



THE
POWER
TO KNOW.

**SAS[®] Enterprise Miner[™] and
SAS[®] Text Miner Procedures
Reference for SAS[®] 9.1.3
The ASSOC 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).

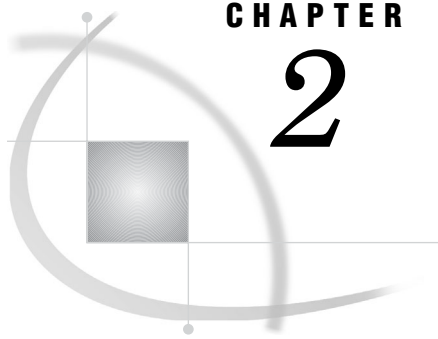
SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513.

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 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

2

The ASSOC Procedure

<i>Overview: ASSOC Procedure</i>	33
<i>Syntax: ASSOC Procedure</i>	34
<i>PROC ASSOC Statement</i>	34
<i>CUSTOMER Statement</i>	35
<i>TARGET Statement</i>	35
<i>Details: ASSOC Procedure</i>	36
<i>Output Processing</i>	36
<i>Example: ASSOC Procedure</i>	37
<i>Example 1: ASSOC Procedure</i>	37
<i>References</i>	37

Overview: ASSOC Procedure

Association discovery is the identification of items that occur together in a given event or record. This technique is also known as market basket analysis. Online transaction processing systems often provide the data sources for association discovery. Associations rules are based on frequency counts of the number of times items occur alone and in combination in the database. The rules are expressed as “if item A is part of an event then item B is also part of the event X percent of the time.” The rules should not be interpreted as a direct causation but as an association between two or more items. Identifying creditable associations can help the business technologist make decisions such as when to distribute coupons, when to put a product on sale, or how to layout items in a store.

Hypothetical association discovery rules include: If a customer buys shoes, then 10% of the time he also buys socks. A grocery chain might find that 80% of all shoppers are apt to buy a jar of salsa when they also purchase a bag of tortilla chips. When “do-it-yourselfers” buy latex paint they, also buy rollers 85% of the time. Forty percent of investors holding an equity index fund will have a growth fund in their portfolio.

An association rule has a left side (antecedent) and a right side (consequent). Both sides of the rule can contain more than one item. The confidence factor, level of support, and lift are three important evaluation criteria of association discovery. The strength of an association is defined by its confidence factor, which is the percentage of cases in which a consequent appears with a given antecedent. The level of support is how frequently the combination occurs in the market basket (database). Lift is equal to the confidence factor divided by the expected confidence. A creditable rule has a large relative confidence factor, a relatively large level of support, and a value of lift greater than 1. Rules having a high level of confidence but little support should be interpreted with caution.

The maximum number of items in an association determines the maximum size of the item set to be considered. For example, the default of four items indicates that up to 4-way associations are performed.

Syntax: ASSOC Procedure

```
PROC ASSOC <option(s)>;
  CUSTOMER variable-list;
  TARGET variable;
```

PROC ASSOC Statement

Invokes the ASSOC procedure.

```
PROC ASSOC <option(s)>;
```

Required Arguments

DATA=<libref.>SAS-data-set

Identifies the input data source. To perform association discovery, the input data set must have a separate observation for each product purchased by each customer. You must also assign the ID model role to a variable and the TARGET model role to another variable in the Input Data Source. The data set can be DMDB encoded, but if it is, you have to specify option DMDB.

DMDB

This option is required if you use a DMDB encoded data set for the DATA= option.

DMDBCAT=<libref.>SAS-catalog

Identifies the metadata catalog of the input data source.

OUT=<libref.>SAS-data-set

Specifies the output data set that contains the following variables: SET_SIZE, COUNT, ITEM1, ITEM2,...ITEM n (where n is the maximum number of variables). See “Details: ASSOC Procedure” on page 36 for more information.

SET_SIZE:

Variable that contains the total number of transactions in the data set. The first observation has the SET_SIZE equal to 0. SET_SIZE is labeled as *Relations* in the Results Browser.

COUNT:

Contains the number of transactions meeting the rule.

ITEM1, ITEM2,...ITEM n :

Contains the individual items forming the rule including the arrow.

Tip: The OUT= data set created by PROC ASSOC is input to the RULEGEN and SEQ procedures. Run PROC ASSOC and PROC RULEGEN to perform association discovery. Run PROC ASSOC and PROC SEQ to perform sequence discovery.

Options

ITEMS=*integer*

Specifies the maximum number of events or transactions to chain (or associate) together.

PCTSUP= *integer*

Specify the minimum level of support to claim that items are associated (that is, occur together in the database). The support percentage figure that you specify refers to the proportion of the largest single item frequency, and not the end support.

SUPPORT=*integer*

Specifies the minimum number of transactions that must be considered in order for a rule to be accepted. Rules that do not meet this support level are rejected. The level of support represents how frequently the combination occurs in the market basket (input data source).

Default: 5% of the largest item frequency count

For example consider the following PROC ASSOC LOG output. There are 1001 customers in the transaction data set, The largest single item frequency is 600, meaning that a particular item was purchased by 600 customers. No other item was purchased by a higher number of customers. The support in this case defaults to 5% of 600, that is, 30 customers. Only the rules with at least 30 customers therefore will be accepted.

```
----- Potential 1 item sets = 20 -----
Counting items, records read: 7007
Number of customers: 1001
Support level for item sets: 30
Maximum count for a set: 600
Sets meeting support level: 20
Megs of memory used: 0.51
```

CUSTOMER Statement

Specifies the customer(s) to be analyzed.

Alias: CUST

CUSTOMER *variable-list*;

Required Argument

variable-list

Specifies one or more names of customers to be analyzed.

TARGET Statement

Specifies the target to be analyzed.

TARGET *variable*;

Required Argument

variable

Specifies the NOMINAL variable, which contains items usually ordered by customers.

Note: By default, this variable will be normalized, that is left-justified, truncated to NORMLEN characters and upcased. For example, items such as "Blue Socks", "blue socks", "blue SOCKS " will all be mapped to a single normalized item of "BLUE SOCKS". Refer to PROC DMDB chapter for more information on normalization. \triangle

Details: ASSOC Procedure

The input to the ASSOC procedure has the following role variables: ID and TARGET. All records with the same ID values form a transaction. Every transaction has a unique ID value and one or more TARGET values.

You can have more than one ID variable. However, associations analysis can be performed on only one target variable at a time. When there are multiple ID variables, PROC ASSOC concatenates them into a single identifier value during computation.

For numeric target variables, missing values constitute a separate item or target level and show up in the rules as a period (.). For character target variables, completely blank values constitute a separate item (target level) and show up in the rules as a period (.). All records with missing ID values are considered a single valid transaction.

Output Processing

PROC ASSOC makes a pass through the data and obtains transaction counts for each item. It outputs these counts with a SET_SIZE of 1 and the items listed under ITEM1. Items that do not meet the support level are discarded. By default, the support level is set to 5% of the largest item count.

PROC ASSOC then generates all potential 2-item sets, makes a pass through the data and obtains transaction counts for each of the 2-item sets. The sets that meet the support level are output with SET_SIZE of 2 and items listed under ITEM1 and ITEM2.

The entire process is repeated for up to n -item sets. The output from PROC ASSOC is saved as SAS data sets. The data sets enable you to define your own evaluation criteria and/or reports.

Note that the ordering of n -items within an n -item set is not important. Any individual transaction, where each of the n -items occurs in any order, qualifies for a count to that particular set. The support level, once set, remains constant throughout the process.

Caution: The theoretical potential number of item sets can grow very quickly. For example, with 50 different items, you have 1225 potential 2-item sets and 19,600 3-item sets. With 5,000 items, you have over 12 million of the 2-item sets, and a correspondingly large number of 3-item sets.

Processing an extremely large number of sets could cause your system to run out of disk and/or memory resources. However, by using a higher support level, you can reduce the item sets to a more manageable number.

Example: ASSOC Procedure

PROC ASSOC must be executed before PROC RULEGEN or PROC SEQUENCE is run.

Example 1: ASSOC Procedure

Please see the RULEGEN and SEQUENCE procedures syntax for examples of PROC ASSOC code.

References

- Agrawal, R., Imielinski, T., and Swami, A. (1993), "Mining Association Rules between Sets of Items in Large Databases", *Proceedings, ACM SIGMOID Conference on Management of Data*, 207–216, Washington, D. C.
- Berry, M. J. A. and Linoff, G. (1997), *Data Mining Techniques for Marketing, Sales, and Customer Support*, New York: John Wiley and Sons, Inc.