IBM SPSS Data Preparation 26

注释

使用本信息及其支持的产品之前，请阅读 第 5 页的『通知』 中的信息。

产品信息

此版本适用于 IBM® SPSS Statistics V26R0M0 及所有后续发行版和修改版，除非在新版本中另有说明。

目录

数据准备 1
数据准备简介 1
数据准备过程的用法 1
错误变量名称：名称超过 64 个字符，或者前一命令没有对其定义。 1
标识异常个案：输出 2
标识异常个案：保存 3
标识异常个案：缺失值 3
标识异常个案：选项 4
DETECTANOMALY 命令其他功能 4

通知 5
商标 6

索引 9

数据准备

SPSS® Statistics Professional Edition 或"数据准备"选项中包含以下数据准备功能。

数据准备简介

随着计算系统能力的提高，对信息的需要成比例增长，导致收集的数据中出现更多的个案、更多的变量以及更多的数据输入错误。这些错误会损害作为数据仓储最终目标的预测模型的预测，因此您需要使数据保持"干净"。不过，数据仓储中的数据量的增长已经大大超出了手动验证个案的能力，而这对于实现自动化的数据验证过程来说十分关键。

数据准备附加模块允许标识活动数据集中的异常个案、无效个案、变量和数据值，并准备建模数据。

数据准备过程的用法

数据准备过程的用法取决于您的特定需要。加载数据后，典型的过程是：

元数据准备

复查数据文件中的变量并确定其有效值、标签和测量级别。标识不太可能但经常存在编码错误的变量值的组合。根据这些信息定义验证规则。这是一项极为耗时的任务，不过，如果您需要定期验证具有类似属性的数据文件，则完成这项任务是十分值得的。

数据验证

运行基本检查并针对定义的验证规则进行检查，标识无效个案、变量和数据值。找到无效数据时，调查并更正原因。这可能需要另一个通过元数据准备的步骤。

模型准备

使用自动数据准备获得将改进模型构建的原始字段的转换。标识可能导致许多预测模型出现问题的潜在统计离群值。有些离群值是尚未标识的无效变量值导致的结果。这可能需要另一个通过元数据准备的步骤。

数据文件变成"干净"的之后，就可以从其他附加模块构建模型了。

错误变量名称：名称超过 64 个字符，或者前一命令没有对其定义。

异常检测过程查找基于聚类组标准值偏差的异常个案。该过程设计为在探索性数据分析步骤中，快速检测到用于数据审核的异常个案，并优先于任何推论性数据分析。此算法设计为一般"异常检测"；即异常个案的定义不被指定为任何特定应用程序，例如对保健行业中异常付款模式的检测或对金融业中洗钱行为的检测，其中对异常的定义可以被很好地界定。

示例 雇用的构建中风治疗效果预测模型的数据分析人员对数据质量非常关注，因为这类模型对异常观察值十分敏感。某些偏离的观察值表示真正唯一的个案，因此不适合用于预测，而其他观察值是由数据输入错误导致的，其值从技术上说是"正确"的，因此不能被数据验证过程捕获。"标识异常个案"过程找出并报告这些离群值，以便分析人员能够确定如何处理这些值。

统计 该过程生成对等组、连续和分类变量的对等组标准值、基于对等组标准值偏差的异常指标，以及对被视为异常的个案影响最大的变量影响值。

数据注意事项

数据。此过程既处理连续变量也处理分类变量。每行表示一个不同观察值，每列表示一个对等组以其为基础的不同变量。个案标识变量可在用于标记输出的数据文件中获得，但不能用于分析中。允许缺失值。被指定的权重变量可以忽略。

检测模型可用于新检验数据文件。检验数据元素必须与培训数据元素一致。并且，根据算法设置，用于创建模型的缺失值处理方法可适用于优先于评分的检验数据文件。

个案顺序。 注意，解决方案可取决于个案顺序。要使顺序的影响降至最低程度，可随机个案等级排序的顺序。想要验证给定解的稳定性，您可能想要通过以不同随机顺序排序的案例来得到多个不同的解。在文件非常大的情况，可使用以不同随机顺序排序的个案样本运行多次。

假设。算法假设所有变量都为不恒定且独立的，并且没有个案具有含有任何输入变量的缺失值。假设每个连续变量具有正态（高斯）分布，假设每个分类变量具有多项分布。经验内部检验表明，该过程对于违反独立性假设和分布假设均相当稳健，但应了解这些假设符合的程度。

标识异常个案

1. 从菜单中选择：

   数据 > **标识异常个案...**
2. 选择至少一个分析变量。
3. 还可以选择一个个案标识变量用于标记输出。
4. 单击**应用**。

具有未知测量级别的字段

当数据集中的一个或多个变量（字段）未知时，将显示测量级别警报。由于测量级别会影响该过程的计算结果，因此所有变量必须都定义有测量级别。

扫描数据

读取活动数据集中的数据，并分配缺省测量级别给任何具有当前未知测量级别的字段。如果数据集较大，该过程可能需要一些时间。

手动分配

列出所有具有未知测量级别的字段。您可以将测量级别分配给这些字段。您也可以在数据编辑器的"变量列表"窗格中分配测量级别。

由于测量级别对该过程很重要，因此您无法运行该过程，除非所有字段均定义了测量级别。

标识异常个案：输出

"输出"对话框提供用于生成表格输出的选项。

异常个案及其被视为异常的原因的列表

选中时，此选项可生成三个表：

- 异常个案指标列表显示标识为异常的个案，并显示其相应的异常指标值。
- 异常个案 Peer ID 列表显示异常个案及其相应对等组的相关信息。
- 异常原因列表显示个案号、原因变量、变量影响值、变量值以及每个原因的变量的标准值。

所有表都根据异常指标按降序排列。此外，如果在"变量"对话框上指定了个案标识变量，则会显示个案的 ID。

摘要 此组中的控件可生成分布摘要。

对等组标准值

此选项显示连续变量标准值表（如果分析中使用了任何连续变量）以及分类变量标准值表（如果分析中使用了任何分类变量）。连续变量标准值表显示每个对等组的每个连续变量的平均值和标准差。分类变量标准值表显示每个对等组的每个分类变量的众数（最大类别）、频率和频率百分比。连续变量的平均值和分类变量的众数在分析中用作标准值。

异常指标

异常指标摘要显示标识为最不正常个案的异常指标的描述统计。

按分析变量列出出现的原因

对于每个原因，该表将每个变量的出现频率和频率百分比显示为原因。该表还报告每个变量的影响的描述统计。如果在"选项"选项卡上将最大的原因数量设置为 0，则此选项不可用。

已处理的个案数

个案处理摘要显示活动数据集中所有个案的计数和计数百分比、分析中包含和排除的个案，以及每个对等组中的个案。

标识异常个案：保存

"保存"对话框提供变量和模型保存选项。

保存变量

此组中的控件允许您将模型变量保存到活动数据集。您也可以选择将存在名称冲突的现有变量替换为要保存的变量。

异常指标

将每个个案的异常指标值保存到具有指定名称的变量中。

对等组

将对等组 ID、个案计数以及每个个案的以百分比表示的大小保存到具有指定根名称的变量中。例如，如果指定了根名称 *Peer*，则会生成变量 *Peerid*、*PeerSize* 和 *PeerPctSize*。*Peerid* 为个案的对等组 ID，*PeerSize* 为组的大小，而 *PeerPctSize* 为用百分比表示的组大小。

原因 使用指定的根名称保存原因变量集。原因变量集包含作为原因的变量的名称、变量影响测量、变量自身的值以及标准值。变量集的数量取决于在"选项"选项卡上请求的原因的数目。例如，如果指定根名称 *Reason*，则会生成变量 *ReasonVar_k*、*ReasonMeasure_k*、*ReasonValue_k* 和 *ReasonNorm_k*，其中 *k* 是第 *k* 个原因。如果原因数量设置为 0，则此选项不可用。

替换具有相同名称或根名称的现有变量

在选中时，将替换名称与要保存的变量相冲突的现有变量。

导出模型文件

允许将模型保存到外部 XML 文件。

标识异常个案：缺失值

"缺失值"对话框用于控制对用户缺失值和系统缺失值的处理。

从分析中排除缺失值

具有缺失值的个案会从分析中排除。

在分析中包括缺失值

连续变量的缺失值将替换为它们对应的总平均值，分类变量的缺失类别将分组并视为有效类别。处理过的变量随后在分析中使用。或者，您也可以请求创建表示每个个案中缺失变量的比例的附加变量并在分析中使用该变量。

标识异常个案：选项

"选项"对话框包含异常个案条件的设置并定义对等组数量范围。

异常个案的标识条件

以下设置确定在异常列表中包括多少个个案。

具有最高异常指标值的个案所占的百分比

指定一个小于或等于 100 的正数。

具有最高异常指标值的个案的固定数量

指定一个正整数，该整数小于或等于分析中使用的活动数据集的个案总数。

仅标识异常指标值符合或超过最小值的个案

指定一个非负数。如果某个个案的异常指标值大于或等于指定分界点，则将该个案视为异常个案。此选项与**个案百分比**和**个案的固定数量**选项一起使用。例如，如果指定 50 作为固定数量，并指定 2 作为分界值，则异常列表最多可包含 50 个个案，每个个案的异常指标值都大于等于 2。

对等组的数量

该过程搜索指定的最小值和最大值之间的最佳对等组数量。该值必须为正整数，并且最小值不能超过最大值。如果指定的值相等，则该过程假定对等组的数量是固定的。

注：根据数据中的变动量，有时数据可支持的对等组的数量可能小于指定的最小数量。在这种情况下，该过程可能会生成数量较少的对等组。

最大的原因数量

原因包括变量影响测量、此原因的变量名、变量的值以及相应对等组的值。指定一个非负整数，如果此值等于或超过分析中使用的已处理变量的数量，则会显示所有变量。

DETECTANOMALY 命令其他功能

使用命令语法语言还可以：

- 在分析中省略活动数据集中的一些变量，而不显式指定所有分析变量（使用 EXCEPT 子命令）。
- 指定通过调整平衡连续和分类变量的影响（使用 CRITERIA 子命令的 MLWEIGHT 关键字）。

请参阅命令语法参考以获取完整的语法信息。

通知

本信息是为在美国提供的产品和服务编写的。本资料的其他语言版本可以从 IBM 获取。但是，您可能需要拥有该语言的产品副本或产品版本才能访问这些资料。

IBM 可能在其他国家或地区不提供本文档中讨论的产品、服务或功能特性。有关您当前所在区域的产品和服务的信息，请向您当地的 IBM 代表咨询。任何对 IBM 产品、程序或服务的引用并非意在明示或暗示只能使用 IBM 的产品、程序或服务。只要不侵犯 IBM 的知识产权，任何同等功能的产品、程序或服务，都可以代替 IBM 产品、程序或服务。不过，用户应自行负责评估和验证任何非 IBM 产品、程序或服务的工作情况。

IBM 公司可能已拥有或正在申请与本文档内容有关的各项专利。您获得该文档并不意味着授予您任何这些专利许可。您可以将书面许可查询函件发送至：

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

有关双字节 (DBCS) 信息的许可查询，请与您所在国家或地区的 IBM 知识产权部门联系，或用书面方式将查询寄往：

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"按现状"提供本出版物，不附有任何种类的（无论是明示的还是暗含的）保证，包括但不限于暗含的有关非侵权、适销和适用于某种特定用途的保证。某些管辖区域在某些交易中不允许免除明示或暗含的保证。 因此本条款可能不适用于您。

本信息可能含有技术误差或排版错误。此处的信息会定期进行更改；这些更改会体现在本出版物的新版本中。IBM 可以随时对本资料中描述的产品和/或程序进行改进和/或更改，而不另行通知。

本信息中对任何非 IBM Web 站点的引用都只是为了方便起见才提供的， 不以任何方式充当对那些 Web 站点的保证。那些 Web 站点中的资料不是 IBM 产品资料的一部分，使用那些 Web 站点带来的风险将由您自行承担。

IBM 可以按它认为适当的任何方式使用或分发您所提供的任何信息而无须对您承担任何责任。

本程序的被许可方如果要了解有关程序的信息以达到如下目的：(i) 允许在独立创建的程序和其他程序（包括本程序）之间进行信息交换，以及 (ii) 允许对已经交换的信息进行相互使用，请与下列地址联系：

IBM Director of Licensing
IBM Corporation

North Castle Drive, MD-NC119
Armonk, NY 10504-1785
US

此类信息的提供应遵照相关条款和条件，其中包括在某些情况下支付适当费用。

本资料中描述的许可程序及其所有可用的许可资料均由 IBM 依据 IBM 客户协议、IBM 国际软件许可协议或任何同等协议中的条款提供。

所引用的性能数据和客户示例只用于阐述说明。根据具体配置和操作条件，实际性能结果可能有所不同。

涉及非 IBM 产品的信息可从这些产品的供应商、其出版说明或其他可公开获得的资料中获取。IBM 没有对这些产品进行测试，也无法确认其性能的精确性、兼容性或任何其他关于非 IBM 产品的声明。有关非 IBM 产品性能的问题应当向这些产品的供应商提出。

有关 IBM 未来方向或意向的声明均可能未经通知即变更或撤销，并且仅代表目标和目的。

本信息包含日常业务运营中使用的数据和报告的示例。为了尽可能详尽地对其进行说明，示例中包含了人员的姓名、公司、品牌和产品的名称。所有这些名称都是虚构的，如果与实际人员或公司企业有任何类似则纯属巧合。

版权许可：

本信息包括源语言形式的样本应用程序，这些样本说明不同操作平台上的编程方法。如果是为按照在编写样本程序的操作平台上的应用程序编程接口 (API) 进行应用程序的开发、使用、经销或分发为目的，您可以任何形式对这些样本程序进行复制、修改、分发，而无须向 IBM 付费。这些示例并未在所有条件下作全面测试。因此，IBM 不能担保或暗示这些程序的可靠性、可维护性或功能。本样本程序仍然是"按现状"提供的，不附有任何种类的保证。对于因使用样本程序所引起的任何损害，IBM 概不负责。

凡这些实例程序的每份拷贝或其任何部分或任何衍生产品，都必须包括如下版权声明：

商标

IBM、IBM 徽标和 ibm.com 是 International Business Machines Corp., 在全球许多管辖区域注册的商标或注册商标。其他产品和服务名称可能是 IBM 或其他公司的商标。Web 站点 www.ibm.com/legal/copytrade.shtml 上的"Copyright and trademark information"部分中提供了 IBM 商标的最新列表。

Adobe、Adobe 徽标、PostScript 和 PostScript 徽标是 Adobe Systems Incorporated 在美国和/或其他国家或地区的注册商标或商标。

Intel、Intel 徽标、Intel Inside、Intel Inside 徽标、Intel Centrino、Intel Centrino 徽标、Celeron、Intel Xeon、Intel SpeedStep、Itanium 和 Pentium 是 Intel Corporation 或其子公司在美国和其他国家或地区的商标或注册商标。

Linux 是 Linus Torvalds 在美国、其他国家或地区或两者的注册商标。

Microsoft、Windows、Windows NT 和 Windows 徽标是 Microsoft Corporation 在美国、其他国家或地区或两者的商标。

UNIX 是 The Open Group 在美国和其他国家或地区的注册商标。

Java 和所有基于 Java 的商标和徽标是 Oracle 及/或其分支机构的商标和注册商标。

索引

[C]

错误变量名称：名称超过 64 个字符，或者
前一命令没有对其定义。 1
保存变量 3
导出模型文件 3
缺失值 3
输出 2
选项 4

[D]

对等组
在"标识异常个案"中 2, 3

[Q]

缺失值
在"标识异常个案"中 3

[Y]

异常指标
在"标识异常个案"中 2, 3
原因
在"标识异常个案"中 2, 3

Printed in China