IBM SPSS Missing Values 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 81 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。Missing Values 可选附加模块提供本手册中描述的其他分析方法。此 Missing Values 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

章 1

缺失值简介

具有缺失值的个案会引发严重的问题，因为典型的建模过程会简单地从分析中丢弃这些个案。如果存在少量缺失值（大约低于个案总数的 5%），且这些值可以被认为随机缺失，即值的缺失不依赖于其他值，则列表删除的典型方法相对比较“安全”。“缺失值”选项可以帮助确定列表删除是否足够，并在必要时提供其他缺失值处理方法。

缺失值分析与多重插补过程

“缺失值”选项提供了两组处理缺失值的过程。

- 多重插补过程提供了缺失数据模式分析，着眼于最终对缺失值进行多重插补。这意味着会产生多个版本的数据集，它们分别包含各自的插补值集。在执行统计分析时，汇集了针对所有插补数据集的参数估计，因此提供的估计结果通常比单个插补更为准确。
- 缺失值分析提供了略微不同的描述性工具集，用以分析缺失数据（尤其是 Little’s MCAR 检验），并包括多种单一插补方法。注意，多重插补通常被认为优于单一插补。

缺失值任务

可以按照这些基本步骤来开始进行缺失值分析：

- **检查缺失情况。**使用“缺失值分析”和“分析模式”探索数据中的缺失值模式，并确定是否有必要进行多重插补。
- **为缺失值规因。**使用“插补缺失数据值”以对缺失值进行多重插补。
- **分析“完整”的数据。**使用任何支持多重插补数据的过程。分析多重插补数据请参见第 24 页码 以了解有关分析多重插补数据集和支持这些数据的过程列表的详细信息。

章 2

缺失值分析

“缺失值分析”过程执行三个主要功能：

- 描述缺失值的模式。缺失值所在位置。其范围。变量对是否往往在多个个案中具有缺失值？日期值是否为极值？值是否为随机缺失？
- 为不同缺失值方法估计均值、标准差、协方差和相关性：列表法、成对法、回归法、或 EM（期望最大化）法。成对法还可显示成对完整个案的计数。
- 使用回归法或 EM 法用估计值填充（插补）缺失值。但多重插补通常被认为可以提供更准确的结果。

缺失值分析有助于解决由不完整的数据造成的若干问题。如果带有缺失值的个案与不带缺失值的个案有着根本的不同，则结果将被误导。此外，缺失的数据还可能降低所计算的统计量的精度，因为计算时的信息比原计划的信息要少。另一个问题是，很多统计过程背后的假设都基于完整的个案，而缺失值可能使所需的理论复杂化。

示例。在评估白血病治疗方式时，将测量几个变量。但是，并不是针对每个患者都进行所有的测量。缺失数据的模式以表格形式显示出来，表现为随机的。EM 分析用于估计均值、相关性和协方差。它还用来确定数据正在随机完全缺失。缺失值然后将由归因值替换，并保存到新的数据文件中以供进一步分析。

统计量。单变量统计量，包括非缺失值个数、均值、标准差、缺失值个数以及极值个数。使用列表法、成对法、EM 法或回归法的估计均值、协方差矩阵以及相关性矩阵。对 EM 结果进行的 Little 的 MCAR 检验。按各种方法进行的均值总计。对于按缺失和非缺失值定义的组：t 检验。对于所有变量：按个案与变量显示的缺失值模式。

数据注意事项

数据。 数据可以是分类数据或定量数据（刻度或连续）。尽管如此，您只能为定量变量估计统计数据并插补缺失数据。对于每个变量，必须将未编码为系统缺失值的缺失值定义为用户缺失值。例如，如果将对问卷项的回答不知道编码为 5，并且您希望将其视为缺失，则对于此项应将 5 编码为用户缺失值。

频率权重。此过程接受频率（复制）权重。忽略复制权重为负值或零值的个案。非整数权重被截断。

假设。 列表法、成对法和回归法估计都基于这样的假设：缺失值的模式不依赖于数据值。（此条件又称为**完全随机缺失**，即 MCAR。）因此，当数据为 MCAR 时，所有估计方法（包括 EM 法）提供相关性和协方差的一致无偏估计。违反 MCAR 假设可能导致由列表法、成对法和回归法生成的有偏差的估计。如果数据不是 MCAR，则您需要使用 EM 估计。

EM 估计依赖于这样的假设：缺失数据的模式仅与观察数据相关。（此条件又称为**随机缺失**，即 MAR。）此假设允许通过可用信息对估计值进行调整。例如，在一项教育与收入的调查中，受教育程度低的对象可能会有更多收入缺失值。在这种情况下，该数据为 MAR，而不是 MCAR。换句话说，就 MAR 而言，收入被记录的概率取决于对象的受教育水平。概率可能因受教育程度而异但不因在教育水平内的收入而异。如果收入被记录的概

率同样因属于每一教育水平的收入而异（例如，高收入人群不报告其收入），则该数据既不是 MCAR 也不是 MAR。这是一种很普遍的情况，且一旦发生，没有一种方法适合。

相关过程。 很多过程都允许您使用列表或成对估计。“线性回归和因子分析”允许用均值替换缺失值。预测附加模块提供了几种方法，可用于按时间序列替换缺失值。

获取缺失值分析

▶ 从菜单中选择：
分析 > 缺失值分析...

图片 2-1
“缺失值分析”对话框

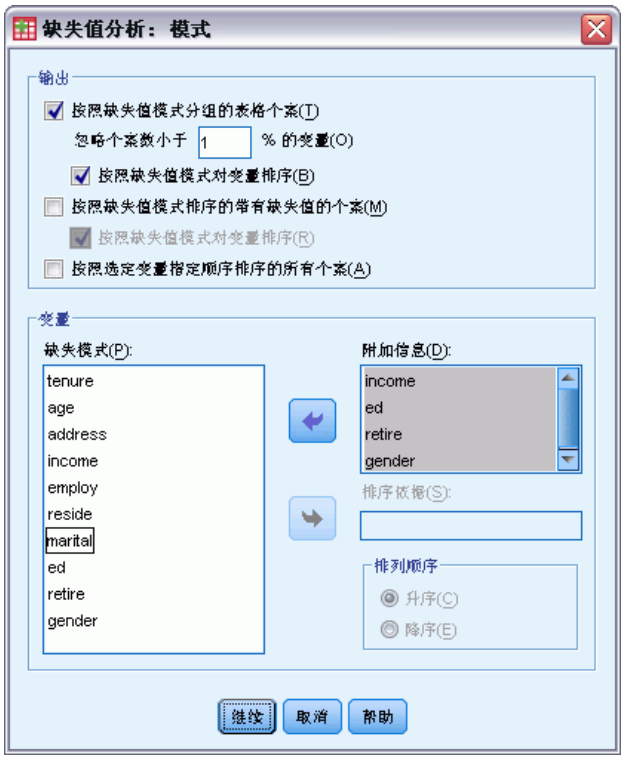
▶ 至少选择一个定量（尺度）变量用于估计统计数据并根据需要插补缺失值。

根据需要，您可以：

- 选择分类变量（数值或字符串）并输入类别个数限制（**最大类别**）。
- 单击**模式**将缺失数据模式制表。有关详细信息，请参阅第 4 页码显示缺失值模式。
- 单击**描述**显示缺失值的描述统计。有关详细信息，请参阅第 5 页码显示缺失值的描述统计。
- 选择一种估计统计（均值、相关性和协方差）和可能插补缺失值的方法。有关详细信息，请参阅第 6 页码估计统计量与插补缺失值。
- 如果选择 EM 或回归法，请单击**变量**以指定将在估计中使用的子集。有关详细信息，请参阅第 9 页码预测的变量与预测变量。
- 选择一个个案标签变量。此变量用于在显示个别个案的模式表格中标注个案。

显示缺失值模式

图片 2-2
“缺失值分析： 模式”对话框



您可以选择显示多种显示缺失数据模式和范围的表格。这些表格能帮助您标识：

- 缺失值位置
- 变量对是否往往在个别个案中具有缺失值
- 数据值是否为极值

输出

可用三种类型的表格显示缺失数据的模式。

制表个案。 分析变量中的缺失值模式，以每种模式中显示的频率被制成表格。使用**按照缺失值模式对变量排序**以指定计数和变量是否按模式相似性排序。使用**省略小于 n % 个案的模式**以删除不经常出现的模式。

具有缺失值的个案。 针对每个分析变量将每一个带有缺失值或极值的个案制表。使用**按照缺失值模式对变量排序**以指定计数和变量是否按模式相似性排序。

全部个案。 对每个个案进行制表且每个变量都被表示为缺失值和极值。如果没指定变量**排序依据**，个案将按其在数据文件中出现的顺序列出。

在显示个别个案的表格中，使用以下符号：

+	极高值
–	极低值
S	系统缺失值
A	用户缺失值的第一种类型
B	用户缺失值的第二种类型
C	用户缺失值的第三种类型

变量

您可以显示分析中所含变量的附加信息。您添加至**附加信息**的变量在缺失模式表格中被逐个显示。对于定量（尺度）变量，显示均值；对于分类变量，显示在每个类别中具有模式的个案数量。

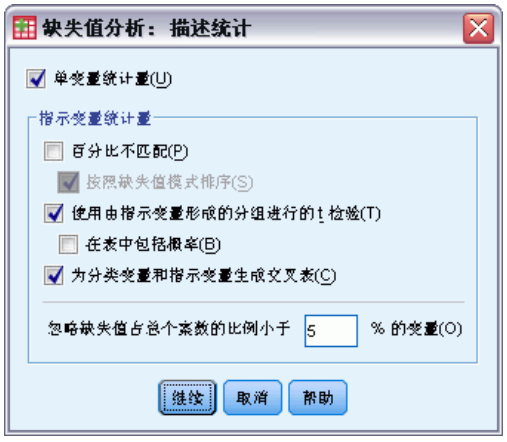
- **排序依据。**个案按照指定变量的值的升序或降序列出。仅适用于**全部个案**的情况。

显示缺失值模式

1. 在缺失值分析主对话框中，选择您想要显示的缺失值模式变量。
2. 单击**模式**。
3. 选择需要显示的模式表格。

显示缺失值的描述统计

图片 2-3
“缺失值分析： 描述”对话框



单变量统计

单变量统计能帮您标识缺失数据的大体范围。对于每个变量，显示以下内容：

- 非缺失值的数量
- 缺失值的数量和百分比

对于定量（尺度）变量，还显示以下内容：

- 均值
- 标准差
- 极高值和极低值的数量

指示变量统计量

对于每个变量，创建一个指示变量。此分类变量指示单个个案的变量存在或缺失。指示变量用于创建不匹配、t 检验与频率表格。

不匹配的百分比。对于每对变量，显示一个变量具有缺失值，另一个变量具有非缺失值的个案数百分比。表中的每个对角元素都包含单个变量具有缺失值的百分比。

使用由指示变量形成的分组进行的 t 检验。 使用 Student t 统计量，比较每个定量变量的两个组的均值。该组指定一个变量存在或缺失。显示两个组的 t 统计量、自由度、缺失和非缺失值计数以及均值。您还可以显示任何与 t 统计量相关的双尾概率。如果您的分析所产生的检验超过一个，则不得将这些概率用于显著性检验。只有当计算单个检验时，此概率才适合。

为分类变量和指示变量生成交叉表。为每个分类变量显示一个表。对于每个类别，该表显示其他变量具有非缺失值的频率和百分比。同时显示每种类型缺失值的百分比。

省略个案数缺失小于 n % 的变量。 为减小表的大小，可以省略仅为少量个案计算的统计量。

显示描述统计

▶ 在缺失值分析主对话框中，选择您想要显示的缺失值描述统计变量。

▶ 单击**描述性**。

▶ 选择需要显示的描述统计。

估计统计量与插补缺失值

您可以使用列表法（仅限完整个案）、成对法、EM（期望最大化）法和/或回归法选择估计均值、标准差、协方差和相关性。您还可以选择插补缺失值（估计替换值）。注意，在解决缺失值问题方面，多重插补通常被认为优于单一插补。Little’s MCAR 检验对于确定是否需要进行插补方面仍然有效。

列表法

此方法仅使用完整个案。一旦任何分析变量具有缺失值，计算中将忽略该个案。

成对法

此方法参见分析变量对，并只有当其在两种变量中都具有非缺失值时才使用个案。频率、均值以及标准差是针对每对分别计算的。由于忽略个案中的其它缺失值，两个变量的相关性与协方差不取决于任何其它变量的缺失值。

EM 法

此方法假设一个部分缺失数据的分布并基于此分布下的可能性进行推论。每个迭代都包括一个 E 步骤和一个 M 步骤。在给定观察值和当前参数估计值的前提下，E 步骤查找“缺失”数据的条件期望值。这些期望值将替换“缺失”数据。在 M 步骤中，即使填写了缺失数据，也将计算参数的最大似然估计值。“缺失”包含在引号中，因为缺失值不是直接填写的。而其函数用于对数似然。

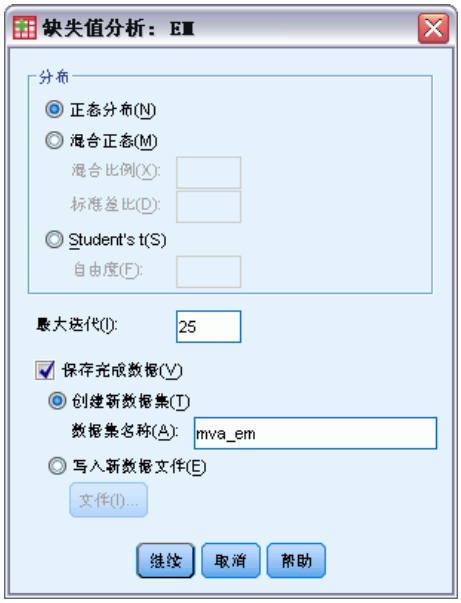
用于检验值是否完全随机丢失（MCAR）的 Roderick J. A. Little 卡方统计量作为 EM 矩阵的脚注印刷。对于此检验，原假设就是数据完全随机缺失且 0.05 水平的 p 值显著。若值小于 0.05，则数据将不会完全随机缺失。数据可能随机缺失（MAR）或不随机缺失（NMAR）。您无法假设一个或其它数据缺失，而是需要分析数据以确定数据是如何缺失的。

回归法

此方法计算多个线性回归估计值并具有用于通过随机元素增加估计值的选项。对于每个预测值，其过程可以从一个随机选择的完整个案中添加一个残差，或者从 t 分布中添加一个随机正态偏差，一个随机偏差（通过残差均值方的平方根测量）。

EM 估计选项

图片 2-4
“缺失值分析： EM”对话框



EM 法使用迭代过程估计具有缺失值的定量（尺度）变量的均值、协方差矩阵及相关性。

分布。 EM 法基于指定分布下的可能性进行推论。默认情况下，假设正态分布。如果您知道分布的尾部比正态分布的尾部要长一些，则您可以要求该过程从自由度为 n 的学生 t 分布中构建似然函数。混合正态分布同样提供具有较长尾部的分布。指定两

个分布的混合正态分布与混合比例的标准偏差比率。混合正态分布假设只有分布标准偏差不同。均值必须相同。

最大迭代次数。 设置最大迭代次数估计真正的协方差。达到此迭代次数后，即使估计值尚未收敛，过程也将停止。

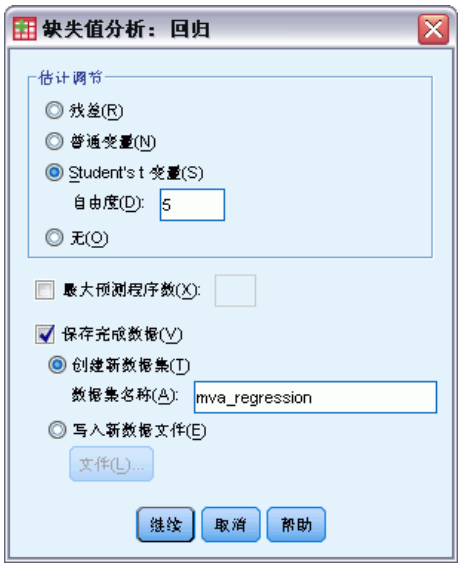
保存已完成的数据。 您可以保存一个有归因值而不是缺失值的数据集。但仍要注意，使用归因值且基于协方差的统计量将会过低估计其各自的参数值。过低估计程度与共同未被观察到的个案数量成比例。

指定 EM 选项

- 在缺失值分析主对话框中，选择您想要使用 EM 法估计的缺失值变量。
- 在“估计”组中选择 EM。
- 要指定预测的变量和预测变量，请单击“变量”。有关详细信息，请参阅第 9 页码预测的变量与预测变量。
- 单击 EM。
- 选择所需 EM 选项。

回归估计选项

图片 2-5
“缺失值分析： 回归”对话框



回归法使用多重线性回归估计缺失值。显示预测变量的均值、协方差矩阵以及相关矩阵。

估计调节。 回归方法可为回归估计添加随机分量。可以选择残差、正态变量、Student t 变量或无调节。

- **残差.** 从要添加到回归估计的完整个案的观察到的残差中，随机选择误差项。

- **正态变量**. 从期望值为 0 且标准差等于回归的均方误差项平方根的分布中，随机抽取误差项。
- **Student t 变量**. 从 t(n) 分布中随机抽取误差项，并按根均方误差 (RMSE) 标度误差项。

最大预测值数目。设置估计过程中使用的预测变量（自变量）的最大数目限制。

保存已完成的数据。将数据集写入当前会话或外部 IBM® SPSS® Statistics 数据文件，将缺失值替换为由回归法估计的值。

指定回归选项

1. 在缺失值分析主对话框中，选择您想要显示的使用回归法估计的缺失值变量。
2. 在“估计”组中选择**回归**。
3. 要指定预测的变量和预测变量，请单击“**变量**”。有关详细信息，请参阅第 9 页码预测的变量与预测变量。
4. 单击**回归**。
5. 选择所需回归选项。

预测的变量与预测变量

图片 2-6
“缺失值分析: EM 和回归的变量”对话框

默认情况下，所有定量变量用于 EM 法和回归法估计。如有需要，您可以选择特定值作为估计中的预测的变量和预测变量。给定变量可存在于两个列表中，但可能会出现您想限制变量使用的情况。例如，有些分析人员不喜欢估计结果变量值。您可能还会想针对不同估计使用不同变量并多次运行过程。例如，如果您有一组护士等级项和一组医生等级项,您可能想使用护士项来运行一次对缺失护士项的估计及再次运行对医生项的估计。

当使用回归法时就会产生另一个考虑。在多重回归中，使用大型自变量子集比使用小型子集生成的预测值要差。因此，变量必须达到 F-to-enter 使用限值 4.0。该限值可通过语法更改。

指定预测的变量和预测变量。

- ▶ 在缺失值分析主对话框中，选择您想要显示的使用回归法估计的缺失值变量。
- ▶ 在“估计”组中选择 **EM** 或**回归**。
- ▶ 单击**变量**。
- ▶ 如果您想使用特定变量而不是全部变量作为预测的变量和预测变量，选择**选择变量**并将变量移至适当列表。

MVA 命令附加功能

使用命令语法语言还可以：

- 使用 MPATTERN、DPATTERN 或 TPATTERN 子命令上的 DESCRIBE 关键字，为缺失值模式、数据模式和制表模式分别指定不同的描述变量。
- 使用 DPATTERN 子命令为数据模式表指定多个排序变量。
- 使用 DPATTERN 子命令为数据模式指定多个排序变量。
- 使用 EM 子命令指定容差和收敛方式。
- 使用 REGRESSION 子命令指定容差和 F-to-enter。
- 使用 EM 和 REGRESSION 子命令，为 EM 和回归指定不同的变量列表。
- 为每个 TTESTS、TABULATE 和 MISMATCH 指定取消显示的个案的不同百分比。

请参阅命令语法参考以获取完整的语法信息。

章 3

多重归因

多重插补的目的是为缺失值生成可能的值，因而创建一些“完整”的数据集。多重插补数据集对应的分析过程为每个“完整”数据集生成输出，并生成包含当原始数据集无缺失值时的结果估计的汇聚输出。这些汇聚结果通常比单一插补方法所提供的结果更准确。

分析变量。分析变量可为：

- **标定.** 当变量值表示不具有内在等级的类别时，该变量可以作为名义变量；例如，雇员任职的公司部门。名义变量的示例包括地区、邮政编码和宗教信仰。
- **有序.** 当变量值表示带有某种内在等级的类别时，该变量可以作为有序变量；例如，从十分不满意到十分满意的服务满意度水平。有序变量的示例包括表示满意度或可信度的态度分数和优先选择评分。
- **刻度.** 当变量值表示带有有意义的度规的已排序类别时，该变量可以作为刻度（连续）变量对待，以便在值之间进行合适的距离比较。刻度变量的示例包括以年为单位的年龄和以千美元为单位的收入。

  该过程假设已经将适当的测量级别分配给所有变量，但您可以通过在源变量列表中右键单击该变量并从上下文菜单中选择测量级别暂时更改变量的测量级别。

变量列表中每个变量旁的图标标识测量级别和数据类型：

	数值	字符串	日期	时间
尺度（连续）		n/a		
有序				
名义				

频率权重。此过程接受频率（复制）权重。忽略复制权重为负值或零值的个案。非整数权重被四舍五入为最接近的整数。

分析权重。分析（回归或抽样）权重被包含进缺失值摘要和拟合插补模型中。排除分析权重为负值或零值的个案。

复杂样本。 多重插补过程不显式处理层次、聚类或其他复杂抽样结构，尽管可以接受以分析权重变量形式的最终抽样权重。同时注意“复杂抽样”过程目前不自动分析多重插补数据集。有关支持汇聚的过程完整列表，请参见 分析多重插补数据第 24 页码。

缺失值。用户缺失值和系统缺失值视为无效值；即两种缺失值在插补值时被替换，且两种缺失值被视为插补模型中用作预测变量的无效值。用户缺失值和系统缺失值在缺失值分析中也被视为缺失。

复制结果（插补缺失数据值）。如果您想准确复制您的插补结果，除了使用相同过程设置以外，还可以使用针对随机数字生成器的相同初始化值、相同数据顺序和相同变量顺序。

- **随机数字生成器。**该过程在插补值计算期间使用随机数字生成器。想要以后再次生成相同的随机结果，在每次运行“插补缺失数据值”过程之前使用随机数字生成器的相同初始化值。
- **个案顺序。**以个案顺序插补值。
- **变量顺序。**完全条件指定（FCS）插补方法以“分析变量”列表中所指定的顺序插补值。

有两个对话框专门用于多重插补。

- 分析模式提供数据中缺失值模式的描述性测量，可用作插补之前的探索步骤。
- 插补缺失数据值用于产生多重插补。可使用支持多重插补数据集的过程分析完整数据集。请参见分析多重插补数据第 24 页码以了解有关分析多重插补数据集和支持这些数据的过程列表的详细信息。

分析模式

分析模式提供数据中缺失值模式的描述性测量，可用作插补之前的探索步骤。

示例。电信供应商想更好理解客户数据库中的服务用途模式。他们拥有客户所使用的服务的完整数据，但是公司收集的人口统计信息有大量缺失值。分析缺失值的模式可以帮助确定插补的下一步。有关详细信息，请参阅第 42 页码第 5 章中的使用多重插补完成并分析数据集。

分析缺失数据模式

从菜单中选择：

分析 > **多重归因** > **分析模式**...

图片 3-1
“分析模式”对话框

▶ 请选择至少两个分析变量。该过程分析这些变量的缺失数据的模式。

可选设置

分析权重。此变量包含分析（回归或抽样）权重。此过程在缺失数据概要中融入分析权重。排除分析权重为负值或零值的个案。

输出。显示下面的可选输出：

- **缺失值概要。**它会显示带面板的饼图，图中显示具有一个或多个缺失值的分析变量、个案或单独数据值的数量及百分比。
- **缺失值模式。**它会显示缺失值的制表模式。每个模式对应于分析变量上具有相同的不完整和完整数据模式的一组个案。您可以使用此输出判断该单调插补方法是否可用于您的数据，如果不能，判断您的数据近似单调模式的程度。该过程对分析变量排序，以揭示或近似单调模式。如果重新排序后不存在非单调模式，则您可以得出结论，如此排序分析变量时数据具有单调模式。
- **缺失值频率最高的变量。**它会按缺失值百分比的降序顺序显示一个分析变量表格。该表格包括刻度变量的描述性统计数据（均值和标准差）。

  您可以控制显示变量的最大数量和显示中可包括的变量的最小缺失百分比。将显示满足两个条件的变量集合。例如，设置最大变量数量为 50 和最小缺失百分比为 25 会要求表格显示缺失值至少为 25% 的最多 50 个变量。如果有 60 个分析变量，但只有其中 15 个的缺失值大于或等于 25%，则输出只包括 15 个变量。

插补缺失数据值

插补缺失数据值用于产生多重插补。可使用支持多重插补数据集的过程分析完整数据集。请参见分析多重插补数据第 24 页码 以了解有关分析多重插补数据集和支持这些数据的过程列表的详细信息。

示例。电信供应商想更好理解客户数据库中的服务用途模式。他们拥有客户所使用的服务的完整数据，但是公司收集的人口统计信息有大量缺失值。此外，这些值并未随机完全缺失，因此多重插补将用于完成数据集。有关详细信息，请参阅第 42 页码第 5 章中的使用多重插补完成并分析数据集。

插补缺失数据值

从菜单中选择：

分析 > 多重归因 > 归因缺失数据值...

图片 3-2
“插补缺失数据值变量”选项卡

- 在插补模型中选择至少两个变量。该过程插补这些变量缺失数据的多个值。
- 指定要计算插补的数量。默认情况下，该值为 5。

▶ 指定写入插补数据的数据集或IBM® SPSS® Statistics格式数据文件。

输出数据集由带有缺失数据的原始数据和带有每次插补归因值的一组个案组成。例如，如果原始数据集有 100 个个案并且您有五个插补，那么输出数据集将有 600 个个案。输入数据集中的所有变量被包括在输出数据集中。现有变量的字典属性（名称、标签等）被复制到新数据集。文件也包含一个新变量 Imputation_，它是一个指示插补的数值变量（原始数据为 0，或具有插补值的个案为 1..n）。

当创建输出数据集时，过程自动定义 Imputation_ 变量为拆分变量。如果过程执行时拆分生效，则输出数据集包括拆分变量值每个组合的一个插补集合。

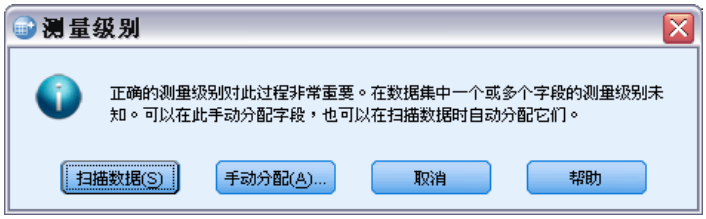
可选设置

分析权重。此变量包含分析（回归或抽样）权重。该过程在用于插补缺失值的回归和分类模型中融入了分析权重。分析权重也用在插补值概要中；例如均值、标准差和标准误差。排除分析权重为负值或零值的个案。

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 3-3
测量级别警报

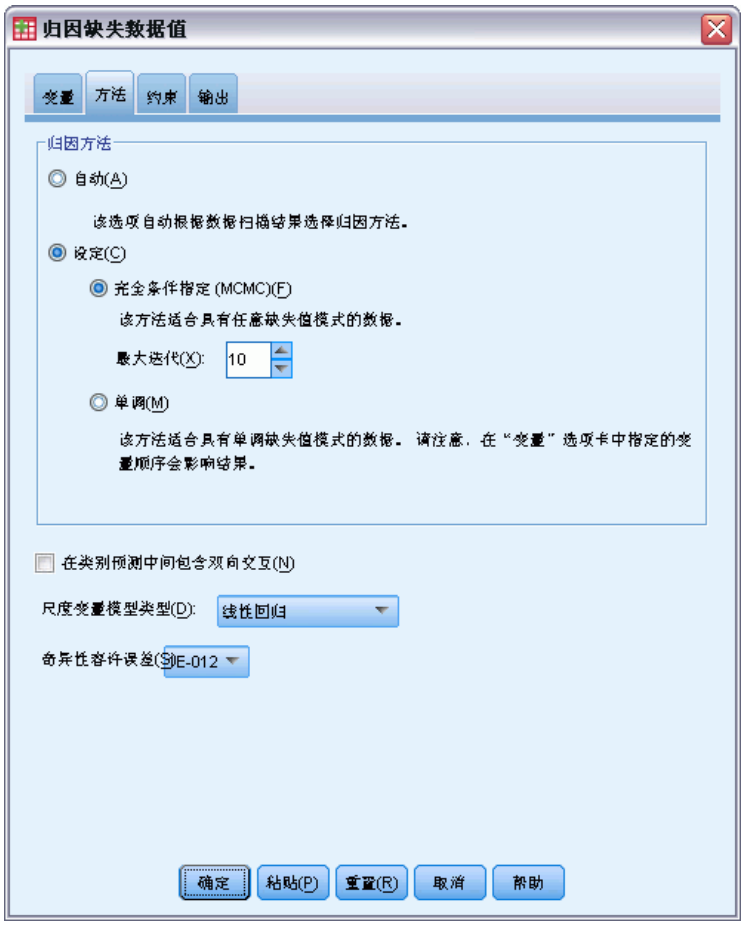


- **扫描数据。** 读取活动数据集中的数据，并分配默认测量级别给任何具有当前未知测量级别的字段。如果数据集较大，该过程可能需要一些时间。
- **手动分配。** 打开列出了所有具有未知测量级别的字段的对话框。您可以使用该对话框将测量级别分配给这些字段。您也可以在数据编辑器的变量视图中分配测量级别。

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。

方法

图片 3-4
“插补缺失数据值方法”选项卡



“方法”选项卡指定如何插补缺失值，包括使用的模型类型。分类预测值是指示符（哑元）编码。

插补方法。自动方法扫描数据，并在数据显示单调缺失值模式时使用单调方法；否则使用完全条件指定。如果您确定应使用的方法，您可以将其指定为**自定义**方法。

- **完全条件指定。**这是一个迭代 Markov 链 Monte Carlo (MCMC) 方法，当缺失数据模式任意（单调或非单调）时可使用该方法。

  对于每个迭代以及对于以变量列表中指定顺序的每个变量，完全条件指定（FCS）方法使用模型中的所有其他可用变量作为预测值，拟合一个单变量（单个因变量）模型，然后为拟合的变量插补缺失值。此方法持续执行，直到达到最大迭代次数，最大迭代的插补值保存到插补数据集中。

  最大迭代次数。它规定 FCS 方法所使用的 Markov 链进行的迭代（或“步骤”）数量。如果自动选择 FCS 方法，则它使用默认 10 次迭代次数。当您明确选择 FCS 时，您可以指定自定义迭代次数。如果 Markov 链不收敛，您可能需要增加

迭代次数。在“输出”选项卡上，您可以保存 FCS 迭代历史记录数据并将其画成曲线，以评估收敛。

- **单调。**这是一种非迭代方法，只有当数据具有单调缺失值模式时才可使用该方法。当您可以排序变量使得（如果变量具有非缺失值）所有先前变量也具有非缺失值时，就表示存在单调模式。当将此指定为**定制**方法时，确保以显示单调模式的顺序指定列表中的变量。

  对于单调顺序的每个变量，单调方法使用模型中的所有前面的变量作为预测值，拟合一个单变量（单个因变量）模型，然后为拟合的变量插补缺失值。这些插补值保存到插补数据集中。

包括二阶交互。当自动选择插补方法时，每个变量的插补模型包括预测变量的常数项和主效应。当选择特定方法时，您也可以在分类预测变量中包括所有可能的二阶交互。

刻度变量的模型类型。 当自动选择插补方法时，线性回归用作刻度变量的单变量模型。当选择特定方法时，您也可以选择预测均值匹配（PMM）作为刻度变量的模型。PMM 是线性回归的一种变型，它将回归模型计算得出的插补值与最接近的观察值匹配。

Logistic 回归总是用作分类变量的单变量模型。无论是哪种模型类型，都使用指示符（哑元）编码处理分类预测值。

奇异性容许误差。奇异（非可逆）矩阵具有线性相关列，对估计算法可能产生严重问题。即使近似奇异的矩阵也可导致不良结果，因此该过程会将行列式小于容许误差的矩阵作为奇异矩阵对待。指定一个正值。

约束

图片 3-5
“插补缺失数据值约束”选项卡

“约束”选项卡能限制插补过程中变量的角色，以及限制刻度变量插补值范围，使其似是而非。此外，您可以将分析限制为具有小于缺失值的最大百分比的变量。

变量概要数据的扫描。单击**扫描数据**使列表显示分析变量和每个分析变量观察到的缺失百分比、最小百分比和最大百分比。概要可基于所有个案或在“个案”文本框中指定限制为前 n 个个案的扫描。单击**重新扫描数据**更新分布概要。

定义约束

- **角色。**它能自定义应插补的和/或视为预测变量的变量集合。通常，每个分析变量被视为插补模型中的因变量和预测值。**角色**可用于关闭您希望**只用作预测变量**的变量的插补，或将变量排除用作预测值（**仅插补**），并因此使预测模型更紧凑。这是可为分类变量或为只用作预测变量的变量指定的唯一约束。

- **最小和最大值**这些列能指定刻度变量的最小和最大允许插补值。如果插补值在此范围之外，过程会抽取另一数值直至其在抽取最大数量范围内找到一个（请参见下文**最大抽取**）。只有在“方法”选项卡上当**线性回归**被选为刻度变量模型类型时这些列才可用。
- **四舍五入。** 一些变量可以用作刻度变量，但是拥有自然进一步限制的值；例如，一家的人数必须是整数，在杂货店消费的金额不能有分数的分。此列允许您指定要接受的最小命名。例如，要获得整数值，您应指定 1 作为舍入命名；要获得舍入到最接近的分的值，您应指定 0.01。通常，值被四舍五入为舍入命名的最近整数倍数。下表显示对归因值 64823 的不同四舍五入值（舍入之前）。

舍入命名	四舍五入 6.64832 的值
10	10
1	7
0.25	6.75
0.1	6.6
0.01	6.65

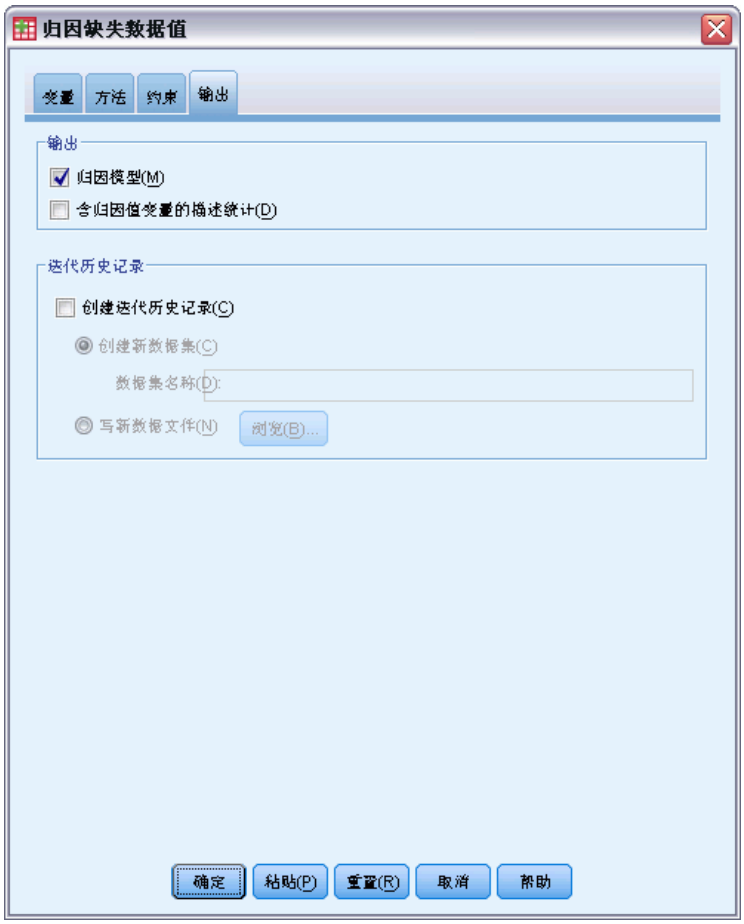
排除具有大量缺失数据的变量。通常，如果分析变量具有估测插补模型的足够数据，则插补分析变量且将其用作预测变量，与缺失值数量无关。您可以选择排除缺失值百分比较高的变量。例如，如果您指定 50 为**最大缺失百分比**，则缺失值超过 50% 的分析变量不会被插补，也不会被用作插补模型中的预测变量。

最大抽取。如果为刻度变量的插补值指定最小或最大值（请参见上文的**最小和最大值**），则过程尝试抽取个案的值，直至其找到指定范围内值的集合。如果在每个个案的指定抽取数量内未找到值的集合，则过程抽取模型参数的另一个集合，并重复个案抽取过程。如果在个案的指定数量和参数抽取内未找到值的集合，则发生错误。

请注意，增加这些数值会增加处理时间。如果过程时间较长，或无法找到适当抽取，则检查所指定的最小和最大值，确保值大小适当。

输出

图片 3-6
“插补缺失数据值输出”选项卡



显示。 控制输出的显示。总是显示整体插补概要，它包括与插补指定、迭代次数（完全条件指定方法）、插补因变量、排除插补的因变量和插补序列相关的表格。如果指定显示分析变量的约束，则也会显示该约束。

- **插补模型。**它显示因变量和预测变量的插补模型，且包括单变量模型类型、模型效应和插补值数量。
- **描述统计。**它显示对其插补值的因变量的描述统计。对于刻度变量，描述统计包括原始输入数据（插补之前）的均值、计数、标准差、最小和最大值及完整数据（插补在一起的—原始数据和插补值）。对于分类变量，描述统计包括原始输入数据（插补之前）的计数和百分比（按分类）、插补值（按插补）和完整数据（插补在一起的—原始数据和插补值）。

迭代历史记录。当使用完全条件指定插补方法时，您可以请求包含 FCS 插补迭代历史记录数据的数据集。该数据集包含对其插补值的每个刻度因变量的均值、标准差（按迭代）和插补。您可以将数据画成曲线，以帮助评估模型收敛。有关详细信息，请参阅第 60 页码第 5 章中的检查 FCS 收敛。

MULTIPLE IMPUTATION 命令附加功能

使用命令语法语言还可以：

- 指定为其显示描述统计的变量子集（IMPUTATIONSUMMARIES 子命令）。
- 指定在过程的一次运行中缺失模式的分析和插补。
- 当插补任何变量（MAXMODELPARAM 关键字）时指定允许的模型参数最大数。

请参阅命令语法参考以获取完整的语法信息。

使用多重插补数据

当创建多重插补（MI）数据集时，添加一个称为 Imputation_ 的带有变量标签 Imputation Number 的变量，并按其升序方式排序数据集。来自原始数据集的个案值为 0，插补值个案是从 1 到 M，其中 M 是插补数量。

当您打开数据集时，存在 Imputation_ 将数据集标识为可能的 MI 数据集。

激活用于分析的多重插补数据集

数据集必须使用**比较各组**选项进行拆分，其中 Imputation_ 为分组变量，以被视为分析中的 MI 数据集。您也可以在其他变量上定义拆分。

从菜单中选择：

数据 > **拆分文件...**

图片 3-7
“拆分文件”对话框

▶ 选择**比较各组**。

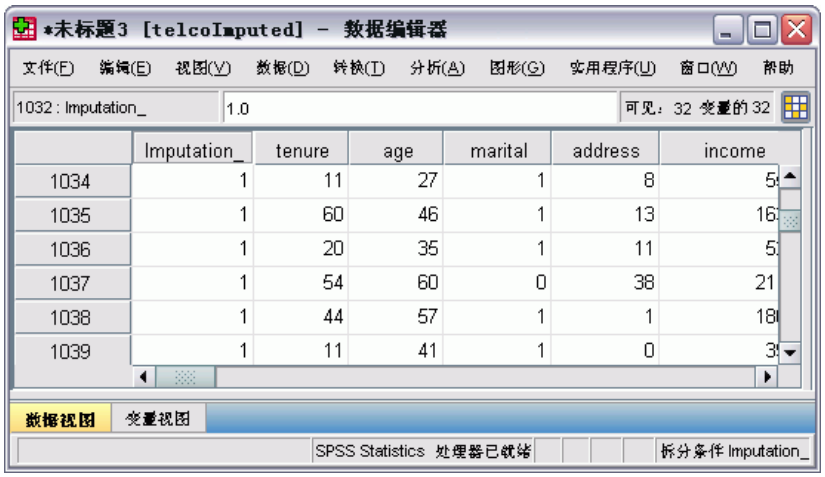
▶ 选择 Imputation Number [Imputation_] 作为对个案分组的变量。

或者，当您打开标记（见下文）时，拆分 Imputation Number [Imputation_] 上的文件。

区分插补值与观察值。

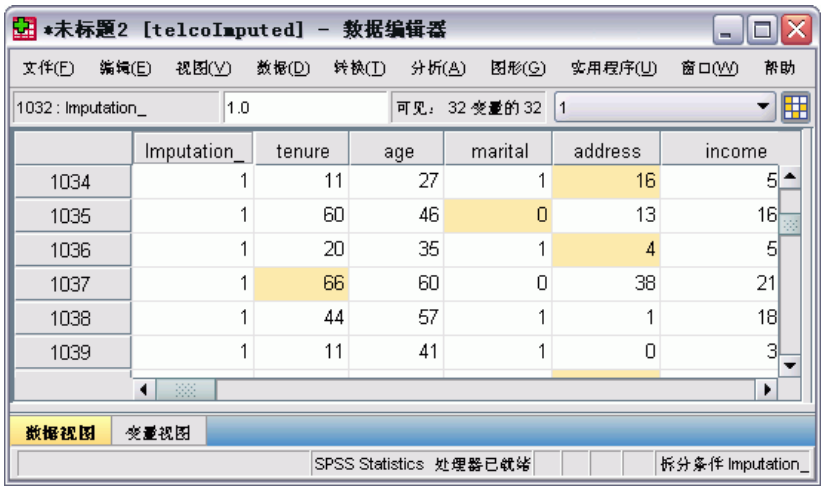
您可以通过单元背景色、字体和加粗类型（用于插补值）来区分插补值与观察值。有关生效标记的详细信息，请参见多重插补选项第 28 页码。当您在当前会话中使用插补缺失值创建一个新数据集，默认情况下打开标记。当您打开一个包括插补的已保存的数据文件时，关闭标记。

图片 3-8
“数据编辑器”（插补标记关闭）



要打开标记，请从“数据编辑器”菜单中选择：
视图 > 标记插补数据...

图片 3-9
“数据编辑器”（插补标记打开）



也可以单击“数据编辑器”的数据视图上编辑条右边的插补标记按钮来打开标记。

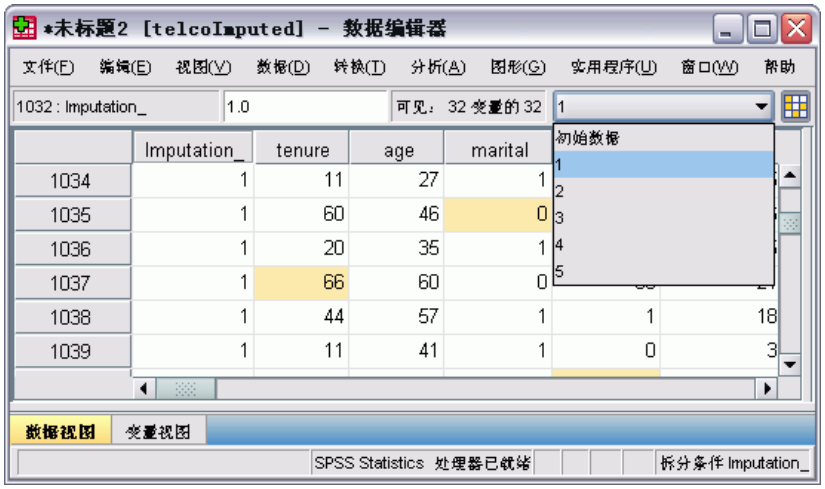
在插补之间移动

- 从菜单中选择：
  编辑 > 转至归因…
- 从下拉列表中选择插补（或原始数据）。

图片 3-10
“转到”对话框

或者，也可从“数据编辑器”的数据视图上编辑栏的下拉列表中选择插补。

图片 3-11
“数据编辑器”（插补标记打开）



在选择插补时相对个案位置被保留。例如，如果在源数据集中有 1000 个个案，个案 1034，即第 1 个插补中的第 34 个个案，显示在网格顶部。如果您从下拉列表中选择插补 2，则个案 2034，即第 2 个插补中的第 34 个个案，将显示在网格顶部。如果您从下拉列表中选择**原始数据**，则个案 34 将显示在网格顶部。在两个插补之间浏览时，列位置也将被保留，这样可以方便地进行值比较。

转换和编辑插补值

有时候您需要对插补数据执行转换。例如，您可能想要将薪金变量的所有值记入日志，并将结果保存在新变量中。如果使用插补数据计算得出的值与使用原始数据计算得出的值不同，则其将被视作被插补的值。

如果您在“数据编辑器”单元中的编辑插补值，则该单元仍将被视作被插补的。不建议以此方式编辑插补值。

分析多重插补数据

许多过程支持多重插补数据集分析结果的汇聚。当打开插补标记时，会在支持汇聚的过程旁边显示一个特殊的图标。在“分析”菜单的“描述统计”子菜单中，例如“频率”、“描述”、“探索”和“交叉表”都支持汇聚，而“比率”、“P-P 图”和“Q-Q 图”却不支持。

图片 3-12
“分析”菜单（插补标记打开）

可以汇聚表格输出和模型 PMML。请求汇聚输出没有新的过程；在“选项”对话框上一个新的选项卡能对多重插补输出进行全局控制。

- **表格输出的汇聚。**默认情况下，当您对多重插补（MI）数据集运行所支持的过程时，会自动为每个插补、原始（未插补）数据和考虑到插补中的偏差的汇聚后（最终）结果生成结果。各个过程中汇聚的统计量各有不同。
- **PMML 的汇聚。**您也可以从导出 PMML 且支持的过程获得汇聚后 PMML。汇聚 PMML 以与非汇聚 PMML 相同的方式进行请求，不同之处在于它可以被保存。

不支持的过程不会产生汇聚后输出，也不产生汇聚后 PMML 文件。

汇聚水平

使用以下两种水平的其中一种汇聚输出：

- **Naïve 组合。**只有汇聚参数可用。
- **单变量组合。**汇聚参数、其标准误、检验统计和有效自由度、p 值，置信区间和汇聚诊断（部分缺失信息、相对有效性、相对偏差增加）可用时会显示。

通常汇聚系数（回归和相关性）、均值（均值差）和计数。当统计量的标准误可用时，则使用单变量汇聚，否则使用 naïve 汇聚。

支持汇聚的过程

以下过程在为每个输出所指定的汇聚水平上支持 MI 数据集。

频率

- “统计”表格支持“单变量”汇聚（如果也需要均值的标准误）时的“均值”和 Naïve 汇聚时的“有效数量”和“缺失数量”。
- “频率”表格支持 Naïve 汇聚时的“频率”。

描述性

- “描述统计”表格支持“单变量”汇聚（如果也需要均值的标准误）时的“均值”和 Naïve 汇聚时的“N”。

交叉表

- “交叉制表”表格支持 Naïve 汇聚时的“计数”。

均值

- “报告”表格支持“单变量”汇聚（如果也需要均值的标准误）时的“均值”和 Naïve 汇聚时的“N”。

单样本 T 检验

- “统计”表格支持“单变量”汇聚和 Naïve 汇聚时的“均值”。
- “检验”表格支持“单变量”汇聚时的“均值差”。

独立样本 T 检验

- “组统计”表格支持“单变量”汇聚和 Naïve 汇聚时的“均值”。
- “检验”表格支持“单变量”汇聚时的“均值差”。

配对样本 T 检验

- “统计”表格支持“单变量”汇聚和 Naïve 汇聚时的“均值”。
- “相关性”表格支持 Naïve 汇聚时的“相关性”和“N”。
- “检验”表格支持“单变量”汇聚时的“均值”。

单因素方差分析

- “描述统计”表格支持“单变量”汇聚和 Naïve 汇聚时的“均值”。
- “对比检验”表格支持“单变量”汇聚时的“对比值”。

线性混合模型

- “描述统计”表格支持 Naïve 汇聚时的“均值”和“N”。
- “固定效应估计值”表格支持“单变量”汇聚时的“估计值”。
- “协方差参数估计值”表格支持“单变量”汇聚时的“估计值”。
- 估计边际均值：“估算值”表格支持“单变量”汇聚时的“均值”。
- 估计边际均值：“成对比较”表格支持“单变量”汇聚时的“均值差”。

“广义线性模型”和“广义估计方程”。 这些过程支持汇聚 PMML。

- “分类变量信息”表格支持 Naïve 汇聚时的“N”和“百分比”。
- “连续变量信息”表格支持 Naïve 汇聚时的“N”和“均值”。
- “参数估计值”表格支持“单变量”汇聚时的系数、B。
- 估计边际均值：“估计系数”表格支持 Naïve 汇聚时的“均值”。
- 估计边际均值：“估算值”表格支持“单变量”汇聚时的“均值”。
- 估计边际均值：“成对比较”表格支持“单变量”汇聚时的“均值差”。

双变量相关

- “描述统计”表格支持 Naïve 汇聚时的“均值”和“N”。
- “相关性”表格支持单变量汇聚时的“相关性”和“N”。注意，在汇聚之前使用 Fisher 的 z 转换来转换相关性，并在汇聚之后执行逆转换。

偏相关

- “描述统计”表格支持 Naïve 汇聚时的“均值”和“N”。
- “相关性”表格支持 Naïve 汇聚时的“相关性”。

线性回归。 此过程支持汇聚 PMML。

- “描述统计”表格支持 Naïve 汇聚时的“均值”和“N”。
- “相关性”表格支持 Naïve 汇聚时的“相关性”和“N”。
- “系数”表格支持“单变量”汇聚时的“B”和 Naïve 汇聚时的“相关性”。
- “相关系数”表格支持 Naïve 汇聚时的“相关性”。
- “残差统计”表格支持 Naïve 汇聚时的“均值”和“N”。

二元 Logistic 回归。 此过程支持汇聚 PMML。

- “方程中变量”支持“单变量”汇聚时的“B”。

多项 Logistic 回归。 此过程支持汇聚 PMML。

- “参数估计值”表格支持“单变量”汇聚时的系数、B。

Ordinal 回归

- “参数估计值”表格支持“单变量”汇聚时的系数、B。

判别分析。 此过程支持汇聚模型 XML。

- “组统计”表格支持 Naïve 汇聚时的“均值”和“有效 N”。
- “汇聚组内矩阵”表格支持 Naïve 汇聚时的“相关性”。
- “典型判别函数系数”表格支持 Naïve 汇聚时的“未标准化系数”。
- “组质心函数”表格支持 Naïve 汇聚时的“未标准化系数”。
- “分类函数系数”表格支持 Naïve 汇聚时的“系数”。

卡方检验

- “描述”表格支持 Naïve 汇聚时的“均值”和“N”。
- “频率”表格支持 Naïve 汇聚时的“观察 N”。

二项式检验

- “描述”表格支持 Naïve 汇聚时的“均值”和“N”。
- “检验”表格支持 Naïve 汇聚时的“N”、“观察到的比例”和“检验比例”。

游程检验

- “描述”表格支持 Naïve 汇聚时的“均值”和“N”。

单样本 Kolmogorov-Smirnov 检验

- “描述”表格支持 Naïve 汇聚时的“均值”和“N”。

两个独立样本检验

- “秩数”表格支持 Naïve 汇聚时的“秩均值”和“N”。
- “频率”表格支持 Naïve 汇聚时的“N”。

多个独立样本检验

- “秩数”表格支持 Naïve 汇聚时的“秩均值”和“N”。
- “频率”表格支持 Naïve 汇聚时的“计数”。

两个关联样本检验

- “秩数”表格支持 Naïve 汇聚时的“秩均值”和“N”。
- “频率”表格支持 Naïve 汇聚时的“N”。

多个关联样本检验

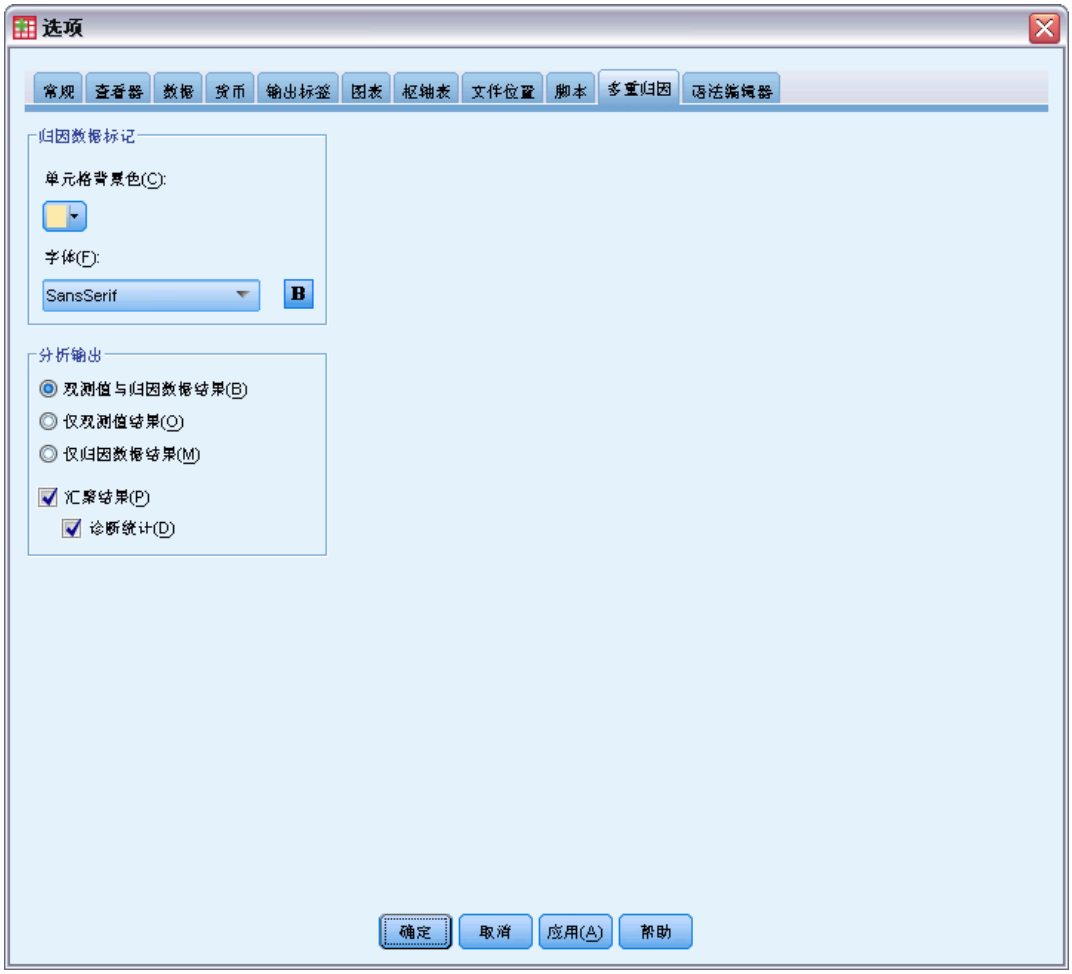
- “秩数”表格支持 Naïve 汇聚时的“秩均值”。

Cox 回归。 此过程支持汇聚 PMML。

- “方程中变量”支持“单变量”汇聚时的“B”。
- “协变量均值”表格支持 Naïve 汇聚时的“均值”。

多重插补选项

图片 3-13
“选项”对话框：“多重插补”选项卡



“多重插补”选项卡控制与多重插补相关的两类首选项：

插补数据外观。缺省情况下，包含插补数据的单元格与包含非插补数据的单元格具有不同的背景颜色。插补数据的直观显示有助于您滚动数据集并找到这些单元格。还可以更改缺省单元格背景颜色、字体，以及使插补数据粗体显示等。

分析输出。这组首选项控制在分析多重插补数据集时产生的浏览器输出类型。缺省情况下，将为每个原始（插补前）数据集和插补数据集产生输出。此外，还为那些支持插补数据汇聚的过程生成最终的汇聚结果。当执行单变量汇聚时，还会显示汇聚诊断结果。不过，您可以隐藏那些不愿看到的输出。

设置多重插补选项

从菜单中选择：
编辑 > 选项

单击“多重插补”选项卡。

部分 II:
示例

章
4

缺失值分析

描述缺失值的模式

电信供应商想更好理解客户数据库中的服务用途模式。公司想确保数据在运行进一步分析之前随机完全缺失。

客户数据库中的随机样本包含在 telco_missing.sav。 有关详细信息，请参阅附录 A 中的样本文件中的IBM SPSS Missing Values 20。

运行分析以显示描述统计

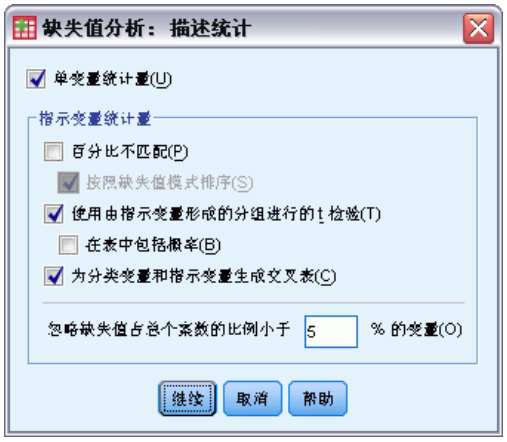
- 要运行缺失值分析，请从菜单中选择：
  分析 > **缺失值分析...**

图片 4-1
“缺失值分析”对话框

- 选择婚姻状况 [marital]、教育程度 [ed]、退休 [retire]和性别 [gender]作为分类变量。
- 选择服务月数 [tenure] 到家庭成员人数 [reside]作为定量（尺度）变量。

  此时，您可以运行过程并获取单变量统计，但我们将选择附加描述统计量。
- 单击**描述性**。

图片 4-2
缺失值分析：“描述”对话框



在“描述”对话框您可以指定各种描述统计量以在输出中显示。缺省单变量统计可帮您确定缺失数据的大体范围，但指示变量统计量提供更多关于一个变量中缺失数据的模型如何影响另一变量的值的信息。

- 选择**对指示变量构成的组进行 t 检验**。
- 选择**为分类变量和指示变量生成交叉表**。
- 单击**继续**。
- 在“缺失值分析”主对话框中，单击**确定**。

估计描述统计

对于本示例，输出包括：

- 单变量统计
- 分离方差表 t 检验，包括分组意味着另一变量存在或缺失。
- 各分类变量表通过各定量（刻度）变量显示每个类别缺失数据的频率。

图片 4-3
单变量统计表

	N	均值	标准差	缺失		极值数目[a]	
				计数	百分比	低	高
MonthsWithService	968	35.56	21.268	32	3.2	0	0
Age	975	41.75	12.573	25	2.5	0	0
YearsAtAddress	850	11.47	9.965	150	15.0	0	9
Income	821	71.1462	83.14424	179	17.9	0	71
YearsWithEmployer	904	11.00	10.113	96	9.6	0	15
PeopleInHousehold	966	2.32	1.431	34	3.4	0	33
MaritalStatus	885			115	11.5		
EducationalLevel	965			35	3.5		
RetirementStatus	916			84	8.4		
Gender	958			42	4.2		

a. 超出范围（Q1 - 1.5*IQR，Q3 + 1.5*IQR）的案例数。

单变量统计供您在缺失数据范围内首先查看，变量与变量。各变量的非缺失值数量显示在 N 列，缺失值数量显示在缺失计数列。缺失百分比列显示具有缺失值的个案的百分比，并针对比较变量间缺失数据的程度提供一种好的测量方法。income（以千元为单位的家庭收入）拥有最多具有缺失值（17.9%）的个案，而age（年龄）拥有最少（2.5%）。income 也拥有最多具有极值。

图片 4-4
分离方差 t 检验表

		MonthsWithService	Age	YearsAtAddress	Income	YearsWithEmployer	PeopleInHousehold
YearsAtAddress	t	.4	.3	.	3.5	1.4	1.0
	df	202.2	192.5	.	313.6	191.1	199.5
	在存在	819	832	850	693	766	824
	缺失	149	143	0	128	138	142
	均值(在存)	35.68	41.79	11.47	74.0779	11.20	2.34
	均值(缺失)	34.91	41.49	.	55.2734	9.86	2.21
Income	t	-5.0	-8.3	-3.9	.	-5.9	3.6
	df	249.5	222.8	191.1	.	203.3	315.2
	在存在	793	801	693	821	741	792
	缺失	175	174	157	0	163	174
	均值(在存)	33.93	40.01	10.67	71.1462	9.91	2.39
	均值(缺失)	42.97	49.73	14.97	.	15.93	2.02
YearsWithEmployer	t	-1.0	-.4	-.7	.5	.	-.3
	df	110.5	110.2	97.6	114.9	.	110.9
	在存在	877	881	766	741	904	874
	缺失	91	94	84	80	0	92
	均值(在存)	35.34	41.69	11.37	71.4953	11.00	2.31
	均值(缺失)	37.70	42.27	12.32	67.9125	.	2.37
MaritalStatus	t	.0	1.8	1.2	-.8	.9	-2.2
	df	148.1	149.5	138.8	121.2	128.3	134.2
	在存在	856	862	748	728	805	857
	缺失	112	113	102	93	99	109
	均值(在存)	35.56	42.00	11.61	70.3887	11.10	2.28
	均值(缺失)	35.57	39.85	10.43	77.0753	10.17	2.61
RetirementStatus	t	-.6	-.4	-.4	.3	.	.2
	df	95.4	94.4	84.0	93.2	.	99.0
	在存在	888	893	777	751	904	885
	缺失	80	82	73	70	0	81
	均值(在存)	35.44	41.70	11.42	71.3356	11.00	2.32
	均值(缺失)	36.89	42.29	11.96	69.1143	.	2.30

分离方差 t 检验表有助于标识缺失值模型可能影响定量（刻度）变量的变量。使用为单个个案指定变量存在或缺失的指示变量计算 t 检验。分组意味着也可以对指示变量制

似乎年纪较长的响应者更不可能报告收入水平。当 income 缺失时，平均 age 为 49.73，与之相比，当 income 未缺失时为 40.01。实际上，income 的缺失似乎影响多个定量（刻度）变量的平均值。此指示数据可能并未完全随机缺失。

图片 4-5
婚姻状况 [marital] 交叉制表

			总计	Unmarried	Married	缺失
						SysMis
YearsAtAddress	存在	计数	850	390	358	102
		百分比	85.0	85.5	83.4	88.7
	缺失	% SysMis	15.0	14.5	16.6	11.3
Income	存在	计数	821	380	348	93
		百分比	82.1	83.3	81.1	80.9
	缺失	% SysMis	17.9	16.7	18.9	19.1
YearsWithEmployer	存在	计数	904	418	387	99
		百分比	90.4	91.7	90.2	86.1
	缺失	% SysMis	9.6	8.3	9.8	13.9
RetirementStatus	存在	计数	916	423	392	101
		百分比	91.6	92.8	91.4	87.8
	缺失	% SysMis	8.4	7.2	8.6	12.2

分类变量 crosstabulations 与指示变量显示与分离方差 t 检验表中发现的相似的信息。再次创建指示变量，只是此次其用于为各分类变量计算各类别的频率。值有助于帮您确定缺失值类别之间是否存在差异。

请查看 marital（婚姻状况）表，指示变量的缺失值数量在 marital 类别之间似乎变化不大。一个人结婚与否似乎并不影响任何定量（刻度）变量的数据缺失情况。例如，85.5% 未婚者报告 address（当前地址居住年限），83.4% 已婚者报告相同变量。差异很小并且很可能是巧合。

图片 4-6
教育程度 [ed] 交叉制表

			总计	Did not complete high school	High school degree	Some college	College degree	Post-undergraduate degree	缺失
									SysMis
YearsAtAddress	存在	计数	850	163	240	175	186	56	30
		百分比	85.0	83.2	85.7	88.4	81.9	87.5	85.7
	缺失	% SysMis	15.0	16.8	14.3	11.6	18.1	12.5	14.3
Income	存在	计数	821	155	229	165	193	50	29
		百分比	82.1	79.1	81.8	83.3	85.0	78.1	82.9
	缺失	% SysMis	17.9	20.9	18.2	16.7	15.0	21.9	17.1
YearsWithEmployer	存在	计数	904	178	254	178	204	60	30
		百分比	90.4	90.8	90.7	89.9	89.9	93.8	85.7
	缺失	% SysMis	9.6	9.2	9.3	10.1	10.1	6.2	14.3
MaritalStatus	存在	计数	885	193	278	148	184	52	30
		百分比	88.5	98.5	99.3	74.7	81.1	81.2	85.7
	缺失	% SysMis	11.5	1.5	.7	25.3	18.9	18.8	14.3
RetirementStatus	存在	计数	916	180	259	180	207	60	30
		百分比	91.6	91.8	92.5	90.9	91.2	93.8	85.7
	缺失	% SysMis	8.4	8.2	7.5	9.1	8.8	6.2	14.3

现在请考虑 ed（教育程度）的交叉制表。如果对象至少接受过大学教育，婚姻状况响应更可能缺失。未接受大学教育的对象中至少 98.5% 报告婚姻状况。另一方面，那些拥有大学学位的人中只有 81.1% 报告婚姻状况。对于那些曾接受大学教育但未获学位者，数量更少。

图片 4-7
退休 [retire] 交叉制表

			总计	No	Yes	缺失 SysMis
YearsAtAddress	存在	计数	850	744	33	73
		百分比	85.0	85.0	80.5	86.9
	缺失	% SysMis	15.0	15.0	19.5	13.1
Income	存在	计数	821	732	19	70
		百分比	82.1	83.7	46.3	83.3
	缺失	% SysMis	17.9	16.3	53.7	16.7
YearsWithEmployer	存在	计数	904	864	40	0
		百分比	90.4	98.7	97.6	.0
	缺失	% SysMis	9.6	1.3	2.4	100.0
MaritalStatus	存在	计数	885	777	38	70
		百分比	88.5	88.8	92.7	83.3
	缺失	% SysMis	11.5	11.2	7.3	16.7

在 retire（退休）中可看到更大差异。那些退休者与那些未退休者相比更不可能报告其收入。退休客户中只有 46.3% 报告收入水平，而那些未退休者报告收入水平的百分比为 83.7。

图片 4-8
性别 [gender] 交叉制表

			总计	Male	Female	缺失 SysMis
YearsAtAddress	存在	计数	850	363	456	31
		百分比	85.0	78.6	91.9	73.8
	缺失	% SysMis	15.0	21.4	8.1	26.2
Income	存在	计数	821	381	406	34
		百分比	82.1	82.5	81.9	81.0
	缺失	% SysMis	17.9	17.5	18.1	19.0
YearsWithEmployer	存在	计数	904	412	457	35
		百分比	90.4	89.2	92.1	83.3
	缺失	% SysMis	9.6	10.8	7.9	16.7
MaritalStatus	存在	计数	885	400	445	40
		百分比	88.5	86.6	89.7	95.2
	缺失	% SysMis	11.5	13.4	10.3	4.8
RetirementStatus	存在	计数	916	420	461	35
		百分比	91.6	90.9	92.9	83.3
	缺失	% SysMis	8.4	9.1	7.1	16.7

gender（性别）的另一差异明显。男性与女性相比，地址信息经常缺失。虽然这些差异可能是巧合，其似乎不可能。数据似乎并非随机完全缺失。

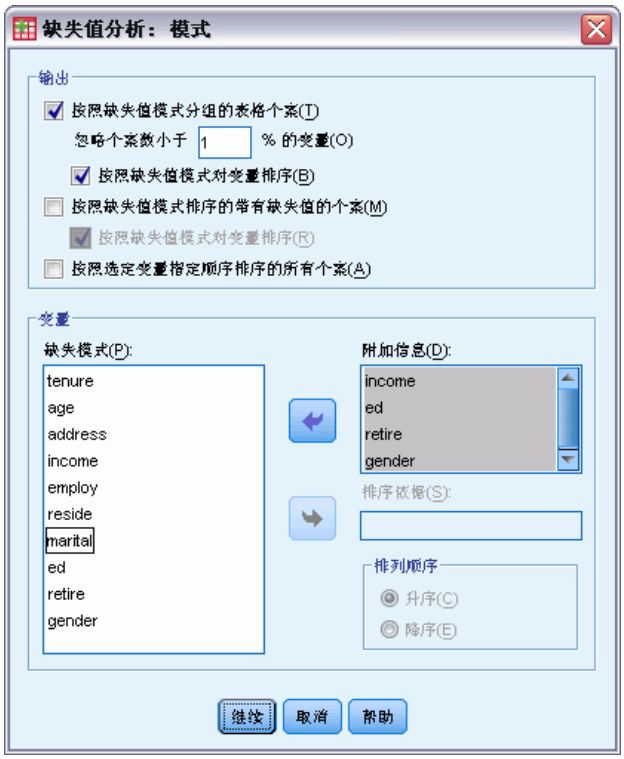
我们将进一步查看缺失数据的模式以进一步探索。

重新运行分析以显示模式

图片 4-9
“缺失值分析”对话框

- “调用缺失值分析”对话框。对话框记住用于之前分析中的变量。不要更改它们。
- 单击**模式**。

图片 4-10
“缺失值分析： 模式”对话框



在“模式”对话框您可以选择各种模式表。我们将显示按缺失值模式分组的制表模式。因为 ed（教育程度）、retire（退休）和 gender（性别）中的缺失模式似乎影响数据，我们将选择显示这些变量的附加信息。由于其大量缺失值，我们也将包括 income（以千元为单位的家庭收入）的附加信息。

- 选择**按缺失值模式分组的制表个案**。
- 选择 income、ed、retire 和 gender 并将其添加至附加信息列表。
- 单击**继续**。
- 在“缺失值分析”主对话框中，单击**确定**。

评估模式表

图片 4-11
制表模式表

案例数	缺失模式[a]										不完整案例数[b]	Income[c]	EducationalLevel[d]					RetirementStatus[d]		Gender[d]	
	Age	PeopleInHousehold	MonthsWithService	EducationalLevel	Gender	RetirementStatus	YearsWithEmployer	MaritalStatus	YearsAtAddress	Income			Did not complete high school	High school degree	Some college	College degree	Post-undergraduate degree	No	Yes	Male	Female
475										X	475	76.5853	99	157	87	101	31	463	12	201	274
109										X	584	.	27	35	19	17	11	95	14	47	62
16									X	X	687	.	5	9	0	1	1	12	4	12	4
87									X		562	54.4368	21	27	9	24	6	85	2	66	21
13		X									488	56.0000	4	3	2	3	1	13	0	4	9
60								X			535	77.2167	1	2	27	24	6	59	1	35	25
16				X							491	47.8125	0	0	0	0	0	16	0	6	10
17			X								492	76.2353	2	7	3	4	1	17	0	7	10
18					X						493	54.1111	3	7	4	4	0	17	1	0	0
16								X		X	660	.	0	0	7	8	1	14	2	6	10
37						X	X				520	59.4595	9	14	5	8	1	0	0	15	22

不显示少于 1% 个（10 个或更少）案例的模式。

a. 以缺失模式排列变量。

b. 完整案例数，如果未使用该模式（用 X 标记）中缺失的变量。

c. 在各个唯一模式处的均值

d. 在各个唯一模式处的频率分布

制表模式表显示在个别个案中多个变量的数据是否往往缺失。也就是说，其可以帮您确定数据是否联合缺失。

在超过 1% 的个案中存在三种模式的联合缺失数据。变量 employ（当前雇方工作年限）和 retire（退休）与其它变量对相比更容易缺失。这并不奇怪，因为 retire 和 employ 记录类似信息。如果您不知道对象是否退休，您很可能也不知道对象为当前雇主工作的年限。

平均 income（以千元为单位的家庭收入）似乎因缺失值模式的不同变化很大。实际上，在 marital（婚姻状况）缺失时，6%（1000 中 60）个案的平均 Income 更高。（当 tenure（服务月数）缺失时其更高，但此模式只占 1.7% 个案。） 请记住，那些接受更高水平教育者更不可能响应婚姻状况相关问题。您可以在 ed（教育程度）频率中看到此倾向。通过假定那些接受更高水平教育者赚更多钱并且更不可能报告婚姻状况，我们可能解释 income 的增加。

考虑缺失数据的描述统计与模式，我们可以得出结论，数据并未随机完全缺失。我们可以通过 Little 的 MCAR 检验确定此结论，并附印 EM 估计值。

重新运行 Little 的 MCAR 检验分析

图片 4-12
“缺失值分析”对话框

- “调用缺失值分析”对话框。
- 单击 EM。
- 单击**确定**。

图片 4-13
EM 均值表

MonthsWithService	Age	YearsAtAddress	Income	YearsWithEmployer	PeopleInHousehold
36.12	41.91	11.58	77.3941	11.22	2.29

a. Little 的 MCAR 检验:卡方 = 179.836，DF = 107，显著性 = .000

Little 的 MCAR 检验结果以脚注形式显示于各 EM 估计表。Little 的 MCAR 检验的原假设为数据完全随机缺失（MCAR）。当缺失数据模式不取决于数据值时，数据为 MCAR。因为在示例中显著性值小于 0.05，我们可以得出结论，数据并未随机完全缺失。这确定了我们从描述统计与制表模式中得出的结论。

此时，因为数据并未随机完全缺失，所以按列表删除具有缺失值的个案或单独地插补缺失值并不安全。不过，您可以使用多重插补来进一步分析此数据集。

章 5

多重插补

使用多重插补完成并分析数据集

电信供应商想更好理解客户数据库中的服务用途模式。他们拥有客户所使用的服务的完整数据，但是公司收集的人口统计信息有大量缺失值。此外，这些值并未随机完全缺失，因此多重插补将用于完成数据集。

客户数据库中的随机样本包含在 telco_missing.sav。 有关详细信息，请参阅附录 A 中的样本文件中的IBM SPSS Missing Values 20。

分析缺失值模式

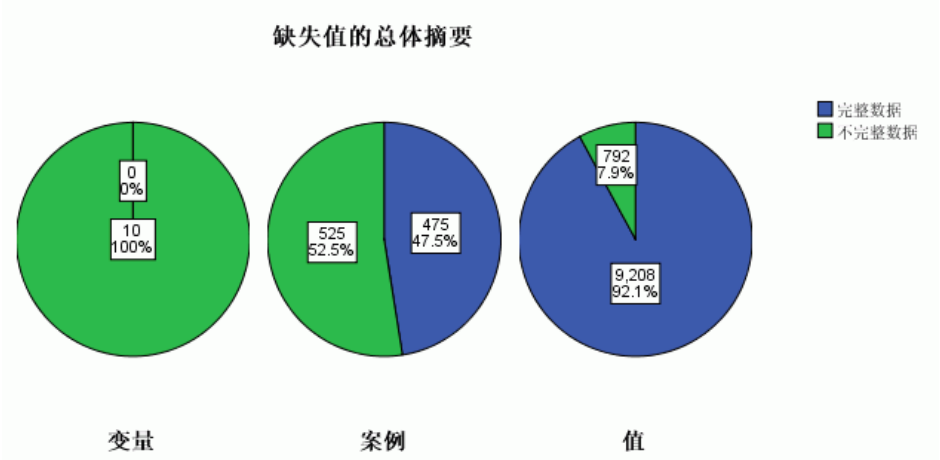
▶ 第一步，查看缺失数据的模式。从菜单中选择：
分析 > 多重插补 > 分析模式...

图片 5-1
“分析模式”对话框

▶ 选择服务月数 [tenure] 到家庭成员人数 [reside]作为分析变量。

总体摘要

图片 5-2
缺失值整体摘要



缺失值的整体摘要显示三个在数据中显示缺失值不同方面的饼图。

- 变量图表显示每 10 个分析变量在个案上至少有一个缺失值。
- 个案图表显示每 1000 个个案中有 525 个在变量上至少有一个缺失值。
- 值图表显示 10000 个值中有 792 个（个案 × 变量）缺失。

带有缺失值的每个个案平均在 10 个变量中有大约 1.5 个上有缺失值。这表明**列表删除**将在数据集中丢失许多信息。

变量摘要

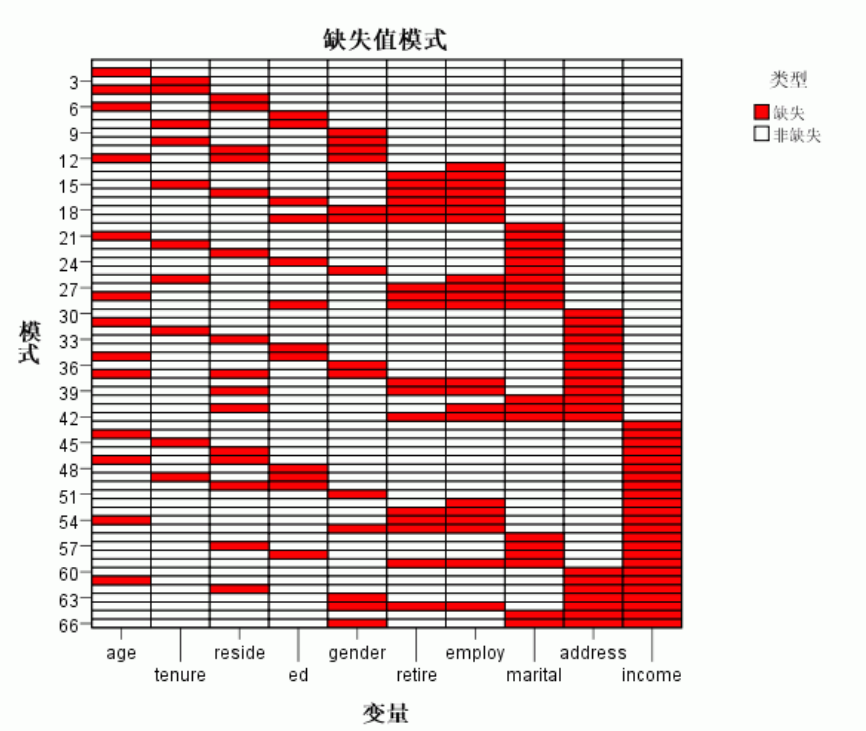
图片 5-3
变量摘要

	缺失		有效的 N	均值	标准 偏差
	N	百分比			
家庭收入（千）	179	17.9%	821	71.1462	83.14424
在现住址居住年数	150	15.0%	850	11.47	9.965
婚姻状况	115	11.5%	885		

为带有至少 10% 缺失值的变量显示变量摘要，且此摘要显示表中的每个变量的缺失值的数量和百分比。还为刻度变量的有效值显示均值和标准差，为所有变量显示有效值的数量。Household income in thousands、Years at current address 和 Marital status 依次有最多的缺失值。

模式

图片 5-4
缺失值模式

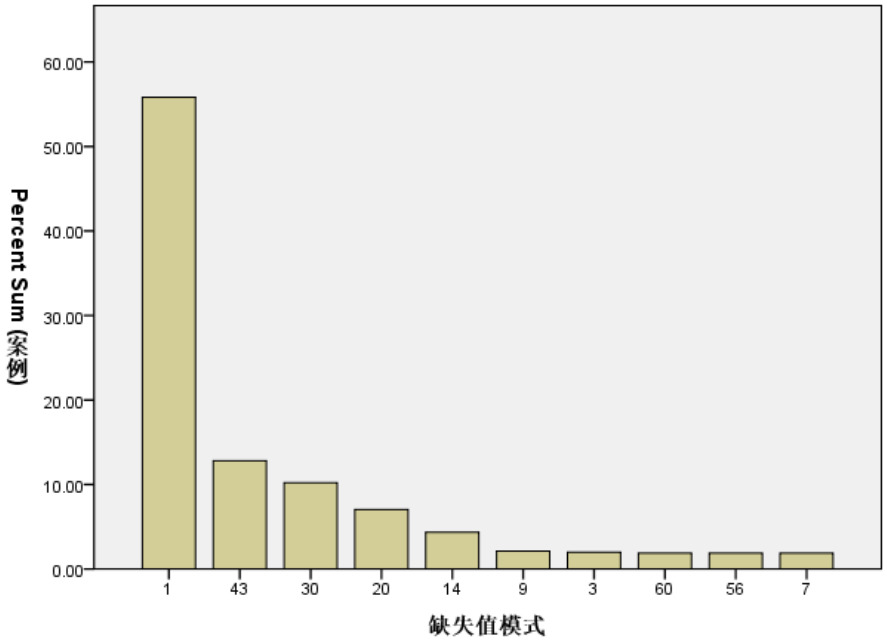


模式图表显示分析变量的缺失值模式。每个模式对应于具有相同的不完整和完整数据模式的一组个案。例如，模式 1 代表没有缺失值的个案，而模式 33 代表在 reside (Number of people in household) 和 address (Years at current address) 上有缺失值的个案，模式 66 代表在 gender (Gender)、marital (Marital status)、address 和 income (Household income in thousands) 上有缺失值的个案。一个数据集可以潜在拥有 2 个变量数模式。对于 10 个分析变量，是 2^{10}=1024；但是，在数据集的 1000 个个案中只代表了 66 个模式。

图表对分析变量和模式进行排序，以揭示所存在的单调性。具体而言，变量以缺失值的顺序从左至右按升序依次排开。然后，模式首先按最后一个变量排序（首先是未缺失值，然后是缺失值），然后按倒数第二个变量排序，依此类推，从左至右进行。这揭示了该单调插补方法是否可用于您的数据，如果不能，判断您的数据近似单调模式的程度。如果数据单调，则图表中的所有缺失单元格和未缺失单元格将变成连续；即，在图表右下角没有未缺失单元格的“岛”，在图表左上角没有缺失值的“岛”。

此数据集非单调，有许多需要插补的值以便达到单调性。

图片 5-5
模式频率



当请求模式时，一个伴随条形图显示每个模式的个案百分比。这显示数据集中超过一半的个案有模式 1，缺失值模式图表显示这是没有缺失值的个案模式。模式 43 代表在 income 上带有缺失值的个案，模式 30 代表在 address 上带有缺失值的个案，模式 20 代表在 marital 上带有缺失值的个案。大多数个案（大约五分之四）由这四个模式代表。模式 14、60 和 56 是十个中最常发生的唯一几个模式，以代表在多个变量上带有缺失值的个案。

缺失模式的分析对多重插补未显示任何特定障碍物，除非使用单调方法并不真正可行。

缺失值的自动插补

现在即可开始插补值；我们将从带有自动设置的运行开始，但是在请求插补之前，我们将设置随机种子。通过设置随机数种子您可以精确复制此分析。

▶ 要设置随机数种子，请从菜单中选择：
转换 > **随机数字生成器**...

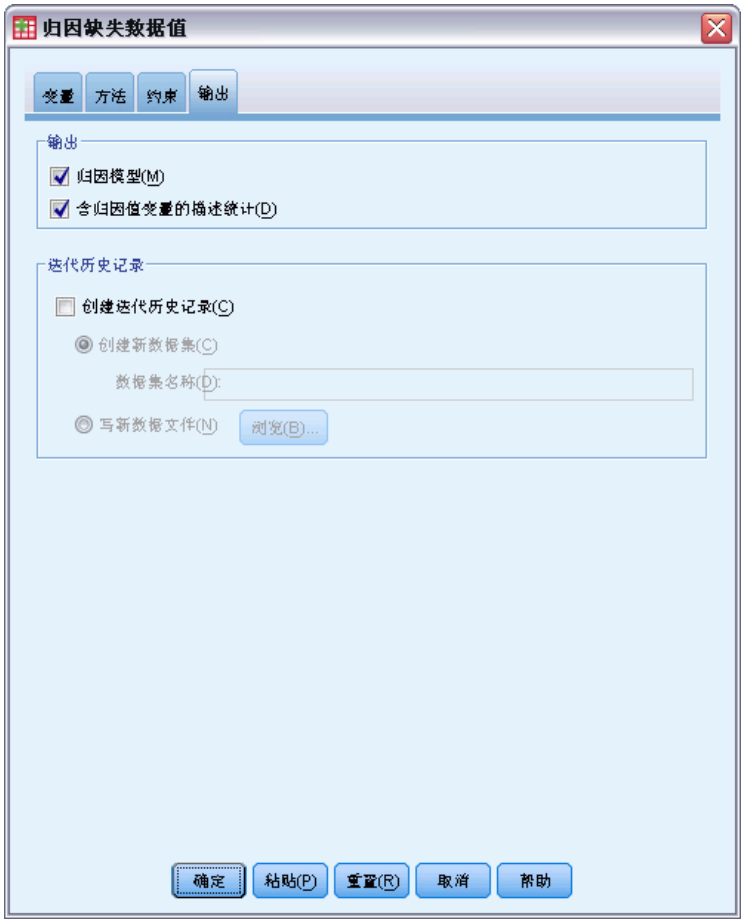
图片 5-6
“随机数生成器”对话框

- 选择设置工作发生器。
- 选择Mersenne 扭曲器。
- 选择设置起点。
- 选择固定值并键入 20070525 作为值。
- 单击确定。
- 要多重插补缺失数据值，请从从菜单中选择：
  分析 > 多重插补 > 插补缺失数据值...

图片 5-7
“插补缺失数据值”对话框

- 在插补模型中选择服务月数 [tenure] 到家庭成员人数 [reside]作为变量。
- 键入 **telcolmputed** 作为应该保存插补数据的数据集。
- 单击**输出**选项卡。

图片 5-8
“输出”选项卡



- 选择带有插补值的变量描述统计。
- 单击确定。

归因指定

图片 5-9
归因指定

归因方法	完全条件指定
归因数	5
刻度变量模型	线性回归
模型中包含的交互数	（无）
最大缺失值百分比	100.0%

插补指定表是一种复查您所请求内容的有用方法，以便可以确认指定正确。

归因结果

图片 5-10
插补结果

归因方法	完全条件指定
完全条件指定方法迭代	10
因变量　已归因	tenure,age,marital,address,income,ed, employ,retire,gender,reside
未归因（太多缺失值）	
未归因（无缺失值）	
归因序列	age,tenure,reside,ed,gender,retire,employ, marital,address,income

插补结果概述了在插补过程期间实际发生的情况。特别要注意的是：

- 指定表中的插补方法是自动的，自动方法选择实际选择的方法是完全条件指定。
- 插补所有请求的变量。
- 插补序列是以缺失值模式图表上 x 轴上变量出现的顺序。

归因模型

图片 5-11
归因模型

	模型		缺失值	归因值
	类型	效果		
年龄	线性回归	ed,gender,retire,marital,tenure, reside,employ,address,income	25	125
服务月数	线性回归	ed,gender,retire,marital,age, reside,employ,address,income	32	160
家庭人数	线性回归	ed,gender,retire,marital,age, tenure,employ,address,income	34	170
受教育水平	Logistic 回归	gender,retire,marital,age,tenure, reside,employ,address,income	35	175
性别	Logistic 回归	ed,retire,marital,age,tenure, reside,employ,address,income	42	210
退休	Logistic 回归	ed,gender,marital,age,tenure, reside,employ,address,income	84	420
现职位工作年数	线性回归	ed,gender,retire,marital,age, tenure,reside,address,income	96	480
婚姻状况	Logistic 回归	ed,gender,retire,age,tenure, reside,employ,address,income	115	575
在现住址居住年数	线性回归	ed,gender,retire,marital,age, tenure,reside,employ,income	150	750
家庭收入（千）	线性回归	ed,gender,retire,marital,age, tenure,reside,employ,address	179	895

归因模型表进一步详细介绍了每个变量是如何插补的。特别要注意的是：

- 变量以插补序列的顺序列出。
- 刻度变量用线性回归建模，分类变量用 Logistic 回归建模。
- 每个模型使用所有其他变量作为主效应。
- 报告每个变量的缺失值数，以及为该变量插补的值的总数（数字缺失 × 插补数）。

描述统计

图片 5-12
tenure（服务月数）的描述统计

数据	归因	N	均值	标准 偏差	极小值	极大值
初始数据		968	35.56	21.268	1.00	72.00
归因值	1	32	38.88	25.522	-2.66	87.33
	2	32	34.42	24.566	-15.03	86.71
	3	32	37.58	19.934	6.75	81.44
	4	32	37.33	24.041	-15.23	93.07
	5	32	36.74	17.118	-.59	78.74
归因后完整数据	1	1000	35.67	21.410	-2.66	87.33
	2	1000	35.52	21.368	-15.03	86.71
	3	1000	35.63	21.220	1.00	81.44
	4	1000	35.62	21.351	-15.23	93.07
	5	1000	35.60	21.142	-.59	78.74

描述统计表显示带有插补值的变量摘要。为每个变量产生一个单独的图表。显示的统计类型取决于是刻度还是分类变量。

刻度变量的统计包括计数、均值、标准差、最小值和最大值，为原始数据、每组插补值和每个完整数据集（结合源数据和插补值）显示。

tenure（服务月数）的描述统计表在大约与原始数据相等的每组插补值中显示均值和标准差；但是，当您查看最小值并发现 tenure 的负值已经插补时立即出现问题。

图片 5-13
marital（婚姻状况）的描述统计

数据	归因	类别	N	百分比
初始数据		0	456	51.5
		1	429	48.5
归因值	1	0	43	37.4
		1	72	62.6
	2	0	47	40.9
		1	68	59.1
	3	0	52	45.2
		1	63	54.8
	4	0	46	40.0
		1	69	60.0
	5	0	45	39.1
		1	70	60.9
归因后完整数据	1	0	499	49.9
		1	501	50.1
	2	0	503	50.3
		1	497	49.7
	3	0	508	50.8
		1	492	49.2
	4	0	502	50.2
		1	498	49.8
	5	0	501	50.1
		1	499	49.9

对于分类变量，统计包括原始数据、插补值和完整数据的计数和百分比（按分类）。marital（婚姻状况）表有一个有趣的结果，因为对于插补值，与原始数据相比，被估计为已婚的个案比例更大。这可能是由于随机变异；或者缺失的几率可能与此变量的值有关。

图片 5-14
income（以千元为单位的家庭收入）的描述统计

数据	归因	N	均值	标准 偏差	极小值	极大值
初始数据		821	71.1462	83.14424	9.0000	944.0000
归因值	1	179	84.0215	91.16694	-127.7096	337.6979
	2	179	96.8150	91.76282	-115.5968	326.9475
	3	179	87.5989	92.18560	-210.5039	304.1452
	4	179	92.5710	102.09050	-195.6540	351.9920
	5	179	101.0354	92.89237	-91.0533	345.4123
归因后完整数据	1	1000	73.4508	84.73254	-127.7096	944.0000
	2	1000	75.7409	85.27348	-115.5968	944.0000
	3	1000	74.0912	85.01951	-210.5039	944.0000
	4	1000	74.9812	87.17159	-195.6540	944.0000
	5	1000	76.4963	85.69261	-91.0533	944.0000

与 tenure 以及所有其他刻度变量一样，income（以千元为单位的家庭收入）清晰显示负的插补值 — ，我们将需要运行一个在某些变量上带有回归的定制模型。但是，income 显示其他潜在问题。每个插补的均值比原始数据高得多，每个插补的最大值比原始数据低得多。收入的分布往往非常向右偏斜，因此这可能是问题的来源。

定制插补模型

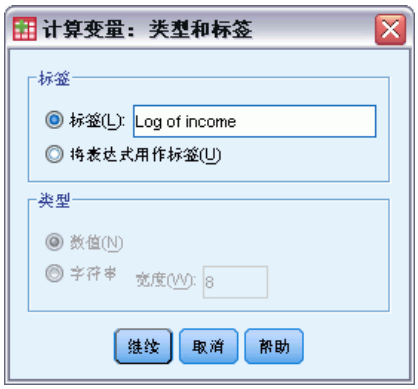
为了防止插补值处于每个变量的合理值的范围以外，我们指定一个在变量上带有回归的定制插补模型。此外，Household income in thousands 非常向右偏斜，进一步分析很可能使用 income 的对数，因此好像直接插补对数收入比较合理。

- 确保原始数据集处于活动状态。
- 要创建对数收入变量，请从菜单中选择：
  转换 > 计算变量...

图片 5-15
“计算变量”对话框

- 键入 lninc 作为目标变量。
- 键入 ln(income) 作为数值表达式。
- 单击类型&标签。

图片 5-16
“类型和标签”对话框

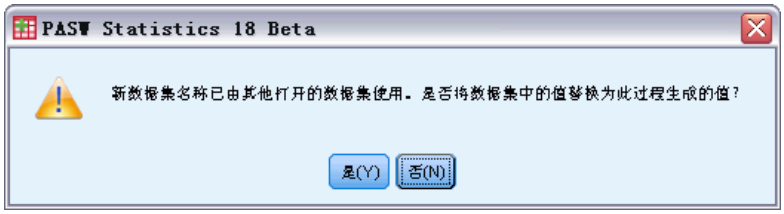


- 键入 Log of income 作为标签。
- 单击**继续**。
- 在“计算变量”对话框中单击**确定**。

图片 5-17
在插补模型中带有 Log of income 的变量选项卡替换 Household income in thousands。

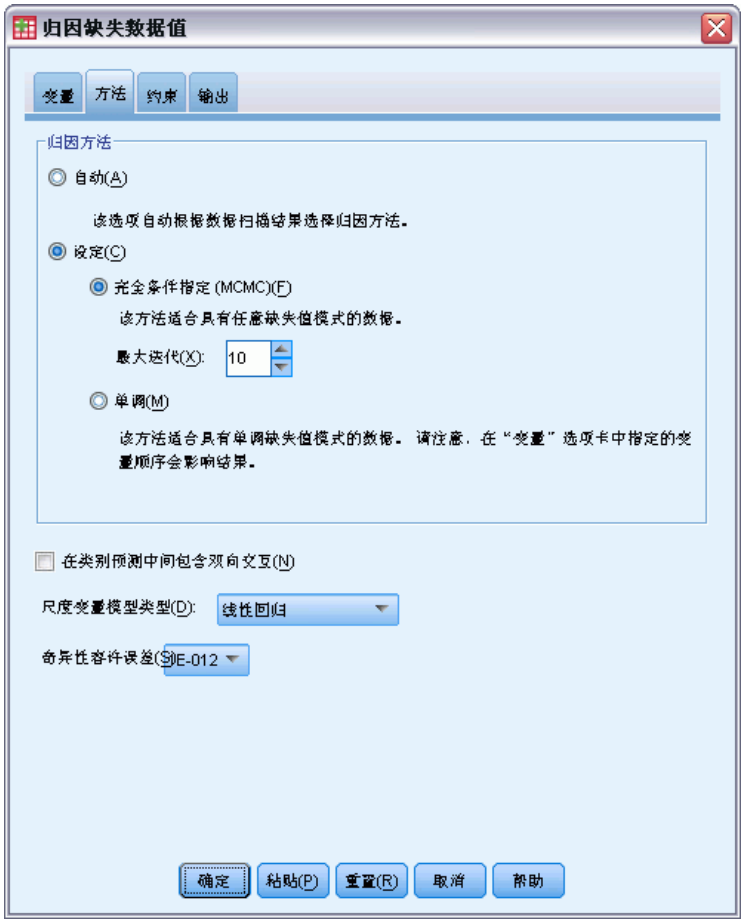
- 调用“插补缺失数据值”对话框并单击**变量**选项卡。
- 在模型中取消选择 Household income in thousands [income] 并选择 Log of income [lninc] 作为变量。
- 单击**方法**选项卡。

图片 5-18
提醒替换现有数据集



- 在出现的提醒中单击**是**。

图片 5-19
“方法”选项卡

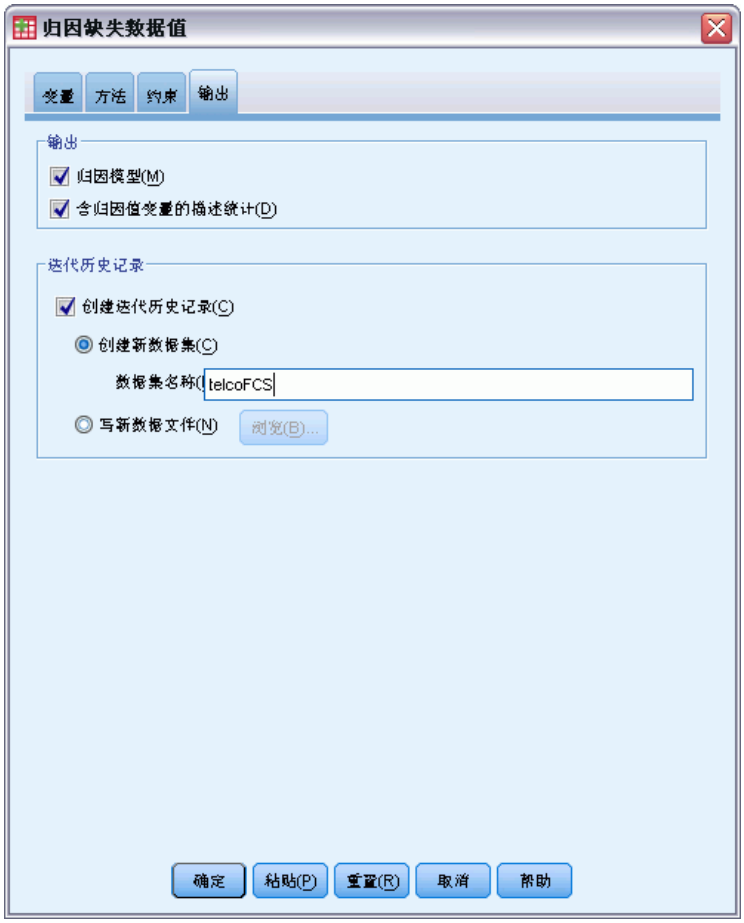


- 选择**定制**并保留**完全条件指定**选作插补方法。
- 单击**约束**选项卡。

图片 5-20
“约束”选项卡

- 单击**扫描数据**。
- 在定义约束网格中，键入 1 作为 服务月数 [tenure] 的最小值。
- 键入 18 作为 age (Age in years) 的最小值。
- 键入 0 作为 address (Years at current address) 的最小值。
- 键入 0 作为 employ (Years with current employer) 的最小值。
- 键入 1 作为最小值，1 作为 reside (Number of people in household) 的四舍五入级别。注意许多其他刻度变量以整数值报告，但是假定某人在当前地址住了 13.8年是合理的，但是不能认为 2.2 个人住在那里。
- 键入 0 作为 lninc (Log of income) 的最小值。
- 单击**输出**选项卡。

图片 5-21
“输出”选项卡



- 选择**创建迭代历史**并键入 `telcoFCS`作为新数据集的名称。
- 单击**确定**。

归因约束

图片 5-22
归因约束

	归因中的角色		归因值		
	因变量	预测值	极小值	极大值	取整
服务月数	是	是	1	（无）	
年龄	是	是	18	（无）	
婚姻状况	是	是			
在现住址居住年数	是	是	0	（无）	
受教育水平	是	是			
现职位工作年数	是	是	0	（无）	
退休	是	是			
性别	是	是			
家庭人数	是	是	1	（无）	整数
收入的对数	是	是	0	（无）	

定制插补模型产生一个复查置于插补模型上的回归的新表。一切都是符合您的指定。

描述统计

图片 5-23
tenure（服务月数）的描述统计

数据	归因	N	均值	标准 偏差	极小值	极大值
初始数据		968	35.56	21.268	1.00	72.00
归因值	1	32	35.48	19.880	5.01	81.87
	2	32	37.74	23.840	2.88	100.98
	3	32	37.78	21.352	5.53	82.63
	4	32	38.99	20.546	8.76	77.21
	5	32	37.78	20.085	1.57	93.50
归因后完整数据	1	1000	35.56	21.216	1.00	81.87
	2	1000	35.63	21.345	1.00	100.98
	3	1000	35.63	21.264	1.00	82.63
	4	1000	35.67	21.244	1.00	77.21
	5	1000	35.63	21.225	1.00	93.50

在带有回归的定制插补模型中 tenure (Months with service) 的描述统计表显示 tenure 的负插补值的问题已解决。

图片 5-24
marital（婚姻状况）的描述统计

数据	归因	类别	N	百分比
初始数据		0	456	51.5
		1	429	48.5
归因值	1	0	44	38.3
		1	71	61.7
	2	0	52	45.2
		1	63	54.8
	3	0	46	40.0
		1	69	60.0
	4	0	48	41.7
		1	67	58.3
	5	0	47	40.9
		1	68	59.1
归因后完整数据	1	0	500	50.0
		1	500	50.0
	2	0	508	50.8
		1	492	49.2
	3	0	502	50.2
		1	498	49.8
	4	0	504	50.4
		1	496	49.6
	5	0	503	50.3
		1	497	49.7

marital (Marital status) 的表现在有一个插补 (3)，其分布更加与原始数据一致，但是大多数仍显示与原始数据相比，被估计为已婚的个案比例更大。这可能是由于随机变异，但是可能要求进一步研究数据才能确定这些值是否随机缺失（MAR）。此处，我们不用继续解决此问题。

图片 5-25
lninc (Log of income) 的描述统计

数据	归因	N	均值	标准 偏差	极小值	极大值
初始数据		821	3.9291	.75305	2.1972	6.8501
归因值	1	179	4.1909	.81638	2.1813	6.3869
	2	179	4.1368	.93216	1.7272	6.5118
	3	179	4.2151	.88185	2.1154	6.2945
	4	179	4.2622	.87908	1.7182	6.6775
	5	179	4.1155	.97928	1.5827	6.4682
归因后完整数据	1	1000	3.9759	.77092	2.1813	6.8501
	2	1000	3.9662	.79161	1.7272	6.8501
	3	1000	3.9803	.78490	2.1154	6.8501
	4	1000	3.9887	.78708	1.7182	6.8501
	5	1000	3.9624	.80091	1.5827	6.8501

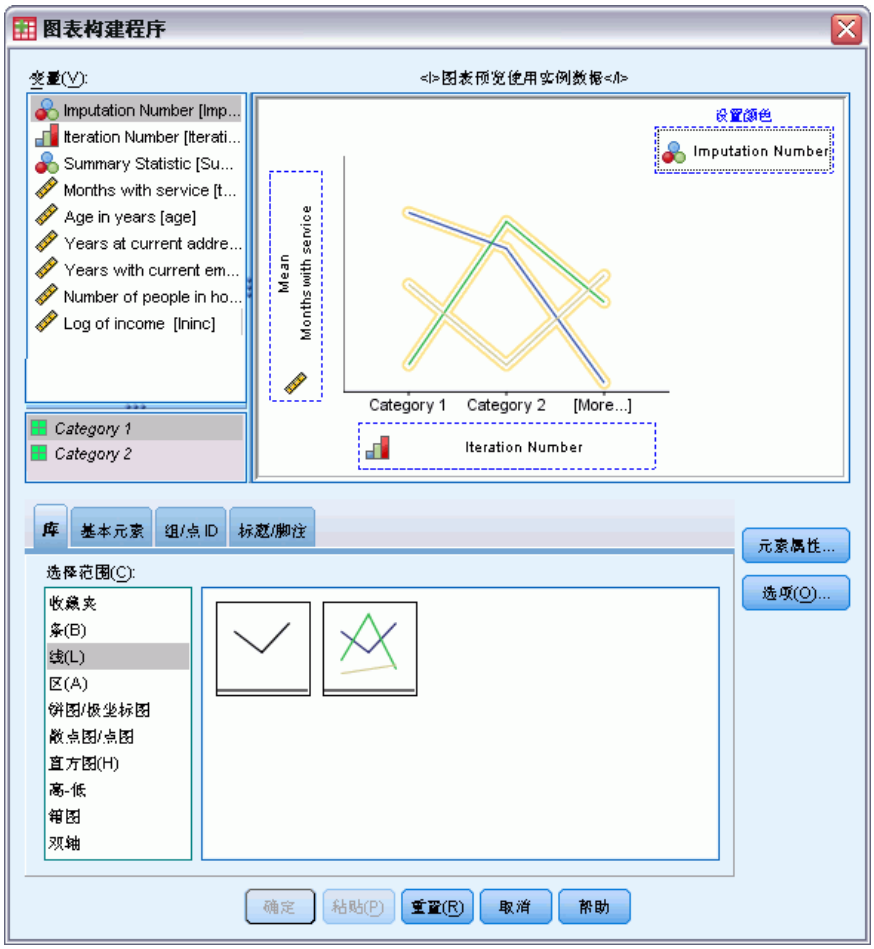
与 tenure 以及所有其他刻度变量一样，lninc (Log of income) 并不显示负的插补值。此外，在 income 刻度中，与自动插补运行 — 相比，插补的均值更接近原始数据的均值，lninc 原始数据的均值大约是 $e^{3.9291}$=50.86，而插补中的典型均值大约是 $e^{4.2}$=66.69。此外，每个插补的最大值更加接近原始数据的最大值。

检查 FCS 收敛

当使用完全条件指定方法时，最好检查均值和标准差（按迭代）的图，插补值的每个刻度因变量的插补以便帮助评估模型收敛。

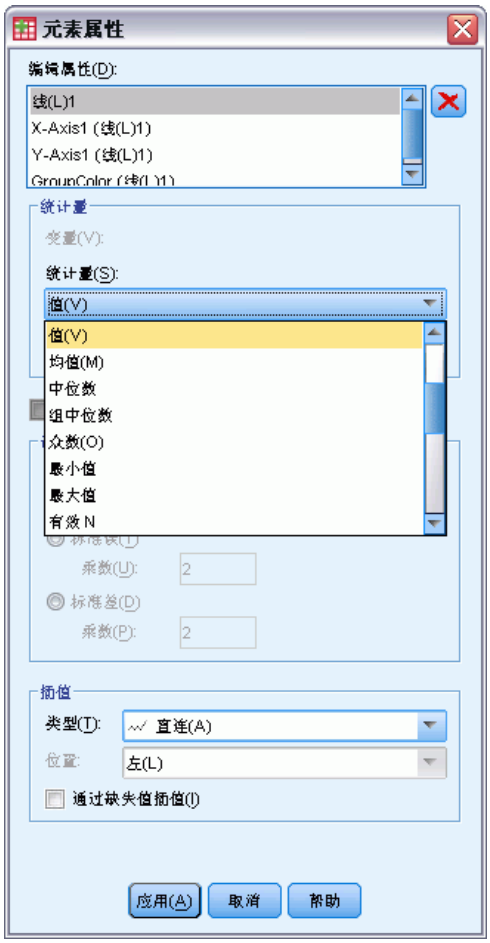
- 要创建此类型的图表，请激活 telcoFCS 数据集，然后从菜单中选择：
  图形 > 图表生成器...

图片 5-26
图表生成器，多行图



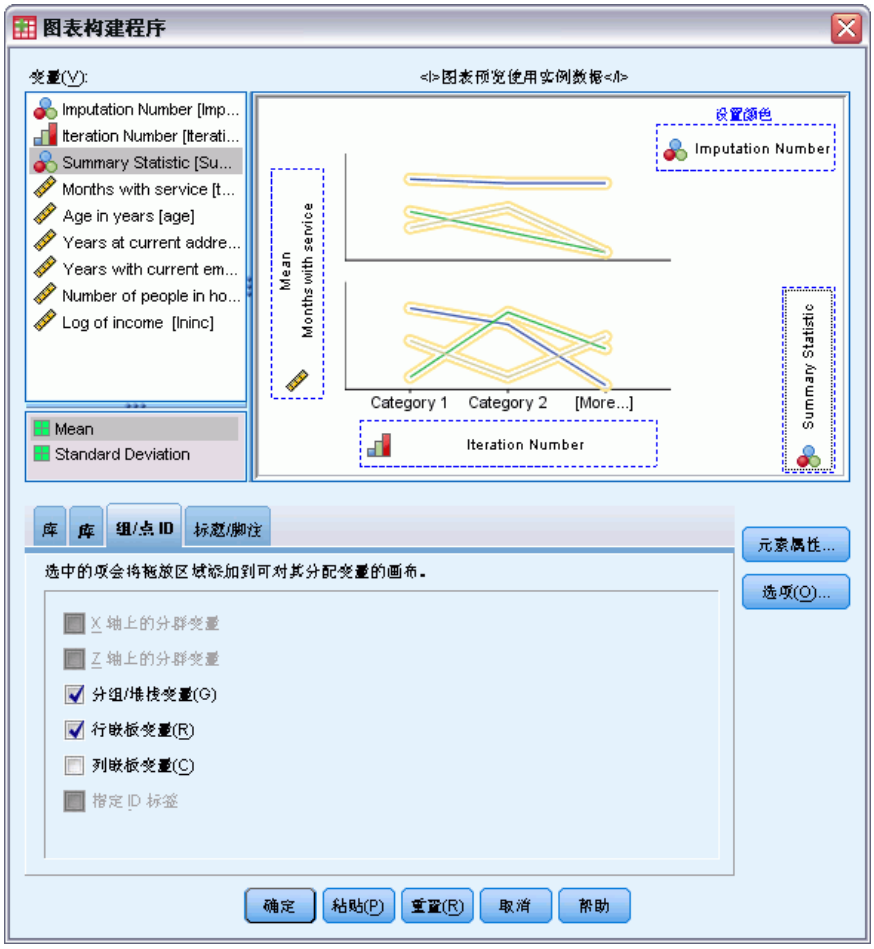
- 选择线图库并选择“多行”。
- 选择服务月数 [tenure] 作为要在 Y 轴上绘制的变量。
- 选择 Iteration Number [Iteration_] 作为要在 X 轴上绘制的变量。
- 选择 Imputation Number [Imputations_] 作为设置颜色的变量。

图片 5-27
图表生成器，元素属性



- 在元素属性中，选择值作为要显示的统计。
- 单击应用。
- 在图表生成器中，单击组/点 ID 选项卡。

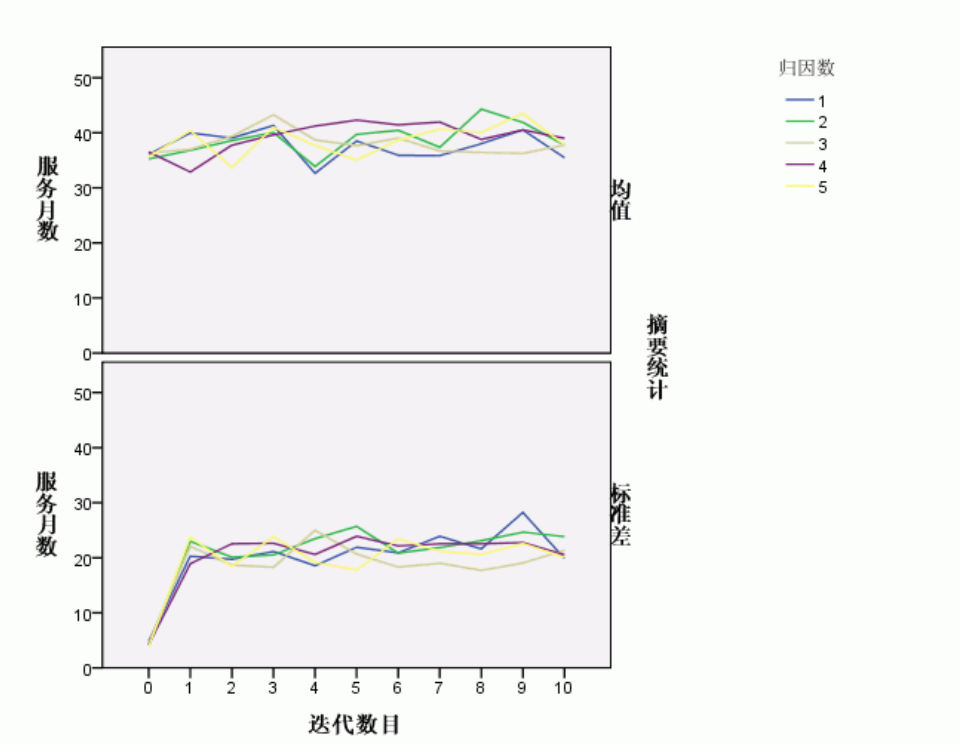
图片 5-28
图表生成器，“组/点 ID”选项卡



- 选择行面板变量。
- 选择 Summary Statistic [SummaryStatistic_] 作为面板变量。
- 单击确定。

FCS 收敛图表

图片 5-29
FCS 收敛图表



您已创建了一对多行表，对于 5 个请求的每个插补在 FCS 插补方法的每次迭代显示 Months with service [tenure] 的插补值的均值和标准差。此图的目的在于在行中查找模式。应该没有，这些看起来是适当的“随机”。您可以为其他刻度变量创建类似的图，注意这些图也不显示可辨别模式。

分析完整的数据

现在，您的插补值好像比较满意，您可以在“完整”的数据上运行分析。数据集包含变量 Customer category [custcat]，按服务用途模式对客户群进行分段，将客户分为四组。如果您可以使用人口统计学信息拟合一个模型以预测组成员资格，则可以为各个潜在客户定制服务。

- 激活 telcoImputed 数据集。要为完整数据创建一个多项 Logistic 回归模型，请从菜单中选择：

  分析 > 回归 > 多项 Logistic...

图片 5-30
“多项 Logistic 回归”对话框

- 选择客户类别作为因变量。
- 选择 Marital status、Level of education、Retired 和 Gender 作为因子。
- 选择 Age in Years、Years at current address、Years with current employer、Number of people in household 和 Log of income 作为协变量。
- 您想将其他客户与预定了基本服务的客户进行比较，因此选择客户类别，然后单击**参考类别**。

图片 5-31
“参考类别”对话框

- 选择**第一个类别**。
- 单击**继续**。
- 在“多项 Logistic 回归”对话框中单击**模型**。

图片 5-32
“模型”对话框

- 选择**定制/步进式**。
- 从“步进项构建项”下拉框中，选择**主效应**。
- 选择从 lninc 到 reside 都作为步进项。
- 单击**继续**。
- 在“多项 Logistic 回归”对话框中单击**确定**。

步骤摘要

图片 5-33
步骤摘要

归因数	模型	操作	效应	模型拟合标准 -2 倍对数似然值	效应选择测试 卡方[a]	df	显著水平
初始数据	0	已输入	截距	1.354E3	.		
	1	已输入	ed	1.261E3	92.583	12	.000
	2	已输入	employ	1.238E3	23.308	3	.000
	3	已输入	marital	1.230E3	7.856	3	.049
1	0	已输入	截距	2.763E3	.		
	1	已输入	ed	2.596E3	166.655	12	.000
	2	已输入	employ	2.559E3	36.469	3	.000
	3	已输入	marital	2.546E3	13.449	3	.004
2	0	已输入	截距	2.763E3	.		
	1	已输入	ed	2.594E3	168.614	12	.000
	2	已输入	employ	2.545E3	48.551	3	.000
	3	已输入	reside	2.531E3	14.553	3	.002
3	0	已输入	截距	2.763E3	.		
	1	已输入	ed	2.594E3	168.386	12	.000
	2	已输入	employ	2.555E3	39.515	3	.000
	3	已输入	marital	2.542E3	12.759	3	.005
4	0	已输入	截距	2.763E3	.		
	1	已输入	ed	2.599E3	164.010	12	.000
	2	已输入	employ	2.557E3	41.914	3	.000
	3	已输入	reside	2.543E3	13.316	3	.004
	4	已输入	address	2.534E3	9.705	3	.021
	5	已输入	lninc	2.524E3	9.383	3	.025
5	0	已输入	截距	2.763E3	.		
	1	已输入	ed	2.610E3	152.374	12	.000
	2	已输入	employ	2.571E3	39.304	3	.000
	3	已输入	marital	2.560E3	11.259	3	.010

步进方法: 向前输入

a. 要输入的卡方基于似然比检验。

多项 Logistic 回归支持汇聚回归系数；但是您会注意输出中的所有表显示每个插补和原始数据的结果。这是因为文件在 Imputation_ 上拆分，因此所有使用拆分变量的表将在单个表中共同显示拆分文件组。

您还会看到参数估计值表并不显示汇聚的估计值；要了解为什么，请查看阶段摘要。我们请求逐步选择模型效应，对于所有插补并未选择相同组的效应，因此不可能执行汇聚。但是，这仍然提供有用信息，因为我们看到 ed (Level of education)、employ (Years with current employer)、marital (Marital status) 和 address (Years at current address) 在插补中经常被逐步选择选中。我们将只使用这些预测变量拟合另外一个模型。

使用预测变量子集运行模型

图片 5-34
“模型”对话框

- 调用“多项 Logistic 回归”对话框，单击**模型**。
- 从步进项列表中取消选择变量。
- 从“强制输入项构建项”下拉框中，选择**主效应**。
- 选择 employ、marital、ed 和 address 作为强制输入项。
- 单击**继续**。
- 在“多项 Logistic 回归”对话框中单击**确定**。

汇聚的参数估计值

此表相当大，但是透视将为我们提供几个不同有用的输出视图。

图片 5-35
汇聚的参数估计值

► 激活（双击）表格，然后从上下文菜单中选择**透视托盘**。

图片 5-36
汇聚的参数估计值

- 将插补数从行移到层。
- 从插补数下拉列表中选择汇聚。

图片 5-37
汇聚的参数估计值

客户类别		B	标准误	显著水平	Exp(B)	Exp(B) 的置信区间 95% 下限	上限	分数缺失信息。	相对增加方差	相对效率
电子服务	截距	.618	.430	.150				.035	.035	.993
	employ	.029	.012	.019	1.029	1.005	1.054	.111	.119	.978
	[marital=0]	-.539	.197	.006	.583	.396	.859	.072	.076	.986
	[marital=1]	0[g]	.	.	.	.	.	.	.	.
	[ed=1]	-2.066	.467	.000	.127	.051	.316	.029	.029	.994
	[ed=2]	-1.380	.448	.002	.251	.105	.605	.049	.051	.990
	[ed=3]	-.809	.454	.075	.445	.183	1.084	.039	.039	.992
	[ed=4]	-.597	.446	.180	.550	.230	1.318	.012	.012	.998
	[ed=5]	0[g]	.	.	.	.	.	.	.	.
	address	.028	.011	.013	1.028	1.006	1.051	.026	.027	.995
附加服务	截距	-1.069	.611	.080				.063	.066	.987
	employ	.052	.011	.000	1.053	1.030	1.076	.121	.130	.976
	[marital=0]	-.313	.184	.089	.731	.509	1.049	.075	.078	.985
	[marital=1]	0[g]	.	.	.	.	.	.	.	.
	[ed=1]	.463	.611	.449	1.588	.479	5.261	.034	.034	.993
	[ed=2]	.679	.607	.263	1.972	.600	6.477	.035	.036	.993
	[ed=3]	.639	.623	.306	1.894	.558	6.427	.043	.044	.991
	[ed=4]	.405	.632	.521	1.500	.434	5.180	.048	.049	.991
	[ed=5]	0[g]	.	.	.	.	.	.	.	.
	address	.014	.010	.168	1.014	.994	1.035	.006	.006	.999
总服务	截距	1.107	.410	.007				.026	.026	.995
	employ	.038	.013	.004	1.039	1.013	1.065	.164	.182	.968
	[marital=0]	-.632	.196	.001	.532	.362	.781	.047	.048	.991
	[marital=1]	0[g]	.	.	.	.	.	.	.	.
	[ed=1]	-3.597	.554	.000	.027	.009	.081	.127	.137	.975
	[ed=2]	-1.777	.429	.000	.169	.073	.392	.041	.042	.992
	[ed=3]	-1.318	.440	.003	.268	.113	.635	.042	.043	.992
	[ed=4]	-.525	.425	.216	.592	.257	1.360	.028	.028	.995
	[ed=5]	0[g]	.	.	.	.	.	.	.	.
	address	.015	.012	.210	1.015	.992	1.039	.077	.080	.985

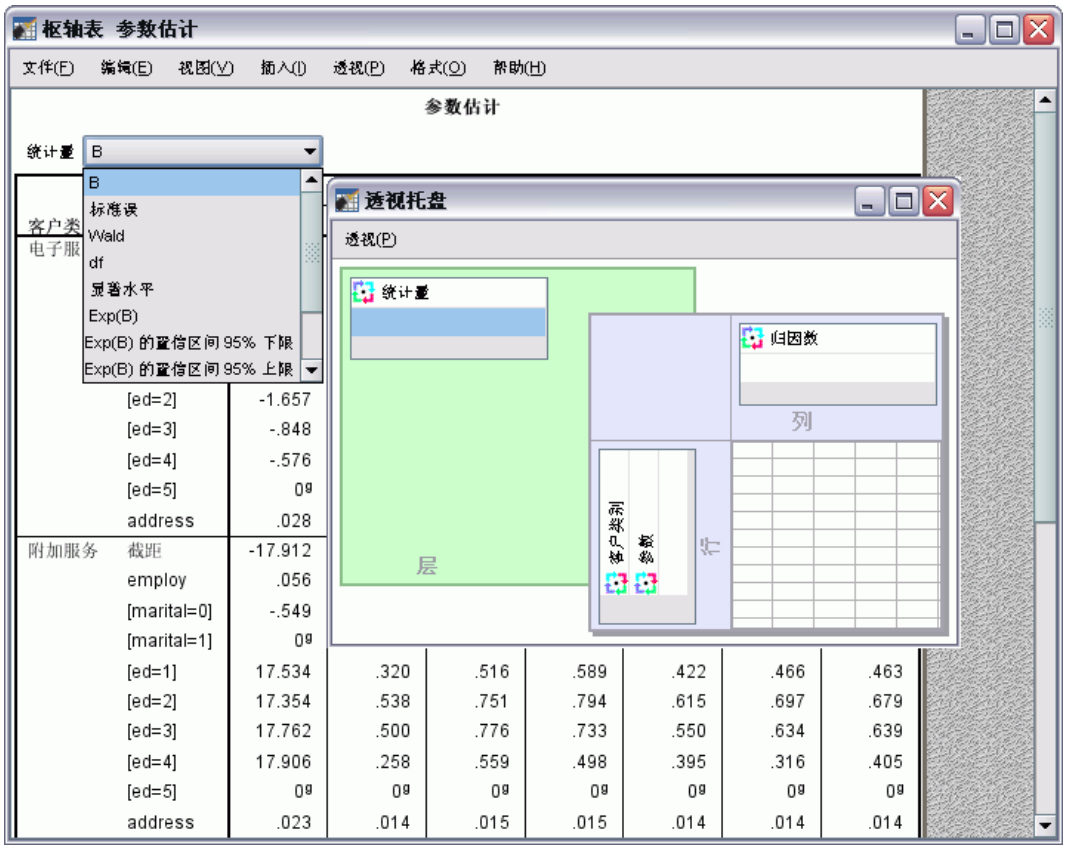
此视图为汇聚的结果显示所有统计。您可以按照与您使用此表用于没有缺失值的数据集相同的方法使用并解释这些系数。

参数估计值表汇总了每个预测变量的作用。系数与其标准误的比率的平方等于 Wald 统计量。如果 Wald 统计量的显著性水平很小（小于 0.05），则该参数不同于 0。

- 具有较大负系数的参数将降低响应类别相对于参考类别的似然性。
- 系数为正的参数将增加该响应类别的似然性。
- 如果给定截距项，则与每个因子的最后一个类别关联的参数为冗余。

在表中有三个附加列，提供有关汇聚输出的更多信息。**部分缺失信息**是由于非回应根据**相对偏差增加**的缺失信息与“完整”信息之比的估计值，而是回归系数的（已修改的）插补之间和平均插补之内方差之比。**相对有效性**比较此估计值与使用插补的无穷数计算的（理论）估计值。相对有效性由部分缺失信息和用于获得汇聚结果的插补数决定；当部分缺失信息较大时，需要一个更大的插补数使相对有效性接近 1，使汇聚的估计值接近理想的估计值。

图片 5-38
汇聚的参数估计值



- 现在重新激活（双击）表格，然后从上下文菜单中选择**透视托盘**。
- 将插补数从层移到列。
- 将统计从列移到层。
- 从统计下拉列表中选择 B。

图片 5-39
汇聚的参数估计值，列中的插补数和层中的统计

统计量= B

客户类别		归因数						
		初始数据	1	2	3	4	5	汇聚
电子服务	截距	.637	.698	.604	.673	.605	.512	.618
	employ	.054	.027	.035	.027	.029	.025	.029
	[marital=0]	-.760	-.588	-.530	-.557	-.559	-.462	-.539
	[marital=1]	0g	0g	0g	0g	0g	0g	0g
	[ed=1]	-2.272	-2.105	-2.119	-2.033	-2.119	-1.955	-2.066
	[ed=2]	-1.657	-1.435	-1.409	-1.409	-1.429	-1.221	-1.380
	[ed=3]	-.848	-.818	-.843	-.854	-.861	-.667	-.809
	[ed=4]	-.576	-.649	-.556	-.630	-.549	-.603	-.597
	[ed=5]	0g	0g	0g	0g	0g	0g	0g
	address	.028	.027	.027	.027	.031	.027	.028
附加服务	截距	-17.912	-.852	-1.210	-1.164	-1.045	-1.076	-1.069
	employ	.056	.048	.057	.050	.053	.051	.052
	[marital=0]	-.549	-.382	-.304	-.331	-.286	-.265	-.313
	[marital=1]	0g	0g	0g	0g	0g	0g	0g
	[ed=1]	17.534	.320	.516	.589	.422	.466	.463
	[ed=2]	17.354	.538	.751	.794	.615	.697	.679
	[ed=3]	17.762	.500	.776	.733	.550	.634	.639
	[ed=4]	17.906	.258	.559	.498	.395	.316	.405
	[ed=5]	0g	0g	0g	0g	0g	0g	0g
	address	.023	.014	.015	.015	.014	.014	.014
总服务	截距	1.266	1.159	1.163	1.125	1.028	1.062	1.107
	employ	.044	.035	.046	.036	.036	.036	.038
	[marital=0]	-.522	-.626	-.697	-.629	-.608	-.599	-.632
	[marital=1]	0g	0g	0g	0g	0g	0g	0g
	[ed=1]	-3.590	-3.714	-3.763	-3.680	-3.478	-3.351	-3.597
	[ed=2]	-2.133	-1.841	-1.850	-1.796	-1.734	-1.662	-1.777
	[ed=3]	-1.214	-1.352	-1.363	-1.407	-1.255	-1.211	-1.318
	[ed=4]	-.468	-.576	-.545	-.525	-.416	-.563	-.525
	[ed=5]	0g	0g	0g	0g	0g	0g	0g
	address	.012	.017	.012	.018	.017	.012	.015

此表的视图对于比较各个插补的值比较有用，可以快速可视化检查从插补到插补的回归系数估计值中的变异，甚至根据原始数据。具体来说，将层中的统计切换到标准错误允许您查看多重插补与列表删除（原始数据）相比是如何降低系数估计值中的变异性的。

图片 5-40
警告

仅使用下列变量: retire, gender, age, reside, lninc 定义子总体，而在构建模型中并不使用这些变量。

对于 归因数 = 初始数据，Hessian 矩阵中有意外的奇异性。这表明应排除某些指示变量，或合并某些类别变量。

尽管有上述警告，PLUM 过程将继续进行。所显示的后续结果将基于最后一次迭代过程。模型拟合的有效性不确定。

但是，在本例中，原始数据集实际上导致了一个错误，这解释了在表格的原始数据列中非常大的附加服务截距的参数估计值和 ed (Level of education) 的非冗余级别。

摘要

通过使用多重插补过程，您分析了缺失值的模式，并发现如果使用简单列表删除，许多信息很可能会丢失。在多重插补初始自动运行之后，您发现需要系数才能将插补值保持在合理的边界内。带有系数的运行生成合适的值，没有证据直接表明 FCS 方法不收敛。使用带有多重插补值的“完整”数据集，您将一个多项 Logistic 回归拟合到数据中，并获得汇聚的估计值，同时发现实际上在原始数据上使用列表删除不可能进行最终模型拟合。

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，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，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，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” 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 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 创建了一个表，用来阐释可能的社会群体。Guttman 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 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， 和 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。
- **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。
- **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。
- **verd1985.sav。**该数据文件涉及某项调查。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。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。

索引