《统计软件杂志》 MMMMMM YYYY，第 VV 卷，第 II 期。



0.i00

arXiv:2304.02740v1 [stat.CO] 5 Apr 2023

# PStrata：用于主分层的 R软件包

**Bo Liu**杜
克大学

**Fan Li**杜
克大学

**摘要**

治疗后混杂是因果推断中的常见问题，包括不依从、死亡截断、替代终点等特殊情形。主分层（Frangakis andRubin2002）是在存在治疗后混杂的情况下定义和估计因果效应的一般框架。一个突出的特殊情形是随机实验中针对不依从的工具变量方法（Angrist,Imbens, and Rubin1996）。尽管主分层具有多功能性，但由于其固有的潜在混合结构需要复杂的推断工具和高度定制化的编程，绝大多数应用研究人员仍无法使用它。我们开发了R 软件包 **PStrata** ，以自动化处理几种常见情景下的主分层统计分析。**PStrata** 同时支持贝叶斯范式和频率学派范式。对于贝叶斯范式，其计算架构结合了R、C++、Stan，其中 R 提供用户界面，Stan 自动执行后验抽样，而 C++ 通过自动生成 Stan代码将二者连接起来。对于频率学派范式，**PStrata** 实现了三重稳健加权估计量。**PStrata** 支持常规结局和生存时间结局，适用于非结构化数据和聚类数据。

关键词：因果推断、依从者平均因果效应、工具变量、主分层、不依从、死亡截断、Stan。

## 1. 引言

治疗后混杂是因果推断中的常见问题。特殊情形包括但不限于随机实验中的不依从、死亡截断、替代终点、治疗转换，以及整群随机试验中的招募偏倚。主分层（PS）（Frangakis and Rubin 2002）是一个通用框架，可用于在存在治疗后混杂的情况下定义和估计因果效应。随机实验中针对不依从的工具变量方法（Angrist*et al.* 1996）可视为 PS 的一个特例。尽管 PS 用途广泛，但大多数领域

科学家难以使用 PS，因为它需要复杂的建模和推断工具，例如混合模型和高度定制化的编程。我们开发了 R软件包 PStrata，以自动化 PS 几种最常见情形的统计分析。PStrata 同时支持贝叶斯和频率学派推断策略。对于贝叶斯推断，PStrata 的计算架构结合了 R、C++ 和 Stan：R 提供用户界面，Stan 自动执行后验抽样，而 C++ 则通过自动生成相应的 Stan代码将二者连接起来。对于频率学派推断，PStrata 实现了三重稳健加权估计量? 。PStrata 支持常规结局和生存时间结局，并可处理非聚类和聚类数据。

本文回顾并介绍**PStrata**软件包。在第2节中，我们回顾主分层的统计框架及其几个最常见的特殊情形。第3 节介绍**PStrata**中的主要函数。第4节通过两个模拟示例介绍这些函数，示例涉及具有连续结局和生存时间结局的不依从情形。

## 2. 主分层概述

在深入了解**PStrata**的细节之前，我们先简要介绍主分层框架的基础知识。治疗后混杂变量位于治疗与结局之间的因果路径上；它们可能受到治疗的影响，也会影响响应。自Angrist 等人于(1996年发表论文，以及Frangakis andRubin (2002年发表论文以来，围绕这一主题已发展出大量文献。治疗后变量的情形涵盖了广泛的具体例子，例如不依从（如Imbens and Rubin1997）、因死亡而删失的结局（如Rubin 2006；Zhang, Rubin, and Mealli 2009）、替代终点（如Gilbert and Hudgens 2008；Jiang, Ding, and Geng 2016）。下面我们首先在随机实验中的不依从背景下回顾这一框架。

对于来自某个总体的独立同分布（iid）样本中的单位 $i$（$i = 1, \ldots\ldots, N$），设 $Z_i$ 为分配给该单位的治疗（治疗为 1，对照为 0），设 $D_i$ 为实际接受的治疗（治疗为 1，对照为 0）。当 $Z_i \neq D_i$ 时，会发生不依从。由于 $D$ 发生在分配之后，因此它有两个潜在取值，即 $D_i$（0）和 $D_i$（1），其中 $D_i = D_i(Z_i)$。结果 $Y_i$ 也有两个潜在结果，即 $Y_i$（0）和 $Y_i$（1）。
Angrist 等人（1996）根据实际治疗的联合潜在状态 $S_i =$（$D_i$（1） $D_i$（0））， 将单位分为四种依从类型：依从者 $S_i =$（1, 0） = co、从不接受者 $S_i =$（0, 0） = nt、总是接受者 $S_i =$（1, 1） = at，以及违从者 $S_i =$（0, 1） = df。后来，这一分类被推广为主分层（Frangakis and Rubin 2002），而这些依从类型是主分层的特例。
主分层 $S_i$ 的关键特性是，它们不受治疗分配的影响，因此可以被视为治疗前变量。因此，在主分层内比较 $Y_i$(1)和 $Y_i$(0)——即主因果效应（主因果效应）——具有因果解释：

$$\tau_u = \mathbb{E}\{Y_i(1) - Y_i(0) \mid S_i = s\} = \mathbb{E}\{Y_i(1) \mid S_i = s\} - \mathbb{E}\{Y_i(0) \mid S_i = s\},$$

对于 $u =$ nt，co、at，df。在不依从情境中，传统的因果估计目标是意向治疗（ITT）效应，它等同于分配的平均因果效应（ATE），并且可以分解为四个主因果效应之和：

$$\mathrm{ITT} = \mathbb{E}\{Y_i(1) - Y_i(0)\} = \sum_s \pi_s \tau_s,$$

其中 $\pi_s = \Pr(S_i = s)$ 是层 $s$ 的比例。

## 2.1. 存在不依从的随机实验

由于因果推断存在根本性问题，个体的依从分层 $S_i$ 无法观测。因此，要识别分层特定效应，还需要额外的假设。除了 $Z_i$的随机化之外，Angrist 等人(1996)还提出了两个额外假设。

假设**1.**（单调性）： $D(1) \geq D(0)$，即不存在10 违从者分层。

在单调性假设下，在单侧不依从情形中（即对照组单位无法接受治疗），只有两个分层——永不接受者和依从者；而在双侧不依从情形中（即对照组单位也可以接受治疗），则有三个分层——永不接受者、依从者和总是接受者。

假设**2.**（排除限制）：对于分层 $s$，其中 $D(0) = D(1)$，即 $s = 00$，$11$，$E[Y_i(1) - Y_i(0) | S_i = s] = 0$。

在不依从的背景下，排除限制指出，对于从不接受者和总是接受者，分配的因果效应 $Z$ 为0。排除限制排除了从所分配治疗到结果的直接效应，该效应不通过实际治疗介导。

在单调性和排除限制下，依从者平均因果效应是非参数可识别的，为

$$\tau_{\text{co}} \equiv \mathbb{E}\{Y(1) - Y(0) \mid S = \text{co}\} = \frac{\mathbb{E}(Y \mid Z = 1) - \mathbb{E}(Y \mid Z = 0)}{\mathbb{E}(D \mid Z = 1) - \mathbb{E}(D \mid Z = 0)},$$

这恰好是两阶段最小二乘估计量的概率极限Angrist *et al.*(1996）。因为在单调性下，只有依从者的实际治疗受分配影响， $\tau_{\text{co}}$ 可以解释为治疗的效应。

## 2.2. 死亡截断

“死亡截断”是指结果在某些中间变量取特定值的单元上未定义的情况。该名称来源于一个原始例子，其中治疗 $Z$ 是一种医学干预，结果 $Y$ 是研究结束时的生活质量，而$D$ 是患者在研究结束时的存活状态。可以说生活质量仅对存活的患者有定义，因此对于死亡的患者，结果被“死亡截断或删失”。Rubin (2006)提出使用主分层来处理死亡截断。这里四个分层有时被命名为“从不幸存者”、“总是幸存者”、“依从者”和“违从者”。治疗效应仅在“总是幸存者”分层中有良好定义，并被称为“幸存者平均因果效应（SACE）”。单调性通常仍然合理，但排除限制不再适用。文献中已提出各种其他假设。

## 2.3. 倾向得分的基于模型的分析

现在我们来介绍工具变量设置的贝叶斯推断，这一方法最初由 ImbensandRubin(1997) 提出。在不作任何假设的情况下， $Z$ 和 $D$ 的观测单元格由多个层中的个体混合构成

这些个体来自不止一个层。例如，被分配到治疗组并接受治疗的个体（$Z = 1$， $D = 1$）可能属于始终接受者，也可能属于依从者。我们必须从观测数据中区分不同依从类型的因果效应。因此，这里的基于模型的推断类似于混合模型的推断。在贝叶斯分析中，根据某些模型假设填补缺失的标签 $S_i$ 是很自然的。具体而言，使用治疗后变量进行贝叶斯推断时，除了为参数 $\theta$指定先验分布外，还需要指定两个模型：

1. S模型，即给定协变量的主分层模型： $S_i \mid X_i$;
2. Y模型，即给定分层、协变量和治疗的结局模型： $Y_i \mid S_i$， $X_i$， $Z_i$

一种常见的模型选择是针对 $S_i$ 的多项逻辑回归，以及针对 $Y_i$的广义线性模型。我们可以通过迭代地从Pr(D$^{\text{mis}}$ | Y$^{\text{obs}}$,D$^{\text{obs}}$, $\boldsymbol{Z}$, $\boldsymbol{X}$, $\theta$)中插补缺失的 $S$，并从 Pr($\theta$ | Y$^{\text{obs}}$,D$^{\text{obs}}$,D$^{\text{mis}}$, $\boldsymbol{Z}$, $\boldsymbol{X}$) 中更新 $\theta$的后验分布，来模拟联合后验分布 Pr($\theta$,D$^{\text{mis}}$ | Y$^{\text{obs}}$, D$^{\text{obs}}$, $\boldsymbol{Z}$, $\boldsymbol{X}$)。更多详细信息请参见 Imbens and Rubin (1997)；Hirano、Imbens、Rubin 和 Zhou（2000年）。

### 2.4. Frequentist approach under principal ignorability

对于频率学派方法，**PStrata** 实现了 Jiang、Yang 和 Ding （2022）中基于加权的多重稳健估计量。其关键假设是主分层可忽略性（Jo 和 Stuart 2009）。

假设 **3.** *(主忽略性)* $\mathsf{E}[Y(1)|s = 11, X] = \mathsf{E}[Y(1)|s = 10, X]$ 和 $E[Y(0)|s = 00, X] = \mathsf{E}[Y(0)|s = 10, X]$

主分层可忽略性假设排除了主分层的混合结构。另一种方法是构建基于逆概率加权的多重稳健估计量。需要三个量。定义 $\pi(X) = \Pr(Z = 1|X)$ 为倾向得分或治疗概率，并定义 $e_u(X) = \Pr(S = s|X)$，用于 $s = 00, 01, 10, 11$，作为主得分。令$\pi = \mathsf{E}[\pi(X)] = P(Z = 1)$ 和 $e_s = \mathsf{E}[e_u(X)] = \Pr(S = s)$ 分别为边际倾向得分和主得分。此外，定义 $p_z(X) = \Pr(W = 1|Z = z, X)$，用于 $z = 0, 1$，表示在给定治疗和协变量条件下中间变量的概率。在单调性假设下，很容易看出

$$e_{10}(X) = p_1(X) - p_0(X), \quad e_{11}(X) = p_0(X), \quad e_{00}(X) = 1 - p_1(X)$$

and

$$e_{10} = p_1 - p_0, \quad e_{11} = p_0, \quad e_{00} = 1 - p_1$$

其中 $p_z = \mathsf{E}[p_z(X)]$。加权估计量 $\hat{}\tau^{mr}$ 是三个得分的逆概率的乘积：倾向得分、主层得分和中间变量的概率。只要这三个模型中任意两个设定正确，估计量 $\hat{}\tau^{mr}$ 就是一致的。在主层可忽略性成立时，我们仍需要单调性，但不一定需要排除限制。更多详情见Jiang 等人 (2022)。

## 3. 软件包概述

贝叶斯推断主要分为三个步骤。首先，定义数据生成模型，这等同于指定给定参数时所有观测数据的联合似然；同时，指定这些参数的先验分布。其次，从参数的后验分布中抽取样本，该后验分布可直接由先验分布和似然函数得到。最后，利用这些后验抽样计算感兴趣的量，这些量是参数的函数。

该软件包的贝叶斯组件**PStrata** 旨在严格遵循这些步骤，具体如图1所示。首先，用户指定数据生成模型以及这些模型中参数的先验分布。正如我们稍后将看到的，这些数据生成模型可以仅通过 S模型、Y模型和假设来构建。然后，**PStrata**依靠**Stan** 语言从后验分布中抽取样本。在用户指定模型和数据集后，**PStrata** 会自动生成相应的**Stan** 代码和**Stan** 数据。代码和数据由**Stan** 处理，以促进后验抽样过程。最后，这些后验样本会被传回**PStrata** 进行后处理，例如计算估计量、汇总和可视化等。

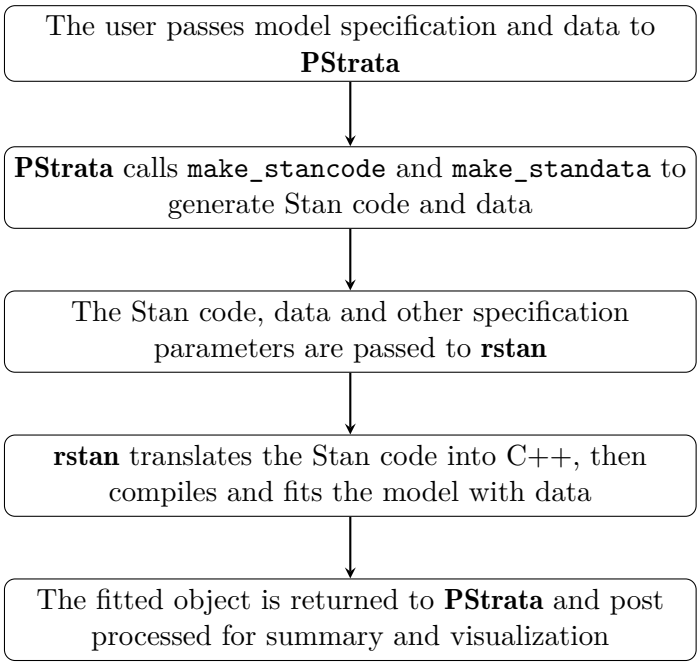


Figure 1: Overview of the model fitting process in **PStrata**

PStrata 中的同名主函数 `PStrata` 被设计为一个连接前端与后端的集成模块。在前端，用户输入模型设定和数据，然后获得后验样本作为输出。后端的抽样过程应与前端完全解耦，以便用户在指定模型时无需担心抽样，这对于 Gibbs 抽样器来说几乎不可能实现。 `PStrata` 包含两个核心子模块 `PSObject`和 `PSSample`以满足这一要求。 `PSObject` 子模块解释用户输入，生成包含模型设定基本信息的主分层对象。 `PSSample`子模块调用 Stan 生成后验样本。这些后验样本在`PStrata` 中返回，并可进一步传递给 `PSOutcome` 和 `PSContrast` 以计算潜在

结果和因果效应。这些模块之间的关系如图2所示。

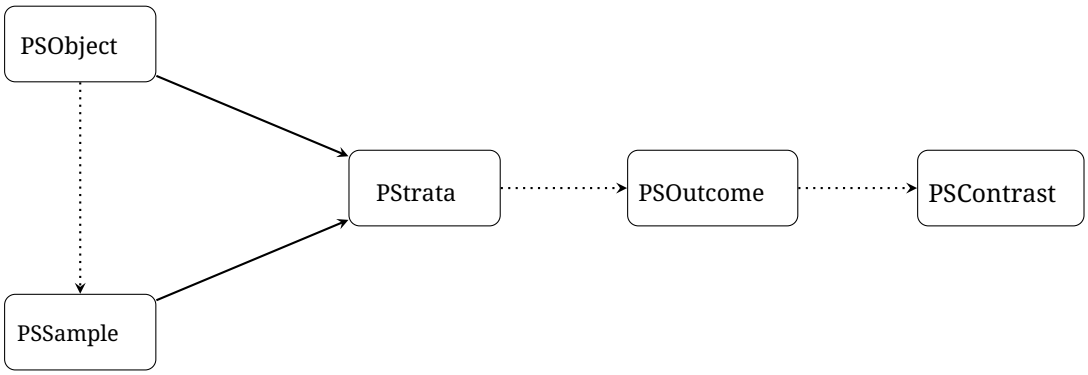


图2：PStrata 中设计的核心模块。实线箭头表示终点模块调用起点模块。虚线箭头表示起点模块的输出作为参数传递给终点模块。

在第 x.x 节中，我们介绍模型设定的细节。在第 x.x 节中，.....

### 3.1. 模型设定

用于主分层的典型数据集包括治疗 $Z$、处理后协变量 $D$、结果 $Y$ 和协变量 $X$。为更好地阐明主要思想，我们假设 $Z$ 和 $D$ 都是二元变量，但这一思想可以很容易地扩展到非二元但离散的 $Z$ 和 $D$的情形，以及包含多个处理后协变量的情形（此时$D$ 是一个向量）。由于 $D$ 是在治疗后观测到的，对 $D$ 进行控制会导致因果效应估计产生偏差，因为这破坏了治疗 $Z$上的随机化。然而，主分层 $S$由 $D$ 在每种治疗下的潜在取值定义，可以被视为治疗前变量（Imbens and Rubin 1997）。当主分层 $S$ 固定时，从 $Z$ 到 $Y$的因果路径上不存在任何变量；因此，可以正确估计分层特异性因果效应（图3）。

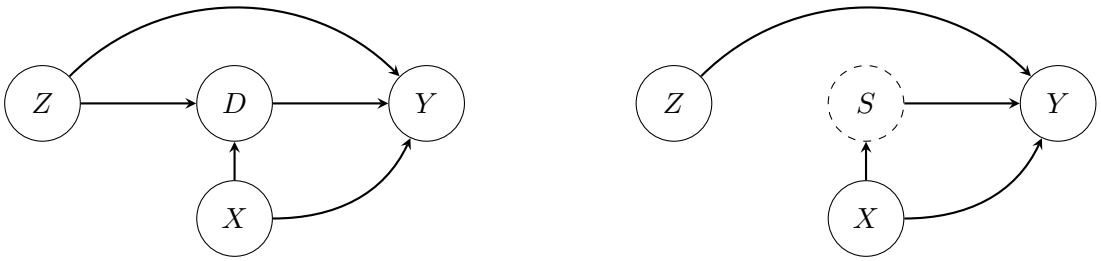


图 3：主分层示意图。左图显示，治疗后协变量 $D$ 位于从治疗 $Z$ 到结果 $Y$的因果路径之间。右图显示，不可观测的主分层 $S$ 将 $D$ 与从 $Z$ 到 $Y$的因果路径分隔开。

If the p如果已知主分层 $S$，那么 Y模型 Pr($Y \mid X$, $Z$, $S$) 就足以进行估计 o

任意因果估计量。然而，由于 $S$ 被定义为 $D$ 的潜在取值，因此只能观测到其中一个实现值。因此，要估计 $S$，就需要建立主分层模型 $\Pr(S \mid X)$。令$\mathcal{S}$ 表示所有可能主分层的集合。数据生成过程可表示如下。

$$\Pr(Z, D, Y \mid X) = \sum_{s \in \mathcal{S}} \Pr(Z \mid X) \Pr(S = s \mid Z, X) \Pr(D \mid S = s, Z, X) \Pr(Y \mid D, S = s, Z, X) \tag{3.1}$$

$$\propto \sum_{s \in \mathcal{S}: D = D(s,Z)} \Pr(S = s \mid X) \Pr(Y \mid S = s, Z, X). \tag{3.2}$$

第一项 $\Pr(Z \mid X)$ 在（3.1）中是随机试验的常数。第三项 $\Pr(D \mid S = s, Z, X)$ 是确定性的，它控制求和所涵盖的主分层集合。$D$ 最后一项中的 $\Pr(Y \mid D, S = s, Z, X)$ 可以省略，因为在已知 $Z$ 和 $S$的情况下，它不包含额外信息。

（3.2）表明，数据生成模型可以表示为仅涉及 S模型和 Y模型的混合模型，我们的目标是从观测数据中解开这种混合。根据实际应用，我们可以施加额外假设，以帮助识别混合模型中的各个组成部分。第一个假设是单调性，它排除了违从者。在某些情况下，如果对照组无法接受治疗，我们也可以排除始终接受者。一般而言，该假设定义了我们纳入（$\mathcal{S}$ 3.1）中的所有分层集合。第二个假设是排除限制（ER）。假设分层 $s$满足 ER 后，我们排除了从 $Z$ 到 $Y$ 的直接因果路径（给定 $S$），因此，

$$\Pr(Y \mid S = s, Z = z_1, X) = \Pr(Y \mid S = s, Z = z_2, X), \quad \text{if } D(s, z_1) = D(s, z_2). \tag{3.3}$$

PStrata 提供了 `PSObject` 函数，用于指定 S模型和 Y模型、设定单调性和 ER 假设，并为模型中出现的参数指定先验分布。该函数的声明如下：

```
PSObject(S.formula, Y.formula, Y.family, data, strata, ER,
+    prior_intercept, prior_coefficient, prior_sigma,+    prior_alpha,
 prior_lambda, prior_theta, survival.time.points)
```

我们将在后续章节详细讨论这些参数，但 `data`和 `survival.time.points`除外。 `data` 参数仅用于提供拟合模型所需的数据集。 `survival.time.points` 参数仅用于生存模型，稍后将在第 x.x 节讨论。

### *S*模型和 *Y*模型

在 PStrata 中，我们将 S模型 $p(S_i \mid X_i)$ 指定为多项模型，

$$\Pr(S_i = s \mid X_i; \{\beta_s\}) \propto \exp(X_i' \beta_s). \tag{3.4}$$

当模型中包含截距时， $X_i$ 在（3.4）中可理解为带有常数 1的增广向量。为确保可识别性，选择一个参考分层 $s_0 \in \mathcal{S}$，并

$\beta_{s_0}$ 固定为零。S模型等价地表示为

$$\Pr(S_i = s \mid X_i; \{\beta_s\}) = \frac{\exp(X_i'\beta_s)}{\sum_{s' \in \mathcal{S}} \exp(X_i'\beta_{s'})}. \tag{3.5}$$

我们将 Y模型指定为广义线性模型，它在对各种类型的数据进行建模方面具有足够的灵活性。Y模型 在数学上可写为

$$Y_i \mid S_i = s, X_i, Z_i = z \sim F(\mu_{sz}(X_i), \theta_{sz}), \tag{3.6}$$

其中 $F$ 是一种通常被称为广义线性模型族的分布。该分布的均值通过连接函数 $g$与 $X_i$ 的线性组合相关:

$$\mu_{sz}(X_i) := \mathsf{E}[Y_i \mid S_i = s, X_i, Z_i = z, \gamma_{sz}] = g^{-1}(X_i'\gamma_{sz}), \tag{3.7}$$

其中 $\{\gamma_{sz}\}$ 是需要估计的参数。某些分布族 $F$ 除了均值之外还可能有其他参数，例如高斯模型中的标准差参数，以及伽马模型和逆高斯模型中的形状参数。当存在这些参数时，我们假设它们与协变量相互独立。但是，我们允许这些参数依赖于分层 $s$ 和治疗 $z$，如（3.6）中 $\theta_{sz}$ 所表示。

S模型和Y模型通过参数 `S.formula`、 `Y.formula` 和 `Y.family` 指定。以下是一个（不完整的）示例。

```
    ...S.formula = Z + D ~ X,
Y.formula = Y ~ X,
Y.family = binomial(link = "logit"), ...
```

`~` 符号之前的术语定义了主层化模型所必需的变量。具体而言， `S.formula` 定义了由 `+` 符号连接的治疗变量 $Z$ 和处理后协变量 $D$，而 `Y.formula` 定义了响应变量 $Y$。当存在多个处理后协变量时，可以按顺序将它们连接起来，例如 `S.formula =Z + D1 + D2 ~ X`。但是，为避免歧义，我们不允许使用多个处理变量。如果存在多个处理变量，可以将它们重新编码为一个包含多个处理组的变量。

符号后的术语定义了包含在 `~` 和3.4中的协变量，其语法与 `lm()`完全一致。回顾基本语法，预测变量之间以 `+` 分隔，如 `Y ~ X1 + X2`所示。默认情况下会包含截距项；使用 `Y ~ X1 + X2 + 0`或 `Y ~ X1 + X2 - 1` 可移除截距项。每个预测变量可以是原始协变量（例如 `X1`）、变换协变量（例如 `I(X1^2)`）或多个协变量的交互作用（例如 `X1:X2`）。

该 `Y.family` 参数指定了广义线性模型的族和连接函数，遵循 `glm()`的约定。PStrata 支持的族和连接函数列于表1。

*Assumptions*

在二元不依从的情形下，有两个重要假设：单调性（假设1）和排除限制（假设2）。单调性排除了违从者存在的可能

| 族 | 连接函数 | $f(y;\mu,\theta)$ |
|---|---|---|
| gaussian | identity, log, inverse | $\frac{1}{\sqrt{2\pi\sigma^2}}$指数$\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$ |
| binomial | logit, probit, cauchit, log, cloglog | $\mu^y(1-\mu)^{1-y}$ |
| Gamma | identity, log, inverse | $\frac{(\alpha/\mu)^\alpha}{\Gamma\alpha} y^{\alpha-1}\exp(-\frac{\alpha y}{\mu})$ |
| poisson | identity, log, sqrt | $\mu^y e^{-\mu}/(y!)$ |
| inverse.gaussian | 1/mu^2, inverse, identity, log | $\sqrt{\frac{\lambda}{2\pi y^3}}\exp\left(-\frac{\lambda(y-\mu)^2}{2y\mu^2}\right)$ |

表1：支持的分布族、连接函数，以及由均值 $\mu$ 和附加参数 $\theta$表示的相应密度/质量函数。

因此，所有需要考虑的分层都是始终接受者、从不接受者和依从者。这可以通过 `strata = c(at = "11", nt = "00", co = "01")`指定。排除限制（ER）假设，对于某些分层， $Z$ 对 $Y$的影响不经过 $D$ ，即不存在直接效应；该假设通过一个与 `strata`长度相同的二元逻辑向量指定，用于指示是否对每个分层采用 ER。例如，`ER = c(True, True, False)` 假设始终接受者和从不接受者满足 ER，但依从者不满足。为简化起见，我们在假设采用 ER 的组别后添加星号（`*`），以简写方式指定这两个假设。例如，上述假设可以通过 `strata = c(at = "11*", nt = "00*", co = "01")`指定，而 `ER` 参数可以留空。

当存在多个中间变量时，主分层由每个中间变量的潜在值定义。要指定这样的主分层，使用 `|` 来分隔每个中间变量的潜在值。例如，由两个中间变量 $D_1(0) = D_1(1) = 0$ 和 $D_2(0) = D_2(1) = 1$ 定义的主分层可以表示为`"00|11"`。

*Priors*

S模型和Y模型都是参数模型。这些参数包括截距（$\alpha_s$ 和 $\gamma_{sz}$）、系数（$\beta_s$ 和 $\eta_{sz}$），以及Y模型族特有的其他参数。如表1所示， `gaussian` 族引入了一个参数 $\sigma$ ，表示误差的标准差；`Gamma` 族引入了形状参数 $\alpha$；而`inverse.gaussian` 族引入了形状参数 $\lambda$。按照惯例，我们将这些参数称为 `sigma`、 `alpha` 和 `lambda` 。

**PStrata** 提供了灵活的方式来指定这些参数（包括截距、系数和模型特定参数）的先验分布。截距和系数的先验分布可分别在 `prior_intercept` 和 `prior_coefficient`中指定。当 Y模型中存在其他参数时，这些变量的先验分布根据其他参数的名称通过 `prior_sigma`、`prior_alpha` 或 `prior_lambda`定义。支持的先验分布完整列表见表2。

默认情况下， `prior_intercept`是整个实数线上的平坦（非正常）分布，`prior_coefficient` 是标准高斯分布 $\mathcal{N}(0,1)$，而 `prior_sigma`、 `prior_alpha`和 `prior_lambda` 是标准逆伽马分布 $\mathcal{IG}(1,1)$。

| Domain | Prior distribution | Specification |
|---|---|---|
| $(-\infty, \infty)$ | Uniform (improper) | `prior_flat()` |
| | Normal | `prior_normal(mean = 0, sigma = 1)` |
| | 学生 t 分布 $t$ | `prior_t(mean = 0, sigma = 1, df = 1)` |
| | 柯西分布 | `prior_cauchy(mean = 0, sigma = 1)` |
| | 双指数分布 | `prior_lasso(mean = 0, sigma = 1)` |
| | Logistic | `prior_logistic(mean = 0, sigma = 1)` |
| $(0, \infty)$ | Chi squared | `prior_chisq(df = 1)` |
| | Inverse Chi squared | `prior_inv_chisq(df = 1)` |
| | Exponential | `prior_exponential(beta = 1)` |
| | Gamma | `prior_gamma(alpha = 1, beta = 1)` |
| | Inverse Gamma | `prior_inv_gamma(alpha = 1, beta = 1)` |
| | Weibull | `prior_weibull(alpha = 1, sigma = 1)` |

Table 2: List of prior distributions.

### 3.2. Other data types

*Survival data*

生存数据是指结果受到右删失的数据类型。假设每个受试者都有真实事件时间 $T_i$ 和删失时间 $C_i$，但由于删失无法观察到它们。观测数据是删失事件时间 $Y_i = \min\{T_i, C_i\}$ 和事件指示符 $\delta_i = \mathbf{1}(T_i < C_i)$。根据数据生成过程，我们假设存在一个潜在分布 $p(T_i, C_i \mid X_i, S_i, Z_i)$，它进而决定观测数量 $p(Y_i, \delta_i \mid X_i, S_i, Z_i)$ 的分布。当删失是条件可忽略的，即 $T_i \perp\!\!\!\perp C_i \mid \{X_i, S_i, Z_i\}$ 时，结果表明 $T_i$ 和 $C_i$ 的分布可以从似然 $p(Y_i, \delta_i \mid X_i, S_i, Z_i)$ 中分离出来:

$$p(Y_i, \delta_i = 1 \mid X_i, S_i, Z_i) = p(T_i \geq Y_i \mid S_i, Z_i, X_i) p(C_i = Y_i \mid S_i, Z_i, X_i), \tag{3.8}$$
$$p(Y_i, \delta_i = 0 \mid X_i, S_i, Z_i) = p(T_i = Y_i \mid S_i, Z_i, X_i) p(C_i \leq Y_i \mid S_i, Z_i, X_i). \tag{3.9}$$

由于我们通常只关心真实事件时间 $T_i$，关于删失时间$C_i$ 的项会从似然函数中消失，因此只需指定 $T_i$ 的分布。

两种广泛使用的模型用于 $T_i$ 分别是 Cox 比例风险模型（Cox 1972）和加速失效时间（AFT）模型（Wei 1992年）。Cox 比例风险模型假设风险函数应当是

$$h(t; S_i = s, Z_i = z, X_i) = h_0(t; s, z) \exp(X_i' \beta_{s,z}), \tag{3.10}$$

其中，$h_0(t; s, z)$ 是基线风险函数。在贝叶斯推断中，$h_0$ 的参数形式是理想的。一种常见形式是威布尔模型（Abrams、Ashby 和 Errington，1996）：$h_0(t; s, z) = \exp(\alpha_{s,z}) t^{\varphi_{s,z}-1}$，$\alpha_{s,z} \in \mathbb{R}$, $\varphi_{s,z} > 0$，在此模型下，Cox 模型 3.10 经重新参数化后可写成参数形式

$$h(t; S_i = s, Z_i = z, X_i) = t^{\exp(\theta_{s,z})-1} \exp(X_i' \beta_{s,z}). \tag{3.11}$$

由风险函数可进一步推导出

$$p(T_i \geq Y_i \mid S_i = s, Z_i = z, X_i) = \exp\left\{-t^{\exp(\theta_{s,z})} \exp(-\theta_{s,z} + X_i'\beta_{s,z})\right\}, \tag{3.12}$$

$$p(T_i = Y_i \mid S_i = s, Z_i = z, X_i) = h(t; s, z, X_i) p(T_i \geq Y_i \mid S_i = s, Z_i = z, X_i). \tag{3.13}$$

AFT 模型假设 $\log(T_i) = X_i'\beta + \epsilon_i$，其中 $\epsilon_i$ 服从给定的分布。通常，我们假设 $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$。然后，可以计算得出

$$p(T_i \geq Y_i \mid S_i = s, Z_i = z, X_i) = 1 - \Phi\left(\frac{\log Y_i - X_i'\beta_{s,z}}{\sigma_{s,z}}\right), \tag{3.14}$$

$$p(T_i = Y_i \mid S_i = s, Z_i = z, X_i) = \phi\left(\frac{\log Y_i - X_i'\beta_{s,z}}{\sigma_{s,z}}\right), \tag{3.15}$$

其中，$\Phi$ 和 $\phi$ 分别是标准正态分布的累积密度函数和概率密度函数。

PStrata 允许在条件可忽略性假设下对生存数据拟合主分层模型。在生存数据上运行模型与处理常规数据相似，但需要进行一些修改。首先，在指定 Y模型时，需要同时包含删失事件时间 $Y$ 和事件指标 $\delta$。例如，`Y.formula = Y + delta ~ X`指定了一个Y模型，其中 `Y` 和 `delta` 都出现在公式的左侧。其次，`Y.family` 对于Weibull-Cox模型应为 `survival(method = "Cox")`，对于AFT 模型应为 `survival(method = "AFT")`。第三，由于拟合的结局——时间 $t$ 处的生存概率——是 $t$的连续曲线，因此需要指定一组用于计算生存概率的时间点。可以将自定义的时间点作为向量指定给参数 `survival.time.points`，或者直接提供一个正整数来表示时间点的数量；此时，时间点将从 0 到观测事件时间的 90% 分位数之间均匀选取。

### 多层数据

在前面的章节中，我们假设各单位是来自总体的独立样本。在一些应用中，如果存在自然形成的集群，这一假设可能不太合理，例如学校、医院和地区。举例来说，在多家医疗中心开展的试验中，来自同一医疗中心的患者，其结局可能存在相关性。层次模型通常用于对这类数据建模：假设每个集群都有一个集群特异性效应（随机效应），该效应来自某个未知分布；而在每个集群内部，各单位的结局服从与该集群特异性效应相关的某种分布。层次结构非常适合贝叶斯分析，PStrata 允许在 S模型和 Y模型中指定多层结构。

我们介绍最常见的情形，即只有一个层级的聚类，不过 PStrata 也能灵活地对各种聚类结构进行建模。设 $j = 1, \ldots, J$ 表示各个聚类，$C_i$ 表示个体 $i$ 所属的聚类。除个体协变量 $X_i$ 外，设 $V_j$ 为聚类 $j$ 的聚类协变量。在假设随机效应为加性的情况下，多层级 S模型假设

$$\Pr(S_i = s \mid X_i, V_j, C_i = j) \propto \exp(X_i'\beta_s + V_j'\widetilde{\beta}_s + \eta_j), \tag{3.16}$$

其中 $\eta_j \sim \mathsf{N}(0, \sigma^2)$ 为聚类随机效应。类似地，Y模型的均值定义为

$$\mathsf{E}(Y_i \mid Z_i = z, X_i, V_j, C_i = j) = g^{-1}(X_i'\gamma_{sz} + V_j'\widetilde{\gamma}_{sz} + \xi_{sz,j}), \tag{3.17}$$

其中$\xi_{sz,j} \sim \mathsf{N}(0, \tau_{sz}^2)$ 是聚类随机效应。
当存在集群时，PStrata 支持纳入这些随机效应，方法是在 `S.formula` 和/或 `Y.formula` 中使用与 `lme4::lmer()` 相同的语法指定这些效应。纳入上述加性随机效应的语法是加入一项 `(1 | C)`，其中 `C`是集群 ID。例如，上文提到的Y模型可以指定为 `Y.formula = Y ~ X + V + (1 | C)`，其中 `C` 是一个向量，表示每个单元所属的集群 ID。

~随机效应的加性假设可以放宽，以允许系数 $\beta_s$、 $\beta_s$、$\gamma_{sz}$、 $\tilde{\gamma}_{sz}$ 在不同集群之间变化。例如，系数 $\beta_s$ 上的随机效应将修改S模型，如下所示

$$\Pr(S_i = s \mid X_i, V_j, C_i = j) \propto \exp(X_i'(\beta_s + \omega_j) + V_j'\widetilde{\beta}_s + \eta_j), \tag{3.18}$$

其中$\eta_j$ 和$\omega_j$ 是分别从分布$\eta_j$ ~ $\mathsf{N}(0, \sigma^2)$$\omega_j$ ~$\mathsf{N}(0, \kappa_j^2)$中抽取的随机效应。在文献中，$\eta_j$通常被称为随机截距，$\omega_j$通常被称为随机斜率。

在 PStrata 中，上述模型可指定为 `S.formula = S ~ X + V + (X | C)`，其中，项 `(X | C)` 表示在 $X$ 上设置随机斜率，并根据群组 $C$ 设置（隐式）随机截距。要移除隐式随机截距，应将模型指定为 `S.formula = S ~ X + V + (X - 1 | C)` 或 `S.formula = S ~ X + V + (X + 0| C)`。有关语法的更多详情，请参阅 `lme4::lmer()`的文档。

### 3.3. Outcome and contrast

PStrata 调用Stan 从参数的后验分布中抽取样本。这些后验样本可用于计算汇总统计量，例如分层概率、分层特异性潜在结果及其对比。

对于每个参数，我们用括号中的上标表示每次迭代中抽取的样本。例如，从后验分布 $\alpha$ 中抽取的样本记为$\alpha^{(1)}, \alpha^{(2)}, \ldots, \alpha^{(B)}$，其中 $B$是后验抽样次数。

*Stratum probability*

回顾多项逻辑S模型（3.4）。在第 $k$ 次迭代中，单位$i$ 属于层 $s$ 的估计概率可通过代入该次迭代的后验抽样结果获得，

$$p_{is}^{(k)} := p(S_i = s \mid X_i) = \frac{\exp(\beta_s^{(k)\prime} X_i)}{\sum_{\varsigma \in \mathcal{S}} \exp(\beta_\varsigma^{(k)\prime} X_i)}. \tag{3.19}$$

通过在每次迭代中计算该量，我们得到一系列估计的分层概率 $p_{is}^{(1)}$，……，$p_{is}^{(B)}$，可将其视为从 $p_{is}$ 的后验分布中抽取的样本。基于这些后验样本，PStrata 会计算并报告估计分层概率的期望、标准差和分位数。

## 结果

令 $\mu_{sz}$ 表示分层 $s$ 在治疗 $z$ 下的平均潜在结局，其定义为

$$\begin{aligned}\mu_{sz} &= \mathsf{E}[Y(z) \mid S = s] \\ &= \mathsf{E}[\mathsf{E}[Y(z) \mid S = s, X] \mid S = s] \\ &= \int \mathsf{E}[Y(z) \mid S = s, X = x] \frac{\Pr(S = s \mid X = x)}{\Pr(S = s)} \mathbb{P}(\mathrm{d}x) \\ &= \frac{\int \mathsf{E}[Y(z) \mid S = s, X = x] \Pr(S = s \mid X = x) \mathbb{P}(\mathrm{d}x)}{\int \Pr(S = s \mid X = x) \mathbb{P}(\mathrm{d}x)}.\end{aligned} \tag{3.20}$$

为了估计这样的量，我们用 $\mathbb{P}(\mathrm{d}x)$ 在数据中的经验分布来近似测度 $X$ ，因此两个积分都可以化简为对每个单位求和。在积分中，第一项$\mathsf{E}[Y(z) \mid S = s, \ X = x]$ 由 Y模型（3.7）给出，

$$\mathsf{E}[Y_i(z) \mid S_i = s, X_i] = g^{-1}(\gamma_{sz}' X_i). \tag{3.21}$$

项 $\Pr(S = s \mid X = x)$ 正是上一节定义的层概率 $p_{is}$ 。这表明可以通过以下方式估计潜在结果 $\mu_{sz}$

$$\mu_{sz} \approx \frac{\sum_{i=1}^{n} g^{-1}(\gamma_{sz}' X_i) p_{is}}{\sum_{i=1}^{n} p_{is}}. \tag{3.22}$$

将参数 $\gamma_{sz}$ 和层概率 $p_{is}$的后验样本代入后，我们得到后验抽样值 $\mu_{sz}^{(1)}$，…，$\mu_{sz}^{(B)}$，其中

$$\mu_{sz}^{(k)} = \frac{\sum_{i=1}^{n} g^{-1}(\gamma_{sz}^{(k)\prime} X_i) p_{is}^{(k)}}{\sum_{i=1}^{n} p_{is}^{(k)}}. \tag{3.23}$$

估计潜在结果的期望、标准差和分位数可直接根据这些后验抽样值计算。

## 对比

许多具有统计意义的估计量都可以表示为潜在结果的对比。例如，分层 $s$ 的 SATE 是两个潜在结果之差:

$$\tau_s = \mathsf{E}[Y(1) - Y(0) \mid S = s] = \mu_{s1} - \mu_{s0}. \tag{3.24}$$

还可以通过以下方式比较分层 $s$ 和 $s'$ 之间的 SATE:

$$\Delta_{s,s'} = \tau_s - \tau_s' = (\mu_{s1} - \mu_{s0}) - (\mu_{s'1} - \mu_{s'0}) = (\mu_{s1} - \mu_{s'1}) - (\mu_{s0} - \mu_{s'0}). \tag{3.25}$$

要获得这些对比的后验分布，可将潜在结果的后验样本 $\mu_{sz}^{(k)}$代入 $\mu_{sz}$，从而得到一系列对比值。这些值可视为来自该对比后验分布的样本，进而可直接求得期望、标准差和分位数。

PStrata 允许对分层或处理组进行对比，并且可以按任意顺序嵌套，从而确保在定义各种有意义的对比时具有灵活性。As

例如， $\tau_s$ 是处理组之间的对比，而$\Delta_{s,s'}$ 是先进行处理组之间的对比，再进行分层之间的对比，或顺序相反。

### 生存数据

对于生存数据，最常见的因果估计量是生存概率，定义为 $\Pr(T_i > t)$。令 $\mu_{sz}(t)$ 表示在治疗 $z$下，分层 $s$ 在时间 $t$ 的平均生存概率：

$$\mu_{sz}(t) = \Pr(T > t \mid S = s, Z = z) \tag{3.26}$$

$$= \frac{\int \Pr(T > t \mid S = s, Z = z, X = x) \Pr(S = s \mid X = x) \mathbb{P}(\mathrm{d}x)}{\int \Pr(S = s \mid X = x) \mathbb{P}(\mathrm{d}x)}. \tag{3.27}$$

对于任意给定的 $t$，术语 $\Pr(T > t \mid S = s, \ Z = z, \ X = x)$ 可通过 Cox 模型下的（3.12）或 AFT 模型下的（3.14）计算。使用 $X$的经验分布求和来近似积分，

$$\mu_{sz}(t) = \frac{\sum_{i=1}^{n} \Pr(T_i > t \mid S_i = s, Z_i = z, X_i) p_{is}}{\sum_{i=1}^{n} p_{is}}. \tag{3.28}$$

在 **PStrata**中，直接将 $\mu_{sz}$ 计算为一个函数并不可行。因此，会预先指定一组时间点 $t_1$、...、$t_m$，并在这些时间点上计算 $\mu_{sz}$，$\mu_{szj} := \mu_{sz}(t_j)$。将参数的后验样本和 $p_{is}$ 代入（3.28），以获得 $\mu_{szj}$的期望、标准差和分位数。除分层和处理组外，生存数据还支持不同时间点之间的对比。

### *Multi-level data*

我们介绍了在 S模型（3.16）和 Y模型（3.17）下多层级数据的结局与对比量计算，但这一思想可以轻松扩展到更复杂的模型形式。在每次迭代中，我们不仅从后验分布中抽样系数，还抽样聚类随机效应。具体而言，令 $\beta_s^{(k)}$， $\widetilde{\beta}_s^{(k)}$为系数，并令 $\eta_j^{(k)}$ 为第 $k$ 次迭代中抽样得到的随机效应。此外，令 $C_i$ 表示受试者 $i$ 所属的聚类。于是，分层概率可估计为

$$p_{is}^{(k)} = \frac{\exp(X_i' \beta_s^{(k)} + V_{C_i}' \widetilde{\beta}_s^{(k)} + \eta_{C_i}^{(k)})}{\sum_{\varsigma \in \mathcal{S}} \exp(X_i' \beta_\varsigma^{(k)} + V_{C_i}' \widetilde{\beta}_\varsigma^{(k)} + \eta_{C_i}^{(k)})}. \tag{3.29}$$

Similarly, the average potential outcome can be estimated by

$$\mu_{sz}^{(k)} = \frac{\sum_{i=1}^{n} g^{-1}(X_i' \gamma_{sz}^{(k)} + V_{C_i}' \widetilde{\gamma}_{sz}^{(k)} + \xi_{sz,C_i}^{(k)}) p_{is}^{(k)}}{\sum_{i=1}^{n} p_{is}^{(k)}}, \tag{3.30}$$

其中， $\gamma_{sz}^{(k)}$、 $\widetilde{\gamma}_{sz}^{(k)}$和 $\xi_{sz}^{(k)}$ . $C_i$ 是第 $k$ 次迭代中获得的后验分布样本。根据这些样本，可以直接计算结果和对比的后验分布的期望、标准差和分位数。

## 4. Case study: Effect of flu vaccination on hospital visits

我们通过一项**PStrata** 案例研究，展示其在正态不依从情形下的应用。该研究利用 Mcdonald、Hui 和 Tierney(1992年) 首先研究的数据估计流感疫苗的效果，随后由 Hirano等人 (2000年) 重新分析。在这项研究中，医生被随机选中并收到一封信，信中鼓励他们为患者接种流感疫苗。关注的治疗是患者的实际疫苗接种状态，结果是一个二元变量，表示他们是否因流感相关原因住院就诊。每位患者都观测到一些协变量；按照 Hirano 等人 (2000年) 的做法，我们在 S模型和 Y模型中纳入年龄以及慢性阻塞性肺疾病（COPD）发生率。

在本案例中，随机分配的处理 $Z_i$ 是鼓励指标，结果 $Y_i$是住院就诊指标。实际疫苗接种状态 $D_i$是随机化后中间变量。传统的意向治疗方法估计 $Z_i$ 对$Y_i$的因果效应，该效应估计的是鼓励的有效性，而不是疫苗接种本身的效力。使用主分层框架，我们根据潜在疫苗接种状态 $S_i = (D_i(0)\ D_i(1))$定义主分层，即始终接种者 $S_i = (1,1)$、从未接种者 $S_i = (0,0)$、服从者接种者 $S_i = (0,1)$和违抗者接种者$S_i = (1,0)$。由于无论是否受到鼓励，都可以进行流感疫苗接种，且并非强制接种，因此可能存在始终接种者或从未接种者患者。我们假设不存在违抗者接种者患者，但其他三个分层中均有患者。此外，我们假设，对于始终接种者和从未接种者患者而言，接种疫苗方面的鼓励对流感相关住院就诊本身没有任何直接效应。我们感兴趣的是估计服从者接种者患者中的 $Z_i$ 因果效应，因为这种因果效应估计的是该亚群的疫苗接种效力。

*Create Principal Stratification Object (PSObject)*

```
R> PSobj <- PSObject(
+     S.formula = encouragement + vaccination ~ age + copd,
+     Y.formula = hospital ~ age + copd,
+     Y.family = binomial(link = "logit"),+     data = flu,
+     strata = c(n = "00*", c = "01", a = "11*"),
+     prior_intercept = prior_normal(0, 1),
+     prior_coefficient = prior_normal(0, 1)+   )
```

在上述代码片段中，我们通过指定 S模型和 Y模型、提供 数据集 并指定我们所作的假设，创建了一个主分层对象 `PSobj`。
在 `S.formula` 中，我们在左侧提供 $Z$ 和 $D$ 的变量名称，在右侧提供 S模型的协变量。这指定了多项模型。

$$\Pr(S = s \mid X_1, X_2) \propto \exp(\alpha_s + \beta_{1,s}X_1 + \beta_{2,s}X_2), \tag{4.1}$$

其中，$X_1$ 和 $X_2$ 分别指年龄和慢性阻塞性肺疾病，$\alpha_s$、$\beta_{1,s}$ 和$\beta_{2,s}$ 为参数。同样，在 `Y.formula`中，我们在等号左侧提供 $Y$ 的变量名，在右侧提供Y模型的协变量。这与 `Y.family` 一起指定了结局

将模型作为广义线性模型

$$Y \mid S = s, Z = z, X_1, X_2 \sim \mathrm{Ber}(\mu_{s,z}(X_1, X_2)), \tag{4.2}$$
$$\operatorname{logit} \mu_{s,z}(X_1, X_2) = \gamma_{s,z} + \delta_{1,s,z} X_1 + \delta_{1,s,z} X_2, \tag{4.3}$$

其中，$\gamma_{s,\ z}$, $\delta_{1,\ s,\ z}$ 和 $\delta_{2,\ s,\ z}$ 是参数。
我们假设存在三个可能的分层，由 $n$ =(0，0)、 $c$ =(0， 1) 和$a$ =(1， 1)定义。我们在 $n$ =(0，0) 和 $a$ =(1，1) 上标注星号，因为我们假设这些分层满足 ER，这意味着

$$(\gamma_{n,0}, \delta_{1,n,0}, \delta_{2,n,0}) = (\gamma_{n,1}, \delta_{1,n,1}, \delta_{2,n,1}), \tag{4.4}$$
$$(\gamma_{a,0}, \delta_{1,a,0}, \delta_{2,a,0}) = (\gamma_{a,1}, \delta_{1,a,1}, \delta_{2,a,1}). \tag{4.5}$$

最后，我们为截距 $\alpha_s$, $\gamma_{s,\ z}$ 以及系数 $\beta_{1,\ s}$, $\beta_{2,\ s}$,$\delta_{1,\ s,\ z}$ 和 $\delta_{2,\ s,\ z}$指定先验分布。在本例中，我们假设所有这些参数的先验分布均为标准正态分布。

### 拟合主分层模型

我们使用函数 PStrata拟合该模型。

```
R> fit <- PStrata(+    PSobj,
+    warmup = 1000, iter = 2000,
+    cores = 6, chains = 6+  )
```

在上述代码片段中，我们通过抽取 6 条链来拟合贝叶斯主分层模型，每条链包含2000年次迭代，其中包括1000 次预热迭代。为提高效率，抽样任务被分配到 6 个核心上。这些参数会直接传递给 rstan::stan()，详细信息请参阅 rstan 手册。

```
R> summary(fit)
```

```
        mean         sd      2.5%       25%    median       75%     97.5%
n 0.6940673 0.01196718 0.6707998 0.6859948 0.6940854 0.7022795 0.7170083
c 0.1137423 0.01621615 0.0803734 0.1034171 0.1138241 0.1246476 0.1455330
a 0.1921904 0.01055071 0.1717857 0.1850637 0.1918401 0.1992098 0.2128891
```

根据拟合模型，69.4%（CI：67.1%至 71.7%）的患者属于从未接种者，19.2%（CI：17.2%至21.3%）的患者属于始终接种者，其余 11.4%（CI：8.0% 至14.6%）属于服从者接种者。

### 预测平均效应

```
R> outcome <- PSOutcome(fit)
R> summary(outcome, "matrix")
```

```
        mean       sd     2.5%      25%   median      75%    97.5%
n:0 0.081546 0.007444 0.067090 0.076531 0.081497 0.086451 0.096345
n:1 0.081546 0.007444 0.067090 0.076531 0.081497 0.086451 0.096345
c:0 0.166652 0.063261 0.060931 0.122136 0.160119 0.204622 0.310435
c:1 0.068589 0.027496 0.024559 0.048624 0.065305 0.085192 0.129120
a:0 0.100001 0.014312 0.073536 0.090090 0.099535 0.109130 0.129424
a:1 0.100001 0.014312 0.073536 0.090090 0.099535 0.109130 0.129424
```

我们使用 `PSOutcome` 计算分层 $s$ 和治疗组 $z$下的平均效应，定义如下

$$\mu_{s,z} = \mathbb{E}[Y(z) \mid S = s]. \tag{4.6}$$

从输出结果可以看出，从未接种者和始终接种者在两个组中的平均结局均为 0.082（置信区间：0.067 至 0.096）和 0.100（置信区间：0.074 至 0.129）。对于这两个分层，治疗组和对照组的平均结局相同，这是 ER 假设的结果。对于服从者接种者，对照组的平均结局为 0.167（置信区间：0.061 至 0.310），治疗组的平均结局为 0.069（置信区间：0.026 至 0.129）。

### 计算对比

要确定治疗组与对照组之间平均结局的差异，可以使用 `PSContrast` 函数。

```
R> contrast <- PSContrast(outcome, Z = TRUE)
R> summary(contrast, "matrix")
```

```
                mean       sd      2.5%       25%    median       75%    97.5%
n:{1}-{0}  0.000000 0.000000  0.000000  0.000000  0.000000  0.000000 0.000000
c:{1}-{0} -0.098063 0.068129 -0.246702 -0.140799 -0.092666 -0.052169 0.022559
a:{1}-{0}  0.000000 0.000000  0.000000  0.000000  0.000000  0.000000 0.000000
```

在上面的代码片段中， `Z = TRUE` 指定我们希望计算治疗 $Z$所有水平之间的对比。由于 ER 假设，从未接种者和始终接种者的对比均为零。服从者的对比也称为服从者平均因果效应，为 -0.098（置信区间：-0.247，0.023）。

### 结论

根据主分层分析，平均而言，鼓励接种流感疫苗会使服从者接种者的住院就诊平均比例从 16.7% 降至 6.9%。在 95% 的置信水平下，这一差异在统计学上并不显著，但置信区间仅以微小幅度包含零。主分层分析还成功识别出由主分层定义的这些亚群之间的差异。例如，从未接种者的住院就诊平均比例为 8.2%，而始终接种者的住院就诊平均比例则反直觉地更高，为 10.0%。这可能是因为无论如何都不会接种疫苗的人认为自己风险较低，因此即使不接种疫苗，出现严重症状的可能性也较低。

# 5. Simulation

在本节中，我们开展模拟研究，以说明如何在不同情景下使用该软件包。

## 5.1. Case 1: Two-sided noncompliance

此模拟研究呈现了一个经典的不依从情境，其中违从者（$S = (1,0)$）被排除在分析之外，并假设总是接受者（$S = (1,1)$）和从不接受者（$S = (0,0)$）中均存在 ER。

我们生成10000个单位。两个协变量 $X_1, X_2$ 独立地从 $\mathcal{N}(0, 1)$中抽样。我们独立地分配 $S \in \{(0, 0),(0, 1),(1, 1)\}$，其概率分别为0.3、0.5和0.2。治疗分配对每个单位都是随机的，具有 $p(Z = 1) = 0.5$，以模拟随机临床试验。给定分层 $S$ 和治疗分配 $Z$，结果 $Y$服从高斯分布，其中

$$
\begin{aligned}
Y \mid S = (0,0), Z = z, X_1, X_2 &\sim \mathcal{N}(X_1 - X_2 + X_1 X_2, 0.3) \\
Y \mid S = (0,1), Z = z, X_1, X_2 &\sim \mathcal{N}(2X_1 - (1+z)X_2 + 2 + 6z, 0.2) \\
Y \mid S = (1,1), Z = z, X_1, X_2 &\sim \mathcal{N}(X_1 + X_2 - 1, 0.2).
\end{aligned}
$$

我们使用6条链运行采样器，每条链进行1000次预热迭代和1000次采样迭代。层概率和因果效应的真实值以及相应的后验分布在表 3.

```
R> fit <- PStrata(+    S.formula = Z + D ~ 1,
+    Y.formula = Y ~ X1 * X2,
+    Y.family = gaussian(),+    data = data,
+    strata = c(n = "00*", c = "01", a = "11*"),
+    prior_intercept = prior_normal(0, 1),
+    warmup = 1000, iter = 2000,+    cores = 6,
 chains = 6, refresh = 10+  )R> summary(fit)
```

```
        mean          sd      2.5%       25%    median       75%     97.5%
n 0.2888397 0.004491325 0.2800518 0.2858500 0.2888168 0.2918257 0.2976134
c 0.5046871 0.004968135 0.4948485 0.5014518 0.5046069 0.5079710 0.5145919
a 0.2064732 0.003959123 0.1988303 0.2037680 0.2064324 0.2091346 0.2143214
```

```
R> outcome <- PSOutcome(fit)
R> contrast <- PSContrast(outcome, Z = TRUE)
R> summary(contrast, "matrix")
```

```
               mean       sd     2.5%      25%  median      75%    97.5%
n:{1}-{0} 0.000000 0.000000 0.000000 0.000000 0.00000 0.000000 0.000000
c:{1}-{0} 5.992352 0.005649 5.981134 5.988609 5.99229 5.996191 6.003188
a:{1}-{0} 0.000000 0.000000 0.000000 0.000000 0.00000 0.000000 0.000000
```

| | True value | Posterior mean | 2.5% quantile | 97.5% quantile |
|---|---|---|---|---|
| $\Pr(S=(0,0))$ | 0.300 | 0.289 | 0.280 | 0.298 |
| $\Pr(S=(0,1))$ | 0.500 | 0.505 | 0.495 | 0.515 |
| $\Pr(S=(1,1))$ | 0.200 | 0.206 | 0.199 | 0.214 |
| $\mathrm{E}[Y(1)-Y(0) \mid S=(0,1)]$ | 6.000 | 5.992 | 5.981 | 6.003 |

Table 模拟1中重要量的后验摘要及真实值。

### 5.2. Case 2: Two post-treatment variables

本模拟研究包含一个更复杂的情景，其中存在两个处理后协变量$D_1$ 和 $D_2$ 。令主分层由 $S=(D_1(0)\ D_1(1)\ D_2(0)\ D_2(1))$ 定义。本研究纳入 16 个分层中的 5 个，即 $\mathcal{S}=\{0000,0001,0101,0011,1111\}$，并假设 0000、0011 和 1111 满足 ER。本研究未纳入基线协变量。

我们模拟了 10,000 个数据，以便更好地识别分层。我们随机将主分层 $S$ 和治疗状态 $Z$ 分配给每个单位，其中分层分配概率为 $p=(0。15，0。3，0。2，0。2，0。15)$，治疗分配概率为 $P(Z=1)=0.5$。结果变量 $Y$ 从表4中给出的高斯分布进行抽样。

| | Probability | $Z=0$ | $Z=1$ |
|---|---|---|---|
| $S=0000$ | 0.15 | $\mathcal{N}(3,1)$ | |
| $S=0001$ | 0.30 | $\mathcal{N}(-1,0.5)$ | $\mathcal{N}(-3,0.5)$ |
| $S=0011$ | 0.20 | $\mathcal{N}(2,0.5)$ | $\mathcal{N}(5,0.5)$ |
| $S=0101$ | 0.20 | $\mathcal{N}(-1,3)$ | |
| $S=1111$ | 0.15 | $\mathcal{N}(1,2)$ | |

Table 4: Outcome specification for Simulation 2

我们使用6 条链运行采样器，每条链进行1000次预热迭代和 1000 次采样迭代。参数的真实值及相应的后验均值见表 5。

```
R> fit <- PStrata(+     S.formula = Z + D ~ 1,+     Y.formula = Y ~ 1,
+     Y.family = gaussian(),+     data = data,+     strata = c(
+       nn = "00|00*", nc = "00|01", cc = "01|01", na = "00|11*", aa = "11|11*"
+     ),+     prior_intercept = prior_normal(0, 1),+     warmup = 1000, iter = 2000,
+     cores = 6, chains = 6+   )R> summary(fit)
```

```
         mean          sd      2.5%       25%    median       75%     97.5%
nn 0.1536845 0.006951125 0.1412590 0.1481628 0.1537814 0.1591534 0.1663185
nc 0.2784839 0.023542175 0.2465955 0.2555127 0.2798529 0.3014315 0.3093388
cc 0.2263757 0.026944918 0.1922943 0.1997387 0.2288498 0.2529571 0.2599586
na 0.1954942 0.005190111 0.1855834 0.1918809 0.1954155 0.1990191 0.2056500
aa 0.1459617 0.003763111 0.1386817 0.1434026 0.1459706 0.1484749 0.1533925
```

```
R> outcome <- PSOutcome(fit)
R> summary(outcome, "matrix")
```

```
          mean       sd      2.5%       25%    median       75%     97.5%
nn:0  3.026044 0.037910  2.952770  3.000045  3.025884  3.052389  3.099262
nn:1  3.026044 0.037910  2.952770  3.000045  3.025884  3.052389  3.099262
nc:0  0.522369 1.518245 -1.016887 -0.995661  0.503381  2.039995  2.077103
nc:1 -2.981742 0.016428 -3.014958 -2.992721 -2.981645 -2.970739 -2.949451
cc:0  0.503428 1.501155 -1.018944 -0.997362  0.490139  2.004243  2.040805
cc:1  5.008498 0.018250  4.971816  4.996631  5.008637  5.021135  5.043137
na:0 -1.081677 0.075050 -1.227409 -1.133591 -1.081889 -1.031386 -0.933170
na:1 -1.081677 0.075050 -1.227409 -1.133591 -1.081889 -1.031386 -0.933170
aa:0  0.940070 0.062138  0.817387  0.897509  0.940663  0.982155  1.061605
aa:1  0.940070 0.062138  0.817387  0.897509  0.940663  0.982155  1.061605
```

```
R> contrast <- PSContrast(outcome, Z = TRUE)
R> summary(contrast, "matrix")
```

```
               mean       sd      2.5%       25%    median       75%     97.5%
nn:{1}-{0}  0.00000 0.000000  0.000000  0.000000  0.000000  0.000000  0.000000
nc:{1}-{0} -3.50411 1.520898 -5.072012 -5.024450 -3.497351 -1.983469 -1.949683
cc:{1}-{0}  4.50507 1.496066  2.963890  3.009801  4.513690  6.000999  6.035906
na:{1}-{0}  0.00000 0.000000  0.000000  0.000000  0.000000  0.000000  0.000000
aa:{1}-{0}  0.00000 0.000000  0.000000  0.000000  0.000000  0.000000  0.000000
```

| | 真实值 | 后验均值 | 2.5% 分位数 | 97.5% 分位数 |
|---|---|---|---|---|
| $\Pr(S = 0000)$ | 0.150 | 0.153 | 0.141 | 0.166 |
| $\Pr(S = 0001)$ | 0.300 | 0.278 | 0.247 | 0.309 |
| $\Pr(S = 0011)$ | 0.200 | 0.226 | 0.192 | 0.260 |
| $\Pr(S = 0101)$ | 0.200 | 0.195 | 0.186 | 0.206 |
| $\Pr(S = 1111)$ | 0.150 | 0.146 | 0.139 | 0.153 |
| $\mathrm{E}[Y(1) - Y(0) \mid S = 0001]$ | -2.000 | -3.504 | -5.072 | -1.950 |
| $\mathrm{E}[Y(1) - Y(0) \mid S = 0101]$ | 3.000 | 4.505 | 2.964 | 6.046 |

Table 5: 模拟2中重要量的后验摘要及其真实值。

### 5.3. 案例3：时间到事件结局下的不依从

我们模拟了一项包含 $N = 10,000$ 个单位的随机实验，设定为存在不依从的情形。我们以 $S \in \{(0,0),(0,1),(1,1)\}$ 分别以 0.25、0.60 和 0.15 的概率独立分配。治疗分配 $Z_i$ 独立地从 $Bernoulli(0.5)$中抽取。真实的未删失失效时间根据 Weibull-Cox 模型（3.11）生成，并针对分层与治疗分配的六种组合分别设定参数。我们不纳入协变量，并包含截距，此时 Weibull-Cox 模型的形式为

$$h(t; S_i = s, Z_i = z) = t^{\exp(\theta_{s,z})-1} \exp(\alpha_{s,z}). \tag{5.1}$$

表 6给出了这两种情形下的T模型参数。我们抽取删失时间 $C_i$ 独立地来自速率为0.3的指数分布，使得事件发生率约为35%。

| | | $\Pr(Y(0) \mid S = s)$ | | $\Pr(Y(1) \mid S = s)$ | |
|---|---|---|---|---|---|
| $s$ | $\Pr(S = s)$ | $\theta_{s,0}$ | $\alpha_{s,0}$ | $\theta_{s,1}$ | $\alpha_{s,1}$ |
| $n$ | 0.25 | $\log(2.0)$ | -4.0 | $\log(1.5)$ | -3.0 |
| $c$ | 0.60 | $\log(1.5)$ | -2.5 | $\log(1.0)$ | -1.5 |
| $a$ | 0.15 | $\log(1.0)$ | -1.0 | $\log(0.6)$ | 0.0 |

表 6：模拟 3 中真实 S模型和 T 模型的参数。

```
R> fit <- PStrata(+    S.formula = Z + D ~ 1,
+    Y.formula = Y + delta ~ 1,
+    Y.family = survival(method = "Cox"),
+    data = data,
+    strata = c(n = "00", c = "01", a = "11"),
+    prior_intercept = prior_normal(0, 1),
+    warmup = 1000, iter = 2000,+    cores = 6,
 chains = 6+  )R> summary(fit)
```

```
        mean          sd      2.5%       25%    median       75%     97.5%
n 0.2337785 0.005976189 0.2221630 0.2296578 0.2337909 0.2378282 0.2453549
c 0.6112725 0.007849057 0.5959362 0.6058670 0.6113284 0.6166199 0.6265693
a 0.1549490 0.005075883 0.1451325 0.1514720 0.1549051 0.1585060 0.1649443
```

```
R> outcome <- PSOutcome(fit)
R> plot(outcome) + xlab("time") + ylab("survival probability")
```

三个分层的估计概率分别为 0.234（置信区间：0.222 至 0.245）、0.611（置信区间：0.596 至 0.627）和 0.155（置信区间：0.145 至 0.165），接近真实值 0.25、0.60 和 0.15。

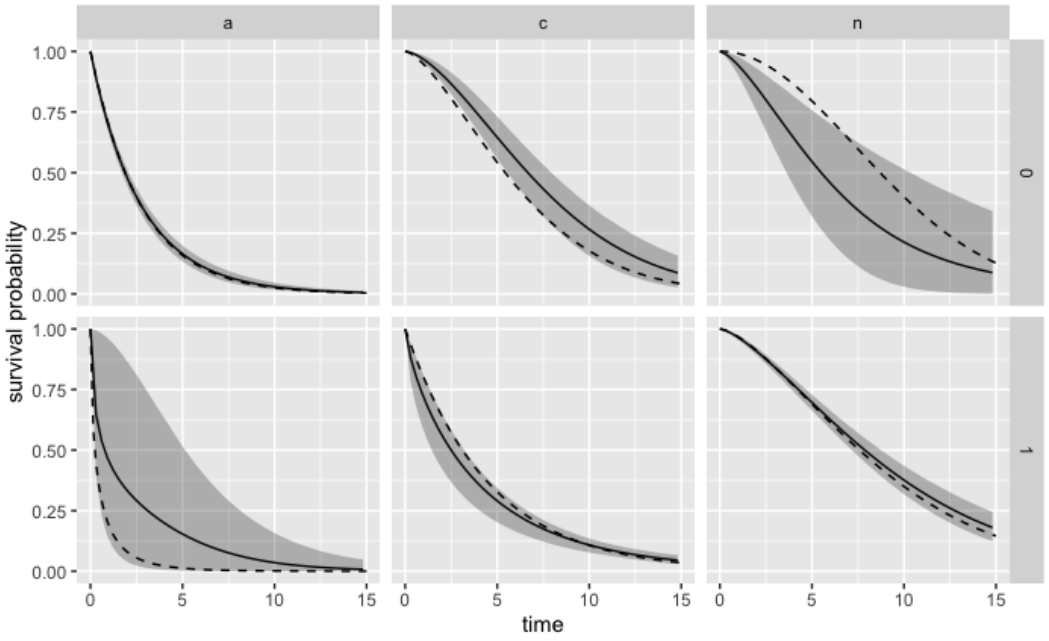


图 4：估计的生存概率曲线（实线）及其 95% 置信区域（阴影）。同时还绘制了真实的生存概率曲线（虚线）。

使用 `plot()`，我们可以直观地查看估计的生存概率曲线及其 95% 置信区域（图 4）。真实的生存概率曲线也已添加以供参考。

### 5.4. Case 4: Data with cluster effects

我们模拟一项包含 $N = 1{,}000$ 个单位的随机实验。我们以分别为 0.25、0.50 和 0.25 的概率，独立地进行 $S \in \{(0,0),(0,1),(1,1)\}$分配。治疗分配$Z_i$ 独立地从 $Bernoulli(0.5)$中抽取。为模拟簇的存在，我们将这些单位随机分配到 10个簇中，记为 $C_i = j$。对于每个簇 $j$，我们从标准正态分布中抽取随机效应 $\xi_j$。在给定分层 $S$ 和治疗分配 $Z$的条件下，结果 $Y$ 服从高斯分布，其中

$$
\begin{aligned}
Y \mid S = (0,0), Z = z, X_1, X_2, C = j &\sim \mathcal{N}(X_1 - X_2 + \xi_j, 0.3) \\
Y \mid S = (0,1), Z = z, X_1, X_2, C = j &\sim \mathcal{N}(2X_1 - (1+z)X_2 + 2 + 6z + \xi_j, 0.2) \\
Y \mid S = (1,1), Z = z, X_1, X_2, C = j &\sim \mathcal{N}(X_1 + X_2 - 1 + \xi_j, 0.2).
\end{aligned}
$$

我们使用 6 条链运行采样器，每条链进行 1000 次预热迭代和 1000 次采样迭代。分层概率和因果效应的真实值及其相应的后验分布见表 7。

```
R> fit <- PStrata(
+     S.formula = Z + D ~ 1,
+     Y.formula = Y ~ X1 + X2 + (1 | C),
+     Y.family = gaussian(),+     data = data,
```

```
+     strata = c(n = "00*", c = "01", a = "11*"),
+     prior_intercept = prior_normal(0, 1),
+     warmup = 1000, iter = 2000,+     cores = 6,
 chains = 6, refresh = 10+   )R> summary(fit)
```

```
        mean         sd      2.5%       25%    median       75%     97.5%
  n 0.2806836 0.01412599 0.2520836 0.2711908 0.2800077 0.2902168 0.3083786
  c 0.5202746 0.01656015 0.4871375 0.5084747 0.5208637 0.5312158 0.5520182
  a 0.1990418 0.01242839 0.1757518 0.1901582 0.1988140 0.2076523 0.2224745

R> outcome <- PSOutcome(fit)
R> contrast <- PSContrast(outcome, Z = TRUE)
R> summary(contrast, "matrix")

              mean        sd     2.5%      25%  median      75%    97.5%
n:{1}-{0} 0.000000 0.0000000 0.000000 0.000000 0.00000 0.000000 0.000000
c:{1}-{0} 5.977509 0.0178666 5.943497 5.965294 5.97707 5.989835 6.012452
a:{1}-{0} 0.000000 0.0000000 0.000000 0.000000 0.00000 0.000000 0.000000
```

| | True value | Posterior mean | 2.5% quantile | 97.5% quantile |
|---|---|---|---|---|
| $\Pr(S = (0,0))$ | 0.300 | 0.281 | 0.252 | 0.308 |
| $\Pr(S = (0,1))$ | 0.500 | 0.520 | 0.487 | 0.552 |
| $\Pr(S = (1,1))$ | 0.200 | 0.199 | 0.176 | 0.222 |
| $\mathrm{E}[Y(1) - Y(0) \mid S = (0,1)]$ | 6.000 | 5.978 | 5.943 | 6.012 |

表 7：模拟 4 中重要量的后验摘要及真实值。

## 6. Summary

主分层是处理后混杂情形下进行因果推断的重要工具。本文介绍PStrata 软件包，并展示其在最常见的单个和两个处理后变量不依从情形中的应用。PStrata 仍在持续开发中；未来版本将涵盖死亡截断和信息性缺失数据情形，以及图形诊断功能。

## Computational details

**PStrata** 0.0.1 基于 R 4.0.3 构建，并依赖于 **Rcpp** 1.0.6软件包、**dplyr** 1.0.7 软件包、**purrr** 0.3.4 软件包、**rstan** 2.21.1 软件包、**lme4** 1.1-27.1 软件包、**ggplot2** 3.3.6 软件包和 **patchwork**1.1.1 软件包。所有这些使用的软件包均可从综合 R 存档网络（CRAN）获取，网址为 https://CRAN.R-project.org/。

## Acknowledgement

本研究得到以患者为中心的结果研究所（PCORI）合同 ME-2019C1-16146 的支持。我们感谢 Laine Thomas 和 Peng Ding 提供的宝贵意见。

## References

Abrams K, Ashby D, Errington D (1996). “Weibull生存模型的贝叶斯方法——应用于癌症临床试验。” *Lifetime Data Analysis*, **2**(2), 159–174. Angrist J, Imbens G, Rubin D ( 1996). “使用工具变量识别因果效应。” *Journal of the American Statistical Association*, **91**(434), 444–455. Cox DR (1972). “回归模型与生命表。” *Journal of the Royal Statistical Society:Series B (Methodological)*, **34**(2), 187–202. Frangakis C, Rubin D (2002). “因果推断中的主分层。” *Biometrics*,**58**(1), 21–29. Gilbert P, Hudgens M (2008). “评估候选主替代终点。” *Biomet-rics*, **64**(4), 1146–1154. Hirano K, Imbens G, Rubin D, Zhou XH (2000). “在鼓励设计中评估流感疫苗的效果。” *Biostatistics*, **1**, 69–88. Imbens G, Rubin D (1997). “存在不依从的随机实验中因果效应的贝叶斯推断。” *The Annals of Statistics*, **25**(1), 305–327. Jiang Z, Ding P, Geng Z (2016). “通过多项试验识别主因果效应并评估替代终点。” *Journal of the Royal Statistical Society: Series B (StatisticalMethodology)*, **78**(4), 829–848. Jiang Z, Yang S, Ding P (2022). “主可忽略性下因果效应的多重稳健估计。” *JRSSB*。 Jo B, Stuart E (2009). “倾向得分在主因果效应估计中的使用。”*Statistics in Medicine*, **28**(23), 2857–2875. Mcdonald C, Hui S, Tierney W (1992). “计算机提醒流感疫苗接种对流感流行期间发病率的影响。” *M.D. computing : computers in medicalpractice*, **9**, 304–12. Rubin D (2006). “通过潜在结果和主分层进行因果推断：应用于因死亡删失的研究。” *Statistical Science*, **91**, 299–321. Wei LJ (1992). “加速失效时间模型：生存分析中Cox回归模型的有用替代。” *Statistics in Medicine*, **11**(14-15), 1871–1879. Zhang J, Rubin D, Mealli F (2009). “使用主分层对职业培训项目因果效应的基于似然分析。” *Journal of the American Statistical Asso-ciation*, **104**, 166–176.

**所属机构：**
Bo Liu, Fan Li 统计科学系 杜克大学 邮政信箱
90251 达勒姆，北卡罗来纳州 27705，美利坚
合众国 电子邮件： bo.liu1997@duke.edu,
fl35@duke.edu

《统计软件杂志》 http://www.jstatsoft.org/
由开放获取统计基金会出版 http://www.foastat.org/
MMMMMM YYYY，第 VV 卷，第 II 期 提交日期：yyyy-mm-dd
doi:10.18637/jss.v000.i00 已接受：yyyy-mm-dd