IBM SPSS Data Preparation 20

注意：使用本信息及其支持的产品之前，请阅读注意事项第 136 页码下的一般信息。

此版本适用于 IBM® SPSS® Statistics 20 及所有后续发布和修订，除非在新版本中另有说明。

Adobe 产品屏幕截图重印已获得 Adobe Systems Incorporated 的许可。

Microsoft 产品屏幕截图重印已获得 Microsoft Corporation 的许可。

前言

IBM® SPSS® Statistics 是一种用于分析数据的综合系统。Data Preparation 可选附加模块提供本手册中描述的其他分析方法。此 Data Preparation 附加模块必须与 SPSS Statistics Core 系统一起使用，并已完全集成到了该系统中。

关于 IBM Business Analytics

IBM Business Analytics 软件提供决策者赖以提高业务绩效的完整、一致和准确的信息。包括业务智能、预测分析、财务状况和战略管理以及分析应用程序在内的一整套产品组合让您即刻、清楚地了解当前绩效并依此采取行动，以及能够预测未来的成果。结合丰富的行业解决方案、被证明的实践经验和专业的服务，无论公司规模大小，都能促使其获得最高的产能、自信自觉地做出决定并得到更好的成绩。

作为产品组合的一部分，IBM SPSS Predictive Analytics 软件帮助公司预测未来实践并采取积极行动，促使其获得更好的业务成果。全世界的商业政府和学术客户依赖 IBM SPSS 技术，因其具有竞争力的优势，能够吸引、留住和发展客户，同时减少欺诈和减轻风险。通过将 IBM SPSS 软件融入日常运营中，公司成为具有预测性的企业，能够引导和自觉做出决策，以满足业务目标，实现可观的竞争优势。欲知更多信息或联系代表，请访问 http://www.ibm.com/spss。

技术支持

我们提供有“技术支持”以维护客户。客户可就 IBM Corp. 产品使用或某一受支持硬件环境的安装帮助寻求技术支持。如要联系技术支持，请访问 IBM Corp. 网站，网址为 http://www.ibm.com/support。在请求协助时，请准备好您和您组织的 ID 以及支持协议。

针对学生的技术支持

如果您是使用任何学生版 IBM SPSS 软件产品的学生，请访问我们专为学生提供的在线教育解决方案 (http://www.ibm.com/spss/rd/students/)页面。如果您是使用大学提供的 IBM SPSS 软件副本的学生，请联系所在大学的 IBM SPSS 产品协调员。

客户服务

如果对发货或帐户存在任何问题，请联系您当地的办事处。请先准备好您的序列号以供识别。

培训讲座

IBM Corp. 提供公开的以及现场的培训讲座。所有讲座都是以实践小组为特色的。讲座将定期在各大城市开展。有关这些讲座的更多信息，请前往 http://www.ibm.com/software/analytics/spss/training。

附加出版物

SPSS Statistics：数据分析指南、SPSS Statistics：Statistical Procedures Companion 和 SPSS Statistics：Advanced Statistical Procedures Companion（由 Marija Norušis 编写，并已由 Prentice Hall 出版）作为建议的补充材料提供。这些出版物涵盖 SPSS Statistics Base 模块、Advanced Statistics 模块和 回归模块中的统计过程。无论您是刚开始从事数据分析工作，还是已准备好使用高级应用程序，这些书籍都将帮助您最有效地利用在 IBM® SPSS® Statistics 产品中找到的功能。有关其他信息，包括出版物的内容和示例章节，请参阅作者的网站： http://www.norusis.com

内容

部分 I: 用户指南

附录

部分 I:
用户指南

章 1

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，导致收集的数据越来越多—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”附加模块允许您标识活动数据集中的异常个案和无效个案、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。

章 2

验证规则

规则用于确定个案是否有效。有两种类型的验证规则：

- **单变量规则。**单变量规则包含一组应用于单个变量的固定检查，例如范围外值的检查。对于单变量规则，有效值可以表示为一个值范围，也可以表示为一个可接受值列表。
- **交叉变量规则。**交叉变量规则是用户定义的规则，可以应用于单个变量，也可以应用于变量组合。交叉变量规则由标记无效值的逻辑表达式定义。

验证规则保存在数据文件的数据字典中。这样指定一次规则后就可以重用规则。

载入预定义的确认规则

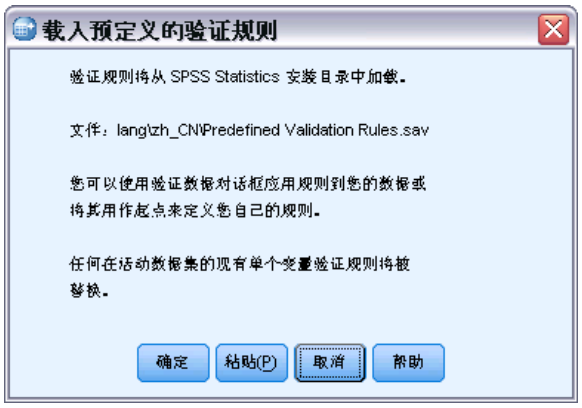
通过从安装中所包含的外部数据文件载入预定义规则可以快速获取一组可供使用的验证规则。

载入预定义的确认规则

▶ 从菜单中选择：

数据 > 验证 > 载入预定义规则...

图片 2-1
载入预定义的确认规则



注意，此过程会删除活动数据集中现有的所有单变量规则。

或者，您也可以使用复制数据属性向导从任何数据文件载入规则。

定义验证规则

“定义验证规则”对话框允许您创建和查看单变量和交叉变量验证规则。

创建和查看验证规则

▶ 从菜单中选择：
数据 > 验证 > 定义规则...

该对话框中包含从数据字典读取的单变量和交叉变量验证规则。如果不存在任何规则，则会自动创建一个新的占位符规则，您可以对其进行修改以满足您的要求。

▶ 在“单变量规则”和“交叉变量规则”选项卡上选择各个规则可查看和修改其属性。

定义单变量规则

图片 2-2
“定义验证规则”对话框，“单变量规则”选项卡

“单变量规则”选项卡允许您创建、查看和修改单变量验证规则。

规则。该列表按名称和规则适用的变量类型显示单变量验证规则。该对话框打开时，它显示在数据字典中定义的规则，或者，如果当前未定义任何规则，则显示名为“Single-Variable Rule 1”的占位符规则。下列按钮将显示在“规则”列表下方：

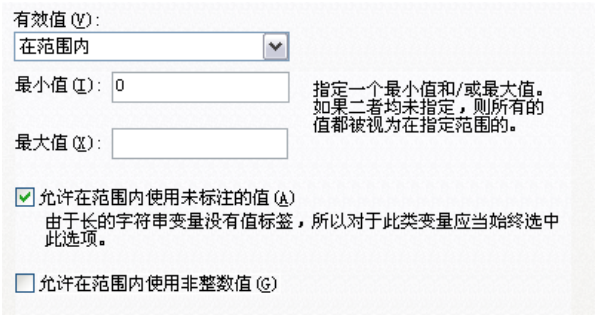
- **新建。**在“规则”列表底部添加一个新的条目。该规则会被选中，并分配名称“SingleVarRule n”，其中 n 是一个整数，这使得新规则的名称在单变量和交叉变量规则中是唯一的。
- **复制。**在“规则”列表底部添加一个所选规则的副本。规则的名称会进行调整，使其在单变量和交叉变量规则中是唯一的。例如，如果复制“SingleVarRule 1”，则第一个复制规则的名称将是“Copy of SingleVarRule 1”，第二个将是“Copy (2) of SingleVarRule 1”，依此类推。
- **删除。**删除所选规则。

规则定义。通过这些控件，可以查看和设置所选规则的属性。

- **名称。**规则的名称在单变量和交叉变量规则中必须是唯一的。
- **类型。**这是规则适用的变量类型。请从**数值**、**字符串**和**日期**中进行选择。
- **格式。**这允许您为可应用于日期变量的规则选择日期格式。
- **有效值。**您可以以值范围或值列表的形式指定有效值。

范围定义控件允许您指定有效范围。该范围以外的值会被标记为无效。

图片 2-3
单变量规则：范围定义



要指定范围，请输入最小值和/或最大值。复选框控件允许您标记范围内的未标注值和非整数值。

列表定义控件允许您定义有效值的列表。未包含在列表中的值会被标记为无效。

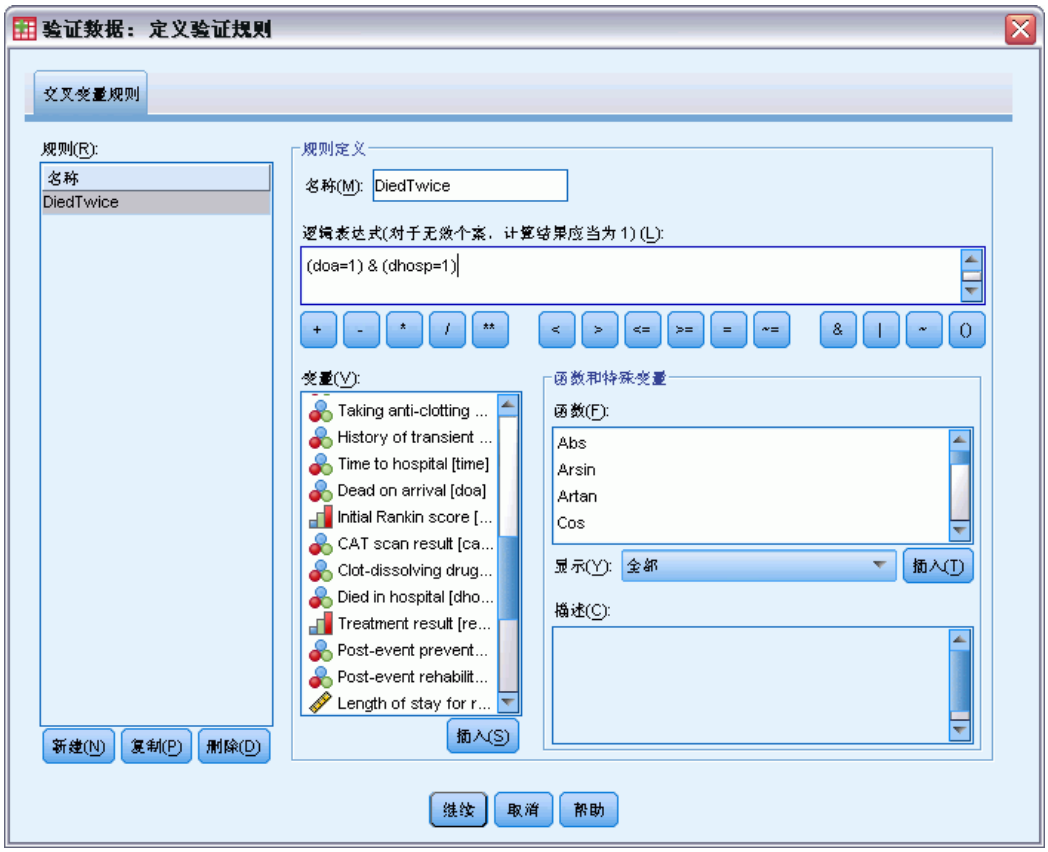
图片 2-4
单变量规则：列表定义

在网格中输入列表值。该复选框确定针对可接受值列表检查字符串数据值时是否区分大小写。

- **允许使用用户缺失值。**控制是否将用户缺失值标记为无效。
- **允许使用系统缺失值。**控制是否将系统缺失值标记为无效。这不适用于字符串规则类型。
- **允许使用空值。**控制是否将空白（也就是完全为空）字符串值标记为无效。这不适用于非字符串规则类型。

定义交叉变量规则

图片 2-5
“定义验证规则”对话框，“交叉变量规则”选项卡



“交叉变量规则”选项卡允许您创建、查看和修改交叉变量验证规则。

规则。该列表按名称显示交叉变量验证规则。该对话框打开时，它显示名为“CrossVarRule 1”的占位符规则。下列按钮将显示在“规则”列表下方：

- **新建。**在“规则”列表底部添加一个新的条目。该规则会被选中，并分配名称“CrossVarRule n”，其中 n 是一个整数，这使得新规则的名称在单变量和交叉变量规则中是唯一的。

- **复制。**在“规则”列表底部添加一个所选规则的副本。规则的名称会进行调整，使其在单变量和交叉变量规则中是唯一的。例如，如果复制“CrossVarRule 1”，则第一个复制规则的名称将是“Copy of CrossVarRule 1”，第二个将是“Copy (2) of CrossVarRule 1”，依此类推。
- **删除。**删除所选规则。

规则定义。通过这些控件，可以查看和设置所选规则的属性。

- **名称。**规则的名称在单变量和交叉变量规则中必须是唯一的。
- **逻辑表达式。**这实际上就是规则定义。您应该编写表达式以使无效个案的计算结果为 1。

构建表达式

▶ 要构建一个表达式，可以将成分粘贴到“表达式”字段中或是在“表达式”字段中直接输入。

- 通过从“函数组”列表中选择组，然后双击“函数和特殊变量”列表中的函数或变量（或者选择函数或变量，然后单击**插入**），可以粘贴函数或常用的系统变量。填充问号指示的任何参数（仅适用于函数）。标记为**全部**的函数组提供所有可用函数和系统变量的列表。对话框的保留区域中显示对当前所选函数或变量的简要描述。
- 字符串常数必须包含在引号或撇号中。
- 如果值包含小数，则必须使用句号 (.) 作为小数指示符。

章 3

验证数据

“验证数据”对话框允许您标识活动数据集中可疑的和无效的个案、变量和数据值。

示例。 数据分析人员每个月必须向客户提供客户满意度报告。她每个月接收到的数据需要进行质量检查，看是否存在不完整的客户标识、超出范围的变量值以及经常错误输入的变量值组合。“验证数据”对话框允许分析人员指定唯一标识客户的变量，为有效变量范围定义单变量规则，并定义交叉变量规则以找出不可能的组合。该过程返回问题个案和变量的报告。此外，每个月的这些数据都具有相同的数据元素，因此分析人员可以将规则应用于下个月的新数据文件。

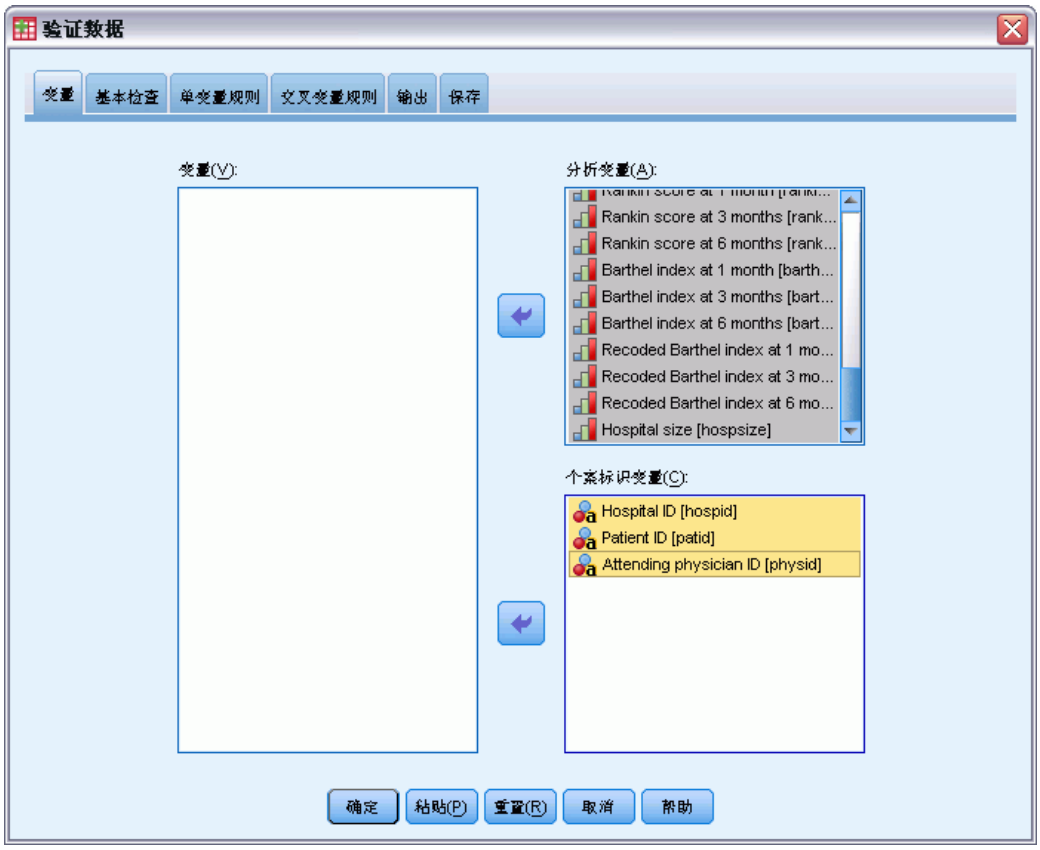
统计量。 该过程生成多项检查失败的变量、个案和数据值的列表，违反单变量和交叉变量规则的次数计数，以及分析变量的简单描述摘要。

权重。 该过程忽略权重变量规范，而是像对待任何其他分析变量一样对待权重变量。

验证数据

- 从菜单中选择：
  数据 > 验证 > 验证数据...

图片 3-1
“验证数据”对话框，“变量”选项卡



- 选择一个或多个分析变量，以便由基本变量检查或单变量验证规则进行验证。

  或者，您可以：

- 单击**交叉变量规则**选项卡并应用一个或多个交叉变量规则。

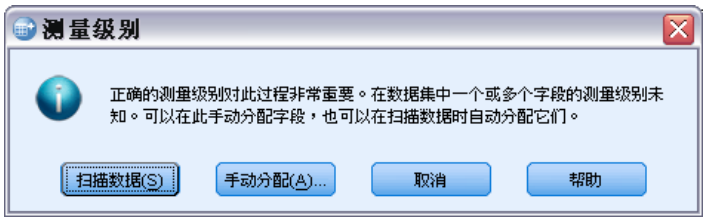
  根据需要，您可以：

  - 选择一个或多个个案标识变量以便检查重复的或不完整的 ID。个案标识变量还可用于标记个案输出。如果指定了两个或更多个案标识变量，则可将其值的组合视为个案标识。

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）的测量级别未知时，将显示测量级别警告。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。

图片 3-2
测量级别警报



- **扫描数据。** 读取活动数据集中的数据，并分配默认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。** 打开列出了所有具有未知测量级别的字段的对话框。您可以使用该对话框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视图中分配测量级别。

由于测量级别对该过程很重要，因此您无法访问运行该过程的对话框，除非所有字段均定义了测量级别。

验证数据：基本检查

图片 3-3
“验证数据”对话框，“基本检查”选项卡

“基本检查”选项卡允许您为分析变量、个案标识和全部个案选择基本检查。

分析变量。 如果在“变量”选项卡上选择了任何分析变量，则可选择以下任意有效性检查。复选框允许您打开或关闭检查。

- **缺失值的最大百分比。** 报告缺失值百分比大于指定值的分析变量。指定的值必须是一个小于等于 100 的正数。
- **单个类别中个案所占的最大百分比。** 如果任何分析变量是分类变量，则此选项报告表示单个非缺失类别的个案的百分比大于指定值的分类分析变量。指定的值必须是一个小于等于 100 的正数。百分比基于具有非缺失变量值的个案。
- **计数为 1 的类别的最大百分比。** 如果任何分析变量是分类变量，则此选项报告仅包含一个个案的变量类别的百分比大于指定值的分类分析变量。指定的值必须是一个小于等于 100 的正数。
- **最小变异系数。** 如果任何分析变量是刻度变量，则此选项报告变异系数的绝对值小于指定值的刻度分析变量。此选项仅适用于均值非零的变量。指定的值必须是一个非负数。指定 0 会关闭变异系数检查。
- **最小标准差。** 如果任何分析变量是刻度变量，则此选项报告标准差小于指定值的刻度分析变量。指定的值必须是一个非负数。指定 0 会关闭标准差检查。

个案标识。 如果在“变量”选项卡上选择了任何个案标识变量，则可选择以下任意有效性检查。

- **标记不完整的 ID。** 此选项报告具有不完整个案标识的个案。对于特定个案，如果任何标识变量的值为空或者缺失，则该标识被视为不完整。
- **标记重复的 ID。** 此选项报告具有重复个案标识的个案。不完整的标识会从可能重复的组中排除。

标记空个案。 此选项报告所有变量均为空或空白的个案。为了标识空个案，您可以选择使用文件中的所有变量（不包括任何标识变量）或者仅使用在“变量”选项卡上定义的分析变量。

验证数据：单变量规则

图片 3-4
“验证数据”对话框，“单变量规则”选项卡

“单变量规则”选项卡显示可用的单变量验证规则，并允许您应用这些规则分析变量。要定义其他单变量规则，请单击**定义规则**。 有关详细信息，请参阅第 3 页码第 2 章中的定义单变量规则。

分析变量。 该列表显示分析变量，汇总其分布，并显示应用于每个变量的规则的数量。注意，用户缺失值和系统缺失值不包含在摘要中。“显示”下拉列表控制显示哪些变量；您可以从**所有变量**、**数值变量**、**字符串变量**和**日期变量**中选择。

规则。 要对分析变量应用规则，请选择一个或多个变量，然后在“规则”列表中选中要应用的所有规则。“规则”列表仅显示适用于所选分析变量的规则。例如，如果选择了数值分析变量，则仅显示数值规则；如果选择了字符串变量，则仅显示字符串规则。如果未选择任何分析变量，或者选择的分析变量具有混合数据类型，则不显示任何规则。

变量分布。 “分析变量”列表中显示的分布摘要可基于所有个案或基于前 n 个个案的扫描，这在“个案数”文本框中指定。单击**重新扫描**可更新分布摘要。

验证数据：交叉变量规则

图片 3-5
“验证数据”对话框，“交叉变量规则”选项卡

“交叉变量规则”选项卡显示可用的交叉变量规则，并允许将其应用于您的数据。要定义其他交叉变量规则，请单击**定义规则**。 有关详细信息，请参阅第 5 页码第 2 章中的定义交叉变量规则。

验证数据：输出

图片 3-6
“验证数据”对话框，“输出”选项卡

个案情况报表。 如果您应用了任何单变量或交叉变量验证规则，则可请求列出每个个案的确认违反规则的报告。

- **最小违规数。** 此选项指定要包括在报告中的个案的最小违规数。指定一个正整数。
- **个案的最大数量。** 此选项指定个案报告中包含的个案的最大数量。请指定一个小于等于 1000 的正整数。

单变量确认规则。 如果已经应用了任何单变量验证规则，则可选择如何显示结果或者是否显示结果。

- **依据分析变量汇总违规数。** 对于每个分析变量，此选项均显示违反的所有单变量验证规则以及违反每个规则的值的数量。它还报告每个变量违反单变量规则的总次数。
- **依据规则汇总违规数。** 对于每个单变量验证规则，此选项均报告违反了该规则的变量以及每个变量的无效值的数量。它还报告全部变量违反每个规则的值的总数。

显示描述统计。 此选项允许您请求分析变量的描述统计。会为每个分类变量生成一个频率表。为刻度变量生成包括均值、标准差、最小值和最大值的摘要统计表。

移动具有确认违反规则的个案。 此选项将违反了单变量规则或交叉变量规则的个案移动到活动数据集的顶部以便于查阅。

验证数据：保存

图片 3-7
“验证数据”对话框，“保存”选项卡

“保存”选项卡允许将记录违规的变量保存到活动数据集。

摘要变量。 这些是可以保存的单个变量。选中一个框可保存该变量。为这些变量提供了默认名称；您可以进行编辑。

- **空个案指示器。** 空个案会分配值 1。所有其他个案都具有代码 0。变量的值反映在“基本检查”选项卡上指定的范围。
- **双 ID 组。** 具有相同个案标识的个案（具有不完整标识的个案除外）会分配有相同的组号。具有唯一标识或不完整标识的个案都具有代码 0。
- **ID 指示器不完整。** 具有空的或不完整的个案标识的个案将分配值 1。所有其他个案的代码都为 0。
- **确认规则违反(总数)。** 这是按个案计数的违反单变量和交叉变量验证规则的总数。

替换现有的摘要变量。 保存到数据文件的变量必须具有唯一的名称，否则就会替换具有相同名称的变量。

保存指示变量。 此选项允许保存确认违反规则的完整记录。每个变量都对应着验证规则的一次应用，如果个案违反了该规则，则值为 1，如果未违反，则值为 0。

章 4

自动数据准备

准备分析数据是任何项目中最重要的步骤之一，而从传统来说也是最耗时的步骤之一。“自动数据准备 (ADP)”为您处理任务，分析您的数据并识别修正，筛选出存在问题或可能无用的字段，并在适当的情况下派生新的属性，通过智能筛选技术改进性能。您可以通过完全**自动**的方式使用算法，这种方式可以允许选择并应用修正；或者也可以通过**交互式**方式使用算法，这种方式可以在做出更改前对其进行预览，并按照需要进行接受或拒绝。

通过使用 ADP，您可以快速、轻松地准备数据以供建模，无需具备相关统计概念的预备知识。您可以更快速地构建模型并进行评分。此外，使用 ADP 还能提高自动化建模过程。

注意：当 ADP 准备字段进行分析时，它将创建包含调整或转换的新字段，而不是替换旧字段的现有值和属性。旧字段不用于进一步分析，其角色被设置为?无?。 同时请注意，任何用户缺失值信息都不会转移到这些新建的字段，而新字段中的任何缺失值将成为系统缺失值。

示例。 在调查业主保险理赔方面拥有有限资源的保险公司希望构建一个模型来标记具有潜在欺骗性的可疑理赔。构建模型前，他们将使用自动数据准备来准备数据进行建模。由于他们希望能够在应用转换前查看建议的转换，他们将在交互模式下使用自动数据准备。 有关详细信息，请参阅第 78 页码第 8 章中的交互式使用自动数据准备。

某汽车集团希望跟踪各类私人汽车的销售情况。为了能够标识表现良好和表现不好的型号，他们希望建立汽车销售和汽车特性之间的关系。他们将使用自动数据准备来准备数据进行分析，同时使用准备“之前”和“之后”的数据构建模型以查看结果的差别。有关详细信息，请参阅第 89 页码第 8 章中的自动使用自动数据准备。

图片 4-1
“自动数据准备目标”选项卡

您的目标是什么？自动数据准备可以推荐能够加快其他算法的建模速度、并增强这些模型的预测能力的数据准备步骤。可包括转换、构建和选择功能。也可对目标进行转换。您可以指定数据准备过程应遵循的建模优先级次序。

- **均衡速度和精确度。**该选项可以准备数据，以使建模算法处理数据的速度和预测的精确度具有同等优先级。
- **优化速度。**该选项可以准备数据，以使建模算法处理数据的速度具有较高优先级。如果您处理非常大的数据集，或要求快速得到结果时，则选择此选项。
- **优化精确度。** 该选项可以准备数据，以使建模算法生成的预测结果的精确度具有较高优先级。
- **自定义分析。** 如果您希望手动修改“设置”选项卡上的算法，请选择此选项。注意，如果您随后在“设置”选项卡上更改了与其他目标之一不一致的选项，则会自动选择该设置。

获得自动数据准备

从菜单中选择：
转换 > 准备建模数据 > 自动...

▶ 单击运行。

根据需要，您可以：

- 在“目标”选项卡上指定目标。

- 在“字段”选项卡上指定字段分配。
- 在“设置”选项卡上指定专家设置。

获得交互式数据准备

从菜单中选择：
转换 > **准备建模数据** > **交互式**...

1. 在对话框顶部工具栏中单击**分析**。
2. 单击“分析”选项卡，并审核建议的数据准备步骤。
3. 如果满意，单击**运行**。否则，单击**清除分析**，更改所需设置，并单击**分析**。

根据需要，您可以：

- 在“目标”选项卡上指定目标。
- 在“字段”选项卡上指定字段分配。
- 在“设置”选项卡上指定专家设置。
- 单击**保存 XML**，将建议的数据准备步骤保存到 XML 文件。

字段选项卡

图片 4-2
“自动数据准备字段”选项卡

“字段”选项卡指定应准备哪些字段以供进一步分析。

使用预定义角色。 此选项使用现有的字段信息。如果存在具有“目标”角色的单个字段，它将用作目标，否则将不存在目标。所有具有预定义角色“输入”的字段将用作输入。需要至少一个输入字段。

使用自定义字段分配。 当您通过将字段从其默认列表中移走来覆盖字段角色时，对话框会自动切换到该选项。当进行自定义字段分配时，请指定以下字段：

- **目标（可选）。** 如果计划构建需要目标的模型，请选择目标字段。这类似于将字段角色设为“目标”。
- **输入。** 选择一个或多个输入字段。这类似于将字段角色设为“输入”。

设置选项卡

“设置”选项卡包含若干组不同的设置，您可以对其进行修改以微调算法如何处理数据。如果您对与其他目标不一致的默认设置进行了更改，则“目标”选项卡会自动更新为选择**自定义分析**选项。

准备日期和时间

图片 4-3
自动数据准备，准备日期和时间

许多建模算法无法直接处理日期和时间细节。这些设置允许您从现有数据中的日期和时间派生新的持续时间数据，以用作模型输入。该字段包含必须采用日期或时间存储类型预定义的日期和时间。不建议在自动数据准备后将原始日期和时间字段用作模型输入。

准备日期和时间以供建模。 取消选择该选项将在保持选择的同时禁用所有其他“准备日期&时间”控件。

计算到参考日期为止已过去的时间。 这将为包含日期的每个变量生成自参考日期后的年/月/日数。

- **参考日期。** 指定以该日期为参考，根据输入数据中的日期信息计算持续时间的日期。如果选择**当前日期**，则 ADP 执行时始终使用当前系统日期。要使用特定日期，选择**固定日期**，并输入所需日期。
- **持续日期的单位。** 指定 ADP 是自动确定持续日期的单位，还是从**固定单位**（年、月或日）中选择。

计算到参考时间为止已过去的时间。 这将为包含时间的每个变量生成自参考日期后的小时/分钟/秒数。

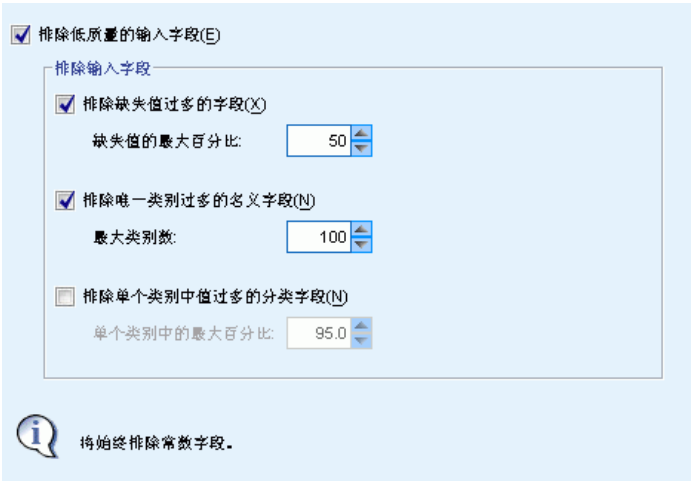
- **参考时间。** 指定以该时间为参考，根据输入数据中的日期信息计算持续的时间。如果选择**当前时间**，则 ADP 执行时始终使用当前系统时间。要使用特定时间，选择**固定时间**，并输入所需具体时间。
- **持续时间的单位。** 指定 ADP 是自动确定持续时间的单位，还是从**固定单位**（小时、分或秒）中选择。

提取循环时间元素。 使用这些设置将单个日期或时间字段分割成一个或多个字段。例如，如果您选择了全部三个日期复选框，则输入日期字段 “1954-05-23” 会被分割成三个字段：1954、5 和 23，分别使用在**字段名称**面板中定义的后缀，原始日期字段则被忽略。

- **从日期提取。** 对于任何日期输入，请指定是否要提取年、月、日或任意组合。
- **从时间提取。** 对于任何时间输入，请指定是否要如果要提取小时、分、秒或任意组合。

排除字段

图片 4-4
自动数据准备排除字段设置



质量较差的数据会影响到预测的准确性，因此需要为输入特征指定可接受的质量级别。所有为常量或缺失值达 100% 的字段自动被排除。

排除低质量的输入字段。 取消选择该选项将在保持选择的同时禁用所有其他“排除字段”控件。

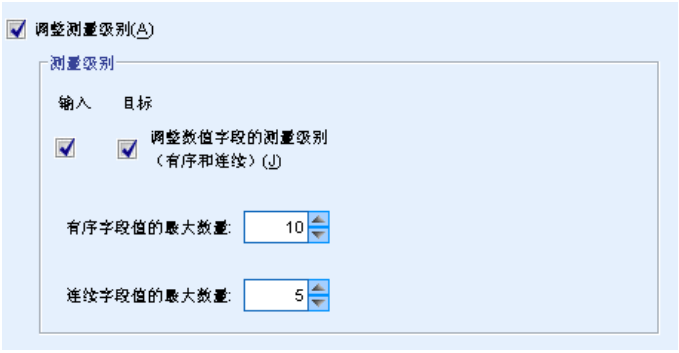
排除缺失值过多的字段。 删除缺失值超过指定百分比的字段，而不会用于进一步分析。指定大于或等于 0 的值等同于取消选择该选项，同时指定小于或等于 100 的值将自动排除具有所有缺失值的字段。默认值为 50。

排除唯一类别过多的名义字段。 删除类别超过个数的字段，而不会用于进一步分析。指定一个正整数。默认值为 100。这对于自动从建模中删除包含记录特有信息（如 ID、地址或名称）的字段非常有用。

排除单个类别中值过多的分类字段。 删除在单个类别中包含超过指定百分比的记录的有序和名义字段，而不会用于进一步分析。指定大于或等于 0 的值等同于取消选择该选项，同时指定小于或等于 100 的值将自动排除常数字段。默认值为 95。

调整测量

图片 4-5
自动数据准备调整测量设置



调整测量级别。 取消选择该选项将在保持选择的同时禁用所有其他“调整测量”控件。

测量级别。 指定是否将“值太少”的连续字段的测量级别调整为有序，并将“值太多”的有序字段的测量级别调整为连续。

- **有序字段值的最大数量。** 具有超过指定类别数目的有序字段将被重新设计为连续字段。指定一个正整数。默认值为 10。该值必须大于或等于连续字段值的最小数量。
- **连续字段值的最小数量。** 具有少于指定唯一值数目的连续字段将被重新设计为有序字段。指定一个正整数。默认值为 5。该值必须小于或等于有序字段值的最大数量。

提高数据质量

图片 4-6
自动数据准备提高数据质量设置

准备要提高数据质量的字段。 取消选择该选项将在保持选择的同时禁用所有其他“提高数据质量”控件。

处理离群值。 指定是否为输入和目标替换离群值；如果是，则指定离群值截断标准（采用标准差测量）和离群值替换方法。可以通过修整（设置为截断值）或将其设置为缺失值来替换离群值。在任何离群值被设置为缺失值后，将按照下面所选的缺失值处理设置进行处理。

替换缺失值。 指定是否替换连续、名义或有序字段的缺失值。

重新排序名义字段。 选中此选项，以按从小（最少出现）到大（最常出现）的类别顺序重新编码名义（集合）字段值。新字段值从 0 开始作为最少出现的类别。注意，如果原始字段为字符串，新字段将为数值。例如，如果名义字段的数据值为“A”、”A”、”A”、”B”、”C”、”C”，那么自动数据准备将把 “B” 重新编码为 0、将 “C” 编码为 1，同时将 “A” 编码为 2。

重新调整字段

图片 4-7
自动数据准备重新调整字段设置

重新调整字段。 取消选择该选项将在保持选择的同时禁用所有其他“重新调整字段”控件。

分析权重。此变量包含分析（回归或抽样）权重。分析权重将作为对目标字段各个水平上方差的差异的一种考量。选择一个连续字段。

连续输入字段。 这将使用 **Z 得分转换**或**最小/最大转换**来标准化连续输入字段。当您在“选择和构建”设置中选择了**执行特征构建**时，重新调整输入特别有用。

- **Z 得分转换。** 以观察到的均值和标准差作为总体参数估计，将字段标准化，然后将 z 得分映射到具有指定**最终均值**和**最终标准差**的正态分布的对应值。为**最终均值**指定一个数字并为**最终标准差**指定一个正数。默认值为 0 和 1，分别对应于标准化重新调整。
- **最小/最大转换。** 以观察到的最小值和最大值作为总体参数估计，将字段映射到具有指定**最小值**和**最大值**的均匀分布的对应值。在指定数字值时，确保**最大值**大于**最小值**。

连续目标。 这将使用 Box-Cox 转换来转换连续目标，其结果字段为近似正态分布，且具有指定的**最终均值**和**最终标准差**。为**最终均值**指定一个数字并为**最终标准差**指定一个正数。默认值分别为 0 和 1。

注意：如果目标已被 ADP 转换，则使用转换后目标构建的后续模型将针对转换后的单位评分。要解释和使用结果，您必须将预测值转换回原始尺度。 有关详细信息，请参阅第 39 页码逆转换得分。

转换字段

图片 4-8
自动数据准备转换字段设置

为提高数据预测能力，您可以转换输入字段。

转换建模字段。 取消选择该选项将在保持选择的同时禁用所有其他“转换字段”控件。

分类输入字段

- **合并松散类别以最大化与目标的关联。** 选中此选项，可以减少与目标关联的需处理的字段数目，得到更简约的模型。通过输入与目标间的关系可以确定类似的类别。无显著差异（即 p 值大于指定值）的类别则被合并。指定一个大于 0 且小于或等于 1 的值。如果将所有类别合并为单个类别，则会从进一步分析中排除字段的原始和派生版本，因为它们没有值作为预测变量。
- **没有目标时，根据以下计数合并松散类别。** 如果数据集没有目标，您可以选择合并有序和名义字段的松散类别。等频法用于合并具有低于指定的总记录数最小百分比的类别。指定一个大于或等于 0 且小于或等于 100 的值。默认值为 10。当不存在具有低于指定最小个案百分比的类别，或只剩下两个类别时，合并停止。

连续输入字段。 如果数据集包含类别目标，则可以采用强关联对连续输入分级，以改进处理性能。块是根据“齐次子集”的属性来创建，后者通过 Scheffe 方法进行确定，并使用指定的 p 值作为确定齐次子集的临界值 alpha。指定一个大于 0 且小于或等于 1 的值。默认值为 0.05。如果特定字段的离散化结果为单个块，则会排除字段的原始和分级版本，因为它们没有值作为预测变量。

注意：ADP 中的离散化与最佳离散化不同。最佳离散化使用熵信息将连续字段转换为分类字段。这需要在内存中对全部数据进行排序和存储。ADP 使用齐次子集来离散化连续字段，这意味着 ADP 离散化不需要在内存中对全部数据进行排序和存储。通过使用齐次子集方法离散化连续字段，离散化后的类别数总是小于或等于目标中的类别数。

选择和构建

图片 4-9
自动数据准备选择和构建设置

为提高数据预测能力，您可以根据现有字段构建新的字段。

执行特征选择。 如果某个连续输入与目标关联的 p 值大于指定的 p 值，则从分析中删除此连续输入。

执行特征构建。 选择该选项从若干现有特征组合派生出新特征。旧特征将不用于进一步分析。该选项仅适用于目标为连续或不存在目标的连续输入特征。

字段名称

图片 4-10
自动数据准备命名字段设置

为方便识别新的和转换后的特征，ADP 可以创建并应用基本新名称、前缀或后缀。您可以更改这些名称，以使其与您的要求和数据更相关。

转换字段和构建字段。 指定要应用到转换目标和输入字段的名称扩展。

此外，还需要指定要应用到通过“选择和构建”设置所构建的任何特征的前缀名称。新名称将通过为此前缀根名称添加数字后缀生成。数字格式取决于生成的特征数目，例如：

- 第 1-9 个构建的特征将命名为：feature1 到 feature9。
- 第 10-99 个构建的特征将命名为：feature01 到 feature99。
- 第 100-999 个构建的特征将命名为：feature001 到 feature999，依此类推。

这可以确保不论有多少个特征，都将按有意义的顺序排列。

从日期和时间算出的持续时间。 指定要应用到从日期和时间计算的持续时间的名称扩展。

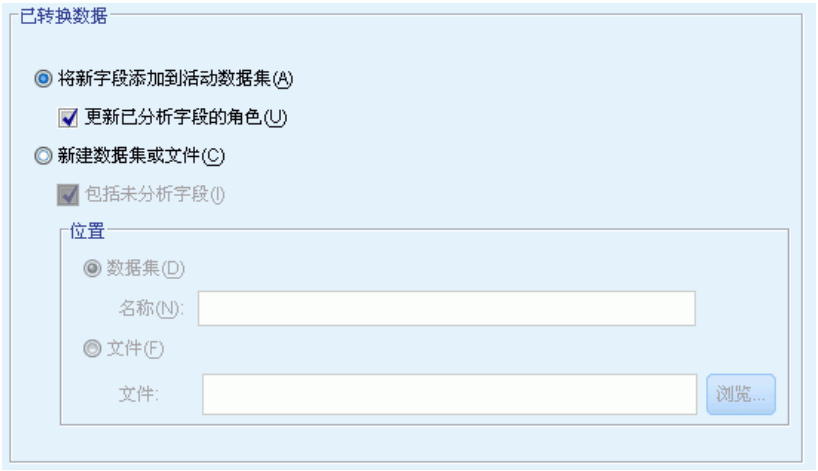
从日期和时间提取的循环元素。 指定要应用到从日期和时间提取的循环元素的名称扩展。

应用和保存转换

根据您在使用交互式还是自动数据准备对话框，应用和保存转换的设置也略有差异。

交互式数据准备应用转换设置

图片 4-11
交互式数据准备应用转换设置



已转换数据。这些设置指定已转换数据的保存位置。

- **将新字段添加到活动数据集。**自动数据准备所创建的任何字段都将作为新字段添加到活动数据集。**更新已分析字段的角色**可将由自动数据准备从进一步分析中排除的任何字段的角色设置为“无”。
- **新建包含已转换数据的数据集或文件。** 自动数据准备所建议的字段都将添加到新数据集或文件中。**包括未分析字段**会将在“字段”选项卡上未指定的原始数据集字段添加到新数据集。这对于将包含未在建模中使用的信息（如 ID、地址或名称）的字段转移到新数据集中非常有用。

自动数据准备应用和保存设置

图片 4-12
自动数据准备应用和保存设置

“已转换数据”组与“交互式数据准备”中的相同。在自动数据准备中，有下列其他选项可用：

应用转换。在“自动数据准备”对话框中，取消选择本选项将在保持选择的同时禁用所有其他“应用和保存”控件。

将转换另存为语法。这可将建议的转换作为命令语法保存到外部文件。“交互式数据准备”对话框则不包含此控件，因为它可在您单击**粘贴**时将转换作为命令语法粘贴到语法窗口。

将转换另存为 XML。这可将建议的转换作为 XML 保存到外部文件，后者可通过 TMS MERGE 与模型 PMML 合并，或通过 TMS IMPORT 应用到其他数据集。“交互式数据准备”对话框则不包含此控件，因为它可在您单击对话框顶部工具栏中的**保存 XML** 时将转换另存为 XML。

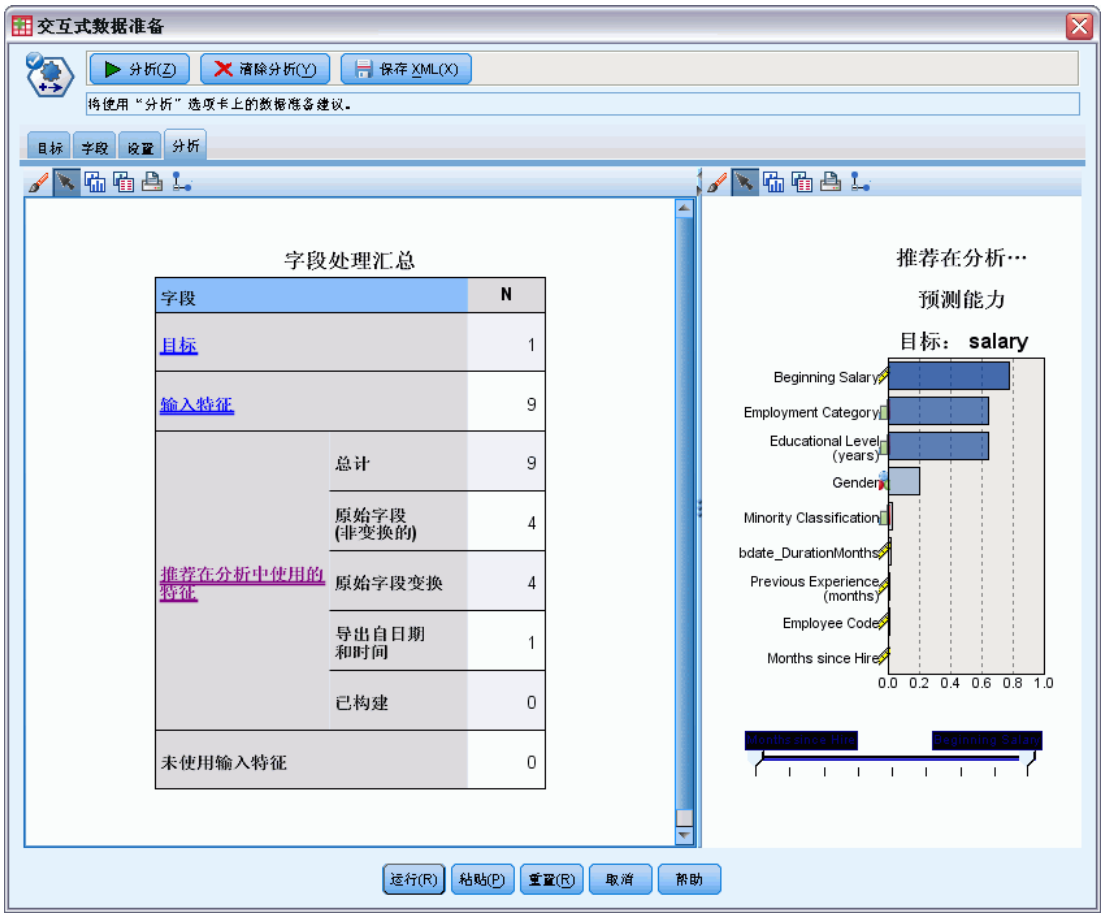
分析选项卡

注意： 在“交互式数据准备”对话框中使用“分析”选项卡，您可以审核建议的转换。“自动数据准备”对话框不包含此步骤。

- 在完成对 ADP 的设置（包括对“目标”、“字段”和“设置”选项卡所作的任何更改）后，单击**分析数据**。算法将设置应用到数据输入，并在“分析”选项卡上显示结果。

“分析”选项卡包含表格和图形输出，其中显示数据处理概要，并显示有关如何修改或改进数据以提高得分的建议。您可以审核这些建议，并加以接受或拒绝。

图片 4-13
“自动数据准备分析”选项卡



“分析”选项卡包含两个面板，主视图位于左侧，链接或辅助视图位于右侧。有三个主视图：

- 字段处理概要（默认视图）。 有关详细信息，请参阅第 29 页码字段处理概要。
- 字段。 有关详细信息，请参阅第 30 页码字段。
- 操作摘要。 有关详细信息，请参阅第 32 页码操作摘要。

有四个链接/辅助视图：

- 预测能力（默认视图）。 有关详细信息，请参阅第 33 页码预测能力。
- 字段表。 有关详细信息，请参阅第 34 页码字段表。
- 字段详细信息。 有关详细信息，请参阅第 35 页码字段详细信息。
- 操作详细信息。 有关详细信息，请参阅第 37 页码操作详细信息。

视图间链接

在主视图内，表格中的下划线文本控制链接视图中的显示。单击文本将显示有关特定字段、字段集合或处理步骤的详细信息。您最近一次选择的链接显示为深色，这可帮助您识别两个视图面板内容间的联系。

重置视图

要重新显示原始分析建议，并放弃对分析视图的任何更改，请单击主视图面板底部的重置。

字段处理概要

图片 4-14
字段处理摘要

字段处理摘要

字段		N
目标		1
输入特征		9
建议在分析中使用的特征	总计	8
	原始字段（未转换）	1
	原始字段的转换	7
	从日期和时间派生	0
	已构建	0
未使用的输入特征		1

“字段处理摘要”表格提供了有关字段处理的预计总体影响的快照，包括对特征状态的更改和构建的特征数目。

请注意，这里不会实际构建模型，因此并不存在总体预测能力在数据准备前后的变化测量或图表，您只能显示单个建议预测变量的预测能力图表。

该表格显示以下信息：

- 目标字段数。
- 原始（输入）预测变量数。

- 在分析和建模中建议使用的预测变量数。其中包括建议的字段总数、建议的原始和未转换的字段数、建议的转换字段数（排除任何字段的中间版本、从日期/时间预测变量派生的字段以及构建的预测变量）、从日期/时间字段派生的建议字段数，以及建议的构建预测变量数。
- 不建议以任何形式（原始、派生字段和构建预测变量的输入）使用的输入预测变量数。

如果任何**字段**信息带有下划线，单击可在链接视图中显示更多信息。在“字段表链接视图”中显示**目标**、**输入特征**和**未使用输入特征**的详细信息。有关详细信息，请参阅第 34 页码字段表。 在“预测能力”链接视图中显示**建议在分析中使用的特征**。 有关详细信息，请参阅第 33 页码预测能力。

字段

图片 4-15
字段

“字段”主视图显示处理过的字段，以及 ADP 是否建议在下游模型中使用它们。您可以覆盖任何字段建议。例如，排除构建的特征或包含 ADP 建议排除的特征。如果字段已转换，您可以决定是接受建议转换，还是使用原始版本。

“字段”视图由两个表格组成，分别显示目标和处理或创建的预测变量。

目标表

仅当数据中定义有目标时，才会显示**目标**表。

该表包含两列：

- **名称。** 此为目标字段的名称或标签。不论字段是否已转换，始终使用原始名称。
- **测量级别。** 此列显示代表测量级别的图标。将鼠标悬停在图标上可以显示数据描述标签（连续、有序、名义等）。

  如果目标已转换，则**测量级别**列将反映最终转换版本。注意：您不能关闭目标转换。

预测变量表格

预测变量表格总是显示。表格的每一行代表一个字段。默认情况下，按预测能力的降序来排列行。

对于普通特征，原始名称始终用作行名称。表格中以单独行显示日期/时间字段的原始和派生版本，此外，还包括构建的预测变量。

注意，在表格中显示的字段转换后版本始终代表最终版本。

默认情况下，在预测变量表中只显示建议的字段。要显示其余字段，选中表格上方的**在表中包括非推荐字段**复选框，这些字段随即显示在表格底部。

该表包含以下列：

- **使用的版本。** 此列显示一个下拉列表，以控制字段是否将在下游使用，以及是否使用建议的转换。默认情况下，下拉列表将反映建议。

  对于已转换的普通预测变量，下拉列表有三个选项：**已转换**、**原始**和**不使用**。

  对于未转换的普通预测变量，下拉列表的选项为：**原始**和**不使用**。

  对于派生的日期/时间字段和构建的预测变量，选项为：**已转换**和**不使用**。

  对于原始日期字段，下拉列表被禁用，并设置为**不使用**。

  注意：对于同时具有原始和已转换版本的预测变量，如果切换**原始**和**已转换**版本，则会自动更新这些特征的**测量级别**和**预测能力**设置。
- **名称。** 每个字段的名称均为链接。单击名称可以在链接视图中显示有关该字段的更多信息。 有关详细信息，请参阅第 35 页码字段详细信息。
- **测量级别。** 此列显示代表数据类型的图标。将鼠标悬停在图标上可以显示数据描述标签（连续、有序、名义等）。
- **预测能力。** 只会对 ADP 建议的字段显示预测能力。如果未定义目标，则不会显示此列。预测能力范围从 0 到 1，其中较大的值表示“更好的”预测变量。通常，预测能力对于比较一个 ADP 分析内的预测变量有用，但不应跨分析比较预测能力值。

操作摘要

图片 4-16
操作摘要

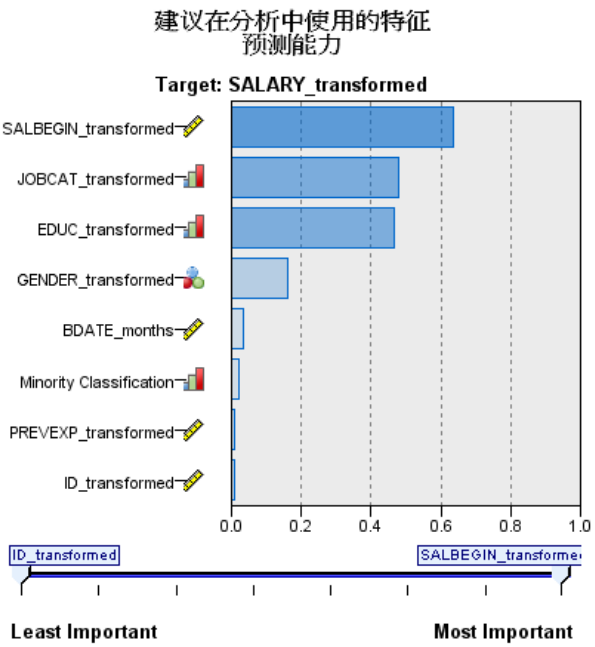
对于自动数据准备的每个操作，将会转换和/或过滤掉输入预测变量。保留的字段将用于下一个操作。在最后步骤中保留的字段将被建议用于建模，转换和构建预测变量的输入则被过滤掉。

“操作摘要”是一张简单列表，列出了 ADP 所执行的处理操作。如果任何**操作**带有下划线，单击可在链接视图中显示有关所执行操作的更多信息。 有关详细信息，请参阅第 37 页码操作详细信息。

注意：只会显示每个字段的原始和最终转换版本，而不会显示在分析过程中使用的任何中间版本。

预测能力

图片 4-17
预测能力



在首次运行分析时默认显示，或者在“字段处理概要”主视图中选择了**建议在分析中使用的预测变量**时显示，该图表显示建议预测变量的预测能力。字段按其预测能力排序，预测能力值最高的字段显示在顶端。

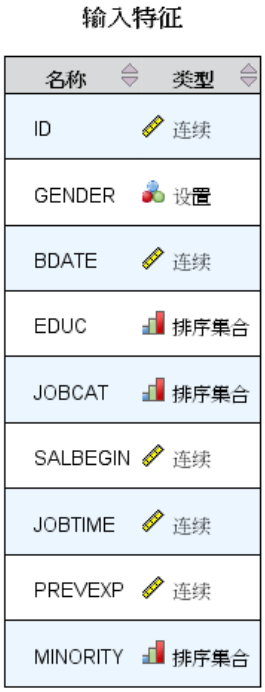
对于普通预测变量的转换后版本，字段名称将反映您在“设置”选项卡的“字段名称”面板中选择的后缀，例如：_transformed。

测量级别图标显示在各个字段名后面。

每个建议预测变量的预测能力通过线性回归或 naïve Bayes 模型进行计算，具体取决于目标是连续还是分类。

字段表

图片 4-18
字段表



在“字段处理摘要”主视图中单击**目标**、**预测变量**或**未使用预测变量**时显示，“字段表”视图显示一个简单表，其中列出了相关特征。

该表包含两列：

- **名称。** 预测变量名。

  对于目标变量，不论其是否已转换，始终使用字段的原始名称或标签。

  对于普通预测变量的转换后版本，其名称将反映您在“设置”选项卡的“字段名称”面板中选择的后缀，例如：_transformed。

  对于从日期和时间派生的字段，将使用最终转换版本的名称，例如：bdate_years。

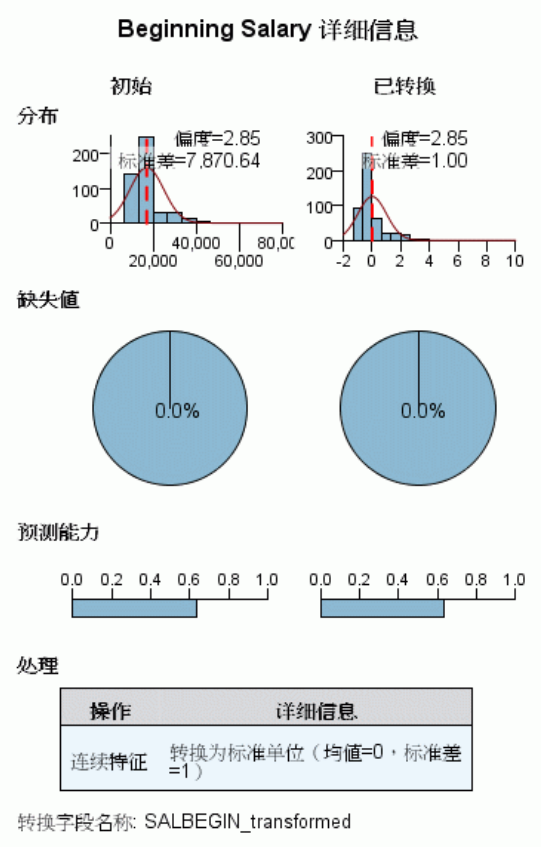
  对于构建的预测变量，将使用构建预测变量的名称，例如：Predictor1。

- **测量级别。** 此列显示代表数据类型的图标。

  对于目标，**测量级别**始终反映转换后的版本（如果目标已转换）。例如，从有序（有序集合）转换为连续（范围、尺度），反之亦然。

字段详细信息

图片 4-19
字段详细信息



在“字段”主视图中单击任何**名称**时显示，“字段详细信息”视图包括选定字段的分布、缺失值和预测能力图表（如果适用）。此外，字段的处理历史和转换后的字段名称也将显示（如果适用）

对于每个图表集，两个版本将并排显示，以比较字段在应用转换前后的情况。如果字段的转换后版本不存在，则只显示原始版本的图表。对于派生的日期或时间字段和构建的预测变量，只显示新预测变量的图表。

注意：如果字段因为类别太多而被排除，则只显示处理历史。

分布图

连续字段分布显示为直方图，并叠放一条正态分布曲线，还有一条均值垂直参考线。类别字段显示为条形图。

直方图带有标签以显示标准差和偏度。不过，如果值个数等于或低于 2，或原始字段的方差低于 10-20，则不会显示偏度。

将鼠标悬停在图表的上方，可以显示直方图的均值，或条形图中类别计数与占记录总数的百分比。

缺失值图表

该图表显示为饼图，以比较在应用转换前后的缺失值百分比。图表标签显示百分比。

如果 ADP 执行了缺失值处理，则转换后的饼图还应包含替换值作为标签，即用于替换缺失值的值。

将鼠标悬停在图表的上方，可以显示缺失值计数和占记录总数的百分比。

预测能力图表

对于建议的字段，以条形图形式显示转换前后的预测能力。如果目标已经过转换，则计算的预测能力对应于转换后的目标。

注意：如果未定义目标，或在“主视图”面板中单击目标，将不会显示预测能力图表。

将鼠标悬停在图表的上方，可以显示预测能力值。

处理历史表

该表格显示字段的转换后版本是如何派生的。ADP 采取的操作按照其执行顺序列出。不过，对于某些步骤，可能对特定字段执行了多个操作。

注意：该表格不显示未转换字段的处理历史。

表中的信息分为二或三列：

- **操作。** 操作的名称。例如，连续预测变量。 有关详细信息，请参阅第 37 页码操作详细信息。
- **详细信息。** 所执行处理的列表。例如，转换成标准单位。
- **函数。** 只针对构建的预测变量显示，其中显示输入字段的线性组合，例如，0.06*age + 1.21*height。

操作详细信息

图片 4-20
ADP 分析 - 操作详细信息

步骤 9: 连续特征

转换	特征个数	标准	
		平均值	标准差
转换为标准单位	5	0	1

特征空间构建	N
特征已构建	0
特征因与目标关联较低而被排除	1
特征因在离散化后为常数而被排除	0

在“操作摘要”主视图中选择任何带有下划线的操作时显示，“操作详细信息”链接视图显示所执行的每个处理步骤的操作相关与通用信息。首先显示操作相关的详细信息。

对于每个操作，描述用作标题位于链接视图的顶部。操作相关详细信息显示在标题下方，可能包括派生预测变量数目、字段重新设计、目标转换、类别合并或重新排序和预测变量构建或排除等详细信息。

在处理每个操作时，在处理过程中使用的预测变量数可能会变化，例如，排除或合并预测变量。

注意：如果某个操作已关闭，或未指定目标，则在“操作摘要”主视图中单击该操作时，会在操作详细信息位置显示一条错误消息。

有 9 个可能的操作，不过对于每个分析而言，这些操作并非都有必要使用。

文本字段表

该表显示下列项的数目：

- 从分析中排除的预测变量。

日期和时间预测变量表

该表显示下列项的数目：

- 从日期和时间预测变量派生的持续时间。

- 日期和时间元素。
- 派生的日期和时间预测变量总数。

如果已计算了任何日期持续时间，则参考日期或时间将显示为脚注。

预测变量筛选表

该表显示从处理中排除的以下预测变量数目：

- 常量。
- 缺失值过多的预测变量。
- 在单个类别中有太多个案的预测变量。
- 类别过多的名义字段（集合）。
- 筛选出的预测变量总数。

检查测量级别表

该表显示重新设计、分解成以下项的字段数目：

- 重新设计为连续字段的有序字段（有序集合）。
- 重新设计为有序字段的连续字段。
- 重新设计总数。

如果输入字段（目标或预测变量）并非连续或有序，这将显示为脚注。

离群值表

该表显示离群值处理方式的计数。

- 发现并修整其离群值的连续字段数，或发现离群值并将其设为缺失值的连续字段数，具体取决于您在“设置”选项卡的“准备输入和目标”面板上的设置。
- 由于在离群值处理后为常量，而被排除的连续字段数。

离群值分界值显示为脚注。如果输入字段（目标或预测变量）不是连续的，还会显示另一个脚注。

缺失值表

该表显示已替换缺失值、分解为以下项目的字段数：

- 目标。如果未指定目标，则不显示此行。
- 预测变量。它将进一步分解为名义（集合）、有序（有序集合）和连续特征数。
- 被替换的缺失值总数。

目标表

该表显示目标是否被转换，显示为：

- 到正态的 Box-Cox 转换。这将进一步分解为显示指定标准（均值和标准差）和 Lambda 的列。
- 对其重新排序以提高稳定性的目标类别。

分类预测变量表

该表显示以下分类预测变量的数目：

- 按最低到最高重新排序其类别以提高稳定性。
- 合并其类别以最大化目标关联。
- 合并其类别以处理松散类别。
- 由于与目标关联程度过低而被排除。
- 由于在合并后为常量而被排除。

如果没有分类预测变量，则显示相应脚注。

连续预测变量表

有两个表。第一个表格显示以下转换数之一：

- 转换成标准单位的预测变量值。此外，还会显示转换的预测变量数、指定的均值和标准差。
- 映射到通用范围的预测变量值。此外，还会显示通过最值法转换的预测变量数，以及指定的最小值和最大值。
- 离散化的预测变量值和预测变量数。

第二个表显示预测变量空间构建详细信息，显示为以下预测变量的数目：

- 已构建。
- 由于与目标关联程度过低而被排除。
- 由于在离散化后为常量而被排除。
- 由于在构建后为常量而被排除。

如果未输入连续预测变量，则显示相应脚注。

逆转换得分

如果目标已被 ADP 转换，则使用转换后目标构建的后续模型将针对转换后的单位评分。要解释和使用结果，您必须将预测值转换回原始尺度。

图片 4-21
逆转换得分

要逆转换得分，从菜单中选择：
转换 > **准备建模数据** > **逆转换得分**...

- 选择要逆转换的字段。此字段应包含转换后目标的模型预测值。
- 为新字段指定后缀。此新字段将包含采用未转换目标的原始尺度的模型预测值。
- 指定包含 ADP 转换的 XML 文件位置。这应当是从交互式或自动数据准备对话框中保存的文件。 有关详细信息，请参阅第 25 页码应用和保存转换。

章 5

标识异常个案

“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。

统计量。 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，可随机排列个案的顺序。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

▶ 从菜单中选择：
数据 > 标识异常个案...

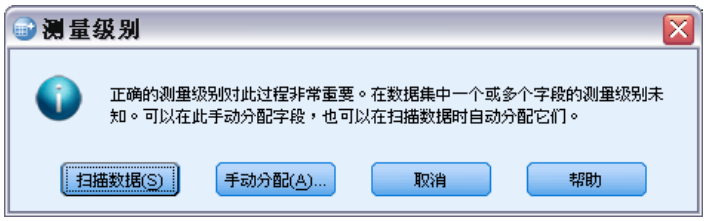
图片 5-1
“标识异常个案”对话框，“变量”选项卡

- 选择至少一个分析变量。
- 还可以选择一个个案标识变量用于标记输出。

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）的测量级别未知时，将显示测量级别警告。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。

图片 5-2
测量级别警报



- **扫描数据。** 读取活动数据集中的数据，并分配默认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。** 打开列出了所有具有未知测量级别的字段的对话框。您可以使用该对话框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视图中分配测量级别。

由于测量级别对该过程很重要，因此您无法访问运行该过程的对话框，除非所有字段均定义了测量级别。

标识异常个案：输出

图片 5-3
“标识异常个案”对话框，“输出”选项卡

异常个案及其被视为异常的原因的列表。 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”选项卡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

图片 5-4
“标识异常个案”对话框，“保存”选项卡

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、变量影响度量、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。

导出模型文件。 允许以 XML 格式保存模型。

标识异常个案：缺失值

图片 5-5
“标识异常个案”对话框，“缺失值”选项卡

“缺失值”选项卡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

图片 5-6
“标识异常个案”对话框，“选项”选项卡

异常个案的标识条件。 这些选择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。 该过程将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。

注意：注意：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。 原因包括变量影响度量、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 命令的附加功能

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。

章 6

最优离散化

“最优离散化”过程通过将每个变量的值分布到块中离散化一个或多个尺度变量（因此称为**离散化输入**变量）。块的构成根据“监督”离散化过程的分类向导变量得以最优化。然后，可以使用块而非原始数据值进行进一步的分析。

示例。 减少变量具有的不同值的数量具有多种用途，包括：

- 其他过程的数据要求。离散化变量可作为分类变量用于需要分类变量的过程。例如，“交叉表”过程要求所有变量均为分类变量。
- 数据隐私。报告离散化值而不是实际值可帮助保护数据源的隐私。“最优离散化”过程可指导块的选择。
- 速度性能。有些过程在处理较少数量的不同值时更加有效。例如，使用离散化变量时“多项 Logistic 回归”的速度会提高。
- 揭示数据的完全分离或准完全分离。

最优离散化与可视离散化。 “可视离散化”对话框提供了多种不使用向导变量创建块的自动方法。这些“未受监督”的规则对于生成描述统计（例如频率表）十分有用，但如果最终目标是生成预测模型，则“最优离散化”更好。

输出。 该过程生成块的分割点以及每个离散化输入变量的描述统计的表。此外，您可以将新变量保存到包含离散化输入变量的离散化值的活动数据集中，并将离散化规则作为命令语法保存以便用于离散化新数据。

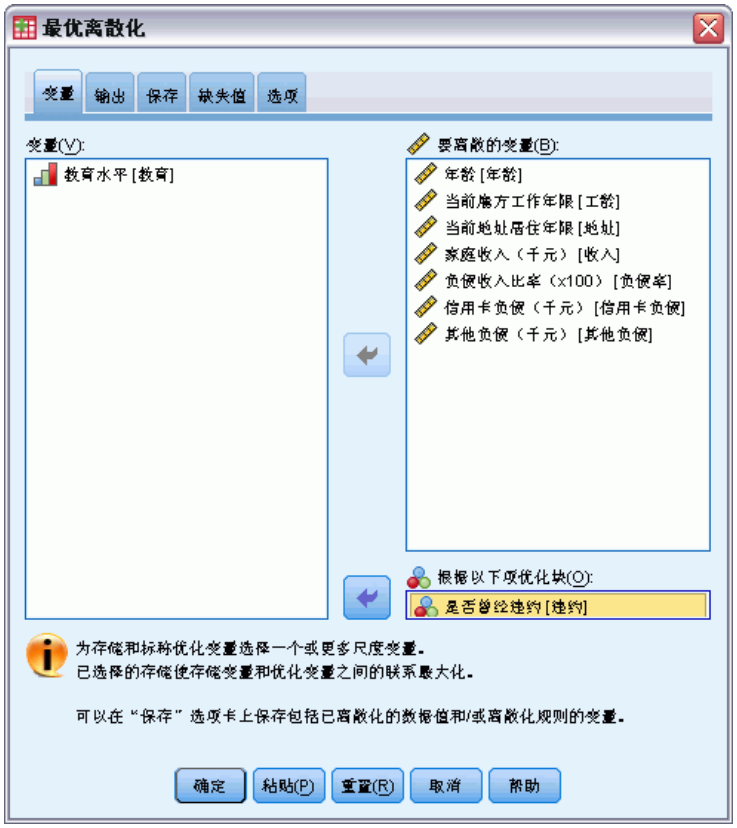
数据。 此过程需要离散化输入变量是数值型刻度变量。向导变量应是分类变量，可以是字符串或数值。

获取最优离散化

从菜单中选择：

转换 > **最优离散化...**

图片 6-1
“最优离散化”对话框，“变量”选项卡



- 选择一个或多个离散化输入变量。
- 选择一个向导变量。

缺省情况下不会生成包含离散化数据值的变量。使用保存选项卡可以保存这些变量。

最优离散化：输出

图片 6-2
“最优离散化”对话框，“输出”选项卡

“输出”选项卡控制结果的显示。

- **块的端点。** 显示每个离散化输入变量的端点集。
- **已离散化变量的描述统计。** 对于每个离散化输入变量，此选项显示具有有效值的个案数、具有缺失值的个案数、不同有效值的数量以及最小值和最大值。对于向导变量，此选项显示每个相关离散化输入变量的类分布。
- **已离散化变量的模型熵。** 对于每个离散化输入变量，此选项显示相对于向导变量的变量预测准确性的测量。

最优离散化：保存

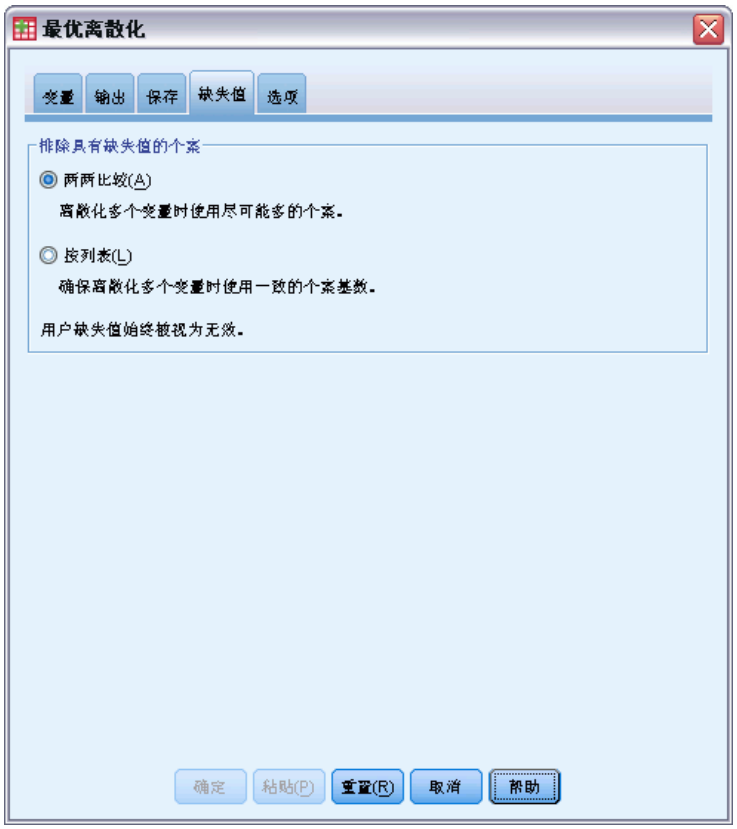
图片 6-3
“最优离散化”对话框，“保存”选项卡

将变量保存到活动数据集。 包含离散化数据值的变量可在进一步分析中代替初始变量。

将离散化规则另存为 语法。生成可用于离散化其他数据集的命令语法。记录的规则基于离散化算法确定的分割点。

最优离散化：缺失值

图片 6-4
“最优离散化”对话框，“缺失值”选项卡



“缺失值”选项卡指定是通过列表删除还是成对删除处理缺失值。用户缺失值总是被视为无效。将初始变量值记录到新变量中时，用户缺失值会转换为系统缺失值。

- **配对方式。** 此选项针对每个向量和离散化输入变量对进行操作。该过程将利用向导和离散化输入变量的具有非缺失值的所有个案。
- **列表**此选项跨“变量”选项卡上指定的所有变量进行操作。如果某个个案的任何变量缺失，则排除整个个案。

最优离散化：选项

图片 6-5
“最优离散化”对话框，“选项”选项卡

预处理。 “预离散化”具有许多不同值的离散化输入变量可缩短处理时间，而不会使最终块的质量发生大幅度下滑。块的最大数量为创建的块的数量设置了一个上限。这样，如果指定 1000 作为最大值，但离散化输入变量的不同值的数量少于 1000，则为离散化输入变量创建的预处理块的数量将等于离散化输入变量中不同值的数量。

稀疏填充的块。 有时候，该过程可能会生成仅具有很少个案的块。下面的方案会删除这些伪分割点：

- 对于给定的变量，假定该算法找到了 n_{final} 个分割点，从而有 $n_{final}+1$ 个块。对于块 i = 2、……、n_{final}（从值第二低的块到值第二高的块），计算

$$\frac{sizeof(b_i)}{\min(sizeof(b_{i-1}),sizeof(b_{i+1}))}$$

  其中 sizeof(b) 是块中的个案数。

- 当此值小于指定的合并阈值时，b_i 被认为是稀疏填充的，并将与 b_{i-1} 或 b_{i+1} 合并，具体取决于哪一个具有较低的类信息熵。

该过程仅穿过这些块一次。

块端点。 此选项指定如何定义区间下限。因为该过程自动确定分割点的值，所以这主要是偏好的问题。

第一个(最低)块/最后一个(最高)块。 这些选项指定如何定义每个离散化输入变量的最小和最大分割点。通常情况下，该过程假设离散化输入变量可采用实数线上的任何值，但是，如果由于某些理论或实际的原因需要限制该范围，则可通过最低值/最高值进行限制。

OPTIMAL BINNING 命令的附加功能

使用命令语法语言还可以：

- 通过均等频率方法执行未受监督的离散化（使用 CRITERIA 子命令）。

请参见命令语法参考以获取完整的语法信息。

部分 II:
示例

章 7

验证数据

“验证数据”过程标识可疑的和无效的个案、变量和数据值。

验证医疗数据库

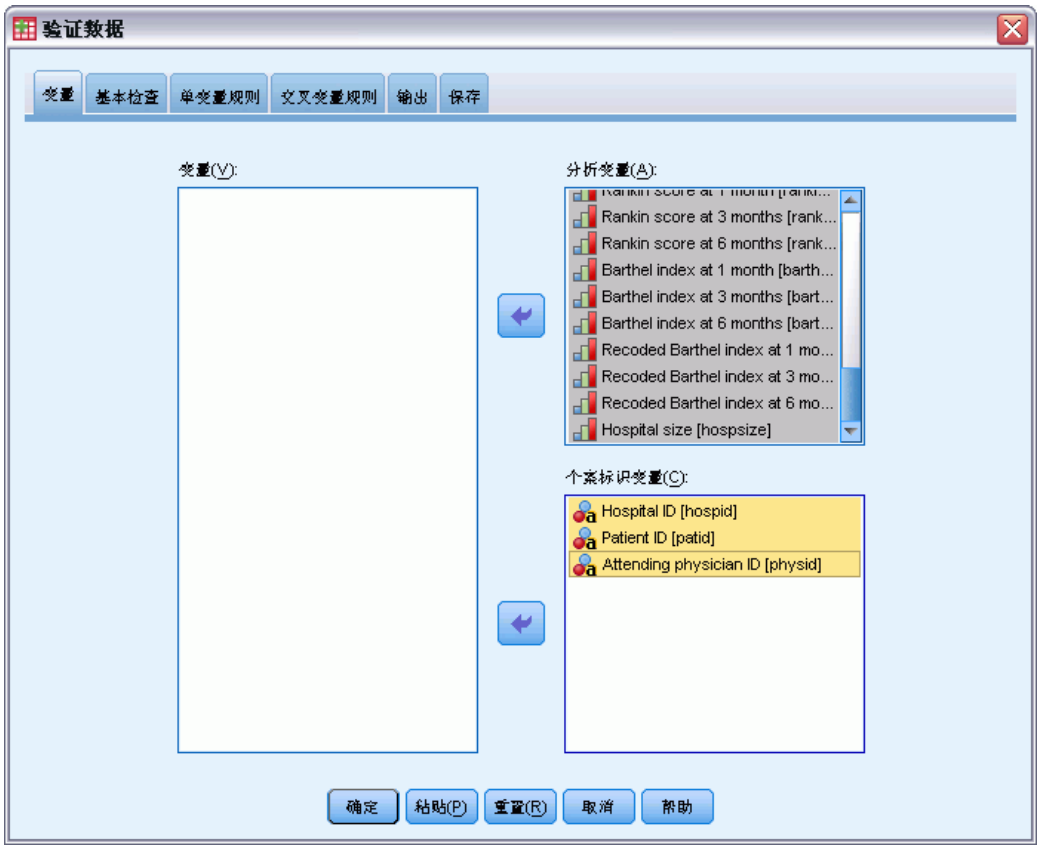
医疗集团雇用的分析人员必须保证系统中的信息的质量。此过程涉及检查值和变量并为数据输入团队的经理准备报告。

数据库的最后状态收集在 stroke_invalid.sav 中。有关详细信息，请参阅第 129 页码附录 A 中的样本文件。 使用?验证数据?过程可获取生成该报告所需的信息。在 validatedata_stroke.sps 中可以找到生成这些分析的语法。

执行基本检查

- 要运行“验证数据”分析，请从菜单中选择：
  数据 > **验证** > **验证数据**...

图片 7-1
“验证数据”对话框，“变量”选项卡



- 选择 Hospital size 和 Age in years 到 Recoded Barthel index at 6 months 作为分析变量。
- 选择 Hospital ID、Patient ID 和 Attending physician ID 作为个案标识变量。
- 单击**基本检查**选项卡。

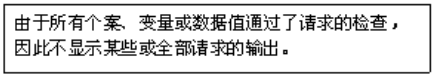
图片 7-2
“验证数据”对话框，“基本检查”选项卡

缺省设置为您想要运行的设置。

▶ 单击**确定**。

警告

图片 7-3
警告



分析变量通过了基本检查，并且不存在空个案，因此会显示一个警告，解释为何不存在与这些检查对应的输出。

不完整标识

图片 7-4
不完整个案标识

	标识符		
观测值	Hospital ID	Patient ID	Attending physician ID
288	OZN		125304
573		6137798782	790697
774		2322241867	176466

个案标识变量中存在缺失值时，该个案将无法正确标识。在此数据文件中，个案 288 缺失 Patient ID，而个案 573 和 774 缺失 Hospital ID。

重复标识

图片 7-5
重复个案标识（显示了前 11 项）

			标识符		
重复的标识符组	重复数	具有重复标识符的个案	Hospital ID	Patient ID	Attending physician ID
1	2	10, 11	PBW	1406462419	355184
2	2	14, 15	PBW	2191527525	355184
3	2	21, 22	PBW	7237535360	616528
4	2	28, 29	NHV	4592215163	942982
5	2	30, 31	NHV	7628592330	371884
6	2	64, 65	NHV	0300750006	371884
7	2	83, 84	QWS	4590625286	215041
8	2	86, 87	QWS	6272818258	817329
9	2	96, 97	QWS	1959349605	215041
10	3	100, 101,102	QWS	5856145337	817329
11	3	104, 105,106	QWS	1543897849	817329

个案应该由标识变量的值的组合唯一标识。这里显示了重复标识表的前 11 个条目。这些重复标识是具有多个事件的患者，他们在每个事件中作为单独的个案被输入。因为这些信息可收集在单个行中，所以这些个案应该清除。

复制和使用其他文件中的规则

分析人员发现此数据文件中的变量与另一个项目中的变量很相似。为那个项目定义的验证规则作为关联数据文件的属性保存，可通过复制该文件的数据属性应用于此数据文件。

▶ 要复制其他文件的规则，请从菜单中选择：
数据 > **复制数据属性**...

图片 7-6
复制数据属性，第 1 步（欢迎）

- 选择从外部 IBM® SPSS® Statistics 数据文件 patient_los.sav 中复制属性。有关详细信息，请参阅第 129 页码附录 A 中的样本文件。
- 单击下一步。

图片 7-7
复制数据属性，第 2 步（选择变量）

这些就是您要将其属性从 patient_los.sav 复制到 stroke_invalid.sav 中的相应变量的变量。

▶ 单击下一步。

图片 7-8
复制数据属性，第 3 步（选择变量属性）

- 取消选择除**定制属性**外的所有属性。
- 单击**下一步**。

图片 7-9
复制数据属性，第 4 步（选择数据集属性）

- 选择**定制属性**。
- 单击**完成**。

现在就可以重用验证规则了。

图片 7-10
“验证数据”对话框，“单变量规则”选项卡

- 要使用复制的规则验证 stroke_invalid.sav 数据，请单击“对话框调用”工具栏按钮并选择**验证数据**。
- 单击**单变量规则**选项卡。

“分析变量”列表显示了在“变量”选项卡上选择的变量，有关其分布的一些摘要信息，以及每个变量附加的规则的数目。属性复制自 patient_los.sav 的变量具有附加到这些变量的规则。

“规则”列表显示数据文件中可用的单变量验证规则。这些规则全部复制自 patient_los.sav。注意，这些规则中的一些适用于在其他数据文件中不具有严格对应项的变量。

图片 7-11
“验证数据”对话框，“单变量规则”选项卡

- 选择 Atrial fibrillation、History of transient ischemic attack、CAT scan result 和 Died in hospital 并应用 0 to 1 Dichotomy 规则。
- 将 0 to 3 Categorical 应用于 Post-event rehabilitation。
- 将 0 to 2 Categorical 应用于 Post-event preventative surgery。
- 将 Nonnegative integer 应用于 Length of stay for rehabilitation。
- 将 1 to 4 Categorical 应用于 Recoded Barthel index at 1 month 到 Recoded Barthel index at 6 months。
- 单击保存选项卡。

图片 7-12
“验证数据”对话框，“保存”选项卡

- 选择**保存用来记录所有确认违规的指示变量**。使用此过程可以更容易地连接导致违反单变量规则的个案和变量。
- 单击**确定**。

规则描述

图片 7-13
规则描述

规则	描述
Nonnegative integer	类型: 数字 域: 范围 标记用户缺失值: 否 标记系统缺失值: 是 极小值: 0 标记范围内未标记的值: 否 标记范围内的非整数值: 是 $VD.SRule[5]: 规则
0 to 1 Dichotomy	类型: 数字 域: 列表 标记用户缺失值: 否 标记系统缺失值: 是 列表: 0, 1 $VD.SRule[1]: 规则
1 to 4 Categorical	类型: 数字 域: 列表 标记用户缺失值: 否 标记系统缺失值: 是 列表: 1, 2, 3, 4 $VD.SRule[4]: 规则

显示至少违反一次的规则。

规则描述表显示对所违反的规则的解释。此功能对于跟踪大量验证规则十分有用。

变量摘要

图片 7-14
变量摘要

	规则	违规数
Age category	1 to 4 Categorical	1
	合计	1
Gender	0 to 1 Dichotomy	1
	合计	1
History of angina	0 to 1 Dichotomy	1
	合计	1
Time to hospital	Nonnegative integer	2
	合计	2
Dead on arrival	0 to 1 Dichotomy	1
	合计	1

变量摘要表列出至少违反了一个验证规则的变量、所违反的规则，以及每个规则和每个变量的违规数。

个案报告

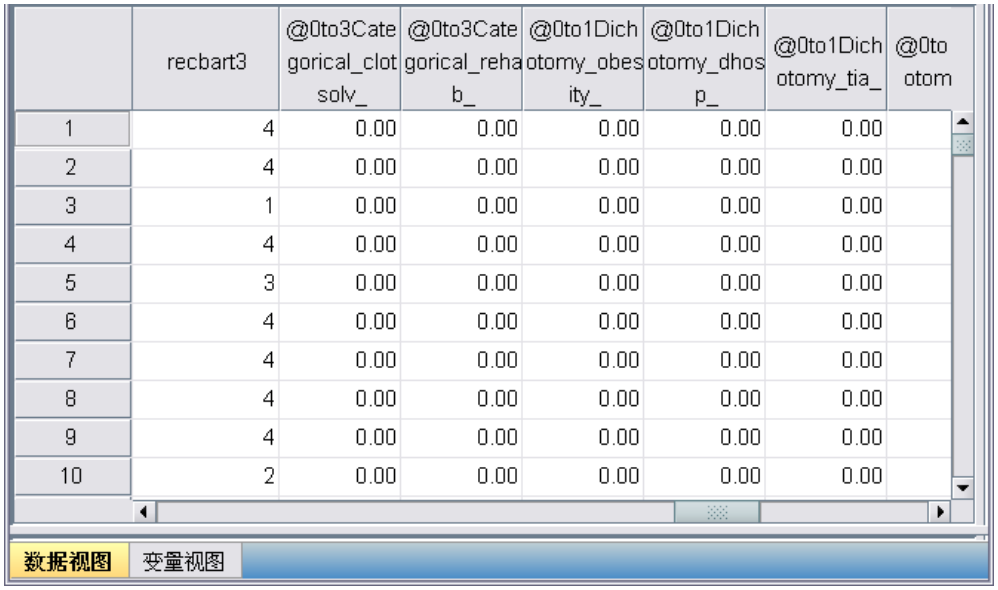
图片 7-15
个案报告

	确认违反规则	标识符		
观测值	单变量[a]	Hospital ID	Patient ID	Attending physidsn ID
175	0 to 1 Dichotomy (1)	OZN	0333204686	883285
274	0 to 1 Dichotomy (1)	OZN	1038840465	103254
310	Nonnegative integer (1)	OZN	2090290204	883285
437	0 to 1 Dichotomy (1)	WPA	2349729006	723384
752	Nonnegative integer (1)	GFG	4993307441	828754
1173	1 to 4 Categorical (1)	ALK	8737661990	185787

a. 违反规则的变量数遵循每个规则。

个案报告表列出至少违反了一个验证规则的个案（包括个案号和个案标识）、所违反的规则，以及个案违反该规则的次数。无效值显示在“数据编辑器”中。

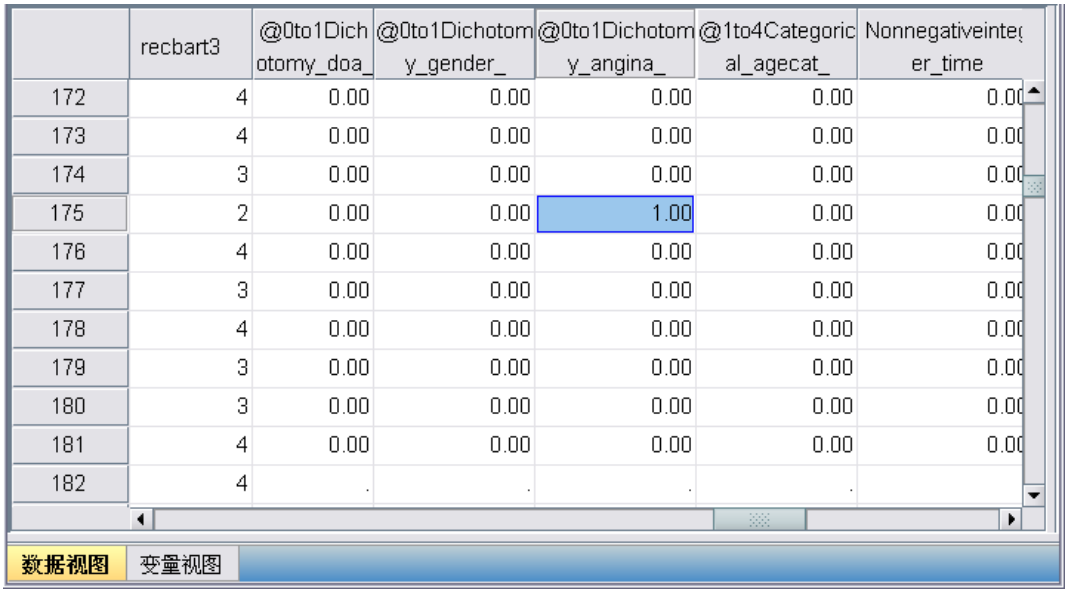
图片 7-16
具有保存的违规指示器的“数据编辑器”



	recbart3	@0to3Categorical_clotsolv_	@0to3Categorical_rehab_	@0to1Dichotomy_obesity_	@0to1Dichotomy_dhosp_	@0to1Dichotomy_tia_	@0tootom
1	4	0.00	0.00	0.00	0.00	0.00	
2	4	0.00	0.00	0.00	0.00	0.00	
3	1	0.00	0.00	0.00	0.00	0.00	
4	4	0.00	0.00	0.00	0.00	0.00	
5	3	0.00	0.00	0.00	0.00	0.00	
6	4	0.00	0.00	0.00	0.00	0.00	
7	4	0.00	0.00	0.00	0.00	0.00	
8	4	0.00	0.00	0.00	0.00	0.00	
9	4	0.00	0.00	0.00	0.00	0.00	
10	2	0.00	0.00	0.00	0.00	0.00	

每次应用验证规则时都会生成一个单独的指示变量。因此，@0to3Categorical_clotsolv_ 表示对变量 Clot-dissolving drugs 应用? to 3 Categorical对于给定个案，指出哪个变量值无效的最简单方法就是扫描指示变量的值。值为 1 表示关联变量的值无效。

图片 7-17
带个案 175 的违规指示器的“数据编辑器”

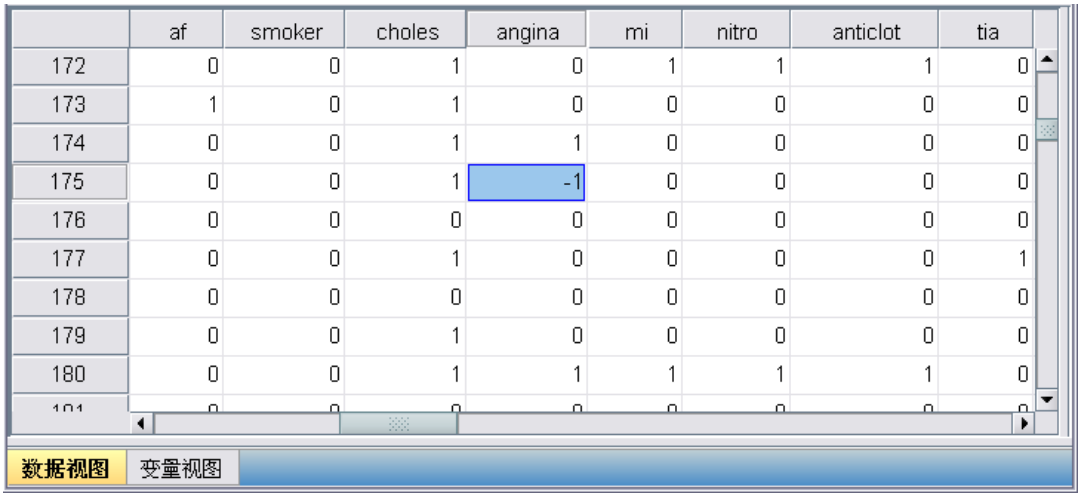


	recbart3	@0to1Dichotomy_doa_	@0to1Dichotomy_gender_	@0to1Dichotomy_angina_	@1to4Categorical_agecat_	Nonnegativeinteger_time
172	4	0.00	0.00	0.00	0.00	0.00
173	4	0.00	0.00	0.00	0.00	0.00
174	3	0.00	0.00	0.00	0.00	0.00
175	2	0.00	0.00	1.00	0.00	0.00
176	4	0.00	0.00	0.00	0.00	0.00
177	3	0.00	0.00	0.00	0.00	0.00
178	4	0.00	0.00	0.00	0.00	0.00
179	3	0.00	0.00	0.00	0.00	0.00
180	3	0.00	0.00	0.00	0.00	0.00
181	4	0.00	0.00	0.00	0.00	0.00
182	4	.	.	.	.	

数据视图 变量视图

转到个案 175，也就是第一个违规的个案。要加快搜索速度，请查看变量摘要表中与这些变量关联的指示器。很容易看到 History of angina 具有无效值。

图片 7-18
具有无效的“History of angina”值的“数据编辑器”



	af	smoker	choles	angina	mi	nitro	anticlot	tia
172	0	0	1	0	1	1	1	0
173	1	0	1	0	0	0	0	0
174	0	0	1	1	0	0	0	0
175	0	0	1	-1	0	0	0	0
176	0	0	0	0	0	0	0	0
177	0	0	1	0	0	0	0	1
178	0	0	0	0	0	0	0	0
179	0	0	1	0	0	0	0	0
180	0	0	1	1	1	1	1	0
181	0	0	0	0	0	0	0	0

数据视图 变量视图

History of angina 具有值 -1。这个值对于数据文件中的治疗和结果变量来说是有效缺失值，它在此处无效的原因是患者病史值当前尚未定义用户缺失值。

定义自己的规则

从 patient_los.sav 复制的验证规则非常有用，但您需要定义更多规则才能完成这项工作。此外，有时到达时已死亡的患者会被错误地标记为在医院死亡。单变量验证规则无法捕获这种情况，因此您需要定义交叉变量规则处理这种情况。

- 单击?对话框调用?工具栏按钮并选择**验证数据**。
- 单击**单变量规则**选项卡。（您需要为 Hospital size、度量 Rankin 得分的变量以及对应于未重新编码的 Barthel 指数的变量定义规则。）
- 单击**定义规则**。

图片 7-19
“定义验证规则”对话框，“单变量规则”选项卡

显示了当前定义的规则，其中?规则?列表中选择了 0 to 1 Dichotomy，并且规则的属性显示在?规则定义?组中。

- 要定义规则，请单击**新建**。

图片 7-20
“定义验证规则”对话框，“单变量规则”选项卡（定义了“1 to 3 Categorical”）

- 键入 1 to 3 Categorical 作为规则名称。
- 对于?有效值?，选择**在列表中**。
- 键入 1、2 和 3 作为值。
- 取消选择**允许使用系统缺失值**。
- 要为 Rankin 得分定义规则，请单击**新建**。

图片 7-21
“定义验证规则”对话框，“单变量规则”选项卡（定义了“0 to 5 Categorical”）

- 键入 0 to 5 Categorical 作为规则名称。
- 对于?有效值?，选择**在列表中**。
- 键入 0、1、2、3、4 和 5 作为值。
- 取消选择**允许使用系统缺失值**。
- 要为 Barthel 指数定义规则，请单击**新建**。

图片 7-22
“定义验证规则”对话框，“单变量规则”选项卡（定义了“0 to 100 by 5”）

- 键入 0 to 100 by 5 作为规则名称。
- 对于?有效值?，选择**在列表中**。
- 键入 0、5、…，以及 100 作为值。
- 取消选择**允许使用系统缺失值**。
- 单击**继续**。

图片 7-23
“验证数据”对话框，“单变量规则”选项卡（定义了“0 to 100 by 5”）

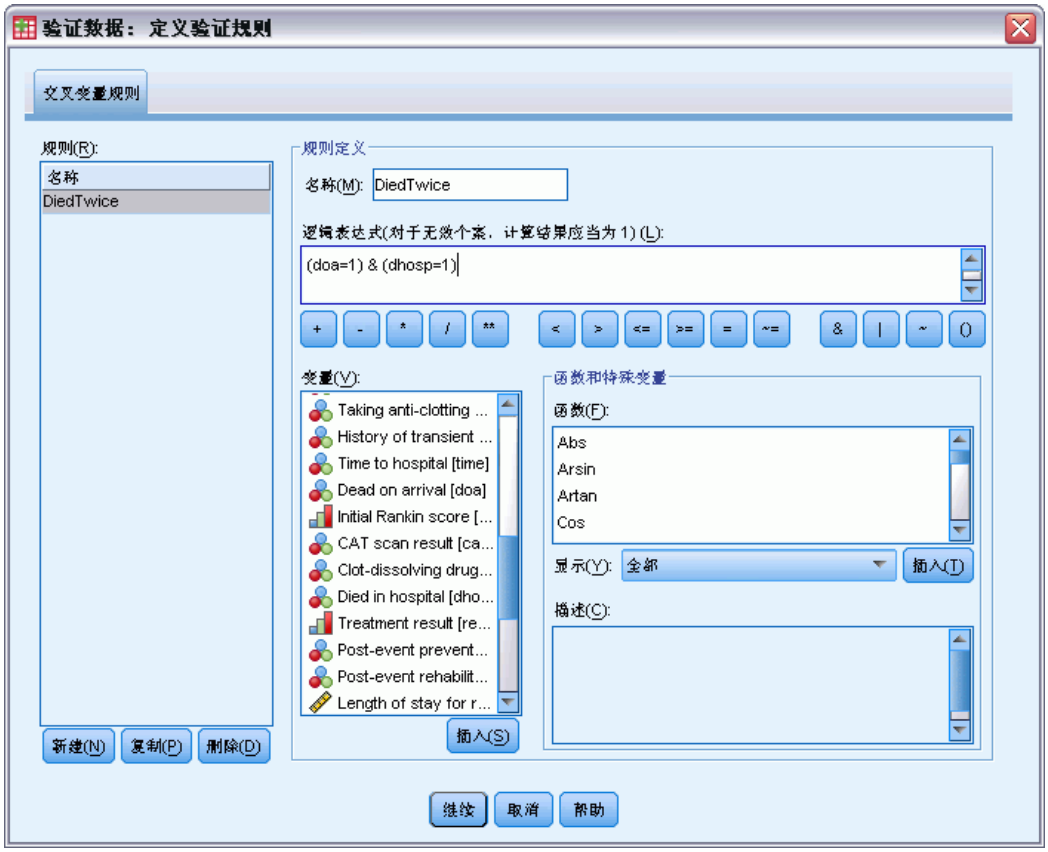
现在您需要将定义的规则应用于分析变量。

- 将 1 to 3 Categorical 应用于 Hospital size。
- 将 0 to 5 Categorical 应用于 Initial Rankin score 和 Rankin score at 1 month 到 Rankin score at 6 months。
- 将 0 to 100 by 5 应用于 Barthel index at 1 month 到 Barthel index at 6 months。
- 单击**交叉变量规则**选项卡。

当前未定义任何规则。

- 单击**定义规则**。

图片 7-24
“定义验证规则”对话框，“交叉变量规则”选项卡



如果不存在任何规则，则会自动创建一个新的占位符规则。

- 键入 DiedTwice 作为规则名称。
- 键入 (doa=1) & (dhosp=1) 作为逻辑表达式。如果患者既记录为达到时已死亡又记录为院内死亡，则将返回值 1。
- 单击继续。

  “交叉变量规则”选项卡中会自动选择新定义的规则。
- 单击确定。

交叉变量规则

图片 7-25
交叉变量规则

规则	违规数	规则表达式
DiedTwice	27	(doa = 1) & (dhosp = 1)

交叉变量规则摘要列出至少违反了一次的交叉变量规则、违规数，以及对所违反的每个规则的描述。

个案报告

图片 7-26
个案报告

	确认违反规则		标识符		
观测值	单变量[a]	交叉变量	Hospital ID	Patient ID	Attending physician ID
20		DiedTwice	PBW	1192970826	355184
49		DiedTwice	NHV	8717862852	237418
129		DiedTwice	QWS	6901932085	215041
138		DiedTwice	RLD	1205005069	695521
162		DiedTwice	OZN	5546809538	125304
175	0 to 1 Dichotomy (1)		OZN	0333204686	883285
274	0 to 1 Dichotomy (1)		OZN	1038840465	103254
310	Nonnegative integer (1)		OZN	2090290204	883285
414		DiedTwice	WPA	3351107142	462020
437	0 to 1 Dichotomy (1)		WPA	2349729006	723384
447		DiedTwice	WPA	7163481282	519548
458		DiedTwice	WPA	9159094175	652070
462		DiedTwice	WPA	2137520354	723384
537		DiedTwice	SLB	5246122506	928076
544		DiedTwice	SLB	1605957462	506108
620		DiedTwice	GFG	8141858966	828754
629		DiedTwice	GFG	3397891610	539412
630		DiedTwice	GFG	3397891610	539412
639		DiedTwice	GFG	3962622031	327422
644		DiedTwice	GFG	4271782383	749432
649		DiedTwice	GFG	0950686750	618069
653		DiedTwice	GFG	0663642766	001448
722		DiedTwice	CFC	0418125500	877354
748		DiedTwice	GFG	8744721380	539412
752	Nonnegative integer (1) 0 to 1 Dichotomy (3)		GFG	4993307441	828754
868		DiedTwice	WWL	9714672452	237547
881		DiedTwice	WWL	6613279456	574275
915		DiedTwice	EFX	2575793702	501318
933		DiedTwice	IZO	2807437472	680253
1010		DiedTwice	BLA	5284009939	657638
1028		DiedTwice	BLA	8021997463	185703
1054		DiedTwice	ALK	0950897644	267830
1173	1 to 4 Categorical (1)		ALK	8737661990	185787

a. 违反规则的变量数遵循每个规则。

个案报告现在包括违反了交叉变量规则的个案以及以前发现的违反了单变量规则的个案。所有这些个案都需要报告到数据输入以便进行更正。

摘要

分析人员具备了向数据输入经理作出初步报告所需的信息。

相关过程

“验证数据”过程是一个有用的数据质量控制工具。

- 标识异常个案过程分析数据中的模式并标识具有一些类型各异的显著值的个案。

章 8

自动数据准备

准备分析数据是任何项目中最重要的步骤之一，而从传统来说也是最耗时的步骤之一。“自动数据准备 (ADP)”为您处理任务，分析您的数据并识别修正，筛选出存在问题或可能无用的字段，并在适当的情况下派生新的属性，通过智能筛选技术改进性能。您可以通过完全**自动**的方式使用算法，这种方式可以允许选择并应用修正；或者也可以通过**交互式**方式使用算法，这种方式可以在做出更改前对其进行预览，并按照需要进行接受或拒绝。

通过使用 ADP，您可以快速、轻松地准备数据以供建模，无需具备相关统计概念的预备知识。您可以更快速地构建模型并进行评分。此外，使用 ADP 还能提高自动化建模过程。

交互式使用自动数据准备

在调查业主保险理赔方面拥有有限资源的保险公司希望构建一个模型来标记具有潜在欺骗性的可疑理赔。他们在 insurance_claims.sav 中收集了之前理赔的信息样本。有关详细信息，请参阅第 129 页码附录 A 中的样本文件。 构建模型前，他们将使用自动数据准备来准备数据进行建模。由于他们希望能够在应用转换前查看建议的转换，他们将在交互模式下使用自动数据准备。

在目标之间选择

- 要交互式运行“自动数据准备”，请从菜单中选择：
  转换 > **准备建模数据** > **交互式...**

图片 8-1
“目标”选项卡

第一个选项卡显示控制缺省设置的目标，但目标之间的实际差别是什么？通过使用每个目标运行过程，我们可以看到结果的差别。

- 确保选择了**均衡速度&精确度**并单击**分析**。

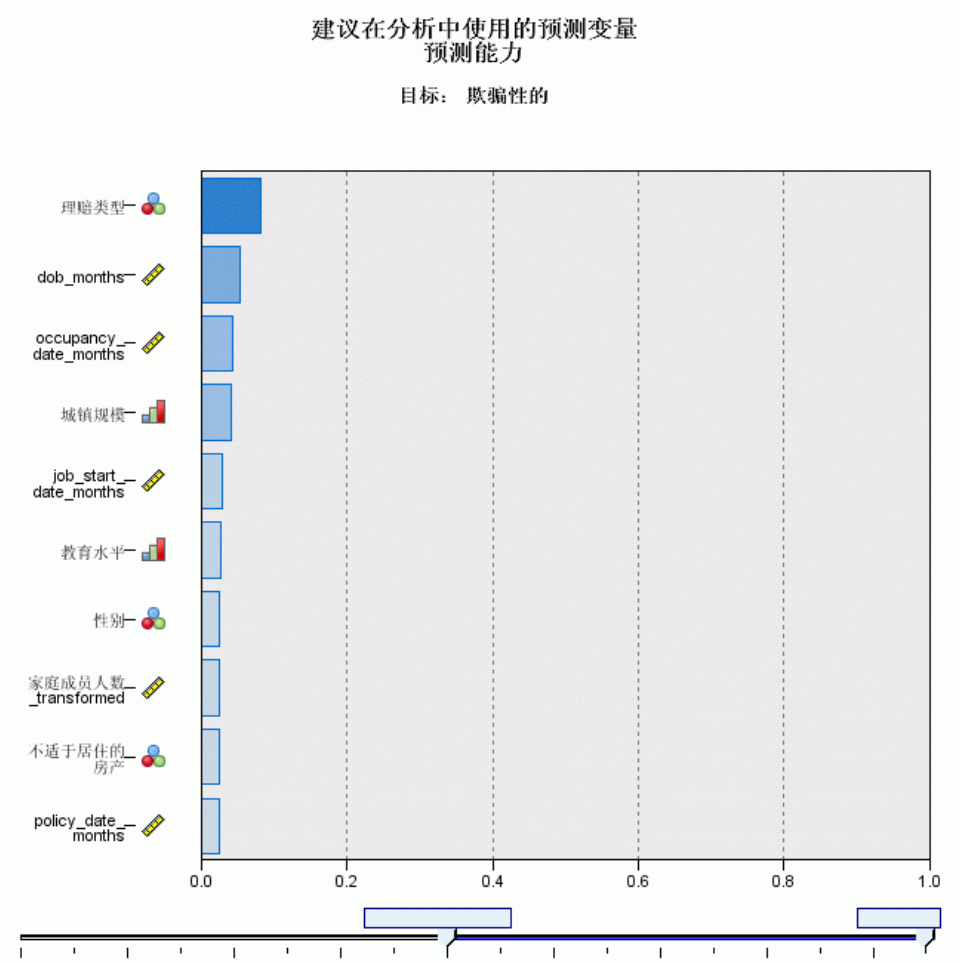
图片 8-2
“分析”选项卡，均衡目标的字段处理摘要

字段处理汇总

字段		N
目标		1
预测变量		18
建议在分析中使用的预测变量	总计	18
	原始字段（未变换）	8
	原始字段变换	5
	导出自日期和时间	5
	已构建	0
未使用预测变量		0

在过程处理数据时，关注自动切换到“分析”选项卡。缺省主视图是“字段处理摘要”，其中概述了自动数据准备如何处理字段。将有一个目标，18 个输入，同时建议使用 18 个字段来建模。对于建议用于建模的字段，9 个为原始输入字段，4 个为原始输入字段的转换，5 个派生自日期和时间字段。

图片 8-3
“分析”选项卡，均衡目标的预测能力



缺省辅助视图是“预测能力”，可使您快速了解哪个建议字段对于建模最有用。注意，当建议 18 个预测变量用于分析时，缺省情况下只有前 10 个预测变量显示在预测能力图表中。要显示更多或更少字段，使用图表下面的幻灯片控件。

将**均衡速度和精确度**作为目标时，理赔类型被标识为“最佳”预测变量，之后是家庭成员人数和索赔人的当前年龄（以月为单位计算的从出生之日起到当前日期的持续时间）。

- 单击**清除分析**，然后单击“目标”选项卡。
- 选择**优化速度**并单击**分析**。

图片 8-4
“分析”选项卡，优化速度时的字段处理摘要

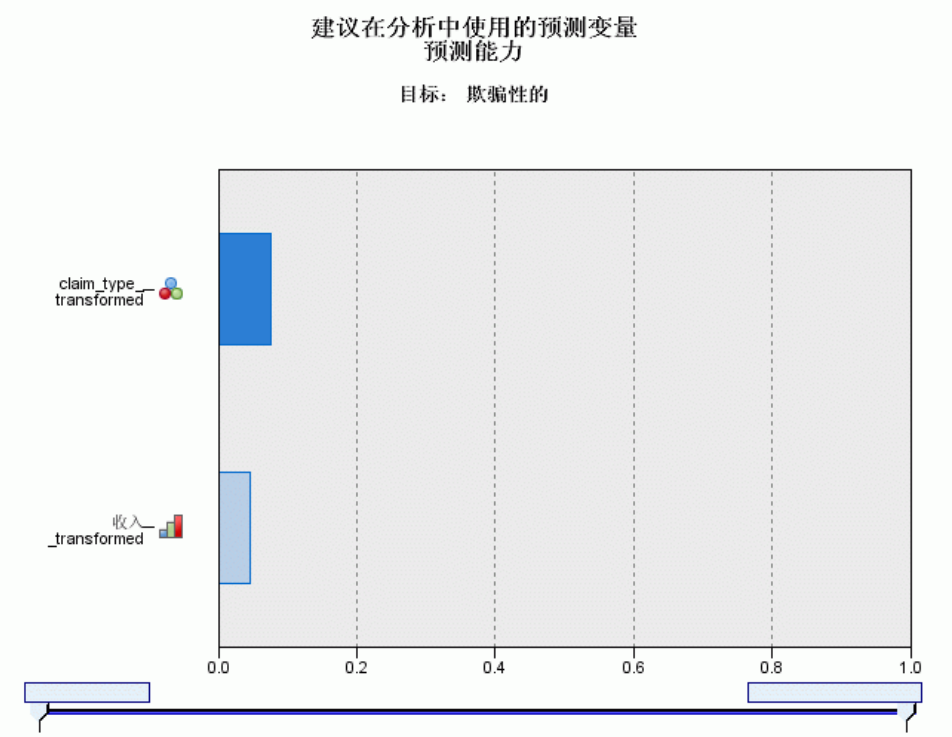
字段处理汇总

字段		N
目标		1
预测变量		18
建议在分析中使用的预测变量	总计	2
	原始字段（未变换）	0
	原始字段变换	2
	导出自日期和时间	0
	已构建	0
未使用预测变量		16

- 无法构建有用的预测变量。最常见的原因是：太少的连续预测变量与目标高度关联，或所有连续预测变量都不相关。

在过程处理数据时，关注再次自动切换到“分析”选项卡。在此个案中，只建议使用 2 个字段来建模，且这两个字段必须都是原始字段的转换。

图片 8-5
“分析”选项卡，优化速度时的预测能力



将**优化速度**作为目标时，claim_type_transformed 被标识为“最佳”预测变量，之后是 income_transformed。

- 单击**清除分析**，然后单击“目标”选项卡。
- 选择**优化精确度**并单击**分析**。

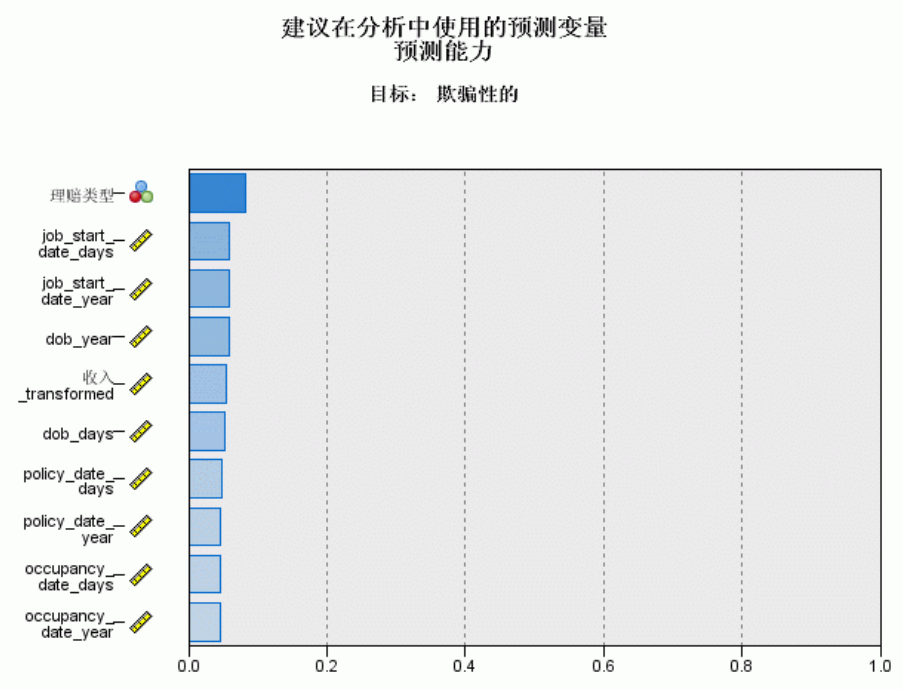
图片 8-6
“分析”选项卡，优化精确度时的预测能力

字段处理汇总

字段		N
目标		1
预测变量		18
建议在分析中使用的预测变量	总计	32
	原始字段（未变换）	8
	原始字段变换	5
	导出自日期和时间	19
	已构建	0
未使用预测变量		0

将**优化精确度**作为目标时，建议使用 32 个字段来建模，因为通过从日期提取日、月和年，从时间提取小时、分钟和秒，可以从日期和时间派生出更多字段。

图片 8-7
“分析”选项卡，优化精确度时的预测能力



理赔类型被标识为“最佳”预测变量，之后是索赔人开始其最近工作的天数（从工作开始日期到当前日期的持续时间）以及索赔人开始其当前工作的年份（从工作开始日期中提取）。

汇总：

- **均衡速度&精确度**将从日期创建可用于建模的字段，同时可转换类似 reside 的连续字段以使它们更加正态地分布。
- **优化精确度**将从日期创建一些额外字段（它还将检查离群值，如果目标是连续的，可将其转换以使其更加正态地分布）。
- **优化速度**不会准备日期，也不会重新调整连续字段，但会在目标为分类时合并分类预测变量的类别并离散化连续预测变量（同时在目标为连续时执行特征选择和构建）。

保险公司决定进一步探索**优化精确度**结果。

▶ 从主视图下拉列表选择**字段**。

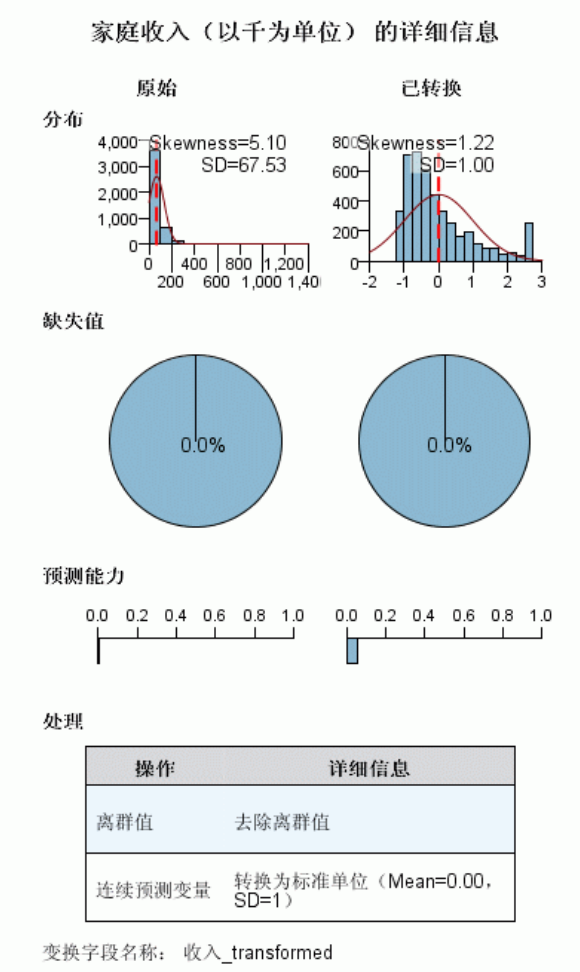
字段和字段详细信息

图片 8-8
字段

“字段”视图显示处理过的字段，以及 ADP 是否建议将它们用于建模。单击任何字段名称可以在链接视图中显示有关该字段的更多信息。

- 单击**收入**。

图片 8-9
“家庭收入（千元）”的字段详细信息

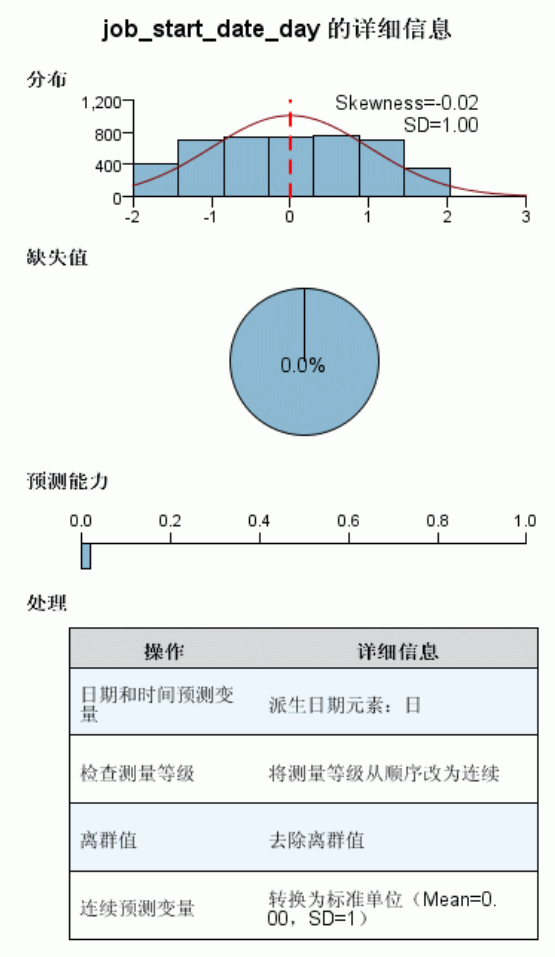


“字段详细信息”视图显示原始和已转换家庭收入（千元）的分布。根据处理表，标识为离群值的记录被删除（通过将它们的值设置为等于分界值以确定离群值），同时该字段被标准化为具有均值 0 和标准差 1。在已转换字段直方图右侧较远处的“增加”显示可能有超过 200 条记录被标识为离群值。收入有很大的偏斜分布，因此这可能是缺省分界值在确定离群值时太有侵略性的个案。

还请注意已转换字段相对原始字段在预测能力上的增强。这似乎是有用的转换。

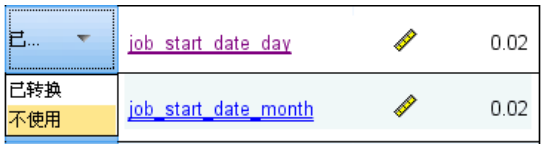
- 在“字段”视图中，单击 job_start_date_day。（注意，这与 job_start_date_days 不同。）

图片 8-10
job_start_date_day 的字段详细信息



字段 job_start_date_day 是从 雇佣开始日期 [job_start_date] 提取的日。该字段对判断理赔是否具有欺骗性几乎没有任何帮助，因此保险公司希望从建模注意事项中将其删除。

图片 8-11
“家庭收入（千元）”的字段详细信息



- 在“字段”视图中，从 job_start_date_day 行的“使用的版本”下拉列表选择**不使用**。对具有 _day 和 _month 后缀的所有字段执行相同的操作。
- 要应用转换，单击**运行**。

数据集现在已经可以用于建模，因为所有建议的预测变量（包括新预测变量和旧预测变量）的角色均已设置为“输入”，而未建议的预测变量的角色都设置为“无”。要创建只包括建议的预测变量的数据集，使用对话框中的“应用转换”设置。

自动使用自动数据准备

某汽车集团希望跟踪各类私人汽车的销售情况。为了能够标识表现良好和表现不好的型号，您希望建立汽车销售和汽车特性之间的关系。这些信息收集在 car_sales_unprepared.sav 中。有关详细信息，请参阅第 129 页码附录 A 中的样本文件。 使用自动数据准备准备数据进行分析。同时使用准备“之前”和“之后”的数据构建模型，以便可以比较结果。

准备数据

- 要在自动模式中运行自动数据准备，请从菜单中选择：
  转换 > **准备建模数据** > **自动...**

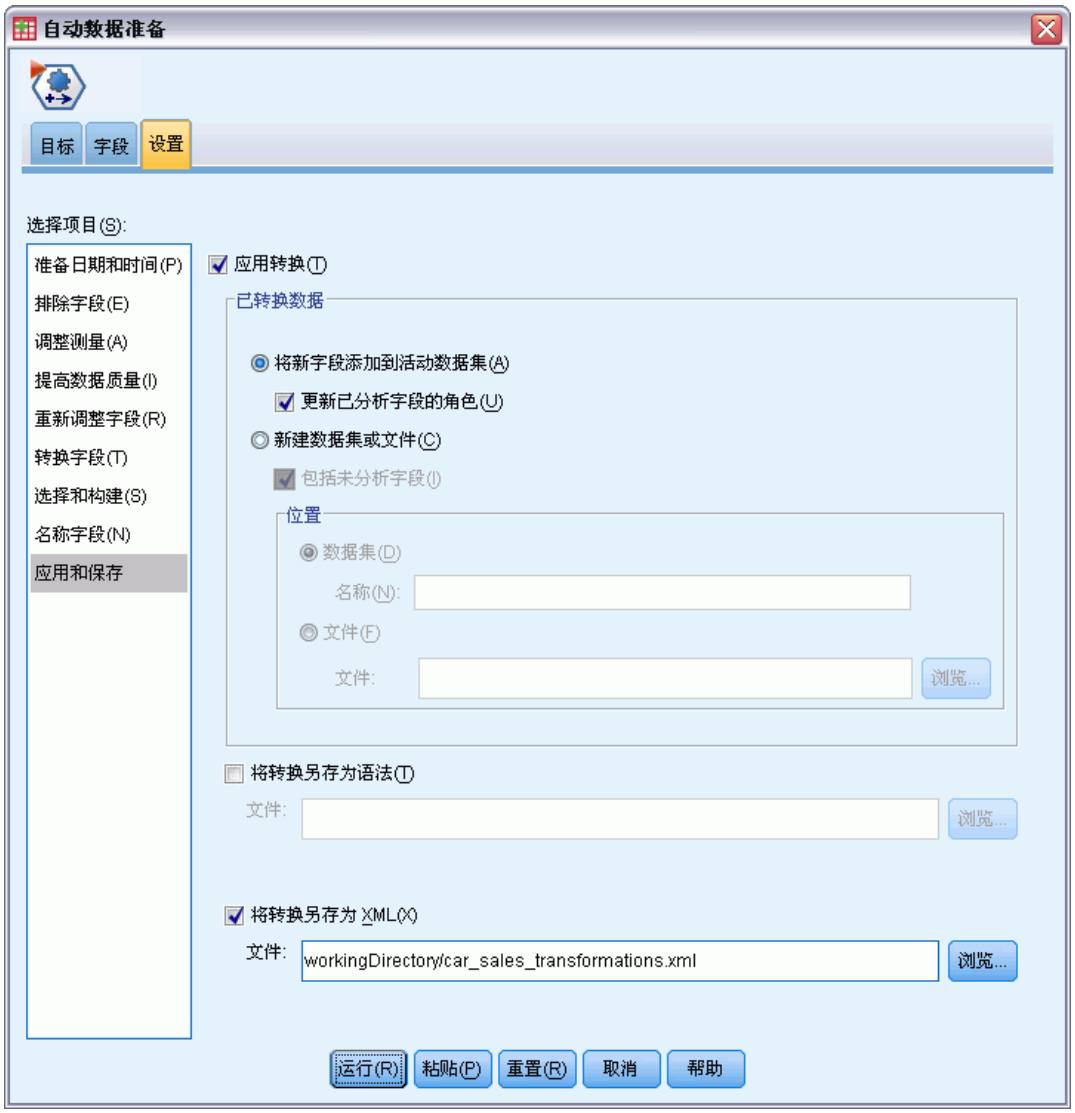
图片 8-12
“目标”选项卡

▶ 选择**优化精确度**。

由于目标字段销售（千元）是连续的，且可以在自动数据准备过程中转换，您希望将该转换保存到一个 XML 文件，以便能够使用“逆转换得分”对话框将已转换目标的预测值转换回原始尺度。

▶ 单击**设置**选项卡，然后单击**应用和保存**设置。

图片 8-13
“应用和保存”设置



- 选择保存转换为 XML 并单击浏览导航到 workingDirectory/car_sales_transformations.xml，将您希望保存文件的路径替换为 workingDirectory。
- 单击运行。

这些选择将生成以下命令语法：

```
*Automatic Data Preparation.
ADP
  /FIELDS TARGET=sales INPUT=resale type price engine_s horsepow wheelbas width length
    curb_wgt fuel_cap mpg
  /PREPDATETIME DATEDURATION=YES(REFERENCE=YMD('2009-06-04') UNIT=AUTO)
    TIMEDURATION=YES(REFERENCE=HMS('08:43:35') UNIT=AUTO) EXTRACTYEAR=YES(SUFFIX='_year')
    EXTRACTMONTH=YES(SUFFIX='_month') EXTRACTDAY=YES(SUFFIX='_day')
    EXTRACTHOUR=YES(SUFFIX='_hour') EXTRACTMINUTE=YES(SUFFIX='_minute')
```

```
    EXTRACTSECOND=YES(SUFFIX='_second')
   /SCREENING PCTMISSING=YES(MAXPCT=50) UNIQUECAT=YES(MAXCAT=100) SINGLECAT=NO
  /ADJUSTLEVEL INPUT=YES TARGET=YES MAXVALORDINAL=10 MINVALCONTINUOUS=5
  /OUTLIERHANDLING INPUT=YES TARGET=NO CUTOFF=SD(3) REPLACEWITH=CUTOFFVALUE
  /REPLACEMISSING INPUT=YES TARGET=NO
  /REORDERNOMINAL INPUT=YES TARGET=NO
  /RESCALE INPUT=ZSCORE(MEAN=0 SD=1) TARGET=BOXCOX(MEAN=0 SD=1)
  /TRANSFORM MERGESUPERVISED=NO MERGEUNSUPERVISED=NO BINNING=NONE SELECTION=NO
    CONSTRUCTION=NO
  /CRITERIA SUFFIX(TARGET='_transformed' INPUT='_transformed')
  /OUTFILE PREPXML='/workingDirectory/car_sales_transformations.xml'.
TMS IMPORT
  /INFILE TRANSFORMATIONS='/workingDirectory/car_sales_transformations.xml'
    MODE=FORWARD (ROLES=UPDATE)
  /SAVE TRANSFORMED=YES.
EXECUTE.
```

- ADP 命令准备目标字段销售和输入字段转售至mpg。
- 指定了 PREPDATETIME 子命令但不使用它，因为没有字段是日期或时间字段。
- ADJUSTLEVEL 子命令将超过 10 个值的有序字段重新设计为连续字段且将少于 5 个值的连续字段重新设计为有序字段。
- OUTLIERHANDLING 子命令用均值的 3 个标准差的值替换超出均值的 3 个标准差的连续输入（非目标）的值。
- REPLACEMISSING 子命令替换缺失的输入值（非目标）。
- REORDERNOMINAL 子命令按发生频率的升序重新编码名义输入的值。
- RESCALE 子命令使用 Z 得分转换标准化连续输入，使其具有均值 0 和标准差 1，并使用 Box-Cox 转换标准化连续目标，使其具有均值 0 和标准差 1。
- TRANSFORM 子命令关闭此子命令指定的所有默认操作。
- CRITERIA 子命令指定用于转换目标和输入的默认后缀。
- OUTFILE 子命令指定转换应保存到 /workingDirectory/car_sales_transformations.xml，其中 /workingDirectory 是您要保存 car_sales_transformations.xml 的路径。
- TMS IMPORT 命令读取 car_sales_transformations.xml 中的转换，并将其应用到活动数据集，更新已转换的现有字段的角色。
- EXECUTE 命令导致转换被处理。在将此用作语法较长流的一部分时，您可去掉 EXECUTE 命令以节省部分处理时间。

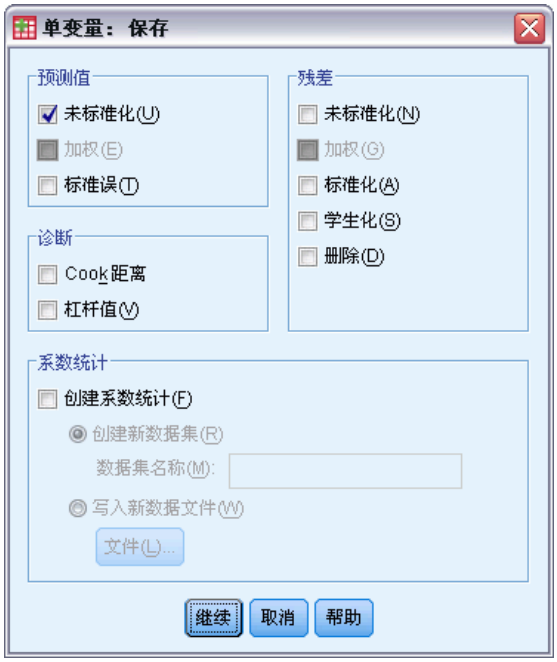
在“未准备数据”上构建模型

▶ 要在未准备数据上构建数据，请从菜单中选择：
分析 > 一般线性模型 > 单变量...

图片 8-14
“GLM 单变量”对话框

- 选择 Sales in thousands [sales] 作为因变量。
- 选择 Vehicle type [type] 作为固定因子。
- 选择 4-year resale value [resale] 到 Fuel efficiency [mpg] 作为协变量。
- 单击保存。

图片 8-15
“保存”对话框



- 在“预测值”组中选择**未标准化**。
- 单击**继续**。
- 在“GLM 单变量”对话框中单击**确定**。

这些选择将生成以下命令语法：

```
UNIANOVA sales BY type WITH resale price engine_s horsepow wheelbas width length
    curb_wgt fuel_cap mpg
  /METHOD=SSTYPE(3)
  /INTERCEPT=INCLUDE
  /SAVE=PRED
  /CRITERIA=ALPHA(0.05)
  /DESIGN=resale price engine_s horsepow wheelbas width length curb_wgt fuel_cap
    mpg type.
```

图片 8-16
基于未准备数据的模型的主体间效应

主体间效应的检验

因变量:Sales in thousands

源	Ⅲ型平方和	df	均方	F	Sig.
校正模型	226123.658[a]	11	20556.696	5.050	.000
截距	12227.688	1	12227.688	3.004	.086
resale	50.702	1	50.702	.012	.911
price	471.630	1	471.630	.116	.734
engine_s	19872.712	1	19872.712	4.882	.029
horsepow	9644.486	1	9644.486	2.369	.127
wheelbas	29824.272	1	29824.272	7.327	.008
width	263.465	1	263.465	.065	.800
length	1374.525	1	1374.525	.338	.562
curb_wgt	32762.692	1	32762.692	8.049	.005
fuel_cap	1124.237	1	1124.237	.276	.600
mpg	337.585	1	337.585	.083	.774
type	17668.779	1	17668.779	4.341	.040
误差	427402.183	105	4070.497		
总计	1062354.955	117			
校正的总计	653525.841	116			

a. R 方 = .346（调整 R 方 = .277）

缺省“ GLM 单变量”输出包括主体间效应，它是方差表的分析。检验模型中的每一项以及模型整体解释因变量中变异的能力。注意，在本表中不显示变量标签。

预测变量显示不同的显著性水平，那些显著性值低于 0.05 的通常被认为对模型有用。

在“已准备数据”上构建模型

图片 8-17
“GLM 单变量”对话框

- 要在已准备数据上构建模型，请调用“GLM 单变量”对话框。
- 取消选择 Sales in thousands [sales] 并选择 sales_transformed 作为因变量。
- 取消选择 4-year resale value [resale] 到 Fuel efficiency [mpg] 并选择 resale_transformed 到 mpg_transformed 作为协变量。
- 单击**确定**。

这些选择将生成以下命令语法：

```
UNIANOVA sales_transformed BY type WITH resale_transformed price_transformed
    engine_s_transformed horsepow_transformed wheelbas_transformed width_transformed
    length_transformed curb_wgt_transformed fuel_cap_transformed mpg_transformed
  /METHOD=SSTYPE(3)
  /INTERCEPT=INCLUDE
  /SAVE=PRED
  /CRITERIA=ALPHA(0.05)
  /DESIGN=resale_transformed price_transformed engine_s_transformed horsepow_transformed
    wheelbas_transformed width_transformed length_transformed curb_wgt_transformed
    fuel_cap_transformed mpg_transformed type.
```

图片 8-18
基于已准备数据的模型的主体间效应

因变量:sales_transformed

源	Ⅲ型平方和	df	均方	F	Sig.
校正模型	78.965[a]	11	7.179	13.512	.000
截距	2.549	1	2.549	4.797	.030
resale	.852	1	.852	1.603	.207
price	8.540	1	8.540	16.075	.000
engine_s	2.943	1	2.943	5.540	.020
horsepow	.054	1	.054	.102	.749
wheelbas	1.148	1	1.148	2.161	.144
width	.026	1	.026	.049	.826
length	.407	1	.407	.766	.383
curb_wgt	.027	1	.027	.051	.822
fuel_cap	.089	1	.089	.168	.682
mpg	3.226	1	3.226	6.073	.015
type	4.268	1	4.268	8.033	.005
误差	77.035	1	.531		
总计	77.035	145			
校正的总计	156.000	157			
	156.000	156			

a. R 方 = .506（调整 R 方 = .469）

构建在未准备数据上的模型和已准备数据上的模型的主体间效应有几个有趣的区别值得注意。首先，注意总自由度增加了；这是因为缺失值在自动数据准备期间被插补值替换，因此按列表从第一个模型中删除的记录对第二个模型可用。更值得注意的是，某些预测变量的显著性可能已改变。虽然这两个模型都同意引擎大小 [engine_s] 和汽车类型 [type] 对于模型有用，轴距 [wheelbas] 和空车重量 [curb_wgt] 不再重要，同时汽车价格 [price_transformed] 和燃料效率 [mpg_transformed] 现在很重要。

为什么会出现这种变化？销售有偏斜分布，因此一旦销售被转换，轴距和空车重量的一些原本受影响的记录变得不再受影响。另一个可能性是，由于缺失值替换更改了这些变量的统计显著性，因此有额外个案可用。在任何个案中，这都需要进一步的调查，我们将不在此继续。

注意，R 方对于构建在已准备数据上的模型更高，但由于销售已经转换，这可能不是比较每个模型性能的最佳度量。相反，您可以计算观察值和两组预测值之间的非参数相关性。

比较预测值

▶ 要获得两个模型预测值的相关性，请从菜单中选择：
分析 > **相关** > **双变量...**

图片 8-19
“双变量相关性”对话框

- 选择 Sales in thousands [sales]、Predicted Value for sales [PRE_1] 和 Predicted Values for sales_transformed [PRE_2] 作为分析变量。
- 取消选择 Pearson 并选择“相关系数”组中的 Kendall 的 tau-b 和 Spearman。

  注意，Predicted Values for sales_transformed [PRE_2] 可以用于计算非参数相关性而无需逆转换到原始尺度，因为逆转换不更改预测值的秩次。
- 单击确定。

  这些选择将生成以下命令语法：

```
NONPAR CORR
  /VARIABLES=sales PRE_1 PRE_2
  /PRINT=BOTH TWOTAIL NOSIG
  /MISSING=PAIRWISE.
```

图片 8-20
非参数相关性

			Sales in thousands	Predicted Value for sales	Predicted Value for sales_ transformed
Kendall's tau_b	Sales in thousands	相关系数	1.000	.376**	.480**
		Sig.（双侧）	.	.000	.000
		N	157	117	157
	Predicted Value for sales	相关系数	.376**	1.000	.659**
		Sig.（双侧）	.000	.	.000
		N	117	117	117
	Predicted Value for sales_transformed	相关系数	.480**	.659**	1.000
		Sig.（双侧）	.000	.000	.
		N	157	117	157
Spearman's rho	Sales in thousands	相关系数	1.000	.530**	.664**
		Sig.（双侧）	.	.000	.000
		N	157	117	157
	Predicted Value for sales	相关系数	.530**	1.000	.835**
		Sig.（双侧）	.000	.	.000
		N	117	117	117
	Predicted Value for sales_transformed	相关系数	.664**	.835**	1.000
		Sig.（双侧）	.000	.000	.
		N	157	117	157

**. 在置信度（双侧）为 0.01 时，相关性是显著的。

第一列显示使用已准备数据构建模型的预测值与使用 Kendall 的 tau-b 和 Spearman 的 rho 方法构建模型的观测值的关联更强。这表明运行自动数据准备改进了模型。

逆转换“预测值”

- 已准备数据包括销售的转换，因此此模型的预测值将不能直接作为得分使用。要将预测值转换回原始尺度，请从菜单中选择：

  转换 > **准备建模数据** > **逆转换得分**...

图片 8-21
“逆转换得分”对话框

- 选择 Predicted Value for sales_transformed [PRE_2] 作为要逆转换的字段。
- 键入 _backtransformed 作为新字段的后缀。
- 键入 workingDirectory\car_sales_transformations.xml，将文件路径替换为 workingDirectory，以使 XML 文件的位置包含转换。
- 单击**确定**。

这些选择将生成以下命令语法：

```
TMS IMPORT
  /INFILE TRANSFORMATIONS='workingDirectory/car_sales_transformations.xml'
    MODE=BACK (PREDICTED=PRE_2 SUFFIX='_backtransformed').
EXECUTE.
```

- TMS IMPORT 命令读取 car_sales_transformations.xml 中的转换并将向后转换应用到 PRE_2。
- 含有向后转换值的新字段名为 PRE_2_backtransformed。
- EXECUTE 命令导致转换被处理。在将此用作语法较长流的一部分时，您可去掉 EXECUTE 命令以节省部分处理时间。

摘要

使用自动数据准备，您可以快速获得可以改进模型的数据转换。如果目标已转换，您可以将转换保存到 XML 文件并使用“逆转换得分”对话框将已转换目标的预测值转换回原始尺度。

章 9

标识异常个案

“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。

标识异常个案算法

此算法分为三个阶段：

建模。该过程创建聚类模型，用来说明数据集中的自然分组（即聚类），如果不说明，这些分组是不明显的。聚类基于一组输入变量。用于计算聚类组标准值的结果聚类模型和足够的统计量会存储起来，供以后使用。

评分。模型应用于每个个案来标识其聚类组，并根据其聚类组为每个个案创建一些指标来度量个案的异常性。全部个案是按异常指标的值进行排序的。个案列表的前面部分标识为异常集。

原因。对于每个异常个案，变量都按其对应的变量偏差指标进行排序。显示排在前面的变量、变量值和对应的标准值，作为个案标识为异常的原因。

标识医疗数据库中的异常个案

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。

该信息收集在 stroke_valid.sav 中。有关详细信息，请参阅第 129 页码附录 A 中的样本文件。使用“标识异常个案”过程可使数据文件变得干净。在 detectanomaly_stroke.sps 中可以找到再次生成这些分析的语法。

运行分析

- 要标识异常个案，请从菜单中选择：
  数据 > **标识异常个案...**

图片 9-1
“标识异常个案”对话框，“变量”选项卡

- 选择 Age category 到 Stroke between 3 and 6 months 作为分析变量。
- 选择 Patient ID 作为个案标识变量。
- 单击输出选项卡。

图片 9-2
“标识异常个案”对话框，“输出”选项卡

- 选择对等组标准值、异常指标、按分析变量列出出现的原因和已处理的个案数。
- 单击保存选项卡。

图片 9-3
“标识异常个案”对话框，“保存”选项卡

- 选择**异常指标**、**对等组**和**原因**。

  保存这些结果使您可以生成汇总结果的有用的散点图。

- 单击**缺失值**选项卡。

图片 9-4
“标识异常个案”对话框，“缺失值”选项卡

- 选择**在分析中包括缺失值**。此过程是必需的，因为要处理在治疗前或治疗中死亡的病人存在许多用户缺失值。一个度量每个个案的缺失值比例的额外变量会作为刻度变量添加到分析中。
- 单击**选项**选项卡。

图片 9-5
“标识异常个案”对话框，“选项”选项卡

- 键入 2 作为认为异常的个案的百分比。
- 取消选择**仅标识异常指标值符合或超过最小值的个案**。
- 键入 3 作为最大的原因数量。
- 单击**确定**。

个案处理摘要 (0)

图片 9-6
个案处理摘要 (S)

		N	组合百分比	总计百分比
对等 ID	1	710	67.7%	67.7%
	2	90	8.6%	8.6%
	3	248	23.7%	23.7%
已组合		1048	100.0%	100.0%
合计		1048		100.0%

每个个案都分类到类似个案的对等组中。个案处理摘要显示创建的对等组的数量，还显示每个对等组中的个案的数量和百分比。

异常个案指标列表

图片 9-7
异常个案指标列表

观测值	patid	异常索引
843	7840326167	2.837
510	0714726620	2.022
623	6553808330	2.014
501	6461046805	2.002
607	1077125669	1.897
884	2260043998	1.889
614	4030164769	1.869
241	1038840465	1.865
13	2191527525	1.826
172	4458028382	1.786
705	1336411777	1.778
651	4103977868	1.767
384	2247641363	1.767
839	0437454972	1.766
861	9746101913	1.757
19	7237535360	1.756
806	4391632997	1.756
871	6961938294	1.739
239	7315965190	1.738
887	6044244232	1.737
245	0816869249	1.736

异常指标用于度量个案相对于其 peer 组的异常程度。将显示异常指标值最高的 2% 的个案及其个案号和 ID。共列出 21 个个案，其值介于 1.736 到 2.837 之间。列表中第一个和第二个个案的异常指标值之间存在相对较大的差异，这表示个案 843 可能是异常个案。其他个案需要逐一进行判断。

异常个案 Peer ID 列表

图片 9-8
异常个案 Peer ID 列表

观测值	patid	对等 ID	对等大小	对等大小百分比
843	7840326167	3	248	23.7%
510	0714726620	3	248	23.7%
623	6553808330	3	248	23.7%
501	6461046805	3	248	23.7%
607	1077125669	3	248	23.7%
884	2260043998	3	248	23.7%
614	4030164769	3	248	23.7%
241	1038840465	3	248	23.7%
13	2191527525	3	248	23.7%
172	4458028382	3	248	23.7%
705	1336411777	1	710	67.7%
651	4103977868	1	710	67.7%
384	2247641363	3	248	23.7%
839	0437454972	3	248	23.7%
861	9746101913	3	248	23.7%
19	7237535360	1	710	67.7%
806	4391632997	1	710	67.7%
871	6961938294	1	710	67.7%
239	7315965190	3	248	23.7%
887	6044244232	1	710	67.7%
245	0816869249	3	248	23.7%

显示潜在异常个案及其对等组成员资格信息。前 10 个个案（共 15 个个案）属于对等组 3，其余个案属于对等组 1。

异常个案原因列表

图片 9-9
异常个案原因列表

原因: 1

观测值	patid	原因变量	变量影响	变量值	变量范数
843	7840326167	cost	.411	200.51	19.8273
510	0714726620	cost	.120	96.59	19.8273
623	6553808330	cost	.175	114.01	19.8273
501	6461046805	barthel1	.084	80	(缺失值)
607	1077125669	cost	.126	96.11	19.8273
884	2260043998	cost	.138	99.73	19.8273
614	4030164769	barthel1	.085	45	(缺失值)
241	1038840465	barthel1	.115	25	(缺失值)
13	2191527525	barthel1	.118	40	(缺失值)
172	4458028382	barthel1	.120	100	(缺失值)
705	1336411777	cost	.244	198.25	42.4673
651	4103977868	barthel1	.064	30	95
384	2247641363	barthel1	.122	20	(缺失值)
839	0437454972	barthel1	.109	95	(缺失值)
861	9746101913	barthel1	.102	70	(缺失值)
19	7237535360	barthel3	.080	5	100
806	4391632997	barthel2	.088	10	100
871	6961938294	barthel1	.094	5	95
239	7315965190	barthel1	.092	45	(缺失值)
887	6044244232	barthel1	.066	40	95
245	0816869249	barthel1	.124	5	(缺失值)

原因变量是对将个案分类为异常个案贡献最多的变量。显示每个异常个案的主要原因变量，以及其影响、个案的值和对等组标准值。分类变量的对等组标准值(Missing Value)指示对等组中的大多数个案对于该变量都具有缺失值。

变量影响统计是原因变量对个案与其对等组偏差的成比例贡献。如果分析中有 38 个变量（包括缺失的比例变量），则变量的期望影响为 1/38 = 0.026。变量 cost 对个案 843 的影响是 0.411，这是一个相对较大的值。个案 843 的 cost 的值为 200.51，而对等组 3 中的个案的平均值为 19.83。

对话框选择请求顶部的三个原因的结果。

- 要查看其他原因的结果，请通过双击激活该表。
- 将原因从层维度移动到行维度。

图片 9-10
异常个案原因列表（前 8 个个案）

观测值	原因	patid	原因变量	变量影响	变量值	变量范数
843	1	7840326167	cost	.411	200.51	19.8273
	2	7840326167	barthel1	.076	65	(缺失值)
	3	7840326167	rankin1	.044	2	(缺失值)
510	1	0714726620	cost	.120	96.59	19.8273
	2	0714726620	barthel1	.083	80	(缺失值)
	3	0714726620	rehab	.068	3	(缺失值)
623	1	6553808330	cost	.175	114.01	19.8273
	2	6553808330	surgery	.089	2	(缺失值)
	3	6553808330	barthel1	.089	70	(缺失值)
501	1	6461046805	barthel1	.084	80	(缺失值)
	2	6461046805	rehab	.068	3	(缺失值)
	3	6461046805	rankin1	.063	1	(缺失值)
607	1	1077125669	cost	.126	96.11	19.8273
	2	1077125669	barthel1	.094	85	(缺失值)
	3	1077125669	rehab	.072	3	(缺失值)
884	1	2260043998	cost	.138	99.73	19.8273
	2	2260043998	barthel1	.114	65	(缺失值)
	3	2260043998	rehab	.072	3	(缺失值)
614	1	4030164769	barthel1	.085	45	(缺失值)
	2	4030164769	rankin1	.085	3	(缺失值)
	3	4030164769	recbart1	.062	2	(缺失值)

此配置使您可以方便地比较每个个案的前三个原因的相对贡献。就像怀疑的一样，个案 843 被视为异常个案，因为它具有异常大的 cost 值。而对于个案 501，没有任何单个原因对异常情况的贡献是超过 0.10 的。

刻度变量标准值

图片 9-11
刻度变量标准值

		对等 ID			
		1	2	3	已組合
Length of stay for rehabilitation	均值	16.55	16.39	15.91	16.39
	标准偏差	12.596	.000	6.834	10.887
Total treatment and rehabilitation costs in thousands	均值	42.4673	3.5089	19.8273	33.7641
	标准偏差	26.45401	.50997	20.17309	27.31266
缺失比例	均值	.006	.541	.354	.134
	标准偏差	.021	2.9E-016	.083	.197

刻度变量标准值报告每个对等组中的每个变量以及全体变量的均值和标准差。比较这些值可以向您指出哪些变量对对等组的构成做出了贡献。

例如，Length of stay for rehabilitation 的均值在所有三个对等组中都是常数，这意味着此变量不会对对等组的构成做出贡献。而 Total treatment and rehabilitation costs in thousands 和 Missing Proportion 均提供了有关对等组成员资格的信息。对等组 1 具有最高的平均费用和最少的缺失值。对等组 2 具有很低的费用以及大量缺失值。对等组 3 具有中等水平的费用和缺失值。

此组织建议对等组 2 中包含到达时已死亡的病人，因此费用很少，并且所有治疗和康复变量均缺失。对等组 3 中看似包含许多在治疗过程中死亡的病人，因此存在治疗费用但不存在康复费用，因此康复变量缺失。对等组 1 看似包含通过治疗和康复活下来的几乎所有病人，因此费用最高。

分类变量标准值

图片 9-12
分类变量标准值（前 10 个变量）

		对等 ID			
		1	2	3	已組合
Age category	最受欢迎的类别	2	3	2	2
	频率	277	25	81	383
	百分比	39.0%	27.8%	32.7%	36.5%
Gender	最受欢迎的类别	0	0	1	0
	频率	361	46	126	529
	百分比	50.8%	51.1%	50.8%	50.5%
Physically active	最受欢迎的类别	1	0	0	0
	频率	373	55	139	531
	百分比	52.5%	61.1%	56.0%	50.7%
Obesity	最受欢迎的类别	0	0	0	0
	频率	555	67	178	800
	百分比	78.2%	74.4%	71.8%	76.3%
History of diabetes	最受欢迎的类别	0	0	0	0
	频率	665	80	219	964
	百分比	93.7%	88.9%	88.3%	92.0%
Blood pressure	最受欢迎的类别	1	1	1	1
	频率	445	49	139	633
	百分比	62.7%	54.4%	56.0%	60.4%
Atrial fibrillation	最受欢迎的类别	0	0	0	0
	频率	641	83	216	940
	百分比	90.3%	92.2%	87.1%	89.7%
Smoker	最受欢迎的类别	0	0	0	0
	频率	578	69	179	826
	百分比	81.4%	76.7%	72.2%	78.8%
Cholesterol	最受欢迎的类别	0	0	0	0
	频率	406	52	136	594
	百分比	57.2%	57.8%	54.8%	56.7%
History of angina	最受欢迎的类别	0	0	0	0
	频率	493	52	167	712
	百分比	69.4%	57.8%	67.3%	67.9%

分类变量标准值所实现的目标与刻度标准值大致相同，但分类变量标准值会报告模态（最常见）类别以及该类别中对等组的个案的数量和百分比。 值的比较可能比较麻烦；例如，初看时 Gender 对聚类形成的贡献似乎比 Smoker 大，因为三个对等组的 Smoker 的模态类别都相同，而对等组 3 的 Gender 存在差异。但是，因为 Gender 仅具有两个值，所以可以推断对等组 3 中 49.2% 的个案的值为 0，这与其他对等组中的百分比非常类似。比较而言，Smoker 的百分比介于 72.2% 到 81.4% 之间。

图片 9-13
分类变量标准值（选定的变量）

		对等 ID			
		1	2	3	已組合
Dead on arrival	最受欢迎的类别	0	1	0	0
	频率	710	90	248	958
	百分比	100.0%	100.0%	100.0%	91.4%
Initial Rankin score	最受欢迎的类别	0	(缺失值)	5	5
	频率	166	90	104	193
	百分比	23.4%	100.0%	41.9%	18.4%
CAT scan result	最受欢迎的类别	0	(缺失值)	0	0
	频率	607	90	184	791
	百分比	85.5%	100.0%	74.2%	75.5%
Clot-dissolving drugs	最受欢迎的类别	2	(缺失值)	0	2
	频率	318	90	129	394
	百分比	44.8%	100.0%	52.0%	37.6%
Died in hospital	最受欢迎的类别	0	(缺失值)	1	0
	频率	710	90	171	787
	百分比	100.0%	100.0%	69.0%	75.1%
Treatment result	最受欢迎的类别	1	(缺失值)	1	1
	频率	524	90	96	620
	百分比	73.8%	100.0%	38.7%	59.2%
Post-event preventative surgery	最受欢迎的类别	0	(缺失值)	(缺失值)	0
	频率	323	90	171	369
	百分比	45.5%	100.0%	69.0%	35.2%
Post-event rehabilitation	最受欢迎的类别	0	(缺失值)	(缺失值)	0
	频率	278	90	171	314
	百分比	39.2%	100.0%	69.0%	30.0%

刻度变量标准值引发的怀疑在分类标准值表中的下面的位置得到确认。对等组 2 完全由到达时已死亡的病人组成，因此所有治疗和康复变量都缺失。对等组 3 中的大多数病人 (69.0%) 都在治疗期间死亡，因此康复变量的模态类别为 (Missing Value)。

异常指标摘要

图片 9-14
异常指标摘要

	异常列表中的 N	最小值	最大值	均值	标准偏差
异常索引	21	1.736	2.837	1.872	.240

异常列表中的 N 由指定确定：异常百分比为 2%

该表提供异常列表中个案的异常指标值的摘要统计。

原因摘要

图片 9-15
原因摘要（治疗和康复变量）

	出现次数作为原因		变量影响统计量			
	频率	百分比	最小值	最大值	均值	标准偏差
Dead on arrival	0	.0%	.	.	.	.
Initial Rankin score	0	.0%	.	.	.	.
CAT scan result	0	.0%	.	.	.	.
Clot-dissolving drugs	0	.0%	.	.	.	.
Died in hospital	0	.0%	.	.	.	.
Treatment result	0	.0%	.	.	.	.
Post-event preventative surgery	0	.0%	.	.	.	.
Post-event rehabilitation	0	.0%	.	.	.	.
Rankin score at 1 month	0	.0%	.	.	.	.
Rankin score at 3 months	0	.0%	.	.	.	.
Rankin score at 6 months	0	.0%	.	.	.	.
Barthel index at 1 month	13	61.9%	.064	.124	.100	.021
Barthel index at 3 months	1	4.8%	.088	.088	.088	.
Barthel index at 6 months	1	4.8%	.080	.080	.080	.
Recoded Barthel index at 1 month	0	.0%	.	.	.	.
Recoded Barthel index at 3 months	0	.0%	.	.	.	.
Recoded Barthel index at 6 months	0	.0%	.	.	.	.
Stroke between release and 1 month	0	.0%	.	.	.	.
Stroke between 1 and 3 months	0	.0%	.	.	.	.
Stroke between 3 and 6 months	0	.0%	.	.	.	.
Length of stay for rehabilitation	0	.0%	.	.	.	.
Total treatment and rehabilitation costs in thousands	6	28.6%	.120	.411	.202	.112
缺失比例	0	.0%	.	.	.	.
整体	21	100.0%	.064	.411	.127	.076

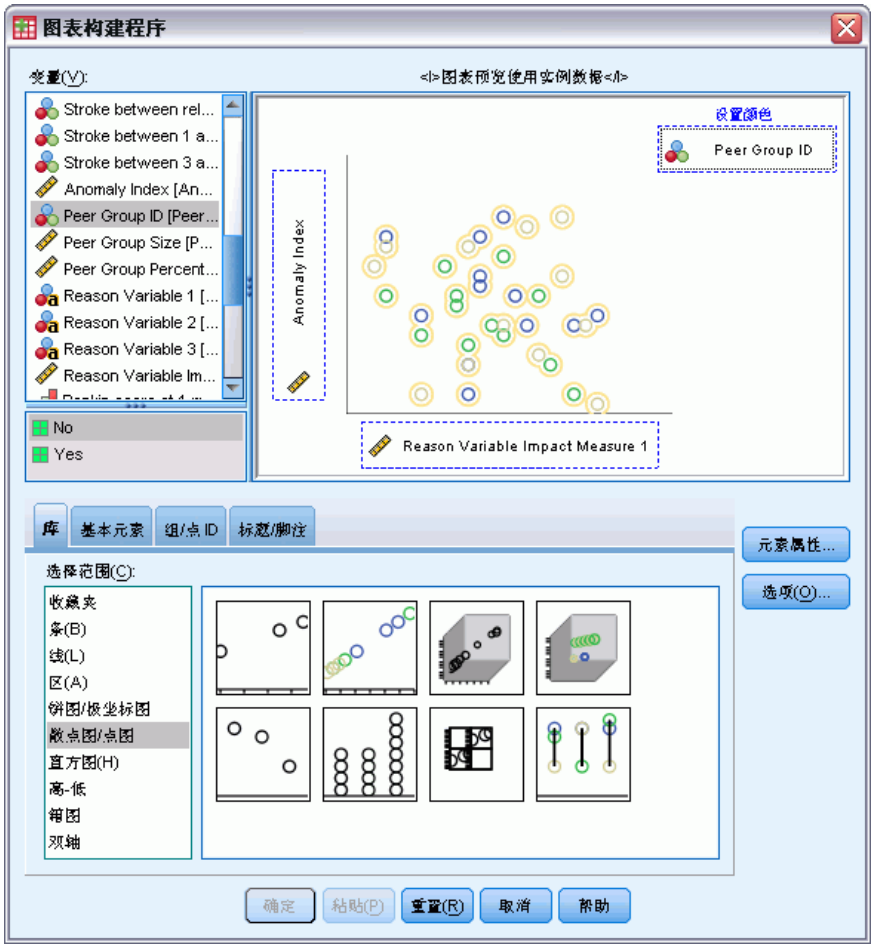
对于分析中的每个变量，该表都将变量的角色作为主要原因汇总。大多数变量（例如从 Dead on arrival 到 Post-event rehabilitation 的变量）都不是异常列表上的任何个案的主要原因。Barthel index at 1 month 是最常见的原因，然后是 Total treatment and rehabilitation costs in thousands。会汇总变量影响统计，报告每个变量的最小值、最大值和均值影响，以及作为多个个案的原因的变量的标准差。

按变量影响显示的异常指标的散点图

该表包含许多有用信息，但相互关系很难把握。通过使用保存的变量，就可以构造一个使此过程变得简单的图形。

▶ 要构造此散点图，请从菜单中选择：
图形 > 图表构建程序...

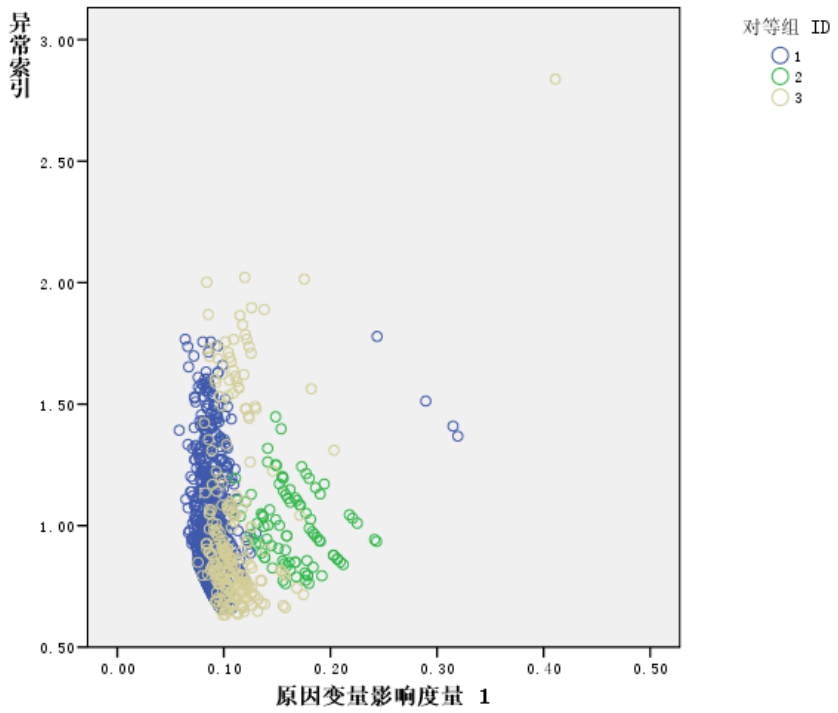
图片 9-16
“图表生成器”对话框



- 选择**散点图/点图**库，并将“分组散点”图标拖到画布上。
- 选择异常指标作为 y 变量，选择原因变量影响度量 1 作为 x 变量。
- 选择对等组 ID 作为用于设置颜色的变量。
- 单击**确定**。

通过这些选择将生成散点图。

图片 9-17
按第一个原因变量的影响度量显示的异常指标的散点图



观察该图形可得出几个结论：

- 右上角的个案属于对等组 3，该个案既是最异常的个案，也是单变量做出最大贡献的个案。
- 沿 y 轴向下移动，我们发现三个属于对等组 3 的个案，它们的异常指标值稍高于 2.00。这些个案作为异常个案，应进行更准确的调查。
- 沿 x 轴移动，我们发现四个属于对等组 1 的个案，它们的变量影响度量近似介于 0.23 到 0.33 之间。这些个案应该进行更全面的调查，因为这些值将这些个案与图中的点的主体分离开来。
- 对等组 2 看来具有较强的齐性，其异常指标和变量影响值与集中趋势的偏离都不太大。

摘要

通过使用“标识异常个案”过程，就可以找出可以进行进一步检查的多个个案。这些个案不能由其他验证过程标识，因为要使用变量之间的关系（而不仅仅是变量本身的值）确定异常个案。

比较让人失望的是，对等组主要是基于两个变量构建的：Dead on arrival 和 Died in hospital。在进行进一步分析时，您可以了解强制创建大量对等组的效果，或者可以执行仅包含经过治疗活下来的病人的分析。

相关过程

“标识异常个案”过程是一个检测数据文件中的异常个案的有用工具。

- 验证数据过程标识活动数据集中可疑的和无效的个案、变量和数据值。

章
10

最优离散化

“最优离散化”过程通过将每个变量的值分布到块中而离散化一个或多个尺度变量（称为**离散化输入**变量）。块的构成根据“监督”离散化过程的分类向导变量得以最优化。然后就可以在需要使用或者最好使用分类变量的过程中使用块而不是初始数据值进行进一步的分析。

最优离散化算法

“最优离散化”算法的基本步骤具有如下特征：

预处理（可选）。 离散化输入变量分为 n 个块（其中 n 由您指定），并且每个块都包含相同数量的个案或者尽可能相同数量的个案。

标识潜在分割点。 不与离散化输入变量的第二大不同值属于相同的向导变量类别的离散化输入的每个不同值就是潜在的分割点。

选择分割点。 获得最多信息的潜在分割点使用 MDLP 接受准则进行评估。一直重复此过程，直到潜在分割点被接受。接受的分割点定义块的端点。

使用最优离散化离散贷款申请数据

作为银行在降低贷款拖欠率方面的一项举措，信贷员收集了过去和现在的客户的财务和统计信息，希望能够创建一个模型来预测拖欠贷款的概率。有几个潜在预测变量是刻度变量，但信贷员想要构建一个最适合分类预测变量的模型。

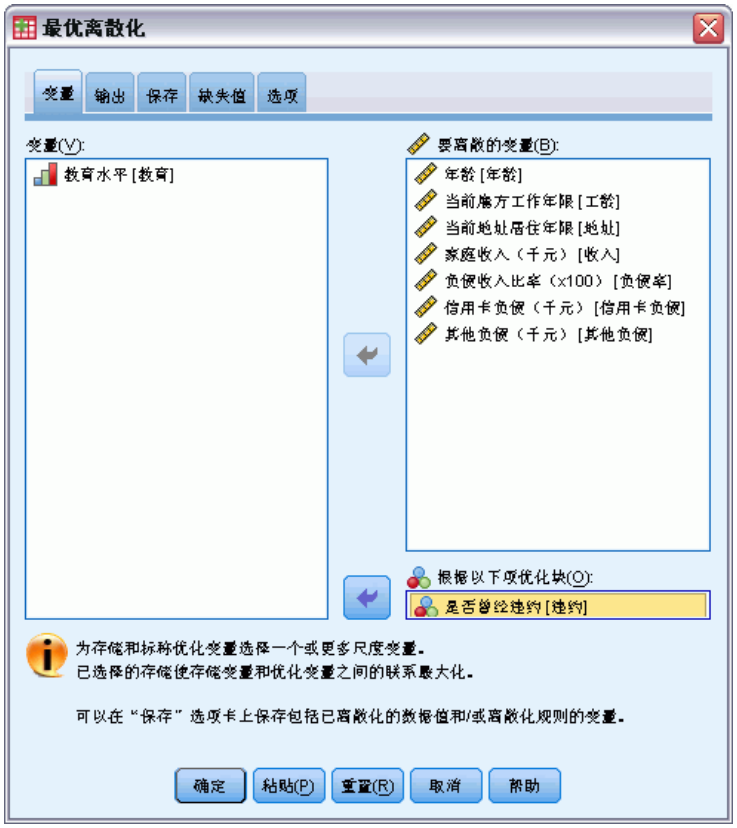
有关 5000 个过去的客户的信息收集在 bankloan_binning.sav 中。有关详细信息，请参阅第 129 页码附录 A 中的样本文件。使用“最优离散化”过程生成刻度预测变量的离散化规则，然后使用生成的规则处理 bankloan.sav。然后就可以使用处理过的数据集创建预测模型。

运行分析

- 要运行“最优离散化”分析，请从菜单中选择：

  转换 > **最优离散化...**

图片 10-1
“最优离散化”对话框，“变量”选项卡



- 选择年龄和当前雇方工作年限到其他负债（千元）作为块的变量。
- 选择是否曾经违约作为向导变量。
- 单击**输出**选项卡。

图片 10-2
“最优离散化”对话框，“输出”选项卡

- 选择**描述统计**和**模型熵**作为离散化的变量。
- 单击**保存**选项卡。

图片 10-3
“最优离散化”对话框，“保存”选项卡

- 选择**创建包含已离散化的数据值的变量**。
- 为要包含生成的离散化规则的语法文件输入一个路径和文件名。在本例中，我们使用了 /bankloan_binning-rules.sps。
- 单击**确定**。

这些选择将生成以下命令语法：

```
* Optimal Binning.
OPTIMAL BINNING
/VARIABLES GUIDE=default BIN=age employ address income debtinc creddebt
othdebt SAVE=YES (INTO=age_bin employ_bin address_bin income_bin debtinc_bin
creddebt_bin othdebt_bin)
/CRITERIA METHOD=MDLP
PREPROCESS=EQUALFREQ (BINS=1000)
FORCEMERGE=0
LOWERLIMIT=INCLUSIVE
LOWEREND=UNBOUNDED
UPPEREND=UNBOUNDED
/MISSING SCOPE=PAIRWISE
/OUTFILE RULES='/bankloan_binning-rules.sps'
/PRINT ENDPOINTS DESCRIPTIVES ENTROPY.
```

- 此过程将使用 MDLP 离散化和向导变量 default 对离散化输入变量 age、employ、address、income、debtinc、creddebt 和 othdebt 进行离散化。
- 这些变量的离散化值将存储在新变量 age_bin、employ_bin、address_bin、income_bin、debtinc_bin、creddebt_bin 和 othdebt_bin 中。
- 如果一个离散化输入变量具有超过 1000 个不同值，则在执行 MDLP 离散化之前，均等频率方法会将不同值的数量减少到 1000。
- 表示离散化规则的命令语法保存在文件 /bankloan_binning-rules.sps 中。
- 离散化输入变量要求块端点、描述统计和模型熵值。
- 其他离散化标准设置为其缺省值。

描述统计

图片 10-4
描述统计(S)

描述统计

	N	最小值	最大值	相异值数	块个数
年龄	5000	20	58	39	2
当前雇方工作年限	5000	0	38	39	4
当前地址居住年限	5000	0	37	38	3
家庭收入(千元)	5000	12.10	2461.70	1100	2
负债收入比率(x100)	5000	.08	44.62	2060	5
信用卡负债(千元)	5000	.01	139.58	5000	4
其他负债(千元)	5000	.01	416.52	4999	2

描述统计表提供了有关离散化输入变量的摘要信息。前四个列与预先离散化的值有关。

- **N** 是分析中使用的个案的数量。如果使用缺失值的列表删除，则此值在所有变量中都应该相同。使用成对缺失值处理时，此值可能不是常数。因为此数据集不具有缺失值，所以该值就是个案数。
- **最小值**和**最大值**列显示每个离散化输入变量在数据集中（预先离散化）的最小值和最大值。除了了解每个变量的值的观察范围，它们还可用于捕获期望范围外的值。
- **不同值的数目**指出哪些变量是使用均等频率算法预处理的。缺省情况下，具有超过 1000 个不同值的变量（家庭收入（千元）到其他负债（千元））会预先离散化为 1000 个不同的块。然后这些预处理的块就会使用 MDLP 针对向导变量进行离散化。您可以在选项选项卡上控制预处理功能。
- **块个数**是该过程生成的块的最终个数，该数目大大小于不同值的数目。

模型熵

图片 10-5
模型熵

模型熵

	模型熵
年龄	.788
当前雇方工作年限	.754
当前地址居住年限	.781
家庭收入(千元)	.803
负债收入比率(x100)	.711
信用卡负债(千元)	.776
其他负债(千元)	.801

模型熵越小表示参照变量 是否曾经违约 上的离散化变量的预测准确性越高。

模型熵可告诉您预测模型中每个变量对于拖欠概率的作用。

- 对于生成的每个块，最可能的预测变量是包含与向导变量具有相同值的个案的变量，这样，向导变量就能够完全预测。这样的预测变量具有未定义的模型熵。这在真实情况下通常是不存在的，并且可能表明数据质量有问题。
- 最不可能的预测变量是不好过猜测的变量；此模型熵的值取决于数据。在此数据集中，总共 5000 个客户中有 1256 个客户（或 0.2512）拖欠了贷款，有 3744 个客户（或 0.7488）未拖欠；因此，最不可能的预测变量的模型熵可能为 $-0.2512 \times \log_2(0.2512) - 0.7488 \times \log_2(0.7488) = 0.8132$。

我们最有把握的就是，模型熵值较低的变量应该得出较好的预测变量，因为模型熵值是否合适取决于应用程序和数据。在本例中，生成的块的数目较大（相对于不同类别的数目）的变量看似具有较低的模型熵值。应该使用具有更广泛的变量选择工具的预测建模过程，将这些离散化输入变量作为预测变量进行进一步评估。

离散化摘要

离散化摘要按向导变量的值报告生成的块的边界以及每个块的频率计数。会为每个离散化输入变量生成一个单独的离散化摘要表。

图片 10-6
“年龄”的离散化摘要

块	端点		水平 是否曾经违约 的个案数		
	下限	上限	否	是	合计
1	[a]	32	1129	639	1768
2	32	[a]	2615	617	3232
合计			3744	1256	5000

Premature end of pattern reached - probably illegal character: ':' (pattern is: "每个块的计算方法为: 下限 <= ^1 < 上限。")

a. 无限制

年龄的摘要显示 1768 个客户（年龄不大于 32 岁）放在了“块 1”中，而剩余的 3232 个客户（年龄大于 32 岁）则放入“块 2”。“块 1”中曾经违约的客户的比例 (639/1768=0.361) 大于“块 2”中曾经违约的客户的比例 (617/3232=0.191)。

图片 10-7
“家庭收入（千元）”的离散化摘要

块	端点		水平 是否曾经违约 的个案数		
	下限	上限	否	是	合计
1	[a]	26.70	1054	513	1567
2	26.70	[a]	2690	743	3433
合计			3744	1256	5000

Premature end of pattern reached - probably illegal character:
':' (pattern is: "每个块的计算方法为: 下限 <= ^1 < 上限。")
a. 无限制

家庭收入（千元）的摘要显示了类似的模式，在 26.70 处存在一个唯一的分割点，“块 1”中曾经违约的客户的比例（513/1567=0.327）大于“块 2”中曾经违约的客户的比例（743/3433=0.216）。如同根据模型熵统计量进行的预测，这两个比例的差异小于年龄的比例差异。

图片 10-8
“其他负债（千元）”的离散化摘要

块	端点		水平 是否曾经违约 的个案数		
	下限	上限	否	是	合计
1	[a]	2.19	2161	539	2700
2	2.19	[a]	1583	717	2300
合计			3744	1256	5000

Premature end of pattern reached - probably illegal character:
':' (pattern is: "每个块的计算方法为: 下限 <= ^1 < 上限。")
a. 无限制

其他负债（千元）的摘要显示了相反的模式，在 2.19 处存在唯一的一个分割点，“块 1”中曾经违约的客户的比例（539/2700=0.200）大于“块 2”中曾经违约的客户的比例（717/2300=0.312）。同样，如同根据模型熵统计量进行的预测，这两个比例的差异小于年龄的比例差异。

图片 10-9
“当前雇方工作年限”的离散化摘要

块	端点		水平 是否曾经违约 的个案数		
	下限	上限	否	是	合计
1	[a]	3	629	478	1107
2	3	8	1066	461	1527
3	8	18	1471	268	1739
4	18	[a]	578	49	627
合计			3744	1256	5000

Premature end of pattern reached - probably illegal character:
':' (pattern is: "每个块的计算方法为: 下限 <= ^1 < 上限。")
a. 无限制

当前雇方工作年限的摘要显示了拖欠贷款的客户的比例随块数的增加而降低的模式。

块	拖欠者比例
1	0.432
2	0.302
3	0.154
4	0.078

图片 10-10
“当前地址居住年限”的离散化摘要

块	端点		水平 Previously defaulted 的个案数		
	下限	上限	No	Yes	总计
1		7	1652	829	2481
2	7	14	1184	313	1497
3	14		908	114	1022
总计			3744	1256	5000

每个块的计算方法为: 下限 <= Years at current address < 上限。

a. 无限制

当前地址居住年限的摘要显示了类似的模式。如同根据模型熵统计量进行的预测，当前雇方工作年限的各个块中拖欠贷款的客户比例的差异比当前地址居住年限的更明显。

块	拖欠者比例
1	0.334
2	0.209
3	0.112

图片 10-11
“信用卡负债（千元）”的离散化摘要

块	端点		水平 是否曾经违约 的个案数		
	下限	上限	否	是	合计
1	[a]	.97	2169	466	2635
2	.97	1.91	848	307	1155
3	1.91	6.05	643	352	995
4	6.05	[a]	84	131	215
合计			3744	1256	5000

Premature end of pattern reached - probably illegal character:
':' (pattern is: "每个块的计算方法为: 下限 <= ^1 < 上限。")

a. 无限制

信用卡负债（千元）的摘要显示了相反的模式，其中拖欠贷款的客户的数目随块数的增加而增加。当前雇方工作年限和当前地址居住年限能够更好地标识不拖欠贷款的概率较高的客户，而信用卡负债（千元）能够更好地标识拖欠贷款的概率较高的客户。

块	拖欠者比例
1	0.177
2	0.266
3	0.354
4	0.609

图片 10-12
“负债收入比率（x100）”的离散化摘要

块	端点		水平 是否曾经违约 的个案数		
	下限	上限	否	是	合计
1	[a]	4.39	912	88	1000
2	4.39	12.09	2006	437	2443
3	12.09	18.71	625	386	1011
4	18.71	31.00	198	303	501
5	31.00	[a]	3	42	45
合计			3744	1256	5000

Premature end of pattern reached - probably illegal character:
':' (pattern is: "每个块的计算方法为: 下限 <= ^1 < 上限。")

a. 无限制

负债收入比率（x100）的摘要显示的模式与信用卡负债（千元）显示的模式相似。此变量具有最低的模型熵值，因此是贷款拖欠概率的最佳预测变量。对于分类拖欠贷款的概率较高的客户，该变量好过信用卡负债（千元），对于分类拖欠贷款的概率较低的客户，该变量的效果和当前雇方工作年限一样好。

块	拖欠者比例
1	0.088
2	0.179
3	0.382
4	0.605
5	0.933

离散化的变量

图片 10-13
“数据编辑器”中 bankloan_binning.sav 的离散化变量

	default	age_bin	ed	employ_bin	address_bin	income_bin	debtinc_bin	creddebt_bin
1	0	2	3	3	2	2	2	4
2	0	1	1	3	1	2	3	2
3	0	2	1	3	3	2	2	1
4	0	2	1	3	3	2	1	3
5	0	1	2	1	1	2	3	2
6	1	2	2	2	1	1	2	1
7	0	2	1	4	2	2	4	3
8	0	2	1	3	2	2	1	1
9	0	1	1	2	1	1	4	2
10	0	2	1	1	2	1	4	3
11	0	1	1	1	1	1	1	1
12	1	1	1	2	1	1	2	1

数据视图 变量视图

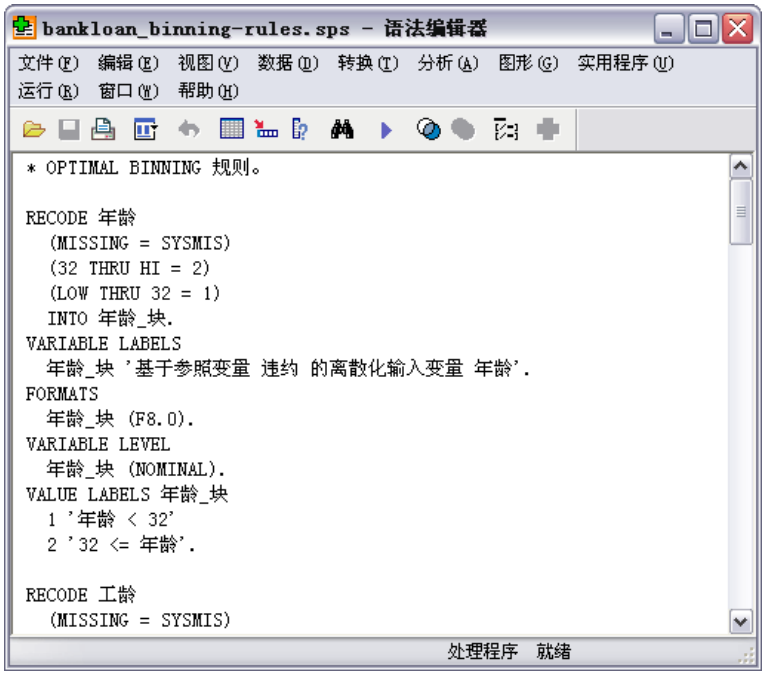
在“数据编辑器”中，此数据集的离散化过程的结果非常明显。如果要使用描述性或报告过程生成定制离散化结果摘要，则这些离散化变量很有用，但不建议使用此数据集构建预测模型，因为离散化规则是使用这些个案生成的。较好的计划是将该离散化规则应用于包含其他客户的信息的另一个数据集。

应用语法离散化规则

运行“最优离散化”过程时，您曾请求将该过程生成的离散化规则保存为命令语法。

- 打开 bankloan_binning-rules.sps。

图片 10-14
语法规则文件

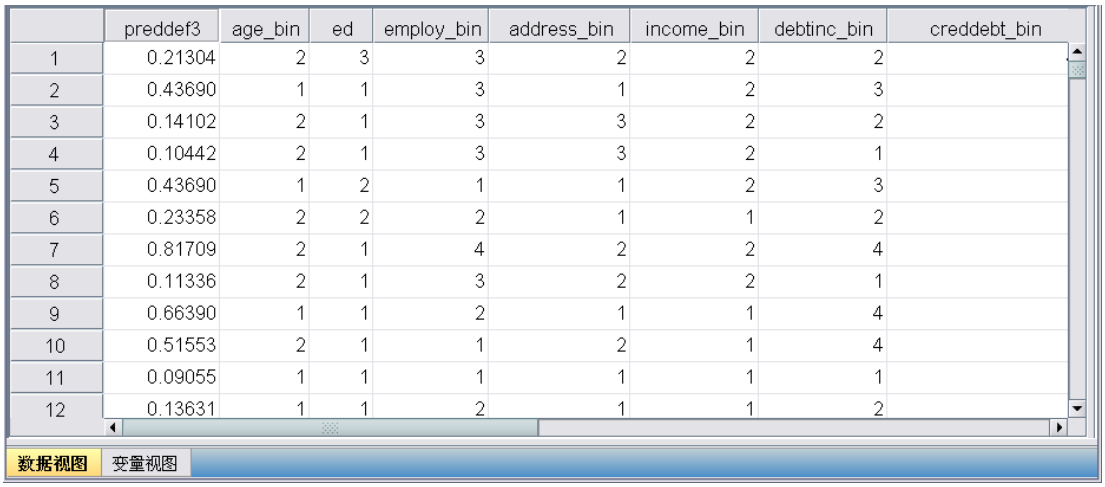


对于每个离散化输入变量，总是有一个命令语法块来执行离散化；设置变量标签、格式和水平；并设置块的值标签。这些命令可应用于具有与 bankloan_binning.sav 相同的变量的数据集。

- 打开 bankloan.sav。有关详细信息，请参阅第 129 页码附录 A 中的样本文件。
- 返回 bankloan_binning-rules.sps 的“语法编辑器”视图。

▶ 要应用这些离散化规则，请从“语法编辑器”菜单中选择：
运行 > 全部…

图片 10-15
“数据编辑器”中 bankloan.sav 的离散化变量



	preddef3	age_bin	ed	employ_bin	address_bin	income_bin	debtinc_bin	creddebt_bin
1	0.21304	2	3	3	2	2	2	
2	0.43690	1	1	3	1	2	3	
3	0.14102	2	1	3	3	2	2	
4	0.10442	2	1	3	3	2	1	
5	0.43690	1	2	1	1	2	3	
6	0.23358	2	2	2	1	1	2	
7	0.81709	2	1	4	2	2	4	
8	0.11336	2	1	3	2	2	1	
9	0.66390	1	1	2	1	1	4	
10	0.51553	2	1	1	2	1	4	
11	0.09055	1	1	1	1	1	1	
12	0.13631	1	1	2	1	1	2	

数据视图 变量视图

bankloan.sav 中的变量已经根据通过对 bankloan_binning.sav 运行“最优离散化”过程生成的规则进行了离散化。此数据集现在可以用于构建最好使用或者要求使用分类变量的预测模型。

摘要

通过使用“最优离散化”过程，我们生成了用于尺度变量（这些刻度变量可作为贷款拖欠概率的潜在预测变量）的离散化规则，并将这些规则应用于分离数据集。

在离散化过程中，您发现离散化的当前雇方工作年限和当前地址居住年限能够更好地标识不拖欠贷款的概率较高的客户，而信用卡负债（千元）能够更好地标识拖欠贷款的概率较高的客户。为贷款拖欠概率构建预测模型时，这个有趣的发现可以给您一些额外的启示。如果主要关注的是避免坏账，则信用卡负债（千元）将比当前雇方工作年限和当前地址居住年限更为重要。如果首要任务是扩展客户群，则当前雇方工作年限和当前地址居住年限将更为重要。

附录
A

样本文件

随产品一起安装的样本文件可以在安装目录的 Samples 子目录中找到。对于以下每种语言在“样本”子目录中有单独的文件夹：英语、法语、德语、意大利语、日语、韩语、波兰语、俄语、简体中文、西班牙语和繁体中文。

并非所有样本文件均提供此处的全部语言版本。如果样本文件未提供某种语言的版本，则相应语言文件夹中包含该样本文件的英语版本。

描述

以下是对在整个文档的各种示例中使用的样本文件的简要描述。

- **accidents.sav。** 该假设数据文件涉及某保险公司，该公司正在研究给定区域内汽车事故的年龄和性别风险因子。每个个案对应一个年龄类别和性别类别的交叉分类。
- **adl.sav。**该假设数据文件涉及在确定针对脑卒中患者的建议治疗类型的优点方面的举措。医师将女性脑卒中患者随机分配到两组中的一组。第一组患者接受标准的物理治疗，而第二组患者则接受附加的情绪治疗。在进行治疗的三个月时间里，将为每个患者进行一般日常生活行为的能力评分并作为原始变量。
- **advert.sav。**该假设数据文件涉及某零售商在检查广告支出与销售业绩之间的关系方面的举措。为此，他们收集了过去的销售数据以及相关的广告成本..
- **aflatoxin.sav。**该假设数据文件涉及对谷物的黄曲霉毒素的检测，该毒素的浓度会因谷物产量的不同（不同谷物之间及同种谷物之间）而有较大变化。谷物加工机从 8 个谷物产量的每一个中收到 16 个样本并以十亿分之几 (PPB) 为单位来测量黄曲霉毒素的水平。
- **anorectic.sav。**在研究厌食/暴食行为的标准症状参照时，研究人员(Van der Ham, Meulman, Van Strien, 和 Van Engeland, 1997)对 55 名已知存在进食障碍的青少年进行了调查。其中每名患者每年都将进行四次检查，因此总观测数为 220。在每次观测期间，将对这些患者按 16 种症状逐项评分。但 71 号和 76 号患者的症状得分均在时间点 2 缺失，47 号患者的症状得分在时间点 3 缺失，因此有效观测数为 217。
- **bankloan.sav。**该假设数据文件涉及某银行在降低贷款拖欠率方面的举措。该文件包含 850 位过去和潜在客户的财务和人口统计信息。前 700 个个案是以前曾获得贷款的客户。剩下的 150 个个案是潜在客户，银行需要按高或低信用风险对他们进行分类。
- **bankloan_binning.sav。**该假设数据文件包含 5,000 位过去客户的财务和人口统计信息。
- **behavior.sav。**在一个经典示例中(Price 和 Bouffard, 1974)，52 名学生被要求以 10 分的标度对 15 种情况和 15 种行为的组合进行评价，该 10 分的标度介于 0 = 平均值在个人值之上，值被视为相异性。
- **behavior_ini.sav。**该数据文件包含 behavior.sav 的二维解的初始配置。

- **brakes.sav。**该假设数据文件涉及某生产高性能汽车盘式制动器的工厂的质量控制。该数据文件包含对 8 台专用机床中每一台的 16 个盘式制动器的直径测量。盘式制动器的目标直径为 322 毫米。
- **breakfast.sav。**在一项经典研究中(Green 和 Rao, 1972)，21 名 Wharton School MBA 学生及其配偶被要求按照喜好程度顺序对 15 种早餐食品进行评价，从 1 =他们的喜好根据六种不同的情况加以记录，从“全部喜欢”到“只带饮料的快餐”。
- **breakfast-overall.sav。**该数据文件只包含早餐食品喜好的第一种情况，即“全部喜欢”。
- **broadband_1.sav。**该假设数据文件包含各地区订制了全国宽带服务的客户的数量。该数据文件包含 4 年期间 85 个地区每月的订户数量。
- **broadband_2.sav。**该数据文件和 broadband_1.sav 一样，但包含另外三个月的数据。
- **car_insurance_claims.sav。**在别处被提出和分析的(McCullagh 和 Nelder, 1989)关于汽车损坏赔偿的数据集。平均理赔金额可以当作其具有 gamma 分布来建模，通过使用逆联接函数将因变量的均值与投保者年龄、车辆类型和车龄的线性组合关联。提出理赔的数量可以作为尺度权重。
- **car_sales.sav。**该数据文件包含假设销售估计值、订价以及各种品牌和型号的车辆的物理规格。订价和物理规格可以从 edmunds.com 和制造商处获得。
- **car_sales_uprepared.sav。**这是 car_sales.sav 的修改版本，不包含字段的任何已转换版本。
- **carpet.sav。**在一个常用示例中(Green 和 Wind, 1973)，一家公司非常重视一种新型地毯清洁用品的市场营销，希望检验以下五种因素对消费者偏好的影响—包装设计、品牌名称、价格、优秀家用品 标志和退货保证。包装设计有三个因子水平，每个因子水平因刷体位置而不同；有三个品牌名称（K2R、Glory 和 Bissell）；有三个价格水平；最后两个因素各有两个级别（有或无）。十名消费者对这些因素所定义的 22 个特征进行了排序。变量优选包含对每个特征的平均等级的排序。低排序与高偏好相对应。此变量反映了对每个特征的偏好的总体度量。
- **carpet_prefs.sav。**该数据文件所基于的示例和在 carpet.sav 中所描述的一样，但它还包含从 10 位消费者的每一位中收集到的实际排列顺序。消费者被要求按照从最喜欢到最不喜欢的顺序对 22 个产品特征进行排序。carpet_plan.sav 中定义了变量 PREF1 到 PREF22 包含相关特征的标识符。
- **catalog.sav。**该数据文件包含某编目公司出售的三种产品的假设每月销售数据。同时还包括 5 个可能的预测变量的数据。
- **catalog_seasfac.sav。**除添加了一组从“季节性分解”过程中计算出来的季节性因子和附带的日期变量外，该数据文件和 catalog.sav 是相同的。
- **cellular.sav。**该假设数据文件涉及某便携式电话公司在减少客户流失方面的举措。客户流失倾向分被应用到帐户，分数范围从 0 到 100。得到 50 分或更高分数的帐户可能会更换提供商。
- **ceramics.sav。**该假设数据文件涉及某制造商在确定新型优质合金是否比标准合金具有更高的耐热性方面的举措。每个个案代表对一种合金的单独检验；个案中会记录合金的耐热极限。
- **cereal.sav。**该假设数据文件涉及一份 880 人参于的关于早餐喜好的民意调查，该调查记录了参与者的年龄、性别、婚姻状况以及生活方式是否积极（根据他们是否每周至少做两次运动）。每个个案代表一个单独的调查对象。

- **clothing_defects.sav。**这是关于某服装厂的质量控制过程的假设数据文件。检验员要对工厂中每次大批量生产的服装进行抽样检测并清点不合格的服装的数量。
- **coffee.sav。**这是关于六种冰咖啡的认知品牌形象(Kennedy, Riquier, 和 Sharp, 1996)的数据文件。对于 23 种冰咖啡特征属性中的每种属性，人们选择了由该属性所描述的所有品牌。为保密起见，六种品牌用 AA、BB、CC、DD、EE 和 FF 来表示。
- **contacts.sav。**该假设数据文件涉及一组公司计算机销售代表的联系方式列表。根据这些销售代表所在的公司部门及其公司的秩来对每个联系方式进行分类。同时还记录了最近一次的销售量、最近一次销售距今的时间和所联系公司的规模。
- **creditpromo.sav。**该假设数据文件涉及某百货公司在评价最新信用卡促销的效果方面的举措。为此，随机选择了 500 位持卡人。其中一半收到了宣传关于在接下来的三个月内降低消费利率的广告。另一半收到了标准的季节性广告。
- **customer_dbase.sav。** 该假设数据文件涉及某公司在使用数据仓库中的信息来为最有可能回应的客户提供特惠商品方面的举措。随机选择客户群的子集并为其提供特惠商品，同时记录下他们的回应。
- **customer_information.sav。**该假设数据文件包含客户邮寄信息，如姓名和地址。
- **customer_subset.sav。**来自 customer_dbase.sav 的拥有 80 个个案的子集。
- **debate.sav。**该假设数据文件涉及在某政治辩论前后对该辩论的参与者所做的调查的成对回答。每个个案对应一个单独的调查对象。
- **debate_aggregate.sav。**该假设数据文件分类汇总了 debate.sav 中的回答。每个个案对应一个辩论前后的偏好的交叉分类。
- **demo.sav。**这是关于购物客户数据库的假设数据文件，用于寄出每月的商品。将记录客户对商品是否有回应以及各种人口统计信息。
- **demo_cs_1.sav。**该假设数据文件涉及某公司在汇编调查信息数据库方面的举措的第一步。每个个案对应不同的城市，并记录地区、省、区和城市标识。
- **demo_cs_2.sav。**该假设数据文件涉及某公司在汇编调查信息数据库方面的举措的第二步。每个个案对应来自第一步中所选城市的不同的家庭单元，并记录地区、省、区、市、子区和单元标识。还包括设计前两个阶段的抽样信息。
- **demo_cs.sav。**该假设数据文件包含用复杂抽样设计收集的调查信息。每个个案对应不同的家庭单元，并记录各种人口统计和抽样信息。
- **dmdata.sav。**该假设数据文件包含直销公司的人口统计学和购买信息。dmdata2.sav 包含收到测试邮件的联系人子集的信息，dmdata3.sav 包含未收到测试邮件的其余联系人的信息。
- **dietstudy.sav。**该假设数据文件包含对 “Stillman diet” (Rickman, Mitchell, Dingman, 和 Dalen, 1974) 的研究结果。每个个案对应一个单独的主体，并记录其在实行饮食方案前后的体重（磅）以及甘油三酸酯的水平（毫克/100 毫升）。
- **dvdplayer.sav。**这是关于开发新的 DVD 播放器的假设数据文件。营销团队用原型收集了焦点小组数据。每个个案对应一个单独的被调查用户，并记录他们的人口统计信息及其对原型问题的回答。
- **german_credit.sav。**该数据文件取自加州大学欧文分校的 Repository of Machine Learning Databases (Blake 和 Merz, 1998)中的 “German credit” 数据集。
- **grocery_1month.sav。**该假设数据文件是在数据文件 grocery_coupons.sav 的基础上加上了每周购物“累计”，所以每个个案对应一个单独的客户。所以，一些每周更改的变量消失了，而且现在记录的消费金额是为期四周的研究过程中的消费金额之和。

- **grocery_coupons.sav。**该假设数据文件包含由重视顾客购物习惯的杂货连锁店收集的调查数据。对每位顾客调查四周，每个个案对应一个单独的顾客周，并记录有关顾客购物地点和方式的信息（包括那一周里顾客在杂货上的消费金额）。
- **guttman.sav。** Bell (Bell, 1961)创建了一个表，用来阐释可能的社会群体。Guttman (Guttman, 1968)引用了该表的一部分，其中包括五个变量，用于描述以下七个理论社会群体的社会交往、对群体的归属感、成员的物理亲近度以及关系正式性：观众（比如在足球比赛现场的人们）、听众（比如在剧院或听课堂讲座的人们）、公众（比如报纸或电视观众）、组织群体（与观众类似但具有紧密的关系）、初级群体（关系密切）、次级群体（自发组织）及现代社区（因在物理上亲近而导致关系松散并需要专业化服务）。
- **health_funding.sav。**该假设数据文件包含关于保健基金（每 100 人的金额）、发病率（每 10,000 人的比率）以及保健提供商拜访率（每 10,000 的比率）的数据。每个个案代表不同的城市。
- **hivassay.sav。**该假设数据文件涉及某药物实验室在开发用于检测 HIV 感染的快速化验方面的举措。化验结果为八个加深的红色阴影，如果有更深的阴影则表示感染的可能性很大。用 2,000 份血液样本来进行实验室试验，其中一半受到 HIV 感染而另一半没有受到感染。
- **hourlywagedata.sav。**该假设数据文件涉及在政府机关和医院工作的具有不同经验水平的护士的时薪。
- **insurance_claims.sav。**该假设数据文件涉及某保险公司，该公司希望构建一个模型用于标记可疑的、具有潜在欺骗性的理赔。每个个案代表一次单独的理赔。
- **insure.sav。**该假设数据文件涉及某保险公司，该公司正在研究指示客户是否会根据 10 年的人寿保险合同提出理赔的风险因子。数据文件中的每个个案代表一副根据年龄和性别进行匹配的合同，其中一份记录了一次理赔而另一份则没有。
- **judges.sav。**该假设数据文件涉及经过训练的裁判（加上一个体操爱好者）对 300 次体操表演给出的分数。每行代表一次单独的表演；裁判们观看相同的表演。
- **kinship_dat.sav。**Rosenberg 和 Kim (Rosenberg 和 Kim, 1975) 开始分析 15 个亲属关系项（伯母、兄弟、表兄妹、女儿、父亲、孙女、祖父、祖母、孙子、母亲、侄子或外甥、侄女或外甥女、姐妹、儿子和叔叔）。他们让四组大学生（两组女同学，两组男同学）根据相似程度将各项排序。他们让其中的两组同学（一组女同学，一组男同学）进行了两次排序，第二次排序和第一次排序采取的标准不同。这样，一共得到六组“源”。每个源对应一个 15×15 的相似性矩阵，其单元格中的值等于源中的人数减去此源中对象被划分的次数。
- **kinship_ini.sav。**该数据文件包含 kinship_dat.sav 的三维解的初始配置。
- **kinship_var.sav。**该数据文件包含自变量 gender、gener(ation) 和 degree (of separation)，这些变量可用于解释 kinship_dat.sav 的解的维数。具体而言，它们可用来将解的空间限制为这些变量的线性组合。
- **marketvalues.sav。**该数据文件涉及 1999 - 2000 年间 Algonquin，Ill. 地区新的房屋开发中的住房销售。这些销售仅仅来自公众记录。
- **nhis2000_subset.sav。**美国健康访问调查 (NHIS) 是针对美国全体公民的大型人口调查。该调查对美国的具有全国代表性的家庭样本进行了面对面的访问，并获取了每个家庭的成员的健康行为和健康状态的人口统计信息和观察数据。该数据文件包含取自 2000 年调查信息的子集。国家健康统计中心。2000 年美国健康访问调查。公用数据文件和文档。

ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Datasets/NHIS/2000/。2003 年发布。

- **ozone.sav。**这些数据包含了用来根据其余变量预测臭氧浓度的六个气象变量的 330 个观察值。在以前的研究人员中，(Breiman 和 Friedman(F), 1985) 和 (Hastie 和 Tibshirani, 1990) 发现了这些变量之间的非线性，这妨碍了标准回归方法。
- **pain_medication.sav。**该假设数据文件包含用于治疗慢性关节炎疼痛的抗炎药的临床试验结果。我们感兴趣的是该药见效的时间以及它和现有药物的比较。
- **patient_los.sav。**该假设数据文件包含被医院确诊为疑似心肌梗塞（即 MI 或“心脏病发作”）的患者的治疗记录。每个个案对应一位单独的患者，并记录与其住院期有关的一些变量。
- **patlos_sample.sav。**该假设数据文件包含在治疗心肌梗塞（即 MI 或“心脏病发作”）期间收到溶解血栓剂的患者的样本治疗记录。每个个案对应一位单独的患者，并记录与其住院期有关的一些变量。
- **poll_cs.sav。**该假设数据文件涉及民意测验专家在确定正式立法前公众对法案的支持水平方面的举措。个案对应注册的选民。每个个案记录选民居住的县、镇、区。
- **poll_cs_sample.sav。**该假设数据文件包含在 poll_cs.sav 中列出的选民的样本。该样本是根据 poll.csplan 中指定的设计来选取的，而且该数据文件记录包含概率和样本权重。请注意，由于该抽样计划使用与大小成正比 (PPS) 方法，因此，还有一个文件 (poll_jointprob.sav) 包含联合选择概率。在选取了样本之后，对应于选民人群统计信息及其对提交法案的意见的附加变量将被收集并添加到数据文件。
- **property_assess.sav。**该假设数据文件涉及某县资产评估员在利用有限的资源不断更新资产价值评估方面的举措。个案对应过去一年中县里所出售的资产。数据文件中的每个个案记录资产所在的镇、最后评估资产的评估员、该次评估距今的时间、当时的估价以及资产的出售价格。
- **property_assess_cs.sav。**该假设数据文件涉及某州资产评估员在利用有限的资源不断更新资产价值评估方面的举措。个案对应该州的资产。数据文件中的每个个案记录资产所在的县、镇和区，最后一次评估距今的时间以及当时的估价。
- **property_assess_cs_sample.sav。**该假设数据文件包含在 property_assess_cs.sav 中列出的资产的样本。该样本是根据 property_assess.csplan 中指定的设计来选取的，而且该数据文件记录包含概率和样本权重。在选取了样本之后，附加变量 Current value 将被收集并添加到数据文件。
- **recidivism.sav。**该假设数据文件涉及某政府执法机构在了解其管辖区域内的屡犯率方面的举措。每个个案对应先前的一名罪犯，并记录其人口统计信息和第一次犯罪的详细资料；如果在第一次被捕后两年内又第二次被捕，则还将记录两次被捕间隔的时间。
- **recidivism_cs_sample.sav。**该假设数据文件涉及某政府执法机构在了解其管辖区域内的屡犯率方面的举措。每个个案对应在 2003 年 6 月期间第一次被捕释放的先前的一名罪犯，并记录其人口统计信息和第一次犯罪的详细资料，及其第二次被捕的数据（如果发生在 2006 年 6 月底之前）。根据 recidivism_cs.csplan 中指定的抽样计划从抽样部门选择罪犯；该计划使用与大小成正比 (PPS) 方法，因此，还有一个文件 (recidivism_cs_jointprob.sav) 包含联合选择概率。
- **rfm_transactions.sav。** 此假设数据文件包含购买交易数据，即每笔交易的购买日期、购买商品和消费金额。

- **salesperformance.sav。**这是关于评估两个新的销售培训课程的假设数据文件。60 名员工被分成 3 组且都接受标准的培训。另外，组 2 接受技术培训；组 3 接受实践教程。在培训课程结束时，对每名员工进行测验并记录他们的分数。数据文件中的每个个案代表一名单独的受训者，并记录其被分配到的组以及测验的分数。
- **satisf.sav。**该假设数据文件涉及某零售公司在 4 个商店位置所进行的满意度调查。总共对 582 位客户进行了调查，每个个案代表一位单独客户的回答。
- **screws.sav。**该数据文件包含关于螺钉、螺栓、螺母和图钉的特征的信息(Hartigan, 1975)。
- **shampoo_ph.sav**这是关于某发制品厂的质量控制的假设数据文件。在规定的时间间隔对六批独立输出的产品进行检测并记录它们的 pH 值。目标范围是 4.5 - 5.5。
- **ships.sav。**在别处被提出和分析的(McCullagh 等。, 1989)关于波浪对货船造成的损坏的数据集。在给定了船的类型、建造工期和服务期后，可以根据以泊松比率发生来为事件计数建模。在因子交叉分类构成的表格中，每个单元格的分类汇总服务月数提供遇到风险的值。
- **site.sav。**该假设数据文件涉及某公司在为扩展业务而选择新址方面的举措。该公司聘请了两名顾问分别对选址进行评估，除了提供长期报告外，他们还要以“前景颇佳”、“前景良好”或“前景不佳”来对每个选址进行总结。
- **smokers.sav。**该数据文件摘自 1998 年全国家庭药物滥用调查并且是美国家庭的概率样本。(http://dx.doi.org/10.3886/ICPSR02934) 因此，分析该数据文件的第一步应该是对数据进行加权以反映总体趋势。
- **stocks.sav** 该假设数据文件包含某一年的股票价格和成交量。
- **stroke_clean.sav。**该假设数据文件包含某医学数据库在经过“数据准备”选项中的过程清理后的状态。
- **stroke_invalid.sav。**该假设数据文件包含某医学数据库的初始状态及一些数据输入错误。
- **stroke_survival。**此假设数据文件涉及正在研究结束缺血性中风后复元计划的患者存活时间的研究人员面临着很多挑战。中风后，记录心肌梗塞、缺血性中风或出血性中风的发生及其时间。样本为左侧截短，因为只包含在中风后管理的复元计划结束后存活的患者。
- **stroke_valid.sav。**该假设数据文件包含在使用“验证数据”过程检查值后，某医学数据库的状态。它仍包含潜在异常个案。
- **survey_sample.sav。**此数据文件包含调查数据，包括人口统计学数据和各种态度测量。它基于 1998 NORC 综合社会调查的变量子集，但某些数据值已经过修改，并添加了其他虚拟变量以供演示用途。
- **telco.sav。**该假设数据文件涉及某电信公司在减少客户群中的客户流失方面的举措。每个个案对应一个单独的客户，并记录各类人口统计和服务用途信息。
- **telco_extra.sav。**该数据文件与 telco.sav 数据文件类似，但删除了 “tenure” 和经对数转换的客户消费变量，代替它们的是标准化的对数转换客户消费变量。
- **telco_missing.sav。** 该数据文件是 telco.sav 数据文件的子集，但某些人口统计数据值已被缺失值替换。

- **testmarket.sav。**该假设数据文件涉及某快餐连锁店为其菜单添加新项目的计划。有三种可能的促销新产品的活动，所以会在多个随机选择的市场中的地点引入新的项目。在每个地点采用不同的促销方式，并记录新项目前四周的每周销售情况。每个个案对应单独地点的一周。
- **testmarket_1month.sav。**该假设数据文件是在数据文件 testmarket.sav 的基础上加上了每周销售“累计”，所以每个个案对应一个单独的地点。所以，一些每周更改的变量消失了，而且现在记录的销售是为期四周的研究过程中的销售之和。
- **tree_car.sav。**该假设数据文件包含人口统计和车辆购买价格数据。
- **tree_credit.sav。**该假设数据文件包含人口统计和银行贷款历史数据。
- **tree_missing_data.sav** 该假设数据文件包含具有大量缺失值的人口统计和银行贷款历史数据。
- **tree_score_car.sav。**该假设数据文件包含人口统计和车辆购买价格数据。
- **tree_textdata.sav。**这是一个只有两个变量的样本数据文件，主要打算在指定测量级别和值标签之前显示变量的默认状态。
- **tv-survey.sav。**该假设数据文件涉及由某电视演播室进行的一项关于是否要继续制作一档成功的节目的调查。906 位调查对象被问及他们在各种情况下是否会收看该节目。每行代表一位单独的调查对象；每列代表一种单独的情况。
- **ulcer_recurrence.sav。**此文件包含某项研究的部分信息，该研究旨在比较两种用来防止溃疡复发的治疗的功效。它提供了区间数据的优秀示例并且已在别处被提出和分析(Collett, 2003)。
- **ulcer_recurrence_recoded.sav。**该文件重新组织 ulcer_recurrence.sav 中的信息以允许为研究的每个区间的事件概率建模而不是简单地为研究结束事件概率建模。它已在别处被提出和分析(Collett 等。, 2003)。
- **verd1985.sav。**该数据文件涉及某项调查(Verdegaal, 1985)。该调查记录了 15 个主体对 8 个变量的响应。需要处理的变量被分成 3 个集。数据集 1 包含 年龄 和 婚姻；数据集 2 包含 宠物 和 新闻；数据集 3 包含 音乐 和 居住。宠物被尺度化为多名义而年龄被尺度化为有序；所有其他变量都被尺度化为单名义。
- **virus.sav。**该假设数据文件涉及某因特网服务提供商 (ISP) 在确定病毒对其网络的影响方面的举措。他们从发现病毒到威胁得以遏制这段时间内跟踪其网络上受感染的电子邮件的流量的（近似）百分比。
- **wheeze_steubenville.sav。**这是关于空气污染对儿童健康影响的纵向研究的一个子集(Ware, Dockery, Spiro III, Speizer, 和 Ferris Jr., 1984)。这些数据包含儿童的气喘状况的重复二分类测量（这些儿童来自 Steubenville，Ohio，年龄为 7 到 10 岁），以及母亲在研究的第一年中是否为吸烟者的固定记录。
- **workprog.sav。**该假设数据文件涉及一份尝试为弱势群体提供较好的工作的政府工作计划。文件后还有一个潜在计划参与者的样本，其中一些参与者是被随机选择来参加该计划的，而其他参与者则不是。每个个案代表一位单独的计划参与者。
- **worldsales.sav** 该假设数据文件包含按不同大洲和产品列出的销售收入。

附录
B

注意事项

这些信息开发用于在全球提供的产品和服务。

IBM 可能在其他国家/地区中不提供在本文档中讨论的产品、服务或功能。请咨询您当地的 IBM 代表以了解有关您所在地区当前可用产品和服务的信息。任何对 IBM 产品、程序或服务的引用，并不意味着仅可使用这些 IBM 产品、程序或服务。作为替代，可以使用任何功能相当的产品、程序或服务，前提是不侵犯任何 IBM 知识产权。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。

在本文档中介绍的主题可能涉及 IBM 的专利或申请中的专利。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, Armonk, NY 10504-1785, U.S.A.

要查询双字节字符集 (DBCS) 相关许可证信息，请联系所在国家/地区中的 IBM 知识产权部门，或者以书面形式将查询函件发送至：

Intellectual Property Licensing, Legal and Intellectual Property Law, IBM Japan Ltd., 1623-14, Shimotsuruma, Yamato-shi, Kanagawa 242-8502 Japan.

以下段落不适用于英国或此类条款与当地法律不符的其他国家/地区： INTERNATIONAL BUSINESS MACHINES 公司“按原样”提供本出版物，不保证任何明示或暗示，包括但不限于对非侵权性、适销性或对特定用途适用性的暗示担保。有些州/省不允许针对某些交易的明示或暗示免责条款，因此本声明可能不适用于您。

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 可能随时对本出版物中所述的产品和/或程序进行改进和/或更改，恕不另行通知。

在本信息材料中对任何非 IBM 网站的引用仅为了方便用户，并不以任何方式表明对这些网站的认可。这些网站上的材料并非本 IBM 产品材料的一部分，您对这些网站的使用需自担风险。

IBM 可以自认为适当并且不会对您构成任何约束的任何方式使用或分发您提供的任何信息。

如果本程序的受许可方试图了解有关程序的信息以启用：(i) 在独立创建的程序和其他程序（包括本程序）之间交换信息；(ii) 相互使用交换的信息，则应联系：

IBM Software Group, Attention:Licensing, 233 S. Wacker Dr., Chicago, IL 60606, USA.

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。

在本文档中介绍的受许可保护程序，及其所有受许可保护材料由 IBM 在双方签署的“IBM 客户协议”、“IBM 国际程序许可证协议”或任何其他等同协议下提供。

非 IBM 产品的相关信息来自这些产品的供应商，及其发布的公告或其他公开来源。IBM 并未测试过这些产品，无法确认有关非 IBM 产品的性能准确性、兼容性或任何其他声明。有关非 IBM 产品功能的问题应由这些产品的供应商负责。

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。所有这些名称均为虚构，与真实商业企业使用的名称和地址的任何雷同纯属巧合。

如果您正在查阅此信息的软拷贝，照片和彩色插图可能不会显示。

商标

IBM、IBM 徽标、ibm.com 和 SPSS 是 IBM Corporation 的商标，在全球许多司法辖区注册。有关最新的 IBM 商标列表，请访问网页 http://www.ibm.com/legal/copytrade.shtml。

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。

Java 和所有基于 Java 的商标和徽标是 Sun Microsystems, Inc. 在美国、其他国家或地区或两者的商标。

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。

此产品使用 Polar 工程咨询公司的 WinWrap Basic，版权 1993 年-2007 年，http://www.winwrap.com。

其他产品和服务名称可能是 IBM 或其他公司的商标。

Adobe 产品屏幕截图重印已获得 Adobe Systems Incorporated 的许可。

Microsoft 产品屏幕截图重印已获得 Microsoft Corporation 的许可。

参考书目

Bell, E. H. 1961. Social foundations of human behavior:Introduction to the study of sociology. New York: Harper & Row.

Blake, C. L., 和 C. J. Merz. 1998. "UCI Repository of machine learning databases." Available at http://www.ics.uci.edu/~mlearn/MLRepository.html.

Breiman, L., 和 J. H. Friedman(F). 1985. Estimating optimal transformations for multiple regression and correlation. Journal of the American Statistical Association, 80, .

Collett, D. 2003. Modelling survival data in medical research, 2 ed. Boca Raton: Chapman & Hall/CRC.

Green, P. E., 和 V. Rao. 1972. Applied multidimensional scaling. Hinsdale, Ill.: Dryden Press.

Green, P. E., 和 Y. Wind. 1973. Multiattribute decisions in marketing:A measurement approach. Hinsdale, Ill.: Dryden Press.

Guttman, L. 1968. A general nonmetric technique for finding the smallest coordinate space for configurations of points. Psychometrika, 33, .

Hartigan, J. A. 1975. Clustering algorithms. New York: John Wiley and Sons.

Hastie, T., 和 R. Tibshirani. 1990. Generalized additive models. London: Chapman and Hall.

Kennedy, R., C. Riquier, 和 B. Sharp. 1996. Practical applications of correspondence analysis to categorical data in market research. Journal of Targeting, Measurement, and Analysis for Marketing, 5, .

McCullagh, P., 和 J. A. Nelder. 1989. Generalized Linear Models, 2nd ed. London: Chapman & Hall.

Price, R. H., 和 D. L. Bouffard. 1974. Behavioral appropriateness and situational constraints as dimensions of social behavior. Journal of Personality and Social Psychology, 30, .

Rickman, R., N. Mitchell, J. Dingman, 和 J. E. Dalen. 1974. Changes in serum cholesterol during the Stillman Diet. Journal of the American Medical Association, 228, .

Rosenberg, S., 和 M. P. Kim. 1975. The method of sorting as a data-gathering procedure in multivariate research. Multivariate Behavioral Research, 10, .

Van der Ham, T., J. J. Meulman, D. C. Van Strien, 和 H. Van Engeland. 1997. Empirically based subgrouping of eating disorders in adolescents:A longitudinal perspective. British Journal of Psychiatry, 170, .

Verdegaal, R. 1985. Meer sets analyse voor kwalitatieve gegevens (in Dutch). Leiden: Department of Data Theory, University of Leiden.

Ware, J. H., D. W. Dockery, A. Spiro III, F. E. Speizer, 和 B. G. Ferris Jr.. 1984. Passive smoking, gas cooking, and respiratory health of children living in six cities. American Review of Respiratory Diseases, 129, .

索引