IBM SPSS Bootstrapping 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 35 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。

受许可保护材料 - IBM 所有

Copyright IBM Corporation 1989, 2011.

美国政府用户受限权利 - 使用、复制或披露受与 IBM Corp. 签订的 GSA ADP Schedule Contract 的限制。

前言

IBM® SPSS® Statistics 是一种用于分析数据的综合系统。Bootstrapping 可选附加模块提供本手册中描述的其他分析方法。此 Bootstrapping 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。

Copyright IBM Corporation 1989, 2011.

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

附录

部分 I:
用户指南

章 1

Bootstrap 简介

在收集数据时，您通常对从中抽取样本的总体的属性感兴趣。您通过从样本计算得到的估计值来做出有关这些总体参数的推论。例如，如果随产品附带的 Employee data.sav 数据集为来自更大的职员总体的随机样本，则当前工资的样本均值 $34,419.57 为职员总体的当前平均工资的估计值。并且，此估计值对大小为 474 的样本具有标准误 $784.311，因此职员总体当前平均工资的 95% 置信区间为 $32,878.40 至 $35,960.73。但是这些估计量有多可靠呢？对于某些“已知”总体和表现良好的参数，我们非常了解样本估计值的属性，因此可以相信这些结果。Bootstrap 寻求发现有关“未知”总体和异常参数的估计量的更多属性信息。

图片 1-1
做出有关总体平均值的参数推论

			Statistic	标准 误差
当前薪金	均值		$34,419.57	$784.311
	95% 置信区间	下限	$32,878.40	
		上限	$35,960.73	

Bootstrap 的工作原理

考虑最简单的情况，对于样本大小为 N 的数据集，您可以通过放回方式从原始数据集中取得 B 个 bootstrap 样本（大小为 N），并为这 B 个 bootstrap 样本中的每个样本计算估计量。这 B 个 bootstrap 估计值为大小为 B 的样本，您可从中做出有关估计量的推论。例如，如果从 Employee data.sav 数据集取得 1000 个 bootstrap 样本，则当前工资样本均值的 bootstrap 估计标准误 $778.76 可以替代估计值 $784.311。

此外，Bootstrap 可进一步提供标准误和中位数的置信区间，而参数估计则对此不适用。

图片 1-2
做出有关样本均值的 bootstrap 推论

					Bootstrap			
							95% 置信区间	
			Statistic	标准 误差	偏差	标准 误差	下限	上限
当前薪金	均值		$34,419.57	$784.311	$18.52	$776.91	$32,990.38	$36,026.06
	95% 置信区间	下限	$32,878.40					
		上限	$35,960.73					
	中值		$28,875.00		$-13.22	$536.63	$27,750.00	$29,850.00

a. 除非得到提示，否则 bootstrap 结果基于 1000 个 bootstrap 样例

产品对 Bootstrap 的支持

Bootstrap 作为子对话框包含在支持 bootstrap 的过程中。请参见支持 Bootstrap 的过程，以了解哪些过程支持 bootstrap。

Copyright IBM Corporation 1989, 2011.

当在对话框中请求 bootstrap 时，将在对话框生成的常规语法外粘贴新的单独 BOOTSTRAP 命令。BOOTSTRAP 命令按照您的指定创建 bootstrap 样本。产品内部对这些 bootstrap 样本的处理方式与拆分类似，尽管它们不会明确显示在“数据编辑器”中。这意味着，内部有效存在 B*N 个个案，因此在 bootstrap 期间处理数据时，状态栏上的个案计数器将从 1 计数到 B*N。输出管理系统 (OMS) 用于收集在每个“bootstrap 拆分”上运行分析的结果。这些 bootstrap 结果在汇聚后与过程生成的其余常规输出一起显示在“查看器”中。在某些个案中，您可能会看到对“bootstrap 拆分 0”的引用；这是原始数据集。

章 2

Bootstrap

Bootstrap 方法可以导出稳健的标准误估计值，并能为诸如均值、中位数、比例、几率比、相关系数或回归系数等估计值导出置信区间。它还可用于构建假设检验。当参数估计方法的假设存在疑问（例如，异方差残差拟合较小样本的回归模型），参数推论无法执行或需要非常复杂的标准误计算公式（例如，为中位数、四分位数和其他百分位数计算置信区间）时，Bootstrap 是最好的替代选项。

示例。 一家电信公司每月大约会流失 27% 的客户。为了正确实施减少客户流失的举措，管理部门需要知道这一百分比在各个预定义客户组之间是否不同。使用 bootstrap 可以确定单一的客户流失率是否充分描述了四个主要的客户类型。有关详细信息，请参阅第 3 章中的使用 Bootstrap 获得比例的置信区间中的IBM SPSS Bootstrapping 20。

在查看员工记录时，管理部门对员工的以往工作经验比较感兴趣。工作经验向右偏斜，这使得平均值在作为员工“典型”以往工作经验的估计方面不如中位数理想。然而，在产品中无法获得中位数的参数置信区间。有关详细信息，请参阅第 3 章中的使用 Bootstrap 获得中位数的置信区间中的IBM SPSS Bootstrapping 20。

管理部门还有兴趣使用线性模型拟合当前和起始工资之间的差异，以确定哪些因素与员工工资上升存在关联。在 bootstrap 线性模型时，可以使用特殊的重新抽样方法（残差和狂野 bootstrap）以获得更准确的结果。有关详细信息，请参阅第 3 章中的使用 Bootstrap 选择更好的预测变量中的IBM SPSS Bootstrapping 20。

许多过程支持 bootstrap 抽样和对 bootstrap 样本分析结果的汇聚。指定 bootstrap 分析的控件作为公共子对话框直接集成在支持 bootstrap 的过程中。在 bootstrap 对话框上的设置会在不同过程中保留，因此如果您通过对话框使用 bootstrap 运行频率分析，则对支持此功能的其他过程而言，bootstrap 默认打开。

获取 Bootstrap 分析

- 从菜单中，选择一个支持 bootstrap 的过程，并单击 **Bootstrap**。

Copyright IBM Corporation 1989, 2011.

图片 2-1
Bootstrap 对话框

▶ 选择执行 bootstrap。

此外，您还可以控制下列选项：

样本数。对于生成的百分位数和 BCa 区间，建议使用至少 1000 个 bootstrap 样本。指定一个正整数。

为 Mersenne 扭曲器设置种子。设置种子允许您复制分析。使用此控件类似于将“Mersenne 扭曲器”设为活动生成器并在“随机数生成器”对话框中指定固定起始点，两者的重大差别在于在此对话框中设置种子会保留随机数生成器的当前状态并在分析完成后恢复该状态。

置信区间。指定一个大于 50 且小于 100 的置信水平。百分位数区间简单地使用对应于所需置信区间百分位数的有序 bootstrap 值。例如，一个 95% 的百分位数置信区间使用 bootstrap 值的第 2.5 个和第 97.5 个百分位数作为区间的下限和上限（必要时插 bootstrap 值）。偏差修正加速 (BCa) 区间为调整区间，它更加准确，但代价是需要更长的计算时间。

抽样。简单方法为通过放回方式从原始数据集进行个案重新取样。分层方法为通过放回方式从原始数据集进行个案重新取样，但在层次变量的交叉分类定义的层内。如果层中的单元相对均一，且不同层间的单元相差较大，则分层 bootstrap 抽样非常有用。

支持 Bootstrap 的过程

下列过程支持 bootstrap。

注意：

- Bootstrap 不能用于多重插补数据集。如果在数据集中存在 Imputation_ 变量，Bootstrap 对话框将被禁用。
- Bootstrap 使用列表删除来确定个案基础；也就是说，在任何分析变量上具有缺失值的个案将从分析中删除，因此当 bootstrap 生效时，列表删除也处于生效，即使分析过程指定了其他缺失值处理方式。

Statistics Base 选项

频率

- 统计表支持均值、标准差、方差、中位数、偏度、峰度和百分位数的 bootstrap 估计。
- 频率表支持百分比的 bootstrap 估计。

描述性

- 描述统计表支持均值、标准差、方差、偏度和峰度的 bootstrap 估计。

探索

- 描述表支持均值、5% 切尾均值、标准差、方差、中位数、偏度、峰度和内距的 bootstrap 估计。
- M 估计量表支持 Huber 的 M 估计量、Tukey 的双权重、Hampel 的 M 估计量和 Andrew 的 Wave 的 bootstrap 估计。
- 百分位数表支持百分位数的 bootstrap 估计。

交叉表

- 定向测量表支持 Lambda、Goodman 和 Kruskal Tau、不定性系数和 Somers 的 d 的 bootstrap 估计。
- 对称度量表支持 Phi、Cramer 的 V、列联系数、Kendall 的 tau-b、Kendall 的 tau-c、Gamma、Spearman 相关性和 Pearson 的 R 的 bootstrap 估计。
- 风险评估表支持几率比的 bootstrap 估计。
- Mantel-Haenszel 一般几率比表支持 ln(Estimate) 的 bootstrap 估计和显著性检验。

均值

- 报告表支持均值、中位数、组内中位数、标准差、方差、峰度、偏度、调和均值和几何均值的 bootstrap 估计。

单样本 T 检验

- 统计表支持均值和标准差的 bootstrap 估计。
- 检验表支持平均值差值的 bootstrap 估计和显著性检验。

独立样本 T 检验

- 组统计表支持均值和标准差的 bootstrap 估计。
- 检验表支持平均值差值的 bootstrap 估计和显著性检验。

配对样本 T 检验

- 统计表支持均值和标准差的 bootstrap 估计。
- 相关性表支持相关性的 bootstrap 估计。
- 检验表支持均值的 bootstrap 估计。

单因素方差分析

- 描述统计表支持均值和标准差的 bootstrap 估计。
- 多重比较表支持平均值差值的 bootstrap 估计。
- 对比检验表支持对比值的 bootstrap 估计和显著性检验。

GLM 单变量

- 描述统计表支持均值和标准差的 bootstrap 估计。
- 参数估计值表支持系数、B 的 bootstrap 估计和显著性检验。
- 对比结果表支持差值的 bootstrap 估计和显著性检验。
- 估计边际均值：估计值表支持均值的 bootstrap 估计。
- 估计边际均值：成对比较表支持平均值差值的 bootstrap 估计。
- 两两比较检验：多重比较表支持平均值差值的 bootstrap 估计。

双变量相关

- 描述统计表支持均值和标准差的 bootstrap 估计。
- 相关性表支持相关性的 bootstrap 估计和显著性检验。

注意：

如果除 Pearson 相关性外，还请求了非参数相关性（Kendall 的 tau-b 或 Spearman），则对话框将分别为其粘贴 **CORRELATIONS** 和 **NONPAR CORR** 命令，以及单独的 **BOOTSTRAP** 命令。将使用相同的 bootstrap 样本计算全部相关性。

在汇聚之前，将 Fisher Z 转换应用于相关性。在汇聚之后，应用逆 Z 转换。

偏相关

- 描述统计表支持均值和标准差的 bootstrap 估计。
- 相关性表支持相关性的 bootstrap 估计。

线性回归

- 描述统计表支持均值和标准差的 bootstrap 估计。
- 相关性表支持相关性的 bootstrap 估计。
- 模型概要表支持 Durbin-Watson 的 bootstrap 估计。

- 系数表支持系数、B 的 bootstrap 估计和显著性检验。
- 相关系数表支持相关性的 bootstrap 估计。
- 残差统计表支持均值和标准差的 bootstrap 估计。

Ordinal 回归

- 参数估计值表支持系数、B 的 bootstrap 估计和显著性检验。

判别分析

- 标准化典则判别函数系数表支持标准化系数的 bootstrap 估计。
- 典则判别函数系数表支持非标准化系数的 bootstrap 估计。
- 分类函数系数表支持系数的 bootstrap 估计。

Advanced Statistics 选项

GLM 多变量

- 参数估计值表支持系数、B 的 bootstrap 估计和显著性检验。

线性混合模型

- 固定效应估计值表支持估计值的 bootstrap 估计和显著性检验。
- 协方差参数估计值表支持估计值的 bootstrap 估计和显著性检验。

广义线性模型

- 参数估计值表支持系数、B 的 bootstrap 估计和显著性检验。

Cox 回归

- 方程中的变量表支持系数、B 的 bootstrap 估计和显著性检验。

Regression 选项

二元 Logistic 回归

- 方程中的变量表支持系数、B 的 bootstrap 估计和显著性检验。

多项 Logistic 回归

- 参数估计值表支持系数、B 的 bootstrap 估计和显著性检验。

BOOTSTRAP 命令附加功能

使用命令语法语言还可以：

- 执行残差和狂野 bootstrap 抽样（SAMPLING 子命令）

请参阅命令语法参考以获取完整的语法信息。

部分 II:
示例

章
3

Bootstrap

Bootstrap 方法可以导出稳健的标准误估计值，并能为诸如均值、中位数、比例、几率比、相关系数或回归系数等估计值导出置信区间。它还可用于构建假设检验。当参数估计方法的假设存在疑问（例如，异方差残差拟合较小样本的回归模型），参数推论无法执行或需要非常复杂的标准误计算公式（例如，为中位数、四分位数和其他百分位数计算置信区间）时，Bootstrap 是最好的替代选项。

使用 Bootstrap 获得比例的置信区间

一家电信公司每月大约会流失 27% 的客户。为了正确实施减少客户流失的举措，管理部门需要知道这一百分比在各个预定义客户组之间是否不同。

该信息收集在 telco.sav 中。有关详细信息，请参阅第 28 页码附录 A 中的样本文件。使用 bootstrap 来确定单一的客户流失率是否充分描述了四个主要的客户类型。

注意：本示例使用了“频率”过程，并且需要 Statistics Base 选项。

准备数据

您必须首先按客户类别拆分文件。

- 要拆分文件，请从“数据编辑器”菜单中选择：
  数据 > 拆分文件...

图片 3-1
“拆分文件”对话框

- 选择比较各组。

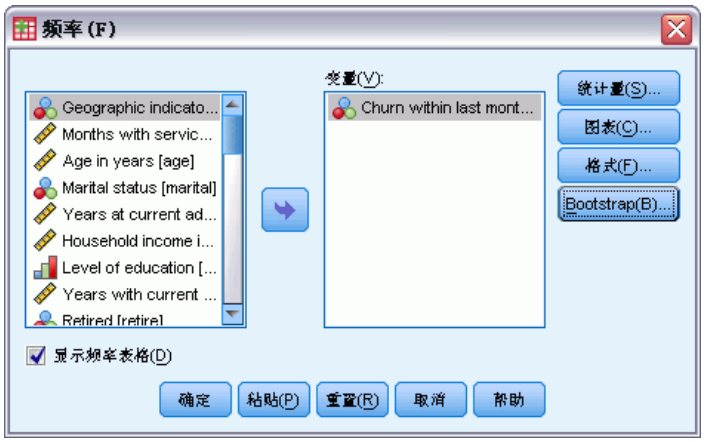
Copyright IBM Corporation 1989, 2011.

- 选择客户类别作为建立组的变量。
- 单击**确定**。

运行分析

- 要获得比例的 bootstrap 置信区间，请在菜单中选择：
  分析 > **描述统计** > **频率...**

图片 3-2
“频率”主对话框

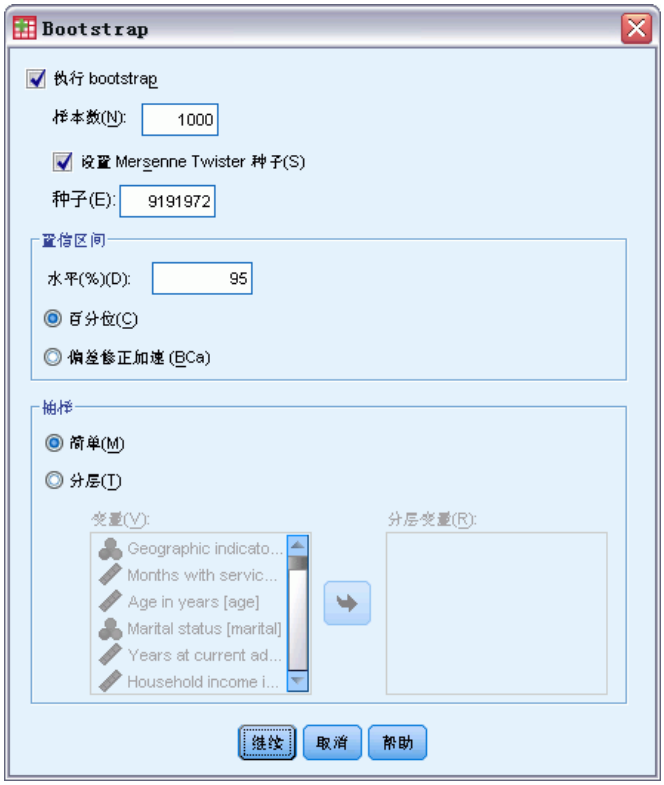


- 选择上月内流失 [churn] 作为分析变量。
- 单击**统计量**。

图片 3-3
“统计量”对话框

- 选择“集中趋势”组中的**均值**。
- 单击**继续**。
- 单击“频率”对话框中的 **Bootstrap**。

图片 3-4
Bootstrap 对话框



- 选择**执行 bootstrap**。
- 为准确复制本例中的结果，请选择**为 Mersenne 扭曲器设置种子**并键入 9191972 作为种子。
- 单击**继续**。
- 在“频率”对话框中单击**确定**。

这些选择将生成以下命令语法：

```
SORT CASES  BY custcat.
SPLIT FILE LAYERED BY custcat.
PRESERVE.
SET RNG=MT MTINDEX=9191972.
SHOW RNG.
BOOTSTRAP
  /SAMPLING METHOD=SIMPLE
  /VARIABLES INPUT=churn
  /CRITERIA CILEVEL=95 CITYPE=PERCENTILE  NSAMPLES=1000
  /MISSING USERMISSING=EXCLUDE.
FREQUENCIES VARIABLES=churn
  /STATISTICS=MEAN
  /ORDER=ANALYSIS.
RESTORE.
```

- SORT CASES 和 SPLIT FILE 命令在变量 custcat 上拆分文件。

- PRESERVE 和 RESTORE 命令“记住”随机数生成器的当前状态，并在 bootstrap 结束后将系统恢复到该状态。
- SET 命令将随机数生成器设置成 Mersenne 扭曲器，并将索引设置成 9191972，以便准确复制 bootstrap 结果。SHOW 命令在输出中显示索引以供参考。
- BOOTSTRAP 命令使用简单重新取样请求 1,000 个 bootstrap 样本。
- 变量 churn 用于确定重复取样的个案基础。在此变量上具有缺失值的记录将从分析中删除。
- 在 BOOTSTRAP 后的 FREQUENCIES 过程会在每个 bootstrap 样本上运行。
- STATISTICS 子命令在原始数据上为变量 churn 生成平均值。此外，还将为频率表中的平均值和百分比生成汇聚的统计量。

Bootstrap 指定

图片 3-5
Bootstrap 指定

采样方法	简单箱图
样本数	1000
置信区间度	95.0%
置信区间类型	百分位

Bootstrap 指定表包含在重新取样期间使用的设置，并可在检查是否执行了符合您意图的分析时作为有用的参考。

统计量

图片 3-6
带比例 bootstrap 置信区间的统计表

Churn within last month

Customer category			Statistic	Bootstrap[a] 偏差	标准 误差	95% 置信区间 下限	上限
Basic service	N	有效	266	0	0	266	266
		缺失	0	0	0	0	0
	均值		.31	.00	.03	.26	.37
E-service	N	有效	217	0	0	217	217
		缺失	0	0	0	0	0
	均值		.27	.00	.03	.21	.34
Plus service	N	有效	281	0	0	281	281
		缺失	0	0	0	0	0
	均值		.16	.00	.02	.12	.20
Total service	N	有效	236	0	0	236	236
		缺失	0	0	0	0	0
	均值		.37	.00	.03	.31	.44

a. 除非得到提示，否则 bootstrap 结果基于 1000 个 bootstrap 样例

统计表针对每个客户类别级别，显示上月内流失的平均值。由于上月内流失仅取 0 和 1 值，其中 1 代表流失的客户，因此平均值等于流失比例。**统计量**列显示通常由“频率”生成的值，其中使用了原始数据集。**Bootstrap** 列由 bootstrap 算法生成。

- **偏差**是此统计量在 bootstrap 样本中的平均值与**统计量**列中的值之间的差值。在本例中，针对全部 1000 个 bootstrap 样本计算上月内流失的均值，然后再计算这些均值的平均值。
- 标准误是 1000 个 bootstrap 样本中上月内流失均值的标准误。
- 如果以升序排列 1000 个 bootstrap 样本，则 95% bootstrap 置信区间的下限为上月内流失的第 25 与 26 个平均值的插值。其上限为第 975 与 976 个平均值的插值。

表中结果显示流失率因客户类型而异。具体来说，附加服务客户的置信区间未与任何其他区间重叠，说明这些客户在正常情况下不太可能流失。

在使用只有两个值的分类变量时，这些置信区间可以作为由单样本非参数检验过程或单样本 T 检验过程所生成结果的替代。

频率表

图片 3-7
带比例 bootstrap 置信区间的频率表

Customer category			频率	百分比	有效百分比	累积百分比	百分比 Bootstrap[a]			
							偏差	标准 误差	95% 置信区间 下限	95% 置信区间 上限
Basic service	有效	No	183	68.8	68.8	68.8	.0	2.8	63.2	74.4
		Yes	83	31.2	31.2	100.0	.0	2.8	25.6	36.8
		合计	266	100.0	100.0		.0	.0	100.0	100.0
E-service	有效	No	158	72.8	72.8	72.8	.1	3.1	66.4	78.8
		Yes	59	27.2	27.2	100.0	-.1	3.1	21.2	33.6
		合计	217	100.0	100.0		.0	.0	100.0	100.0
Plus service	有效	No	237	84.3	84.3	84.3	.0	2.1	80.1	88.3
		Yes	44	15.7	15.7	100.0	.0	2.1	11.7	19.9
		合计	281	100.0	100.0		.0	.0	100.0	100.0
Total service	有效	No	148	62.7	62.7	62.7	.0	3.2	56.4	69.1
		Yes	88	37.3	37.3	100.0	.0	3.2	30.9	43.6
		合计	236	100.0	100.0		.0	.0	100.0	100.0

a. 除非得到提示，否则 bootstrap 结果基于 1000 个 bootstrap 样例

频率表针对每个类别显示百分比的置信区间（比例 × 100%），因此对所有分类变量可用。在产品的其他部分中未提供相当的置信区间。

使用 Bootstrap 获得中位数的置信区间

在查看员工记录时，管理部门对员工的以往工作经验比较感兴趣。工作经验向右偏斜，这使得平均值在作为员工“典型”以往工作经验的估计方面不如中位数理想。然而，如果不使用 bootstrap，在产品的统计过程中通常无法获得中位数的置信区间。

这些信息收集在 Employee data.sav 中。有关详细信息，请参阅第 28 页码附录 A 中的样本文件。使用 Bootstrap 获得中位数的置信区间。

注意：本示例使用了“探索”过程，并且需要 Statistics Base 选项。

运行分析

▶ 要获得中位数的 bootstrap 置信区间，请在菜单中选择：
分析 > 描述统计 > 探索...

图片 3-8
“探索”主对话框

▶ 选择以往经验（月数）[prevexp] 作为因变量。

▶ 选择“显示”组中的统计量。

▶ 单击 Bootstrap。

图片 3-9
Bootstrap 对话框

- 选择**执行 bootstrap**。
- 为准确复制本例中的结果，请选择**为 Mersenne 扭曲器设置种子**并键入 592004 作为种子。
- 要获得更精确的区间（但需要花费更多处理时间），请选择**偏差修正加速 (BCa)**。
- 单击**继续**。
- 在“探索”对话框中单击 **确定**。

这些选择将生成以下命令语法：

```
PRESERVE.
SET RNG=MT MTINDEX=592004.
SHOW RNG.
BOOTSTRAP
  /SAMPLING METHOD=SIMPLE
  /VARIABLES TARGET=prevexp
  /CRITERIA CILEVEL=95 CITYPE=BCA  NSAMPLES=1000
  /MISSING USERMISSING=EXCLUDE.
EXAMINE VARIABLES=prevexp
  /PLOT NONE
  /STATISTICS DESCRIPTIVES
  /CINTERVAL 95
  /MISSING LISTWISE
  /NOTOTAL.
RESTORE.
```

- PRESERVE 和 RESTORE 命令“记住”随机数生成器的当前状态，并在 bootstrap 结束后将系统恢复到该状态。
- SET 命令将随机数生成器设置成 Mersenne 扭曲器，并将索引设置成 592004，以便准确复制 bootstrap 结果。SHOW 命令在输出中显示索引以供参考。
- BOOTSTRAP 命令使用简单重新取样请求 1000 个 bootstrap 样本。
- VARIABLES 子命令指定使用变量 prevexp 来确定重新取样的个案基础。在此变量上具有缺失值的记录将从分析中删除。
- CRITERIA 子命令除请求 bootstrap 样本数量外，还请求偏差修正和加速 bootstrap 置信区间以替代缺省百分位数区间。
- EXAMINE 过程在 BOOTSTRAP 后，它在每个 bootstrap 样本上运行。
- PLOT 子命令关闭图输出。
- 所有其他选项设置为其缺省值。

描述性

图片 3-10
带 bootstrap 置信区间的描述表

描述

					Bootstrap[a]			
							BCa 95% 置信区间	
			统计量	标准误	偏差	标准误	下限	上限
经验（以月计）	均值		95.86	4.804	-.01	4.86	86.39	105.20
	均值的 95% 置信区间	下限	86.42					
		上限	105.30					
	5% 修整均值		84.64		.02	4.94	75.38	94.21
	中值		55.00		-.11	3.66	50.00	60.00
	方差		10938.281		18.783	977.081	8954.509	13057.229
	标准差		104.586		-.015	4.689	94.644	114.245
	极小值		0					
	极大值		476					
	范围		476					
	四分位距		121		-1	10	103	137
	偏度		1.510	.112	.006	.110	1.284	1.768
	峰度		1.696	.224	.040	.463	.823	2.876

a. 除非得到提示，否则 bootstrap 结果基于 1000 个 bootstrap 样例

描述表包含多个统计量以及这些统计量的 bootstrap 置信区间。平均值的 bootstrap 置信区间 (86.39, 105.20) 与参数置信区间 (86.42, 105.30) 接近，表明“典型”员工具有大概 7-9 年的以往工作经验。然而，以往经验（月数）具有偏斜的分布，这使得平均值在作为“典型”当前工资指标方面不如中位数理想。中位数的 bootstrap 置信区间 (50.00, 60.00) 与平均值置信区间相比值更窄和更低，它表明“典型”员工具有大概 4-5 年的以往工作经验。通过使用 bootstrap，可以获得更能代表典型以往经验的值范围。

使用 Bootstrap 选择更好的预测变量

在查看员工记录时，管理部门有兴趣使用线性模型拟合当前和起始工资之间的差异，以确定哪些因素与员工工资上升存在关联。在 bootstrap 线性模型时，可以使用特殊的重新抽样方法（残差和狂野 bootstrap）以获得更准确的结果。

这些信息收集在 Employee data.sav 中。有关详细信息，请参阅第 28 页码附录 A 中的样本文件。

注意：本示例使用了“GLM 单变量”过程，并且需要 Statistics Base 选项。

准备数据

您必须首先计算当前工资与起始工资的差值。

▶ 从菜单中选择：
转换 > 计算变量...

图片 3-11
“计算变量”对话框

▶ 键入 diff 作为目标变量。

▶ 键入 salary-salbegin 作为数值表达式。

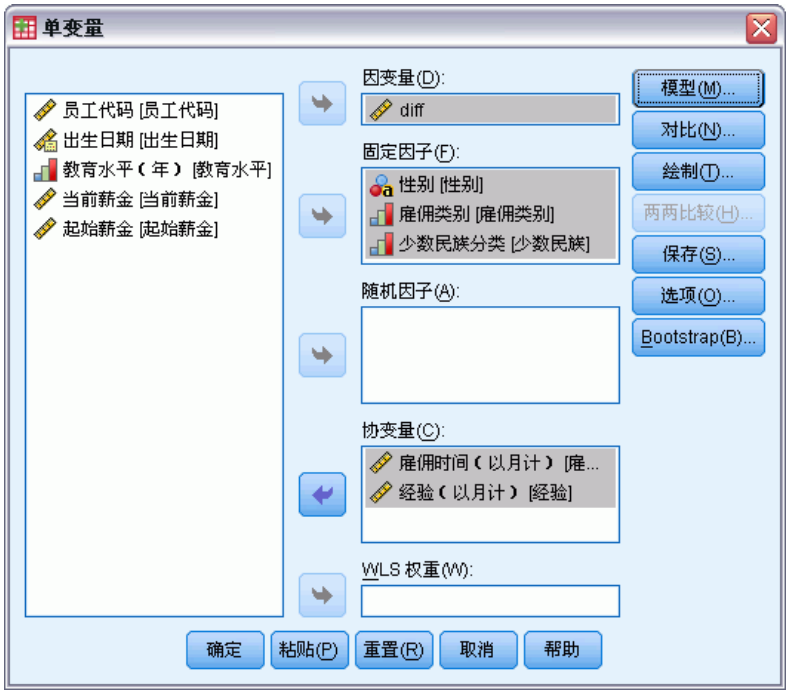
▶ 单击**确定**。

运行分析

要使用狂野残差 bootstrap 运行“GLM 单变量”，首先需要创建残差。

▶ 从菜单中选择：
分析 > **一般线性模型** > **单变量...**

图片 3-12
“GLM 单变量”主对话框



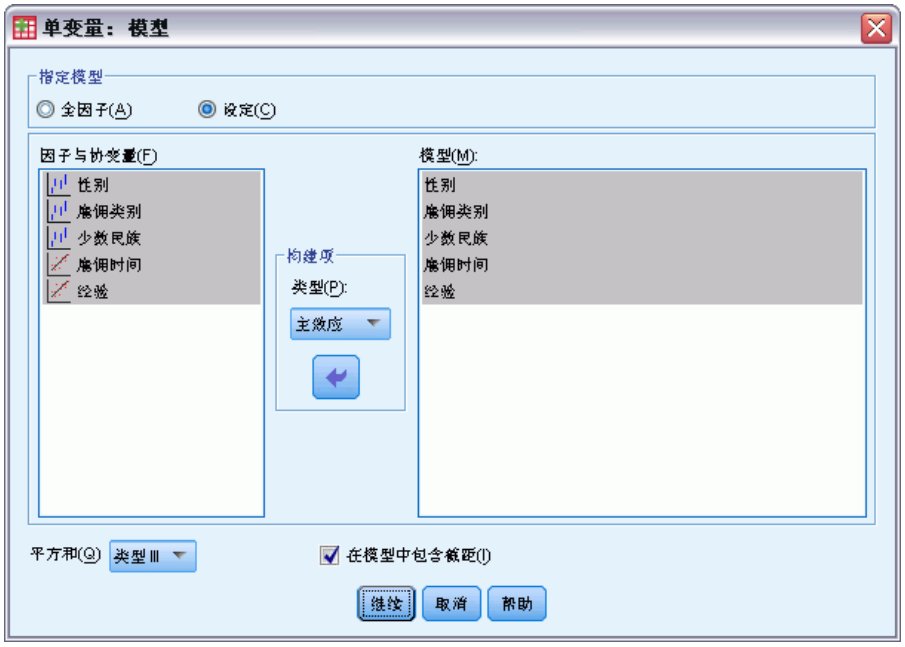
▶ 选择 diff 作为因变量。

▶ 选择性别 [gender]、雇佣类别 [jobcat] 和少数民族分类 [minority] 作为固定因子。

▶ 选择雇用时间 [jobtime] 和以往经验（月数） [prevexp] 作为协变量。

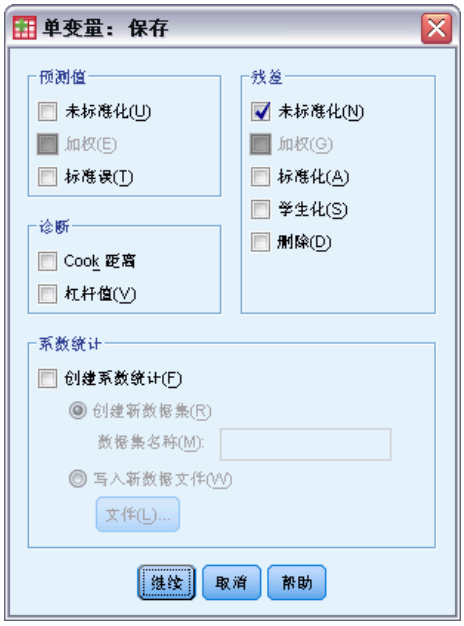
▶ 单击**模型**。

图片 3-13
“模型”对话框



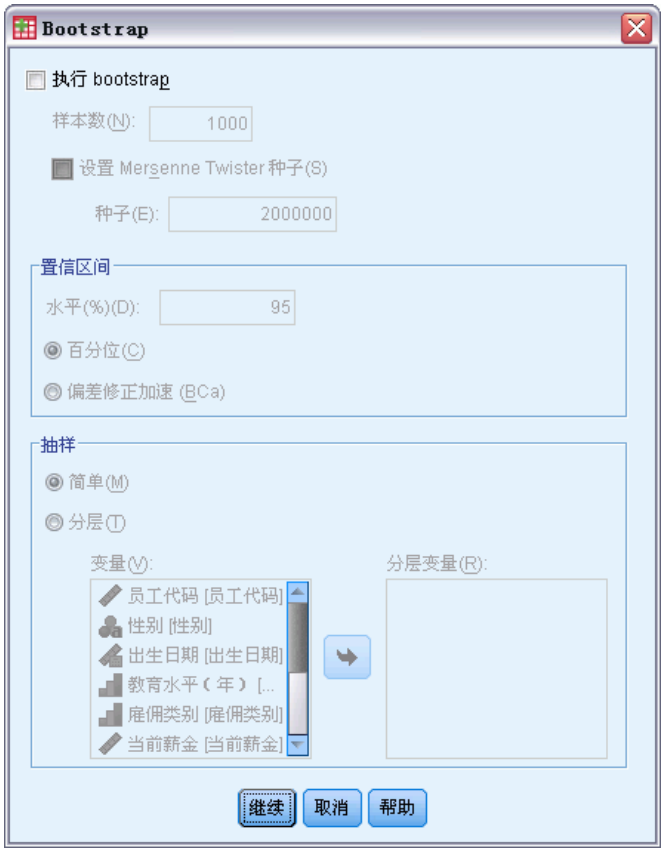
- 选择**定制**，并从“构建项”下拉菜单中选择**主效应**。
- 选择从 gender 到 prevexp 作为模型项。
- 单击**继续**。
- 在“GLM 单变量”对话框中单击**保存**。

图片 3-14
“保存”对话框



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

图片 3-15
Bootstrap 对话框



Bootstrap 设置会在支持 bootstrap 的不同对话框中保留。当 bootstrap 生效时，不支持将新变量保存到数据集，因此您需要确定已关闭 bootstrap 功能。

- 需要时可取消选中**执行 bootstrap**。
- 在“GLM 单变量”对话框中单击**确定**。数据集现在包含新变量 RES_1，其中包含此模型中的未标准化残差。
- 重新调用“GLM 单变量”对话框，并单击**保存**。

▶ 取消选中**未标准化**，然后单击**继续**，并在“GLM 单变量”对话框中单击**选项**。

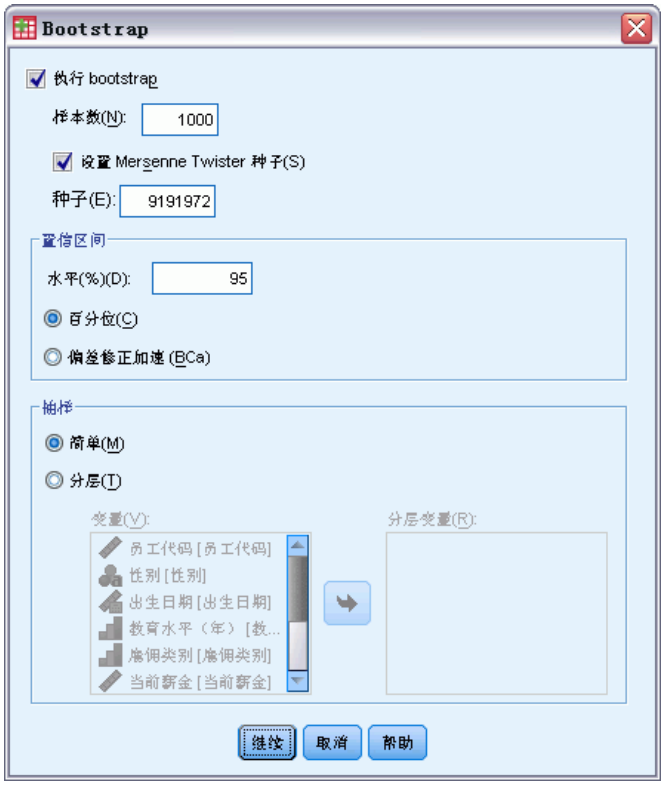
图片 3-16
“选项”对话框

▶ 选择“显示”组中的**参数估计**。

▶ 单击**继续**。

▶ 在“GLM 单变量”对话框中单击 **Bootstrap**。

图片 3-17
Bootstrap 对话框



- 选择**执行 bootstrap**。
- 为准确复制本例中的结果，请选择**为 Mersenne 扭曲器设置种子**并键入 9191972 作为种子。
- 由于不存在通过对话框执行狂野 bootstrap 的选项，因此请单击**继续**，然后在“GLM 单变量”对话框中单击**粘贴**。

这些选择将生成以下命令语法：

```
PRESERVE.
SET RNG=MT MTINDEX=9191972.
SHOW RNG.
BOOTSTRAP
  /SAMPLING METHOD=SIMPLE
  /VARIABLES TARGET=diff INPUT=gender jobcat minority jobtime prevexp
  /CRITERIA CILEVEL=95 CITYPE=PERCENTILE  NSAMPLES=1000
  /MISSING USERMISSING=EXCLUDE.
UNIANOVA diff BY gender jobcat minority WITH jobtime prevexp
  /METHOD=SSTYPE(3)
  /INTERCEPT=INCLUDE
  /PRINT=PARAMETER
  /CRITERIA=ALPHA(.05)
  /DESIGN=gender jobcat minority jobtime prevexp.
RESTORE.
```

为了执行狂野 bootstrap 取样，编辑 SAMPLING 子命令的 METHOD 关键字，使其读取 METHOD=WILD(RESIDUALS=RES_1)。

“最终”命令语法如下所示：

```
PRESERVE.
SET RNG=MT MTINDEX=9191972.
SHOW RNG.
BOOTSTRAP
  /SAMPLING METHOD=WILD(RESIDUALS=RES_1)
  /VARIABLES TARGET=diff INPUT=gender jobcat minority jobtime prevexp
  /CRITERIA CILEVEL=95 CITYPE=PERCENTILE  NSAMPLES=1000
  /MISSING USERMISSING=EXCLUDE.
UNIANOVA diff BY gender jobcat minority WITH jobtime prevexp
  /METHOD=SSTYPE(3)
  /INTERCEPT=INCLUDE
  /PRINT=PARAMETER
  /CRITERIA=ALPHA(.05)
  /DESIGN=gender jobcat minority jobtime prevexp.
RESTORE.
```

- PRESERVE 和 RESTORE 命令“记住”随机数生成器的当前状态，并在 bootstrap 结束后将系统恢复到该状态。
- SET 命令将随机数生成器设置成 Mersenne 扭曲器，并将索引设置成 9191972，以便准确复制 bootstrap 结果。SHOW 命令在输出中显示索引以供参考。
- BOOTSTRAP 命令采用狂野取样方法，并将 RES_1 作为包含残差的变量，请求 1000 个 bootstrap 样本，
- VARIABLES 子命令指定 diff 为线性模型中的目标变量；它与变量 gender、jobcat、minority、jobtime 和 prevexp 用于确定重新取样的个案基础。在这些变量上具有缺失值的记录将从分析中删除。
- CRITERIA 子命令除请求 bootstrap 样本数量外，还请求偏差修正和加速 bootstrap 置信区间以替代缺省百分位数区间。
- UNIANOVA 过程在 BOOTSTRAP 后，它在每个 bootstrap 样本上运行并生成原始数据的参数估计。此外，还为模型系数生成汇聚的统计量。

参数估计值

图片 3-18
参数估计

因变量:diff

参数	B	标准 误差	t	Sig.	95% 置信区间	
					下限	上限
截距	22789.014	2920.700	7.803	.000	17049.673	28528.355
[性别=f]	-4085.253	726.416	-5.624	.000	-5512.701	-2657.804
[性别=m]	0[a]	.	.	.	.	.
[雇佣类别=1]	-17717.706	939.798	-18.853	.000	-19564.463	-15870.949
[雇佣类别=2]	-13101.918	1780.683	-7.358	.000	-16601.061	-9602.776
[雇佣类别=3]	0[a]	.	.	.	.	.
[少数民族=0]	1332.363	819.349	1.626	.105	-277.705	2942.431
[少数民族=1]	0[a]	.	.	.	.	.
雇佣时间	145.539	32.586	4.466	.000	81.505	209.572
经验	-21.423	3.575	-5.993	.000	-28.447	-14.398

a.此参数为冗余参数，将被设为零。

参数估计表显示模型项的常规、非 bootstrap 参数估计。[minority=0] 的显著性值 0.105 大于 0.05，表明少数民族分类对工资增长没有影响。

图片 3-19
Bootstrap 参数估计

因变量:diff

参数	B	Bootstrap[a]				
		偏差	标准 误差	显著性水平（双侧）	95% 置信区间	
					下限	上限
截距	22789.014	-95.084	3280.762	.001	16079.630	28835.063
[性别=f]	-4085.253	32.480	622.971	.001	-5365.321	-2892.131
[性别=m]	0	0	0		0	0
[雇佣类别=1]	-17717.706	46.324	1454.230	.001	-20671.451	-14889.507
[雇佣类别=2]	-13101.918	47.958	1753.311	.001	-16658.596	-9671.891
[雇佣类别=3]	0	0	0		0	0
[少数民族=0]	1332.363	-10.592	651.144	.012	57.831	2642.534
[少数民族=1]	0	0	0		0	0
雇佣时间	145.539	.707	35.285	.001	79.081	217.761
经验	-21.423	-.065	2.859	.001	-27.533	-16.055

a. 除非得到提示，否则 bootstrap 结果基于 1000 个 bootstrap 样例

现在来看 Bootstrap 参数估计表。在“标准误”中，您可以看到某些系数的参数标准误（例如截距）相对于 bootstrap 估计太小，因此置信区间更宽。对于某些系数，例如 [minority=0]，参数标准误太大，因此在 bootstrap 结果中报告的显著性值 0.006 小于 0.05，这表明观察到的少数民族和非少数民族员工之间的工资增长差异并不是偶然的。管理部门现在知道这种差异值得进一步研究以确定可能的原因。

推荐参考

有关 bootstrap 的更多信息，请参见以下内容：

Davison, A. C., 和 D. V. Hinkley. 2006. Bootstrap Methods and their Application. : Cambridge University Press.

Shao, J., 和 D. Tu. 1995. The Jackknife and Bootstrap. New York: Springer.

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。

Copyright IBM Corporation 1989, 2011.

- **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。

Copyright IBM Corporation 1989, 2011.

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.

Davison, A. C., 和 D. V. Hinkley. 2006. Bootstrap Methods and their Application. : Cambridge University Press.

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, .

Shao, J., 和 D. Tu. 1995. The Jackknife and Bootstrap. New York: Springer.

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.

Copyright IBM Corporation 1989, 2011.

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