IBM SPSS Data Preparation 26

Note

Before using this information and the product it supports, read the information in "Notices" on page 7.

Product Information

This edition applies to version 26, release 0, modification 0 of IBM® SPSS Statistics and to all subsequent releases and modifications until otherwise indicated in new editions.

Contents

Data preparation

The following data preparation features are included in SPSS® Statistics Professional Edition or the Data Preparation option.

Introduction to data preparation

As computing systems increase in power, appetites for information grow proportionately, leading to more and more data collection—more cases, more variables, and more data entry errors. These errors are the bane of the predictive model forecasts that are the ultimate goal of data warehousing, so you need to keep the data "clean." However, the amount of data warehoused has grown so far beyond the ability to verify the cases manually that it is vital to implement automated processes for validating data.

The data preparation add-on module allows you to identify unusual cases, invalid cases, variables, data values in your active dataset, and prepare data for modeling.

Usage of data preparation procedures

Your usage of data preparation procedures depends on your particular needs. A typical route, after loading your data, is:

Metadata preparation
Review the variables in your data file and determine their valid values, labels, and measurement levels. Identify combinations of variable values that are impossible but commonly miscoded. Define validation rules based on this information. This can be a time-consuming task, but it is well worth the effort if you need to validate data files with similar attributes on a regular basis.

Data validation
Run basic checks and checks against defined validation rules to identify invalid cases, variables, and data values. When invalid data are found, investigate and correct the cause. This may require another step through metadata preparation.

Model preparation
Use automated data preparation to obtain transformations of the original fields that will improve model building. Identify potential statistical outliers that can cause problems for many predictive models. Some outliers are the result of invalid variable values that have not been identified. This may require another step through metadata preparation.

Once your data file is "clean," you are ready to build models from other add-on modules.

Identify Unusual Cases

The anomaly detection procedure searches for unusual cases based on deviations from the norms of their cluster groups. The procedure is designed to quickly detect unusual cases for data-auditing purposes in the exploratory data analysis step, prior to any inferential data analysis. This algorithm is designed for generic anomaly detection; that is, the definition of an anomalous case is not specific to any particular application, such as detection of unusual payment patterns in the healthcare industry or detection of money laundering in the finance industry, in which the definition of an anomaly can be well-defined.

Example
A data analyst hired to build predictive models for stroke treatment outcomes is concerned about data quality because such models can be sensitive to unusual observations. Some of these outlying observations represent truly unique cases and are thus unsuitable for prediction, while other observations are caused by data entry errors in which the values are technically "correct"

and thus cannot be caught by data validation procedures. The Identify Unusual Cases procedure finds and reports these outliers so that the analyst can decide how to handle them.

Statistics
The procedure produces peer groups, peer group norms for continuous and categorical variables, anomaly indices based on deviations from peer group norms, and variable impact values for variables that most contribute to a case being considered unusual.

Data considerations

Data. This procedure works with both continuous and categorical variables. Each row represents a distinct observation, and each column represents a distinct variable upon which the peer groups are based. A case identification variable can be available in the data file for marking output, but it will not be used in the analysis. Missing values are allowed. The weight variable, if specified, is ignored.

The detection model can be applied to a new test data file. The elements of the test data must be the same as the elements of the training data. And, depending on the algorithm settings, the missing value handling that is used to create the model may be applied to the test data file prior to scoring.

Case order. Note that the solution may depend on the order of cases. To minimize order effects, randomly order the cases. To verify the stability of a given solution, you may want to obtain several different solutions with cases sorted in different random orders. In situations with extremely large file sizes, multiple runs can be performed with a sample of cases sorted in different random orders.

Assumptions. The algorithm assumes that all variables are nonconstant and independent and that no case has missing values for any of the input variables. Each continuous variable is assumed to have a normal (Gaussian) distribution, and each categorical variable is assumed to have a multinomial distribution. Empirical internal testing indicates that the procedure is fairly robust to violations of both the assumption of independence and the distributional assumptions, but be aware of how well these assumptions are met.

Identifying unusual cases

1. From the menus choose:

   Data > **Identify Unusual Cases...**
2. Select at least one analysis variable.
3. Optionally, choose a case identifier variable to use in labeling output.
4. Click **Apply**.

Fields with unknown measurement level

The measurement level alert displays when the measurement level for one or more variables (fields) in the dataset is unknown. Since measurement level affects the computation of results for this procedure, all variables must have a defined measurement level.

Scan Data
Reads the data in the active dataset and assigns default measurement level to any fields with a currently unknown measurement level. If the dataset is large, that may take some time.

Assign Manually
Lists all fields with an unknown measurement level. You can assign measurement level to those fields. You can also assign measurement level in the Data Editor's Variable List pane.

Since measurement level is important for this procedure, you cannot run this procedure until all fields have a defined measurement level.

Identify Unusual Cases: Output

The Output dialog provides options for generating tabular output.

List of unusual cases and reasons why they are considered unusual

When selected, this option produces three tables:

- The anomaly case index list displays cases that are identified as unusual and displays their corresponding anomaly index values.
- The anomaly case peer ID list displays unusual cases and information concerning their corresponding peer groups.
- The anomaly reason list displays the case number, the reason variable, the variable impact value, the value of the variable, and the norm of the variable for each reason.

All tables are sorted by anomaly index in descending order. Moreover, the IDs of the cases are displayed if the case identifier variable is specified on the Variables dialog.

Summaries

The controls in this group produce distribution summaries.

Peer group norms

This option displays the continuous variable norms table (if any continuous variable is used in the analysis) and the categorical variable norms table (if any categorical variable is used in the analysis). The continuous variable norms table displays the mean and standard deviation of each continuous variable for each peer group. The categorical variable norms table displays the mode (most popular category), frequency, and frequency percentage of each categorical variable for each peer group. The mean of a continuous variable and the mode of a categorical variable are used as the norm values in the analysis.

Anomaly indices

The anomaly index summary displays descriptive statistics for the anomaly index of the cases that are identified as the most unusual.

Reason occurrence by analysis variable

For each reason, the table displays the frequency and frequency percentage of each variable's occurrence as a reason. The table also reports the descriptive statistics of the impact of each variable. If the maximum number of reasons is set to 0 on the Options tab, this option is not available.

Cases processed

The case processing summary displays the counts and count percentages for all cases in the active dataset, the cases included and excluded in the analysis, and the cases in each peer group.

Identify Unusual Cases: Save

The Save dialog provides variable and model save options.

Save Variables

Controls in this group allow you to save model variables to the active dataset. You can also choose to replace existing variables whose names conflict with the variables to be saved.

Anomaly index

Saves the value of the anomaly index for each case to a variable with the specified name.

Peer groups

Saves the peer group ID, case count, and size as a percentage for each case to variables with the specified rootname. For example, if the rootname *Peer* is specified, the variables *Peerid*, *PeerSize*, and *PeerPctSize* are generated. *Peerid* is the peer group ID of the case, *PeerSize* is the group's size, and *PeerPctSize* is the group's size as a percentage.

Reasons

Saves sets of reasoning variables with the specified rootname. A set of reasoning variables consists of the name of the variable as the reason, its variable impact measure, its own value, and the norm value. The number of sets depends on the number of reasons requested on the Options tab. For example, if the rootname *Reason* is specified, the variables *ReasonVar_k*, *ReasonMeasure_k*, *ReasonValue_k*, and *ReasonNorm_k* are generated, where *k* is the *k*th reason. This option is not available if the number of reasons is set to 0.

Replace existing variables that have the same name or root name

When selected, existing variables whose names conflict with the variables to be saved are replaced.

Export Model File

Allows you to save the model to an external XML file.

Identify Unusual Cases: Missing Values

The Missing Values dialog is used to control handling of user-missing and system-missing values.

Exclude missing values from analysis

Cases with missing values are excluded from the analysis.

Include missing values in analysis

Missing values of continuous variables are substituted with their corresponding grand means, and missing categories of categorical variables are grouped and treated as a valid category. The processed variables are then used in the analysis. Optionally, you can request the creation of an additional variable that represents the proportion of missing variables in each case and use that variable in the analysis.

Identify Unusual Cases: Options

The Options dialog includes settings for unusual case criteria and defining a range for the number of peer groups.

Criteria for Identifying Unusual Cases

These following settings determine how many cases are included in the anomaly list.

Percentage of cases with highest anomaly index values

Specify a positive number that is less than or equal to 100.

Fixed number of cases with highest anomaly index values

Specify a positive integer that is less than or equal to the total number of cases in the active dataset that are used in the analysis.

Identify only cases whose anomaly index value meets or exceeds a minimum value

Specify a non-negative number. A case is considered anomalous if its anomaly index value is larger than or equal to the specified cutoff point. This option is used together with the **Percentage of cases** and **Fixed number of cases** options. For example, if you specify a fixed number of 50 cases and a cutoff value of 2, the anomaly list will consist of, at most, 50 cases, each with an anomaly index value that is larger than or equal to 2.

Number of Peer Groups

The procedure searches for the best number of peer groups between the specified minimum and maximum values. The values must be positive integers, and the minimum must not exceed the maximum. When the specified values are equal, the procedure assumes a fixed number of peer groups.

Note: Depending on the amount of variation in your data, there may be situations in which the number of peer groups that the data can support is less than the number specified as the minimum. In such a situation, the procedure may produce a smaller number of peer groups.

Maximum Number of Reasons

A reason consists of the variable impact measure, the variable name for this reason, the value of the variable, and the value of the corresponding peer group. Specify a non-negative integer; if this value equals or exceeds the number of processed variables that are used in the analysis, all variables are shown.

DETECTANOMALY command additional features

The command syntax language also allows you to:

- Omit a few variables in the active dataset from analysis without explicitly specifying all of the analysis variables (using the `EXCEPT` subcommand).
- Specify an adjustment to balance the influence of continuous and categorical variables (using the `MLWEIGHT` keyword on the `CRITERIA` subcommand).

See the *Command Syntax Reference* for complete syntax information.

Notices

This information was developed for products and services offered in the US. This material might be available from IBM in other languages. However, you may be required to own a copy of the product or product version in that language in order to access it.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:

IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
US

For license inquiries regarding double-byte (DBCS) information, contact the IBM Intellectual Property Department in your country or send inquiries, in writing, to:

Intellectual Property Licensing
Legal and Intellectual Property Law
IBM Japan Ltd.
19-21, Nihonbashi-Hakozakicho, Chuo-ku
Tokyo 103-8510, Japan

INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some jurisdictions do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM websites are provided for convenience only and do not in any manner serve as an endorsement of those websites. The materials at those websites are not part of the materials for this IBM product and use of those websites is at your own risk.

IBM may use or distribute any of the information you provide in any way it believes appropriate without incurring any obligation to you.

Licensees of this program who wish to have information about it for the purpose of enabling: (i) the exchange of information between independently created programs and other programs (including this one) and (ii) the mutual use of the information which has been exchanged, should contact:

IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
US

Such information may be available, subject to appropriate terms and conditions, including in some cases, payment of a fee.

The licensed program described in this document and all licensed material available for it are provided by IBM under terms of the IBM Customer Agreement, IBM International Program License Agreement or any equivalent agreement between us.

The performance data and client examples cited are presented for illustrative purposes only. Actual performance results may vary depending on specific configurations and operating conditions.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBMproducts. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

Statements regarding IBM's future direction or intent are subject to change or withdrawal without notice, and represent goals and objectives only.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to actual people or business enterprises is entirely coincidental.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs. The sample programs are provided "AS IS", without warranty of any kind. IBM shall not be liable for any damages arising out of your use of the sample programs.

Each copy or any portion of these sample programs or any derivative work, must include a copyright notice as follows:

© IBM 2019. Portions of this code are derived from IBM Corp. Sample Programs.

Trademarks

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corp., registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the web at "Copyright and trademark information" at www.ibm.com/legal/copytrade.shtml.

Adobe, the Adobe logo, PostScript, and the PostScript logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.

Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

UNIX is a registered trademark of The Open Group in the United States and other countries.

Java and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.

Index

Printed in USA