinary least squares or logistic regression methods. (Logistic regression is a form of regression analysis in which the response variable represents a binary or ordinal-level response.)

PROC DMINE and PROC DMSPLIT are underlying procedures for the Variable Selection node.

---

### Syntax: DMINE Procedure

```
PROC DMINE <option(s)>;
    FREQ variable;
    TARGET variable;
    VARIABLES variable-list;
    WEIGHT variable;
```

---

## PROC DMINE Statement

Invokes the DMINE procedure.

**PROC DMINE** <option(s)>;

### Required Arguments

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

Identifies the input data set.

**Default:** none

**DMDBCAT**=<libref.> SAS-catalog

Identifies an input catalog of metadata catalog of the input data source. The catalog contains important information (for example, the range of variables, number of missing values of each variable, moments of variables) that is used by many other Enterprise Miner procedures that require a DMDB data set.

**Default:** None.

### Options

**COVOUT**

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

**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.

**MAXROWS**=*value*

Specifies the upper bound for the number of independent variables selected for the model. This is an upper bound for the number of rows and columns of the X'X matrix of the regression problem.

**Default:** 3000. This means that for most models, the MINR2 and STOPR2 settings will determine the number of selected independent variables. The X'X matrix used for the stepwise regression requires  $n * \left(\frac{n+1}{2}\right)$  double precision values storage in RAM, where *n* is the number of rows in the matrix. (This corresponds to 3000 \* 1500 \* 8 bytes (which is about 36 megabytes) of RAM needed for storage.)

**MINR2**=*value*

Specifies a lower bound for the individual R-square value of a variable to be eligible for the model selection process. Variables with R-square values greater than or equal to *value* are included in the selection process.

**Definition:** R-square is the ratio of the model sum of squares (SS) to the total sum of squares. It measures the sequential improvement in the model as input variables are selected.

**Default:**  $10^{-5}$

**NOAOV16**

By default, the DMINE procedure creates the AOV16 variables, calculates their R-squares with the target variable, and then uses the remaining significant variables in the final forward stepwise selection process. The interval scaled variables are grouped into categories to create the AOV16 variables. The range of interval scaled variables can be equally divided into 16 categories and each observation (value) of the variable is then mapped into one of these categories. The NOAOV16 option prevents the procedure from including the AOV16 variables in the final stepwise selection process. Note that the R-square value is calculated for each AOV16 variable even if you specify the NOAOV16 option.

**Definition:** The DMINE procedure organizes numeric variables into 16 equally spaced groups or bins called AOV16 variables. The AOV16 variables are created to help identify non-linear relationships with the target. Bins that have zero observations are eliminated; therefore, an AOV16 variable can have fewer than 16 bins.

**Default:** Create the AOV16 variables. Note that there is not an AOV16 option, only a NOAOV16 option to prevent these variables from being used in the final forward stepwise selection process.

### NOINTER

Specifies not to consider interactions between categories (that is, a two-way interaction) of CLASS variables in the process of variable selection.

**Definition:** A two-way interaction measures the effect of a classification input variable across the levels of another classification variable. For example, credit worthiness might not be consistent across job classifications. The lack of uniformity in the response might signify a credit worthiness by job interaction.

**Default:** Two-way interactions between categories of the class variables are considered in the variable selection process. Note that the two-way interactions can dramatically increase the processing time of the DMINE procedure.

### NOMONITOR

Suppresses the output of the status monitor that indicates the progress made in the computations.

**Default:** The output of the status monitor is displayed.

### NOPRINT

Suppresses all output printed in the output window.

**Default:** The output is printed to the output window.

### OUTEST=<libref.> SAS-data-set

Specifies the name of an output data set which contains the estimation and fit statistics.

### OUTGROUP=<libref.> SAS-data-set

Contains information on:

- 1 AOV16\_intervalvar variables to show individual bin values for the intervalvar.
- 2 G\_classvar variables to show group numbers for individual class levels of classvar.
- 3 GI\_classvar1\_classvar2 variables showing groupings for all interactions between levels of classvar1 and classvar2.

The OUTGROUP data set can be useful in interpreting some parts of the generated DATA step score code (see CODE statement).

The OUTGROUP data set contains the following variables:

Variable	Type
Bin	Numeric
Group	Numeric
Level	Character
Level2	Character
Name	Character
Variable	Character
Variable2	Character
Vartype	Character

**STOPR2=value**

Specifies a lower value for the incremental model R-square value at which the variable selection process is stopped.

**Default:**  $5 * 10^{-5}$

**THREADS=n, where n is an integer,  $1 \leq n \leq 64$ ; [FORCE]**

Specifies the number of computational threads to use. If used with the FORCE option, the exact number of threads specified will be used. If FORCE is not specified, this number is used as a hint. The actual number of threads is chosen based on various factor such as the size of the problem, number of computers available, memory requirement, and so on. PROC DMINE also uses threads for reading partitioned data sets created by the SPD Engine and performing preliminary computations. This is not controlled by the THREADS option.

**USEGROUPS**

PROC DMINE automatically tries to reduce the levels of each class variable to a group variable based on the relationship with the target. By doing so, observations of class variables with many categories (for example, ZIP codes) can be mapped into groups of fewer categories. If you specify the USEGROUPS option, and a class variable can be reduced to a group variable, then only the group version of the variable is considered in the model. If you omit the USEGROUPS option, then both the group variable and the original class variable are allowed in the model.

---

## CODE Statement

**CODE** *code-option(s)*;

**CODE Options****CATALOG | CAT | C= *library.catalog.entry.type***

Specifies where to write the output SAS DATA step code using the form of *library.catalog.entry.type*. The compound name can have one to four levels. The default library is determined by the SAS system option USER=, which is usually set to WORK. The default entry is SASCODE, and the default type is SOURCE.

*Note:* You cannot specify both FILE= and CATALOG= in the same CODE statement. If you specify neither, the code is written to the SAS log unless the PMML option is specified. △

**FILE=***file-name*

Specifies the file to be used for the output SAS DATA step code.

When enclosed in a quoted string, FILE= provides the path specification to an external file. For example, FILE="c:\mydir\scorecode.sas".

FILE= can also use unquoted SAS filenames of no more than eight characters. If the filename is assigned as a fileref in a FILENAME statement, the file that is specified in the FILENAME statement is opened. The special filerefs LOG and PRINT are always assigned. If the specified name is not an assigned fileref, then the specified name value is concatenated with a .txt extension before opening. For example, if FOO is not an assigned fileref, FILE=FOO would cause FOO.txt to be opened. If the specified filename has more than eight characters, an error message is printed.

**LINESIZE | LS=** *integer-value*

Use LINESIZE= to specify the line size for generated code. The permissible integer range for LINESIZE = is 64 to 254.

**Default:** 72

**PMML | XML**

Produces scoring code in Predictive Modeling Markup Language, an XML-based standard for representing data mining results. For more information, see the PMML Support in Enterprise Miner section in the Enterprise Miner 5.3 Java Help.

**PREDICT**

Use the PREDICT option to additionally generate score code to compute predicted target values, the P\_\* variables.

**RESIDUAL | NORESIDUAL**

Use the RESIDUAL option to generate residual values for the variables R\_\*, F\_\*, CL\_\*, CP\_\*, BL\_\*, BP\_\*, and ROI\_\*. If you request code for residuals and then score a data set that does not contain target values, the residuals will have missing values.

**Default:** NORESIDUAL

---

## FREQ Statement

**Alias:** FREQUENCY

**Tip:** You can specify the FREQ variable in PROC DMDB or PROC DMINE. Specify the FREQ variable in PROC DMDB so that the information is saved in the catalog and so that the variable is automatically used as a FREQ variable in PROC DMINE. This also ensures that the FREQ variable is automatically used by all other Enterprise Miner procedures in the project.

---

**FREQ** *variable*;

### Required Argument

*variable*

Specifies one numeric (interval-scaled) FREQUENCY variable.

**Range:** Any integer. A noninteger value is truncated.

**CAUTION:**

If the **FREQ** variable specified in **PROC DMINE** differs from that in **PROC DMDB**, no **FREQ** variable will be used.  $\triangle$

## TARGET Statement

**TARGET** *variable*;

### Required Argument

*variable*

Specifies the output variable. One variable name can be specified identifying the target (response) variable for the two regressions.

## VARIABLES Statement

Alias: VAR

---

**VARIABLES** *variable-list*;

### Required Argument

*variable-list*

Specifies all the variables (numeric and categorical, that is, **INTERVAL** and **CLASS**) that can be used for independent variables in the prediction or modeling of the target variable.

## WEIGHT Statement

Alias: WEIGHTS

**Tip:** Specify the **WEIGHT** variable in **PROC DMDB** so that the information is saved in the catalog and so that the variable is used automatically as a **WEIGHT** variable in **PROC DMINE**.

---

**WEIGHT** *variable*;

### Required Argument

**variable**

Specifies one numeric (interval-scaled) variable that is used to weight the input variables.

---

## Details: DMINE Procedure

PROC DMINE performs the following two tasks:

- 1 PROC DMINE first computes a forward stepwise least squares regression. In each step, an independent variable is selected, which contributes maximally to the model R-square value. Two parameters, MINR2 and STOPR2, can be specified to guide the variable selection process.

**MINR2**

If a variable has an individual R-square value smaller than MINR2, the variable is not considered for selection into the model.

**STOPR2**

A second test is performed using the STOPR2 value: the remaining independent variable with the largest contribution to the model R-square is added to the model. If the resulting global R-square value changes from its former value by less than the STOPR2 value, then the stepwise regression is terminated.

- 2 For a binary target (CLASS response variable), a fast algorithm for (approximate) logistic regression is computed in the second part of PROC DMINE. The independent variable is the prediction from the former least squares regression. Since only one regression variable is used in the logistic regression, only two parameters are estimated, the intercept and slope. The range of predicted values is divided into a number of equidistant intervals (knots), on which the logistic function is interpolated.

If NOPRINT is not specified, a table is printed indicating the accuracy of the prediction of the target.

---

## Fit Statistics for OUTEST and OUTFIT Data Sets

The OUTEST= 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.

**Table 7.1** Fit statistics for the Training Data

Fit Statistic	Training Data
_AIC_	Train: Akaike's Information Criterion
_ASE_	Train: Average Squared Error
_AVERR_	Train: Average Error Function
_DFE_	Train: Degrees of Freedom for Error
_DFM_	Train: Model Degrees of Freedom

<b>Fit Statistic</b>	<b>Training Data</b>
_DFT_	Train: Total Degrees of Freedom
_DIV_	Train: Divisor for ASE
_ERR_	Train: Error Function
_FPE_	Train: Final Prediction Error
_MAX_	Train: Maximum Absolute Error
_MSE_	Train: Mean Square Error
_NOBS_	Train: Sum of Frequencies
_NW_	Train: Number of Estimate Weights
_RASE_	Train: Root Average Sum of Squares
_RFPE_	Train: Root Final Prediction Error
_RMSE_	Train: Root Mean Squared Error
_SBC_	Train: Schwarz's Bayesian Criterion
_SSE_	Train: Sum of Squared Errors
_SUMW_	Train: Sum of Case Weights Times Frequency
_MISC_	Train: Misclassification Rate

**Table 7.2** Fit statistics for the Test Data

<b>Fit Statistic</b>	<b>Test Data</b>
_TASE_	Test: Average Squared Error
_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
_TMISL_	Test: Lower 95% Confidence Limit for TMISC
_TMISU_	Test: Upper 95% Confidence Limit for TMISC

**Table 7.3** Fit Statistics for Validation Data

Fit Statistic	Validation Data
_VASE_	Valid: Average Squared Error
_VAVERR_	Valid: Average Error Function
_VDIV_	Valid: Divisor for VASE
_VERR_	Valid: Error Function
_VMAX_	Valid: Maximum Absolute Error
_VMSE_	Valid: Mean Square Error
_VNOBS_	Valid: Sum of Frequencies
_VRASE_	Valid: Root Average Squared Error
_VRMSE_	Valid: Root Mean Square Error
_VSSE_	Valid: Sum of Square Errors
_VSUMW_	Valid: Sum of Case Weights Times Frequency
_VMISC_	Valid: Misclassification Rate

## Missing Values

Missing values are handled differently, depending on the type of variable.

- Missing values in categorical variables are replaced with a new category that represents missing values.
- Missing values in noncategorical variables are replaced with the mean.
- Observations with missing target values are dropped from the data.

## Examples: DMINE Procedure

The following examples were executed on the Windows XP Professional operating system; the version of the SAS System was 9.1.3.

### Example 1: Modeling a Continuous Target with the DMINE Procedure (Simple Selection Settings)

#### Features:

- Setting the MINR2= and STOPR2= cutoff values.
- Specifying the target and input variables.
- Excluding the AOV16 variables by specifying the NOAOV16 option.
- Excluding the two-way class interactions by specifying the NOINTER option.

As a marketing analyst at a catalog company, you want to quickly identify the inputs that best predict the dollar amount that customers will purchase from your new fall

outerwear catalog. The fictitious catalog mailing data set is named SAMPSIO.DMEXA1 (stored in the sample library). The data set contains 1,966 customer cases. The interval target AMOUNT contains the purchase amount in dollars.

There are 48 input variables available for predicting the target. Note that PURCHASE is a binary target that is modeled in “Example 3: Modeling a Binary Target with the DMINE Procedure”. ACCTNUM is an id variable, which is not a suitable input variable.

## Program

Before you analyze the data using the DMINE procedure, you must create a DMDB catalog that contains the metadata for the input data set being analyzed. For more information about how to do this, see “Example 1: Getting Started with the DMDB Procedure” in the DMDB procedure documentation.

```
proc dmdb batch data=sampsio.dmexal out=dmbexal dmdbcat=catexal;
  id acctnum;
  var amount income homeval frequent recency age
      domestic apparel leisure promo7 promo13 dpml2
      county return mensware flatware homeacc lamps
      linens blankets towels outdoor coats wcoat
      wappar hhappar jewelry custdate numkids travtime job;
  class purchase(desc) marital ntitle gender telind
      aprtmnt snglmom mobile kitchen luxury dishes tmktord
      statecod race origin heat numcars edlevel;
run;
```

The PROC DMINE 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.

```
proc dmine data=dmbexal dmdbcat=catexal
```

The MINR2= option specifies a lower bound for the individual R-square value to be eligible for the model selection process. Variables with R2 values less than the MINR2 cutoff are not entered into the model. The STOPR2 specifies a lower value for the incremental model R-square value at which the forward selection process is stopped.

```
  minr2=0.020 stopr2=0.0050
```

The NOAOV16 option prevents the DMINE procedure from including the AOV16 variables in the final forward stepwise selection process.

```
  noaov16
```

The NOINTER option prevents the use of two-way interactions between categories of class variables in the selection process.

```
  nointer;
```

The VAR statement lists the numeric and categorical inputs (independent variables).

```
var  income homeval frequent recency age
     domestic apparel leisure promo7 promo13 dpm12
     county return mensware flatware homeacc lamps
     linens blankets towels outdoor coats wcoat
     wappar hhappar jewelry custdate numkids travtime job
     marital ntitle gender telind aprtmnt snglmom mobile
     kitchen luxury dishes tmktord statecod race origin heat
     numcars edlevel;
```

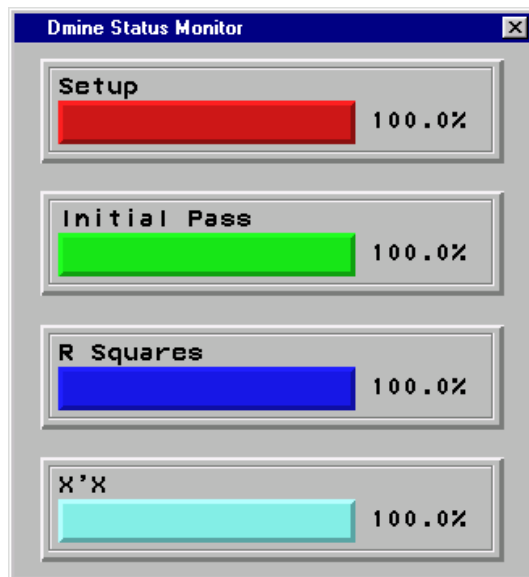
The TARGET statement defines the target (response) variable.

```
target amount;
title 'DMINE: Continuous Target';
run;
```

## Output

### DMINE Status Monitor

When you invoke the DMINE procedure, the Dmine Status Monitor window appears. This window monitors the execution time of the procedure. To suppress the display of this window, specify the NOMONITOR option on the PROC DMINE statement.



### Partial Listing of R-Squares for Target Variable

This section of the output ranks all model effects by their R-square values. The degrees of freedom (DF) associated with each effect is also listed. Effects that have an R-square value less than the MINR2 = cutoff value are not chosen for the model. These effects are labeled as "R2 < MINR2" in the table. The remaining significant variables are analyzed in a subsequent forward selection regression.

There are four types of model effects:

- Class* effects are estimated for each class variable and all possible two-factor interactions. The R-square statistic is calculated for each class effect using a

one-way analysis of variance. Two-factor interaction effects are constructed by combining all possible levels of each class variable into one term. Because the NOINTER option was specified, the two-factor interactions are not used in the final forward stepwise regression. The degrees of freedom for a class effect is equal to: (the number of unique factor levels minus 1). For two-factor interactions, the degrees of freedom is equal to: (the number of levels in factor A multiplied by the number of levels in factor B minus 1).

- *Group* effects are created by reducing each class effect through an analysis of means. The degrees of freedom for each group effect is equal to the number of levels.
- *VAR* effects are estimated from interval variables as standard regression inputs. A simple linear regression is performed to determine the R<sup>2</sup> statistic for interval inputs. The degrees of freedom is always equal to 1.
- *AOV16* effects are calculated as a result of grouping numeric variables into a maximum of 16 equally spaced buckets. AOV16 effects might account for possible non-linearity in the target variable AMOUNT. The degrees of freedom are calculated as the number of groups. Because the NOAOV16 option was specified, the AOV16 variables are not used in the final forward stepwise regression.

Note that the original input LEISURE has the largest R-square statistic with the target. Several AOV16 and group variables have large R-square values, but these effects are not used in the final forward stepwise regression.

#### The DMINE Procedure

##### R-Squares for Target Variable: AMOUNT

Effect	DF	R-Square
Var: LEISURE	1	0.482651
AOV16: APPAREL	12	0.476210
Class: KITCHEN*STATECOD	197	0.427107
AOV16: LEISURE	8	0.426806
Group: KITCHEN*STATECOD	9	0.420965
Class: KITCHEN*LUXURY	16	0.401859
Var: APPAREL	1	0.400057
Group: KITCHEN*LUXURY	5	0.395858
AOV16: DOMESTIC	15	0.386944
Var: DOMESTIC	1	0.365188
AOV16: FREQUENT	11	0.341752
Class: LUXURY*STATECOD	101	0.333484
Group: LUXURY*STATECOD	6	0.328393
Class: LUXURY*TMKTORD	7	0.321248
Group: LUXURY*TMKTORD	4	0.317663
Class: KITCHEN*DISHES	42	0.312778
Class: MARITAL*KITCHEN	17	0.308421
Group: KITCHEN*DISHES	5	0.306628
Var: FREQUENT	1	0.304805
Class: TMKTORD*STATECOD	110	0.304564
Group: MARITAL*KITCHEN	5	0.303297
Group: TMKTORD*STATECOD	8	0.299534
Class: KITCHEN*TMKTORD	26	0.292056
Group: KITCHEN*TMKTORD	5	0.286952
AOV16: DPM12	12	0.277018
Class: NTITLE*KITCHEN	28	0.271750

Group: NTITLE*KITCHEN	6	0.269033
Class: KITCHEN*RACE	25	0.259373
Class: KITCHEN*EDLEVEL	26	0.257153
Group: KITCHEN*RACE	5	0.256847
Class: LUXURY*DISHES	15	0.254661
Group: KITCHEN*EDLEVEL	6	0.253342
Group: LUXURY*DISHES	3	0.250377
Class: LUXURY*RACE	8	0.245464
Group: LUXURY*RACE	2	0.244735
Class: KITCHEN*ORIGIN	35	0.243767
Class: LUXURY*ORIGIN	11	0.241906
Class: NTITLE*LUXURY	7	0.241903
Group: KITCHEN*ORIGIN	6	0.240676
Group: NTITLE*LUXURY	2	0.240577
Class: LUXURY*NUMCARS	6	0.239431
Class: APRTMNT*LUXURY	3	0.238892
Class: TELIND*LUXURY	3	0.238632
Class: LUXURY*EDLEVEL	7	0.238391
Class: LUXURY*HEAT	7	0.238017
Class: MOBILE*LUXURY	3	0.237998
Class: SNGLMOM*LUXURY	3	0.237730
Group: LUXURY*ORIGIN	1	0.237513
Class: KITCHEN*NUMCARS	23	0.237350
Class: MARITAL*LUXURY	3	0.237237
Class: GENDER*LUXURY	3	0.237082

## The DMINE Procedure

## R-Squares for Target Variable: AMOUNT

Effect	DF	R-Square
Group: LUXURY*EDLEVEL	1	0.236986
Group: LUXURY*HEAT	1	0.236986
Class: LUXURY	1	0.236986
Group: APRTMNT*LUXURY	1	0.236986
Group: LUXURY*NUMCARS	1	0.236986
Group: MARITAL*LUXURY	1	0.236986
Group: GENDER*LUXURY	1	0.236986
Group: TELIND*LUXURY	1	0.236986
Group: SNGLMOM*LUXURY	1	0.236986
Group: MOBILE*LUXURY	1	0.236986
Class: KITCHEN*HEAT	27	0.235776
Group: KITCHEN*NUMCARS	5	0.233440
Class: GENDER*KITCHEN	16	0.232768
Group: KITCHEN*HEAT	5	0.231237
Class: APRTMNT*KITCHEN	15	0.229631
Class: MOBILE*KITCHEN	15	0.229531
Group: GENDER*KITCHEN	5	0.228979
Class: SNGLMOM*KITCHEN	15	0.227912
Group: APRTMNT*KITCHEN	5	0.226338
Group: MOBILE*KITCHEN	5	0.225471
Class: TELIND*KITCHEN	14	0.225469

Group: SNGLMOM*KITCHEN	4	0.223912
Class: KITCHEN	9	0.223201
Group: TELIND*KITCHEN	4	0.222322
Class: TMKTORD*RACE	14	0.221117
Group: KITCHEN	4	0.220665
Group: TMKTORD*RACE	3	0.217392
Class: TMKTORD*ORIGIN	17	0.189869
Group: TMKTORD*ORIGIN	5	0.187283
Class: DISHES*TMKTORD	23	0.172790
AOV16: OUTDOOR	11	0.170436
Group: DISHES*TMKTORD	5	0.170070
Var: DPM12	1	0.161643
Class: MARITAL*TMKTORD	8	0.157953
Group: MARITAL*TMKTORD	3	0.156439
AOV16: HHAPPAR	13	0.156192
AOV16: RECENCY	14	0.156007
Class: NTITLE*TMKTORD	14	0.153028
Group: NTITLE*TMKTORD	3	0.150849
Class: APRTMNT*TMKTORD	7	0.148450
Group: APRTMNT*TMKTORD	4	0.146693
Class: TMKTORD*EDLEVEL	12	0.142642
Group: TMKTORD*EDLEVEL	4	0.141424
Class: DISHES*STATECOD	169	0.139466
Group: DISHES*STATECOD	10	0.137017
AOV16: JEWELRY	10	0.134423
Class: TMKTORD*HEAT	12	0.133375
Class: MOBILE*TMKTORD	7	0.133325
Class: STATECOD*EDLEVEL	146	0.133091
Group: MOBILE*TMKTORD	4	0.132432
Class: TMKTORD*NUMCARS	10	0.132326

Additional effects are not listed

### SS and R2 Portion for Effects Chosen for the Target

This section lists the chosen input variables from the forward stepwise regression. The table is divided into the following five columns:

- **Effect** lists the sequentially selected effects, which are ranked by the R2 statistic.
- **DF** shows the degrees of freedom associated with each model effect.
- **R--Square** measures the sequential improvement in the model as input variables are selected. Multiply the R2 statistic by 100 to express it as a percentage. You can interpret the R2 statistic for the LEISURE effect as: "48.27% of the variation in the target AMOUNT is explained by its linear relationship with LEISURE". The R2 statistic for APPAREL indicates that this effect accounts for an additional 13.44% of the target variation.
- **Sum of Squares** lists the sums of squares for each model effect.
- **Error Mean Square** lists the Error Mean Square, which measures variation due to either random error or to other inputs that are not in the model. The EMS should get smaller as important inputs are added to the model. Note that although STATECOD has an R2 value greater than the STOPR2 cutoff value of 0.005, the error mean square becomes larger when this effect enters the model.

## Effects Chosen for Target: AMOUNT

Effect	DF	R-Square	F Value	p-Value	Sum of Squares
Var: LEISURE	1	0.482651	1832.276768	<.0001	532690709
Var: APPAREL	1	0.134371	688.734726	<.0001	148302146
Class: LUXURY	1	0.091226	613.481401	<.0001	100683600
Var: DOMESTIC	1	0.048854	394.413801	<.0001	53918925
Var: DPM12	1	0.010139	85.379981	<.0001	11190481
Class: KITCHEN	9	0.008596	8.312566	<.0001	9486960
Class: STATECOD	54	0.005128	0.822385	0.8184	5659245

## Effects Chosen for Target: AMOUNT

Effect	Error Mean Square
Var: LEISURE	290726
Var: APPAREL	215325
Class: LUXURY	164118
Var: DOMESTIC	136706
Var: DPM12	131067
Class: KITCHEN	126809
Class: STATECOD	127435

*Note:* Note that the AOV16, GROUP, and two-way class interaction effects are not considered in the forward stepwise regression. Including these effects can produce a better model, but it will also increase the execution time of the DMINE procedure. To learn how to include these effects into the analysis, see Example 2.  $\Delta$

**Final ANOVA Table for the Target**

The ANOVA table is divided into the following four columns:

- Effect** labels the source of variation as Model, Error, or Total.
- DF** lists the degrees of freedom for each source of variation.
- R--Square** is the model R<sup>2</sup>, which is the ratio of the model sums of squares (SS) to the total sums of squares. In this example, the inputs collectively explain 78.10% of the total variability in the target AMOUNT.
- Sum of Squares** partitions the total target variation into portions that can be attributed to the model inputs and to error.

## The DMINE Procedure

## The Final ANOVA Table for Target: AMOUNT

Effect	DF	R-Square	Sum of Squares
Model	68	0.780964	861932067
Error	1897	.	241744735
Total	1965	.	1103676802

**SS and R2 portion for Effects not chosen for Target**

The final section lists the sums of squares and the R-square values for the effects that are not chosen in the final model.

The DMINE Procedure

Effects Not Chosen for Target: AMOUNT

Effect	DF	R-Square	F Value	p-Value	Sum of Squares
Var: FREQUENT	1	0.000191	1.658422	0.1980	211268
Group: KITCHEN	0	0	0	1.0000	0
Var: RECENCY	1	0.000000297	0.002575	0.9595	328.303498
Class: TMKTORD	4	0.000484	1.047691	0.3811	533997
Group: TMKTORD	3	0.000216	0.623363	0.5999	238457
Var: OUTDOOR	1	0.001063	9.246947	0.0024	1173287
Var: JEWELRY	1	0.004469	39.492809	<.0001	4932686
Var: HHAPPAR	1	0.002179	19.054693	<.0001	2405347
Var: TOWELS	1	0.000193	1.674163	0.1959	213272
Var: LINENS	1	0.000018064	0.156377	0.6926	19937
Var: HOMEACC	1	0.000007748	0.067067	0.7957	8550.941493
Var: LAMPS	1	0.000318	2.760456	0.0968	351453
Var: PROMO7	1	0.001217	10.589546	0.0012	1342694
Var: MENSWARE	1	0.002733	23.960472	<.0001	3016894
Var: WAPPAR	1	0.002255	19.724999	<.0001	2489091
Var: BLANKETS	1	0.000469	4.068428	0.0438	517624
Group: STATECOD	0	0	0	1.0000	0
Var: PROMO13	1	0.000582	5.047509	0.0248	641861
Class: DISHES	9	0.001805	1.742596	0.0746	1991598
Group: DISHES	4	0.001196	2.598493	0.0346	1320109
Var: COATS	1	0.001546	13.477360	0.0002	1706268

---

## Example 2: Including the AOV16 and Grouping Variables into the Analysis (Detailed Selection Settings)

**Features:**

- Omitting the NOAOV16 option to include the AOV16 variables into the analysis.
  - Specifying the USEGROUPS option to include only the group variables in the final model. If the original class variable can be reduced into a group variable that contains fewer levels, then only the group variable is considered in the final model.
  - Omitting the NOINTER option to include the two-way class interactions into the analysis.
  - Specifying the NOMONITOR option to suppress the monitor window, which displays the execution time of the procedure.
-

This example expands on the previous example by including the AOV16, GROUP, and two-way class interaction effects into the final forward stepwise analysis. Including these effects into the analysis can produce a better model, but it will also increase the execution time of the DMINE procedure.

If you have not already done so, you should submit the PROC DMDB step from Example 1 before you submit the example PROC DMINE step.

## Program

```
proc dmine data=WORK.dmbexal dmbcat=catexal
  minr2=0.020 stopr2=0.0050
```

The USEGROUPS option specifies to retain only the reduced group variables (the original class variables are not included in the model selection process). Since the NOAOV16 option was not specified, the AOV16 variables that have an R-square value less than the MINR2 cutoff are used in the final forward stepwise regression. Because the NOINTER option is omitted, the two-factor class interactions are also evaluated in the forward stepwise regression.

```
  usegroups
```

The NOMONITOR option suppresses the monitor that displays the execution status of the DMINE procedure.

```
  nomonitor;

  var  income homeval frequent recency age
       domestic apparel leisure promo7 promo13 dpml2
       county return mensware flatware homeacc lamps
       linens blankets towels outdoor coats wcoat
       wappar hhappar jewelry custdate numkids travtime job
       marital ntitle gender telind aprtmnt snglmom mobile
       kitchen luxury dishes tmktord statecod race origin heat
       numcars edlevel;

  target amount;
  title 'DMINE: Continuous Target';
  title2 'Add AOV16, GROUP, and 2--Way Interactions Effects';
run;
```

## Output

### Partial Listing of the R-Squares for the Target Variable

The DMINE Procedure

R-Squares for Target Variable: AMOUNT

Effect	DF	R-Square
Var: LEISURE	1	0.482651
AOV16: APPAREL	12	0.476210
Class: KITCHEN*STATECOD	197	0.427107
AOV16: LEISURE	8	0.426806
Group: KITCHEN*STATECOD	9	0.420965
Class: KITCHEN*LUXURY	16	0.401859

Var:	APPAREL	1	0.400057
Group:	KITCHEN*LUXURY	5	0.395858
AOV16:	DOMESTIC	15	0.386944
Var:	DOMESTIC	1	0.365188
AOV16:	FREQUENT	11	0.341752
Class:	LUXURY*STATECOD	101	0.333484
Group:	LUXURY*STATECOD	6	0.328393
Class:	LUXURY*TMKTORD	7	0.321248
Group:	LUXURY*TMKTORD	4	0.317663
Class:	KITCHEN*DISHES	42	0.312778
Class:	MARITAL*KITCHEN	17	0.308421
Group:	KITCHEN*DISHES	5	0.306628
Var:	FREQUENT	1	0.304805
Class:	TMKTORD*STATECOD	110	0.304564
Group:	MARITAL*KITCHEN	5	0.303297
Group:	TMKTORD*STATECOD	8	0.299534
Class:	KITCHEN*TMKTORD	26	0.292056
Group:	KITCHEN*TMKTORD	5	0.286952
AOV16:	DPM12	12	0.277018
Class:	NTITLE*KITCHEN	28	0.271750
Group:	NTITLE*KITCHEN	6	0.269033
Class:	KITCHEN*RACE	25	0.259373
Class:	KITCHEN*EDLEVEL	26	0.257153
Group:	KITCHEN*RACE	5	0.256847
Class:	LUXURY*DISHES	15	0.254661
Group:	KITCHEN*EDLEVEL	6	0.253342
Group:	LUXURY*DISHES	3	0.250377
Class:	LUXURY*RACE	8	0.245464
Group:	LUXURY*RACE	2	0.244735
Class:	KITCHEN*ORIGIN	35	0.243767
Class:	LUXURY*ORIGIN	11	0.241906
Class:	NTITLE*LUXURY	7	0.241903
Group:	KITCHEN*ORIGIN	6	0.240676
Group:	NTITLE*LUXURY	2	0.240577
Class:	LUXURY*NUMCARS	6	0.239431
Class:	APRTMNT*LUXURY	3	0.238892
Class:	TELIND*LUXURY	3	0.238632
Class:	LUXURY*EDLEVEL	7	0.238391
Class:	LUXURY*HEAT	7	0.238017
Class:	MOBILE*LUXURY	3	0.237998
Class:	SNGLMOM*LUXURY	3	0.237730
Group:	LUXURY*ORIGIN	1	0.237513
Class:	KITCHEN*NUMCARS	23	0.237350
Class:	MARITAL*LUXURY	3	0.237237

## The DMINE Procedure

R-Squares for Target Variable: AMOUNT

Effect	DF	R-Square
Class: GENDER*LUXURY	3	0.237082
Group: LUXURY*EDLEVEL	1	0.236986

Group: LUXURY*HEAT	1	0.236986
Class: LUXURY	1	0.236986
Group: APRTMNT*LUXURY	1	0.236986
Group: LUXURY*NUMCARS	1	0.236986
Group: MARITAL*LUXURY	1	0.236986
Group: GENDER*LUXURY	1	0.236986
Group: TELIND*LUXURY	1	0.236986
Group: SNGLMOM*LUXURY	1	0.236986
Group: MOBILE*LUXURY	1	0.236986
Class: KITCHEN*HEAT	27	0.235776
Group: KITCHEN*NUMCARS	5	0.233440
Class: GENDER*KITCHEN	16	0.232768
Group: KITCHEN*HEAT	5	0.231237
Class: APRTMNT*KITCHEN	15	0.229631
Class: MOBILE*KITCHEN	15	0.229531
Group: GENDER*KITCHEN	5	0.228979
Class: SNGLMOM*KITCHEN	15	0.227912
Group: APRTMNT*KITCHEN	5	0.226338
Group: MOBILE*KITCHEN	5	0.225471
Class: TELIND*KITCHEN	14	0.225469
Group: SNGLMOM*KITCHEN	4	0.223912
Class: KITCHEN	9	0.223201
Group: TELIND*KITCHEN	4	0.222322
Class: TMKTORD*RACE	14	0.221117
Group: KITCHEN	4	0.220665
Group: TMKTORD*RACE	3	0.217392
Class: TMKTORD*ORIGIN	17	0.189869
Group: TMKTORD*ORIGIN	5	0.187283
Class: DISHES*TMKTORD	23	0.172790
AOV16: OUTDOOR	11	0.170436
Group: DISHES*TMKTORD	5	0.170070
Var: DPM12	1	0.161643
Class: MARITAL*TMKTORD	8	0.157953
Group: MARITAL*TMKTORD	3	0.156439
AOV16: HHAPPAR	13	0.156192
AOV16: RECENCY	14	0.156007
Class: NTITLE*TMKTORD	14	0.153028
Group: NTITLE*TMKTORD	3	0.150849
Class: APRTMNT*TMKTORD	7	0.148450
Group: APRTMNT*TMKTORD	4	0.146693
Class: TMKTORD*EDLEVEL	12	0.142642
Group: TMKTORD*EDLEVEL	4	0.141424
Class: DISHES*STATECOD	169	0.139466
Group: DISHES*STATECOD	10	0.137017
AOV16: JEWELRY	10	0.134423
Class: TMKTORD*HEAT	12	0.133375
Class: MOBILE*TMKTORD	7	0.133325
Class: STATECOD*EDLEVEL	146	0.133091

Additional effects are not listed

**SS and R2 Portion for Effects Chosen for the Target**

As in the Example 1 analysis, the original input LEISURE appears to be the most important predictor of how much a customer spends. Notice that the remaining chosen effects are either AOV16 variables or group variables.

The DMINE Procedure

Effects Chosen for Target: AMOUNT

Effect	DF	R-Square	F Value	p-Value	Sum of Squares
Var: LEISURE	1	0.482651	1832.276768	<.0001	532690709
Group: KITCHEN*LUXURY	5	0.159019	173.871642	<.0001	175505155
AOV16: APPAREL	12	0.111100	72.911713	<.0001	122618569
AOV16: DOMESTIC	15	0.041460	25.951392	<.0001	45758285
AOV16: WAPPAR	7	0.012199	17.330293	<.0001	13463471
Group: LUXURY*STATECOD	6	0.009566	16.626635	<.0001	10557347
AOV16: LEISURE	6	0.007401	13.360939	<.0001	8168061

Effects Chosen for Target: AMOUNT

Effect	Error Mean Square
Var: LEISURE	290726
Group: KITCHEN*LUXURY	201879
AOV16: APPAREL	140145
AOV16: DOMESTIC	117549
AOV16: WAPPAR	110982
Group: LUXURY*STATECOD	105828
AOV16: LEISURE	101890

**Final ANOVA Table for the Target**

The R-square value of 0.8234 is slightly larger than the 0.7810 value obtained from the initial analysis, which did not include the AOV16 or group variables.

The DMINE Procedure

The Final ANOVA Table for Target: AMOUNT

Effect	DF	R-Square	Sum of Squares
Model	52	0.823395	908761597
Error	1913	.	194915205
Total	1965	.	1103676802

**Partial Listing of the SS and R2 portion for Effects Not Chosen for Target Table**

The DMINE Procedure

Effects Not Chosen for Target: AMOUNT

Effect	DF	R-Square	F Value	p-Value	Sum of Squares
Group: KITCHEN*STATECOD	7	0.004018	6.339610	<.0001	4434939
Var: APPAREL	1	0.002421	26.578454	<.0001	2672342
Var: DOMESTIC	1	0.003524	38.931050	<.0001	3889555
AOV16: FREQUENT	11	0.003076	3.065222	0.0004	3395145
Group: LUXURY*TMKTORD	4	0.000735	1.994483	0.0929	811183
Group: KITCHEN*DISHES	4	0.001166	3.172136	0.0131	1286988
Var: FREQUENT	1	0.000691	7.510696	0.0062	762668
Group: MARITAL*KITCHEN	4	0.001050	2.855330	0.0225	1159219
Group: TMKTORD*STATECOD	6	0.000987	1.787006	0.0980	1089776
Group: KITCHEN*TMKTORD	5	0.001704	3.718159	0.0024	1880850
AOV16: DPM12	11	0.004189	4.200942	<.0001	4623270
Group: NTITLE*KITCHEN	6	0.000859	1.552696	0.1572	947581
Group: KITCHEN*RACE	5	0.000594	1.287343	0.2665	655343
Group: KITCHEN*EDLEVEL	6	0.000983	1.778125	0.0998	1084391
Group: LUXURY*DISHES	3	0.000234	0.844223	0.4696	258116
Group: LUXURY*RACE	2	0.000380	2.060645	0.1277	419452
Group: KITCHEN*ORIGIN	6	0.001012	1.832568	0.0891	1117402
Group: NTITLE*LUXURY	2	0.000378	2.051720	0.1288	417640
Group: LUXURY*ORIGIN	1	0.000011145	0.120670	0.7283	12301
Group: LUXURY*EDLEVEL	1	0.000029538	0.319844	0.5718	32600
Group: LUXURY*HEAT	1	0.000029538	0.319844	0.5718	32600
Class: LUXURY	1	0.000029538	0.319844	0.5718	32600
Group: APRTMNT*LUXURY	1	0.000029538	0.319844	0.5718	32600
Group: LUXURY*NUMCARS	1	0.000029538	0.319844	0.5718	32600
Group: MARITAL*LUXURY	1	0.000029538	0.319844	0.5718	32600
Group: GENDER*LUXURY	1	0.000029538	0.319844	0.5718	32600
Group: TELIND*LUXURY	1	0.000029538	0.319844	0.5718	32600
Group: SNGLMOM*LUXURY	1	0.000029538	0.319844	0.5718	32600
Group: MOBILE*LUXURY	1	0.000029538	0.319844	0.5718	32600
Group: KITCHEN*NUMCARS	5	0.000523	1.132970	0.3406	576990
Group: KITCHEN*HEAT	4	0.000827	2.246110	0.0619	913044
Group: GENDER*KITCHEN	4	0.000526	1.426289	0.2227	580780
Group: APRTMNT*KITCHEN	4	0.000476	1.290712	0.2714	525722
Group: MOBILE*KITCHEN	4	0.000475	1.286757	0.2730	524116
Group: SNGLMOM*KITCHEN	3	0.000448	1.617937	0.1832	494075
Group: TELIND*KITCHEN	3	0.000452	1.632653	0.1798	498557
Group: KITCHEN	3	0.000448	1.617565	0.1833	493962
Group: TMKTORD*RACE	2	0.000422	2.290479	0.1015	466124
Group: TMKTORD*ORIGIN	5	0.001013	2.200761	0.0518	1117668
AOV16: OUTDOOR	11	0.002135	2.115533	0.0166	2355951
Group: DISHES*TMKTORD	5	0.001168	2.540563	0.0267	1289097
Var: DPM12	1	0.003065	33.766034	<.0001	3382479
Group: MARITAL*TMKTORD	3	0.000785	2.841335	0.0366	866008
AOV16: HHAPPAR	13	0.001816	1.518707	0.1030	2004564
AOV16: RECENCY	14	0.002323	1.808118	0.0324	2564039
Group: NTITLE*TMKTORD	3	0.000899	3.258263	0.0208	992437
Group: APRTMNT*TMKTORD	4	0.000834	2.264073	0.0601	920311
Group: TMKTORD*EDLEVEL	4	0.000714	1.937697	0.1016	788181
Group: DISHES*STATECOD	10	0.001055	1.143145	0.3257	1163877

Additional effects not listed.

## Example 3: Modeling a Binary Target with the DMINE Procedure

### Features:

- Setting the MINR2= and STOPR2= cutoff values.
- Specifying the target and input variables.

As a marketing analyst at a catalog company, you want to determine the inputs that best predict whether a customer will make a purchase from your new fall outerwear catalog. The fictitious catalog mailing data set is named SAMPSIO.DMEXA1 (stored in the sample library). The data set contains 1,966 customer cases. The binary target (PURCHASE) contains a formatted value of “Yes” if a purchase was made and a formatted value of “No” if a purchase was not made.

There are 48 input variables available for predicting the target. Note that AMOUNT is an interval target that is modeled in Examples 1 and 2 of this chapter. ACCTNUM is an id variable, which is not a suitable input variable.

### Program

Before you analyze the data using the DMINE procedure, you must create the DMDB catalog that contains the metadata for the input data set being analyzed. For more information about how to do this, see “Example 1: Getting Started with the DMDB Procedure” in the DMDB procedure documentation. Since the (DESCENDING) ORDER option is specified for the target PURCHASE on the CLASS statement, the DMINE procedure reads this encoded information from the metadata and then models the probability that a customer will make a purchase (PURCHASE = 'Yes'). The default ORDER is set to ASCENDING for all class variables.

```
proc dmdb batch data=sampsio.dmxal out=dmbexal dmbcat=catexal;
  id acctnum;
  var amount income homeval frequent recency age
      domestic apparel leisure promo7 promo13 dpml2
      county return mensware flatware homeacc lamps
      linens blankets towels outdoor coats wcoat
      wappar happar jewelry custdate numkids travtime job;
  class purchase(desc) marital ntitle gender telind
      aprtmnt snlglmom mobile kitchen luxury dishes tmktord
      statecod race origin heat numcars edlevel;
run;
```

The PROC DMINE statement invokes the procedure. The DATA= option identifies the training data set that is used to fit the model. The DMDBCAT= option identifies the DMDB training data catalog.

```
proc dmine data=WORK.dmbexal dmbcat=catexal
```

The MINR2= option specifies a lower bound for the individual R-square value to be eligible for the model selection process. Variables with R2 values less than the MINR2 cutoff are not entered into the model. The STOPR2 specifies a lower value for the incremental model R-square value at which the forward selection process is stopped.

```
minr2=0.020 stopr2=0.0050;
```

The VAR statement specifies the numeric and categorical inputs (independent variables).

```
var income homeval frequent recency age
    domestic apparel leisure promo7 promo13 dpml2
    county return mensware flatware homeacc lamps
    linens blankets towels outdoor coats wcoat
    wappar hhappar jewelry custdate numkids travtime job
    marital ntitle gender telind aprtmnt snglmom mobile
    kitchen luxury dishes tmktord statecod race origin heat
    numcars edlevel;
```

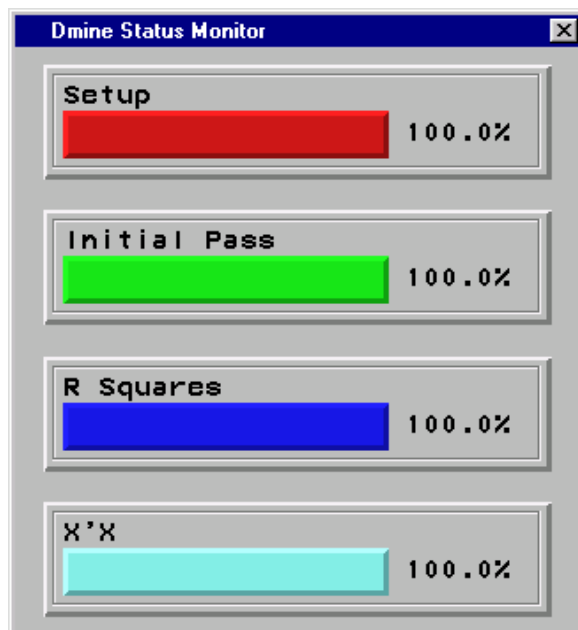
The TARGET statement defines the target (response) variable.

```
target purchase;
title 'DMINE: Binary Target';
run;
```

## Output

### DMINE Status Monitor

When you invoke the DMINE procedure, the Dmine Status Monitor window appears. This window monitors the execution time of the procedure.



### Partial Listing of the R-Squares for the Target Variable

This section of the output ranks all model effects by their R-square values. The degrees of freedom (DF) associated with each effect is also listed. The significant variables are analyzed in a subsequent forward stepwise regression. Non-significant variables are labeled as having an R2 value less than the MINR2 cutoff; these variables are not chosen in the final model.

There are four types of model effects:

- *Class* effects are estimated for each class variable and all possible two-factor class interactions. The R-square statistic is calculated for each class effect using a one-way analysis of variance. Two-factor interaction effects are constructed by combining all possible levels of each class variable into one term. The degrees of freedom for a class effect is equal to: (the number of unique factor levels minus 1). For two-factor interactions, the degrees of freedom is equal to: (the number of levels in factor A multiplied by the number of levels in factor B minus 1). You can omit the two-factor interaction effects from the final stepwise analysis by specifying the NOINTER option on the PROC DMINE statement.
- *Group* effects are created by reducing each class effect through an analysis of means. The degrees of freedom for each group effect is equal to the number of levels. Since the USEGROUPS option was not specified in the PROC DMINE statement, the group effects and the original class effects can be used in the final model.
- *VAR* effects are estimated from interval variables as standard regression inputs. A simple linear regression is performed to determine the R2 statistic for interval inputs. The degrees of freedom is always equal to 1.
- *AOV16* effects are calculated as a result of grouping numeric variables into a maximum of 16 equally spaced buckets. AOV16 effects might account for possible non-linearity in the target variable PURCHASE. The degrees of freedom are calculated as the number of groups. The AOV16 variables can be expensive to compute. You can prevent these variables from being evaluated in the forward stepwise regression by specifying the NOAOV16 option in the PROC DMINE statement.

For this example, the class KITCHEN\*STATECOD interaction has the largest R2 with the target PURCHASE. Class and group interactions composed of the same terms have very similar R2 values. Of all the AOV16 variables, FREQUENT has the largest R2 value with the target.

#### The DMINE Procedure

##### R-Squares for Target Variable: PURCHASE

Effect	DF	R-Square
Class: KITCHEN*STATECOD	197	0.104515
Group: KITCHEN*STATECOD	8	0.102622
Class: NTITLE*STATECOD	191	0.095532
Group: NTITLE*STATECOD	9	0.094020
Class: STATECOD*ORIGIN	166	0.091665
Group: STATECOD*ORIGIN	8	0.089885
Class: DISHES*STATECOD	169	0.087521
Group: DISHES*STATECOD	8	0.086178
Class: STATECOD*EDLEVEL	146	0.073597
Group: STATECOD*EDLEVEL	9	0.072271
Class: STATECOD*HEAT	141	0.064292
Group: STATECOD*HEAT	9	0.063314

Class: LUXURY*STATECOD	101	0.060728
Class: STATECOD*NUMCARS	121	0.059681
Group: LUXURY*STATECOD	10	0.059665
AOV16: FREQUENT	11	0.059157
Group: STATECOD*NUMCARS	9	0.058523
Class: STATECOD*RACE	105	0.057520
Class: TMKTORD*STATECOD	110	0.057061
Group: STATECOD*RACE	9	0.056617
Group: TMKTORD*STATECOD	8	0.056178
Class: MARITAL*STATECOD	107	0.053660
Group: MARITAL*STATECOD	10	0.052603
Class: GENDER*STATECOD	104	0.051421
Group: GENDER*STATECOD	10	0.050504
Var: FREQUENT	1	0.049836
Class: MOBILE*STATECOD	103	0.046757
Group: MOBILE*STATECOD	10	0.046089
Class: SNGLMOM*STATECOD	88	0.045005
Group: SNGLMOM*STATECOD	9	0.044277
Class: APRTMNT*STATECOD	87	0.041434
Group: APRTMNT*STATECOD	8	0.040680
Class: TELIND*STATECOD	79	0.039984
Group: TELIND*STATECOD	8	0.039344
AOV16: RECENCY	14	0.037393
AOV16: DOMESTIC	15	0.035231
AOV16: APPAREL	12	0.034342
Class: KITCHEN*DISHES	42	0.030692
Group: KITCHEN*DISHES	6	0.030357
Var: RECENCY	1	0.029513
Var: DOMESTIC	1	0.028953
Class: KITCHEN*HEAT	27	0.027990
Group: KITCHEN*HEAT	6	0.027514
Class: KITCHEN*LUXURY	16	0.027035
Group: KITCHEN*LUXURY	3	0.026644
Class: STATECOD	54	0.026485
Group: STATECOD	9	0.025986
Var: APPAREL	1	0.025248
Class: MARITAL*KITCHEN	17	0.023754
Class: NTITLE*KITCHEN	28	0.023727
Group: MARITAL*KITCHEN	4	0.023332

## The DMINE Procedure

R-Squares for Target Variable: PURCHASE

Effect	DF	R-Square
Group: NTITLE*KITCHEN	6	0.023318
Class: KITCHEN*TMKTORD	26	0.023173
Group: KITCHEN*TMKTORD	6	0.022876
Class: KITCHEN*ORIGIN	35	0.022162
Class: KITCHEN*NUMCARS	23	0.022065
Group: KITCHEN*ORIGIN	6	0.021850
Class: KITCHEN*EDLEVEL	26	0.021744

Group: KITCHEN*NUMCARS	6	0.021660	
Group: KITCHEN*EDLEVEL	7	0.021648	
Class: KITCHEN*RACE	25	0.021408	
Class: TELIND*KITCHEN	14	0.021264	
Group: TELIND*KITCHEN	5	0.021058	
Group: KITCHEN*RACE	6	0.021017	
AOV16: TOWELS	10	0.020751	
AOV16: OUTDOOR	11	0.020078	
Class: GENDER*KITCHEN	16	0.019668	R2 < MINR2
Group: GENDER*KITCHEN	5	0.019475	R2 < MINR2
Class: LUXURY*DISHES	15	0.019254	R2 < MINR2
Group: LUXURY*DISHES	5	0.019026	R2 < MINR2
Class: APRTMNT*KITCHEN	15	0.018868	R2 < MINR2
Class: SNGLMOM*KITCHEN	15	0.018747	R2 < MINR2
Group: APRTMNT*KITCHEN	4	0.018564	R2 < MINR2
AOV16: HHAPPAR	13	0.018516	R2 < MINR2
AOV16: JEWELRY	10	0.018476	R2 < MINR2
Group: SNGLMOM*KITCHEN	4	0.018415	R2 < MINR2
Class: MOBILE*KITCHEN	15	0.018367	R2 < MINR2
Group: MOBILE*KITCHEN	5	0.018160	R2 < MINR2
AOV16: LAMPS	12	0.017806	R2 < MINR2
Class: KITCHEN	9	0.017419	R2 < MINR2
AOV16: LINENS	13	0.017313	R2 < MINR2
AOV16: PROMO13	15	0.017200	R2 < MINR2
Group: KITCHEN	3	0.017077	R2 < MINR2
AOV16: BLANKETS	11	0.016874	R2 < MINR2
Class: LUXURY*ORIGIN	11	0.016788	R2 < MINR2
Group: LUXURY*ORIGIN	5	0.016662	R2 < MINR2
Var: OUTDOOR	1	0.016601	R2 < MINR2
Class: DISHES*HEAT	24	0.016527	R2 < MINR2
Var: HHAPPAR	1	0.016393	R2 < MINR2
Group: DISHES*HEAT	8	0.016353	R2 < MINR2
AOV16: PROMO7	15	0.015782	R2 < MINR2
Var: LEISURE	1	0.015420	R2 < MINR2
Class: DISHES*ORIGIN	31	0.015407	R2 < MINR2
Class: LUXURY*EDLEVEL	7	0.015258	R2 < MINR2
Group: DISHES*ORIGIN	7	0.015221	R2 < MINR2
Class: NTITLE*DISHES	23	0.015182	R2 < MINR2
Group: LUXURY*EDLEVEL	5	0.015145	R2 < MINR2
Group: NTITLE*DISHES	7	0.015059	R2 < MINR2
Class: DISHES*TMKTORD	23	0.014731	R2 < MINR2
Group: DISHES*TMKTORD	5	0.014576	R2 < MINR2
Class: TELIND*LUXURY	3	0.014330	R2 < MINR2
Group: TELIND*LUXURY	2	0.014322	R2 < MINR2

Additional Effects Are Not Listed

### SS and R2 Portion for Effects Chosen for Target

This section lists the chosen input variables from the forward stepwise regression. The table is divided into the following five columns:

- **Effect** lists the sequentially selected effects, which are ranked by the R-square statistic.
- **DF** shows the degrees of freedom associated with each model effect.

- **R-Square** measures the sequential improvement in the model as input variables are selected. Multiply the R2 statistic by 100 to express it as a percentage. You can interpret the R2 statistic for the KITCHEN\*STATECOD interaction as "10.45% of the variation in the target PURCHASE is explained by its linear relationship with this effect". The R2 statistic for NTITLE\*STATECOD indicates that this two-factor interaction accounts for an additional 6.38% of the target variation.
- **Sum of Squares** lists the sums of squares for each model effect.
- **Error Mean Square** lists the Error Mean Square, which measures variation due to either random error or to other inputs that are not in the model. The EMS should get smaller as important inputs are added to the model.

The DMINE Procedure

Effects Chosen for Target: PURCHASE

Effect	DF	R-Square	F Value	p-Value	Sum of Squares
Class: KITCHEN*STATECOD	197	0.104515	1.047456	0.3201	51.355464
Class: NTITLE*STATECOD	132	0.068297	1.023316	0.4141	33.559295
Class: STATECOD*ORIGIN	106	0.066019	1.251912	0.0470	32.439763
Class: DISHES*STATECOD	100	0.047252	0.946470	0.6295	23.218111
Class: STATECOD*EDLEVEL	83	0.043738	1.059160	0.3412	21.491715
AOV16: FREQUENT	9	0.033188	7.745761	<.0001	16.307663
Class: STATECOD*HEAT	73	0.036939	1.066755	0.3323	18.150719
Class: STATECOD*NUMCARS	55	0.023741	0.906285	0.6690	11.665595
Class: STATECOD*RACE	40	0.022761	1.202693	0.1827	11.183920
Class: LUXURY*STATECOD	37	0.018769	1.074704	0.3513	9.222419
Class: MARITAL*STATECOD	39	0.017208	0.932652	0.5899	8.455638
Group: TMKTORD*STATECOD	8	0.009270	2.475593	0.0116	4.554812
Var: RECENCY	1	0.006623	14.323348	0.0002	3.254244

Effects Chosen for Target: PURCHASE

Effect	Error Mean Square
Class: KITCHEN*STATECOD	0.248877
Class: NTITLE*STATECOD	0.248444
Class: STATECOD*ORIGIN	0.244454
Class: DISHES*STATECOD	0.245313
Class: STATECOD*EDLEVEL	0.244473
AOV16: FREQUENT	0.233930
Class: STATECOD*HEAT	0.233081
Class: STATECOD*NUMCARS	0.234034
Class: STATECOD*RACE	0.232477
Class: LUXURY*STATECOD	0.231929
Class: MARITAL*STATECOD	0.232468
Group: TMKTORD*STATECOD	0.229986
Var: RECENCY	0.227199

### The Final Anova Table for the Target

The ANOVA table is divided into the following four columns:

- **Effect** labels the source of variation as Model, Error, or Total.
- **DF** lists the degrees of freedom for each source of variation.
- **R-Square** is the model R<sup>2</sup>, which is the ratio of the model sums of squares (SS) to the total sums of squares. In this example, the selected inputs collectively explain 49.83% of the total variability in the target PURCHASE.
- **Sum of Squares** partitions the total target variation into portions that can be attributed to the model inputs and to error.

The Final ANOVA Table for Target: PURCHASE

Effect	DF	R-Square	Sum of Squares
Model	880	0.498320	244.859360
Error	1085	.	246.510427
Total	1965	.	491.369786

### SS and R2 portion for Effects Not Chosen for the Target: PURCHASE

Effects Not Chosen for Target: PURCHASE

Effect	DF	R-Square	F Value	p-Value	Sum of Squares
Group: KITCHEN*STATECOD	0	0	0	1.0000	0
Group: NTITLE*STATECOD	0	0	0	1.0000	0
Group: STATECOD*ORIGIN	0	0	0	1.0000	0
Group: DISHES*STATECOD	0	0	0	1.0000	0
Group: STATECOD*EDLEVEL	0	0	0	1.0000	0
Group: STATECOD*HEAT	0	0	0	1.0000	0
Group: LUXURY*STATECOD	0	0	0	1.0000	0
Group: STATECOD*NUMCARS	0	0	0	1.0000	0
Group: STATECOD*RACE	0	0	0	1.0000	0
Group: MARITAL*STATECOD	0	0	0	1.0000	0
Var: FREQUENT	1	0.001038	2.246918	0.1342	0.509911
Var: DOMESTIC	1	0.001172	2.537689	0.1114	0.575743
Var: APPAREL	1	0.001021	2.211639	0.1373	0.501921

### Estimating Logistic

When the target is binary, predicted values or SUPERX's are computed from the forward stepwise regression. The SUPERX's are then grouped into 256 equally spaced intervals, which are used as the independent variable in a final logistic regression analysis. The logistic regression helps you decide the cutoff of the binary response. Since there is one input, only two parameters are estimated (the intercept and the slope).

The first table shows the iteration history for estimating the intercept (alpha) and the slope (beta) for the approximate logistic regression.

The second table contains the predicted values, which are bucketed into the 256 equally sized sub-intervals. The table contains the following columns:

- **NO** – number of observations that have an observed value of PURCHASE = 'No'
- **N1** – number of observations that have an observed value of PURCHASE = 'Yes'
- **Nmiss** – number of missing values.

- **X** – center value of the corresponding interval.
- **P** – predicted value for the center value X of the sub-interval.

Iter	Alpha	Beta
0	-2.051976	4.099820
1	-3.124154	6.177697
2	-3.697468	7.272698
3	-3.823559	7.512664
4	-3.828418	7.521912
4	-3.828418	7.521912

N0	N1	Nmiss	X	P
46	1	0	-0.135	0.0078
2	0	0	-0.130	0.0081
1	0	0	-0.125	0.0084
1	0	0	-0.120	0.0087
4	0	0	-0.115	0.0091
4	0	0	-0.110	0.0094
1	0	0	-0.105	0.0098
1	0	0	-0.100	0.0101
2	0	0	-0.095	0.0105
3	0	0	-0.090	0.0109
3	0	0	-0.085	0.0113
4	0	0	-0.080	0.0118
1	0	0	-0.075	0.0122
2	0	0	-0.070	0.0127
4	0	0	-0.065	0.0132
6	0	0	-0.060	0.0137
4	0	0	-0.055	0.0142
5	0	0	-0.050	0.0147
4	0	0	-0.045	0.0153
3	0	0	-0.040	0.0158
1	0	0	-0.035	0.0164
2	0	0	-0.030	0.0171
1	0	0	-0.025	0.0177
5	0	0	-0.015	0.0191
9	0	0	-0.010	0.0198
5	0	0	-0.005	0.0205
174	0	0	0.000	0.0213
3	0	0	0.005	0.0221
7	0	0	0.010	0.0229
4	0	0	0.015	0.0238
2	0	0	0.020	0.0247
3	0	0	0.025	0.0256
3	0	0	0.030	0.0265
3	0	0	0.035	0.0275
3	0	0	0.040	0.0285
4	0	0	0.045	0.0296
7	0	0	0.050	0.0307
2	0	0	0.055	0.0318
3	0	0	0.060	0.0330
1	0	0	0.065	0.0342

2	0	0	0.070	0.0355
3	0	0	0.075	0.0368
4	1	0	0.080	0.0382
4	0	0	0.085	0.0396
8	0	0	0.090	0.0410
1	0	0	0.095	0.0425
1	0	0	0.100	0.0441
4	0	0	0.105	0.0457
4	1	0	0.110	0.0474
9	0	0	0.115	0.0491
4	0	0	0.125	0.0527
6	0	0	0.130	0.0547
1	0	0	0.135	0.0566

Additional Sub intervals not listed.

0	3	0	0.935	0.9610
1	2	0	0.940	0.9624
0	1	0	0.945	0.9637
0	3	0	0.950	0.9650
0	6	0	0.955	0.9663
0	7	0	0.960	0.9675
0	3	0	0.965	0.9686
0	4	0	0.970	0.9698
0	6	0	0.975	0.9708
0	6	0	0.980	0.9719
0	3	0	0.985	0.9729
0	5	0	0.990	0.9739
1	0	0	0.995	0.9748
0	147	0	1.000	0.9757
0	2	0	1.005	0.9766
0	1	0	1.010	0.9774
0	2	0	1.015	0.9783
0	1	0	1.020	0.9790
0	1	0	1.025	0.9798
0	1	0	1.030	0.9805
0	4	0	1.035	0.9812
0	6	0	1.040	0.9819
0	3	0	1.045	0.9826
0	5	0	1.050	0.9832
0	3	0	1.055	0.9838
0	3	0	1.060	0.9844
0	2	0	1.070	0.9855
0	1	0	1.075	0.9860
0	2	0	1.080	0.9866
0	4	0	1.085	0.9870
0	3	0	1.090	0.9875
0	2	0	1.095	0.9880
0	1	0	1.100	0.9884
0	4	0	1.105	0.9888
0	2	0	1.110	0.9892
0	1	0	1.115	0.9896
0	1	0	1.125	0.9904
0	4	0	1.130	0.9907
0	23	0	1.140	0.9914