IBM SPSS Missing Values 26

注释

使用本信息及其支持的产品之前，请阅读 第 17 页的『通知』 中的信息。

产品信息

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

目录

缺失值

SPSS® Statistics Premium Edition 或"缺失值"选项中包含以下缺失值功能。

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，则成列删除的典型方法相对比较"安全"。"缺失值"可以帮助确定成列删除是否足够，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"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：

1. **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。
2. **插补缺失值。** 使用"插补缺失数据值"以对缺失值进行多重插补。
3. **分析"完整"的数据。** 使用任何支持多重插补数据的过程。请参阅第 12 页的『分析多重插补数据』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。

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平均值、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）。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 分析用于估计平均值、相关性和协方差。它还用来确定数据正在随机完全缺失。缺失值然后将由插补值替换，并保存到新的数据文件中以供进一步分析。

统计。 单变量统计，包括非缺失值个数、平均值、标准差、缺失值个数以及极值个数。使用列表法、成对法、EM 法或回归法的估计平均值、协方差矩阵以及相关性矩阵。对 EM 结果进行的 Little 的 MCAR 检验。按各种方法进行的平均值总计。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（重复）权重。忽略重复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，所有估算方法（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"允许用平均值替换缺失值。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

1. 从菜单中选择：

   分析 > **缺失值分析...**
2. 至少选择一个定量（刻度）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。请参阅主题『显示缺失值模式』，了解更多信息。
- 单击**描述**显示缺失值的描述统计。请参阅主题第 3 页的『显示缺失值的描述统计』，了解更多信息。
- 选择一种估计统计（平均值、相关性和协方差）和可能插补缺失值的方法。请参阅主题第 4 页的『估计统计与插补缺失值』，了解更多信息。
- 如果选择 **EM** 或**回归**，请单击**变量**以指定要在估计中使用的子集。请参阅主题第 6 页的『预测的变量与预测变量』，了解更多信息。
- 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

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

显示 可用三种类型的表格显示缺失数据的模式。

个案表（按缺失值模式分组）
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。

按照缺失值模式排序的带有缺失值的个案
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。

所有个案（可以选择按选定变量排序）
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(A)
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（刻度）变量，显示平均值；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**所有个案（可以选择按选定变量排序）**。

显示缺失值模式

1. 在"缺失值分析"主对话框中，选择一个或多个要显示缺失值模式的变量。
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

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（刻度）变量，还显示以下内容：

- 平均值(E)
- 标准差
- 极高值和极低值的数量

指示符变量统计

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

不匹配百分比

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。

使用由指示符变量形成的分组进行的 t 检验

使用 Student *t* 统计，比较每个定量变量的两个组的平均值。该组指定一个变量存在或缺失。显示两个组的 *t* 统计、自由度、缺失和非缺失值计数以及平均值。您还可以显示任何与 *t* 统计相关的双尾概率。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。

对分类变量和指示符变量进行交叉制表

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。

省略缺失值占个案数的比例小于 n %

为减小表的大小，可以省略仅为少量个案计算的统计。

显示描述统计

1. 在"缺失值分析"主对话框中，选择要显示缺失值描述统计的变量。
2. 单击**描述性**。
3. 选择需要显示的描述统计。

估计统计与插补缺失值

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/或回归法选择估计平均值、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、平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（通过残差平均值方的平方根测量）。

EM 估计选项

EM 法使用迭代过程估计具有缺失值的定量（刻度）变量的平均值、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。缺省情况下，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率的标准偏差比率。混合正态分布假设只有分布标准偏差不同。平均值必须相同。

最大迭代数

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性，过程也将停止。

保存完成数据

您可以保存一个有插补值而不是缺失值的数据集。但仍要注意，使用插补值且基于协方差的统计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

1. 在"缺失值分析"主对话框中，选择要使用 EM 法估计缺失值的变量。
2. 在**估计**组中选择 **EM**。
3. 要指定预测的变量和预测变量，请单击"**变量**"。请参阅主题第 6 页的『预测的变量与预测变量』，了解更多信息。
4. 单击 **EM**。
5. 选择您想要的 EM 选项。

回归估计选项

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

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 *(Residuals)*

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 *(Normal Variates)*

从具有期望值 0 和等于回归的均方误差项平方根的标准差的分布中，随机抽取误差项。

Student t 变量 *(Student's t Variates)*

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。

最大预测变量数

设置估计过程中使用的预测变量（自变量）的最大数目限制。

保存完成数据

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，选择要使用回归法估计缺失值的变量。
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**"。请参阅主题『预测的变量与预测变量』，了解更多信息。
4. 单击**回归**。
5. 选择您想要的回归选项。

预测的变量与预测变量

缺省情况下，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

1. 在"缺失值分析"主对话框中，选择要使用回归法估计缺失值的变量。
2. 在"估计"组中选择 **EM** 或**回归**。
3. 单击**变量**。
4. 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性方式。
- 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。

多重插补

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。

多重插补数据注意事项

分析变量。 分析变量可为：

- 名义 *(Nominal)*. 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 *(Ordinal)*. 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 *(Scale)*. 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弹出菜单中选择测量级别暂时更改变量的测量级别。要永久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：

频率权重。 此过程接受频率（重复）权重。忽略重复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。对于支持汇聚的过程完整列表，请参阅第 12 页的『分析多重插补数据』。

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。

有两个过程专门用于多重插补。

- 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。请参阅第 12 页的『分析多重插补数据』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。

从菜单中选择：

分析 > **多重插补** > **分析模式...**

1. 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：

缺失值摘要

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（平均值和标准差）。

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。请参阅第 12 页的『分析多重插补数据』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。

从菜单中选择：

分析 > **多重插补** > **插补缺失数据值...**

1. 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。
2. 指定要计算插补的数量。缺省情况下，该值为 5。
3. 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的插补值的一组个案组成。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；例如平均值、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，就会显示测量级别警告。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。

扫描数据

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

手动指定

列出测量级别未知的所有字段。您可以对这些字段指定测量级别。此外，也可以在数据编辑器的"变量列表"窗格中指定测量级别。

由于测量级别对于此过程很重要，因此除非所有字段均定义有测量级别，否则您无法运行此过程。

方法

"方法"对话框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 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。

定制 当您确定所要使用的方法时，请选择此项。

完全条件指定 (MCMC)

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，则它使用缺省 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 链不收敛性，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，以评估收敛性。

Monotone

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。

在类别预测中间包含双阶交互

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双向交互。

标度变量的模型类型

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，您也可以选择预测平均值匹配（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差的矩阵作为奇异矩阵对待。指定一个正值。

约束

"约束"对话框允许您限制插补过程中变量的角色，以及限制标度变量插补值范围，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。

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。

输出

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，那么也会显示该约束。

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。

具有插补值的变量的描述统计

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）的平均值、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。该数据集包含每个已插补值的标度因变量的平均值、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性。

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。

处理多重插补数据

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。

激活多重插补数据集以用于分析

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...**

1. 选择**比较各组**。
2. 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

您可以通过单元格背景色、字体和加粗类型（用于插补值）来区分插补值与观察值。当您在当前会话中使用插补缺失值创建一个新数据集，缺省情况下打开标记。当您打开一个包括插补的已保存的数据文件时，关闭标记。

要打开标记，请从"数据编辑器"菜单中选择：

查看 > **标记插补数据**

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

1. 从菜单中选择：

   编辑 > **转至插补...**
2. 从下拉列表中选择插补（或原始数据）。

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。

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"单元格中的编辑插补值，则该单元格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"交叉表格"都支持汇聚，而"比率"、"P-P 图"和"Q-Q 图"却不支持。

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。** 缺省情况下，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、其标准误差、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）、平均值（平均值差值）和计数。当统计的标准误差可用时，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。以下功能受支持：

- "统计"表格支持"单变量"汇聚（如果也需要平均值的标准误差）时的"平均值"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。以下功能受支持：

- "描述统计"表格支持"单变量"汇聚（如果也需要平均值的标准误差）时的"平均值"和 Naïve 汇聚时的"N"。

交叉表。以下功能受支持：

- "交叉制表"表格支持 Naïve 汇聚时的"计数"。

平均值。以下功能受支持：

- "报告"表格支持"单变量"汇聚（如果也需要平均值的标准误差）时的"平均值"和 Naïve 汇聚时的"N"。

单样本 T 检验。以下功能受支持：

- "统计"表格支持"单变量"汇聚和 Naïve 汇聚时的"平均值"。
- "检验"表格支持"单变量"汇聚时的"平均值差值"。

独立样本 T 检验。以下功能受支持：

- "组统计"表格支持"单变量"汇聚和 Naïve 汇聚时的"平均值"。

- "检验"表格支持"单变量"汇聚时的"平均值差值"。

配对样本 T 检验。以下功能受支持：

- "统计"表格支持"单变量"汇聚和 Naïve 汇聚时的"平均值"。
- "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"平均值"。

单因素 ANOVA。以下功能受支持：

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

UNIANOVA。以下功能受支持：

- Naïve 汇聚时的描述。
- Naïve 汇聚时的主体间。
- 单变量汇聚时的参数估计。
- 单变量汇聚时的估计边际均值。
- 单变量汇聚时的估计边际均值比较。

GLM 单变量。以下功能受支持：

- "参数估计值"表格支持"单变量"汇聚时的系数、B。

线性混合模型。以下功能受支持：

- "描述统计"表格支持 Naïve 汇聚时的"平均值"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"平均值"。
- 估计边际均值："成对比较"表格支持"单变量"汇聚时的"平均值差值"。

"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"平均值"。
- "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"平均值"。
- 估计边际均值："估算值"表格支持"单变量"汇聚时的"平均值"。
- 估计边际均值："成对比较"表格支持"单变量"汇聚时的"平均值差值"。

双变量相关性。以下功能受支持：

- "描述统计"表格支持 Naïve 汇聚时的"平均值"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。

偏相关性。以下功能受支持：

- "描述统计"表格支持 Naïve 汇聚时的"平均值"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"平均值"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"平均值"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。

序数回归。以下功能受支持：

- "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"平均值"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。以下功能受支持：

- "描述"表格支持 Naïve 汇聚时的"平均值"和"N"。
- "频率"表格支持 Naïve 汇聚时的"观察 N"。

二项式检验。以下功能受支持：

- "描述"表格支持 Naïve 汇聚时的"平均值"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。以下功能受支持：

- "描述"表格支持 Naïve 汇聚时的"平均值"和"N"。

单样本 Kolmogorov-Smirnov 检验。以下功能受支持：

- "描述"表格支持 Naïve 汇聚时的"平均值"和"N"。

两个独立样本检验。以下功能受支持：

- "等级"表格支持 Naïve 汇聚时的"等级平均值"和"N"。
- "频率"表格支持 Naïve 汇聚时的"N"。

多个独立样本检验。以下功能受支持：

- "等级"表格支持 Naïve 汇聚时的"等级平均值"和"N"。
- "频率"表格支持 Naïve 汇聚时的"计数"。

两个相关样本检验。以下功能受支持：

- "等级"表格支持 Naïve 汇聚时的"等级平均值"和"N"。
- "频率"表格支持 Naïve 汇聚时的"N"。

多个关联样本检验。以下功能受支持：

- "等级"表格支持 Naïve 汇聚时的"等级平均值"。

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"。
- "协变量平均值"表格支持 Naïve 汇聚时的"平均值"。

多重插补选项

"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查看器输出类型。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。

通知

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

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

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

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

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

Intellectual Property Licensing
Legal and Intellectual Property Law
IBM Japan Ltd.
19-21, Nihonbashi-Hakozakicho, Chuo-ku
Tokyo 103-8510, Japan

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

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

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

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

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

IBM Director of Licensing
IBM Corporation

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

此类信息的提供应遵照相关条款和条件，其中包括在某些情况下支付适当费用。

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

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

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

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

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

版权许可：

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

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

商标

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

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

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

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

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

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

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

索引

Printed in China