

提高数据准备之工作，从而获得更准确的分析结果

在做数据分析之前，所有分析者必须先要准备他们的数据。SPSS Statistics的基础模块(Base)含有数据准备的许多工具，但是有时候数据准备需要更多专用的技巧。SPSS Statistics的附加的Data Preparation模块能简单便捷地识别可疑或无效的观测值或变量，以及数据值；它能够了解数据缺失的模式，总结变量的分布；更精确地和名义变量的算法一起工作。这些流程化的数据准备过程，使您更快的进入数据分析阶段，有助于达到更精确的结论。

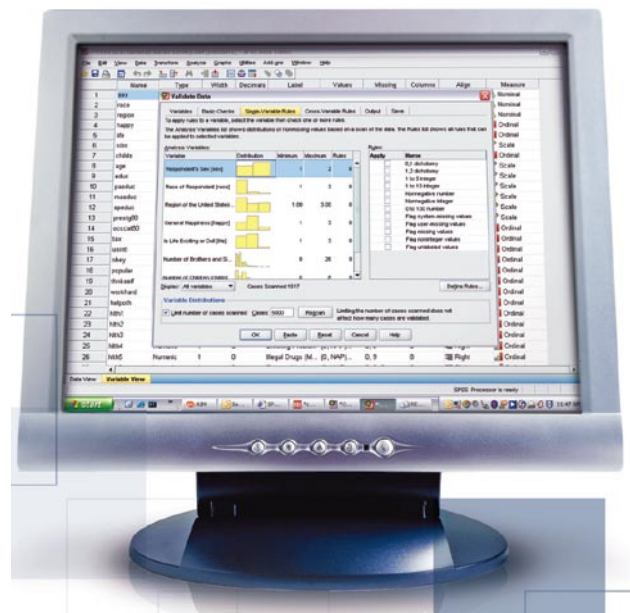
SPSS数据准备模块(Data Preparation)可以仅作为客户端软件安装，为提高性能和扩展性，也可以和SPSS Statistics服务器一起作为C/S结构安装。

执行数据检验

以前通过人工校验的方式，例如，你也许先运行一个频数分析，根据输出的频数表，圈出需要改正的观测，再查找具体观测的记录。无需说，这是非常费时的工作，并且由于同一个组织的不同的分析人员可能应用稍微不同的方法，要维持不同项目之间处理方法的一致性是十分困难的。应用SPSS Data Preparation中的数据校验方法，免除了人工校验的繁琐，简化了数据校验过程，并且让它效率化、流程化。

SPSS Data Preparation过程，可依据变量的测度水平，套用相应的规则对数据进行检验（无论是分类还是连续）。例如，调查数据的变量为五个尺度的李克特(Likert)量，通过数据校验程序您可以对该变量应用5个取值的变量规则，然后标记出那些取值在规定的5个值范围以外的观测。这样，您可以知道哪些记录为无效案例以及它们违反了哪些规则、有多少观测值为无效的报告。您可以为单个变量指定校验规则（例如取值范围校验）及多变量间的交叉校验规则（例如，“怀孕的男性”）。

在数据分析前，你根据数据校验结果，凭你的知识自主决定数据的有效性，是否删除或者更正可疑的记录。



迅速发现多个异常值

异常值探测程序可以提前找出偏态分析中的离群值。异常探察程序能够基于数据与数据集中相似观测的偏离情况而探察出异常值，并给出偏离的原因。您可以通过创建新变量来标识异常值。找到异常记录后，您可以进一步检查它们，决定是否在分析中包含该部分记录。

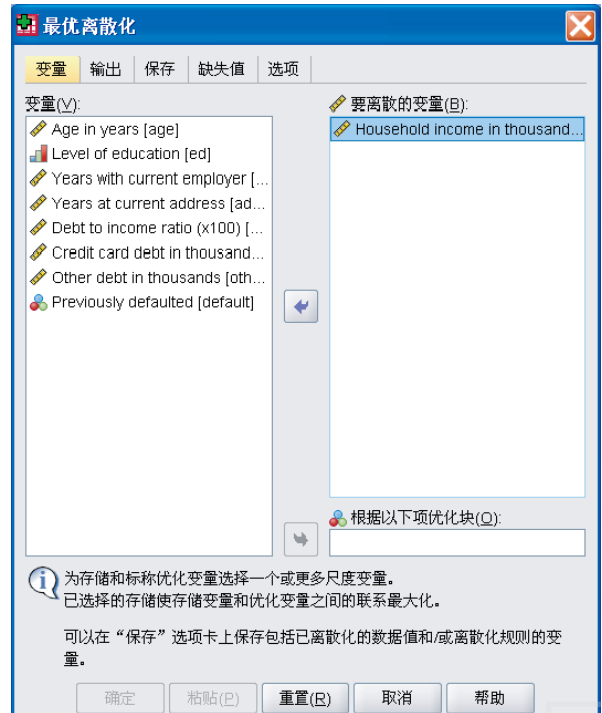
建模前预处理数据

为了应用专用于名义变量的算法（例如，Naïve Bayes, Logit模型），建模前你需要先对尺度变量分段。否则，对大数据集，象多元逻辑回归分析等算法会处理极长的时间，或者算法根本不能收敛。另外，得到的分析结果可能难以阅读或解释。

最优离散化（或最优分段），可以让您决定分段点，让专用于名义变量的算法在尺度变量上达到尽可能好的结果。

在建模前，你可以从三种不同的分段算法中选择分段方法来预处理数据：

- 无监督的—每段中的频数相等。
- 有监督的—由目标变量来决定分界点。该方法比无监督方法更精确，但是计算量更大。
- 混合方法—结合以上两种方法。当变量有大量互不相同的取值时，该方法尤其有用。



最优离散化使您更精确的使用为名义变量设计的算法

功能

验证数据有效性

利用数据校验程序验证工作数据集的数据有效性

- 基本校验：设定应用于变量和观测的基本校验。例如，获得标识具有高缺失比例或空观测变量的报告。
 - 最大缺失值比例
 - 单一类别观测最高比例
 - 观测中技术1出现的最高比例
 - 最小变异系数
 - 最小标准偏差
 - 标记不完整IDs
 - 标记重复IDs
 - 标记空观测
- 标准规则：数据描述，查看单一变量规则，并把规则应用于分析变量
 - 数据描述
 - 分布：通过微型条形图显示分类变量的分布、通过直方图显示连续变量的分布
 - 显示数据的最大最小值
 - 单变量规则
 - 把规则应用于单一变量来识别缺失值或无效值，如超出有效范围的值
 - 用户也可自定义单一变量规则
 - 定制规则：定义交叉规则表达式表示响应违反逻辑规则（如，“怀孕的男性”）
 - 输出：报告描述无效数据
- 个案情况报告，依据个案列出确认违规数
 - 要包括的个案最小违规数
 - 报告中个案的最大违规数量
- 标准校验规则报告
 - 依据分析变量汇总
 - 依据规则汇总
 - 输出描述统计量
- 保存：使您能够保存表示规则违反的变量，并利用它们清洗数据、过滤不好的个案
- 摘要变量
 - 空个案指示变量
 - 重复ID指示变量
 - 不完整ID指示变量
 - 被违反的规则（总数）
- 用来记录所有确认为违规的指示变量

标识异常个案

异常探索过程，基于个案与其对等组的偏离程度识别异常，并给出异常的原因

- 通过变量子命令指定该过程中要用到的变量。指定分类的，连续的，和ID变量，并列出不在于分析中的变量
- 通过缺失值处理子命令指定该过程处理缺失值的方法
 - 应用缺失值处理。如果选择了该选项，对于连续型变量的缺失值，缺失值将被替换为变量的均值；对于分类变量，缺失值将被合并作为一有效的新类别包含在分析中。经过处理后的变量将被包括到分析中。
- 创建新的变量表示每个个案中缺失值的比例，把将其用于分析中。如果选择了该选项，该过程将会创建一称为“缺失值比例”的变量，用以表示每个个案中的缺失值比例，并在包含在分析过程中
- 通过CRITERIA子命令可以进行如下设置：
 - 对等组的最小和最大数量
 - 调整测量水平的权重
 - 异常列表中原原因的数目
 - 指定识别出为异常值的个案百分比
 - 指定识别出为异常值的个案数目
 - 设定某个案是否被识别为异常的异常指标阈值
- 通过保存子命令保存新变量到工作数据集
 - 异常指标
 - 对等组ID
 - 对等组大小
 - 对等组中个案数所占百分比
 - 原因变量名称
 - 原因变量的影响度量
 - 原因变量的值
 - 原因变量的范数
- 利用OUTFILE子命令将模型以XML格式导出
- 利用PRINT子命令控制输出，你可以输出
 - 个案处理摘要
 - 异常指标，异常值的对等组ID，异常原因
 - 如果分析中涉及到任何连续变量，则输出连续变量范数；如果是分类变量，则输出分类变量的范数
 - 异常指标摘要
 - 每个异常原因的摘要表
- 抑制所有输出，除了附注和任何警告之外

最优离散化：

用最优离散化（或最优分段）来预处理数据，把连续型变量的值分配到不同的段中，从而变为分类变量。该过程在减少要离散化变量的取值个数中十分有用。当应用某些最优离散化算法时，指导变量可以帮助你决定区间端点，可以最大化分段变量和指导变量的联系。

■ 可供选择的方法如下：

– 无监督分段：该分段应用等频数算法来离散化分段变量。指导变量不是必需的。

– 有监督分段：该分段算法应用MDLP（最短表达长度原理）算法，它应用MDLP算法来离散化待分段变量二不需任何预处理，适用于记录数不多的数据集。指导变量是必需的。

– 混合MDLP分段算法：该方法先用等频数算法预处理数据，然后应用MDLP算法。对于具有大量记录的数据集合，该方法特别适用。

■ 指定以下的准则：

- 对每一个待分段变量如何定义最小分段点
- 如何定义区间的左端点
- 是否强制合并稀疏分布的取值

■ 保留以下变量

- 含离散化的数据值的变量
- 离散化规则的语法命令文件

■ 用PRINT命令来控制输出结果的显示方式

- 分段端点
- 已离散化变量的描述统计量
- 已离散化变量的模型熵

系统要求

■ 软件: SPSS Base 17.0

■ 其它的系统需求根据平台的不同而异



广州博脉信息技术有限公司 网址：<http://www.pomine.com>
地址：广州市天河区龙口东路科技大厦 1207 室 邮编：510630
电话：020-22644217, 22644215 传真：020-22644215